**Finding a template**

*Methods*

For each structure from the original Ez-Beta dataset, the sequence extracted from its aligned PDB file was entered as searches on the HHOMP website. From the list of clusters and scores returned, it was noted which cluster had the highest score, and, if the cluster was a composite of other clusters, what its subclusters were.

Pairs in which both structures had the same top cluster, or in which the top cluster of one was a subcluster of the top cluster of another, will be called “co-clustered” pairs in this article. A case in which two structures had different top clusters, but those clusters were composed out of the exact same subclusters, was also considered to be a co-clustered pair

For 4 of the 9 co-clustered pairs, their structures were aligned in PyMOL and compared visually.

*Results and Discussion*

The transmembrane regions of the pairs always aligned very well. In all for pairs, the number of strands was the same. The pair of 1AF6 and 1A0S each have two major irregularities in the transmembrane region: they have two short unusually short strands, and the angle of the strands is very different between the oligomerization interface and the opposite face of the barrel. However, 1A0S and 1AF6 share these quircks, and their transmembrane regions line up very exactly.

This indicates that a threading method that chooses templates using HHOMP could potentially predict the coordinates of the transmembrane region very well. Because these proteins were drawn from the E­­z-β dataset, they have no more than 26% pairwise sequence identity. Very similar sequences are not required for this method of finding templates.

The extracellular regions in the pairs did not match. As they mostly consist of floppy loops, it was not expected that their exact position would match. However, there were also differences in loop length and secondary structure. Two pairs that had beta sheets where the strands extend into the extracellular space had different length strands, according to PyMOL's secondary structure prediction.

In three pairs, the short periplasmic loops aligned very well. However, in the pair of 3EFM and 1QFG, 3EFM had two unusually long periplasmic loops that were not long in 1QFG.

These differences outside of the membrane are of limited impact, since Ez-β and Ez-β-based methods are based upon the chemical gradient of the membrane. Differences in z-coordinate that are simply indications of how far into water the residue is, where there is not a chemical gradient, are not very relevant. However, the differences in periplasmic loop size are probably chemically significant. Differences in strand length may be important to keep in mind for methods such as the Ez-β moment in which secondary structure plays a role in deciding which residues are included in the calculation.

**Threading to the template**

*Methods*

Swiss-PDB Viewer was used to make a structural alignment of 1A0S and 1AF6, and to generate a sequence alignment from this structural alignment. ClustalX was used to make all other sequence alignments.

The sequences of 1A0S and 1AF6 were retrieved from the Protein Databank. The BBTMout and BBTMall matrices from (ref), as well as the Gonnet series of matrices which is hardcoded into ClustalX, were used to generate three pairwise alignments of these two sequences.

The sequences in cluster73, the top cluster for 1A0S, were retrieved from the HHOMP database. Alignments of these sequences, the 1A0S sequence, and the 1AF6 sequence were produced using BBTMout, BBTMall, and the Gonnet series of matrices.

The residue parameter z­diff was used to evaluate the effectiveness of these alignment methods for threading. If the structure of 1AF6 is predicted from sequence using the structure of 1A0S as a template, zdiff is a measure of vertical distance between a residue in 1AF6, and the residue in 1A0S onto which it is mapped. zdiff is calculated with a digital implementation of the following algorithm **(put this in supplementary info?)**:

There are four decks of cards. Two of them are called "sequence decks". The left sequence deck has all the single-letter residue names from the aligned sequence of 1AF6, as well as cards with dashes representing gaps. Likewise for the right sequence deck and 1A0S. The residue names and gaps are in order such that the top card of each deck are at the first position in the alignment, the next to top card are at the second position in the alignment, etc.

There are also two "structure decks". Each card contains the z-coordinate of a residue in one of the structures, in order, with the top card representing the N terminal residue.

To begin the algorithm, cards are drawn simultaneously from all four decks.

* If neither of the sequence draws are gaps, the 1A0S structure card minus the 1AF6 structure card is recorded as a zdiff. All cards are discarded and replaced with new draws.
* If only the 1A0S sequence draw is a gap, this means that there is a residue in 1AF6 that is not mapped onto any residue in the template. Since the 1A0S sequence draw was a gap but there is a 1A0S z-coordinate on the table, the residue type corresponding to this z-coordinate has not yet been drawn; therefore, all cards are discarded and replaced with new draws except for the 1A0S structure card. No zdiff is recorded.
* If only the 1AF6 sequence draw is a gap, this means that there is a position in 1A0S, described by both a z-coordinate and a residue type, to which no residue in 1AF6 is mapped. All cards are discarded and replaced except for the 1AF6 structure card, since its corresponding residue type has not yet been drawn from the 1AF6 sequence deck. No z­diff is recorded.

