Study of the interaction between a novel, protein stabilizing dipeptide and Interferon-alpha-2a by construction of a Markov State Model from Molecular Dynamics simulations

Andreas Tosstorff[1*], Günther H.J. Peters[2], Gerhard Winter[1]

[1]Department of Pharmacy, Pharmaceutical Technology and Biopharmaceutics, Ludwig-Maximilians-Universität München, Munich, Germany

[2]Department of Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[*]Contact: andreas.tosstorff@cup.uni-muenchen.de

Abstract

We recently reported the discovery of a novel protein stabilizing dipeptide, glycyl-D-asparagine, through a structure-based approach. As the starting hypothesis leading to the discovery, we postulated a stabilizing effect achieved by binding of the dipeptide to an aggregation prone region on the protein's surface. Here we present a detailed study of the interaction mechanism between the dipeptide and Interferon-alpha-2A (IFN) through the construction of a Markov state model from molecular dynamics trajectories. We identify multiple binding sites and compare these to aggregation prone regions. Additionally, we calculate the lifetime of the protein-excipient complex. If the excipient remained bound to the IFN after administration, it could alter the protein's therapeutic efficacy. We establish that the lifetime of the complex between IFN and glycyl-D-asparagine is extremely short. Under these circumstances, stabilization by stoichiometric binding is consequently no impediment for a safe use of an excipient.

Keywords

Interferon-alpha-2a, Excipient, Protein Aggregation, Protein Formulation, Markov State Model

Introduction

Small molecules are commonly found in therapeutic protein drug formulations as co-solutes with the intend to stabilize the drug product among other against chemical degradation or aggregation of the therapeutic protein. Opposed to native self-association, protein aggregation proceeds by multiple steps that among other involve a partial or complete unfolding of the protein[1].

Two commonly accepted mechanisms of stabilization of a protein against aggregation by a small molecular co-solute are preferential exclusion and stoichiometric binding [2–6]. Preferential exclusion describes an entropically driven rise of chemical potential of both, protein and co-solute molecules relative to their separate solutions. The increase in chemical potential manifests by a reduced concentration of co-solute in proximity to the protein surface relative to the bulk solution. Protein unfolding will lead to an increased exposure of protein surface, increasing the unfavorable exclusion of co-solute. The protein's native state is therefore preferred to the non-native. The stabilizing effect of a diverse group of co-solutes such as sugars, polyols, amino acids, methylamines and inorganic salts on proteins has been well established and traced back to preferential exclusion as mechanism of action [2,7]. Preferential exclusion is observed for weakly interacting co-solutes that require to be present at high concentration (above 200 mM) in order to benefit protein stability.

Stoichiometrically interacting co-solutes are known to stabilize proteins by binding preferentially to the native protein structure relative to the unfolded one, which can for example be determined by differential scanning fluorimetry or calorimetry and results in a shift of the infliction point of the characteristic unfolding curve (T_m) [8].

The large majority of pharmaceutical excipients act through the mechanism of preferential exclusion, which has the intrinsic benefit that their application is not limited to a single protein but across many if not all. Developing excipients that act as stoichiometric stabilizers has largely been neglected, despite the potential to provide a complementary mean to stabilize a protein[9].

We previously described the discovery of an outstanding stabilizing effect of the dipeptide glycyl-D-asparagine at low concentrations against aggregation of Interferon-alpha-2A upon exposure to freezing-thawing and shaking stress[10]. We found that the dipeptide would bind to the protein at a µM affinity and reduces particle formation at low concentration (6.25 mM), hinting at a stabilization through a stoichiometric interaction. The compound was discovered through a virtual screen that targeted the hydrophobic and solvent exposed residue Phe27. This residue is involved in the interaction between interferon-alpha-2 and interferon-alpha-receptor 2 (Figure 1, PDB entry 3S9D) [11]. A potential risk of stoichiometrically acting excipients is that the protein drug-excipient complex does not disassociate after drug administration, thus potentially altering the drug's efficacy. The lifetime of the protein-excipient complex is therefore a crucial parameter to consider when developing stoichiometrically binding excipients. A short lifetime means that the protein-excipient complex disassociates rapidly. As the excipient is much smaller than the protein, it will distribute, metabolize and clear much faster than the protein after administration. In the case of the dipeptide presented here, its metabolism is facilitated further due to the presence of a peptide bond prone to enzymatic hydrolysis [12]. Its low molecular weight compared to that of IFN will lead to a fast clearance through the kidneys [13]. A long lifetime of the protein-excipient complex would instead result in a permanent occupation of the protein surface by one or more excipient molecules, potentially altering the proteins interaction with its target molecule, and consequently its efficacy.

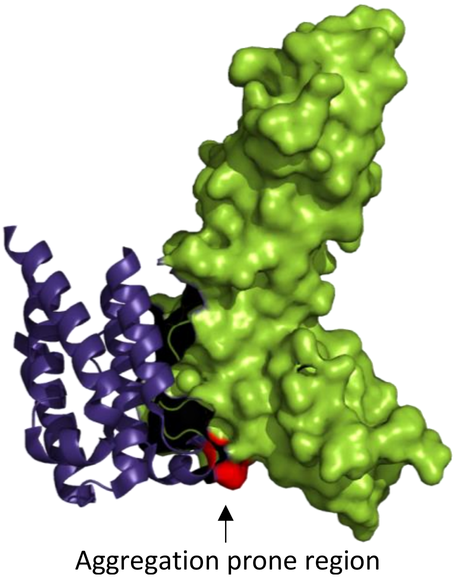


Figure 1: Complex between Interferon-alpha-2 (violet) and Interferon-alpha-receptor-2 (green) (PDB entry 3S9D). The aggregation prone region (APR) targeted by the excipient (red) to inhibit interferon-alpha-2a aggregation coincides with the binding site to the receptor.

