Paulina Panek

1.

a. I think this is one subunit, but we would have to know how the two domains assemble, fold, and if they are stable by themselves.

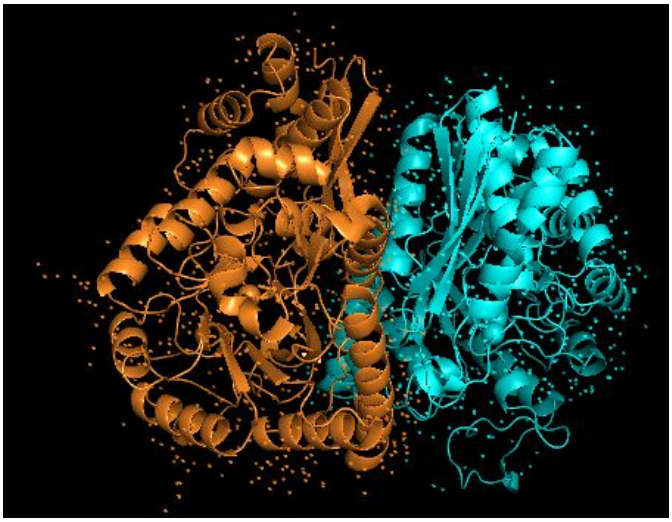


Figure 1: Two Domains of 1ONE

b. There are two domains

c. There is anti-parallel beta sheet, that I think is part of an open twisted alpha/beta structure.

d. The active site is probably at the carboxy edge of the beta sheet. The structure and present motif reminds me of Figure 4.14c from Branden & Tooze textbook that shows hexokinase, so I think this protein is also some sort of enzyme that might be catalyzing phosphorylation.

2.

a. Antiparallel β-sheet and an α-helix with Zn atoms
b. The motif is zinc finger; Involved residues are cysteins (red) and histidines (pink) (figure 1 left). Zn^{2+} ions stabilize the fold. Two Cys and two His residues coordinate zinc and form ββα fold.
c. The protein binds to the major groove of DNA. Many transcription factors have zinc fingers.The α-helix together with coordinated residues (figure 1 right - spheres) fit perfectly into the grooves.

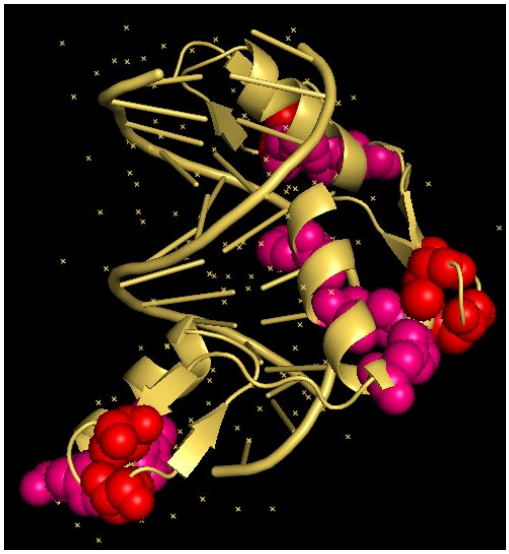


Figure 2: Zinc Finger
Left: Coordination of His and Cys showed with dots

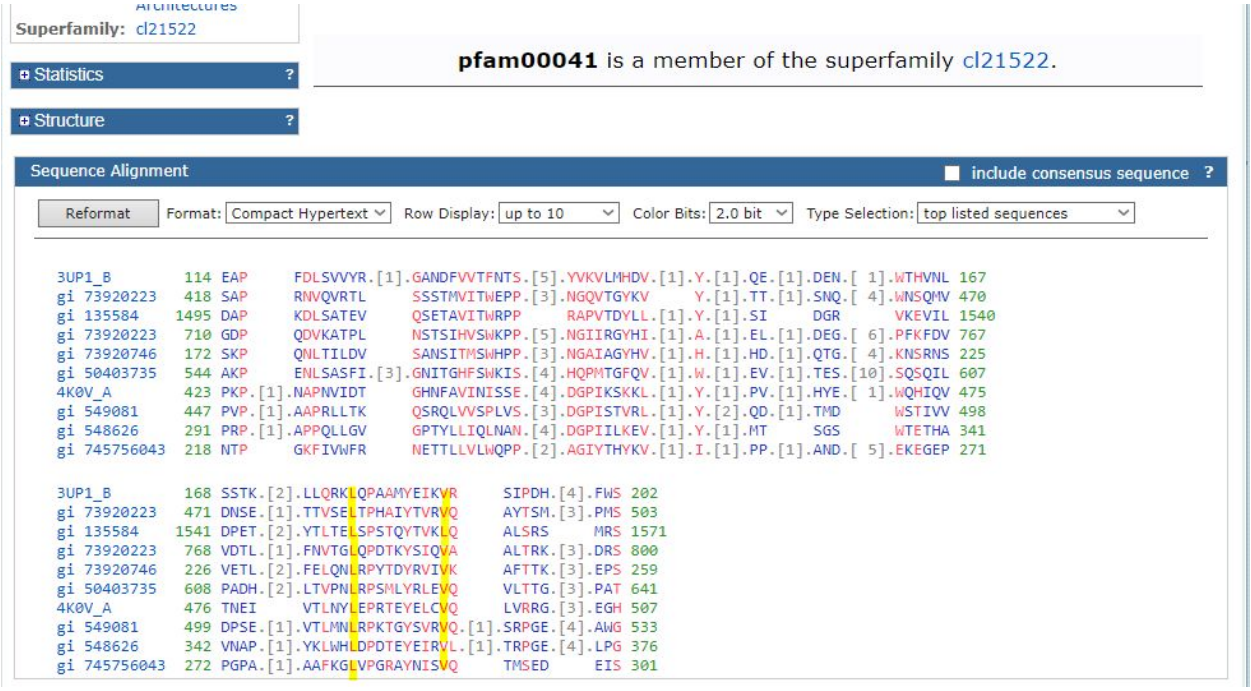
Right: Binding to the DNA groves

3.

a.

Figure 3: Fibronectin III
Green spheres represent conserved prolines
Red spheres represent non-conserved prolines

Yes, there seem to be regions where leucine and to some extent valine are highly conserved as well (highlighted yellow on screenshot below)



b. Non-conserved prolines are mostly on random coil.

To find aggregation propensity, I used AGGRESCAN[1] and found that the unmutated protein is more prone to aggregation than one where all prolines are changed to alanines (fig. 4).

