Rotamer Dynamics: Analysis of Rotamers in Molecular Dynamics Simulations of Proteins

HADDAD, Y.; ADAM, V.; HEGER, Z.

Biophysical Journa
2019, vol. 116, iss. 11, June 2019, pp. 2062-2072

ISSN: 0006-3495

DOI: https://doi.org/10.1016/j.bpj.2019.04.017

Rotamer Dynamics: Analysis of Rotamers in Molecular Dynamics Simulations of Proteins

Yazan Haddad, Vojtech Adam, Zbynek Heger*

[1]Department of Chemistry and Biochemistry, Mendel University in Brno, Zemedelska 1, CZ-613 00 Brno, Czech Republic

[2]Central European Institute of Technology, Brno University of Technology, Purkynova 656/123, CZ-612 00 Brno, Czech Republic

***Corresponding author**

Zbynek Heger, Department of Chemistry and Biochemistry, Mendel University in Brno, Zemedelska 1, CZ-613 00 Brno, Czech Republic; E-mail: heger@mendelu.cz; phone: +420-5-4513-3350; fax: +420-5-4521-2044.

Abstract

Given by Chi torsional angles, rotamers describe the side-chain conformations of amino acid residues in a protein based on the rotational isomers (hence the word rotamer). Constructed rotamer libraries, based on either protein crystal structures or dynamics studies, are the tools for classifying rotamers (torsional angles) in a way that reflect their frequency in nature. Rotamer libraries are routinely used in structure modeling and evaluation. In this perspective article, we would like to encourage researchers to apply rotamer analyses beyond their traditional use. Molecular dynamics (MD) of proteins highlight the *in silico* behavior of molecules in solution, and thus can identify favorable side-chain conformations. In this article, we used simple computational tools to study rotamer dynamics (RD) in MD simulation. First, we isolated each frame in the MD trajectories in separate pdb files via cpptraj module in Amber. Then, we extracted torsional angles via bio3d module in R language. The classification of torsional angles was also done in R according to the penultimate rotamer library. RD analysis is useful for various applications such as protein folding, study of rotamer-rotamer relationship in protein-protein interaction, real-time correlation between secondary structures and rotamers, study of flexibility of side-chains in binding site for molecular docking preparations, use of RD as guide in functional analysis and study of structural changes caused by mutations, providing parameters for improving coarse-grain MD accuracy and speed, and many others. Major challenges facing RD to emerge as a new scientific field involve the validation of results *via* easy inexpensive wet-lab methods. This realm is yet to be explored.

Keywords: Rotamer dynamics, molecular dynamics, protein structure, rotamer library, torsional angle, dihedral angle.

Introduction

Proteins play major structural and functional roles in all living beings. The molecular understanding of protein structures at the atomic level is important for explaining cellular mechanisms as well as for various medical and non-medical applications. At the moment of writing this paper, there are already more than 137,000 biological macromolecular structures deposited at the Protein Data Bank (PDB). Among these macromolecules, over 127,000 structures have been annotated as proteins (www.rcsb.org). This experimentally-verified crystal structure reservoir is a thriving environment for bioinformatics research and exploitation via computational study. Molecular dynamics (MD) analysis is a well-established computational method for study of structures in controlled environment settings. Based on pre-defined parameters, known as force fields, it is possible to apply laws of physics to calculate the effect of molecular forces (*i.e.* bond lengths, bond angles, torsional/dihedral angles, electrostatic and Van der Waals bonds), and predict the trajectories of all atoms in a system. MD analysis of proteins employs these calculations for many times to simulate the behavior of molecules in well-controlled aqueous environment *in silico*. More than 20,000 ISI-indexed articles per year have been published on the topic of MD in the past three years (webofscience.com).

On the other hand, as early as the first protein structures were discovered, it was shown that particular bonds and angles were restricted towards ideal ranges. Torsional angles can either describe the dihedral rotation in the backbone (Phi/φ between Cα and N, Psi/ψ between Cα and C, omega/ω between C and N) or side-chain (Chi1/χ1 between Cα and Cβ, Chi2/χ2 between Cβ and Cγ, etc.) as shown in **Figure 1A**. Rotamers, the rotations of the side-chain torsional angles, were intensively studied by various methods to identify the ideal rotamer ranges occurring in nature. To construct a rotamer library, one method depends on collection of protein structures and statistical analysis of side-chain conformations, while another method applies clustering approach of the three possible carbon sp^3–sp^3 rotations (i.e. +60°, 180°, and -60°). These three torsional angles can be represented by IUPAC gauche/trans nomenclature (*g+*, *t*, *g-*, respectively); however, previous researchers used *g+* and *g-* to represent either +60° or -60° without consistency. An alternative nomenclature (*p*, *t*, *m* for +60°, 180°, and -60°, respectively; shown in **Figure 1B**) was proposed by Lovell *et al.* (1).

Rotamers usually represent a local energy minimum of torsional angles, thus the backbone torsional angles phi and psi can also be involved (2). The *Dunbrack backbone-dependent rotamer library* is among the widely used rotamer libraries (3). Rotamer libraries that are dependent on secondary structure are useful for homology modeling (4). However, Dunbrack argues that his data are in favor of backbone-dependency rather than explicit secondary structures (2). In this paper, we used the term “secondary structure” to broadly describe backbone torsional angles (Phi and Psi) as well as inter-chain hydrogen bonds.

The *penultimate rotamer library* by the Richardson laboratory (1) attempted to avoid the internal atomic clashes resulting from ideal hydrogen atoms in the structure and uncertain residues with high B-factor. Thus, this library provided a higher quality, coverage and low number of rotamer classes (nearly 153 rotamers). The latter advantage was ideal for analysis and graphical representation, which is why this library was chosen for the protocol presented in this article.

Both libraries mentioned earlier focused on high-quality crystal structures of proteins, and thus little information is known about rotamer flexibility in solution. The *dynameomics rotamer library* employs MD simulation for at least 31 ns at 25 °C to predict rotamers of proteins in solution environment (5). The library was compared to several structure datasets and the researchers investigated role of buried *vs*. surface residues in both crystal and dynamic structures. Furthermore, the library was supported by experimental data from nuclear magnetic resonance (NMR) relaxation to measure S^2 side-chain order parameters of Ala Cβ, Ile Cγ and Cδ, Leu Cδ, Met Cε, Thr Cγ, and Val Cγ on a ps to ns time scale. This NMR-based method is routinely used to directly probe methyl group mobility in the side-chains of proteins (6).

Previous studies employing MD analysis and focusing on rotamers often represented their data *via* plotting changes in χ dihedral angles over time (7, 8) or through principal component analysis (9). While few dihedral angles are easy to plot, a simple classification scheme is required when dealing with large number of heterogeneous residues. We hope that our work will address the graphical challenges faced by previous researchers. Watanabe and others resolved to decomposing rotamer histograms from MD simulation into Gaussians to make dihedral populations (10), which is equivalent to construction of new rotamer library! It is important to develop a simple, comparative, benchmarked, and easy-to-visualize rotamer analysis method in MD simulations that can be widely adapted by researchers. The penultimate rotamer library is ideal choice for use in MD analysis since it is backbone-independent (hence all possible rotamers are included at once) with countable number of rotamers (thus easy to classify for graphical visualization) and simple nomenclature (for instance, ptp rotamer of Met residue describes torsions for Chi angles in the order p then t then p for Chi1, Chi2 and Chi3, respectively). The ptm-85 and ptm180 rotamers of Arg residue describe the Chi4 angle to be around the mean -85° and 180°, respectively. The penultimate rotamer library also describes all possible rotamer ranges predicted from a very stringent collection of highly resolved and refined structures.