The occupation of a protein by a ligand is the result of the simultaneously occurring binding and unbinding processes [14]. When considering the equilibrium reaction between Protein P and ligand L to form a complex PL (Equation 1), the rates of binding, r_{on}, and unbinding, r_{off} can be defined as the product of a rate constant k and the concentration of the reactants (Equation 2, Equation 3).

$$P + L \rightleftharpoons PL \qquad \textit{Equation 1}$$

$$r_{on} = k_{on} \cdot [P] \cdot [L] \qquad \textit{Equation 2}$$

$$r_{off} = k_{off} \cdot [PL] \qquad \textit{Equation 3}$$

Here, k_{on} and k_{off} are the rate constants for the corresponding binding and unbinding reactions and $[P]$, $[L]$, $[PL]$ are the concentration of the protein, ligand and protein ligand complex respectively.

In order to estimate the lifetime of a protein-ligand complex, the residence time τ can be calculated from the inverse of the off-binding rate constant k_{off} (Equation 4) [15].

$$\tau = \frac{1}{k_{off}} \qquad \textit{Equation 4}$$

Computational simulations are a popular mean to study protein-ligand interactions, as they allow to gain insights on the interaction with atomic detail. Interactions between the excipients mannitol, sucrose, trehalose and sorbitol and a ligase and a Fab fragment have previously been studied by docking calculations [16]. In this work a correlation between calculated binding affinity of excipients to the protein and T_m was observed. The T_m experiments were, however, conducted at excipient concentrations ranging from 145 to 220 mM, which may hint at a stabilization by preferential exclusion. A method that combines protein-protein and protein-excipient docking combined with molecular dynamics (MD) simulations to discover new excipients is described as well in a patent application [17]. It aims at identifying excipients that bind to regions involved in protein self-association in order to reduce protein aggregation. It does not state how protein aggregation is measured experimentally and does not relate simulation data to data from experiments on protein aggregation. It does furthermore not yield any novel excipients but is limited to commonly employed stabilizing substances such as amino acids.

MD simulations are a mean to study protein ligand interactions at atomic detail, where each atom is treated as a classical particle and interactions between these particles are defined in force fields[18]. Shukla and Trout used MD simulations to determine the preferential interaction coefficient of a protein in aqueous arginine solutions of 250 to 2500 mM [19]. The study of stoichiometric binding by molecular dynamics is most commonly reported in the context of small molecule drug discovery. Analysis of molecular trajectories, which are often collected in parallel setups is challenging and can introduce errors due to biases in the starting structure and introduced restraints intended to enhance sampling of rare transitions. Markov state theory has been used in trajectory analysis to eliminate these biases and accurately describe the mechanism of protein-ligand interactions [20]. Markov state theory serves to model transitions between discrete states. The modeled process is considered memoryless, meaning that if the system is in a specific state, its future state does not depend on the system's history. The publicly available EMMA and HTMD programs drastically facilitate the construction of Markov state models from MD trajectories [21]. In order to construct a Markov state model, MD trajectories have to be discretized. Discretization for Markov state model generation has been shown to work best when the dimensionality of the trajectory data is reduced. In an MD simulation, each simulated atom is described by three cartesian coordinates indicating its position, and three cartesian velocities, indicating its current movement in three-dimensional space. One frame of an MD trajectory therefore consists of 6N dimensions, where N is the number of simulated atoms. By identifying a set of features of lower dimensionality, such as dihedral angles or the distance between the ligand and each protein residue, the complexity of an MD trajectory can be reduced, but the information of interest, e.g. protein conformation or protein ligand binding, is preserved. Mathematical approaches to reduce the dimensionality of a matrix are principle component analysis, which preserves the highest degree of variance or time lagged independent component analysis which preserves the highest degree of kinetic variance. The first principal component will therefore describe the motion of highest amplitude, while the first time lagged independent component will describe the slowest transition. After reduction of dimensionality, the data set of reduced dimensionality has to be discretized by a clustering algorithm. Finally, the transition probabilities between the clustered states can be calculated and experimental observables can be derived. Detailed descriptions on the workflow to construct a Markov model has been published by the developers of EMMA[22].

To our knowledge, Markov state theory has not yet been employed to describe protein-excipient interactions. Here we use a Markov state model to investigate the mechanism of interaction between the stabilizing dipeptide glycyl-D-asparagine and Interferon-alpha-2A, to elucidate interaction sites and to estimate the residence time of the formed protein-excipient complex.

Methods

System setup and simulation

Each randomized starting systems was constructed using HTMD [23]. One of 24 structures from PDB entry 1ITF was randomly selected using NumPy's random.choice function [24]. The protonation states of the protein were adjusted to pH 7.0. The protein was centered and randomly rotated. Subsequently the ligand was centered, randomly rotated and placed at a random distance between 6 to 11 Å away from the furthest protein atom along the x-axis (Figure S 1). The ligand was again rotated randomly around the origin. The system was then solvated with an additional 5 Å buffer. Finally, two disulfide bridges were built.

The ligand was parametrized using GAFF2 for bonded and non-bonded parameters. Atomic partial charges were calculated with Gaussian 16 (Gaussian Inc., Wallingford, CT, U.S.A.) and fitted with the RESP procedure in antechamber. Each system was minimized and equilibrated prior to the production run. Minimization was performed using pmemd on CPUs, whereas molecular dynamics simulations were performed on GPUs using pmemd.cuda implemented in Amber 18 [25–28]. A cutoff of 9 Å was defined for nonbonded interactions. The first 5000 cycles of minimization used the steepest descent algorithm, followed by 5000 cycles using the conjugate gradient algorithm. MD simulations were run using Langevin dynamics with a collision frequency of 1 ps^{-1}[29]. The SHAKE algorithm was used to allow for timesteps of 2 fs[30].

