BIOINFORMATICS

Vol. 23 ECCB 2006, pages e212–e218
doi:10.1093/bioinformatics/btl325

Prediction and simulation of motion in pairs of transmembrane α-helices

Angela Enosh[1,*], Sarel J. Fleishman[2], Nir Ben-Tal[2] and Dan Halperin[1]

[1]School of Computer Science and [2]Department of Biochemistry Tel Aviv University, Ramat Aviv 69978, Israel

ABSTRACT

Motivation: Motion in transmembrane (TM) proteins plays an essential role in a variety of biological phenomena. Thus, developing an automated method for predicting and simulating motion in this class of proteins should result in an increased level of understanding of crucial physiological mechanisms. We have developed an algorithm for predicting and simulating motion in TM proteins of the α-helix bundle type. Our method employs probabilistic motion-planning techniques to suggest possible collision-free motion paths. The resulting paths are ranked according to the quality of the van der Waals interactions between the TM helices. Our algorithm considers a wide range of degrees of freedom (dofs) involved in the motion, including external and internal moves. However, in order to handle the vast dimensionality of the problem, we employ some constraints on these dofs in a way that is unlikely to rule out the native motion of the protein. Our algorithm simulates the motion, including all the dofs, and automatically produces a movie that demonstrates it.

Results: Overexpression of the RTK ErbB2 was implicated in causing a variety of human cancers. Recently, a molecular mechanism for rotation-coupled activation of the receptor was suggested. We applied our algorithm to investigate the TM domain of this protein, and compared our results with this mechanism. A motion pathway that was similar to the proposed mechanism ranked first, and motions with partial overlap to this pathway followed in rank order. In addition, we conducted a negative-control computational-experiment using Glycophorin A. Our results confirmed the immobility of this TM protein, resulting in degenerate paths comprising native-like conformations.

Supplementary information: Supplementary data are available at http://www.cs.tau.ac.il/~angela/EGFR.html

Contact: angela@post.tau.ac.il

1 INTRODUCTION

In total, approximately 20–30% of proteins encoded by the genome are transmembrane (TM). They form pumps and channels that control and guide the transportation of ions and metabolites across the membrane. Other TM proteins function as receptors and are responsible for molecular recognition of hormones and neurotransmitters. Despite recent advances, it is extremely difficult to crystallize these proteins, and even when a high-resolution structure is determined, much effort is required to elucidate the protein's mechanism of action. So far, cartoon-resolution mechanisms have been suggested for only a few TM-proteins, e.g. the lactose permease (Abramson *et al*., 2003) and ErbB2 (Fleishman *et al*., 2002). However, molecular details for these mechanisms are not defined yet. These molecular details include, for instance, the following questions: What exactly are the conformational changes that occur in each step along the reaction coordinate? Whether, and to what extent do the helices move as rigid bodies? Which torsion angles and side-chains alter during the conformational change? Thus, one of the challenging tasks in computational studies of TM-protein structures is to define these molecular details as continuous motion that goes beyond the cartoon-level resolution published so far in order to gain insight into these mechanisms.

Proteins display a broad range of motions, from the fast and localized motions (e.g. side-chain movements) to the slow large-scale motions (e.g. domain movements). An important characteristic of biomolecules is that the different types of motion are interdependent and coupled to one another. Thus, in the investigation of slow large-scale motions as we propose to find, ignoring the fast small-scale motions might obscure the overall conformational changes.

Many large-scale motions take place on time scales beyond the accessibility of time-dependent methods, such as molecular dynamics (MD) (Karplus *et al*., 2002). Normal-mode analysis (NMA), the Gaussian Network Model (GNM) and the Anisotropic Network Model (ANM) (Bahar *et al*., 2005) are fast time-independent methods used for computing vibrational modes and estimating the flexibility of the protein. However, these techniques are not ideally suited to deal with energy barriers and multiple minima in the potential-energy surface. Monte Carlo simulations provide a useful alternative, but to the best of our knowledge, they were not used to study large-scale motions in TM proteins.

Motion planning (MP) is a fundamental problem, originally studied in robotics and computational geometry, but with implications in numerous other fields (Latombe, 1991, 1999; Sharir, 2004). The MP problem can be stated as follows: given a robot in an environment with obstacles, find a collision-free path connecting the current (start) configuration of the robot to a desired (goal) configuration. A class of randomized-path planning methods, known as Probabilistic Road Map (PRM) methods have been successfully applied to complicated high-dimensional problems (Kavraki *et al*., 1996; Hsu *et al*., 1999; Choset *et al*., 2005). PRM techniques sample the robot's configuration space at random, and retain the collision-free samples as milestones. Then, pairs of milestones are connected with local paths that serve as collision-free connectors of the generated milestones. The result is an undirected graph, called a probabilistic roadmap, whose nodes are the milestones and the edges are the local paths.

A distinction exists between multi-query strategies (e.g. Kavraki *et al*., 1999) and single-query ones (e.g. Hsu *et al*., 1999). In a single-query strategy the goal is typically to find a collision-free path between the two query configurations by exploring as little space as possible. Single-query strategies often build a new road map for each query by growing trees of sampled milestones rooted at the initial and goal configurations (Hsu *et al*., 1999). Rapidly-exploring Random Trees (RRT) (LaValle *et al*., 2001;

*To whom correspondence should be addressed.

