Mark: 5.5/6

Very good report with a few mistakes - see comments below for more detail.

Structure Refinement of ed GCN4p16-31 with unstrained, instantaneous restrains and time-averaged restrains simulations

Guido Putignano[a], Lorenzo Tarricone[a]

[a] *Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland*

Abstract

Structure refinement refers to the process of improving the accuracy and precision of a molecular model to better align with experimental data. This paper presents an in-depth analysis of peptide structure refinement using molecular dynamics (MD) simulations, emphasizing the role of distance restraints derived from nuclear Overhauser enhancement (NOE) observed in nuclear magnetic resonance (NMR) experiments. Employing the GROMOS program, we conducted three distinct MD simulations: a standard unrestrained simulation and two simulations implementing NOE-derived distance restraints, applied either instantaneously or on a time-averaged basis.

The study primarily aims to evaluate how these restraints and their application methods influence the simulation outcomes, particularly in terms of compatibility with experimental NMR data. We explored three critical areas: (i) understanding the mechanisms and implications of instantaneous restraints (IR) and time-averaged restraints (TAR); (ii) setting up MD simulations with these restraining techniques; and (iii) analyzing the simulation results against experimental NMR data, focusing on both NOE-derived distances and 3J-coupling constants.

Our findings provide valuable insights into the peptide behavior under different restraint scenarios. We observed that the choice of restraint method significantly affects the structural conformations in the simulations. This study not only furthers our understanding of peptide dynamics in the context of structural biology but also offers practical guidance for researchers in selecting and implementing restraining methods in MD simulations for more accurate peptide structure predictions.

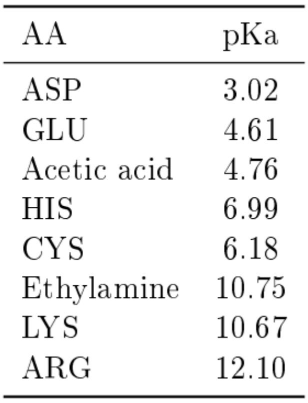
Keywords: Unstrained, Instantaneous Restrains, Time-Averaged Restrains, Structure Refinement.

1. Introduction

A. Peptide has a +1e charge. His is not protonated - only one of the ring Ns (the NE2) has a H and Lys is protonated, therefore the pH must be between 6 and 10.

Question A: The residues are numbered 16-31 because the peptide sequence starts from residue 16 (Asn) and ends at residue 31 (Gly). The N-terminal part of the peptide sequence (residues 1-15) is not included or relevant for the study, as we are studying the C-terminal coiled-coil trigger sequence of the yeast transcriptional activator GCN4, therefore we are ignoring the first residues. "Ac-" stands for acetyl, indicating that the N-terminus has an acetyl group. "-NH2" indicates an amino group at the C-terminus. When considering the question of the charge, there were different opinions on that, given some elements of the text that were not fully consistent with the make_top_protein.arg file. If we follow what is written in the text: "The His residue is protonated at NE2, the Arg and Lys side chains are protonated with charge +e", we can consider the overall charge of the peptide in the simulations is +2e. This is because His19 Arg25 and Lys27 and Lys28 residues are protonated and have a charge of +1e each, while Glu21 and Glu23 have a charge of -1e. This means that the overall charge of the peptide is positive. The protonation states of histidine (His), arginine (Arg), and lysine (Lys) suggest that the simulation corresponds to a low pH environment where these residues are likely to be positively charged. If we refer to the table from ex3 1, we can notice that we need a pH lower than the lowest of pKa values in order for the three groups to be all three protonated. Therefore we need a pH ≤ 6.99. In order to have the negative charge on Glu we need a value ≥ 4.61

In any case, file on Gromos may have some points of inconsistency. The text doesn't explicitly mention the addition of counter-ions. However, looking at the argument file of the function make top, we see that no counter-ions have been added. The initial configuration was equilibrated at constant volume during the 100 ps equilibration, as there is no mention of constant pressure. The force constant for positional restraints was gradually reduced, and the temperature was raised from 60 to 278 K during this equilibration. The box volume was kept constant at 89.925 nm^3. Thus, we performed an NVT ensemble and kept the volume constant.



