BIOINFORMATICS

Vol. 00 no. 00 2003
Pages 1–7

Assigning Transmembrane Segments to Helices in Intermediate-Resolution Structures

Angela Enosh [a,*], *Sarel J. Fleishman* [b], *Nir Ben-Tal* [b] *and Dan Halperin* [a]

[a]*School of Computer Science,* [b]*Department of Biochemistry, Tel Aviv University, Ramat Aviv, 69978, Israel*

ABSTRACT

Motivation: Transmembrane (TM) proteins that form α-helix bundles constitute approximately 50% of contemporary drug targets. Yet, it is difficult to determine their high-resolution ($<$ 4Å) structures. Some TM proteins yield more easily to structure determination using cryo electron microscopy (cryo-EM), though this technique most often results in lower resolution structures, precluding an unambiguous assignment of TM amino-acid sequences to the helices seen in the structure. We present computational tools for assigning the TM segments in the protein's sequence to the helices seen in cryo-EM structures.

Results: The method examines all feasible TM helix assignments and ranks each one based on a score function that was derived from loops in the structures of soluble α-helix bundles. A set of the most likely assignments is then suggested. We tested the method on eight TM chains of known structures such as bacteriorhodopsin and the lactose permease. Our results indicate that many assignments can be rejected at the outset, since they involve the connection of pairs of remotely placed TM helices. The correct assignment received a high score, and was ranked highly among the remaining assignments. For example, in the lactose permease, which contains 12 TM helices, most of which are connected by short loops, only 12 out of 479 million assignments were found to be feasible, and the native one was ranked first.

Availability: The program and the non-redundant set of protein structures used here are available at:
`http://www.cs.tau.ac.il/~angela`

Contact: angela@post.tau.ac.il

1 INTRODUCTION

In recent years, the pace of structure determination of TM proteins has increased, but technical problems related to protein purification and crystallization still hamper TM protein structure determination. Thus, notwithstanding their biomedical importance, less than 30 distinct folds of TM proteins have been solved to date by high-resolution methods such as X-ray crystallography.

Eukaryotic TM proteins form predominantly α-helix bundles in the membrane. These proteins are composed of TM helices and loops, which are typically located on the internal or external sides of the membrane, and connect pairs of consecutive helices. Structure prediction in this class of proteins often relies conceptually on the two-stage model for their assembly in the membrane (Popot and Engelman, 1990). According to this model, TM protein folding begins with the insertion of the TM segments into the membrane as α-helices. In the second stage these helices assemble to form a bundle (reviewed in Popot and Engelman, 2000; White and Wimley, 1999).

Some of the factors stabilizing TM protein structures have been elucidated in recent years on the basis of solved structures and biochemical experiments (e.g., Choma *et al*., 2000; Eilers *et al*., 2000; MacKenzie and Engelman, 1998; Russ and Engelman, 2000). A number of computational methods have been suggested for positioning and orienting the helices comprising the TM domain with respect to one another (e.g., Adams *et al*., 1995; Fleishman and Ben-Tal, 2002; Kim *et al*., 2003; Pellegrini-Calace *et al*., 2003).

Here, we consider a situation in which the locations of the TM helices in $3D$-space can be deduced experimentally. The challenge is then to assign the TM segments in the protein sequence into the corresponding helices in $3D$-space. For concreteness, let us focus on proteins that were solved at intermediate in-plane resolution ($5-10$Å) (Unger, 2001). From these data, one can derive helix positions, as well as their tilt and azimuthal angles with respect to the membrane. However, the individual amino acids cannot be identified, so that the correspondence between the TM segments and the cryo-EM helices cannot be decided unambiguously. So far, no method has tackled this problem.

Providing a solution to the helix-assignment problem is a first step toward modeling of TM proteins. That is, by assigning the TM segments to the helices in the cryo-EM data, conformation space in a modeling exercise can be limited substantially. In addition, helix assignment is directly useful for

*To whom correspondence should be addressed.

structural studies of membrane proteins, as it reveals which helices are in contact with each other, and outlines helices that are located in critical positions, such as around a pore in channels and pumps.

We show here that many putative helix assignments can be eliminated based on the (estimated) maximal lengths of each of the loops in the protein. In addition, we present a novel score function, that was derived on the basis of conformations of loops in α-helix bundles (of soluble proteins), in order to rate the capability of loops to connect each pair of helices. Based on this score function, we ranked assignments of 8 TM-protein chains of known structures taken from the Protein Data Bank (`http://www.rcsb.org/pdb/`), and our results show that the native-state assignment ranks high in many cases.

