Title of this paper

Alexander Davis, Vikas Nanda

**ABSTRACT**

**INTRODUCTION**

There are only about a hundred structures of transmembrane beta barrels (TMBs) solved at atomic resolution (ref opm). The accuracy of knowledge-based potentials for TMBs, such as TMSIP and Ezβ, is limited by the number of structures. More structures would also make it possible to compare Ezβ moments for a family of TMBs, and discern whether there are consistent trends.

One solution to this problem is to computationally predict the structures of TMBs from the vast library of recorded protein sequences. Different applications would require different degrees of speed and accuracy. Perceiving trends in the Ezβ moments of families of TMBs would require many structures per family, so speed of structure prediction would be a necessity. Constructing TMSIP requires knowledge of contacts between amino acids, so the structures would need a high degree of accuracy. Ezβ only requires knowledge of the distance of each residue from the center of the membrane, so even rough methods of structure prediction can still provide it with a larger dataset and increased accuracy. Since it only considers Cα's,

There are two well-developed methods of predicting the structure of a TMB. There is TMBpro, by Randall et. al, which finds a template with the correct number of strands, and then fits the sequence to the template by minimizing energy (ref Structural Bioinformatics of Membrane Proteins). Naveed et. al have developed a template-free method, which they justify with the recent discoveries of TMBs with unexpected numbers of strands, whose structures could not have been predicted by any template-based method (ref Naveed).

For the next edition of Ezβ, the authors plan to use a simple method of homology modeling: template-based, like TMBpro, but using evolutionary information to map the sequences onto the structure, rather than an energy minimization.

To find TMB sequences, as well as appropriate templates, the plan is to use HHOMP, a database of TMB sequences assembled by using hidden Markov models (HMMs) to find sequences homologous to TMBs with known structure (ref HHOMP). These sequences are divided into clusters using CLANS: points representing sequences were placed on a plane, where they attract each other based upon pairwise BLAST e-values (ref HHOMP). The clusters of points that result were used to categorize the sequences. The database has an HMM based search that assigns a given sequence a score for each cluster, representing the degree to which it matches that cluster. For Ezβ 2, sequences of known structure are to be matched with an HHOMP cluster of sequences, and used as a template to predict their structures.

One aim of this study was to test whether this structure prediction meets Ezβ's modest demand of reasonably good predictions of distance from the center of the membrane - or, to put it another way, reasonably close z-coordinates of each Cα, if the center of the membrane is considered the z=0 plane.

The other aim of this study was, given a good template, to quantitatively evaluate different methods of mapping a sequence onto it. Pairwise sequence alignments are compared to using multiple sequence alignments with a whole cluster, and alignments with matrices based upon TMB substitution rates are compared with more general matrices.

**FINDING A TEMPLATE**

The best way to evaluate the above described method of finding templates would be to use a predict a structure for sequence in the HHOMP database with a solved structure and low sequence identity with the template. Then, compare the predicted z-values with the actual ones.

However, because I only thought of that very recently, what was actually done in this study is more indirect. The pairs of structures that were compared were pairs that HHOMP matched with the same cluster, and which had low sequence identity with each other.

***Methods***

For each structure from the original Ezβ dataset (ref), the sequence extracted from its aligned PDB file was entered as a search on the HHOMP website. From the list of clusters and their corresponding match scores that this search returned, it was noted which cluster had the highest score; this was considered the "top cluster" for that structure.

Pairs in which both structures had the same top cluster were considered "co-clustered". Some clusters contain other clusters within them; if the top cluster of one sequence was contained within the top cluster of another sequence, then these sequences were also considered co-clustered.

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

No two structures in the dataset, and therefore no co-clustered pair, have more than 26% pairwise sequence identity.

***Results and Discussion***

The transmembrane regions of the pairs always aligned very well. In all four 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 unusually short strands, and the angle of their strands is very different between the oligomerization interface and the opposite side of the barrel. This is a level of correspondence beyond what would be expected of two unrelated TMBs with the same number of strands.

The extracellular regions of the pairs did not match. As they mostly consist of floppy loops, it was not expected that their exact position would match. However, beyond this, there werre also differences in loop length. When the extracellular region contained α helices, β sheets not contiguous with the transmembrane strands, these secondary structural elements were not necessarily shared between the pairs. Two pairs had transmembrane β strands that extended into the extracellular region, and in these cases, the length of the strands was different within the pairs.

In three pairs, the short periplasmic loops aligned very well. However, in the pair of 3EFM and 1QFG, 3EFM had two unusually long loops whose equivalents in 1QFG were short.

If the close alignment of the transmembrane region seen in these few examples is the rule for co-clustered proteins, then, given a sequence, finding a protein of known structure with which it co-clusters would be an excellent method of structure prediction for the purposes of Ezβ. The imporance divergence in the loop regions depends on to what extend there exists a chemical gradient in those regions. If they are chemically uniform then it does not matter for Ezβ. If they are not, then it is unfortunate that the gradient would not be measurable with the enhanced Ezβ, but much can still be accomplished with only the transmembrane region.

