Discovering the determinants of outer membrane protein insertion is a task of basic scientific interest and technological value. It is an investigation with potential for insight both into the molecular mechanisms of bacterial function, and into the chemistry of biological membranes. The technological value comes from the prospect of designing mutants of soluble proteins that will insert into membranes. Outer membrane proteins have a β-barrel fold that is similar to that of many soluble proteins, most importantly GFP. If the determinants of outer membrane protein insertion were fully understood, it may be possible to design a mutant GFP along the same principles that would insert into membranes, which would have important applications as a voltage sensor in neuroscience experiments (Blunck et al., 2005).

Outer membrane proteins (OMPs) are insoluble in water, and if unfolded can spontaneously refold and insert into vesicles (Surrey and Jähnig, 1992). Together this suggests that at equilibrium, OMPs are in membranes, rather than being forced in and held there by some kinetic barrier. It is likely, then, that negative of folding is a necessary condition of insertion.

Under the hypothesis that for a whole OMP is a sum of contributions from solvent-exposed residues on its surface, the problem of estimating for an arbitrary OMP becomes the problem of finding free energies of transfer for individual amino acids. Once this is achieved, the calculation of is as simple as summing the transfer energies. There have been three broad categories of approaches to this problem.

One is experimental. 's for each amino acid have been derived through a mutation study (Moon and Fleming, 2011). Another approach is through simulation. Molecular dynamics simulations have been used to derive 's which are very close to those estimated from experiment (Gumbart and Roux, 2012).

This work builds upon the knowledge-based Ezβ potential (Hsieh et al., 2012). Unlike the above referenced experiment- and simulation- based values, the Ezβ potential is *depth-dependent*: it estimates the energy of transfer to a given depth in the membrane, not just to the center. Depth is represented by a number, z, that represents the distance from the center: at z=0, Ezβ should, and does, correlate with the experimental values. This depth dependence is likely to increase accuracy because the membrane has been experimentally shown to have varying hydrophobicity; and because each residue has a distinct depth-dependent frequency profile, sometimes with a smooth transition or even a peak partway through, that seem to reflect a dependence of insertion energy upon depth that is more complicated than a simple distinction between "in" and "out" (Hsieh et al., 2012).

Membrane protein structures are scarce, though. Using only the currently solved structures, there is not enough data to capture the asymmetry of the two leaflets of the outer membrane in the potential (Hsieh et al., 2012). Also, methionine is too rare to characterize the depth-dependence of its frequency, but it is common enough that it might be characterizable with a larger dataset.

There are plenty of available sequences, though, from which additional structures can be guessed. There are a variety of methods that can make decent guesses of the structure of the transmembrane strands (Forrest et al., 2006; Randall et al., 2008; Hayat and Elofsson, 2012; Naveed et al., 2012). This study concerns the evaluation of a specific protocol for one of those methods: homology modeling.

This study will explore a method of homology modeling based purely on evolutionary information. There are two distinct advantages to this. One is that it is very easy and can bed done with off-the-shelf general purpose software, with a little scripting to tie it together. This also makes it easy to understand. The other advantage is that it embodies no prior assumptions about energies. Some other methods involve a knowledge-based potential at some point during the process. In the homology modeling procedure, biases may emerge due to the way the sequences are handled, but at least we will not end up with energies of transfer that reflect other energy functions that we fed into the structure prediction process.

There are two fundamental steps in homology modeling. The first is finding a template for each sequence, and the second is mapping the structure onto each template. In this study I test investigate the effectiveness of using the HHOMP database's search function to match sequences with templates (Remmert et al., 2009), and the sequence alignment program ClustalW (Larkin et al., 2007).

## Experiment 1: Determination of the evolutionary distance between sequences in an HHOMP cluster

### Introduction

The HHOMP database is a database of sequences predicted to be OMPs (Remmert et al., 2009). The sequences have been divided into clusters by the program CLANS, to produce clusters of proteins related proteins (Frickey and Lupas, 2004). Many clusters contain a sequence of known structure.

