JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

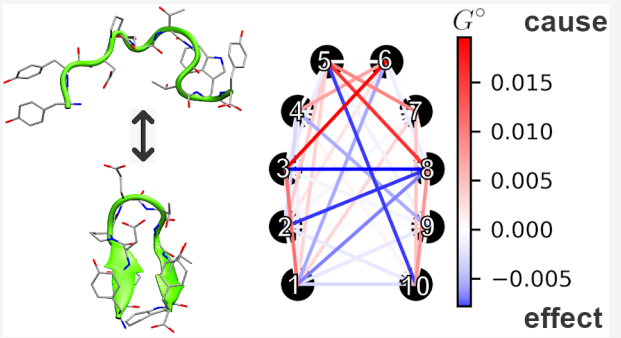
Granger Causality Analysis of Chignolin Folding

Marcin Sobieraj and Piotr Setny*

Cite This: J. Chem. Theory Comput. 2022, 18, 1936−1944 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Constantly advancing computer simulations of biomolecules provide huge amounts of data that are difficult to interpret. In particular, obtaining insights into functional aspects of macromolecular dynamics, often related to cascades of transient events, calls for methodologies that depart from the well-grounded framework of equilibrium statistical physics. One of the approaches toward the analysis of complex temporal data which has found applications in the fields of neuroscience and econometrics is Granger causality analysis. It allows determining which components of multidimensional time series are most influential for the evolution of the entire system, thus providing insights into causal relations within the dynamic structure of interest. In this work, we apply Granger analysis to a long molecular dynamics trajectory depicting repetitive folding and unfolding of a mini β-hairpin protein, CLN025. We find objective, quantitative evidence indicating that rearrangements within the hairpin turn region are determinant for protein folding and unfolding. On the contrary, interactions between hairpin arms score low on the causality scale. Taken together, these findings clearly favor the concept of zipperlike folding, which is one of two postulated β-hairpin folding mechanisms. More importantly, the results demonstrate the possibility of a conclusive application of Granger causality analysis to a biomolecular system.



1. INTRODUCTION

Molecular dynamics (MD) simulations provide increasingly comprehensive insights into the functioning of biomolecular systems.[1−3] One prominent area which has been fruitfully explored by means of MD is the problem of protein folding.[4−7] Numerous studies have demonstrated that polypeptide chains described by atomistic force fields can successfully reach experimentally determined native states guided solely by sequence-based effects.[8−11] Although less powerful in terms of practical ability to deliver sequence-based structure predictions compared to specialized approaches, in particular those based on extremely successful application of machine learning techniques,[12] MD simulations are unique in that they provide means to trace and, possibly, explain the details of the actual folding process. Still, even having access to atomistic, time-resolved folding trajectories does not always ensure unambiguous, objective interpretation of events that occur along the folding pathway or a clear understanding of the underlying biophysical driving forces.[13−15]

In this respect, even the folding of short β-hairpin structures is far from being clear.[16,17] Regarded as minimalistic protein models for their ability to achieve well-defined native states while having as few as 10 amino acids, they have been extensively investigated.[18−23] So far, two major theories concerning the sequence of events have been formulated. The first one assumes that the folding pathway starts with the appearance of a turn in the middle of the polypeptide chain and advances by outward propagation of hairpin contacts in a zipperlike manner.[24] The second one postulates the collapse of hydrophobic residues within hairpin arms as the primary event, followed by series of structural rearrangements that lead to turn stabilization and the formation of interstrand contacts.[25,26] Notably, both views received support from experimental and computational studies of a few β-hairpin structures, which possibly implies that in fact there is no single, uniform mechanism of β-hairpin formation.

Particularly well-suited systems for computer-based investigation of hairpin folding are a human-designed miniprotein called chignolin[27] and its later variant CLN025.[28] Both comprise only 10 amino acids, form stable hairpin structures amenable for nuclear magnetic resonance and X-ray crystallography, and, with experimentally confirmed folding times of only a few hundred nanoseconds, can be exhaustively sampled by fully atomistic simulations.[29−31] Accordingly, a number of studies have addressed chignolin and CLN025 reversible folding in an explicit aqueous solvent, finding support for both the zipperlike[29,32−34] and the hydrophobic

Received: September 20, 2021
Published: February 15, 2022

collapse driven[35] mechanisms, but also suggesting variability[9] and possible force field dependence of available pathways.[31,36]

Certainly the lack of consensus concerning chignolin and CLN025 folding, at least to some extent, stems from discrepancies between force field models. Whereas the native state is generally well captured by all major models, only very subtle differences are enough to change the properties of the unfolded ensemble[37] as well as the nature and sequence of intermediate states, as has been demonstrated in a recent study.[31] In addition to that, however, there exist only limited analysis options providing insights into detailed temporal characteristics of biomolecular structure rearrangements. Major efforts in this respect have been devoted to the development of Markov state models (MSMs) and associated methodologies for objective determination of relevant system representation.[38] The resulting framework allows assembling information from multiple, relatively short simulations into a complete kinetic model which can then be used to characterize significantly longer relaxation processes and their associated structural changes. While extremely powerful in determining pathways interconnecting meaningful system states and associated time scales, MSMs do not reveal, however, the significance of temporal relations between the occurring events. As a result, even having captured the kinetically relevant reaction coordinate, it is still difficult to determine which elements of system dynamics are of importance for its propagation along the path.

