

This document contains supplementary material for the application note "iMembrane: Homology-Based Membrane-Insertion of Proteins" by Sebastian Kelm, Jiye Shi and Charlotte M. Deane.

1 ALGORITHM

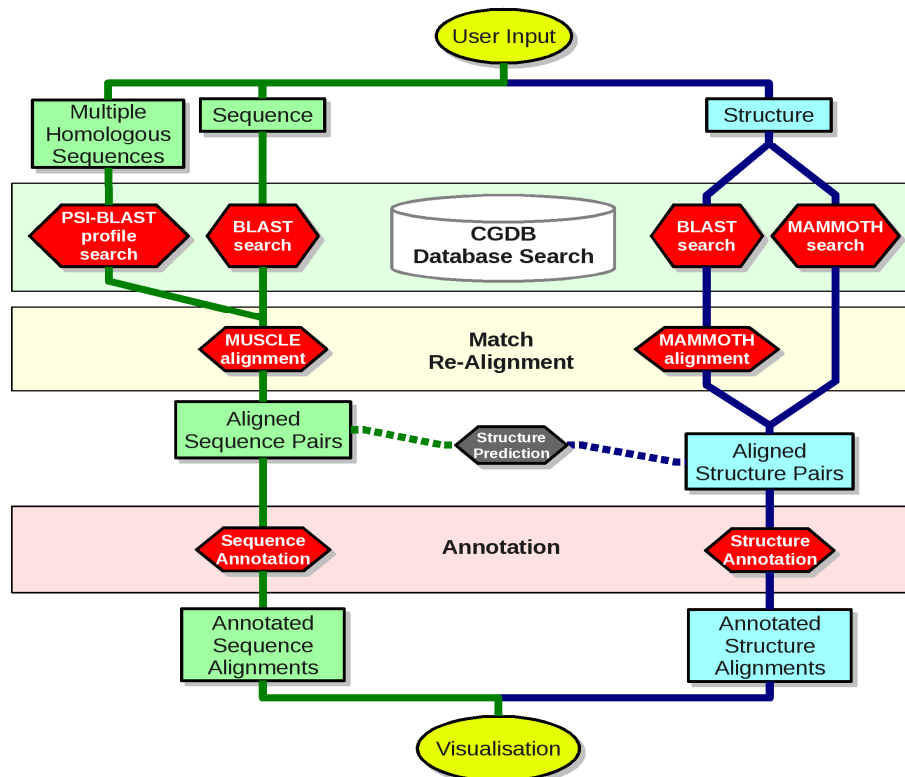


Figure 1: Flow chart of iMembrane's algorithm. Beginning at the top, users currently have three main choices of input. They may upload a structure (right branch) or a sequence (left branch). The sequence may be accompanied by a set of homologous sequences, which will increase the sensitivity of the following search step. The user's input is used to search the Course-Grained Database (CGDB) of membrane proteins for homologous proteins. The query is then re-aligned to its database match using either MUSCLE (for sequences) or MAMMOTH (for structures). This alignment is then annotated using the molecular dynamics data available for the CGDB protein. This annotated alignment is finally visualised in the iMembrane web interface. In the future, an additional structure prediction step will be implemented, such that we will be able to annotate every residue of a sequence-only input, as well as give back its proposed structure.

2 ACCURACY

Figures 2 and 3 show the relationship of accuracy with sequence identity (**A**), BLAST E-value (**B**) and MAMMOTH Z-score (**C**). **Figure 2** shows Q3 accuracy scores, **Figure 3** shows Q2 accuracy scores.

(Q3 is the fraction of residues having been correctly assigned N, H or T. Q2 is the fraction of residues having been correctly assigned N or "anything but N".)

The test results were obtained by performing a leave-one-out cross-validation of iMembrane using the CGDB proteins. For every protein, the top hit with the same number of chains as the query was chosen (hits were ranked using the BLAST E-value).

Below are further details about the different types of data shown in the plots.

Two types of annotation can be calculated, for a given input protein.