LaValle, 2006), briefly described in Section 3.1, have been recognized as a very useful tool for designing efficient single-query paths in highly constrained spaces.

Probabilistic techniques combined with optimization and clustering have been used to sample conformational spaces of ligands and identify their low-energy conformations (Finn *et al*., 1996). Randomized path-planning methods were used successfully in computational biology by replacing the collision detection, used in robotic applications, with a molecular force field. Singh *et al*. (1999) applied PRM techniques to the ligand-binding problem. Apaydin *et al*. (2001) and Amato *et al*. (2003) applied PRM techniques to study protein folding. Recently, Cortes *et al*. (2005) developed an algorithm to compute large-amplitude motions in flexible molecular models. They applied RRTs to compute protein loop conformational changes and ligand trajectories.

We extend the RRT framework to predict TM α-helix bundle motions and the conformational changes of the helices in the bundle. Eukaryotic TM proteins form predominantly α-helix bundles in the membrane. Considering the α-helices as rigid bodies may reduce the conformational space substantially. However, owing to the large spectrum of motion scales, we do not assume that the helices are completely rigid. Therefore, in addition to movements of the helices as rigid bodies in three-dimensional (3D) space, we consider also changes in torsion angles and side-chain flexibility within these helices, while using constraints on these degrees of freedom (dofs) in a way that the conformational space will not exceed reasonable computational limits. Our algorithm is divided into two main stages. The first stage filters out many infeasible pathways using purely geometric considerations resulting in collision-free paths. In the second stage, these paths are analyzed using an energy-based criterion. The direct output of the algorithm is several movies that simulate the feasible paths that can be further examined, while taking into account functional data on the protein under study.

We tested the effectiveness of the algorithm with an application to the receptor tyrosine kinase (RTK) ErbB2 and Glycophorin A. Our results comply with previous data on these proteins. It is encouraging to note that motion paths for ErbB2 suggested by our algorithm are similar to the mechanism proposed by Fleishman *et al*. (2002) although we used very different methods to suggest and simulate the motion path.

2 A TM PROTEIN MODEL

A protein can be described as a long linkage with side-chains attached to the C_α atoms on its backbone. Using a standard modeling assumption for proteins, bond lengths and angles are often treated as fixed during motion. However, torsion angles can change significantly when the protein's conformation changes. Thus, in our model, a protein is considered as an articulated mechanism with revolute joints corresponding to the torsion angles along the protein backbone.

TM proteins of the α-helix bundle type comprise helices that are embedded in the membrane. Although helices are often considered as rigid bodies, for motion prediction purposes we cannot treat them as entirely rigid. Thus, when moving from one conformation to another, there might be slight changes in the (ϕ, ψ) torsion angles of amino acids in the helices. We model a helix as a kinematic chain using the chain tree hierarchy introduced by Lotan *et al*. (2004). In the chain tree hierarchy, the rotatable bonds, around which the (ϕ, ψ) torsion angles are defined, cut the protein backbone into rigid groups of atoms, called links. There are two types of links. The first includes the C_{i-1}, O_{i-1} and N_i atoms, where i stands for the position of amino acids along the protein backbone. The second group includes $C\alpha_i$ and all side-chain atoms attached to it (Fig. 1). A reference frame is attached to each link in the chain and the relative location of consecutive frames is defined by a homogeneous transformation matrix, which is a function of the torsion-angle between them. As the conformations of a helix change, the update of the torsion angles of its backbone is done quickly by updating the matrices corresponding to these torsion angles instead of updating the Cartesian coordinates of the atoms. Collision detection with R rigid links, takes $O(R^{\frac{4}{3}})$ time, which is not optimal in the worst case, but performs well in practice.

The algorithm of Lotan *et al*. (2004) assumes that the side-chains are rigid, whereas in our implementation, under some criteria (as explained below), we do allow side-chains to move.

2.1 Structural constraints

On the one hand, one of the driving forces behind motion in TM proteins is to keep the helices tight together in a way that the interactions between these helices do not decrease dramatically. On the other hand, the helices cannot pack so closely as to generate steric clashes between atoms. A steric clash occurs, when the distance between the centers of two non-bonded atoms is significantly smaller than the sum of these atoms' van der Waals (vdW) radii. We partly allow penetration between atoms using a cutoff parameter $\mathcal{K}$, which is the percentage of the vdW radii, namely the centers of two non-bonded atoms of vdW radii r_1 and r_2 must be at least $\mathcal{K}(r_1{+}r_2)$ apart. For our experiments, we used $\mathcal{K} = 60\%$. Thus, a fine combination of the two contradicting forces, tightness and steric-clash avoidance, is considered in our model.

2.2 Problem statement

Given a set of helices represented as kinematic chains and an initial spatial conformation of these helices, we aim to find a feasible motion path (or paths) that simulate the native motion towards goal conformations (that may not be given in advance). We denote the set of n TM helices by $\{h_1 \ldots h_n\}$. Each helix has six dofs corresponding to its position and orientation.

