**Blue Waters Illinois Allocation Report 2016**

**Primary contact email and name:**
NAME: Dr. Gustavo Caetano-Anolles
EMAIL: gca@illinois.edu

# Title: "Evolutionary dynamics of the protein structure-function relation and the origin of the genetic code"

**PI: Gustavo Caetano-Anolles,** Department of Crop Sciences, University of Illinois at Urbana-Champaign, 1101 W Peabody Drive, Urbana IL 61801. Phone: 217-333-8172. Email: gca@illinois.edu

**Co-PI(s): Frauke Grater,** Molecular Biomechanics, Heidelberg Institute for Theoretical Studies (HITS) gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany. Phone: +49-6221-533267. Email: frauke.graeter@h-its.org

**Collaborator(s): Fizza Mughal,** University of Illinois at Urbana-Champaign.

**Subject:** BIOLOGY, CHEMISTRY, & HEALTH

| | |
|---|---|
| Molecular | X |
| Cellular | |
| Medicine | |
| Biophysics | |
| Other (list) | |

**Number of images:** 3

**Executive Summary**

The flexibility of the unstructured regions of proteins links protein dynamics to function. These unstructured regions have distinct set of motions that can be mapped onto the corresponding functions. However, the structure-to-function paradigm in proteins and its dynamics remains an uncharted territory for molecular dynamic exploration. Structural flexibility has been shown to be evolutionarily conserved and to be the product of molecular evolution. Our investigation of flexible loop regions of proteins strives to decipher the evolutionary signal embedded in the dynamics of structural change. With this goal in mind, we simulated the molecular dynamics of well-annotated loops from proteins domains of aminoacyl-tRNA synthetase enzymes, and in the later phase, of metaconsensus enzymes, amounting to ~9000 nanoseconds of simulation time. The analyses of the resulting molecular trajectories, while in its infancy, draws attention to patterns that may help explain the evolutionary relation between protein dynamics and molecular function.

**Key Challenges**

The protein loops, the major source of overall flexibility in the structure of proteins, have been found to be important for function and structural stability of these macromolecules [1]. By virtue of this flexibility, it is possible that there may be characteristic motions associated with loop structures and functions. In addition, protein dynamics and flexibility have been observed to be strongly conserved [2]. Protein dynamics and function are also correlated, evident by the similar motions of non-homologous enzymes with similar functions [3]. Therefore, the conserved nature of protein flexibility and dynamics may hold the key to unlocking the underlying evolutionary drivers of protein function. The identification of evolutionary drivers is possible by reconstruction of evolutionary trajectories and is of considerable importance to synthetic biology and translational medicine [4]. In order to investigate this further, we performed proof-of-concept molecular dynamics simulations of protein loops in Blue Waters. The study concluded that the length of flexible intrinsically disordered protein regions impacts folding dynamics [5]. We followed up these experiments with simulations of the loops associated with the aminoacyl-tRNA synthetase (aaRS) enzymes, as a candidate to study the protein structure-function paradigm at the heart of the genetic code. This is because the evolutionary origins of the genetic code and aaRS enzymes are thought to be linked to each other [6]. A preliminary global analysis of variables such as root-mean-square deviation (RMSD) and radius of gyration (R) show a tendency towards an overall decrease in their values. Here we extend the initial analysis.

**Accomplishments**

We used the current allocation to complete our previous round of simulations that required a total of 87 protein loops corresponding to the domains of the aaRS enzymes. These enzymes are responsible for specifying the specificities of the genetic code, i.e. the recognition and highly specific loading of each amino acid on the the 3' end of the specific tRNA molecule. The simulations were performed using the NAM 2.9 platform with the forcefield parameters described by CHARMM36 on a timescale of 10 nanoseconds for each loop. The loops were classified by the 'Density Search' system of classification by the ArchDB database. They were annotated with molecular functions directly derived from the gene ontology (GO) database. Global parameters such as RMSD and R was computed for the resulting simulations. Additionally, local parameters were calculated using RMS fluctuations, principal component analyses (PCA) and community behavior captured via networks based on residue movements in the trajectory. The use of PCA is motivated by our objective to map specific motions of the protein associated with the respective function. The residues in the unstructured region possessed higher values of RMS fluctuation compared to the bracing secondary structures of the loop. The communities of residues in our network analyses (figure 3) exhibit interesting behavior in terms of negatively cross-correlated motions (figure 2). Remarkably, we noted that the betweenness values plotted from the resulting networks for each of the loops showed high values for residues in the C-terminus secondary structure for the majority of the structures that were analyzed.