A possible way to draw conclusions about temporal relations in complex processes is provided by Granger causality analysis (GC). It was first proposed in 1969[39] and has found applications predominantly in economics, finance, and neuroscience.[40−45] Given a process described by multidimensional time series, GC determines whether, based on the knowledge of one time series, X, it is possible to probabilistically predict the behavior of another time series, Y. Causality considered in this way, expressed in the following as X Granger-causes (G-causes) Y, avoids the deeply philosophical question of the "true cause" of a given phenomenon and, more importantly, provides an effective statistical procedure for measuring the strength of temporal relationships. The original formulation of GC was founded on the framework of multivariate autoregressive models, thus relying on the existence of linear couplings within the system. The general idea was further extended to include nonlinear effects by means of information theory, employing transfer entropy[46] instead of time-shifted correlation to measure temporal dependencies between data channels.[47] Both formalisms have been applied to the analysis of molecular dynamics simulations utilizing source data in the form of time series describing fluctuations of atomic positions in Cartesian space,[48−50] residue-based fraction of native contacts,[51] or custom molecular descriptors.[52,53] Most recently, causal relations have been inferred from a transfer entropy measure which, instead of directly using time-resolved signals, operated on probability distributions involving elements of transition matrices representing local Markov state models, constructed for disjoined regions of a protein structure.[54]

In this article, we perform GC analysis of CLN025 folding based on a 106 μs long atomistic simulation performed by Lindorff-Larsen et al.[9] The resulting trajectory contains multiple folding and unfolding events allowing for parametrization of a converged GC model. Instead of focusing on individual residues, we consider inter-residue distances, as putative causal relations between them can be interpreted in terms of dependences between the formation and disruption of particular physical interactions in the course of the (un)folding process. We demonstrate that without any preliminary knowledge-based assumptions the model clearly favors one of the debated chignolin folding mechanisms and points to the importance of transient structural motifs in the unfolded ensemble for subsequent steps toward the native state.

2. METHODS

2.1. Multivariate Autoregressive Model. Multivariate autoregressive models (MVARs) are used to describe and analyze multidimensional temporal signals exploiting the existence of time-shifted correlations between individual components. In particular, they can be applied to forecast signal evolution based on linear combination of past values in respective channels. Given a multidimensional time series $\mathbf{x}(t) = \{x_1(t), ..., x_K(t)\}$ propagated with a time step Δt, a prediction, $\mathbf{x}_{\text{pred}}(t)$, of signal value at time t can be attempted based on P preceding steps in the following manner:

$$\mathbf{x}_{\text{pred}}(t) = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{x}(t - p\Delta t) \tag{1}$$

Here, $\mathbf{A}_p$ are $K \times K$ real matrices containing parameters defining an MVAR model of order P, which are determined under the assumption that the residual difference between real and predicted signals, $\Delta\mathbf{x}(t) = \mathbf{x}(t) - \mathbf{x}_{\text{pred}}(t)$, remains a white noise with stationary variance.

One of the possible ways to estimate model parameters is provided by the Yule−Walker method.[55,56] Having obtained matrices of mixed second statistical moments of signal components $\mathbf{\Gamma}(r) = \mathbf{\Gamma}^{\mathrm{T}}(-r) = \langle \mathbf{x}(t)\, \mathbf{x}^{\mathrm{T}}(t - r\Delta t)\rangle$ for $r \in \{0, ..., P\}$, the parameters are calculated by solving the following set of $P + 1$ linear equations (see the Supporting Information for practical details):

$$\mathbf{\Gamma}(r) = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{\Gamma}(r - p) + \delta_{r,0}\mathbf{V} \tag{2}$$

where $\mathbf{V} = \langle \Delta\mathbf{x}(t)\, \Delta\mathbf{x}^{\mathrm{T}}(t)\rangle$ is the white noise covariance matrix and $\delta_{r,0}$ is the Kronecker delta function.

2.2. Granger Causality. The elements of the residual white noise covariance matrix, $\mathbf{V}$, in a parametrized MVAR model (eq 1) can be considered as quality indicators of model predictions for their respective signal components, based on data contained in preceding steps, in all signal channels. The Granger measure for a causal relation between two channels i and j is based on the assessment of how much the prediction error for the component X_i increases once the component X_j is excluded from model parametrization. Formally, the Granger causality from channel j to i, denoted as $J_{j\to i}$, is expressed as

$$J_{j\to i} = 1 - \frac{V_i}{V_i^{(j)}} \tag{3}$$

where V_i and $V_i^{(j)}$ are ith diagonal elements of covariance matrices $\mathbf{V}$ and $\mathbf{V}^{(j)}$, obtained for the MVAR model parametrized with all the channels and excluding the jth channel, respectively. $J_{j\to i}$ constitute elements of the generally asymmetric Granger causality matrix, $\mathbf{J}$, and can vary between 0 and 1. $J_{j\to i} = 0$ indicates no G-causal relationship between given signal components, meaning that the omission of the jth channel does not affect the model's ability to predict the ith

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944

channel. In turn, $J_{j\to i} = 1$ reveals full coupling, implying that the model loses its ability to predict the ith channel, if the jth channel is excluded from parametrization.

2.3. Simulation Data and Processing Methods. The source trajectory representing 106 μs of the CLN025 β-hairpin fully atomistic molecular dynamics simulation in explicit solvent was obtained on request from D. E. Shaw Research.[9] The trajectory comprised 534 743 frames saved with a time step of 0.2 ns. The peptide sequence Tyr-Tyr-Asp-Pro-Glu-Thr-Gly-Thr-Trp-Tyr was parametrized with the CHARMM22* force field, and the TIP3P model was used for water. Unassisted, spontaneous folding into a stable structure in good agreement with the crystallographic geometry (0.1 nm of Cα root-mean-square deviation, RMSD) was observed. A simulation temperature of 340 K was chosen to obtain multiple, reversible (un)folding events. Technical details of the simulation are given in the original report.[9]

In order to select a representative folded structure, the subset of trajectory frames spaced every 20 ns was clustered according to an all-atom RMSD using the Gromos algorithm as implemented in the Gromacs package, with a cutoff of 0.3 nm, and the reference geometry was determined as the central structure of the largest cluster. Subsequently, in order to determine a continuous reaction coordinate, ξ, for the (un)folding process, the all-atom RMSD with respect to this structure was used together with sines and cosines of peptide backbone ϕ and ψ angles as components in time independent component analysis (TICA).[57,58] The TICA analysis in the resulting 37 dimensions was conducted with the kinetic mapping weighting scheme and a lag time of 120 ns, and its dominant independent component (IC) was used as a reaction coordinate. The choice of particular parameters was based on the method proposed by Best and Hummer,[59] which provides objective criteria to optimize a reaction coordinate to properly capture reactive trajectories between stable states of interest (see Results and Discussion). To this end, we considered a number of trial reaction coordinates and selected the best among them (see the Supporting Information for details). We note that GC analysis itself does not require the definition of any reaction coordinate; however, it is useful to do so for the interpretation of results.

