The Raymond and Beverly Sackler Faculty of Exact Sciences
School of Computer Science

Algorithmic Infrastructure for the Prediction of Structure and Motion in Transmembrane Proteins

Thesis submitted for the degree of "Doctor of Philosophy"

By

Angela Enosh

This work was carried out under the supervision of Prof. Dan Halperin and Prof. Nir Ben-Tal

Submitted to the Senate of Tel-Aviv University
August 2008
Acknowledgements

I am deeply grateful to my advisors, Professor Dan Halperin and Professor Nir Ben-Tal for their caring supervision, guidance and advices from the early stage of my research. This interesting combination of computer science and biochemistry filled me with extraordinary experiences and enriched my knowledge substantially in the fields of motion planning and TM-proteins which eventually lead to this thesis. Above all, I am grateful for their constant encouragement and support in various ways.

My thanks to the members of Dan Halperin’s laboratory, who have been both my friends and colleagues.

I am thankful to Barak Raveh, who has been a scientific collaborator and a friend, for profound discussions and ideas and a wonderful enjoyable collaboration.

I am deeply and forever grateful to my parents, Galia and Yair, for their love, support and encouragement throughout my entire life. My deepest thanks to my sisters, Shuli, Yael and Hagar, for their constant love and support. I am grateful to my husband, Sharon, for believing in me and for his invariable help and support for the last ten years. I am grateful to my beloved son, Bary, who is my source of strength and inspiration. I am deeply indebted to my beloved grandfather, Shmuel, who died last year - his remarkable personality, his love and care will live with me forever.
Preface

This thesis is based on the following collection of papers that were published throughout the PhD period in scientific journals.


5. Barak Raveh‡, Angela Enosh‡, Ora Furman-Schueler and Dan Halperin. Rapid sampling of molecular motions with prior information constraints. Submitted. (‡These authors contributed equally to this work)
**Table of Contents**

**Abstract**

**Introduction**

**Manuscripts**

Assigning transmembrane segments to helices in intermediate-resolution structures  M1

Quasi-symmetry in the cryo-EM structure of EmrE provides the key to modeling its transmembrane domain  M2

Prediction and simulation of motion in pairs of transmembrane α-helices  M3

Generation, comparison and merging of pathways between protein conformations: Gating in K-channels  M4

Rapid sampling of molecular motions with prior information constrains  M5

**Discussion**

**Hebrew Introduction**

**Hebrew Abstract**
Abstract

Transmembrane (TM) proteins play pivotal roles in many cellular processes including cell-to-cell signaling, solute transport and cellular organization, and are therefore attractive targets for drug discovery. It is estimated that 20–30% of proteins encoded by the genome are transmembrane. Despite their significance, however, only a small fraction of the distinct folds of TM proteins have been solved to date by high-resolution methods such as X-ray crystallography and NMR. Attempts to determine TM protein structures by these methods are acutely hindered by technical problems related to their purification and crystallization. This intricacy emphasizes the urgent need to develop automated computational methods for structural and mechanistic analysis in TM proteins.

TM proteins yield more easily to structure determination using cryo-electron microscopy (cryo-EM). Cryo-EM often results in intermediate resolution structures, where amino acids cannot be viewed but the locations of the TM helices can be deduced experimentally. Thus, the challenge is to assign the TM segments in the protein sequence into the corresponding helices in the cryo-EM structure. To this end, I have developed a computational tool that reduces the possible assignments between TM-sequence segments and the helices seen in cryo-EM substantially. This is a first step toward modeling of TM proteins based on cryo-EM data.

Motion in TM proteins plays an essential role in a variety of biological phenomena. TM proteins form pumps and channels that control and guide the transportation of ions and metabolites across the membrane. They also function as receptors and are responsible for molecular recognition of hormones and neurotransmitters. Thus, when a high-resolution structure is determined, it raises many questions about its mechanism of action. Up to now, approximated mechanisms have been suggested for only a few TM-proteins, e.g., cartoon-level resolution for the lactose permease, and helical movements for the tyrosine kinase receptor ErbB2 (also called HER2). My goal here was to propose mechanisms in molecular-detail.

In my thesis, I have developed a general framework for the generation, comparison, and clusterization of motion pathways with application to helical TM proteins. The method is rooted in probabilistic motion-planning techniques in robotics. Within the framework, many collision-free motion pathways are generated efficiently, while considering a wide range of degrees of freedoms, including external and internal moves. The pathways are then clustered and hybridized using a novel method, which I developed, for pathways comparison, which is based on dynamic programming. Using these methods, I investigated the TM domains of the ErbB2 receptor and the KcsA potassium-channel. My studies yield the shortest pathway, which comprises minimal-energy conformations among the many generated pathways. In both cases, the resulting motion-pathways suggested molecular interpretations of experimental data. In addition, in a joint work with Barak Raveh, we extended the framework to soluble proteins and integrated the software within the Rosetta package. This extension allows the incorporation of partial biological information (on the pathways or the target conformation) to the generation of motion pathways.
Introduction

Transmembrane (TM) proteins are extremely important for a broad range of processes and function in all biological systems. They function, for example, as ion channels, transporters and receptors. Structural and mechanistic knowledge about them should, therefore, significantly advance our understanding of these functions. They are today’s prime drug targets and this knowledge may also help in rational drug design.

In this research I have developed algorithmic tools towards structure prediction, as well as, frameworks for motion-pathways elucidation, for the alpha-helix bundle class of TM proteins.

Towards Structure Prediction

Eukaryotic TM proteins form predominantly alpha-helix bundles in the plasma 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 on the two-stage model for protein assembly in the membrane (Popot and Engelman, 1990). According to this model, the first step in TM-protein folding is the insertion of the TM domains into the membrane as alpha-helices. Only in the second stage do these helices associate to form helix bundles (White and Wimley, 1999; Popot and Engelman, 2000). One of the implications of the two-stage model is that, overall, the stability of individual TM domains is independent of that of other domains. Hence, prediction of TM-protein structure may begin with prediction of TM-helix locations in the amino-acid sequence of the protein. The most common approach to this effect is based on the use of hydrophobicity scales (White and Wimley, 1999; Kessel and Ben-Tal, 2002). These scales determine which amino acids are positioned inside or outside the membrane respectively (Bowie, 2006; Fleishman and Ben-Tal, 2006; von Heijne, 2006).

Structural data of several TM proteins at intermediate-resolution of approximately 5 −10Å in-plane is available from cryo-EM studies of 2D crystals (see, e.g., Fleishman et al., 2006). 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. Thus, the correspondence between the TM-sequence segments and the cryo-EM helices cannot be decided unambiguously.

Structure prediction of TM proteins based on cryo-EM data may comprise the following steps: (i) The derivation of helical positions from cryo-EM maps experimentally or by computational means (e.g., Wrigger et al., 1999; Jiang et al., 2001). (ii) Assigning TM-sequence segments to the derived helices, this step may result in a number of feasible assignments. (iii) The prediction of favorable packing orientations of alpha-helices in TM proteins. A number of computational methods have been suggested for positioning and orienting the helices comprising the TM domain with respect to one another (see, e.g., Fleishman and Ben-Tal, 2002; Kim et al., 2003; Pellegrini-Calace et al., 2003; Fleishman et al., 2004; Adamian and Liang, 2006; Yarov-Yarovoy et al., 2006; Park and Helms, 2007). (iv) Predicting the conformations of the extramembrane segments which connect these helices using various methods for modeling loop-closure conformations (e.g., Wedemeyer and Scheraga, 1999; Xiang et al., 2002; Canutescu and Dunbrack, 2003; Cortes et al., 2004; Coutsias et al., 2004) and (v) adding side chains to the proposed backbone model (by e.g., Dunbrack and Karplus, 1994; Bower et al., 1997; Xiang and Honig, 2001; Gordon et al., 2002; Canutescu et al., 2003; Chazelle et al., 2004; Kingsford et al., 2005).

Consequently, providing a solution to the helix-assignment problem is a first step toward modeling of TM proteins. That is, the assignment of TM-sequence segments to the helices in the
cryo-EM data can limit the conformation space substantially. In addition, helix assignment is directly useful for structural studies of TM 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.

In this research, I have developed a novel method for assigning TM segments to helices observed in intermediate-resolution cryo-EM structures based on geometric considerations (M1: Enosh et al., 2004). In a joint work with our collaborators, we predicted the structure and mechanism of substrate translocation of the small multidrug resistance antiporter from *E. coli*, EmrE (M2: Fleishman et al., 2006).

**Motion-Pathways Prediction**

Proteins are viewed as dynamic structures with varied types of motion ranging from the fast localized movements, e.g., side-chain flexibility to the large-scale motions that extend over larger distances as domain movements. Large-scale motions are thought to relate to functional motions, however, the different types of motion are interdependent and coupled to one another, thus, in the study of slow large-scale motions it is important to consider fast small-scale motions that eventually are the ones that impose limitations on the simulation time (Becker and Watanabe, 2001). The following paragraphs introduce the concept of a forcefield and summarize the mainstream techniques that are used to simulate, investigate and predict the dynamics in molecules.

**Forcefields with Application to Motion Simulation in Proteins.** A vital key to the success of any computational approach to the study of dynamics in molecules is the quality of the potential functions in use. These functions are used to evaluate the feasibility of a particular conformation of a molecule. Ideally, this would be done with quantum mechanics. However, these methods are strictly limited to small molecules. As a compromise, empirical functions have been developed based on classical physics to approximate the true energy of molecular systems, called molecular forcefields. A forcefield is a combination of a potential energy function and a set of parameters that are atom-type-dependent values defined within the potential function. Forcefields are usually combined with computational methods for motion simulations with the aim to scale conformations according to their energy, where low energy conformations are preferable.

A potential energy function is a mathematical equation that is used to calculate the potential energy of a molecular conformation based on its three-dimensional structure. The equation comprises terms that describe a variety of physical interactions that dictate the structure, such as, bond lengths, bond angles, torsion angles, van der Waals interactions and electrostatic interactions (MacKerell, 2001; Kavraki, 2007). Forcefields can be applied to all-atoms models or extended-atom models, in which the hydrogens are not explicitly represented and instead are treated as part of the nonhydrogen atom to which they are covalently bound. Extended-atom models are particularly useful for simulations that involve large sampling of the conformational space.

There is a wide range of forcefields, of which the majority of molecular simulations are performed with the CHARMM (Chemistry at Harvard Molecular Mechanics, e.g., Brooks et al., 1983; MacKerell et al., 1998), AMBER (Assisted Model Building with Energy Refinement, e.g., Pearlman et al., 1991; Pearlman et al., 1995; Case, et al., 2002), and GROMOS (Groningen Molecular Simulation, e.g., Berendsen, et al., 1995; Gunsteren et al., 1996) packages. This variety yielded many papers that compare between the performance of the different forcefields (e.g., Hall and Pavitt, 1984; Whitlow and Teeter, 1986; Kini and Eveans, 1992; Hobza et al., 1997). All the potentials are approximations and therefore one forcefield may be more suitable than others for a particular study. For an overview on empirical forcefields for biological macromolecules, see the paper by MacKerell, 2004.
Molecular Dynamics Simulations. Proteins are dynamic systems. The extent of their motion becomes more evident with the accumulation of structures of the same molecule in different conformations. The first detailed atomic motions in a protein were provided in 1977 via a molecular dynamics (MD) simulation by McCammon et al. MD simulation is a powerful and widely used tool that allows prediction of the dynamic properties of proteins directly from their three-dimensional structures by numerically solving Newton’s equations of motion (Becker and Watanabe, 2001). It can be described as an iterative process by which the Newtonian equations of motion are solved on a step-by-step basis using small time intervals. Consequently, slow motions which may extend over larger distances cannot be computed completely by MD simulations on biologically relevant timescales.

To facilitate conformational transitions, different approaches were integrated with MD, e.g., NonEquilibrium Molecular Dynamics (NEMD), Targeted Molecular Dynamics (TMD) and Steered Molecular Dynamics (SMD). NEMD (Evans and Morriss, 1990) involves the simulation of a classical system of atoms or molecules interacting via interatomic potential forces where an external field drives the system away from thermodynamic equilibrium. SMD (Izrailev et al., 1998) allows large conformational changes in molecules on time scales accessible to MD simulations by applying external forces to the system under study. It has been applied to investigate binding mechanisms (e.g., Izrailev et al., 1997) and recently it has been applied to TM proteins, e.g., to investigate the gating mechanism of KvAP voltage-gated potassium channel (Monticelli et al., 2004); to study the opening of the gate in the KcsA channel (Biggin and Sansom, 2002); and to investigate the gating of MscL channel (Gullingsrud and Schulten, 2003). TMD simulation can be applied when two conformations are known, by enforcing a conformational transition from an initial structure towards the given final structure (Schlitter et al., 1994). In the context of TM proteins, it has been applied, e.g., to study the pore opening of the KcsA channel (Compoint et al., 2005) and to investigate the gating mechanism of MscL (Schlitter et al., 1995).

Monte-Carlo Simulations. Monte Carlo (MC) simulation is a standard technique for sampling in conformational space and for computing protein motion (Binder and Heerman, 1992; Becker, 2001). MC simulations perform a series of steps in conformational space, starting from a given conformation. In each step, a new random “trial conformation” is generated by perturbing some degrees of freedom (e.g., Lotan et al., 2004) of the previous conformation, then, a decision is made whether the new conformation should be accepted or rejected. If rejected, new trial conformations will be created until one of them is accepted. If accepted, the search process continues from the new conformation. The trial conformation is usually accepted or rejected based on the energy-value difference between the trial and the previous conformation according to Metropolis criterion (Leach, 1996). If the energy of the new trial conformation is lower than that of the previous conformation, it is always accepted. But even if the energy of the trial conformation is higher, there is a certain probability proportional to the change in the energy
value, that it will be accepted; the goal of accepting higher energy conformations is to escape from local minima. Thus, MC simulations tend to move toward low energy states.

At high temperatures there is a significant probability to cross high energy barriers and this probability becomes significantly smaller at low temperatures. Thus, at low temperatures, MC simulations often get trapped in local minima. Efficient MC-based techniques to overcome this local minima trapping have been suggested, e.g., in (Swendsen and Wang, 1986; Li and Scheraga, 1987; Hansmann, 1997; Zhang and Skolnick, 2001), where the replica sampling algorithm was used in which several simulations are implemented at different temperatures. Another example is the parallel hyperbolic MC algorithm (Zhang et al., 2002) that can speed up the thermalization of the protein-folding process by flattening the local high-energy barriers. Stochastic roadmap simulation (SRS) is another variant of MC that allows the analysis of an ensemble of trajectories simultaneously (Apaydin et al., 2002; Apaydin et al., 2003; Chiang et al., 2006). It is an efficient simulation tool to study molecular motion that overcomes the local minima problem encountered in MC simulations. SRS is based on the construction of a roadmap that approximates a large number of simultaneous MC-simulation trajectories. These trajectories are then processed together, which greatly reduces computation time.

MC based techniques have been applied to diverse purposes, such as protein folding (e.g., Li and Scheraga, 1987; Shimada et al., 2001), identification of global minimum conformations (e.g., Abagyan and Totrov, 1999; Zhang and Skolnick, 2001), side chain optimization (e.g., Holm and Sander, 1992) and study of molecular motions (e.g., Apaydin et al., 2002; Apaydin et al., 2003; Tikhonov and Zhorov, 2004).

In my work I have generated trajectories by applying the RRT algorithm (i.e., Rapidly-exploring Random Trees; see section “Motion Planning Algorithms” below). The algorithm is similar to MC searches that are biased into the unexplored regions of the conformation space. Unlike MC simulation, the RRT algorithm records the topology of the energetically feasible conformational space in a tree, so that paths in the tree correspond to low energy pathways.

Normal Mode Analysis. Normal mode analysis (NMA) is one of the most powerful techniques used to probe large-scale dynamics of biological molecules. It was applied to proteins from the early 1980s (Brooks and Karplus, 1983) with the aim to predict functionally-relevant motions. NMA is based on the harmonic approximation of the potential energy function around a stable equilibrium structure. The derivation of the normal modes begins with the minimization of the conformational potential energy. At the minimum, the potential energy is expanded in a Taylor series in terms of mass-weighted coordinates (Hayward, 2001). The second derivatives of the Taylor expansion are assumed to describe the energy surface over fluctuations that are far from the minimum and they are the elements of the Hessian matrix. Normal mode directions and their frequencies are determined by the eigenvectors and eigenvalues of the Hessian matrix. Thus, a molecule with N atoms has 3N normal modes.