The purpose of this perspective is to create simple and easy RD analysis strategy that can be adapted and developed by researchers in the MD field. However, for a full proof of concept exploitation it is also important to develop biophysically relevant graphical visualization. In the following protocol and example sections, we will also point out to several unforeseen technical challenges and discuss the best ways to overcome them.

A wide range of programs have been developed for MD simulations. Many programs can already perform analysis related to RD. The CHARMM program uses a correlation function to study average and RMS fluctuation for Chi1 and Chi2 angles (11). In GROMACS, it is possible to prepare an index file with the preferred dihedral angles and extract these data from a simulation in the form of trigonometric functions and perform principal component analysis (12). In this case, rotamer classification is performed post-analysis. RD analysis can still be done in the same way as our protocol; however, it might be more laborious particularly in preparing index file and performing rotamer classification that is more biophysically relevant. We have noticed that extracting dihedral angles and assigning them to residues pose a challenge when performing analysis *via* many programs because it requires either selecting the four atoms per dihedral angle or defining them in an index file like in GROMACS. We do not have practical experience in these programs but the list includes LAMMPS (13) and Python modules. Another example is VMD Timeline plugin (University of Illinois at Urbana-Champaign, USA) which can produce some dihedral angle graphical representation in a trajectory. At this point, the Bio3D module (Grant lab, University of California, San Diego, USA) in R language (The R foundation for Statistical Computing, Austria) was a very attractive choice since we only had to define residues (not dihedral angles) for extraction of dihedral angles. The only limitation was that it can perform this for a single structure at a time and in pdb format.

Protocol

In brief, a MD simulation was first done using sander module in AMBER 14 program (University of California, San Francisco, USA). Since torsional angles calculations, in Bio3D module in R language, can only be performed for a single structure at a time, the process was automated to perform calculations for all simulation frames. Firstly, the trajectory file was converted to pdb format, and all frames were saved as single pdb file per frame using cpptraj module in AMBER (**Figure 2A**). Torsional angles were calculated and saved for each residue using Bio3D module in R. The data were transformed by collecting each angle value for each frame to final format of angles (in columns) and frames (in rows). Using the Penultimate rotamer library, the torsional angle data were classified into rotamers using if/else statements (**Figure 2B** and **supplementary file Rotamers.R**).

To show a practical example of this protocol, we applied RD analysis on peptide-protein interaction in the next section. In this example, implicit-water MD simulations of neurotrophic pNGF peptide (SSSHPIFHRGEFSV$_{\text{-NH2}}$) and its receptor (Ig2 extracellular domain of tropomyosin receptor kinase TrkA) were done in free and bound states. Atomistic coordinates were derived from protein databank model (PDB-ID 2IFG) and modified in UCSF Chimera (University of California, San Francisco, USA). Input coordinate and topology files were prepared using H++ server (14) with protonation states optimal for physiological pH (7.4). Canonical ensemble (NVT) MD simulations were performed using improved Generalized Born solvent model for protein simulations (15). The structures were minimized (maximum 5000 cycles), equilibrated for 500 ps at 25 °C, and production was performed in Langevin dynamics (25 °C, 16 Å non-bonded cutoff, 0.002 ps time steps) for several consecutive periods of 50 ns. Coordinates were printed every 1000 steps. The bound complex separated after ~150 ns therefore only part of the entire simulation will be shown for demonstrative purposes. Approximately 50,000 frames, each representing 2 ps time step and corresponds to total 100 ns of MD simulation were used for the example.

There are few important notes to take care of when saving trajectories in multiple pdb files (**Figure 2A**). The number of frames (reflecting the number of files) is decided by the time interval, which we expect the rotamers convergence to be visible. This can be controlled by taking strides in steps between frames, for instance, using “offset 5” in the trajout command to skip 5 frames between each output (offset 1 is the default). In the case of large molecule simulations, computational load is important (Quick tip: opening a folder with large number of files will slow down the computer). Removal of non-protein atoms can speed up the processing. Further, to process large numbers of frames in large molecules we also recommend selecting tens of residues at a time and saving them separately (e.g. the command “strip !(:9-26)” can be used in the line before “trajout …” to select residues 9 to 26). Residues of Ala and Gly can be avoided since they have no rotamers. The reader is referred to AMBER manual for further customization. In this example, we saved 50,000 frames of three simulations (peptide free, receptor free, peptide-receptor bound) in pdb format.

To open the pdb files in R language, the bio3d library module was loaded in step 1 (**Figure 2B**). Commands for reading each pdb frame were produced via FOR statement and captured in variable x, which was in turn saved in a file and executed. In step 2, the torsional angles of residue number 3 were collected for all frames by commands produced via FOR statement and captured in variables y supplemented with z which was in turn saved in a file and executed. Similar code automation can be done for collecting data for all residues (Shown in step 2-3a). Previously, the angles are saved in a variable called tor_residue3. In step 3, the angles are saved in a tab-delimited text file.

The classification of rotamers according to the penultimate rotamer library was justified in the previous section. In step 4, the angles which were saved in step 3 are read into variable called tor. A variable for rotamers is created. In step 5, an IF/ELSE statement is used to classify and save the rotamers variable Rota_residue1 in a tab-delimited text file. IF/ELSE statement scripts for the rest of amino acid rotamers are shown in Supplement file (**Rotamers.R**). If the computational load is larger than processing capacity, which might happen with Arg residues containing up to 34 groups, we recommend dividing the groups into two or more steps and saving each calculation in separate column, e.g. Rota_residue1[i,1] , Rota_residue1[i,2] , etc… Alternatively, it is possible to combine groups together for easier visualization (e.g. focusing on Chi1 and Chi2).

The rotamer groups are represented by data strings (in accordance to nomenclature in **Figure 1**). There is an array of methods that can be used for data visualization and analysis. Graphical representation includes rotamer frequency distribution, distribution over time (time evolution), and other correlations with time, secondary structure, energy, ligands, and other rotamer combinations.

Example: RD analysis of pNGF peptide binding to TrkA receptor

Neurotrophic peptides are a new generation of synthetic neurotrophic factors derived from neurotrophins, and can be used to induce neuron differentiation and prevent or reverse neuronal degeneration for treatment of various diseases (16). Due to the vast cross-interactions between neurotrophins and the three tropomyosin receptor kinases (TrkA, TrkB and TrkC), the therapeutic selectivity and specificity pose a challenge in controlling their side effects (17). Understanding the peptide binding process in detail is important for optimization and development of selective therapeutics. Here we show a study of the binding between pNGF peptide (**Figure 3A**) and its counter Ig2 extracellular domain of tropomyosin receptor kinase TrkA (**Figure 3D**). This example is shown for the purpose of graphical visualization only and should not be used to derive conclusions without experimental validation. Among the residues on the interaction interface from the peptide are H4 and P5, which face the residues S304 and H343 on the receptor respectively. In free form, H4 exhibits a *Bend* or no secondary structure and a variety of rotamers (m-70, m80, m170, and t-80), while in bound form; H4 exhibits a predominant t-80 rotamer (**Figure 3B**). Similarly, P5 residue shifts towards the Cγ exo rotamer stabilized with formation of alpha helix secondary structure (**Figure 3C**). On the other side, the stabilization of both H343 residue m-70 rotamer and H4 residue t-80 rotamer resulted in nearly equal dynamic interaction between the two histidines and S304 residue forming both m and t rotamers (**Figure 3E and 3H**). The latter was a rare case where a rotamer (S304 residue) is more fixed when the protein is free, while the other examples show the expected fixation of rotamers upon binding; namely, t rotamer in V294 (**Figure 3F**), m-70 in H298 (**Figure 3G**) and p90 in F329 (**Figure 3I**) residues. The distribution of rotamer frequency provides a summary for RD and its relationship with secondary structure frequency.