The potential of mean force (PMF) as a function of ξ was obtained based on the probability, $p(\xi)$, of finding the trajectory at ξ, assuming $\mathrm{PMF}(\xi) = -\ln p(\xi) + F_0$, with F_0 constant and chosen such that the global minimum was 0. The PMF uncertainty is reported as plus or minus one standard deviation, based on calculations for the original trajectory split into five consecutive blocks.

To characterize intermediate steps and their representative structures, illustrating the (un)folding process, the ξ region between free energy minima corresponding to the folded, F, and the unfolded, U, states was split using six equidistant centers, and each trajectory frame was assigned to its closest center. Representative structures for such defined folding steps were determined as medoids of respective sets of frames, using distances in kinetically weighted, 37-dimensional TICA space.

For the calculation of the transition matrix and commitor function (CF), the trajectory was converted into an integer sequence based on discrete state numbers assigned to each frame according to the procedure described above, and it was processed using a median filter of 1 ns width. The CF was calculated as a probability that, upon leaving the state of interest, the trajectory reaches the F state prior to visiting the U state. Uncertainty of CF estimation at each point is reported as a range between minimal and maximal values obtained in independent calculations for the trajectory split into five consecutive blocks. The calculations of TICA, CF, and the transition matrix and their visualization were performed with the PyEMMA Python package.[60]

For the calculation of conditional probability, $p(\mathrm{TP}|\xi)$, of the system being on a transition path (TP) between F and U states at a given ξ,[59] TPs were identified as continuous trajectory fragments connecting those two states in either direction, with no recrossing. The reaction coordinate region between F and U states was discretized into 50 bins, ξ_i, and numbers of frames in each bin, N_i and N_i^{TP}, were recorded for the entire trajectory, as well as for TP fragments only, respectively, and were used to estimate $p(\mathrm{TP}|\xi_i) = N_i^{\mathrm{TP}}/N_i$. An error of the estimate is reported as plus or minus one standard deviation obtained for independent calculations involving N_i, and N_i^{TP} evaluation for five consecutive trajectory blocks.

Inter-residue distances were calculated as the shortest separation between the respective heavy atom sets. Hydrogen bonds were defined on the basis of geometric criteria involving an acceptor−donor distance of $\leq$0.32 nm and an acceptor−hydrogen donor angle in the range [130, 180] deg.[61]

The overlap, $\Omega(\xi) \in [0, 1]$, between the distribution of inter-residue distances at a given ξ value, discretized in 21 bins between F and U free energy minima, and their ensemble at state F (corresponding to the bin centered on the global PMF minimum) was calculated as an overlap area of two normal distributions of respective mean values and standard deviations using an overlap function implemented in the Python statistics library. It was further normalized such that the range of its variation along the transition path spanned the entire range between 0 and 1.

2.4. Granger Causality Analysis. To provide useful data for Granger analysis, the original molecular dynamics trajectory was featurized into a set of all 45 inter-residue distances, as defined above. The trajectory was split into two consecutive, equal parts, and each of them was processed independently to assess the convergence of results. Prior to performing the Granger analysis, the distances were normalized to have 0 mean and variance 1. MVAR parametrization and residual covariance matrix evaluation was carried out by using the Yule−Walker method as implemented in the Time Series 1.4 package of Mathematica 7.0. The optimal order, P, of the MVAR model was determined on the basis of the Schwarz−Bayes criterion,[62] which seeks to minimize the following expression:

$$\mathrm{SBC}(P) = 2\log(\det(\mathbf{V})) + \frac{KP\log(N)}{N} \tag{4}$$

with N being the number of time steps considered. Here, the first term on the right-hand side accounts for possibly accurate model predictions, while the second term penalizes model complexity. In our case, $\mathrm{SBC}(P)$ was stable with growing P, so we adopted $P = 1$. The low order of the resulting optimal MVAR model may be related to the relatively long trajectory time step.

In order to obtain insights into the involvement of particular contacts in G-causal relations, rather than to consider dependencies between all 990 possible contact pairs, we introduced the following measures:

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944

- $G_i^{\leftarrow} = \sum_k J_{k\to i}$, called *predictability to*, indicating the extent to which the behavior of the ith contact can be predicted based on the evolution of all remaining contacts
- $G_i^{\rightarrow} = \sum_k J_{I\to k}$, called *predictability from*, indicating the extent to which information encoded by the evolution of the ith contact is useful for the prediction of all remaining contacts
- $G_i^{\circ} = \frac{1}{2}(G_i^{\rightarrow} - G_i^{\leftarrow})$, called *origin predictability*, indicating whether a contact is a source, or an initiator, of events ($G_i^{\circ} > 0$), or rather a sink, or a terminator, of events ($G_i^{\circ} < 0$)
- $G_i^{+} = \frac{1}{2}(G_i^{\rightarrow} + G_i^{\leftarrow})$, called *predictability indicator*, indicating the extent of general participance in G-causal relations

3. RESULTS AND DISCUSSION

3.1. Folding Pathway. Even though GC analysis itself does not depend on prior determination of representative states or processes within the system of interest, their knowledge is essential for meaningful interpretation of the results. Thus, in order to capture relevant intermediate configurations of CLN025 (un)folding, we first devised a reaction coordinate, ξ, that would possibly well follow transition paths, TPs. To this end, we considered a number of descriptors characterizing peptide structure and validated their suitability to serve as a reaction coordinate using the method proposed by Best and Hummer.[59] Briefly, the approach is based on the estimation of conditional probability that the system is on a TP, given its particular position along the reaction coordinate, $p(\mathrm{TP}|\xi)$. The extent to which the maximum of the resulting probability profile is able to reach the theoretical limit of 0.5 is then used to gauge the quality of the underlying reaction coordinate.