Previous studies on NMA have shown that the vibrational normal modes with the lowest frequencies and largest fluctuations are the ones that are functionally relevant (e.g., Ma and Karplus, 1998; Krebs et al., 2002; Cui et al., 2004; Trakhanov et al., 2005). Implied that the large space of normal modes for a biological molecule can be defined by a smaller subspace, comprising M lowest frequency normal modes, where M is between 10% to 20% of the total normal modes (as suggested by Hayward, 2001). Krebs et al. (2002) showed that over half of 3814 known protein motions can be approximated by their two lowest-frequency normal modes. In a recent review, Wynsbergh and Cui (2006) suggested that a much larger number of modes is required for properly investigating correlated motions.

NMA is a coarse-grained approach due to the omission of high-frequency modes and unharmonic motions even when the normal modes are calculated based on all-atoms molecular structures. Thus, coarse-grained NMA method (Bahar et al., 1997; Bahar and Rader, 2005) has become a major trend in recent years and it has been applied to large molecular complexes in
order to reduce the complexity of the systems (e.g., Atilgan et al., 2001; Tama et al., 2002). A review on the use of coarse-grained NMA methods can be found in a paper by Bahar and Rader, 2005.

This method is very attractive due to its simplicity, the relative low computational cost and the clear visualization of the resulting motions. However, this technique has limitations, resulting from the use of the harmonic approximation while the nature of protein motions is known to be unharmonic. The limitations, such as solvent damping, the absence of information about energy barriers and multiple minima in the potential energy surface are discussed in details in a review on NMA by Ma (2005). Thus, the correspondence between low-frequency normal modes and functional motions might be surprising. Despite these limitations, various NMA applications have been developed in the study of the dynamics of biological molecules. NMA results for molecular-dynamics study are particularly meaningful when interpreted in conjunction with experimental data.

NMA has been hybridized with other methods, e.g., MD (Zhang et al., 2003; Tatsumi et al., 2004) and motion-planning algorithms (Kirillova et al., 2007).

Several web applications enabling traditional NMA approaches have been constructed, e.g., ProMode (Wako et al., 2004), ElNémo (Suhre and Sanejouand, 2004), MolMovDB (Echols et al., 2003) and recently a web application for the analysis of transmembrane alpha-helical movement has been constructed by Skjaerven et al. (2007).

NMA has been applied to TM proteins to investigate their gating mechanism for, e.g., potassium channels (Shrivastava and Bahar, 2006), the bacterial large conductance mechanosensitive channel (Mscl; Valadie et al., 2003) and the TM domain of the nicotinic acetylcholine receptor (nAChR; Taly et al., 2005).

**Motion Planning Algorithms.** A protein can be viewed as a kinematic chain of amino-acid residues, each of which contributes two degrees of freedom (dofs), (corresponding to its $\phi$ and $\psi$ torsion angles) to the main chain of the protein. In addition, a side chain may contribute up to five rotatable dihedral angles but these dofs are usually reduced to a small number of rotameric states for each amino acid (Dunbrack and Karplus, 1994). Under the assumption that bond lengths and angles are almost constant across conformations at room temperature, these rotatable torsion angles and the side-chain rotameric states are the only dofs that guide the motion in proteins. Therefore, proteins are very similar to a class of robotic-articulated mechanisms with revolute joints corresponding to the torsion angles along the protein backbone, suggesting that robotic motion planning and protein motion prediction are very similar problems. Due to this analogy, motion planning (MP) algorithms from robotics have been applied to study protein dynamics. The main difference, however, between a robotic manipulator and a protein chain is that the protein is subject to physical forces while a robotic manipulator is analyzed in terms of colliding and non-colliding configurations. A brief overview of how path planning techniques have been applied to protein motion analysis can be found in a recent survey by Kavraki, 2007.

The basic motion-planning problem can be stated as follows: Given a robot moving in an environment cluttered with obstacles, and given a start and a goal configuration for the robot, find a collision-free path connecting these two configurations if one exists. (*Configuration* is the common robotics term for *conformation*.) Otherwise, determine that no such path exists.

A configuration of a robot is a full description of the robot's state, including its position in three-dimension, orientation and its internal dofs (e.g., revolute joint angles). MP is a fundamental problem, originally studied in robotics and computational geometry, but with implications in numerous other fields (Latombe, 1991; Latombe, 1999; Halperin et al., 2004; Sharir, 2004).

In MP problems, the configuration parameters of the robot (its dofs) define the configuration space, $C_{space}$, i.e., the union of all possible configurations of the robot, annotated by whether the robot is in collision or collision-free at each configuration. In essence, all motion-
planning problems are equivalent; once they are formulated in the configuration space, they reduce to the problem of finding a connected sequence of points between the start and the goal configurations. The dimension of $C_{\text{space}}$ depends on the number of dofs, thus, for high-dimensional problems, MP is a challenging research problem. (For very low-dimensional problems, the MP problem can be solved by using some basic methods such as, e.g., Minkowski sums; de Berg et al., 1997). Numerous methods have been developed for MP, e.g., skeleton (visibility graph or the Voronoi diagram), cell decomposition and potential field. Introduction to the MP problem and algorithms can be found in (Howie et al., 2005). The generalized MP problem in three-dimensional space is shown to be PSPACE-hard (Reif, 1979), therefore, in complex MP cases, heuristic algorithms are preferable upon exact algorithms and play an important role in engineering applications.

For high dimensional problems, the class of randomized-path planning methods, known as Probabilistic Roadmap (PRM) methods have shown great potential (Kavraki et al., 1996; Hsu et al., 1999; Bohlin and Kavraki, 2000; Hsu, 2000). The costly computation of an explicit representation of the $C_{\text{space}}$ is replaced by sampling random configurations and using a fast collision checker to determine whether they are in collision.

The basic PRM framework comprises the following steps. Initially, a large number of points (milestones) in $C_{\text{space}}$ are sampled and tested for a collision, keeping those that are free of collisions. This step results in a collision-free configuration set in $C_{\text{space}}$. Then, attempt is made to connect pairs of collision-free samples that are relatively close to each other by a local planner, i.e., by sampling and collision-checking configurations between them, resulting in a graph data structure (roadmap) that capture the connectivity of the collision-free space. Given initial and goal configurations, an attempt is made to connect these configurations to existing configurations stored in the roadmap. In case of success, any standard graph search method can be applied to the roadmap in order to find feasible collision-free paths.

Randomized path-planning techniques provide an incremental planning scheme which does not depend on the dimensionality of the configuration space. The analysis of the convergence rate of several PRM algorithms reveals the strength of this technique (e.g., Kavraki et al., 1998; Hsu et al., 2002), each new milestone added to a probabilistic roadmap refines the connectivity relationship captured in the graph and reduces the probability that the planner fails to find a solution path, when one exists. Various applications of randomized planners are reviewed in (Latombe, 1999) including robotics, design for manufacturing and servicing, graphic animation of digital actors and more.

A distinction exists between multi-query strategies (as PRM, where the roadmap is preprocessed once and can be queried many times, e.g., Kavraki et al., 1996) 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 roadmap for each query by growing trees of sampled milestones rooted at the initial and goal configurations (Hsu et al., 1999). There are two common and similar types of single-query strategies: Expansive Spaces Trees (ESTs) and Rapidly-exploring Random Trees (RRTs) which are basically similar to one another.

The EST method (Hsu et al., 1999; Sanchez and Latombe, 2001) tries to explore new parts of the space. At each step, a node in the tree is selected for expansion. The node is selected with a probability that is inversely proportional to the number of other nodes that lie within its neighborhood. Thus, isolated nodes in sparsely sampled parts of $C_{\text{space}}$ are more likely to be chosen. Some bias may also be present for nodes closer to the goal. Attempts are made to expand the chosen node by sampling configurations in the vicinity of this node, and the configurations that can be connected to it by the local planner are added to the tree.

RRT method (LaValle and Kuffner, 2001; LaValle, 2006) has been recognized as a very useful tool for designing efficient single-query paths in highly constrained spaces. The RRT algorithm attempts to find a collision-free path between initial and goal conformations using a
greedy heuristic that biases the conformational exploration from the initial toward the goal conformation and vice versa, though at the same time avoids the pitfalls of local minima. At each step, a random configuration is generated in the configuration space, with some bias toward the goal configuration. Its nearest node in the tree is expanded toward the random configuration for a predefined number of steps or until an obstacle is encountered, whichever comes first. The last collision-free configuration on this path is added to the tree. It was shown (LaValle and Kuffner, 2001) that this method leads to Voronoi-biased growth of the tree, i.e., vertices with large Voronoi cells that correspond to the unexplored areas of the conformation space have a larger probability of being extended.

Probabilistic MP techniques have been applied in the context of molecular biology from the late 1990s. These 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 in computational biology by replacing the collision detection, used in robotic applications, with a molecular forcefield. Singh et al. (1999) applied PRM techniques to the ligand-binding problem. PRM techniques have been applied to study protein folding (Apaydin et al., 2001; Amato and Song, 2002; Amato et al., 2003; Shawna et al., 2005). Cortes et al. (2005) developed an algorithm to compute large-amplitude motions in flexible molecular models using RRT. Later, the same group integrated normal modes analysis with path-planning for the study of large conformational changes in proteins (Kirillova et al., 2007).

I have applied RRTs to predict stable conformations and motion pathways in pairs of transmembrane helices (M3: Enosh et al., 2007). In a recent work (M4: Enosh et al., 2008), I have applied RRTs to investigate and simulate the motion pathway of the KcS potassium channel. Later, in a joint work with Barak Raveh (M5: Raveh* et al., 2008), we have integrated the RRT framework within the Rosetta software (Rohl et al., 2004) to suggest motion pathways for proteins which enhances state of the art energy function that incorporates VDW forces, hydrogen bonds, solvation and electrostatic terms. In that work we present Path-Rover which is an extension of the general RRT framework that incorporates biological information to the exploration of the conformation space. This biological information can be deduced by various experimental techniques, such as spin labeling, NMR and cross-linking that may suggest pairwise distance between specific atoms or by, e.g., Circular Dichroism spectroscopy that may suggest the formation of secondary structures during the motion pathway. In addition, homologue structures or simply expert intuition may suggest biological information that can be easily integrated with our software in the form of predicates and direct the RRT towards the conformational regions that fulfill these predicates.

Probabilistic MP techniques, and RRT in particular, may yield many plausible pathways between two given configurations. As part of my thesis (M4: Enosh et al., 2008), I have developed an efficient technique to compare and fit pairs of pathways, using dynamic programming (DP). The fitting of motion pathways can be combined with general clustering techniques (e.g., hierarchical clustering; Johnson, 1967) to meaningfully cluster similar pathways. This technique, when applied to protein-motion analysis, is enhanced to hybridize better motion pathways than found by the, e.g., RRT algorithm, namely, pathways that are energetically favorable. The path-fitting algorithm that I propose is reminiscent of curve-fitting techniques previously used to match curves in two- and three-dimensions (Wolfson 1990), which in turn are close to the DP solution to the longest common subsequence problem (Cormen et al., 2001). Unlike the latter problems, my path-fitting algorithm works in high-dimensional spaces and may be useful in various applications.

* Angela Enosh and Bark Raveh contributed equally to this work.
Assigning transmembrane segments to helices in intermediate-resolution structures

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

Received on January 15, 2004; accepted on March 1, 2004

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 (< 4 Å) 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 40 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 comprised of TM helices and loops, which are typically located on the internal or external sides of the membrane, and connect pairs of consecutive helices.

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 three-dimensional (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 positions of 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 structural studies of TM 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...
Assignment of transmembrane segments to helices

**Definition of helices.** Positions, tilt and azimuthal angles of each helix can be extracted from intermediate-resolution cryo-EM maps (Unger, 2001). Canonical \( \alpha \)-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, \( C_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\} \). 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 \( c_{ik} \) is the closest atom to the outer plane of the membrane. We denote the internal \( C\alpha \) atom by \( \text{internal}(C_i) = c_{i1} \), and the external \( C\alpha \) atom by \( \text{external}(C_i) = c_{ik} \).

It should be noted that the positions of helices deduced from cryo-EM in this manner suffer from imprecision. First, the orientations of the helices around their principal axes cannot be derived from cryo-EM maps 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 \( S = \{T_1, X_1, T_2, \ldots, X_{n-1}, T_n\} \) and a set of helix locations in 3D-space \( C = \{C_1, C_2, \ldots, C_n\} \), 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 \( \sigma \) such that for each \( 1 \leq i \leq n \), \( T_i \) is assigned to \( C_{\sigma(i)} \), and (ii) attribute a score to each feasible 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 \( \max_{\text{dist}}(X_i) \) is longer than the distance between \( \text{external}(C_k) \) and \( \text{external}(C_m) \). 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

---

**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 as follows: \((A0, B1, C2), (A0, B2, C1), (A2, B0, C1)\) and \((A2, B1, C0)\). Note 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.

---

1.1 Terminology and formal statement of the problem

The sequence of a TM protein of the \( \alpha \)-helix bundle type, denoted by \( S \), comprises TM and extra-membrane segments, which connect TM segments that are consecutive in the sequence (Fig. 1a). 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, \( T_i \subset S \), by \( T_i = \{t_{i1}, t_{i2}, \ldots, t_{ik}\} \), as an ordered sequence of amino acids from the N- to the C-terminus. Similarly, we denote an extra-membrane segment, \( X_i \subset S \), by \( X_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\} \), 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 \( \text{length}(X_i) \), 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 \( \text{max}_{\text{dist}}(X_i) = \text{length}(X_i) + 1 \times \text{dist}(C\alpha, C\alpha) \), where \( \text{dist}(C\alpha, C\alpha) \) is the distance between two consecutive \( C\alpha \) atoms, which is typically taken as 3.8 Å (Creighton, 1993).

---

**Function.,** which was derived on the basis of conformations of loops in \( \alpha \)-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 (PDB, http://www.rcsb.org/pdb/), and our results show that the native-state assignment ranks high in many cases.
between internal($C_i$) and 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 (Fig. 1). This stage is followed by Ranking the feasible assignments—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 $G(V, E_{int} \cup E_{ext})$, such as the one in Figure 1c, where:

$$ V = \{(T_i, C_j) \mid 1 \leq i, j \leq n\}, $$

$$ E_{int} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \text{dist}[\text{internal}(C_j), \text{internal}(C_m)] \leq \text{max} \cdot \text{dist}(X_i)\}, $$

$$ E_{ext} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \text{dist}[\text{external}(C_j), \text{external}(C_m)] \leq \text{max} \cdot \text{dist}(X_i)\}, $$

where $V$ 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 external($C_j$) and external($C_m$) on the external side of the membrane. The same applies to the internal edges where $X_i$ is sufficiently long to connect internal($C_j$) and 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 $n$-th to the 1st level (where $n$ is the number of TM segments in the protein). The $k$-th level in the graph consists of vertices comprising $T_k$, namely $\{(T_k, C_j) \mid 1 \leq j \leq n\}$. Given the set of nodes $\{(T_k, C_j) \mid 1 \leq j \leq n\}$ in the $k$-th level, we construct the $(k-1)$th level as follows. For each vertex $(T_k, C_j)$, we go over all the helices $C_i \in C \setminus \{C_j\}$ and if $X_{k-1}$ can connect the two helices $C_i$ and $C_j$ on the external or internal side of the membrane, we add the vertex $(T_{k-1}, C_i)$ (if it is still missing) to the $(k-1)$th level, and a directed edge $e = [(T_{k-1}, C_i), (T_k, C_j)]$, where $e \in E_{ext}$ or $e \in E_{int}$. Thus, a directed edge $e \in E_{ext} \cup E_{int}$ can appear only between two consecutive levels. At the beginning, all the vertices $(T_n, C_j)$ in the $n$-th level are examined against the pairs $(T_{n-1}, C_i)$ where $C_i \in C \setminus \{C_j\}$, and created if and only if the above condition holds. After construction of the graph $G$ we can eliminate all the nodes between the second to the $n$-th 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 $n$-th level of $G$, and is comprised of an alternating sequence of external and internal edges (either $e_k \mid k$ even) are external and $e_k \mid k$ odd are internal, or vice versa. In addition, we require that $\pi$ 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 $\pi$ 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 <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 $\sigma^k$, which assigns the TM segments $T_1, \ldots, T_n$ to the helices $C_{\sigma_1(i)}, \ldots, C_{\sigma_n(i)}$, where $1 \leq k \leq n!$. We define the score function $F$ of a permutation $\sigma^k$ 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 conformations of the extra-membrane loops as in the classic loop-closure problem, since the native extra-membrane segments that connect pairs of helices are unknown. Hence, we seek to define a score for matching an extra-membrane segment to connect a given pair of 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 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).
2.2.1 The preprocessing phase. We denote the two consecutive helices from the \( N \)- to the \( C \)-terminus 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 the motifs overlap. Transforming these motifs to the common reference frame yields a set of points in 3D-space each of which corresponds to the location of the beginning of the second helix in the helix-loop-helix motifs (i.e. \( B \)'s) relative to the common first helix (i.e. the overlapping \( A \)'s).