We extended this methodology of simulation to include protein loops associated with domains found in metaconsensus enzymes [7]. These metaconsensus enzymes belong to EC groups that carry consensus of three comparative bioinformatics methods based on sequence [8], structure [9] and metabolic reactions [10]. We performed 116 simulations, each of a duration of about 70-75 nanoseconds. In addition to our objectives of finding correlations between function and flexibility, the longer timescales for this dataset provides the grounds to test whether the relatively smaller timescale of 10 nanoseconds can effectively capture the conformational landscape of the protein loops compared to the longer timescale. We are now planning to test the presence of patterns in the community structures, refining our approach to map loop motions onto protein function. Ultimately, we will be plotting results from these analyses onto robust phylogenomic timelines to study the effect of evolution on protein function [11].

**Why it matters**

Understanding the structure-function paradigm from an evolutionary perspective represents a fundamental but uncharted territory of exploration in molecular biology. Since "*nothing in biology makes sense except in the light of evolution*" [12], our exploration has the potential to unravel basic knowledge about mechanisms and dynamics needed for engineering and medical applications. Our preliminary work substantiates future and more ambitious proposals.

**Why Blue Waters**

The immense computing power of Blue Waters has been key to enable our structure-function studies of proteins at the intersection of the evolution of the genetic code and metabolic networks and the structural dynamics of proteins. Using our current allocation for simulating 13 initial and current 116 protein loops, including some simulations used for benchmarking and tests, we have performed nearly ~9000 nanoseconds (9 milliseconds) worth of molecular dynamic simulations. Our new set of simulations provides a means of comparing how loops belonging to proteins with multiple domains behave (including aaRS enzymes) relative to those possessing single domains among the metaconsensus enzyme groups. In order to keep pace with the growing availability of proteomics data, it is imperative that substantially powerful machines such as Blue Waters continue to be developed to explore this "land of opportunity" [13]. This includes studying complex evolutionary processes that otherwise may not be possible.

**Next Generation Work**

The results obtained from our preliminary analyses suggests the presence of patterns in the dynamics of protein structure that are linked to function. There is potential to cover uncharted territory of evolutionary drivers in protein function and solve one of the most basic yet difficult conundrums in nature: does "form follow function" or vice-versa?

**References**


1. Espadaler J, Querol E, Aviles FX, Oliva B. Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics. 2006;22: 2237–43. doi:10.1093/bioinformatics/btl382

2. Marsh JA, Teichmann SA. Protein flexibility facilitates quaternary structure assembly and evolution. PLoS Biol. Public Library of Science; 2014;12: e1001870. doi:10.1371/journal.pbio.1001870

3. Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle LM, Kroon G, et al. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat Struct Mol Biol. 2013;20: 1243–9. doi:10.1038/nsmb.2676

4. Wilke CO. Bringing molecules back into molecular evolution. PLoS Comput Biol. Public Library of Science; 2012;8: e1002572. doi:10.1371/journal.pcbi.1002572

5. Caetano-Anollés G, Gräter F, Debès C, Mercadante D, Mughal F. The dynamics of protein disorder and its evolution: Understanding single molecule FRET experiments of disordered proteins. Blue Waters Annu Rep. Urbana, Illinois; 2014; 100–101.

