Changepoint Detection in Protein Dynamics Using Inter-residue Interactions

Anthony Ma
akma327@stanford.edu

Brian Do
bdo@stanford.edu

Irving Hsu
irvhsu@stanford.edu

Abstract

Changepoint detection is the problem of detecting significant changes in time series data over possibly many observables. In many cases, the feature space may be very large, but a particular change point can be attributed to a smaller subspace. Previous work has been published for developing effective changepoint detection algorithms over multivariate data sets. These methods can be applied in conjunction with molecular dynamics to determine significant protein conformational changes in a given trajectory.

We wondered whether certain feature spaces are better than others when trying to determine significant protein structural changes. Whereas previous literature has utilized large feature sets such as atomic coordinates and pairwise distance, we propose that smaller sets of more informative features such as the presence or absence of noncovalent side chain interactions would yield better results. In this paper, we explore the efficacy of inter-residue interactions for predicting changepoints and discuss the tradeoffs between the size of a feature space and informativeness of individual features.

1. Introduction

Understanding the dynamic physical structure of biological macromolecules such as proteins is essential for studying and manipulating their function. In recent years, molecular dynamics (MD) simulations have been used to model proteins and their surroundings in the presence of a molecular mechanics force field that follows Newtonian physics [2, 8]. These simulations have been used alongside crystallography to yield insights into protein folding, small molecule binding of their targets, the mechanism of rotary motor enzymes, and many other processes [1, 5, 6].

One important piece of information that MD simulations can provide is a list of the major structural transitions, or changepoints, that occur over the interval being simulated [4]. These transitions often correspond to important structural changes that enable a chemical reaction to be catalyzed or a molecule to be transported, as opposed to random fluctuations. As another example, protein folding often occurs in a punctuated manner, with long intervals of random noise occasionally punctuated by a large abrupt conformational change [1]. Thus, identifying changepoints can give us valuable information about how a protein functions, and about the domains that are critical for its function.

Historically, changepoints were manually identified from MD simulations through visual inspection, a time-consuming process because each simulation can have thousands to millions of frames [4]. Recent work has sought to automate this task. However, automatic changepoint detection has proved difficult for MD in particular because each frame contains thousands of data points in the form of atomic coordinates, whereas most existing algorithms are for univariate data.

To reduce the feature space, previous work has exploited the observation that the locations of atoms obey an anharmonic (super-Gaussian) distribution [9] or that distances between high-variance pairs of atoms can provide more relevant information than the individual locations [4]. In addition, machine learning methods such as dimensionality reduction, clustering, and Markov chains have been used to identify characteristic states as well as the transitions between these states [4]. Unfortunately, all of these methods are still very computationally resource intensive.

In this paper, we show that a small number of functional side-chain interactions between pairs of residues can be used to detect changepoints in MD trajectories much more quickly than using pan-atomic data. Changes in interactions such as salt bridges, pi stacking, disulfide bonds, or hydrogen bonds often correspond to large transitions in tertiary structure [10]. By binarizing the data to indicate simply the presence or absence of each interaction, we are able to sparsify the matrix and further reduce the computational resources that are required to identify changepoints. Importantly, since the algorithm also identifies the features that contribute to each putative changepoint, we are able to use the gain or loss of specific interactions to gain insight into the biology of each structural transition.

2. Methods

2.1. Changepoint Detection

Changepoints were detected using an algorithm called SIMPLE [4]. Briefly, SIMPLE chooses changepoints in multivariate data to maximize the likelihood that the observations between changepoints come from the same probability distribution. We ran SIMPLE on matrices where rows represented features (each method and sample had a different number of features) and columns represented 1,000 time points. α and β were set at 0.7 (the default values). To assess changepoints that were detected at different levels of regularization (the λ parameter), we first performed a binary search to find the minimum λ that would return exactly one changepoint. We then ran SIMPLE on 20 equally spaced λ values between 0 and this maximum λ value and saved the changepoints detected in each run.