Further information about the dynamics of rotamer-rotamer interaction over time can be obtained *via* time-evolution plots (**Figure 4A**). Here, a detailed time scale of 2 picosecond per frame showed a stable rotamer conformation over scales of ten(s) nanoseconds. The frequency of rotamer combination for the four listed residues showed highest occurrence of the t-80, Cγ exo, m, m-70 rotamers of H4, P5, S304 and H343, respectively (**Figure 4B**).

In contrast to the principal component analysis for torsional angles that we mentioned in the introduction, a special type of factor analysis for nominal and mixed types of variables is used (18). Here, the whole residue (i.e. side chain) is studied as one unit instead of a heterogeneous index of dihedral angles. Multiple factor analysis for mixed data showed direct correlation between pNGF His4 and TrkA His343 in the first two dimensions (**Figure 4C**). A detailed correlation of the possible rotamer combinations can be visualized as well (**Figure 4D**). For example, in the negative panel of the first component dimension four interesting rotamers are correlated together: t-80, Cγ exo, m, m-70 rotamers of H4, P5, S304 and H343, respectively (**Figure 4D**). Trimming of the noise from the data (removing the first and last 10 ns) improved the component dimension 1 (from 8.86 to 12.03%) and 2 (from 6.37 to 11.39%) with similar outcomes. This highlights the benefit of time-evolution plots used in **Figure 4A**.

We believe that this factor analysis approach to the study of dihedral angles *via* RD is more relevant to the researchers nowadays than factor analysis of heterogeneous indices of dihedral angles. In the previous example, the objective is to understand the binding process in order to produce more selective mutant peptide for therapy. We think that rotamers provide a biophysically relevant representation of the structure, particularly when they are studied in association with energetics and thermodynamics.

Applications and Future Prospects

Before discussing the applications of RD analysis, it is important to give this subject its modest and unexaggerated weight. When performing a MD simulation, the major forces are calculated from two kinds of energy: bonded and non-bonded. The changes in dihedral angles belong mostly to the bonded energy, while a great contribution to the total energy of the system comes from interactions mediated by non-bonded, *viz.* non-covalent interactions. These include the Van der Waals and the electrostatic (ionic and hydrogen bonds). In protein-protein interactions and protein-ligand interactions, water also plays a significant role through networks of hydrogen bonds (19). During globular protein folding, hydrophobic side-chains are confined inside the protein in tightly packed and more rigid fashion than the rest of the protein, while hydrophilic residues protrude to face the water surface (20). We hope that the usefulness of RD analysis in providing real-time insight on protein folding could be evaluated by analyzing the flexibility of rotamers in large datasets. It is worth a note that implicit water models might provide more rotamer flexibility than explicit water models since the latter would provide more realistic water-based hydrogen bonding. Comparative RD analysis can give new perspective to improvements on implicit water models that are more representative to explicit water in the future.

As mentioned earlier, the relationship between the backbone and side-chains shifts towards the energy minimum. Thus, it is possible to visualize the correlation between secondary structures and rotamers (**Figure 3**). Similarly, protein-protein interactions can involve rotamer-rotamer relationships, which can either involve switching or fixing of movement (**Figure 3E**). RD analysis can show these changes in real time (**Figure 4A**).

Molecular docking is a computational method used to study both protein-protein interactions and protein-ligand interactions where the interaction is often scored using scoring functions based on free energy estimates *via* molecular mechanics or other methods (21). Most docking softwares employ predefined parameters that can increase the accuracy of the prediction. These include; existing water molecules, protonation states of some residues, and “explicit flexibility” of amino acid side-chains in the binding site or binding interface. However, the aforementioned parameters are the roots to many challenges in producing a universally accurate and reliable solution *via* MD (22). We hope that by controlling more factors involved in the interaction, and by defining the restraints involved in rotamers, this method can improve its accuracy to some extent. In fact, similar innovations have been implemented to integrate MD in molecular docking aside from the usual post-docking validation procedures (23). The protein’s intrinsic flexibility is a major drawback for docking, however a combination of MD method and sampling of multiple receptor conformations (MRC) were shown to match and possibly outperform crystal structures in retrospective virtual screening experiments (23). MRC-based methods are often referred to as ensemble docking which implements “implicit flexibility” in both side-chains and backbone (24). The role of side-chain flexibility is highlighted in peptide-protein docking (25).

Another plausible application of RD can be found in mutational scans that are used for functional analysis and also for exploratory industrial development of recombinant proteins and enzymes. The specificity and binding affinity are determined by structural and physico-chemical properties at the interaction interface or in a binding site even with a small number of amino acid substitutions (26). Nevertheless, laboratory mutagenesis methods can be time-consuming and highly costly. RD analysis can give a different detailed map of the changes in the three dimensional landscape surrounding the mutated residue to provide further insight into its desired functionality. The same can be said regarding post-translational modifications of amino acids and addition of sugars, lipids, etc. On the other hand, increasing protein stability *aka* protein engineering is a desirable goal for different life science purposes ranging from basic research to clinical and industrial applications (27). We hope that further analysis of rotamers in MD simulation will help identify the local and distant factors that contribute to residues with rotamers of highly restricted dihedral angles range. The development of highly rigid protein structures is important for improving thermo-stability, non-degradability and pressure tolerance.

Rotamers analysis in MD simulations was previously shown to be useful in predicting side-chain packing, which is important for developing coarse-grained MD (a technique that differ from atomistic MD by grouping atoms or residues into grains/beads of various sizes, thus reducing computational load). In fact, predicting $\chi1$ rotamer states alone increased the speed of MD calculations, and thus reduced MD simulation time significantly (28).

The number of articles on the topic of rotamer or rotamers rarely exceeded the range of 70–90 articles per year, compared to more than 20,000 articles published on the topic of MD (webofscience.com). Clearly, the development of state-of-the-art RD tools and awareness among scientists of possible applications of rotamers in MD analysis are required to narrow this gap. We hope that by improving the tools for RD study (computationally and experimentally), and by dissemination of knowledge of the issue, the researchers will have better background in this field and will be able to extend their analysis of MD beyond the backbone/secondary structure into more detailed side-chain structure study. One strategy is to study convergence of rotamers over time as triggered by certain events. Another strategy is to gain functional information from fluctuations in side-chains from comparative study as shown in the example (e.g. ligand free vs. ligand bound, native proteins vs. protein interaction, or any comparative conditions).

