Probabilistic models for conformational changes in protein structures

Thach Nguyen & Michael Habeck

Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen

cnguyen1@gwdg.de

Abstract

Conformational changes are often implicated in the biological function of biomolecules. To better understand large-scale structural transitions in biomolecules, we have developed two probabilistic models. Our first model describes a conformational change in terms of rigid-body dynamics. We have developed a probabilistic mixture model for automatically segmenting a set of input structures into rigid domains (Nguyen and Habeck (2016)). Here we present several improvements of the published algorithm. We simplify our model by reducing the number of model parameters and improve the robustness and convergence of the algorithm. By using methods for detecting communities in large networks, we obtain an approximate segmentation and an estimate of the number of rigid domains. These estimates provide useful initial values for our segmentation algorithm, which converges rapidly also for a large number of input structures. Our second model is based on a contact network. In contrast to existing elastic network models, we use non-Gaussian springs and allow for the rupture of network bonds during a conformational transition. Our adaptive network model is useful for flexible fitting where a high-resolution structure is deformed so as to fit structural data showing an alternative conformational state. We benchmark our adaptive network model on a large set of conformational transitions and illustrate its power on flexible fitting guided by sparse cross-linking data.

Generative model, Algorithm and Methods

The structure of each of the K rigid domains is represented by a $N \times 3$ matrix Y_k. We assume that the structure ensemble X is generated by rigid transformations of the domains:

$$X_{mn} \simeq R_{mk} Y_{kn} + t_{mk} \quad \text{if} \quad z_n = k\,. \tag{1}$$

We account for deviations due to experimental errors or shortcomings of the model by assuming a Gaussian error model, where each domain has its own error parameter σ_k:

$$p(X_{mn}|Y_k, R_{mk}, t_{mk}, \sigma_k, Z_{nk} = 1) = \mathcal{N}(R_{mk}Y_{kn} + t_{mk}, \sigma_k^2) \tag{2}$$

We develop an efficient Markov chain Monte Carlo algorithm to estimate the model parameters within a Bayesian framework (Nguyen and Habeck, 2016). The algorithm estimates the three-dimensional structures of the rigid domains as well as their location. For a particular choice of the prior probability over the segmentation variables, we obtain a Gaussian mixture model for protein ensembles. We also demonstrate that our sampling algorithm can be used to detect the number of rigid domains in a data-driven fashion, circumventing the need to choose the number of rigid domains beforehand.

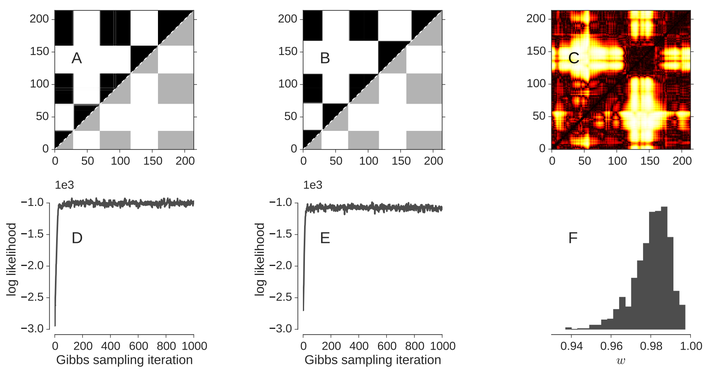

Figure 1: Segmentation analysis of Adenylate kinase. A: Segmentation based on independent segmentation labels using the label forgetting representation (upper diagonal matrix). The segmentation found in the literature is shown in gray below the diagonal. B: Label forgetting representation of the segmentation using sequentially correlated segmentation labels. C: Difference distance matrix between open and closed conformation of AdK. (Shown is the exponentially transformed absolute deviation between the distances to improve the visibility). D,E: Evolution of the log likelihood during Gibbs sampling for both independent and sequential segmentation. F: Estimated probability w that two successive domain assignments z_n, z_{n+1} are identical.

To assess the quality of our probabilistic segmentation algorithms in a systematic way, we ran both Gibbs samplers on more than 3000 examples from the DynDom database (Hayward and Berendsen, 1998). Each entry comprises a pair of protein structures showing a varying degree of conformational heterogeneity. The entries were downloaded and processed automatically using a Python script. For each entry, we ran 500 iterations of Gibbs sampling based on the prior probability independent segmentation and sequential correlated segmentation and a maximum of ten components ($K = 10$). Before starting the Gibbs samplers, the structures were centered and superimposed onto their average structure. The last 50 segmentations generated with the Gibbs sampler were used to assess the quality of the segmentation in terms of the overlap and the segmentation error).