[1] http://bioinf.uab.es/aggrescan/

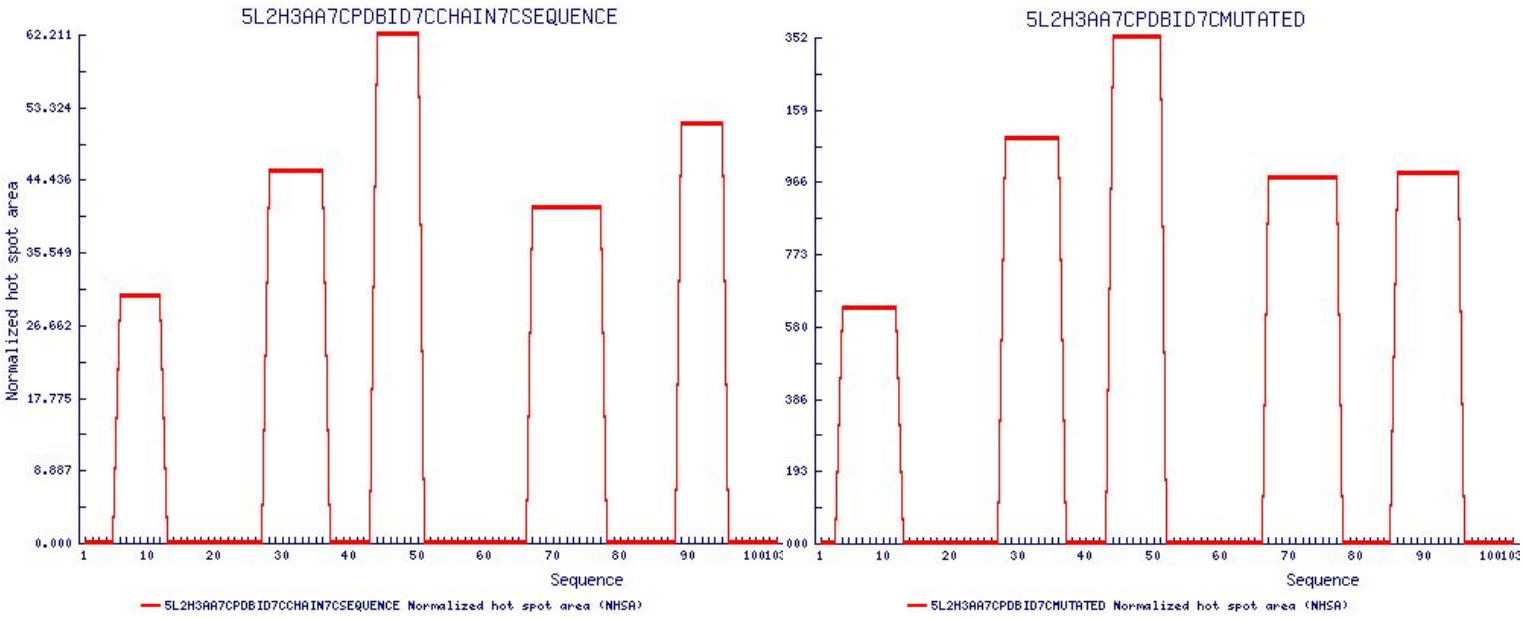


Figure 4: Normalized aggregation propensity
Left: Original sequence for 5L2H
Right: All prolines changed to alanines

The same regions still have tendency to form aggregates. However, after we mutate the structure each normalized area becomes smaller, implying that part of proline's role in the sequence is to form aggregates.

c. A cross-beta structure is found in amyloid that can form fibrils. The most common test for presence of this structure is placing a sample in X-ray diffraction beam and the diffraction lines should form a characteristic cross. Another simpler way is to use aromatic dyes such as congo red. This method is in clinical setting in tests for amyloidosis.

4.

a. The protein consists of four anti-parallel β-sheets connected via random coil to two α-helices on the other side. There is a copper cofactor attached to CYS 15.
b. According to PDB, this protein originates from Homo Sapiens, i.e. humans.
c. I am using NCBI BLAST because I have an option there to search for homologs in chosen organism, including the right strain. I changed the word size to 2 since our sequence is short and otherwise did not provide good results. The bacterial homolog is called heavy metal binding protein (NCBI Reference Sequence: YP_144984.1).

d. For sequence alignment, I used Clustal Omega, it found only 23% identity between the sequences. * under aligned residues symbolized conserved amino acids

```
Percent Identity  Matrix - created by Clustal2.1


1: YP_144984.1                  100.00   23.08
2: 1TL4_A|PDBID|CHAIN|SEQUENCE   23.08  100.00
```

Results for job clustalo-I20181031-212611-0658-17353634-p1m

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Hide Colors | View result with Jalview | Send to Simple Phylogeny | Se

```
CLUSTAL O(1.2.4) multiple sequence alignment

YP_144984.1                 -MLKLKVEGMTCNHCVMAVTKALKKVPGVEKVEVSLEKGEALVEGTADPKALVQAVEEEG   59
1TL4:A|PDBID|CHAIN|SEQUENCE MPKHEFSVDMTCGGCAEAVSRVLNKLGGVKY-DIDLPNKKVCIESEHSMDTLLATLKKTG   59
                               :    .***. *. **::.*:*: **:  ::.* : :. :*.  . .:*: :::: *

YP_144984.1                 YKAEVLA--      66
1TL4:A|PDBID|CHAIN|SEQUENCE KTVSYLGLE      68
                              ... *.
```

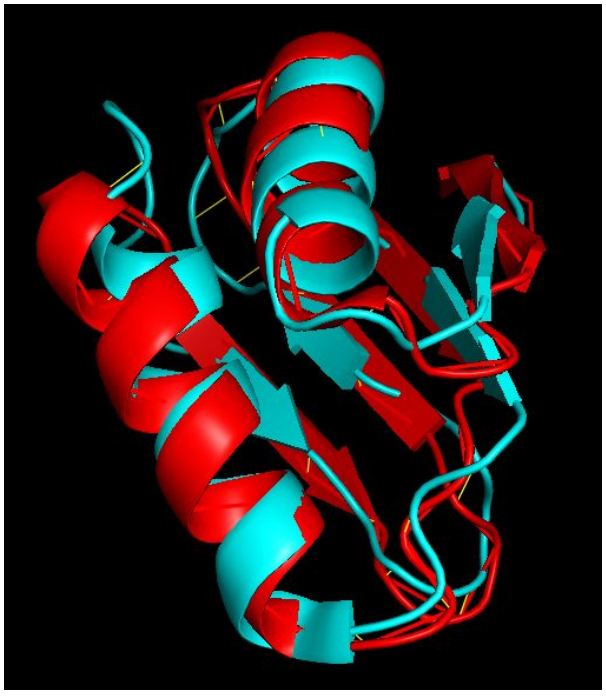


Figure 5: Overlaying 1TL4 and 2ROE
To compare structure, I use PyMol to overlay 1TL4 and 2ROE from pdb (blue is 1TL4, red is 2ROE)