From various considered candidates for ξ (see the Supporting Information for details), we chose a combination of heavy atom RMSD from the representative native state structure with sines and cosines of backbone dihedral angles, transformed with the use of TICA.[58] A two-dimensional free energy map as a function of the two first ICs of the TICA solution with a lag time of 120 ns (Figure 1A) indicates two unique minima corresponding to the folded and the unfolded states. The minimum free energy path between them leads predominantly along the first IC, and it was adopted as a one-dimensional reaction coordinate. The resulting PMF (Figure 1B) indicates folding free energy in the range of $-2.5k_BT$, in agreement with previous studies based on the same trajectory.[9,31] The height of the free energy barrier for unfolding, $\sim 5k_BT$, is, however, higher by $\sim 1k_BT$ than that reported by Lindorff-Larsen in the original study.[9] We attribute this difference to the possibility that the originally considered reaction coordinate based on the fraction of native contacts, Q, may not provide optimal resolution for the case of 10 residues only miniprotein, likely enriching the transition region with configurations that, in fact, do not belong to reactive trajectories. Indeed, whereas the maximum in $p(\mathrm{TP}|\xi)$ obtained for our proposed reaction coordinate is close to 0.5 (Figure 1C), this is not the case for our trial reaction coordinate based on Q (see the Supporting Information for details), in line with similar result reported for the same trajectory.[63]

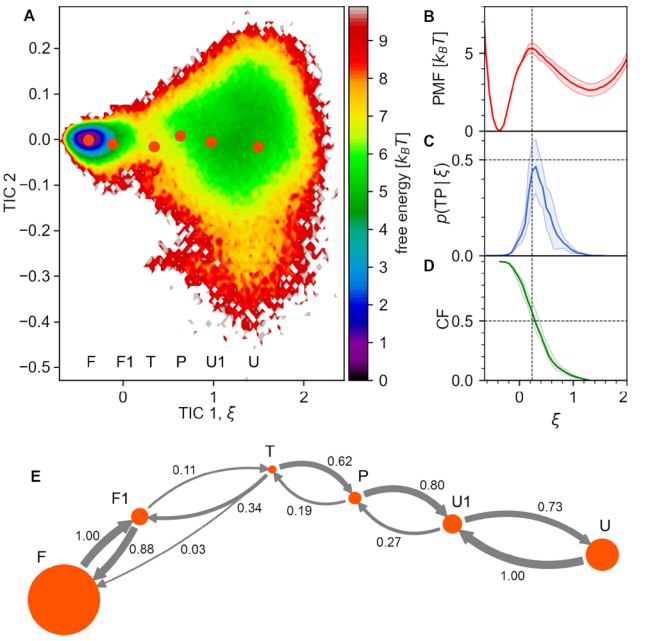


Figure 1. (A) Free energy map as a function of two dominant TICA ICs. Red dots indicate the location of representative frames for six (un)folding steps. (B–D) Descriptors of the folding process as functions of the reaction coordinate. Shaded areas indicate estimation errors (see text for details). (E) Transition matrix between six folding steps, with circle areas proportional to corresponding populations of simulation frames (transitions with probability < 0.03 not shown).

In order to obtain representative structures of CLN025 along its (un)folding paths that would aid in the interpretation of GC results, we chose to select a set of states that (a) covers the reaction coordinate between the free energy minima for folded and unfolded states, (b) captures a geometry typical for the transition state region, and (c) is sufficiently small to enable visual analysis. We note that, as opposed to models aimed at the analysis of system kinetics, it is not necessary that the sets of states and related transition probabilities pass the Chapman–Kolmogorov test.[64] Instead, as we seek to interpret putative causal relations in continuous processes such as (un)folding, it is desirable to obtain a set of steps that occur sequentially one by one during reactive trajectories. Given the above, we arbitrarily partitioned the reaction coordinate into six bins with centers evenly spaced between free energy minima and determined their representative structures as medoids in TICA space (Figure 1A). The resulting steps along the (un)folding pathway characterize folded (F), close to folded (F1), transition (T), preliminary (P), close to unfolded (U1), and unfolded (U) states. Assuming that hairpin folding is a process that leaves state U and reaches state F prior to coming back, we found in total 46 such folding events in the entire trajectory, with a mean duration of 9 ± 1 ns. In turn, the average duration of 46 unfolding processes was 14 ± 3 ns.

The relevance of the adopted reaction coordinate and chosen discrete states is further supported by the fact that the value of the reaction coordinate at which the CF calculated directly from trajectory passes the value of 0.5 (Figure 1D) is consistent with the location of the transition state suggested by the PMF maximum (Figure 1A). In addition, as evidenced by inspection of the accompanying transition matrix (Figure 1E), jumps between nonadjacent states are infrequent, with the highest rate of 0.03 observed for the transition from T directly

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936–1944

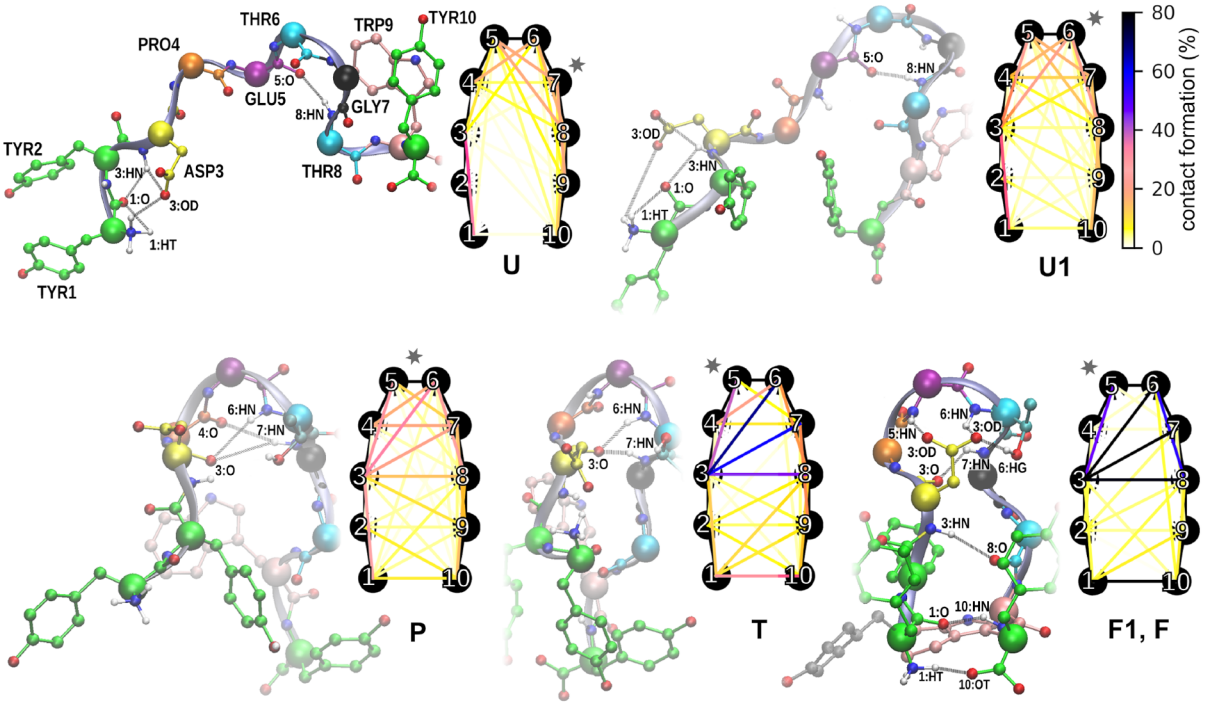


Figure 2. Representative structures for subsequent folding steps and schematic depiction of inter-residue contact frequencies. Shown are major hydrogen bonds captured in particular structures. Gray stars indicate turn locations in polypeptide backbone. Detailed frequencies of contact formation and the distribution of turn angles are provided in the Supporting Information.

