A PCA-guided Search Algorithm to Probe the Conformational Space of the Ras Protein

Rudy Clausen
Department of Computer Science
4400 University Drive
Fairfax, VA 22030
rclausen@gmu.edu

Amarda Shehu *
Department of Computer Science
School of Systems Biology
4400 University Drive
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT
We present a search algorithm to probe the conformational space of the RAS protein, a critical enzyme that employs conformational switching for its biological activity in the cell. The algorithm is guided by available experimental data on crystallographic structures of wildtype and mutant Ras. A principal component analysis (PCA) over these structures provides directions for exploration, which are used in combination with energetic refinement to sample low-energy conformations of Ras. Our results show that experimental structures are reproduced and the space is further populated with novel structures, warranting further investigation into structural characterization of Ras.

Categories and Subject Descriptors
G.3 [Probability and Statistics]: Probabilistic algorithms;
J.3 [Computer Applications]: Life and Medical Sciences

General Terms
Algorithms

Keywords
Ras; structure space; search; PCA; multiscaling

1. BACKGROUND
While proteins use specific three-dimensional structures to dock onto molecular partners and perform critical functions in the cell, proteins are also known to be highly flexible [11]. In particular, the Ras enzyme, which is the object of our preliminary investigation here, exploits its structural flexibility to switch between a GTP-bound active state and a GDP-bound inactive state. Ras is critical to signaling pathways in cell growth and proliferation. Many studies have shown that a large range of mutations in Ras affect its conformational switching, often leading to cancer and developmental diseases. In fact, mutations in Ras are found in more than 25% of human cancers [1].

We present here a probabilistic search algorithm to model the structural flexibility in Ras beyond what is documented by experimental investigation of stable wildtype and mutant Ras structures. Effectively, we propose an algorithm to obtain a discrete representation of low-energy regions in the underlying energy surface in terms of an ensemble of low-energy conformations. These conformations are representatives of stable and semi-stable structural states and may present interesting novel structures not currently probed in the wet laboratory. The motivation for doing so in Ras, in particular, is to provide further understanding of structural modulations that may possibly lead to new binding sites, compounds, or interacting molecular partners for Ras.

2. METHODS
The proposed algorithm employs knowledge about Ras in order to feasibly explore a high-dimensional conformational space. Unlike previous studies on Ras based on Molecular Dynamics [2], which tend to have limited exploration capability [11], the algorithm avoids issues of timescale by operating in a low-dimensional space obtained by PCA on X-ray structures of Ras. The PCA-guided approach here is similar to how Normal Mode Analysis has been used in other contexts and implementations to explore flexibility in peptides, protein loops, or protein binding sites [4]. The algorithm proposed here conducts biased random walks in conformational space and exploits multiscaling, the employment of various representations of different resolution, in order to balance between exploration and energetic refinement of promising regions in the Ras structure space.

A desired number of low-energy conformations is specified a priori, and biased random walks are conducted until this number is met. All walks start from a given (energetically-refined) X-ray structure of Ras at Cα level of detail. Consecutive conformations \( C_i \) and \( C_{i+1} \) in a current biased random walk are generated through a perturbation move followed by a local optimization move. The perturbation move makes use of a PCA-based direction vector to modify \( C_i \). The vector is a linear combination of the \( k \) PCs with highest eigenvalues, where \( k \) is selected to capture \( \geq 80\% \) of the structural variability in experimental data while control-
ling dimensionality. The weight for the first and top PC is sampled at random in $[-1, 1]$, whereas the others are scaled based on the ratio of their eigenvalue to that of the first PC. The obtained conformation is then subjected to a local optimization move that adds representational detail as follows. The conformation is first added backbone atoms and then refined with an all-atom energy function. If its energy is within a threshold, $C_{i+1}$ is added to the current walk. Otherwise, it is considered a dead end, and a new biased random walk is started.

3. RESULTS

Analysis of the PCA over 46 representative X-ray structures of Ras and analysis of deformations between actual structures and structures reconstructed with $k \in \{5, 7, 10\}$ PCs allow concluding that $k = 10$ PCs are most effective (data not shown). Analysis of the backbone reconstruction process over the actual X-ray structures also allows concluding that BBQ [3] is the most suitable protocol to recover original structures. We have benchmarked the algorithm over various initial conditions, and what we show below highlights the exploration power of the algorithm. Figure 1 projects 10,000 conformations obtained when initiated from the X-ray GTP-bound Ras structure under PDB id 1qra and 10,000 conformations obtained when initiated from the X-ray GDP-bound Ras structure under PDB id 4q21. The projection is over the top two PCs for ease of visualization. Figure 1 additionally shows all X-ray structures (including those not used for the PCA). The projection shows that the algorithm has high sampling capability, covering regions of the space also populated by X-ray structures and even new regions not probed in experiment.

Figure 1: Projection of space explored by the algorithm on the top two PCs. Conformations obtained when initiated from PDB id 1qra are in red. Those obtained when initiated from PDB id 4q21 are in green. X-ray structures of Ras are in blue.

4. CONCLUSION

The algorithm presented here can be considered similar in spirit to basin hopping frameworks that make repeated use of perturbation and local search [7, 6, 10]. Ongoing work enhances the preliminary framework presented here with tree-based search based on previous work showing such frameworks have high sampling capability in diverse molecular modeling settings [8, 9, 5]. The proposed algorithm is an important step towards detailed modeling of the space of stable and semi-stable structures of a protein, as it promises to feasibly extend our computational capabilities to modeling such large spaces all the while directly employing available experimental data.

5. ACKNOWLEDGMENTS

This work is supported in part by NSF CCF Grant No. 1016995 and NSF IIS CAREER Award No. 1144106.

6. REFERENCES