Molecular dynamics of sense and sensibility in processing and analysis of data
Wassenaar, Tsjerk

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2006

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 4

The Effect of an External Condition on the Simulation Outcomes; the Influence of the Box Type and the Force Field
1 Introduction

The objective of many studies using molecular dynamics is to investigate to what degree differences in the simulation conditions influence the outcome of the simulation. The term condition can be interpreted broadly. For example, one could investigate the effect of the temperature, the pressure or the type of solvent. Another interesting class of problem is the assessment of the effect of a ligand on a protein or the effect of the interaction between two or more proteins. In such cases the condition is interpreted as the presence or absence of the associated molecule(s). Here we present a method for the statistical evaluation of the similarity or otherwise of two or more sets of simulations, performed under different conditions.

The comparison of simulations is non-trivial. First, as was outlined in Chapter 2, the results obtained from a simulation form a complex data set. Second, to allow a statistical evaluation it is necessary to perform simulations under each (combination) of conditions multiple times to be able to estimate the variation between two otherwise identical simulations. The computational cost involved in performing replicate runs is the primary reason why the use of rigorous statistical methods for the analysis of molecular simulations has been limited. With advances in computing resources, these problems are slowly being overcome.

The subjects of analysis in this chapter are the effect of the box type used for the simulation and the effect of changes in the force field. The influences of these factors are subtle and multifaceted, which makes it difficult to assess their effects using other methods. As discussed in the previous chapter, the box type used can influence the behaviour of a simulation in many ways. Likewise, changes in the force field may lead to unpredictable differences in the results from simulations. For this reason it is of interest to know how sensitive or insensitive a simulation of a macromolecule is to changes in the force field parameters.

In the first section ANOVA and MANOVA are used to investigate the effect of box shape on the dynamic properties of three proteins simulated under periodic boundary conditions (PBC). The second section presents an assessment of the sensitivity of molecular simulations to changes in the force field on a set of 216 simulations on 36 proteins, using MANOVA and multiple regression.
2 The effect of the box shape on the dynamic properties of proteins simulated under periodic boundary conditions

2.1 The choice of box type and possible consequences thereof

As was shown in Chapter 3, several different box types may be used to perform simulations under periodic boundary conditions. For example, if the simulation box is based on the near-densest lattice packing a significant reduction in the volume of the box can be achieved for non-spherical proteins. However, one consequence of performing a simulation in a tight-fitting periodic unit cell is that rotational degrees of freedom must be constrained to prevent neighbouring periodic images from interacting directly. This can be achieved, e.g. by using the method previously proposed by Amadei et al.\[1\]. Another consequence of performing a simulation in a minimal volume unit cell is that the box shape itself may influence the nature of the motions the system can undergo by restricting sampling in certain directions and thus affect the outcome of the simulation. In addition, the distribution of solvent around the solute may influence the configurations sampled. The notion that the shape of the unit cell may affect the results of a simulation is not new. For example, it has been argued previously that the orientation of the lattice vectors spanning the box can affect the freezing and boiling point of water in a simulation\[2\]. Given that no space-filling shape is completely spherically symmetric, all boxes will affect the outcome of the simulation to a greater or smaller degree. However, these effects and the possible consequences thereof for molecular simulations have largely been ignored.

The development of a method to base the shape of the unit cell on the geometry of the solute together with the application of methods to constrain rotational degrees of freedom without affecting the statistical mechanical ensemble

Figure 4.1: Simulation systems for all proteins and all box types included in the study. Proteins are shown in cartoon representation, simulation boxes as wire frames. Solvent is arranged around the protein according to the shortest distance from the central image in the periodic system, revealing the Voronoi region of that protein in the lattice, defined by the box. Solvent is coloured according to the shortest distance to any of the periodic images of the protein. Colours range from blue/cyan ( < 1 nm ) to red ( > 3 nm ).
provides an opportunity to directly evaluate the effect of the box shape, the degree of solvation and the distribution of the solvent on the dynamic properties of proteins. The aim of this study was to determine if it were possible to detect statistically significant differences between simulations performed in any of four space-filling box types, namely the truncated octahedron, the rhombic dodecahedron, the rectangular box and the NDLP box. The first requirement for such an assessment is the availability of a method to compare sets of simulations. In the present study two approaches were taken. While these are based on different assumptions, they converge in their results. The first approach was based on the investigation of a selected set of instantaneous properties derived from the trajectories, which were then compared using multivariate analysis of variance (MANOVA, see Chapter 2.5). The second approach was based on pair-wise comparisons between the regions of conformational space sampled in the simulations, which were subsequently combined and analyzed using (univariate) analysis of variance (ANOVA, see Chapter 2.4).

As was discussed in Chapter 2, expressing a trajectory in terms of a series of instantaneous properties is an efficient way to reduce the dimensionality of the data set from a molecular simulation. The properties selected for this study were the root mean square deviation (RMSD), the radius of gyration, the total number of intra-protein hydrogen bonds and the number of hydrogen bonds associated with assigned secondary structure elements, the solvent accessible surface and the number of residues involved in α-helix and β-sheet secondary structure elements. Note that many of these properties are expected to have a non-normal distribution. For this reason time averages were taken for each of the properties, rather than instantaneous values, as described in Chapter 2.

The second approach used for the comparison of simulations was aimed at assessing the statistical significance of differences in the directionality of sampling of conformational space between different sets of simulations. This approach is based on principal component analysis. In particular, the similarity in sampling between any two simulations of the same system can be compared on the basis of the correspondence of their eigenvectors. This is done by calculating the root mean square inner product (RMSIP)[3] of all combinations of two sets of eigenvectors obtained from two simulations. Each combination of simulations thus yields a degree of similarity. This similarity index is assumed to be characteristic for a macromolecule simulated over a certain time and is subsequently used for further analysis.

These methods of analysis were used to investigate the effect of the box type on simulations of three different proteins of varying shape and secondary structure content. These proteins were Chymotrypsin Inhibitor II (PDB code 2CI2)[4], the GAG polyprotein M-domain of the rous sarcoma virus (PDB code 1A6S)[5], and Lysozyme (PDB code 1AKI)[6].

\section{Results}

\subsection{Simulations}

A total of 105 simulations were performed on the three different proteins in each of four different box types; a truncated octahedron, a rhombic dodecahedron, a rectangular box and the appropriate NDLP box. For each protein – box type combination simulations were performed both with and without roto-translational constraints, with the exception of the simulations performed in a NDLP box, which were exclusively performed with application of these constraints. The reason for this was that the NDLP box corresponds to a specific orientation of the solute in the box. Allowing rotational motion is likely to lead to direct interactions between periodic images in this type of box.
Each combination (protein – box type – roto-translational constraints) was used to perform five independent simulations. The total simulation time for this study amounted to 1.2 µs. The simulations are summarized in Table 4.1 and the simulation systems are shown in Figure 4.1.

From each trajectory a set of instantaneous properties was derived, the time average of which from 4 ns to 5 ns was determined. Each time average was taken as a single measurement. The resulting property vectors are characteristic of the systems and the simulations. These vectors are shown in Figure 4.2 as a parallel plot.

In addition, for each protein, the root mean square inner products (RMSIPs) were determined for each pair of simulations. The distributions of the RMSIP values are shown in the boxplot in Figure 4.3.

### 2.2.2 Statistical analysis

The first stage in the analysis of the data involved the question whether the different samples, which were obtained under different combinations of conditions, were drawn from the same population. In other words, the first stage involved testing the equality of the means. This was done using three-way fixed effects MANOVA for the sets of descriptive properties and using three-way fixed effects ANOVA for the comparison of the sampling of conformational space. The results for these tests are given in Tables 4.2 and 4.3, respectively.

Table 4.2 shows that the main factors effecting the deviation from the common mean are the protein used in the simulation ($p < 2 \times 10^{-16}$), the box type used for the simulation ($p = 9 \times 10^{-6}$) and the interaction between the protein and the box type ($p = 7 \times 10^{-4}$). The application of rotational constraints did not give rise to statistically significant differences ($p = 0.87$). The statistical significance of interaction effects between the protein and the box-type used indicates that the effect of the box-type on the results depend on the protein used. To quantify this effect, one way MANOVA was performed on...
each of the proteins separately, assessing the box-effect only. These tests revealed that the p-values were 0.81, 0.10 and 0.008 respectively for 2CI2, 1AKI and 1A6S.

Diagnostic analysis of three-way MANOVA results (as well as for the three-way ANOVA results of the next section) showed that the samples were heteroscedastic, i.e. of unequal variances. Since homoscedasticity is a basic assumption of (M)ANOVA, this can in some cases indicate that the results of the analysis may be invalid. However, the differences in variance were found to be due to the protein. Analysis of each of the proteins separately showed that each of these models itself was valid. Separate analysis did not lead to any change in the overall conclusions made and are therefore not presented.

Results similar to those for the MANOVA were obtained for the analysis of variance of collective modes from the simulation (Table 4.3). Again, the main determining factor was the protein ($p < 2 \times 10^{-16}$). Statistically significant differences were also found for combinations of box types ($p = 2 \times 10^{-8}$) and for interaction terms between the protein and the combination of box types ($p = 9 \times 10^{-10}$). Again, no statistically significant effects were observed related to the application or not of rotational constraints ($p = 0.63$).

Similar to the analysis of sets of instantaneous properties, the interaction between the box type and the protein was found to be statistically significant. Therefore, the next step was again to investigate the results of different proteins separately. For the RMSIP samples, the respective p-values from individual one way ANOVA were $4 \times 10^{-9}$, $2 \times 10^{-5}$ and $6 \times 10^{-2}$.