- **Membrane Contact:** Here the annotation of the CGDB hit (H, T and N) is transferred directly to the input sequence via the sequence alignment. CGDB residues are partitioned into these 3 groups using the time which they spend in each state, i.e. if a residue spent the majority of its simulation time in contact with polar head groups, it would be labeled H.
- **Membrane Layer:** Here a 3D abstraction of the contact data is created, positioning planes in optimal positions so that the majority of residues retain their original contact (H, T and N) labels. This new consistent annotation of the CGDB entry can then be projected onto the input sequence via the sequence alignment. Membrane Layer (extended) is a subtype of the Membrane Layer annotation type. It can be computed only when a structure input is given. The layer annotation can then be transferred using each residue's 3D

coordinates, instead of relying on the sequence alignment. This tends to be more accurate and allows the annotation of unaligned residues.

Two kinds of input are accepted by iMembrane: a protein structure or a protein sequence.

- **Structure-based annotation:** An input structure is annotated by performing rapid sequence alignments (BLAST) to find initial hits, followed by structure alignments (MAMMOTH). In the case of the “Membrane Layer (extended)” prediction, the annotation is transferred from the CGDB hit to the target using the 3D coordinate match and not the corresponding sequence alignment. In all other cases the transfer is made using the sequence alignment. High accuracy is consistently achieved when %ID is greater than about 20%, the BLAST E-value is below $1e-10$ or the MAMMOTH Z-score is above 20.
- **Sequence-based annotation:** When the input to iMembrane is a sequence, it is used to query CGDB with BLAST and is then re-aligned to hits using MUSCLE. All annotation is transferred using the sequence alignment. High accuracy is consistently achieved when %ID is greater than 35% or when the BLAST E-value is below $1e-25$.

A summary of the five kinds of annotation shown in the plots (Fig. 2 and 3):

Annotation Label	Input Type	Annotation Type	Annotation Method	Consistently High Accuracy Achieved At...
seq: contact	sequence	contact	alignment	ID>35%, E<e-25
seq: layer	sequence	layer	alignment	ID>35%, E<e-25
str: contact	structure	contact	alignment	ID>20%, E<e-10, Z>20
str: layer	structure	layer	alignment	ID>20%, E<e-10, Z>20
str: layer (extended)	structure	layer	3d coordinates	ID>20%, E<e-10, Z>20

Outliers:

Figures 2A, 2B, 3A and 3B show a re-occurring group of outliers: clusters of three data points observed in all alignment-based annotation types (“Membrane Contact” and “Membrane Layer”, but not “Membrane Layer (extended)”). These outliers score only around 60% and 70% accuracy, even though they have a high sequence identity of 75-100% and a BLAST E-value around $3e-80$.

All three outlier hits involve one particular CGDB protein: 2A0L, a voltage gated potassium channel from the hyperthermophilic archaea *Aeropyrum pernix*. The CGDB simulations of 2A0L show that the protein severely deforms the membrane around it, (possibly while undergoing conformational change itself), to a far greater degree than most other CGDB proteins. The twisted protein and surrounding membrane result in a different membrane contact pattern from that of the almost identical proteins tested against.

This is, however, not necessarily a problem for an observant user of iMembrane. Our accuracy test queried CGDB with other CGDB proteins. Thus, should a user query iMembrane with a third similar protein, both of the CGDB proteins that produced an inaccurate hit in the test will be presented as possible hits for the user’s third protein. It is then up to the user to decide, which of the membrane contact patterns would be the more likely choice.