All 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 \( \text{KD}_l, 2 \leq l \leq 7 \), 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 \( l = 4 \).

2.2.2 The scoring phase. We compute \( f[\mathbf{X}_i, C_{\sigma(i)}, C_{\sigma(i+1)}] \) as follows. We place the two helices \( C_{\sigma(i)} \) and \( C_{\sigma(i+1)} \) 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 \( C_{\sigma(i+1)} \). 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 points and \( q \).

Let \( Q \) be a cube centered at \( q \) with side size \((10 \cdot x) \text{ Å}\). We query KD\(_q\) 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,

\[
f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)}) = \sum_{r \in Q} e^{-\text{dist}(q,r)}
\]

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 \( f[\mathbf{X}_i, C_{\sigma(i)}, C_{\sigma(i+1)}] = 0 \).

---

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

### Table 1. Helix–loop–helix motifs classified by loop length

<table>
<thead>
<tr>
<th>Loop length</th>
<th>Number of motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>456</td>
</tr>
<tr>
<td>3</td>
<td>260</td>
</tr>
<tr>
<td>4</td>
<td>171</td>
</tr>
<tr>
<td>5</td>
<td>167</td>
</tr>
<tr>
<td>6</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
</tr>
</tbody>
</table>

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

3 THE DISTRIBUTION OF END POINTS OF SHORT LOOPS IS HIGHLY NON-UNIFORM

Structures of helix–loop–helix motifs (resolution of 2 Å 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 PDB, only proteins whose sequences were <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 2–7 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.

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 \( l = 4 \) (Fig. 2d–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 \( \phi \) and \( \psi \)). 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, Figure 2d–f show that the locus of the free end in
Fig. 2. The distribution of the starting points of helices $B$ 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. (a) and (d) display the points together with their least-mean-square (LMS) plane. The view point of (b) and (e) is the normal to the LMS plane. (c) and (f) present side view. It can be seen that the starting points of helices $B$ in motifs with loops of length three form a torus-like shape in 3D-space.
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 eight 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 PDB 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 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 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 comprised 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 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 and 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 as 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 seven 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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>PDB</th>
<th>Loop lengths</th>
<th>(n_f)</th>
<th>(n_{pos})</th>
<th>(i) Accurate Rank</th>
<th>(ii) Noisy Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteriorhodopsin</td>
<td>1c3w</td>
<td>3,14,2,3,10,4</td>
<td>7</td>
<td>5040</td>
<td>44</td>
<td>948</td>
</tr>
<tr>
<td>Sensory rhodopsin</td>
<td>1h68</td>
<td>7,12,2,3,3,4</td>
<td>7</td>
<td>5040</td>
<td>84</td>
<td>512</td>
</tr>
<tr>
<td>Cytochrome c oxidase</td>
<td>l0cccC</td>
<td>3,5,19,2,7,7</td>
<td>7</td>
<td>5040</td>
<td>74</td>
<td>335</td>
</tr>
<tr>
<td>Cytochrome c oxidase</td>
<td>l0ccE</td>
<td>5,6,1,1</td>
<td>5</td>
<td>120</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Cytochrome c oxidase</td>
<td>l0ccH</td>
<td>7,2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Glycerol facilitator</td>
<td>1fx8</td>
<td>6,19,24,8,4</td>
<td>6</td>
<td>720</td>
<td>236</td>
<td>352</td>
</tr>
<tr>
<td>Halorhodopsin</td>
<td>1e12</td>
<td>2,20,2,4,1,5</td>
<td>7</td>
<td>5040</td>
<td>34</td>
<td>73</td>
</tr>
<tr>
<td>Lactose permease</td>
<td>1pv6</td>
<td>3,2,1,3,1,24,3,1,3,1,1</td>
<td>12</td>
<td>&gt;10^8</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

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 \(n_f\), their loop lengths, and the number of possible assignments \(n_{pos} = n_f!\). The results are categorized by the number of feasible assignments \(n_{feas}\) that remained following the pruning phase and by the rank of the native assignment. We ran the program on PC Intel Pentium IV, CPU 2.4 GHz, 256 MB RAM, and the running time using the accurate data was below 2 s for each of the proteins. When using the noisy data, the running time varied between 8 s for 1occ subunit H and 6.5 min for 1pv6. Currently we are working on additional TM proteins. The results obtained from processing these cases are available at http://www.cs.tau.ac.il/~angela.
each other. In the lactose permease, for instance, 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 extramembrane 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 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 biology and genetic studies. Finally, forward-looking experiments 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) (Jakobsson et al., 1999). MGST1 is a homotrimer, in which each monomer is comprised of four TM segments. The 3D structure of MGST1 was determined at an in-plane resolution of 6 Å using cryo-EM (Holm et al., 2002; Schmidt-Krey et al., 2000). The electron-density map shows 3 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).

ACKNOWLEDGEMENTS

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. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation.

REFERENCES


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 (Fig. A1): (i) the unknown orientation of the helix with respect to its axis and (ii) the unknown translation of the helix along its axis.

We now redefine the score function $f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)})$ that was introduced in Section 2.2 to account for 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 Cα 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 $p = \text{external} [C_{\sigma(i)}]$ and $q = \text{external} [C_{\sigma(i+1)}]$, whose locations are known only approximately.

$p'$ and $q'$ are restricted to bounded regions as shown below.

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 (Fig. A1, 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 (Fig. A1, ii). It follows that $E(p)$ has a cylindrical envelope shape with radius 2.5 Å (typically, radius of a helix) and its height is set to 5 Å.

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 Cα 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: $\max_{p \in E(p), q \in E(q)} f(X_i, T_k[C_{\sigma(i)}], T_j[C_{\sigma(i+1)}])$.

It can be shown that in order to cover the envelope $E(p)$ adequately, we need to sample 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 $E(p)$.
Quasi-symmetry in the Cryo-EM Structure of EmrE Provides the Key to Modeling its Transmembrane Domain

Sarel J. Fleishman¹, Susan E. Harrington¹, Angela Enosh²
Dan Halperin², Christopher G. Tate³ and Nir Ben-Tal¹*

¹Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel
²School of Computer Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel
³MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Small multidrug resistance (SMR) transporters contribute to bacterial resistance by coupling the efflux of a wide range of toxic aromatic cations, some of which are commonly used as antibiotics and antiseptics, to proton influx. EmrE is a prototypical small multidrug resistance transporter comprising four transmembrane segments (M1–M4) that forms dimers. It was suggested recently that EmrE molecules in the dimer have different topologies, i.e. monomers have opposite orientations with respect to the membrane plane. A 3-D structure of EmrE acquired by electron cryo-microscopy (cryo-EM) at 7.5 Å resolution in the membrane plane showed that parts of the structure are related by quasi-symmetry. We used this symmetry relationship, combined with sequence conservation data, to assign the transmembrane segments in EmrE to the densities seen in the cryo-EM structure. A Cα model of the transmembrane region was constructed by considering the evolutionary conservation pattern of each helix. The model is validated by much of the biochemical data on EmrE with most of the positions that were identified as affecting substrate translocation being located around the substrate-binding cavity. A suggested mechanism for proton-coupled substrate translocation in small multidrug resistance antiporters provides a mechanistic rationale to the experimentally observed inverted topology.

© 2006 Elsevier Ltd. All rights reserved.

*Corresponding author

Keywords: dual topology; protein structure prediction; structural bioinformatics; cryo-EM; mechanism of action

Introduction

Bacterial multidrug resistance is a growing challenge to medical treatment, with previously harmless bacteria inducing life-threatening infections. One of the mechanisms for the acquirement of multidrug resistance is the active extrusion of toxic compounds from the bacterial cell through membrane transporters. Efflux of toxic compounds is driven either by ATP hydrolysis, as in the ABC transporter superfamily, or by coupling the extrusion of toxic compounds to the inward movement of protons down their electrochemical gradient, as in the small multidrug resistance (SMR) family of antiporters. Of the SMRs, EmrE is a representative from Escherichia coli, which has been extensively characterized structurally, phylogenetically, and biochemically. These analyses have provided evidence that EmrE contains four transmembrane (TM) segments that form α-helices.

A recent electron cryo-electron microscopy (cryo-EM) analysis of 2D crystals of EmrE bound to one of its substrates, tetraphenylphosphonium (TPP⁺), clearly resolved the eight α-helices comprising the EmrE dimer at an in-plane resolution of 7.5 Å and 16 Å perpendicular to the membrane plane. However, at this resolution, the individual amino acid residues were not observed, and the TM segments could not be assigned unambiguously to...
the densities representing the α-helices. The 2D crystals of EmrE bind TPP⁺ with the same high affinity as detergent-solubilized EmrE, and EmrE in the native E. coli membrane,4,8 so it is thought that the cryo-EM structure of EmrE is a faithful representation of the protein’s physiological conformation. Quasi-symmetry between six helices was detected around an axis lying within the plane of the membrane, suggesting that the EmrE monomers might assume dual topology in the membrane, with the monomers arranged in an inverted or upside-down manner with respect to one another.7 In contrast, no obvious symmetry relationship was observed around axes perpendicular to the membrane plane in either the 3D structure or a previous 2D projection map.9

Two atomic-resolution X-ray structures of EmrE have been solved in recent years. The first structure at 3.8 Å resolution appears to have trapped the molecule in an unphysiological state,16 and is incompatible with much of the biochemical data on this protein.11 Recently, another X-ray structure of EmrE was solved at 3.7 Å resolution,12 which included one molecule of bound substrate TPP⁺ per dimer. However, it has been argued that this structure too may not be physiologically relevant,13 for three main reasons. (i) The X-ray structure is very different from the cryo-EM structure of EmrE.12 (ii) Several key residues that were shown to be critical for substrate binding are not in a position to bind substrate in the structure. For instance, it was demonstrated by different experimental approaches that Glu14 residues from both monomers are crucial for translocation, participate in substrate and proton binding,14–19 and are in proximity to one another.20 By contrast, the X-ray structure shows that Glu14 from only one monomer forms partial contact with substrate and the two glutamate residues are over 20 Å apart. (iii) Evolutionary conservation has been shown to be a powerful predictor of helix orientations in integral membrane proteins, with conserved amino acid positions usually occupying locations that are buried in the protein core, whereas lipid-facing positions are evolutionarily variable.21–28 the X-ray structure of EmrE orients many conserved positions (Figure 1(a)) towards lipid, and conversely, variable amino acids are placed at helix–helix interfaces.29

The difficulties that have arisen in determining a high-resolution structure of EmrE that accounts for the body of experimental evidence and recent data supporting the dual topology of EmrE and other members of the SMR family30,31 gave us the impetus to try to understand the cryo-EM structure through modeling strategies. The proposal of the dual topology architecture of EmrE contradicts previous experimental data that suggested EmrE had one unambiguous topology,32 but could obviously have crucial implications for structural modeling. Here, we show that the most straightforward structural interpretation of dual topology, i.e. that EmrE is arranged as an anti-parallel homodimer, provides the key for determining a model of EmrE based on the cryo-EM structure.

Results

Quasi-symmetry and helix assignment

The assignment of the two sets of four hydrophobic segments seen in the sequence of EmrE to the eight helices observed in the cryo-EM structure is potentially the most significant hurdle in the structural modeling (theoretically having 4 × 8! = 161,280 different permutations).2 However, if a symmetry relationship existed between two parts of the structure, this problem could be greatly simplified (to 2 × 4! = 48 permutations). The previous analysis of the cryo-EM structure of EmrE identified symmetry between two parts of the structure around an axis of symmetry within the plane of the membrane, but there were no symmetry relationships around axes perpendicular to the membrane plane.7 Recent data suggesting dual topology in EmrE molecules provide additional support for the in-plane 2-fold symmetry axis.30,31 Indeed, several integral membrane proteins contain two structurally related domains that are related by a rotational axis of quasi-symmetry within the membrane plane (e.g. GlpF,33 ClC,34 and SecYE13).2

To derive the most likely helix arrangement for EmrE, four pieces of experimental data were used. (1) Positions of α-helices were based on the cryo-EM structure.7 (2) The continuous density between the ends of helices F and H suggested that they were adjacent in the amino acid sequence (Figure 1(b)).7 (3) The two monomers in the EmrE dimer are represented by A-D and E-H, based upon the symmetrical relationship between A-B-C and H-G-F, correspondingly (Figure 1(b)). (4) Densities A-B-C-F-G-H that form the substrate-binding chamber are composed of helices M1, M2, and M3, because amino acid residues that are involved in substrate binding and translocation are found only in these three helices (Table 1). These data alone were insufficient to give a conclusive model, so evolutionary conservation was used to guide the assignment of sequence segments to helices. The rationale behind the use of evolutionary conservation for helix assignment is that residues that are packed against other helices are conserved during evolution, since even minor substitutions in such positions often weaken interhelix contacts and adversely affect protein function.21–25 Conversely, lipid-exposed positions are expected to be generally accommodating to sequence variability. Hence, we correlated the conservation of sequences with the extent of burial of each of the helices observed in the cryo-EM structure against other helices to constrain the possible assignments.

We found that the most informative helices in the cryo-EM structure were C and F, which are related to one another by the in-plane symmetry axis.
These helices are unique in EmrE, because one half of each helix is buried on all sides by other helices and the other half is exposed to lipid on only one of its faces; of the four hydrophobic segments, only M3 contains conserved amino acid residues in this identical pattern. The N terminus of M3 is highly conserved, implying it is packed on all sides by other helices, but its C terminus shows a helical periodic pattern of variable residues, suggesting that one face is lipid-exposed (Figure 1(a)). In Figure 1(b), the most lipid-exposed, C-terminal portion of M3 is represented by the right-hand (distant) end of C and the left-hand (near) end of F.