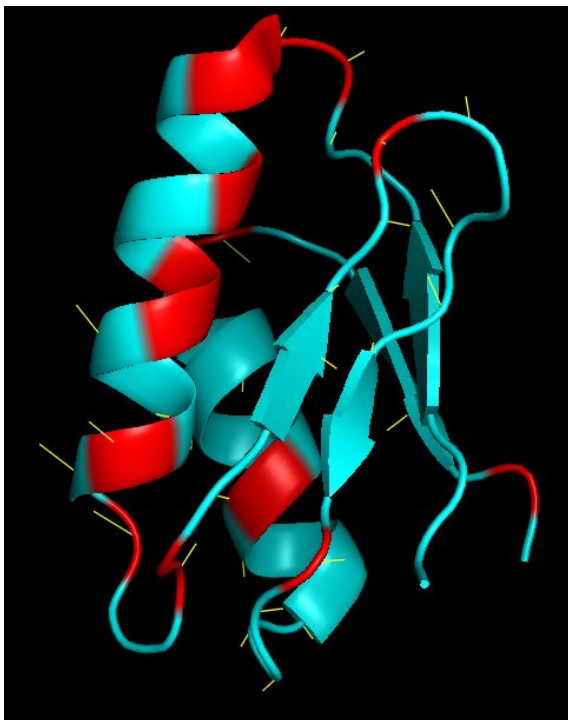


Figure 6: Conserved Residues in 1TL4
The overall structure of the two is very similar. The conserved residues are shown on 1TL4 below (red represents conserved residue).

The conserved residues seem to be located mostly on the helix facing the “inside” of the 3D structure. My guess is the active site is there and having that particular sequence there is crucial for proper binding of the substrate.

e. Frustration index is a measure of how much energy is contributed by a residue or residue pair relatively to their contribution in a molten globule.[2] High local frustration means the energy is not close to the minimal possible energy. High local frustration seems to be commonly found around binding sites, so it could be used to localize those[3].

[2] Nucleic Acids Res. 2016 Jul 8; 44(Web Server issue): W356–W360. Published online 2016 Apr 29. doi: [10.1093/nar/gkw304]

[3] Proc Natl Acad Sci U S A. 2014 Sep 30; 111(39): 14141–14146. Published online 2014 Sep 16. doi: [10.1073/pnas.1405233111]

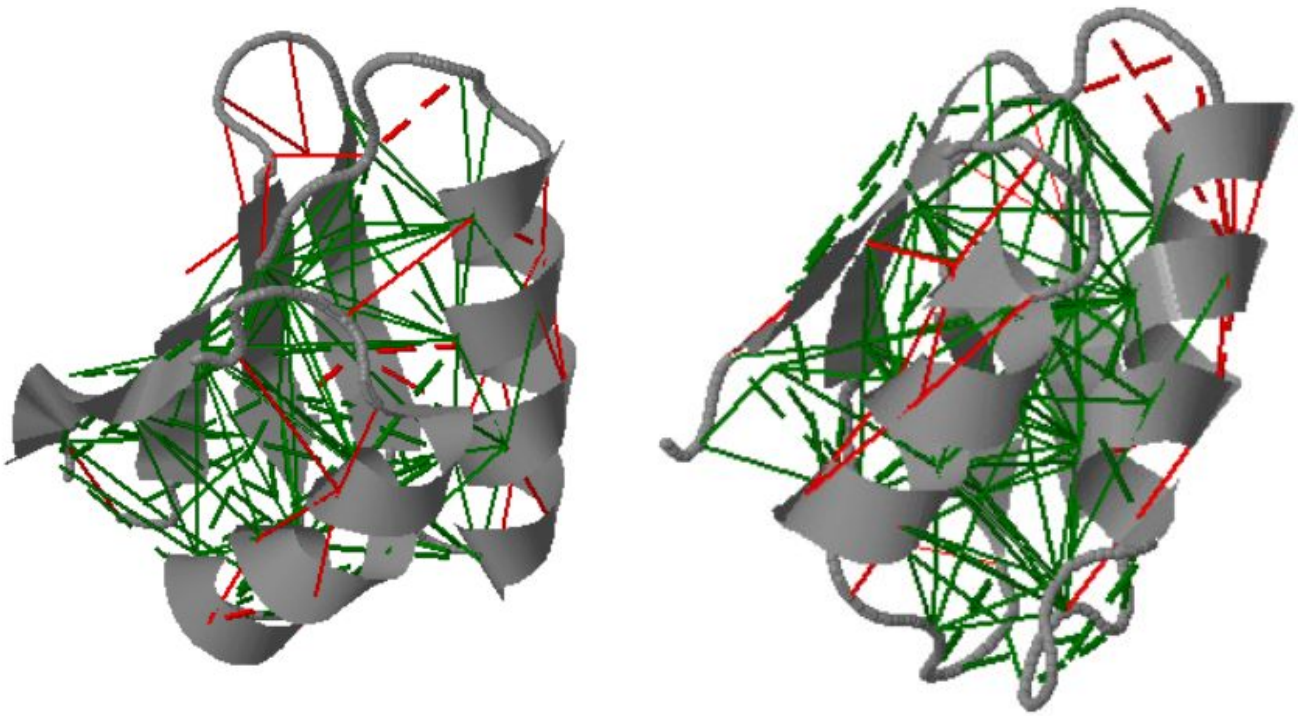


Figure 7: Local Frustration Map[4]
Left: 1TL4 Right: 2ROE
Red represents high local frustration

In both molecules it seems to be mostly localized on two α-helices which is consistent with my theory that those make the active site for binding.