AA	pKa
ASP	3.02
GLU	4.61
Acetic acid	4.76
HIS	6.99
CYS	6.18
Ethylamine	10.75
LYS	10.67
ARG	12.10

Figure 1: pKa for different residues

Amino Acid	pKa Value
Glu (Glutamic Acid)	4.61
His (Histidine)	6.99
Lys (Lysine)	10.67
Arg (Arginine)	12.10

Table 1: pKa Values of Selected Amino Acids

Question B: B. Correct reasoning up to the restrained distances. There, a correction term, given in Table 1 in the script, has to be applied.

#IDR1	JDR1	KDR1	LDR1	ICDR	IDR2	JDR2	KDR2	LDR2	ICDR	R0	W0	NRAH
171	0	0	0	0	164	162	165	168	1	0.35	1	1

Table 2: First line of noe.dsr file

NOESPEC	1	16	HN	15	HA	1.8	0.0	1.7

Table 3: First line of the NOElist.exp

R0 = can be obtained through the sum of the first and last column in the noe.dsr file. Example of the first line 1.7 + 1.8 = 3.5 / 10. There is a scaling factor given a difference in unit measure between files. The numbers 16 HN 15 NA are also present in both files. They are commented on and in the case of the first row correspond to **16 GLY H HN 1 15 VAL CA HA**. I also know that when considering the atom number 16, I can check the `$coord/peptide_gch.cnf$` file to find the molecule it's referring to. In this case, 16 is referring to GLY

Nr. 2

The line in the NOElist.exp file is:
2 16 HN 15 HN 1.8 0.0 3.2
The corresponding line in the now.dsr file is:

#	1	1	16	GLY	H	HN	1	15	VAL	H	HN		
	171	0	0	0	0	163	0	0	0	0	0.5	1	1

We can clearly assess a correspondence between the values.

Nr. 46
The line in the NOElist.exp file is:
46 5 HN 4 HA 1.8 0.0 1.7
The corresponding line in the now.dsr file is:

```
#   1   1   7   GLU    H    HN    1   4   LEU   CA   HA       78
0   0   0   0    49    47    50   54   1   0.5    1    1
```

The line in the NOElist.exp file is:
64 6 HA 9 HB@ 1.8 0.0 1.7
The corresponding line in the now.dsr file is:

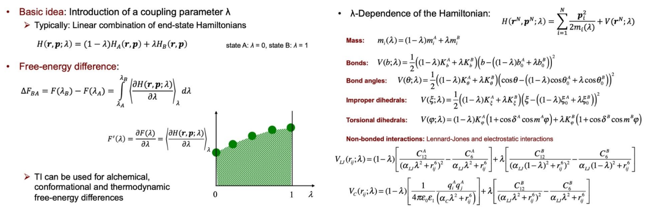
```
 #    1    1   6   ASN   CA   HA   1   9   ALA   CB   HB@      68
66   69   75   1    98   97    0   0   5   0.45   1     1
```

Here R0 should be 0.35 but it's 0.45

Nr. 179 The line in the NOElist.exp file is:
179 8 HG@@ 5 HA 1.8 0.0 3.2
The corresponding line is:

```
 #    1    1   8   VAL   CB   HG@@    1   5   GLU    CA   HA      90
91   92    0   6    58   56    59    64   1   0.72    1    1
```

Also in this case there is a different R0 value. One reason may be the virtual atoms that make the results different from what is expected. In this case, I provided the question to the Teaching Assistant and the answer was what I wrote here.



- Basic idea: Introduction of a coupling parameter λ
  - Typically: Linear combination of end-state Hamiltonians

$$H(\boldsymbol{r},\boldsymbol{p};\lambda) = (1-\lambda)H_A(\boldsymbol{r},\boldsymbol{p}) + \lambda H_B(\boldsymbol{r},\boldsymbol{p})$$

state A: λ = 0, state B: λ = 1

- Free-energy difference:

$$\Delta F_{BA} = F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda$$

$$F'(\lambda) = \frac{\partial F(\lambda)}{\partial \lambda} = \left\langle \frac{\partial H(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial \lambda} \right\rangle_{\lambda}$$

  - TI can be used for alchemical, conformational and thermodynamic free-energy differences