The assignment of M3 to C and F is supported partly by the observation that M3 is predicted by our analysis to be the longest hydrophobic stretch (23 residues compared to 18 or 19 for the other TM segments, Figure 1(a)) paralleling its assignment to the most tilted helices in the structure. Further
Table 1. Summary of experimental data gathered on residues of helices M1 to M4 of EmrE, and locations of those residues in the model structure

<table>
<thead>
<tr>
<th>Residue</th>
<th>Environment predicted from model</th>
<th>Activity data</th>
<th>Environment of label</th>
<th>Activity data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr4</td>
<td>Substrate chamber</td>
<td>– 2,4,5</td>
<td>Accessible</td>
<td>4</td>
</tr>
<tr>
<td>Ile5</td>
<td>Lipid facing</td>
<td>++ 2,4</td>
<td>Accessible</td>
<td>4</td>
</tr>
<tr>
<td>Tyr6</td>
<td>Lipid facing</td>
<td>++ 2,3,4,5</td>
<td>Partly accessible</td>
<td>4</td>
</tr>
<tr>
<td>Leu7</td>
<td>Chamber</td>
<td>– 2,3,4</td>
<td>Accessible</td>
<td>2</td>
</tr>
<tr>
<td>Gly8</td>
<td>Lipid facing</td>
<td>++ 2,4</td>
<td>Lipid-facing</td>
<td>4</td>
</tr>
<tr>
<td>Gly9</td>
<td>Lipid facing</td>
<td>++ 2,3,4</td>
<td>Accessible</td>
<td>4</td>
</tr>
<tr>
<td>Ala10</td>
<td>Interhelix contact</td>
<td>– 2,3,4</td>
<td>Accessible</td>
<td>2</td>
</tr>
<tr>
<td>Ile11</td>
<td>Chamber</td>
<td>++ 2</td>
<td>Water-exposed</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– 1,3,4</td>
<td>Partly accessible</td>
<td>2</td>
</tr>
<tr>
<td>Leu12</td>
<td>Lipid facing</td>
<td>++ 1,2,3,4</td>
<td>Accessible</td>
<td>4</td>
</tr>
<tr>
<td>Ala13</td>
<td>Lipid facing</td>
<td>++ 1,2,3,4,6</td>
<td>Accessible*</td>
<td>4</td>
</tr>
<tr>
<td>Glu14</td>
<td>Chamber</td>
<td>– 2,3,4,6</td>
<td>Proximal to E14</td>
<td>2</td>
</tr>
<tr>
<td>Val15</td>
<td>Interhelix contact</td>
<td>++ 2,3,4</td>
<td>Inaccessible</td>
<td>4</td>
</tr>
<tr>
<td>Ile16</td>
<td>Lipid facing</td>
<td>++ 2,3,4</td>
<td>Lipid-facing</td>
<td>2</td>
</tr>
<tr>
<td>Gly17</td>
<td>Interhelix contact</td>
<td>+ 1,2</td>
<td>Partly accessible</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– 3,4</td>
<td>Water-exposed</td>
<td>2</td>
</tr>
<tr>
<td>Thr18</td>
<td>Chamber</td>
<td>– 1,2,3,4</td>
<td>Accessible</td>
<td>2</td>
</tr>
<tr>
<td>Thr19</td>
<td>Interhelix contact</td>
<td>++ 2,4</td>
<td>Inaccessible</td>
<td>4</td>
</tr>
<tr>
<td>Leu20</td>
<td>Interhelix contact</td>
<td>++ 2,4</td>
<td>Constrained</td>
<td>4</td>
</tr>
<tr>
<td>Met21</td>
<td>Interhelix contact</td>
<td>++ 2,3,4</td>
<td>Inaccessible</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Constrained</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

**M2**

<table>
<thead>
<tr>
<th>Residue</th>
<th>Environment predicted from model</th>
<th>Activity data</th>
<th>Environment of label</th>
<th>Activity data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val34</td>
<td>Substrate chamber</td>
<td>+ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly35</td>
<td>Lipid facing</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr36</td>
<td>Interhelix contact</td>
<td>+ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile37</td>
<td>Chamber</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile38</td>
<td>Lipid facing</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cys39</td>
<td>Interhelix contact</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr40</td>
<td>Chamber</td>
<td>– 1,5</td>
<td>Proximal to substrate</td>
<td>4</td>
</tr>
<tr>
<td>Cys41</td>
<td>Chamber</td>
<td>+ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ala42</td>
<td>Lipid facing</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser43</td>
<td>Interhelix contact</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe44</td>
<td>Chamber</td>
<td>– 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp45</td>
<td>Interhelix contact</td>
<td>++ 1,7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu46</td>
<td>Interhelix contact</td>
<td>++ 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
support is provided by the observation that the M3 N terminus contains several sequence signatures that would favor flexibility of the helix backbone, correlating with a kink in helix C observed in the cryo-EM structure (Figure 1(b)). These sequence signatures include (Figure 1(c)): (a) the presence of two highly conserved glycine residues in positions 65 and 67; (b) the observed substitution of position Ser64 with glutamate residues in other SMR members; and (c) the fact that position Ala61 is substituted by proline in several homologues. Notably, proline residues in multiple-sequence alignments of TM domains have been shown to be indicators for kinks, even in cases where the sequence of the protein, for which a structure is available, does not exhibit a proline. Although these sequence features would favor flexibility of the helix backbone, the segment does not necessarily exhibit a kink and, thus, helix F is seen to be mostly straight in the cryo-EM structure (Figure 1(b)).

Given the assignment of M3 to helices C and F, and the experimental constraints listed above, there is only one solution for the assignment of the remaining helices. As the termini of helices F and H are apparently connected by density on the side of the structure away from the viewer in Figure 1(b) (indicated by an arrow), and based on the assignment of the portion of F near the connection to H to be the N terminus of M3, then helix H must be M2. Since the M2–M3 interconnecting loop is predicted to contain only five amino acid residues (Figure 1(a)), it might be rigid and could, therefore, be visible in the cryo-EM structure. If M2 is helix H, then, by symmetry, helix A is also M2. Multiple sources of biochemical data have implicated residues on M1 as crucial for substrate binding and translocation; given the assignment of M2 and M3 above, M1 must occupy the symmetry-related B and G helices around the translocation chamber. In contrast, amino acid residues in M4 are not involved directly in substrate binding or translocation (Table 1). The lack of data implicating residues on M4 in substrate binding is in agreement with the location of helices D and E, separated from the substrate-binding chamber by helices C and F. Finally, the helix assignment suggested here (M1=B,G; M2=A,H, M3=C,F, and M4=D,E) is consistent with constraints imposed by the short interconnecting loops observed in the EmrE sequence (Figure 1(a)) on the distances between the helix ends seen in the cryo-EM structure. In addition to this most likely helix assignment, we tested each of the 47 other permutations against the known functional data on EmrE, the interconnecting loop lengths, and SMR evolutionary conservation. None of these other permutations fit the aggregate data on EmrE nearly as well as the suggested assignment (data not shown).

We note that domain swapping, where helices from one monomer interpenetrate between helices in the other monomer, could confound the proposed helix assignment. However, this possibility would connect helices that are distant from one another, and is therefore made unlikely by the very short lengths of the interconnecting loops (Figure 1(a)). The only loop that would allow domain swapping is between M1 and M2. However, the swapping of these domains involves conformations in which the loop blocks substrate entry to the binding chamber.

### Structural modeling

Canonical α-helices were constructed to fit the helix axes extracted from the cryo-EM structure. For each helix, all the rotations around its principal axis were sampled in 5° increments; each conformation was scored according to a rule that favors situations in which evolutionarily conserved amino acid positions were packed inside the protein core, with variable positions facing the lipid. Following the orientation of each of the helices, we introduced a kink into helix C to account for the deviation from α-helical regularity observed for this helix in the cryo-EM structure (Figure 1(b)). At the vertical resolution of the cryo-EM structure (16 Å), the position of the kink cannot be determined unambiguously. We therefore estimated this position on the basis of the direction of the kink observed in the cryo-EM structure and features observed in SMR
sequences of M3 that imply backbone flexibility at the N terminus of this helix (Figure 1(c)), and placed the kink so that it affects mainly the backbone hydrogen bond between positions Ser64 and Ile68. We note that this approximate location for the kink within helix C does not affect the conclusions we draw below on the support that biochemical and biophysical data provide to the model structure.

The computed conformation fits the conservation profile of each helix quite closely (Figure 2) with all of the variable residues facing the lipid, and the conserved residues facing the protein core. It is important to note that no experimentally derived information was used to constrain the orientations of the helices. As shown below, these orientations nevertheless provide a structural framework for understanding most of the biochemical data on EmrE.

Comparison of the EmrE model with data from biochemical and biophysical experiments

It is difficult to interpret the pertinent biochemical and biophysical data on EmrE on the basis of the model structure, because the model does not contain side-chains. Furthermore, we estimate that the orientations of the individual α-helices around their principal axes might vary by up to 20°, and that the positions of Cα atoms on the terminal turns of each of the helices might diverge from the positions specified in our model.23 Even with this level of uncertainty, however, it is possible to provide a rough account of the majority of the experimental data on the basis of the model structure.

The structure of EmrE has been probed using a number of biophysical and biochemical techniques. The model presented here does not seriously conflict with any of these data and, in fact, can be used to rationalize and simplify a number of observations. The experiments discussed in this section used spin labels to probe the environment of helix M1,20 and site-directed mutagenesis to define amino acid residues important for folding and transport activity.

The TM region M1 of EmrE contains Glu14, which is essential for substrate binding and translocation.14–16,18,19 Therefore, this region has been studied intensively using a number of biophysical and biochemical approaches. Site-directed spin-labeling experiments were applied to all the residues of M1 to infer which of them are packed against other helices, exposed to lipid, or are in the vicinity of M1 residues of the neighboring monomer.20 All of the residues that were identified as lipid-exposed by the spin-labeling experiments are predicted to be lipid facing in our EmrE model (dark blue spheres in Figure 3(a) and Table 1). Interestingly, lipid-exposed positions were identified by spin-labeling to be restricted mainly to the N-terminal part of M1; side-chains in its C terminus were found to be motionally more restricted (light blue spheres in Figure 3(a)). These results are in close agreement with the model structure, in which the N-terminal part is more exposed to lipid and the C terminus is packed against other helices from almost all directions. Thus, the spin-labeling data20 verify the assignment of M1 to helices B and G as well as the helices’ orientations around their principal axes. The spin-labeling experiments also identified only two residues on M1 (Glu14 and Thr18) that are vicinal to their counterparts on the other monomer. Indeed, these two residues face one another according to the model structure, and these two pairs have the closest Cα–Cα distances of all residues on M1 in the model (~16 Å). The model proposed by Koteiche et al. for the relative orientations of the two M1s considered only parallel helix packing,20 however, their results fit equally well with our antiparallel model shown here.

Hsmr is a homologue of EmrE from the archaeon Halobacterium salinarium, which is unique among SMR members, in that approximately 40% of its sequence is comprised of Ala and Val residues.40

Figure 2. A view of the EmrE model perpendicular to the membrane bilayer color-coded according to evolutionary conservation. On all helices, the conservation signal closely matches the pattern of exposure of residues to lipid, with conserved residues buried at interhelix contact regions, and variable residues placed in membrane-exposed positions. The Glu14 residues on both monomers are shown as red spheres. The two monomers are distinguished by the presence or the absence of an apostrophe. The Figure was generated with MOLSCRIPT60 and rendered with Raster3D.61
Presumably, this composition reflects an evolutionary pressure to increase the G+C content of the genome, while maintaining the relatively high hydrophobicity necessary for a TM protein. Positions that are not Ala or Val in Hsmr are, therefore, considered important for structure or function; conversely, positions that are Ala or Val in Hsmr, but not Ala or Val in EmrE, can be presumed to be unimportant. As expected, the vast majority of these positions are lipid exposed according to the model structure (Figure 3(b)). This observation suggests that, despite the low level of sequence identity between SMR proteins, the overall fold of the homologous proteins is conserved.

The cryo-EM structure of EmrE was derived from crystals of the transporter bound to TPP⁺. The position of TPP⁺ in the plane of the membrane is clear from the 3D structure and from comparisons of projection maps of EmrE with and without TPP⁺. However, the position of TPP⁺ along the axis perpendicular to the membrane plane is less certain due to the low resolution along this axis. To provide

Figure 3 (legend on opposite page)
rough constraints for which residues are located around the substrate, we docked TPP\(^+\) manually (Figure 3(c)) based on the constraint that the position of TPP\(^+\) with respect to Glu14 should roughly match that seen in the atomic-resolution structure of the water-soluble multidrug receptor BmrR,\(^{42}\) which was also crystallized in a TPP\(^+\)-bound form. In harmony with various experimental assays, the Glu14 residues from both monomers are in position to form contact with the substrate.\(^{17,19}\) It is also notable that several aromatic residues are found in the vicinity of the modeled TPP\(^+\), providing partners for aromatic interactions with substrate. In particular, positions Tyr40, Tyr60, and Trp63 were shown experimentally to bind substrate,\(^{36}\) and their C\(^{\alpha}\) atoms are located within 6 Å from carbon atoms of the modeled TPP\(^+\) (Figure 3(c)). Other aromatic residues are located between \(\alpha\)-helices, where they might increase structural stability. Table 1 lists 52 amino acid residues that line the translocation chamber or mediate interhelix contacts in our model; 21 of these residues have been mutated and implicated experimentally in substrate binding and translocation. It should be noted, however, that it is likely that the specific residues mediating EmrE binding to substrates other than TPP\(^+\) might vary from those specified here, in analogy to the differences observed in the binding modes of different substrates to bacterial multidrug gene regulators,\(^{43}\) and bacterial multidrug resistance transporters.\(^{44}\)

Mordoch et al.\(^{45}\) conducted an extensive substituted cysteine-accessibility method analysis of Cys-less EmrE (Table 1). They replaced 48 positions throughout the protein with cysteine, and tested the mutant transport properties. Only five positions were absolutely sensitive to replacement, in that the mutants were incapable of conferring resistance to known EmrE substrates. Of these five, two mutations (Ile11Cys and Thr18Cys) led to good expression of EmrE, but transport was reduced severely; these positions are on the same face of M1 as the essential Glu14 and are probably involved in substrate transport (Figure 1(d)). The other three mutations (Tyr40Cys, Phe44Cys, and Leu93Cys) resulted in no expression of EmrE, so it may be that these residues are important in the folding or stability of EmrE in the membrane. Subsequent mutational analysis of Tyr40 showed that this residue is also important in substrate recognition.\(^{18}\) Both Phe44 and Leu93 are predicted to be at the interfaces between helices (Figure 3(d)), which may explain their effect on protein folding and/or stability. This substituted cysteine-accessibility method study also identified ten residues in the TM regions that showed decreased resistance to only one of the antiporter cognate substrates.\(^{45}\) Eight out of these ten residues are positioned at or around the substrate-binding chamber (Table 1 and light blue spheres in Figure 3(d)). Presumably, substituting these positions alters the properties of the protein surface that lines the channel, hence reducing substrate affinity. As Mordoch et al. note, the residues on M2 are distributed on two helical faces, and indeed the two sensitive positions (Ala42 and Gln49), which the model does not place at or around the substrate-binding chamber, are located on this helix. Notably, mutants of these two residues are sensitive only to acrylflavine among the cognate substrates that were tested,\(^{45}\) implying that these residues might be involved in the binding of only certain substrates.\(^{43,44}\)

There are two reports of experimental data that conflict with the model we present here, because they both suggest that the monomers within EmrE have an identical orientation in the membrane with the N and C termini in the cytoplasm.\(^{32,46}\) A third study reported single topology for the QacC homologue of EmrE, although the data were inconclusive regarding the localization of the C terminus.\(^{47}\) These results are, however, contrary to