6. Caetano-Anollés G, Wang M, Caetano-Anollés D. Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. PLoS One. Public Library of Science; 2013;8: e72225. doi:10.1371/journal.pone.0072225

7. Goldman AD, Baross JA, Samudrala R. The enzymatic and metabolic capabilities of early life. PLoS One. Public Library of Science; 2012;7: e39912. doi:10.1371/journal.pone.0039912

8. Tatusov RL. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. Oxford University Press; 2001;29: 22–28. doi:10.1093/nar/29.1.22

9. Caetano-Anollés G, Caetano-Anollés D. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. J Mol Evol. 2005;60: 484–98. doi:10.1007/s00239-004-0221-6

10. Srinivasan V, Morowitz HJ. The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol Bull. 2009;216: 126–30. Available: http://www.biolbull.org/content/216/2/126.abstract

11. Kim KM, Caetano-Anollés G. The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol. BioMed Central Ltd; 2012;12: 13. doi:10.1186/1471-2148-12-13

12. Dobzhansky T. Nothing in biology makes sense except in the light of evolution. American Biology Teacher 1973;35(3): 125-129.

12. Berezovsky IN, Guarnera E, Zheng Z, Eisenhaber B, Eisenhaber F. Protein function machinery: from basic structural units to modulation of activity. Curr Opin Struct Biol. 2017;42: 67–74. doi:10.1016/j.sbi.2016.10.021

**Publications and data sets**

Mughal, F., G. Caetano-Anollés and F. Gräter. Mining the evolutionary dynamics of protein loop structure and its role in biological functions. *Blue Waters Annual Report* (Urbana, Illinois, 2015), pp.130-131.

Caetano-Anollés, G., F. Gräter, C. Debès, D. Mercadante, and F. Mughal, The dynamics of protein disorder and its evolution: Understanding single molecule FRET experiments of disordered proteins. *Blue Waters Annual Report* (Urbana, Illinois, 2014), pp. 100-101.

FIGURE 1. Protein loop 1B7Y_B_408 corresponding to the a.6.1.1 SCOP domain with residue connections based on motions during the trajectory. Red denotes positive correlations while blue signifies negative correlations.

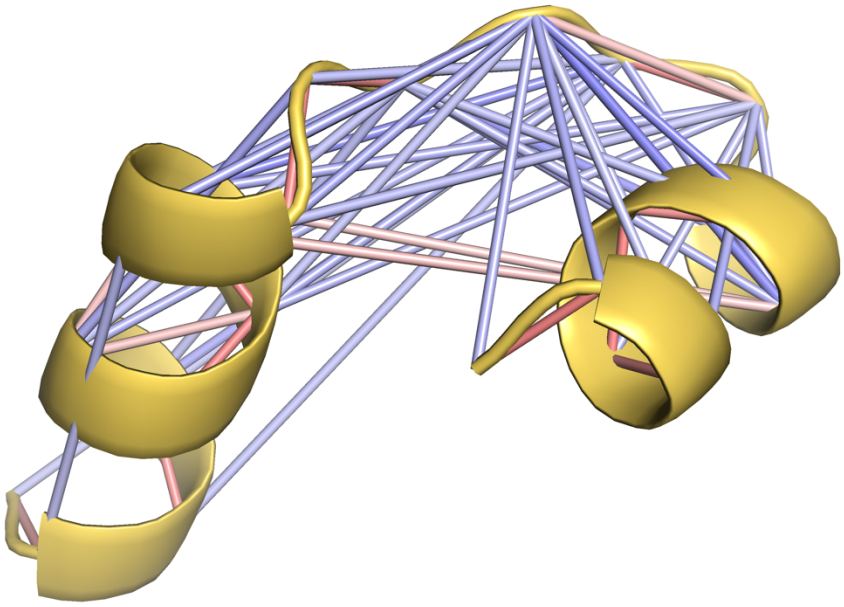

FIGURE 2. Dynamical Cross Correlational Map of the protein residue backbone of 1B7Y_B_408.