The MANOVA and ANOVA results clearly show that there are differences between the sets of samples obtained for proteins 1A6S and 1AKI. However, this analysis does not indicate which combinations of sets should be regarded different from each other. To investigate which samples caused the rejection of the hypothesis of equal means, a multiple comparison was performed to assess the statistical significance of differences between each pair of samples. This was done using the method of Tukey’s Honest Significant Differences[7] for the
analysis of the RMSIP data and using a Roy union-intersection approach[8, 9] for analysis of instantaneous properties.

The results of the tests comparing the instantaneous properties of box types for 1A6S and 1AKI are given in Tables 4.4 and 4.5. Since an effect of the box type on these properties was effectively ruled out for 2CI2, this protein was not included in the further (post-hoc) analysis. In addition, as only one out of the 35 simulations of 1A6S showed any β-sheet content (0.11 residues averaged over 1 ns), this property was excluded from the analysis of this protein. The results show that the most significant differences observed for 1AKI (p-values 0.002, 0.009 and 0.05) were between simulations performed in a rectangular box with respect to the other box types. For 1AKI the most significant differences were found between simulations performed in a truncated octahedron box and simulations performed in other box types (p-values 0.02, 0.03, 0.11).

Multiple comparisons of the RMSIP sets for 1A6S and 1AKI revealed that for 1AKI seven pairs of RMSIP samples differ from each other on the 95% level, whereas for 1A6S thirteen pairs differ. The results for 1A6S show that the overlap between simulations performed in a rectangular box or between simulations performed in a rectangular and a rhombic dodecahedron box tend to be higher than for simulations from other pairs of box types. For simulations of 1AKI, RMSIP samples from simulations in a truncated octahedron and either a rectangular or a NDLP box are lower than samples involving simulations performed in other box types. These results are in good agreement with the results obtained from the analysis of configurational properties.

**Figure 4.3: Boxplots of overlaps of conformational subspaces sampled.** Root mean square inner products of pairs of simulations from all combinations of box types for each of the proteins included in the study. Horizontal bars indicate the grand mean for each protein.

**Figure 4.4: Differences between average structures.** A. Stereo view of superimposed average structures from simulations of protein 1A6S in a rectangular (blue) and a NDLP (purple) box. B. Stereo view of superimposed average structures from simulations of protein 1AKI in a rectangular (blue) and a truncated octahedron (purple) box.
2.3 Discussion

The application of (M)ANOVA on a number of instantaneous properties obtained from simulation trajectories demonstrates the possibility to use statistical techniques in the evaluation of results from molecular dynamics simulations, and notably in the comparison thereof.

The results presented show that the application of roto-translational constraints using the method of Amadei et al.[1] does not give rise to statistically significant differences in either descriptive properties or the main collective modes, as reflected in the RMSIP values. This is reassuring, but was expected as the method should not affect the statistical mechanical ensemble[1].

The main aim of the present study was to assess the effect of box type on the results of the simulations. The results show that the choice of box type can have a statistically significant effect on the outcome of a simulation. The magnitude of the effect is protein dependent. The box type in effect can act as a constraint on the dynamic behaviour of the solute, restricting the conformational ensemble of the system. It is possible that this is a direct consequence of the orientation of the solute in the

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Wilks $\lambda$</th>
<th>$F$ (appx)</th>
<th>$df_1$</th>
<th>$df_2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>2</td>
<td>1.3·10^{-5}</td>
<td>1877</td>
<td>22</td>
<td>148</td>
<td>2e-16 *</td>
</tr>
<tr>
<td>Box</td>
<td>3</td>
<td>0.37</td>
<td>2.70</td>
<td>33</td>
<td>219</td>
<td>9e-06 *</td>
</tr>
<tr>
<td>RTC</td>
<td>1</td>
<td>0.93</td>
<td>0.54</td>
<td>11</td>
<td>74</td>
<td>0.87</td>
</tr>
<tr>
<td>Protein × Box</td>
<td>6</td>
<td>0.62</td>
<td>1.75</td>
<td>66</td>
<td>401</td>
<td>7e-04 *</td>
</tr>
<tr>
<td>Protein × RTC</td>
<td>2</td>
<td>0.72</td>
<td>1.23</td>
<td>22</td>
<td>148</td>
<td>0.24</td>
</tr>
<tr>
<td>Box × RTC</td>
<td>2</td>
<td>0.72</td>
<td>1.21</td>
<td>22</td>
<td>148</td>
<td>0.25</td>
</tr>
<tr>
<td>Protein × Box × RTC</td>
<td>4</td>
<td>0.59</td>
<td>0.96</td>
<td>44</td>
<td>285</td>
<td>0.55</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>2</td>
<td>11.48</td>
<td>5.74</td>
<td>5815</td>
<td>2e-16 *</td>
</tr>
<tr>
<td>Box</td>
<td>9</td>
<td>0.05</td>
<td>0.01</td>
<td>6.15</td>
<td>2e-08 *</td>
</tr>
<tr>
<td>RTC</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>Protein × Box</td>
<td>18</td>
<td>0.08</td>
<td>0.00</td>
<td>4.55</td>
<td>9e-10 *</td>
</tr>
<tr>
<td>Protein × RTC</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.58</td>
<td>0.56</td>
</tr>
<tr>
<td>Box × RTC</td>
<td>13</td>
<td>0.01</td>
<td>0.01</td>
<td>0.93</td>
<td>0.52</td>
</tr>
<tr>
<td>Protein × Box × RTC</td>
<td>13</td>
<td>0.01</td>
<td>0.00</td>
<td>0.97</td>
<td>0.48</td>
</tr>
<tr>
<td>Residuals</td>
<td>1320</td>
<td>1.30</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: MANOVA results. Results from three-way MANOVA on descriptive properties from simulations, with conditions Protein, box type (Box) and rotational constraints (RTC). Interaction terms are indicated with a multiplication sign. $df$ is the number of degrees of freedom for the specific term, $F$ is the approximate $F$ statistic determined from Wilk's $\lambda$, $df_1$ and $df_2$ are the numerator and denominator degrees of freedom, respectively. $p$ is the probability that the variance due to the source is equal to the residual variance. An asterix marks effects which are statistically significant at the 95% confidence level ($p < 0.05$).

Table 4.3: ANOVA results. Results from three way ANOVA on RMSIP values determined from pairs of simulations, with conditions Protein, box type (Box) and rotational constraints (RTC). Interaction terms are indicated with a multiplication sign. $df$ is the number of degrees of freedom for the specific term, $SS$ is the sum of squares and $MS$ the mean of squares, equal to the sum of squares divided by the degrees of freedom. $F$ is the $F$ statistic, equal to the ratio between the mean of squares and the mean of squares of the residuals, $p$ is the probability that this ratio is equal to one. An asterix marks effects which are statistically significant at the 95% confidence level ($p < 0.05$).
box or its relation to its periodic images. It is also possibly that this arises from indirect effects due to constraints imposed on the solvent distribution around the solute in the specific box type (lattice) chosen. We note that both the analysis of the descriptive properties, which mainly reflect the structural properties of a mean configuration, as well as analysis of the degree of overlap of motions between different simulations gave comparable results.

The protein 2CI2 was largely unaffected by the type of box in which it was simulated, reflecting the intrinsic stability of this protein. For 1A6S and 1AKI the results show a different picture. For 1A6S the simulations performed in a rectangular box were significantly different from simulations performed in other box types. This was reflected in higher average α-helix content and in the number of hydrogen bonds related to secondary structure elements as well as in the hydrophilic solvent accessible surface. On the other hand, the average hydrophobic solvent accessible surface was smaller. The rectangular box favoured the formation of secondary structure and suppressed fluctuations.

The analysis of the RMSIP values suggested that simulations performed in a rectangular box showed more overlap than those in other box types. This also implies restricted motility. The differences observed were in the range of 5-10%. The differences between simulations of 1A6S in a rectangular and an NDLP box are illustrated in Figure 4.4A, which shows a stereo image of superimposed average structures from the simulations. Figure 4.4B gives a similar image of 1AKI, which is discussed later. Figure 4.4A shows that the protein 1A6S was flexible, as indicated by the large spread in the averages obtained from one box type. This corresponds well with the finding that the

<table>
<thead>
<tr>
<th>Property</th>
<th>Average</th>
<th>DH - MB</th>
<th>DH - OH</th>
<th>DH - RC</th>
<th>MB - OH</th>
<th>MB - RC</th>
<th>OH - RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>bbRMSD</td>
<td>0.435</td>
<td>0.036</td>
<td>0.006</td>
<td>0.002</td>
<td>-0.030</td>
<td>-0.033</td>
<td>-0.004</td>
</tr>
<tr>
<td>ssRMSD</td>
<td>0.405</td>
<td>0.047</td>
<td>0.011</td>
<td>0.011</td>
<td>-0.036</td>
<td>-0.036</td>
<td>0.000</td>
</tr>
<tr>
<td>totHbnd</td>
<td>60.167</td>
<td>-3.254</td>
<td>-1.462</td>
<td>0.194</td>
<td>1.792</td>
<td>3.449</td>
<td>1.656</td>
</tr>
<tr>
<td>ssHbnd</td>
<td>31.845</td>
<td>0.098</td>
<td>-0.522</td>
<td>-3.258</td>
<td>-0.620</td>
<td>-3.356</td>
<td>-2.737</td>
</tr>
<tr>
<td>Gyr</td>
<td>1.179</td>
<td>0.010</td>
<td>0.006</td>
<td>0.011</td>
<td>-0.004</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>SASphob</td>
<td>28.606</td>
<td>0.364</td>
<td>0.064</td>
<td>1.134</td>
<td>-0.301</td>
<td>0.770</td>
<td>1.070</td>
</tr>
<tr>
<td>SASphil</td>
<td>24.391</td>
<td>0.378</td>
<td>0.506</td>
<td>-0.871</td>
<td>0.127</td>
<td>-1.249</td>
<td>-1.377</td>
</tr>
<tr>
<td>α-Helix</td>
<td>37.468</td>
<td>2.696</td>
<td>-1.744</td>
<td>-2.585</td>
<td>-4.440</td>
<td>-5.281</td>
<td>-0.842</td>
</tr>
<tr>
<td>p-value</td>
<td>0.16</td>
<td>0.82</td>
<td>0.05</td>
<td>0.10</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 4.5: Averages and contrasts (differences) of the various descriptive properties described in Figure 4.2 for 1AKI.

