On inference of causality for discrete state models in a multiscale context

Susanne Gerber and Illia Horenko[1]

Institute of Computational Science, Università della Svizzera Italiana, 6900 Lugano, Switzerland

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved August 28, 2014 (received for review June 5, 2014)

Discrete state models are a common tool of modeling in many areas. E.g., Markov state models as a particular representative of this model family became one of the major instruments for analysis and understanding of processes in molecular dynamics (MD). Here we extend the scope of discrete state models to the case of systematically missing scales, resulting in a nonstationary and nonhomogeneous formulation of the inference problem. We demonstrate how the recently developed tools of nonstationary data analysis and information theory can be used to identify the simultaneously most optimal (in terms of describing the given data) and most simple (in terms of complexity and causality) discrete state models. We apply the resulting formalism to a problem from molecular dynamics and show how the results can be used to understand the spatial and temporal causality information beyond the usual assumptions. We demonstrate that the most optimal explanation for the appropriately discretized/coarse-grained MD torsion angles data in a polypeptide is given by the causality that is localized both in time and in space, opening new possibilities for deploying percolation theory and stochastic subgridscale modeling approaches in the area of MD.

multiscale systems | probabilistic networks | Granger causality | nonstationarity | regularization

Discrete state modeling is a powerful tool in many areas of science such as in computational biophysics [where it is mostly used in a form of Markov state models (1–4)], materials science [e.g., deployed in percolation theory and Ising models (5)], bioinformatics [e.g., as probabilistic Boolean models for analysis and control of complex biological networks (6)], and geosciences [e.g., used in the form of the generalized linear regression models (7)]. A central issue of discrete state modeling is the identification of an optimal model for the discrete quantity of interest y (e.g., being a Boolean variable or a probability measure) expressed as a function of other available discrete quantities $x_1, x_2, \ldots, x_n$ (being also Boolean variables or probability measures) and of all other potentially relevant quantities u (being discrete and/or continuous variables). Inference of causality then implies identification of all x_i that have a statistically significant impact on y and distinguishing them from all those x_j that are insignificant for y. To give a concrete example, in the context of molecular dynamics variable y may describe a probability for a certain torsion angle (e.g., from the protein backbone) to be in one of the discrete conformational states; $x_1, x_2, \ldots, x_n$ can be the values of probabilities for all torsion angles of this protein in previous times and variable u may represent all of the positions and velocities of individual atoms, simulation settings (e.g., temperature), and force-field and solvent properties, etc. Understanding the causality in this situation will mean, for example, identification of the memory depth (e.g., in the context of Markov state models, where the number of previous time steps is determined and is needed to explain/predict the current state y) and identification of proper order of the neighbor interactions (i.e., in context of Ising and percolation models, determining how many and which spatially neighboring torsion angles characterized by x_i have a statistical impact on the torsion angle corresponding to a variable y). In addition to giving an additional insight into the system, determining the correct causality and the proper order of interactions allows us also to construct the much simpler (in terms of their computational/numerical complexity) computational methods for such systems. The main reason for this is explained by the fact that the computational complexity for most of the discrete state model methods grows polynomially with respect to (w.r.t.) the overall number of causality interactions in the model (8).

To describe and to understand a given set of data y^t (or time series if t denotes, e.g., the time index of the data, i.e., $t = 0, \ldots, N_T$), either the standard settings of discrete state modeling rely on the explicit availability of u^t or, by imposing a priori assumptions on different components of u^t (e.g., stationarity, statistical independence, etc.) and deploying appropriate central limit theorems, the whole u^t is modeled as some stationary (in time) and homogeneous (in space) stochastic process. However, in a context of multiscale and multiphysics models, the presence of unresolved scale quantities u^t (that are not statistically independent or identically distributed) may result in the nonstationarity and nonhomogeneity of the resulting data-driven discrete state models and may manifest itself in the presence of secular trends and/or in regime-transition behavior (9). Application of the standard stationary discrete state modeling approaches common to machine learning and statistics (e.g., methods like artificial neuronal networks, support vector machines, and generalized linear models) may lead to biased results (9) and wrong inference of underlying causality (i.e., in the attribution of regressors x_i^t in terms of their importance or unimportance for explaining the model variable y). Moreover, the standard continuous instruments of causality identification based on correlation [e.g., cross-correlation and cross-covariance (10)] or linear predictability [such as the concept of Granger causality (11–13)]

Significance

The presented framework is capable of parameter identification and optimal causality inference for discrete/Boolean-valued processes in a multiscale context, allowing us to understand such processes beyond the usual statistical assumptions of standard approaches. By applying this framework to appropriately coarse-grained molecular dynamics data it is demonstrated that the optimal causality of the considered process is localized in time and space. This offers new opportunities for applying localized parameterization tools from other areas in a molecular dynamics context, thus opening up new possibilities to bridge the gap toward long timescales in molecular dynamics simulations. The new methodology is expected to become very useful in various scientific fields (bioinformatics, geoscience) where large amounts of multiscale discrete/Boolean data have been accumulated.

