PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Fragment-based modeling of membrane protein loops: successes, failures, and prospects for the future

Sebastian Kelm,[1]* Anna Vangone,[1,2] Yoonjoo Choi,[1] Jean-Paul Ebejer,[1] Jiye Shi,[3,4] and Charlotte M. Deane[1]

[1] Department of Statistics, University of Oxford, Oxford, United Kingdom
[2] Department of Chemistry and Biology, University of Salerno, Fisciano (SA), Italy
[3] Research and Development, UCB Pharma, Slough, United Kingdom
[4] Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, China

ABSTRACT

Membrane proteins (MPs) have become a major focus in structure prediction, due to their medical importance. There is, however, a lack of fast and reliable methods that specialize in the modeling of MP loops. Often methods designed for soluble proteins (SPs) are applied directly to MPs. In this article, we investigate the validity of such an approach in the realm of fragment-based methods. We also examined the differences in membrane and soluble protein loops that might affect accuracy. We test our ability to predict soluble and MP loops with the previously published method FREAD. We show that it is possible to predict accurately the structure of MP loops using a database of MP fragments (0.5–1 Å median root-mean-square deviation). The presence of homologous proteins in the database helps prediction accuracy. However, even when homologues are removed better results are still achieved using fragments of MPs (0.8–1.6 Å) rather than SPs (1–4 Å) to model MP loops. We find that many fragments of SPs have shapes similar to their MP counterparts but have very different sequences; however, they do not appear to differ in their substitution patterns. Our findings may allow further improvements to fragment-based loop modeling algorithms for MPs. The current version of our proof-of-concept loop modeling protocol produces high-accuracy loop models for MPs and is available as a web server at http://medeller.info/fread.

Proteins 2014; 82:175–186.

Key words: database loop structure prediction; homology modeling; FREAD; PyFREAD; MEDELLER.

INTRODUCTION

Membrane proteins (MPs) represent about one third of all known proteins. They regulate the transport of molecules and information into and out of every living cell. Due to their involvement in many medically relevant processes, they comprise over half of current drug targets.[1]

Unlike globular soluble proteins (SPs), whose natural environment is an aqueous solution (such as the cytoplasm), MPs sit inside a lipid bilayer. Thus, a large proportion of a MP's amino acids are in direct contact with the hydrophobic fatty acid tails of the membrane lipids. The presence of the membrane around the protein creates a very different physicochemical environment that has direct effects on a MP's three-dimensional (3D) structure. Transmembrane (TM) segments are usually one of two structure types: α helices or β strands. These TM segments are connected to each other by stretches of amino acids with irregular structure, known as loops. Especially in helical TM proteins, the geometry of secondary structure elements is often well-conserved, with approximately parallel helices being oriented perpendicular to the membrane plane (parallel to the membrane normal) and spanning the entire width of the membrane. The structure of the loop regions connecting the TM segments can vary greatly between homologues.[2]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: 2020 Science project (http://www.2020science.net).

*Correspondence to: Sebastian Kelm; Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom.
E-mail: kelm@stats.ox.ac.uk

Received 8 November 2012; Revised 22 February 2013; Accepted 26 March 2013
Published online 15 April 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24299

Therefore, loops tend to be the parts of MPs that are the hardest to model.

In MPs, loops can interact with the polar head groups of the membrane lipids as well as with water molecules and thus tend to contain many hydrophilic and charged residues. Positively charged amino acids such as lysine and arginine are especially common in loops protruding into the cytosol (the positive inside rule).[3,4] In addition to their chemical properties, MP loops may also have characteristic shapes. The typical MP loop connects two roughly parallel TM segments and protrudes from the membrane into a polar environment.

Due to the physical crowding of the membrane, the loop tends not to interact with other parts of the protein, except other loops, but instead may be found touching the polar head groups of the membrane lipids. Loops in SPs, on the other hand, can interact with sequentially distant residues and often lie on the surface of the protein rather than protruding from it.

Three-dimensional modeling of loops occurs under the constraints of the rest of the structure, in particular the two anchor regions of the loop (here the anchors are defined as the two residues on either side of the loop, i.e., four residues in total). Loop modeling approaches are often split into two types: *ab initio* and database methods.

Ab initio methods typically sample dihedral angles from empirical distributions specific to each amino acid type to create a starting conformation. These sampled conformations are unlikely to connect the two anchors and hence a separate step is needed to close the gap. This “loop closure” step is often performed by modifying the dihedral and bond angles of the loop to minimize the size of the gap between the free end of the loop and the corresponding anchor atoms. One such loop closure technique is the cyclic coordinate descent algorithm,[5] which iteratively modifies one dihedral angle at a time, but there are alternatives that optimize all dihedral angles simultaneously.[6–8]

Clashing loop structures are removed and the remaining ones are clustered to reduce redundancy. Such starting conformations can be further refined by optimizing energy functions, such as the CHARMM or AMBER molecular dynamics force fields. Finally, the remaining conformations can be ranked using an energy function, which is not necessarily the same as was used for refinement.

Protein Local Optimization Program (PLOP)[9,10] RAPPER,[11,12] and MODELLER[13] are examples of *ab initio* methods. The Internal Coordinate Mechanics (ICM) modeling package[14,15] also contains loop modeling procedures, which have recently been updated to use the Internal Coordinate Mechanics Force Field (ICMFF) force field.[16]

The three main problem areas for *ab initio* methods are: a) achieving a sufficient sampling of conformational space; b) achieving an accurate ranking of these decoys with the energy function; c) keeping the computational cost manageable. In a 2010 benchmark, the *ab initio* methods achieved a sampling on par with our database method FREAD and often succeeded in generating near-native conformations.[17] Unfortunately, *ab initio* methods generally had problems identifying the best conformations, resulting in poor average accuracies. In this 2010 comparison of loop modeling methods, RAPPER (and similarly PLOP and MODELLER) achieved average accuracies of around 1, 3, 6, and 10 Å root-mean-square deviation (RMSD) for loops of lengths 4, 9, 15, and 20 residues.[17] This dependence of accuracy on the length of the loop is to be expected, as the computational cost of sampling and energy calculations tends to scale exponentially with the length of the loop. In the original publications of PLOP, the authors reported sub-Ångström accuracies for loops of up to eight residues in length and around 1–2 Å RMSD up to 13 residues. It is likely that this difference in observed accuracy was due to multiple factors, one of which was the choice of dataset. The 2010 assessment[17] possibly contained many loops which were less suitable to *ab initio* prediction than the dataset used by the authors of PLOP. Since then, the authors of ICM and the authors of PLOP have demonstrated accuracies around 1 Å for sets of loops up to 17 residues in length.[16,18] The computational time required to perform these predictions ranged from 2–115 CPU days for PLOP,[18] with averages of about 10 days (on Intel/AMD processors with clock speeds of 900 MHz–1.4 GHz). ICMs computational cost was an average of 2.5 h for eight residue loops and 55 h for 12 residue loops (on Intel Core2 2.13-GHz processors).