Figure 2: Q3 Accuracy

Figure 2 A

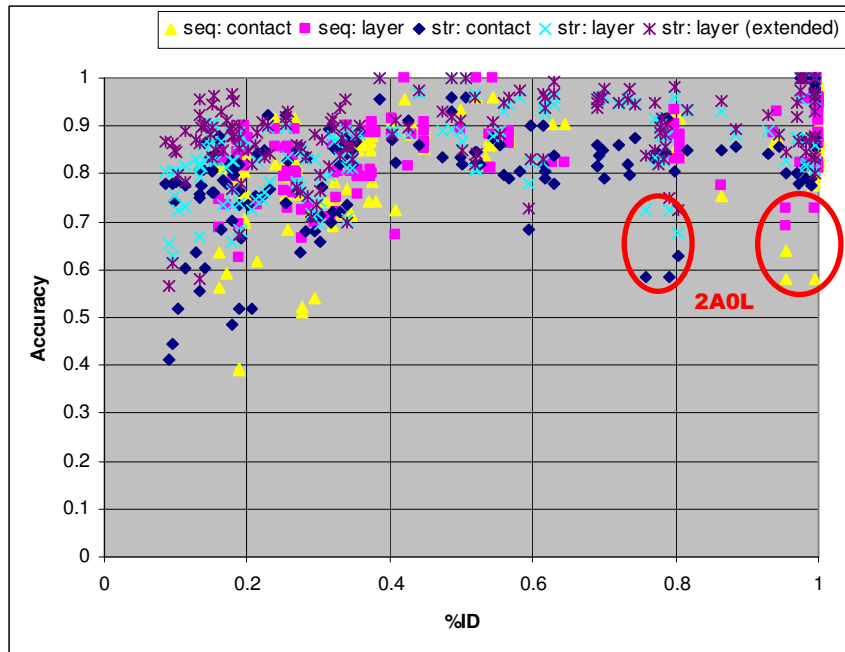


Figure 2 B

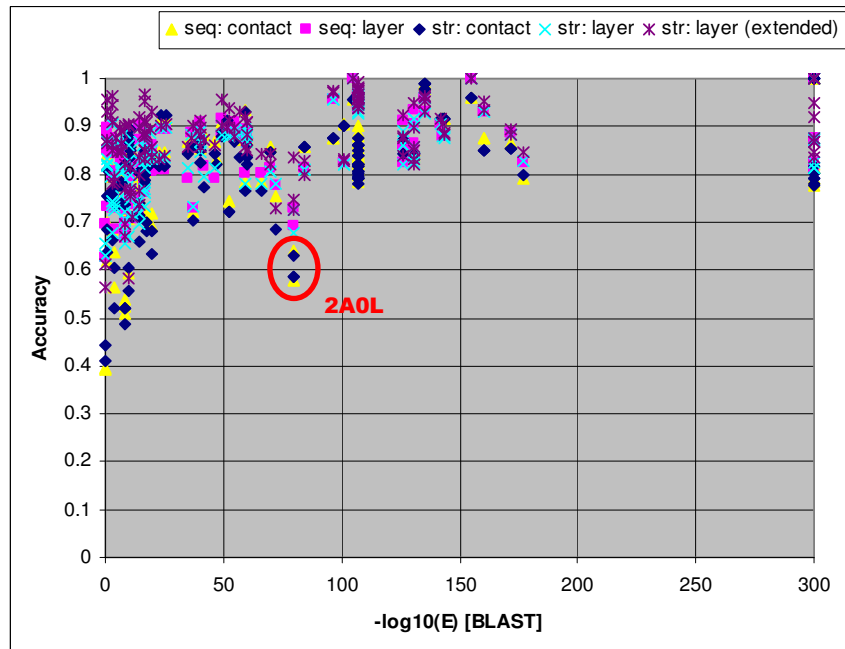


Figure 2 C