- λ-Dependence of the Hamiltonian:

$$H(\boldsymbol{r}^N,\boldsymbol{p}^N;\lambda) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m_i(\lambda)} + V(\boldsymbol{r}^N;\lambda)$$

Mass: $m_i(\lambda) = (1-\lambda)m_i^A + \lambda m_i^B$

Bonds: $V(b;\lambda) = \frac{1}{2}\left((1-\lambda)K_b^A + \lambda K_b^B\right)\left(b - \left((1-\lambda)b_0^A + \lambda b_0^B\right)\right)^2$

Bond angles: $V(\theta;\lambda) = \frac{1}{2}\left((1-\lambda)K_\theta^A + \lambda K_\theta^B\right)\left(\cos\theta - \left((1-\lambda)\cos\theta_0^A + \lambda\cos\theta_0^B\right)\right)^2$

Improper dihedrals: $V(\xi;\lambda) = \frac{1}{2}\left((1-\lambda)K_\xi^A + \lambda K_\xi^B\right)\left(\xi - \left((1-\lambda)\xi_0^A + \lambda\xi_0^B\right)\right)^2$

Torsional dihedrals: $V(\varphi;\lambda) = (1-\lambda)K_\varphi^A\left(1+\cos\delta^A\cos m^A\varphi\right) + \lambda K_\varphi^B\left(1+\cos\delta^B\cos m^B\varphi\right)$

Non-bonded interactions: Lennard-Jones and electrostatic interactions

$$V_{LJ}(r_{ij};\lambda) = (1-\lambda)\left[\frac{C_{12}^A}{(\alpha_{LJ}\lambda^2 + r_{ij}^6)^2} - \frac{C_6^A}{\alpha_{LJ}\lambda^2 + r_{ij}^6}\right] + \lambda\left[\frac{C_{12}^B}{(\alpha_{LJ}(1-\lambda)^2 + r_{ij}^6)^2} - \frac{C_6^B}{\alpha_{LJ}(1-\lambda)^2 + r_{ij}^6}\right]$$

$$V_C(r_{ij};\lambda) = (1-\lambda)\left[\frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i^A q_j^A}{\left(\alpha_C\lambda^2 + r_{ij}^6\right)}\right] + \lambda\left[\frac{C_{12}^B}{(\alpha_{LJ}\lambda^2 + r_{ij}^6)^2} - \frac{C_6^B}{\alpha_{LJ}\lambda^2 + r_{ij}^6}\right]$$

Figure 2: Hamiltonian and its dependencies

2. Analysis

C. Thorough analysis.

C: When considering the Root Mean Square Deviation (RMSD), we found several differences between our simulations and the given simulations. In order to make the description of this exercise easier, we will divide the simulation into ”our trajectories” and ”provided trajectories” as presented in the figure. Before describing the main differences between the trajectories, it may be helpful to understand the different types of simulations.

Table 4: Types of Molecular Dynamics (MD) Simulations

Simulation Type	Description
Standard (Unrestrained) Simulation	In a standard MD simulation, there are no restraints applied to the atoms or groups of atoms. This means that all parts of the molecule or system are free to move according to the forces calculated based on the molecular mechanics force fields. The simulation mimics the natural dynamics of the molecule in solution, assuming the force field accurately represents the physical forces present.
Instantaneous Restraints (IR)	When applying instantaneous restraints in an MD simulation, distance restraints derived from NOE data are imposed on the atoms for every time step of the simulation. This means that at every point in the simulation, the distances between certain atom pairs are kept within the ranges determined by the experimental NOE data. These restraints are applied instantaneously, ensuring that the molecule conforms to the NOE-derived distances at all times. This can be useful for keeping parts of the molecule in a conformation that is known to be experimentally valid.
Time-Averaged Restraints (TAR)	In contrast to IR, time-averaged restraints are applied over a certain period. The NOE-derived distance restraints are enforced not at every instant but as an average over a defined window of time. This allows for some fluctuations around the NOE-derived distances but requires that the average structure over time complies with the NOE data. This method acknowledges that molecules are dynamic and that NOE data represents an average of structures present in solution. TAR can help to stabilize the simulated structure around an experimentally determined average without preventing local dynamics and fluctuations that might occur naturally.

