Supporting Information

Determination of Structural Ensembles of Flexible Molecules in Solution from NMR Data Undergoing Spin Diffusion

Francesca Vasile\textsuperscript{1} and Guido Tiana\textsuperscript{2}\textsuperscript{*}

\textsuperscript{1}Department of Chemistry, Università degli Studi di Milano, via Golgi 19, 20133 Milano, Italy
\textsuperscript{2}Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milano, Italy.

*guido.tiana@unimi.it
S1. Obtaining a force field from the principle of maximum entropy

If one already has a *a priori* model for the system which provides a probability distribution \( p_0(r) \), a way of defining the minimum-biased model is that of minimizing the Kullback-Leibler divergence between the two distributions

\[
D_{KL}[p||p_0] = \sum_r p(r) \log \frac{p(r)}{p_0(r)}.
\]  

(S1)

that can be constrained so that the averages performed with the model \( \langle I_{ij} \rangle = \sum_r I_{ij}(r)p(r) \) match the experimental data \( I_{ij}^{exp} \) with the method of Lagrange multipliers. One has then to minimize the function

\[
\sum_r p(r) \log \frac{p(r)}{p_0(r)} - \sum_{ij} \lambda_{ij} \langle \sum_r p(r) I_{ij}(r) \rangle - I_{ij}^{exp}
\]

obtaining

\[
p(r) = p_0(r) \exp\left[ -\sum_{ij} \lambda_{ij} I_{ij}(r) \right],
\]

(S3)

where \( Z' \) is the normalizing partition function. If \( p \) and \( p_0 \) are equilibrium distributions following Boltzmann statistics

\[
p(r) = \frac{1}{Z} \exp \left[ -\frac{U(r)}{kT} \right],
\]

(S4)

and

\[
p_0(r) = \frac{1}{Z_0} \exp \left[ -\frac{U_0(r)}{kT} \right],
\]

(S5)

with \( Z = \sum_r \exp \left[ -U(r)/kT \right] \), then Eq. (S11) becomes

\[
U(r) = U_0(r) + kT \sum_{ij} \lambda_{ij} I_{ij}(r).
\]

(S6)

Consequently, one can correct any potential \( U_0 \) to match the experimental data in a minimally-biased way, with a correction that has the same functional dependence on the conformation of the system as that of the forward model. Here two problems arise. First, the correcting potential obtained substituting Eq. (1) in Eq. (S6) is not two-body, but the interaction between pairs of atoms depend on the position of all other atoms, like in polarizable force fields. Moreover, the relation

\[
\frac{\partial \log Z}{\partial \lambda_{ij}} = I_{ij}^{exp}
\]

(S7)

that relates the Lagrange multipliers, which now act as parameters of the potential, to the experimental NOEs is an implicit equation involving the partition function, and thus is of little practical use.

The former problem is solved assuming a two-body functional form, as usually done by standard force fields (but still calculating the forward model, as described in the next Section, with the many-body Eq. (1)), that is using

\[
U(r) = U_0(r) + \sum_{ij} \frac{\lambda_{ij}}{(r_{ij}-\bar{r}_{ij})^6},
\]

(S8)

where \( \lambda'_{ij} \) includes the Lagrange multipliers, the thermal contribution \( kT \) and the contribution of the forward model, assumed constant. The numerical values of \( \lambda'_{ij} \) are calculated minimizing the \( \chi^2 \) between the NOE intensities calculated from the simulation and the experimental ones, as described in the next Section, thus avoiding the use of Eq. (S7). The use of a correction to the potential that has the same functional form of the forward model guarantees that we are searching among parametrizations that minimize the Kullback-Leibler divergence; the minimization of the \( \chi^2 \) guarantees that the Lagrange multipliers are satisfying the constraints on the thermodynamic averages. Since the expression in Eq. (S2) is convex, it displays a unique solution; thus, if one is able
to find a distribution $p(r)$ that minimizes the Kullback-Leibler divergence and matches the experimental averages, this will be the only possible solution.

