Conformational Analysis of Antibody Binding Site Using EDMA Methods

Xiaohong Chen

A thesis submitted to the Department of Chemistry
In partial fulfillment of the requirements for
the degree of Master of Science in Chemistry

Brock University
St. Catharines, Ontario
January, 2008

ABSTRACT

Euclidean distance matrix analysis (EDMA) methods are used to distinguish whether or not significant difference exists between conformational samples of antibody complementarity determining region (CDR) loops, isolated L1 loop and L1 in three-loop assembly (L1, L3 and H3) obtained from Monte Carlo simulation. After the significant difference is detected, the specific inter-C_α distance which contributes to the difference is identified using EDMA.

The estimated and improved mean forms of the conformational samples of isolated L1 loop and L1 loop in three-loop assembly, CDR loops of antibody binding site, are described using EDMA and distance geometry (DGEOM). To the best of our knowledge, it is the first time the EDMA methods are used to analyze conformational samples of molecules obtained from Monte Carlo simulations. Therefore, validations of the EDMA methods using both positive control and negative control tests for the conformational samples of isolated L1 loop and L1 in three-loop assembly must be done.

The EDMA-I bootstrap null hypothesis tests showed false positive results for the comparison of six samples of the isolated L1 loop and true positive results for comparison of conformational samples of isolated L1 loop and L1 in three-loop assembly. The bootstrap confidence interval tests revealed true negative results for comparisons of six samples of the isolated L1 loop, and false negative results for the conformational comparisons between isolated L1 loop and L1 in three-loop assembly. Different conformational sample sizes are further explored by combining the samples of isolated L1 loop to increase the sample size, or by clustering the sample using self-organizing map (SOM) to narrow the conformational distribution of the samples being compared

molecular conformations. However, there is no improvement made for both bootstrap null hypothesis and confidence interval tests. These results show that more work is required before EDMA methods can be used reliably as a method for comparison of samples obtained by Monte Carlo simulations.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor Heather Gordon for giving me this opportunity to realize my dream. It is only because of your support, understanding and unlimited patience that I am able to complete this project.

Great thanks go to Professor van der Est and Professor McLaughlin for your enthusiastic interest in my project, and for your time, encouragement and wonderful suggestions.

A great deal of thanks goes to my friends in Chemistry and biotechnology programs for your caring. Thank you Pat for your endless patience and help through the course of this project.

Finally, thanks go to my parents and my family for you always believing in me, and for your unconditional love. Thank you so much for your constant support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Antibodies

1.1.1 Introduction to Antibodies

Antibody molecules, also called immunoglobulin molecules, are produced by B lymphocytes in response to antigens.[1] Antibodies specifically bind to the epitope (region of the antigen that directly contacts the antibody) portions of the antigens and form antibody-antigen complexes. The subsequent immune response leads to the destruction of the antigens.

An antibody consists of two light peptide chains and two heavy peptide chains which form a "Y" shape (**Figure 1(a)**)[2] and are connected by disulphide bridges and non-covalent interactions. The light and heavy chains are defined by their molecular weights. The molecular weight for light chains is about 25 kDa and for heavy chains is 50-75 kDa.[3] Both light and heavy chains are composed of a variable region and a constant region. In the constant region, the amino acid sequence is conserved among the various immunoglobins. In the variable region, the amino acid sequence is different. Both heavy and light chains contain domains, which are independently folded, functional units. The variable regions are the N-terminal domains (V_H and V_L) whereas the constant regions are the C-terminal domains (C_H and C_L). Each heavy chain has four domains and the light chain has two domains. There are four disulfide bonds connecting light to heavy chains and connecting the two heavy chains. The fragment for antigen binding (Fab) includes V_H, V_L, C_{H1} and C_L domains. The very tips of the variable regions are called antibody binding sites, which are referred to as the complementarity determining regions (CDR) or hypervariable regions (**Figure 1(b)**).

An antibody always has two identical heavy chains and two identical light chains. Therefore an antibody has two identical CDR binding sites. Each antibody can bind to two

antigens. The CDR directly contacts the epitope portion of the antigen's surface. The entire CDR is comprised of six peptide loops: three loops in the heavy chain (H1, H2, H3) and three loops in the light chain (L1, L2, L3).[4]

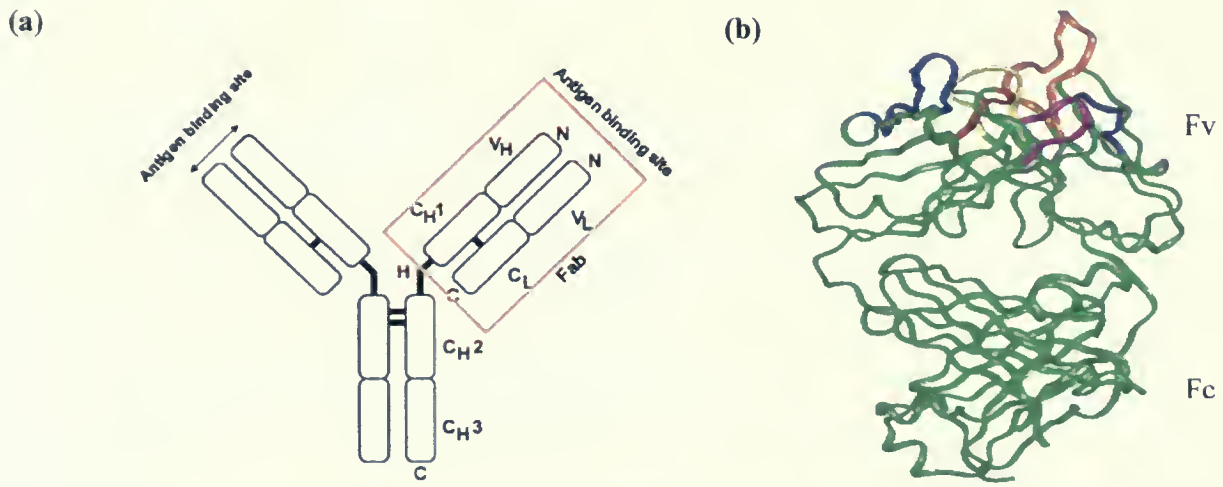

Figure 1. (a) General structure of an antibody,[2] the portion in the frame is called Fab; (b) Enlarged image of a Fab generated with Insight II,[5] Complementarity Determining Regions (CDRs) form the antigen binding site: orange loop is L1, blue is L2, yellow is L3, red is H1, purple is H2 and pink is H3.

The Kabat definition[6] of each CDR loop is adopted in this paper. The Kabat definition of a CDR describes where the loop begins and ends in an amino acid sequence of the light and heavy chains based on the amino acid sequence variability.

1.1.2 Secondary structure in a Fab

An amino acid includes a chiral carbon, called the alpha carbon or C_α, a carboxylic acid, amino groups and an R group, which is the side chain. A general structure of an alpha amino acid is shown in **Figure 2**. Amino acids in proteins are connected by peptide bonds. The secondary structure of a peptide consists of α-helices and β-sheets formed through hydrogen bonding. The prevalent secondary structures in a Fab are β-sheets.[7]

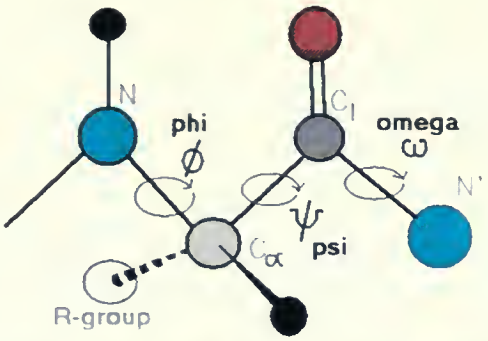

Figure 2.[8] Basic amino acid structure within a peptide. There are three torsion angles along the backbone chain, ψ φ and ω. Backbone atoms include the carbonyl carbon (grey), C1; oxygen (red), O; the alpha carbon (grey), Cα; and the amide group nitrogen (blue), N.

The β-strands are extended with the backbone dihedral angles φ and ψ approximately ±180°. Polypeptide chains can form parallel β-strands or antiparallel β-strands (**Figure 3**).[9] In an antiparallel β sheet, the directions of the N-terminus to the C-terminus along two adjacent chains alternate. Every amino acid forms two hydrogen bonds with an amino acid in another chain. In a parallel β sheet, the directions of the N-terminus to the C-terminus are the same in these two chains.

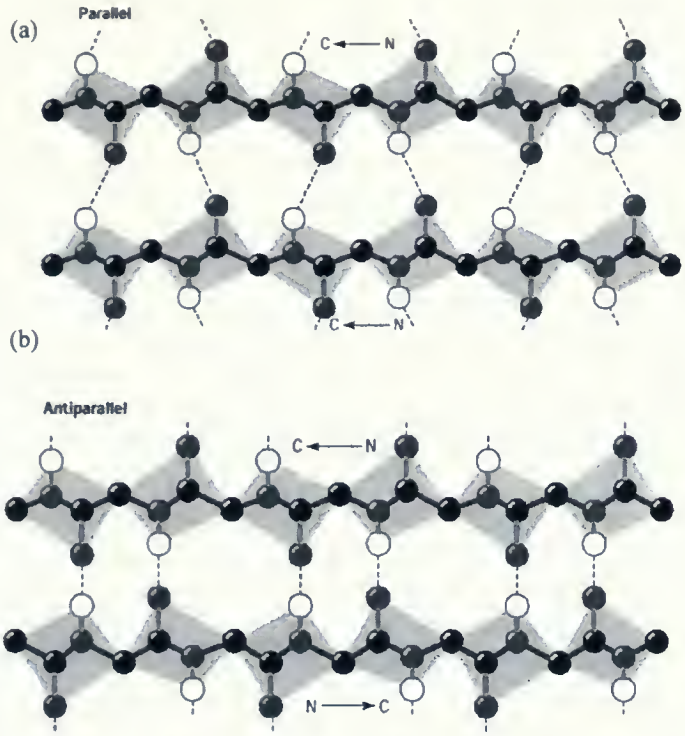

Figure 3.[9] General structures of β-sheets. (a) Parallel β-strands. (b) Antipallel β-strands.

1.1.3 Mechanisms of antibody-antigen interaction

The general goal of this research project is to investigate the conformational diversity of antibody binding sites. A protein that adopts more than one conformation has "conformational diversity". Millions of conformations of the six peptide loops of a model CDR may be produced using Monte Carlo and molecular dynamics simulation methods. This project is focused on the analysis of the conformational diversity of antibody binding sites and the conformational changes observed when the antibody is bound to an antigen relative to its unbound state. Recent experimental results suggest that some antibody binding sites have pre-existing equilibrium isomers, which form the conformational diversity of the antibody.[10] Therefore, the antibody can adopt different binding site conformations and maybe bind unrelated antigens. This model gives a possible explanation for the cross-reactivity of pathogen-specific antibody that leads to autoimmune diseases. [11]

There are three mechanisms that describe how a protein, such as an antibody might bind its ligand: lock and key, induced fit and pre-existing equilibrium.[12]

The lock and key mechanism is described in **Figure 4(a).** There is no conformational change involved when the antigen binds to an antibody and forms a complex. That an antibody binds only to a specific antigen is well explained by this mechanism. Experimental evidence of lock and key binding was supplied by the Poljak research group in 1986 using X-ray crystallography.[13] In this experiment, the three-dimensional structure of a complex of an antigen (lysozyme) and the antibody (myeloma immunoglobulin) Fab was compared with the unbound antibody. The antibody-antigen interface was tightly packed. No conformational changes were observed, which support the lock and key mechanism.

The induced-fit model is one explanation for when the conformation of an antibody changes when an antigen binds to it. The induced fit mechanism involves antigen-induced conformational changes, which include conformational change of complementarity determining region loops or side chain conformational changes.[14-16] The induced-fit mechanism is illustrated in **Figure 4(b).** The experimental evidence was provided by Wilson et al.[14] using X-ray crystallography. The three-dimensional structures of CDR H3 and L1 loops are changed when an antibody, Fab 17/9, binds to an antigen, peptide immunogen (Tyr^{P100}-Leu^{P108}) from influenza virus hemagglutinin.

Recently, James et al.[10-12] have provided experimental structural evidence for the pre-existing equilibrium mechanism. The pre-existing equilibrium model (**Figure 4(c)**) assumes there is more than one pre-existing conformational isomer. The specific conformation was selected from the ensemble of pre-existing conformations when it binds to an antigen. The James and Tawfik group found two isomeric conformations on the antigen binding site for the same antibody (SPE7), which they call Ab1 and Ab2. These two conformations exist and interconvert quickly in solution. Ab1 binds to the recombinant protein, antigen TrxShear3.

Because Ab2 has a different conformation, it does not bind to antigen TrxShear3 but binds to hapten, 2, 4-dinitrophenyl (DNP). The pre-existing equilibrium binding mechanism may explain why multiple structures of an antibody could enable it to bind to more than one partner.

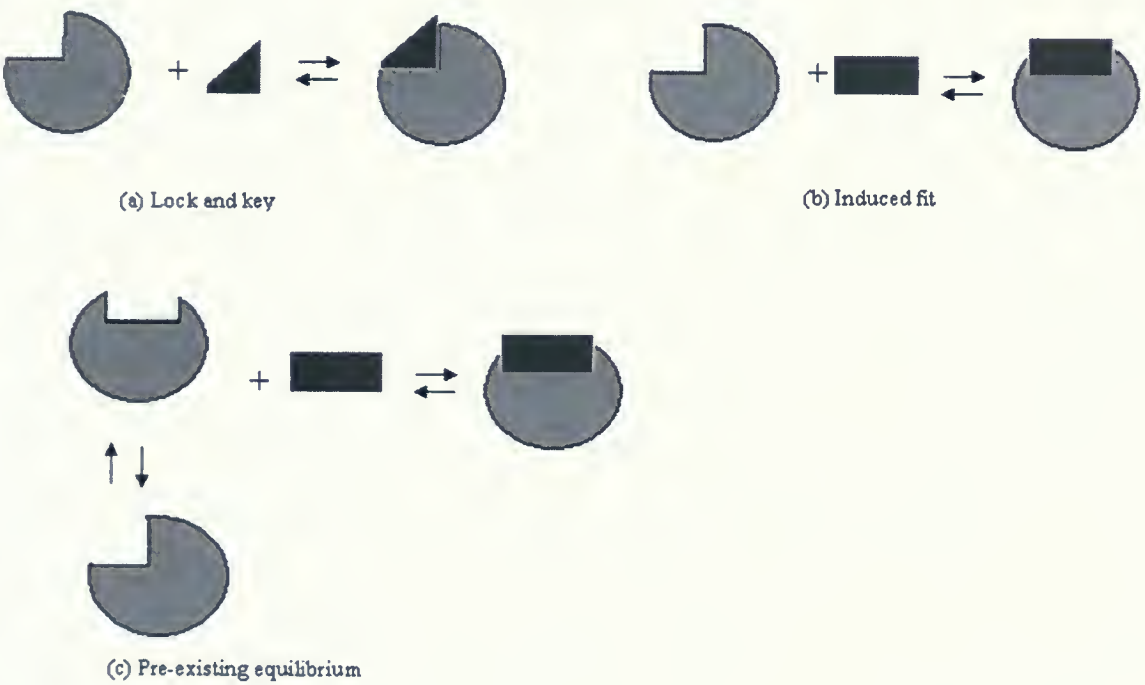

Figure 4. Antibody-antigen interaction models: (a) Lock and key. (b) Induced fit. (c) Pre-equilibrium. Protein is in grey; ligand is in dark grey.

1.1.4 Models Investigated

In this project, the conformational changes between a model CDR L1 loop and this L1 loop in a three loop assembly of the antibody 8F5 binding site[17, 18] are explored using Euclidean distance matrix analysis (EDMA) and distance geometry (DGEOM) methods.[19, 20] If the EDMA method can identify conformational changes of CDR loops in the antibody binding

site, this method will eventually be used to provide computational evidence for the pre-existing equilibrium or induced fit mechanisms.

An X-ray crystallographic conformation of the 8F5 Fab is found in the RCSB Protein DataBank.[21] This conformation is used as a starting structure for Monte Carlo simulations. Here, the conformational diversity of a portion of the antibody 8F5 binding site, as obtained from Monte Carlo simulations, is analyzed using Euclidean distance analysis (EDMA) methods. Antibody 8F5 is a neutralizing antibody to human rhinovirus serotype 2, which is a main causative agent of the common cold. The Kabat amino acid composition of the six CDR loops of 8F5 is shown in **Table 1**. Molecular dynamics simulation studies of the shape of antibody binding sites have shown that the conformations of unbound antibody and the antibody-antigen complex of 8F5 have significant fluctuations.[22] Conformational changes of the backbone of each of the CDR loops are observed. The especially significant motions were undergone by L1 and H3. The smallest range of movement was displayed by L3.

Table 1. Kabat amino acid composition of antibody 8F5's six CDR loops[17]

CDR Loop	Amino Acid Composition	Kabat Loop Length
L1	Lys24-Ser-Ser-Gln-Ser-Leu-Leu-Asn-Ser-Arg-Thr-Gln-Lys-Asn-Tyr-Leu-Thr40	17
L2	Trp56-Ala-Ser-Thr-Arg-Glu-Ser62	7
L3	Gln95-Asn-Asn-Tyr-Asn-Tyr-Pro-Leu-Thr103	9
H1	Asp31-Ile-Tyr-Ile-His35	5
H2	Arg50-Leu-Asp-Pro-Ala-Asn-Gly-Tyr-Thr-Lys-Tyr-Asp-Pro-Lys-Phe-Gln65	16
H3	Tyr99-Tyr-Ser-Tyr-Tyr-Asp-Met-Asp-Tyr107	9

It is very difficult to simulate all conformations for the entire set of six loops of the CDR using the Monte Carlo method. Therefore, the starting point was to investigate one loop (**Figure 5(a)**): the isolated L1 loop, and subsequently, L1 in a three-loop assembly, which includes L1, L3 and H3 (**Figure 5(b)**). The X, Y and Z coordinate data sets, which represent the positions of the alpha carbons (C_α) of the isolated L1 loop and L1 in the three-loop assembly, were obtained from Monte Carlo simulations.[23] During these Monte Carlo simulations, CDR loops are simplified to one or two atoms based on the weight and size of the amino acids. In this paper, the C_α is the only atom to be selected to represent the backbone structure or the conformations of the CDR loop. The L1 loop of 8F5 has 17 C_α based on the Kabat definition[6] (**Table 1**).

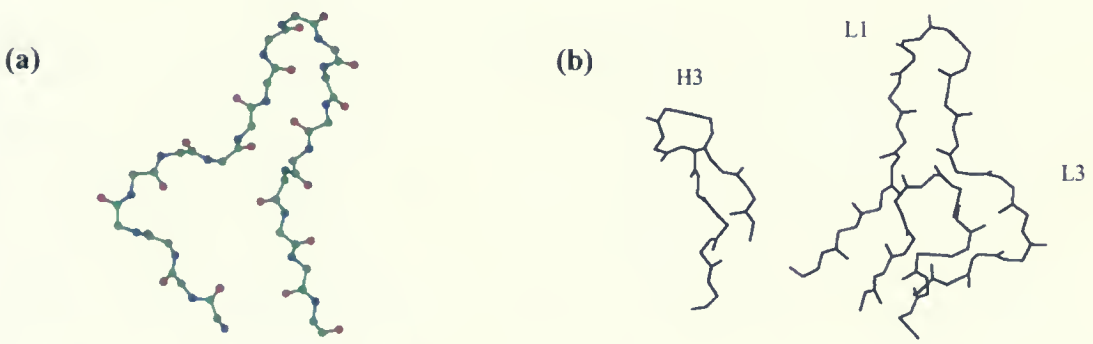

Figure 5. (a) The isolated light chain L1 loop backbone structure of antibody 8F5 (PDB: 1BBD), which includes C (green), O (red) of carbonyl, Cα (green) and N (blue) atoms. There are 19 amino acid residues in L1 loop including two feet. (b) L1 in three-loop assembly. These figures were generated by Insight II.

Two extra C_α are added to the L1 loop as two "feet" and positioned at Z=0 (**Figure 6**). There are 19 C_α atoms including two "feet", which are $C_{\alpha 23}$ and $C_{\alpha 41}$ whose coordinates are taken from the amino acids immediate preceding and following the L1 loop. The distance between $C_{\alpha 23}$ and $C_{\alpha 41}$ is 9.41 Å. These two "feet" were fixed during the Monte Carlo simulations and constrain the L1 in a loop. These two "feet" belong to the framework of the variable region, which support the CDR loops in place. The mobile portion of the loop may not move below Z=0 during the Monte Carlo simulation.

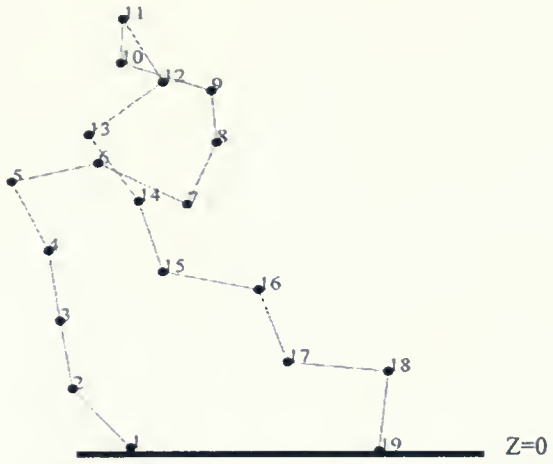

Figure 6. Alpha carbon structure of L1 loop for 8F5 generated by Insight II.

Because it is very difficult to effectively sample all conformations of the entire six loops of the binding site using the Monte Carlo method, a research strategy is developed. The starting point in this work is for one loop, which is the isolated L1 loop and further, L1 in a three-loop assembly, which involves L1, L3 and H3. Each loop of the three-loop assembly has two "feet" which are positioned at Z=0. So, all these three loops are able to move above the XY plane. The goal of the comparison of conformational samples is to distinguish if differences exist between isolated L1 loop and L1 in three-loop assembly. We expect there are significant differences between isolated L1 and L1 in the presence of other CDR loops because of the non-covalent interactions between CDR loops. The Monte Carlo simulations were performed by other members of the Gordon research group.

1.2 *Monte Carlo simulation*

1.2.1 Metropolis Monte Carlo

Computational simulation may be used to provide detailed information of the conformational diversity of biological molecules.[24, 25] Monte Carlo (MC) simulation is a major method in the computer simulation field.

Monte Carlo simulation is a stochastic technique. MC simulations generate a pseudorandom sequence of molecular conformations to represent the Boltzmann distribution[26] of the system under investigation.[27] In 1953, Metropolis and his co-workers described the process of a Monte Carlo method on calculating the properties of a protein, which was composed of interacting molecules.[28]

Consider a molecule or a collection of molecules consisting of N atoms at a temperature T and contained in a volume V. The Metropolis MC simulation is carried out as follows.

