Exploring Protein Folding and Unfolding with Graph Theory

Jenny Xu, Itai Brand-Thomas, and Nevon Song
Amy Wagaman and Sheila Jaswal, *Amherst College*

Departments of Chemistry and Mathematics & Statistics

Introduction

- Proteins exist in folded or unfolded states and transition back and forth in a state of equilibrium
- Diseases such as Alzheimer's and ALS are linked to protein misfolding and subsequent aggregation
- Understanding protein folding and unfolding mechanisms will provide insight into the causes of these diseases

Protein Background

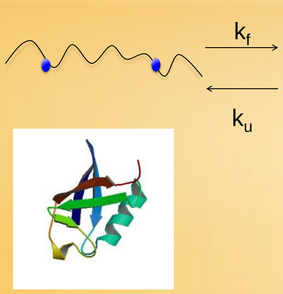
- Proteins can have different 3D secondary structures
- Human growth hormone has alpha helices (pink)

Unfolded **Folded**

Chain of amino acids folds into 3D functional form

- CAMPATH-1 antibody has beta sheets (blue/green)

Unfolded **Folded**

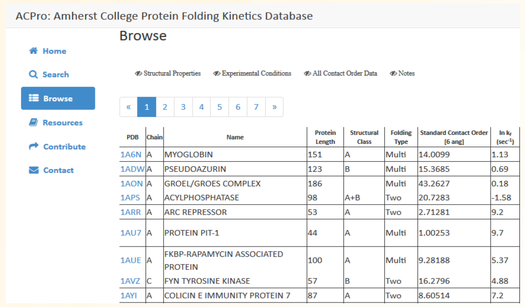


- Helices and sheets can also be combined (Ubiquitin, above left)

Methods

ACPro

- Amherst College Protein Folding Kinetics Database assembles un/folding and structural data for over 100 proteins
- ACPro is the most comprehensive and largest such database
- We helped populate ACPro, located at https://www.ats.amherst.edu/protein/

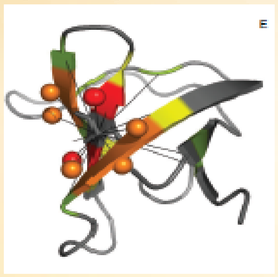


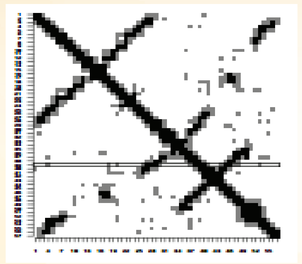
PDB	Chain	Name	Protein Length	Structural Class	Folding Type	Standard Contact Order (6 ang)	ln k_f (sec^{-1})
1A6N	A	MYOGLOBIN	151	A	Multi	14.0099	1.13
1ADW	A	PSEUDOAZURIN	123	B	Multi	15.3685	0.69
1AON	A	GROEL/GROES COMPLEX	186		Multi	43.2627	0.18
1APS	A	ACYLPHOSPHATASE	98	A+B	Two	20.7283	-1.58
1ARR	A	ARC REPRESSOR	53	A	Two	2.71281	9.2
1AU7	A	PROTEIN PIT-1	44	A	Multi	1.00253	9.7
1AUE	A	FKBP-RAPAMYCIN ASSOCIATED PROTEIN	100	A	Multi	9.28188	5.37
1AVZ	C	FYN TYROSINE KINASE	57	B	Two	16.2796	4.88
1AYI	A	COLICIN E IMMUNITY PROTEIN 7	87	A	Two	8.60514	7.2

Graph Theory

- Graph theory is useful to quantify characteristics of the folded protein structures
- For protein graph:
  - Each amino acid is a vertex
  - Amino acids in contact form an edge
  - Multiple methods for determining contacts exist (CA, CB, AA)
  - Various distances (6, 8, 10, 12 Angstroms) can define a contact
  - The protein graph can be viewed via its adjacency matrix (right)

Example: 1SHG (Alpha-Spectrin SH3 Domain)





1SHG visual and adjacency matrix at 6 Angstroms using both CA and AA methods

Variables of Interest

- Graph theory variables include:
  - *Average path length (APL)* –average of shortest path lengths between all pairs of amino acids
  - *Clustering coefficient (CC*) – measures extent to which connections form triangles
  - *Average degree (AD)* – average number of links each amino acid establishes with its neighbors
- Protein folding kinetics variables:
  - *Lnku* – natural log of unfolding rate constant
  - *Lnkf* – natural log of folding rate constant

Example: 1SHG vs. 1A6N (Myoglobin)

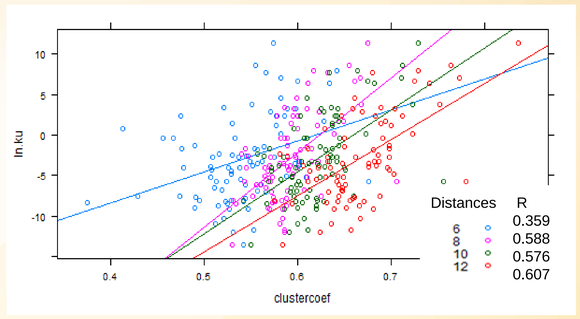
	APL	CC	AD	lnku	lnkf
1SHG	3.630	0.529	4.947	-4.83	1.05
1A6N	6.622	0.601	6.026	-3.77	1.13

Comparison of 1SHG (length 57) and 1A6N (length 151) using CA method at 6 Angstroms

Preliminary Findings

- Using R, we calculated proteins' various graph theory variables for all methods and distances
- We can subset our data to investigate relationships for specific graph constructions and distances

Example: CC vs. lnku



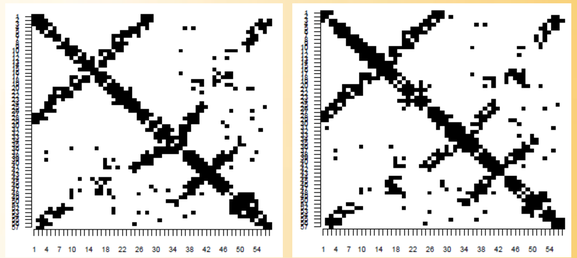
Next Step:

Randomized Assessment of Stability (RAS)

- We define a new method, RAS, to examine the stability of graph theory variables as the graph is perturbed
- Motivated by Jung et al. (2005)

Example: RAS using APL as variable of interest

- Our method involves:
  - Randomly dropping 1-20% of edges and examining the resulting graph



Two 1SHG adjacency matrices with dropped edges. Graphs constructed using AA method, distance of 6 Angstroms, and randomly dropping 20% of edges.

	Original	Trial 1	Trial 2
Edges	389	319	324
CC	0.578	0.438	0.465
APL	2.168	2.318	2.306

- Repeat 100 times per percentage dropped and compute mean APL
- Plot average change in APL against percentage dropped
- Fit OLS regression and extract slope
- The slope, if exists, is RAS of APL for that protein
- RAS can be plotted against lnkf/lnku to determine if correlation exists
- RAS will be performed with other graph variables of interest

References:
[1] Berman, H.M., et al. "The Protein Data Bank." Nucleic Acids Research, 28.01 (2000). Web. 22 July 2014. < http://rcsb.org/pdb/home/home.do>.
[2] Jung, J., Lee, J. and Moon, H-T. " Topological Determinants of Protein Unfolding Rates." PROTEINS: Structure, Function, and Bioinformatics, 58 (2005): 389-395.
[3] Wagaman, A.S. and Jaswal, S.S. "Capturing protein folding-relevant topology via absolute contact order variants." Journal of Theoretical and Computational Chemistry, 13.01 (2014).