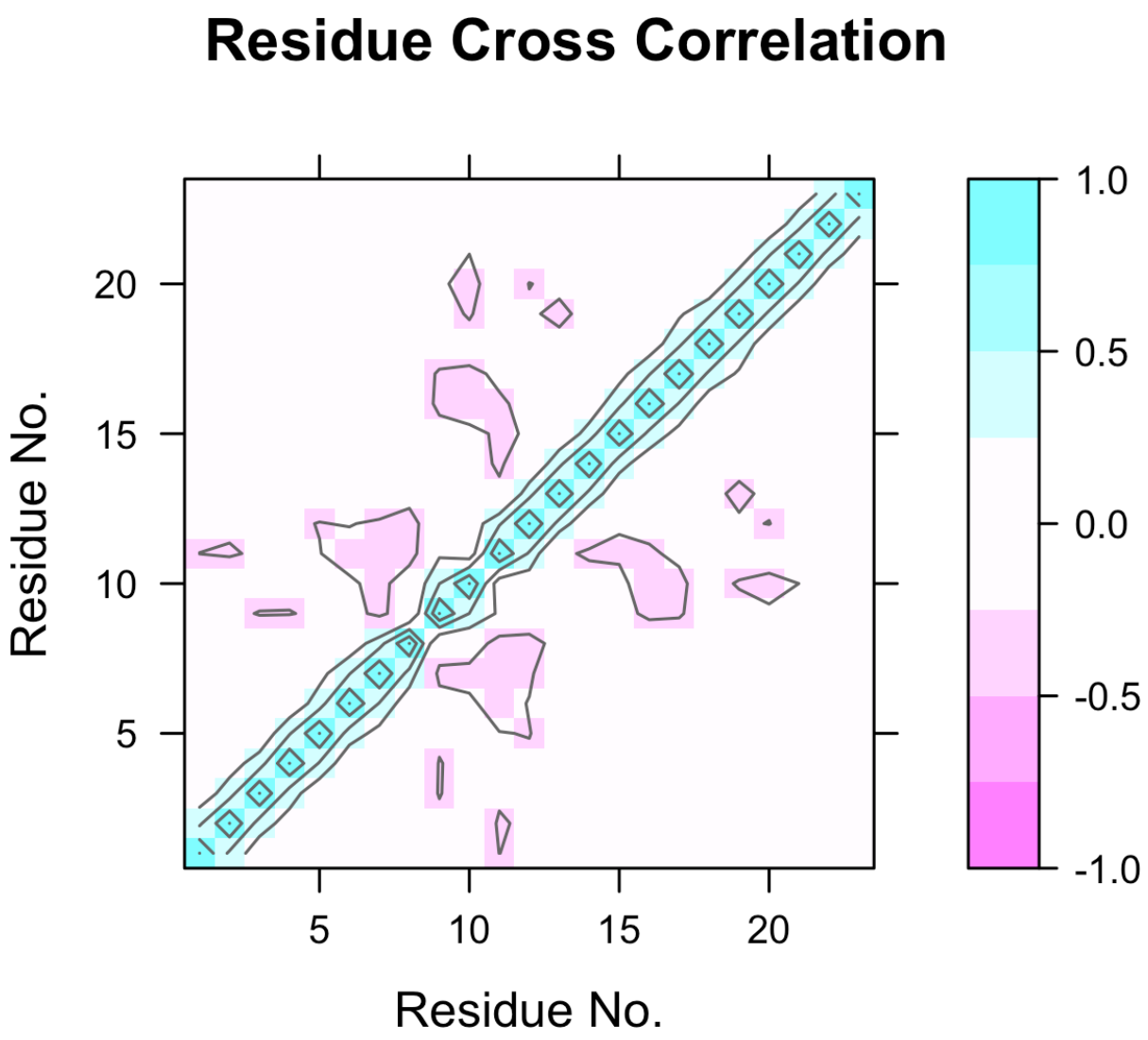

FIGURE 3. All-residue network of 1B7Y_B_408 with highlighted community structures.