This is repeated until all four decks are empty, which happens simultaneously.

*Results and Discussion*

An MSA contains more information than a pairwise alignment. There are more cues as to the correct positioning of residues. It may be expected that threading guided by an MSA with the whole cluster would result in a lower zdiff than guided by a pairwise alignment.

The BBTM matrices were derived using transmembrane beta barrel substitution rates, whereas the Gonnet matrices are general. Thus it may be expected that alignment with the BBTM matrices will perform better than the Gonnet matrices for aligning evolutionarily analogous, and thus presumably structurally analogous, parts of the proteins.

**(shouldn't this stuff be in the introduction? Seems a pretty widely spaced nasty way of formatting the paper really though)**

The Swiss-PDB viewer structural alignment has an average zdiff of 1.30 Å. This can be taken as a theoretical upper bound on quality of the thread.

The Gonnet series alignments were actually much better than the BBTM alignments. This is probably because the Gonnet series contains different matrices to be used at different levels of sequence identity. The BBTM matrices used were single matrices copied directly from Jie Liang's paper, designed for an evolutionary distance of 40 evolutionary time units. 1A0S and 1AF6 have a sequence identity of between # percent, so a BBTM matrix designed for a greater evolutionary distance will likely produce a better pairwise alignment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Swiss-PDB Viewer | Gonnet series | BBTMall | BBTMout |
| Pairwise |  | 3.84 |  | 7.80 |
| MSA |  | 2.48 | 8.05 | 7.60 |
| Structural | 1.30 |  |  |  |

Table X blah blah blah (don't forget unit!)

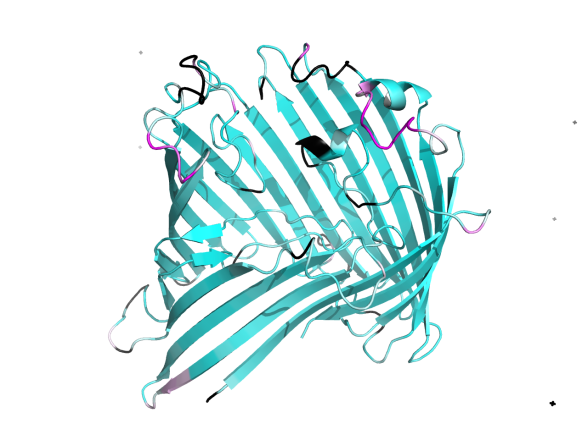
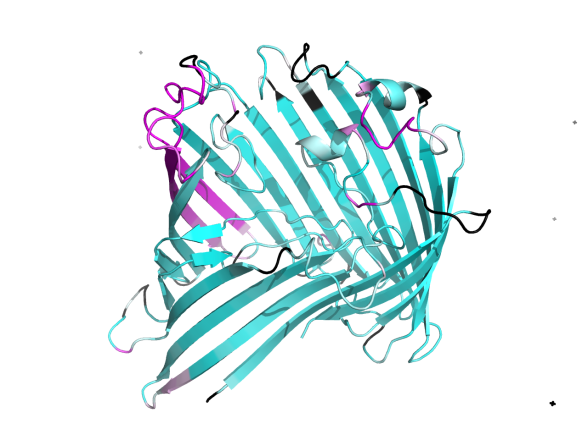
Within the Gonnet alignments, the MSA did, in fact, have a lower average zdiff than the pairwise alignment. A closer look reveals that this is not due to disorderly random factors scattered across the alignment. Examining a plot of residue-by-residue zdiff, it is apparent that the main source of error is in short misaligned segments. The small difference in average zdiff is due to the multiple alignment having significantly fewer such segments.

Figure X blah blah blah, more of these graphs in supplementary info perhaps

To see what structural features caused trouble for the alignments, the structure of 1A0S was colorized in PyMOL using the absolute value of zdiff (Figure X). Unsurprisingly, zdiff was high in the extracellular region. This was true even for the structural alignment. The structural alignment also had some residues with zdiff around 4 Å in the periplasmic space - these residues were some of the farthest from the membrane center on the periplasmic side. However, the low zdiff along the entire transmembrane region is an exact statement of what was seen earlier just by aligning structures: 1A0S is a very good template for 1AF76, and a good alignment has the potential to provide an excellent model for its structure.

The average zdiff in the Gonnet alignment turned out to be mostly from two strands, strand 8 and strand 9. Otherwise, it lived up to the promise implicit in the structural alignment.

