Chapter 4: Future Directions

Table of Contents

[4.1 Improving the protein design algorithm 3](#_Toc168064742)

[4.2 Heterodimer design 5](#_Toc168064743)

[4.3 Advancing forces research 6](#_Toc168064744)

[4.4 Detecting protein concentration in high-throughput 7](#_Toc168064745)

4.1 Improving the protein design algorithm

In chapter 3, I detailed how I developed my protein design algorithm. I utilized known literature and improved my algorithm over time with the knowledge at hand. However, most of my algorithm was finished prior to the development of AlphaFold, a machine learning based algorithm that predicts both known and unknown protein structures to near atomic precision. According to the data in my paper in Chapter 2, many of my structures appear to be associating according to the design interface. However, our energetic terms are unable to predict the ability for these proteins to associate. By studying AlphaFold and other machine learning based algorithms, we can improve the ability of the algorithm to better design structures and our energetics for predicting association.

One of the advantages of AlphaFold is the ability to predict structures without energetics. AlphaFold (describe AlphaFold in detail here; maybe take the image?). However, many groups aim to design proteins based on structures found in nature rather than sequence. I am unable to predict the structures of my design proteins due to the sequence dependency of AlphaFold. But the way that AlphaFold is able to connect sequence information to structure is relevant for improving design.

By extracting sequence information from known membrane protein structures, we can develop a prediction algorithm that takes into account the frequency of amino acids and their nearby partners. The current iteration of the design algorithm uses energetics to decide on amino acids at the interface. But utilizing natural sequence interfaces for design could improve the algorithm, as well as give insight into how to improve the energy terms. By tuning this new predictor term and designing sequences with different values from 0-100%, we may see that tuning closer to natural sequences at the interface results in increased association. In the current algorithm, we have developed a term similar to this idea: …explain membrane sequence similarity term here. Using this similarity score, our sequences were designed to resemble membrane protein sequences based on overall frequency. However, with this new term it is possible that we see that the amino acids near each other in weaker sequences are not as frequently found in MP structures as in stronger sequences. This gives credence to why our energetics might be better able to predict GASright sequences. The GASright sequence signature is maintained in our structures, and the frequency at the interface likely resembles sequences more often found in nature.

Concluding paragraph here on how we could develop this term and how to assess it against our sequences that we’ve currently designed and others; make this all sound scientific and be really easy to follow with the ideas I have in my head.

* Energetics doesn’t seem to work outside of GASright. Design with other energetic terms/reparameterize the van der Waals term using the information that we have for these sequences that associate primarily through packing.
* Random ideas:
  + Alphafold is really good at predicting structures without energetics using multiple sequence alignments and …(shape fitting basically?).
    - See how well Alphafold is able to predict structures in different dimer geometric regions
      * Use this as a way to evaluate the designability of the space, effectively replacing our density plot of frequency with how well it’s likely that an area is to be computationally predicted (basically breaking down areas into pockets that we have a really good understanding of and ones that we don’t)
  + Shape fitting
    - Energetics didn’t work very well, but many of our interfaces were suggested to be confirmed by our mutational data. Structurally, it seems like we’re able to design proteins at least semi-accurately
    - See if there are design papers that do shape fitting type stuff and see how they do it. We currently only have SASA implemented in MSL, but it doesn’t really work well.
    - If there is something, instead of using energetics during a monte carlo backbone optimization, use the fitting ability instead. Can then take that structure and try to calculate the energy post (maybe using something based on alphafold: Use energy as a way to say how designable the protein is, rather than how well it associates)
* Other energetics terms or reparameterizing our energetics for the other regions:
  + Is it possible that GASright is accurate with these energy terms because they are simply the most involved? What terms might we be missing? Does that Calpha-hbond somehow mitigate the error that we’re getting in our other designs? Could hbonding and packing actually be contributing less in other instances, as if the hbond is acting as a multiplier of sorts? Why don’t we break the hbonding component into regular and Calpha?

4.2 Heterodimer design

In order to …talk about the reason we chose to do homodimers. However, …explain that there are many heterodimers … I am currently working with another graduate student in the lab to develop our protein design algorithm for heterodimer sequences. Heterodimer design adds multiple variables of complexity for design: When designing homodimer sequences, we are able to simplify our algorithm by setting individual helices at symmetric geometric terms. However, with heterodimers, two of the geometric terms are not longer symmetric: z-shift and axial rotation. Each of these terms is a representative of the alignment between interfacial amino acids and their placement within the membrane respective to each helix. The z-shift defines the placement within the membrane while the axial rotation defines the part of the face of each helix at the interface. With this increased geometric search space, much of the algorithm that relies on symmetry is no longer applicable and must be developed to apply these terms to individual helices.