Root Mean Square Deviation (RMSD) serves as a key indicator, measuring the average distance between the atoms of superposed protein structures. Theoretically, applying restraints is expected to result in a lower RMSD, thereby maintaining the structural integrity of the protein closer to its initial conformation.

In our study's trajectory analysis, the unrestrained simulation displayed the highest RMSD. This was followed in ascending order of RMSD by the Time-Averaged Restraint and the Instantaneous Restraint simulations. This trend aligns with our initial hypothesis, suggesting that restraints, particularly Instantaneous Restraints, effectively stabilize the protein structure near its initial state. Instantaneous Restraints are designed to continuously enforce conformational consistency, in contrast to Time-Averaged Restraints that rely on averaging values over time.

However, our findings revealed some unexpected patterns when compared to the provided trajectories. Notably, the unrestrained simulation in our study exhibited the lowest variance, while the Instantaneous Restraint simulation had the lowest mean RMSD. This deviates from our initial expectations, where we anticipated a higher mean RMSD in unrestrained simulations. The divergence in results, especially in the unrestrained simulations, highlights the complexity and variability inherent in protein dynamics simulations. Based on our trajectories, it's possible to see that the provided trajectories have different values, especially for the unrestrained case.

What we can see from our results is

Dataset	Mean	Variance
Unstrained	0.111768	0.0002453566
rmsd_NOE_TAR.out	0.119808	0.001081832
rmsd_NOE_IR.out	0.087139	0.001541292

Table 5: Mean and Variance of RMSD for provided trajectories in nm

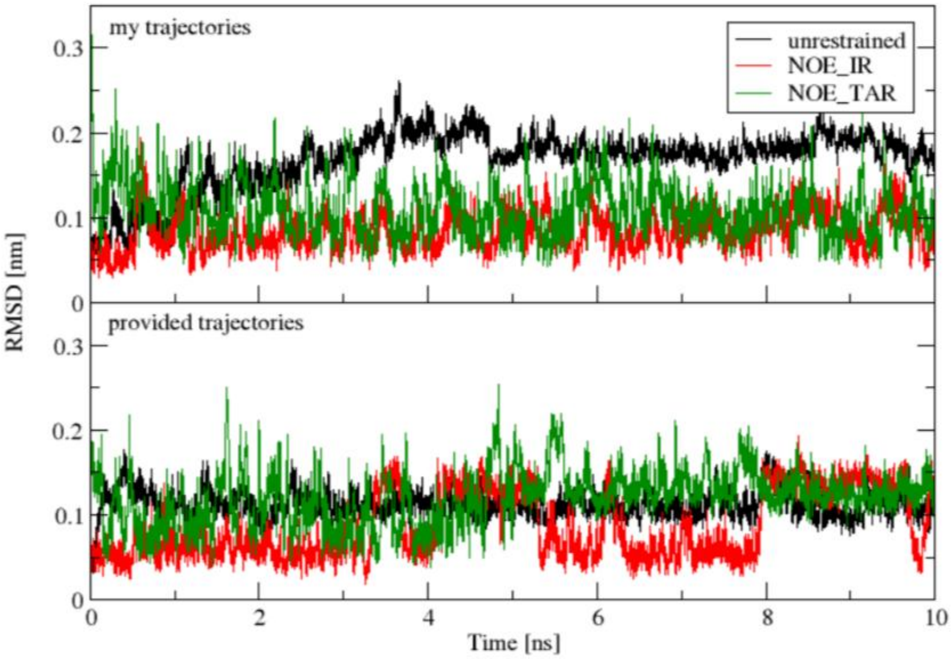


Figure 3: RMSD comparaison between two different simulations

D: Root Mean Square Fluctuation is a measure used to determine the flexibility or fluctuation of individual residues over the course of a simulation. The simulation has RMSF values that go from 0 to 0.4. The overall fluctuation (flexibility) of the residues tends to increase towards the beginning and the end of the peptide chain, particularly for the last residue (Gly). This suggests that the terminal residue is particularly flexible, showing a similar behaviour in the three cases. For the initial residue (Asn), instead, we can notice the same pattern in the three cases, while describing a different initial value (0.1 in NOE IR, 0.2 in unrestrained and 0.25 in NOE TAR). The NOE IR and NOE TAR lines generally track below the black line (unrestrained), indicating that the application of NOE restraints reduces the flexibility of the residues compared to the unrestrained condition. This is expected, as the restraints limit the range of motion allowed by the residues. The NOE IR and NOE TAR lines are quite close to each other, suggesting that the difference in restraint application (instantaneous vs. time-averaged) does not drastically alter the overall flexibility profile of the peptide in this simulation.