As mentioned earlier, validation of RD analysis with NMR measurement of methyl side-chain order parameter values (S^2-values) is the gold standard; however, it is not the cheapest, easiest nor most feasible among researchers. Such difficulty was obviously not an issue for experimental validation of torsional angles of secondary structures. Indicators of secondary structure can be detected by circular dichroism spectroscopy (29), Fourier transform infrared spectroscopy (30), Raman spectroscopy (31) and other methods. However, some relevant information is still attainable regarding side-chains of proteins, but with more laborious work in interpretation (32). For some aromatic residues like tyrosine and tryptophan, the fluorescence lifetime (i.e. decay) can accurately analyze rotamer distributions (33, 34). Recent studies *via* vibrational spectrometry –complemented with computational chemistry– reported detailed assignments of the side-chains torsional vibrations of dipeptides such as Ala-Gln

(35), Gly-Val (36), Gly-Leu (37), Gly-Tyr (38), Met-Ser (39), His-Phe (40). Torsional vibrations were mostly featured in the spectral range below 1000 cm^{-1}. On the other hand, alternative computational methods for prediction of rotamers without MD are less common, yet one noticeable example is Dead-End Elimination algorithm (DEE) method (41). The DEE method – often used for protein 3D structure prediction – relies on calculating and minimizing potential energy, and then limiting side-chain conformations to discrete set of rotamers. As the name implies, rotamers that cannot be grouped in the global minimum energy conformation are eliminated. On a different level, it is possible to derive information related to RD from the profiles of root-mean-square deviations (RMSD) and root-mean-square fluctuations (RMSF) of atomic coordinates. Unlike RD, these mathematical methods require a defined reference, and while they can give quick indications on the fixed *vs*. flexible residues, the study of RD gives a better chemical and geometric description of the system.

As with all MD studies, the possible bias in results that originates from using certain force field (FF) is also a concern for RD analysis. FF bias has been a critical problem particularly for the studies of protein folding, even when using different versions of the same FF! For instance, Shao and Zhu reported that some versions of Amber FF can have preference to certain secondary structures in contrast to others (42). This issue has been previously addressed in the Biophysical journal (43, 44). The accuracy of FF can be further improved by analyzing both backbone and side-chain torsional angles (45). Hopefully, in protein folding studies, RD analysis can be useful for assessment of MD bias resulting from FF. The field of protein structural biology is on the verge of accurate sequence/crystal-structure prediction, as shown in fast advances in *de novo* and homology modeling techniques (46). However, we are yet far from approaching the sequence/dynamic-structure prediction model, which better describes proteins in physiological conditions. We believe that a dynamic-structure model(s) can be established for proteins based on probabilistic distributions of torsional angles alone, in which case, RD will play a pivotal role. Such quantitative and descriptive models will better exploit the infinite landscape of protein folding. Moreover, in 2014, a group of researchers was able to expand the genetic alphabet to include unnatural nucleotide base pairs (47). It is hoped that in the future such expansion can be reflected in codons and eventually the expression of as many as 172 different synthetic amino acids (48). The development and understanding of both natural and synthetic amino acid side-chains will require more attention by researchers in the coming decades.

In conclusion, computational methods have wide interest among researchers, and they are becoming more feasible, accessible, integrable and accurate day by day. Here we have questioned the feasibility and proof of concept of performing RD analysis using freely available tools. RD analysis is very descriptive, chemically and biophysically relevant when compared to torsional angle description, the same way a secondary structure is relevant when compared to backbone torsional angles. We think the time has come for a benchmarked and adaptable approach for performing RD analysis. The development of fast, cheap and reliable experimental methods that validate rotamers in solution will make the breakthrough for this field.

Author Contributions: YH performed the computation and writing. VA reviewed the manuscript, and ZH was principle investigator and contributor to scheme and organization of work. All authors have given approval to the final version of the manuscript.

Conflict of Interest: The authors declare no conflict of interest.

Funding Sources: European Research Council (ERC) under the European Union´s Horizon 2020 research and innovation programme (grant agreement No. 759585) and The Czech Science Agency (project No. 18-10251S) are gratefully acknowledged.

Acknowledgements: Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures".

References

1. Lovell, S. C., J. M. Word, J. S. Richardson, and D. C. Richardson. 2000. The penultimate rotamer library. Proteins 40(3):389-408.
2. Dunbrack, R. L. 2002. Rotamer libraries in the 21(st) century. Curr. Opin. Struct. Biol. 12(4):431-440.
3. Dunbrack, R. L., and M. Karplus. 1993. Backbone-dependent rotamer library for proteins - application to side-chain prediction. J. Mol. Biol. 230(2):543-574.
4. Bates, P. A., and M. J. E. Sternberg. 1999. Model building by comparison at CASP3: Using expert knowledge and computer automation. Proteins:47-54.
5. Scouras, A. D., and V. Daggett. 2011. The dynameomics rotamer library: Amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water. Protein Sci. 20(2):341-352.
6. Carbonell, P., and A. del Sol. 2009. Methyl side-chain dynamics prediction based on protein structure. Bioinformatics 25(19):2552-2558.
7. Engh, R. A., L. X.-Q. Chen, and G. R. Fleming. 1986. Conformational dynamics of tryptophan: a proposal for the origin of the non-exponential fluorescence decay. Chem. Phys. Lett. 126(3-4):365-372.
8. Das, S., S. Das, A. Roy, U. Pal, and N. C. Maiti. 2016. Orientation of tyrosine side chain in neurotoxic Aβ differs in two different secondary structures of the peptide. R. Soc. Open Sci. 3(10):160112.
9. Altis, A., P. H. Nguyen, R. Hegger, and G. Stock. 2007. Dihedral angle principal component analysis of molecular dynamics simulations. J. Phys. Chem 126(24):244111.
10. Watanabe, H., M. Elstner, and T. Steinbrecher. 2013. Rotamer decomposition and protein dynamics: efficiently analyzing dihedral populations from molecular dynamics. J. Comput. Chem. 34(3):198-205.
11. Schleif, R. 2013. A concise guide to charmm and the analysis of protein structure and function. WWW.
12. Abraham, M. J., D. v. d. Spoel, E. Lindahl, B. Hess, and t. G. d. team. 2018. GROMACS User Manual version 2018.
13. Sandia-National-Laboratory. 2018. LAMMPS user's manual: compute dihedral/local command.
14. Gordon, J. C., J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev. 2005. H++: a server for estimating p K as and adding missing hydrogens to macromolecules. Nucleic Acids Res. 33:368-371.
15. Nguyen, H., D. R. Roe, and C. Simmerling. 2013. Improved generalized born solvent model parameters for protein simulations. J. Chem. Theory Comput. 9(4):2020-2034.
16. Travaglia, A., A. Pietropaolo, R. Di Martino, V. G. Nicoletti, D. La Mendola, P. Calissano, and E. Rizzarelli. 2015. A small linear peptide encompassing the NGF N-terminus partly mimics the biological activities of the entire neurotrophin in PC12 cells. ACS Chem. Neurosci. 6(8):1379-1392.
17. Haddad, Y., V. Adam, and Z. Heger. 2017. Trk receptors and neurotrophin cross-interactions: New perspectives toward manipulating therapeutic side-effects. Front. Mol. Neurosci. 10:1-7.
18. Chavent, M., V. Kuentz-Simonet, A. Labenne, and J. Saracco. 2014. Multivariate analysis of mixed data: The PCAmixdata R package. arXiv preprint arXiv:1411.4911.
19. Ferreira, L., R. dos Santos, G. Oliva, and A. Andricopulo. 2015. Molecular docking and structure-based drug design strategies. Molecules 20(7):13384-13421.
20. Hurley, J. H. 1994. The role of interior side-chain packing in protein folding and stability. The Protein Folding Problem and Tertiary Structure Prediction. Springer, pp. 549-578.
21. Liu, J., and R. Wang. 2015. Classification of current scoring functions. J. Chem. Inf. Model. 55(3):475-482.