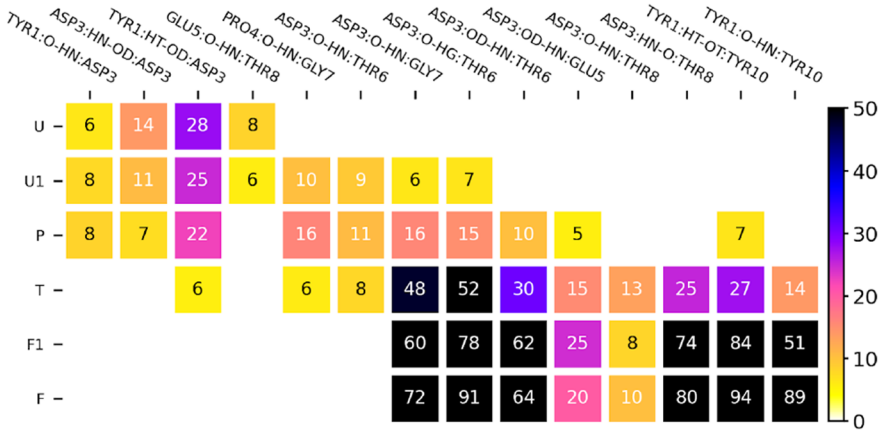


Figure 3. Fractions of hydrogen bonds formed in subsequent CLN025 (un)folding steps. Bonds formed in <5% of respective simulation frames are not shown.

to F, bypassing F1. This implies that the majority of (un)folding events indeed proceeds by visiting all consecutive steps.

3.2. Structural Characterization of Folding Steps. Structural rearrangements between subsequent folding steps are analyzed on the basis of a set of representative geometries together with corresponding fractions of inter-residue contacts (Figure 2) as well as the most stable hydrogen bonds (Figure 3). As can be expected for a short peptide, the ensemble of unfolded structures is rather wide with practically no trace of the native geometry. Notably, though, there are two structural elements that are unique for the unfolded state (Figure 2U). The first one is a network of temporary hydrogen bonds at the N-terminus between Tyr1 and Asp3 that engages the aspartic acid side chain and pulls it away from its nativelike configuration. The second is a hydrogen bond between Glu5 and Thr8, which is responsible for the stabilization of a shallow turn in the region of residue 7, which is displaced with respect to the native hairpin turn located around residue 5.

In the part of the unfolded ensemble that is closer to the transition state (Figure 2U1), the dominant turn region shifts to residue 6, being supported by increasingly frequent main chain hydrogen bonds between Pro4 and Gly7, additionally augmented by new interactions between the Asp3 main chain oxygen atom and Thr6 and Gly7 amide nitrogen atoms.

This repositioning of the turn region toward residue 5 is continued in the preliminary step (Figure 2P), in which the interaction between Glu5 and Thr8 disappears in favor of further enhanced hydrogen bonds between Pro4 and Gly7 as well as the Asp3 main chain with Thr6 and Gly7. Notably, the

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944

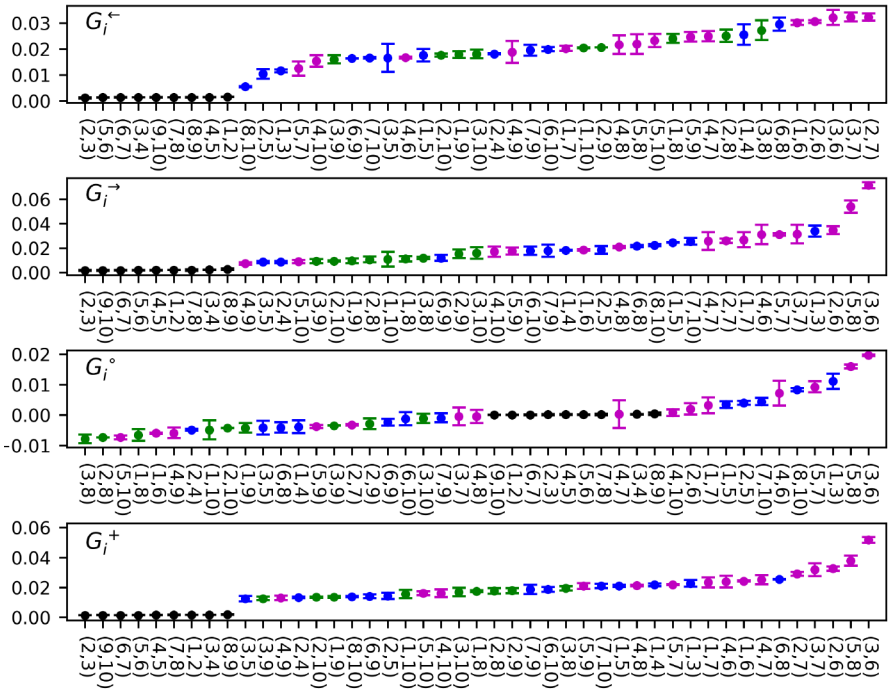


Figure 4. Contact-based descriptors of Granger causality. Color codes for contact groups: magenta, turn; blue, arms; green, ladder; black, direct.

prevalence of the latter increases along with the destabilization of Tyr1−Asp3 interaction, illustrating gradual shifting of Asp3 engagement from the N-terminus to the turn region.