[4] http://frustratometer.qb.fcen.uba.ar/

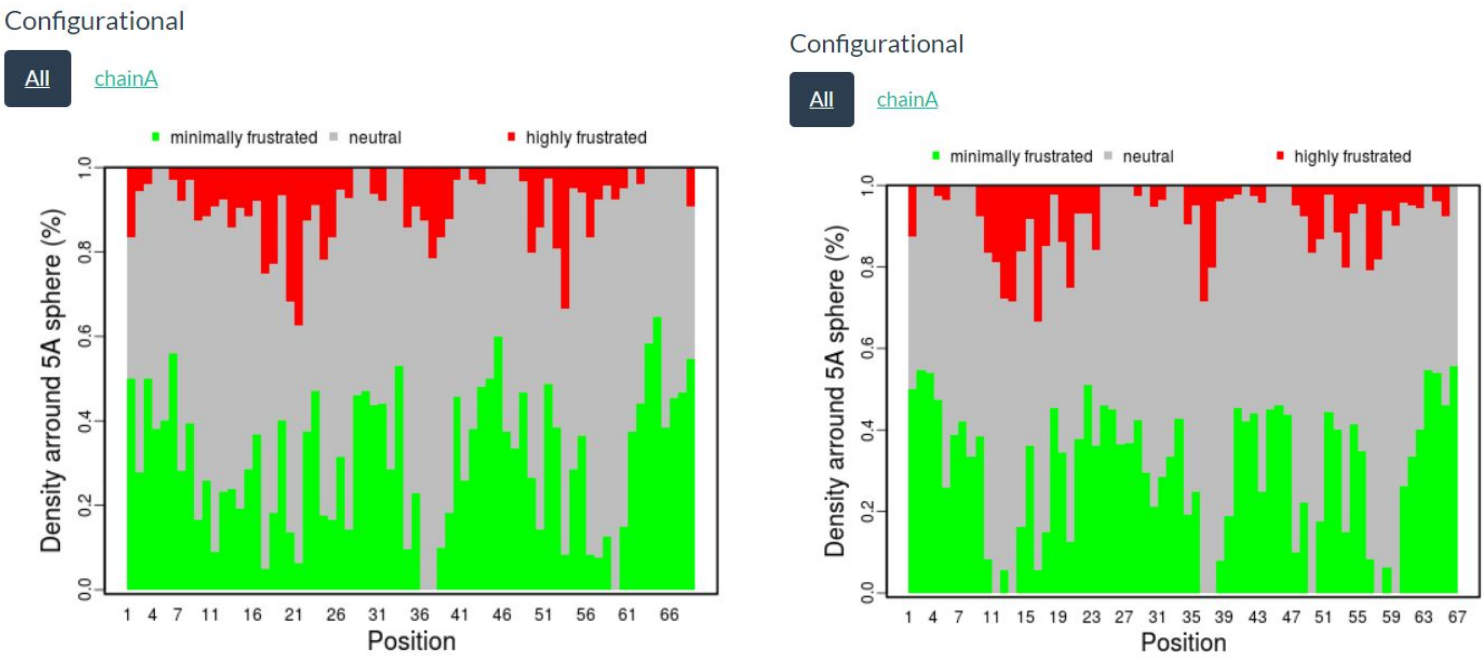


Figure 8: Fraction of contacts in each frustration class
Left: 1TL4 Right: 2ROE

Both proteins seem to have 3 clusters of highly frustrated contacts - around residues 16-21, 36, and 55 which is consistent with conserved residues.

g) 2ROE (bacterial) seems to have more acidic residues on the surface than 1TL4 (human).

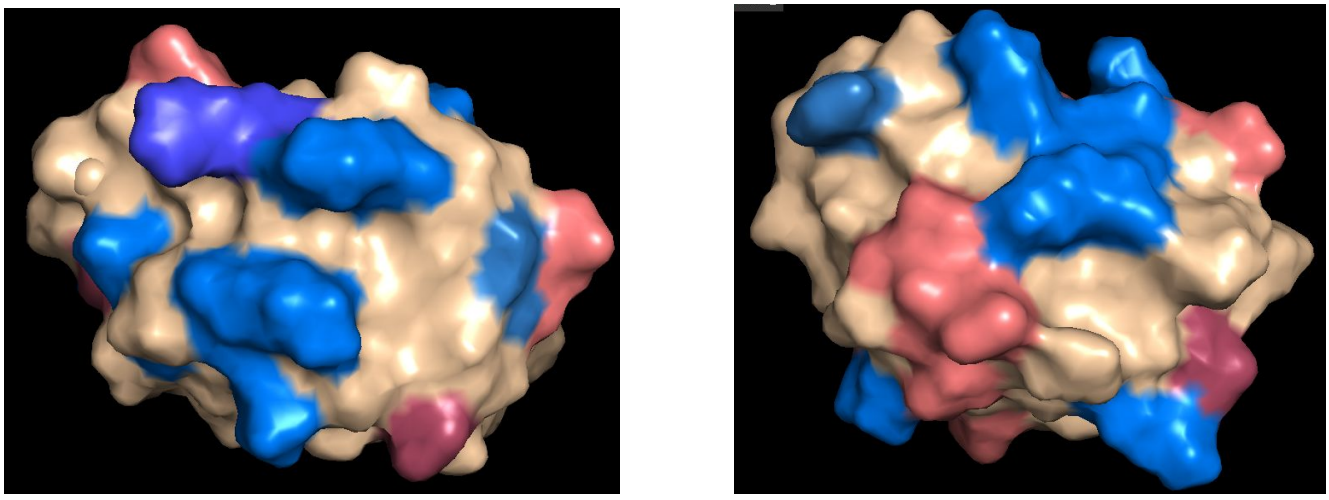


Figure 9 : Aligned 1TL4 (left, human) and 2ROE (right, bacteria) surfaces
Pictures show the same face of the molecule after alignment
Basic residues are in shades of blue (Arg, His, Lys)
Acidic residues in shades of pink (Glu, Asp)

h) At physiological pH, Glu and Asp tend to be negative while Arg and Lys tend to be positive. Histidine is mostly neutral and for the purpose of this exercise I will treat it as such.[5]

Protein	# Glu	# Asp	# Arg	# Lys	Net Charge
1TL4	5	4	1	8	-9 + 9 = **0**
2ROE	10	1	0	9	-11 +9 = **-2**

The net charge is not the same, because the pH inside the prokaryotic vs eukaryotic cell is different. This finding agrees with what was estimated in part g). The role of this protein is copper transport, so it has to be ensured that Cu rather than get oxidized in the presence of water.

5. Globular proteins are most often enzymes and in order to catalyze a reaction, they need their active site somewhat close to the surface and exposed to solvent, since substrates generally don't penetrate the protein - there is a lot of repulsive forces preventing that.

According to our textbook by Branden & Tooze, the following features indicate an active site:

- Carboxy edge of beta sheet in open sheet alpha/beta structures
- In alpha/beta barrel at the bottom of a funnei-shaped pocket made by the loops that connect the carboxy end of the beta strands with the amino end of the alpha-helices
- Crevices at the topological switchpoints
- Cleft region between two alpha/beta domains

According to our textbook by Fersht, there is a possible trade-off between stability and activity of enzymes. This is because active and binding sites tend to be sources of instability in molecules. Some very thermally stable enzymes of bacteria living in high temperatures show little activity in low temperatures. It should be also noted that stability beyond the turnover times of proteins is not always advantageous. Mutation studies of barnase showed possible mutations that increase stability without compromising activity. If such structures were very advantageous, they'd be selected for during evolution.

6.

a. GroEL and GroES form a large complex that includes 14 subunits coming from GroEL and 7 subunits from GroES. GroEL apical domain binds misfolded or incompletely folded proteins that tend to have exposed hydrophobic regions. Those regions bind to hydrophobic patch inside the GroEL cavity. Then ATP binds to the equatorial domain near the hinge point inducing a conformational change. GroES binds then to the apical domain acting like a lid, enclosing the

[5] http://www.russelllab.org/aas/charged.html

misfolded chain inside, in the hydrophilic chamber that favors proper folding and hiding the exposed hydrophobic regions inside the protein. This is called cis conformation. ATP is hydrolyzed and another ATP molecule binds to the trans ring, causing release of the GroES and of the now folded protein.