The initial molecular conformation is chosen for the molecule, for example, the X, Y and Z coordinates (X_0, Y_0, Z_0) of an atom with potential energy E_0. The potential energy is discussed later. A new conformation is produced by MC simulations by changing the position of each atom. The atom (X_0, Y_0, Z_0) is changed to new positions X_1,Y_1, Z_1. The relationship between these Cartesian coordinates is as follows:

$$\begin{aligned} X_1 &= X_0 + r\xi_1 \\ Y_1 &= Y_0 + r\xi_2 \\ Z_1 &= Z_0 + r\xi_3 \end{aligned} \qquad (1)$$

where r is the maximum allowed displacement. The ξ_1, ξ_2 and ξ_3 are random numbers, each of which lies between -1 and 1. The new energy E_1 corresponding to the new molecular conformation is calculated. The change of the energy of the system is ΔE:

$$\Delta E = E_1 - E_0 \qquad (2)$$

Whether the move is rejected or accepted is based on the Boltzmann test. If the move does not change or decrease the energy, $\Delta E \leq 0$, then this move is accepted. However, if the move increases the energy, there are two situations: if $\xi_4 \leq \exp(-\Delta E/kT)$, where ξ_4 is another random number, $0 \leq \xi_4 \leq 1$, k is Boltzmann constant and T is temperature, the move will be allowed. Otherwise, if the $\xi_4 > \exp(-\Delta E/kT)$, the move will be rejected and the atom will stay at (X_0, Y_0, Z_0). The Monte Carlo process is described in **Figure 7.**

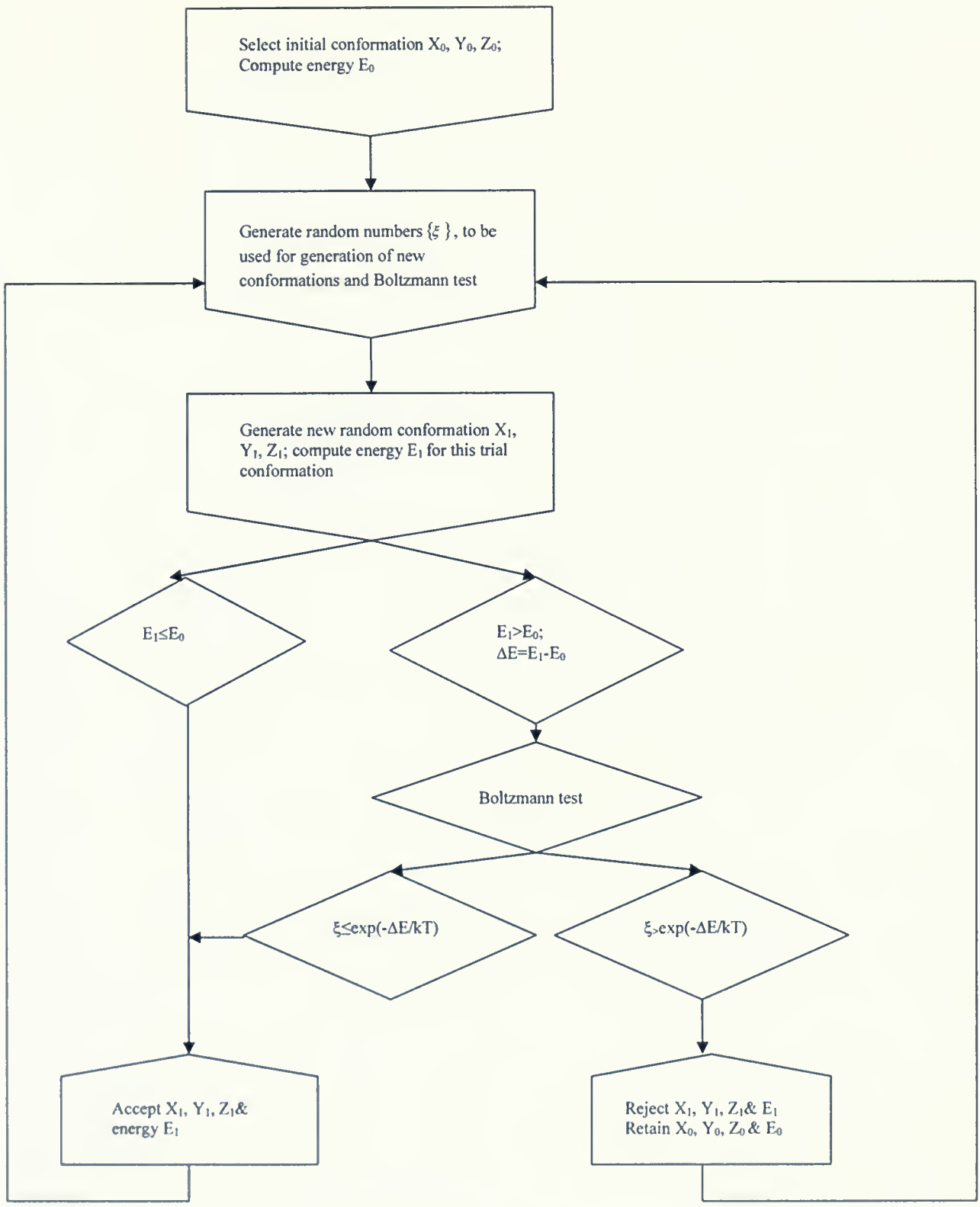

Figure 7. Flow chart for Metropolis Monte Carlo sampling process

The Boltzmann distribution is one of the classic distributions which predicts the distribution energy function for a number of particles. [26] The probability of finding an atom in state a can be expressed as:

$$p(a) = \frac{g_a e^{-Ea/kT}}{\sum_a g_a e^{-Ea/kT}} \qquad (3)$$

where g_a is the degeneracy, E_a is the energy at state a, k is Boltzmann's constant, T is temperature in Kelvin.

In order to calculate the potential energy for the model peptide loops examined here (**Figure 5**), the Wallqvist and Ullner potential energy function is used:[29]

$$E_{total} = E_{bs} + E_{ba} + E_{st} + E_{bb} + E_{hd} \qquad (4)$$

where E_{bs} is the bond stretching potential energy, E_{ba} is the bond angle-bending potential, E_{st} is a potential that orients the amino acid side chain relative to backbone direction, E_{bb} is peptide backbone potential and E_{hd} is the hydrophobic or hydrophilic interaction existing between sidechains.

The conformational samples used in this project were obtained from independent Monte Carlo simulations. Independent simulations mean either that simulations start from a different initial conformation or a different sequence of the random numbers ξ is generated. Therefore, conformational samples obtained from independent MC simulations contain different individual coordinates but are still drawn from the same Boltzmann distribution. In order to represent the total conformational distribution of the L1 loop, as many different conformations as possible should be collected. The conformations were saved after every 2000 attempted moves of all the sites within the L1

loop or three-loop assembly in order to ensure that subsequent saved conformations were independent, that is, not correlated to each other.

1.3 Euclidean distance matrix analysis

1.3.1 Euclidean distance matrix

Euclidean distance matrix analysis (EDMA) is a method for comparing the shapes or structures of three-dimensional objects.[19] The shape of an object is described using inter-landmark distances. EDMA can identify whether a significant difference exists between two samples and detect where the difference is located. The EDMA method was introduced to the conformational study of biological structures in 1990's by Lele and Richtsmeier.[30] The craniofacial appearances of normal mice and of mice with a genetic mutation Ts65Dn that is a model for Down syndrome were compared and the specific locations which contributed to the difference were detected using the EDMA method.[31, 32] The models of craniofacial morphology in orofacial clefting were studied using EDMA by Ayoub et al.[33] and McIntyre.[34] When this method was applied to the comparison of molecular structures of the human insulin protein, significant differences between human insulin wild type and mutant type, which cause diabetes, were detected successfully.[19] Two insulin conformations were compared, one from wild type and another from a mutant that substitutes serine for phenylalanine in the 24th position on the B chain. The conformational difference was detected by comparing statistics T for inter-C_α distances. It is important to note that this was the only reported use of EDMA on examining protein structure that we have found. However, in this case only two protein conformations were compared and the EDMA methods we will use, EDMA-I bootstrap null hypothesis and confidence interval tests were not used. We want

to see if these two EDMA methods are appropriate for using on the conformational samples obtained from MC simulations.

In this work, Euclidean distance matrix analysis, EDMA, is explored as a statistical tool for detecting whether or not two conformational samples are significantly different, and if so, the inter-atomic distances responsible for that difference. The presence of nuisance parameters, such as translation and rotation, do not affect EDMA results. Optimal superimposition is also widely adopted for detecting bimolecular conformational changes. For example, superimposition describes a molecular conformation based on either minimizing the sum of squared distances or the sum of distances.[35] The most obvious difference from EDMA is that superposition can only be performed between two molecules at a time and two collections of molecular conformations cannot be superimposed. On the other hand, EDMA can handle samples with large quantities of conformations at a time. For doing the comparison between conformational samples, EDMA is much more efficient than superimposition.[36, 37]

The shape of a protein can be represented by a Euclidean distance matrix [38, 39]or form matrix, which is the collection of all of the inter-C_α distances in the protein. This project starts with many conformational samples, which were obtained from Monte Carlo simulations.

The Euclidean distance D_{jm} between $C_{\alpha j}$ and $C_{\alpha m}$ of amino acid residues j and m is calculated as follows:[19]

$$D_{jm} = \sqrt{(X_j - X_m)^2 + (Y_j - Y_m)^2 + (Z_j - Z_m)^2} \qquad (5)$$

where (X_j, Y_j, Z_j) and (X_m, Y_m, Z_m) are the Cartesian coordinates of $C_{\alpha j}$ and $C_{\alpha m}$, respectively.

The Euclidean distance matrix for a single molecular conformation i can be expressed:

$$DM(i) = (D_{jm}(i))_{j=1,2,.....k,m=1,2.....k} \qquad (6)$$

where j and m represent the amino acid sequence number and K is the total number of amino acid residues. The distance matrix for conformation i is symmetric with all diagonal elements equal to zero. Therefore, only the numbers in the upper triangle are used to represent the distance matrix. There are a total of k(k-1)/2 entries in this non-redundant representation of the distance matrix.

In order to describe the n conformations of a conformational sample A, the mean conformation FM (A) for sample A is estimated as follows. The squared Euclidean distance between $C_{\alpha j}$ and $C_{\alpha m}$ of conformation i is:

$$e_{jm,i} = D_{jm}^2(i) \qquad (7)$$

The average squared Euclidean distance $\overline{e_{jm}}$ for the n conformations of sample A is estimated as follows:

$$\overline{e_{jm}} = n^{-1}\sum_{i=1}^{n} e_{jm,i} \qquad (8)$$

The variance σ_{jm} of the squared distance between $C_{\alpha j}$ and $C_{\alpha m}$ in sample A is calculated:

$$\sigma_{jm} = n^{-1}\sum_{i=1}^{n} (e_{jm,i} - \overline{e_{jm}})^2 \qquad (9)$$

Finally, the estimate of the mean form matrix for sample A is calculated:

$$FM(A) = (\delta_{jm}{}^{0.5})_{J=1,2.....K,m=1,2.....K} \qquad (10)$$

where δ_{jm} is given by

$$\delta_{jm} = (\overline{e_{jm}}^{\,2} - \frac{3}{2}\sigma_{jm})^{0.5} \qquad (11)$$

The mean form is computed from δ_{jm}, instead of $\overline{e_{jm}}$ in equation (8) because $(X_j - X_m)^2$, $(Y_j - Y_m)^2$ and $(Z_j - Z_m)^2$ for the n conformations are each distributed as non-central chi-squared (χ^2) random variables.[26] Likewise the sum of the squared distance e_{jm}, equation (7), has a non-central χ^2 distribution with three degrees of freedom (**Figure 8**). The cumulative probability for such a non-central χ^2 distribution is shown in **Figure 9**.

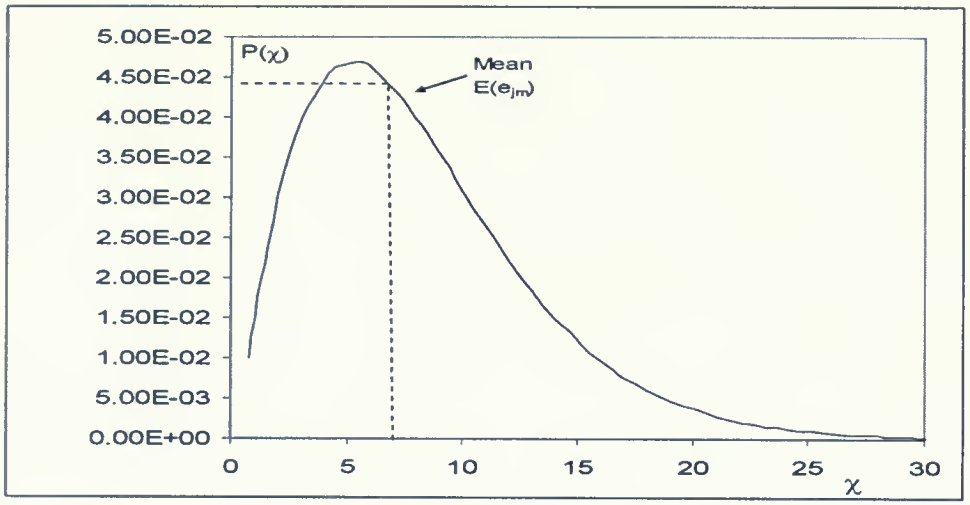

Figure 8. The non-central χ^2 distribution with 3 degrees of freedom; Pr is probability density.

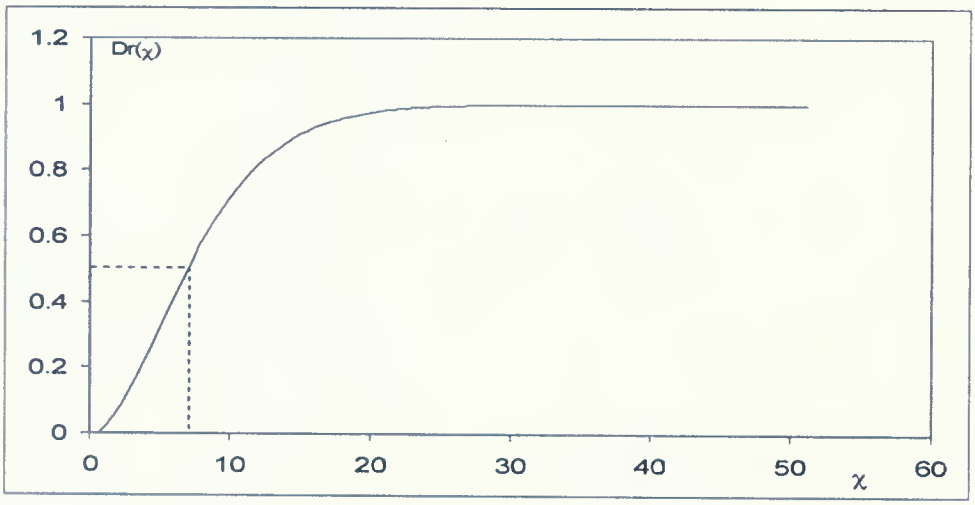

Figure 9. The cumulative probability Dr for non-central χ^2 distribution with 3 degrees of freedom.

The first moment of non-central χ^2 distribution is the mean of the distribution:

$$E(e_{jm}) = 3\varphi_{jm} + \delta_{jm} \qquad (12)$$

where φ_{jm} is the scaling parameter, δ_{jm} is noncentrality parameter.[31]

The second moment is the variance:

$$\sigma(e_{jm}) = 6\varphi_{jm}{}^2 + 4\delta_{jm}\varphi_{jm} \qquad (13)$$

The mean form σ_{jm} can be obtained from equations (8) and (9):

Substitute $\varphi_{jm} = \frac{E(e_{jm}) - \delta_{jm}}{3}$ from equation (12) into (13), then we will get:

$$\sigma(e_{jm}) = \frac{2E(e_{jm})^2 - 2\delta_{jm}^2}{3} \qquad (14)$$

let $E(e_{jm}) \approx \overline{e_{jm}}$. Equation (14) can be rearranged to equation (11) and the estimate of the square of the mean form matrix δ_{jm} is obtained.

Euclidean distance matrix analysis (EDMA) is applied to detect if there is a significant difference between two conformational samples A and B and to identify the Euclidean distances responsible for that difference. There are two tests: the EDMA-I bootstrap null hypothesis test and the confidence interval test.[19, 36] The EDMA-I bootstrap null hypothesis test is used to detect whether or not significant conformational difference exists between two samples and the confidence interval test is used to identify the origin or individual inter-C_α distances contributing to that conformational difference.

1.3.2 Bootstrap null hypothesis test using EDMA-I method

Null hypothesis testing is a standard requirement for statistical analyses. The EDMA-I bootstrap null hypothesis test is used to detect if a significant difference exists between two conformational samples.

Three components are required in this hypothesis test: a null hypothesis, an observed test statistic value T_{obs}, and the distribution of the test statistic T for the null hypothesis. If the observed value of the test statistic T_{obs} lies in the extreme tails of the null distribution, the null hypothesis is rejected. Otherwise, it will be retained. To generate values of the test statistic T

for estimating the null distribution of T, the bootstrap method is applied, which is a random sampling statistical tool.[40]

The null hypothesis of the EDMA-I test assumes that the mean shapes of the compared samples are identical. For example: let $A_1, A_2, A_3, ..., A_n$ and $B_1, B_2, B_3, ..., B_m$ represent the molecular conformations of a protein having k amino acids in sample one and sample two, respectively.[19] As discussed in section 1.3.1, the estimated mean form matrix for the inter-C_α distances of sample one FM(1), with n conformations and k amino acids is as follows:

$$FM(1) = \begin{vmatrix} D_{11} & D_{12} & \cdots & D_{1k} \\ D_{21} & D_{22} & \cdots & D_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ D_{k1} & D_{k2} & \cdots & D_{kk} \end{vmatrix} \qquad (15)$$

where D_{jm} are given by equation (5) and $D_{11} = D_{22} = ... = D_{kk} = 0$.

Based on the mean form matrices FM(1) and FM(2) for the two conformational samples, the form difference matrix for samples one and two, $FDM(1,2)$ is calculated, equation (16), which is the ratio of form matrix one to form matrix two.

$$FDM(1,2) = \frac{FM(1)}{FM(2)} \qquad (16)$$

Let R_{jm} represent the elements of $FDM(1, 2)$. Then,

$$FDM(1,2) = \begin{vmatrix} R_{11} & R_{12} & \cdots & R_{1k} \\ R_{21} & R_{22} & \cdots & R_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R_{k1} & R_{k2} & \cdots & R_{kk} \end{vmatrix} \qquad (17)$$

The entries along the diagonal must be $R_{11} = R_{22} = ... = R_{kk} = 1$. $FDM(1,2)$ is symmetric about the diagonal as are FM(1) and FM(2). The value of all values of R_{jm} would be 1.0 if

FM(1) ≡ FM(2). The EDMA-I test begins by converting the elements of the upper triangle of the *FDM* (1,2) into a vector. The vector is sorted in an ascending order:

$$\left|R_{12}, R_{13}, \cdots, R_{k-1,k}\right| \xrightarrow{\text{sort}} \left|R_{\min}, \cdots, R_{\max}\right|$$

The observed statistic T_{obs} is the ratio of the maximum to the minimum value in the *FDM* (1,2):

$$T_{obs} = \frac{\max\left(FDM(1,2)\right)}{\min\left(FDM(1,2)\right)} = \frac{R_{\max}}{R_{\min}} \qquad (18)$$

The value of T_{obs} should be a real number larger or equal (when FM(1) ≡ FM(2)) to 1.0. Because in practice the conformational samples one and two have a finite size, $T_{obs} \neq 1.0$, even if they are drawn from the same underlying distribution. A bootstrap method is an empirical way to generate the probability distribution of T for two samples of size n and m, drawn from the same underlying population. Comparison of T_{obs} (equation (18)) to this bootstrap distribution of T for the null hypothesis (i.e. that samples one and two are drawn from the same underlying population) allows us to determine whether or not there is significant difference between samples one and two. The null distribution is the probability distribution of the test statistic when the compared samples are identical.[41] The null distribution of the statistic T is generated by using a bootstrap approach. First, n conformations are randomly selected from sample one to form sample 1'. Then, m conformations are randomly selected, also from sample one, to form sample 2' (**Figure 10**). Note that sampling replacement is used, so that individual conformations may be selected more than once.

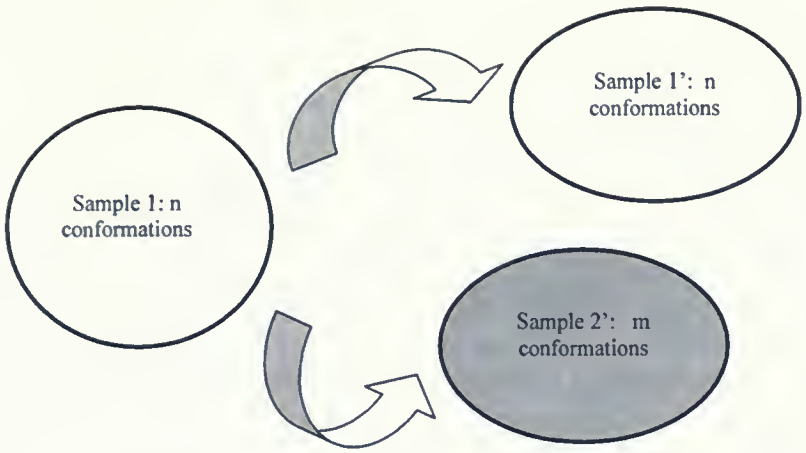

Figure 10. Take random samples from sample one and form samples 1' and 2'.

Therefore, samples 1' and 2' are drawn from the same underlying population of molecular conformations. The form difference matrix $FDM(1', 2')$ based on samples 1' and 2' is calculated as in equation (16).

The T value for samples 1' and 2' is calculated:

$$T = \frac{\max\ (FDM(1',2'))}{\min\ (FDM(1',2'))} \qquad (19)$$

The above process of generating a value of T for the comparison of two random samples of sizes n and m drawn from the same underlying population is repeated for W=200 to 1000 times. The W values of T form the null distribution of the bootstrap statistic T.

The last step is to plot the W values of the bootstrap statistic T by using a histogram. If T_{obs}, computed using equation (18) from the original samples one and two, falls in the upper tail of the null distribution, the null hypothesis that the mean forms of sample one and two are identical, will be rejected. Here is an example. The EDMA null hypothesis test of Ts65Dn mouse model from Lele's book[19] was reproduced. Down's syndrome is a human genetic disease, caused by the third copy of chromosome21. This disease is accompanied a change in craniofacial appearance. The Ts65Dn mouse has an extra copy of segmental

chromosome16 and has comparable phenotypes of human Down's syndrome, such as shape changes in the mandible jaw bone and delayed maturation. Therefore, the Ts65Dn mouse model is widely applied to the research of human Down's syndrome disease.[42, 43] Here, the mandibles from skulls of normal mice and the Ts65Dn mice are compared. In this model, there are 13 normal mouse mandibles and seven Ts65Dn mouse mandibles. The number of points or landmarks (LM) in the mouse mandible is 11 (**Figure 11**). Therefore, there are 55 inter-landmark distances ($C_{11}^2 = 11 \times 10 \div 2 = 55$).

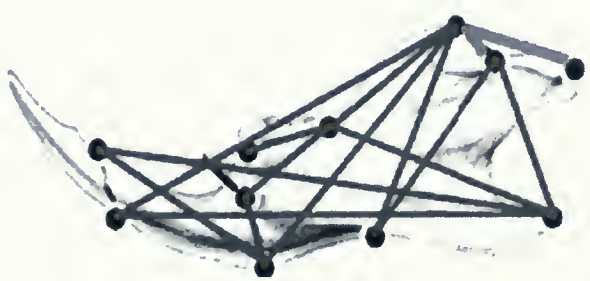

Figure 11.[19] **Mouse hemi-mandible with 11 landmarks.**

The mean form difference matrix for the mean form of mandibles from13 normal mice divided by the mean form of seven Ts65Dn mice is shown in **Table 2**. The observed T value is the ratio of maximum to minimum entries, $T_{obs} = \frac{1.258}{0.787} = 1.598$. Two hundred bootstrap values of T were then calculated from the sample of normal mice. The distribution of the bootstrap T statistics is plotted in a histogram using Excel in **Figure 12**. The results computed here are in agreement with those presented in ref.19.

Table 2. The mean form difference matrix for normal mouse and Ts65Dn mouse mandibles

	LM 2	LM 3	LM 4	LM 5	LM 6	LM 7	LM 8	LM 9	LM 10	LM11
LM1	1.091	0.787	0.884	1.148	1.141	1.133	1.150	1.170	1.196	1.258
LM 2		1.146	1.155	1.108	1.044	1.051	1.067	1.070	1.068	1.085
LM 3			1.019	1.031	1.017	1.024	1.036	1.025	1.018	1.011
LM 4				1.033	1.015	1.024	1.035	1.026	1.019	1.014
LM 5					0.991	1.013	1.032	1.024	0.999	1.007
LM 6						1.055	1.068	1.105	1.052	1.053
LM 7							1.055	1.022	1.041	1.041
LM 8								1.041	1.077	1.060
LM 9									1.007	1.051
LM 10										1.028

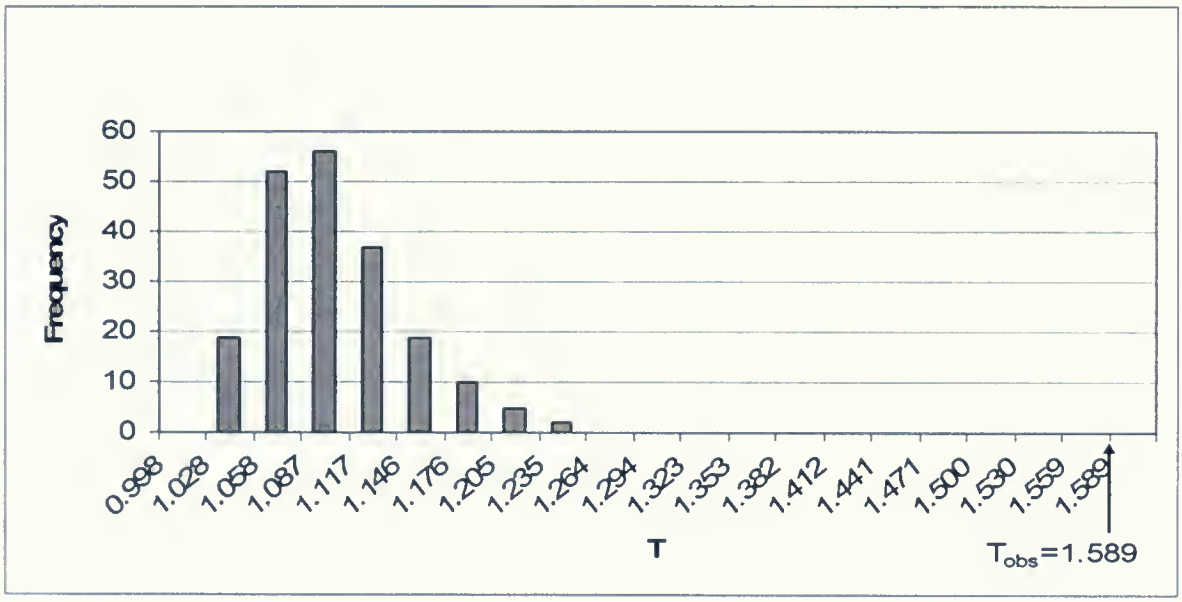

Figure 12. Distribution of bootstrap T of normal mouse mandibles.

In **Figure 12**, the range of the T distribution is between 1.028 and 1.264. That is, these values of T were computed from the comparison of two samples drawn from the same population, that being the set of 13 normal mouse mandibles. The deviation in T from 1.0 is due to the finite size of mouse mandibles being sampled. The observed T_{obs} =1.598, calculated by comparison of Ts65Dn mouse mandibles to those of normal mice, falls outside the

distribution range and lies beyond the upper tail of the bootstrap null distribution. The result of this null hypothesis test reveals that there is a significant conformational difference between the mandibles of normal mice and Ts65Dn mice.