**Figure 3.** Structural interpretation of biochemical and phylogenetic data on EmrE and its homologues. (a) Spin-labeling experiments identified lipid-exposed residues (dark blue), and motionally restricted (light blue) positions.\(^{32}\) Red spheres identify Glu14 and Thr18, which were shown to be close to their counterparts on the other monomer, as indeed they are in the model. Notice that in both monomers, the motionally restricted residues on the N-terminal turn of M1 are surrounded by other helices from almost all sides, whereas positions that were identified experimentally as lipid-exposed are indeed located in lipid-facing parts of the protein. (b) Green spheres mark positions on EmrE that are aligned with Ala and Val in the Hsmr homologue from *H. salinarium*, but are not Ala and Val in EmrE. Such positions are thought to have little structural or functional importance,\(^{30}\) and indeed the majority face the lipid environment. (c) Docking of a molecule of TPP\(^+\) in the EmrE model structure. TPP\(^+\) was docked manually such that it approximately fits the orientation seen in a crystal structure of the cytoplasmic receptor for TPP\(^+\) BmrR.\(^{42}\) The two Glu14 residues (red spheres) are in proximity to aromatic rings from the substrate TPP\(^+\). Aromatic residues in the TM domain of EmrE are marked by purple spheres. Some of these residues surround TPP\(^+\), thus providing possible interaction partners for the substrate. Positions Tyr40, Tyr60, and Trp63, which are marked on the Figure, have been implicated directly in substrate binding.\(^{18}\) Others are placed in spacious regions of the structure, where they might serve to enhance the interactions between helices (e.g. the M2/M2′, M1/M3, and M1′/M3′ interfaces). The substrate TPP\(^+\) molecule is shown in space-filling spheres, with light blue corresponding to carbon and yellow to phosphate atoms. (d) Blue spheres indicate four positions where mutations to Cys abolish functionality, and green spheres indicate positions that change resistance to only one of the transporter’s cognate substrates.\(^{45}\) Orange spheres mark positions that are involved in substrate binding.\(^{18,19}\) All of the blue spheres map to positions at the interfaces between the helices, where mutations might disrupt protein folding or oligomerization, or around the binding chamber. Most of the light blue spheres map to positions around the translocation chamber at least in one of the monomers, where changes to the surface of the protein might modify substrate recognition. The orange spheres are all located around the Glu14 residue. A listing of all residues in the TM domain and their experimentally determined structural or functional roles is provided in Table 1.
the topology analysis reported by Daley et al.,30 who showed that the predominant orientation of EmrE has the N and C termini in the periplasm. The cross-linking data identifying helix–helix interactions46 are difficult to reconcile with our model, and will require an atomic-resolution structure to be determined before the conflict can be resolved, as was the case for the cross-linking data for the lactose permease.48 The most internally consistent cross-linking data showing that helix M4 lies parallel with and adjacent to M4 from the neighboring monomer could be explained easily by suggesting that EmrE is a tetramer in the membrane, related by a 2-fold perpendicular to the membrane,4 a proposal that is supported by recent data from studying the interaction between peptides representing individual TM regions of Hsmr, the archaeal homologue of EmrE.49

**An alternate-access mechanism for substrate translocation**

Transport of drug substrates from the cytoplasm or cytoplasmic leaflet of the lipid bilayer out of the bacterium is thought to occur in essentially two steps3,4 (Figure 4(a)). First, the drug substrate binds

![Figure 4](image)

**Figure 4.** (A) A mechanism for proton-coupled translocation of substrates by the SMR family of proteins. (1) Two substrate-bound forms of the protein interconvert between conformations, in which the substrate, marked by an asterisk (*) faces the cytoplasm or the periplasm due to conformational changes. (2) In the periplasmic-facing conformation, the substrate is supplanted by the binding of two protons to the Glu14 positions (marked by $E^−$) on both monomers, thus driving the equilibrium towards substrate translocation. (3) A conformational change reorients the binding site towards the cytoplasm. (4) Substrate binding on the cytoplasmic side forces the protons out of the translocation chamber into the cytoplasm. (b) A suggestion for the conformational change represented by Step 1 in Figure 4(a). Periplasmic-facing and cytoplasmic-facing conformations of the EmrE dimer based on the cryo-EM structure. The transition between the two conformations involves a reorientation of the M1–M3 helices in both monomers by approximately 20° with respect to the in-plane axis of symmetry; a kinking and straightening of M3; and a small translation of M1–M3 in both monomers with respect to the M4 helices. As these changes occur in the protein dimer, the TPP$^+$ substrate, which is accessible from the near end in the conformation on the left, moves downwards and becomes accessible from the far end of the EmrE dimer in the right-hand conformation. Thus, interconversion between these two conformations could alter the accessibility of substrate from cytoplasmic facing to periplasmic facing. The conformation on the right was obtained by rotating the conformation on the left by 180° with respect to the in-plane axis of quasi-symmetry, so that the two conformations are completely superimposable. Thus, inverted topology would reproduce a single substrate-binding mode as two conformations, one of which is accessible to the periplasm and the other to the cytoplasm. The two monomers are arbitrarily marked A and B.
to EmrE, which induces a conformational change, so that the inward-facing binding pocket is opened to the periplasm and closed to the cytoplasm. The high concentration of protons in the periplasm competes directly with the drug for binding at the two Glu14 residues, so protonation results in release of the drug. A further conformation change then re-orients the protein to face the cytoplasm, where it may bind another drug molecule.

The nature of these conformation changes is uncertain, but the cryo-EM structure and our model suggest its basic features. To explain the substrate-translocation process, we propose that M1-M4 from monomer A adopt the conformation of M1-M4 observed in monomer B in our model and vice versa (Figure 4(b)) during the step marked as (1) in Figure 4(a). Due to the in-plane symmetry, this transition results in a structure identical with the original model rotated by 180° with respect to the in-plane axis of symmetry; the two symmetry-related structures are shown on both sides of the chemical equilibrium in Figure 4(b). To analyze the details of this transition, it is useful to divide the model structure into three subunits: (1) M1, M2 and M3 (monomer A); (2) M1′, M2′ and M3′ (monomer B); and (3) M4 and M4′.

Helices M1–M3 from monomer A are virtually superimposable on M1′–M3′ of monomer B, except for the kink in M3, suggesting that M1–M3 move as one unit during the transition described in Figure 4(b). The two M4 helices are seen to make minimal movements with respect to one another during the transition, suggesting that they are stable as a helix pair. Indeed, the cryo-EM and model structures show this pair to be closely packed with glycine residues (positions 90 and 97) lining the interhelix interface, which can stabilize helix packing.50 By contrast, the interfaces between helices M3 and M4 are small in both monomers in comparison to any of the other pairs of interacting helices in the structure (Figure 4(b)). It therefore comes as no surprise that the most significant conformational change that occurs during the transition can be localized to the contact region between M3 and M4, with the crossing angles between these helices changing by approximately 20° around the in-plane axis of symmetry in order to switch the M3-M4 packing from that observed in monomer A to that observed in monomer B, and vice versa. The kinking and straightening of the two M3 helices and a small translation of the M1-M3 helices in both monomers with respect to the M4 helices, coupled to the movement of the TPP+ molecule perpendicular to the membrane plane would then complete the transition. Thus, although residues on M4 have so far not been recognized as important for substrate binding and translocation (Table 1), this putative mechanism suggests a crucial role for M4 in stabilizing the dimer interface during the translocation process. The short M3-M4 loop, consisting of six residues in the SMR family, would hold the two parts of the structure together in the face of these relative motions. It is interesting to note in this connection that recent results have suggested a role for the M4 helix in mediating the formation of SMR tetramers.49

The sum of these conformational changes would alternately open the substrate-translocation chamber to the cytoplasmic and periplasmic media, allowing substrate to bind in the cytoplasmic-facing conformation, and then to be replaced by protons when the protein faces the periplasm (Figure 4(a)). Interestingly, this mechanism suggests that the periplasmic- and cytoplasmic-facing conformations of substrate-bound EmrE are essentially identical, and would thus require a single substrate-binding mode to be optimized structurally, which would then by symmetry be reproduced in both cytoplasmic-facing and periplasmic-facing conformations. Currently, there are only two conformations of EmrE (Figure 4(a), the upper panels) for which we have structural information (Figure 4(b)),7,8 and the structure of further transport intermediates (i.e. Figure 4(a), lower panel) will be essential to identify conformational changes that occur during the transport cycle. However, the availability of our model will now allow the design of specific experiments, such as using site-specific spin labels to monitor movements in EmrE during the transport cycle.

**Discussion**

The suggestion of dual topology of EmrE,7 and the recent support for this from global-topology analyses,30,31 were the key for the successful modeling of EmrE presented here. The presence of 2-fold quasi-symmetry between the monomers of the EmrE dimer within the plane of the membrane (Figure 1(b)) implied an antiparallel orientation of the EmrE monomers. Our previous modeling attempts (not described), which were not guided by the in-plane pseudo 2-fold axis, were unsuccessful in providing explanations for the biochemical and biophysical observations on EmrE. In contrast, our model with the monomers in an antiparallel orientation explains virtually all the biochemical and biophysical data. The model makes many predictions about the structure of EmrE that will provide a platform for further experimental work, such as the identification of other residues in the translocation pathway that have not yet been studied (Table 1), and residues that may be important in mediating helix packing, and therefore could be involved in the conformational changes.

The suggestion of oppositely oriented monomers in EmrE has been made only recently,7 and is reinforced by global topology analysis of bacterial proteins,30,31 but, so far, mechanistic advantages of dual topology have not been proposed. The mechanism of translocation that we invoked above suggests a potential advantage. That is, if the cytoplasmic-facing and periplasmic-facing conformations of substrate-bound EmrE are essentially identical, then only one mode of substrate binding should be
devised by evolution, which would be replicated as two conformations, one facing the cytoplasm and another facing the periplasm. This might also provide partial solution to a long-standing puzzle in SMR research; namely, how these small proteins consisting of roughly 100 amino acid residues can catalyze the coupled translocation of substrate and protons, a feat that is accomplished in other antiporter families, such as the major-facilitator family, by much larger proteins. Thus, inverted topology might be a parsimonious evolutionary solution to the problem of vectorial transport. In this connection, it is interesting to note that two of the five proteins identified as having the dual-topology architecture are from the SMR family (the others have not been fully characterized mechanistically), and a sixth case of dual topology was identified involving two homologous proteins (YdgE and YdgF), which are also SMR members that are likely to have arisen from a gene-duplication event.

Although much of the biochemical and biophysical data gathered on EmrE are in harmony with the model structure, there are one or two pieces of data that are not in agreement. The topology of the protein is clearly the most important point of disagreement, because dual topology provided the basis for the model structure reported here and for the suggested mechanism of substrate translocation; ultimately, if inverted topology for EmrE is incorrect, then so is the model. We have found that dual topology provides the most satisfactory model for EmrE, but Ninio et al. predict, on the basis of labeling data, that the monomers have identical topology, conflicting with other lines of experimental data that suggest inverted topology for EmrE. The reason for this discrepancy among different lines of experimental data are unclear. The possible conflicts of our EmrE model with the cross-linking data have not been fully characterized mechanistically.

Despite many years of structural studies of the SMR transporter EmrE, an atomic-resolution structure of this representative protein that can explain much of the biochemical and structural data has not emerged. Here, we have used phylogenetic analysis combined with constraints obtained from a cryo-EM structure of EmrE and some biochemical experiments in order to produce a model structure specifying the approximate positions of individual amino acid residues for EmrE and its homologues. Although this model was constrained only by some biophysical data on EmrE, it is encouraging that the model is capable of accounting for so much of the biochemistry. By revealing the locations of individual amino acid residues in the membrane-spanning regions, the model can be used in order to plan and interpret experiments aimed at deciphering the molecular details of the substrate-translocation mechanism in EmrE and its homologues.

**Methods**

**Sequence data**

An initial alignment of a few tens of EmrE homologues was constructed using CLUSTAL W. On the basis of this alignment, we then constructed a hidden Markov model (HMM), which was then calibrated and used to search SWISSPROT and TrEMBL for additional sequence homologues. Sequences showing over 90% identity with other sequences in the set were removed to obtain 98 sequences, which were then aligned. Conservation scores were then computed for each amino acid position using the ConSeq server and the Rate4Site algorithm. The sequence alignment was inspected to identify hydrophobic stretches that correspond to the hydrophobic cores of the helices in forming the TM domain. Starting from the secondary structure assignment derived from NMR, we manually modified the N and C termini of each hydrophobic domain so that the longest stretches of hydrophobic residues would be included. The following segments of EmrE were used as the hydrophobic stretches: TM1, 4–21; TM2, 34–52; TM3, 58–80; TM4, 87–104. The conservation scores and the hydrophobic segments are shown in Figure 1(b).

**Conformation scoring function**

The method for conformational search was as described. In brief, this scoring function favors the burial of evolutionarily conserved amino acid positions in the protein core and the exposure of variable positions to the lipid, without biasing helix orientations according to experimentally derived data. Conformations that expose charged amino acids to the lipid milieu are penalized (in EmrE, this applies only to M1 due to position Glu14). The following scoring function is used to score each conformation:

\[
\text{Score} = \sum_i (2(B'_i - 1/2)(H'_i - C'_i))
\]

where \(B'_i\) quantifies the extent of burial of amino acid residue \(i\) in the protein core. It assumes values of 0 to 1, with 1 signifying complete burial against another helix, and 0 signifying complete exposure to the lipid or the pore lumen. The function is computed by iterating over all of the helices in the structure other than the one on which \(i\) is located, and taking into account distance from, and orientation of \(i\) with respect to each of these helices. \(B'_i\) is then taken as the maximum of the values calculated for each of the helices. Thus, high values of \(B'_i\) imply that \(i\) is in close contact with another helix, whereas low values indicate that it is not interacting with any of the helices.

The \(C'_i\) values are the normalized evolutionary-rate scores assigned by Rate4Site. High-through-low values of \(C'_i\) are assigned to variable-through-conserved positions, respectively. \(H'_i\) is the free energy of transfer from water to lipid of amino acid \(i\) according to the Kessel and Ben-Tal scale. \(H'_i\) values are taken into account only if they are greater than 7 kcal/mol, and only for residues \(i\) that are exposed to the membrane, i.e. for which the burial scores \(B'_i\) are less than 0.5. Thus, the hydrophobicity scale

† The multiple-sequence alignment of SMR proteins can be downloaded from http://ashtoret.tau.ac.il/~sarel/EmrE.html
serves as a significant penalty on the exposure of the most polar residues to the membrane environment.

Conformational search

Canonical Cα-trace models of eight α-helices were constructed according to the helix axes parameters derived from helical models that were made to fit the cryo-EM structure, and their geometric centers were placed at the hypothetical membrane midplane. The amino acid identities of positions in the hydrophobic segments M1–M4 were assigned to the relevant positions on these helices.

Each helix was rotated around its principal axis independently, in 5° steps, and its optimal orientation was derived. Then, the optimal orientations of all helices were superimposed to yield the optimal conformation of the entire complex.

Data Base accession number

The cryo-EM structure is available from the EM data bank with accession code 10877. The coordinates of the model structure of a dimer of EmrE containing backbone atoms has been deposited in the PDB with accession number 2i68.

Acknowledgements

The authors gratefully acknowledge many comments and discussions with S. Schuldiner regarding biochemical and biophysical data, and their implications, for providing unpublished data. 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. was supported in part by the IST Programme of the EU as shared-cost RTD (FET Open) Project under Contract No IST-006413 (ACS - Algorithms for Complex Shapes), and by the Hermann Minkowski-Minerva Center for Geometry at Tel-Aviv University.

References


‡ http://www.ebi.ac.uk/msd/index.html


Edited by J. Bowie

(Received 19 July 2006; received in revised form 25 August 2006; accepted 25 August 2006)

Available online 30 August 2006
Prediction and simulation of motion in pairs of transmembrane α-helices

Angela Enosh1,*, Sarel J. Fleishman2, Nir Ben-Tal2 and Dan Halperin1
1School of Computer Science and 2Department 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) (Karlus 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;
Motion in pairs of transmembrane α-helices

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. Apaydın 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 to the receptor tyrosine kinase (RTK) ErbB2 and Glycophorin A.

2 A TM PROTEIN MODEL

A protein can be described as a long linkage with side-chains attached to the Cα 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).

The conformational 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 to the receptor tyrosine kinase (RTK) ErbB2 and Glycophorin A.

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 K, which is the percentage of the vdW radii, namely the centers of two non-bonded atoms of vdW radii r1 and r2 must be at least K(r1+r2) apart. For our experiments, we used 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 {h1…hn}. Each helix has six dofs corresponding to its position and orientation.

2.3 Relaxations applied to the TM helices

If a helix h, has mₜ torsion angles, the dimensionality of the configuration space in our problem is enormous with 6n + ∑ₙₜ=1 (mᵣ - 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
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, $C_{\text{space}}$, is the union of all these possible conformations. $C_{\text{space}}$ is divided into feasible, $C_{\text{feasible}}$, and forbidden, $C_{\text{forbidden}}$, regions. $C_{\text{forbidden}}$ contains all the conformations that involve steric clashes between atoms (both within and between helices). In addition, $C_{\text{forbidden}}$ contains conformations that involve low vdW interactions between the helices. $C_{\text{feasible}}$ is simply $C_{\text{space}} \setminus C_{\text{forbidden}}$.