Database methods, unlike *ab initio* methods, perform conformational sampling by selecting fragments from a predefined database of known structures. These fragments must share certain local properties with the query loop, such as having similar anchor structures and amino acid sequence. After insertion into the model, clashing fragments are discarded and the remaining ones ranked using a similarity measure comparing the database fragment to what is known of the query loop (anchors and sequence). Examples of database methods are LIP,[19] the works of Fernandez-Fuentes *et al.*,[20] and Peng and Yang,[21] SuperLooper,[22] and FREAD.[17,23] So far, FREAD is the only published database method that achieves constant accuracy around 1 Å for loops up to Length 8 and around 1–3 Å up to Length 20,[17] independent of loop length. The greatest advantage of database methods is their unmatched speed-to-accuracy ratio. Near-native accuracies, on par with or better than the best available *ab initio* methods, can be achieved within seconds or minutes of CPU time. The inherent disadvantage of any database method, however, is the lack of complete coverage. Due to the many degrees of freedom of loop structure (and sequence) and the difficulty in obtaining structural information, we currently do not have experimentally determined structures covering every possible loop conformation for loops of any significant length. It is thus possible that a database method will be unable to give a good prediction for a given loop, simply because the "correct" conformation is not present in the database. The performance of database methods is known to depend on several factors: the search algorithm used, database completeness, and the choice of anchor structures,[24] as well as the choice of the database itself.[23]

In practice, one would tend to use a combination of database and *ab initio* methods. Large-scale proteomics studies might use only database methods, due to their high speed, whereas scientists studying a single protein in great detail might favor a combination of database predictions and compute intensive *ab initio* modeling.

In this article, we focus on database search methods, represented by our fragment-based loop modeling method FREAD. When considering a database method for MPs, two problems immediately become apparent. The first problem is a lack of MPs of known structure. It is well-established that the more complete the database, the better the results will be. This, one might expect, would result in a generally poor performance, both in terms of coverage and accuracy, of any method relying on a MP database.

In 2012, we now know more than 80,000 protein structures from 1393 Structural Classification of Proteins (SCOP)[25] folds. Fernandez-Fuentes *et al.*[26] showed that conformations of short protein fragments were becoming saturated for fragments up to 14 residues in length. For each such fragment of known structure, there was at least one other fragment in the database with over 50% sequence identity and a similar conformation. At that time (in 2006), there were less than 20 MPs in the Protein Data Bank (PDB) (Stephen White's database of known MPs structures*). These results can thus be thought of as applying only to SPs. More recently, Choi and Deane[17] showed that soluble protein loops can now be modeled using a fragment-based approach at a high accuracy (around 1 Å global RMSD, see MATERIALS AND METHODS) and a coverage of around 60%.

This recent success of fragment-based loop modeling methods is due to the increasing saturation of soluble protein structures in the PDB.[27] When such methods were first conceived,[28] they showed promising results but were extremely limited by the low availability of proteins of known structure. According to statistics published on the PDB website,† in 1986 the PDB contained 213 protein structures from 50 SCOP folds. This situation was similar to the current state of affairs in the world of MP structures today. In 2011, the PDB contained 308 unique MP structures (Stephen White's database*), a number that is likely to keep increasing. We anticipate that the success of MP-specific methods will grow with database size in a similar way as it did for soluble protein methods.

The second problem in benchmarking a database method is that of homology: even if good loop modeling accuracies are achieved, are these simply due to copying the loop structure from a 99% identical protein? If so, one would not be assessing the true performance of the loop modeling method and similar results should be achievable irrespective of the method used. To overcome this benchmarking problem, we explicitly filtered out homologous database hits for each modeled loop before assessing modeling performance.

The FREAD method has previously been shown to produce high accuracy models of SP loops[17] and antibody complementarity determining regions,[29] independent of loop length. In this article, we demonstrate that models of MP loops built in the same way, through the use of SP fragments, are far less accurate. In contrast, we show that modeling MP loops using fragments of MPs can yield high accuracy models. We explore the reasons for the successes and failures of these different strategies by analyzing statistical properties of MP and soluble protein loops.

MATERIALS AND METHODS

Native loop test sets

This study uses two sets of X-ray structures: one containing only SPs, another containing only MPs.

An initial list of potential MPs was created by listing the union of PDB[27] codes contained within the Protein Data Bank of Transmembrane Proteins (PDB_TM),[30] Orientations of Proteins in Membranes (OPM),[31] and Coarse-Grained Database (CGDB)[32] databases and those in the SCOP[25] category "membrane and cell surface proteins and peptides."

This list was then run through the Protein Sequence Culling Server (PISCES) server,[33] to keep only X-ray structures with resolution $\leq$ 3 Å, R factor $\leq$ 0.3, and length $\geq$ 40. The remaining structures were split into component chains, and duplicate chains (with 100% sequence identity) were removed. Each structure was then run though iMembrane[34] and only those chains with an iMembrane hit were retained. This ensured that only true MPs were part of the dataset.

Residues annotated by JOY[35] as being anything but helices and sheets were treated as loop residues. Only loops of length >3 and within the membrane, or close to it, were considered. Loops close to the membrane were defined as those which start/end within four residues of the nearest TM residue.

For each loop length, loops were clustered by sequence identity using Unweighted Pair Group Method with Arithmetic Mean (UPGMA)[36] and made nonredundant

*http://blanco.biomol.uci.edu/mpstruc/

†http://www.pdb.org

at the 40% identity level. From each cluster, the representative loop with the lowest average B factor was chosen.

For the purpose of testing FREAD prediction accuracy and coverage, 20 representative loops of lengths 4–17 (residues) were randomly chosen. Only 20 loops were chosen as the number of examples in the database of known MP structures is low at longer loop lengths.

A list of nonredundant loops in SPs had previously been compiled in analogs fashion[17] and made available for download.[‡] From this list, we randomly chose 20 representatives for each loop length from 4 to 17. For further analyses, a second test set of 90 loops per loop length, with lengths ranging from four to nine residues, was chosen from the membrane and soluble loop clusters. The higher number of examples allows more reliable observations to be made at each loop length. Due to the small size of the membrane database, however, this amount of nonredundant examples is currently only available for loop lengths up to nine residues.