This null hypothesis test is used to answer the question of whether or not the mean form of sample one is significantly different from the mean form of sample two. Because this test is a one way test, the baseline group has to be chosen.[19] The baseline group is the one for which the null distribution is computed. The baseline sample is the numerator in equation (16). Generally, the sample with a larger sample size is chosen to be the baseline group.

In order to know whether two compared samples are different or not for certain, the one way test should be run twice, once using each sample as a baseline.

1.3.3 Bootstrap confidence interval test

A confidence interval refers to an interval estimate that has a certain probability of including the true value of the parameter.[44] Because the test presents information about confidence or reliability in an estimate, the interval test is a powerful statistical tool. The objective of this bootstrap confidence interval test is to identify the specific distances responsible for the significant difference between two conformational samples.

The bootstrap confidence interval approach uses a bootstrap random sampling method to generate samples directly from two sample datasets. The following example briefly describes how the bootstrap confidence interval approach works. Let $A_1, A_2, A_3, \ldots, A_n$ and $B_1, B_2, B_3, \ldots, B_m$ represent the molecular conformations in samples one and two respectively. Each conformation is of a protein having k amino acids.

First, the form difference matrix *FMD*(1,2) is calculated using the estimated mean forms for the two conformational samples: FM(1) and FM(2), equation (16). Second, n conformations are randomly selected from sample one, $A_1^*, A_2^*, A_3^*, \ldots, A_n^*$ to form sample 1*and m conformations are randomly selected from sample two, $B_1^*, B_2^*, B_3^*, \ldots, B_m^*$ to form sample 2*(**Figure 13**). Note that sampling replacement is used, so that any conformation may be selected more than once.

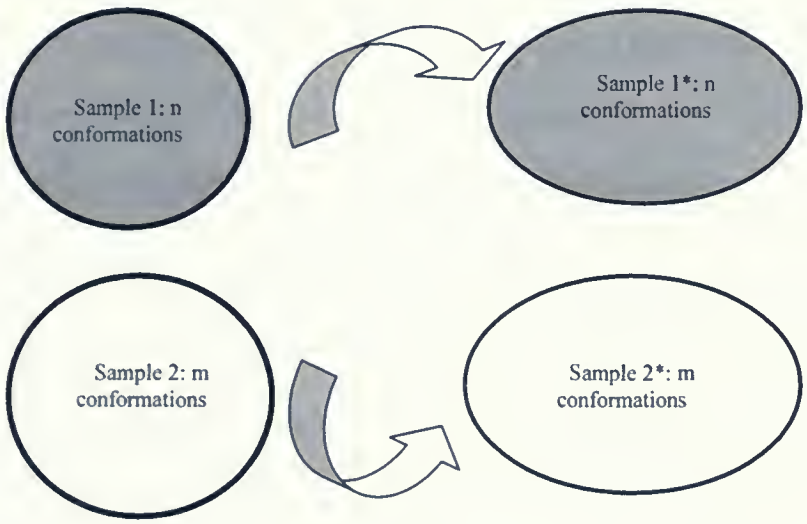

Figure 13. Take random samples from samples one and two to form samples 1* and 2*.

Third, the form difference matrix *FDM* (1*,2*) is calculated based on the mean forms FM(1*) and FM(2*) of random samples 1* and 2*. The above two steps are repeated for 200 to 1000 times (W). *FDM* (1*,2*) are collected and put into vector format. All *FDM* (1*,2*) will form a matrix with ((k-1) k)/2 rows and W columns in **Figure 14**.

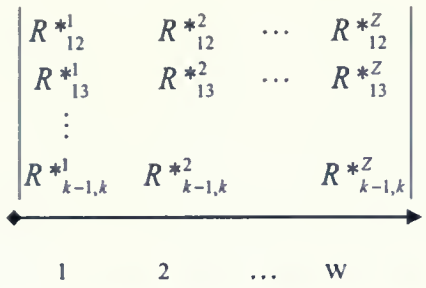

Figure 14. Super matrix for W values of *FDM* (1*,2*).

Each column of the super matrix in **Figure 14** is a form difference matrix that is expressed as a vector. Each row contains W values of R_{jm}, the mean distance ratio between $C_{\alpha j}$ and $C_{\alpha m}$ in bootstrap samples obtained from the original samples one and two. The confidence interval is constructed by sorting the ratios in each row in increasing order (**Figure 15**). The maximum and the minimum values of R^*_{jm} are the upper and lower confidence limits for R^*_{jm}.

$$\begin{vmatrix} (R^*_{12})_{min} & \cdots & (R^*_{12})_{max} \\ (R^*_{13})_{min} & \cdots & (R^*_{13})_{max} \\ \vdots & & \\ (R^*_{k-1,k})_{min} & \cdots & (R^*_{k-1,k})_{max} \end{vmatrix}$$

Figure 15. Each row of the super matrix is sorted in increasing order.

The range of the confidence intervals reflects the similarity of two compared samples. If the confidence interval includes the value 1.0 between the lower and upper limits, then the specific distance is identical in sample one and two. Otherwise, the specific distance in two samples is different. To illustrate this, **Figure 16** is a frequency plot of the values of R^*_{jm} for three idealized rows of the sorted matrix in **Figure 15** in bootstrap samples of *FDM*(1,2). In **Figure 16**, when R^*_{jm}=1.0 is included in the range of the confidence interval for the distance

between C_{aj} and C_{am}, the mean distance between C_{aj} and C_{am} for sample one, $D_{jm}(1)$ is not different to the distance for sample two, $D_{jm}(2)$. If R^*_{jm} is always less than 1.0 over the range of the confidence interval, the mean distance between C_{aj} and C_{am}, for sample one is less than sample two, $D_{jm}(1) < D_{jm}(2)$. Otherwise, if $R^*_{jm} > 1.0$ for all W, the mean distance between C_{aj} and C_{am} for sample one is larger than sample two, $D_{jm}(1) > D_{jm}(2)$.

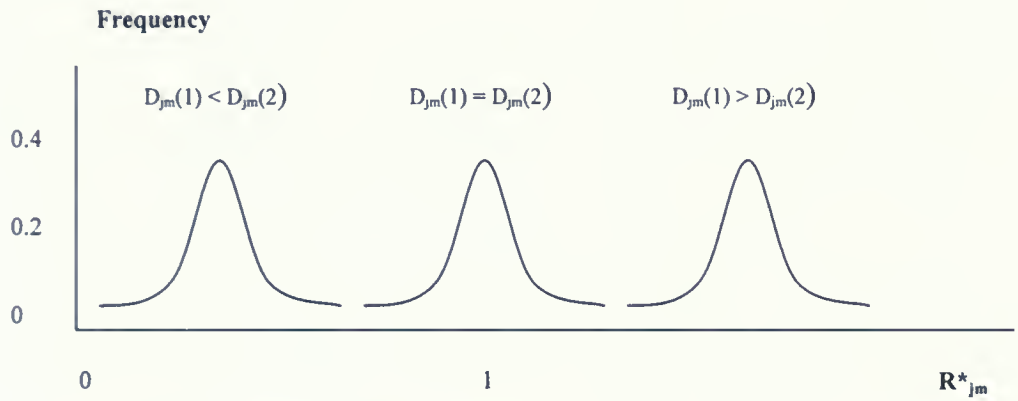

Figure 16. Distribution of R^*_{jm} for samples one and two.

In section 1.3.2, an example that compares the mandibles of normal mice to those of the Ts65Dn mouse model from Lele's book[19] was reproduced using the EDMA null hypothesis test. Here, the confidence interval test for the normal mice and the Ts65Dn mice is reproduced. There are 55 inter-landmark distances measured within a mouse mandible. A subset of the confidence interval results are displayed in **Table 3.** The specific inter-landmark distance is marked by *, if the value 1.0 is included between its lower and upper limits. There are only 13 of the 55 inter-landmark distances, which include the value 1.0 in the range of lower and upper limits. The mean inter-landmark distances for those 13 are not significantly different. The other 42 inter-landmark distances of the mouse mandible show significant differences between those for normal mice and Ts65Dn mice.

Table 3 Subset of 55 confidence intervals for normal mouse compared to Ts65Dn mouse mandibles

No.	Inter-LM distance	Lower limit	Upper limit	* if lower limit <1.0< upper limit
1	LM1 and LM2	1.066	1.113	
2	LM1 and LM3	0.781	0.817	
3	LM1 and LM4	0.872	0.895	
4	LM1 and LM5	1.120	1.170	
5	LM1 and LM6	1.118	1.166	
6	LM1 and LM7	1.110	1.158	
7	LM1 and LM8	1.123	1.185	
8	LM1 and LM9	1.144	1.204	
9	LM1 and LM10	1.178	1.224	
10	LM1 and LM11	1.211	1.313	
11	LM2 and LM3	1.125	1.183	
12	LM2 and LM4	1.113	1.216	
13	LM2 and LM5	1.061	1.158	
14	LM2 and LM6	1.030	1.060	
15	LM2 and LM7	1.036	1.066	
16	LM2 and LM8	1.049	1.087	
17	LM2 and LM9	1.054	1.086	
18	LM2 and LM10	1.058	1.084	
19	LM2 and LM11	1.063	1.104	
20	LM3 and LM4	0.979	1.041	*
21	LM3 and LM5	1.000	1.064	*
22	LM3 and LM6	1.006	1.031	
23	LM3 and LM7	1.014	1.039	
24	LM3 and LM8	1.021	1.055	
25	LM3 and LM9	1.016	1.040	
26	LM3 and LM10	1.013	1.029	
27	LM3 and LM11	0.999	1.031	*
28	LM4 and LM5	1.004	1.070	
29	LM4 and LM6	1.003	1.030	
30	LM4 and LM7	1.009	1.044	
31	LM4 and LM8	1.018	1.057	
32	LM4 and LM9	1.013	1.041	
33	LM4 and LM10	1.010	1.032	
34	LM4 and LM11	0.998	1.033	*
35	LM5 and LM6	0.899	1.084	*
36	LM5 and LM7	0.957	1.072	*
37	LM5 and LM8	0.976	1.086	*
38	LM5 and LM9	0.934	1.100	*
39	LM5 and LM10	0.937	1.070	*
40	LM5 and LM11	0.943	1.073	*
41	LM6 and LM7	1.025	1.095	
42	LM6 and LM8	1.038	1.103	
43	LM6 and LM9	1.075	1.149	
44	LM6 and LM10	1.034	1.085	
45	LM6 and LM11	1.036	1.086	
46	LM7 and LM8	1.001	1.093	
47	LM7 and LM9	0.975	1.050	*
48	LM7 and LM10	1.004	1.072	

1.4 *Distance geometry*

1.4.1 Background of distance geometry

Distance geometry is used to convert a set of inter-atomic distances or distance bounds to X, Y and Z coordinates of the atoms.[45, 46] The goal of distance geometry is to obtain a molecular conformation or set of available conformations consistent with the set of upper and lower bounds of the inter-atomic distance matrix. All previous applications of EDMA first obtained an "improved" mean form matrix from FM (A), equation (10), whose elements correspond to a real three-dimensional shape or conformation.[46] The improved mean form matrix can be obtained by inputting the estimated mean form matrix into a distance geometry program.

Generally, there are three situations that distance geometry deals with when converting a set of inter-atomic distances into Cartesian coordinates. The first one is when the inter-atomic distances are exact, in other words there is no measurement error. Therefore, the original distance matrix does correspond to a real three-dimensional object. The second one is when only a sparse set of inter-atomic distance is given. Not all of the inter-atomic distances are known. The third one is when distance geometry deals with a distance range or lower and upper bounds. Distance geometry is best known as a powerful method to determine the solution conformations of molecules in NMR experiments.[47, 48] For example, to determine a protein structure in solution using NMR, the distance geometry algorithm is applied to generate a molecular conformation that satisfies the distance between two protons within a molecule as given by Nuclear Overhauser Effect (NOE) data. NOE between nuclear spins is used to describe the correlation of two protons and indicates that two protons are separated, on average, by a distance of less than 5.0 Å. Instead of an exact distance, NOEs give a range of the inter-atomic distances, or the maximum and minimum inter-atomic distances.[49-51] Distance geometry searches for X, Y and Z coordinates of all atoms of molecule that are

consistent with the upper and lower bounds of the available inter-atomic distances. Note that only a sparse set of these distance bounds are available.

A molecular conformation can be expressed by a set of inter-atomic distances, or Euclidean distance matrix. For example, the Euclidean distance matrix for butane in **Table 4** summarizes the inter-atomic distances for two different idealized conformations, *cis* and *trans*, as in **Figure 17**. With the expectation of the correct configuration about any chiral centres existing in the molecule, there is no loss of conformational information because the three-dimensional shape of the molecule can be re-generated from a complete distance matrix.[48]

Table 4 Euclidean distance matrix for butane (Å):

atoms	1	2	3	4
1	0.0	1.5	2.5	**3.8**
2	1.5	0.0	1.5	2.5
3	2.5	1.5	0.0	1.5
4	**2.6**	2.5	1.5	0.0

In **Table 4**, all diagonal entries are 0.0Å. The upper triangle represents the upper bounds or largest allowed inter-atomic distances, i.e. for the *trans* conformation, and the lower triangle represents the lower bounds or the lowest limit of allowed inter-atomic distances, i.e. for the *cis* conformation. Here, the upper and lower bounds for atoms 1 and 4 are different; the remaining upper and lower bounds are identical for both geometric conformations, assuming rigid bond lengths of 1.5Å and rigid bond angles of 109.5°.

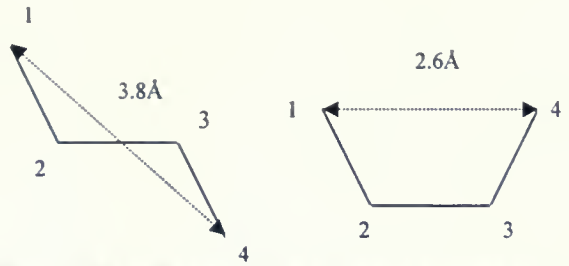

Figure 17. The *trans* and *cis* conformations of butane.

The metric matrix approach is one of the major algorithms used for distance geometry calculation. This method was introduced to the analysis of molecular conformation by Crippen and Havel,[52-55] where the X, Y and Z atomic coordinates of the molecular conformation are directly generated from inter-atomic distances or distance bounds. The metric matrix approach algorithms for distance geometry are simply described as follows.[45]

A metric matrix is a matrix that can deduce the three-dimensional coordinates, which are consistent with the inter-atomic distances. The most important point to distance geometry is that the metric matrix can be generated from the distance matrix D. K is the total number of atoms. The element d_{jm} in the distance matrix D represents the distance between atoms j and m. To convert the distance matrix D to the metric matrix G, first, d_{jo} is calculated, which is the distance between point j and the centre of mass o:

$$d^2_{jo} = \frac{1}{K}\sum_{m=1}^{K} d^2_{jm} - \frac{1}{K^2}\sum_{m=2}^{K}\sum_{n=1}^{m-1} d^2_{mn} \qquad (20)$$

The elements g_{jm} of metric matrix G are calculated using:

$$g_{jm} = (d^2_{jo} + d^2_{mo} - d^2_{jm})/2 \qquad (21)$$

Once the metric matrix G is obtained, the eigenvalues and eigenvector of the matrix can be calculated:

$$G = VL^2V^T \qquad (22)$$

where L^2 are the eigenvalues and the V are the eigenvectors of G. The metric matrix G can be converted into the three-dimensional coordinate matrix C and be expressed as:

$$G = CC^T \qquad (23)$$

where C is a matrix containing the coordinates of each atom. From equations (22) and (23), the X, Y and Z coordinate matrix can be generated based on the eigenvalues and eigenvectors of the metric matrix G:

$$C=VL \qquad (24)$$

Therefore, the X, Y and Z coordinates of the atoms are expressed in matrix C, which is converted from the initial inter-atomic distance matrix D.

In this work, the distance geometry program DGEOM[20] was used to obtain Cartesian coordinates for the L1 loop C_α's from the inter-atomic distances in the estimated mean form matrices, D, obtained from Monte Carlo simulations.

1.4.2 DGEOM program

The distance geometry program DGEOM, developed by Blaney and co-workers,[20] is used to convert the mean form matrix of the inter-C_α distances for the L1 loop into corresponding X, Y and Z coordinates. The estimated mean form matrix of conformational sample A, FM (A), equation (10), does not always correspond to a real shape in a three-dimensional Euclidean space either because of measurement error or because the K(K-1)/2 inter-atomic distances are mean values that collectively cannot be achieved in a real molecule. That is, while each member of the conformational sample is a real three-dimensional shape, the collection of means of the inter-atomic distances may not be.

Distance geometry is not always able to produce a structure based on the mean form matrix because of a geometrical impossibility, called "triangle inequality". [45, 56] The triangle inequality states that for points A, B and C forming a triangle, the length of any side of the triangle is not larger than the sum length of the other two sides and is not less than the difference between the other two sides. **Figure 18** shows that, for example, $BC \leq AC+AB$, $BC \geq |AC-AB|$.

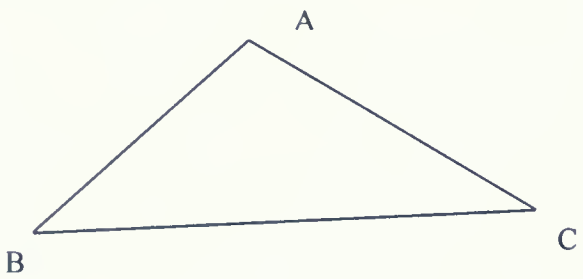

Figure 18. Triangle inequality relationship: $BC \leq AC+AB$, $BC \geq |AC-AB|$.

If a violation of the triangle inequality relationship is detected within the distance matrix, DGEOM will not produce corresponding three-dimensional structures; instead it will give a warning message. An example of a triangle inequality problem that shows inconsistent bounds and geometrical impossibility is in **Table 5**. Three atoms $C_{\alpha 6}$, $C_{\alpha 2}$ and $C_{\alpha 10}$ cannot form a triangle because $D_{2,6} + D_{6,10} < D_{2,10}$. This example was obtained when I ran the DGEOM program to get a real structure for a conformational sub-sample of the L1 loop, to be discussed in section 2.6.

Table 5 Triangle inequality problem of three inter-C_α distances

Inter-C_α distance	Inter-C_α distances (Å)
$C_{\alpha 6}$- $C_{\alpha 2}$: $D_{2,6}$	5.424
$C_{\alpha 10}$- $C_{\alpha 2}$: $D_{2,10}$	10.330
$C_{\alpha 6}$- $C_{\alpha 10}$: $D_{6,10}$	4.890

The process that distance geometry uses to avoid the triangle inequality problem is called bounds smoothing. If the bounds smoothing works, DGEOM will generate X, Y and Z coordinates corresponding to the input distance matrix. Otherwise, DGEOM may not produce a real structure.

Here, the purpose of using DGEOM is to find a real three-dimensional conformation closest to that represented by the mean form of the CDR L1 loop obtained from conformational samples, which were generated by MC simulations. To achieve this purpose, there are some requirements that need to be satisfied before DGEOM starts to work, which include an input file and a constraint file. The input file, which includes X, Y and Z coordinates of an arbitrary conformation, supplies the bond lengths and bond angles. The estimated mean form matrix of a conformational sample obtained from Monte Carlo simulation is also supplied to DGEOM via the constraint file. A constraint file involves giving lower and upper bounds for each inter-atomic distance. There is flexibility in how to choose the constraints. If the constraint file has only an upper bound for some inter-atomic distance, the lower bound will be assigned as the sum of the van der Waals radii or the hydrogen bonding distance. Different conditions for our inter-atomic constraints are discussed in section 2.2.2.

DGEOM produces the corresponding structures based on the requirements of the constraint and input files. The distance error function is used to determine which generated

conformations will be rejected and which will be accepted. The distance error is the sum of the differences between the inter-atomic distances of the conformers produced by DGEOM and those of the lower and upper bounds in the constraint file. The distance error can be calculated as follows:

$$\text{Distance error} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left| D_{ij} - lowerbound(upperbound)_{ij} \right|^2 \quad (25)$$

where k is the total number of atoms. The distance error is the sum of the squared minimum differences between the inter-C_α distance of $C_{\alpha i}$ and $C_{\alpha j}$, generated by DGEOM, and the corresponding lower or upper bound in the constraint file. The distance error function is minimized by DGEOM. Any structure with a distance error larger than some assignable default value is rejected. **Figure 19** illustrates the process of DGEOM.

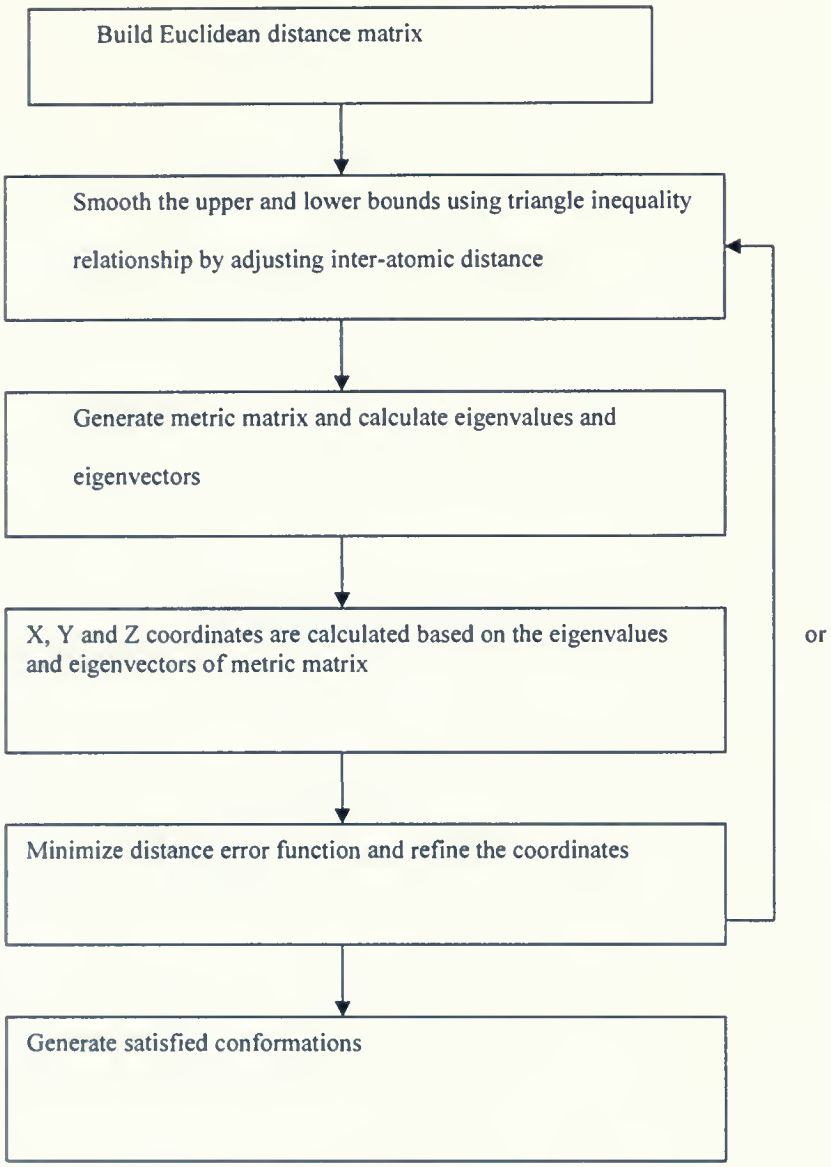

Figure 19. Flowchart of DGEOM process.

1.5 Goals of research

The overall objective of this research is to use EDMA to distinguish the differences between conformational samples obtained from Monte Carlo simulations of isolated CDR loop L1 and L1 in the presence of other CDR loops, L3 and H3 (**Figure 5**).

The first goal of our research is to describe a mean shape for a conformational sample of a part of an antibody binding site, the L1 loop described in section 1.1.4, obtained by Monte Carlo simulations. The second goal is to compare mean shapes obtained from two different samples and determine whether or not they differ. Third is to describe why the conformational distributions differ and to identify the contributing inter-C_α distances by using Euclidean distance matrix analysis (EDMA) methods. Since this is the first application of EDMA to the analysis of molecular conformational samples obtained by Monte Carlo simulations, both positive and negative control tests are designed to assess the specificity and selectivity of this method.

2 Experimental methodologies and results

2.1 Experimental models and conformational samples

In this section, the experimental samples are described. The samples are eight molecular conformational samples representing parts of the antibody 8F5 binding site. The antigen-binding site or CDR of 8F5 shows conformational diversity in previous experimental studies and molecular dynamic simulations.[22]

The conformational samples are of the isolated L1 loop (**Figure 5a**) and L1 in the three-loop assembly (**Figure 5b**) obtained from Monte Carlo simulations. The conformational diversity and differences between the various samples are investigated using EDMA methods. The programs for EDMA calculation were developed using the FORTRAN 90 language[57] and the UNIX operating system. The visualization and animation of molecular conformations are made using the Insight II 2005 molecular modelling software.[5]

The conformational samples of the isolated L1 loop (samples one, two, four, five, six, and seven (**Table 6**)) were obtained from six independent Monte Carlo simulations. Each sample consists of 8000 conformations, which are selected from six independent twenty million step-long Monte Carlo simulations. **Figure 20** shows the cumulative mean energies of each simulation has converged at twenty million steps, which means that the cumulative mean energy will no longer change as the number of steps of MC simulation increases. This evidence supports that the Monte Carlo simulations have run long enough to sample all of conformation space for the isolated L1 loop. The energy distributions of 8000 conformations in the six samples of isolated L1 loop are plotted in **Figure 21**. There is no significant difference among the energy distributions of these six samples. In order to focus on the conformations with higher energy, the scale of energy axis is decreased (**Figure 22**). For each of the samples, there are five or fewer conformations in the two highest energy ranges, which

should not have a large affect on the structural averages taken over the entire 8000 conformations. Therefore, these six samples of isolated L1 loop were expected to have no significant differences between them.