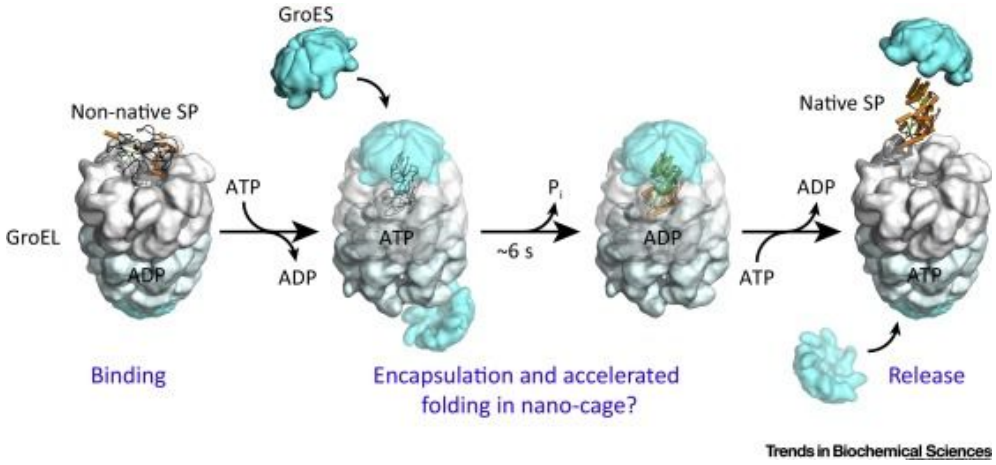


Figure 10: GroEL-GroES Mechanism[6]

b. When a cell is exposed to high temperatures, more of chaperons are expressed as a response. Their role is to take care of the misfolded proteins. Heat can denature proteins because it interferes with the hydrogen bonds and other interactions. Heat increases kinetic energy causing the molecules to vibrate and eventually break those bonds.
c. Nuclear proteins enter nucleus via pore and what is interesting is that they accumulate there despite the concentration gradient by a mechanism that is not fully understood.[7] Some small proteins can passively enter the nucleus via diffusion, but larger ones require ATP and factors in order to get through the membrane.[8] During a heat shock, the heat shock proteins start accumulating in the nucleus probably to ensure proper folding of histones and thus protection of the genetic material. I think the crowded nucleus affects proteins in the sense that they cannot passively enter and they need to be compact so they don't take more space than necessary. I imagine highly organized structures, rather than IDPs.

[6] Hayer-Hartl, Bracher, & Hartl. (2016). The GroEL–GroES Chaperonin Machine: A Nano-Cage for Protein Folding. Trends in Biochemical Sciences, 41(1), 62-76.
[7] J. Biosci., Vol. 21, Number 2, April 1996, pp 123-132.
[8] The Journal of Biological Chemistry 276, 20261-20266.

d. Some chaperons assist transport across membranes. For example, mitochondrial hsp70 is needed for transport across inner membrane of mitochondria.[9]
e. Hsp70 can assist folding to the native state, unfolding and refolding of misfolded proteins, and stops the proteins from aggregating.[10] Hsp70 has two domains - C-terminal where the substrate is bound and N-terminal ATPase domain. It is smaller than GroEL-GroS.

   Hsp70 is bound to ATP when no peptide is bound to it. Similarly to GroEL, Hsp70 recognizes the hydrophobic regions of misfolded peptides and binds to them which induces the ATPase activity. When ADP binds to a specific site of the Hsp70 protein, the binding pocket closes and again, similarly to GroEl-GroES, traps the peptide inside its cavity assisting folding.

   In contrast to GroEL-GroES complex, Hsp70 can also tightly bind to partially synthesized peptides to prevent them from aggregating before they are complete.
f. Hsp100 uses a different overall mechanism to chaperone folding. It is a hexameric protein, arranged in a ring structure with a pore in the middle. In the presence of ATP, the misfolded protein is threaded through that pore in the middle and gets unfolded. This allows the peptide to fold again. The active site is "on the inside" of the hexameric ring and is not exposed to the solvent much.[11]

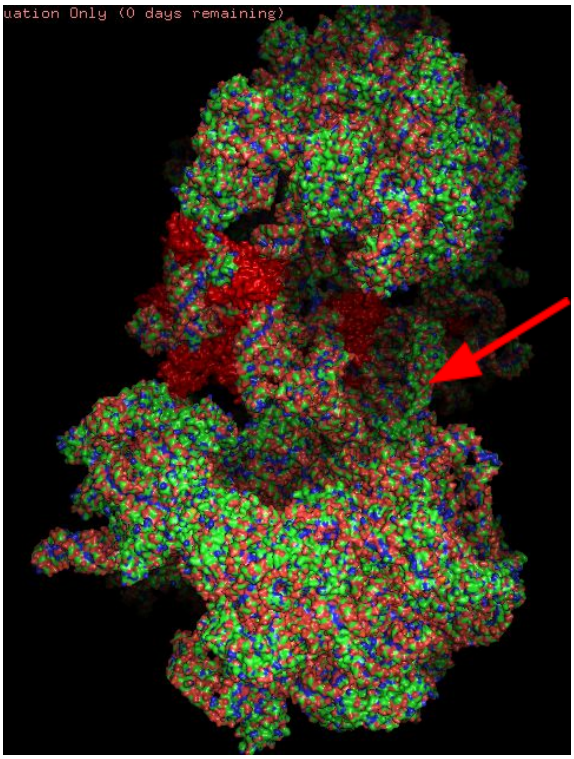
7.

a. Ribosomes are so big, they can be observed under a microscope and that's how they were discovered. They were initially thought to be organelles. Developments of modern techniques, however showed that ribosomes are actually just very big structures consisting of proteins and RNA.

   Suggested definition:
   Structures found inside of cells that consist of proteins and rRNA that serve as the sites of translation.

[9] Craig, Gambill, Voos, & Pfanner. (1995). The role of molecular chaperones in transport of proteins across membranes. Membrane Protein Transport, 2(C), 1-28.
[10] Cell Mol Life Sci. 2005 Mar; 62(6): 670–684. doi: [10.1007/s00018-004-4464-6]

[11] Nat Rev Mol Cell Biol. 2013 Oct; 14(10): 630–642.



b. To answer this question I will use PDB 5VPO, the 70S P-site ASL SufA6 complex, because I took a Bacterial Genetics course and I am very familiar with prokaryotic translation.

In the attached image the top part is the 30s subunit and the bottom is the larger 50s subunit. mRNA is threaded through the middle between those subunits (shown with arrow) where the active sites are (A, P, E).