Since our energy terms work best with GASright structures, as an initial approach we are aiming to design heterodimeric GASrights. As geometries to start with, we have the geometries from GASright sequences found within the membrane protein PDB space. …test applying to the algorithm, changing (what parts specifically/functions need to be changed?)

* Currently working with another graduate student in the lab to develop the current algorithm for heterodimers.
  + Talk about the complexity of heterodimers (increased parameters, needing to calculate the monomer energy for each helix, moving each helix individually rather than simultaneously with homodimers)
  + An initial start:
    - Take the heterodimer space from geometries within the GASright region
      * However, we also have that data from the MPs. It’s just currently a bit more difficult to interpret (I’m not sure how axial rotation and z are decided actually; maybe dependent on where the starting point of the helix is?)
        + What we might be able to do is just use the differences between axial rotations of the two helices and z shifts
        + OR we could actually make it standardized, by first choosing the closest AAs at the interface to basically define the interfacial parallelogram for these structures

4.3 Advancing forces research

To better understand the driving forces involved in membrane protein association, we can rerun our design algorithm and test different variables. The goal of my research was to better understand the impact of van der Waals packing on association. Although we made marginal improvements to that knowledge, by designing additional proteins with more stringency in sequence will allow us to better determine the impact of packing as well as other potential forces.

One of the caveats of my conclusions to packing impacting membrane protein association is that there are multiple amino acids present in my sequences that could result in other forces driving association, including hydrogen bonding and electrostatics. To further the research on packing, we can rerun our design algorithm only allowing amino acids that would result in packing and no other forces. By rerunning design using… I have rerun this and already have a set of x proteins available and ready to test in high throughput.

Additionally, we can further assess the impact of packing on association by altering the backbone sequence. By assessing the association differences for the same interface between poly-Leucine, poly-Alanine, and poly-Valine backbones, we can gain insight into the impact of how minute changes in a protein results in differences in association. Although these backbone AAs are not often found at the interface, the slight loss in steric bulk (Leu->Val->Ala) may result in changes in association. Van der Waals is … Therefore, there may be a way to correlate the changes in association with change in AA size, allowing us to tune our van der Waals energy term by assessing these differences in bulk. We can also further fine tune our algorithm for other energy terms as well by designing with only AAs that result in hydrogen bonding and electrostatics. I have run the previous version of the algorithm using these parameters and designed x proteins to be tested in high throughput sort-seq.

* Machine learning on the terms to better predict, adding terms, etc.

4.4 Detecting protein concentration in high-throughput

One of the weaknesses of our TOXGREEN assay is the inability to accurately assess the expression levels of each our proteins in the membrane. Currently, much of the research is based on the assumption that our proteins express at the same level since much of the sequence is similar. However, this did not turn out to be true between sequences from different regions of design. Right-handed designs expressed noticeably less than both left and GASright designs when tested using western blots. Additionally, the western blot only assesses total protein concentration, not taking into account just the proteins that are inserted into the membrane and that result in fluorescence output. Therefore, it is possible that our Right-handed designs express and insert the same in the membrane as other designed sequences. However, we currently do not have the tools to assess this. To improve our assay and accuracy in assessing membrane protein association, developing a way to assess the concentration of our proteins in high-throughput would be beneficial, allowing us to normalize the fluorescence yield of each sequence to the protein concentration.

* I think there was a paper I read for chem bio seminar that was pretty relevant

So I just realized that there’s truly a problem with not knowing the protein concentration. In Samantha’s study, she normalized to %GpA by TOXCAT, done by lots of people previously. But in Gladys’ study, she saw that the proteins that she chose from Samantha’s study were all weaker than GpA in vitro, WHEN NORMALIZED BY CONCENTRATION. This means that GpA inherently expresses lower than those other constructs, which is a big part of the reason why our proteins look like they’re associating better. And this is likely the same with G83I. GpA has a deltaG of -6 while G83I is -2.5. Hilariously enough, most of my proteins are right around G83I, a known monomer. So this also means that since it’s expressing lower, we’re likely seeing better results in TOXGREEN and it only gets enhanced when doing it in high-throughput sort-seq. This is a big reason why sort-seq needs to be normalized for protein concentration in the future. It also might mean that a lot of my designed proteins that I think are associating aren’t actually associating. Well…actually I just rethought it a bit. My designs show pretty good association in TOXGREEN, suggesting that they’re fine. Sort-seq actually might be kind of cool though: it’s simultaneously supposed to be more accurate AND more sensitive. But I think we get a bit of noise along the way: converting counts to reconstructed fluorescence to TOXGREEN to TOXCAT to Kd to deltaG. So I actually think that although it’s not 100% good, it’s a fair way to assess association.

Address the above somehow