A strategy that immediately springs to mind is to solve the template-finding step is to use each protein of known structure as a template for all sequences with which it shares a cluster. In previous work, I found that proteins of known structure in nearby clusters are highly structurally similar suggesting that structural divergence in OMPs is slow enough for homology modeling to be viable. However, if the sequence identities with the template are too low, it may nevertheless be impossible to map the structures onto the template properly.

To address this concern, I picked a cluster, considered a known protein within it to be a template for the others, and checked the sequence identity of each other sequence in the cluster with this template.

### Methods

I downloaded a sequence alignment of each sequence in the cluster OMP.12.6.1 from HHOMP. This sequence alignment contained a sequence of chain A from the PDB entry 1FW3. A script written using Biopython (Cock et al., 2009) was used to calculate the fraction of positions in each sequence that were aligned with identical residues in the 1FW3 sequence.

### Results

There are 85 sequences in OMP.12.6.1. As shown in Figure 1, almost all of the sequences have very low sequence identity with the template.

|  |
| --- |
|  |
| **Figure 1.** A histogram showing fraction identity with the template in the cluster OMP.12.6.1. |

### Discussion

The most common sequence identities are beyond the point at which homology modeling of membrane proteins tends to become less accurate (Forrest et al., 2006). This seems to suggest that the homology modeling scheme cannot be successful. However, in Experiment 2 I found that even at such low sequence identities homology models can be successfully built.

## Experiment 2: The error of predicted z values from homology models

### Introduction

The structural information used to derive Ezβ is, for each amino acid, the distance between its Cα and the center of the membrane, with a positive or a negative sign to represent that the Cα is on the extracellular or the periplasmic side, respectively. This can be imagined as the z coordinate in a 3 dimensional coordinate system in which the xy plane is the center of the membrane.

The most direct possible way to evaluate the effectiveness, for this purpose, of a modeling procedure is to carry out the procedure on a protein whose structure is already known, and compare the predicted z coordinates to the real ones.

### Methods

Outer membrane proteins of known structure were entered into HHOMP's search function, which returned the cluster with which each sequences most closely matched. A list was compiled of proteins that matched clusters containing a distinct, known structure. The protein entered in the search was used as the template for homology modeling, and the protein of known structure already in the database was the target structure to be predicted.

The z coordinates of each protein in the template structure was determined by alignment of the barrel along the z axis (Hsieh et al., 2012). The true z coordinates of the target structure were determined by structural alignment to the z-aligned template protein using PyMOL's "super" command.

The sequences in the cluster obtained by downloading the complete multiple sequence alignment from the cluster's page. The sequence of the template protein was extracted from its PDB structure and added to this alignment. Though the cluster annotation reported which protein of known structure was included in the cluster, the sequences of such proteins were not found in the alignments. Therefore, the sequence of the target structure had to be added to the alignment in the same way as that of the template structure. Then, all these sequences—cluster, template, and target—were realigned using ClustalW2 with default parameters.

z coordinates of Cα's in the target were predicted by checking the z coordinates of the Cα of the position in the template to which they were aligned by ClustalW2. The absolute values of the differences between these predicted coordinates and the actual ones taken from the structural alignment of the target structure were recorded. The results were visualized using PyMOL.

### Results

The accuracy of the models in the transmembrane strands was very high, even for very distant sequence. A large number of sequences in each of their clusters have more sequence identity with the template than the representative tested (Table 1).

When the alignment is off by a single position, the error in z is very roughly 3 Angstroms. In several of the proteins, a great deal of the residues that were off by 3 Angstroms were not off by 6 Angstroms, suggesting that parts of the protein were misaligned by a single position.

|  |
| --- |
| C:\cygwin\home\alex\beta-barrel-oligomerization\fall 2012 final paper\2mpr.png |
| **Figure 2.** The true structure of 2MPR, colored according to the deviations in the z values predicted for each residue. The colors are on a spectrum from cyan to magenta, with cyan corresponding to no error, magenta corresponding to 3 Å error or greater, and lighter shares or white for intermediate errors. Black residues are those for which no prediction could be made. The pattern illustrates that misalignments in the transmembrane strands tend to cover entire strands. |

Errors in the transmembrane strands tended to be misalignments of entire strands. Elsewhere, they were more fragmented (Figure 2).