Future work in this area will be a similar comparison between proteins of known structure not in the HHOMP database, and proteins in the top cluster with known structure and low sequence identity. This will be a direct measure of the effectiveness of using a protein of known structure as a template for the proteins in its top cluster. For now, though, the above results indirectly suggest that when this is done, it will be found that the method does in fact produce close enough templates.

**MAPPING RESIDUES ONTO A TEMPLATE**

An MSA contains more information than a pairwise alignment. There are more cues as to the correct positioning of residues. On this basis I predicted that a mapping defined by an MSA with the whole cluster would result in lower zdiff than one defined by a pairwise alignment.

Though sequence alignments that are aided by secondary structure predictions are an intriguing possibility for this application, all that was tested was the method of aligning using substitution matrices, because there are many widely available, easily found resources useful for learning to align in this way, and for automating the process. Substitution matrices from three different sources were tested. One is the Gonnet series of matrices, which was made from substitution rates in all available sequences - that is, it is not specifically tailored to TMBs. The other is the work of Morales and Liang, who measured substitution rates specifically in TMBs. This study tested the BBTMall matrix, made using all substitutions, and the BBTMout matrices, made using only substitutions in lipid-facing residues.

The BBTM matrices were derived using substitution rates from the transmembrane regions of TMBs (ref), whereas the Gonnet matrices are general. On the reasonable hypothesis that evolutionarily analogous parts of TMBs are strucutrally analogous, whichever matrix best reflects the evolution of the TMBs should do better. The Gonnet series may be expected to do better than BBTM when aligning the loop regions, and the BBTM matrices better when aligning the strands. Tests performed by Jimenez-Morales and Liang indicate that the BBTM matrices should do better. They tested how many homologs could be picked out of a dataset of TMB sequences using the BBTM matrices, compared to the BLOSUM62 and PAM250 matrices, which, like the Gonnet matrices, are not membrane-protein specific. More homologs were detected using the BBTM matrices. On this basis, I predicted that an alignment with the BBTM either of the matrices would have lower zdiff than an alignment with the Gonnet series.

***Methods***

The co-clustered pair of 1AF6 and 1A0S were used as a model to investigate the effectiveness of different methods of mapping a sequence onto a template, with 1A0S acting as the template.

Swiss-PDB Viewer was used to make a structural alignment of 1A0S and 1AF6, and to generate a sequence alignment that reflects this structural alignment. All other sequence alignments were made with ClustalX.

The sequences of 1A0S and 1AF6 were retrieved from the Protein Databank, and it was verified that they were the same as the sequences of the aligned structures used in the derivation of Ezβ. The BBTMall matrix from (ref) and the Gonnet series of matrices hardcoded into ClustalX were used to create pairwise sequence alignments.

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

The residue parameter zdiff was used to evaluate the effectiveness of these alignments for homology modeling. For each residue in 1AF6, zdiff is the difference in z coordinate between that residue's Cα and the Cα of the residue in 1A0S with which it is aligned. If 1A0S was used as a template, and the sequence of 1AF6 mapped onto it with the correspondence defined by that alignment, then zdiff  would be the error in the height of each residue. The algorithm used to calcualte zdiff is described in the supplementary information accompanying this paper.

***Results and Discussion***

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

The Gonnet series alignments actually had much lower zdiff than the BBTM alignments (Table 1). This is probably because the Gonnet series contains different matrices for different evolutionary distances - ClustalX chooses which one to use by using sequence identity as a rough measure of evolutionary distance. In this study, only one BBTMall matrix and one BBTMout matrix. These matrices were designed for an evolutionary distance of 40 evolutionary time units (ref). 1A0S and 1AF6 have a sequence identity between 20% and 30%, so these were not the appropriate matrices to use for the pairwise alignment, and a series of BBTM matrices for different evolutionary distances would have resulted in a better multiple sequence alignment. I predict that if a series of BBTM matrices were used, zdiff would be lower than that achieved with the Gonnet series.

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

Table 1. Average zdiff in angstroms for all alignments tested.

Within the Gonnet alignments, the MSA did, in fact, have a lower average zdiff than the pairwise alignment. Examining plots of residue-by-residue zdiff, reveals that the main source or error is in short misaligned segments, not small, disorlery errors scattered throughout the sequence (Figure 1). The difference in average zdiff, though small, is due to the multiple sequence alignment having significantly fewer misaligned segments.