22. Cheng, T., Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant. 2012. Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J. 14(1):133-141.
23. De Vivo, M., M. Masetti, G. Bottegoni, and A. Cavalli. 2016. Role of molecular dynamics and related methods in drug discovery. J. Med. Chem. 59(9):4035-4061.
24. Bonvin, A. M. 2006. Flexible protein–protein docking. Curr. Opin. Struct. Biol. 16(2):194-200.
25. Dagliyan, O., E. A. Proctor, K. M. D'Auria, F. Ding, and N. V. Dokholyan. 2011. Structural and dynamic determinants of protein-peptide recognition. Structure 19(12):1837-1845.
26. Allam, A., L. Maigre, M. Alimi, R. A. de Sousa, A. Hessani, E. Galardon, J. M. Pages, and I. Artaud. 2014. New Peptides with Metal Binding Abilities and Their Use as Drug Carriers. Bioconjugate Chemistry 25(10):1811-1819.
27. Buß, O., J. Rudat, and K. Ochsenreither. 2018. FoldX as protein engineering tool: Better than random based approaches? Comput. Struct. Biotechnol. J. 16:25-33.
28. Jumper, J. M., K. F. Freed, and T. R. Sosnick. 2016. Rapid calculation of side chain packing and free energy with applications to protein molecular dynamics. arXiv preprint arXiv:1610.07277.
29. Greenfield, N. J. 2006. Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. 1(6):2876.
30. Yang, H., S. Yang, J. Kong, A. Dong, and S. Yu. 2015. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. Nat. Protoc. 10(3):382-396.
31. Rygula, A., K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska. 2013. Raman spectroscopy of proteins: a review. J. Raman Spectrosc. 44(8):1061-1076.
32. Barth, A. 2007. Infrared spectroscopy of proteins. Biochim. Biophys. Acta Bioenerg. 1767(9):1073-1101.
33. Clayton, A. H., and W. H. Sawyer. 1999. Tryptophan rotamer distributions in amphipathic peptides at a lipid surface. Biophys. J. 76(6):3235-3242.
34. Saraiva, M. A., C. D. Jorge, H. Santos, and A. L. Maçanita. 2016. Earliest events in α-synuclein fibrillation probed with the fluorescence of intrinsic tyrosines. J. Photochem. Photobiol. B 154:16-23.
35. Kecel, S., A. E. Ozel, S. Akyuz, and S. Celik. 2010. Conformational analysis and vibrational spectroscopic investigation of L-alanyl-L-glutamine dipeptide. J. Spectrosc. 24(3-4):219-232.
36. Celik, S., A. E. Ozel, S. Akyuz, S. Kecel, and G. Agaeva. 2011. Conformational preferences, experimental and theoretical vibrational spectra of cyclo (Gly–Val) dipeptide. J. Mol. Structure 993(1-3):341-348.
37. Celik, S., A. E. Ozel, and S. Akyuz. 2016. Comparative study of antitumor active cyclo (Gly-Leu) dipeptide: A computational and molecular modeling study. Vib. Spectrosc. 83:57-69.
38. Çelik, S., S. Akyuz, and A. E. Ozel. 2017. Vibrational spectroscopic and structural investigations of bioactive molecule Glycyl-Tyrosine (Gly-Tyr). Vib. Spectrosc. 92:287-297.
39. Kecel-Gunduz, S., B. Bicak, S. Celik, S. Akyuz, and A. E. Ozel. 2017. Structural and spectroscopic investigation on antioxidant dipeptide, L-Methionyl-L-Serine: A combined experimental and DFT study. J. Mol. Structure 1137:756-770.
40. Celik, S., A. E. Ozel, S. Kecel, and S. Akyuz. 2012. Structural and IR and Raman spectral analysis of cyclo (His-Phe) dipeptide. Vib. Spectrosc. 61:54-65.
41. Maglia, G., A. Jonckheer, M. De Maeyer, J. M. Frère, and Y. Engelborghs. 2008. An unusual red-edge excitation and time-dependent Stokes shift in the single tryptophan mutant protein DD-carboxypeptidase from Streptomyces: The role of dynamics and tryptophan rotamers. Protein Sci. 17(2):352-361.
42. Shao, Q., and W. Zhu. 2018. Assessing AMBER force fields for protein folding in an implicit solvent. Phys. Chem. Chem. Phys. 20(10):7206-7216.
43. Mittal, J., and R. B. Best. 2010. Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. Biophys. J. 99(3):26-28.
44. Freddolino, P. L., S. Park, B. Roux, and K. Schulten. 2009. Force field bias in protein folding simulations. Biophys. J. 96(9):3772-3780.

45. Best, R. B., X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr. 2012. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles. J. Chem. Theory Comput. 8(9):3257-3273.

46. Krieger, E., S. B. Nabuurs, and G. Vriend. 2003. Homology modeling. Methods Biochem. Anal. 44:509-524.

47. Malyshev, D. A., K. Dhami, T. Lavergne, T. Chen, N. Dai, J. M. Foster, I. R. Corrêa, and F. E. Romesberg. 2014. A semi-synthetic organism with an expanded genetic alphabet. Nature 509(7500):385-388.

48. Service, R. F. 2014. Synthetic biology. Designer microbes expand life's genetic alphabet. Science 344(6184):571.

Figure Legends:

Figure 1

(A) Nomenclature of the torsional angles in the backbone and side-chain structure. (B) Representation of the first three Chi torsional angles in the side-chain and the p,t,m nomenclature.