<table>
<thead>
<tr>
<th>Property</th>
<th>Average</th>
<th>DH - MB</th>
<th>DH - OH</th>
<th>DH - RC</th>
<th>MB - OH</th>
<th>MB - RC</th>
<th>OH - RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>bbRMSD</td>
<td>0.144</td>
<td>0.033</td>
<td>0.026</td>
<td>0.026</td>
<td>-0.007</td>
<td>-0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>ssRMSD</td>
<td>0.115</td>
<td>0.009</td>
<td>0.032</td>
<td>0.006</td>
<td>0.023</td>
<td>-0.004</td>
<td>-0.026</td>
</tr>
<tr>
<td>totHbnd</td>
<td>109.937</td>
<td>-3.364</td>
<td>0.229</td>
<td>-0.944</td>
<td>3.593</td>
<td>2.420</td>
<td>-1.173</td>
</tr>
<tr>
<td>ssHbnd</td>
<td>28.786</td>
<td>-0.242</td>
<td>-0.583</td>
<td>-0.314</td>
<td>-0.341</td>
<td>-0.072</td>
<td>0.268</td>
</tr>
<tr>
<td>Gyr</td>
<td>1.372</td>
<td>0.003</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.004</td>
<td>-0.002</td>
</tr>
<tr>
<td>SASphob</td>
<td>26.485</td>
<td>0.479</td>
<td>0.449</td>
<td>0.546</td>
<td>-0.030</td>
<td>0.068</td>
<td>0.097</td>
</tr>
<tr>
<td>SASphil</td>
<td>40.257</td>
<td>0.108</td>
<td>-0.051</td>
<td>0.050</td>
<td>-0.160</td>
<td>-0.059</td>
<td>0.101</td>
</tr>
<tr>
<td>β-Sheet</td>
<td>9.778</td>
<td>-1.100</td>
<td>0.884</td>
<td>0.480</td>
<td>1.984</td>
<td>1.579</td>
<td>-0.404</td>
</tr>
<tr>
<td>α-Helix</td>
<td>46.288</td>
<td>0.404</td>
<td>-0.327</td>
<td>0.874</td>
<td>-0.731</td>
<td>0.470</td>
<td>1.201</td>
</tr>
<tr>
<td>p-value</td>
<td>0.48</td>
<td>0.11</td>
<td>0.62</td>
<td>0.03</td>
<td>0.70</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>
RMSIP values from simulations performed in the same box types were relatively low. In the figure, it is possible to distinguish consistent differences between the sets. Especially with regards to the terminal regions, the behaviour of the protein seems to depend on the box type.

For lysozyme (1AKI) the results show that simulations performed in a truncated octahedron differed from simulations performed in other box types, with the largest (and statistically significant) differences found between this box and a rectangular or a NDLP box. Again, this difference is reflected in both analysis of descriptive properties and the RMSIP values. For 1AKI most of the properties showed differences less than 5%. The superimposed average structures from simulations in a rectangular and a truncated octahedron box are given in figure 4.4B. This figure shows that the spread is much smaller than for 1A6S. The difference between the sets of averages is found to be small, but consistent, and is in accordance with the results from the statistical analysis. Lysozyme is known to undergo well defined domain motions, which are believed to be functionally important. In this case it is clear that the shape of the box could constrain such motions. This will also be true in other proteins which undergo allosteric changes.

Though the cause of the difference is here suggested to be the different distribution of solvent around the solute in the various box types, one can also think of other sources. For example, a difference in the volume will result in a different solute concentration or, when using counterions, in a different concentration of these. Both effects may influence the solute dynamics. In the present study such effects were implicitly present as part of the difference between the box types. The intention of the work was to compare different box types as they would commonly be simulated. In this regard, we note that no statistically significant differences were detected between the rhombic dodecahedron and NDLP box types with regards to the results obtained. The difference in both the solute and ion concentrations was larger between these two box types than between box types which did yield statistically significant differences suggesting that at least for the solute and ion concentrations involved in this study the effects are not significant.

Another area of difference between the simulations is that, although the protocol used to set up the simulations was identical in all cases, there are significant differences in the water density in specific cases. This artefact is due to slight differences in the overlap between water molecules, the protein and/or the edges of the box due to differences in the orientation of the protein and the shape of the box. This effect can be severe when creating minimal box sizes and is normally corrected by equilibrating at constant pressure. For example, in the case of 1A6S the rectangular box solvent density (±940 kg m$^{-3}$) is significantly lower than the other box types (±970 kg m$^{-3}$). In this case the possibility that density of the solvent may in fact be the dominant factor determining the differences in results cannot be excluded. For protein 2CI2 the differences in solvent density did not lead to differences in simulation results. Finally, differences in solvent density can be excluded as the primary cause for the differences found for 1AKI. Rather, the NDLP box shows a significantly lower density, but gives results similar to those obtained in a rectangular or rhombic dodecahedron box. The truncated octahedron, which has a solvent density similar to those in a rectangular and rhombic dodecahedron box type, is responsible for differences in the simulation results. The solvent densities are included in Table 4.1.

Note that together the results demonstrate that the use of an NDLP box can significantly improve the efficiency of simulations, without the introduction of major artefacts. The present results also show that the use of any box type including a rectangular box type or a truncated octahedron carries some risk of introducing artefacts. It is evident that in a rectangular box rotational motion may lead to direct interactions between periodic images. However, in this study it was found that the effect of the box type on the dynamics was independent of the use of rotational constraints. This shows that the effect of the box type is not caused by direct interactions between periodic images.
2.4 Conclusion

Here an assessment has been presented of the influence of the box type used in a given simulation on a number of properties which can be regarded descriptive for the behaviour of a protein A set of 105 simulations, corresponding to three proteins in four box types and five replicate runs, was analyzed with respect to sampling of conformational space and configurational properties, such as the number of hydrogen bonds, the solvent accessible surface and the number of residues involved in secondary structure elements. It has been shown that a small but statistically significant effect may be attributed to the box type used. The nature and magnitude of the effect are, however, strongly dependent on the protein studied. The study also shows that the use of an NDLP box can increase the efficiency of a simulation without introducing major artefacts. In most cases the optimal box will be the rhombic dodecahedron. Because of its approximate spherical symmetry, this box minimizes effects related to the box shape or the resulting distribution of solvent. In addition, the results demonstrate the possibility of using ANOVA and MANOVA to compare sets of simulations. These methods enable one to evaluate conditions which may have a statistically significant effect on a simulation. Together with multiple comparison tests, ANOVA and MANOVA allow one to make assessments on the effects of a wide range of conditions.

In the next section of this chapter, another application of MANOVA for the comparison of simulations is demonstrated. In addition, the analysis is extended to investigate the origin of the differences observed.

3 The effect of changes in the force field on the dynamic properties of proteins in molecular simulation

3.1 Force fields and force field development

In molecular dynamics simulations the complex interactions between particles in a biological system are approximated by a set of analytical functions and parameters, collectively referred to as the force field. Between different biomolecular force fields (e.g. AMBER[10], CHARMM[11], GROMOS[12, 13] and OPLS[14]) there is a high degree of similarity with regard to the interaction function. However, the values of the parameters, which must be determined empirically, differ considerably. In general, such parameters are obtained by fitting a range of molecular properties of small molecules (e.g. geometric, energetic, dynamic and dielectric) against different sets of quantum mechanical and experimental data. Such fitting can be done in a variety of ways and different emphasis can be given to different aspects of the data.

As the range of molecular properties that can be included in the parameterization is limited, force field parameterization is an underdetermined problem. In this regard, different strategies in force field development may lead to comparable approximations of the overall behaviour of macromolecules. This is especially true for short time scales, during which the effect of differences between force fields does not exceed noise levels. This consideration sheds new light on previous studies aimed at comparing different force fields, such as the recent study from Price and Brooks[15]. They compared
2 ns simulations of a set of three proteins performed with AMBER, CHARMM and OPLS and concluded that none of these force fields showed any consistent trend in the variations for most of the properties monitored. However, from an analytical point of view, it must be considered that differences between force fields will lead to differences between simulation results. The questions are how large are these differences and over what time scales can they be detected?

The effects of changes in the force field on the behaviour of macromolecules in simulation are expected to be subtle at short time scales. To assess such effects, it is necessary to use statistically robust methods for comparing simulation results. Using the framework and methods for such analysis, outlined in Chapter 2, we attempt for the first time to make a robust assessment of the effect of the force field on the outcome of a simulation.