In the transition state phase (Figure 2T), the N-terminal hydrogen bond network, which is observed in the U−P states, becomes entirely absent, liberating the Asp3 side chain, which now repositions and starts maintaining key hydrogen bonds through its carboxyl group with residues 5−7. These interactions take over the role in the stabilization of the turn region from previously mentioned Pro4 and Gly7 which now breaks due to main chain reorientation. In this phase, also both hairpin arms start to interact through main chain hydrogen bonds between Asp3 and Thr8, as well as yet transient hydrogen bonds that connect the N-terminal Tyr1 to the C-terminal Tyr10, and are accompanied by a hydrophobic contact between Tyr2 and Trp9 side chains.

In the folded phases (Figure 2F1,F) the aforementioned interactions solidify and fix hairpin arms in their native configuration. The transition between close to folded, F1, and fully native, F, structures is related to the permanent stabilization of the hydrophobic interaction formed by Tyr2 and Trp9 and rotation of the Tyr1 phenol ring such that it covers hairpin termini, thus sealing the terminal salt bridge through its dehydration.

3.3. Granger Causality for CLN025. The values of descriptors calculated for individual inter-residue contacts based on the complete Granger causality matrix, **J** (Supporting Information), are presented in Figure 4.

In general, the obtained amplitudes are rather low, indicating only subtle G-causal relations within the peptide. Nevertheless, independent results based on the first half and the second half of the trajectory (error bars in Figure 4) appear to be consistent, suggesting that the analysis has indeed converged. Notably, all contacts between directly neighboring residues (Figure 4, black) invariably receive 0 scores, as should be expected based on the independence of their contact distances determined by covalent bonds from peptide conformations. In order to further validate the significance of the results, we repeated all calculations for a trajectory in which frames were randomly shuffled, thus destroying all existing correlations. Having checked that the distribution of such determined G values in their respective categories is Gaussian, we used the Student's t test to estimate the probability of a null hypothesis that the true results belong to the same ensemble (see the Supporting Information for details). Such obtained p values turned out to be significantly lower than 0.01 in the case of all contacts in all categories, except for contacts between neighboring residues.

An upper range of the $G^{\rightarrow}$ and $G^{\leftarrow}$ parameter spectrum is dominated by contacts belonging to the hairpin turn region (Figure 4, magenta). It is consequently reflected by their dominant G^{+} predictability indicator values, manifesting the highest overall involvement in G-causal relations. The above findings imply that rearrangements within this region foreshadow downstream conformational changes (high $G^{\rightarrow}$) and also that they are a culmination of preceding steps (high $G^{\leftarrow}$). Taken together, this suggests that turn rearrangements may play a major role during the transition state, constituting the threshold between folded and unfolded states. Indeed, one of the highest scoring contacts, Asp3−Thr6, which is (un)formed during the transition phase (Figure 2), has been identified to play a major role in the turn nucleation and stabilization in our analysis (section 3.2) as well as other[19,31] analyses of CLN025 folding.

On the contrary, ladderlike contacts between opposite hairpin arms (Figure 4, green) are characterized by overall much lower G^{+} values. In particular, their $G^{\rightarrow}$ indices are low, implying that on the basis of events occurring within this group of contacts little can be predicted concerning general peptide dynamics. This leads to a conclusion that neither unfolding nor folding processes proceed as the sole result of fluctuations in ladderlike interactions until changes within the turn region take

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944

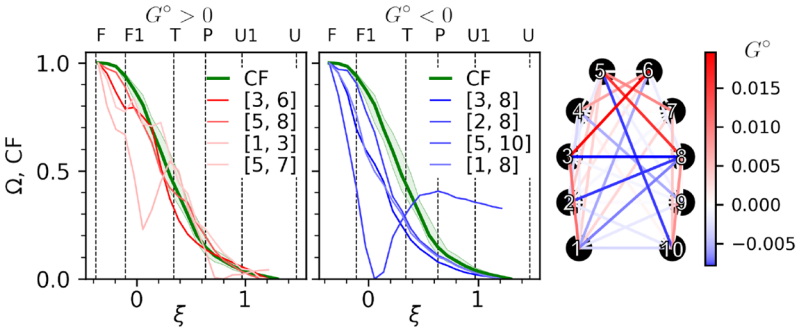


Figure 5. Degree of distance overlap with the native ensemble, $\Omega(\xi)$, for four contacts with highest (left plot) and lowest (right plot) G° values. CF, commitor function. Hairpin scheme: G° values for all contacts.

place. Notably, Tyr1−Tyr10 and Tyr2−Trp9 contacts, considered in the context of chignolin and CLN025 folding as a proxy for the hydrophobic core,[27,28,31−33] do not achieve distinct scores in any category. Thus, in addition to relatively high predictability indices, in particular $G^{\rightarrow}$ for contacts in the turn region described above, our analysis clearly supports the zipperlike rather than the hydrophobic collapse driven mechanism of CLN025 folding, at least based on the simulation under study.

A parameter that is meant to distinguish between contacts that are rather the source of G-causal relations (i.e., their time evolution carries information useful in the prediction of other contact behavior) from those that are predominantly a sink of such relations (i.e., their behavior can be predicted based on the evolution of other contacts) is G° (section 2.4), which adopts positive values in the former case and negative values in the latter case.

It seems reasonable that structural changes followed by most distinct downstream effects should take place while the system is crossing a transition state barrier since their occurrence is expected to trigger downhill evolution along the free energy gradient, typically of higher magnitude than fluctuations within any stable state. Indeed, if the state of four contacts with the highest G° values is followed as a function of the folding reaction coordinate (Figure 5, left plot), it can be observed that most significant changes on their route from the U state to the F state (measured as a change in the degree of overlap, Ω) with the folded state ensemble occur in the T state, closely following the CF. Furthermore, the degree of similarity of Ω(RMSD) to CF(RMSD) apparently correlates positively with the G° value. In contrast, major changes within four contacts at the negative end of the G° spectrum are consistently shifted with respect to CF toward the folded state (Figure 5, right plot).