2.2. Molecular Dynamics Data

We generated binary heat maps for different noncovalent side chain interactions for time series data derived from published full length trajectories for GPCRs from Dror et al [3]. In this paper we utilized Ma's prior noncovalent sidechain interaction calculator, to determine all salt bridges, pi cation, face-to-face pi stacking, and T-stacking aromatic interactions for the entire protein over the first 1000 frames of each trajectory, using pairwise distances between each residue. The geometric criterion was adopted from published literature values [7]. In this paper, we defined salt bridges as a cutoff of 4 Å between positively and negatively charged amino acid residues. Pi cation interactions were defined by a 6 Å cutoff between aromatic center and cationic amino acid, as well as a 60° deviation between orthogonal and center to cation vector. Pi stacking interactions have a 7 Å aromatic center distance cutoff, 30° deviation between the orthogonals to each aromatic plane, and will have an additional cutoff for the distance between projected aromatic center 1 upon aromatic plane 2 and center 2. The criteria for T-Stacking is similar except a 5 Å aromatic center distance cutoff is used instead.

Because SIMPLE relies on probability distributions in estimating changepoints, we added 1% jitter to the binarized data.

2.3. Comparison of Different Methods

To obtain significant changepoints for each sample, we ranked every observed changepoint by how many times it showed up across all of the λ values (i.e. if a changepoint was detected at all 20 λ values, it would rank more highly than a changepoint that was detected at the first 19 λ values but not the highest one). Since λ values were equally spaced across the range of meaningful values for each sample, we are able to use these ranks to compare the significance of changepoints. We then kept the top ranking 15 changepoints for each sample.

We then used permutation tests to assess the significant changepoints detected using the interaction method and the atomic location method against our baseline (pairwise distance). To do this, we randomly generated 1000 sets of 15 changepoints from a uniform $[0, 1000]$ distribution, then calculated for each set the root mean square distance (RMSD) from our baseline pairwise distance set. In calculating RMSD, we used the distance between each random changepoint and its closest neighbor in the pairwise changepoint set. As an example, if all observed changepoints correspond exactly to a baseline changepoint, then RMSD = 0. We then calculated the RMSD of the changepoint sets outputted by the interaction method or the atomic location method, and used its location in the random RMSD distribution to compute a p-value.

2.4. Visualization of Changepoints

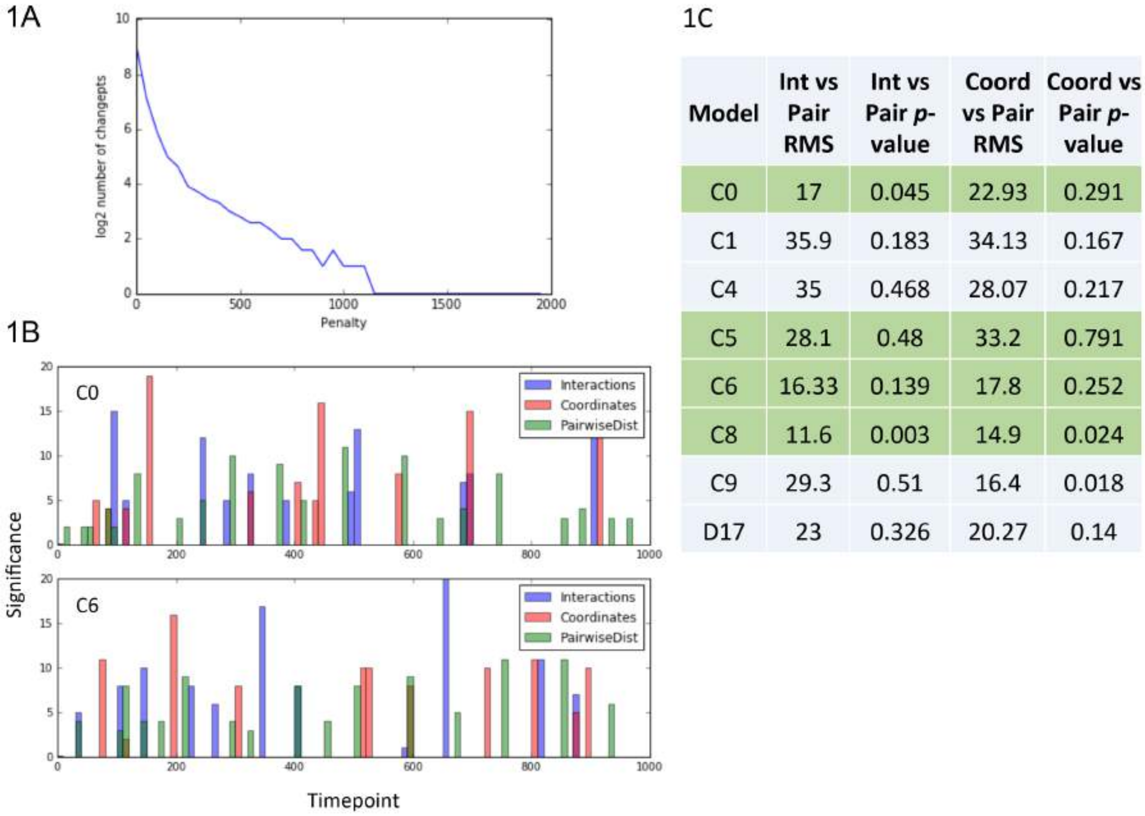
Visualization of the noncovalent side chain interaction for full trajectories was done through VMD utilzing Ma's graphicalTimelapseDisplay plugin. In the "simple" mode of the display software, lines are drawn between α-carbons of the pair of residues that share an interaction. In "detail" mode, the lines are drawn either between specific atoms or centroids of aromatic groups that share an interaction. Salt bridges, Pi-Cationic, Pi-Stacking, and T-Stacking interactions are represented by red, blue, green, and purple lines respectively.