The aim of this study was to compare three different versions of the GROMOS force field. These versions were the original GROMOS96 force field set 43a1[12] and two more recent versions, 53a5[13] and 53a6[13]. In this series, the first number denotes the number of atom types and the last number is the version number. The 53a5 and 53a6 sets are major revisions of the 43a1 set, for which many parameters were redefined. This was done using experimental data regarding the free energy of solvation of small molecule analogues of the amino acids. For the set 53a6 hydration enthalpies were used in addition to free energies. These changes were aimed at improving the partition behaviour of side chains, which has been shown to be wrong in some cases for 43a1[16].

To allow the comparison of the different parameter sets with regards to their effects on the simulation outcome, a large data set was generated. Two independent simulations of five nanoseconds were performed on each of 36 proteins[4, 17-51] in combination with each of the three different parameter sets. The total number of simulations performed in this study thus amounted to 216. The statistical analysis of this data set consisted of three stages. The first two stages correspond to the strategy used to assess the effect of the box type. First, an assessment was made of the similarity or otherwise of results obtained with the three different parameter sets. When this test revealed that at least two of the sets should be considered different, a multiple comparison procedure was used to identify the source(s). Finally, for each pair of parameter sets showing statistically significant differences in the simulation outcomes, regression analysis was performed to understand the difference in terms of specific properties of the protein.

Following the procedure in the previous section, each simulation was characterized by a number of instantaneous properties, averaged over the time period from 4 to 5 ns. For this study the properties used were (1) the root mean square deviation (RMSD) of the average structure against the experimentally determined structure, (2) the average radius of gyration (RG), (3) the number of intramolecular hydrogen bonds corresponding to hydrogen bonds associated with secondary structure elements defined in the experimental structure (HB), (4) the hydrophobic solvent accessible surface (PSA), and (5) the hydrophilic solvent accessible surface (HSA).

To assess the effect of the force field on the outcome of the simulation, the data was normalized to eliminate the contribution from the different proteins as described in the methods section. In this way, a modified linear model could be applied, excluding the protein term and allowing concentration on the force field effect and the effects due to specific interactions between the protein and the force field.
Table 4.6: Proteins included in the assessment of the effect of three force fields on the outcome of simulations.

a) PDBID denotes the PDB entry name. b) The contents of secondary structures follow the definitions in the corresponding PDB files. In the case of 2ci2 and 3ci2, 1pgb and 2gb1, 1shg and 1aey, 1ubi and 1d3z, 1a19 and 1bta, each pair of structures describes a protein produced by X-ray and NMR respectively. For these pairs, the overlaps of the PDB definitions of secondary structures are used. For 2gb1, we used the definition in the PDB file of 1pgb because there is no definition in the PDB file of 2pgb. The definition in the PDB file of 1aey is used for 1shg because of the same reason. Among all the NMR determined structures, only a single minimized average structure was given for 1sap, 1a1z, 1bw6, 1coo, 1lea, 2af8, 2ezh, 2gb1 and 1bta in the PDB. In the remaining cases, where multiple structures have been deposited in the PDB, the first structure in each set was chosen to represent the molecule. Two C-terminal residues of 2ci2 were removed for compatibility with 3ci2. The C-terminal residue of 1aey was removed for compatibility with 1shg. c) The net charge of each structure in aqueous solution at neutral pH.

<table>
<thead>
<tr>
<th>ID</th>
<th>PDBID</th>
<th>Description</th>
<th>Sec. struc.</th>
<th>Res.</th>
<th>Charge</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1vif</td>
<td>dihydrofolate reductase</td>
<td>α / β</td>
<td>60</td>
<td>0</td>
<td>[17]</td>
</tr>
<tr>
<td>2</td>
<td>1tuc</td>
<td>α-Spectrin SRC homology 3 domain, circular permutant</td>
<td>α / β</td>
<td>61</td>
<td>0</td>
<td>[18]</td>
</tr>
<tr>
<td>3</td>
<td>1vcc</td>
<td>N-terminal fragment of vaccinia virus DNA topoisomerase</td>
<td>α / β</td>
<td>77</td>
<td>-1</td>
<td>[19]</td>
</tr>
<tr>
<td>4</td>
<td>1ail</td>
<td>unique RNA-binding domain of the influenza virus NS1 protein</td>
<td>α</td>
<td>70</td>
<td>2</td>
<td>[20]</td>
</tr>
<tr>
<td>5</td>
<td>1cei</td>
<td>colicin E7 immunity protein</td>
<td>α</td>
<td>85</td>
<td>-9</td>
<td>[21]</td>
</tr>
<tr>
<td>6</td>
<td>1ctf</td>
<td>C-terminal domain of the ribosomal protein L7/L12</td>
<td>α / β</td>
<td>68</td>
<td>-2</td>
<td>[22]</td>
</tr>
<tr>
<td>7</td>
<td>1pgx</td>
<td>B2 immunoglobulin-binding domain of streptococcal protein G</td>
<td>α / β</td>
<td>70</td>
<td>-4</td>
<td>[23]</td>
</tr>
<tr>
<td>8</td>
<td>1tlf</td>
<td>N-terminal domain of translational initiation factor IF3</td>
<td>α / β</td>
<td>76</td>
<td>5</td>
<td>[24]</td>
</tr>
<tr>
<td>9</td>
<td>2acy</td>
<td>acyl-phosphatase</td>
<td>α / β</td>
<td>98</td>
<td>1</td>
<td>[25]</td>
</tr>
<tr>
<td>10</td>
<td>2fxb</td>
<td>ferredoxin</td>
<td>α / β</td>
<td>81</td>
<td>-17</td>
<td>[26]</td>
</tr>
<tr>
<td>11</td>
<td>1r69</td>
<td>N-terminal domain of phage 434 repressor</td>
<td>α</td>
<td>63</td>
<td>4</td>
<td>[27]</td>
</tr>
<tr>
<td>12</td>
<td>2ci2</td>
<td>serine proteinase inhibitor CI-2</td>
<td>α / β</td>
<td>63</td>
<td>-1</td>
<td>[4]</td>
</tr>
<tr>
<td>13</td>
<td>1pgb</td>
<td>B1 immunoglobulin-binding domain of streptococcal protein G</td>
<td>α / β</td>
<td>56</td>
<td>-4</td>
<td>[29]</td>
</tr>
<tr>
<td>14</td>
<td>1shg</td>
<td>a Src-homology 3 (SH3) domain</td>
<td>α / β</td>
<td>57</td>
<td>1</td>
<td>[30]</td>
</tr>
<tr>
<td>15</td>
<td>1ubi</td>
<td>ubiquitin</td>
<td>α / β</td>
<td>76</td>
<td>0</td>
<td>[31]</td>
</tr>
<tr>
<td>16</td>
<td>1a19</td>
<td>barstar</td>
<td>α / β</td>
<td>89</td>
<td>-6</td>
<td>[32]</td>
</tr>
<tr>
<td>17</td>
<td>1aoy</td>
<td>DNA-binding domain of arginine repressor</td>
<td>α / β</td>
<td>78</td>
<td>3</td>
<td>[33]</td>
</tr>
<tr>
<td>18</td>
<td>1stu</td>
<td>dsRNA binding domain of Drosophila staufen protein</td>
<td>α / β</td>
<td>68</td>
<td>5</td>
<td>[34]</td>
</tr>
<tr>
<td>19</td>
<td>1sro</td>
<td>S1 RNA binding domain of polyribo-nucleotide phosphorylase, PNPase</td>
<td>α / β</td>
<td>76</td>
<td>1</td>
<td>[51]</td>
</tr>
<tr>
<td>20</td>
<td>1sap</td>
<td>DNA-binding protein Sac7d</td>
<td>α / β</td>
<td>66</td>
<td>6</td>
<td>[35]</td>
</tr>
<tr>
<td>21</td>
<td>1afi</td>
<td>Mercuric ion binding protein, MerP</td>
<td>α / β</td>
<td>72</td>
<td>3</td>
<td>[36]</td>
</tr>
<tr>
<td>22</td>
<td>1bb8</td>
<td>DNA-binding domain of trn916 integrase</td>
<td>α / β</td>
<td>71</td>
<td>5</td>
<td>[37]</td>
</tr>
<tr>
<td>23</td>
<td>2bby</td>
<td>DNA-binding domain of rap30</td>
<td>α / β</td>
<td>69</td>
<td>3</td>
<td>[38]</td>
</tr>
<tr>
<td>24</td>
<td>2frmr</td>
<td>KH1 domain of Fragile X protein</td>
<td>α / β</td>
<td>65</td>
<td>-4</td>
<td>[39]</td>
</tr>
<tr>
<td>25</td>
<td>1a1z</td>
<td>FADD (Mort1) death-effector domain</td>
<td>α</td>
<td>83</td>
<td>-3</td>
<td>[40]</td>
</tr>
<tr>
<td>26</td>
<td>1bw6</td>
<td>DNA-binding domain of centromere binding protein B</td>
<td>α</td>
<td>56</td>
<td>6</td>
<td>[41]</td>
</tr>
<tr>
<td>27</td>
<td>1cco</td>
<td>C-terminal domain of RNA polymerase alpha subunit</td>
<td>α</td>
<td>81</td>
<td>-3</td>
<td>[42]</td>
</tr>
<tr>
<td>28</td>
<td>1lea</td>
<td>LexA repressor DNA binding domain</td>
<td>α / β</td>
<td>72</td>
<td>2</td>
<td>[43]</td>
</tr>
<tr>
<td>29</td>
<td>2af8</td>
<td>actinorhodin polyketide synthase acyl carrier protein</td>
<td>α</td>
<td>86</td>
<td>-12</td>
<td>[44]</td>
</tr>
<tr>
<td>30</td>
<td>2ezh</td>
<td>I gamma subdomain of the Mu end DNA-binding domain of phage Mu transposase</td>
<td>α</td>
<td>65</td>
<td>-2</td>
<td>[45]</td>
</tr>
<tr>
<td>31</td>
<td>3ci2</td>
<td>serine proteinase inhibitor CI-2</td>
<td>α / β</td>
<td>63</td>
<td>0</td>
<td>[46]</td>
</tr>
<tr>
<td>32</td>
<td>2gb1</td>
<td>B1 immunoglobulin-binding domain of streptococcal protein G</td>
<td>α / β</td>
<td>56</td>
<td>-4</td>
<td>[47]</td>
</tr>
<tr>
<td>33</td>
<td>1aey</td>
<td>a Src-homology 3 (SH3) domain</td>
<td>α / β</td>
<td>57</td>
<td>1</td>
<td>[48]</td>
</tr>
<tr>
<td>34</td>
<td>1d3z</td>
<td>ubiquitin</td>
<td>α / β</td>
<td>76</td>
<td>0</td>
<td>[49]</td>
</tr>
<tr>
<td>35</td>
<td>1bta</td>
<td>barstar</td>
<td>α / β</td>
<td>89</td>
<td>-6</td>
<td>[50]</td>
</tr>
</tbody>
</table>
3.2 Results