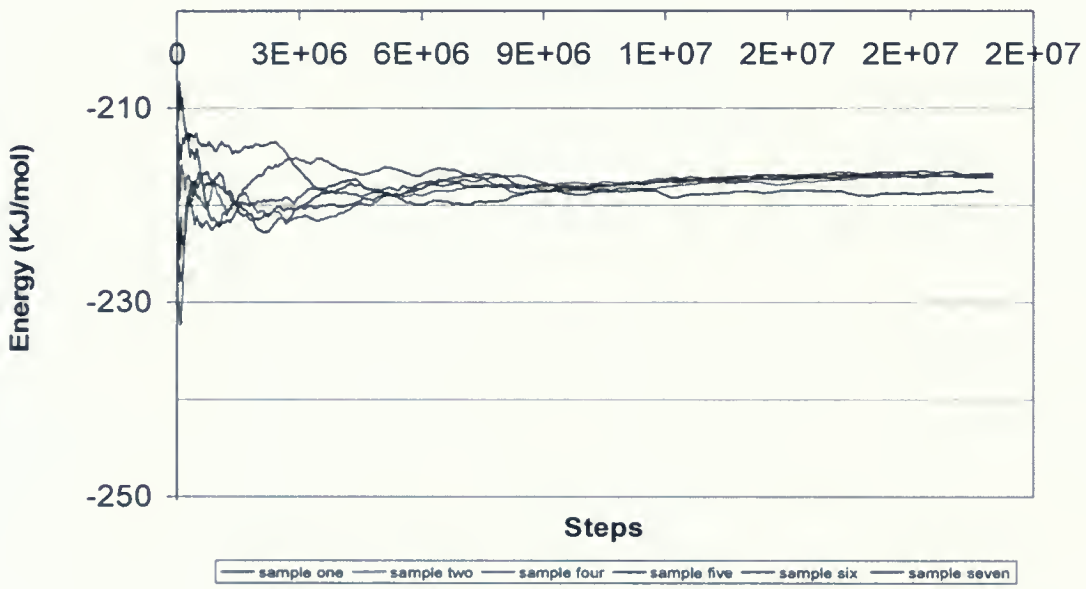

Figure 20. Cumulated mean energies vs. steps of Monte Carlo simulation.

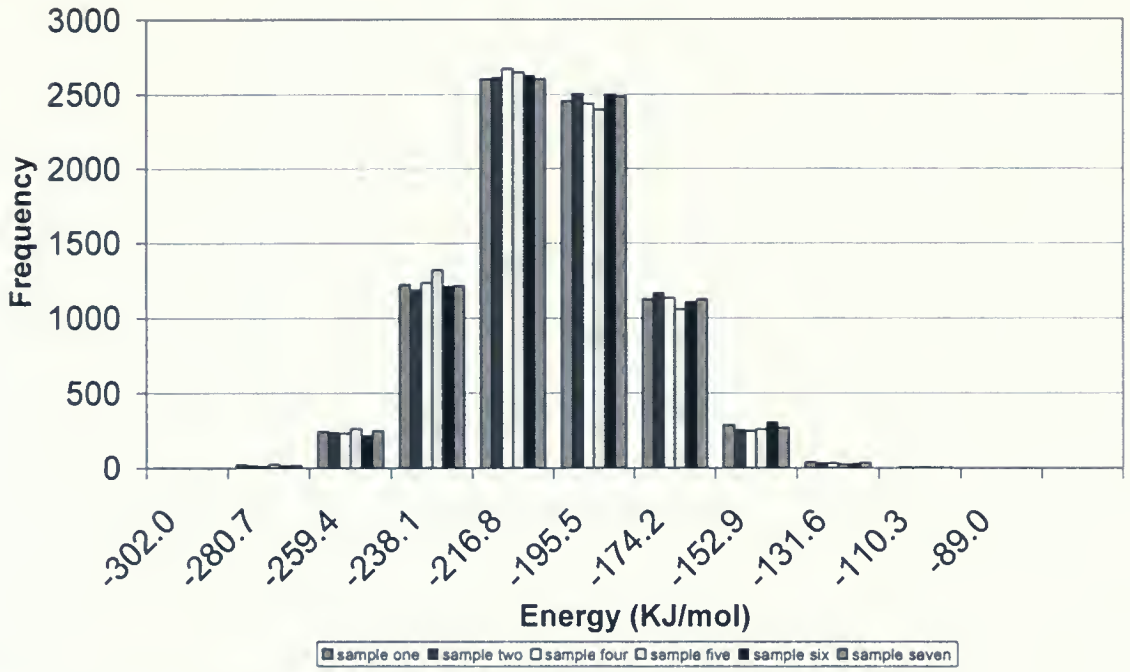

Figure 21. Energy distributions of six samples of isolated L1 loop.

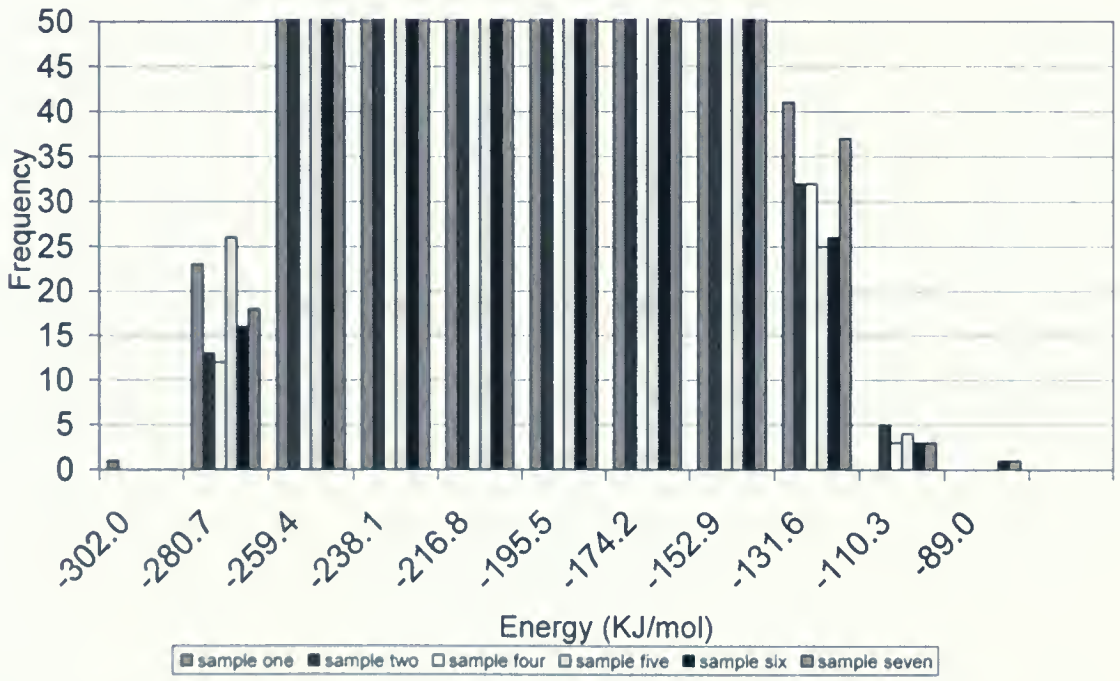

Figure 22. Part of enlarged image for energy distributions of six samples of isolated L1 loop.

Two other conformational samples, samples nine and ten (**Table 6**), are from two independent Monte Carlo simulations of L1 in a three-loop assembly, which consists of two loops L1 and L3 from the light chain and the H3 loop from the heavy chain. These two conformational samples each contain 4000 conformations. The two conformational samples should not have any significant differences because they were obtained from same system under the same simulation conditions. Although the conformations in the samples are not identical, their distributions should be the same. Between the single L1 loop conformations and the conformations of the L1 loop in the three loop assembly, significant differences should be detected because the non-covalent interactions, such as van der Waals forces between loops, will affect the conformation of the L1 loop.

Table 6 Conformational samples

Sample Number	Loop	Conformational Sample Sizes
1, 2, 4, 5, 6, 7	Isolated L1 Loop	8000
9, 10	L1 in three loop assembly	4000

2.2 Statistical analyses of conformational samples

The shape of a protein can be represented by a Euclidean distance matrix or form matrix, which is the collection of all of the inter-C_α distances in the protein. In this research, FORTRAN programs were designed to compare the conformational samples, described in **Table 6**. For each sample, the Euclidean distance between all inter-C_α distances are calculated based on equation (5) in section 1.3.1. In total, there are 171 inter-C_α distances

(C_{19}^2=171) in the L1 loop. In order to compare any two samples, the mean form for each sample must first be calculated.

2.2.1 Estimating the mean form for conformational samples

The estimator of the mean form matrix FM (1) for sample one is calculated using equation (10). For example, the mean form matrices of conformational samples one and two (**Table 6**), each of which has 8000 conformations, are shown in **Tables 7** and **8**. Because the mean form matrices are square symmetric, we use only the elements of the upper triangle to describe them. Note that the values of the diagonal elements in the matrix are zero and therefore are not shown. The off-diagonal represents the adjacent inter-C_α distances, which are always close to 3.8 Å in a protein, regardless of the values of the φ, ψ and ω torsional angles of the backbone.

Table 7 Mean form matrix of conformational sample one of isolated L1*

	Ca2	Ca3	Ca4	Ca5	Ca6	Ca7	Ca8	Ca9	Ca10	Ca11	Ca12	Ca13	Ca14	Ca15	Ca16	Ca17	Ca18	Ca19
Ca1	3.808	5.959	7.382	8.810	9.373	9.917	10.728	11.556	11.799	12.529	12.757	14.142	14.896	14.142	12.625	10.614	9.374	9.410
Ca2		3.801	5.874	7.122	7.070	7.380	8.406	9.486	9.853	10.667	10.917	12.268	12.923	12.153	10.783	9.197	8.783	9.857
Ca3			3.796	5.932	6.160	5.326	5.971	7.511	7.885	8.538	8.526	9.855	10.514	9.734	8.307	6.403	5.956	7.840
Ca4				3.796	5.780	6.036	5.660	6.818	7.572	8.384	8.360	9.732	10.284	9.493	8.219	6.577	5.854	7.710
Ca5					3.796	5.855	6.607	7.264	7.600	8.589	8.709	9.967	10.307	9.497	8.496	7.710	7.754	9.743
Ca6						3.795	5.956	7.134	6.868	7.317	7.339	8.446	8.638	7.867	7.220	6.994	7.715	10.107
Ca7							3.790	5.946	6.375	8.142	5.529	6.654	7.073	6.496	6.105	5.657	6.757	9.554
Ca8								3.793	5.777	8.324	5.431	8.363	7.101	6.984	6.835	6.182	7.179	9.839
Ca9									3.796	5.968	6.745	7.659	8.080	8.242	8.535	7.935	8.634	10.810
Ca10										3.799	6.014	7.368	7.993	8.217	8.481	7.855	8.414	10.532
Ca11											3.798	6.069	7.327	7.846	8.152	7.676	8.185	10.422
Ca12												3.797	5.997	6.937	6.900	6.414	7.073	9.705
Ca13													3.804	6.009	6.825	6.945	7.971	10.540
Ca14														3.806	6.060	7.149	8.726	11.235
Ca15															3.803	5.980	7.935	10.483
Ca16																3.799	6.145	8.728
Ca17																	3.800	6.235
Ca18																		3.807

* **Units are in Å**

Table 8 Mean form matrix of conformational sample two of isolated L1*

	Ca2	Ca3	Ca4	Ca5	Ca6	Ca7	Ca8	Ca9	Ca10	Ca11	Ca12	Ca13	Ca14	Ca15	Ca16	Ca17	Ca18	Ca19
Ca1	3.811	5.989	7.482	8.646	8.947	9.630	10.566	11.391	11.675	12.546	12.775	14.326	14.977	14.184	12.765	10.700	9.462	9.410
Ca2		3.801	5.896	7.008	6.796	7.261	8.301	9.253	9.687	10.637	10.906	12.411	13.008	12.285	11.037	9.369	8.915	9.899
Ca3			3.794	5.915	6.113	5.368	5.968	7.370	7.866	8.656	8.617	10.015	10.565	9.788	8.410	6.440	6.008	7.813
Ca4				3.798	5.763	6.013	5.617	6.710	7.391	8.331	8.275	9.667	10.208	9.514	8.377	6.894	6.389	8.134
Ca5					3.797	5.855	6.597	7.187	7.420	8.539	8.646	9.991	10.505	9.900	8.966	7.969	8.093	9.928
Ca6						3.794	5.929	7.021	6.750	7.280	7.419	8.708	9.101	8.500	7.770	7.042	7.675	9.825
Ca7							3.790	5.936	6.418	6.248	5.825	7.015	7.424	6.798	6.308	5.519	6.585	9.195
Ca8								3.794	5.771	6.320	5.329	6.169	6.703	6.474	6.461	5.804	6.789	9.529
Ca9									3.794	5.968	6.695	7.492	7.911	8.125	8.531	8.006	8.579	10.790
Ca10										3.797	6.006	7.416	8.166	8.487	8.741	8.128	8.496	10.604
Ca11											3.798	6.070	7.441	8.014	8.178	7.745	8.222	10.474
Ca12												3.797	6.022	6.929	6.719	6.413	7.130	9.784
Ca13													3.805	5.999	6.746	7.121	8.213	10.790
Ca14														3.806	6.048	7.313	8.825	11.325
Ca15															3.804	6.026	7.843	10.330
Ca16																3.798	6.142	8.663
Ca17																	3.798	6.217
Ca18																		3.808

* **Units are in Å**

2.2.2 Improving the mean form matrix using distance geometry

The estimated mean form matrix does not always represent a real three-dimensional shape because the means of the inter-atomic distances may violate the triangle inequality relationship (section 1.4.2). Distance geometry techniques are used to build conformations of molecules by converting a set of inter-atomic distances or distance bounds into a set of three-dimensional Cartesian coordinates. In this work, the three-dimensional conformation of L1 loop whose form matrix is closest to the estimated mean form matrix is obtained using the distance geometry program DGEOM, which is applied to convert the inter-C_α distances in the estimated mean form matrix into the X, Y and Z coordinates of C_α atoms.

One of the conformations from the isolated L1 loop obtained from the Monte Carlo simulation is used as a DGEOM input file, which defines the bond angles and bond lengths between C_α atoms. The constraint file is the crucial factor that affects the quality of the output

conformations produced by DGEOM. In order to get a structure as close as possible to the estimated mean form matrix and which has a realistic loop conformation, different constraint conditions were tested (**Table 9**).

Table 9 Results for different constraint conditions

	Exp 1.a[1,2] Lower bound is 3.8 Å	Exp 1.b[1,2] Lower bound is sum of van der Waals radii	Exp 2[1,3] Lower ≡ Upper bounds	Exp 3[1,4] Lower ≠ Upper bounds
Test 1: Bond angle assessment	40°-130° Max frequency at 50°-60°	30°-160° Max frequency at 30°-40°	80°-160° Max frequency at 100°-130°	68° -176° Max frequency at 100°-130°
Test 2: Sum of distance errors of inter-C_α (Å)	262 ± 25	599 ± 73	134.0 ± 0.3	24 ± 3
Test 3: Sum of distance errors of adjacent inter-C_α (Å)	2 ± 1	18 ± 3	7.93 ± 0.06	14.4 ± 0.5
Test 4: Sum of distance errors of $C_{\alpha 1}$- $C_{\alpha 19}$ (Å)	1.4 ± 1.5	2.9 ± 1.5	0.22 ± 0.01	0.6 ± 0.3
Test 5: Structure realistic, e.g. no bonds overlapping	90%	0	100%	100%

[1] Results based on 10 DGEOM-produced structures for each experiment
[2] Upper bound ≡ entries in mean form matrix (FM)
[3] Upper ≡ lower bound ≡ entries in FM
[4] Lower bound ≡ entries in FM- 0.83Å, Upper bound ≡ entry in FM + 0.83Å

Several criteria were applied to judge whether an output structure produced by DGEOM was acceptable or not. For example, the structure should form a loop with two "feet" fixed and the rest of the loop staying above the XY plane; there should be no overlaps within a loop (see **Figures 5a** and **6**). The sum of distance errors and specific distance errors were very important parameters reflecting the quality of the output structures of DGEOM. The sum of distance errors refer to the sum of all individual inter-C_α distance differences between the output structure and the distance bounds given by the constraint file. The specific distance

errors are the sum of the adjacent inter-C_α distance differences between the output structure and the adjacent inter-C_α distances, which should be about 3.8 Å, or the inter-C_α distance difference between $C_{\alpha 1}$ and $C_{\alpha 19}$, which is 9.41 Å. The smaller the distance error the better is the output structure produced by DGEOM.

The best constraint conditions are those for which the DGEOM program provides one or more structures, which form a loop shape without overlaps, have correct C_α-C_α-C_α bond angles corresponding to a β-strand (100°-160°) and have the smallest sum error and specific errors. In **Table 9,** the different constraints of inter-C_a distance for Exp 1.a and 1.b, Exp 2 and Exp 3 are compared. In Exp 3, the upper and lower bound are equal to the entries in mean form matrix (FM) ±0.83Å, which is an arbitrary number and less than half of the van der Waals radius of C_α. Test 1 shows the Exp 2 and Exp 3 have correct bond angle of β-strand, which is about 100°-130°, while the bond angles in Exp1.a and 1.b are outside this range. All loop structures in Exp 2 and Exp 3 are realistic without any bonds overlapping while Exp 1.a and 1.b have some or all bonds overlapping. The sum of distance errors in Exp 2 is larger than Exp 3. However, the sums of distance errors of adjacent inter-C_α and $C_{\alpha 1}$- $C_{\alpha 19}$ in Exp 2 are smaller than those in Exp 3. The DGEOM-produced structures in Exp 2 are the closest structures to the estimated mean form matrix. In Exp 2, the constraints are that the lower bound is equal to upper bound. Both of them are equal to the entries in estimated the mean form matrix. Therefore, the best result in **Table 9** is Exp 2 based on the above criteria. Therefore, the constraint conditions for Exp 2 were adopted for use with DGEOM for the comparison of two conformational samples. The improved mean form matrix for sample one obtained from the best three-dimensional conformation produced by DGEOM corresponding to the constraints given in the mean form matrix in **Table 7** are shown in **Table 10**. The values in **Table 10** sacrificed some accuracy in order to represent a real three-dimensional

shape. For example, the entries on off diagonal are not always about 3.8 Å (e.g. see $C_{\alpha 7}$-$C_{\alpha 8}$ distance).

Table 10 Improved mean form matrix for sample one obtained from DGEOM*

	Ca2	Ca3	Ca4	Ca5	Ca6	Ca7	Ca8	Ca9	Ca10	Ca11	Ca12	Ca13	Ca14	Ca15	Ca16	Ca17	Ca18	Ca19
Ca1	3.812	5.513	6.380	8.158	9.856	10.246	11.341	13.539	12.537	13.323	13.539	15.997	16.677	14.985	12.758	10.559	8.895	[illegible]38
Ca2		3.445	5.583	5.999	6.620	7.661	9.164	11.264	11.440	12.458	12.164	14.089	14.108	12.056	10.128	8.724	8.272	9.415
Ca3			3.408	5.130	5.290	4.881	6.266	8.870	8.763	9.268	8.759	10.797	11.211	9.502	7.510	5.623	5.097	7.038
Ca4				3.854	5.921	5.214	5.489	7.123	6.199	7.319	8.007	10.596	11.322	10.599	9.342	7.222	5.803	8.074
Ca5					3.720	5.113	5.509	6.019	6.783	9.029	9.652	11.510	11.113	10.281	9.955	9.030	8.862	11.486
Ca6						3.252	4.560	6.064	7.962	9.488	8.926	9.871	8.730	7.175	7.168	7.384	8.613	11.498
Ca7							2.040	5.081	6.233	6.791	5.753	6.928	6.599	5.488	5.136	4.773	6.082	9.293
Ca8								3.386	4.410	5.107	4.612	6.017	5.986	6.024	6.334	5.784	6.594	9.929
Ca9									3.341	5.621	6.508	7.588	7.074	8.174	9.411	9.145	8.611	12.879
Ca10										3.292	5.632	7.848	8.722	9.995	10.423	9.118	8.462	11.414
Ca11											3.196	5.871	7.945	9.469	9.509	7.870	7.057	9.796
Ca12												3.195	5.766	6.953	6.800	5.546	5.834	8.725
Ca13													3.391	5.415	6.366	6.621	8.163	10.936
Ca14														3.433	6.098	7.703	10.001	13.032
Ca15															3.393	6.049	9.048	11.889
Ca16																3.283	6.609	9.007
Ca17																	3.343	5.866
Ca18																		3.349

* **Units are in Å**

2.3 Methodologies for Euclidean distance matrix analysis assessment

In this project, Euclidean distance matrix analysis, EDMA, is used to detect whether or not two conformational samples are significantly different, and if so, the inter-atomic distances responsible for that difference. To the best of our knowledge, it is the first time the EDMA methods are used to analyze conformational samples of molecules obtained from Monte Carlo simulations. Therefore, before the EDMA methods can be confidently applied to our system, the first thing was to develop a methodology to assess the reliability of the EDMA method. Both the EDMA-I bootstrap null hypotheses test and the confidence interval test would be required to pass both positive and negative control tests. Only if the EDMA

methods pass all the positive and negative control tests can they be applied with confidence to the investigation of conformational changes of the complete antibody binding site.

2.3.1 Detecting significant differences using EDMA-I bootstrap null hypotheses tests

EDMA-I bootstrap null hypotheses tests were used to detect if there were significant differences between samples one and two. Because samples one and two were obtained from two independent Monte Carlo simulations for the same system, i.e. isolated L1 loop, the conformational distribution from the two samples should not have any significant differences.

The EDMA comparison methods were applied to samples one and two in order to test the specificity and selectivity of the methods. The definition of selectivity is the ability to distinguish small differences, whereas specificity was how well a test can correctly identify the negative cases in a study.[58]

2.3.1.1 Negative control tests for samples of isolated L1 loop

In this paper, samples one, two, four, five, six and seven (**Table 6**) are obtained from independent Monte Carlo simulations for the same system, the isolated L1 loop. So, there should be no significant differences between any of these conformational samples. The purpose of comparing these conformational samples using the EDMA-I bootstrap null hypothesis test is to create a negative control test. The negative control test should show a true negative result that is that there is no significant difference between the two compared samples. If all the negative control tests for these conformational samples result in true negatives, then we will be confident that the EDMA-I bootstrap null hypothesis test is not too selective.

The null hypothesis in our project is that the mean conformations of two samples are not significantly different. If the observed value of the test statistic T_{obs} (equation 18) computed

for the two samples falls within the range of the null distribution of T, the null hypothesis is retained, and the result of the control test is a true negative. Otherwise, if the statistic T_{obs} lies in the extreme tail of the null distribution, the null hypothesis will be rejected and the result of the control test is a false negative.

The statistic T_{obs} is the ratio of the maximum to the minimum values in the *FDM (1, 2)* (equation (16)). Here, T_{obs} is calculated using FORTRAN and DGEOM programs for all pairwise comparisons of conformational samples of the isolated L1 loop. For each comparison, we require DEGOM to produce ten different structures corresponding to the distance constraints provided by each input mean form matrix. Therefore, 100 (10×10) values of T_{obs} are calculated for each sample comparison from the improved mean form matrices obtained from the DGEOM produced conformations. Only the smallest T_{obs} is selected from these 100 values since when T_{obs}=1, the sample means are identical. To illustrate the range of values of T_{obs} found for comparison of samples one and two, 200 values of T_{obs} are plotted in **Figure 23**.

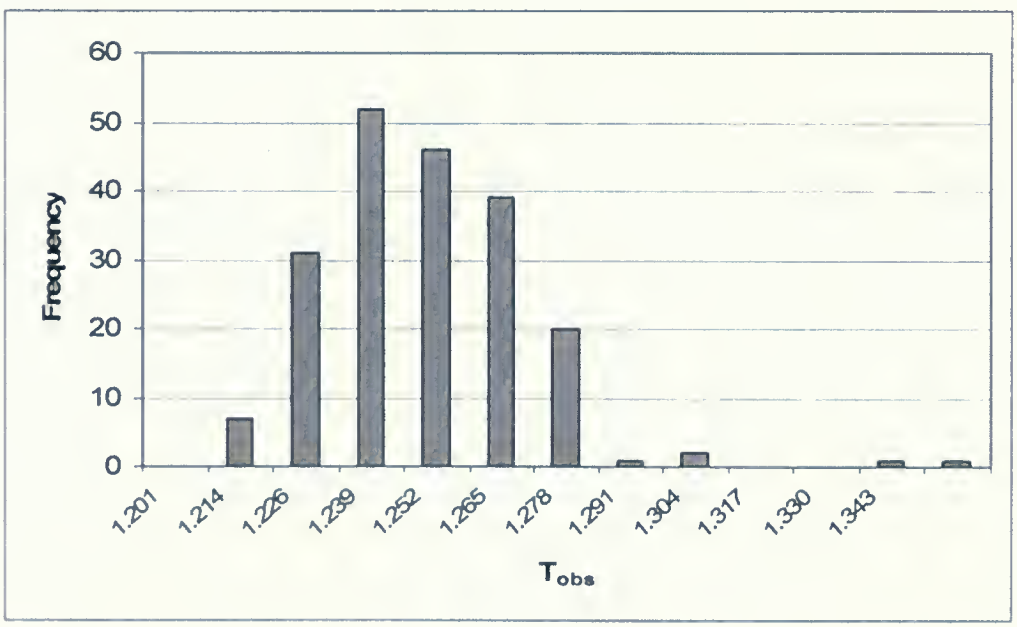

Figure 23. Distribution of T_{obs} for the comparison of sample one and sample two.