Dataset	**Mean**	**Variance**
Unstrained	0.092794397	0.002729834
rmsd_NOE_TAR.out	0.121806151	0.002300772
rmsd_NOE_IR.out	0.086130771	0.001682640

Table 6: Mean and Variance of RMSF in nm

D. Good analysis.

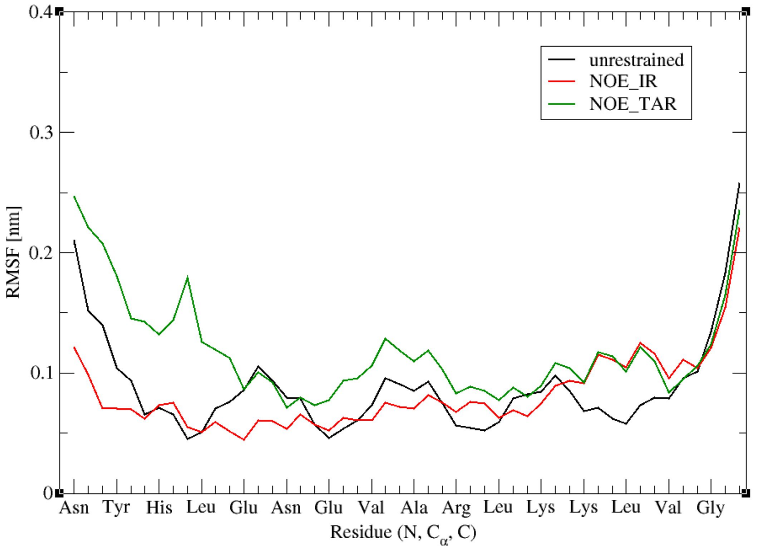


Figure 4: RMSF simulation

E. Number of positive violations: unrestrained > NOE_IR > NOE_TAR.

E: The Nuclear Overhauser Effect (NOE) manifests in nuclear magnetic resonance (NMR) spectroscopy, particularly during interactions among nuclear spins. This effect is instrumental in revealing the spatial closeness of nuclear spins, playing a pivotal role in the structural analysis of molecules, especially in structural biology for deciphering the architecture of biomolecules like proteins and nucleic acids.

During NOE experiments, the proximity of nuclear spins can either amplify (positive NOE) or diminish (negative NOE) the strength of NMR signals. The term 'upper bound violation' in NOE context typically denotes instances where the measured NOE values surpass a predefined or theoretical threshold, such as deviations from 10-ns simulations in our scenario. These discrepancies might stem from various factors, including the dynamics of molecular conformations, motion at the molecular level, or inaccuracies in the experimental arrangement. In this specific instance, the instances of upper bound violations display lower values when considering restraints, as expected.