3.2.1 Simulations

The data set for this study was provided by Dr. Alessandra Villa from the Goethe University in Frankfurt am Main (Germany) and Fan Hao from the University of Groningen. The data consisted of a total of 216 simulations performed on 36 different proteins performed using each of three different versions of the GROMOS force field, namely versions 43a1[12], 53a5[13] and 53a6[13]. Each protein – force field combination was used for two replicate runs of 5 ns. For each protein – force field combination one of the simulations was extended to 10 ns. The longer simulations were used to test for convergence, but were not used for the statistical analysis presented below. The simulations are summarized in Table 4.6 and the proteins included in the assessment are given in Figure 4.5. This figure shows the range of sizes and shapes of the proteins and the differences in secondary structure content.

Figure 4.5: Proteins included in the assessment of the effect of the force field on the simulation outcome. Proteins are referred to by their PDB ID codes.
3.2.2 Statistical analysis

The observation vectors of the instantaneous properties included in the analysis are shown as profiles for each protein under each of the three force fields in Figure 6. Though the distributions of values are similar between the force fields, the figure shows that some proteins have a different profile under the different conditions. The similarity was assessed using MANOVA on the transformed observation vectors, the results of which are given in Table 4.6. From this table it can be seen that the probability that the force fields are equal with regards to the results is negligible ($p < 2.20\times10^{-16}$). It can also be seen that there is a statistically significant contribution due to interactions between the force field and the protein ($p < 2.20\times10^{-16}$).

To identify the force field(s) responsible for rejection of the hypothesis of equal force field effects, contrasts between force fields were tested using the Roy union-intersection method[8, 9]. Differences between the mean vectors of each pair of force fields are given in Table 4.7, together with the $p$-values denoting the probability that the vector is equal to the null vector. The table shows that the difference between the newer force fields, 53a5 and 53a6, is not statistically significant at the 95% confidence level ($p = 0.55$). However, force field 43a1 is found to be different from both 53a5 ($p < 4.5e-34$) and 53a6 ($p < 1.1e-35$).

To gain further insight in the source(s) of the differences the data was subjected to regression analysis with a number of characteristic properties as independent variables. These properties were concerned with charge, secondary structure and the content of specific residue types, such as aromatic and sulphur containing residues.

### Table 4.7: MANOVA table of results

<table>
<thead>
<tr>
<th>Source</th>
<th>dof</th>
<th>Wilk's λ</th>
<th>$F$</th>
<th>df1</th>
<th>df2</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force field</td>
<td>2</td>
<td>0.15</td>
<td>70.3</td>
<td>10</td>
<td>208</td>
<td>&lt; 2.20E-016</td>
</tr>
<tr>
<td>Interaction</td>
<td>105</td>
<td>3.10E-006</td>
<td>11.7</td>
<td>525</td>
<td>525</td>
<td>&lt; 2.20E-016</td>
</tr>
<tr>
<td>Residuals</td>
<td>108</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4.8: Averages and contrasts (differences) of the relative deviations of descriptive properties for the different force fields

Properties included are the root mean square deviation (RMSD), the radius of gyration (RGYR), the number of intramolecular hydrogen bonds corresponding to hydrogen bonds associated with secondary structure elements defined in the experimental structure (HBND), the hydrophobic solvent accessible surface (PHOB), and the hydrophilic solvent accessible surface (PHIL). The $p$-values were determined using the Roy union-intersection approach and give the probability that the difference vector is equal to the null vector.

<table>
<thead>
<tr>
<th>Property</th>
<th>Average</th>
<th>43a1 - 53a5</th>
<th>43a1 - 53a6</th>
<th>53a5 - 53a6</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSD</td>
<td>0.194</td>
<td>-0.023</td>
<td>-0.041</td>
<td>-0.018</td>
<td>4.47E-034</td>
</tr>
<tr>
<td>RGYR</td>
<td>0.008</td>
<td>-0.038</td>
<td>-0.034</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>HBND</td>
<td>-0.221</td>
<td>-0.009</td>
<td>-0.004</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>PHOB</td>
<td>0.040</td>
<td>-0.063</td>
<td>-0.051</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>PHIL</td>
<td>0.047</td>
<td>-0.136</td>
<td>-0.135</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>4.47E-034</td>
<td>1.14E-035</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.1: Results for multi-linear regressions of differences of selected properties between force fields 53a5 and 43a1 against a number of properties characterizing the proteins. Each line in the table summarizes the results of a multiple regression of the five measured properties against the regressor on the left. Results which were found to be statistically significant are shown in bold face type. The table shows how the differences between the relative deviations from the references for force fields 53a5 and 43a1 are related to properties characterizing the proteins. A positive slope (β) indicates that the differences become larger with increasing values of the regressor. A negative value for β indicates that the differences become smaller with increasing values of the regressor.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>R (adj)</th>
<th>β</th>
<th>p</th>
<th>R^2</th>
<th>R^2 adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSD</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Rg</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>MCC</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N_res</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N_pos.cha</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>N_neg.cha</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Charge</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>net</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>pos.cha</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>neg.cha</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>α-helix%</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>β-sheet%</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>loop%</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Sulfur%</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Amide%</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Notes:
- The regressors include N_res (the number of residues in a structure), N_pos.cha (the number of positively charged residues), N_neg.cha (the number of negatively charged residues, N_net (the number of charged residues at neutral pH), Charge (the net charge), CHG charged (the number of positively charged residues in a structure), α-helix (the percentage of residues in an α-helix), β-sheet (the percentage of residues in a β-sheet), loop (the percentage of residues in a loop), Sulfur (the percentage of Met and Cys residues in a structure), Amide (the percentage of residues containing aromatic groups).
- The structural properties include RMSD (the positional root mean square deviation of backbone atoms in the secondary structure elements), Rg (the radius of gyration), HSA (the hydrophilic surface area), PSA (hydrophobic surface area), HB (the number of mainchain hydrogen bonds in secondary structure elements).
- β denotes the slope of the regression line.
- p denotes the probability that the slope equals 0 (no regression relation).
- R (adj) denotes the adjusted correlation coefficient.
Table 4.10. Results for multi-linear regressions of differences of selected properties between force fields 53a6 and 43a1 against a number of properties characterizing the proteins. Each line in the table summarizes the results of a multiple regression of the five measured properties against the regressor on the left. Results which were found to be statistically significant are shown in bold face type. The table shows how the differences between the relative deviations from the references for force fields 53a6 and 43a1 are related to properties characterizing the proteins. A positive slope (β) indicates that the differences become larger with increasing values of the regressor. A negative value for β indicates that the difference becomes smaller with increasing values of the regressor.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>RMSD</th>
<th>Rg</th>
<th>HSA</th>
<th>PSA</th>
<th>HB</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_res</td>
<td>0.002</td>
<td>0.134</td>
<td>0.037</td>
<td>-0.001</td>
<td>0.038</td>
<td>0.095</td>
</tr>
<tr>
<td>N_cha_res</td>
<td>0.008</td>
<td>0.018</td>
<td>0.130</td>
<td>0.001</td>
<td>0.188</td>
<td>0.023</td>
</tr>
<tr>
<td>Charge_net</td>
<td>-0.008</td>
<td>0.007</td>
<td>0.175</td>
<td>-0.002</td>
<td>0.004</td>
<td>0.195</td>
</tr>
<tr>
<td>N_pos_cha</td>
<td>-0.002</td>
<td>0.660</td>
<td>-0.024</td>
<td>-0.002</td>
<td>0.220</td>
<td>0.016</td>
</tr>
<tr>
<td>N_neg_cha</td>
<td>0.016</td>
<td>0.000</td>
<td>0.307</td>
<td>0.003</td>
<td>0.003</td>
<td>0.207</td>
</tr>
<tr>
<td>N_cha_res_%</td>
<td>0.398</td>
<td>0.205</td>
<td>0.019</td>
<td>-0.028</td>
<td>0.710</td>
<td>-0.025</td>
</tr>
<tr>
<td>N_pos_res</td>
<td>-0.452</td>
<td>0.206</td>
<td>0.019</td>
<td>-0.179</td>
<td>0.028</td>
<td>0.109</td>
</tr>
<tr>
<td>α-helix</td>
<td>0.415</td>
<td>0.004</td>
<td>0.195</td>
<td>0.087</td>
<td>0.012</td>
<td>0.149</td>
</tr>
<tr>
<td>β-sheet</td>
<td>0.002</td>
<td>0.224</td>
<td>0.015</td>
<td>-0.001</td>
<td>0.008</td>
<td>0.168</td>
</tr>
<tr>
<td>loop</td>
<td>0.000</td>
<td>0.845</td>
<td>-0.028</td>
<td>0.001</td>
<td>0.093</td>
<td>0.054</td>
</tr>
<tr>
<td>α-helix %</td>
<td>0.110</td>
<td>0.155</td>
<td>0.031</td>
<td>0.031</td>
<td>0.084</td>
<td>0.058</td>
</tr>
<tr>
<td>β-sheet %</td>
<td>-0.122</td>
<td>0.187</td>
<td>0.023</td>
<td>-0.057</td>
<td>0.006</td>
<td>0.180</td>
</tr>
<tr>
<td>loop %</td>
<td>-0.073</td>
<td>0.586</td>
<td>-0.020</td>
<td>0.026</td>
<td>0.400</td>
<td>-0.008</td>
</tr>
<tr>
<td>Sulfur %</td>
<td>1.261</td>
<td>0.186</td>
<td>0.023</td>
<td>0.386</td>
<td>0.881</td>
<td>0.060</td>
</tr>
<tr>
<td>Aromatic %</td>
<td>0.569</td>
<td>0.161</td>
<td>0.029</td>
<td>0.123</td>
<td>0.196</td>
<td>0.021</td>
</tr>
<tr>
<td>Amide %</td>
<td>-0.991</td>
<td>0.049</td>
<td>0.083</td>
<td>-0.173</td>
<td>0.149</td>
<td>0.033</td>
</tr>
</tbody>
</table>