Author contributions: I.H. designed research; S.G. and I.H. performed research; I.H. contributed new reagents/analytic tools; S.G. and I.H. analyzed data; and S.G. and I.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: horenkoi@usi.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1410404111/-/DCSupplemental.

(an explanation of Granger causality and related standard concepts is in *SI Text*) are not applicable to discrete/Boolean objects like y^t and $x_1^t, x_2^t, \ldots, x_n^t$. Moreover, even in the completely stationary and homogeneous setting, the resulting problems can be ill-posed (e.g., meaning that a small perturbation of y or x_i might result in a large change of the inferred parameters and a completely different causality understanding) (7, 10).

Our aim is to present a computationally tractable approach for causality inference with missing scales that allows us to address the above issues of nonstationarity, nonhomogeneity, and ill-posedness.

The Model of Causality with Unresolved Scales

We start with defining the probability vectors $\Lambda(t) = (\mathbb{P}[y^t | x_1^t \text{ and } u^t], \ldots, \mathbb{P}[y^t | x_n^t \text{ and } u^t], \mathbb{P}[y^t \text{ and not one of } x_i^t])$, $P_x^t = (\mathbb{P}[x_1^t \text{ and } u^t], \ldots, \mathbb{P}[x_n^t \text{ and } u^t], 1)$, and a variable $P_y^t = \mathbb{P}[y^t]$. Then the probability of observing y^t together with $x_1^t, \ldots, x_n^t$ and u^t can be written as a scalar product of these two vectors,

$$P_y^t = \Lambda^{\dagger}(t) P_x^t, \qquad [1]$$

where † denotes the transposition. In a case of pairwise disjoint events $x_1, \ldots, x_n$ the discrete state model [**1**] represents the law of a total probability and is thereby exact. In a case when the components of x are not pairwise disjoint, [**1**] represents a linearized probabilistic model. We call this (a priori unknown) conditional probabilities vector $\Lambda(t)$ a causality vector, to distinguish it from the commonly used concepts of correlation and Granger causality that are introduced in terms of the Euclidean metric.

If only the series of binary/Boolean/probabilistic observations P_x^t and a realization sequence of y^t are available for different values of t, the problem of causality inference will be in identifying both the causality vector $\Lambda(t)$ and the sequence of probabilities P_y^t in [**1**]. To give a simple example, if y^t is a sequence of coin flips at different times and x^t is a single ($n = 1$) binary stochastic process describing the outcomes of coin flips at times $t - 1$, then in the case of a fair coin the optimal causality inference based on [**1**] should provide a stationarity causality model with $P_y^t \equiv 0.5$ and $\Lambda(t) \equiv (0, 0.5)$.

Assuming y^t being statistically independent in t (conditioned on the knowledge of variables x and u) sequence of binary variables or observed probabilities, inference of both the unknown causality vector $\Lambda(t)$ and the unknown probability process P_y^t for the discrete state model [**1**] can be done via a maximization w.r.t. $\Lambda(t)$ of the following log-likelihood functional,

$$\mathcal{L} = \sum_{t=0}^{N_T} \left[(1 - y^t) \ln\left(1 - \Lambda^{\dagger}(t) P_x^t\right) + y^t \ln\left(\Lambda^{\dagger}(t) P_x^t\right) \right], \qquad [2]$$

subject to the constraint

$$0 < \Lambda^{\dagger}(t) P_x^t < 1, \quad \text{for all } P_x^t. \qquad [3]$$

The resulting inference problem [**2**, **3**] is nonstationary (because parameter vector Λ is time dependent) and nonhomogeneous (because different elements of this vector can have different values). It is a direct consequence of the impact from time-dependent variables u^t, e.g., coming from the unresolved scales in a multiscale problem. This problem has a trivial yet useless solution $\Lambda(t) = (0, \ldots, 0, \alpha^t)$ (where $\alpha^t \equiv y_t$), meaning that all of the covariates $x_1 \ldots, x_n$ are unimportant for explaining y and that y is completely explained by the unresolved effects given in the form of the noise process α^t. This demonstrates that [**2**, **3**] is an ill-posed problem. A standard way of dealing with ill-posed problems is based on imposing additional information or additional assumptions. E.g., in the context of stationary/homogeneous modeling, it is assumed that Λ is a time-independent vector (6, 14, 15). A resulting problem becomes well-posed, robust, and uniquely solvable and the impact of unresolved scales is then essentially modeled via a stationary and homogeneous Bernoulli process with a time-independent probability α. If the impact of u is significantly time-dependent, this assumption might be overstringent and can lead to biased results (9).

Numerical Inference of the Optimal Causality

The central methodological contribution of this paper is in finding that one can transform the above ill-posed problem [**2**, **3**] into the well-posed clustering problem formulation, resulting in a computationally tractable yet completely nonstationary and nonhomogeneous discrete model. This can be achieved by deploying the following very mild assumption,

$$\Lambda(t) = \sum_{i=1}^{\mathbf{K}} \gamma_i^t \Lambda_i, \qquad [4]$$

with $\sum_{i=1}^{\mathbf{K}} \gamma_i^t = 1$, $\gamma_i^t \geq 0$ for all times t.

As explained above, changes in the unresolved scales can induce a temporal change of the intrinsic causality relations between the resolved scales variables y and $x_1, x_2, \ldots, x_n$. From the viewpoint of physics this additional assumption [**4**] means that these changes are explained through $\mathbf{K}$ different (unknown) configuration sets of the unresolved scales that give rise to the $\mathbf{K}$ different (unknown) causality vectors Λ_i for the observed and analyzed scales. In a sense, [**4**] defines a decomposition of the whole configuration space of the system into $\mathbf{K}$ a priori unknown domains, each of which is defined by its unique causality relations for the resolved variables. Based only on the available observed time series for y^t and $x_1^t, x_2^t, \ldots, x_n^t$ ($t = 0, \ldots, N_T$), the challenge now is to identify the optimal number $\mathbf{K}$ of these domains as well as the optimal values of causality vectors Λ_i and optimal indicator functions γ_i^t for each of the domains. These indicator functions can then tell at which moment of time which observed causality domain in configuration space is visited by the system.

From the machine-learning and probability theory perspectives, assumption [**4**] means that the true time-dependent parameters $\Lambda(t)$ can be represented as a probabilistic mixture of $\mathbf{K}$ time-independent (or stationary) parameter vectors $\Lambda_i = (\lambda_i^{(1)}, \ldots, \lambda_i^{(n)}, \alpha_i)$ with time-dependent mixing probabilities $(\gamma_1^t, \ldots, \gamma_{\mathbf{K}}^t)$ ($\mathbf{K}$ is a priori unknown). I.e., with assumption [**4**] the nonstationarity of discrete state model [**1**] can be described and understood as a regime-transition process γ, switching between $\mathbf{K}$ different causality regimes, each one of them being described via a fixed causality vector Λ_i. Setting $\mathbf{K} \equiv 1$ is equivalent to the stationarity assumption of standard discrete state modeling mentioned above. It is also very important to mention that no further mathematical or probabilistic assumptions are required; e.g., it is from now on needless to assume that the switching process γ is time homogeneous or stationary or that it belongs to Markovian, Bernoulli, or Poisson families, etc. In this sense the description of γ and its numerical treatment will remain nonparametric throughout the methodology that is proposed.

It can be demonstrated (detailed proof in *SI Text*) that based just on this assumption [**4**] and using the fact that the available data have a finite size (i.e., that $N_T < +\infty$), the original ill-posed problem [**2**, **3**] can be transformed to

$$l^{\epsilon} = \sum_{i=1}^{\mathbf{K}} \left[\sum_{t=0}^{N_T} \gamma_i^t g(t, \Lambda_i) - \epsilon^2 \sum_{j=1}^{n} \lambda_i^{(j)} \right] \to \max_{\gamma_i^t,\ \Lambda_i} \qquad [5]$$

subject to constraints

$$\begin{aligned} &\sum_{i=1}^{\mathbf{K}} \gamma_i^t = 1, \gamma_i^t \geq 0 \quad \text{for all } t \text{ and } i, \\ &0 < \sum_{j=1}^{n} \lambda_i^{(j)} < 1, \quad 0 \leq \lambda_i^{(j)} \leq 1, \quad \text{for all } i \text{ and } j, \\ &\sum_{t_1, t_2 = 0}^{N_T} \left| \gamma_i^{t_1} - \gamma_i^{t_2} \right| \leq \overline{\mathbf{C}}(N_T), \quad \text{for all } i, \end{aligned} \qquad [6]$$

with $g(t,\Lambda_i)=(1-y^t)\ln(1-\Lambda_i^\dagger P_x^t)+y^t\ln(\Lambda_i^\dagger P_x^t)$. A key feature of this formulation [**5**, **6**] is that it represents a lower bound for the original problem [**2**, **3**], meaning that maximizing this transformed lower bound problem w.r.t. γ_i^t, Λ_i (for fixed values of constants $\mathbf{K}$, $\overline{\mathbf{C}}$, and ε^2) simultaneously maximizes the original problem w.r.t. $\Lambda(t)$. Detailed discussion of this and the other mathematical and numerical properties of the transformed problem [**5**, **6**] can be found in *SI Text*. The problem of identifying the most optimal values for the constants $\mathbf{K}$, $\overline{\mathbf{C}}$, and ε^2 is tackled below.

From the machine-learning and probability theory perspectives, the transformed problem [**5**, **6**] is essentially a regularized clustering problem (for $\mathbf{K}$ clusters with γ_i^t being cluster affiliations) with a distance function defined as $g(t, \Lambda_i)$. For ε^2 it is a special case of the nonstationary and nonparametric model family called finite element models with bounded variation of parameters (FEM-BV) that has been introduced recently (16, 17) (more details in *SI Text*). As explained in *SI Text*, an iterative numerical scheme can be deployed to solve [**5**, **6**] w.r.t. γ_i^t, Λ_i for fixed values of $\mathbf{K}$, ε^2, and $\overline{\mathbf{C}}(N_T)$. Another useful feature of [**5**, **6**] that naturally comes from the above derivation is related to the "sparsifying"/regularizing effect of the second term in Eq. **5**. I.e., increasing ε^2 will result in "zeroing out" or "shrinkage" of the statistically unimportant values $\lambda_i^{(j)}$, thereby allowing us to identify only the statistically significant causality relations between x and y. This specific feature of [**5**] is known under the name LASSO regularization and is very widely used in signal/image processing (18) and compressed sensing (19).

There remains a question of finding the most optimal values of $\mathbf{K}$, $\overline{\mathbf{C}}$, and ε^2 for [**5**, **6**], because different combinations of these values will result in different optimal causality models (characterized by different causality vectors Λ_i and affiliations γ_i^t). E.g., increasing the parameter $\mathbf{K}$ will result in the overall increase in the total number of parameters Λ_i and may lead to overfitting as $\mathbf{K}$ approaches N_T. To avoid this problem we suggest using the concepts from information theory, e.g., the information criteria that allow us to access a statistical significance of different candidate models by simultaneously minimizing the number of free model parameters and maximizing the log-likelihood. A short introduction into information-theoretical concepts as well as the exact formula of the Bayesian information criterion (BIC) for the casualty inference problem [**5**, **6**] can be found in *SI Text*. As we illustrate below with our example (see Fig. 2), information-theoretical approaches like the BIC (20) can be straightforwardly used to rank different causality models obtained from [**5**, **6**] in terms of their statistical significance for the available data and to compute the posterior probabilities for different candidate models. Models with the lowest BIC values (meaning the highest posterior model probabilities and highest statistical significance, details in refs. 17 and 20) are the most optimal models in the sense of information theory. Such models are simultaneously most simple in terms of causality relations and the number of involved model parameters as well as most qualitative/probable in terms of the respective log-likelihood and statistical significance for the considered data. Deployment of the FEM-BV framework for numerical solution of [**5**, **6**] together with the BIC allows a completely automated parameter identification and causality inference for discrete state models [**1**] that are essentially free of any tunable user-defined parameters.

Application to Causality Inference from Biomolecular Molecular Dynamics Data

We now proceed with an illustrative application of our approach to a problem of causality inference in molecular dynamics (MD) data analysis. The considered dataset represents an output of the 0.5-μs simulation (with 2-fs time step) of a 10-alanine (10-ALA) polypeptide in explicit water at room temperature, performed with the Amber99sb-ildn force field (21). This MD dataset was produced and provided by Antonia Mey from Freie Universität Berlin, Germany. For further analysis only the values of torsion angles ϕ_i and ψ_i ($i = 1, \ldots, 8$) inside of the molecular backbone (i.e., ignoring the two end groups and the ω_i angles) have been considered with a time-step resolution of 100 ps, resulting in 16 torsion angles time series with 5,000 time points each. Fig. 1 shows the Granger causality matrix that is directly inferred from the shifted torsion angles data (details of Granger causality matrix computations in *SI Text*). Because Granger causality is based on the Euclidean distance, shifting is done to avoid the discontinuity of the ±180° crossings. The Granger causality is measured in terms of predictability gain that is inferred with linear and stationary stochastic models that are fitted to the analyzed data. It is one of the most popular methods for data-driven causality inference and is very widely deployed in areas ranging from economics (11) to ecology (12) and climate

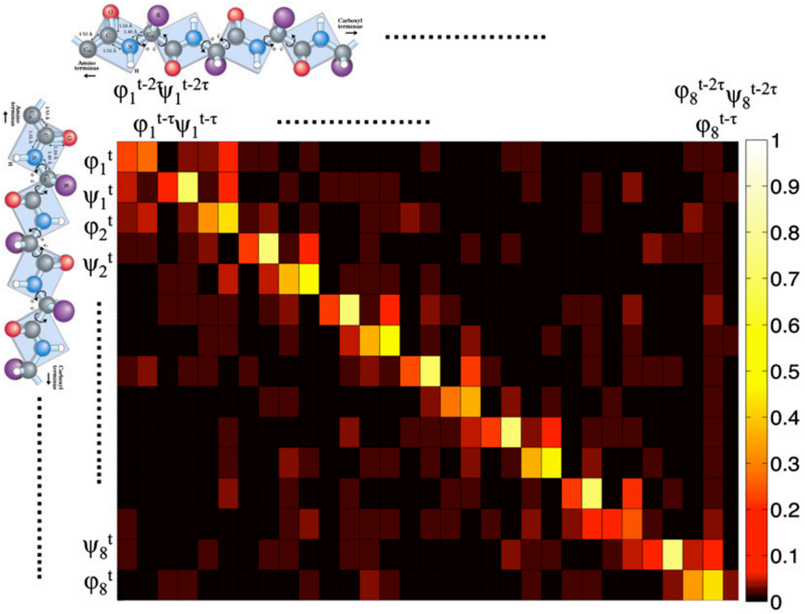

Fig. 1. Granger causality matrix (11) computed for up to 2τ lagged delays (τ = 100 ps) of the shifted Ramachandran torsion angles data $\phi_i(t)$, $\psi_i(t)$.

research (13). Inspection of Fig. 1 reveals a lot of nonlocal Granger causality relations between different time instances and different torsion angles that are far away from each other in the peptide chain. However, this result might be also biased by the nonlinearity or nonstationarity of the data as well as by a not completely appropriate shifting of the angles data and remaining discontinuities at ±180° crossing. To avoid the preliminary shifting of the data (that was required for the computation of Granger causality in Fig. 1), each of the 16 torsion angles was independently discretized/clustered (Fig. S1), deploying the FEM-BV–K-means clustering algorithm (17) with a torus metric (because Euclidean distance, as deployed, e.g., in a standard K-means clustering is not applicable to the angular data). A detailed discussion of the deployed clustering method and its relation to alternative approaches is in *SI Text*. The resulting Boolean time series $\underline{\phi}_i(t), \underline{\psi}_i(t)$ are thereby free from all fluctuations on the shorter timescales and contain only the essential local metastable conformational flipping processes for each of the torsion angles along the peptide backbone (Fig. S2).

Next, we deploy the standard Markov state modeling paradigm that is widely used in analysis and modeling of bimolecular systems. This paradigm can be essentially performed in three steps (1–4): step *i*, decomposition of the full configuration space (that is 16 dimensional for a considered torsion angles data example) into n (disjoint) boxes; step *ii*, computation of the Markovian transition probabilities between these boxes; and step *iii*, spectral decomposition of the resulting transfer operator and identification of conformations as metastable states of this operator. This approach can be seen as a special case of the discrete state model [**1**] for x_i^t, taking the value 1 if the system visits a configuration box i at time t (and 0 otherwise), with n being the total number of configuration boxes and for $\Lambda(\cdot)$ a priori assumed to be time-independent. Taking the binary discretization $\underline{\phi}_i(t), \underline{\psi}_i(t)$ obtained above for each of the 16 torsion angles, we can define the resulting discretization of the full 16-dimensional configuration space through a subset of 2^{16} possible combinations of these variables. Maximizing [**5**, **6**] then results in a Markov state model with the BIC value of $6.48 \cdot 10^4$. Next, we compare this model (where different x_i represent up to 2^{16} global configurations that are pairwise disjoint) with the model where x_i are built from up to $16m$ local values of the variables $\underline{\phi}_i(t), \underline{\psi}_i(t)$, where m indicates the memory (and $m = 1$ for the Markov model).

We go through all of the 16 Boolean time series $\underline{\phi}_i(t), \underline{\psi}_i(t)$ and for each one of them we set y^t respectively as either $\underline{\phi}_i(t)$ or $\underline{\psi}_i(t)$ and set $(x_1^t, \ldots, x_n^t)$ as $(\underline{\phi}_1(t-m\tau), \ldots, \underline{\phi}_1(t-\tau), \underline{\psi}_1(t-m\tau), \ldots, \underline{\psi}_8(t-2\tau), \underline{\psi}_8(t-\tau))$. In each of these 16 cases we solved the problem [**5**, **6**] for all possible combinations of the values $\mathbf{K} = [1, 2]$, $\varepsilon^2 = [0, 1, \ldots, 100]$ and memory $m = [1, 2, \ldots, 20]$, resulting in a total of 64,000 complete maximizations of [**5**, **6**] for all of the binary variables jointly and 4,000 different optimal causality models [**1**]. Optimizations have appeared to be rather insensitive to variation of the persistency bound $\overline{\mathbf{C}}(N_T)$ so to reduce the number of total causality models that need to be computed we set $\overline{\mathbf{C}}(N_T) = N_T$, thereby essentially switching the respective persistency constraint off.

The BIC-optimal model (denoted with a black cross in Fig. 2) has a BIC value of $1.89 \cdot 10^4$ and a posterior Bayesian probability of 0.9999 (20), making it the most appropriate causality model in terms of information theory and "Occam's razor" principle of these 4,000 models and also making it more statistically significant then the standard Markov state model considered above. This optimum corresponds to the combination of $K = 1$ (i.e., it is a stationary model), $m = 2$ (i.e., it is a non-Markovian model with memory depth of two time steps and 200 ps), and $\varepsilon^2 = 10$; i.e., it is a regularized model. Inspection of the resulting optimal parameters Λ put together into a matrix (Fig. 3) further reveals the effect of this optimally sparsifying regularization: In contrast to the Granger causality from Fig. 1 and the cross-correlation matrices from Figs. S3 and S4 in *SI Text*, causality relations in the optimal discrete state model [**1**] are localized both in time and in the neighborhood impacts. Inspection of the confidence intervals for the optimal Λ as well as the repetition of the whole estimation procedure with 75% and 90% of the data showed that the obtained Λ is robust and does not change noticeably when further increasing the size of the underlying statistics. In contrast, application of standard tools (Fig. 1 and Figs. S3 and S4 in *SI Text*) does not reveal any causality patterns and indicates global interactions/correlations. This comparison highlights the conceptual difference of the introduced methodology from standard approaches (e.g., based on Granger causality and Euclidean cross-correlations) and stresses the importance of verifying the validity of underlying mathematical assumptions in analysis of physical data. Moreover, in contrast to the standard Euclidean tools, causality matrix Λ from Fig. 3 also explicitly incorporates the impact of the unresolved scales α^t (visualized as the first column of this matrix).

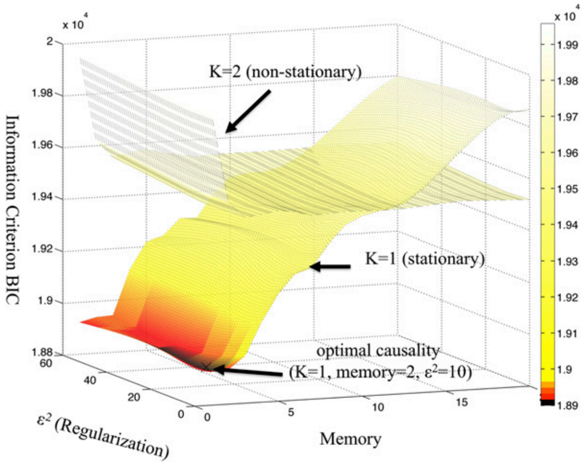

Fig. 2. Values of BIC for optimal solutions of [**5**, **6**] obtained for different combinations of **K**, ε^2, and m. Shown is a joint information content for all of the causality relations between all of the 16 torsion angles together, measured in terms of 4,000 different joint causality models (every model is a single point on this 3D plot).

This particular result seems to be counterintuitive at first glance, because one would expect a lot of nonlocal effects in MD data (e.g., induced by the solvent and long-range interactions like hydrogen bonds). However, application of the methodological framework presented in this paper and information-theoretical comparison of 4,000 candidate models really demonstrate that most of the nonlocal effects are getting "filtered out" in the first step of data processing, i.e., when the torsion angles are individually clustered, revealing the metastable conformational flipping process on a longer timescale. The binary discrete state model characterized by the matrix in Fig. 3 that was identified to be most optimal in terms of local causality relations can now be used, e.g., (*i*) for comparing the statistical significance and/or uncertainty of predictions obtained by this model with the other causality models and (*ii*) for reproducing the 3D structure of the peptide from the vector consistent with binary realizations $(\underline{\phi}_1(t), \underline{\psi}_1(t), \underline{\phi}_2(t), \underline{\psi}_2(t), \ldots)$ from [**1**]. As for the significance, deployment of the Bayesian information criterion in identification of the most optimal causality model (as shown in Fig. 2) directly involves the comparison of statistical significance of different discrete causality models for given data. More specifically, the ranking of models with respect to their posterior model probabilities (as obtained in our analysis) is directly based on the statistical significance measurement and comparison of different candidate models, computed in terms of such information-theoretical measures as relative entropy and Fisher's information (more details in *SI Text* and ref. 20). As for the problem of reproducing the 3D structures, direct inspection of the binary vectors $(\underline{\phi}_1(t), \underline{\psi}_1(t), \underline{\phi}_2(t), \underline{\psi}_2(t), \ldots)$ can already reveal some

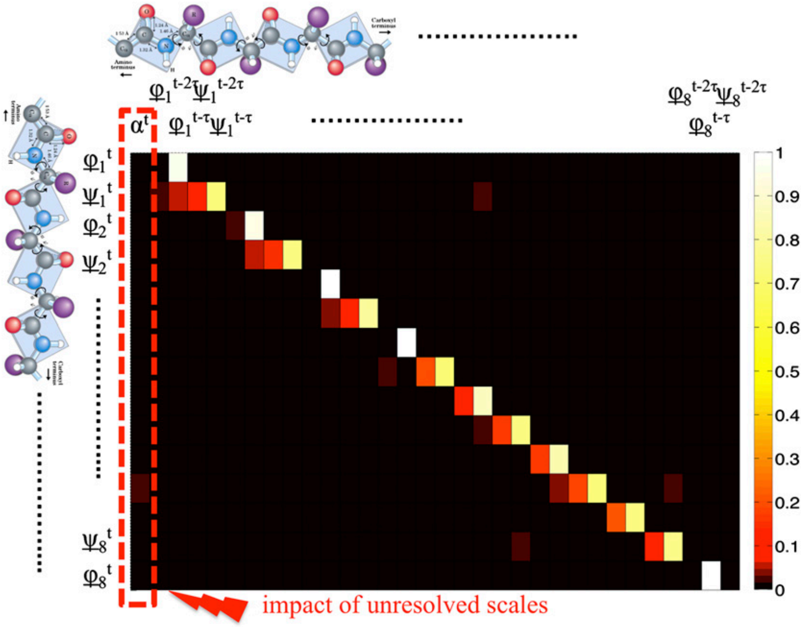

Fig. 3. Discrete causality matrix Λ corresponding to the BIC-optimal model from Fig. 2. For comparison, the color scale is chosen to be the same as in Fig. 1.

information about the 3D structures. E.g., for the 10-ALA example considered in this paper the binary combination of alternating 0 and 1, i.e., (0, 1, 0, 1, 0, 1, 0, 1, . . .), corresponds to a "pure" α-helical structure. The sequence of all zeroes, i.e., (0, 0, 0, 0, 0, 0, 0, 0, . . .), describes a coil-like structure (the 0 state corresponds to the β/coil region of the Ramachandran map). Discussion of further possible approaches to the 3D reconstruction problem (*ii*) can be found in *SI Text*.

The optimal discrete causality model can be also used for predictions of proportions for each of these molecular conformations in a conformational ensemble that could otherwise be obtained only from the very long MD simulations or from the experiments. E.g., performing a 0.1-ms run with the optimal local percolation model [**1**], we get that the relative weight of the statistical coil structure (0, 0, 0, 0, 0, 0, 0, 0, . . .) is predicted to be 0.03 ± 0.002, whereas the weight of a purely α-helical conformation is predicted to be around $0.15 \cdot 10^{-3} \pm 0.11 \cdot 10^{-3}$. The predicted empirical distribution of 3D conformational weights obtained from the optimal percolation model is shown in Fig. 4. Sizes of the obtained confidence intervals for predicted weights mainly result from the uncertainty of the optimal local causality matrix from Fig. 3 (measured in terms of the Fisher information and indirectly quantified by the deployed Bayesian information criterion) and decrease with the increasing sizes of the available statistics. Respective uncertainties for conformational probabilities estimated directly from the available MD data or from the global causality models are significantly bigger. However, the overall quality of these prediction results is very much dependent on the quality/accuracy of the force field that was deployed in the MD simulation. Current investigations presented in this paper do not take this source of possible errors into consideration.

Concluding Discussion

Causality belongs to the most fundamental concepts in science. As discussed above, besides gaining a better insight into a system, appropriate discarding of the insignificant causality relations reduces the computational complexity of the corresponding models and methods. In computational molecular biophysics and MD (an area chosen to illustrate the introduced approach), one of the main current challenges is in bridging the computational gap toward long simulations for realistic bimolecular systems, e.g., toward their reliable computation on timescales of milliseconds or even seconds. In this context, deploying the modern hardware supercomputing facilities and elaborate parallel MD software, spectacular results for millisecond and beyond folding of large biological molecules have been reported in the literature (22). Doing MD, in every step and for every atom one in principle needs to account for all of the forces coming from interactions with all of the other atoms in the system. Even after deploying long-range truncation methods the causality in such models remains global, implying huge computational cost and a lot of all-to-all communications during the computation. The results from Fig. 3 indicate that the following two-step procedure (that is currently also widely adopted in the context of standard Markov state modeling of biomolecular dynamics (1–4)) could help to overcome this difficulty and to bridge the gap toward a long-time MD simulation: step *i*, parameterization of the discrete model [**1**] for the appropriately coarse-grained MD time series on a computationally feasible short timescale, inferring the optimal causality vectors Λ together with their confidence intervals; and step *ii*, inferred optimal values and confidence intervals for Λ can be used to run the coarse-grained

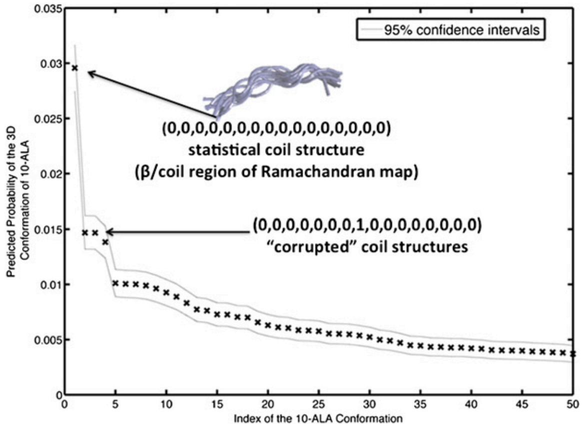

Fig. 4. Conformational weights predicted from a 0.1-ms run of the optimal percolation model [**1**] with the local causality matrix from Fig. 3.

model [1] for the desired long simulation times, computing the statistical quantities of interest (e.g., relative populations of different conformational states, transition rates between them, etc.) and their uncertainties from the output of this model run. An example of such predictions is given in the Fig. 4.

A reduced description of a peptide through a vector of N discrete/Boolean variables (where N is the number of considered torsion angles) can be used to recover the full 3D geometric configuration/conformation of the molecule and is much cheaper to compute using the local discrete state model [1] than the full global MD simulation. Discussion of possible strategies for reproducing the 3D structure of the peptide from these binary vectors can be found in *SI Text*.

One of the main advantages of such a localized causality model compared with the Markov state models based on the global discretization of the molecular configuration space is a much lower number of free model parameters Λ that results from the much smaller number n of possible configurations. If N is the total number of considered torsion angles and m is the memory, then the number of free parameters n for a single y in a standard Markov state model is $n \le 2^{\mathrm{N}} - 1$, whereas in the localized causality model $n \le \mathbf{K}\mathrm{N}m$. If N is large, then $\mathbf{K}\mathrm{N}m \ll 2^{\mathrm{N}} - 1$, meaning that localized models can be reliably estimated without a danger of overfitting from a much shorter time series (as was also demonstrated in the MD example above by the comparison of BIC values for both models). Moreover, discarding the irrelevant causality relations can also significantly increase the efficiency of the model computations. One 0.1-ms run of the "local" causality model from Fig. 3 (involving 48 causality relations between the torsion angles) could be performed 11 times faster than the run of the model with a full global causality (involving 528 causality relations). The lower bound of this efficiency ratio η (i.e., the ratio of run time complexities for the local and global causality models) can be straightforwardly estimated analytically as $\eta\,(\mathrm{N}, m) = \mathcal{O}\,(\mathrm{N}m)$. This means that for realistic proteins with N being of the order of thousands, the efficacy gain due to causality locality might be much more significant than in a considered relatively simple peptide example. Computational complexity of the regularized clustering problem [5, 6] allows us to deal with up to several thousand binary/Boolean variables y and x at the same time, even with the current sequential algorithmic implementation on a PC that is not deploying any form of parallelization.

Locality of the identified optimal model is also very important because it may open new ways of modeling the coarse-grained MD deploying the localized techniques and tools from other areas (such as percolation models (5, 23) and stochastic subgridscale parameterization approaches from fluid mechanics and geophysics (24)).

The presented MD application example with a relatively simple and short peptide molecule aims to demonstrate the importance of assessing the validity of underlying mathematical and probabilistic assumptions in analysis of real data. As demonstrated in the example above, applicability and performance of different standard tools of correlation and causality inference crucially depends on the validity of these assumptions in every particular situation. If these mathematical and statistical assumptions (such as the Euclideanity of the distance measure or the linearity and stationarity of the underlying model) are not fulfilled, assessment of causality might be biased or completely corrupted, even for such simple physical systems as the 10-ALA polypetide considered in this paper. In addition to being a fully nonstationary discrete state model, in the case of the disjoint regressor variables x, model [1] is equivalent to the nonstationary formulation of the law of the total probability and is thereby exact. It means that it can also be directly deployed to understand a discretized behavior of very nonlinear continuous dynamical systems, without imposing further mathematical or probabilistic assumptions. We expect this method of discrete causality inference to be useful also in other areas of science where large amounts of binary, probabilistic, or Boolean data are needed to be processed and understood, e.g., in biophysics/bioinformatics, geophysics, etc. Computer code with the numerical implementation of the introduced methods can be provided by the authors upon request and will be also made available online.

ACKNOWLEDGMENTS. The authors thank Dr. Antonia Mey from Freie Universität Berlin for the provided MD data. This work was funded by the Swiss National Research Foundation (Grant 200021_152979).

1. Schütte C, Sarich M (2013) *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*, Courant Lecture Notes (Courant Institute of Mathematical Sciences, New York University, New York), Vol 24.
2. Gu C, et al. (2013) Building Markov state models with solvent dynamics. *BMC Bioinformatics* 14(Suppl 2):S8.
3. Bowman G, Pande V, Noé F (2013) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology (Springer, Dordrecht, The Netherlands).
4. Chodera JD, Noé F (2014) Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* 25:135–144.
5. Bollobas B, Riordan O (2006) *Percolation* (Cambridge Univ Press, Cambridge, UK).
6. Shmulevich I, Dougherty E (2009) *Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks* (SIAM Press, Philadelphia).
7. Grafarend E, Awange J (2012) *Applications of Linear and Nonlinear Models* (Springer Geophysics, Heidelberg).
8. Chandler D (1987) *Introduction to Modern Statistical Mechanics* (Oxford Univ Press, New York).
9. de Wiljes J, Putzig L, Horenko I (2014) Discrete non-homogenous and non-stationary logistic and Markov regression models for spatio-temporal data with unresolved external influences. *Comm Appl Math Comp Sci* 9(1):1–46.
10. Brockwell P, Davis R (2002) *Introduction to Time Series and Forecasting* (Springer, Berlin).
11. Granger CWJ (1988) Some recent development in a concept of causality. *J Econom* 39(1-2):199–211.
12. Sugihara G, et al. (2012) Detecting causality in complex ecosystems. *Science* 338(6106): 496–500.
13. Mosedale TJ, Stephenson DB, Collins M, Mills TC (2012) Granger causality of coupled climate processes: Ocean feedback on the North Atlantic Oscillation. *J Clim* 19: 1182–1194.
14. Boros E, Hammer P, Hooker J (1995) Boolean regression. *Ann Oper Res* 58:201–226.
15. Crama Y, Hammer P (2011) *Boolean Functions: Theory, Algorithms, and Applications* (Cambridge Univ Press, New York).
16. Horenko I (2010) Finite element approach to clustering of multidimensional time series. *SIAM J Sci Comput* 32(1):62–83.
17. Metzner P, Putzig L, Horenko I (2012) Analysis of persistent non-stationary time series and applications. *Comm Appl Math Comp Sci* 7(2):175–229.
18. Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *J R Stat Soc B* 58: 267–288.
19. Candes E, Wakin M (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25(2):21–30.
20. Burnham K, Anderson D (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, Berlin).
21. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8):1950–1958.
22. Lane TJ, Shukla D, Beauchamp KA, Pande VS (2013) To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr Opin Struct Biol* 23(1):58–65.
23. Smirnov S (2010) Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model. *Ann Math* 172:1435–1467.
24. Majda A, Timofeyev I, Vanden-Eijnden E (2002) A priori tests of a stochastic mode reduction strategy. *Physica D* 170:206–252.