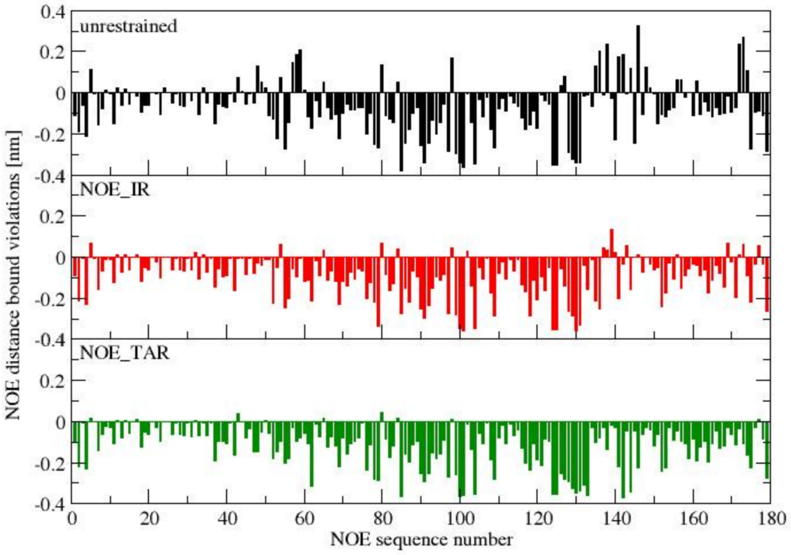


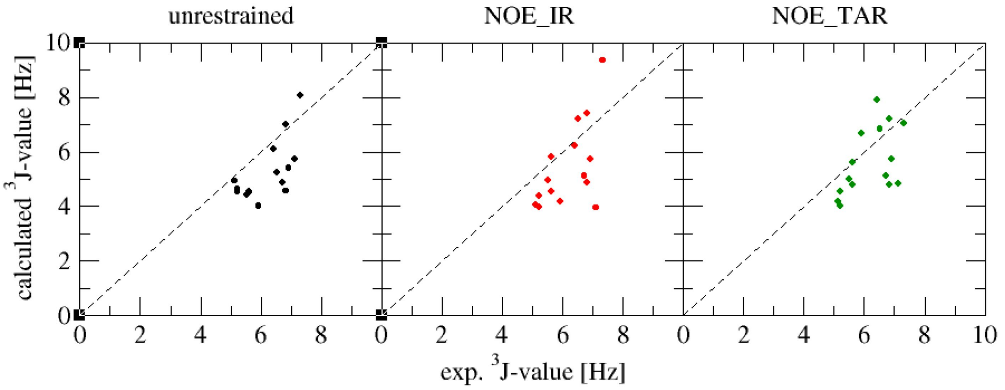
Figure 5: NOE Distance

F: In the context of nuclear magnetic resonance (NMR) spectroscopy, a ${}^3J$-value typically refers to a J-coupling

F. Good.

constant, also known as scalar or spin-spin coupling. J-coupling constants are indicative of the interaction between nuclear spins on different atoms within a molecule. These interactions are mediated through the chemical bonds connecting the atoms and provide valuable information about the molecular structure, conformation, and stereochemistry.

The panel contains points scattered around the dashed line, indicating the correspondence between the experimental and calculated ^3J values without any restraints applied. There is a difference between the unrestrained, instant restrained and time averaged restrained conditions. The spread of the points suggests a moderate correlation between the calculated and experimental values, with several points deviating significantly from the reference line, indicating discrepancies in the calculations. Compared to the values from the unrestrained condition, the IR and TAR have a better correlation between the calculated and experimental values. Moreprecisely, NOE_TAR shows the most accurate correlation between calculated and experimental values. Moving from the unrestrained to the NOE_TAR and NOE_IR plots, there's a visible increase in accuracy as points cluster closer to the diagonal line, indicating that the calculated values are more consistently reflecting the experimental values.



G. Ok.

Figure 6: Calcuation 3J-values

G: We can divide this answer based on the two restraints considered.
Given the third restraint:
a- The restrained hydrogen atoms are nitrogen one of Valine (residue 15) and the aliphatic carbon of (14)Leucine's C carbon b- The value of distance under which no potential (and relative force) is added is $R_0 = 0.35nm$
c- The first hydrogen atom (VAL) is explicitly modelled, while for the second (LEU) we use the positions of: LEU-15 Nitrogen, LEU-15 C carbon, LEU-15's first caron atom of the residual chain and LEU-15's carboxylic group's carbon.

For the seventieth restraint:
a- The restrained hydrogen atoms are the aliphatic carbon of (11) Leucine's C_α carbon one and (14) Leucine's second carbon of the radical group (pseudo atom).
b- The value of distance under which no potential (and relative force) is added is $R_0 = 0.44nm$
c- The first hydrogen is defined with: (11) Leucine's C_α carbon, (11) Leucine's nitrogen, (11) Leucine's first atom of the radical group and (11) Leucine's carboxylic carbon. For the second hydrogen, we have (14) Leucine's first carbon of the radical group, (14) Leucine's C_α carbon, (14) Leucine's second carbon of the radical group

H: Given the experimentally measured 3J-coupling constant and our simulation results, we have:

H. Correct. I. Well explained. J. Ok.

Parameter	Value
Experimental 3J-Coupling Constant (J0)	6.70 Hz
Calculated 3J-Coupling Constant (J ave)	4.90 Hz
Atoms Defining Dihedral Angle	C, N, CA, C of ASN residue (atoms 64, 66, 68, 75)

Table 7: 3J-Coupling Constants and Dihedral Angle Definition for Nr.6

In this case, the experimental measured $3J$-coupling constant is 6.70 Hz. Moreover, it's possible to see that the calculated $3J$-coupling constant is 4.90 ± 1.005 Hz. Thus, the experimental and the calculated $3J$-coupling constants differ by 1.8 Hz. For the dihedral angle, it is the angle between the planes made up of C_α, N and H atoms of the first amino acid and the N, C_α, and H_α atoms of the second amino acid.

I: Time-Averaged Restraining (TAR) is arguably the preferred method for several compelling reasons. Firstly, the data derived from NMR experiments are inherently averages of real quantities over the duration of the simulation. This aligns well with the TAR methodology, which focuses on constraining the average values rather than imposing restraints at every single timepoint.

Moreover, allowing for the averaging of constraints accommodates the inherent mobility of certain molecular components, such as sidechains. This is particularly relevant in the context of peptides, which often do not adopt a single rigid structure. Instead, they exist in a dynamic equilibrium of multiple conformations, with probabilities dictated by the underlying ensemble.

In essence, TAR offers a more realistic representation of the molecular system under study, acknowledging and incorporating its dynamic nature and structural variability.

J: We will have that the restraining force keeps pushing the observable at point t $Q_n(t)$ beyond the experimental value Q_n^0 as long as the average value $\overline{Q}_n(t)$ is bigger than the experimental value. This would be a problem in the case of J-value restraining because dihedral angles can rotate further as there's no repulsion to "tell the angle to don't go further". Moreover, the function connecting the structure to observables is multiple-valued (i.e. taking an horizontal line in some intervals will have more than one intersection) and we would restrain the system from reaching the other possible values (we wouldn't allow the system to go over the barriers)

3. Thinking Questions

K: K. Ok. L. Well explained.

Method	Fisibility	Advantages	Shortcomings
r^{-n}	One step less than what is done when calculating the distances	Closer measure to the observed intensities value taken experimentally and might do a better job dealing with long-distance relationships	There's not a precise way of selecting the exponent, just clear when the molecule is very big or very small
Dihedral angle	Doable, but you would need in any case to pass through the Karplus equation, which is not an invertible function (may require additional information?)	Would be a more intuitive measure with respect to the J-Couplings	Nonlinearity of the Karplus relation

Table 8: Comment of feasibility of alternative methods

L: It might be a good idea to use different parameterizations of the Karplus equation for different dihedral angles when we have for example a complex molecule, made by different subunits. The parameterization of the Karplus equation is always empirical and it might be that we have in the literature different parametrizations for the different subunits (maybe also with a QM correction). Additionally, given some particular experimental conditions, tweaking these hyperparameters for some dihedral angles might help to better fit the data.

M: We have these three conditions:

M. Correct. N. Ok.

- **Unrestrained simulation**: Here we have a fixed number of particles (N), box volume (at least in the simulation) (V) and Temperature (T) (at least in our simulation). Therefore the resulting ensemble is canonical. If we had not considered the thermostat we would be in the NVE, therefore in a microcanonical ensemble.
- **IR**: The presence of the restraint modifies the Hamiltonian, but importantly the energy is conserved. We would therefore also be in a microcanonical ensemble.
- **TAR**: Here we are still adding a restraining term to the Hamiltonian, but the difference with IR is that now the Hamiltonian is time-dependent and the energy is therefore not conserved.

N: To do ensemble averaging, instead of averaging through time we want to average through the ensemble. Therefore we want our observable to be the weighted average of the observables derived experimentally, where the weights of this weighted sum are given by the (Boltzmann) probability of a given conformed. The equation is

$$\overline{q(\mathbf{r}(t))} = \sum_{k=1}^{nconformers} p_k q(\mathbf{r}(t))$$

4. Appendix

Summary 1

You can find the list for the Google Colab Jupyter here.