If T_{obs} is equal to 1.0, the mean forms from two compared samples are identical. If T_{obs} is much larger than 1.0, these two samples are very different. The smallest value of T_{obs} obtained from comparison of samples one and two is 1.214. **Table 11** contains the smallest values of T_{obs} obtained for pairwise comparisons of all six conformational samples of the isolated L1 loop. There are in total 15 comparisons (**Table 11**) between any two samples among the six samples (one, two, four, five, six and seven). Because the null hypothesis test using EDMA-1 is a one way test,[19] a conformational sample baseline (numerator in equation (16)), which serves as the null distribution sample, was chosen. The sample with larger conformations serves as the baseline. In this case, because all conformational samples for isolated L1 loop have the same size, choosing either sample to be the baseline should not affect the testing results. The comparison between any two samples is conducted twice by changing the null distribution sample (baseline exchange). In **Table 11**, all T_{obs} are larger than 1.0. The largest T_{obs} is 2.905 shows the biggest difference exists between sample six and seven. These results of T_{obs} do not meet our initial expectation because these six samples should express the same conformational population of isolated L1 loop.

Table 11 The best T_{obs} values for the comparison of six samples of isolated L1 loop

Samples	Sample two	Baseline exchange	Sample four	Baseline exchange	Sample five	Baseline exchange	Sample six	Baseline exchange	Sample seven	Baseline exchange
Sample one	1.214	1.922	1.301	1.288	1.489	1.522	1.682	1.856	2.358	1.685
Sample two			1.304	1.285	1.869	1.871	1.918	1.870	2.082	2.066
Sample four					2.086	2.099	2.270	2.127	2.236	2.229
Sample five							2.489	2.541	1.168	1.104
Sample six									2.905	2.876

However, the null hypothesis will not be rejected if these values of T_{obs} fall within the range of the null distributions generated for these samples. The steps of bootstrap null hypothesis testing for sample one and sample two are described in **Figure 10**. First, 8000 conformations are randomly selected from sample one to form samples 1' and 2' respectively. The estimated mean form matrices for samples 1' and 2' are calculated. (equation (10)). Second, the DGEOM program was used in the null hypothesis algorithm to produce ten real three-dimensional structures for each input estimated mean form matrix. The 100 possible T values from each of the ten improved mean form matrices for samples 1' and 2' were calculated. The two structures with smallest T are selected for the null distribution. These two structures likely represent the most similar estimated mean form matrices for the two samples. Third, the above steps were repeated 200 times and 200 T values were calculated (equation (19)) for samples 1' and 2'. Then, a histogram was plotted using EXCEL to describe the null distribution of the bootstrap statistic T based on these 200 T values. If the T_{obs} computed using equation (18) for samples one and two, falls beyond the upper tail of the null distribution, the null hypothesis will be rejected. Otherwise, if T_{obs} falls within the range of the null distribution, the null hypothesis will be retained.

First, sample one is chosen as a baseline sample. The null distribution of sample one is plotted in **Figure 24**. As mentioned before, the value of T_{obs} obtained from comparison of samples one and two is 1.214, which falls beyond the range of the null distribution of sample one. The null hypothesis is rejected. There are significant differences between samples one and two. The result of the negative control test shows false positive.

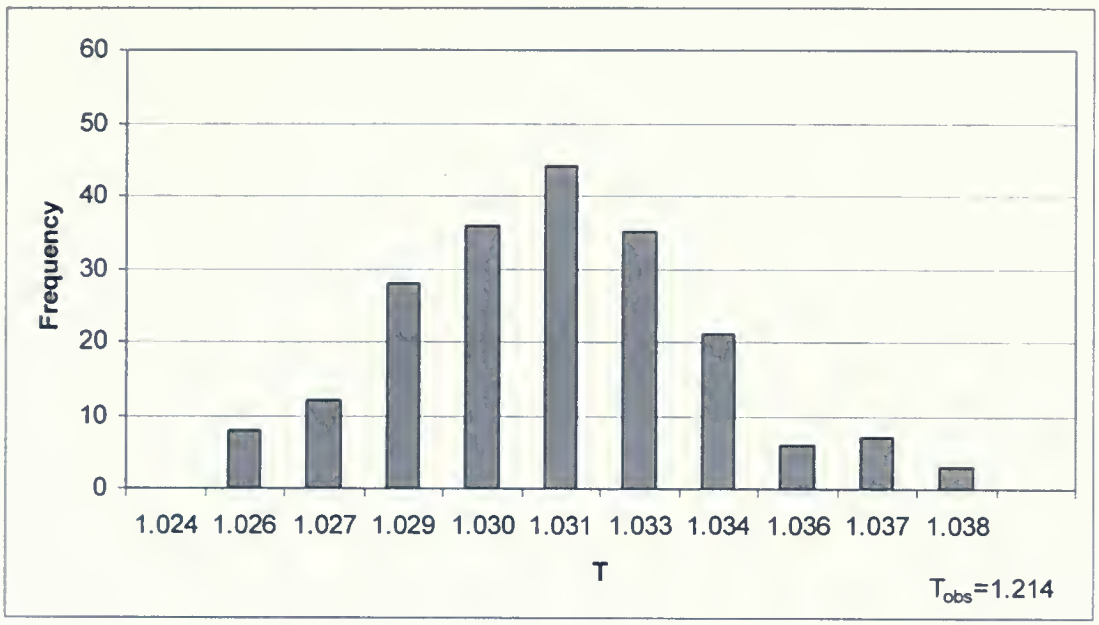

Figure 24. The null distribution of T of sample one, T_{obs} is for samples one and two.

This time sample two is chosen to be the baseline and the maximum and minimum values of *FDM (1, 2)* are 1.393 and 0.725 (in **Table 12**), so that T_{obs}=1.922. The null distribution of sample two is plotted in **Figure 25**. Note that T_{obs}=1.922 falls beyond the null distribution range of sample two. So, the negative control test shows false positive.

Table 12 Form difference matrix *FDM (1, 2)* for samples one and two

	Ca2	Ca3	Ca4	Ca5	Ca6	Ca7	Ca8	Ca9	Ca10	Ca11	Ca12	Ca13	Ca14	Ca15	Ca16	Ca17	Ca18	Ca19
Ca1	1.043	1.011	1.010	0.997	0.940	1.013	1.017	1.054	1.082	1.063	1.033	1.004	0.993	0.963	0.917	0.979	0.955	0.992
Ca2		0.981	1.064	1.012	0.784	0.973	1.068	1.205	1.393	1.286	1.122	1.038	0.979	0.876	0.798	0.881	0.957	0.968
Ca3			1.010	0.948	0.799	0.971	1.071	1.140	1.283	1.241	1.123	1.026	0.978	0.893	0.810	0.866	0.962	1.000
Ca4				1.050	1.317	1.146	0.914	0.873	0.725	0.765	0.868	0.943	1.009	1.106	1.094	1.230	0.911	1.036
Ca5					1.160	1.019	0.923	0.937	0.888	0.921	0.946	0.983	0.999	1.020	0.971	1.078	0.954	17
Ca6						0.943	1.022	1.059	1.021	1.013	1.001	1.018	1.028	1.013	0.923	1.103	0.998	1 045
Ca7							1.263	1.065	1.080	1.069	1.033	1.000	0.970	0.922	0.888	1.057	1.040	1.055
Ca8								0.934	0.953	0.997	1.006	1.045	1.024	1.030	1.035	1.061	1.017	1.022
Ca9									1.019	1.000	0.964	1.035	1.001	0.996	0.997	1.009	0.980	0.990
Ca10										0.981	0.964	1.041	1.006	0.998	1.008	0.967	0.940	0.942
Ca11											0.974	1.060	1.009	1.002	1.042	0.922	0.930	0.910
Ca12												0.976	0.969	0.988	1.094	0.891	0.998	0.955
Ca13													0.995	0.959	1.046	0.872	0.976	0.945
Ca14														0.991	1.053	0.970	1.012	1.001
Ca15															0.995	1.033	1.042	1.046
Ca16																1.021	1.108	1.091
Ca17																	0.986	1.043
Ca18																		0.980

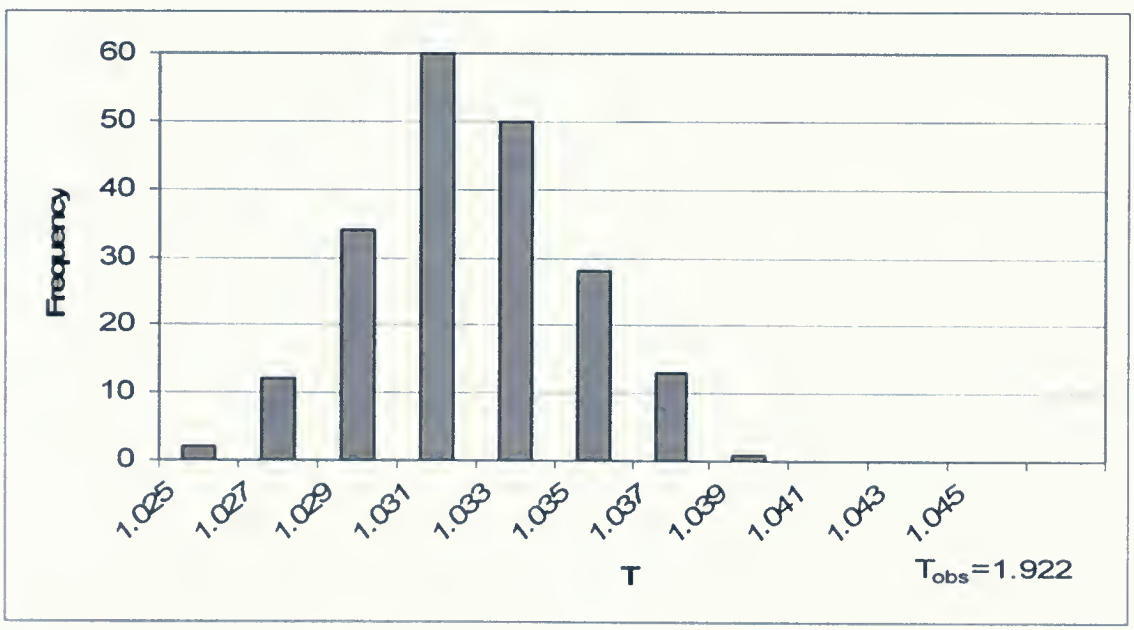

Figure 25. The null distribution of T of sample two, T_{obs} is for samples two and one.

The above steps of bootstrap null hypothesis testing were repeated for each pairwise comparison of six samples. Null distributions of T for all six samples were generated and

plotted (**Figures 26-39**). T_{obs} for all pairwise comparisons of six samples were calculated as described in **Table 11**. To summarize, for all comparisons of six samples (one, two, four, five, six and seven), all T_{obs} fall out of the range of the null distribution in **Figures 26** to **39.** The results of the null hypothesis testing using EDMA-I method showed that the comparisons of any two mean conformations of six samples obtained from independent Monte Carol simulations for a same system, isolated L1, were significantly different. Therefore, the negative control tests of EDMA-I null hypothesis test gave false positive results in all cases. In section 2.5, we explore the effect of increasing the sample size on these results.

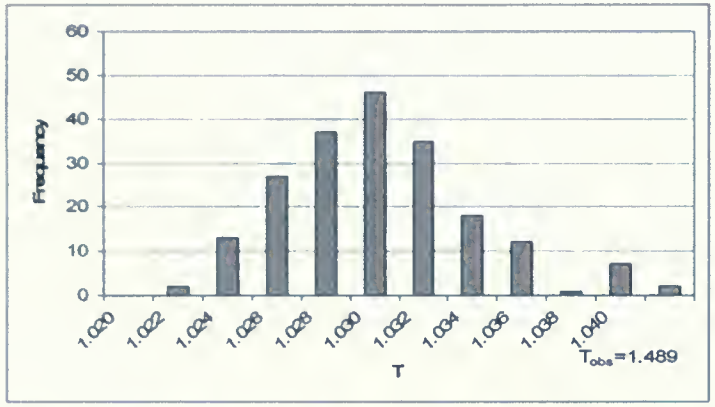

Figure 26. The null distribution of T of sample one, T_{obs} is for samples one and five.

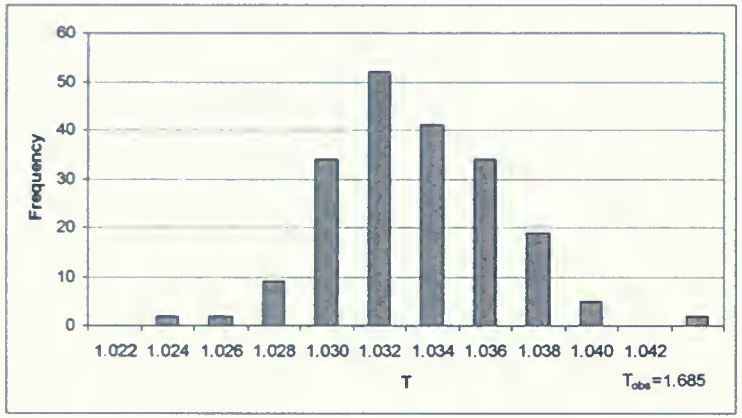

Figure 27. The null distribution of T of sample seven, T_{obs} is for samples one and seven.

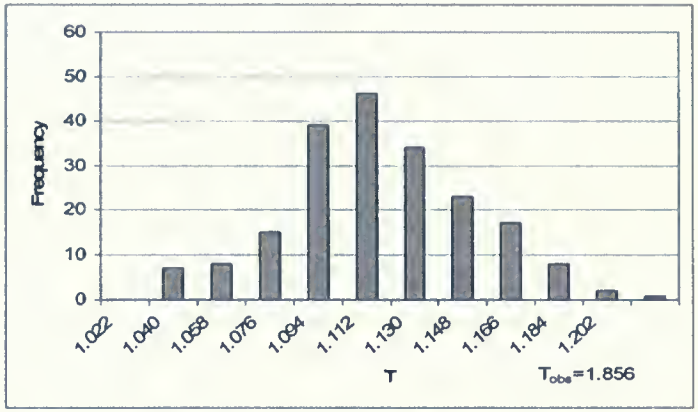

Figure 28. The null distribution of T of sample six, T_{obs} is for samples one and six.

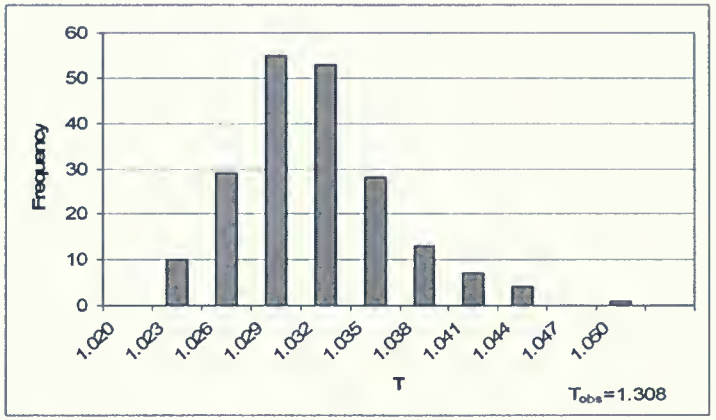

Figure 29. The null distribution of T of sample four, T_{obs} is for samples one and four.

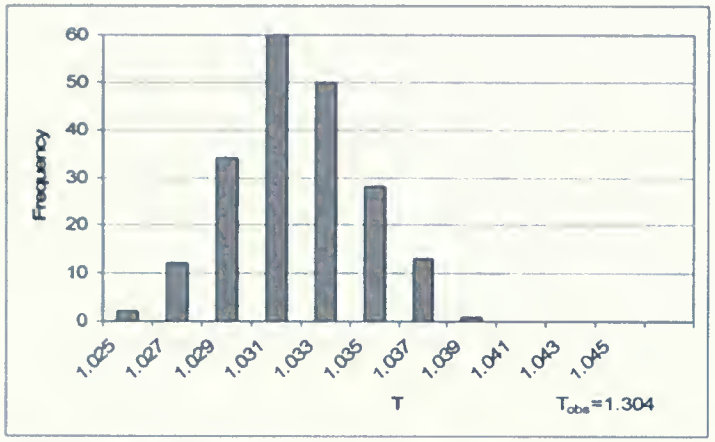

Figure 30. The null distribution of T of sample four, T_{obs} is for samples two and four.

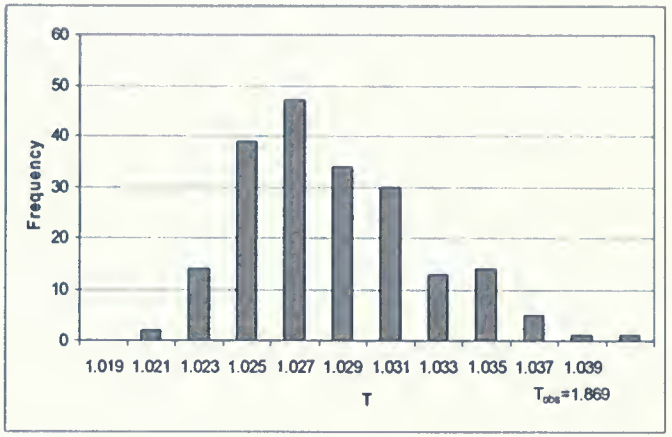

Figure 31. The null distribution of T of sample five, T_{obs} is for samples two and five.

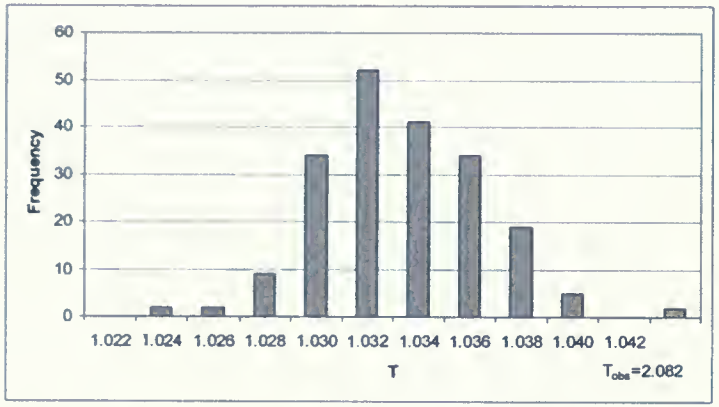

Figure 32. The null distribution of T of sample seven, T_{obs} is for samples two and seven.

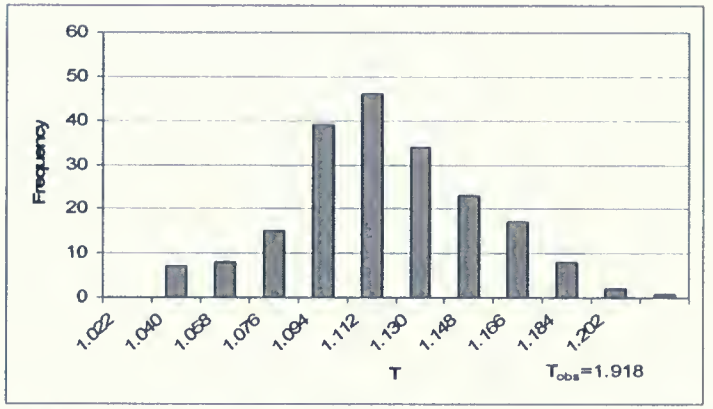

Figure 33. The null distribution of T of sample six, T_{obs} is for samples two and six.

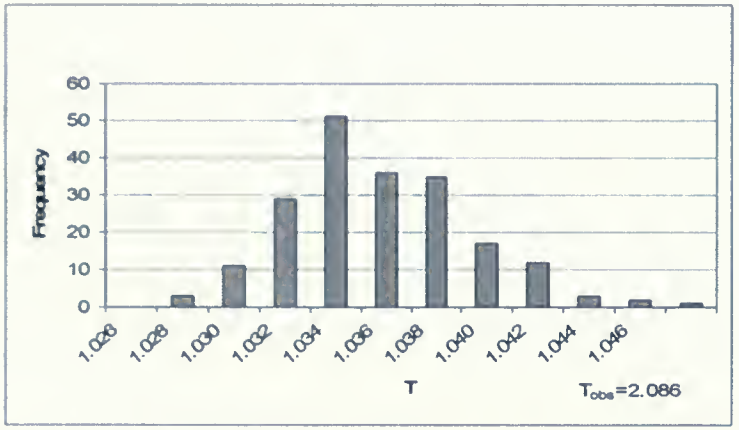

Figure 34. The null distribution of T of sample four, T_{obs} is for samples five and four.

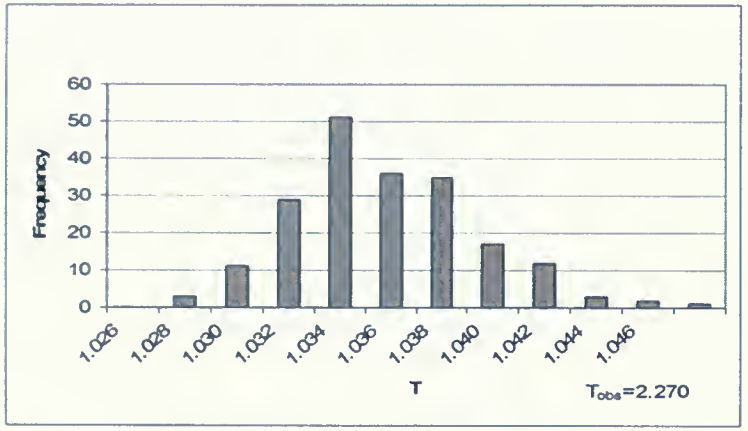

Figure 35. The null distribution of T of sample four, T_{obs} is for samples four and six.

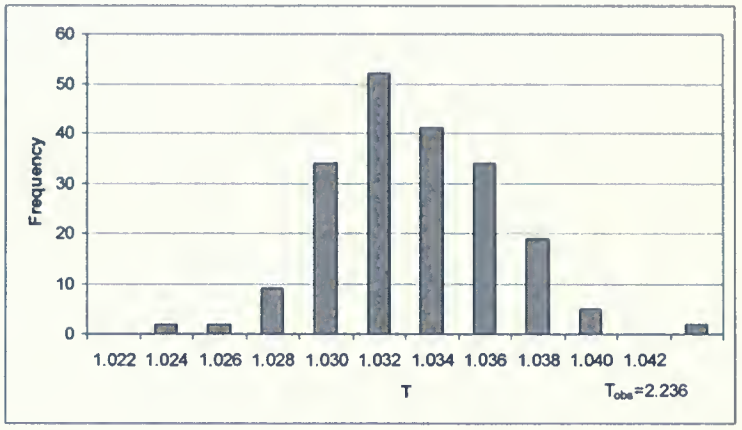

Figure 36.The null distribution of T of sample seven, T_{obs} is for samples four and seven.

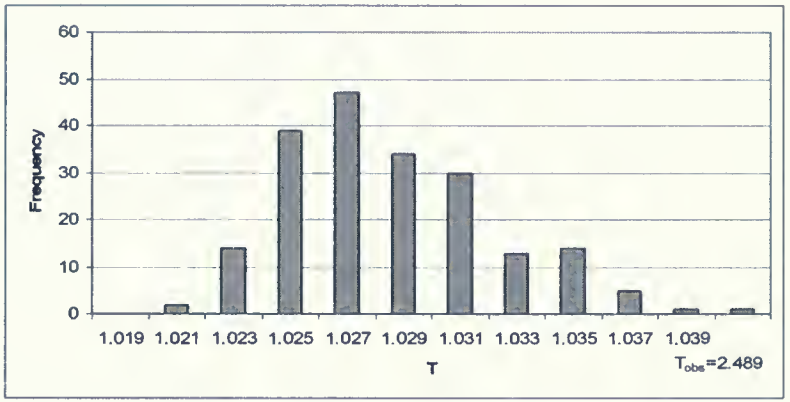

Figure 37. The null distribution of T of sample five, T_{obs} is for samples five and six.

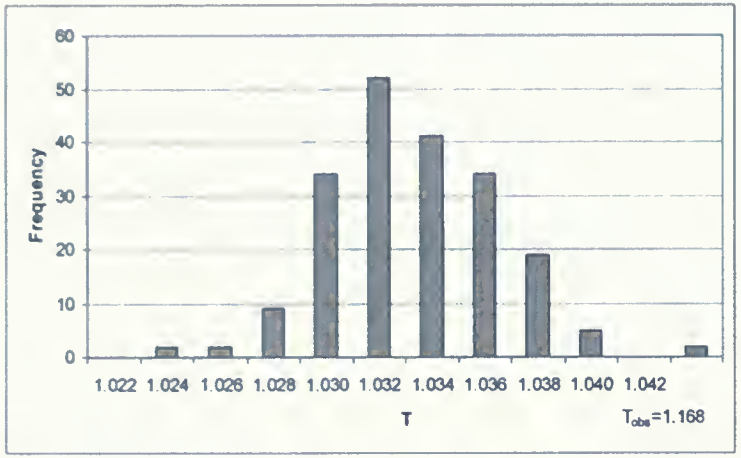

Figure 38. The null distribution of T of sample seven, T_{obs} is for samples seven and five.