### Discussion

The results are very encouraging. The method of alignment as it stands could probably provide a good approximation of the real depth-dependent frequencies.

There is, of course, room for improvement. The most obvious is correcting a mistake in the procedure. The sequences of the proteins of known structure were extracted from the PDB structure files, which means that any parts of the protein too flexible or disordered to be resolved are missing from the sequence.

Another easy fix is using any sequence alignment program besides ClustalW. ClustalW is the least accurate among popular sequence alignment programs (Forrest et al., 2006; Nuin et al., 2006).

The best tool for the job is ClustalΩ (Sievers et al., 2011). The most accurate sequence alignment programs are also the slowest, and will not be able to handle alignment of the larger clusters. Among programs with enough speed, ClustalΩ is the most accurate.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **a)** PDBID | **b)** %id w/ template | **c)** TM strands w/in 3 Å | **d)** TM strands w/in 6 Å | **e)** TM strands % no pred | **f)** All w/in 3 Å | **g)** All w/in 6 Å | **h)** All % no pred | **i)** Seqs above | | 1BY5 | 13.3 | 93.4 | 97.1 | 2.8 | 81.2 | 92.0 | 46.3 | 338 | | 1T16 | 12.4 | 98.6 | 100 | .8 | 74.9 | 85.7 | 25.4 | 108 | | 2MPR | 19.1 | 89.4 | 100 | 1.9 | 77.0 | 92.5 | 29.6 | 19 | | 1PHO | 56.0 | 100 | 100 | .6 | 96.0 | 100 | 7.9 | 18 | | 2OMF | 55.6 | 100 | 100 | 0 | 95.1 | 95.6 | 24.0 | 18 | |
| **Table 1. a)** PDBID of the target structure. **b)** Percent sequence identity between the target structure and the template structure. **c)** Percent of predicted z coordinates within 3 Å vertical distance of the real value in the transmembrane strands. **d)** Percent of predicted z coordinates within 6 Å vertical distance of the real value. **e)** Percent of transmembrane strand residues for which no prediction could be made. **f)** Percent of predicted z coordinates within 3 Å vertical distance of the real value in the whole protein. **g)** Percent of predicted z coordinates within 6 Å vertical distance of the real value in the whole protein. **h)** Percent of residues in the whole protein for which no prediction could be made. **i)** Sequences in the cluster with more sequence identity with the template. |

Another opportunity for improvement is in the measure of performance. It is difficult to explain the z error; people tend to get confused. A simpler or more familiar measure, such as RMSD, or percentage of correctly aligned positions taking a structure-based sequence alignment as an ideal, would make it easier to communicate the results. Since z error varies in discontinuous jumps, a small cutoff of z error is essentially just a measurement of the number of correctly aligned positions anyway.

## Other future work

### Detection of protein-protein interfaces

Ezβ has been applied to the detection of protein-protein interfaces in the outer membrane (Hsieh et al., 2012). However, my previous work showed that it is incapable of detecting interfaces using only the strand regions as data. Since these are the regions that are reliably predictable from sequence, this rules out any practical application.

There are two reasons to expect insertion energy to be elevated at permanent OMP interfaces. One reason - it could be called the stickyness argument - is that polar residues are expected to be responsible for oligomerization. The stickyness argument extends intuition from the study of transmembrane helices, whose association is driven by polar residues (Tatko et al., 2006). In fact, polar contacts are very important to the oligomerization of porins, but these contacts are not formed by the outward-facing residues on transmembrane strands (Phale et al., 1998). Crystallographers who discover porin structures often comment on the hydrophobicity of the interface (Cowan et al., 1992; Forst et al., 1998; Baslé et al., 2006), but never mention any polar contacts in the strands. In fact, a computational study found that porin interfaces are just as hydrophobic as the rest of the monomer surface (Hayat et al., 2011). If Ezβ can, in fact, be used to detect interfaces, it is not established by the stickyness argument.