S2. Computational details

The initial potential $U_0(r)$ is the Amber 03 force field in implicit solvent, modelled with GBSA. At each iteration, replica-exchange MD simulations at four temperatures (T=300K, 330K, 370K and 420K) are performed to sample the conformational space for 50 ns each replica, recording 5000 conformations at 300K. Calculations are carried out with a tailor-made code calling Gromacs 4.5.5 for the replica-exchange part.

At the end of the sampling, the parameters of the interactions between the pairs of hydrogen atoms and of heavy atoms bound to the hydrogen atoms and the involved in the NOE signals undergo 500 random updates, accepting only the changes that decrease the $\chi^2$ between the calculated and the experimental NOE intensities. The same kind of update is applied to the parameters of the Ryckaert-Bellemans torsional potential associated with the observed $i$-($i+4$) NOE, that were added to the potential with all parameters set to zero at the beginning of the simulation. A new MD simulation is then started with the new potential, starting from the last conformation of the previous run. The simulation is explicit solvent used as comparison are performed in 990 TIP/3P water molecules. The code is freely available at https://github.com/guidotiana/ffoptim.

S3. NOESY experiments

All NMR spectra were registered on Bruker Avance III 400 MHz using a solution 3.5 mM of the peptide in water (with 10% D2O). The water suppression was carried out by excitation sculpting. The assignment was performed through one- and two-dimensional $^1$H-NMR spectra by standard method. The assignment of the molecule is reported in Table S3. For the conformational analysis three independent NOESY spectra (with 32 scans and 256 increments) were collected using a mixing time of 700 ms (Figure S1) and the intensities of cross peaks were measured. To obtain the NOE build up curves, NOESY spectra with 100, 200, 400, 700 and 1000 ms were used (Figure 1). The cross-peaks intensities were calculated and plotted as function of mixing time. Diagonal intensities are extrapolated at zero mixing time with a least-square cubic fit.

S4. Calculation of the rotational correlation time

In Table S1 we report the intensities obtained for six crosspeaks of the peptide. The value of $\tau_c$, although being of the same order of magnitude, is different for each pair, because it is affected not only by the rotational motion of the molecule as a whole, but also the internal motion of the single groups, since the molecule is flexible. Not being able to find the value of $\tau_c$ for each pair, because not all crosspeaks can be detected at such a low mixing time, we used for the forward model the average of the values reported in Table S2, that is $\tau_c = 135$ ps.

S5. Approximation of the spectral density function

The main approximation used in the present approach is that of using the expression of a rigid molecule which tumbles isotropically to calculate the spectral density in Eq. (6). This approximation allowed us to perform replica-exchange simulation instead of fixed-temperature MD simulations that would be required to calculate explicitly the spectral density.
To check the validity of this approximation, we performed two independent 10 µs simulations from two different initial conformations with the (same) optimized potential, and back-calculated the correlation functions responsible for the NOEs, as described in ref. 26.

The first result is that the tumbling of the molecule is found to be approximately isotropic, with a correlation time which is independent on the initial conformation and has the same order of magnitude of the experimental one (see Fig. S8 in the SI). The rotational correlation time \( \tau_c \approx 25 \) ps is smaller than those obtained experimentally (cf. Sect. S4 above). This is not completely unexpected due to the fact that simulations are done in implicit solvent.

Moreover, we calculated the NOE intensity without approximating the spectral density as in Eq. (6) but calculating it explicitly from

\[
C_{\text{alt}}(\tau) = \left( \frac{P_2(\cos \alpha_{tt+\tau})}{r^3(\tau) r^3(t+\tau)} \right),
\]

(S9)

Where \( P_2(x) = \frac{3}{2} x^2 - \frac{1}{2} \) is the second-order Legendre polynomial and \( \alpha_{tt+\tau} \) is the angle between the inter-spin vectors at the two times. The spectral function \( J(\omega) \) is the calculated as Fourier transform of Eq. (S9). It is important to note that this approach requires a real, fixed-temperature MD simulation to calculate the time correlation functions, while cannot be used with trajectories obtained from a replica-exchange simulation, as those used in the main part of our work.