Equilibration followed the scheme described by Henriksen et al. and consisted of three steps [31]. For 1 ps, no pressure scaling was used and the temperature was set to 10 K. The system was then heated to 300 K within 100 ps. The last stage consisted of 50 equilibration cycles of 100 ps, each using a Monte

Carlo barostat set to atmospheric pressure. Production was performed using the NVT ensemble, running 60 ns per trajectory. 600 trajectories were generated in total.

Data analysis

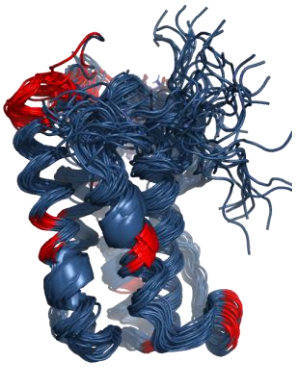
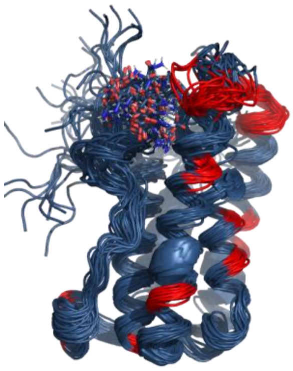
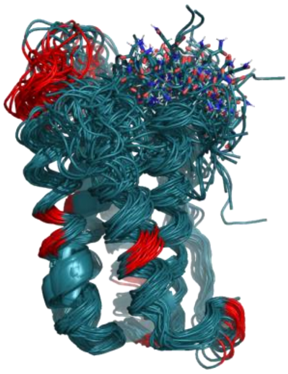
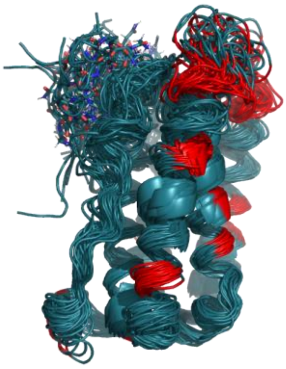
A Markov State Model was constructed using HTMD which builds on PyEMMA. We followed a stepwise approach based on the multiple tutorials accompanying HTMD and PyEMMA. Trajectories were first stripped of all water, sodium and chloride. The selected featurization scheme to study the protein-ligand interaction was the pairwise, residual, minimum distance between each protein residue and the dipeptide, considering only heavy atoms. The data was then projected on the first 10 time-lagged independent components with a lag time of 1 ns. The projected data was then clustered into 60 micro-states using the k-means algorithm. A Markov state model was constructed with a lag time of 10 ns and the micro-states were clustered to 5 macro-states using PCCA++. The model was validated using the Chapman-Kolmogorov (CK) criterion. If the model fulfills the CK criterion, the occupation of future states is independent of past states, i.e. the model is markovian. (Figure S 2). Statistical errors of thermodynamic and kinetic quantities were obtained from 1000 bootstrapping cycles retaining 80% of the data. Structures were rendered using PyMOL.

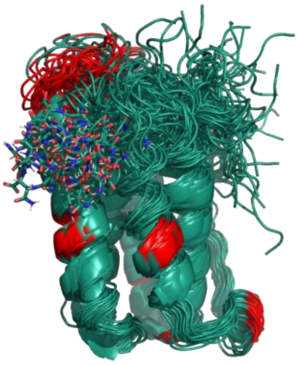
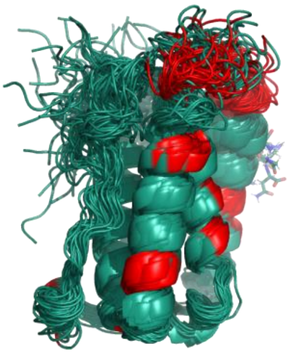
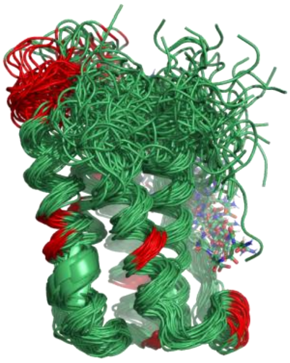
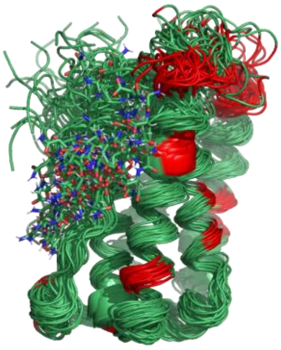
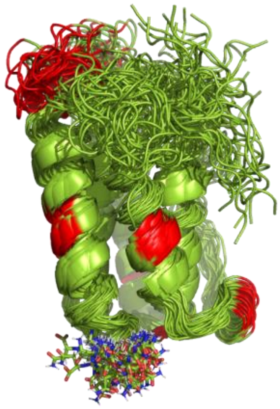
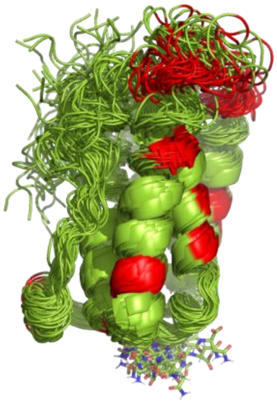
Identification of aggregation prone regions

Three different methods were used to identify aggregation prone regions on the surface of Interferon-alpha-2A: Aggrescan3D[32], AggScore[33] and CamSol[34]. For Aggrescan3D and CamSol the scores were calculated by submitting the first frame of PDB entry 1ITF to the corresponding webserver. The aggregation propensity according to the AggScore method was calculated using Schrödinger's Maestro software using the same structure file as for the 2 other methods. Aggregation prone residues identified through any of the methods are residues 16, 27, 61,65, 86, 89, 98, 99, 100, 101, 102, 103, 106, 109, 110, 111, 116, 117, 128, 129, 137.