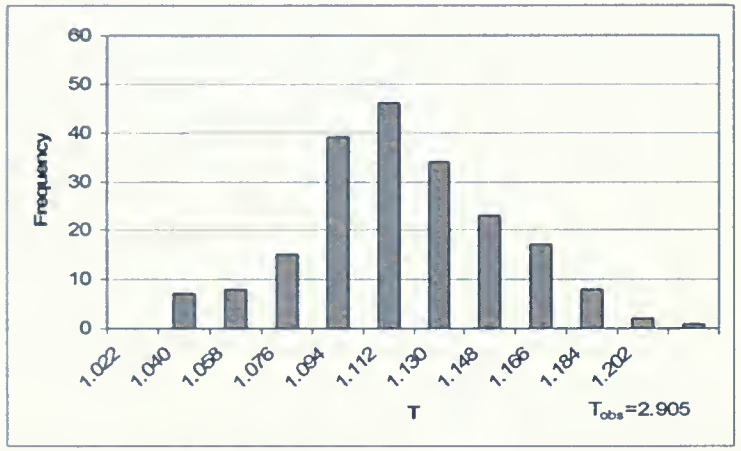

Figure 39. The null distribution of T of sample six, T_{obs} is for samples seven and six.

2.3.1.1.1 Effect of bootstrap sample size on null distribution of T for EDMA-I

In previously conducted negative control tests (Section 2.3.1.1), the number of bootstrap samples used to compute T values for the null hypothesis distribution was 200. In order to see whether increasing the number of bootstrap samples affected the null distribution, one trial of using W=1000 was used to construct the null distribution of sample one (see **Table 6**). **Figure 40** shows the comparison between two null distributions of T for bootstrap samples of 200 and 1000. Note that the comparison between those two null distribution was conducted using the initial estimated form matrices, and not those corresponding to the subsequent 'improved' mean form matrices generated from the best conformations obtained from DGEOM. Using conformations generated by DGEOM employing the constraints of the estimated mean form matrices, the larger values of T seen here are no longer contaminating the null distribution (compare **Figure 40** with **Figure 24**). Based on visual inspection, there is no apparent difference between the two null distributions in **Figure 40**. **Figure 41** is an expanded plot of **Figure 40** covering the range of T occupied by the majority of the population. Here too, there is no apparent difference between the normalized null distributions obtained from bootstrap samples of 200 and 1000. Therefore, we do not believe that our conclusions based on the EDMA-I null hypothesis testing described in Section 2.3.1.1 would be affected by increasing the size of bootstrap sample.

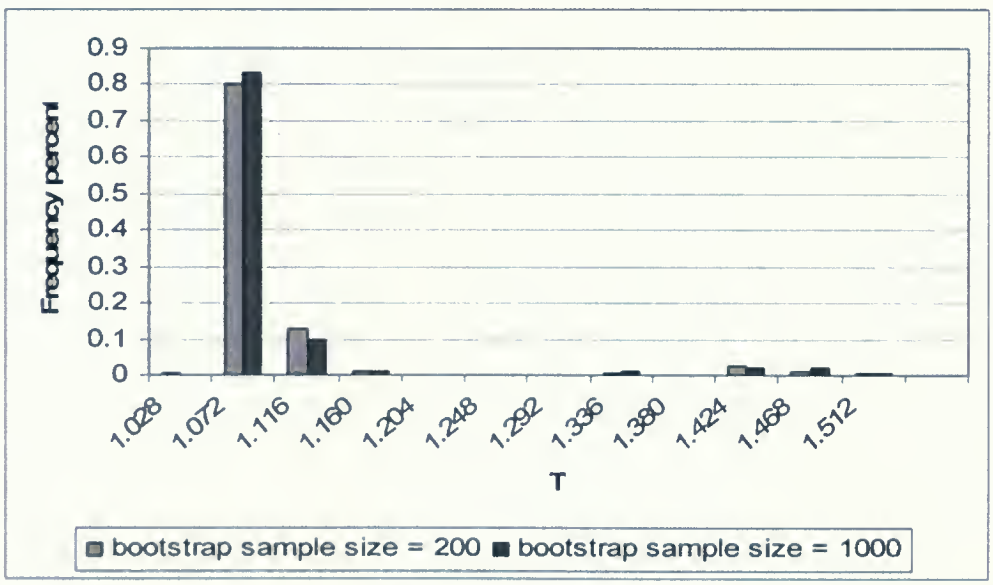

Figure 40. The comparison of two normalized null distributions for T generated from sample one for bootstrap sample sizes of 200 and 1000.

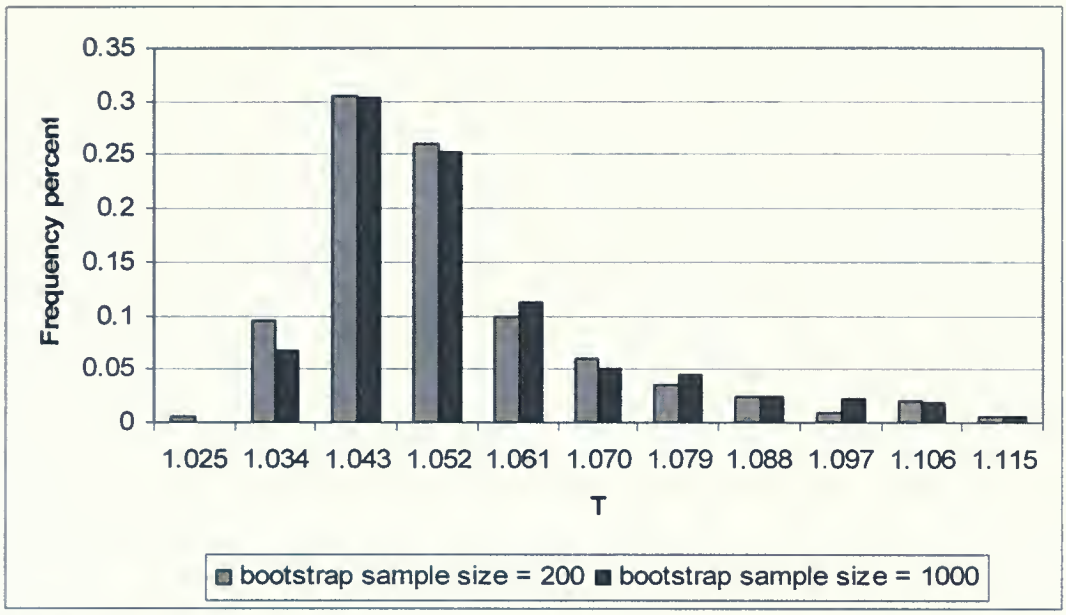

Figure 41. The comparison of two null distributions for T generated from sample one for bootstrap sample sizes of 200 and 1000. Expansion of Figure 40.

2.3.1.2 Negative control tests for samples of L1 in three-loop assembly

The negative control test of EDMA-I null hypothesis test of isolated L1 loop gave false positive results we believe because the conformations for L1 loop have a broad distribution that is discussed in section 2.4. The conformations of L1 in the three-loop assembly are supposed to have a narrower distribution than those of the isolated L1 loop because the L1 in the three-loop assembly is a subset of isolated L1 loop conformations[59] by analysis using clustering the self-organizing map (SOM) method.[60] Samples nine and ten, each has 4000 conformations (**Table 6**), and were obtained from independent Monte Carlo simulations for the L1 in three-loop assembly. So there are should be no significant difference between these two samples. The negative control test of EDMA-I bootstrap null hypothesis test should show a true negative result that there is no significant difference between samples nine and ten.

The process of negative control testing of six samples of isolated L1 loop in section 2.3.1.1 is followed for samples nine and ten. The null distribution of T of sample nine are

plotted and T_{obs} = 1.439 for samples nine and ten (**Figure 42**). Compared to the null distribution of T of six samples of isolated L1 loop (**Figures 24- 39**), the null distribution of sample nine has a narrower distribution, which meets our initial expectation for the conformational samples of L1 in three-loop assembly. T_{obs} falls beyond the range of the null distribution of T of sample nine. The result of the EDMA-I null hypothesis test is a false positive. Because samples nine and ten have the same size, choosing either sample to be the baseline should not affect the testing result.

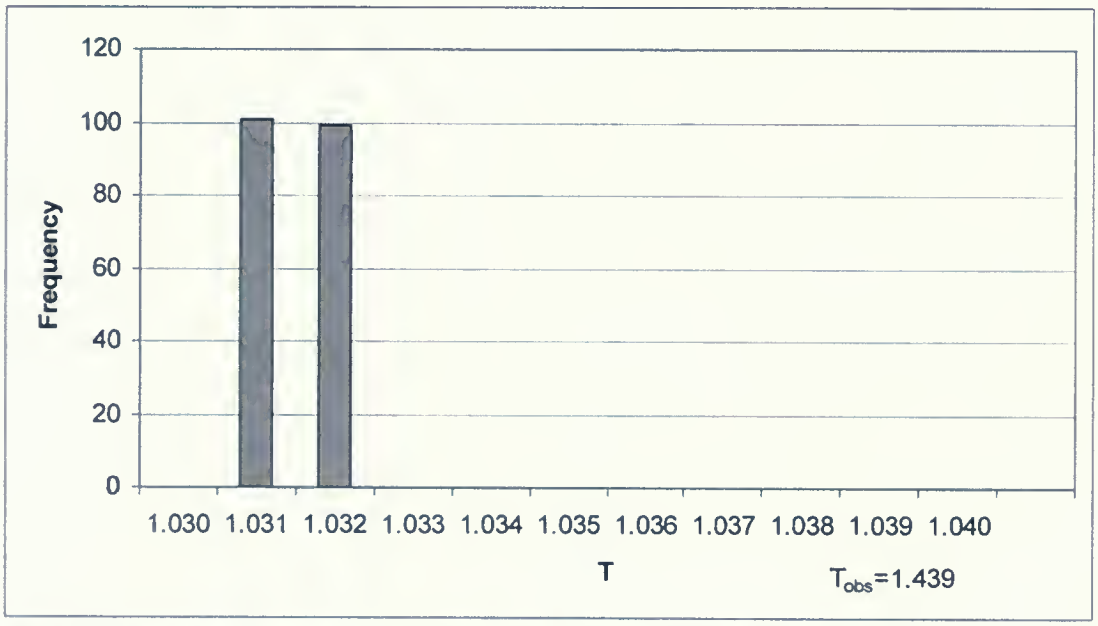

Figure 42. The null distribution of T of sample nine. T_{obs} is for samples nine and ten.

2.3.1.3 Positive control tests for comparison between isolated L1 and L1 in three-loop assembly samples

The conformations of L1 in the three-loop system, **Figure 5b**, have been shown to be a subset of isolated L1 loop conformations by analysis using clustering the self-organizing map (SOM) method.[59] Therefore, we expect there to be significant differences between conformational samples obtained for the isolated L1 loop and those for L1 in the three-loop assembly. The goal for the comparison of isolated L1 and L1 in the three-loop assembly was

to assess the EDMA-I bootstrap null hypothesis test using a positive control test. A true positive result is that a significant difference detected between compared samples that are known to be different. A false negative result is that no significant difference is detected between compared samples that are known to be different. For this positive control test, sample one for the isolated L1 loop and sample nine for L1 in the three loop assembly, **Table 6**, were chosen.

The null hypothesis in our project was that the mean conformations of samples one and nine are identical. If the observed value of the test statistic T_{obs} lies in the extreme tails of the null distribution of T, the null hypothesis is rejected and the result is a true positive. Otherwise, the null hypothesis will be retained and the result is a false negative.

The conformational sample for the isolated L1 loop, sample one (**Table 6**) served as baseline because it had the larger sample size (N=8000) as compared to sample nine for the L1 loop in the three loop assembly (N=4000). The value of T_{obs} was calculated as described in section 1.3.2. The distribution of T for the null hypothesis was obtained using sample one which is shown in **Figure 43**. T_{obs}=3.028 for samples one and nine falls beyond the null distribution of the bootstrap T of sample one, which means there is significant difference between mean conformations of sample one for isolated L1 loop and sample nine for L1 in three-loop assembly. T_{obs}=2.858 for samples one and ten also falls out of the extreme of this null distribution of sample one (**Figure 43**). Therefore, the null hypothesis of the mean conformational similarity for samples one and nine was rejected, which agreed with our expectations. The values of T_{obs} for sample one and samples nine and ten are larger than those for six samples in **Table 11** that reveal the mean conformations of isolated L1 loop and L1 in three loop assembly are more different than those all from isolated L1 loop. The EDMA-I null hypothesis tests for sample one and samples nine and ten have given true positive results.

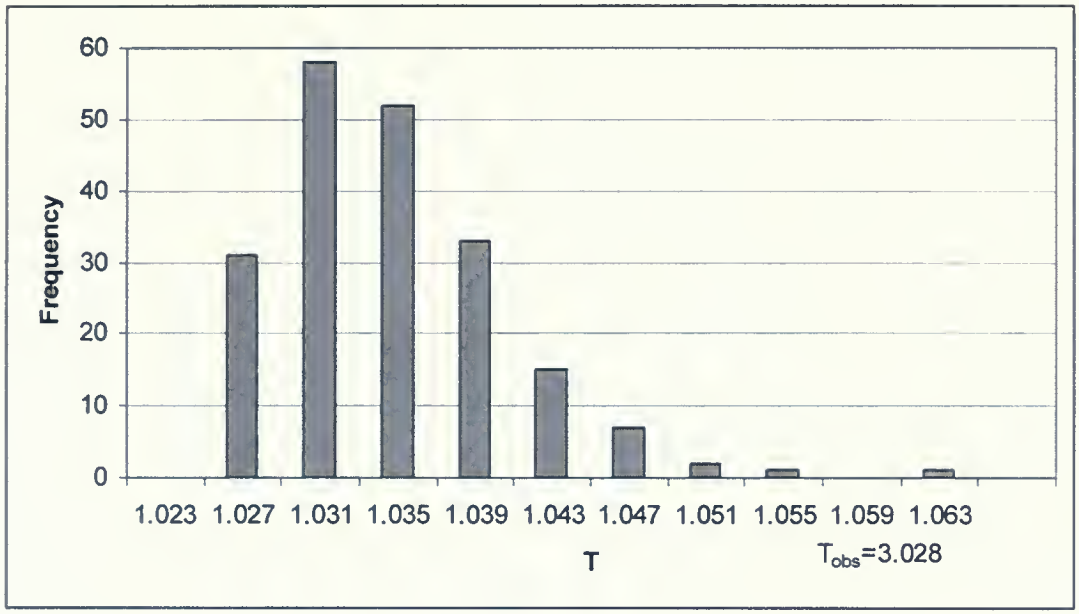

Figure 43. The null distribution of T of sample one. T_{obs} is for samples one and nine.

T_{obs}= 2.972 for samples two and nine falls beyond the null distribution of the bootstrap T of sample two (**Figure 44**), which means there is significant difference between mean conformations of sample two for isolated L1 loop and sample nine for L1 in three-loop assembly.

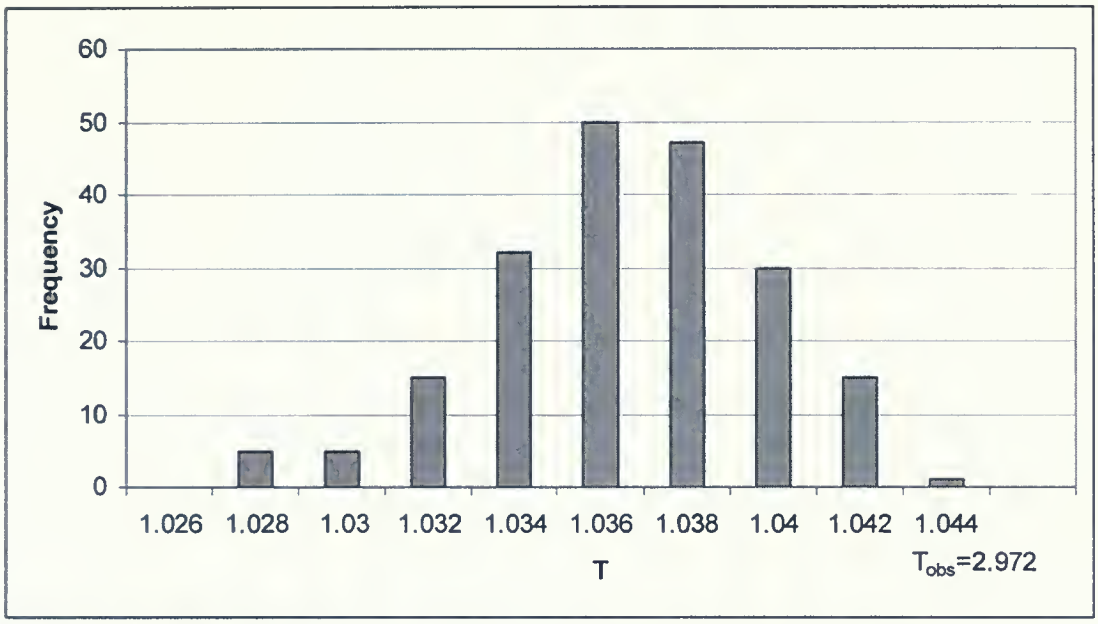

Figure 44. The null distribution of T of sample two. T_{obs} is for samples two and nine.

To summarize, because the conformations of L1 in three-loop assembly are a subset of those for the isolated L1 loop, we expected that the mean conformations of samples one and nine should not be similar. The results of positive control tests of the EDMA-1 null hypothesis for samples one and nine, samples one and ten, samples two and nine agreed with what we expected. The positive control tests for EDMA- 1 bootstrap null hypothesis show true positive results.

2.3.2 Localizing significant differences using confidence interval test

The more different the mean values of the inter-C_α distances are between two conformational samples, the further the individual values in the *FDM* (e.g. **Table 2**) deviate from 1.0. However, checking each value in the *FDM* is tedious work. In this paper, we use a more efficient method, the bootstrap confidence interval test, to pinpoint individual inter-C_α distances that give rise to conformational difference. Both negative control and positive

control tests were designed to assess the success of the bootstrap confidence interval method. Only if the bootstrap confidence interval method passes all the control tests can it be applied to the comparisons of conformational samples obtained from Monte Carlo simulations. The null hypothesis is that the inter-C_α distances of the two conformational samples are statistically identical to each other.

2.3.2.1 Negative control tests for comparison of isolated L1 loop conformational samples

The result of a negative control test should show that no significant difference is detected between compared samples, which were known to be identical. There should be no significant differences between the six conformational samples (one, two, four, five, six and seven) in **Table 6** because they were obtained from independent Monte Carlo simulations for the isolated L1 loop. The purpose of comparing these six conformational samples was to assess the bootstrap confidence interval with a negative control test. A true negative result would be that there is no significant difference detected between comparisons of these six conformational samples. Otherwise, if significant differences are detected between the compared samples, the bootstrap confidence interval test has given a false positive result.

The algorithm for computing the bootstrap confidence intervals, described in section 1.3.3, between two conformational samples was written using a Unix script.

1. The estimated mean forms for two conformational samples, e.g. one and two, were calculated and were input into the DGEOM program to produce 100 structures, which are consistent with the requirement of the distance constraints. FORTRAN programs were developed to pick out two conformations, which result in the smallest T_{obs} since when samples are identical, T_{obs}=1.0. These two picked conformations are most likely to represent the estimated mean form matrix conformations. The chosen mean

conformations for each of samples one and two were used to compute "improved" mean form matrices and subsequently to calculate the *FDM(1, 2)*(equation(16)).

2. 8000 conformations from sample one were randomly selected and called sample 1*; 8000 conformation from sample two were randomly selected and called sample 2* (**Figure 13**). Note that sampling replacement was used, so that any conformation may be selected more than once.
3. "Improved" mean form matrices for samples 1* and 2* were produced from the closest corresponding conformations given by DGEOM and the distance constraints from FM (1*) and FM (2*). The form difference matrix *FDM* (1*2*) was calculated.
4. Steps 2 and 3 were repeated 200 times.

A total of 200 *FDM* (1*, 2*) matrices were collected. All values in those 200 *FDM* (1*, 2*) were written into a matrix with 171 rows and 200 columns (**Figure 14**). Each column was a form difference matrix in vector format and each row contains 200 form difference ratios for a specific inter-C_α distance. In order to obtain the confidence intervals for the inter-C_α distances, the values in each row were sorted in increasing order (**Figure 15**). The minimum and maximum values are the 100% lower and upper confidence limits for that particular inter-C_α distance. Note that if *FDM* (1*, 2*) = 1 for a given inter-C_α distance, that mean inter-C_α distance is identical between the two samples. If *FDM* (1*, 2*) < 1.0, the particular mean inter-C_α distance of sample one is larger than that in sample two. If *FDM* (1*, 2*) > 1.0, the mean inter-C_α distance of sample one is smaller than sample two (**Figure 16**). Therefore, if the value 1.0 is between the lower and upper values of confidence intervals, the given inter-C_α distance is not significantly different between the two compared conformational samples. Otherwise, these values of the inter-C_α distances are different and

the null hypothesis that there is no significant difference between the compared samples is rejected.

Table 13 shows the 100% lower and upper confidence limits for a few of the 171 inter-C_α distances between samples one and two, each of which has 8000 conformations of the isolated L1 loop. For example, the 100% confidence interval between $C_{\alpha 1}$ and $C_{\alpha 2}$ includes 1.0 between the lower and upper limits of 0.9632 and 1.0185, respectively. This means that the mean distance between $C_{\alpha 1}$ and $C_{\alpha 2}$ is the same for both conformational samples. In fact, all the confidence intervals in **Table 13** show that there was no significant difference between samples one and two for this subset of inter-C_α distances.

Table 13 A subset of confidence intervals for comparison of mean inter-C_α distances of conformational samples one and two

Inter-Cα distance	Lower Limit	Upper Limit
$C_{\alpha 1}$ and $C_{\alpha 2}$	0.9632	1.0185
$C_{\alpha 1}$ and $C_{\alpha 3}$	0.9432	1.0265
$C_{\alpha 1}$ and $C_{\alpha 4}$	0.9505	1.0409
$C_{\alpha 1}$ and $C_{\alpha 5}$	0.9525	1.0441
$C_{\alpha 1}$ and $C_{\alpha 6}$	0.9257	1.0451
$C_{\alpha 1}$ and $C_{\alpha 7}$	0.9358	1.0507
$C_{\alpha 1}$ and $C_{\alpha 8}$	0.9516	1.1249
$C_{\alpha 1}$ and $C_{\alpha 9}$	0.9626	1.0569
$C_{\alpha 1}$ and $C_{\alpha 10}$	0.9522	1.1326
$C_{\alpha 1}$ and $C_{\alpha 11}$	0.9279	1.0613
$C_{\alpha 1}$ and $C_{\alpha 12}$	0.9226	1.0993
$C_{\alpha 1}$ and $C_{\alpha 13}$	0.9751	1.1208
$C_{\alpha 1}$ and $C_{\alpha 14}$	0.9632	1.0185
$C_{\alpha 1}$ and $C_{\alpha 15}$	0.9432	1.0265
$C_{\alpha 1}$ and $C_{\alpha 16}$	0.9505	1.0409
$C_{\alpha 1}$ and $C_{\alpha 17}$	0.9525	1.0441
$C_{\alpha 1}$ and $C_{\alpha 18}$	0.9257	1.0451
$C_{\alpha 1}$ and $C_{\alpha 19}$	0.9358	1.0507
$C_{\alpha 2}$ and $C_{\alpha 3}$	0.9516	1.1249
$C_{\alpha 2}$ and $C_{\alpha 4}$	0.9626	1.0569
$C_{\alpha 2}$ and $C_{\alpha 5}$	0.9522	1.1326
$C_{\alpha 2}$ and $C_{\alpha 6}$	0.9496	1.0438
$C_{\alpha 2}$ and $C_{\alpha 7}$	0.9630	1.0422
$C_{\alpha 2}$ and $C_{\alpha 8}$	0.9066	1.0171
$C_{\alpha 2}$ and $C_{\alpha 9}$	0.8991	1.0080
$C_{\alpha 2}$ and $C_{\alpha 10}$	0.9609	1.0983

Table 14 summarizes the results of the bootstrap confidence interval test on all pairwise comparisons of the six conformational samples obtained from Monte Carlo simulations of the isolated L1 loop. The "*" was used to mark the inter-C_α distances that did not include the value 1.0 between the lower and upper limits of the confidence intervals. For example, when samples one and two were compared (column 1 in **Table 14**), the value of 1.0 was found between the lower and upper confidence limits, with the exception of the inter-C_α distance

between $C_{\alpha 16}$ and $C_{\alpha 19}$ (No. 168, **Table 14**). The frequencies of different inter-C_α distances for all 15 comparisons of six samples obtained from isolated L1 loop are calculated and presented in the last column of **Table 14**.