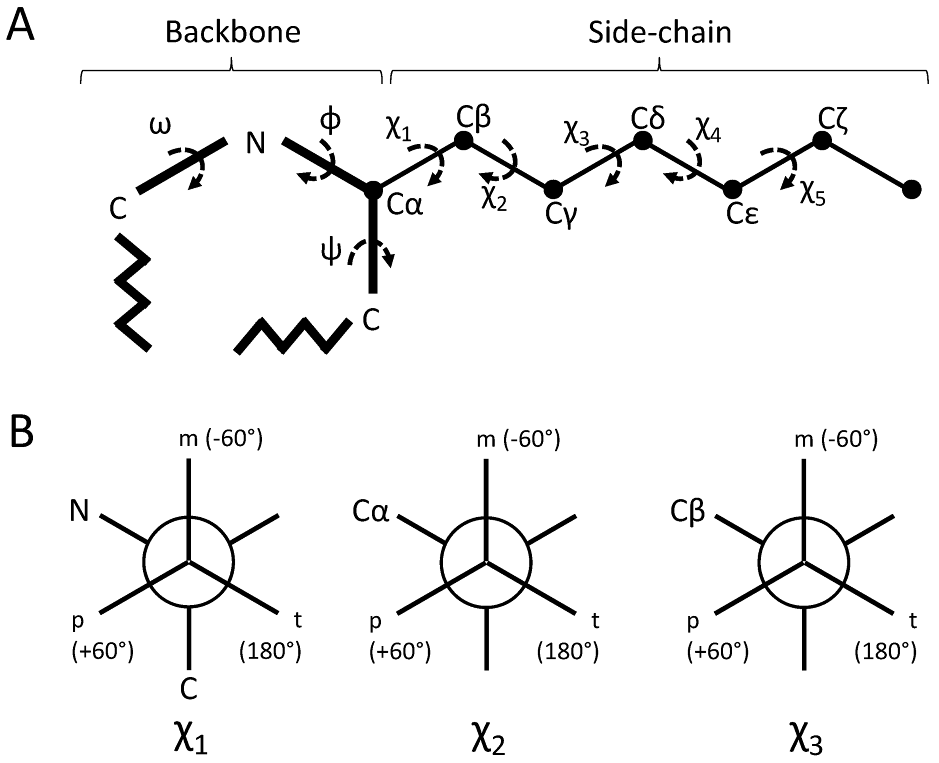
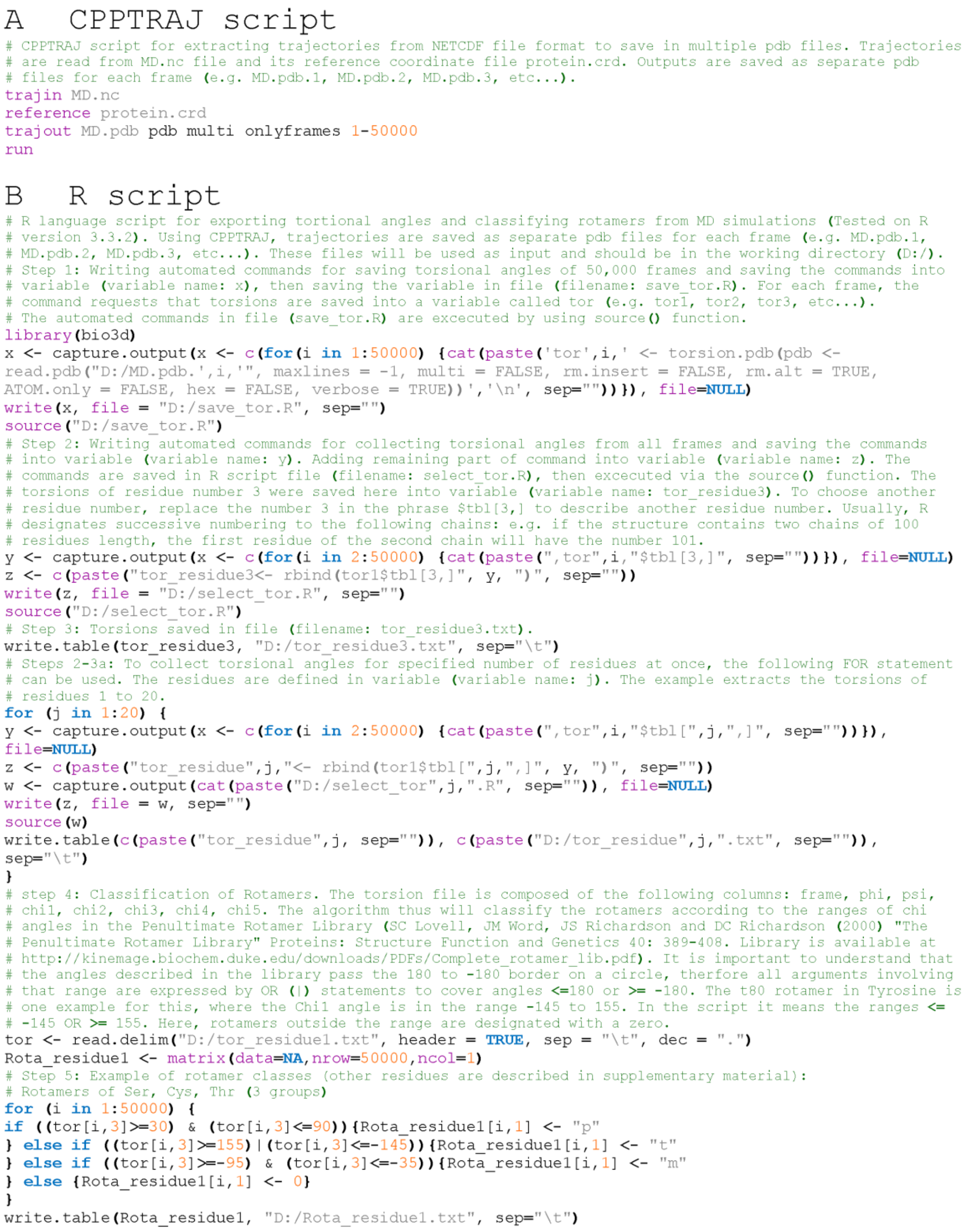


Figure 2

(A) CPPTRAJ script. (B) R script. Comments are shown in green, counts and numbers in orange, strings in grey, commands in purple and logic in blue.



A   CPPTRAJ script

```
# CPPTRAJ script for extracting trajectories from NETCDF file format to save in multiple pdb files. Trajectories
# are read from MD.nc file and its reference coordinate file protein.crd. Outputs are saved as separate pdb
# files for each frame (e.g. MD.pdb.1, MD.pdb.2, MD.pdb.3, etc...).
trajin MD.nc
reference protein.crd
trajout MD.pdb pdb multi onlyframes 1-50000
run
```