Results

From the constructed Markov model, 5 macro states were identified. State 5 comprises mostly unbound and non-specifically associated structures. States 0 to 4 show specific regions of interaction between the dipeptide and INF with different degrees of fuzziness. Macro-state 0 involves interactions with residues 41, 42, 43, 46, 48, 51, 114, 115, 164. Macro-state 1 can be characterized by interactions with residues 3, 40, 41, 45-49, 155-165. For Macro-state 2, residues 5-10, 13, 90, 91, 93, 94, 96, 147 were identified. In macro-state 3, the dipeptide is in contact with residues 33-38, 40, 41, 42, 46, 114, 118, 121, 122, 125, 146, 149, 165. Macro-state 4, which is the least fuzzy one, only involves residues 22, 23, 73, 75-78 (Figure 2). While the study of protein conformation was not the scope of this study, we observed high flexibility in the N-terminal and the C-terminal loop region as was already described previously [24]. Interactions with the C-terminus are consequently present in multiple of the macro-states. When comparing the binding sites to aggregation prone regions identified on the protein surface, we find that macro-state 0 and 2 show an interaction close to the aggregation prone residues 98 to 100 (predicted by Aggrescan3D, AggScore, CamSol). Macro-state 3 shows an interaction in close proximity to the aggregation prone residues 27 (predicted by Aggrescan3D, AggScore, CamSol), 128 and 129 (predicted by AggScore). Macro-state 4 shows binding in proximity to aggregation prone residue 137 (predicted by AggScore).

Macro-state 0		
Macro-state 1		

Macro-state 2		
Macro-state 3		
Macro-state 4		

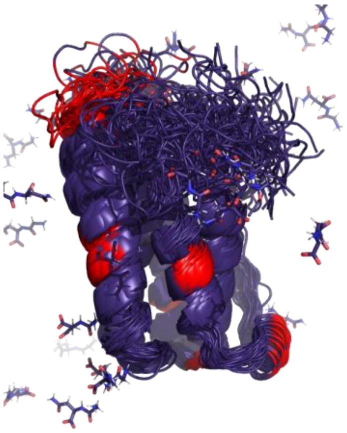
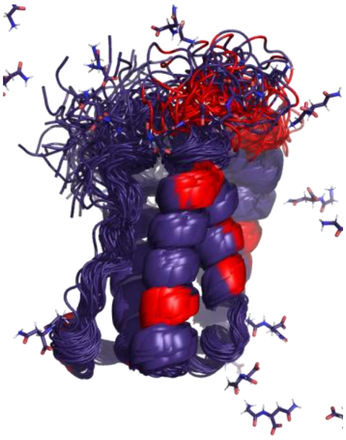
Macro-state 5		

Figure 2: Representative structures from two perspectives at 180° rotation of macro-states defined by the constructed Markov model. Aggregation prone residues: 16, 27, 61,65, 86, 89, 98, 99, 100, 101, 102, 103, 106, 109, 110, 111, 116, 117, 128, 129, 137.

When comparing the docked structure that led to the discovery of the dipeptide as protein stabilizing substance, one observes a similarity to macro-state 3. In both, the docked pose as well as in macro-state 3, interactions with residues 33, 34 and 146 are observed. The interaction between ARG 33 and the dipeptide in both cases consists of a salt bridge between the residue's side chain and the dipeptide's carboxyl group (not depicted for macro-state 3). In the docked pose, the interaction with residue 34 is between the backbone carbonyl group and the dipeptides N-terminal amine. In the MD simulation, the amide nitrogen of residue 34 interacts with the dipeptide's amide carbonyl group. The docked pose suggests a hydrogen bond between the side chain carboxyl of GLU146 and the dipeptide's amine, which is also observed in the third macro-state. The docked pose shows the ASN side chain of the dipeptide forming a hydrogen bond with the backbone carbonyl of ALA145, which is not the case in the structures sampled from macro-state 3 (Figure 3).

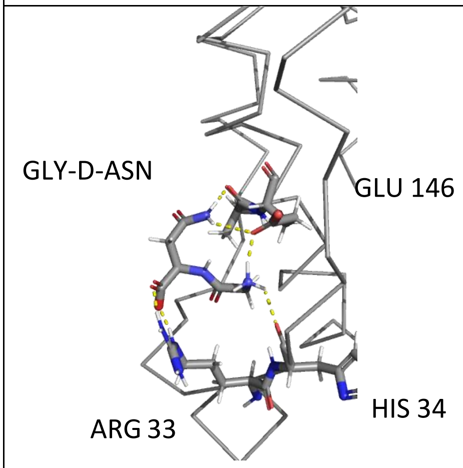
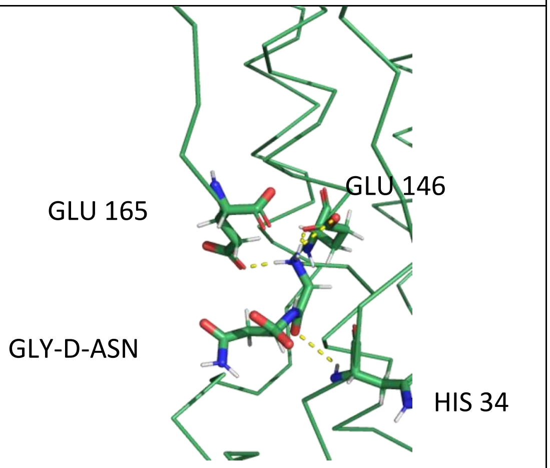
Docked pose	Structure sampled from macro-state 3
	

Figure 3: Comparison of docked pose with the most similar structure of those sampled from macro-state 3. Interacting residues are represented as sticks.