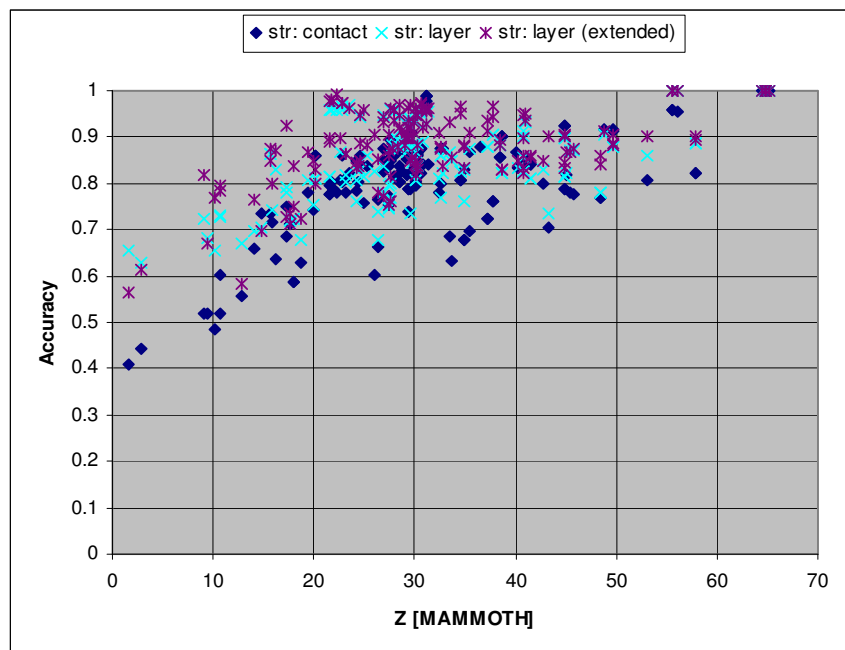


Figure 3: Q2 Accuracy

Figure 3 A

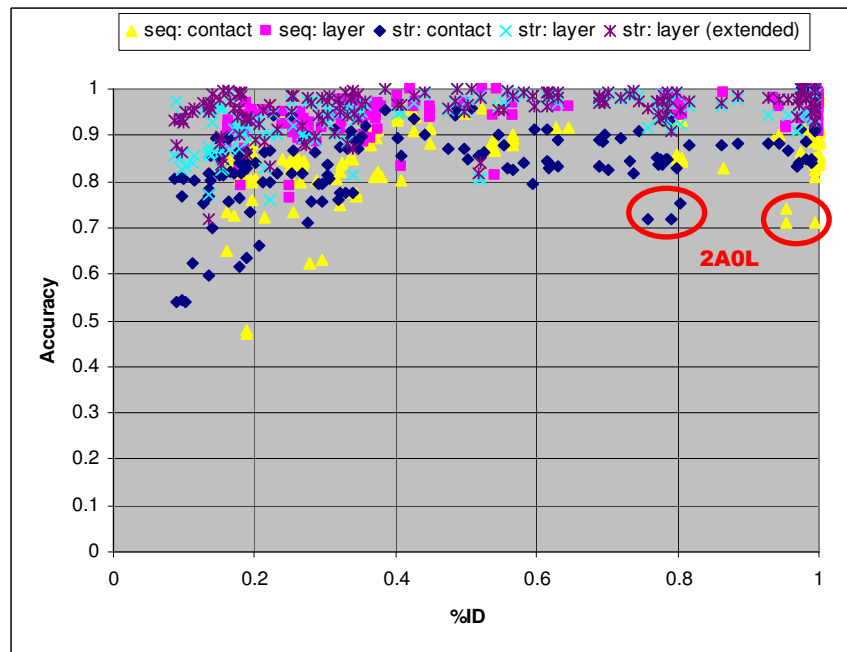


Figure 3 B

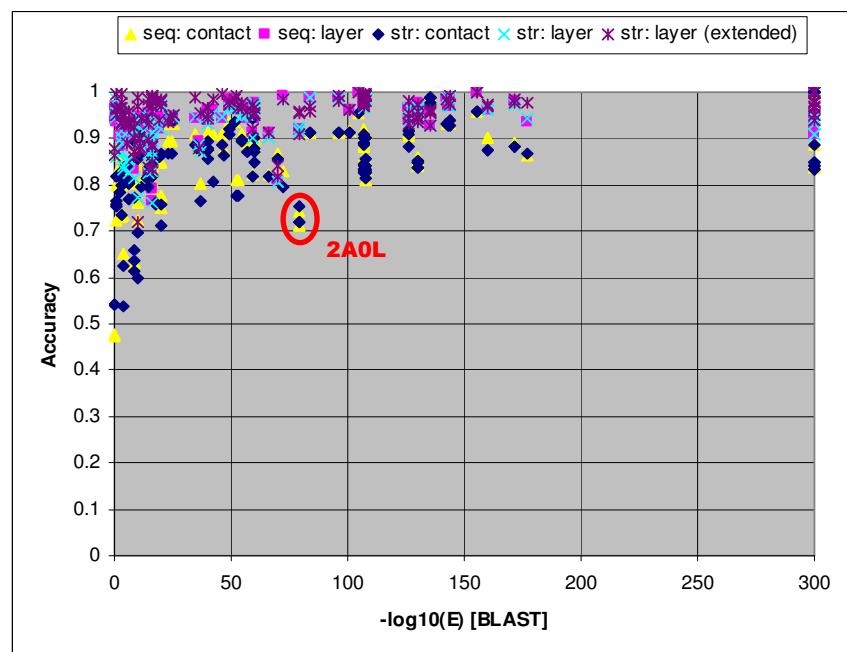


Figure 3 C