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 $T$ rooted at the initial conformation $q_{\text{init}}$. $T$ is incrementally grown to efficiently explore the feasible conformation space in order to find a feasible path connecting $q_{\text{rand}}$ to a goal conformation. In each iteration, a random conformation, $q_{\text{rand}}$, is generated and the nearest node, $q_{\text{near}}$, in $T$ (according to some appropriate distance metric $M$) is expanded towards $q_{\text{rand}}$. If no collision is found on the way towards the random conformation, then $q_{\text{rand}}$ becomes a new vertex in the tree and an edge is added between $q_{\text{near}}$ and $q_{\text{rand}}$. Otherwise, $q_{\text{near}}$ expands as close as possible towards $q_{\text{rand}}$. In this case, the last feasible conformation (unless it is too close to $q_{\text{near}}$) becomes a vertex in $T$ and an edge is added between $q_{\text{near}}$ and the new vertex (Fig. 2). It was shown (LaValle et al., 2001) that this method leads to Voronoi-biased growth of $T$. This means that vertices with large Voronoi cells 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 $q_{\text{init}}$. 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 $q_{\text{near}}$ towards $q_{\text{rand}}$ a forbidden conformation, $q_{\text{forbidden}}$, may occur. $q_{\text{forbidden}}$ 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 $D$ (we use $D = 14$ 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 $q_{\text{rand}}$. In a way that the adopted conformation will be free of collisions. In case of a success, $q_{\text{near}}$ continues to expand towards $q_{\text{rand}}$. 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 $\phi$ and $[-30, -50]$ for $\psi$. 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 $T$ for a threshold number of consecutive iterations, it implies that the sampled conformations in $T$ cover $C_{\text{feasible}}$ sufficiently, and the expansion of $T$ 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.
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 \( \alpha \)-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 \( \alpha \)-helix bundle conformation was calculated as

\[
E_{vdw} = \sum_{i,j} \varepsilon_{ij}\left[\frac{r_{ij}}{r_{ij}^*}\right]^12 - 2\left[\frac{r_{ij}}{r_{ij}^*}\right]^6,
\]

where \( r_{ij} \) is the distance between atoms \( i \) and \( j \), \( \varepsilon_{ij} \) is the energy-well depth and \( r_{ij}^* \) is the atomic radii sums. The parameters were taken from CHARMM19 (Neria et al., 1996). Thus, a weight was assigned to each node in \( T \) 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 \( P \) that equals to the highest value of the potential that is observed along the nodes and edges in the pathway, and a goal function \( G \) 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 \( e_i \) for an edge, \( P(\pi) = \max_{0 \leq i \leq k, \sigma \in \mathbb{R}^d, 0 \leq \ell \leq k} \{W(v_\ell) | W(e_\ell)\} \), where \( W \) is the weight of the nodes or edges in \( T \), and \( G(\pi) = W(v_k) \).

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 \( Q \); in our experiments we use \( Q=1.4 \ \text{Å} \)]. For a cluster \( C_i = \{\pi_1, \ldots, \pi_m\} \), a representative path \( \pi^* \) was chosen to be the one that minimizes the LJ potential in the conformations stored on the path edges and nodes, i.e. \( P(\pi^*) = \min_{\pi \in C_i} \{P(\pi)\} \). Different paths may comprise different lengths (number of nodes in the path), still, the above criterion (minimizing \( P \)) is more dominant than the path lengths. However, if several paths in a cluster had the same values \( P(\pi^*) \), 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 \( G(\pi^*) \) 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 \( F(C_i) = \sum_{\pi \in C_i} e^{-G(\pi^*)} \). 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, over-expression 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 the 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 \( C_n \) atoms forming canonical \( \alpha \)-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
which corresponds to the TM domain of ErbB2 [LTSIVSAVV-GILLVVVLGVVFGIL], were built as canonical \( \alpha \)-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. \( K = 60\% \), 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 first by 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 the Ser656-xxx-Gly660 motif as suggested previously (Fleishman et al., 2002).

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 \( \pi \) 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 \( \alpha \) 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 \( \ell_1 \) and \( \ell_2 \) of the helices, we choose an axis of symmetry, namely a line \( \ell \) such that rotation of \( \pi \) radians around \( \ell \) will align \( \ell_1 \) with \( \ell_2 \). Further details can be found in the Supplementary Material.
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 $35^\circ$. 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 $-45^\circ$. 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$1.4 Å 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 Glycoporin 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 generalized to TM proteins with an arbitrary number of helices and

References


Conflict of Interest: none declared.
Generation, comparison and merging of pathways between protein conformations: Gating in K-channels

Angela Enosh\textsuperscript{1,2}, Barak Raveh\textsuperscript{1,3}, Ora Furman-Schueler\textsuperscript{3}, Dan Halperin\textsuperscript{1} and Nir Ben-Tal\textsuperscript{2,*}
\textsuperscript{1}School of Computer Science, Tel Aviv University, Ramat Aviv 69978, Israel.
\textsuperscript{2}Department of Biochemistry, Tel Aviv University, Ramat Aviv 69978, Israel.
\textsuperscript{3}Department of Molecular Genetics and Biotechnology, The Hebrew University, Hadassah Medical School.

Keywords: Structural bioinformatics, robotics, motion planning, conformational changes, Probabilistic RoadMap (PRM), Rapidly-exploring Random-Tree (RRT)

Running title: Exploring conformational changes

*CORRESPONDING AUTHOR AT: PHONE (972-3) 640-6709; FAX (972-3) 640-6834; E-MAIL: \texttt{BENTAL@ASHTORET.TAU.AC.IL}; WEB: \texttt{HTTP://ASHTORET.TAU.AC.IL}
ABSTRACT

We present here a general framework for the generation, alignment, comparison and hybridization of motion pathways between two known protein conformations. The framework, which is rooted in probabilistic motion-planning techniques in robotics, allows the efficient generation of collision-free motion pathways, while considering a wide range of degrees of freedom involved in the motion. Within the framework, we provide the means to hybridize pathways, thus producing, e.g., the motion-pathway of lowest energy barrier out of the many pathways proposed by our algorithm. The method for comparison and hybridization of pathways is modular and may be used also within the context of molecular dynamics and Monte Carlo simulations. The whole framework was implemented within the Rosetta software suite, where the protein is represented in atomic detail. K-channels switch between open and closed conformations, and we used the overall framework to investigate this transition. Our analysis suggested that channel-opening may follow a three-phase pathway: First, the channel unlocks itself from the closed state; second, it opens; and third, it locks itself in the open conformation. A movie that depicts the proposed pathway is available in Supplementary Material and at: http://www.cs.tau.ac.il/~angela/SuppKcsA.html.
INTRODUCTION

K-channels are found, in essence, in all kingdoms of life and all types of cells (1). They are best known for their function in excitable cells. For example, they are involved in the generation and propagation of nerve impulses in the synapse and neuron (1). Mutations in the proteins that form the channel may lead to diseases, such as multiple sclerosis, cystic fibrosis and cardiac arrhythmia (2). Because of their involvement in these and other channelopathies, i.e., channel-related diseases, they are major drug targets. K-channels are diverse in their sequence and mechanism of gating, i.e., the processes by which the pore opens and closes. However, they all share a similar pore structure, and exhibit very similar ion permeability characteristics. Here, we focused on the pore domain. The structure of the domain was revealed by X-ray crystallography studies of the bacterial K-channel from *Streptomyces lividans* (KcsA; (3, 4)). Since then, several more K-channel structures of various sources emerged (e.g., (5-9)), and we now know the structure of some K-channels in their open and others in their closed conformations. For example, the voltage-dependent K-channels from *Aeropyrum pernix* (KvAP; (10)) and the *Methanobacterium thermoautotrophicum* (MthK; (11)) are considered to be in open conformations, while KcsA is known in a closed conformation. Here we present a novel framework to explore conformational changes in proteins, and use it to suggest a pathway between a closed state of the KcsA channel (4) and a model structure of the open state of the channel.

The pore-forming region in the K-channel is composed of four identical monomers that oligomerize around the channel pore (Fig. 1, Fig. 2a). Each monomer contributes a pair of transmembrane helices TM1 (blue, Fig. 1) and TM2 (green, Fig. 1) that are connected by a re-entrant loop. This loop, located at the extracellular end of the channel, contains the selectivity filter (yellow, Fig. 1), which is tuned to select potassium over, e.g., sodium, ions. The pore gate is found at the intracellular region of the TM2 helices (Fig. 1).

There is much evidence that the KcsA channel switches between open and closed conformations (12-17). The computational effort to provide molecular models of the conformational changes includes coarse-grained normal-modes analysis, (e.g., (18)). In one particularly interesting study, using the Gaussian-network method, Shrivastava and Bahar (19) suggested that channel-opening follows a corkscrew motion of the intracellular regions of the channel. This suggestion recently gained strong experimental support from innovative single molecule studies (20). The experimental results were interpreted to indicate a corkscrew motion of the intracellular ends of the TM2 helices, and provided the first direct insight into the large-scale nature of the motion. However, experimental methods are still unable to give a detailed atomic account of the motion, which is of great interest for mutational analysis and drug targeting of transition states.

The transition between the open and closed conformations involves major conformational changes and takes place within microseconds or more (21). The large amplitude of the motion, both in time and space, is beyond the reach of standard molecular dynamics simulations, and external biasing was used to trace it. For example, Biggin and Sansom used steered molecular dynamics and simulated channel-opening by slowly inflating a sphere which was placed at the center of the (closed) gate (21). Tikhonov and Zhorov used a similar approach, by applying full-atom Monte-Carlo simulations while iteratively inflating and deflating a cylinder at the pore axis (22).
In this study, we have developed a very different framework to rapidly explore many putative conformational trajectories, align them using a novel dynamic-programming approach and cluster them into representative motion pathways. Within this general framework, we were able to find plausible low-energy trajectories of the channel-opening either with or without a biasing constraint.

The basic motion-planning problem can be stated as follows: Given a robot moving in an environment cluttered with obstacles, and given a start and a goal configurations for the robot, find a collision-free path connecting these two configurations. (Notice that configuration is the common robotics term for conformation.) The motion-planning problem is fundamental, originally studied in robotics and computational geometry, but with implications in numerous other fields (23-25). A class of randomized-path planning methods, known as Probabilistic RoadMap (PRM) methods, has been successfully applied to complicated high-dimensional problems (26-28).

A variant of the probabilistic techniques focuses on single-query strategies, where 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 roadmap for each query by growing trees of sampled milestones rooted at the start and goal configurations (27). Rapidly-exploring Random-Trees (RRTs) have been recognized as a useful tool for designing efficient single-query paths in highly constrained spaces (29, 30). We used RRTs in the current work to predict plausible motion pathways of the KcsA channel.

Probabilistic motion-planning techniques have been applied in the context of molecular biology (31-34). For example, Cortes et al. (35) developed an algorithm to compute large-amplitude motions in flexible molecular models, using RRTs to compute protein loop conformational changes and ligand trajectories. Later, the same group integrated normal modes analysis with path-planning techniques for the study of large conformational changes in proteins (36). In an earlier work we applied RRTs to predict stable conformations and motion in pairs of transmembrane helices (37).

Here we employ RRTs in a more demanding application, involving a full-fledged protein; we deal with a significantly larger number of dofs and search in a much higher dimensional space. In order to handle the vast dimensionality, we build the open conformation of the KcsA channel by homology modeling, using the open-conformation of KvAP as a template (Fig. 2a, right panel). Then, we used the RRT technique to generate many collision-free pathways that connect the known structure of KcsA in its closed state (Fig. 2a, left panel) to this (putative) conformation. Specifically, two conformation-trees, rooted at each of the two KcsA states, were grown simultaneously, with the aim to merge these trees. The trees were grown using energetic considerations. Thus, each path in the merged tree, between the nodes that are the roots of the original trees, corresponds to a feasible motion pathway of KcsA.

The RRT method resembles Monte-Carlo simulations in that both generate conformations at random. However, RRT 'looks back' at all the previous conformations when generating new ones. Thus, it is biased towards unexplored regions of conformational space. In addition, RRT may yield more than one trajectory in each run. Overall, the sampling using RRT usually provides a more complete picture of the conformational space. It is particularly suitable to the search in constrained environments, such as torsion-angle space of a large protein, where motion is highly restricted due to ample steric clashes.

The major innovation in our current work is the introduction of an efficient technique to cluster and hybridize multiple pathways between the initial and goal conformations. We are not aware of any similar feature within the context of path planning. To this end, we devel-
oped the means to compare and align pairs of pathways, using dynamic programming. In this respect, the alignment of pathways is analogous to the alignment of sequences in BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi). The fitting of paths is used in our work to meaningfully cluster similar pathways. The path-aligning technique is also used to hybridize paths as a means of enhancing the performance of the RRT search further towards finding pathways that are short and energetically-favorable.

We note here that the path-fitting algorithm that we propose is reminiscent of curve-fitting techniques previously used to match curves in two- and three-dimensions (38) which, in turn, are close to the dynamic programming solution to the longest common subsequence problem (39). Unlike these other problems, our path-alignment algorithm works in a very high-dimensional space and requires various technical adaptations.

THEORY

Pathways clustering
Due to the random nature of the RRT algorithm, different runs produce different pathways. The pathways may partially overlap with each other and it is challenging to analyze them in search for the shortest, energetically-favorable, pathway between the initial and goal conformations among them. A first step towards this goal is the clustering of the pathways. We present here a general-purpose approach to this end. Pathways are clustered based on their similarity, which is defined according to the distance between the conformations along the pathways and their sequence of appearance within the pathway. A dynamic programming algorithm is developed for that. The pathways within each cluster are automatically hybridized to produce a unique, short and representative pathway. The pathways are ranked using an energetic criterion, and the energetically most-favorable ones are selected.

A dynamic-programming algorithm for measuring similarity between pathways
We use a variant of dynamic-programming to find the best-alignment between two pathways (Fig. 3a). A pathway is defined as an ordered sequence of conformations and its length is the number of conformations that it contains. For two pathways, P and Q, of lengths m and n, respectively, we define M to be an mxn root-means-square deviation (rmsd) matrix, where the rmsd between the conformation at position i along the pathway P and conformation at position j along the pathway Q is stored in entry M_{i,j}. The M_{i,j} values are normalized to be within the interval (31), and the normalized matrix is denoted by M'_{i,j}, where M'_{i,j}=0 corresponds to rmsd of zero between the corresponding conformations, and M'_{i,j}=1 to conformations that have the lowest resemblance, i.e., their rmsd is above a similarity threshold (1.5Å here).

Let p_1,p_2,...,p_m and q_1,q_2,...,q_n denote the conformations along the pathways P and Q, respectively. We regard each pathway as a word with the constituting conformations as letters, and we allow zero or more spacing between letters in each word (Fig. 3a). A valid fitting between the pathways P and Q is obtained by aligning the two respective words one on top of the other, such that in each column at least one letter is not a space (depicted below as a dash). Here are examples of valid fittings of

P={p_1,p_2,p_3} and Q={q_1,q_2,q_3,q_4} (see Fig. 4):

\begin{align*}
p_1 & \quad p_2 & \quad p_3 & \quad - \quad , \quad p_1 & \quad - & \quad p_2 & \quad - & \quad p_3 & \quad - & \quad - \\
q_1 & \quad q_2 & \quad q_3 & \quad q_4 & \quad - & \quad q_1 & \quad - & \quad q_2 & \quad - & \quad q_3 & \quad q_4
\end{align*}
We next describe all possible valid fittings between two pathways P and Q as above (Fig. 4). Let G = (V,E) be a directed graph, whose vertex set V consists of vertices vi,j (for i=0...m, j=0...n) arranged in a grid – and that has three outward directed edges from each vertex vi,j (i=0...m-1, j=0...n-1): (i) toward vi+1,j, (ii) toward vi,j+1, and (iii) toward vi+1,j+1. The vertices vm,j (j=0...n-1) have one outward directed edge toward vm,j+1, and the vertices vi,n (i=0...m-1) have one outward directed edge toward vi+1,n. Each directed path from v0,0 to vm,n describes a valid fitting of P and Q in the following manner: a vertical edge (vi,j, vi+1,j) corresponds to aligning the conformation pi with a gap, a horizontal edge (vi,j, vi,j+1) corresponds to aligning the conformation qi with a gap, and the diagonal edge (vi,j, vi+1,j+1) corresponds to aligning the conformation pi with qi. It is possible to see that every valid fitting of the pathways P and Q corresponds to a directed path from v0,0 to vm,n in G and vice versa (Fig. 4). We will use fittings and directed paths in G interchangeably.