a) The regressors include N_res (the number of residues in a structure), N_cha_res (the number of charged residues at neutral pH), Charge_net (the net charge), N_neg_cha (the number of negatively charged residues), N_cha_res % (the percentage of charged residues in a structure), α-helix (the number of residues belonging to α-helical structures), Sulfur% (the percentage of Met and Cys residues in a structure), Aromatic% (the percentage of residues containing aromatic groups), Amide% (the percentage of Asn and Gln residues in a structure).

b) The structural properties include RMSD (the positional root mean square deviation of backbone atoms in the secondary structure elements), Rg (the radius of gyration), HSA (the hydrophilic surface area), PSA (hydrophobic surface area), HB (the number of mainchain hydrogen bonds in secondary structure elements).

c) β denotes the slope of the regression line.

d) p denotes the probability that the slope equals 0 (no regression relation).

e) $R^2_{adj}$ denotes the adjusted correlation coefficient.

f) MCC is the multi-linear correlation coefficient and indicates to which extent the joint observed values are correlated to the regressor on the left.
The results for the regression analysis are shown in Tables 4.9 and 4.10. In these tables each line summarizes the results of performing multiple regression analyses on the five descriptive properties obtained from the simulations against the single independent variable indicated on the left. Here, \( \beta \) indicates the estimated slope of the regression relation, \( p \) denotes the probability that this slope is equal to zero (no regression relation) and \( R^2_{\text{adj}} \) denotes the adjusted correlation coefficient. MCC denotes the multiple correlation coefficient, indicating how well the five properties together correlate to the characteristic property of the protein. Statistically significant regression relations (\( \beta \neq 0 \) at 95% confidence level) are indicated in typeface bold. Note that the dependent variable is the difference in the dynamic property (RMSD, RG, HSA, PSA, HB) between 53a5 and 43a1 or between 53a6 and 43a1. Thus, a positive relation indicates that the measured value for a property is increasing for the 53a series parameter sets relative to 43a1 with increasing values for the characteristic property of the protein. Likewise, a negative relation indicates that the measured property is increasing for 43a1 relative to the 53 series parameter sets with increasing fixed variate. The tables show that the highest multiple correlation coefficients are obtained for characteristic properties related to the charges on the protein. Notably, for both sets 53a5 and 53a6 the difference with set 43a1 correlates well to the net charge in the protein. In the case of 53a5 it is the best regressor included (MCC = 0.410). For 53a6 the net charge is the second best regressor (MCC = 0.484) and only the absolute number of negative charges yields a higher MCC (0.520). From the tables 4.9 and 4.10 it can also be seen that the secondary structure content of a protein is a lesser determinant for deviations between the two newer parameter sets 53a5 and 53a5 and the set 43a1. For 53a5 the only relations found to be statistically significant were between the difference in the radius of gyration and the absolute number (\( \beta = 0.320, p = 0.019 \)) as well as the proportion (\( \beta = 0.498, p = 0.044 \)) of residues not involved in secondary structure elements and the proportion of residues involved \( \beta \)-sheets (\( \beta = -0.531, p = 0.040 \)). In the case of 53a6 the correlations with regards to the secondary structure content are generally lower than for 53a5, and though the regression slopes are statistically significant non-zero for the radius of gyration, the low correlation coefficients indicate that there is no serious relation between this property and any of the proteins characteristic properties related to secondary structure. Of the residue types which were most extensively reparameterized, namely the sulphur, aromatic and amide containing amino acids, the only significant correlation was between the number of hydrogen bonds and the proportion of sulphur containing residues (53a5: \( \beta = -0.519, p = 0.005 \) and 53a6: \( \beta = -1.650, p = 0.005 \)). Given the definition of the number of hydrogen bonds as the number corresponding to hydrogen bonds present in the reference structure, these negative slopes indicate that the number of “native” hydrogen bonds observed decrease for either 53a5 or 53a6 as compared to 43a1 for proteins with increasing amounts of sulphur containing residues.

### 3.3 Discussion

This study had two aims. The first aim was to determine if the different versions of the Gromos force field led to detectable differences in the results of the simulations on the time scales studied. A second aim was to further validate the method proposed in the previous section for the comparison of molecular simulations. Clearly the study shows that it is possible to use the methods proposed to detect consistent differences between simulations performed with different parameter definitions for non-bonded interactions. The procedure used suggests that on a 5 ns time scale two out of three force fields do not show consistent differences. Taken together, this shows that the proposed method is suitable for statistical assessment of similarity and dissimilarity of molecular simulations.
The main aim of this study, however, was the assessment of the effect of the force field on the results obtained from molecular simulations of proteins. Differences between force fields essentially give rise to different energy landscapes and consequently affect how the system evolves in time. However, a complex energy landscape can be sufficiently approximated in different ways such that the actual differences between force fields do not necessarily lead to detectable differences in the behaviour on the time scales accessible. With this in mind, the objective of the current assessment is to determine whether different parameter definitions lead to detectable differences in the outcome of simulations of proteins. In other words, how sensitive are molecular simulations to changes in the force field?

The results show that changes in the force field can lead to detectable changes in protein behaviour in silico, even on relatively short time scales. In that sense, the protein is not ‘robust’. Differences were found are between the two new parameter sets 53a5 and 53a6 and the old parameter set 43a1. The inclusion of the enthalpy of hydration for the parameterization of 53a6 does not lead to statistically significant differences between this parameter set and 53a5 on the current time scales and with the protocol used. This does not mean that there are no differences in behaviour, but that possible differences are negligible on the time scale of several nanoseconds.

Regression analysis showed that the differences between the old and the new parameter sets seem to be related to the presence of charged residues in the protein. Apparently, the new force fields tend to destabilize the structure, leading to higher RMSD values, larger radii of gyration and larger solvent accessible surfaces. This effect is more profound in proteins with more charged residues. Other characteristic properties are less predictive for the difference, e.g. the secondary structure content of a protein is only a poor determinant for explanation of the difference observed between parameter sets 53a5/53a6 and 43a1.

Interestingly, the parameters for charged residues have largely remained unaltered going from 43a1 to 53a5/53a6. A possible reason for the finding that the differences between the force field sets are best explained by the presence and number of charged residues in the protein is that the charges of polar residues are enhanced relative to these in the 43a1 force field. Changes made in the parameters of one residue or atom type can affect properties of other residues or atom types. In this way, choosing wrong parameters for one residue type can lead to the assignment of wrong parameters for another residue type, in effect causing compensation of errors in overall behaviour. The reparameterization of uncharged residue types for force field sets 53a5 and 53a6 may have unveiled problems in the parameters for the charged residues. It is well possible that these problems were masked in force field set 43a1 through the presence of compensating errors.

Note, the best determinants were expected to be directly related to the amino acids which were most extensively revised (aromatic, amide and sulphur-containing residues). However, no significant relations between the proportion of residues of these types and the difference observed between the parameter sets was found, except for a tendency to obtain lower numbers of hydrogen bonds in proteins with a higher content of sulphur containing residues.

### 3.4 Conclusions

In this section the results were presented of an assessment to the influence of the force field parameters on the outcome of simulations. To this purpose a series of 216 simulations on 36 different proteins was performed in combination with each of three selected parameter sets, with a total sampling of 1.08 µs. The results show that the new parameter sets, 53a5 and 53a6, yield results which are indistinguishable, but simulations performed with each of these two sets are different...
from simulations performed with the parameter set 43a1. The differences seem to be mostly related to the presence of charged residues, which were not reparameterized going from the 43 to the 53 series force field sets.

4 Overall conclusions

The two studies presented in this chapter demonstrate the application of statistical methods, notably ANOVA and MANOVA, to the purpose of comparing simulations. In both studies the methods are shown to be able to reveal differences between external conditions applied to the simulations, whether this condition is the box type in which the simulation is performed or the force field used to describe the interactions between the particles in the system. In addition, the methods also show that certain external conditions, such as the use of roto-translational constraints or the inclusion or not of enthalpies of solvation for the parameterization of non-bonded interactions, do not give rise to detectable differences. The classification of both effecting and non-effecting external conditions provides further support for the applicability of the proposed methods.