Another reason—this one could be called the entropy argument—is that most of the surface should have selection pressure for low insertion energy, but the part that's buried in the interface, where such selection pressure is lessened or absent, will have higher insertion energy, simply due to a lack of selection. The entropy argument can recover from the observation of hydrophobic interfaces with the theory that they are hydrophobic for a different reason, besides insertion - and therefore, will still be distinguishable by higher Ezβ. Evidence for this comes in the form of the observation that the strict segregation of aromatic residues into two aromatic belts tends to break down at interfaces (Cowan et al., 1992), suggesting that despite the shared hydrophobicity, the depth-dependence of the frequencies is not quite the same.

It is possible, then, that there is a signal, but it is a weak signal, because of the hydrophobicity of the interface. However, by looking for a trend among the homology models of an OMP family, it may be amplified.

### Interpretation of Knowledge-Based Potentials

The justification for using the Boltzmann law to derive energies from frequencies is difficult and technical, and it may not, in fact, be justified in the case of an insertion potential (Borg et al., 2012). To a certain degree, it does not matter. Documenting the depth-dependence of frequencies has its own interest, and this knowledge can guide design without being formalized as a scoring function.

There are certain questions that seem to demand an energy function. For example: are the surfaces of OMPs naturally selected to keep the OMPs steady in the membrane, so they do not fall out? This can be translated to: is the increase in insertion energy as the protein moves up or down seem steeper than it would be by chance? However, even in this case, it seems that some heuristic scoring rule loosely related to the energy, combined with a statistical test, can do the job.

However, whatever a heuristic scoring rule might be capable of, it is hard to compare to other predictors of insertion, and hard to fit into our physical understanding of membrane proteins. For example, one might wonder, can we sum the score with an energy, calculated in another way? Is summation the appropriate operation, and do we double-count information by doing this? Also, it would be interesting to compare the scoring rule to other measures of insertion, such as the experimentally derived free energies. But are we looking for a linear relationship, or a parabolic or some other relationship?

There is another reason to look for an energetic interpretation, or at least some other interpretation. Humans are analogical reasoners, not calculators; and it is hard to use math that we cannot relate to something familiar. I am bothered that I do not have a good analogy, and I suspect that without one, I will make mistakes when I try to reason about them. For example, I used to mistakenly think of Ezβ as an energy of interaction with solvent, rather than a change in interaction energy after transfer from water. Because of this, I thought that the Ezβ moment would be sensitive to the hydrophobic residues found at the top and bottom of the ScrY interface, when in fact their Ezβ is little different than it would be if they were all aspartates. **(true?)**

We could be making such errors of interpretation in our application of Ezβ. In fact, I think I have a plausible candidate. Jim used Ezβ to calculate changes in energy as a protein moves up and down in a membrane, designing so that these changes would be steep. However, Ezβ does not really measure frequency purely as a function of the depth dependent nature of the membrane. It confounds influences due to the varying membrane environment and the varying protein environment. For example, it may be the case that while tyrosine is enriched in the aromatic belt due to the energetic favorability of the snorkeling position, phenylalanine only appears to be enriched in the aromatic belt because it is suppressed in the hydrophobic core by the strong FV antimotif (Jackups and Liang, 2005). Then, if the protein were moved so that the aromatic belt were in the center of the membrane, the energy of the tyrosines would go up, because they left their lowest energy spot in the membrane; but the energy of the phenylalanines would go down, because they are in a more hydrophobic environment but still free of valines. The protein environment moves with the protein; the lipid environment does not; the two are inseparable in Ezβ, and therefore the potential is actually far less trustworthy when used to model proteins that are not in their natural orientations. However, this is obscured by our interpretation of it as a transfer energy.

So aside from easy comparison with other energies, I do not think that there is anything that can be done with a well justified energy function that cannot be done with a heuristic scoring rule. But I do expect a well justified energy function to be easier to use appropriately. It would be a tool superior not in its cutting edge but in its handle.