The Markov model exposes the binding path of the dipeptide, which most frequently transitions from macro-state 5 to 4, occasionally passing through state 3, which acts as an intermediate. The very infrequently occupied states 0, 1 and 2 are all connected to state 3 and are occasionally visited before the dipeptide moves along to states 3 and 4. The predicted residence time is calculated to be 0.03 µs and the equilibrium dissociation constant shows a weak binding of 29 mM compared to the µM affinity observed experimentally[10].

Table 1: Observables derived from the Markov model and experimentally observed dissociation constant for the interaction between IFN and Gly-D-Asn.

k_{on}	313 ± 201 µM^{-1} s^{-1}
k_{off}	30 ± 16 µ s^{-1}
τ	0.03 ± 0.02 µ s
K_D	29 ± 12 mM
K_D experimental	0.11 mM ± 0.02 mM

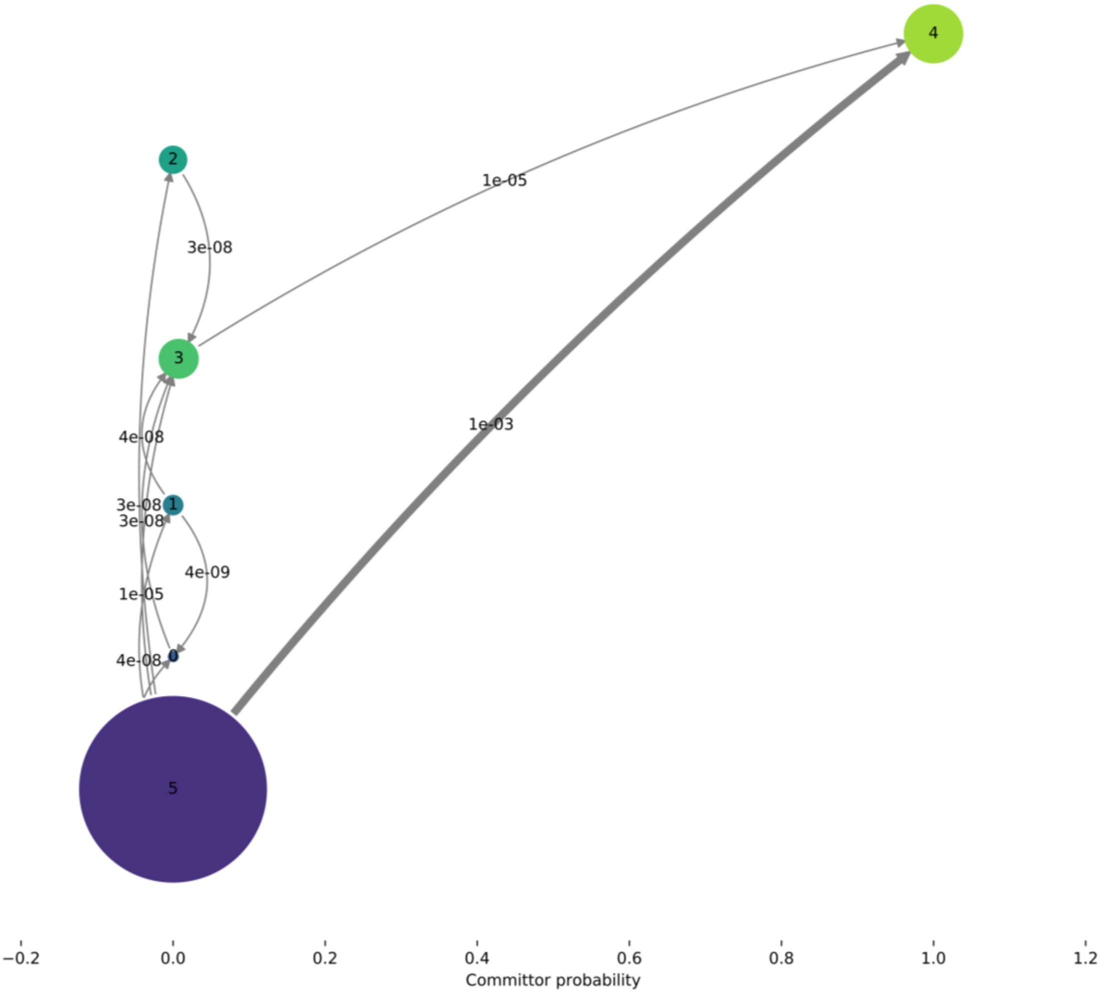


Figure 4: Markov processes can be visualized as a network of macro-states. Each circle represents a macro-state, which in our case corresponds to the ligand occupying a specific binding site (macro-states 0-4) or being unbound (macro-state 5). The areas of the circles are proportional to the stationary probability of the macro-state. Transitions between macro-states are visualized by arrows. Their thickness represents the probability of the transition to occur. The transition probability is also written on top of the arrows. The committor probability describes how likely it is that the system changes to the target state 4 (sink), or to the original state 5 (source). If the committor probability is close to 1, the system will move towards the sink. If it is close to 0, the system will move towards the source. One can therefore conclude that when the ligand is bound to the protein in one of the four macro-states, it will most likely unbind (i.e. transition to macro-state 5) before occupying another bound macro-state.