A basic assumption underlying the analysis in the studies presented is that external effects may cause a redistribution of conformational densities, which can be expressed in terms of descriptive properties and atomic fluctuations. Such redistribution can also be seen as the mechanism underlying allosteric effects. Thus, the methods of analysis of molecular simulations proposed here may also be applicable in a much broader context, where the binding of a ligand to a target protein for example could be treated as an external factor exerting its influence on the system.

5 Methods

5.1 The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions

5.1.1 Simulations

Three proteins of varying shape and secondary structure content were used in this study: Chymotrypsin Inhibitor II (PDB code 2CI2)[4], the GAG polyprotein M-domain of the rous sarcoma virus (PDB code 1A6S)[5], and Lysozyme (PDB code 1AKI)[6]. Starting structures were taken from the Protein Data Bank[52]. For each protein five simulations were performed in each of four box types (rhombic dodecahedron, truncated octahedron, rectangular, and the appropriate NDLP box, as determined using the method presented in Section 3.2.1) either with or without the application of roto-translational constraints. Note that simulations in a NDLP box were performed exclusively with rotational constraints. This is because the NDLP box corresponds to a specific orientation of the solute in the box, as explained in Chapter 3. Allowing rotational motion would rapidly and inevitably lead to direct interactions between periodic images. Simulation boxes were chosen such that the minimal distance to the wall was 1.0 nm in all box types, resulting in a minimal distance between periodic images of at least 2.0 nm. Production runs for 1A6S and 2CI2 were 10 ns each. Production runs for 1AKI were 20 ns for simulations performed in a NDLP box with rotational constraints and in a rhombic dodecahedron, a truncated octahedron or a rectangular box without constraints. Simulations of 1AKI performed in the latter three box types with rotational constraints were 5 ns in length. Note, while the longer simulations (20ns 1AKI and 10ns 1A6S and 2CI2) were
The Effect of an External Condition on the Simulation Outcome

required to allow sufficient convergence of principal components [53], in terms of the analysis of structural properties it was found that 5 ns was sufficient to allow the statistical assessment. As this was the maximum time available for all systems, for consistency, only the analysis based on simulations of 4-5 ns is presented.

All simulations were performed using a modified version of the Gromacs 3.1.4 [54-56] simulation package, in which the roto-translational constraint algorithm of Amadei et al. [1] had been implemented. The interatomic interactions were described using the GROMOS96 43a2 united atom force field [12, 57]. Water molecules were modelled explicitly using the Simple Point Charge (SPC) model [58]. The protonation state of ionizable groups was chosen appropriate for pH 7.0. Counterions were added to neutralize the net charge of the system. Non bonded interactions were evaluated using a twin range cut off of 0.9 and 1.4 nm. Interactions within the shorter range cut off were evaluated at every step whereas interactions within the longer range cut off were evaluated every 10 steps. To correct for the neglect of electrostatic interactions beyond the longer range cut off, a Reaction Field (RF) correction [59] was used with $\varepsilon_{RF} = 78.0$. In all simulations the system was kept at a constant temperature of 300K by applying a Berendsen thermostat [60]. Protein and solvent molecules were independently coupled to the heat bath with a coupling time of 0.1 ps. Simulations were performed at constant volume. Note, the volume was set such that the pressure was on average approximately 1 atmosphere ensuring that the water density was the same in all cases. The time step used for the integration of the equations of motion was 0.002 ps. The bond lengths and angle of the water molecules were constrained using the SETTLE algorithm [61]. Bond lengths within the protein were constrained using the SHAKE [62] algorithm. Starting velocities were randomly assigned from a Maxwellian distribution with different random seeds for each of the simulations. Each of the systems was equilibrated for 10 ps before starting production runs.

5.1.2 Analysis

The effect of the different box shapes and volumes on the result of the simulations was assessed by analyzing the similarity or otherwise of a number of properties derived from the trajectories and of the collective fluctuations in the system.

Properties that were included in the analysis were: (1) the root mean square deviation (RMSD) of the average structure against the experimentally determined structure (bbRMSD), (2) the RMSD, excluding flexible regions of the protein (i.e. residues not assigned to secondary structure elements; ssRMSD), (3) the average radius of gyration (Rgyr), (4) the number of intramolecular hydrogen bonds (totHbnd), (5) the number of hydrogen bonds associated with secondary structure elements defined in the experimental structure (ssHbnd), (6) the hydrophobic solvent accessible surface (PSA), (7) the hydrophilic solvent accessible surface (HSA), (8) the number of residues involved in β-sheet formation (β-Sheet) and (9) the number of residues involved in α-helical elements (α-Helix). These properties were determined using routines available in the Gromacs package and are described in the documentation of Gromacs [56]. These properties were determined over a time window of 1 ns, from 4-5 ns for all trajectories.

The set of instantaneous properties considered were expressed as observation vectors for all simulations and were used for a three-way fixed-effects multivariate analysis of variance (MANOVA) with provision for interactions between applied conditions. These conditions, or factors, were the protein, the type of box used for the simulation and the application of roto-translational constraints.

The first step in the eigenvector analysis was the generation of the positional covariance matrices
for each of the individual simulations. These were diagonalized to obtain the eigenvectors and corresponding eigenvalues. The ten eigenvectors with the largest associated eigenvalues were used to compare pairs of simulations. This was achieved by determining the root mean square inner product (RMSIP)[3] of the selected sets of eigenvectors from both simulations, which is a measure of the overlap of the subspaces sampled in these simulations. The approach used was to determine all pairwise overlaps. These were separated into two classes: from simulations performed under identical conditions, but with different starting configuration, and from simulations performed under different box conditions.

The assumption was made that RMSIP values for simulations performed under identical conditions were approximately normally distributed with a specific mean \( \mu \) and variance \( \sigma^2 \). The mean and the variance were estimated from multiple simulations. If \( a \) and \( b \) denote two sets of simulations performed under different conditions, all pairwise RMSIP values within each set were determined (sets A and B) from which \( \mu_A, \mu_B, \sigma^2_A, \) and \( \sigma^2_B \) could be estimated. All pairwise RMSIPs between simulations from set \( a \) and set \( b \) were then determined (set A×B). This provided the estimates for the mean RMSIP \( \mu_{A:B} \) and the corresponding variance \( \sigma^2_{A:B} \) between any two simulations performed under the different conditions. Samples of RMSIPs calculated from pairs of simulations performed under the same treatments will be called type I or within treatment RMSIPs, samples obtained from pairs of simulations from two different treatments will be called type II or between treatments RMSIPs.

The resulting RMSIP values were analyzed using three-way fixed-effects ANOVA with provision for interactions between conditions. The conditions included in the model were the same as for the MANOVA. Two sets of simulations, \( a \) and \( b \), were regarded to be equal (indistinguishable) with respect to the collective motions governing the dynamics if the corresponding RMSIP samples \( A \) and \( B \) were equal and were both equal to \( A \times B \). If two sets of RMSIPs were found to be equal at the 95% confidence interval, the corresponding simulations were considered to give rise to the same essential motions and any differences in the simulation conditions were regarded as not significantly affecting the sampling of conformational space on the time scales probed.

### 5.1.3 Statistical analysis

All statistical analysis was performed using the program R, a language and environment for statistical computing (http://cran.r-project.org/)[63].

The following model was used for three-way fixed effects MANOVA with provision for interaction is:

\[
x_{ijkh} = \mu_h + \alpha_{ih} + \tau_{jh} + \eta_{ijk} + \epsilon_{ijkh}
\]  

(4.1)

Here, \( x_{ijkh} \) is the \( k \)th observation of the \( h \)th response. \( \mu_h \) is a general-level effect for the \( h \)th response, comparable to the grand mean of all observations. \( \alpha_{ih} \) and \( \tau_{jh} \) are deviations of the \( h \)th response from \( \mu_h \) due to effects of external conditions \( i \) and \( j \). \( \eta_{ijk} \) is a deviation from that response due to an interaction between external conditions \( i \) and \( j \), and \( \epsilon_{ijkh} \) is a random variable term of the \( h \)th response with mean zero. For the evaluation of the hypothesis of equality of the mean vectors Wilk’s lambda was used (Chapter 2.5).

If the ANOVA and MANOVA results suggested a significant difference between at least two sets of simulations, multiple comparisons were made to investigate the source(s) of these differences. For the univariate case this was done using Tukey’s Honest Significant Differences (HSD) method[7], multivariate observations were further analyzed using the Roy union-intersection approach[8, 9].
5.2 The effect of changes in the force field on the dynamic properties of proteins in molecular simulation

5.2.1 Simulations

The simulations using the GROMOS 43a1[12] force field were performed by Dr. Fan Hao from the University of Groningen, The Netherlands. Simulations using GROMOS force fields 53a5[13] and 53a6[13] were performed by Dr. Alessandra Villa from the Goethe University in Frankfurt am Main, Germany. Extraction of structural properties from the simulations was performed by the original researchers. For the present study their data was combined into one large data set.

The 36 starting structures corresponding to 31 different proteins were taken from the Protein Data Bank (PDB)[52]. For each protein two simulations were performed with each of the three parameter sets of the GROMOS force field, 43a1, 53a5 and 53a6. Production runs had a simulation time of 5 ns of which the last 1 ns was used for analysis.