Figure 11: 70S Ribosome P-site ASL SufA6 complex

c. The ribosome structure in bacteria and eukaryotes is very different. This means that we can find molecules that block translation in bacteria to be innocuous in humans since our translational machinery is different. An example can be chloramphenicol that blocks peptidyl transferase activity of bacterial ribosome and is an antibiotic used for cholera or typhoid fever.[12]

8.

a. GFP form a beta barrel with anti-parralel beta sheets on the outside and alpha-helices on the inside.
b. The barrels protect the chromophore (red) from being quenched by water, so it retains its fluorescent properties from aromatic bonds. The unusual thing about GFP is that the chromophore is synthesized from the amino acids in the chain rather than binded to it. The fluorescence is caused by a transition from the first singlet excited state to ground state. GFP shows both expected blue emission as well us green. The mechanism of green emission is not fully understood, but the figure below depicts a suggested mechanism.

---

[12] 23S rRNA cleavage by a mycobacterial MazF toxin
Jason M. Schifano, Regina Edifor, Jared D. Sharp, Ming Ouyang, Arvind Konkimalla, Robert N. Husson, Nancy A. Woychik
Proceedings of the National Academy of Sciences May 2013, 110 (21) 8501-8506; DOI: 10.1073/pnas.1222031110

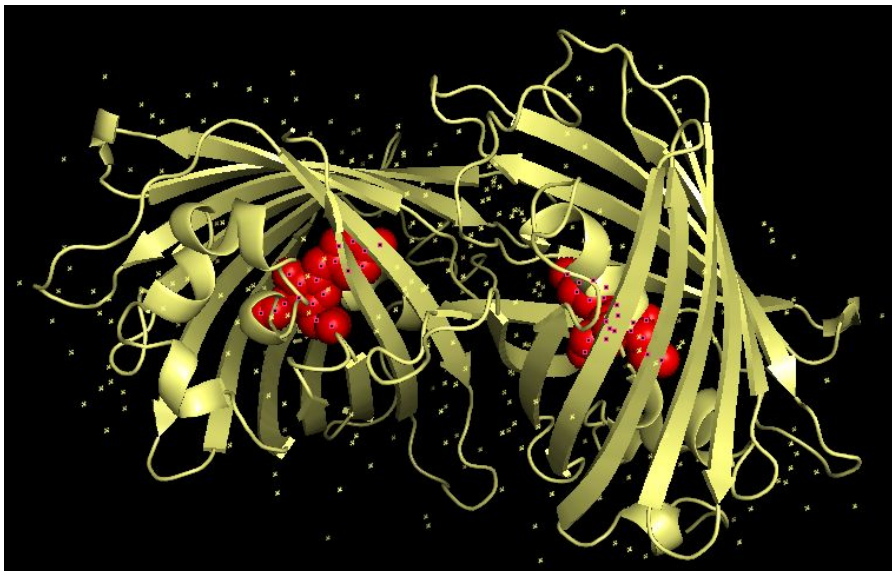


Figure 12: GFP Structure

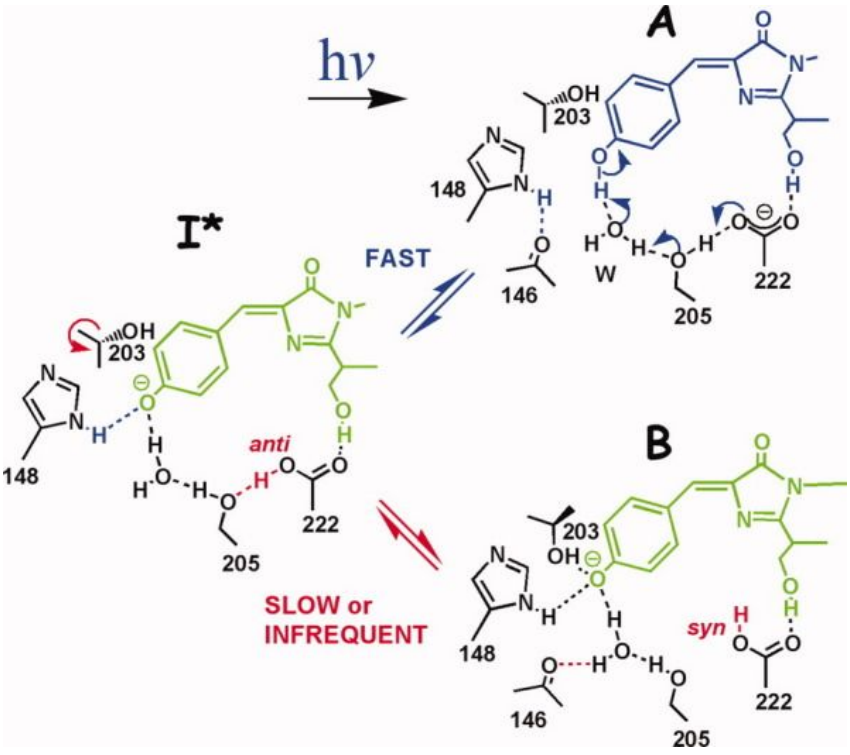


Figure 13: Proposed mechanism of Excited State Proton Transfer (ESPT)[13]

[13] Protein Sci. 2011 Sep; 20(9): 1509–1519. Published online 2011 Jun 28. doi: [10.1002/pro.684]

c. The advantage of using fluorescent proteins, such as GFP, for cancer imagining is that it is a non-invasive technique of observing the dynamics of cancer in vivo. Processes such as mitosis, differential behavior of nucleus and cytoplasm or apoptosis can be monitored with the fluorescent labels. Some successful research efforts included labeling proteins that are expressed only during certain cell cycle phases.[14] Hoffman et. al. conducted an interesting study where they managed to apply a GFP label only to tumor cells. What is interesting, the fluorescent properties of cancer cells remained even after fluorescence-guided surgery of removing the labeled tissue.[15] Introducing the label allows for monitoring reoccurance of cancer. A group from Japan investigated labeling human colorectal tumor cells to use along regular diagnostics.[16]

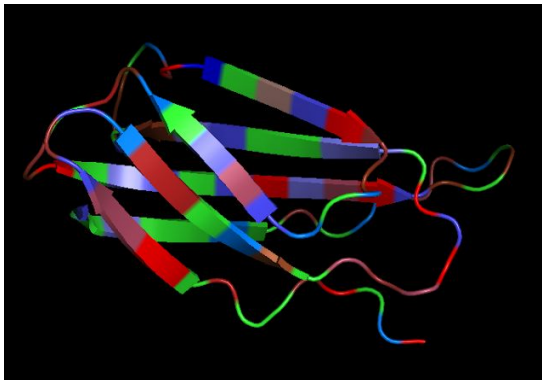
d. My protein (fn3) is important in embryo development, so we could use GFP to study where in embryo and in what quantities my protein is expressed. We could also monitor how that changes as the embryo develops. We would have to implement GFP tags and make sure they do not interfere with fn3 folding (there are different GFP variants that could work better than others). Once optimal labels introduced, we could study fn3 expression in embryo development in vivo or grow tagged cell cultures of interest.