Terminology and Formal Statement of the Problem. The sequence of a TM protein of the α-helix bundle type, denoted by S, is composed of TM and extra-membrane segments, which connect TM segments that are consecutive in the sequence (Figure 1(a)). The locations of TM segments in protein sequences can be predicted fairly precisely on the basis of sequence data alone (Chen *et al.*, 2002). We denote a TM segment, $T_i \in S$, by $T_i = \{t_{i1}, t_{i2} \dots t_{ik_i}\}$, as an ordered sequence of amino acids from the N- to the C-terminus. Similarly, we denote an extra-membrane segment, $X_i \in S$, by $X_i = \{x_{i1}, x_{i2} \dots x_{ik_i}\}$, as an ordered sequence of amino acids from the N- to the C-terminus. The length of an extra-membrane segment X_i, denoted by $\mathrm{length}(X_i)$, is the number of amino acids in the segment. The maximal distance between two points that can be connected by X_i is denoted by $\mathrm{max\_dist}(X_i) = (\mathrm{length}(X_i) + 1) \times \mathrm{dist}(C\alpha, C\alpha)$, where $\mathrm{dist}(C\alpha, C\alpha)$ is the distance between two consecutive $C\alpha$ atoms, which is typically taken as 3.8Å (Creighton, 1993).

A Helix, $C_i \in C$. Positions, tilt and azimuthal angles of each helix can be extracted from intermediate-resolution cryo-EM maps (Unger, 2001). Canonical α-helices are constructed, and made to fit the cryo-EM map. We represent each such helix by a sequence of coordinates of its $C\alpha$ atoms, $C_i = \{c_{i1}, c_{i2} \dots c_{ik_i}\}$. The membrane can be regarded as a region in $3D$ bounded by two planes, to which we refer as the inner and the outer planes of the membrane. We define an order on a helix C_i in the sense that c_{i1} is the closest atom to the inner plane of the membrane, and c_{ik_i} is the closest atom to the outer plane of the membrane. We denote the internal $C\alpha$ atom by $\mathrm{internal}(C_i) = c_{i1}$, and the external $C\alpha$ atom by $\mathrm{external}(C_i) = c_{ik_i}$.

It should be noted that the positions of helices deduced from cryo-EM in this manner suffer from imprecision. First, the orientation of the helices around their principal axes cannot be derived from the cryo-EM map due to the limited in-plane resolution (typically, $5 - 10$Å (Unger, 2001)). Moreover, the low resolution along the axis normal to the membrane plane ($12 - 30$Å) entails a large distortion in the positions of helices along this axis. For simplicity we avoid dealing with these inaccuracies in the description of our algorithm. However, as described in Appendix A, our program takes the noisiness that results from the limited resolution into account by also testing helix positions that are in the vicinity of those seen in the cryo-EM data.

Formal Definition of Our Goals. Given the secondary structure classification of a TM protein sequence $S = \{T_1, X_1, T_2, \dots X_{n-1}, T_n\}$ and a set of helix locations in $3D$-space $C = \{C_1, C_2, \dots C_n\}$, derived from the cryo-EM map, *(i)* find all the *feasible* assignments between the $T_i's$ and the $C_i's$, namely find a permutation σ such that for each $1 \le i \le n$, T_i is assigned to $C_{\sigma(i)}$, and *(ii)* attribute a score to each assignment based on its compatibility with the locations of the helices in $3D$-space.

In principle, a TM segment can be assigned to a helix in $3D$-space with its N- and C-termini on the inner and outer sides of the membrane, respectively, or vice versa. However, it is possible to resolve this ambiguity experimentally. Hence, the number of all the assignments is $n!$. A brute-force approach would require the generation of all these assignments. To reduce this immense computational burden, at the outset we exploit the maximal lengths of the extra-membrane segments to filter out impossible assignments. Suppose we want to match two consecutive segments T_i and T_{i+1} to the helices, C_k and C_m, correspondingly, such that the extra-membrane segment X_i lies on the external side of the membrane. A necessary condition for this assignment to be valid is that the maximal length of the extra-membrane segment ($\mathrm{max\_dist}(X_i)$) is longer than the distance between $\mathrm{external}(C_k)$ and $\mathrm{external}(C_m)$). In the same manner, if X_i should connect the helices on the internal side of the membrane, its maximal length should be larger than the distance between $\mathrm{internal}(C_k)$ and $\mathrm{internal}(C_m)$. Consequently, if this condition does not hold, the assignment should be ignored from the outset.

2 THE ALGORITHM

Our algorithm proceeds in two stages: *Pruning by Distance Constraints* — construction of an assignment graph that contains *only* the set of feasible assignments, i.e., assignments in which the maximal lengths of the extra-membrane segments are longer than the distances between the helices that they connect (Figure 1). This stage is followed by *Loop Conformation Scoring* — attributing scores to the feasible assignments based on their compatibility with the locations of the helices in $3D$-space.

2.1 Pruning by Distance Constraints

We wish to filter out as many assignments as possible, without eliminating the right one. For this purpose we construct a directed acyclic graph $G(V, E_{int} \bigcup E_{ext})$, such as the one in Figure 1(c), where:

$$V = \{(T_i, C_j) \mid 1 \le i, j \le n\},$$
$$E_{int} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \mathrm{dist}(\mathrm{internal}(C_j), \mathrm{internal}(C_m)) \le \mathrm{max\_dist}(X_i)\},$$
$$E_{ext} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \mathrm{dist}(\mathrm{external}(C_j), \mathrm{external}(C_m)) \le \mathrm{max\_dist}(X_i)\}$$

V stands for the vertices and E stands for the edges in G. There are two kinds of edges in G: external (E_{ext}) and internal (E_{int}). There is an edge $e \in E_{ext}$, if and only if the two consecutive TM segments T_i and T_{i+1} can be matched congruently to C_j and C_m. Namely, the extra-membrane segment X_i between T_i and T_{i+1} is sufficiently long to connect the two points $\mathrm{external}(C_j)$ and $\mathrm{external}(C_m)$ on the external side of the membrane. The same applies to the E_{int} edges where X_i is sufficiently long to connect $\mathrm{internal}(C_j)$ and $\mathrm{internal}(C_m)$ on the internal side of the membrane.