Homology model test sets

All the above test sets contain known X-ray structures of proteins, taken from the PDB. In addition, we created a test set of homology models of MPs, which is a subset of the MEDELLER test set.[37] We grouped the 616 pairs of MPs found in the MEDELLER set by their target protein. For each target protein, we chose the most sequence-similar template, excluding any template above 90% sequence identity (to ensure that there were gaps in most of the models). This resulted in a set of 156 loops, with lengths ranging from four to 17 residues, from 59 models of varying accuracy (average core model RMSD 2.2 Å; range 0.6–5.4 Å).

The FREAD loop modeling algorithm

The details of the FREAD algorithm have previously been published.[17] We use a reimplementation of the algorithm that runs an order of magnitude faster than the original.

The input to the FREAD algorithm is the incomplete model of a target protein, lacking the coordinates of a particular query loop. The amino acid sequence of the query loop is known and is provided as a second input. From the model, the main chain coordinates of the loop "anchors" (two residues on each side of the loop) are extracted. A given database of known protein structures is then searched for fragments of the required length with a similar sequence and anchor geometry.

The anchor geometry match is performed by superimposing the query and database anchor coordinates and calculating the RMSD for all main chain atoms (N, Cα, C, O). In this work, only loops with an anchor RMSD match below 1 Å are accepted.

The sequence match is made using environment-specific substitution tables (ESSTs). These tables are different from standard ESSTs in several ways: First, they are based on structural environments defined using the dihedral angles of a protein's backbone coordinates. FREAD uses six classes of dihedral (φ and ψ) angles, which are partitions in the Ramachandran plot (see Supporting Information, Fig. S1). Second, only residues within loop regions are considered when building the tables. Third, only substitutions between residues with identical dihedral angle classes are counted. The potential database match sequence is scored against the query loop sequence by summing the individual subscores. A total score of 25 or more is taken to signify a "good" match.[17]

Once suitable database loops have been identified, they are inserted into the model and a clash check is performed.[38] Loops clashing with the model framework are discarded. This step replaces the Samudrala–Moult pseudoenergy calculation used in the original FREAD program.[17] The clash check has the advantage of being applicable to any type of protein, without needing to be trained on a set of known protein structures.

Finally, database loops matching the search criteria are ranked by their anchor RMSD and returned to the user, along with their 3D coordinates.

As in the original FREAD publication, we assess the accuracy of loop models using the "global" all-backbone-atom RMSD of the loop residues compared to the native X-ray structure.

Calculating loop modeling accuracy

A loop is modeled given the loop's amino acid sequence and the 3D coordinates of its anchor residues. The anchor residues are the two amino acids on either side of the loop (four residues in total). To calculate a loop model's accuracy, its anchor residues are superimposed onto the anchor residues of the native protein structure. Then, the RMSD is calculated between the loop's backbone atoms (N, Cα, C, O) in the native structure and the model. This measure is termed the "global RMSD," as it does not involve any "local" superposition of the loop residues themselves.

Fragment databases

As described above, FREAD requires a database of known structures from which to identify its prediction. The "soluble" database was created by filtering the entire PDB using PISCES, to keep only X-ray structures with resolution ≤ 3 Å, R factor ≤ 0.3, and length ≥ 40, and making the resulting list nonredundant at a sequence identity level of 99%. Those chains included in the initial list of potential MPs were removed from the "soluble" database.

The "membrane" database was created from the same set of iMembrane-annotated structures described in the "native loop test set" section.

[‡] http://opig.stats.ox.ac.uk/sites/fread/

Both membrane and soluble databases include entire proteins, not just the loop regions. The first and last five residues in each protein were discarded.

As the fragment databases and test sets were created from the same dataset, cases where the query and database loop were identical were always eliminated from the FREAD predictions and, as discussed below, homologues were also eliminated in some tests.

Substitution tables

Two substitution table sets were tested: tables created from SP loops and tables built from MP loops.

These sets of ESSTs were created from the membrane and soluble alignment datasets described by Hill *et al.*[39] The major difference between this dataset and the one used to build the original FREAD tables is that the original dataset contained pure structure alignments, where the dihedral angle class of each amino acid in every aligned protein was known. The dataset used here contains alignments between a single protein of known structure and multiple homologous sequences (due to the low availability of MP structures). We thus only know the dihedral angles for a single sequence in the alignment. It is therefore impossible to constrain the tables such that they only contain substitutions between residues of identical dihedral angle class, as was done in the original FREAD tables.

To avoid introducing a large amount of errors into the tables, we counted substitutions in only select parts of each alignment, as follows. Before generating the substitution tables, homologous sequences with identity $<20\%$ to their corresponding protein of known structure were discarded, to avoid errors introduced via low quality alignments. In all remaining sequences, only loops with fewer than four gaps were used to count substitutions, to further ensure only well-aligned loops were considered.

These parameters were chosen from a set of similar parameters by identifying those that minimized the difference between our new soluble tables and those from the original FREAD implementation. The same set of parameters was then used to build equivalent tables for MPs.

Homology filtering

For several of the tests presented in this article, homologous proteins were filtered out of FREAD's database hits before selecting one of the remaining fragments as the prediction. The decision of whether or not to discard a database fragment was made by aligning the complete structures of the query protein to the database protein providing the fragment. A structure superposition was performed using the TM-align program,[40] and the percentage identity was calculated from the implicit sequence alignment, normalized by the number of aligned residues. Proteins with over 40% identity were deemed homologous and the database fragment discarded. We chose 40% as our cut-off to find a compromise between nonhomology and dataset size. In addition, we discarded fragments of all proteins which were annotated as being part of the same SCOP superfamily as the query protein (SCOP version 1.75B). As loops are the least conserved parts of a protein, we expect that the loop regions will have different conformations, even in those cases where potentially homologous proteins remain in the dataset.

RESULTS

FREAD prediction and the importance of database choice

The focus of our first set of experiments was the ability of FREAD to predict the structure of MP loops, as opposed to SP loops. Using our dataset of 20-MP loops per loop length, with lengths ranging from 4 to 17 residues, and our equivalent test set for SP loops, FREAD was run using the default set of cut-off parameters (anchor RMSD $\leq$ 1.0 Å, environment-specific substitution score $\geq$ 25).

Other than the query loop's sequence and anchor structures, which are provided as input, FREAD relies on two pieces of data to perform its predictions: a fragment database and a set of ESSTs. We tested the effect of using either a MP-specific or SP-specific version of both database and substitution tables.

Large differences in accuracy and coverage resulted from the change in database. As shown in Figure 1(B), predicting SP loops using a soluble database consistently yields accuracies close to 1 Å. The same is true when predicting MP loops using a MP database (A). However, predicting MP loops with a soluble database results in low accuracies (A) and coverage (C), both of which worsen with increasing loop length. Predicting soluble loops using a MP database yields close to zero coverage (D). The latter failure could be ascribed to the comparatively small size of the MP database. The reverse, on the other hand, that is, failure to predict MP loops with a SP database, cannot be due to database size.

