# February 15, 2013

Copied directory "automated\_sequence\_alignment\gonnet aligned" to directory "family moments". See "automated\_sequence\_alignment/automated sequence alignment log.docx" for the story of that directory's creation. In short each file is a multiple sequence alignment containing an HHOMP cluster and a protein of known structure.

## Amount of Data Available

I took a look at the length of these alignments. Here's an excerpt from the Python 2.7 session:

>>> alignments = [(i, Bio.AlignIO.read(i, 'clustal')) **for** i **in** glob.glob('\*.clu')]

>>> **for** name, alignment **in** sorted(alignments, key=lambda x: len(x[1]))[::-1]:

**print**('{}: {}'.format(name, len(alignment)))

3EFM with cluster18.clu: 532

**1E54 with cluster28.clu: 320**

1QFG with 22.1.4.clu: 225

2VQI with nn.2.2.clu: 207

3BS0 with cluster71.clu: 195

3JTY with nn.9.1.clu: 173

1T16 with 14.1.1.clu: 163

2QDZ with cluster53.clu: 153

2F1V with 8.2.1.clu: 142

1QD6 with 12.6.1.clu: 86

1FEP with cluster8.clu: 82

1QJP with cluster75.clu: 79

**2J1N with cluster99.clu: 78**

1P4T with cluster144.clu: 77

**2O4V with 16.4.2.clu: 76**

3CSL with 22.4.6.clu: 68

2GUF with 22.4.5.clu: 62

1UYN with 12.1.6.clu: 58

**1A0S with cluster73.clu: 52**

2ERV with 8.4.1.clu: 45

1TLY with cluster108.clu: 37

2WJR with nn.36.1.clu: 36

1QJ8 with 8.3.1.clu: 32

1KMO with 22.2.4.clu: 32

3EMN with nn.54.1.clu: 26

1I78 with 10.1.1.clu: 16

1THQ with 8.5.1.clu: 15

1K24 with 10.2.1.clu: 3

I bolded some, maybe all of the trimers, whatever stood out at me.

I think that for every transmembrane residue, there's about 3 other transmembrane residues that share x and y coordinates with it. Out of this group of four, on average two of them will face outward. But if you multiply those two by only 15 sequences, that's already a decent statistical sample size.f

So, consider each (x,y) position along the rim of the barrel to be a sample. You get some sense of the average energy of this sample, and then you compare them to get a sense of where the interface might be.

There are about 23 non-trimeric porins and 4 trimeric porins with 15 or more sequences. Unless there's a really obvious difference it's going to be hard to draw any conclusions.

## Notes on calculating Ezβ

The code for calculating Ezβ is in "modules/zenergy.py". This module defines the "calculator" class, which has a single method, "calculate", that is a function of a residue type and a depth. See the docstrings of the class and the method for more information.

This module requires a parameter file. One that I have is "family moments/published params.csv". I don't seem to have any logs describing how I created it. However, I have compared it to Daniel, Vik and I's 2012 Protein Science paper and it's the parameters from that.

## Learning to use DSSP

Previously I've used DSSP taken from spreadsheets that Daniel made. I'd like to get away from using Daniel's data since I'm not always sure how to read it, I don't know where it came from, it makes my methods harder for others to understand since can't talk directly about where the information comes from, and I'd rather learn how to generate it myself.

I downloaded a the PDB structure file of 1A0S as " family moments\trying out dssp/1A0S.pdb". I then downloaded the corresponding DSSP file from the DSSP database, and generated a DSSP file myself, using the following code (excerpt from an iPython log, Python 2.7):

**import** subprocess

x = subprocess.output(['dssp', ***'1A0S.pdb'***])

x = subprocess.check\_output(['dssp', ***'1A0S.pdb'***])

x = subprocess.check\_output(['dssp', ***'1A0S.pdb'***, ***'1a0s home calculated.dssp'***])

x

#[Out]# ***''***

urlretrieve(***'ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp/1a0s.dssp'***, filename="1a0s from dssp database.dssp")

urllib.urlretrieve(***'ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp/1a0s.dssp'***, filename="1a0s from dssp database.dssp")

#[Out]# (***'1a0s from dssp database.dssp'***, <mimetools.Message instance at 0x036DD878>)

As you can see, not all of that was purposeful: it took me a few tries to generate the DSSP file.

I compared the two files, and there are only minor differences, numbers off by 1 or 2. At least, I think that's minor, but I didn't look closely at what the numbers mean. Apparently the version in the database was generated with a newer version of DSSP than I use.