Discussion

Here, we use Markov theory for the first time to describe the interaction between a stabilizing small molecule and a therapeutic protein. The use of molecular dynamics simulations to study the interaction had two purposes. On the one hand, we wanted to identify the excipient's favored interaction sites and compare it to the protein's aggregation prone regions. On the other hand, we wanted to estimate the residence time of the protein-ligand complex to rule out any effect of the excipient on the drug protein's efficacy after administration.

We identified five meta-stable interaction sites showing hydrogen bonding and salt-bridges between the protein and the dipeptide, supporting the finding of stoichiometric binding between thee protein and the ligand. The protein-ligand complex formed in macro-state 3 is similar to the one that was proposed by our previously reported virtual screen[10]. In our Markov model, the macro-state 3 is, however, only a weakly populated intermediate state. Despite substantial sampling, we were not able to reproduce the experimentally observed dissociation constant. We can consequently conclude, that the simulations do not elucidate the interaction process in its entirety.

We find that in all 5 bound macro-states, the binding site is in proximity to at least one aggregation prone region. Considering the overall hydrophobicity of Interferon-alpha-2A and the implied presence of multiple of such aggregation prone regions, it seems difficult to consider this observation to be significant, since almost any binding site is likely to be close to an aggregation prone region. Therefore, the simulations on the one hand support our hypothesis of stabilization by stoichiometric binding, on the other hand it neither proves nor disproves that the proximity to an aggregation prone region is the cause for the stabilization. Obtaining a crystal structure of the protein-ligand complex would be highly desirable to further evaluate the model.

The residence time estimated by our model is extremely low, indicating that there is no threat to an altered efficacy caused by a specific protein-excipient interaction since the complex will rapidly disassemble after administration. Since diffusion and distribution of small molecules is of orders of magnitudes faster than that of proteins, equilibrium conditions after administration are no longer given. Considering the underestimation of the dissociation constant, a higher residence time than the one calculated could nevertheless be plausible.

Conclusion

We studied the interaction between the stoichiometric stabilizer glycyl-D-asparagine and Interferon-alpha-2A through the construction of a Markov state model from MD simulations. The binding mechanism is complex and involves interaction sites in proximity to aggregation prone regions. The calculated residence time is of 0.03 µs and does therefore emphasize the improbability of a distorted efficacy of the drug protein caused by a stoichiometric stabilizer.

Acknowledgements

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 675074. Molecular dynamics simulations were carried out on the Steno GPU cluster at DTU chemistry.

Competing interests

The authors declare no competing interests.

Corresponding authors

Correspondence to Andreas Tosstorff.

References

[1] C.J. Roberts, Therapeutic protein aggregation: Mechanisms, design, and control, Trends Biotechnol. 32 (2014) 372–380. doi:10.1016/j.tibtech.2014.05.005.

[2] J.F. Carpenter, J.H. Crowe, The mechanism of cryoprotection of proteins by solutes, Cryobiology. 25 (1988) 244–255. doi:10.1016/0011-2240(88)90032-6.

[3] S.N. Timasheff, Tsutomu Arakawa, Mechanism of Protein Precipitation and Stabilization by co-solvents, J. Cryst. Growth. 90 (1988) 39–46.

[4] S.N. Timasheff, Protein hydration, thermodynamic binding, and preferential hydration, Biochemistry. 41 (2002) 13473–13482. doi:10.1021/bi020316e.

[5] S.N. Timasheff, Thermodynamic binding and site occupancy in the light of the Schellman exchange concept, Biophys. Chem. 101–102 (2002) 99–111. doi:10.1016/S0301-4622(02)00188-6.

[6] J.A. Schellman, The Thermodynamic Stability of Proteins, Annu. Rev. Biophys. Biophys. Chem. 16 (1987) 115–137. doi:10.1146/annurev.bb.16.060187.000555.

[7] K. Shikama, T. Yamazaki, Denaturation of Catalase by Freezing and Thawing, Nature. 190 (1961) 83–84. doi:10.1038/190083a0.

[8] T.T. Waldron, K.P. Murphy, Stabilization of proteins by ligand binding: Application to drug screening and determination of unfolding energetics, Biochemistry. 42 (2003) 5058–5064. doi:10.1021/bi034212v.

[9] D. Matulis, J.K. Kranz, F.R. Salemme, M.J. Todd, Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor, Biochemistry. 44 (2005) 5258–5266. doi:10.1021/bi048135v.

[10] A. Tosstorff, H. Svilenov, G.H.J. Peters, P. Harris, G. Winter, Structure-based discovery of a new protein-aggregation breaking excipient, Eur. J. Pharm. Biopharm. 144 (2019) 207–216. doi:10.1016/j.ejpb.2019.09.010.

[11] C. Thomas, I. Moraga, D. Levin, P.O. Krutzik, Y. Podoplelova, A. Trejo, C. Lee, G. Yarden, S.E.

Vleck, J.S. Glenn, G.P. Nolan, J. Piehler, G. Schreiber, K.C. Garcia, Structural linkage between ligand discrimination and receptor activation by Type i interferons, Cell. 146 (2011) 621–632. doi:10.1016/j.cell.2011.06.048.

[12] H. Lochs, E.L. Morse, S.A. Adibi, Mechanism of hepatic assimilation of dipeptides. Transport versus hydrolysis, J. Biol. Chem. 261 (1986) 14976–14981.