We construct G in a bottom-up fashion, i.e., the levels in G are constructed from the nth to the 1st level (where n is the number of TM segments in the protein). The kth level in the graph consists of vertices composed of T_k, namely $\{(T_k, C_j) \mid 1 \le j \le n\}$. Given the set of nodes $\{(T_k, C_j) \mid 1 \le j \le n\}$ in the kth level, we construct the $(k-1)$st level as follows. For each vertex (T_k, C_j) we go over all the helices $C_t \in C \setminus \{C_j\}$ and if X_{k-1} can connect the two helices C_t and C_j on the external or internal side of the membrane, we add the vertex (T_{k-1}, C_t) (if it is still missing) to the $(k-1)$st level, and a directed edge $e = ((T_{k-1}, C_t), (T_k, C_j))$, where $e \in E_{ext}$ or $e \in E_{int}$. Thus, a directed edge $e \in \{E_{ext} \bigcup E_{int}\}$ can appear only between two consecutive levels. At the beginning, all of the vertices (T_n, C_j) in the nth level are examined against the pairs (T_{n-1}, C_t) where $C_t \in C \setminus \{C_j\}$, and created if and only if the above condition holds. After construction of the graph G we can eliminate all of the nodes between the second to the nth level that do not have at least one entering edge.

A path $\pi = \{v_1, e_1, v_2, e_2, v_3 \ldots e_{n-1}, v_n\}$ in the graph G is considered valid if it starts at the first level of G, ends at the nth level of G, and it is comprised of an alternating sequence of external and internal edges (either $\{e_k \mid k \text{ even}\}$ are external and $\{e_k \mid k \text{ odd}\}$ are internal, or vice versa). In addition, we require that π does not contain two vertices with the same helix (the C_k's in all the vertices $v_i = (T_i, C_k)$ are distinct). Each valid path π defines a feasible assignment between the TM segments of S and the helices in C. It will be shown that this pruning phase eliminates many infeasible assignments when the protein contains short loops (namely, loops whose lengths are less than 6).

2.2 Ranking the Feasible Assignments

In the following stage, a score is assigned to the feasible assignments that are stored in G based on the suitability of the loops to connect helices in the structure. Each feasible assignment is a permutation σ^k which assigns the TM segments $T_1 \ldots T_n$ to the helices $C_{\sigma^k(1)} \ldots C_{\sigma^k(n)}$, where $1 \le k \le n!$. We define the score function F of a permutation σ^k as follows:

$$F(\sigma^k) = \sum_{i=1}^{n-1} f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)})$$

where f scores the suitability of assigning the consecutive TM segments T_i and T_{i+1} to helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$. Namely, f defines the feasibility of connecting the two helices in $3D$-space by X_i.

The problem of adjusting an extra-membrane segment to connect two fixed secondary structures is related to the well-known kinematics problem of loop-closure (Canutescu and Dunbrack, 2003; Manocha and Zhu, 1994; Wedemeyer and Scheraga, 1999; Wojcik *et al.*, 1999; Xiang *et al.*, 2002). However, our problem is slightly different. We wish to rank the assignments instead of predicting the conformation of the extra-membrane loops as in the classic loop-closure problem, since the native extra-membrane segment, which connects the two helices, is unknown. Hence, we seek to define a *score* for matching the extra-membrane segment to connect the two helices in a way that the native match is assigned the highest score.

The evaluation of f is based on the length of the extra-membrane segment X_i and on a statistical analysis we have conducted on solved structures of soluble

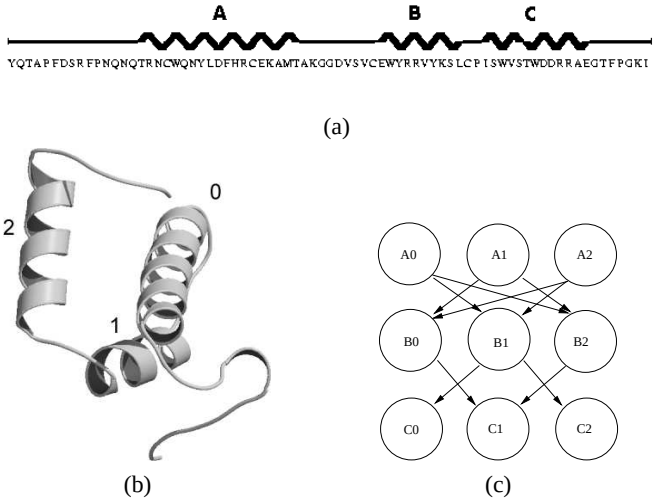

Fig. 1. (a) The locations of the three TM segments in the sequence of chain H of the cytochrome c oxidase. (b) The corresponding $3D$ structure. (c) The assignment graph of this chain. The numbers represent the helices and the letters represent the TM segments. There are four valid paths (feasible assignments) in the graph which are: $(A0, B1, C2)$, $(A0, B2, C1)$, $(A2, B0, C1)$ and $(A2, B1, C0)$. Notice that there is no edge between $(B0)$ and $(C2)$, for example, since the loop between the TM segments B and C is too short to connect helices 0 and 2.