In general, no significant differences were observed with regard to the type of substitution tables used (see Supporting Information, Fig. S3). This is not entirely unexpected, considering the current choice of structural environments in FREAD, which are based solely on dihedral angles. In previous work,[39] we observed no obvious difference between the "coil" (loop) substitution tables in MPs and SPs. All the remaining results presented in this manuscript were thus produced using the same set of substitution tables, created from alignments of SPs.

It thus appears that the FREAD approach is valid for both soluble and MPs separately. However, the poor

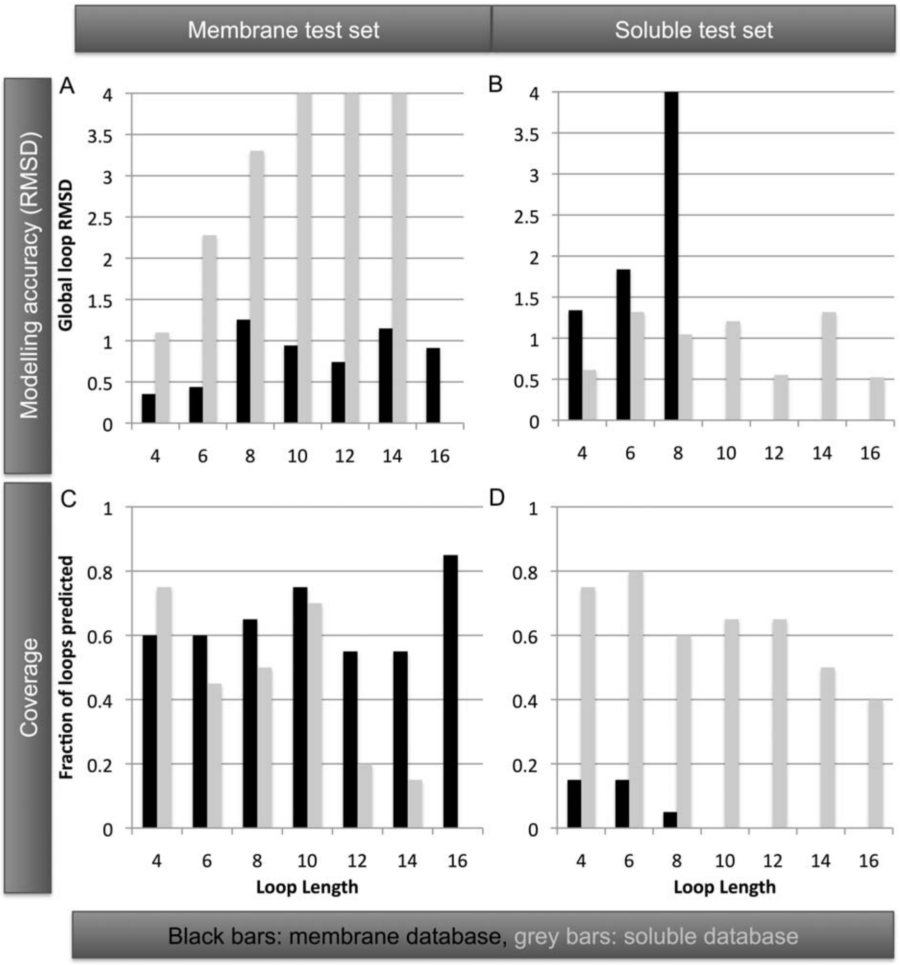

Figure 1

Effect of database choice on loop prediction accuracy (RMSD) and coverage at each loop length. Accuracy (**A**, **C**; left column) and coverage (**B**, **D**; right column) for the soluble (A, B; top row) and membrane (C, D; bottom row) loop test sets. Gray bars represent predictions made using a soluble database, black bars are predictions made using a membrane database. Thus, for example, in (A) it can be seen that membrane loops of Length 6 are predicted better with the membrane database than with the soluble database, whereas (C) shows that the membrane database achieves slightly higher coverage at the same loop length. A version of this figure showing all loop lengths from 4 to 17 can be found in Supporting Information, Figure S2.

accuracies and especially the poor coverage when predicting MP loops using a soluble database seem to suggest the existence of an intrinsic difference between soluble and membrane loops. This difference may be purely structural but could also be sequence composition-related, though it appears not to be due to differences in amino acid substitutions.

In the following sections, we explore the factors that may affect the accuracy of FREAD predictions and the potential differences between SP and MP loops.

Why does FREAD work, given the right database? Why does it matter?

A question that often arises when considering fragment-based loop prediction is that of the fragment's origin. One has to consider the possibility that the 1 Å accuracies achieved by FREAD might only be observed because of database redundancy. If every prediction were actually made using a database protein highly sequence similar to the query protein, this could mean that the method is unable to perform well in the absence of close homologues.

To address this question, we used two new test sets for SP and MP loops each with 90 loops per loop length and lengths ranging from four to nine residues. This increased number of examples at each loop length allows for more reliable analyses. The drawback is that we cannot include loops longer than nine residues in this dataset due to the low number of available examples in the database of known MP structures.

Rather than run the complete FREAD algorithm, we applied only the anchor RMSD filter, which selects database loops with anchor atoms within 1 Å of the query protein. We then identified the best-matching database loop, whose 3D coordinates had the lowest RMSD to the correct conformation. This procedure allows us to assess the best-possible RMSDs theoretically achievable by the FREAD algorithm, given a specific database.

The resulting distribution of "best possible RMSDs" for each loop length are shown in Figure 2(A–C). When querying the soluble database with SP loops [Fig. 2(A)], the median RMSD is always far lower than when querying the same database with MP loops [Fig. 2(B)]. For instance, at Length 8, SP loops achieve a median RMSD of about 0.8 Å versus about 1.7 Å for MP loops. Querying the membrane database with MP loops also achieves median RMSDs of about 0.8 Å [Fig. 2(C)], at Length 8. This means that we commonly find highly similar SP loops as well as highly similar MP loops, but that it is harder to find SP loops with a similar shape to MP loops. The question now arises as to whether this difference is due to intrinsic properties of soluble and membrane loops, or whether it can be explained by the presence of homologues in both the SP and MP databases.

To answer this question, we needed to remove the effect of homologues on our RMSD measurements. We thus repeated the same procedure as above but disregarded any database loops originating from a protein homologous to the query protein [Fig. 2(D–F)]. For this purpose, we defined two proteins as being "homologous" if they shared over 40% global sequence identity or were annotated as being part of the same SCOP superfamily.

Although the use of homologues allows sub-Ångström RMSDs across the tested loop lengths [Fig. 2(A,C)], the removal of the homologous hits results in a strong length dependence of the achievable RMSD [Fig. 2(D,F)]. In other words, the use of homologues has a larger effect on accuracy at higher loop lengths. This length dependence is to be expected, as any structure database is more likely to contain all possible conformations of a Length-4 fragment than a Length-9 fragment (due to the increased degrees of freedom for the structure of longer loops). This result suggests that FREAD's consistent 1 Å accuracy across loop lengths (Fig. 1) is made possible by the existence of homologues within the database and that there are many loops (in both MPs and SPs) that cannot be predicted to the same level of accuracy without the use of homology information.

Furthermore, as is illustrated by Figure 2(D,E), the removal of homologous hits has a similar effect on the two test sets. When using a soluble database, Length-8 soluble and membrane loops have a median achievable RMSD of about 1.4 and 1.7 Å, respectively. This indicates that the shapes of MP and SP loops can be found in the SP database with about equal frequency. In other words, it is possible to find SP loops that are equally distant from the "correct" shape of a query MP loop as the best nonhomologous MP loop. Median RMSDs achievable when using the MP database on MP loops [Fig. 2(F)] are somewhat higher, at about 2.6 Å for Length-8 loops. It is likely that this is due to the far smaller size of the MP database.

The above observations lead us directly to the question: what accuracy and coverage can we expect of FREAD in the absence of homologues? We thus reintroduced the environment-specific substitution score (ESSS) cut-off and filtered out all database loops with an ESSS <25. This can be seen as the removal of all database loops that may or may not have the correct shape but are too sequence-dissimilar to be identified by the FREAD algorithm. Again, we picked the most accurate solution from the remaining database loops.

The introduction of the substitution score cut-off resulted in a reduction in coverage (from 100%, across the board, to the values shown in Table I) but also had a large effect on accuracy, as shown in Figure 2(G–I). For SPs (G), accuracy was between 1 and 2.5 Å for all loop lengths and coverage was 46/90 (51%) for Length-9 loops. For MPs, a constant accuracy of 1 Å across loop lengths could now be observed, when predicting with a MP database (I) although coverage was reduced to 16/90 (18%) for Length-9 loops. Predicting MPs with a SP database yielded far worse accuracies with the median RMSD for Length-9 loops being about 4 Å and a coverage of 34/90 (38%).

From these results, it is tempting to suggest that membrane loops may in fact be easier to predict than soluble loops, in the absence of homologues, as long as one uses the correct database. However, considering the low number of examples this is only speculation. The most striking observation here is, once again, that the FREAD substitution score cut-off successfully filters out inaccurate loop structures, resulting in good accuracies independent of loop length, as long as the query and database loops originate from the same type of protein (MP or SP). This is true even in the absence of homologues in the database. The ESSS cut-off does not work as expected when query and database loops come from different protein types. There thus appears to be an intrinsic difference between membrane and soluble loops causing the FREAD approach to fail. As mentioned earlier, this difference did not appear to lie in the substitution patterns of MP and SP loops.

Finally, we added the last stage of the FREAD algorithm, the ranking by anchor RMSD. This is now equivalent to a true modeling situation, as no use of prior knowledge of the correct query loop conformation is made at any point. As shown in Figure 2(J–L), all results were near-identical to those in the previous step.

We can thus conclude that 1–2.5 Å accuracy across all tested loop lengths is made possible by the FREAD substitution score cut-off (assuming the anchors are already restricted), even in the absence of homologues, as long

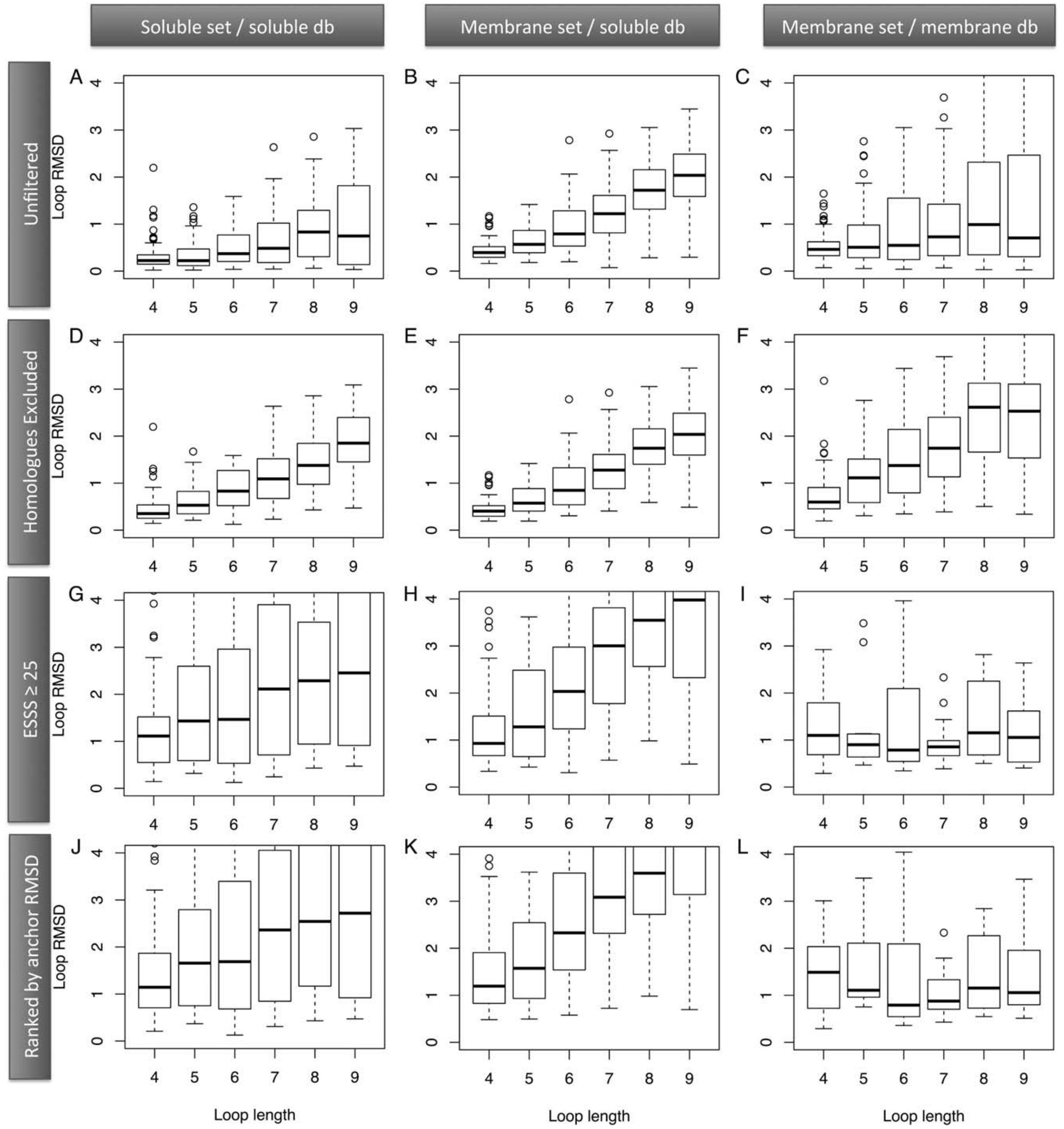

Figure 2

Boxplots of achievable loop modeling accuracies. Every plot shows the loop accuracy by RMSD (Y axis) at each loop length (X axis) achievable on the soluble test set (left column; **A**, **D**, **G**, **J**) and membrane test set when using soluble protein fragments (middle column; **B**, **E**, **H**, **K**), or when predicting membrane loops using a membrane protein database (right column; **C**, **F**, **I**, **L**). In the top row (A–C), only the best-matching fragment in the database is selected for each of the 90 query loops in each length category (number of residues, X axis). In the second row (D–F), the best-matching fragment are selected (as in the top row) except all hits to homologous proteins (>40% sequence identity or annotated as pertaining to the same SCOP superfamily) are removed. Third row (G–I): as in the second row, but now database loops are restricted to those with an environment-specific substitution score $\geq$ 25. Bottom row (J–L): as in the third row, but database loops are now ranked by their anchor RMSD, instead of using knowledge of their similarity to the correct loop conformation.

Table I
Coverage in the Absence of Homologues

Loop length	4	5	6	7	8	9
Sol/sol	52 (58%)	56 (62%)	56 (62%)	61 (68%)	44 (49%)	46 (51%)
Mem/sol	57 (63%)	42 (47%)	51 (57%)	52 (58%)	33 (37%)	34 (38%)
Mem/mem	12 (13%)	11 (12%)	16 (18%)	14 (16%)	12 (13%)	16 (18%)

Coverage of the FREAD loop modelling method, when homologues (protein with >40% sequence identity or annotated as pertaining to the same SCOP superfamily as the query) have been excluded from the database. Loop lengths of 4–9 residues are considered (columns), when predicting the Soluble or Membrane test set (first word in row labels), using the Soluble or Membrane database (second word in row labels). Values are total numbers of predicted loops out of a total of 90.

as the correct database is used. The presence of homologues in the database improves this accuracy to about 1 Å and dramatically increases coverage, especially in the case of MPs (from about 18%, Table I, to 60%, Fig. 1). This coverage difference is due to the high redundancy of MP structures in the current database. The majority of MPs of known structure are members of a small number of structural families and often also share considerable sequence similarity.

Homologues thus play a large role in loop structure prediction, but even more important is the difference between membrane and soluble protein loops. Being a membrane, protein loop has many implications in terms of the physicochemical environment around the loop residues, as well as the structure of the protein itself. This may bias the loop to adopt very different shapes from a soluble loop with a similar anchor structure and amino acid sequence. To obtain good loop models, one should therefore first match the type of protein in the fragment database to the type of protein being modeled. Second, improvements in accuracy and coverage can be achieved by making the database as redundant as possible through the inclusion of homologous proteins.

Differences between membrane and soluble protein loop structures

We showed that predicting MP loops by applying the standard FREAD protocol to a SP database fails. This is illustrated in Figure 3. We also showed that the SP database does contain fragments with similar shapes to MP loops, although these become less frequent at higher loop lengths, and that the FREAD substitution score was unable to identify these "good" fragments. Although using the MP database for prediction virtually guarantees high accuracies, this is not practically useful when modeling a protein of a novel fold or sequence family, as coverage values below 20% are expected (Table I). In other words, four out of five loops would remain unmodeled. Identifying the differences between membrane and soluble protein loop structures is thus of great interest, as this would suggest possible ways of using the much larger SP database to predict MP loops.

Our hypothesis was that the shapes of membrane loops tend to be biased, due to the presence of nearly parallel TM segments and the crowded environment of the membrane lipids. We propose that MP loops will favor a straight conformation, sticking out of the membrane away from the remainder of the proteins TM domain. In contrast, although some SP loops might have similar shapes, they are not confined by the membrane and will be more often able to "lie down" on the surface of the protein, in contact with sequentially distant residues and forming a more globular shape. To test this hypothesis, we compared membrane and soluble protein loops and found the expected trends but almost none of the differences were statistically significant. Further details can be found in supplementary material (Supporting Information, Figs. S4–S6).

Prediction on models of MPs

All previous tests were performed on X-ray structures. Although the loop structures were unknown, the anchor structures were known to be correct. This is the usual test case for most loop modeling methods, including the reportedly highly accurate ICM[16] and PLOP.[18] In a real modeling situation, however, the anchor structures are

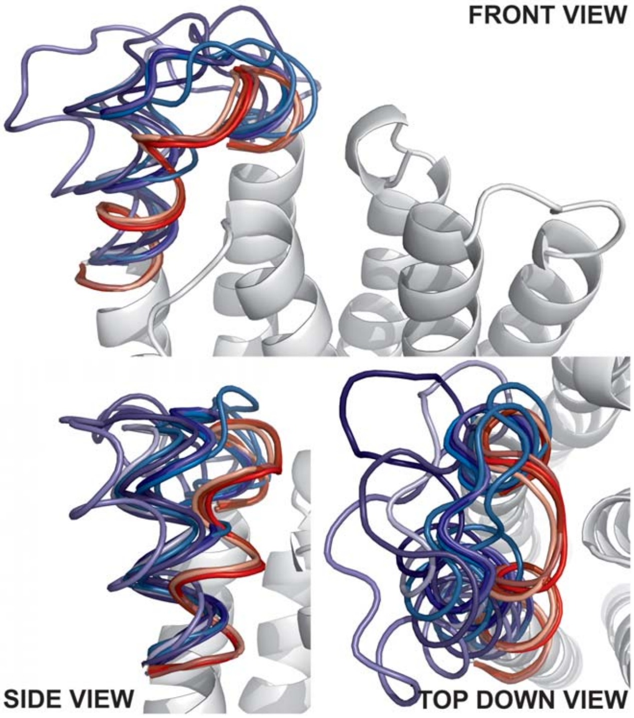

Figure 3
Predictions of a membrane protein loops structure. Loops in shades of blue originate from a database of soluble protein structures, loops in shades of red come from a membrane protein database. Note that not a single soluble loop has a similar shape to the membrane loops. The latter are all close to the native structure.