Only 32 of the 171 inter-C_α distances were detected as being different between the two samples at least once over these 15 comparisons. The inter-C_α distance that was most frequently tested as being different was $C_{\alpha 3}$ and $C_{\alpha 17}$ with only five instances out of 15 comparisons. Therefore, despite the EDMA results in section 2.3.1.1 suggesting that all of these conformational samples differ, there is no inter-C_α distance consistently different between the samples. This is what is expected for comparison of samples that are not significantly different.

Table 14 Comparison of mean inter-C_α distances from different samples of isolated L1 loop[11]

No.	Inter-C_α distances	Samples 1-2	Samples 1-4	Samples 1-5	Samples 1-6	Samples 1-7	Samples 2-4	Samples 2-5	Samples 2-6	Samples 2-7	Samples 4-5	Samples 4-6	Samples 4-7	Samples 5-6	Samples 5-7	Samples 6-7	Total No.
1	1 & 2																
2	1 & 3																
3	1 & 4																
4	1 & 5																
5	1 & 6																
6	1 & 7																
7	1 & 8																
8	1 &9																
9	1 & 10																
10	1 & 11																
11	1 & 12																
12	1 & 13				*												1
13	1 & 14																
14	1 & 15												*				1
15	1 & 16										*	*	*				3
16	1 & 17																
17	1 & 18																
18	1 & 19																
19	2 & 3																
20	2 & 4																
21	2 & 5																
22	2 & 6																
23	2 & 7																
24	2 & 8																
25	2 &9																
26	2 & 10																
27	2 & 11																
28	2 & 12																
29	2 & 13																
30	2 & 14																
31	2 & 15												*				1
32	2 & 16											*	*				2
33	2 & 17																
34	2 & 18					*											1
35	2 & 19																

No.	Inter-C_α distances	Samples 1-2	Samples 1-4	Samples 1-5	Samples 1-6	Samples 1-7	Samples 2-4	Samples 2-5	Samples 2-6	Samples 2-7	Samples 4-5	Samples 4-6	Samples 4-7	Samples 5-6	Samples 5-7	Samples 6-7	Total No.
36	3 & 4																
37	3 & 5																
38	3 & 6																
39	3 & 7																
40	3 & 8																
41	3 & 9													*			1
42	3 & 10																
43	3 & 11																
44	3 & 12																
45	3 & 13																
46	3 & 14																
47	3 & 15																
48	3 & 16			*									*				2
49	3 & 17		*		*				*			*	*				5
50	3 & 18																
51	3 & 19		*			*											2
52	4 & 5																
53	4 & 6		*														1
54	4 & 7						*			*							2
55	4 & 8																
56	4 & 9																
57	4 & 10																
58	4 & 11																
59	4 & 12																
60	4 & 13																
61	4 & 14						*			*							2
62	4 & 15																
63	4 & 16											*					1
64	4 & 17																
65	4 & 18																
66	4 & 19				*							*					2
67	5 & 6																
68	5 & 7		*			*											2
69	5 & 8																
70	5 & 9																
71	5 & 10																
72	5 & 11																

No.	Inter-C_α distances	Samples 1-2	Samples 1-4	Samples 1-5	Samples 1-6	Samples 1-7	Samples 2-4	Samples 2-5	Samples 2-6	Samples 2-7	Samples 4-5	Samples 4-6	Samples 4-7	Samples 5-6	Samples 5-7	Samples 6-7	Total No.
73	5 & 12																
74	5 & 13																
75	5 & 14																
76	5 & 15																
77	5 & 16																
78	5 & 17		*						*						*		3
79	5 & 18																
80	5 & 19																
81	6 & 7																
82	6 & 8																
83	6 & 9		*		*												2
84	6 & 10											*					1
85	6 & 11																
86	6 & 12																
87	6 & 13																
88	6 & 14																
89	6 & 15																
90	6 & 16																
91	6 & 17																
92	6 & 18																
93	6 & 19																
94	7 & 8																
95	7 & 9																
96	7 & 10																
97	7 & 11																
98	7 & 12																
99	7 & 13												*				1
100	7 & 14												*				1
101	7 & 15																
102	7 & 16																
103	7 & 17																
104	7 & 18																
105	7 & 19																
106	8 & 9																
107	8 & 10																
108	8 & 11																
109	8 & 12																

No.	Inter-C_α distances	Samples 1-2	Samples 1-4	Samples 1-5	Samples 1-6	Samples 1-7	Samples 2-4	Samples 2-5	Samples 2-6	Samples 2-7	Samples 4-5	Samples 4-6	Samples 4-7	Samples 5-6	Samples 5-7	Samples 6-7	Total No.
110	8 & 13																
111	8 & 14																
112	8 & 15																
113	8 & 16																
114	8 & 17																
115	8 & 18																
116	8 & 19																
117	9 & 10				*												1
118	9 & 11																
119	9 & 12			*			*			*							3
120	9 & 13																
121	9 & 14																
122	9 & 15																
123	9 & 16																
124	9 & 17																
125	9 & 18																
126	9 & 19																
127	10 &11																
128	10 &12					*											1
129	10 &13																
130	10 &14																
131	10 &15																
132	10 &16																
133	10 &17																
134	10 &18																
135	10 &19										*				*		2
136	11 &12															*	1
137	11 &13																
138	11 &14																
139	11 &15																
140	11 &16																
141	11 &17																
142	11 &18																
143	11 &19																
144	12 & 13																
145	12 & 14					*											1
146	12 & 15								*								1

No.	Inter-C_α distances	Samples 1-2	Samples 1-4	Samples 1-5	Samples 1-6	Samples 1-7	Samples 2-4	Samples 2-5	Samples 2-6	Samples 2-7	Samples 4-5	Samples 4-6	Samples 4-7	Samples 5-6	Samples 5-7	Samples 6-7	Total No.
147	12 & 16																
148	12 & 17																
149	12 & 18																
150	12 & 19																
151	13 & 14																
152	13 & 15																
153	13 & 16																
154	13 & 17			*													1
155	13 & 18																
156	13 & 19																
157	14 & 15											*					1
158	14 & 16																
159	14 & 17																
160	14 & 18							*							*		2
161	14 & 19																
162	15 & 16																
163	15 & 17																
164	15 & 18																
165	15 & 19																
166	16 & 17																
167	16 & 18																
168	16 & 19	*											*				2
169	17 & 18																
170	17 & 19																
171	18 & 19																
Total number of differences		**1**	**6**	**4**	**5**	**5**	**3**	**1**	**3**	**3**	**2**	**6**	**9**	**1**	**3**	**1**	

The total number of different inter-C_α distances between any two samples of isolated L1 loop detected by the bootstrap confidence interval test is shown in both the last row of **Table 14** and in **Table 15.**

As mentioned before, there are should be no significant difference between six samples for the isolated L1 loop. Therefore, we expect that there should be no differences in inter-C_α distances. However, in this work, because these six conformational samples have finite sizes, some different inter-C_α distances are detected during the comparisons. Here the criterion is made: if the number of different inter-C_α distance is larger than 10% of total 171 inter-C_α distances (17), there is significant difference between two compared samples. **Table 15** shows the largest number of different inter-C_α distances detected was 9 between samples four and seven. Because the numbers of the different inter-C_α distances detected for all comparisons of six samples of isolated L1 are less than 10% of 171 inter-C_α distances, we accept that there is no significant difference detected among six samples. Therefore, the results of the bootstrap confidence interval tests show that there is no significant difference between the six compared samples obtained from independent Monte Carlo simulations that meet our expectation. The negative control tests for six samples of isolated L1 loop show true negative.

Table 15 Number of inter-C_α distances that are detected as being different

Sample No.	Sample 2	Sample 4	Sample 5	Sample 6	Sample 7
Sample 1	1	6	4	5	5
Sample 2		3	1	3	3
Sample 4			2	6	9
Sample 5				1	3
Sample 6					1

2.3.2.2 Negative control tests for comparison of isolated L1 loop conformational samples

The conformations of L1 in the three-loop assembly are supposed to have a more narrow distribution than those of the isolated L1 loop because the L1 in the three-loop assembly is a subset of isolated L1 loop conformations which were shown using clustering with the self-organizing map (SOM) method.[59]. Samples nine and ten (in **Table 6**), are obtained from independent Monte Carlo simulations for the L1 in the three-loop assembly. So there are should be no significant difference between these two samples. The negative control test of bootstrap confidence interval should show a true negative result that there is no significant difference between samples nine and ten.

The process of negative control testing of confidence intervals for six samples of isolated L1 loop in section 2.3.2.1 is followed for samples nine and ten. A subset of the confidence intervals of samples nine and ten are shown in **Table 16**. In this table, all 26 inter-C_α distances include the value 1.0 between the lower and upper confidence limits. In fact, all 171 inter-C_α distances include the value 1.0 between the lower and upper limits in the confidence interval test for samples nine and ten. Even no occasional differences are detected between samples nine and ten as with the comparisons of six samples indicated in section 2.3.2.1. Therefore, the results of the confidence interval test of conformational samples nine and ten show that no significant difference exists between samples nine and ten of L1 in three-loop assembly. The negative control test of confidence interval for samples of L1 in three-loop assembly is true negative.

Table 16 A subset of confidence intervals for comparison of mean Inter-C_α distances of conformational samples nine and ten

Inter-C_α distance	lower limit	upper limit
$C_{\alpha1}$ and $C_{\alpha2}$	0.8246	1.1678
$C_{\alpha1}$ and $C_{\alpha3}$	0.8042	1.2122
$C_{\alpha1}$ and $C_{\alpha4}$	0.8067	1.2006
$C_{\alpha1}$ and $C_{\alpha5}$	0.7888	1.2078
$C_{\alpha1}$ and $C_{\alpha6}$	0.8102	1.2032
$C_{\alpha1}$ and $C_{\alpha7}$	0.7960	1.2193
$C_{\alpha1}$ and $C_{\alpha8}$	0.8145	1.1741
$C_{\alpha1}$ and $C_{\alpha9}$	0.8275	1.1990
$C_{\alpha1}$ and $C_{\alpha10}$	0.7969	1.2036
$C_{\alpha1}$ and $C_{\alpha11}$	0.8177	1.1881
$C_{\alpha1}$ and $C_{\alpha12}$	0.8235	1.1705
$C_{\alpha1}$ and $C_{\alpha13}$	0.8211	1.1902
$C_{\alpha1}$ and $C_{\alpha14}$	0.8243	1.1824
$C_{\alpha1}$ and $C_{\alpha15}$	0.8469	1.1363
$C_{\alpha1}$ and $C_{\alpha16}$	0.8160	1.1777
$C_{\alpha1}$ and $C_{\alpha17}$	0.7914	1.2127
$C_{\alpha1}$ and $C_{\alpha18}$	0.8006	1.1837
$C_{\alpha1}$ and $C_{\alpha19}$	0.8446	1.1489
$C_{\alpha2}$ and $C_{\alpha3}$	0.8233	1.1719
$C_{\alpha2}$ and $C_{\alpha4}$	0.8290	1.1678
$C_{\alpha2}$ and $C_{\alpha5}$	0.7964	1.2214
$C_{\alpha2}$ and $C_{\alpha6}$	0.8053	1.1980
$C_{\alpha2}$ and $C_{\alpha7}$	0.8190	1.1643
$C_{\alpha2}$ and $C_{\alpha8}$	0.8176	1.1742
$C_{\alpha2}$ and $C_{\alpha9}$	0.8027	1.1968
$C_{\alpha2}$ and $C_{\alpha10}$	0.8158	1.1919

2.3.2.3 Positive control tests for comparison of isolated L1 loop conformational samples and L1 in three loop assembly samples

A test for a positive outcome of the bootstrap confidence interval tests is required, which is when the null hypothesis is rejected, i.e. the test predicts significant differences between the compared samples, which are known to be different. As mentioned before, it is known that there are differences in the conformational samples between isolated L1 loop and L1 in the three-loop assembly.

The algorithm for computing bootstrap confidence intervals that was previously used, section 2.3.2.1, was followed to calculate the bootstrap confidence interval for sample one of the isolated L1 loop, and for the L1 loop in the three loop assembly, sample nine, **Table 6**.

Table 17 contains a subset of the lower and upper values of the form difference calculated for 200 bootstrap samples, where sample one was used as the baseline. All these confidence intervals include 1.0 between the lower and upper limits. Therefore, the bootstrap confidence interval found no significant difference between compared samples. In fact, this was found for all 171 inter-C_α distances.

Table 17 A subset of confidence intervals for comparison of mean Inter-C_α distances of conformational samples one and nine

Inter-C_α distance	Lower Limit	Upper Limit
$C_{\alpha 1}$ and $C_{\alpha 2}$	0.5828	1.7805
$C_{\alpha 1}$ and $C_{\alpha 3}$	0.6004	1.5597
$C_{\alpha 1}$ and $C_{\alpha 4}$	0.5777	1.7390
$C_{\alpha 1}$ and $C_{\alpha 5}$	0.6108	1.7414
$C_{\alpha 1}$ and $C_{\alpha 6}$	0.5776	1.7338
$C_{\alpha 1}$ and $C_{\alpha 7}$	0.5704	1.7415
$C_{\alpha 1}$ and $C_{\alpha 8}$	0.6090	1.5829
$C_{\alpha 1}$ and $C_{\alpha 9}$	0.6088	1.5416
$C_{\alpha 1}$ and $C_{\alpha 10}$	0.6075	1.5710
$C_{\alpha 1}$ and $C_{\alpha 11}$	0.6089	1.7655
$C_{\alpha 1}$ and $C_{\alpha 12}$	0.5797	1.7936
$C_{\alpha 1}$ and $C_{\alpha 13}$	0.5849	1.5591
$C_{\alpha 1}$ and $C_{\alpha 14}$	0.6102	1.5026
$C_{\alpha 1}$ and $C_{\alpha 15}$	0.5822	1.7685
$C_{\alpha 1}$ and $C_{\alpha 16}$	0.5870	1.7508
$C_{\alpha 1}$ and $C_{\alpha 17}$	0.5943	1.7275
$C_{\alpha 1}$ and $C_{\alpha 18}$	0.5713	1.5257
$C_{\alpha 1}$ and $C_{\alpha 19}$	0.5828	1.7805
$C_{\alpha 2}$ and $C_{\alpha 3}$	0.6004	1.5597
$C_{\alpha 2}$ and $C_{\alpha 4}$	0.5777	1.7390
$C_{\alpha 2}$ and $C_{\alpha 5}$	0.6108	1.7414
$C_{\alpha 2}$ and $C_{\alpha 6}$	0.5776	1.7338
$C_{\alpha 2}$ and $C_{\alpha 7}$	0.5704	1.7415
$C_{\alpha 2}$ and $C_{\alpha 8}$	0.6090	1.5829
$C_{\alpha 2}$ and $C_{\alpha 9}$	0.6088	1.5416
$C_{\alpha 2}$ and $C_{\alpha 10}$	0.6075	1.5710

The results of bootstrap confidence interval tests reflect that there were no significant differences detected between isolated L1 and L1 in the three-loop assembly, which are known to be different. Therefore, this positive control test for the bootstrap confidence interval method gave false negative results. The range of values seen in **Table 17** are much larger than the ranges values in **Table 13**, which are the confidence intervals computed representing the samples from same conformational population, isolated L1 loop. For example, lower and upper confidence limits for the inter-Cα distance between $C_{\alpha 1}$ and $C_{\alpha 2}$ are 0.5828 and 1.7805 in

Table 17; the corresponding values in **Table 13** are 0.9632 and 1.0185, respectively. The larger range value reflects the relative conformational diversity of the compared samples. The result of confidence interval shows evidence that the conformations from different conformational populations, isolated L1 loop and L1 in three-loop assembly, are more diverse than the conformations from same conformational population, isolated L1 loop. The results of confidence interval test for samples two and nine are displayed in **Table 18.** The ranges of the lower and upper limits of confidence interval for samples two and nine are similar to those of samples one and nine because samples one and two are obtained from same system of isolated L1 loop.

Table 18 A subset of confidence intervals for comparison of mean Inter-C_α distances of conformational samples two and nine

Inter-C_α distance	lower limit	upper limit
$C_{\alpha1}$ and $C_{\alpha2}$	0.5929	1.7327
$C_{\alpha1}$ and $C_{\alpha3}$	0.5620	1.6418
$C_{\alpha1}$ and $C_{\alpha4}$	0.5911	1.7144
$C_{\alpha1}$ and $C_{\alpha5}$	0.5817	1.7472
$C_{\alpha1}$ and $C_{\alpha6}$	0.5694	1.7283
$C_{\alpha1}$ and $C_{\alpha7}$	0.6027	1.5867
$C_{\alpha1}$ and $C_{\alpha8}$	0.6513	1.5885
$C_{\alpha1}$ and $C_{\alpha9}$	0.5901	1.7227
$C_{\alpha1}$ and $C_{\alpha10}$	0.6431	1.7195
$C_{\alpha1}$ and $C_{\alpha11}$	0.5926	1.7448
$C_{\alpha1}$ and $C_{\alpha12}$	0.5917	1.6138
$C_{\alpha1}$ and $C_{\alpha13}$	0.6418	1.5892
$C_{\alpha1}$ and $C_{\alpha14}$	0.5824	1.7368
$C_{\alpha1}$ and $C_{\alpha15}$	0.6206	1.5749
$C_{\alpha1}$ and $C_{\alpha16}$	0.5780	1.7362
$C_{\alpha1}$ and $C_{\alpha17}$	0.5637	1.7449
$C_{\alpha1}$ and $C_{\alpha18}$	0.5958	1.7175
$C_{\alpha1}$ and $C_{\alpha19}$	0.6449	1.6026
$C_{\alpha2}$ and $C_{\alpha3}$	0.6373	1.5764
$C_{\alpha2}$ and $C_{\alpha4}$	0.6535	1.6099
$C_{\alpha2}$ and $C_{\alpha5}$	0.5875	1.7588
$C_{\alpha2}$ and $C_{\alpha6}$	0.6373	1.6037
$C_{\alpha2}$ and $C_{\alpha7}$	0.6340	1.6173
$C_{\alpha2}$ and $C_{\alpha8}$	0.6456	1.5823
$C_{\alpha2}$ and $C_{\alpha9}$	0.6432	1.5955
$C_{\alpha2}$ and $C_{\alpha10}$	0.5798	1.7435

2.4 EDMA assessment results

In summary, the EDMA-I bootstrap null hypothesis tests revealed false positive results for the comparison of samples one, two, four, five, six and seven of the isolated L1 loop (section 2.3.1.1). The tests also showed true positive results for conformational samples of isolated L1 loop and L1 in three-loop assembly (section 2.3.1.2).

The bootstrap confidence interval tests showed true negative for comparisons of the samples of the isolated L1 loop, and false negative for the conformational comparisons between isolated L1 loop and L1 in three-loop assembly. The EDMA-I null hypothesis was too selective for our comparisons while the confidence interval tests method are not selective enough for our system. **Table 19** summarizes our findings. Moreover, the bootstrap confidence interval appears to be unable to locate the reason (inter-C_α distances) for the difference in mean forms detected by the bootstrap null hypothesis test on the same samples.

Table 19 Summary of negative control and positive control tests for assessing EDMA methods

Control tests / EDMA methods	Negative control	Positive control
Bootstrap null hypothesis	False positive	True positive
Bootstrap confidence interval	True negative	False negative

The control tests for EDMA-I null hypothesis and confidence interval methods failed. The initial results of control tests did not give us enough confidence to proceed to compare other conformational samples of the antibody using EDMA methods. In the following sections, other methods to improve the EDMA results were explored.

2.5 EDMA methods for samples with increasing sizes

Some factors could affect the control test results of EDMA methods. The number of conformations in each sample, or sample size, may contribute. One possibility was that the sample size of 8000 conformations is not

large enough to represent adequately the entire conformational distribution of the isolated L1 loop. In order to understand how the sample size affects the results of the bootstrap null hypothesis and bootstrap confidence interval tests, larger sample sizes with 16000 and 24000 conformations were tested using EDMA methods. The larger conformational samples were created based on the six samples of isolated L1 loop (**Table 6**) that were used in previous tests are described in **Table 20**. New samples dat12, dat45 and dat67, pooled two conformational samples from samples one and two, four and five, six and seven. Each of the new samples had 16000 conformations. Samples dat124 and dat567 combined three samples of 8000 and so each of them had 24000 conformations.

Table 20 Pooled conformational samples of the isolated L1 loop

Sample name	Composition (samples taken from Table 6)	Conformational sample size
dat12	Samples one + two	16000
dat45	Samples four + five	16000
dat67	Samples six + seven	16000
dat124	Samples one + two + four	24000
dat567	Samples five + six + seven	24000

2.5.1 Bootstrap null hypothesis test

In the previous tests discussed in section 2.3.1.1, EDMA-I bootstrap null hypothesis test was assessed using six samples of the isolated L1 loop, resulting in false positive results. The negative control test was not passed.

Therefore, the number of conformations in each sample was increased in order to evaluate if sample size was a crucial factor.

Samples dat12, dat45 and dat67 were compared and each sample has double the original conformational number. The procedure of the null hypothesis test is as described in section 2.3.1.1. The null distributions of T for the larger samples of 16000 are plotted in **Figure 45-47.** Results show the T_{obs} values for sample sizes of 16000 conformations (samples dat12, dat45 and dat67) were slightly smaller than for the samples with 8000 conformations (samples one, two, four, five, six and seven) in **Table 6.** However, the T_{obs} values were still outside of the range of the null distributions for all comparisons, which means that the EDMA test indicates that there are significant differences between compared samples.

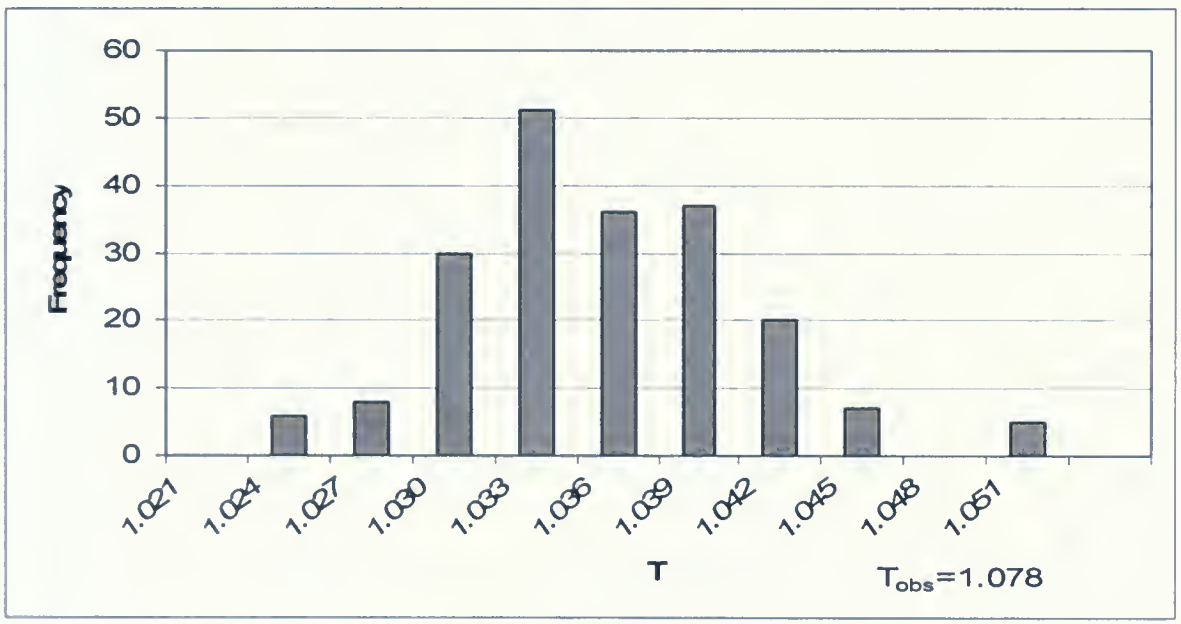

Figure 45. The null distribution of T of dat12 for the comparison of samples dat12 and dat45.

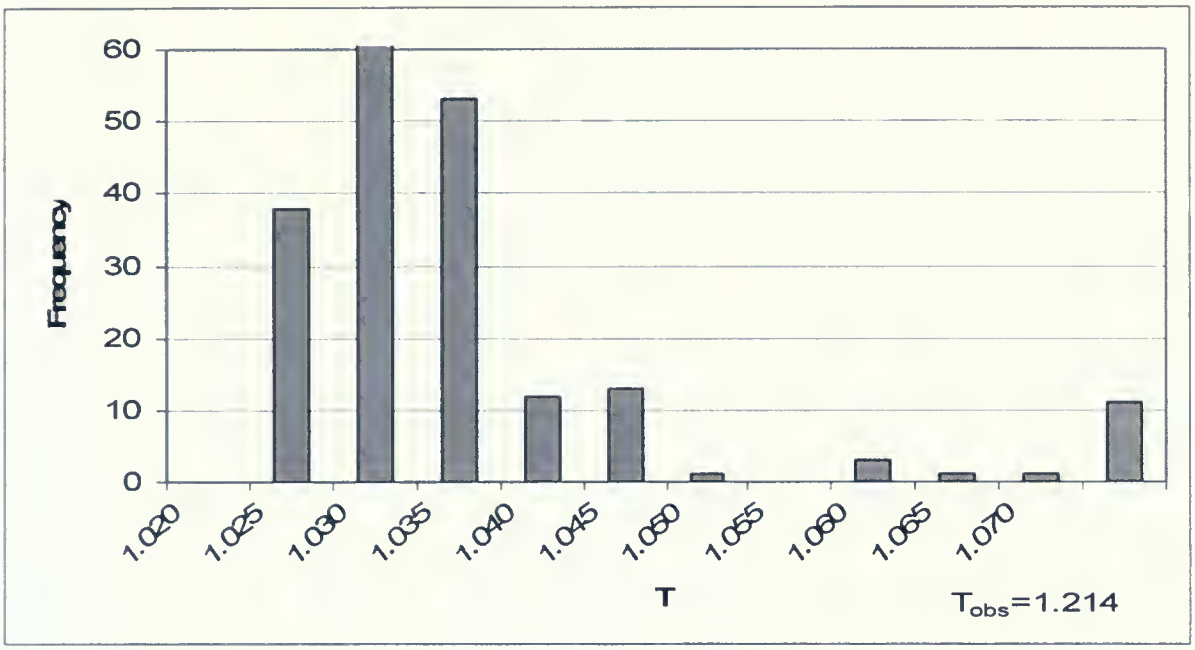

Figure 46. The null distribution of T of dat67 for the comparison of samples dat12 and dat67.