Three out of five contacts with the highest G° values, including the already mentioned Asp3−Thr6 interaction, are located within the hairpin turn (Figure 5), again highlighting the importance of this region in simulated CLN025 folding under study. Notably, contacts of Glu5 with both Gly7 and Thr8 residues are involved in non-native turn stabilization (section 3.2) and need to break before the turn properly centers at Glu5. Similarly, a highly scored Tyr1−Asp3 interaction is involved in a hydrogen bond network within the N-terminal hairpin arm observed in the U, U1, P, and T states, and its vanishing enables Asp3 repositioning that is necessary for turn nucleation. These observations underscore the fact that Granger analysis is sensitive just to changes in signal components and, being also agnostic to the actual direction of the (un)folding process, does not distinguish between contact making and contact breaking.

The group of contacts with $G^\circ < 0$ is clearly dominated by residue pairs that form parallel interactions between opposite hairpin arms, including the components of the hydrophobic core (Figure 5). This suggests that the formation of this structure completes rather than initiates folding and that its fluctuations in the folded state, e.g., temporal disruption, do not imply the commencement of the unfolding process.

There remains a question to what extent the above characterization of the folding process depends on the choice of the particular reaction coordinate. To this end we analyzed a set of six representative structures obtained for the worst (i.e., having the lowest maximum $p(\mathrm{TP}|\xi)$ value) of the considered trial reaction coordinates, which was based on the RMSD with respect to the native hairpin geometry. We found that it captures essentially the same sequence of structural rearrangements as the original one (see Figures S2 and S8), providing for the same interpretation of causal relations.

4. CONCLUSIONS

Temporal relations in complex biomolecules are inherently difficult to capture and to express in a quantitative manner. Our application of Granger causality analysis to chignolin allowed ranking its structural elements according to their contributions to the (un)folding process. We found that most determinant for chignolin dynamics are residues that form the β hairpin turn. Their high scores in descriptors that express involvement in causal relations are reflected in the independent observation that the major part of their transformation between folded and unfolded states occurs when the system traverses the transition state barrier. In contrast, the dynamics of contacts that are formed between hairpin arms, including those contributing to the peptide's hydrophobic core, turn out to encode comparably little information concerning the time evolution of the system. Taken together, these results support the conclusion that the molecular dynamics trajectory under study depicts chignolin folding in agreement with a zipperlike rather than a hydrophobic collapse mechanism.

The above findings indicate the potential of Granger causality analysis to provide objective measures useful in the interpretation of biomolecular dynamics in the context of already existing hypotheses, such as the debated mechanism of β hairpin folding. Notably, however, the method is not dependent on prior knowledge or assumptions concerning the system of interest that in the case of typical MD analysis are necessary to devise suitable descriptors. This gives a possibility for obtaining insights that otherwise might be left unnoticed,

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944

such as the apparently important role of contacts within the N-terminal hairpin arm, whose breaking proved to be necessary to initiate the repositioning of the hairpin turn and subsequent folding.

A limitation of the considered approach is the need for long MD trajectories that contain multiple realizations of the process under study. Obtaining sufficient sampling for more complex systems than the one considered here will remain challenging. In this respect, an interesting route may involve the adaptation of multiple, short runs in conjunction with Markov state models to provide proper weighting of individual transitions between metastable system states. Another possible improvement may be based on the replacement of the autoregressive model used to determine the Granger causality matrix with a more powerful approach. Aside from already mentioned entropy transfer, machine learning based forecasting methods, which are constantly gaining advantage over classic statistical approaches,[65] may be considered to increase the sensitivity of the causality analysis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00945.

Derivation of Yule−Walker method for MVAR parametrization; reaction coordinate selection and validation; contact formation and turn location during CLN025 folding; Granger causality matrix; statistical validation of contact-based causality descriptors (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Piotr Setny − *Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland;* orcid.org/0000-0002-0769-0943; Email: p.setny@cent.uw.edu.pl

Author

Marcin Sobieraj − *Faculty of Physics, University of Warsaw, 02-093 Warsaw, Poland; Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00945

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge Prof. B. Lesyng for his comments concerning the manuscript. This work was supported by EMBO Installation Grant 3051/2015 to P.S.

■ REFERENCES

(1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. a.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−6.

(2) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116*, 6516−6551.

(3) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129−1143.

(4) Best, R. B. Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.* **2012**, *22*, 52−61.

(5) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(6) Gruebele, M.; Dave, K.; Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *Annu. Rev. Biophys.* **2016**, *45*, 233−251.

(7) Gershenson, A.; Gosavi, S.; Faccioli, P.; Wintrode, P. L. Successes and challenges in simulating the folding of large proteins. *J. Biol. Chem.* **2020**, *295*, 15−33.

(8) Freddolino, P. L.; Schulten, K. Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.* **2009**, *97*, 2338−47.

(9) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science (80-.)* **2011**, *334*, 517−520.

(10) Liu, Y.; Strümpfer, J.; Freddolino, P. L.; Gruebele, M.; Schulten, K. Structural Characterization of λ-Repressor Folding from All-Atom Molecular Dynamics Simulations. *J. Phys. Chem. Lett.* **2012**, *3*, 1117−1123.

(11) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 5915−20.

(12) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(13) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat. Phys.* **2010**, *6*, 751−758.

(14) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.* **2016**, *67*, 669−690.

(15) Jung, H.; Covino, R.; Hummer, G. Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations. *arXiv (Physics.Chemical Physics)*, January 14, 2019, 1901.04595, ver. 1. https://arxiv.org/abs/1901.04595v1.

(16) Hughes, R. M.; Waters, M. L. Model systems for β-hairpins and β-sheets. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514−524.

(17) Xiao, Y.; Chen, C.; He, Y. Folding Mechanism of Beta-Hairpin Trpzip2: Heterogeneity, Transition State and Folding Pathways. *Int. J. Mol. Sci.* **2009**, *10*, 2838−2848.

(18) Best, R. B.; Mittal, J. Microscopic events in -hairpin folding from alternative unfolded ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11087−11092.

(19) Davis, C. M.; Xiao, S.; Raleigh, D. P.; Dyer, R. B. Raising the speed limit for β-hairpin formation. *J. Am. Chem. Soc.* **2012**, *134*, 14476−82.

(20) Jones, K. C.; Peng, C. S.; Tokmakoff, A. Folding of a heterogeneous -hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 2828−2833.

(21) Shao, Q. Folding or Misfolding: The Choice of β-Hairpin. *J. Phys. Chem. B* **2015**, *119*, 3893−3900.

(22) Zerze, G. H.; Uz, B.; Mittal, J. Folding thermodynamics of β -hairpins studied by replica-exchange molecular dynamics simulations. *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 1307−1315.