The correlation functions \( C_{\text{alt}}(\tau) \) for the two contacts 60-74 and 48-82, taken respectively as representative of short-range and long-range contacts, are displayed in Fig. S9. Also, the value of the NOEs calculated by Eq. (S9) is indicated in the figure. The overall correlation function \( C_{\text{alt}}(\tau) \) (cf. Fig. S9 in the SI), whose Fourier transform gives the spectral density \( J(\omega) \) that controls the transition rates of Eqs. (3-4), is approximately single-exponential, suggesting that the flexibility of the molecule does not necessarily invalidate Eq. (6)

The two correlation functions calculated for each contact in the two independent simulations appear quite different from each other. This difference can be better appreciated noticing that the associated NOE intensities (estimated in the small-\( \tau_m \) limit, neglecting spin diffusion) are markedly different from each other, more than typical error bars. This difference suggests that plain MD simulations at fixed temperature, at variance with replica-exchange simulations, cannot reach thermodynamic equilibrium and thus are affected by a strong dependence on the initial conditions. In other words, although the calculation of NOEs from a plain MD simulation using Eq. (S9) is in principle more correct than using the approximation of Eq. (6) with a replica-exchange simulation, in practice a plain MD simulation is not able to calculate the thermodynamic average that appears in Eq. (S9).

Nonetheless, we tested the approximation of the spectral density on the plain-MD data, at least to verify internal consistency. For this purpose, we postulated that the plain-MD trajectories described the thermodynamic equilibrium of the system (something which is clearly not true, see above), and calculated the NOE intensities both with the true spectral density of Eq. (S9) and with the approximated one of Eq. (6), using in the latter case the rotational correlation time obtained from Fig. S8. The calculated values are listed in Table S4. The difference between the NOE intensities calculated from the exact and the approximated spectral densities is of the order of 10\%, comparable with the experimental error bars.

The NOEs calculated in this way are approximately one order of magnitude smaller than the experimental ones. The reason for that lies in the low value of the simulated rotational correlation time. In fact, calculating the NOEs from the same plain-MD simulations with the spectral density of Eq. (6), but using now the experimental value of the rotational correlation time gives values (cf. Table S4) that display the same order of magnitude of the experimental ones (still being different in the two independent simulations and different from the experimental values, for the reason explained above).
### Supporting Tables

<table>
<thead>
<tr>
<th>atom $i$</th>
<th>atom $j$</th>
<th>average $I_{ij}$</th>
<th>$\sigma_{ij}$</th>
<th>48-50</th>
<th>82,87</th>
<th>11298</th>
<th>1476</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>23</td>
<td>26400</td>
<td>1734</td>
<td>60</td>
<td>58</td>
<td>11966</td>
<td>379</td>
</tr>
<tr>
<td>25,44</td>
<td>42</td>
<td>20600</td>
<td>2224</td>
<td>60</td>
<td>72</td>
<td>24333</td>
<td>2190</td>
</tr>
<tr>
<td>27</td>
<td>23</td>
<td>23200</td>
<td>692</td>
<td>58</td>
<td>62.63</td>
<td>13500</td>
<td>1768</td>
</tr>
<tr>
<td>36-38</td>
<td>23</td>
<td>12100</td>
<td>1209</td>
<td>74.95</td>
<td>93</td>
<td>24900</td>
<td>1650</td>
</tr>
<tr>
<td>36-38</td>
<td>80,89</td>
<td>6732</td>
<td>1536</td>
<td>80.89</td>
<td>74</td>
<td>28666</td>
<td>4700</td>
</tr>
<tr>
<td>44</td>
<td>58</td>
<td>21200</td>
<td>1907</td>
<td>80.89</td>
<td>76.77</td>
<td>44500</td>
<td>4500</td>
</tr>
<tr>
<td>46</td>
<td>58</td>
<td>7833</td>
<td>929</td>
<td>76.77</td>
<td>82.87</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>46</td>
<td>42</td>
<td>13033</td>
<td>1069</td>
<td>76.77</td>
<td>72</td>
<td>13400</td>
<td>2300</td>
</tr>
<tr>
<td>48-50</td>
<td>58</td>
<td>9021</td>
<td>944</td>
<td>72</td>
<td>80.89</td>
<td>10046</td>
<td>4030</td>
</tr>
<tr>
<td>48-50</td>
<td>42</td>
<td>22633</td>
<td>2253</td>
<td>93</td>
<td>97.98</td>
<td>30232</td>
<td>2400</td>
</tr>
</tbody>
</table>