e. Due to its structure GFP is more stable than other fluorescent proteins and can be relied on for long-term studies. It can be combined with other molecules without losing the fluorescent properties. It is very attractive for in vivo studies as it is non-toxic. Mutations in Tyr66 allows for obtaining different colors which opens up many research possibilities.[17]

---

[14] Lab Invest. 2015 Apr; 95(4): 432–452. Published online 2015 Feb 16. doi: [10.1038/labinvest.2014.154]
[15] Cell Cycle. 2011 Aug 15; 10(16): 2737–2741. Published online 2011 Aug 15. Doi: [10.4161/cc.10.16.16756]
[16] Gut. 2015 Apr;64(4):627-35. doi: 10.1136/gutjnl-2014-306957. Epub 2014 May 28.
[17] https://www.microscopemaster.com/green-fluorescent-protein.html (accessed 10/31/18)

9. .



a. My molecule consists of one subunit. I recognize Greek key motifs that make a β-sandwich domain.

The beta sheets and the coil both seem to contain a combination of hydrophobic and hydrophilic amino acids.

Figure 14: Protein 5
Hydrophobic amino acids in shades of red
Polar uncharged amino acids in shades of blue
Remaining in green

Amino acid composition:

Ala (A)	7	7.4%
Arg (R)	5	5.3%
Asn (N)	2	2.1%
Asp (D)	5	5.3%
Cys (C)	0	0.0%
Gln (Q)	1	1.1%
Glu (E)	3	3.2%
Gly (G)	8	8.5%
His (H)	0	0.0%
Ile (I)	6	6.4%
Leu (L)	4	4.3%
Lys (K)	3	3.2%
Met (M)	0	0.0%
Phe (F)	1	1.1%
Pro (P)	8	8.5%
Ser (S)	11	11.7%
Thr (T)	12	12.8%
Trp (W)	1	1.1%
Tyr (Y)	6	6.4%
Val (V)	11	11.7%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

I analyzed the FASTA sequence with ExPASy ProtParam tool[18] and found the following composition:

The most abundant residues being Thr, Ser, and Val.

b. I used APSSP: Advanced Protein Secondary Structure Prediction Server (see Q9b.txt for complete output file). This tool predicted my protein would consist of random coils and beta sheets, no alpha helices, which is consistent with what I see in PyMol.

[18] https://web.expasy.org/cgi-bin/protparam/protparam

c. My protein belongs to fn3 family which has 501 members.

Pfam Matches

Advanced

	Family		Clan	Description	Cross-references	Start	End	Domain E-values	
	Id	Accession						Ind.	Cond.
>	fn3	PF00041.21	CL0159	Fibronectin type III domain		3	84	1.0e-22	5.6e-27

Your search took: 0.02 secs

Family: *fn3* (PF00041)

6438 architectures | 138558 sequences | 33 interactions | 3076 species | 501 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...

Structures

For those sequences which have a structure in the Protein DataBank, we use the mapping between UniProt, PDB and Pfam coordinate systems from the PDBe group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **fn3** domain has been found. There are 501 instances of this domain found in the PDB. Note that there may be multiple copies of the domain in a single PDB structure, since many structures contain multiple copies of the same protein sequence.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A0A075B5G3_HUMAN	4 - 81	4OV6	F	3 - 79	Jmol OpenAstexViewer
			G	3 - 79	Jmol OpenAstexViewer
BOC_HUMAN	716 - 802	3N1G	C	716 - 802	Jmol OpenAstexViewer
			D	716 - 802	Jmol OpenAstexViewer
		3N1M	C	716 - 802	Jmol OpenAstexViewer
		3N1P	C	716 - 802	Jmol OpenAstexViewer
BOC_MOUSE	602 - 688	1X4Z	A	19 - 105	Jmol OpenAstexViewer
	711 - 797	1X4Y	A	12 - 98	Jmol OpenAstexViewer
		3D1M	D	826 - 916	Jmol OpenAstexViewer

d. According to Pfam my protein is considered evolutionarily conserved. To better understand what that means I run a quick BLAST search of my sequence, since it provides a very clear interface when it comes to matching alignments. See Q9d.pdf for printout of the results. In a summary, there has been many high scoring alignments, many at 100%, coming from different species, so it is definitely a well conserved domain. The fn3 has many functions such as adhesion, migration, differentiation and proliferation of cells in embryonic development.[19] Studies in mice showed that disruptions of the FN gene resulted in high embryonic deaths[20], which explains why this protein has been conserved through so many species. If mutation cause death at embryo stage, they cannot be passed to offsprings.

[19] J Mol Biol. 2007 Mar 23; 367(2): 303–309. Published online 2006 Oct 11. doi: [10.1016/j.jmb.2006.10.017]

[20] George, E., Georges-Labouesse, E., Patel-King, R., Rayburn, H., & Hynes, R. (1993). Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. Development (Cambridge, England), 119(4), 1079-91.

e. If pKa < pH, the functional group will lose protons to the solution leaving an amino acid with a negative charge. If pKa > pH, the functional group will gain a proton and become positive. I used the Protein Calculator tool[21] to find the following:

		Charge		
Residue	**# present**	**pH2**	**pH5.5**	**pH8**
Ala (A)	7	1	0	-0.5
Arg (R)	5	2	1	0.5
Asn (N)	2	1	0	-0.5
Asp (D)	5	1	-1	-1.5
Gln (Q)	1	1	0	-0.5
Glu (E)	3	1	-1	-1.5
Gly (G)	8	1	0	-0.5
Ile (I)	6	1	0	-0.5
Leu (L)	4	1	0	-0.5
Lys (K)	3	2	1	0.5
Phe (F)	1	1	0	-0.5
Pro (P)	8	1	0	-0.5
Ser (S)	11	1	0	-0.5
Thr (T)	12	1	0	-0.5
Trp (W)	1	1	0	-0.5
Tyr (Y)	6	1	0	-0.5
Val (V)	11	1	0	-0.5
Total protein charge:		9	0.5	-0.5

f. From ExPasy ProtParam tool again:

```
Instability index:

The instability index (II) is computed to be 24.69
This classifies the protein as stable.
```

[21] http://protcalc.sourceforge.net/cgi-bin/protcalc

Since the index is lower than 40, we expect the protein to be stable in a test tube. My protein does not contain cysteine. Cysteine usually forms disulfide bridges when present in a protein which stabilize the tertiary structure.

g. According to AGGRESCAN referenced earlier, this protein has several aggregation hotspots:

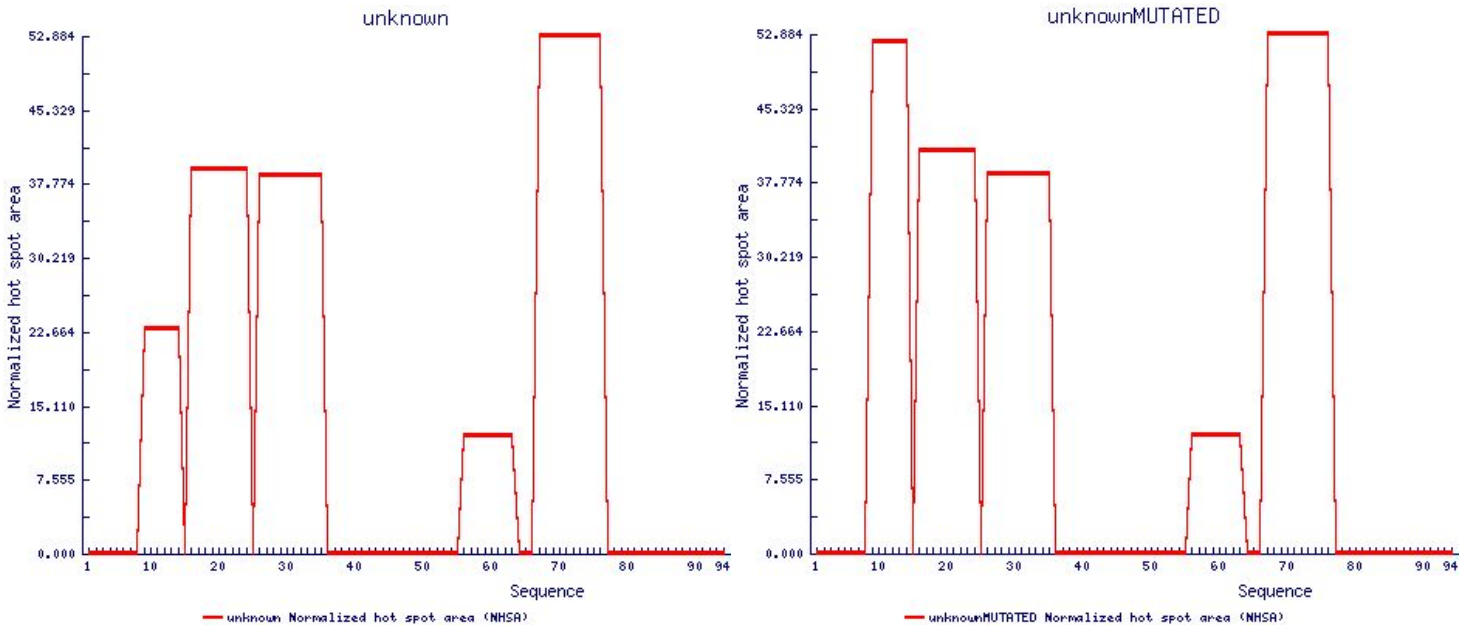


Figure 15: Normalized aggregation hotspots
Before and after mutation

Aggregation is caused by beta sheets, so I would introduce more tryptophan into the sequence, e.g. by changing A12 and A13 to WW:

VSDVPRDLEVV**AA**TPTSLLISWDAPAVTVRYYRITYGETGGNSPVQEFTVPGSKSTATISGLKP
GVDYTITVYAVTGRGDSPASSKPISINYRT

⇩

VSDVPRDLEVV**WW**TPTSLLISWDAPAVTVRYYRITYGETGGNSPVQEFTVPGSKSTATISGLKP
GVDYTITVYAVTGRGDSPASSKPISINYRT

h. My protein consists of random coil and beta sheets. To investigate folding transition state, it'd be more interesting to introduce mutations in the beta sheet rather than in the random coil that doesn't have defined structure. I'd be interested in introducing a mutation at Thr56 and Val50 as shown in .pse file (Q9h.pse). I'd change them to some more alpha-helical amino acids, e.g. alanines. I am curious how the two beta sheets come together and those residues connect the two sheets and are also at the very ends of their respective sheets. We could learn if the two beta sheets fold independently and then come together or if the process is connected. I think adding alpha helical amino acids in those locations would generally disrupt folding, but considering those residues are next to random coils rather than well defined structures, I think the protein would still fold and retain its shape.

10. I'd first check if the protein is soluble in water. If yes, I could do solution NMR, if not solid state NMR. NMR data would give me information about different populations present, the neighboring residue effects and conformations of the residues present. If the protein can be crystallized, X-ray crystallography studies could be conducted for further investigation and validation. Once we had the structural data, we could conduct a BLAST search for the sequence and see if the structure is conserved in the homologs. Highly conserved structure usually implies important role in the proper functioning of the protein.

    As far as in silico studies go, it'd be good to do them after we have a pdb file with a resolved structure. Some research groups try to simulate protein folding starting from completely unfolded state, but those are limited to only smaller proteins and generally might not be as reliable. If a protein is large and we wanted to do such study, we could try the replica exchange models.

11. The assumption in the simplest SBMs is that the structure obtained via NMR or X-ray crystallography is a global energy minimum.Perfectly funneled SBMs assume energy landscape as shown in a) while the reality looks more like b).

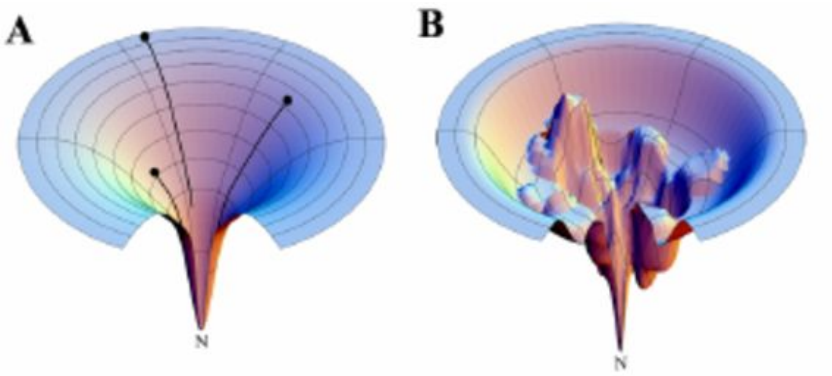


Figure 16: Perfectly funneled vs rugged energy landscape[22]

I am using shadow map because it is recommended for proteins. Cut off map is better for DNA/RNA or mixed systems.

.top topology file; defines structure based Hamiltonian

.gro coordinate file for gromacs, includes all information about atoms, periodic boundary condition and box size

.settings contains information about some basic parameters used for the simulation

[22] Protein Folding, Binding and Energy Landscape: A Synthesis - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Different-types-of-free-energy-landscapes-A-An-idealized-funnel-shaped-landscape-with_fig1_221925547 [accessed 2 Nov, 2018]