(23) Razavi, A. M.; Voelz, V. A. Kinetic Network Models of Tryptophan Mutations in β-Hairpins Reveal the Importance of Non-Native Interactions. *J. Chem. Theory Comput.* **2015**, *11*, 2801−2812.

(24) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. Folding dynamics and mechanism of β-hairpin formation. *Nature* **1997**, *390*, 196−199.

(25) Dinner, A. R.; Lazaridis, T.; Karplus, M. Understanding -hairpin formation. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9068−9073.

(26) Pande, V. S.; Rokhsar, D. S. Molecular dynamics simulations of unfolding and refolding of a -hairpin fragment of protein G. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9062−9067.
(27) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 Residue Folded Peptide Designed By Segment Statistics. *Structure* **2004**, *12*, 1507−18.
(28) Honda, S.; Akiba, T.; Kato, Y. S.; Sawada, Y.; Sekijima, M.; Ishimura, M.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K. Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* **2008**, *130*, 15327−31.
(29) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Lett.* **2006**, *580*, 3422−6.
(30) van der Spoel, D.; Seibert, M. M. Protein Folding Kinetics and Thermodynamics from Atomistic Simulations. *Phys. Rev. Lett.* **2006**, *96*, 238102−4.
(31) McKiernan, K. A.; Husic, B. E.; Pande, V. S. Modeling the mechanism of CLN025 beta-hairpin formation. *J. Chem. Phys.* **2017**, *147*, 104107.
(32) Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.* **2005**, *354*, 173−83.
(33) Harada, R.; Kitao, a. Exploring the Folding Free Energy Landscape of a beta-Hairpin Miniprotein, Chignolin, Using Multiscale Free Energy Landscape Calculation Method. *J. Phys. Chem. B* **2011**, *115*, 8806−8812.
(34) Enemark, S.; Rajagopalan, R. Turn-directed folding dynamics of β-hairpin-forming de novo decapeptide Chignolin. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12442−12450.
(35) Suenaga, A.; Narumi, T.; Futatsugi, N.; Yanai, R.; Ohno, Y.; Okimoto, N.; Taiji, M. Folding dynamics of ID-residue β-hairpin peptide chignolin. *Chem. - An Asian J.* **2007**, *2*, 591−598.
(36) Kamenik, A. S.; Handle, P. H.; Hofer, F.; Kahler, U.; Kraml, J.; Liedl, K. R. Polarizable and non-polarizable force fields: Protein folding, unfolding, and misfolding. *J. Chem. Phys.* **2020**, *153*, 185102.
(37) Kührová, P.; De Simone, A.; Otyepka, M.; Best, R. B. Force-field dependence of chignolin folding and misfolding: comparison with experiment and redesign. *Biophys. J.* **2012**, *102*, 1897−906.
(38) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386−2396.
(39) Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424.
(40) Sims, C. A. Macroeconomics and Reality. *Econometrica* **1980**, *48*, 1−48.
(41) Campbell, J. Y.; Lo, A. W.; MacKinlay, A. C. *The Econometrics of Financial Markets*; Princeton University Press: Princeton, NJ, 2012.
(42) Mills, T. C.; Markellos, R. N. *The Econometric Modelling of Financial Time Series*; Cambridge University Press: Cambridge, NY, 2008.
(43) Tsay, R. S. *Analysis of Financial Time Series*; Wiley: Hoboken, NJ, 2010.
(44) Kaminski, M. J.; Blinowska, K. J. A new method of the description of the information flow in the brain structures. *Biol. Cybern.* **1991**, *65*, 203−210.
(45) Blinowska, K. J.; Kuś, R.; Kamiński, M. Granger causality and information flow in multivariate processes. *Phys. Rev. E* **2004**, *70*, 050902.
(46) Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461−464.
(47) Barnett, L.; Barrett, A. B.; Seth, A. K. Granger causality and transfer entropy Are equivalent for gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701.
(48) Kamberaj, H.; Van Der Vaart, A. Extracting the causality of correlated motions from molecular dynamics simulations. *Biophys. J.* **2009**, *97*, 1747−1755.
(49) Vatansever, S.; Gümüş, Z. H.; Erman, B. Intrinsic K-Ras dynamics: A novel molecular dynamics data analysis method shows causality between residue pair motions. *Sci. Rep.* **2016**, *6*, 37012.
(50) Hacisuleyman, A.; Erman, B. Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin. *PLoS Comput. Biol.* **2017**, *13*, e1005319.
(51) Qi, Y.; Im, W. Quantification of drive-response relationships between residues during protein folding. *J. Chem. Theory Comput.* **2013**, *9*, 3799−3805.
(52) Gorecki, A.; Trylska, J.; Lesyng, B. Causal relations in molecular dynamics from the multi-variate autoregressive model. *Europhys. Lett.* **2006**, *75*, 503−509.
(53) Daniluk, P.; Dziubiński, M.; Lesyng, B.; Hallay-Suszek, M.; Rakowski, F.; Walewski, Ł. From experimental, structural probability distributions to the theoretical causality analysis of molecular changes. *Comput. Assist. Methods Eng. Sci.* **2012**, *19*, 257−276.
(54) Hempel, T.; Plattner, N.; Noé, F. Coupling of Conformational Switches in Calcium Sensor Unraveled with Local Markov Models and Transfer Entropy. *J. Chem. Theory Comput.* **2020**, *16*, 2584−2593.
(55) Yule, G. U. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philos. Trans. R. Soc. A* **1927**, *226*, 267−298.
(56) Walker, G. T. On periodicity in series of related terms. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **1931**, *131*, 518−532.
(57) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101.
(58) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(59) Best, R. B.; Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.
(60) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(61) Desiraju, G. R.; Steiner, T. *The Weak Hydrogen Bond In Structural Chemistry and Biology*; Oxford University Press: Oxford, U.K., 2001.
(62) Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461−464.
(63) Zheng, W.; Best, R. B. Reduction of All-Atom Protein Folding Dynamics to One-Dimensional Diffusion. *J. Phys. Chem. B* **2015**, *119*, 15247−15255.
(64) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.
(65) Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecast* **2018**, *34*, 802−808.

https://doi.org/10.1021/acs.jctc.1c00945
J. Chem. Theory Comput. 2022, 18, 1936−1944