[13] M. Ohlson, J. Sörensson, K. Lindström, A.M. Blom, E. Fries, B. Haraldsson, Effects of filtration rate on the glomerular barrier and clearance of four differently shaped molecules, Am. J. Physiol. Physiol. 281 (2001) F103–F113. doi:10.1152/ajprenal.2001.281.1.F103.

[14] J. Corzo, Time, the forgotten dimension of ligand binding teaching, Biochem. Mol. Biol. Educ. 34 (2006) 413–416. doi:10.1002/bmb.2006.494034062678.

[15] N. Ferruz, G. De Fabritiis, Binding Kinetics in Drug Discovery, Mol. Inform. 35 (2016) 216–226. doi:10.1002/minf.201501018.

[16] T.S. Barata, C. Zhang, P.A. Dalby, S. Brocchini, M. Zloh, Identification of protein-excipient interaction hotspots using computational approaches, Int. J. Mol. Sci. 17 (2016). doi:10.3390/ijms17060853.

[17] F. Insaidoo, D. Roush, In silico process for selecting protein formulation excipients, WO2017155840A1, 2017.

[18] A.T. Hagler, E. Huler, S. Lifson, Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals, J. Am. Chem. Soc. 96 (1974) 5319–5327. doi:10.1021/ja00824a004.

[19] D. Shukla, B.L. Trout, Preferential interaction coefficients of proteins in aqueous arginine solutions and their molecular origins, J. Phys. Chem. B. 115 (2011) 1243–1253. doi:10.1021/jp108586b.

[20] N. Plattner, F. Noé, Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models, Nat. Commun. 6 (2015). doi:10.1038/ncomms8653.

[21] M.K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, F. Noé, PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models, J. Chem. Theory Comput. 11 (2015) 5525–5542. doi:10.1021/acs.jctc.5b00743.

[22] C. Wehmeyer, M.K. Scherer, T. Hempel, B.E. Husic, S. Olsson, F. Noé, Introduction to Markov

state modeling with the PyEMMA software [Article v1.0], Living J. Comput. Mol. Sci. 1 (2019) 1–8. doi:10.33011/livecoms.1.1.5965.

[23] S. Doerr, M.J. Harvey, F. Noé, G. De Fabritiis, HTMD: High-Throughput Molecular Dynamics for Molecular Discovery, J. Chem. Theory Comput. 12 (2016) 1845–1852. doi:10.1021/acs.jctc.6b00049.

[24] W. Klaus, B. Gsell, a M. Labhardt, B. Wipf, H. Senn, The three-dimensional high resolution structure of human interferon alpha-2a determined by heteronuclear NMR spectroscopy in solution., J. Mol. Biol. 274 (1997) 661–675. doi:10.1006/jmbi.1997.1396.

[25] R. Salmon-Ferrer, A.W. Goetz, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with AMBER - Part II: Particle Mesh Ewald, J. Chem. Theory Comput. 9 (2013) 3878–3888. https://pubs-acs-org.emedien.ub.uni-muenchen.de/doi/pdf/10.1021/ct400314y (accessed March 24, 2018).

[26] S. Le Grand, A.W. Götz, R.C. Walker, SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations, Comput. Phys. Commun. 184 (2013) 374–380. doi:10.1016/j.cpc.2012.09.022.

[27] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, I. T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, C.L. P. Li, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, R. H. Wei, D.M. York, P.A. Kollman, Amber 18, (2018).

[28] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, P. Kollman, AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules, Comput. Phys. Commun. 91 (1995) 1–41. doi:10.1016/0010-4655(95)00041-D.

[29] S. Chandrasekhar, Stochastic Problems in Physics and Astronomy, Rev. Mod. Phys. 15 (1943) 1–89. doi:10.1103/RevModPhys.15.1.

[30] S. Miyamoto, P.A. Kollman, Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models, J. Comput. Chem. 13 (1992) 952–962. doi:10.1002/jcc.540130805.

[31] N.M. Henriksen, A.T. Fenley, M.K. Gilson, Computational Calorimetry: High-Precision Calculation of Host-Guest Binding Thermodynamics, J. Chem. Theory Comput. 11 (2015) 4377–4394. doi:10.1021/acs.jctc.5b00405.

[32] R. Zambrano, M. Jamroz, A. Szczasiuk, J. Pujols, S. Kmiecik, S. Ventura, AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures, Nucleic Acids Res. 43 (2015) W306–W313. doi:10.1093/nar/gkv359.

[33] K. Sankar, S.R. Krystek, S.M. Carl, T. Day, J.K.X. Maier, AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches, Proteins Struct. Funct. Bioinforma. 86 (2018) 1147–1156. doi:10.1002/prot.25594.

[34] P. Sormanni, F.A. Aprile, M. Vendruscolo, The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility, J. Mol. Biol. 427 (2015) 478–490. doi:10.1016/J.JMB.2014.09.026.