3. Results

3.1. Comparing changepoints detected by different methods

We obtained eight full-length molecular dynamics simulations of GPCRs with and without drug binding from Dror et al [3]. At each frame in the simulation, the coordinates of each atom are recorded. We sought to find changepoints in these full length simulations using an algorithm called SIMPLE [4]. Briefly, SIMPLE takes in a matrix where rows represent features and columns represent timepoints. We reasoned that using different features derived from the raw coordinates could result in different timepoints being detected.

It was previously noted that pairwise distances between atoms provide the most optimal detection of changepoints when compared against a human-annotated baseline [4]. However, using this approach, the feature vector scales as n^2, making computation intractable even with parallelization. We term this approach the "pairwise distance method". To try to reduce the size of the feature space, we decided to attempt two other methods. In the first method, we simply input the raw coordinate values into SIMPLE. We term this



Model	Int vs Pair RMS	Int vs Pair *p*-value	Coord vs Pair RMS	Coord vs Pair *p*-value
C0	17	0.045	22.93	0.291
C1	35.9	0.183	34.13	0.167
C4	35	0.468	28.07	0.217
C5	28.1	0.48	33.2	0.791
C6	16.33	0.139	17.8	0.252
C8	11.6	0.003	14.9	0.024
C9	29.3	0.51	16.4	0.018
D17	23	0.326	20.27	0.14

Figure 1: (A) Representative plot of $\log_2$(number of changepoints detected) as a function of λ (penalty score). (B) Changepoints detected for C0 (top) and C6 (bottom) and their significance using three datasets. (C) Table showing root mean square (RMS) distances for the interaction and the atomic coordinates method when both are compared to the pairwise distance method. P values represent closeness of each method to the pairwise distance method.

the "coordinate method" or the "trajectory method". In the second method, we calculate the presence or absence of a small number of critical residue-residue interactions at each timestep (see Methods). These interactions can include salt bridges, pi-stacking, pi-cation interactions, or T-stacking interactions. For a protein with 300 residues, this matrix will contain on the order of 100 interactions. We term this the "interaction method." While this dataset is derived from the pairwise distance matrix, the time needed to calculate it is orders of magnitude smaller than the additional time needed to run SIMPLE.

We wanted to compare the changepoints detected by the trajectory method and the interaction method, compared to the changepoints detected by the pairwise method as a gold standard. When calculating changepoints, SIMPLE uses a penalty (a lambda value) where higher values increasingly restrict the number of changepoints outputted (Figure 1a). In order to compare changepoints from all three methods in an unbiased manner, we ranked changepoints by their significance, which we calculated by summing the number of penalty values where the changepoints were detected (Figure 1b). More significant changepoints will continue to be detected at higher (more restrictive) penalties, so this serves as a simple but useful heuristic. To normalize each dataset's tradeoff curve, we calculated the minimum penalty value at which only one changepoint was detected, and used evenly spaced penalty values between 0 and this cutoff as the values where changepoints were calculated.

Using the above strategy, for each dataset we calculated the top 15 changepoints for each of the three methods. We developed a root-mean-square distance (RMSD) approach to quantify the difference between these lists. In particular, we compared the RMSD between the interaction and pairwise methods with the RMSD between the coordinate and pairwise methods. If the interaction method performed significantly better than the coordinate method at predicting the changepoints detected by the pairwise method, the RMSD of the former would be much lower than the RMSD

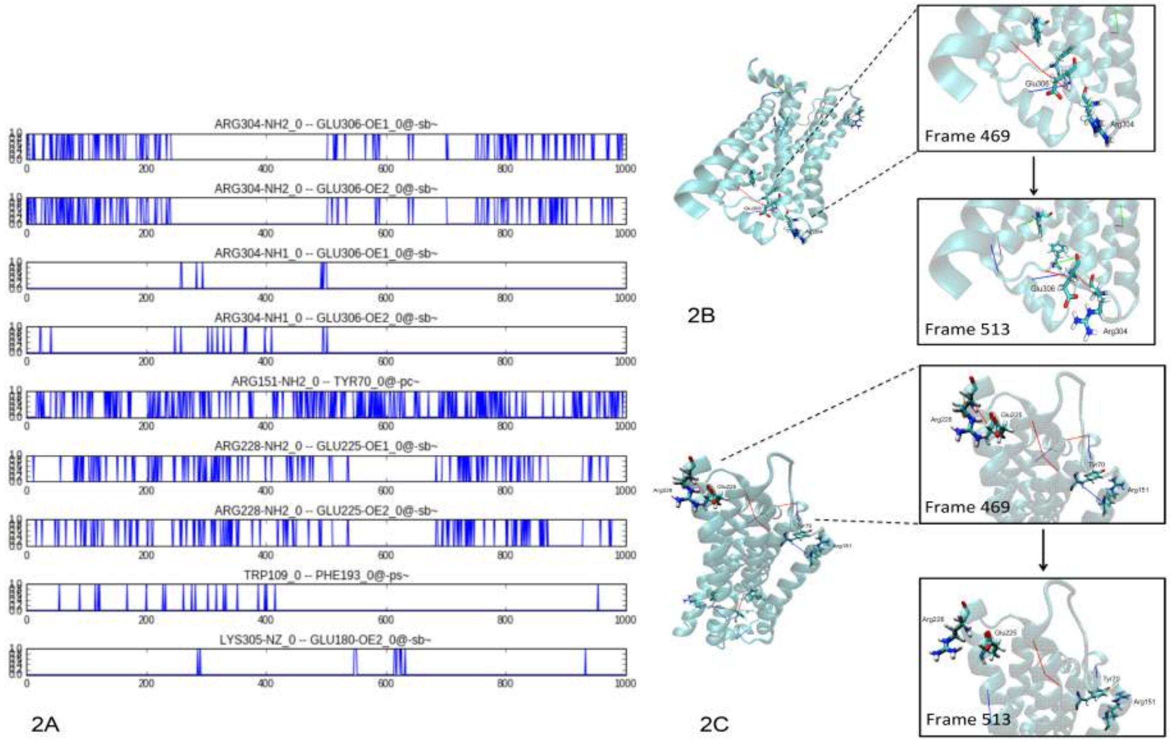


Figure 2: (A) Heatmap showing the presence or absence of the most significant features associated with changepoint at frame 503 for pnas2011a-C-0-all trajectory. (B) Formation of Glu306 $--$ Arg304 saltbridge before (frame 469) and after (frame 513). (C) Breaking of Arg228 $--$ Glu225 salt bridge and Tyr70 $--$ Arg151 salt bridge before (frame 469) and after (frame 513).

of the latter. We ran this for eight simulations and found that the interaction method performed better for four, while the coordinate method performed better for the other four (Figure 1c). Thus, while the interaction method does not always work better than the coordinate method at predicting gold standard changepoints, it does well in several cases with a feature space that is two orders of magnitude smaller than the pairwise method.

3.2. Changepoint analysis using interaction data reveals significant interactions involved in conformational change

Model C0 corresponds to a $1.0\mu s$ simulation of alprenolol binding to Beta-2 adrenergic receptor (B2AR) [3]. Since the interaction method performs better than the trajectory method at predicting the changepoints in the C0 simulation, we studied its changepoints in further detail.

Each changepoint identified by SIMPLE is outputted along with the noncovalent interactions that contribute most to the changepoint. In the C0 simulation, timepoint 503 was identified as the most significant changepoint. We plotted the presence or absence of nine most important interactions that changed around timepoint 503 (Figure 2a). Strikingly, the four interactions involved in the salt bridges between ARG304 $--$ GLU306, do not form until shortly after frame 503, whereas the salt bridge between ARG228 $--$ GLU225 and pi cation interaction between TYR70 $--$ ARG151 are constantly present until frame 503, where they break.

To see whether these changes in interactions corresponded to true conformational changes, we watched the simulations using the protein visualization software VMD. Shortly before the changepoint (at frame 469), we observed the presence of ARG228 $--$ GLU225 salt bridge and TYR70 $--$ ARG151 pi stacking interaction(Figure 2c). By frame 513, these interactions break in exchange for the formation of ARG304 $--$ GLU306 salt bridge in the intracellular region (Figure 2b). This interaction brings together the two transmembrane helices where ARG304 and GLU306 are located.

As further proof that conformational changes can be detected using the interaction method, we generated binary feature mappings for Model C6, which represents another simulation of alprenolol binding, and ran SIMPLE. This trajectory displays an even more significant changepoint (Figure 3a). Pi cation interaction between ARG131 $--$ TYR141 persists from frame 200 to approximately the major changepoint at $t^* = 656$. When $t > t^*$, ARG131 simultaneously forms a new salt bridge interaction with the neighboring

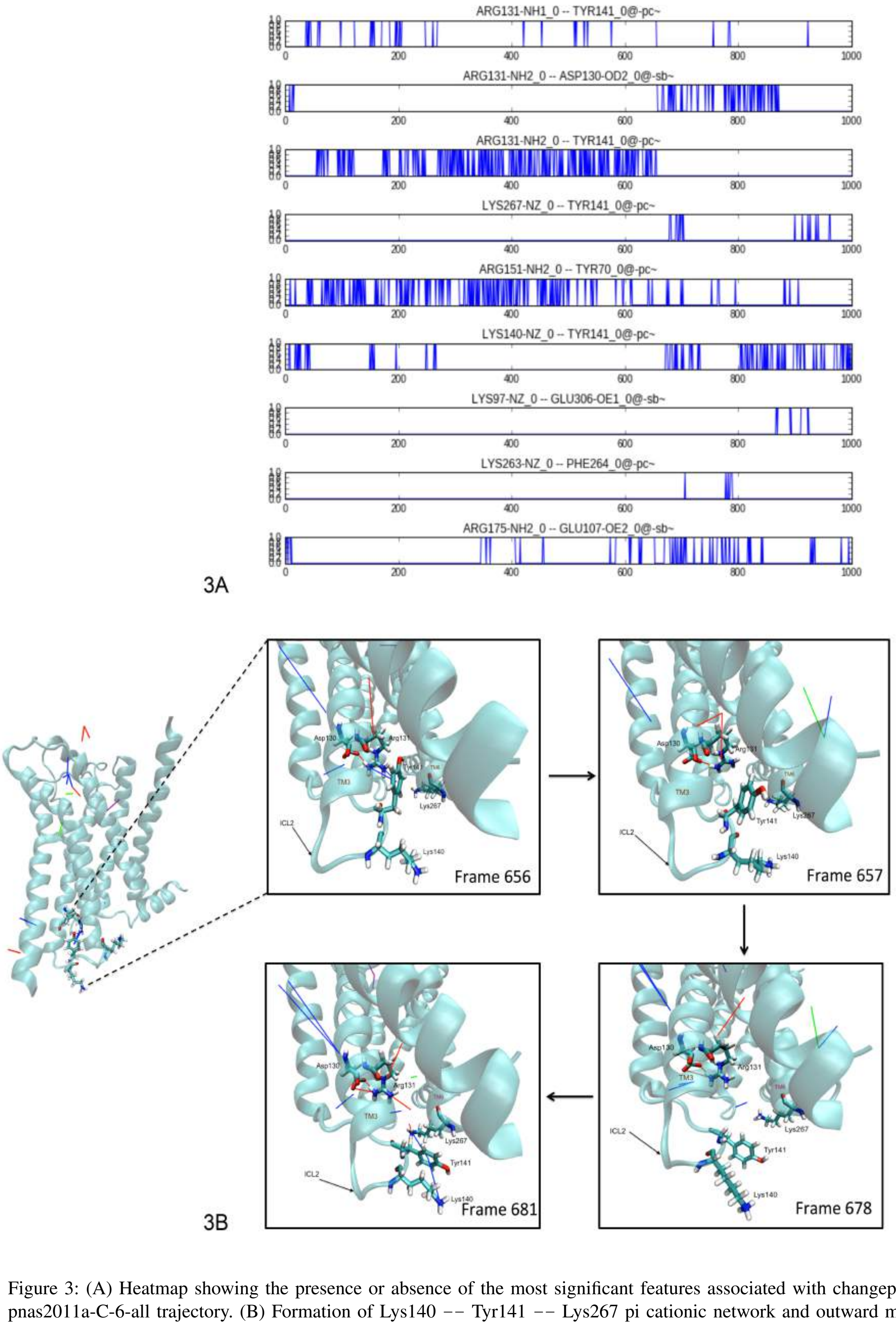


Figure 3: (A) Heatmap showing the presence or absence of the most significant features associated with changepoint at frame 656 for pnas2011a-C-6-all trajectory. (B) Formation of Lys140 −− Tyr141 −− Lys267 pi cationic network and outward movement of TM6 via interaction with ICL2 from frames 656 to 681.

ASP130 residue, and breaks the pi cation interaction with TYR141, thereby allowing the tyrosine residue to move freely and form simultaneous pi cation interactions with LYS267 and LYS140 as shown by the blue stretch starting at frame 681.

This change in residue-residue interactions potentially underlies a significant conformational change. Within a single frame after the changepoint, a salt bridge is formed between ASP130 and neighboring ARG131 in TM3 (Figure 3b). With the assistance of neighboring pi stacking interactions, the NH2 group in ARG131 is tilted enough to break the pi cation interaction between ARG131 and the TYR141 residue located on second intracellular loop (ICL2) between TM3 and TM4. This allows the free tyrosine residue to transition downwards and align itself with the LYS140 located on ICL2 and LYS267 located on TM6. By frame 678, the orthogonal to the aromatic plane in TYR141 aligns sufficiently well to the vectors from aromatic center to the cationic regions of the lysine residues. The results in a persistent pi cationic network between LYS140 -- TYR141 -- LYS267. As a result, the TM6 domain is displaced towards ICL2.

Certain GPCRs like the uOR receptor undergo a 10 Å outward displacement upon activation which would be deemed as a very significant event or change in the structure and function of the GPCR [11]. Given that the ICL2 is also known to be important for G protein binding, it would make sense that the activation of the GPCR via a displacement in TM6 works in tandem with the conformation change of ICL2, which would allow docking of G Protein complex to initiate a downstream phosphorylation cascade to carry out the appropriate ligand binding effect. Thus, using the interaction dataset, SIMPLE successfully identifies this major change in GPCR structure and function without the need for pairwise distance measurements.

4. Discussion

In this study, we sought to use residue-residue interaction data to provide a more efficient way to discover the same changepoints that are detected by using distances between pairs of atoms. This hypothesis rested on the fact that there are a much smaller number of functional noncovalent interactions than there are pairwise distances, which scale quadratically. In addition, we hoped that by using interaction data to detect changepoints, we would also be able to uncover the interactions that are associated with each changepoint, thus enabling us to quickly identify both a potential mechanism for a conformational change.