We now assign a cost to a fitting by assigning weights to edges along the directed path in the graph. Horizontal and vertical edges are assigned equal weights, which we denote by GapPenalty. (In our experiments, GapPenalty = 1 for reasons to be clarified shortly.) A diagonal edge (vi,j, vi+1,j+1) is assigned the cost of the normalized rmsd M’i,j, defined above. M’i,j measures the similarity between the conformations pi and qi that have been aligned, and is a real value between 0 (identity) to 1 (greatest dissimilarity). In our experiments, we have chosen to penalize gaps with the same score as that of greatest dissimilarity, although other choices can be incorporated.

Rephrased in terms of G, our best pathway fitting problem is to find the path of minimum weight between v0,0 and vm,n in the directed graph. Let Pi denote the pathway consisting of the first i conformations of the pathway P. Qi is similarly defined. Our optimal pathway-fitting problem has the following optimal substructure property. The optimal fitting is either (i) the optimal fitting of Pm-1 and Qn-1 concatenated with the fitting of pm and qn, or (ii) the optimal fitting of Pm and Qn-1 concatenated with the fitting of a gap and qn, or (iii) the optimal fitting of Pm-1 and Qn concatenated with the fitting of pm and a gap. The same arguments apply to any pair of sub-pathways Pi and Qi for i>0, j>0. This leads to the following dynamic programming formulation. DP is an (m+1)×(n+1) matrix, whose element DP[i,j] stores the score of the optimal fitting of the sub-pathways between Pi and Qi. Therefore, our goal is to compute DP[m,n] and deduce the optimal fitting.

\[
\begin{align*}
\text{for } i = 0 \ldots m & \quad DP[i,0] = i \times \text{GapPenalty} \\
\text{for } j = 0 \ldots n & \quad DP[0,j] = j \times \text{GapPenalty} \\
\text{for } i = 1 \ldots m & \\
\text{for } j = 1 \ldots n & \\
\quad \text{if } (M_{i,j} \leq \text{SimilarityThreshold}) & \quad DP[i,j] = \min (DP[i-1,j-1]+ M’_{i,j}, \\
\quad & \quad \quad \quad \quad \quad \quad \quad \quad DP[i-1,j]+\text{GapPenalty}, \\
\quad & \quad \quad \quad \quad \quad \quad \quad \quad DP[i,j-1]+\text{GapPenalty}) \\
\quad \text{else} & \quad DP[i,j] = \min (DP[i-1,j]+\text{GapPenalty}, \quad \quad \quad \quad \quad \quad \quad \quad DP[i,j-1]+\text{GapPenalty}) \\
\text{return } DP[m,n] 
\end{align*}
\]
The optimal fitting can be easily deduced from the table DP without extra storage by tracing back the optimal solution from the last alignment backwards. The matrix DP can be computed in O(mn) time, using O(mn) space (or less space, using standard tricks; (39)).

**Clustering**

Having defined a similarity score between pathways, one can cluster the pathways using, e.g., hierarchical clustering (40). Given n objects, hierarchical-clustering assigns each object to a cluster. Then, it applies a series of fusion steps to the n clusters until all the objects are clustered into a single cluster of size n. In each fusion step, the two most similar clusters are merged. There are several methods to define distance (i.e., dissimilarity) between clusters.

Here we use the complete linkage clustering method. Distances between clusters of pathways are computed by the farthest neighbor method. In other words, the distance between two clusters C_k and C_s is defined as: d(C_k, C_s) = max{PD[i,j] | pathway i is in C_k, pathway j is in C_s}. Given n pathways, PD^1 (a distance-pathway table) is an n×n table, where PD[i,j] is the distance between pathways i and j as defined by the dynamic-programming algorithm; clearly, PD[i,i]=0 and PD[i,j]=PD[j,i].

In each fusion step of the hierarchical clustering procedure, the clusters C_k and C_s of minimal d(C_k,C_s) are merged. Clustering proceeds until the clusters are dissimilar from each other based on a predefined criterion. Suppose that the clustering method attempts to merge clusters C_k and C_s. If there exist two pathways: P of length m in C_k and Q of length n in C_s, such that the number of matching conformations along them is below N×min(m,n), the hierarchical-clustering algorithm stops. A value of N=0.7 is used here. In other words, each pair of pathways in the same cluster contains overlapping conformations (within rmsd of 1.5Å) of at least 70% of the shortest pathway. Preliminary examination showed that similar pathways are obtained using slightly different cutoff values.

**Hybridization of the pathways within a cluster and the selection of the best pathway**

After the clustering step, the pathways are ranked and a representative prominent pathway for each cluster is constructed. Intuitively, the representative pathway can be selected based on the following min-max criterion: The pathway, whose conformation of maximal energy is assigned the lowest value amongst all pathways. However, it may be possible to construct a better pathway by merging energetically-preferable sub-pathways. To this end, we exploited the alignment that was produced by the dynamic programming step above.

A pathway-merger was built, in search for a short pathway that compromises low energy conformations (Fig. 3). There are two cases in the alignment between two pathways that can help in the production of a better representative pathway:

(i) An alignment of gaps in one pathway to a sequence of conformations on the other pathway

(Fig. 3b). For example, let us examine the following alignment that was computed for pathways P and Q:

\[
\begin{align*}
\text{p}_1 & \quad \text{p}_2 & \quad \text{p}_3 & \quad \text{p}_4 & \quad \text{p}_5 & \quad \text{p}_6 & \quad \text{p}_7 & \quad - & \quad \text{p}_8 & \quad \text{p}_9 \\
\text{q}_1 & \quad \text{q}_2 & \quad \text{q}_3 & \quad - & \quad - & \quad \text{q}_4 & \quad \text{q}_5 & \quad \text{q}_6 & \quad \text{q}_7
\end{align*}
\]

It is possible to shorten these pathways and obtain the following pathways: pq = \{p_1, p_2, p_3, p_7, p_8, p_9\}, and qp = \{q_1, q_2, q_3, q_4, q_6, q_7\}. These pathways are feasible provided that the local planner will succeed to connect p_3 to p_7, and q_4 to q_6. And if so, they are preferable since they are shorter and associated with energy that is equal to, or lower than, the energy of the pathways.

---

1 PD stands for Distance Pathway, to distinguish from the Dynamic Programming matrix introduced earlier and denoted DP.
original pathways. The merger computes the maximum energy potential of the conformations in these two short alternative-pathways and selects the one with the lowest value. Based on the alignment it is also possible to choose the best conformation at each step. For example, if energy ($p_1$) < energy ($q_1$) and energy ($q_2$) < energy ($p_2$), we will assign $p_1$ to the first conformation in the representative pathway and $q_2$ to the second, provided that the local planner finds a transition between these conformations.

(ii) Two consecutive gaps, one in each pathway (Fig. 3c). For example, let us examine the following alignment:

```
p_1 p_2 p_3 p_4 p_5 p_6 - - p_7 p_8
q_1 q_2 q_3 - - q_4 q_5 q_6 q_7
```

The following hybridized pathways may be produced: $pq = \{p_1, p_2, p_3, q_4, q_5, p_7, p_8\}$ and $qp = \{q_1, q_2, q_3, p_4, p_5, p_6, q_6, q_7\}$. In other words, the two consecutive gaps, one on each pathway, suggest the possibility of choosing a sub-pathway to be within the representative pathway. Again, the merger selects the sub-pathway (among $\{p_4, p_5, p_6\}$ and $\{q_4, q_5\}$) that minimizes the maximal energy of its conformations.

In summary, it is possible to use the gaps in the best fitting procedure using the dynamic-programming approach to detect short and energetically-favorable pathways. In practice, a representative-pathway was selected, iteratively aligned and hybridized with all the other pathways as described above.

**METHODS**

A scheme, describing the overall approach is presented in Fig. 2. The different steps in the scheme are described below.

**A model structure of the KcsA channel in its open state**

The KcsA channel has been solved in its closed conformation (PDB-id 1k4c, (4)). We modeled the open form of KcsA using the structure of KvAP (PDB-id 1orq, (10)), which was determined in an open conformation, as a template. Of the available templates, KvAP has the highest sequence similarity in the pore domain (~30% sequence identity). KvAP has been previously suggested as a plausible template for modeling KcsA in its open form (41). More information about modeling and validation procedures, as well as the model structure (Fig. 2a) is provided in Supplementary Material:S1.

**The types of motions that were considered**

The KcsA tetramer includes 412 amino acids, each of which contributes two torsional backbone degrees of freedom to the overall dimensionality of the conformational space. Additionally, side-chain flexibility throughout the gating phase is considered. Thus, the number of dofs is enormously high, and we used the following measures in order to reduce the computational load while maintaining an appropriate representation that accounts for relevant degrees of freedom. First, following a widespread modeling assumption, bond lengths and angles were held fixed throughout the search in conformational space (42).

Second, all known structures of potassium channels, including those of KcsA (Fig. 2a) and KvAP, manifest a four-fold rotation symmetry around the pore axis. We assumed that this symmetry, which is anticipated for homotetramers, is retained throughout the pathway, thus, reducing the number of dofs four-fold.
Third, the structures of the open and closed states of the channel are very different from each other, and the root-mean-square distance (rmsd) between their C-alpha atoms is 5.95Å. However, the selectivity filter regions of these structures are virtually the same (rmsd of 0.33Å between the C-alpha atoms of amino acids T^{75}VGYG^{79}, see Supplementary Material:S2). Thus, the backbone of the selectivity filter was held fixed. Following the same logic, we also fixed the backbone of the turret loop between ARG52 and THR61. This loop is located in the extracellular region farthest from the gate region and connects the TM1 helix to the pore helix, which is in turn connected to the selectivity filter and TM2 helix (Fig. 1, gray). The two ends of this loop, remain fixed between the open and closed conformations and do not seem to relate to the inter-helices orientation (see Supplementary Material:S2). As proper treatment of the motion would require sampling of many conformations of this loop, we fixed this region as well. Significant differences between the structures are apparent only within TM1 and TM2. Hence, we allowed full movement of all backbone torsion angles of these helices, and of all the side-chains, including regions with a fixed-backbone.

In total, we allowed for 104 backbone degrees of freedom per monomer in our simulation (because of symmetry, this is also the total number of backbone dofs for the entire channel). In addition, all side-chain \( \chi \) angles were allowed to rearrange by sampling conformations from a backbone-dependent rotamer library (43): Each residue was represented by a number of dofs that corresponds to the number of \( \chi \) angles needed to define the side-chain conformation. This number ranges from 0 for small amino acids, such as GLY and ALA, to 4 for large and flexible amino acids, such as LYS, defined by 4 \( \chi \) angles. Since the side-chains were not treated symmetrically, this adds to each of the four monomers 83 residues with sampled side-chain dofs, totaling to 4*83=332 residues with sampled side chain dofs.

Algorithm for motion-paths generation
Pathways were generated based on a method that was presented in our earlier publication within different context (37). One major difference between these studies is that instead of constructing a single tree, rooted at the initial conformation (i.e., the native, closed conformation of KcsA), we followed a prevalent practice in robotics and constructed here two trees, rooted in the initial and goal (Fig. 2a) conformations. In short, two conformation trees were rapidly grown starting from both initial and target conformations and each path in the tree stands for a putative motion pathway. A single tree version of the algorithm was also implemented, in which only the initial conformation was used as input.

The algorithm was implemented within the Rosetta modeling project (42). Energy scoring of generated conformations was calculated with a version of Rosetta score12 that included attractive and repulsive VDW and hydrogen-bonding terms, and a statistical bias for prevalent Ramachandran torsion angles and side-chain rotamers. Side-chains of initial and goal conformations were optimized using Rosetta full-atom repack procedure, and side-chain intermediate conformations were optimized using Rosetta greedy rotamer-trial procedure (44). A detailed description of the algorithm is provided in Supplementary Material:S3.

RESULTS

Pathways between the open and closed conformations of KcsA
We applied the RRT algorithm, described in detail in Supplementary Material:S3, in search for pathways between the (model of the) open and closed conformations of the KcsA channel (Fig. 2a). KcsA was represented in atomic detail, and the conformational space was explored,
subjected to the Rosetta energy-function, as described in METHODS above. Low-energy conformations were added to the tree in order to efficiently-explore the free conformational space. Side-chain clashes were resolved and their interactions optimized using the Rosetta greedy rotamer-trial procedure (44). Two motion pathways, one from the open and the other from the closed conformation, were generated and connected through a pair of similar conformations (Fig. 2a, red dashed line in middle panel). A threshold of C-alpha rmsd of 1.6Å was used to connect nodes from the different trees. In practice, many pathways were connected between much closer conformations of up to 1.15Å apart.

Starting from the initial and goal conformations, we generated 100 pathways from 12 RRT trees. The overall run took less than one hour to complete (on 3 dual core dual CPU, AMD 2.2Ghz/1MB computers, on a Sun cluster). The RRT algorithm examined, on average, 50082 conformations per tree, and the final tree comprised 28300 feasible conformations, on average. Subsequently we aligned, merged and clustered the pathways using the hierarchical approach described in THEORY. A final pathway of lower energy (Fig. 5b) was generated by merging low-energy lags of motion from aligned input pathways. The overall procedure is outlined in Fig. 2d.

**A possible three-phase mechanism of channel opening**

In the simulation, the TM helices are readily observed following a curved rotational motion that resembles the opening of an iris, combined with a corkscrew motion of the intracellular ends of the helices to open the pore (Fig. 5a and Supplementary Material:S4). This is in agreement with previous computational predictions (19, 22) and a very recent single-molecule study that supports the suggested rotational mechanism for channel-opening (20). However, our analysis provides new insight: We suggest here that channel opening follows a three-phase mechanism, involving safe-lock (Fig. 5). The suggested phases, described from an intracellular perspective, are as follows:

- Following a phase of Brownian motion (conformations 1-5), the channel is unlocked from its closed state by a slight clockwise movement (conformations 6-12) that involves bridging over a significant energy barrier (Fig. 5b).
- Second, the channel opens through a bending-motion, in which the intracellular part of the TM2 helix slides over the neighboring TM2 helix, moving counter-clockwise and laterally away from the pore axis (conformations 13-18)
- Finally, the channel is locked back in the open conformation by a counter motion (conformations 19-23)

While the suggested mechanism obviously calls for further validation, it is interesting to examine its function as a control mechanism for channel-gating. By this mechanism, channels in their closed form are trapped in an energy minimum. Channel-opening can be triggered by an unlocking mechanism that involves a preemptive small movement, which allows a subsequent large-scale motion for opening of the channel. Such ‘security mechanism’ may reduce the frequency of unwanted events of channel-opening, which may have negative effects on a living cell.

In order to further characterize the predicted pathway, a detailed analysis of the channel-opening in the gate-region was performed. We used the program HOLE (45) to calculate the opening radii of the main axis of the channel. Analysis of two bottlenecks (Fig. 5c and 5d) along the channel-gate demonstrates the way that such three-phase mechanism might work. The channel-opening throughout the motion pathway is plotted at two bottlenecks that deter-
mine the minimum constriction of the pore, at depths of 23Å (red) and 35Å (green). The importance of these regions in the predicted motion pathway is observed when comparing these plots to the minimal channel-opening over the whole gating region (below 40Å depth), shown as a black curve, and representing the overall opening of the channel. After a phase of Brownian motion (conformations 1-5), the channel starts to unlock at depth of 35Å (conformations 6-12). In agreement with our suggested mechanism, the widening of the cavity at depth 35Å does not manifest itself in an overall opening of the channel (black curve). However, this cavity might provide the greater freedom of motion needed to allow the next phase: A sharp opening of a cavity at depth of 23Å (conformations 13-18), opening a bottleneck, and allowing free flow of the ions.