Supplementary Information

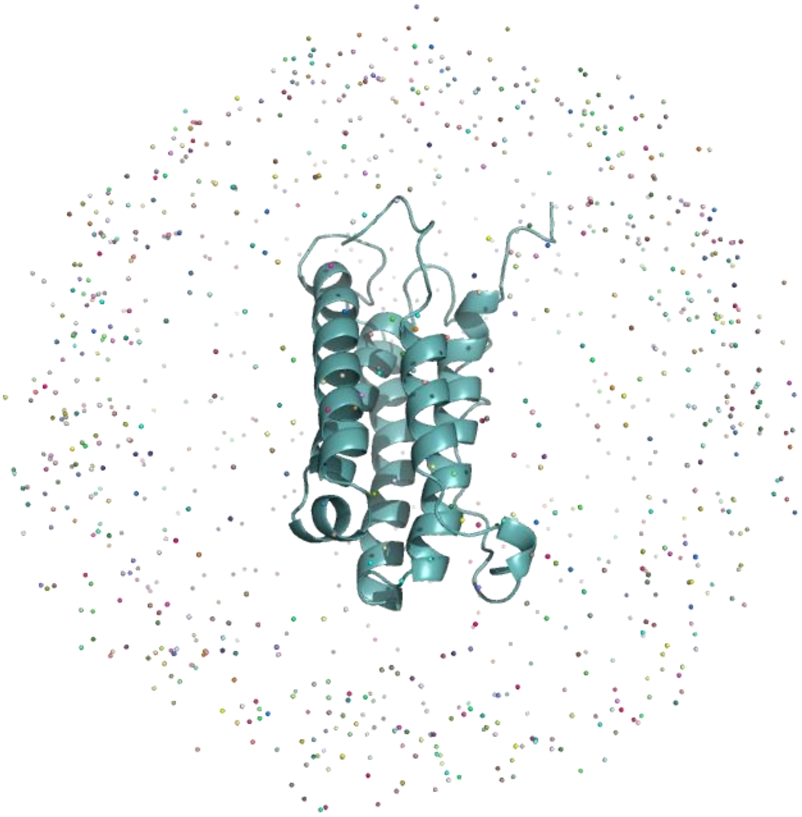


Figure S 1: Overlay of the position of the dipeptide in the starting structure for all 1000 simulations. Each dot represents the starting position of the ligand in the respective simulation.

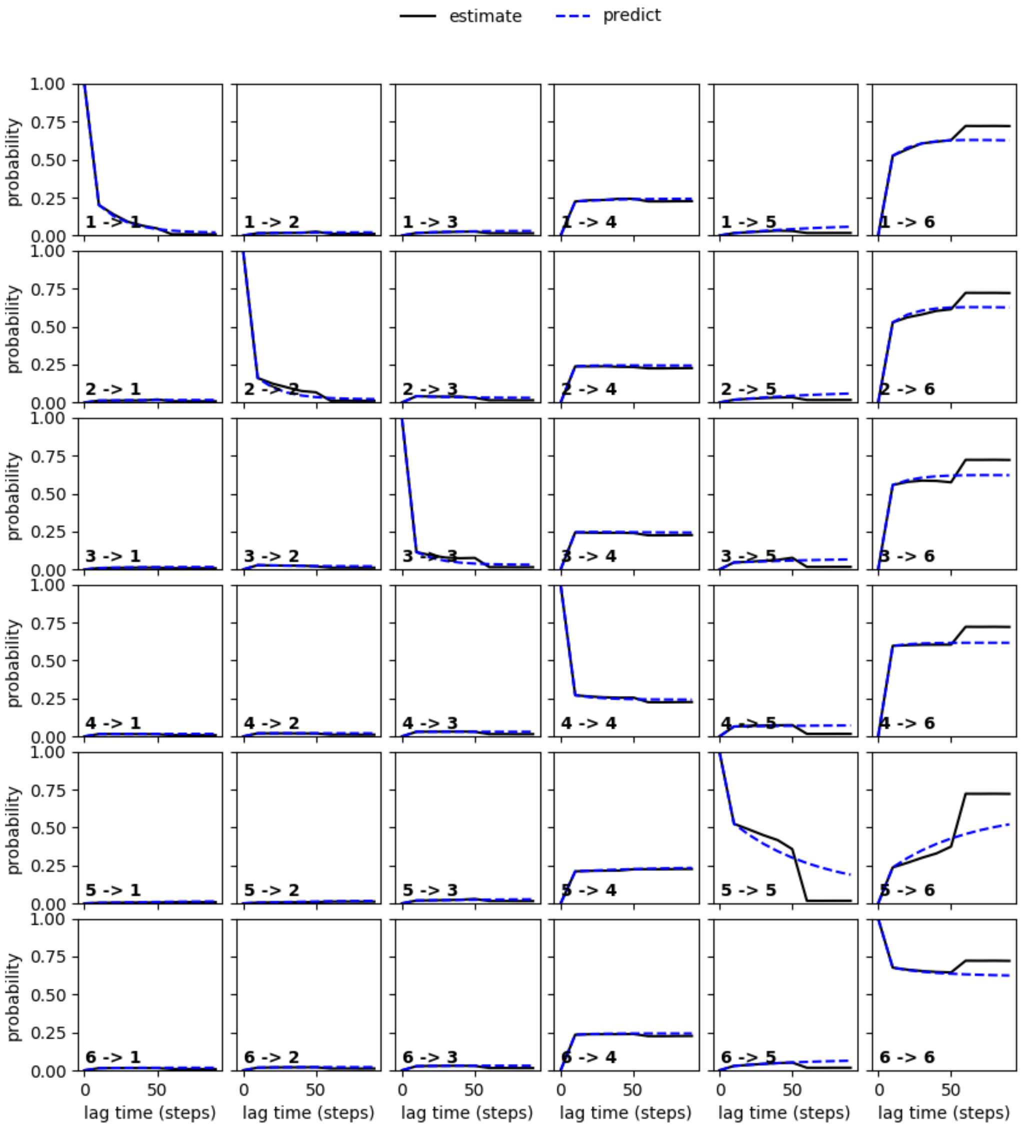


Figure S 2: Chapman-Kolmogorov (CK) test to confirm markovianity of the constructed model. The CK test reveals that the markovianity of the model is given for 50 time steps.