We compared the RMSD between the interaction method and the pairwise method, our gold standard, with the RMSD between the coordinate method and the pairwise method, on eight datasets. Out of these eight, the interaction method performed better on four, suggesting that at least in some cases we could recover similar changepoints from the pairwise data using a much smaller data sample. One limitation in our method is that the pairwise method was used as the gold standard, while in reality it is itself an output of an algorithm that produces false positives and false negatives compared with changepoints detected by humans [4]. In addition, we used the same alpha and beta parameters for all three methods, but because of the order-of-magnitude differences in the number of features it could have been useful to vary these parameters and assess the changepoints detected with each.

Because our method uses interaction data, it inherently biases for changepoints that have changes in more than one interaction. We used this as a proxy for functional conformational changes, but both do not always have to correspond. For instance, in Model C0 we observed a coordinated loss of the Arg228 -- Glu225 salt bridge and a gain in the Arg304 -- Glu306 salt bridge, but when we watched the simulation we observed only very minor conformational changes. On the other hand, in Model C6, the shift in Tyr141 from a salt bridge interaction network in TM3 with a double pi-cationic interaction with lysines in ICL2 and TM6 immediately signaled a large conformational shift that seems to correspond to a functional change in the intracellular region of the G protein. Thus, changepoints detected using interaction data (and all other kinds of inputs) need to be validated; the utility of the algorithm is to propose a list of candidate changepoints that can then be manually curated.

In future work, we hope to run SIMPLE upon full length trajectories rather than smaller segments and compare detected change points to the significant structural changes discovered in corresponding literature. Because one of the strengths of our method is that the detected changepoints are outputted along with the interaction features that contribute, we envision being able to run changepoint analysis on a large number of samples and using machine learning to derive general rules that dictate whether certain kinds of interactions correspond to true conformational changes. It would be interesting to compare change point detection accuracy using interaction data with other existing changepoint algorithms. Finally, we plan to explore other small-size feature spaces, potentially customized to the protein of interest. In the case of GPCRs, the data matrix can include center and mass and end to end vector representation of the seven transmembrane helices and curvature of the intra/extracellular loops. In this case, even compared to noncovalent side chain interaction, the feature space is drastically reduced, but information per feature could increase. It will also be interesting to evaluate the performance of changepoint detection using the same data but increasingly large feature spaces.

References

[1] Dill KA, MacCallum JL. The protein-folding problem, 50 Years On. *Science*, 338:1042–1046, November 2012.

[2] Dror RO, Dirks M, Grossman JP, Xu H, Shaw DE. Biomolecular simulation: a computational microscope for molecular biology. *Annual Review of Biophysics*, 41:429–452, June 2012.

[3] Dror RO, et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13118–13123, June 2011.

[4] Fan Z, Dror RO, Mildorf TJ, Piana S, Shaw DE. Identifying Localized Changes in Large Systems: Change-Point Detection for Biomolecular Simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(24):7454–7459, April 2015.

[5] Huang W, et al. Structural insights into μ-opioid receptor activation. *Nature*, 524:315–321, August 2015.

[6] Ito Y, Ikeguchi M. Molecular dynamics simulations of F_1-ATPase. *Advances in Experimental Medicine and Biology*, 805:411–440, December 2013.

[7] Marcou G, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling*, 47(1):195–207, December 2006.

[8] Monticelli L, Tieleman DP. Force fields for classical molecular dynamics. *Methods in Molecular Biology*, 924:197–213, 2013.

[9] Ramanathan A, Savol AJ, Agarwal PK, Chennubhotla CS. Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: Application to enzyme adenylate kinase. *Proteins: Structure, Function, and Bioinformatics*, 80(11):2536–2551, November 2012.

[10] Tina KG, Bhadra R, Srinivasan N. PIC: Protein Interactions Calculator. *Nucleic Acids Research*, 35:473–476, May 2007.

[11] Wheatley M, et al. Lifting the lid on GPCRs: the role of extracellulcar loops. *British Journal of Pharmacology*, 165(6):1688–1703, March 2012.