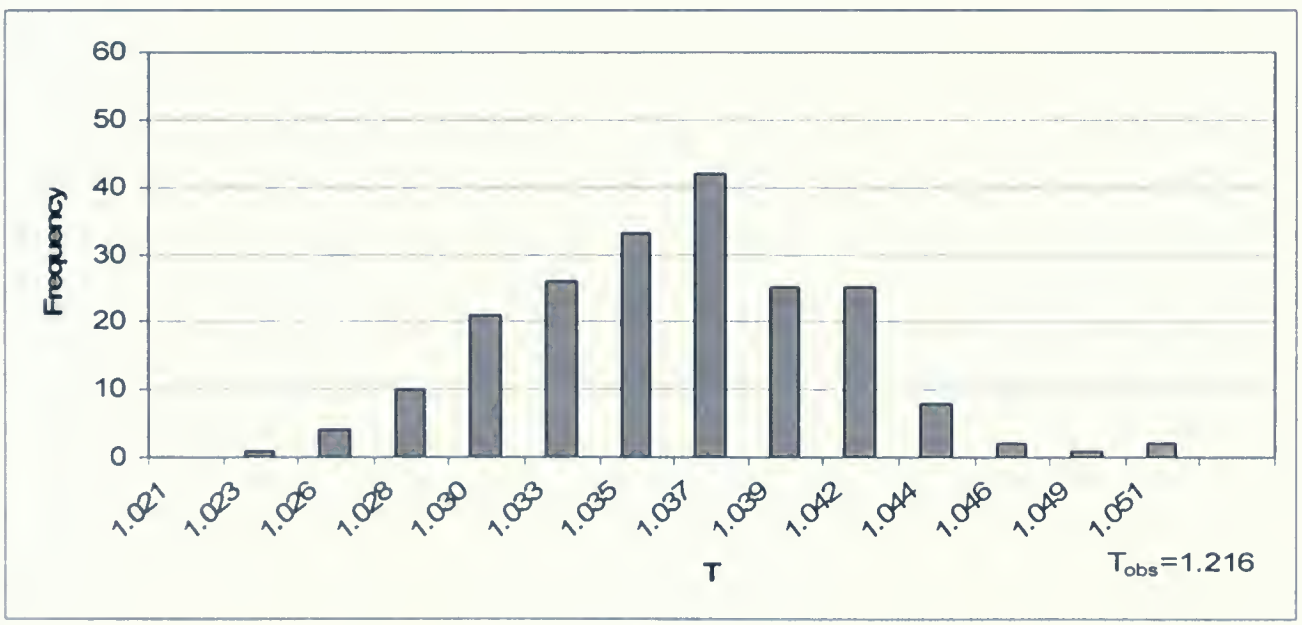

Figure 47. The null distribution of T of dat67 for the comparison of samples dat45 and dat67.

Because the T_{obs} values were slightly decreased as the sample size increased from 8000 to 16000, even larger samples were tested. Samples sizes were increased to 24000 conformations for samples dat124 and dat567, **Figure 48** shows that the null distribution did not change significantly, nor did the T_{obs} value. The results still showed a false positive, which is that the significant difference exists between the compared samples which were known to be identical.

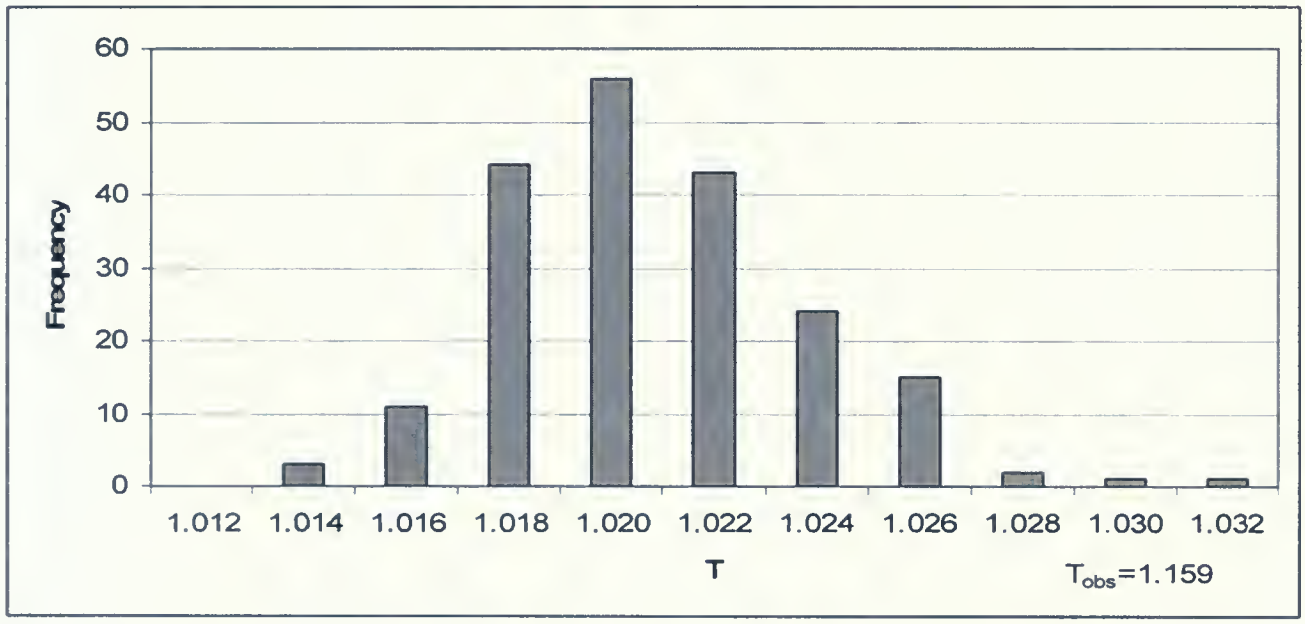

Figure 48. The distribution of bootstrap T statistics of sample dat567 for the comparison of samples dat124 and dat567.

A summary of these null hypothesis test results are shown in **Table 20**. The observed T values, T_{obs}, for compared samples with larger conformational numbers were slightly decreased compared to the samples with 8000 conformations in **Table 11.** However, there was no evidence that showed the observed T would fall within the range of the null distributions for any comparisons. The negative control tests failed for the compared samples for L1 loop with 16000 and 24000 conformations. The negative control tests were not improved by increasing sample sizes.

Table 21 Null hypothesis test results of isolated L1 loop with increased sample sizes

Compared samples	baseline	T_{obs}	Range of T values T_{min} - T_{max}	Frequency peak
dat12-dat45	dat12	1.078	1.022-1.049	1.026/1.027
	dat45	1.084	1.022-1.051	1.033/1.038
dat45-dat67	dat 67	1.234	1.019-1.421	1.025/1.031
	dat 45	1.216	1.022-1.051	1.033/1.038
dat12-dat67	dat 67	1.214	1.019-1.421	1.025/1.031
	dat12	1.212	1.022-1.049	1.026/1.027
dat124-dat567	dat124	1.152	1.013-1.032	1.018/1.020
	dat567	1.159	1.017-1.042	1.026/1.028

The conformations of isolated L1 loop were visualized by using Insight II 2005 molecular modeling software.[5] The three-dimensional coordinates obtained from Monte Carlo simulations were input into the visualized software Insight II. As mentioned before, L1 loop (**Figure 6**) had two feet ($C_{\alpha 1}$ and $C_{\alpha 19}$) and were fixed during the Monte Carlo simulations. The rest of the C_α can move above the XY plane and all 19 C_α form a loop. A broad distribution of these conformations was found.

The EDMA method should be able to quantitatively identify whether or not there is significant difference between compared samples and pinpoint the location of the difference. The successful applications include morphologic investigation on the research of Ts65Dn mouse model, craniofacial morphology in orofacial clefting and molecular conformational analysis of insulin obtained from the solution state NMR introduced in section 1.3. In all cases, the distributions of the conformations were much narrower compared to our system that the conformations of L1 loop were obtained from Monte Carlo simulations. In this project, we believe that the broad distribution of conformations in each sample causes the improved mean forms computed from DGEOM to vary significantly from each other even for sample sizes of 24,000. This may be the reason for poor negative control test results of EDMA tests.

One possible solution is to divide each conformational sample into some sub-samples. Each sub-sample will have more similar conformations and the distribution of each sub-sample will be narrower.

2.6 EDMA tests for samples with narrow conformational distribution

The self-organizing map (SOM) is a tool to convert input data into groups, which have similar data.[60] The goal of the SOM is to classify and visualize the input data on a two-dimensional map.

The SOM was used to divide the conformations in sample one (**Table 6**) into 100 subgroups. Twenty of these 100 clusters are listed in **Table 22**.

The process of generating clustering data using SOM is described as follows:

1. The 8000 conformations were scattered artificially into 10 X 10 bins.
2. The mean conformation in each bin was calculated.
3. The Euclidean distance differences between each conformation and each mean conformation for each bin were evaluated. The conformations are assigned to the bin for which the minimum Euclidean difference occurred.
4. Recalculate the mean conformation for each bin.
5. Repeat step 3 and 4 until no changes occur in the bin memberships.

The conformations in each cluster should be similar to each other. The conformational memberships in the first 20 bins are shown in **Table 22.** Neighboring bins should contain conformations that are more similar to each other than distant bins based on the SOM algorithm.[59]

Table 22 Samples sizes for twenty of 100 conformational clusters obtained by self-organizing map

Samples Name	*Sample 11*	*Sample 12*	*Sample 13*	*Sample 14*	*Sample 15*	*Sample 16*	*Sample 17*	*Sample 18*	*Sample 19*	*Sample 110*
Cluster size	204	110	94	90	141	136	66	132	55	213
Samples Name	*Sample 21*	*Sample 22*	*Sample 23*	*Sample 24*	*Sample 25*	*Sample 26*	*Sample 27*	*Sample 28*	*Sample 29*	*Sample 210*
Cluster size	112	53	65	33	65	78	54	94	47	66

Here, we call the memberships of the first bin *sample 11* and the second bin *sample 12.* The comparisons between samples in the first row of **Table 22** were made by a UNIX script. A Fortran 90 program is used to do the control tests for bootstrap null hypothesis as described in section 2.3.1.1. Difficulties occurred when the improved mean form was generated for each cluster using DGEOM. For the 10 clusters of the first row in **Table 22,** the estimated mean form matrices of samples 11, 13, 15, 18 and 110 cannot be converted to real structures by DGEOM because of the triangle inequality problem (section 1.4.2). Until improved mean form matrices can be generated from these sub-clusters, EDMA test cannot proceed.

3 Discussion

The goals of this project include describing the mean shape of conformational samples obtained from Monte Carlo simulations, to compare the mean shapes and to pinpoint specific inter-C_α distances that contribute to conformational difference. In order to compare mean shapes obtained from two different samples, the mean forms have been estimated using EDMA. The mean form matrix is further improved by locating three-dimensional conformations of our system that are closest to the estimated mean forms of the compared samples using the DGEOM program. Validation of the two EDMA methods, the EDMA-I bootstrap null hypothesis test and confidence interval test, is conducted using both negative control and positive control tests described in previous sections (section 2.3-section 2.6).

Only if both the EDMA-I bootstrap null hypothesis and confidence interval method pass the positive and negative control tests, can the EDMA methods can be applied to the conformational investigation of the antibody binding site model obtained from Monte Carlo simulations. So far, the results from these experiments show that validation of the EDMA methods has failed.

The six conformational samples (samples one, two, four, five, six and seven) obtained from Monte Carlo simulations are known to be drawn from the same distribution. These six conformational samples underwent negative control tests for both the EDMA-I bootstrap null hypothesis test and bootstrap confidence interval test. The results of EDMA- I null hypothesis tests for the conformational comparison of six samples of the isolated L1 loop show that all T_{obs} fall beyond the upper tail of the null distribution of T statistics (**Figures 24 - 39**). Therefore, these tests report that significant difference exists for the comparisons of any two samples of six samples obtained from independent Monte Carlo simulations. The results of EDMA-I bootstrap null hypothesis tests for conformational samples of L1 in three-loop assembly also show T_{obs} falls outside the range of the null distribution (**Figures 42**). For both isolated L1 loop and L1 in three-loop assembly systems, the assessments of the EDMA-I null hypothesis test using negative control display false positive results.

On the other hand, the bootstrap confidence interval tests for six samples of isolated L1 loop were validated using negative control tests. For all comparisons of six samples from isolated L1 loop, the numbers of different inter-C_α distances are less than 10% of total 171 inter-C_α distances **(Table 15)**. So, these results report that no significant difference exists among the six conformational samples of the isolated L1 loop. The negative control test of confidence interval test for L1 in three-loop assembly also shows that no significant difference exists between samples of L1 in three-loop assembly. So, the assessment of bootstrap confidence interval test using negative control displays true negative result.

Then, the EDMA-I bootstrap null hypothesis tests are assessed by positive control tests for the conformational comparison between the isolated L1 loop and L1 in three-loop assembly samples. The conformational samples of isolated L1 loop and L1 in three-loop assembly are known to be significantly different. In **Figures 43** and **44,** the observed T_{obs} falls outside the upper tail of the null distribution of T statistic. Therefore, there is significant difference exist between isolated L1 Loop and L1 in three-loop assembly samples. The assessment of EDMA-I bootstrap for the comparison of isolated L1 loop and L1 in three-loop assembly using a positive control test displayed a true positive result.

The confidence interval tests are also evaluated by a positive control test for the conformational comparison between samples of the isolated L1 loop and L1 in three-loop assembly. For all 171 inter-C_α distances, there is no significant difference detected. The result of this positive control test is a false negative. The EDMA test indicates that there is no significant difference identified between isolated L1 loop and L1 in three-loop assembly samples, which are known to be different.

Some efforts of improving the experimental results have been put into this work by increasing the number of conformations in each sample. Five new samples were obtained by combining six samples of isolated L1 loop. Three samples have 16000 conformations each and the other two samples have 24000 conformations each. The EDMA-I bootstrap null hypothesis tests are re-assessed using a negative control test by the comparison of the conformational samples of isolated L1 loop with increased conformation numbers. As

shown in **Figures 45-48**, the T_{obs} do not fall within the range of the null distribution of T. The results of the negative control for the bootstrap confidence interval test show a false positive. So, there is no improvement found in the on negative control test by increasing the size of each sample.

Two hundred conformations from isolated L1 loop samples were stored and visualized using Insight II. Broad distributions of these conformations were detected. The broad distribution may explain the false positive results of negative tests. In previous study, EDMA methods are successfully applied on the craniofacial investigations of Down syndrome disease because the distribution of the conformations for the mouse models is much narrower as compared to our case.

One possible solution is to narrow down the conformational distribution of each sample. The self-organizing map is applied to cluster a sample into sub-group, in which the similar conformations are grouped together. The EDMA methods are likely suitable for analysis of these smaller, more narrowly distributed conformational clusters. However, difficulties occurred when the improved mean form was generated for each cluster using DGEOM. For 10 clusters on the first row in **Table 22,** samples 11, 13, 15, 18 and 110 can't be converted to real structures by DGEOM because of the triangle inequality problem (section 1.4.2). Therefore, the EDMA tests still can not be evaluated using these clusters produced by SOM.

4 Future work

This is the first time EDMA to be used on the analysis of conformational samples obtained from Monte Carlo simulation. The assessments of EDMA-I null hypothesis and confidence interval tests on CDR loops using positive control and negative control were not successful. There is space to improve the results before the EDMA techniques are applied on the investigation of the molecular conformations of complimentary determine region (CDR) loops at antibody binding site.

Our understanding of the EDMA method may be advanced as the conformational distribution of each sample is within a suitable range. However, the mean form matrices of clusters cannot be converted to obtain a real three-dimensional structure by DGEOM in section 2.6 because of the triangle inequality problem. The constraints in this case are that the lower limits are set equal to the upper limits, which are the values in estimated mean form matrix of each cluster. In order to obtain an improved mean form for each compared sample, the constraint conditions of DGEOM can be further explored by adjusting the lower and upper limits.

In Section 2.2.2, the four experiments with different constraint are tested using four criteria (**Table 9**). The constraint is chosen based on better result in those tests. These experiments and test criteria are relatively limited so that our design for selecting the most suitable constrains for this project may not be good enough. More experiments and tests can be further explored. For example, various lower or upper limits, even different test criteria can be tried which may give us more choices and better results than we obtained before.

As both the EDMA-I null hypothesis and confidence interval tests pass the positive control and negative control, the EDMA tools can more effectively work for our research investigation of the conformational diversity of antibody binding site.

References

(1) Voet, D.; Voet, J. G. In *Biochemistry;* John Wiley & Sons, Inc.: New York, 1994; pp 1360.

(2) University of Medicine and Dentistry, New Jersey Antibody structure and function. *http://www2.umdnj.edu/mimmweb/Instruction/week22/07%20antibody%20str.pdf.*

(3) Pathak, S.; Palan, U. In *Immunology: Essential and Fundamental;* Science Publishers, Inc.: Enfield, New Hampshire, 2005; pp 411.

(4) Sundberg, E. J.; Mariuzza, R. A. In *Antibody Structure and Recognition of Antigen.* Section Title: Immunochemistry; 2004; pp 491-509.

(5) Insight*II*, Biosym Technology **2005**.

(6) Kabat, E. A.; Wu, T. T.; Perry, H. M.; Gottesman, K. S.; Foeller, C. In *Sequences of Proteins of Immunological Interest;* Diane Books Publishing Company: Collingdale, 1992; pp 2719.

(7) Kuby, J. In *Immunology;* W. H. Freeman and Company: New York, 1992; pp 585.

(8) Wampler, J. E. Tutorial on peptide and protein structure. *http://www.bmb.uga.edu/wampler/tutorial/prot2.html.*

(9) Jakubowski, H. Protein structure. *http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olunderstandconfo.html.*

(10) James, L. C.; Roversi, P.; Tawfik, D. S. *Science* **2003**, *299*, 1362-1367.

(11) Foote, J. *Science* **2003**, *299*, 1327-1328.

(12) James, L. C.; Tawfik, D. S. *Trends Biochem. Sci.* **2003**, *28*, 361-368.

(13) Amit, A. G.; Mariuzza, R. A.; Phillips, S. E. V.; Poljak, R. J. *Science* **1986**, *233*, 747-753.

(14) James, M.; Rini, U. S.; Wilson, I. A. *Science* **1992**, *255*, 959-965.

(15) Stanfield, R. L.; Kamimura, M. T.; Rini, J. M.; Profy, A. T.; Wilson, I. A. *Curr. Biol.* **1993**, *1*, 83-93.

(16) Stanfield, R. L.; Fieser, T. M.; Lerner, R. A.; Wilson, O. A. *Science* **1990**, *248*, 712-719.

(17) Tormo, J.; Blaas, D.; Fita, I. *Protein Sci.* **1992**, *1*, 1154-1161.

(18) Tormo, J.; Blaas, D.; Parry, N. R.; Rowlands, D.; Stuart , D.; Tita, I. *EMBO J.* **1994**, *13*, 2247-2256.

(19) Lele, S. R.; Richtsmeier, J. T. In *An Invariant Approach to Statistical Analysis of Shapes;* Interdisciplinary Statistics Series; Chapman & Hall/CRC: Boca Raton, 2001; pp 308.

(20) DGEOM program, Blaney, J. M.; Crippen, G. M.; Dearing, A.; Dixon, J. S. **1990**.

(21) Berman, H. M.; Westbrook, J.; Feng, Z. G.,G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235-242.

(22) De la Cruz, X.; Mark, A. E.; Tormo, J.; Fita, I.; Van Gunsteren, W. F. *J. Mol. Biol.* **1994**, *236*, 1186-1195.

(23) Sergas, A.; Gordon, H. L. *submitted to J. Chem. Phys.*

(24) Bouzida, D.; Kumar, S.; Swendsen, R. H. *Phys. Rev. A* **1992**, *45*, 8894-8901.

(25) Leach, A. R. In *Molecular Modelling, Principles and Applications;* Pearson Education Limited: Harlow, 2001; pp 744.

(26) Johnson, N. L.; Kotz, S. In *Distributions in Statistics: Continuous Univariate Distributions-2;* John Wiley and Sons: New York, 1970; pp 306.

(27) Landau, D. P.; Binder, K. In *A Guide to Monte Carlo Simulations in Statistical Physics;* Cambridge University Press: Cambridge, 2005; pp 448.

(28) Metropolis, N.; Resenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

(29) Wallqvist, A.; Ullner, M. *Proteins.* **1994**, *18*, 267-280.

(30) Richtsmeier, J. T.; Lele, S. *J. Craniofac. Genet. Dev. Biol.* **1990**, *10*, 39-62.

(31) Lele, S. *Math. Geol.* **1993**, *25*, 573-602.

(32) Lele, S.; Richtsmeier, J. T. *Am. J. Phys. Anthropol.* **1991**, *86*, 415-427.

(33) Mao, Z. L.; Siebert, J. P.; Cockshott, W. P.; Ayoub, A. F. *Med. Imag.* **2004**, *5270*, 1395-1402.

(34) McIntyre, G. T.; Mossey, P. A. *Eur. J. Orthodont.* **2004**, *26*, 375-384.

(35) Richtsmeier, J. T.; Burke, D. V.; Lele, S. R. *Am. J. Phys. Anthropol.* **2002**, *119*, 63-91.

(36) Lele, S.; Richtsmeier, J. T. *Am. J. Phys. Anthropol.* **1995**, *98*, 73-86.

(37) Hay, A. D.; Ayoub, A. F.; Moos, K. F.; Singh, G. D. *Cleft. Palate-Cran. J.* **2000**, *37*, 497-502.

(38) Estrada, E., *J. Comput. Chem.* **2007**, 28, 4, 767-777.

(39) Bostick, D. L.; Shen, M.; Vaisman, II. *Proteins* **2004**, 56, 3, 487-501

(40) Wehrens, R.; Putter, H.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 35-52.

(41) Fisher, R. A. In *The Design of Experiments;* Hafner: Edinburgh, 1966; pp 22.

(42) Baxter, L. L.; Moran, T. H.; Richtsmeier, J. T.; Troncoso, J.; Reeves, R. H. *Human Molecular Genetics* **2000**, *9*, 195-202.

(43) Roper, R. J.; St. John, H. K.; Lawler, A.; Reeves, R. H. *Genetics* **2006**, *172*, 437-443.

(44) Brandstatter, E.; Linz, J. K. U. *Methods of Psychological Research Online* **1999**, *4*, 1-14.

(45) Blaney, J. M.; Dixon, J. S. *Rev. Comput. Chem.* **1994**, *5*, 299-335.

(46) Blumenthal, L. M. In *Theory and Applications of Distance Geometry;* Clarendon Press: Oxford, 1953; pp 347.

(47) Havel, T. F.; Wuthrich, K. *J. Mol. Biol.* **1985**, *182*, 281-294.

(48) Wagner, G.; Braun, W.; Havel, T. F.; Schaumann, T.; Go, N.; Wuthrich, K. *J. Mol. Biol.* **1987**, *196*, 611-639.

(49) Wuthrich, K. *Science* **1989**, *243*, 45-50.

(50) Weber, P. L.; Morrison, R.; Hare, D. *J. Mol. Biol.* **1998**, *204*, 483-487.

(51) Kuszewski, J.; Nilges, M.; Brunger, A. T. *J. Biomol. NMR* **1992**, *2*, 33-56.

(52) Crippen, G. M. In *Distance Geometry and Conformational Calculations;* Bawden, D., Ed.; Research Studies Press: Letchworth, 1981; pp 57.

(53) Crippen, G. M.; Havel, T. F. In *Distance Geometry and Molecular Conformation;* John Wiley & Sons: New York, 1988; pp 554.

(54) Havel, T. F.; Snow, M. E. *J. Mol. Biol.* **1991**, *217*, 1-7.

(55) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. *B. Math. Biol.* **1983**, *45*, 665-720.

(56) Blaney, J. M.; Dixon, J. S. *Annu. Rep. Med. Chem.* **1991**, *26*, 281-285.

(57) Nyhoff, L. R.; Leestma, S. C. In *FORTRAN 90 for Engineers and Scientists;* Prentice Hall: New Jersey, 1997; pp 952.

(58) Altman, D. G.; Bland, J. M. *BMJ* **1994**, *308*, 1552-1553.

(59) Gordon, H. L. **2007**, *unpublished work.*

(60) Kohonen, T. In *Self-Organizing Maps;* Kohonen, T., Ed.; Spring Series in Information Sciences; Springer-Verlag: Berlin, 2001; pp 501.