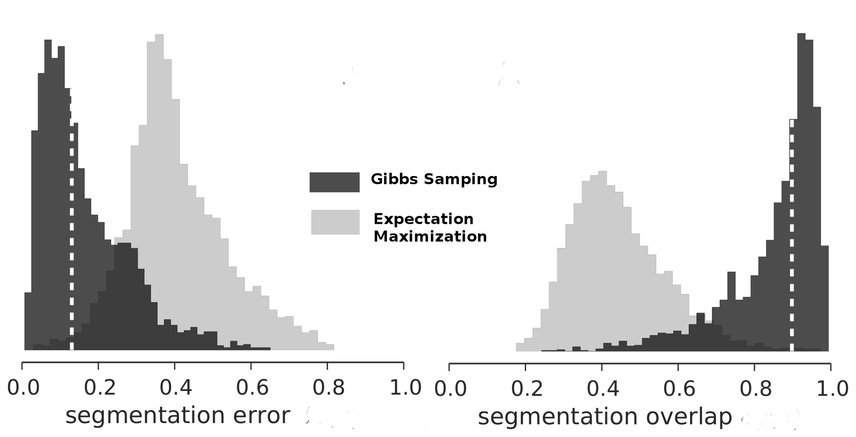

Figure 2: Large-scale benchmark on > 3000 entries from the DynDom database. Median values are indicated as dashed vertical lines. The light gray histograms are the results obtained with expectation maximization.

Improvements

Instead of using spectral clustering (Ng et al., 2001) for initial segmentation, we use Louvain methods for communities detection (Newman and Girvan, 2004) with Reichardt and Bornholdt configuration

$$H = \sum_{ij} (A_{ij} - \gamma(k_i k_j)/2m)\delta(s_i, s_j) \tag{3}$$

Where A_{ij} is adjacency matrix of full graph using negative exponental kernel , $\gamma(k_i k_j)$ is resolution parameter of weighted graph and m is number of edges.

We simplify our model by using single Y_n instead of Y_{kn} in the original model.

Probabilistic network model for structural transitions

We develop a probabilistic model for conformational transitions in biomolecules. The model can be viewed as a network of anharmonic springs that break, if the experimental data support the rupture of bonds. Hamiltonian Monte Carlo in internal coordinates is used to infer structural transitions from experimental data, thereby sampling large conformational transitions without distorting the structure. The model is benchmarked on a large set of conformational transitions from Sfriso et al. (2012). Moreover we demonstrate the use of the probabilistic network model for integrative modeling of macromolecular complexes based on data from crosslinking / mass spectrometry.

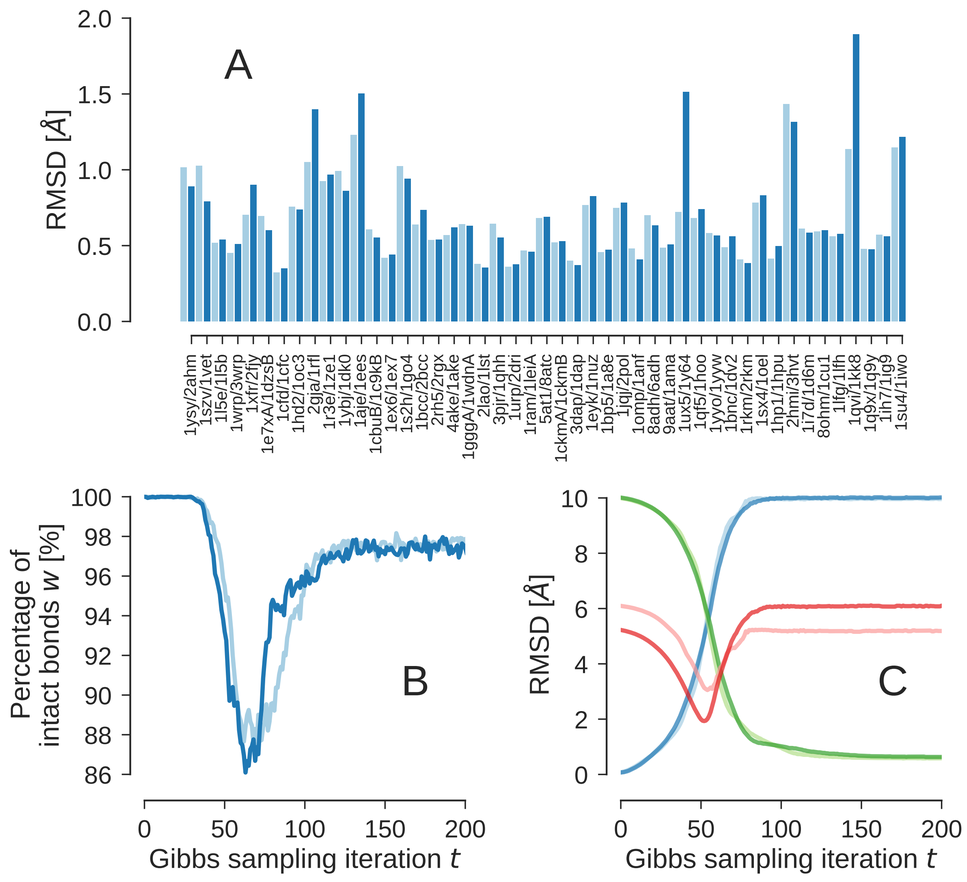

Figure 3: Simulation results for a benchmark of 94 structural transitions (light blue: forward transition, dark blue: reverse transition). **(A)** Cα RMSD between the final structure after 5000 steps of Gibbs sampling and the target state. **(B)** Evolution of the percentage of intact bonds during Gibbs sampling of the conformational change of 5'-NTase. **(C)** RMSD to the initial (blue lines), target (green lines) and an intermediate structure of 5'-NTase (red lines). The RMSDs of the forward transition (1hp1 $\rightarrow$ 1hpu) are shown in light colors, the reverse transition (1hpu $\rightarrow$ 1hp1) is shown in dark colors.

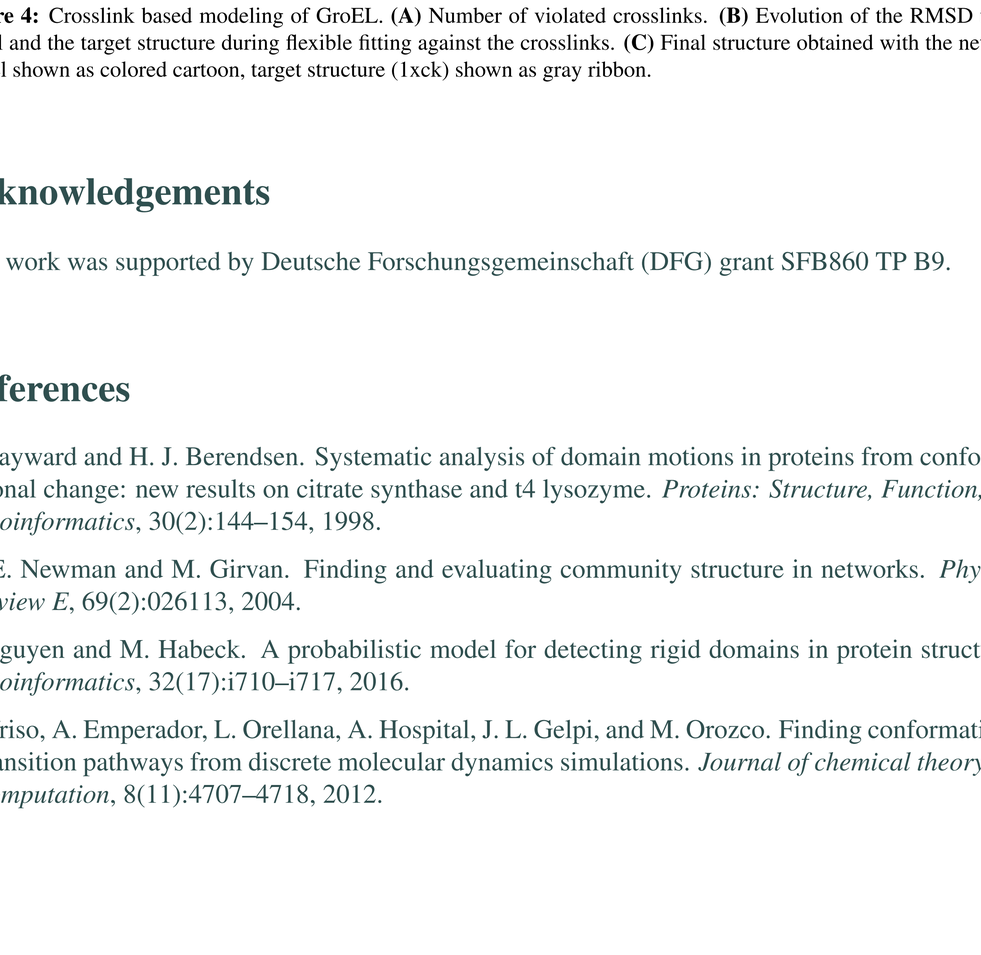

Figure 4: Crosslink based modeling of GroEL. **(A)** Number of violated crosslinks. **(B)** Evolution of the RMSD to the initial and the target structure during flexible fitting against the crosslinks. **(C)** Final structure obtained with the network model shown as colored cartoon, target structure (1xck) shown as gray ribbon.

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft (DFG) grant SFB860 TP B9.

References

S. Hayward and H. J. Berendsen. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and t4 lysozyme. *Proteins: Structure, Function, and Bioinformatics*, 30(2):144–154, 1998.

M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

T. Nguyen and M. Habeck. A probabilistic model for detecting rigid domains in protein structures. *Bioinformatics*, 32(17):i710–i717, 2016.

P. Sfriso, A. Emperador, L. Orellana, A. Hospital, J. L. Gelpi, and M. Orozco. Finding conformational transition pathways from discrete molecular dynamics simulations. *Journal of chemical theory and computation*, 8(11):4707–4718, 2012.