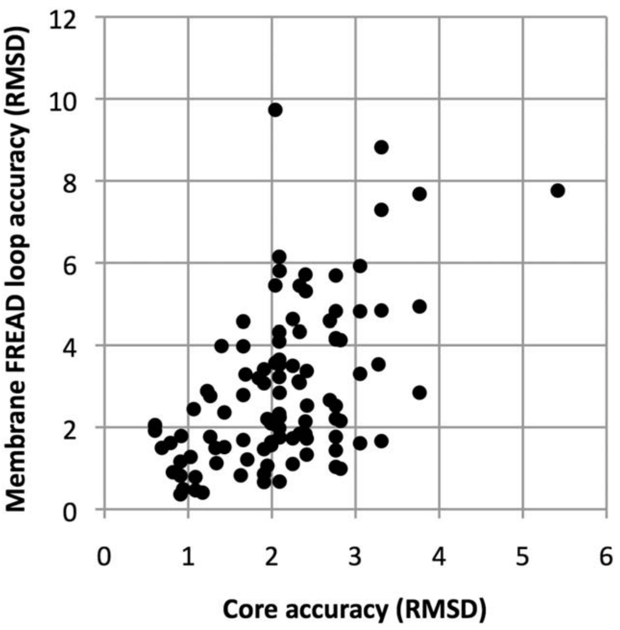

Figure 4

Loop prediction accuracy depends on the accuracy of anchor coordinates. Plot of loop prediction accuracy by RMSD (Y-axis) versus modeling accuracy of the core model, without the loops, by RMSD (X-axis).

nonnative, making it harder to correctly predict the loop's structure. The loop's global RMSD (see MATERIALS AND METHODS) directly depends on the anchor RMSD (Fig. 4). FREAD has previously been shown to perform well in such an imperfect scenario.[17,29,37] For SPs, the relationship between a loop's local sequence similarity and its expected accuracy can be deduced from Figure 1 of Ref. 17. Figure 2 of the same paper shows that FREAD's accuracy does not depend on global similarity between the target protein and the database protein donating the loop fragment. More recently, PLOP was the first physics-based *ab initio* method to be tested in this way,[41] achieving good accuracies on several predicted GPCR loops. In this section of our article, we thus verify once more our main conclusions regarding the importance of database choice, this time applied to a more realistic modeling scenario where loop anchors are incorrect.

The model test set consists of 156 loops, ranging in length from four to 17 residues, taken from 59 homology models of varying accuracy (average core model RMSD 2.2 Å; range 0.6–5.4 Å). Here, we predict the loop structures with either the soluble database ("Soluble FREAD") or membrane database ("Membrane FREAD"). Membrane FREAD was run using the two strategies already used in the original MEDELLER publication[37]: the "high accuracy" strategy, which uses FREADs default cut-offs, and the "high coverage" strategy where, if the default FREAD run yields no prediction, FREAD is run once more with a looser ESSS cut-off of 0. The latter generally results in higher coverage at the expense of accuracy. As a "worst case accuracy" comparison, we also built MODELLER models and compared the accuracy of the corresponding loop coordinates. All models were superimposed onto the target's native X-ray structure using only those atoms present in MEDELLER's "core" model. Then, the global RMSD was calculated for the loop coordinates without resuperimposing them.

Table II shows a summary of the test results (a direct comparison between Membrane FREAD and MODELLER is shown in Supporting Information, Fig. S7). Methods created with SPs in mind (MODELLER and Soluble FREAD) achieved average global RMSDs above 6 Å. Membrane FREAD, using the high-accuracy protocol, achieves an average accuracy of about 3 Å at a coverage of 102/156 (without nonredundancy filtering). This coverage is higher than that of soluble FREAD (79/156). MODELLER always gives complete coverage as it is an *ab initio* method.

These values clearly show that, in homology modeling, the choice of fragment database may be even more important than when modeling loops based on native anchor coordinates, as errors in the anchor structures greatly affect the accuracy of the loop prediction. Good average accuracies can currently only be achieved when using a database of MP fragments to predict MP loops.

Table II

Accuracy and Coverage on Homology Models

	MODELLER	Soluble FREAD	Membrane FREAD (high accuracy)	Membrane FREAD (high coverage)
Average RMSD	6.65	6.55 (6.43)	2.93 (6.20)	4.63 (6.59)
Loops predicted	156 (100%)	79 (51%)	102 (65%)	150 (96%)
Wins/losses vs. MODELLER	–	41/38	89/13	109/41
Wins/losses vs. Soluble FREAD	38/41	–	49/9	63/16
Wins/losses vs. Membrane FREAD (high accuracy)	13/89	9/49	–	0/0
Wins/losses vs. Membrane FREAD (high coverage)	41/109	16/63	0/0	–

Accuracy and coverage on homology models. Several loop modelling tools (columns) were run on 156 loops from 59 homology models of varying accuracy built with MEDELLER. The average backbone RMSD for the core models, used as input to the various loop modelling methods, was 2.2 Å, with a range of 0.6–5.4 Å. Soluble and Membrane refer to the database type used. "Membrane FREAD (high accuracy)" refers to the default FREAD parameters, which are used in the MEDELLER "high accuracy" loop modelling strategy. "Membrane FREAD (high coverage)" refers to MEDELLERs "high coverage" loop modelling strategy, where FREAD is first run with default parameters and, if no suitable fragments are found for a particular loop, FREAD is run again with a looser ESSS cut-off of 0. RMSD values are averaged over all loops where the particular method gave a prediction. For easy comparison, the value in square brackets refers to MODELLERs accuracy over the same set of loops.

Software availability

The new version of the FREAD loop modeling software (PyFREAD) used in this manuscript is available as a web server at http://medeller.info/fread. A command-line version of the program is available on request.

DISCUSSION

In this work, we have used FREAD, a fragment-based loop prediction method, which has been shown to produce accurate predictions of soluble protein loop structure.[17] We have demonstrated that MP loop structures can be predicted just as accurately (about 1 Å on average) using a database of MP fragments. Conversely, predictions using fragments of SPs gave consistently lower accuracies.

We show that the soluble database does contain shapes similar to MP loops (Fig. 2) of lengths up to nine residues, although the average best achievable RMSD increases with loop length. However, even though relatively similar shapes exist in the SP database, the current FREAD algorithm is unable to identify them consistently using anchor RMSD and environment-specific substitution score. The parameters used by FREAD only work as an effective filter when searching a database of the same protein type as the query. It is thus important to choose the appropriate fragment database for the prediction problem at hand if accurate loop models are to be achieved.

In their recent work on prediction of MP structure from sequence, Nugent and Jones[42] used a fragment assembly method (FILM3) to satisfy putative residue–residue contacts (predicted by precise structural contact prediction using sparse inverse covariance estimation (PSICOV)), achieving TM scores above 0.5 for 25 out of 28 target proteins. In their paper, the authors discuss the difficulty of obtaining accurate loop models with their current protocol, which uses exclusively fragments of SPs identified by local sequence similarity, among other measures. As we have shown in this work, local sequence similarity is a poor discriminator for similar loop shape when using a database of soluble protein fragments to predict MP loop structure. In addition, the authors mention reduced accuracy of contact predictions in loop regions, which further reduce loop modeling accuracy. Finally, their use of MODELLER as a refinement function may distort the entire model[37] and especially the loop regions (Table II). Judging from our own observations, better accuracies should be achievable by using fragments of MPs to model the loop regions.

In our search for the reasons for FREAD's failure to predict membrane loops using soluble fragments, we observed a decline in coverage with increasing loop length (Fig. 1, Table I). However, this decline was less obvious when using the "correct" database. Furthermore, no such behavior is observed when sequence is completely disregarded (always 100% coverage) and the best-matching fragment is selected using prior knowledge of the correct structure rather than using a sequence-based score. This indicates a lack of soluble loops with both a similar anchor structure and a similar loop sequence to membrane loops.

Amino acid substitution patterns did not appear to differ sufficiently between MPs and SPs to affect prediction accuracy. We showed this by replacing FREAD's substitution tables with ones built from either a set of SPs or MPs. We applied the SP substitution tables to predict the structures of both SPs and MPs and obtained similarly accurate results in both cases (as long as the correct database was used). Exchanging the tables for MP-based ones had no obvious effect. It is possible that redefining the structural environments used to build the FREAD substitution tables could produce different results. ESSTs built with other structure environments have been shown to improve MP alignment.[39,43]

When using the appropriate database, FREAD achieves prediction accuracies of around 1 Å (Fig. 1). Predictions of MPs yield about 60–80% coverage when the entire database is used. Even when filtering out sequence homologues (>40% sequence identity or annotated as pertaining to the same SCOP superfamily), the FREAD algorithm gives highly accurate predictions (about 1–2 Å), thereby demonstrating that the FREAD approach is indeed valid and its success not simply due to the presence of homologues in the database.

The removal of homologues, however, means a reduction in coverage, for SPs and especially for MPs. In addition, the effect of removing homologues is greater at longer loop lengths, resulting in slightly higher RMSDs for long loops compared to short ones. The reduction in coverage reflects the fact that currently known MP structures are clustered into dense groups that are highly similar within-cluster but very different across-cluster. In other words, the effective number of currently known MP structures is low. This is likely to change as the number of known protein structures grows. Nevertheless, it is possible that some loops are unique to a single protein family and can only be predicted using homologous protein structures or *ab initio* methods.

When the FREAD algorithm was first invented,[23] the number of available SP structures was low. This resulted in low coverage of database methods, making it necessary to combine them with *ab initio* loop modeling methods. Ten years later, FREAD's fragment-based approach was shown to have far higher coverage and accuracy.[17] The results presented here suggest that the same may happen with regards to the prediction of MP loops.

We performed further investigations into the differences between the shapes of MP and SP loops. Our results indicate that short MP loops tend to "stick out" away from the rest of the protein, rather than lying flat against it. It should be possible to engineer a statistical scoring scheme that can be used to identify fragments of SPs that mimic the shape of MP loops.

ACKNOWLEDGMENT

The authors would like to thank the members of the Oxford Protein Informatics Group for software testing and feedback.

REFERENCES

1. von Heijne G. The membrane protein universe: what's out there and why bother? Intern Med 2007;261:543–557.
2. Forrest L, Tang C, Honig B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 2006;91:508–517.
3. von Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 1986;5:3021–3027.
4. von Heijne G. Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992;225:487–494.
5. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003;12:963–972.
6. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ring-like structures. Biopolymers 1987;26:2053–2085.
7. Hurst T. Flexible 3D searching: the directed tweak technique. J Chem Inf Comput Sci 1994;34:190–196.
8. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins 2010;78:3428–3436.
9. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
10. Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65:438–452.
11. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
12. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model. Proteins 2003;51:21–40.
13. Fiser A, Do RKG, Šali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
14. Abagyan R, Totrov M, Kuznetsov D. ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 1994;15:488–506.
15. Abagyan RA, Totrov M. Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. J Comput Phys 1999;151:402–421.
16. Arnautova YA, Abagyan RA, Totrov M. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. Proteins 2011;79:477–498.
17. Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. Proteins 2010;78:1431–1440.
18. Zhao S, Zhu K, Li J, Friesner RA. Progress in super long loop prediction. Proteins 2011;79:2920–2935.
19. Michalsky E, Goede A, Preissner R. Loops in proteins (lip)—a comprehensive loop database for homology modelling. Protein Eng 2003;16:979–985.
20. Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 2006;34:2085–2097.
21. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics 2007;23:2836–2842.
22. Hildebrand PW, Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, Preissner R. Superlooper—a prediction server for the modeling of loops in globular and membrane proteins. Nucleic Acids Res 2009;37:W571–W574.
23. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
24. Lessel U, Schomburg D. Importance of anchor group positioning in protein loop prediction. Proteins 1999;37:56–64.
25. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
26. Fernandez-Fuentes N, Fiser A. Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol 2006;6:15.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Nucleic Acids Res 2000;28:235–242.
28. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO 1986;5:819–822.
29. Choi Y, Deane CM. Predicting antibody complementarity determining region structures without classification. Mol BioSyst 2011;7:3327–3334.
30. Tusnády GE, Dosztányi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33:D275–D278.
31. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: orientations of proteins in membranes database. Bioinformatics 2006;22:623–625.
32. Scott KA, Bond PJ, Ivetac A, Chetwynd AP, Khalid S, Sansom MSP. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. Structure 2008;16:621–630.
33. Wang G, Dunbrack R. Pisces: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
34. Kelm S, Shi J, Deane CM. iMembrane: homology-based membrane-insertion of proteins. Bioinformatics 2009;25:1086–1088.
35. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. JOY: protein sequence-structure representation and analysis. Bioinformatics 1998;14:617–623.
36. Sokal R, Michener C. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 1958;38:1409–1438.
37. Kelm S, Shi J, Deane CM. MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 2010;26:2833–2840.
38. Bugalho M, Oliveira A. Constant time clash detection in protein folding. J Bioinf Comput Biol 2009;7:55–74.
39. Hill JR, Kelm S, Shi J, Deane CM. Environment specific substitution tables improve membrane protein alignment. Bioinformatics 2011;27:i15–i23.
40. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
41. Goldfeld DA, Zhu K, Beuming T, Friesner RA. Loop prediction for a GPCR homology model: algorithms and results. Proteins 2013;81:214–228.
42. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences 2012;109:E1540–E1547.
43. Hill JR, Deane CM. MP-T: improving membrane protein alignment for structure prediction. Bioinformatics 2012;29:54–61.