2.3 Relaxations applied to the TM helices

If a helix h_i has m_i torsion angles, the dimensionality of the configuration space in our problem is enormous with $6n + \sum_{i=1}^{n} (m_i - 1)$ dofs, where n is the number of helices. In addition, we consider side-chain flexibility, leading to more dofs. However, we may use some relaxations on the dimensionality of the problem when considering TM helices. The relaxations we use are as follows: (1) The TM helices cannot be fully buried in the membrane and therefore their axes are limited to maximal tilt angles of 50° with respect to the membrane normal. (2) The lateral movements of the helices as a group in the membrane is not considered by our motion analysis, implying that a specific rigid link of one helix can be placed at a fixed location in 3D. (3) Canonical helices have (ϕ=−60, ψ=−40) torsion angles along the backbone. Since we want to limit helix distortion, we allow each angle to deviate by less than ±10° from torsion angles of a canonical helix. (4) Side-chain movements may be important players in the motion-prediction problem. However, for the purposes

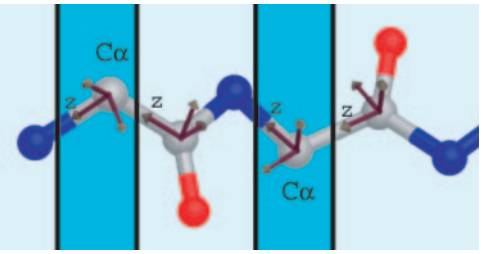

Fig. 1. The backbone degrees of freedom represented on a diglycine peptide. The two-color background shows the partition of the atoms into links. Reference frames are attached to each link origin at the $C\alpha$ and C atoms of the backbone. The z-axis of each frame is the vector along the rotatable bond; the other two axes complete the frame to form an orthogonal right-hand coordinate system.

of obtaining an approximation of the large-scale motions of the protein, it seems reasonable to consider side-chain movements only when they interfere with the way to a desired conformation. Thus, each time we derive motion from one conformation to another, we allow movements only in side-chains that are in conflict with this motion.

3 THE ALGORITHM

We have developed a motion-planning algorithm to predict motion in TM α-helix bundles. For a set of TM helices in 3D space, a conformation of an α-helix bundle comprises all the geometric information related to these helices, namely, the six dofs of helix positions and orientations in 3D space, the torsion angles of each amino acid and the conformations of the side-chains within these helices. The conformation space, C_{space}, is the union of all these possible conformations. C_{space} is divided into feasible, C_{feasible}, and forbidden, C_{forbid}, regions. C_{forbid} contains all the conformations that involve steric clashes between atoms (both within and between helices). In addition, C_{forbid} contains conformations that involve low vdW interactions between the helices. C_{feasible} is simply $C_{\text{space}} \backslash C_{\text{forbid}}$.

Our algorithm proceeds in two stages: Growing RRT—construction of a tree (RRT) that contains the set of feasible collision-free pathways emerging from a given initial conformation, using the constraints described in Section 2.1 applied to the TM helices. This stage is followed by Energy Analysis—assigning weights to the generated nodes and edges in the RRT, corresponding to the energy of a conformation (see Section 3.2 for details) and the energy associated with the move from one conformation to another, respectively. The rationale behind this division is that the first stage uses purely geometric terms to efficiently filter out unlikely pathways and reduces the search space on which the more intricate energy analysis should be applied. Following the two-stage algorithm, several weighted RRTs are built and clustering is performed on the emerging pathways. The energetically favorable pathways are chosen to produce movies.

3.1 Growing RRT

In its general form, the RRT algorithm is based on growing a conformation-space tree $\mathcal{T}$ rooted at the initial conformation q_{init}. $\mathcal{T}$ is incrementally grown to efficiently explore the feasible conformation space in order to find a feasible path connecting q_{init} to a goal conformation. In each iteration, a random conformation, q_{rand}, is generated and the nearest node, q_{near}, in $\mathcal{T}$ (according to some appropriate distance metric M) is expanded towards q_{rand}. If no collision is found on the way towards the random conformation, then q_{rand} becomes a new vertex in the tree and an edge is added between q_{near} and q_{rand}. Otherwise, q_{near} expands as close as possible towards q_{rand}. In this case, the last feasible conformation (unless it is too close to q_{near}) becomes a vertex in $\mathcal{T}$ and an edge is added between q_{near} and the new vertex (Fig. 2). It was shown (LaValle *et al*., 2001) that this method leads to Voronoi-biased growth of $\mathcal{T}$. This means that vertices with large Voronoi cells[1] have a larger probability of being extended. This is a useful property as large Voronoi cells represent unexplored areas of the conformation space.

In our implementation, each node in the tree represents an α-helix bundle conformation. In the beginning, the tree contains a given initial conformation q_{init}. During the expansion process, new conformations are sampled uniformly at random while satisfying the relaxations stated in Section 2.3. While growing an edge from q_{near} towards q_{rand} a forbidden conformation, q_{forbid}, may occur. q_{forbid} is either a conformation with steric clashes, or it contains highly remote helices, i.e. the distance between the helix axes are above a given cutoff $\mathcal{D}$ (we use $\mathcal{D} = 14$ Å in the experiments reported below). In the latter case the expansion is stopped and the algorithm continues as usual. However, when collision between side chains occurs during the expansion toward the sampled conformation, the algorithm tries to adopt a new conformation only for the colliding side-chains that obstruct the way to q_{rand}, in a way that the adopted conformation will be free of collisions. In case of a success, q_{near} continues to expand towards q_{rand}. Otherwise, a new node is generated for the last feasible conformation that was found.

Using the chain-tree hierarchy, the colliding side-chain can easily be detected and examined. We employed a fairly simple procedure that finds the set of collision free rotamers using the backbone-dependent rotamer library from Dunbrack *et al*. (1994), considering rotamers in the range $[-50, -70]$ for ϕ and $[-30, -50]$ for ψ. The backbone-dependent rotamer library evaluates each rotamer by a probability term. Our algorithm preferentially selects high-probability rotamers, while keeping the conformation free of clashes. This step can be computationally expensive, but the number of colliding side-chains in each iteration is relatively small. The algorithm continues to grow the tree till a stopping criterion is fulfilled. In our algorithm, the stopping criterion is reached if novel conformations are not added to the tree after several iterations. In other words, if the algorithm fails to expand $\mathcal{T}$ for a threshold number of consecutive iterations, it implies that the sampled conformations in $\mathcal{T}$ cover C_{feasible} sufficiently, and the expansion of $\mathcal{T}$ is stopped.

When a goal conformation is given, RRT strategies often try to grow two trees rooted at the initial and goal conformations (LaValle, 2006). However, we anticipate that, owing to the paucity of structural information regarding TM proteins, we may often encounter a case whereby only one conformation is known, and so a goal conformation is unavailable. Therefore, after the generation of the tree, our

[1] A Voronoi cell of a vertex v is the set of all points in space that are closer to v than to any other vertex, under the given metric.

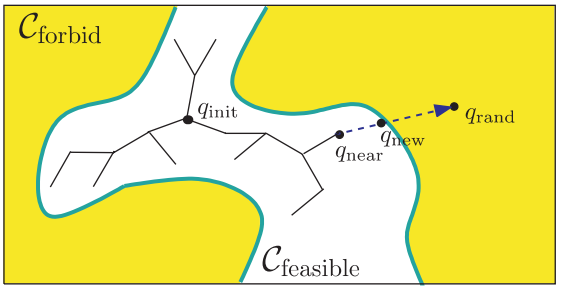

Fig. 2. Expansion of $\mathcal{T}$ using an RRT-based algorithm. The edge from q_{near} travels toward q_{rand} up to the boundary of the C_{forbid} region.

algorithm suggests a goal conformation as well as the path that leads to it.

3.2 Energy analysis

So far, we have considered only geometric constraints imposed on the motion of TM helices, resulting in a tree with collision-free paths. Our next goal is to incorporate energetic considerations into the generation of the tree. It has been suggested that tight packing of α-helices in TM proteins plays a considerable role in stabilizing these proteins (Curran and Engelman, 2003), implying that vdW forces are important descriptors of inter-helix interactions. We calculated the vdW interactions between the helices using the Lennard-Jones (LJ) 6–12 potential. The vdW energy of an α-helix bundle conformation was calculated as

$$E_{\text{vdW}} = \sum_{i>j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \qquad (1)$$

where r_{ij} is the distance between atoms i and j, ϵ_{ij} is the energy-well depth and σ_{ij} is the atomic radii sums. The parameters were taken from CHARMM19 (Neria *et al*., 1996). Thus, a weight was assigned to each node in $\mathcal{T}$, based on the LJ potential of its respective conformation. In the same manner, we added a penalty-weight to each edge between two conformations that corresponds to the maximal LJ potential observed along the local path between them.

Given a weighted RRT, we wish to find paths that minimize the weights along the pathway, and more importantly, lead to a goal conformation that is associated with a low value of the potential. We rely on a common assumption that a pathway may have some energetically unfavorable conformations that may lead to a more favorable conformation, and our aim is to capture these goal conformations. We define two different energy functions for each path: a pathway function $\mathcal{P}$ that equals to the highest value of the potential that is observed along the nodes and edges in the pathway, and a goal function $\mathcal{G}$ that corresponds to the value of the potential of the last conformation in the path, which we refer to as the goal conformation. Formally, for a path $\pi=\{v_0, e_0, v_1, e_1 \ldots e_{k-1}, v_k\}$, where v_i stands for a node and e_j for an edge, $\mathcal{P}(\pi)= \max_{0\le i\le k, 0\le j\le k-1}\{\mathcal{W}(v_i),\mathcal{W}(e_j)\}$, where $\mathcal{W}$ is the weight of the nodes or edges in $\mathcal{T}$, and $\mathcal{G}(\pi) = \mathcal{W}(v_k)$.

3.2.1 Path clustering Different sequences of randomly sampled conformations lead to different trees (RRTs). Thus, instead of growing one tree, several RRTs have been grown in the same way as described in Section 3.1, and clustering is performed on the paths derived from these trees. Each cluster comprises a set of paths that end with the same goal conformation [i.e. the root-mean-square deviation (rmsd) between the atoms of any two goal conformations in a cluster is below a predefined cutoff $\mathcal{Q}$; in our experiments we use $\mathcal{Q}$=1.4 Å]. For a cluster $C_j = \{\pi_1, \ldots, \pi_m\}$, a representative path π^* was chosen to be the one that minimizes the LJ potential in the conformations stored on the path edges and nodes, i.e. $\mathcal{P}(\pi^*) = \min_{1\le i\le m}\{\mathcal{P}(\pi_i)\}$. Different paths may comprise different lengths (number of nodes in the path), still, the above criterion (minimizing $\mathcal{P}$) is more dominant than the path lengths. However, if several paths in a cluster had the same values $\mathcal{P}(\pi^*)$, then the representative path was chosen to be the shortest path among them.

Clusters with a goal conformation that is close to the initial conformation were ignored. A score was assigned to the remaining clusters based on the LJ potential of the goal conformation $\mathcal{G}(\pi^*)$ and the number of paths in the cluster. We integrated the two terms into a form of the colony function (Xiang *et al*., 2002). Thus, the score of a cluster is $\mathcal{F}(C_j) = \sum_{\pi_i \epsilon C_j} e^{-\mathcal{G}(\pi_i)}$. In other words, the score favors clusters comprising many paths leading to a mutual energetically favorable conformation. The representative paths of the highest-score clusters were selected to produce movies that simulate the motion of the TM helices.

4 RESULTS

To explore the utility of the motion-planning algorithm in suggesting possible pathways for conformational changes in proteins, we used it to investigate the TM domain of the RTK ErbB2, overexpression of which has been implicated in many types of cancer [reviewed in Burgess *et al*. (2003)]. The protein, which is a member of the epidermal growth factor-receptor (EGFR) family, includes large extra- and intra-cellular domains that are connected by a single TM helix. It is known to form homo- and heterodimers with other EGFRs. It was proposed that ErbB2 activation involves a rotation in the relative orientation of the cytoplasmic kinase domains within a receptor dimer that is driven by a rotation of the TM helices (Jiang *et al*., 1999). A molecular mechanism for such rotation-coupled activation was suggested based on a computational exploration of conformations of the ErbB2 TM domain (Fleishman *et al*., 2002), yielding two symmetrical, and apparently stable, conformations. The more stable of the two conformations, involved packing of the helices with Gly668 and Gly672 on consecutive helical turns, invoking the Gly-xxx-Gly sequence motif (Curran and Engelman, 2003), at the inter-helix interface. In the less stable conformation, the interface was composed of Ser656 and Gly660 residues on consecutive turns. Based on these calculations it was suggested that activation of the ErbB2 receptor involves rotation of the helices within the TM domain in switching between these two conformations (Fleishman *et al*., 2002), in harmony with the proposition of rotation-coupled activation (Jiang and Hunter, 1999).

The aforementioned computations that served as the basis for suggesting a molecular model for rotation-coupled activation of ErbB2 (Fleishman *et al*., 2002) used a drastically simplified representation of the helices, which comprised solely C_α atoms forming canonical α-helices. To test the feasibility of the suggested molecular mechanism in a more realistic context, we used the method presented in this paper starting from the stable conformation involving the Gly668 and Gly672 residues. Two peptides, each of

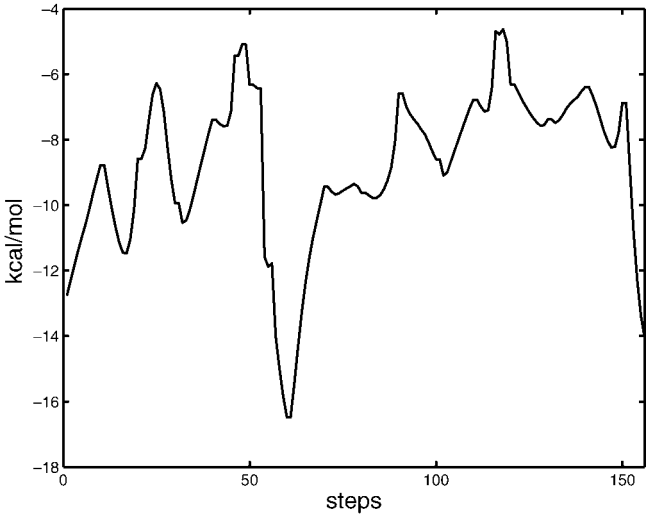

Fig. 3. The LJ potential curve of the conformations along the motion pathway of ErbB2. The curve shows the energy of the preferred pathway according to the colony energy function (Section 3.2). Step 0 corresponds to the initial conformation where the helices were packed via the glycine residues in positions 668 and 672, whereas step 156 corresponds to the goal conformation where the helices interacted through Ser656 and Gly660. The energy minimum in step 60 refers to packing via the Gly668-xxx-Gly672 motif in a conformation that is energetically more favorable than the initial conformation. As expected, it was assigned a lower potential than in step 156, suggesting that packing via Gly668-xxx-Gly672 is more stable than via Ser656-xxx-Gly660 motif as suggested previously (Fleishman *et al*., 2002).

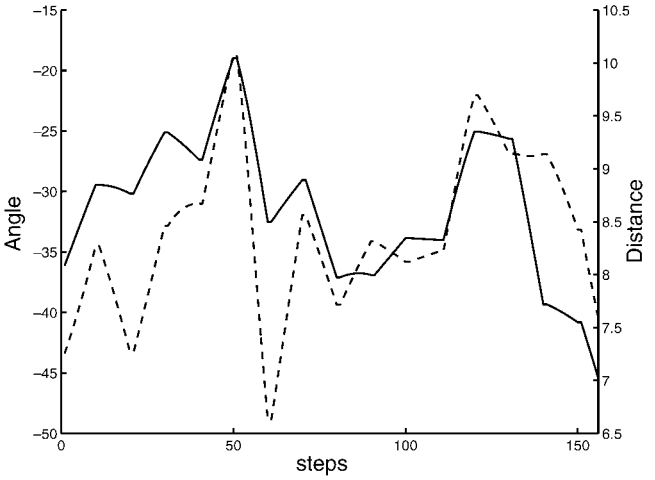

Fig. 4. Crossing angles (°) and interaxial distance (Å) between the helices axes along the most favorable motion pathway simulating the motion in the ErbB2 homodimer. Crossing angles are marked by the continuous curve whereas interaxial distances are marked by the dashed curve. Step 0 corresponds to the initial conformation where the helices were packed via the glycine residues in positions 668 and 672, whereas step 156 corresponds to the goal conformation where the helices interacted through the Ser656-xxx-Gly660 motif.

which corresponds to the TM domain of ErbB2 [LTSIVSAVV-GILLVVVLGVVFGILI], were built as canonical α-helices. They were assembled in a structure that resembled the stable conformation, and side-chains were added to the structure using the SCWRL software (Canutescu *et al*., 2003). Each atom was assigned a vdW radius according to the CHARMM19 forcefield (Neria *et al*., 1996), and the conformational space (external and internal dofs) was explored using the RRT procedure, subjected to two opposing constraints on the distance between the helices. The first was self avoidance: vdW clashes between atoms were not allowed beyond 40% overlap between their radii (i.e. $\mathcal{K} = 60\%$, Section 2.1). An opposing constraint was imposed on the maximal distance between the helices: conformations in which the LJ potential was above a pre-defined cutoff of -5 kcal/mol were excluded. The cutoff value was empirically found to facilitate an efficient exploration of the conformational space. It was the lowest cutoff that yielded motion pathways, i.e. a cutoff value of -6 kcal/mol resulted in paths comprising conformations in the vicinity of the initial state only, and larger values of up to -2 kcal/mol gave similar pathways to those using the -5 kcal/mol cutoff, but also sampled many irrelevant conformations, in which the helices formed little if any contact with one another. We also tried other measures of the helix tightness instead of the LJ potential. For example, each conformation was ranked by the buried-surface area of the helices (calculated with a probe sphere of 1.4 Å) or the number of pairs of atoms that were in contact. The resulting pathways were similar to those obtained by the LJ potential (data not shown), implying that the method is quite robust to the choice of energy function.

A homodimer, such as the ErbB2 TM domain simulated here, is expected to show some degree of symmetry in its conformations. To verify that our implementation retrieves this tendency towards symmetric conformations, we did not impose symmetry on the helices. Nevertheless, the resulting pathways showed that the two helices were symmetry-related throughout all of the simulations. In fact, superimposition of one helix over the other, using a rotation of π radians around the axis of symmetry of the helices' principal axes[2], gave a mean rmsd of 0.57Å (Supplementary Material, Fig. 6). These results encouraged us to impose symmetry on all dofs during the exploration of the conformational space, resulting in a reduction of the number of dofs.

Starting from the initial conformation of the helices, 10 random trees were generated, each of which contained $\sim$320 nodes, i.e. conformations. The conformations were clustered based on the rmsd between the α carbons, and 29 different clusters were found. The next step was to rank the clusters according to their stability. Two different criteria, the total number of conformations in each cluster and the value of the potential of the goal conformation in each cluster, were used to this end. A cluster that contained 79 conformations was ranked first by the colony function (Section 3.2). Encouragingly, the representative conformation of this cluster corresponded to the less stable conformation suggested by Fleishman *et al*. (2002). Each of the pathways was assigned a feasibility score as described in Section 3.2, and the pathway that was assigned the best score was presented in the movie (Supplementary Material, Movie 1). The optimal pathway was composed of a sequence of the most stable conformations. This is in analogy to the path of minimum energy in chemical kinetics. Other

[2]For the two axes ℓ_1 and ℓ_2 of the helices, we choose an axis of symmetry, namely a line ℓ such that rotation of π radians around ℓ will align ℓ_1 with ℓ_2. Further details can be found in the Supplementary Material.

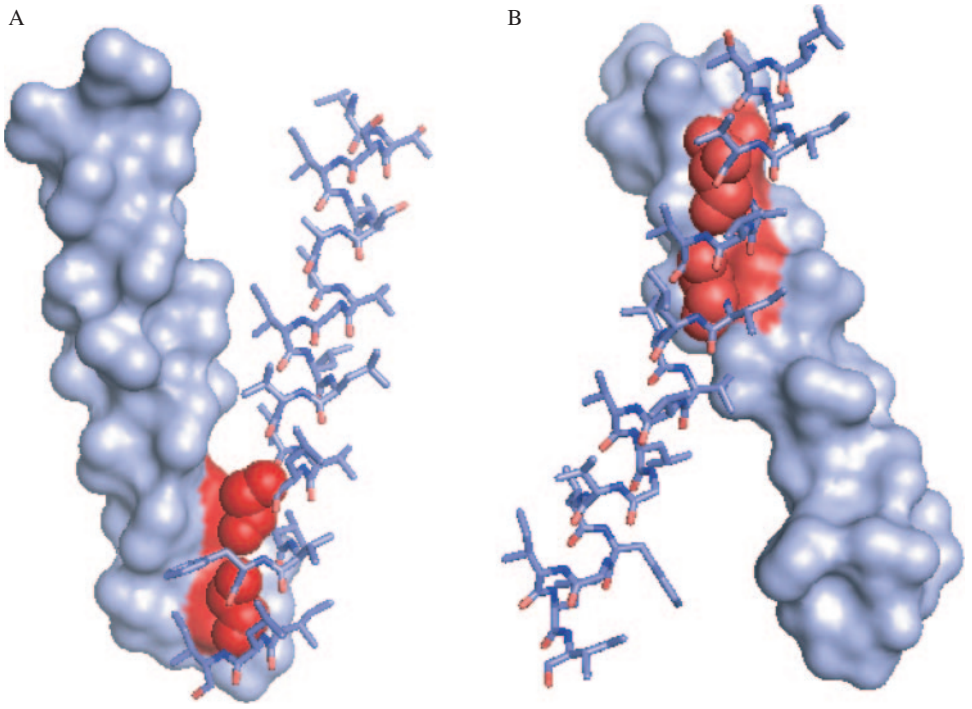

Fig. 5. The initial Gly668-xxx-Gly672 (**A**) and final Ser656-xxx-Gly660 (**B**) conformations of the TM domain of the ErbB2 homodimer. The Gly-xxx-Gly (A) and Ser-xxx-Gly (B) interfaces are marked in dark gray on the molecular surface of the helix at the back. The helix on the front is presented using a balls-and-sticks model, and the glycine and serine residues that comprise the motifs are presented using space-filled model.

characteristics of this pathway are presented in Figures 3 and 4, and representative snapshots from this pathway are provided in Figure 4. It is interesting to note that pathways that were ranked below this one partially overlapped with it.

Figure 3 shows the potential curve of the pathway that was ranked first according to the colony function. The pathway starts from the stable conformation involving the Gly668 and Gly672 residues (Fig. 5A) towards the less stable conformation involving the Ser656 and Gly660 residues (Fig. 5B). The energy is indicative of the stability of the conformation, e.g. in step 60, the pathway leads to the energetically most favorable conformation of packing via the Gly668-xxx-Gly672 motif where the distance between the helices is very small (6.5 Å) and the crossing angle is around $-35°$. The path ends in a conformation where the helices are packed via the Ser656-xxx-Gly660 motif. This conformation is associated with a less pronounced trough in the curve, where the interaxial distance between the helices is 7.5 Å and the angle is around $-45°$. Both this and the initial conformation (Fig. 5A) correspond to ridges-into-groves packing between the helices (Chothia *et al*., 1981) via the Ser-xxx-Gly and Gly-xxx-Gly motifs, respectively. In fact, it is evident from the movie (Supplementary Material, Movie 1) that the helices move subjected to the ridges-into-groves packing and that the stability at each step along the pathway is determined by the steric properties of the residues that mediate the inter-helix contact. For example, the least stable conformation (around Step 120) corresponds to the packing via Val664 residues. As suggested by Fleishman *et al*. (2002), the bulkiness of this residue interferes with the ridges-into-groves packing and this conformation, which determines the height of the energy barrier between the initial and final conformations in our suggested motion pathway. It is encouraging that the search, which started from a conformation that was in the vicinity of the most stable conformation, yielded both the most stable conformation (step 60) and a less favorable, but stable, conformation (step 156).

In addition, we examined the backward motion from a conformation where the helices are packed via the Ser656-xxx-Gly660 motif towards the conformation in which the helices are packed via the Gly668-xxx-Gly672 motif. The results (Supplementary Material, Movie 2) showed that the motion that was ranked first was very similar (in reverse order) to the original path. It ended in a goal conformation with an rmsd of $\sim$1.4Å from the initial conformation of the original path.

Glycophorin A is a bitopic TM protein that forms stable homodimers, and the NMR structure of this protein shows that the two TM helices are packed together via Gly79 and Gly83, similar to the Gly-xxx-Gly motif in one of the conformations suggested for ErbB2 above (MacKenzie *et al*., 1997). We carried out calculations using the NMR structure as the initial conformation. The calculation, which can be thought of as a negative control experiment, resulted in a few redundant pathways, comprising of native-like conformations (Supplementary Material, Movie 3).

5 DISCUSSION

A new RRT algorithm for the detection of stable conformations in TM proteins and putative pathways between them was presented here. In its pure form, the algorithm is based on geometric considerations, and energetic criteria may be added in a flexible

way. The current implementation is based on the LJ potential [Equation (1)].

It should be noted, however, that the calculated energy is unrealistically large in magnitude (e.g. Fig. 3), which is typical for force fields. Thus, the results should be examined only qualitatively. The reason for the apparent success of the potential of Equation (1) to provide reasonable pathways may be indicative of the significance of vdW interactions in stabilizing the conformations. Alternatively, the success of such a rudimentary potential, that excludes all other components of the inter-protein interactions, as well as the effects of the lipids and membrane structure, may be fortuitous. This issue will be clarified as more examples are investigated.

The calculations are very fast. For example, the 10 trees that were used to investigate the ErbB2 dimer (Section 4) were produced within <4 h on a standard desktop PC, which is significantly faster than typical molecular dynamics simulations of a similar system. The short simulation time and the flexible nature of the algorithm enable testing many aspects of the system, including the effects of changes in the energy function. Given a TM protein of interest, one can conduct a few test runs to converge to a reasonable procedure, as we demonstrated here for the TM domain of the ErbB2 and Glycophorin A homodimers.

In this preliminary work, we have focused on simple systems comprising pairs of α-helices, thus circumventing the complexities of modeling loops that connect pairs of helices. Our method can be generalized to TM proteins with an arbitrary number of helices and possibly also to water-soluble proteins of the α-helix bundle class. The addition of more helices will obviously increase the number of dofs. However, it will also reduce C_{feasible} owing to self-avoidance effects. C_{feasible} may be reduced further because many conformations of the helices may be incompatible with the lengths of the loops that connect them (Enosh *et al*., 2004).

ACKNOWLEDGEMENTS

This study was supported by a grant 222/04 from the Israel Science Foundation to N.B.-T. S.J.F was supported by a doctoral fellowship from the Clore Israel Foundation. Work by A.E. and D.H. has been supported in part by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University.

Conflict of Interest: none declared.

REFERENCES

Abramson,J. *et al*. (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**, 610–615.

Amato,N.M. *et al*. (2003) Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, **10**, 239–255.

Apaydin,M.S. *et al*. (2001) Capturing molecular energy landscapes with probabilistic conformational roadmaps. *Proceedings of IEEE International Conference Robotation Automation,* Seoul, pp. 932–939.

Bahar,I. and Rader.,A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin Struct. Biol*., **15**, 1–7.

Burgess,A.W. *et al*. (2003) An open-and-shut case? Recent insights into the activation of EGF/ErbB receptors. *Mol. Cell*, **12**, 541–552.

Canutescu,A.A. *et al*. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*., **12**, 2001–2014.

Choset,H. *et al*. (2005) *Principles of Robot Motion: Theory, Algorithms, and Implementations*. The MIT Press, chapter 7.

Chothia,C. *et al*. (1981) Helix to helix packing in proteins. *J. Mol. Biol*., **145**, 215–250.

Cortes,J. *et al*. (2005) A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, **21**, i116–i125.

Curran,A.R and Engelman,D.M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Op. in Struct. Biol*., **13**, 412–417.

Dunbrack,R.L.O. and Karplus,M. Jr (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains. *Nat. Struct. Biol*., **1**, 334–340.

Enosh,A. *et al*. (2004) Assigning transmembrane segments to helics in intermediate-resolution structures. *Bioinformatics*, **20**, i122–i129.

Finn,P.W. *et al*. (1996) Geometric manipulation of flexible ligands. *Proceedings of Workshop on Applied Computational Geometry,* Berlin, pp. 67–78.

Fleishman,S.J. *et al*. (2002) A putative molecular-activation switch in the transmembrane domain of erbb2. *Proc. Natl Acad. Sci*., **99**, 15937–15940.

Hsu,D. *et al*. (1999) Path planning in expansive configuration spaces. *Int. J. Comput. Geometry Appl*., **9**, 495–512.

Jiang,G. and Hunter,T. (1999) Receptor signaling: when dimerization is not enough. *Curr Biol*., **9**, 568–571.

Karplus,M. and McCammon,J.A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol*., **9**, 646–652.

Kavraki,L. *et al*. (1996) Probabilistic roadmaps for path planning in high dimensional configuration spaces. *In Proceedings of IEEE Transactions on Robotics and Automation*, Vol. 12, 566–580.

Latombe,J.-C. (1991) *Robot Motion Planning*. Kluwer Academic Publishers Boston, MA.

Latombe,J.-C. (1999) Motion planning: A journey of robots, molcules, digital actors, and other artifacts. *Int. J. Robotics Res*., **10**, 1119–1128.

LaValle,S.M. (2006) *Planning Algorithms. Cambridge University Press,* chapter 5. http://msl.cs.uiuc.edu/planning/.

LaValle,S.M. and Kuffner,J.J. (2001) Rapidly-exploring random trees: progress and prospects. In Donald,B.R., Lynch,K.M. and Rus,D. (eds), *Algorithmic and Computational Robotics: New Directions.* A.K. Peters, Wellesley, MA, 293–308.

Lotan,I. *et al*. (2004) Algorithm and data structures for efficient energy maintenance during Monte Carlo simulation of proteins. *J. Comput. Biol*., **11**, 902–932.

MacKenzie,K.R. *et al*. (1997) A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.

Neria,E. *et al*. (1996) Simulation of activation free energies in molecular systems. *J. Chem. Phys*., **105**, 1902–1921.

Sharir,M. (2004) Algorithmic motion planning. In Goodman,J.E. and O'Rourke,J. (eds), *Handbook of Discrete and Computational Geometry*. 2nd edn., 1037–1064. Chapman and Hall/CRC Press, Boca Raton.

Singh,A.P. *et al*. (1999) A motion planning approach to flexible ligand binding. *Int. Sys. for Molec. Biol*., 252–261.

Xiang,Z. *et al*. (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.