B   R script

```
# R language script for exporting tortional angles and classifying rotamers from MD simulations (Tested on R
# version 3.3.2). Using CPPTRAJ, trajectories are saved as separate pdb files for each frame (e.g. MD.pdb.1,
# MD.pdb.2, MD.pdb.3, etc...). These files will be used as input and should be in the working directory (D:/).
# Step 1: Writing automated commands for saving torsional angles of 50,000 frames and saving the commands into
# variable (variable name: x), then saving the variable in file (filename: save_tor.R). For each frame, the
# command requests that torsions are saved into a variable called tor (e.g. tor1, tor2, tor3, etc...).
# The automated commands in file (save_tor.R) are excecuted by using source() function.
library(bio3d)
x <- capture.output(x <- c(for(i in 1:50000) {cat(paste('tor',i,' <- torsion.pdb(pdb <-
read.pdb("D:/MD.pdb.',i,'", maxlines = -1, multi = FALSE, rm.insert = FALSE, rm.alt = TRUE,
ATOM.only = FALSE, hex = FALSE, verbose = TRUE))','\n', sep=""))}), file=NULL)
write(x, file = "D:/save_tor.R", sep="")
source("D:/save_tor.R")
# Step 2: Writing automated commands for collecting torsional angles from all frames and saving the commands
# into variable (variable name: y). Adding remaining part of command into variable (variable name: z). The
# commands are saved in R script file (filename: select_tor.R), then excecuted via the source() function. The
# torsions of residue number 3 were saved here into variable (variable name: tor_residue3). To choose another
# residue number, replace the number 3 in the phrase $tbl[3,] to describe another residue number. Usually, R
# designates successive numbering to the following chains: e.g. if the structure contains two chains of 100
# residues length, the first residue of the second chain will have the number 101.
y <- capture.output(x <- c(for(i in 2:50000) {cat(paste(",tor",i,"$tbl[3,]", sep=""))}), file=NULL)
z <- c(paste("tor_residue3<- rbind(tor1$tbl[3,]", y, ")", sep=""))
write(z, file = "D:/select_tor.R", sep="")
source("D:/select_tor.R")
# Step 3: Torsions saved in file (filename: tor_residue3.txt).
write.table(tor_residue3, "D:/tor_residue3.txt", sep="\t")
# Steps 2-3a: To collect torsional angles for specified number of residues at once, the following FOR statement
# can be used. The residues are defined in variable (variable name: j). The example extracts the torsions of
# residues 1 to 20.
for (j in 1:20) {
y <- capture.output(x <- c(for(i in 2:50000) {cat(paste(",tor",i,"$tbl[",j,",]", sep=""))}),
file=NULL)
z <- c(paste("tor_residue",j,"<- rbind(tor1$tbl[",j,",]", y, ")", sep=""))
w <- capture.output(cat(paste("D:/select_tor",j,".R", sep="")), file=NULL)
write(z, file = w, sep="")
source(w)
write.table(c(paste("tor_residue",j, sep="")), c(paste("D:/tor_residue",j,".txt", sep="")),
sep="\t")
}
# step 4: Classification of Rotamers. The torsion file is composed of the following columns: frame, phi, psi,
# chi1, chi2, chi3, chi4, chi5. The algorithm thus will classify the rotamers according to the ranges of chi
# angles in the Penultimate Rotamer Library (SC Lovell, JM Word, JS Richardson and DC Richardson (2000) "The
# Penultimate Rotamer Library" Proteins: Structure Function and Genetics 40: 389-408. Library is available at
# http://kinemage.biochem.duke.edu/downloads/PDFs/Complete_rotamer_lib.pdf). It is important to understand that
# the angles described in the library pass the 180 to -180 border on a circle, therfore all arguments involving
# that range are expressed by OR (|) statements to cover angles <=180 or >= -180. The t80 rotamer in Tyrosine is
# one example for this, where the Chi1 angle is in the range -145 to 155. In the script it means the ranges <=
# -145 OR >= 155. Here, rotamers outside the range are designated with a zero.
tor <- read.delim("D:/tor_residue1.txt", header = TRUE, sep = "\t", dec = ".")
Rota_residue1 <- matrix(data=NA,nrow=50000,ncol=1)
# Step 5: Example of rotamer classes (other residues are described in supplementary material):
# Rotamers of Ser, Cys, Thr (3 groups)
for (i in 1:50000) {
if ((tor[i,3]>=30) & (tor[i,3]<=90)){Rota_residue1[i,1] <- "p"
} else if ((tor[i,3]>=155)|(tor[i,3]<=-145)){Rota_residue1[i,1] <- "t"
} else if ((tor[i,3]>=-95) & (tor[i,3]<=-35)){Rota_residue1[i,1] <- "m"
} else {Rota_residue1[i,1] <- 0}
}
write.table(Rota_residue1, "D:/Rota_residue1.txt", sep="\t")
```

Figure 3

Representative example of rotamer analysis. An implicit MD simulation of neurotrophic peptide and its receptor done in free and bound states. (A) pNGF peptide (SSSHPIFHRGEFSV$_{\text{-NH2}}$) structure shown in green ribbon. (B) Secondary structure-Rotamer relationship in H4 residue from peptide. (C) Secondary structure-Rotamer relationship in P5 residue from peptide. (D) Part of the TrkA receptor (in orange ribbon) binding to the pNGF peptide (in green ribbon). (E) Four residues at the binding interface with distinct rotamer relations. H4 from peptide acquired the t-80 rotamer. P5 from peptide acquired the Cγ exo rotamer. S304 from the receptor acquired both m and t rotamers to accommodate both adjacent histidines. H343 from the receptor acquired the m-70 rotamer. (F) Secondary structure-Rotamer relationship in V294 residue from receptor. (G) Secondary structure-Rotamer relationship in H298 residue from receptor. (H) Secondary structure-Rotamer relationship in S304 residue from receptor. (I) Secondary structure-Rotamer relationship in F329 residue from receptor. Crystal structure (PDB ID 2IFG) was used. The structure was edited using UCSF Chimera and processed via H++ server. MD simulations were performed in implicit water in AMBER 14. Rotamer analysis was done in R language as described in the text.