Proteins were solvated in a rectangular box with a minimal distance between adjacent periodic images of 1.0 nm. Water molecules were modelled explicitly using the Simple Point Charge (SPC) model[58]. The protonation state of ionizable groups was chosen appropriate for pH 7.0. Arginine and lysine residues were protonated, whereas aspartate and glutamate residues were unprotonated. Histidine residues were neutral and in all cases the tautomer with the δ1 nitrogen atom protonated was chosen. The N-termini were modelled as NH$_3^+$ or l-pro-NH$_2^+$, whereas the C-terminus was modelled as a carboxylate. Note that no counterions were added and that in some cases the simulation systems had a net charge.

All simulations were performed using the GROMACS 3.1 simulation package[54-56]. Non-bonded interactions were evaluated using a twin-range cut-off. Interactions within the shorter-range cut-off (0.9 nm) were evaluated every step, whereas interactions within the longer-range cut-off (1.4 nm) were evaluated every 10 steps, together with the pair-list. To correct for the neglect of electrostatic interactions beyond the 1.4 nm cut-off, a reaction-field correction[59] was used with $\varepsilon_{RF} = 78$. The temperature of the system was kept at 300K by coupling to an external heat-bath using a Berendsen thermostat[60]. The protein and the solvent were independently coupled to the heat-bath with a coupling time of 0.1 ps. The pressure was kept constant at 1 bar by the application of a Berendsen barostat[60] with an isothermal compressibility of $4.5 \times 10^{-5}$ bar$^{-1}$ and a coupling time of 0.5 ps. The time step for the integration of the equations of motion was 0.002 ps. The bond lengths and the angle of the SPC water molecules were constrained using the SETTLE algorithm[61]. Bond lengths within the protein were constrained using the LINCS algorithm. Before starting a production run, a steepest descent energy minimization was performed to get rid of bad contacts. In addition short runs of 10 ps were performed with position restraints on the protein and at a temperature of 250K to gently relax the system. Production runs consisted of 5 ns simulations at a temperature of 300K. For each combination of protein and force field parameter set one of the two replicate simulations was extended to 10 ns.

5.2.2 Analysis

Analysis of the trajectories was performed with tools from the GROMACS package[54-56]. The root mean square deviation was determined for the average structure obtained from 4-5 ns against the starting structure after a least-squares fit. The other properties included in the study were the solvent-accessible surface area, the radius of gyration and the number of hydrogen bonds. These properties were averaged over the time span from 4-5 ns. The solvent-accessible surface area was determined
numercially according to the method of Eisenhaber. The atomic radii used for the calculation were 0.16 nm for carbon, 0.13 nm for oxygen, 0.14 nm for nitrogen, 0.20 nm for sulphur and 0.10 nm for hydrogen atoms. The atomic radius used for the solvent molecules was 0.14 nm. Atoms with a charge within the range from -0.2e to +0.2e were considered hydrophobic. The definitions used for the classification of hydrogen bonds were a maximum hydrogen-acceptor distance of 0.25 nm and a minimum donor-hydrogen-acceptor angle of 60 degrees. Only hydrogen bonds originally present in the reference structures were counted.

5.2.3 Statistical analysis

The set of structural properties considered were expressed as observation vectors for all simulations. The general linear model assumed to describe the original obervations is given by

\[ x_{ijkl} = \mu_{il} + \tau_{jl} + \gamma_{ijl} + \epsilon_{ijkl} \]  (4.2)

Here \( x_{ijkl} \) is the \( k \)th observation of property \( l \) for protein \( i \), under condition (force field) \( j \). \( \mu_{il} \) is a general location parameter for property \( l \) specific to protein \( i \) and is comparable to the protein mean for that property. \( \tau_{jl} \) is a deviation from \( \mu_{il} \) due to the force field used and \( \gamma_{ijl} \) is a deviation specific to the combination of the force field used and the protein. \( \epsilon_{ijkl} \) is a random error with mean zero and a variance specific to the protein.

The main objective was to test for differences between simulations, attributable to a difference in the force field used. Differences between proteins were of no interest in the present study. It should also be noted that the protein specific error terms disallow a direct comparison using an analysis of variance approach, as this method requires the residual variances to be equal (requirement of homoscedasticity). For these reasons the data is transformed in such a way that the protein effect is removed from the model, i.e. the results were normalized with regards to protein effects. The normalization was done by scaling the results according to

\[ x'_{ijkl} = \frac{(x_{ijkl} - a_{il})}{b_{il}} \]  (4.3)

with

\[ a_{il} = \bar{x}_{il} = \frac{1}{cr} \sum_{j=0}^{c} \sum_{k=0}^{r} x_{ijkl} \]  (4.4)

and

\[ b_{il} = \frac{1}{\sqrt{\text{var}(\epsilon_{ijkl})}} = \left[ \frac{1}{cr - 1} \sum_{j=0}^{c} \sum_{k=0}^{r} (x_{ijkl} - \bar{x}_{il})^2 \right]^{-\frac{1}{2}} \]  (4.5)

Here, \( c \) is the number of force fields (=3) and \( r \) denotes the number of simulations per protein – force field combination (=2). The linear model describing the transformed results is then given by

\[ x_{ij} = \mu + \tau_{j} + \gamma_{ij} + \epsilon_{ij} \]  (4.6)

This model was the basis for a fixed-effects multivariate analysis of variance (MANOVA) with provision for force field effects and interactions between protein and force field. Wilk's lambda was used as the test statistic for the evaluation of the equality of the sets.

If the MANOVA results suggested a significant difference between at least two sets of simulations,
multiple comparisons were made to investigate the source(s) of these differences. This post-hoc analysis was done using the Roy union-intersection approach (Appendix A).

All statistical analysis was performed using the program R, a language and environment for statistical computing (http://www.r-project.org/).

6 Appendix A: Multiple comparisons using the Roy union intersection method in R

The program R is a powerful, freely available open-source package for the statistical processing and analysis of data. The standard tests, including ANOVA and MANOVA applied in this work are available in the standard distribution. Other tests, such as the multiple comparisons using Tukey's Honest Significant Differences, were performed using the package multcomp, acquired from the R website. However, the program at present does not provide a method to test multiple contrasts for the multivariate ANOVA model. For this reason an implementation is given for tests of contrasts using a method based on the Roy union-intersection method. This implementation, which was kindly provided by Dr. Yves Rosseel from the University of Ghent, Belgium, is given as a function definition below, which can be loaded (sourced) from within R.

The function is a general method to test multivariate linear hypotheses of the type $LBM = K$, where $L$ is a contrast vector and $B$ is the parameter matrix of the linear model underlying the MANOVA. $M$ is usually the identity matrix and $K$ the null matrix, both of appropriate dimensions. It should be noted that R reparameterizes linear models, such that the intercept includes the first level of each of the factors. Given the linear model for observations obtained with several levels of two different conditions.

Figure 4.6: Profiles of descriptive properties for different proteins in simulation, using different force field parameters. Properties are shown as deviations from the properties in the experimentally known structures and are given as time averages from 4 – 5 ns. Properties included are the hydrophobic solvent accessible surface (PSA), the hydrophilic solvent accessible surface (HAS), the number of hydrogen bonds (HB), the radius of gyration (Rg) and the RMSD. The different colours and line styles indicate the different proteins.
reparameterization in R leads to the equation for the intercept
\[ \mu_0 = \mu + \tau_i + \nu_j + \eta_{ij} \] (4.8)
and effects are represented as deviations from this intercept:
\[ \tau^*_i = \tau_i - \tau \] (4.9)
Reducing the model to a one way analysis of variance for explanatory purposes, a contrast of the parameters \( \tau_i \) is defined as any linear function
\[ \sum_{i=1}^{k} c_i \tau_i \] (4.10)
where the coefficients have the property
\[ \sum_{i=1}^{k} c_i = 0 \] (4.11)
These contrasts can be represented by vector notation as:
\[ (c_1, c_2, \ldots, c_k)(\mu_0, \tau^*_2, \ldots, \tau^*_k)' \] (4.12)
From this it can be understood that the contrast for treatments 1 and 2 from a condition with four levels in R is given by \((0, -1, 0, 0)\), whereas a contrast between treatments 3 and 4 is given by \((0, 0, 1, -1)\).

The function to evaluate contrasts of this form in R is given by the following code, which was written by Dr. Yves Rosseel:

```r
mlh <- function(fit, L, M)
{
  if (!inherits(fit, "maov"))
    stop("object must be of class \"manova\" or \"maov\"")

  if (is.null(dim(L)))
    L <- t(L)

  rss.qr <- qr(crossprod(fit$residuals %*% M))

  X <- as.matrix(model.matrix(fit))
  B <- as.matrix(fit$coef)

  LB <- L %*% B
  LXXL <- as.matrix(solve(L %*% solve(t(X) %*% X) %*% t(L)))
  H <- t(M) %*% t(LB) %*% LXXL %*% LB %*% M

  eig <- Re(eigen(qr.coef(rss.qr, H), symmetric=FALSE)$values)

  q <- nrow(L)
}```
df.res <- fit$df.residual

test <- prod(1 / (1 + eig))
p <- length(eig)
tmp1 <- df.res - 0.5 * (p - q + 1)
tmp2 <- (p * q - 2) / 4
tmp3 <- p^2 + q^2 - 5
tmp3 <- if (tmp3 > 0) sqrt(((p * q)^2 - 4) / tmp3) else 1

wilks <- test
df1 <- p * q
df2 <- tmp1 * tmp3 - 2 * tmp2
F <- ((test^(-1 / tmp3) - 1) * df2) / df1
Prob <- pf(F, df1, df2, lower.tail=FALSE)

out <- list(wilks=wilks, F=F, df1=df1, df2=df2, Prob=Prob)
out

7 References

13. Oostenbrink, C., et al., A biomolecular force field based on the free enthalpy of hydration and solvation:


