Coupling of conformational switches in calcium sensor unraveled with local Markov models and transfer entropy

Tim Hempel^{1,2}, Nuria Plattner¹ and Frank Noé^{1,2,3*}

E-mail: frank.noe@fu-berlin.de

1: Department of Mathematics and Computer Science, Arnimallee 6, FU Berlin, 14195 Berlin, Germany

2: Department of Physics, Arnimallee 6, FU Berlin, 14195 Berlin, Germany

3: Rice University, Department of Chemistry, Houston, Texas 77005, United States

Abstract

Proteins often have multiple switching domains that are coupled to each other and to binding of ligands in order to realize signaling functions. Here we investigate the C2A domain of Synaptotagmin-1 (Syt-1), a calcium sensor in the neurotransmitter release machinery and a model system for the large family of C2 membrane binding domains. We combine extensive molecular dynamics (MD) simulations and Markov modeling in order to model conformational switching domains, their states and their dependence on bound calcium ions. Using transfer entropy, we characterize how the switching domains are coupled via directed or allosteric mechanisms and give rise to the calcium sensing function of the protein. Our switching mechanism contributes to the understanding of the neurotransmitter release machinery and the proposed methodological approach serves as a template to analyze conformational switching domains and study their coupling in macromolecular machines in general.

1 Introduction

Molecular modeling comes with challenges. Even though the advent of graphical processing units (GPUs) has delivered vast amounts of data from increasingly large molecular systems, modelers still have to cope with two fundamental problems: The sampling problem and the curse of dimensionality. In particular, a limiting factor of molecular modeling is the (possibly exponential) growth of the number of metastable states with system size. Sampling all metastable states reversibly can thus be unfeasible, especially if the described processes are extremely slow. Recently, dynamical graphical models have been proposed to overcome this problem by treating the evolution of configurations of many small molecular switches, e.g., dihedral rotamers, in an Ising-model like fashion.¹

Here we develop a complementary approach that, instead of working with very small spin-like molecular switches that can be readily identified from the structure, we identify conformational switching domains and model their kinetics with local Markov state models (MSMs). To this end, we partition the protein into subsystems that are modeled separately as conformational switching domains or, short, conformational switches. We therefore do not need to parametrize an MSM of the full structure and thus require less extensive molecular dynamics (MD) sampling. This decomposition approach is supported by experimental evidence for “independent dynamic [protein] segments”²

that have undergone evolutionary development in an independent fashion.² Further, we quantify the coupling between the local conformational switches by using transfer entropy in the time series of Markov states. Our approach is summarized in Fig. 1.

Our study is motivated by our interest for the role that protein conformational dynamics plays to determine spatio-temporal regulation of cellular processes and signaling. An abundant class of proteins that play important roles in cellular signaling (among other processes) is the C2 family consisting of 96,833 members³ of small beta-sandwich structured protein domains. C2 domains are membrane binders that in many cases require ion binding for activation.

One of the most exciting, yet unresolved problems connected to C2 domains is neurotransmitter exocytosis. Here, C2 domains play roles in vesicle recruitment (MUNC13) as well as in the actual release mechanism (Synaptotagmin-1, 2, 7 and 9).⁴ More specifically, synchronous neurotransmitter release is triggered by the double-C2-motif Synaptotagmin-1 (Syt-1) as a reaction to the electrically driven calcium concentration change. Syt-1 is specifically known to cause fast, concerted neurotransmitter release⁵ and hence is a good candidate for investigating calcium induced signaling with MD. Its structure and environment is summarized in Fig. 2. Even though the chain of reactions seems well understood, it is unclear how the ion concentration change triggers signaling on an atomistic scale. Possible mechanisms include a passive reaction due to the change in electrostatic potential, changes of protein structure or adapted conformational dynamics. While the first mechanism has been studied extensively, our aim is to investigate the role of Syt-1 conformational dynamics for its signaling function. Short preliminary test simulations of both Syt-1 domains, C2A and C2B, with and without membrane revealed that even single domain conformational dynamics is complex. Extensively sampling the whole system's rare transition events between all possible membrane binding modes and protein conformations is prohibitively expensive with direct MD. We thus propose to break the problem down into

obtaining kinetic models for individual components that will be coupled later.¹ As a first step, we thus decided to analyze the Syt-1 C2A domain as it showed the most interesting behavior in our tests.

The difficulties in understanding the inherently complex signaling process arise partially due to a lack of experimental methods capable of capturing the functional processes at a high spatial resolution over sufficiently long time. Molecular dynamics (MD) simulations are complementary here. They enable us to analyze C2 signaling function by assessing differences between the calcium bound and apo C2 domains using the Syt-1 C2A domain at atomistic resolution as an example. Markov modeling is then used to analyze the equilibrium reweighted conformational dynamics of the system, its Ca^{2+} configuration, and the calcium dependent information exchange between its different conformational switches.

2 Determining conformational switches and their coupling

We employ two main methods in order to analyze protein dynamics as a coupling of conformational switching domains: (1) Decomposing the protein structure into conformational switching domains and modeling their kinetics using local MSMs, and (2) Characterizing how these conformational switches are coupled by computing mutual information and transfer entropy between local MSMs. Here we outline the main concepts, see Fig. 1 for a visual summary and Methods for details.

2.1 Local Markov State Models identify conformational switches

Conformational switches are conceptually local protein regions that can exist in different discrete states, whose switching may be either spontaneous (i.e., by thermal noise), or induced by binding other changes of the equilibrium.

Figure 1: **Method workflow at simplified example.** From left to right: Two conformational switches are defined (top and bottom of protein) and raw MD trajectories of two ensembles (depicted by colors blue and black) are discretized into a (micro-) state space that is common to all ensembles of a particular conformational switch. Markov models are estimated separately, yielding transition matrices between metastable states and Markov state trajectories. In a following step, metastable states for each conformational switch can be identified between ensembles by comparing metastable state probability distributions, exploiting the common state space. Further, Markov state trajectories are used to estimate directional networks between conformational switches, exploiting the common time frame.

Here we outline an approach to identify conformational switches that undergo spontaneous transitions in the simulation data.

We first decompose the protein or protein complex into conformational switching domains. When these are not obvious as in the present application, one can choose from a variety of methods to find optimal decompositions of the full system into sub-systems, e.g. Refs.^{7,8}

Secondly, we discretize the conformational switch state space into metastable states using MSMs.^{9–15} In the context of MD, MSMs model exchange between usually very fine-grained protein states by counting transitions between them. Based on this empirical count matrix, a transition probability matrix under reversibility constraints is estimated. Equilibrium probabilities of protein conformations follow from an eigendecomposition of this matrix. Not only does Markov modeling provide a means to combine arbitrary numbers of short off-equilibrium trajectories, but it also has turned out to be an

efficient approach to learn about the slow and biologically relevant processes in vast MD data sets.^{16–18} We further note that kinetic modeling is necessary because in the finite data regime, MD datasets are no equilibrium samples. Thus, simple time averages in general do not represent thermodynamic ensemble averages. These can however be estimated using MSMs.

While various MSM approaches can be employed in order to identify metastable sets, such as Perron Cluster Analysis¹⁹ or VAMPnets,²⁰ we here estimate local Markov models via the hidden Markov model (HMM) approach described in Ref.²¹ HMMs yield a mapping of the complex dynamics into an easily interpretable set of few “hidden” or metastable states of the local conformational switches. Besides the kinetic model, metastable trajectories (Viterbi paths, cf. Methods) are used for the analysis.

Figure 2: Syt-1 C2A with interaction partners for synaptic exocytosis. **a:** Structure of Syt-1 C2A, active sites used in later analysis highlighted by circles; calcium ions are depicted as blue spheres. **b:** Syt-1 C2A (green) in primed fusion complex with its sister domain, Syt-1 C2B, and SNARE/Cpx (gray) based on PDB 5W5C.⁶ **c:** cellular context with plasma membrane (PM) and vesicle membrane (VM).⁶

2.2 Coupling conformational switches via transfer entropy

To analyze pairwise directed influences between conformational switches, we exploit (a) the fact that metastable trajectories of different local protein features are time-synchronous, and (b) that the systems in the HMM formulation are Markovian by definition. We can thus simplify Schreiber’s original definition of transfer entropy²² using the HMM transition matrix $p(x_{n+1}|x_n)$ and its stationary distribution $\pi_X(x_n)$ that are defined for states x_n in a time series X . From time series Y to time series X , transfer entropy $T(Y \rightarrow X)$ hence reads

$$T(Y \rightarrow X) = \sum_{\substack{x_n, x_{n+1} \in X \\ y_n \in Y}} \pi_{X,Y}(x_n, y_n) \cdot p(x_{n+1}|x_n, y_n) \cdot \log_2 \left(\frac{p(x_{n+1}|x_n, y_n)}{p(x_{n+1}|x_n)} \right). \quad (1)$$

Joint probabilities in X and Y , $\pi_{X,Y}$, are obtained from a transition matrix estimate in the combinatorial state space. The transition probability $p(x_{n+1}|x_n, y_n)$ is computed from the combinatorial transition matrix by marginalization over y_{n+1} . This takes into account the common time frame and probes for causal effects of the current state onto the state that

follows one Markov step in the future. Choosing the lag time of the underlying MSM ensures that our analysis indeed captures cross-talk of processes that our MSMs describe, even though other choices are theoretically possible.

Transfer entropy can be interpreted as a measure of the “incorrectness of [the] assumption” that “the state of $[Y]$ has no influence on the transition probabilities on system $[X]$ ”.²² It probes dependency in a direction-dependent fashion, i.e. is a heuristics for the amount of influence the process Y has onto the trajectory of a process X . According to this interpretation, we take $T(X \rightarrow Y) \gg T(Y \rightarrow X)$ as a statistical evidence for directed influence from process X to process Y .

3 Results

3.1 Syt-C2A consists of multiple metastable conformational switches

The analysis of conformational dynamics is conducted with focus on potentially functional protein regions. The Syt-C2A protein body adopts a β -sandwich fold which is naturally very rigid while the connections between the β -sheets are flexible. Furthermore, the largest flexible parts coincide with the calcium binding region (CBR), but there are several other flexible regions at various locations in the protein.

A simple analysis of root mean square fluctuations (SI) shows that only the CBR-1 loop becomes significantly stabilized with calcium binding. In the CBR-2, the opposite seems to happen, which hints towards an opening of the rigid body structure for this particular loop. CBR-2 is closely attached to the protein body in the crystal structure.

In order to obtain a human-readable model, we exploit our observation that Syt-C2A loops operate seemingly independent, and sub-divide Syt-C2A into three metastable regions of interest that are henceforth described as conformational switches: CBR-1, CBR-2 and a site opposite to the CBR (Fig. 2). The latter region will be called C78 since it connects β -sheets 7 and 8

(according to nomenclature of Ref.²³). According to our analysis, the remaining regions express no metastability and will thus not be analyzed. We employ a Markov state model analysis, here using Hidden Markov Models (HMMs) in order to identify the metastable states within these conformational switches (Methods).

3.2 Local Markov models reveal calcium mediated loop dynamics

Our models predict that CBR-1 can form an α -helix as depicted in Fig. 3a. The free energy difference between the helical and other states is $\Delta G_{\alpha} - \Delta G_{\text{other}} = 1.08^{+0.26}_{-0.26}$ kcal/mol (without calcium) and slightly shifts to $\Delta G_{\alpha}^* - \Delta G_{\text{other}}^* = -0.02^{+0.8}_{-0.11}$ kcal/mol (with calcium). The lifetime of this helical state becomes higher, from about $2.34^{+1.15}_{-0.48}$ μ s (without calcium) to $7.19^{+3.23}_{-1.46}$ μ s (with calcium). This means that the calcium binding enables a much more stable α -helical CBR-1 conformation.

This population change is mediated by the three calcium ions inside the CBR which interact with the CBR-1 loops by Coulomb interactions, attracting the polar acidic residue D172 (Fig. 3a). This residue is located at the center of the α -helix sequence; Coulomb attraction towards the protein core stabilizes this structure. We corroborate this observation by noting that D172 has a very low solvent exposure in its α -helical conformation (SI). The second most probable structure in the calcium bound protein is a β -hairpin-like structure (Fig. 3a). In general, it is very rigid and gives high solvent exposure to M137.

However, even without calcium, stabilized configurations exist, as depicted in Fig. 3a). One of them is characterized by M173 being buried within the CBR (red structure in Fig. 3a). The stationary distribution seems tilted towards the M173-burying state in the calcium unbound protein. Besides a disordered state (denoted by D in Fig. 3a), CBR-1 can form another α -helix located towards the N-terminal end of this loop with D172 at its center (depicted in blue).

The CBR-2 loop in the crystal structures appears tightly bound to the protein body. Even though this conformation is populated in both our MD datasets, we find metastable states that are characterized by re-arrangement of CBR-2-stabilizing salt bridges (Fig. 3b).

In particular we note that the salt bridge R199-D178 which is found in the crystal structure and in all of the calcium-unbound trajectories (Fig. 3b, state A) can be opened by calcium binding. R199 becomes solvent exposed in this highest populated calcium bound state. This change goes along with a particular Ca^{2+} arrangement within the binding pocket (Fig. 4b,c). Mechanistically, the CBR-2 residues are influenced by the calcium ion charges. This explains why in the calcium bound trajectories, positively charged residues such as R199 can loosen from the protein body. The calcium binding thus enables R199 to be solvent accessible. Concerning the activation time, the conformational changes between the tightly bound state and the loosely bound ones is in the order of 12 μ s with one exception: In the calcium bound case, it takes $50.65^{+61.45}_{-22.61}$ μ s to attach the loop to the protein body again (result not shown in Fig. 3b).

The C78 loop undergoes slow processes between three metastable states which could be identified in both, calcium bound and unbound data sets. With calcium binding, a shift of probability towards a V250-exposed state (cf. red structure in Fig. 3c) can be observed. Fig. 3c shows that two of the macrostates are characterized by releasing Valine residues from the protein body. These are V250 and V255 which can be attached independently of calcium binding. In order to switch between the V250 to a V255 exposing state, both residues must be attached to the protein body as an intermediate state.

3.3 Ca^{2+} distribution depends upon protein conformation

When analyzing the calcium configuration within the binding pocket we find that the Ca^{2+} distribution relaxes from the crystal structure positions into a broader distribution due to

Figure 3: Markov models of Syt-1 C2A conformational switches in the apo state (left) and calcium-bound state (right). Capital letters denote metastable states. Representative structures are shown and color-coded to distinguish states within bound/apo models. Circle sizes are proportional to state's equilibrium probabilities within each model. Arrows indicate transitions between metastable states, thickness proportional to the transition rate and annotated by pairwise inverse transition rate. **a:** MSMs of CBR-1 including sampled Ca^{2+} positions (black dots). States D and D* denote disordered structures that could not be assigned to unique structural elements. **b:** MSMs of CBR-2 with its loop in closed (crystal structure configuration, blue) and open (red, green, orange) conformations. **c:** MSMs of C78. Stick representations of single residues were drawn where instructive.

Figure 4: Ca²⁺ and binding pocket charge distribution. **a:** projection plane drawn in Ca²⁺-bound Syt-C2A structure. Panels **b** and **c:** Charge density averaged over 100 MSM samples and Ca²⁺ distribution in the calcium binding pocket. Ca²⁺ density is depicted by black points, positively charged amino acids as blue, negatively charged as red filled contour lines. MSM samples were drawn from CBR-2 MSM states A (panel b) and C (panel c), respectively. 2D projection (x, y) into the plane that is depicted in panel a. **d:** Conservation of charges among C2 domain family sequences (PFAM entry PF00168, seed alignment); histogram annotated with Syt-1 C2A domain residue sequence.

Coulomb repulsion between the ions. The crystal structure position of Ca1 acts as the inner edge of the binding pocket which is structured like a funnel along which the ions can collectively move. Therefore, Ca²⁺ distribution is dynamic and is concentrated across multiple possible binding sites that are more or less defined. The residues forming these binding sites are dynamic themselves, i.e. take different conformations in different metastable states as exemplified in Fig. 4b,c.

The importance of the charge distribution created by binding pocket residues becomes evident with a sequence alignment of all known members of the C2 family.³ The histogram of conserved charges derived from the sequence alignment (Fig. 4 d) confirms that charges at Ca²⁺ coordinating residues (Syt-1 C2A residues D172, D178, D232, D230, D238, also reported by Ref.²⁴) are indeed conserved. Further, we find that positive charges at position 199-200 (Syt-1 nomenclature) are conserved, underlining the significance of our findings. Another

distinct feature in the charge conservation histogram is a conserved region of positive residues around K190, which corresponds to the Lysine rich cluster described by Ref.²⁵ for the C2 domains of Syt-1 and rabphilin 3A.

We can further qualitatively generalize our results. Beyond the sequence conservation, C2 domains are highly conserved in structure as the resolved structures from the PFAM C2 family³ have an RMSD of $1.80 \pm 0.36 \text{ \AA}$ to the analyzed crystal structure (1BYN).²⁶ It thus becomes very likely that conformational dynamics plays a role for other C2 domains, too, most likely connected to CBR (or equivalent for non-calcium binders) or other non- β -sheet structural elements such as C78.

3.4 Distant protein features show allosteric behavior

In Markov state modeling and other statistical analysis methods, there is a trade-off between the complexity of the system to analyze and the

amount of detail one can resolve with statistical significance. The presented Syt-C2A example has distinct regions in which conformational switches occur. As commonly observed, obtaining a joint MSM including all combinations of conformational switches proved difficult due to insufficient statistics.¹ However, high-quality MSMs could be obtained for individual conformational switches, in the field of the remaining protein. Here we investigate how these different sub-systems interact using an information theoretic approach to model the network.

A qualitative assessment of un-directional influences using Pearson’s correlation coefficient (of coordinates) and mutual information (of MSMs) shows that there indeed is a weak coupling between conformational switches. As expected, coupling between spatially adjacent sites is higher than between distant sites. In particular, Ca^{2+} ions tighten the coupling between CBR-1 and CBR-2 (Fig. 5).

It is only natural to ask if this coupling is symmetric or if any kind of uni-directional influence between conformational switches can be inferred. A crucial point for the analysis is that a common time frame can be defined from the Markov state trajectories of the individual conformational switches as they occur simultaneously. Transfer Entropy $T(X \rightarrow Y)$ is a well established measure for assessing directional networks between time series X and Y .²² It can be interpreted as a statistical measure for the amount of information transported between two systems and, thus, is direction dependent. We say that X influences Y if $T(X \rightarrow Y) \gg T(Y \rightarrow X)$ (Methods). This attempts to model statistical *causality* (in the Granger-sense) between local protein features. In Granger’s original definition,²⁷ in simple words, a random variable X “causes” Y if knowledge of the past of X reduces the uncertainty of predicting Y ’s future. Transfer entropy relates to this concept;²⁸ therefore we apply a tool corresponding to a very broad definition of allostery.

The transfer entropy estimates from our data are presented in Fig. 5. We can identify a weak non-symmetric coupling between C78 and CBR-1 in the calcium bound case. The transfer entropy between this pair of loops is sig-

Figure 5: Coupling between conformational switching domains. Each link between conformational switches (enumerated 1-3) is quantified by naive computation of Pearson’s correlation coefficient between plain coordinates (r_P), mutual information (MI) and probed for directionality with transfer entropy (TE). TE color code refers to different directions as shown in crystal structure. Bootstrapping histograms for MI and TE are shown to assess the estimation error (MI, TE columns). The correlation coefficients between all pairs of heavy atoms of two protein regions are depicted by histograms to qualitatively assess linear couplings (r_P column).

nificantly higher from C78 to CBR-1 than vice versa, i.e. the prediction error for CBR-1 dynamics can be decreased by adding information about C78. This can be interpreted as a causal influence from C78 onto CBR-1, suggesting an allosteric coupling that is induced by calcium binding. It goes along with a slight increase in mutual information between this pair of conformational switches. We hypothesize that the effect is caused by long-range Coulomb interactions from the ions that is further mediated by the network of chemical bonds of the rigid protein body.

Please note that this result is purely based on statistical analysis and, however desirable, no mechanistic interpretation follows here.

4 Discussion

We have shown that challenges arising from the MD sampling problem and the curse of dimen-

sionality can be met by combining local Markov state models of conformational switches with an information-theoretic analysis of their coupling. We find that the estimation of few-state local Markov models is very robust and avoids commonly experienced statistical problems with estimating Markov models of the global protein state. Further, local models are easy to interpret. We therefore think that the present approach will be instrumental for the modeling of large proteins with loosely coupled local conformational switching domains.

Our results demonstrate that the effect of calcium binding to Syt-C2A is not a simple linear response as might be expected by increasing the charge in a specific part of the system, but rather a complex change of the system kinetics. Even though experimental²⁹ and MD studies^{30,31} suggested that Syt-1 C2A does not switch between well-defined conformations upon calcium binding, sampling the system extensively shows that the C2A domain indeed undergoes subtle and very rare conformational switching which is profoundly impacted by Ca^{2+} .

Further, it is well established that upon calcium binding, the C2A domain effectively switches off its membrane repelling charge,^{32,33} thus being attracted by the membrane while interacting with the SNARE-cpx complex (cf. Fig. 2 and Refs.^{5,6}). This model however does neither specify such an interaction nor does it explain how the signal (calcium ions) spreads to possibly distant interaction sites. Our model suggests that the conformational dynamics is altered upon calcium binding. This could be a potential mechanism for the calcium signal spreading to distant sites.

As any study based on MD simulations, this work relies on the forcefield that was used to generate the data. In particular, we note that the applied calcium ion model is approximate and e.g. does not account for electronic screening of charged moieties or environment-dependent changes in partial charges.^{34,35} Even though this should be taken into account when basing experiments on our results, we are confident that charge-mediated effects such as the ones laid out in this work can be described nev-

ertheless. Comparison to other forcefields, in particular polarizable ones, would improve the prediction quality but is beyond the scope of this conceptual work.

At this point, our system does not include elements of the cellular environment such as phospholipid membranes. We note, however, that the presented methodological framework would be applicable and could be used to quantify effects of membrane binding to C2A conformational dynamics.

Our model predicts a calcium induced population change towards a CBR-1 α -helix (Fig. 3a), burying negatively charged residues such as D172, thereby making the CBR more attractive to the membrane. This is consistent with the charge neutralization argument of Ref.³² Membrane attraction is further enhanced by hydrophobic residues that are not necessary for stabilizing the α -helix. The CBR-1 configurations that, according to our model, are highly populated in the calcium unbound protein (Fig. 3a) might even reinforce the membrane repelling function by moving hydrophobic residue M173, which is exposed in the disordered CBR-1 configuration, into the binding pocket. CBR related polar acidic residues are solvent accessible in this case since they are not bound by calcium ions. This has the side effect that CBR attraction to solvent calcium ions rises.

In the CBR-2, we find that calcium binding enables release of R199 from the protein body (Fig. 3b). As R199 has been reported to be of importance for membrane penetration,²³ ionic interactions with SNARE,³⁶ and forming the C2A-C2B interface,³⁷ our model might be key for understanding the interactions with the fusion machinery. In particular, Syt-1 interaction with SNARE might be triggered by calcium that, given its positive charge, induces the release of R199.

The C78 region generally contains many hydrophobic amino acids which opens the possibility of potential weak membrane interaction at this non-CBR site, a model previously proposed for Syt-1 C2B.³⁸ This interaction would allow the C2A domain to bind both, vesicle and plasma membrane, at the same time. Unfortu-

nately, we are not aware of thorough studies that investigate C78 or adjacent sites. In order to verify our model, we thus propose experimental validation. Such a validation would include assessing the relevance of this protein site e.g. by knock-out mutations of all hydrophobic residues located there. If effects on membrane fusion can be determined, singlet mutants could be used to validate our predictions. We expect especially Valine residues 250 and 255 to alter functionality since their configurations are characteristic for metastable states in the local models.

Our study further suggests cooperativity between the loops. As expected, our correlation and mutual information analysis shows that CBR-1 and CBR-2 loops are weakly coupled. Our model further predicts that calcium binding induces a stronger coupling. Transfer entropy analysis suggests that calcium binding increases the influence that C78 exerts onto CBR-1, even though the absolute magnitude might be rather low. This weak allosteric coupling insinuates that perturbation of C78, e.g. through membrane contact or binding to another protein in the fusion complex, effects on the conformational dynamics of CBR loops. Even though our model does not predict the properties of such an allosterically induced change, we believe that this effect should be taken into account in further studies of C2 domains.

Transfer entropy analysis as conducted in this work is limited to model the interplay of conformational switches. Modeling a high resolution allosteric pathway with this technique is however possible and subject of current research.

We can further qualitatively generalize the mechanisms described here for Syt-1 C2A to a larger spectrum of proteins that are not necessarily connected to neurotransmitter secretion. As sequence alignment of the charges in the C2 family (Fig. 4 d) has shown, especially our newly described calcium dependent mechanism of CBR-2 regulation might indeed be a structural feature of C2 domains in general.

The present approach relied on identifying the conformational switching domains treated by local Markov models with expert knowledge. The development of machine learn-

ing approaches to select these conformational switches automatically is a topic of future research.

5 Methods

5.1 Molecular dynamics simulations

In order to observe all important dynamical processes of the system, we have carried out extensive sampling. 220 trajectories of $2\mu\text{s}$ individual length have been generated, adding up to a cumulated simulation time of $440\mu\text{s}$, which to the best of our knowledge is the largest atomistic MD dataset generated for a C2 domain up to now. Since synaptic vesicle fusion can be triggered by calcium in less than $100\mu\text{s}$ ⁴ the assumption that all relevant processes have been sampled is justified.

Simulations have been carried out for the Syt-C2A domain in its Ca^{2+} -bound and Ca^{2+} -free state using the CHARMM36 force field.³⁹ The setup is based on the C2A structure contained in PDB 2R83 (which contains the double motif C2AB) and Ca^{2+} ion positions are initiated from PDB 1BYN which contains three calcium ions bound to the CBR. Starting structures for MD production runs were randomly drawn from a smaller precursive dataset. For both setups a water box of side length 6 nm was generated with a KCl ion concentration of 0.1 mol/l at neutral total charge using the TIP3P⁴⁰ water model. The setups contain 8363 (calcium free) and 8359 water molecules (with calcium), respectively. The setup as well as equilibration in the NVT and NPT ensembles (100 ps each) were conducted with Gromacs,⁴¹ production run simulations of $2\mu\text{s}$ were carried out in the NPT ensemble at 300K and 1 bar using the OpenMM software package.⁴² We used 1 nm cutoff for non-bonded interactions, PME electrostatics, and rigid water molecules. We further exploited the heavy hydrogen approximation as described in the openMM documentation, allowing a 5 fs integration step with the openMM Langevin integrator. An aggregated simulation time of $184\mu\text{s}$ was obtained

for the calcium free case, 256 μ s for the calcium bound case, in each case the accumulation of new data was continued until converged MSMs were obtained. Visual representation of molecular structures were obtained with VMD.⁴³

5.2 Local MSMs

For modeling local loop dynamics, the full (bound and apo) trajectory data was directly clustered according to a minRMSD norm. This includes superposing the full protein (using MDTraj⁴⁴) and defining discrete states according to the Euclidean norm between all atom coordinates of a specific region. This includes amino acid side chains and hydrogen atoms. We chose A170-T176 (CBR-1), R199-N203 (CBR-2) and T249-E258 (C78) to extract the local conformations, respectively. 15-30 discrete states and the k -means algorithm⁴⁵ were used for each local model. The selection of regions is inspired by a precursive full protein analysis using pairwise minimal residue distances between blocks of two residues and a state discretization in a low dimensional TICA projection.⁴⁶ Besides a relatively low spatial resolution at local sites, this analysis did not yield a valid MSM due to disconnected combinatorial states. It however motivated us to conduct analyses on local sites and to select the ones with metastable dynamics.

Discretizing local protein features after global minRMSD superposition could potentially yield spurious correlations between distant protein sites.⁴⁷ We note that this is unlikely because a) the largest mass of the protein is concentrated in its rigid β -sandwich body and b) the comparably low number of discrete states at local sites is too small to capture such subtle influences. Microstate definitions at example of CBR-2 are shown in the SI. Further and most importantly, the hidden metastable states that we discuss and use for mutual information and transfer entropy analysis reflect major internal loop rearrangements and are therefore most likely not affected by superposition artifacts.

HMMs were estimated separately for calcium bound and unbound data sets based on the common discretization described above. In

contrast to Markov state models, HMMs are not based on the assumption that the microstates obtained by clustering are approximately Markovian. Instead, a number of hidden states (usually less than the observed microstates) are introduced as explained in detail in Ref.⁴⁸ The number of hidden states was chosen such that it yields the best resolution of the slow processes in the protein and fulfills the validation criteria for Markov modeling (see below). All HMMs were estimated at lag time 50 ns using the PyEMMA software package.⁴⁵ We observed that only in relatively small configuration spaces (such as the ones defined by partitioning the protein into conformational switches), a robust, common discretization into Markov states could be obtained.

Corresponding HMM states between calcium bound and unbound HMMs were identified by computing KL-divergences between their observation probabilities per microstate (i.e. in the joint state space) and using a cutoff to identify hidden states across datasets (cf. SI). This operation is only possible because all data is discretized the same way i.e. microstate definitions do not differ. We use the KL-divergence between these distributions, an example (identification of macrostates for C78) is shown in the SI. Note that the KL-divergence becomes zero in the limit of equal distributions, thus one can define a cutoff of 2 below which states are identified for all models.

As our HMMs display high metastability, we can use the approximation $K = T - \mathbb{1}$ to compute the rate matrix K from the transition matrix T . The elements of K are pairwise inverse transition times.

Markov models were validated using two measures.¹⁵ First, the implied time scales were tested for convergence. We further tested for stationary distribution convergence to ensure that the presented results do not depend on the model lag time. Second, the Chapman-Kolmogorow equation was used to check for consistency between predictions of the presented models with direct estimates at higher lag times. Using Bayesian sampling of the posterior, the error was estimated for all the quantities computed by the MSMs/HMMs.⁴⁹ Re-

sults are presented in the SI.

5.3 Free energy validation of ion forcefield

For model validation purposes, we derive binding free energies of each Ca^{2+} using alchemical free energy perturbation methods^{50,51} for the crystal structure conformation.

As a buffer solution contains monovalent ions, it can be expected that calcium binding sites are normally saturated by those ions. This is confirmed by our calcium-free simulations in which we observe Potassium ions in the binding pocket. Alchemical free energy perturbation was hence simplified to a computational transformation of $\text{Ca}^{2+} \rightarrow \text{K}^+$ in the binding pocket and the opposite transformation in the solvent.

We further assume that in experiments, ions fill the binding pocket subsequently, i.e. we compute Ca3 in the presence of (Ca1, Ca2), Ca2 in the presence of Ca1, and Ca1 by itself. As we are interested in the binding free energy difference of a particular protein conformation, we apply harmonic constraints to the full protein. Relative free energies were computed using MBAR.⁵¹

Our results were validated according to Ref.⁵² This includes the validation of alchemical intermediate overlap and convergence analysis. Further, we repeat the same calculation at least 20 times to make sure that our results are reproducible.

The resulting ion binding free energies from the equilibrated crystal structure are comparable to experiment within error, however do not fully match the experimentally observed order (SI).

5.4 Combinatorial Viterbi path model

Since a central concept of this work are hidden state space trajectories or Viterbi paths, the most important concepts⁴⁸ are summarized here. The Viterbi algorithm maximizes the probability P of the hidden state se-

quence $q_1 q_2 \dots q_n$ with the observed sequence $O_1 O_2 \dots O_n$ given our model λ

$$\max_{q_1 q_2 \dots q_n} P(q_1 q_2 \dots q_n, O_1 O_2 \dots O_n | \lambda). \quad (2)$$

To this end, the Viterbi path is a maximum likelihood path on the hidden states. It naturally comes with the time step of the underlying HMM. Its estimation algorithm is also part of PyEMMA.⁴⁵

The degree up to which the local models communicate was estimated by comparing the stationary distribution of independent models to a new estimate. By assuming independence, this property follows directly from the product of local model probabilities per state.

Without assuming independence, a new estimate for the dynamics in combinatorial state space was made based on the local model Viterbi paths. It yields the combinatorial transition matrix or conditional probabilities $p(x_{i+1}, y_{i+1} | x_i, y_i)$ and joint stationary distribution $\pi_{X,Y}(x, y)$ of two processes X and Y .

5.5 Mutual information

Let X and Y be Markov processes that are localized at distinct spatial features of a protein and are well-described by (local, individual) HMMs. The mutual information M of X and Y is defined in terms of the joint probability $p_{X,Y}(x, y)$ and the independent probabilities $p_X(x)$, $p_Y(y)$ of $x \in X$ and $y \in Y$. In this work, we identify these probabilities p as the stationary probabilities of Markov models, most commonly denoted by π . We can thus write

$$M(X, Y) = \sum_{\substack{x \in X \\ y \in Y}} \pi_{X,Y}(x, y) \log_2 \left(\frac{\pi_{X,Y}(x, y)}{\pi_X(x) \pi_Y(y)} \right). \quad (3)$$

We obtain the independent stationary distributions π_X , π_Y from the local hidden Markov models and the joint stationary distribution $\pi_{X,Y}$ from a combined Markov model estimated based on the Viterbi paths. We use Eq. 3 to compute mutual information between all pairs of conformational switches of Syt-1 C2A .

For interpreting mutual information, it can be re-written in terms of the KL-divergence $\mathbf{D}(\cdot\|\cdot)$ as follows:

$$M(X, Y) = \mathbf{D}(\pi_{X,Y}(x, y) \| \pi_X(x) \pi_Y(y)) \quad (4)$$

Mutual information can thus be interpreted as a measure of the incorrectness of the assumption that both systems are independent.²²

Acknowledgement Funding from DFG SFB / TRR186 (project A12) is gratefully acknowledged. We further thank the Einstein Foundation Berlin (project SoOPic to N.P.) and the European Commission (ERC CoG 772230 “ScaleCell” to F.N.). We are grateful for discussions and feedback to Martin Scherer, Simon Bärfuss, Thomas Söllner, Fabian Paul, Shreyas Kaptan, Sebastian Stolzenberg, Guillermo Pérez-Hernández, Katarzyna Ziolkowska, Andreas Mardt.

Supporting Information Available

Method validation measures as well as further information about the Syt-1 C2A system are provided in the SI.

References

- (1) Olsson, S.; Noé, F. Dynamic Graphical Models of Molecular Kinetics. *Proc. Natl. Acad. Sci.* **2019**, *116*, 15001–15006.
- (2) Csermely, P.; Palotai, R.; Nussinov, R. Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. *Trends Biochem. Sci.* **2010**, *35*, 539–546.
- (3) El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A.; Sonnhammer, E. L. L.; Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S. C. E.; Finn, R. D. The Pfam Protein Families Database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432.
- (4) Südhof, T. C. Neurotransmitter Release: The Last Millisecond in the Life of a Synaptic Vesicle. *Neuron* **2013**, *80*, 675–690.
- (5) Chapman, E. R. How Does Synaptotagmin Trigger Neurotransmitter Release? *Annu. Rev. Biochem.* **2008**, *77*, 615–641.
- (6) Zhou, Q.; Zhou, P.; Wang, A. L.; Wu, D.; Zhao, M.; Südhof, T. C.; Brunger, A. T. The Primed SNARE–Complexin–Synaptotagmin Complex for Neuronal Exocytosis. *Nature* **2017**, *548*, 420–425.
- (7) Bernhard, S.; Noé, F. Optimal Identification of Semi-Rigid Domains in Macromolecules from Molecular Dynamics Simulation. *PLoS ONE* **2010**, *5*, e10491.
- (8) Boninsegna, L.; Banisch, R.; Clementi, C. A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures. *J. Chem. Theory Comput.* **2018**, *14*, 453–460.
- (9) Schütte, C.; Fischer, A.; Huisinga, W.; Deuffhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146–168.
- (10) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
- (11) Singhal, N.; Snow, C. D.; Pande, V. S. Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chem. Phys.* **2004**, *121*, 415–425.
- (12) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical Analysis of Conformational Dynamics in Biomolecules:

- Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126*, 155102.
- (13) Noé, F. Probability Distributions of Molecular Observables Computed from Markov Models. *J. Chem. Phys.* **2008**, *128*, 244103.
 - (14) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *PNAS* **2009**, *106*, 19011–19016.
 - (15) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.
 - (16) Plattner, N.; Doerr, S.; Fabritiis, G. D.; Noé, F. Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* **2017**, *9*, 1005.
 - (17) Paul, F.; Wehmeyer, C.; Abualrous, E. T.; Wu, H.; Crabtree, M. D.; Schöneberg, J.; Clarke, J.; Freund, C.; Weikl, T. R.; Noé, F. Protein–Peptide Association Kinetics beyond the Seconds Timescale from Atomistic Simulations. *Nat. Commun.* **2017**, *8*, 1095.
 - (18) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.
 - (19) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.
 - (20) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for Deep Learning of Molecular Kinetics. *Nat. Commun.* **2018**, *9*, 1–11.
 - (21) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139*, 184114.
 - (22) Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
 - (23) Jiménez, J. L.; Smith, G. R.; Contreras-Moreira, B.; Sgouros, J. G.; Meunier, F. A.; Bates, P. A.; Schiavo, G. Functional Recycling of C2 Domains Throughout Evolution: A Comparative Study of Synaptotagmin, Protein Kinase C and Phospholipase C by Sequence, Structural and Modelling Approaches. *J. Mol. Biol.* **2003**, *333*, 621–639.
 - (24) Fernández-Chacón, R.; Königstorfer, A.; Gerber, S. H.; García, J.; Matos, M. F.; Stevens, C. F.; Brose, N.; Rizo, J.; Rosenmund, C.; Südhof, T. C. Synaptotagmin I Functions as a Calcium Regulator of Release Probability. *Nature* **2001**, *410*, 41–49.
 - (25) Guillén, J.; Ferrer-Orta, C.; Buxaderas, M.; Pérez-Sánchez, D.; Guerrero-Valero, M.; Luengo-Gil, G.; Pous, J.; Guerra, P.; Gómez-Fernández, J. C.; Verdaguer, N.; Corbalán-García, S. Structural Insights into the Ca²⁺ and PI(4,5)P₂ Binding Modes of the C2 Domains of Rabphilin 3A and Synaptotagmin 1. *Proc. Natl. Acad. Sci.* **2013**, *110*, 20503–20508.
 - (26) Holm, L.; Laakso, L. M. Dali Server Update. *Nucleic Acids Res.* **2016**, *44*, W351–W355.
 - (27) Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438.
 - (28) Amblard, P.-O.; Michel, O. J. J. On Directed Information Theory and Granger Causality Graphs. *J. Comput. Neurosci.* **2011**, *30*, 7–16.

- (29) Shao, X.; Fernandez, I.; Südhof, T. C.; Rizo, J. Solution Structures of the Ca²⁺-Free and Ca²⁺-Bound C2A Domain of Synaptotagmin I: Does Ca²⁺ Induce a Conformational Change? *Biochemistry* **1998**, *37*, 16106–16115.
- (30) Wu, Z.; Schulten, K. Synaptotagmin's Role in Neurotransmitter Release Likely Involves Ca²⁺-Induced Conformational Transition. *Biophys. J.* **2014**, *107*, 1156–1166.
- (31) Bykhovskaia, M. Calcium Binding Promotes Conformational Flexibility of the Neuronal Ca²⁺ Sensor Synaptotagmin. *Biophys. J.* **2015**, *108*, 2507–2520.
- (32) Striegel, A. R.; Biela, L. M.; Evans, C. S.; Wang, Z.; Delehoy, J. B.; Sutton, R. B.; Chapman, E. R.; Reist, N. E. Calcium Binding by Synaptotagmin's C2A Domain Is an Essential Element of the Electrostatic Switch That Triggers Synchronous Synaptic Transmission. *J. Neurosci.* **2012**, *32*, 1253–1260.
- (33) Corbalan-Garcia, S.; Gómez-Fernández, J. C. Signaling through C2 Domains: More than One Lipid Target. *Biochim. Biophys. Acta - Biomembr.* **2014**, *1838*, 1536–1547.
- (34) Kohagen, M.; Lepšík, M.; Jungwirth, P. Calcium Binding to Calmodulin by Molecular Dynamics with Effective Polarization. *J. Phys. Chem. Lett.* **2014**, *5*, 3964–3969.
- (35) Leontyev, I.; Stuchebrukhov, A. Accounting for Electronic Polarization in Non-Polarizable Force Fields. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2613–2626.
- (36) Zhou, Q.; Lai, Y.; Bacaj, T.; Zhao, M.; Lyubimov, A. Y.; Uervirojnangkoorn, M.; Zeldin, O. B.; Brewster, A. S.; Sauter, N. K.; Cohen, A. E.; Soltis, S. M.; Alonso-Mori, R.; Cholle, M.; Lemke, H. T.; Pfuetzner, R. A.; Choi, U. B.; Weis, W. I.; Diao, J.; Südhof, T. C.; Brunger, A. T. Architecture of the Synaptotagmin-SNARE Machinery for Neuronal Exocytosis. *Nature* **2015**, *525*, 62–67.
- (37) Fuson, K. L.; Montes, M.; Robert, J. J.; Sutton, R. B. Structure of Human Synaptotagmin 1 C2AB in the Absence of Ca²⁺ Reveals a Novel Domain Association. *Biochemistry* **2007**, *46*, 13041–13048.
- (38) Rizo, J. Mechanism of Neurotransmitter Release Coming into Focus. *Protein Sci.* **2018**, *27*, 1364–1391.
- (39) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B* **2010**, *114*, 7830–7843.
- (40) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (41) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
- (42) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Comput. Biol.* **2017**, *13*, e1005659.
- (43) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

- (44) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.
- (45) 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.
- (46) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; Fabritiis, G. D.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.
- (47) Hünenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* **1995**, *252*, 492–503.
- (48) Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286.
- (49) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and Uncertainty of Reversible Markov Models. *J. Chem. Phys.* **2015**, *143*, 174101.
- (50) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632–2647.
- (51) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.
- (52) Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the Analysis of Free Energy Calculations. *J. Comput. Aided Mol. Des.* **2015**, *29*, 397–411.