**a)**

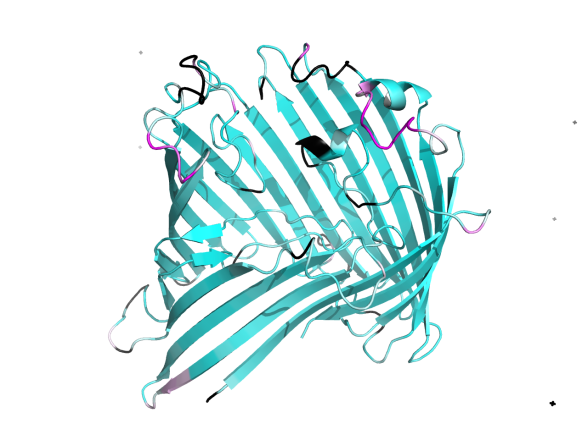
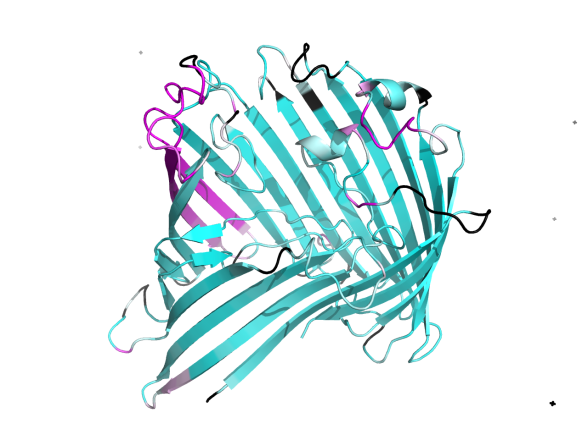
**b)**

Figure 2. Residue-by-residue zdiff for a) pairwise alignment using Gonnet series of matrices b) multiple sequence alignment with cluster73 using Gonnet matrices

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 2). 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 shown previously by aligning structures: 1A0S is a very good template for 1AF6, and combined with a good alignment, can give excellent estimates of 1AF6 z-coordinates in the transmembrane region.

The average zdiff in the Gonnet alignment turned out to be mostly from two strands, strand 8 and strand 9. Everywhere else, it lived up to the promise in the structural alignment. It is hard to tell from this single example why the alignment went wrong there of all places.

In the BBTM alignments, nothing was aligned to strand 8, and, just as in the Gonnet alignments, strand 9 had high zdiff. However, extensive areas of both ad zdiff higher than 7 Å. Because the misalignment is so extensive, it is hard to perceive structural effects on the alignment of the BBTM matrices. It looks like the alignment is the worst near the middle of the sequence. Whether this is because of something structual, such as the misaligned region neing the region that faces away from the trimer, or some effect from being at the middle - rather than the end - of the sequence is beyond my knowledge.

**a) b) **

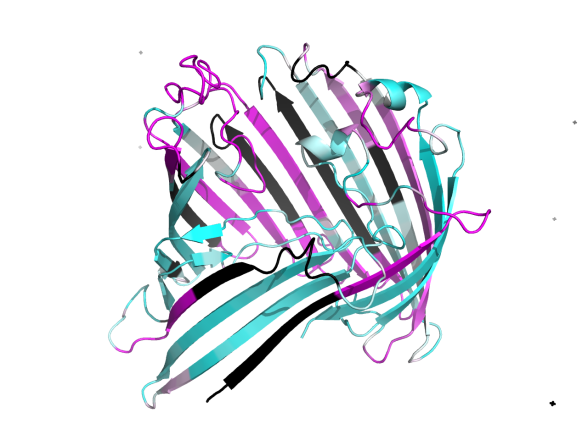
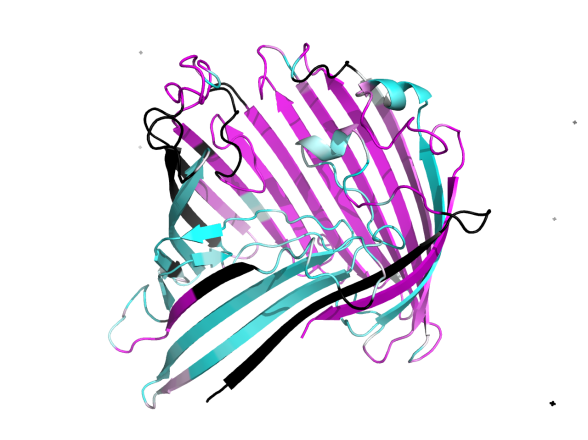
**c)** **d) **

Figure X. Structure of 1A0S colored by absolute value of zdiff. |zdiff| of zero is cyan, and |zdiff|above or equal to 7 is magenta. Between 0 and 7, the colors are a gradient from cyan to white to magenta. a) Swiss-PDB viewer structural alignment b) Gonnet MSA c) BBTMall MSA d) BBTMout MSA

**SUPPLEMENTARY INFORMATION**

**Algorithm for calculating zdiff for an alignment of the sequences of 1A0S and 1AF6**

The inputs to the algorithm are the structures of a monomer of 1A0S, a monomer of 1AF6, and an alignment of their sequences. The output is a list of zdiff's. It is a digital implementation of the following procedure on decks of cards:

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

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.