In the BBTM alignments, nothing was aligned to strand 8, and strand 9 had high zdiff. However, extensive areas of both had zdiffhigher than 7 Å. Strands two and three had low zdiff, but this is not impressive since they are right at the beginning of the protein. It is hard to perceive structural effects on the alignment for the BBTM matrices. It looks more like the alignment is the worst near the middle of the sequence - whether this is because of something structural (the part of the barrel facing away from the trimer) or some effect from being in the middle of the sequence is beyond my knowledge.

**a) b) **

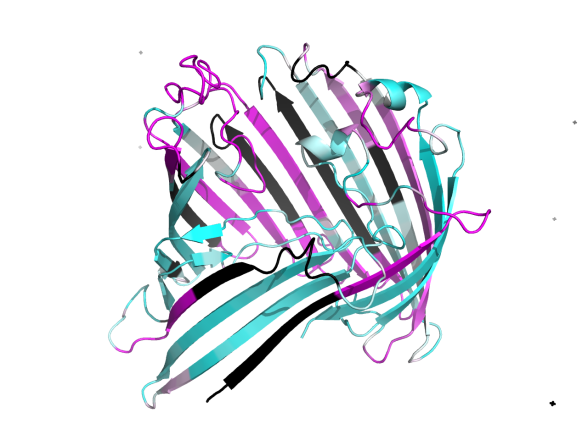
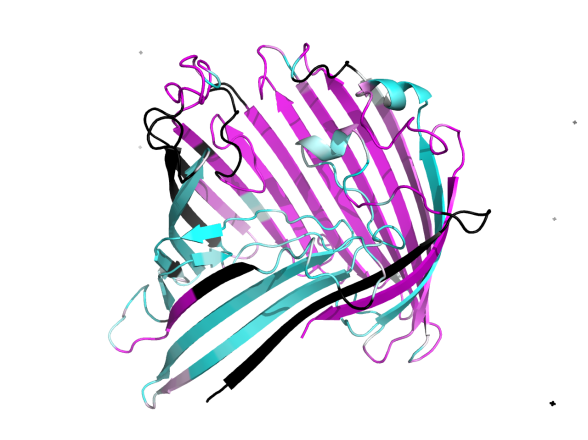
**c)** **d) **

Figure X. Blah blah explain colors. a) Structural b) Gonnet c) BBTMALL d) BBTMOUT

**Evaluating the representativeness of these examples**

*Methods*

A combination of 1A0S and the sequences of cluster73 was aligned in ClustalX using the Gonnet series of matrices. For each sequence, sequence identity with the 1A0S sequence was calculated in the following way: the number of positions in which either is not a gap, and are identical, was divided by the number of positions in which either is not a gap.

A combination of 1A0S, 1AF6, and the sequences of cluster73 was aligned, and the pairwise identity of 1AF6 and 1A0S calculated in the same way.

*Results and Discussion*

1A0S and 1AF6 had about 20% sequence identity. This degree of similarity with 1A0S is typical in cluster73 (Figure X).

Figure X. Blah blah blah.

Though most of the sequences in the cluster are less similar, there are fourteen sequences that are more similar. A method that used 20% sequence identity as a lower bound for inclusion, or an even stricter bound, would still have many sequences to work with. This seems especially true considering that cluster73 is below the median of size of the clusters associated with structures in the Ez-β dataset.

All of the above certainly seems to suggest that threading a sequence using a template co-clustered by HHOMP is a fairly good method of structure prediction. It makes me wonder why I have not seen this commented upon in the literature. Naveed et. al make it clear why they want to use a template-free method - it is probable that not all β-barrel shapes have been discovered. Jacoboni et. al say that " This is because of the fact that unless they belong to the same family, β-barrel membrane proteins share little sequence identity within each other even in the transmembrane spanning regions. It is well documented that in this case, methods based on homology building and threading cannot be successful". However, it is perhaps implied that for proteins in the *same* family, it can work quite well. Perhaps there has not been such a need for methods that provide rough approximations to the coordinates of a β-barrel membrane protein when the structure of another protein in the same family is known, or perhaps I have just not found any reference to it yet.

However, it would be far more convenient to use the sequences that are already in the cluster, than to find new sequences and see if they co-cluster with sequences of known structure. This brings up a tricky point that I do not, at the moment, feel prepared to answer: is the algorithm that was used to place sequences in clusters in the process of creating HHOMP similar to the algorithm used to calculate search scores? Is it similar enough that these results, based upon comparison of structures with high search scores for the same cluster, can be generalized to sequences within clusters?