proteins taken from the Protein Sequence Culling Server (http://www.fccc.edu/research/labs/dunbrack/pisces/) in a preprocessing phase. We restricted our survey to protein sections comprised of two consecutive helices with a loop region between them, namely to *helix-loop-helix* motifs, where secondary-structure elements are assigned according to DSSP (Kabsch and Sander, 1983).

The preprocessing phase. We denote the two consecutive helices in a helix-loop-helix motif, by A and B, and the loop region which connects them by L, and set $l = \text{length}(L)$. Let us examine the helix-loop-helix motifs with the same loop length l ($2 \leq l \leq 7$). All of these motifs (A, L, B) were placed in a common orthogonal reference frame, so that the helices A of all of the motifs overlap. Transforming these motifs to the common reference frame yields a set of points in $3D$-space that represents the starting points of the second helices (i.e., B's) relative to the common first helix (i.e., the overlapping A's).

All of these starting points, denoted by p_i ($1 \leq i \leq N$, where N is the number of helix-loop-helix motifs), were stored in a KD-tree data structure [1]. Since the lengths of the loops in these motifs have a great impact on the locations of the points, p_i's, in $3D$-space, these points were stored in 6 distinct KD-trees which we denote by KD_l, $2 \leq l \leq 7$, one tree per length l. Our results indicate that these points are distributed non-uniformly in $3D$-space. For an illustration, Figure 2 shows the starting points in the common reference frame for $l = 3$ and $l = 4$.

The scoring phase. We compute $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)})$ as follows. We place the two helices $C_{\sigma^k(i)}$ and $C_{\sigma^k(i+1)}$ in the common orthogonal reference frame in the same manner as we have done in the preprocessing phase, and obtain the new starting point q of the helix $C_{\sigma^k(i+1)}$. Given q and the starting points of helix-loop-helix motifs with loop length $x = \text{length}(X_i)$ from the preprocessing phase, the score depends on two criteria: the number of neighboring points in the vicinity of q and the distances between these neighboring points and q.

Let Q be a cube centered at q with side size $(10 \cdot x)$Å. We query KD_x to find the points that were stored in the preprocessing phase which occur in Q. Q represents the region in $3D$-space for the clusters of points in the appropriate KD-tree we wish to examine. The score for this assignment is based on the sum of the distances between q and the derived points that were found inside Q. The score was constructed with the aim of favoring loops that have been observed many times in the protein database we have used. It is, therefore, defined in the form of a colony function (Xiang *et al.*, 2002), whereby loops in the database that are similar to the query make a more significant contribution to the loop's score. Formally, $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)}) = \sum_{r \in Q} e^{-\text{dist}(q,r)}$.

When $x \geq 8$, we do not obtain significant information about the quality of the assignment due to the low frequency of occurrence of long loops in the helix-loop-helix motif in the specified protein database. Thus, for $\text{length}(X_i) \geq 8$, we have set $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)}) = 0$.

Given the assignment graph G that was generated in the pruning phase, we assign a weight, $\text{weight}(e) = f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)})$ to each edge in the graph, namely to each $e = (u, v)$ where $u = (T_i, C_{\sigma(i)})$ and $v = (T_{i+1}, C_{\sigma(i+1)})$. G is an acyclic directed weighted graph. Each valid path in G defines a feasible assignment, and its score is the sum of the weights of the edges in the path, i.e., $F(\pi) = \sum_{e \in \pi} \text{weight}(e)$.

3 THE DISTRIBUTION OF END POINTS OF SHORT LOOPS IS HIGHLY NONUNIFORM

Structures of helix-loop-helix motifs (resolution of 2Å or less, and R-factor of 0.3 or less) of soluble proteins were selected from the Protein Sequence Culling Server (http://www.fccc.edu/research/labs/dunbrack/pisces/). To reduce the bias inherent in the Protein Data Bank, only proteins whose sequences were less than 20% identical were selected. The secondary structures were assigned by DSSP (Kabsch and Sander, 1983). We looked only at helix-loop-helix motifs containing two helical regions of at least 8 amino acids each, which are connected by loops of lengths 2 to 7 amino acids (Table 1). The order of the two helices was specified from the N- to the C-terminus. Entries were classified by the loop lengths. Each loop of length l (where $2 \leq l \leq 7$) contributed to our analysis a point in $3D$-space corresponding to the beginning of helix B. The distribution of the examined points in the common reference frame for short loops (i.e., lengths three and four) is shown in Figure 2. Loops longer than seven were not considered, due to their low frequency of occurrence in our dataset.

Table 1. Helix-loop-helix motifs classified by loop length

Loop Length	2	3	4	5	6	7
Number of motifs	456	260	171	167	98	36

Helix-loop-helix motifs derived from the Protein Sequence Culling Server and classified by their loop lengths.

The scoring function is greatly dependent on this protein database analysis. To understand why our scoring function performs well (as indicated by the results reported below), consider for example the case where $l = 4$ (Figures 2(d-f)), i.e., the loop L has four $C\alpha$'s. In this case L has 8 degrees of freedom (each $C\alpha$ contributes two degrees of freedom ϕ and

[1] KD-trees are orthogonal range-search structures. They are used to store a set P of points in R^d so that the subset of P inside a query axis-aligned hyperbox can be reported efficiently. See, e.g., (de Berg *et al.*, 2000) for details.

ψ). By sheer kinematics considerations, if we fix one end of the loop, the reachable space by the other end (we refer to it as the *free end*) is large, practically limited only by the stretch of the loop (the conformation that has the largest diameter). However, Figures 2(d-f) show that the locus of the free end in length-four loops connecting two helices is limited to a few clusters of points in $3D$-space. Our scoring takes advantage of this phenomenon, which is highly significant in loops of lengths two through five, but is still substantially noticeable in loops of lengths up to seven.

4 IMPLEMENTATION

We verified the scoring function by applying it to 8 TM protein chains, whose structures were solved using X-ray crystallography (Table 2). We restricted our study to those chains, whose TM segments did not contain half-helices or loops (except for the glycerol facilitator, as discussed below). Moreover, we did not consider proteins that contain long extra-membrane segments that could form large domains. It should be noted that the results reported below were derived solely from the solved structures of TM proteins.

The algorithm has been implemented for two distinct cases: (i) using accurate data of the locations of helices as derived from the Protein Data Bank; and (ii) using noisy data, i.e., uncertainty with regard to the positions of the helices. In case (i) the algorithm assumes that the helices are located and oriented in their native conformations. In case (ii), the algorithm assumes that the orientations and locations of the helices are known only approximately. However, in real cases, thanks to the cryo-EM data, we will know that the native helices are located in bounded regions. Therefore, we examine all of the possible orientations and locations of the helices in these bounded regions. The exact definition of these regions is provided in Appendix A.

The two implemented cases (using accurate and noisy data) are examined in Table 2 by the number of feasible assignments that remain after the pruning phase and by the rank of the score of the native assignment with respect to other assignments. In most of the examined TM proteins, the table shows that the native assignment ranks very highly, which implies that the combination of the pruning and scoring phases yields a reliable tool for assigning TM segments to helices.

For example, bacteriorhodopsin (1c3w) is composed of 7 helices, and thus has $7! = 5040$ possible assignments. The number of feasible assignments that remained after the pruning phase is 44. Applying the score function and sorting all of the 44 assignments by their scores, the native assignment was ranked third. When using the noisy data, the list of feasible assignments expanded, but the rank of the native state (13) did not change dramatically, which implies that our score function deals well with this level of noise.

The strength of the pruning phase is clearly shown for the lactose permease (1pv6), where out of 479 million possible assignments, the number of feasible assignments in both cases (i, ii) was below 13 and the running time was relatively short since the assignment graph ruled out many assignments which were not examined. Our method yielded poor results for the glycerol facilitator (1fx8) due to a 24 residue loop which contains a half TM helix. It is rather encouraging that even in this pathological case the algorithm removed approximately half of the potential assignments (352 out of 720) and ranked the native state to be 119.

5 DISCUSSION

A novel method for assigning TM spans in the sequence of an integral membrane protein to the approximate locations of the helices in $3D$-space was presented here. Each of the possible assignments is evaluated based on the compatibility of the extra-membrane segments with the suggested relative locations of the helices. Our results show that in TM proteins with extra-membrane segments of 7 residues or less, the vast majority of the putative assignments can be rejected from the outset, since they involve the connection by short loops of pairs of TM helices that are spatially distant from each other. In the lactose permease, for example, only 12 out of 479 million putative assignments were found to be feasible based on this criterion. The significant reduction in the number of assignments is due to the short lengths of the extra-membrane segments. It demonstrates that, in practice, the complexity of the TM helix assignment problem scales with the lengths of these segments rather than with the number of TM helices.

The feasible assignments are then screened based on the suitability of each of the extra-membrane segments to adopt a conformation that could connect the adjacent TM helices. This is done using a novel knowledge-based score function that was derived from the conformations of loops in helix-loop-helix motifs. Our results show that this function ranks the TM helix assignment of the native structure high among the other feasible assignments. This is best demonstrated with chain H of the cytochrome c oxidase, where the native structure ranks first among the feasible assignments.

In the typical case, the locations of the TM helices in $3D$-space will be determined using medium-resolution data, e.g., from cryo-EM studies at in-plane resolutions of $5 - 10$Å. At such resolution, one can only derive the approximate locations of the TM helices in $3D$-space. The results demonstrate that the method is robust to changes in the locations of the TM helices; the native-state assignment ranks high among the feasible assignments, even when using noisy data (Table 2).

Our results are very encouraging in that the problem of TM helix-assignment is significantly reduced, and yet in the typical case, the analysis is likely to result in several putative assignments rather than only one. We anticipate that the set of potential assignments may be further reduced based on available empirical data, e.g., from biochemical, molecular and genetic studies. Finally, forward-looking experiments

Table 2. The performance of the two-stage (pruning and scoring) algorithm using accurate and noisy data

Name	PDB	Loop Lengths	n_h	$n_{\rm pos}$	(i) Accurate		(ii) Noisy	
					$n_{\rm feas}$	rank	$n_{\rm feas}$	rank
Bacteriorhodopsin	1c3w	3,14,2,3,10,4	7	5040	44	3	948	13
Sensory rhodopsin	1h68	7,12,2,3,3,4	7	5040	84	2	512	48
Cytochrome c oxidase	1occC	3,5,19,2,7,7	7	5040	74	7	335	62
Cytochrome c oxidase	1occE	5,6,1,1	5	120	2	2	2	1
Cytochrome c oxidase	1occH	7,2	3	6	4	1	6	1
Glycerol facilitator	1fx8	6,19,24,8,4	6	720	236	8	352	119
Halorhodopsin	1e12	2,20,2,4,1,5	7	5040	34	5	73	22
Lactose permease	1pv6	3,2,1,3,1,24,3,1,3,1,1	12	$> 10^8$	7	3	12	1

Classification and comparison of the results using (i) accurate helix positions derived from the PDB and (ii) noisy data. The set of TM proteins of known $3D$ structures that were studied are indicated by their names and pdb entries. The subunit is indicated by the last letter. The proteins are classified by the number of TM helices (n_h), their loop lengths, and the number of possible assignments $n_{\rm pos} = n_h!$. The results are categorized by the number of feasible assignments ($n_{\rm feas}$) that remained following the pruning phase and by the position of the native assignment (rank) with respect to other feasible assignments. We ran the program on PC Intel Pentium IV, CPU 2.4GHz, 256 MB RAM, and the running time using the accurate data was below 2 seconds for each of the proteins. When using the noisy data, the running time varied between 8 seconds for 1occ subunit H and 6.5 minutes for 1pv6. Currently we are working on additional TM proteins. The results obtained from processing these cases will be available soon at `http://www.cs.tau.ac.il/~angela`.

may be designed to select the native assignment out of a few possibilities.

The application of the method to oligomeric TM proteins, such as cytochrome c oxidase may complicate the analysis. In the present study, the subunit boundaries were taken as a given, but, if these are unknown, it may be necessary to examine various molecular boundaries, which would entail an increase in the dimensionality of the problem.

To demonstrate the method's usefulness we are applying it to the assignment of the TM helices in the microsomal glutathione transferase 1 (MGST1). This protein is a member of the MAPEG (membrane-associated proteins in eicosanoid and glutathione metabolism) superfamily of TM enzymes (Jakobsson *et al*., 1999).

MGST1 is a homotrimer, in which each monomer is composed of 4 TM segments. The $3D$ structure of MGST1 was determined at an in-plane resolution of 6Å using cryo-EM (Holm *et al*., 2002; Schmidt-Krey *et al*., 2000). The electron-density map shows three repeats of 4 rod-like densities, which presumably correspond to the 12 TM helices of the homotrimer. Our preliminary results show that only a few assignments are consistent with the structure (data not shown).

ACKNOWLEDGEMENT

Work reported in this paper has been supported in part by the IST Programmes of the EU as Shared-cost RTD (FET Open) Projects under Contract No IST-2000-26473 (ECG - Effective Computational Geometry for Curves and Surfaces) and No IST-2001-39250 (MOVIE - Motion Planning in Virtual Environments), by The Israel Science Foundation founded by the Israel Academy of Sciences and Humanities (Center for Geometric Computing and its Applications), by the Hermann Minkowski – Minerva Center for Geometry at Tel Aviv University and by Nofar grant from the Israel Ministry of Trade and Industry. SJF was supported by a doctoral fellowship from the Clore Israel Foundation.

REFERENCES

Adams,P.D., Arkin,I.T., Engelman,D.M. and Brunger,A.T. (1995) Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat. Struct. Biol.*, **2**, 154-162.

de Berg,M., van Kreveld,M., Overmars,M. and Schwarzkopf,O. (2000) Computational Geometry: Algorithms and Applications, 2^{nd} Edition. Springer-Verlag, Berlin.

Canutescu,A.A. and Dunbrack,R.L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.*, **12**, 963-972.

Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774-2791.

Choma,C., Gratkowski,H., Lear,J.D. and DeGrado,W.F. (2000) Asparagine-mediated self-association of a model transmembrane helix. *Nat. Struct. Biol.*, **7**, 161-166.

Creighton,T.E. (1993) Poteins, Structures and Molecular Properties. Freeman, New York.

Eilers,M., Shekar,S.C., Shieh,T., Smith,S.O. and Fleming,P.J. (2000) Internal packing of helical membrane proteins. *Proc. Natl. Acad. Sci. USA.*, **97**, 5796-5801.

Fleishman,S.J. and Ben-Tal,N. (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.*, **321**, 363-378.

Holm,P.J., Morgenstern,R. and Hebert,H. (2002) The 3-D structure of microsomal glutathione transferase 1 at 6 A resolution as determined by electron crystallography of p22(1)2(1) crystals. *Biochim. Biophys. Acta*, **1594**, 276-285.

Jakobsson,P.J., Morgenstern,R., Mancini,J., Ford-Hutchinson,A. and Persson,B. (1999) Common structural features of MAPEG – a widespread superfamily of membrane associated proteins

with highly divergent functions in eicosanoid and glutathione metabolism. *Protein Sci.*, **8**, 689-692.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonding and geometrical features, `http://www.cmbi.kun.nl/gv/dssp/`. *Biopolymers*, **22**, 2577-2637.

Kim,S., Chamberlain,A.K. and Bowie,J.U. (2003) A simple method for modeling transmembrane helix oligomers. *J. Mol. Biol.*, **329**, 831-840.

MacKenzie,K.R. and Engelman,D.M. (1998) Structure-based prediction of the stability of transmembrane helix- helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl. Acad. Sci. USA.*, **95**, 3583-3590.

Manocha,D. and Zhu,Y. (1994) Kinematic manipulation of molecular chains subject to rigid constraints. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 285-293.

Pellegrini-Calace,M., Carotti,A. and Jones,D.T. (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins*, **50**, 537-545.

Popot,J.L. and Engelman,D.M. (1990) Membrane protein folding and oligomerization: The two-stage model. *Biochemistry*, **29**, 4031-4037.

Popot,J.L. and Engelman,D.M. (2000) Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.*, **69**, 881-922.

Russ,W.P. and Engelman,D.M. (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.*, **296**, 911-919.

Schmidt-Krey,I., Mitsuoka,K., Hirai,T., Murata,K., Cheng,Y., Fujiyoshi,Y., Morgenstern,R. and Hebert,H. (2000) The three-dimensional map of microsomal glutathione transferase 1 at 6 A resolution. *EMBO*, **19**, 6311-6316.

Unger,V.M. (2001) Electron Cryomicroscopy Methods. *Curr. Opin. Struct. Biol.*, **11**, 548-554.

White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 319-365.

Wedemeyer,W.J. and Scheraga,H.A. (1999) Exact analytical loop closure in proteins using polynomial equations. *J. Comp. Chem.*, **20**, 819-844.

Wojcik,J., Mornon,J.P. and Chomilier,J. (1999) New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol.*, **289**, 1469-1490.

Xiang,Z., Soto,C.S. and Honig,B. (2002) Evaluating conformational free energies: The colony energy and its application to the problem of protein loop prediction. *Proc. Natl. Acad. Sci.*, **99**, 7432-7437.

APPENDIX A: DEALING WITH THE UNCERTAINTY IN CRYO-EM DATA

Cryo-EM studies at $5-10$Å in-plane resolution provide only the approximate locations of the helix-axes positions and orientations. The uncertainty in $3D$-space is mainly due to two reasons (Figure 3): (i) the unknown orientation of the helix with respect to its axis; (ii) the unknown translation of the helix along its axis.

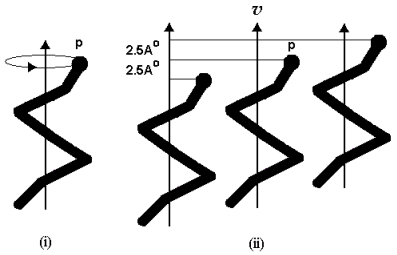

Fig. 3. The detection of helix location in cryo-EM data involves two types of uncertainties: (i) the exact orientation of the helix with respect to its axis is unknown; (ii) the helix derived from the cryo-EM (middle) may shift toward the outer plane ((ii) right) or toward the inner plane ((ii) left) of the membrane.

We now redefine the score function $f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)})$ that was introduced in Section 2.2 to suit the noisiness of the data. For simplicity, we assume that X_i should connect the two helices in the external side of the membrane. We denote by p' and q' the native positions of the external $C\alpha$ atoms of helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$ respectively. The above uncertainties may affect f dramatically, since it strongly depends on the points $p =$ external $(C_{\sigma(i)})$ and $q =$ external $(C_{\sigma(i+1)})$, whose locations are known only approximately. However, the locations of p' and q' are restricted to bounded regions as shown below.

Let us examine the surface where p' can possibly be located accounting for the imprecision in the model. We call this surface the envelope of p and denote it by $E(p)$ (the same discussion applies to q'). $E(p)$ is defined as follows (the numbering corresponds to the numbers of the reasons for imprecision): (i) p' can be located on a circle in $3D$-space centered at the helix axis (Figure $3(i)$); (ii) p' can be located in the range $[p - v \cdot 2.5, p + v \cdot 2.5]$ where v is the unit vector that coincides with the helix axis toward the external side of the membrane (Figure $3(ii)$). It follows that $E(p)$ has a cylindrical envelope shape with radius 2.5Å (typically, radius of a helix) and its height is set to 5Å.

Given p and q as specified above, each pair of points $p_k \in E(p)$ and $q_j \in E(q)$, can be regarded as the external $C\alpha$ atoms of the native helices. We pick uniformly distributed random points $p_k \in E(p)$ and $q_j \in E(q)$ and transform the helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$ so that p and q will coincide with p_k and q_j, respectively (without changing their axes' directions). The transformed helices are denoted by $T_k(C_{\sigma(i)})$ and $T_j(C_{\sigma(i+1)})$. To account for this imprecision, we modify the score function f to be: $\max_{k \in E(p), j \in E(q)} f(X_i, T_k(C_{\sigma(i)}), T_j(C_{\sigma(i+1)}))$.

It can be shown that in order to cover the envelope $E(p)$ adequately, we need to sample $n = 135$ points on $E(p)$. By adequately we mean that with high probability (> 0.98), the native point p' will be at distance $\epsilon = 1$Å at most from at least one of the samples points in $E(p)$.

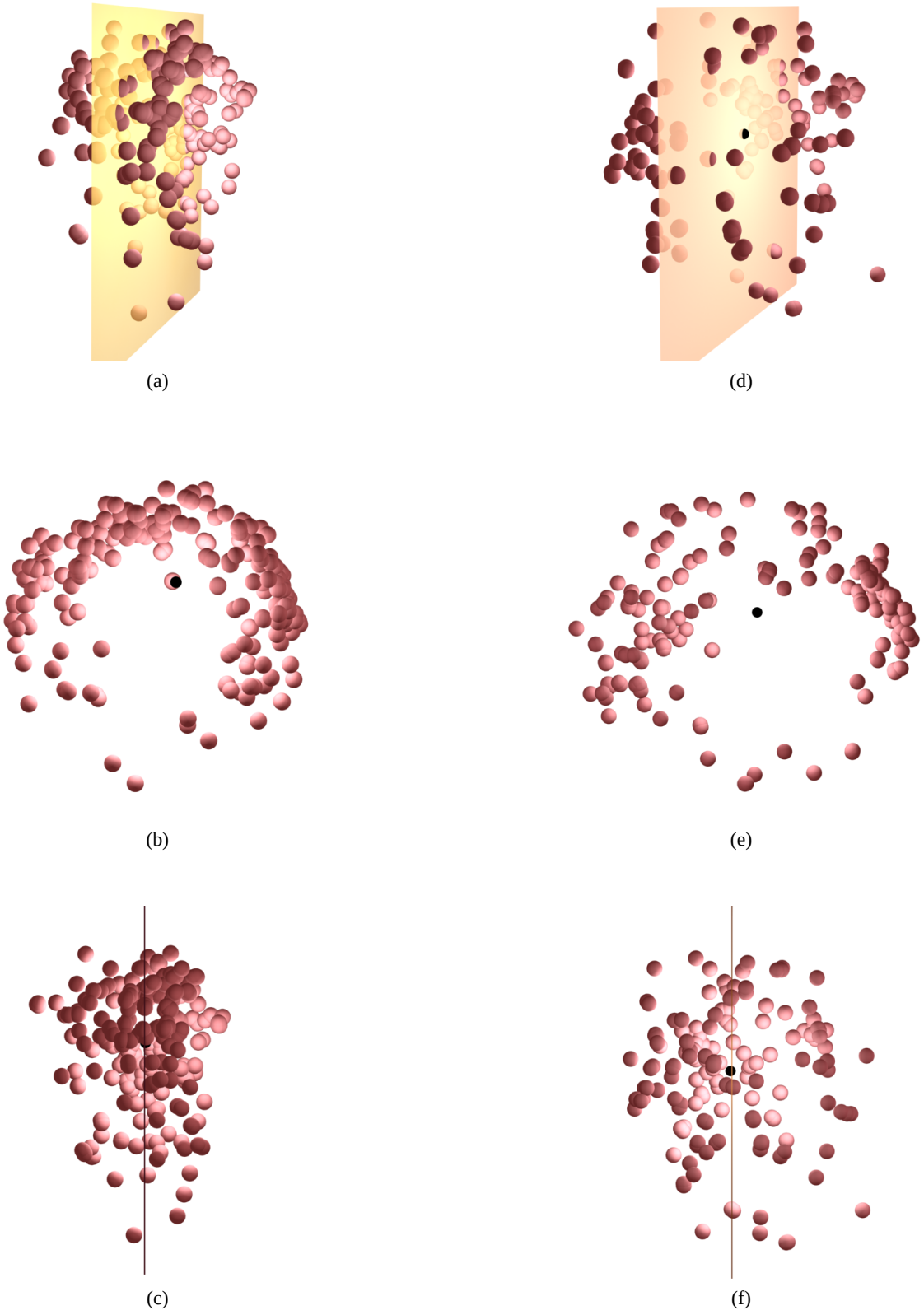

Fig. 2. The distribution of the starting points of helices B's in $3D$-space derived from the helix-loop-helix motifs (A, L, B) with loop lengths 3(a-c) on the left and 4(d-f) on the right. The black spot marks the origin of the common reference frame. Figures (a,d) display the points together with their least-mean-square (LMS) plane. The view point of figures (b,e) is the normal to the LMS plane. Figures (c,f) present a view from the side on the LMS plane. It can be seen that the starting points of 3 amino acids loops create a torus-like shape in $3D$-space.