**Table S1:** The average and standard deviation over three replicate experiments of the NOE intensities recorded at a mixing time of 700 ms. The overlapping spins are indicated in an aggregated way.

<table>
<thead>
<tr>
<th>Atom $i$</th>
<th>Id</th>
<th>Atom $j$</th>
<th>Id</th>
<th>$I_{ij}^{NOE}$</th>
<th>$I_{ij}^{ROE}$</th>
<th>$r$</th>
<th>$\tau_c$ [ps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA1</td>
<td>6</td>
<td>HA2</td>
<td>25</td>
<td>89</td>
<td>61</td>
<td>0.27</td>
<td>181</td>
</tr>
<tr>
<td>HA3</td>
<td>44</td>
<td>HA4</td>
<td>58</td>
<td>55</td>
<td>55</td>
<td>0.18</td>
<td>212</td>
</tr>
<tr>
<td>HA5</td>
<td>25</td>
<td>HB2</td>
<td>46</td>
<td>97</td>
<td>97</td>
<td>0.7</td>
<td>83</td>
</tr>
<tr>
<td>HB5</td>
<td>68</td>
<td>HB7</td>
<td>72</td>
<td>97</td>
<td>97</td>
<td>0.7</td>
<td>83</td>
</tr>
<tr>
<td>HB6</td>
<td>93</td>
<td>HB7</td>
<td>97</td>
<td>97</td>
<td>97</td>
<td>0.7</td>
<td>83</td>
</tr>
</tbody>
</table>

**Table S2:** The crosspeak intensities for selected pairs in the NOESY and in the ROESY experiments at 50 ns. Their ratio $r$ allows one to calculate the rotational time $\tau_c$ through Eq. (S24).
<table>
<thead>
<tr>
<th></th>
<th>NH</th>
<th>$H\alpha$</th>
<th>$H\beta$</th>
<th>$H\gamma$</th>
<th>$H\delta$</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>3.93 (6)</td>
<td>1.59</td>
<td>1.48</td>
<td>0.82-077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.50 (23)</td>
<td>4.13 (25)</td>
<td>1.69</td>
<td>1.4-1.09 (27)</td>
<td>0.75 (36-38)</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>8.22 (42)</td>
<td>3.95 (44)</td>
<td>1.81 (46)</td>
<td>0.66 (48-50 and 52-54)</td>
<td></td>
<td>NH$_2$ 6.73-7.46</td>
</tr>
<tr>
<td>N</td>
<td>8.32 (58)</td>
<td>4.57 (60)</td>
<td>2.59 (62,63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>8.04 (72)</td>
<td>4.43 (74)</td>
<td>2.79-2.94 (76,77)</td>
<td></td>
<td>7.04 (80,89)</td>
<td>6.74 (82,87)</td>
</tr>
<tr>
<td>L</td>
<td>8.12 (95)</td>
<td>4.24 (95)</td>
<td>1.53 (97,98)</td>
<td>1.42</td>
<td>0.78-0.81</td>
<td></td>
</tr>
</tbody>
</table>

**Table S3**: NMR assignment of peptide LIVNYL in water. The number reported in bracket correspond to the numeric id of some of the hydrogens used to discuss the results.