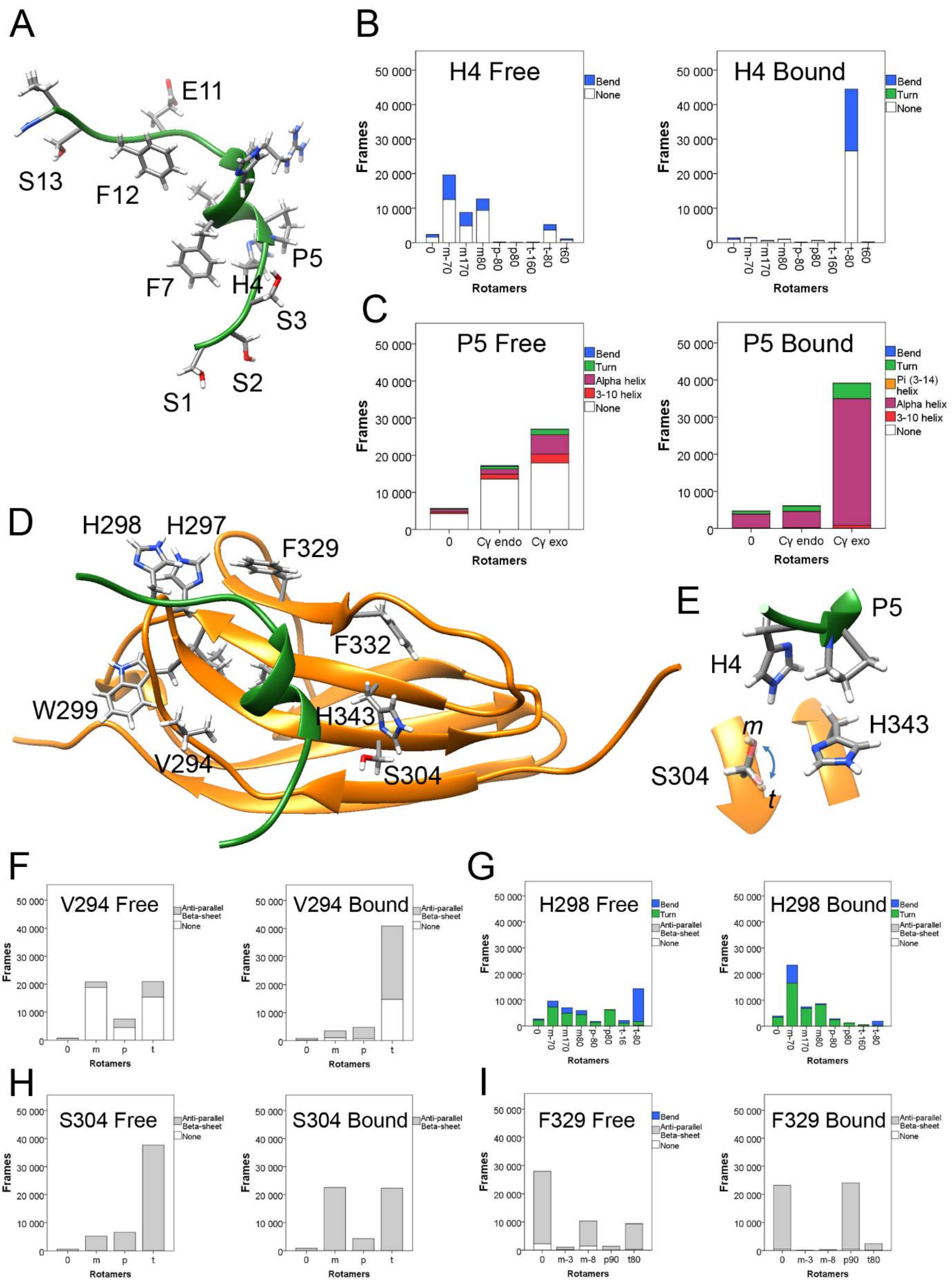
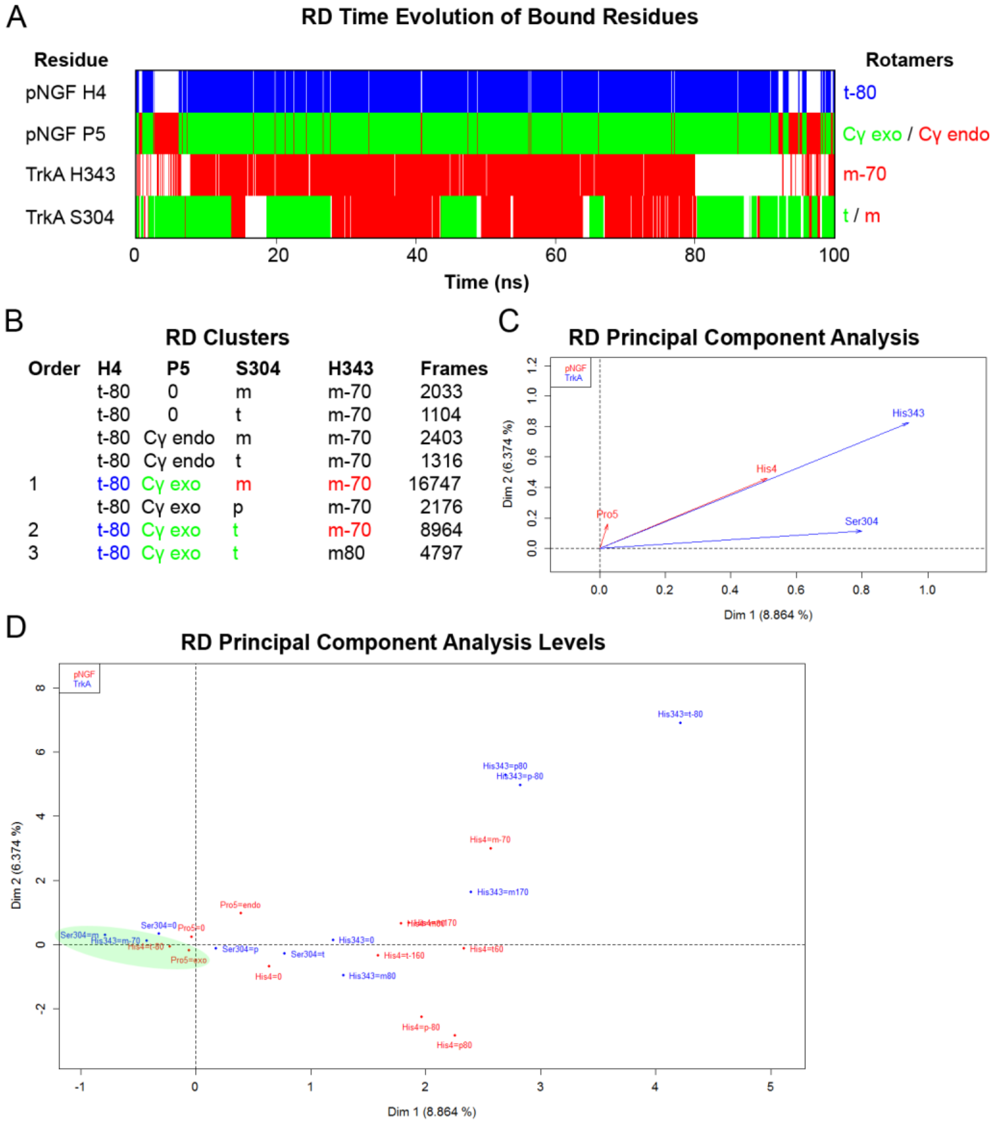

A
E11
S13
F12
F7
H4
P5
S3
S2
S1
B
H4 Free
Bend
None
Frames
50 000
40 000
30 000
20 000
10 000
0
0
m-70
m170
m80
p-80
p80
t-160
t-80
t60
Rotamers
H4 Bound
Bend
Turn
None
C
P5 Free
Bend
Turn
Alpha helix
3-10 helix
None
0
Cγ endo
Cγ exo
Rotamers
P5 Bound
Bend
Turn
Pi (3-14) helix
Alpha helix
3-10 helix
None
D
H298
H297
F329
F332
W299
H343
V294
S304
E
P5
H4
m
H343
S304
t
F
V294 Free
Anti-parallel Beta-sheet
None
0
m
p
t
Rotamers
V294 Bound
Anti-parallel Beta-sheet
None
G
H298 Free
Bend
Turn
Anti-parallel Beta-sheet
None
0
m-70
m170
m80
p-80
p80
t-16
t-80
Rotamers
H298 Bound
Bend
Turn
Anti-parallel Beta-sheet
None
0
m-70
m170
m80
p-80
p80
t-160
t-80
H
S304 Free
Anti-parallel Beta-sheet
None
S304 Bound
Anti-parallel Beta-sheet
I
F329 Free
Bend
Anti-parallel Beta-sheet
None
0
m-3
m-8
p90
t80
Rotamers
F329 Bound
Anti-parallel Beta-sheet
None

Figure 4

RD data visualization and analysis of the bound residues H4 and P5 in pNGF, and H343 and S304 in TrkA. (A) RD time evolution graph for most frequent rotamers. The graph was generated using image() function in gplots module in R language. With exception for TrkA S304 residue, the other residues folded into stable rotamer conformation after 10 ns and continued for most of the simulation. (B) Using count() function in plyr module in R, the number of frames was calculated for each cluster of rotamers. The table shows the highest eight clusters. (C) Multiple factor analysis for mixed data generated using MFAmix() function in PCAmixdata module in R. The graph shows squared loadings of variables. Based on the vector angles, it is very clear that His4 and His343 were much correlated with each other in the two dimensions. (D) Component map of the levels showing individual rotamer. Representative correlated rotamers are shown in green ellipse.



RD Clusters

Order	H4	P5	S304	H343	Frames
	t-80	0	m	m-70	2033
	t-80	0	t	m-70	1104
	t-80	Cγ endo	m	m-70	2403
	t-80	Cγ endo	t	m-70	1316
1	t-80	Cγ exo	m	m-70	16747
	t-80	Cγ exo	p	m-70	2176
2	t-80	Cγ exo	t	m-70	8964
3	t-80	Cγ exo	t	m80	4797