<table>
<thead>
<tr>
<th></th>
<th>$I$, true $J(\omega)$ from MD</th>
<th>$I$, rigid $J(\omega)$ from MD</th>
<th>$I$, rigid $J(\omega)$ from MD with experimental $\tau_c$</th>
<th>$I$, rigid $J(\omega)$ from rep-ex</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-72 (replicate 1)</td>
<td>2.83×10$^3$</td>
<td>3.01×10$^3$</td>
<td>1.82×10$^4$</td>
<td>2.52×10$^4$</td>
</tr>
<tr>
<td>60-72 (replicate 2)</td>
<td>5.82×10$^3$</td>
<td>6.13×10$^3$</td>
<td>4.21×10$^4$</td>
<td></td>
</tr>
<tr>
<td>48-82 (replicate 1)</td>
<td>9.42×10$^2$</td>
<td>9.96×10$^2$</td>
<td>5.74×10$^3$</td>
<td>2.57×10$^3$</td>
</tr>
<tr>
<td>48-82 (replicate 2)</td>
<td>2.14×10$^3$</td>
<td>2.84×10$^3$</td>
<td>1.84×10$^4$</td>
<td></td>
</tr>
</tbody>
</table>

**Table S4**: In the second column, the NOE intensities calculated from the plain MD simulations for a short-range pair of protons (60-72) and for a long range one (48-82) using the exact spectral density and Eq. (S9); in the third column, the NOEs calculated with the rigid-body spectral density of Eq. (6) and the isotropic rotational correlation time obtained from the simulation (cf. Fig. S8); in the fifth column, the NOEs calculated from the rigid-body spectral density using the experimental correlation time (cf. Sect. S4); in the fifth column, the NOEs calculated from the replica-exchange simulations with the algorithm suggested in the main text.
Supporting Figures

**Figure S1**: Finger print region of NOESY spectrum (mixing time = 700ms) of peptide LIVNYL in water. The spectrum is recorded on a Bruker Avance III operating at 400 MHz at 298K.
Figure S2: (above) The comparison between the NOE intensities calculated with Eq. (1) from the ensemble of conformation obtained by restraint minimization with Dyana, and the experimental ones (giving a $\chi^2=73.2$). (Below) The distribution of RMSD between each pair of conformations among the 20 that minimize restraint violations (conformations shown in the inset).
Figure S3: The $\chi^2$ between the calculated and the experimental NOEs as a function of the number of iterations of the optimization algorithms for two independent runs, starting from the Amber03 force field in implicit solvent.
Figure S4: Scatter plots of the \( C_6 \) parameters defining the Lennard-Jones potential (above) and of the parameters \( C_n \) defining the Ryckaert-Bellemans torsional potentials (below) obtained in two different optimizations. The correlation coefficients are 0.37 and 0.18, respectively.
Figure S5: (above) Comparison between calculated and experimental NOE intensity for a simulation in which, in addition to the optimization of the observed NOEs, uNOE are optimized to be zero, with error bars set to a conventional value of 1000. Only observed NOEs are plotted here. The overall $\chi^2$ is 0.26, while that restricted to observed NOEs is 4.1. (Below) The distribution of uNOE obtained by the simulation with standard Amber in explicit solvent and for the optimized potential. The dashed line indicates the width of the error bar. There are 8 erroneously observed (i.e., above 1000) uNOE in the Amber03 simulation and 10 in the optimized simulation.
Figure S6: Comparison of the NOE intensities of Fig. 1 with those obtained disregarding spin diffusion, that is using as forward model for each pair of spin the quantity $\langle 1/d^6 \rangle$, where $d$ is the interatomic distance. The $\chi^2$ between the calculated and the experimental NOEs is 16.4.

Figure S7: The free energy of the system interacting with the potential optimized without taking into account spin diffusion.
Figure S8: The rotational correlation times along the three Cartesian axes (solid colored lines) calculated from two fixed-temperature MD simulations starting from two initial conditions (upper and lower panel, respectively) with the (same) optimal potential found by the iterative algorithm. Dashed lines indicate the single-exponential fits used to obtain the correlation times indicated in the plots.
Figure S9: The overall correlation time (solid curve) calculated with Eq. (26) of ref. 25 in the two replicated fixed-temperature MD simulations for contact 60-74 and 48-82. The dashed lines indicate the single-exponential fit. To appreciate the difference between the various behaviors, in each plot is indicated the NOE intensity $I$ that would be obtained averaging the full correlation function on the two trajectories.