**From the closed to the open conformation using a single RRT tree**
We wanted to examine if an independent growth of an RRT-tree from its initial closed conformation can independently reach the open conformation. We used the RRT method in its original context, using a single tree (37). After exploring 30,000 conformations in a three-hours run, a conformation that is less than 1.6Å from the open-conformation was detected by the algorithm, with a direct motion-pathway from the initial conformation. Considering the vast number of degrees of freedom used in this work, we believe this demonstrates the efficiency of the algorithm and energy-function for sampling the conformational space. Hence, RRTs can be used to complement memory-less search procedures such as Monte-Carlo simulations, and even procedures like tabu-search (46) that use short-term memory.

**DISCUSSION**

We introduced here a unique framework for the very rapid generation, alignment and comparison of pathways between two known protein conformations. We also presented here an algorithmic implementation of these ideas, and used it to study conformational changes in the KcsA channel. The pathways were generated subjected to energetic considerations. The actual run took about an hour. But it is noticeable that the calculations begin with preparations, including the building of the open-state model-structure of KcsA.

**A putative safe-lock mechanism for the opening of the KcsA gate**
Our studies suggest a plausible pathway between the open and closed states of the channel, which might suggest a three-phase safe-lock mechanism for channel-opening. If such mechanism is indeed valid, it might lead to interesting implications, as ion channels take part in vital cellular processes (1). That said, as a computational model, it should be treated with appropriate caution and be subjected to subsequent validations. Even though the conformations along the pathway are stereochemically sound (Supplementary Material:S1), the pathway should best be regarded as a speculation that should be examined in wet-lab experiments. We made the pathway available, as a movie and a series of PDB files, to facilitate the design of such experiments (see Supplementary Material:S4). However, it is important to notice that different runs using different parameters and energy functions might lead to different pathways. In this sense, we believe that the main power of the framework that was presented here is its capacity to rapidly explore a large number of putative pathways under varying constraints. These may be used to generate hypotheses that may be tested in experiments.
A very recent single-molecule study of the KcsA channel was interpreted to indicate an overall corkscrew rotational motion of the intracellular ends of the TM2 helices (20). Our results, as well as the results of previous computational analysis using the Gaussian-network model (19), are in agreement with this new data. And we provide a (hypothetical) molecular interpretation of the single-molecule study.

Tikhonov and Zhorov (22) explicitly assumed lateral outward movement of the channel by inflating and deflating cylinder at the pore-region and applying subsequent Monte-Carlo simulations. Their predictions resemble ours but only to a certain extent (compare Fig. 5a. to Fig. 5 in (22)). A similar approach was used within the context of steered molecular-dynamics (SMD) in a detailed study by Biggin and Sansom (21). Unlike these two studies, we applied a symmetry operator over the four subunits of KcsA and obtained a symmetrical motion. We assumed that the motion is symmetrical due to the symmetry found in all potassium channels, but whether non-symmetrical motion is a methodical artifact or a real property of the KcsA channel is not fully clear. Interestingly, like in the work of Biggin and Sansom, the helices did not disintegrate throughout our simulations, although the change in their internal coordinates is responsible for the movement. Instead, the helices were gently-distorted in small increments, in agreement with the above study.

All in all, this computational investigation is complementary to recent experimental studies of KcsA using NMR (e.g., (47)), EPR (48) and single molecule techniques (e.g., (20)). Ultimately, by the integration of the experimental and computational results we will further our understanding of channel-gating significantly.

The role of the selectivity filter

Here, our motion pathways were restricted to the gate-region. The selectivity-filter region was examined exhaustively by molecular-dynamics simulations (reviewed, e.g., in (49)). Cordero-Morales et al. (50) suggested that the filter undergoes conformation excursions. However, gating events, which are on the order of microseconds, cannot be simulated by pure MD simulations. Biggin and Sansom (21) used steered MD to suggest possible open states for KcsA. In their simulations, the gate region underwent substantial conformational changes, while the selectivity filter region did not alter significantly by the gate-opening, suggesting that the gate-region dynamics is decoupled from that of the selectivity filter. Although the selectivity-filter might change its conformation during the opening of the gate (50), these modifications are relatively minor, albeit their being functionally-important, in comparison to the conformational changes that occur in the gate-reain. In fact, the rmsd between the C-alpha atoms of the source (pdb:1k4c) and target (a model of KcsA according to the KvAP structure) selectivity filters (T^{15}VGYG^{79}) is 0.33Å, in comparison to 5.95Å considering all the C-alpha atoms. Thus, generating motion-pathway between these two conformations by RRTs will naturally lead to immense variations in the gate-region rather than in the selectivity-filter.

The energy-function

Since this study involves a transmembrane (TM) channel, we excluded solvation-related terms from the Rosetta scoring function. We used Rosetta's Van der-Waals potential, hydrogen-bonding score and the statistical Ramachandran and Dunbrack rotamer scores. A designated energy-function for TM-proteins has been recently introduced into Rosetta (51), and we look forward to use it. Unfortunately it was not made publicly available yet. It would be interesting to examine the sensitivity of the suggested motion-pathway to different forcefields. We previously (37) applied the RRT algorithm using very simplistic energy functions that consisted of
the Lennard-Jones term of either GROMOS (52) or CHARMM (53). We repeated these tests
with KcsA. The pathways obtained using these simplistic energy functions differ from each
other in certain aspects and are similar in others. For instance, one of the pathways, observed
when using the CHARMM LJ-potential alone, leads to a fuzzy lateral opening of the TM2
helices (data not shown). This is indicative of the importance of the choice of the energy func-
tion. Since the quality of the Rosetta energy-function is still questionable, our results should
be taken with a grain of salt.

Interestingly, even though we did not explicitly-impose helix-favoring restraints, the heli-
ces retained their regular structures and, in essence, did not deform. This is attributed, in part,
to the fact that the ROSETTA energy-function includes a bias towards prevalent
Ramachandran torsion angles; when starting, as we did, from helical conformations and mov-
ing in small increments, it is difficult to deform the helices significantly. Further investigation
also showed that the low energy cutoff (of zero) that was used here essentially allows only
tightly packed conformations of the type that are found in real proteins and prevents sig-
ificant helix deformations. Control runs that we conducted using higher energy cutoff values (of
600 and 5000; energy is measured in the ROSETTA arbitrary units) showed the emergence of
much less-packed conformations that also included deformed helices (data not shown).

Taking into account the solvation component of the free-energy in an accurate way for a
membrane channel, such as KcsA, is difficult; perhaps even impossible yet. The standard
treatment within the ROSETTA energy-function excludes the membrane environment, and in-
volves a large, and unnecessary, de-solvation free-energy penalty of all the lipid-exposed resi-
dues. It also excludes contributions of the polar head-groups region of the membrane. Thus, it
is very unsatisfying. Moreover, even if the presence of the membrane was taken into account
implicitly, using some mean-field approximation, the treatment of the channel pore itself re-
 mains a major challenge. The small size of the pore may impose very specific arrangement of
the water molecules. Thus, it is not clear that a mean-field treatment, based on bulk water,
would adequately-describe de-solvation in this region. To avoid all this, and since we believe
that the motion is dominated by geometrical and packing restraints, we excluded the solvation
term from the outset.

Following the same line of thinking, we also excluded the effects of lipids. Again, an ex-
plicit and accurate description of protein-lipid interaction is not feasible and we suspect that
any mean-field approximation may be erroneous. Thus, we eliminated these contributions al-
together.

The motion between the open and closed conformations of the channel is driven by exter-
nal effects, such as ligand-binding or changes in the membrane potential. Here, such factors
were not taken into account explicitly. Instead, we used two initial conformations and
searched for a root that connects them. Namely, starting from the initial conformation, the ef-
fects of the external perturbations were replaced by conducting a search in conformational
space with a bias towards the goal conformation. This strategy is beneficial in that it by-
passed the need to describe the external effects in molecular detail. However, because of that
we cannot address questions on the effects of the external factors.

In general, the results, and the suggested pathway for channel opening in particular, are
sensitive to the choice of the energy-function.
Incorporation into the Rosetta package
We integrated our software into the Rosetta open-source project for modeling of proteins (42), and examined the pathways using Rosetta scoring as described in METHODS. After a certain period of testing for stability, our tool will be made available to the public.

General framework for generating, aligning and comparing putative pathways
It is important to emphasize that the framework that we presented above is independent of the particular choice of the energy-function. In particular, the approach that was used to compare, align and cluster the generated pathways, and rank them based on various criteria may be useful in a much broader context. The pathways that were generated by the RRT algorithm comprise conformations, in which each two consecutive conformations are close to each other in 3D in both rmsd and internal coordinates values. Thus, the dynamic-programming algorithm for aligning pathways is well-defined in this case. When using the approach within the context of other applications, the algorithm can be adapted by the generation of intermediate conformations that connect dissimilar consecutive conformations. We demonstrated that this approach can merge low-energy motion-pathways into one or more representative pathways of higher-quality, which can then be subjected to further analysis and comparison.

Future research
The aim of this study was to examine in detail the feasibility of our motion-prediction approach, and our studies of KcsA were sufficient for that. A clear advantage of our approach is speed, where a whole gamut of pathways can be generated, aligned and compared in a matter of minutes to hours. We plan to examine more cases and ultimately establish a database of feasible pathways between all proteins with two or more known conformations. In this respect, we will follow the pioneering work of Echols et al. (54). As experimental methodologies for observing proteins in motion are, for the first time, on the verge of becoming a reality (e.g., (20)), computational methods for motion-prediction are now as relevant as ever, and can both benefit from, and contribute to, our understanding of molecular-motion.

ACKNOWLEDGMENTS
This study was supported by Grants 222/04 and 611/07 from the Israel Science Foundation to N. B-T. Dan Halperin's research is supported in part by the IST Program of the EU as Shared-cost RTD (FET Open) Project under Contract No IST-006413 (ACS - Algorithms for Complex Shapes), by the Israel Science Foundation (grant no. 236/06), and by the Hermann Minkowski--Minerva Center for Geometry at Tel Aviv University.
REFERENCES

Figure Legends

Figure 1. The KcsA channel. Two opposing monomers of the tetrameric channel are shown. Each monomer contains two TM helices (TM1 in blue and TM2 in green). The TM helices are connected to each other by the turret loop (gray), the pore (P)-helix (red), and the selectivity filter (yellow). Two $K^+$ ions within the selectivity filter are shown as purple spheres. The locations of the intracellular and extracellular regions are indicated.

Figure 2. A scheme of the overall algorithm. Finding a motion pathway with minimal energy-barrier between the open and closed conformations. (a) Two RRT conformation-trees were grown simultaneously, starting from the initial (closed) and goal (open) conformations, until the trees could be connected to create a full motion pathway. The goal conformation (right) was modeled based on the known X-ray structure of KvAP (10). (b) A general outline of the methodology.

Figure 3. Alignment between two pathways. The conformations along the optimal pathway (i.e., the pathway with the lowest energy barrier) are in cyan and other conformations are in purple. (a) In order to align two pathways, we regard each pathway as a word with the constituting conformations as letters, which makes the problem analogous to that of the alignment of similar protein sequences. Thus, we can use dynamic programming for the aligning of two words, to align two pathways where two conformations from different pathways are matched based on their rmsd. Zero or more gaps between conformations are allowed in order to maximize the global alignment between the two pathways. (b) In this example, the alignment can be used to shorten the optimal pathway (the upper one) by removing the conformation that is aligned to a gap in the lower pathway. (c) The alignment can be used to hybridize pathways in order to obtain an energetically favorable pathway by replacing sub-pathways. The first and last conformations along the two pathways are aligned with each other, and the other conformations are aligned with gaps. That is, the two sub-pathways that are encircled by the ellipsoids cannot be matched. If the energy barrier of the purple sub-pathway is lower than the cyan sub-pathway, we may exchange between the two sub-pathways and obtain a better energy barrier for the optimal pathway.

Figure 4. The alignment graph for two pathways $P=\{p_1,p_2,p_3\}$ and $Q=\{q_1,q_2,q_3,q_4\}$. Two example alignments, described in the main text, are shown by the solid and dashed arrows. Each path in the graph suggests a pairwise alignment between the two pathways.

Figure 5. A putative three-phase motion-pathway between the closed and open conformations of KcsA. (a) An intracellular perspective of the KcsA channel in its closed conformation (grey cartoon), with the location of the Cα atom of residue V155 along the pathway marked in spacefill representation using different colors. The optimal pathway begins with Brownian motion, which is marked in grey. Phase I of the pathway, where the closed conformation is unlocked, is shown in cyan. Phase II, where the gate opens, is shown in purple, and phase III, where the channel is locked in its open conformation, is marked in red. (b) The energy profile of the pathway; the arbitrary units of the Rosetta energy-function were used (see METHODS). The closed conformation (step 1) was assigned the lowest energy and is the most stable. The first 5 steps are around the initial conformation, and not an integral part of
the opening pathway. It is noteworthy that the pathway from the closed conformation to the (model-structure of the) open conformation (step 23) involves crossing an energy barrier (step 12). This is exactly the transition between phases I and II of panel (a). (c, d) HOLE (45) analysis of the changes in the profile of the channel-pore along the pathway. (c) The magnitude of the radius along the pore-axis (Z). The green and red curves were obtained using the open and closed conformations, and the blue curves were obtained for intermediate conformations along the pathway between these two end-conformations. The gate is around Z=20Å and the selectivity filter, which was held fixed, is around Z=50-to-60Å. (d) The green and red curves mark the changes of the pore radius at Z=35Å and Z=23Å along the pathway, congruently. The black curve marks the changes in the pore radius at the minimal constriction. It is noteworthy that the depth of the minimal constriction along the Z-axis changes along the motion pathway.
Figure 1
(a) Plausible conformations

90°

closed channel

Conformational space

Homology model of open conformation

Generation of 100 plausible pathways between the open and closed conformations

Alignment and hybridization of the pathways

Choose pathway cluster with minimal energy barrier

Figure 2
Figure 3
Figure 4
Figure 5
SUPPLEMENTARY MATERIAL

S1. Modeling the open-conformation of KcsA based on the KvAP template

Pairwise sequence alignment between KcsA and KvAP:

<table>
<thead>
<tr>
<th>KvAP</th>
<th>YHLFGAVMLTVLYGAFAIYIVEYPDPNSSIKSVFDALWAVVTTATTVGYGDVVPATPIGKVIGIAVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KcsA</td>
<td>ALHWRAAGAATVLLVTVLGLSYLAVLAERGAPGCQILITYPRALWNSVEATATTGYGDLYPVTLNGRCVAVVVM</td>
</tr>
<tr>
<td>KvAP</td>
<td>LTGISALTLLGTGTVSNMF------*</td>
</tr>
<tr>
<td>KcsA</td>
<td>VAGITSFGLVTAALATWFVGREG*</td>
</tr>
</tbody>
</table>


**PROCHECK** results for the model structure of KcsA

Ramachandran statistics: 94.2% of the torsion angles are located in favorable core regions of the Ramachandran plot, 4.3% are in allowed regions and 1.4% in disallowed regions. The latter correspond to the four Val84 residues (one per monomer) of the loop that connects the selectivity filter to the TM2 helix. Asp80 of the selectivity filter has a distance of 0.25Å from planarity, as in the X-ray structure. This residue is known to form a salt-bridge with Arg89 of a neighboring monomer, which might be the reason for this large deviation from planarity. Two pairs of residues have minor collisions (Tyr62-Pro63 and Pro83-Val84), the distances between the colliding atoms is 2.4Å-2.6Å; in PROCHECK, pairs of atoms are considered to be colliding if their distances are below/equal to 2.6Å. The colliding atoms share H-bonds contacts that may cause these slight deviations. For comparison, in the Ramachandran statistics of the X-ray structure of KcsA, 98.6% of the torsion angles are located in favorable core regions of the Ramachandran plot and 1.4% are in allowed regions. Thus, the stereochemistry of the model-structure of the open-state is slightly less favorable than the known structure of the closed-state.

**PROCHECK** results for the conformations along the suggested pathway

All the conformations along the pathway have over 92% torsion angles located in the core-favored regions of the Ramachandran plot. Bond-lengths and bond-angles are very close to their ideal values. In the X-ray structure conformation (and all the conformations that emerged from its tree by the RRT algorithm), residues Tyr62 and Asp80 of the selectivity-filter have a distance of 0.2Å-0.35Å from planarity. The X-ray structure is free of collisions; however, some conformations along the pathways involve slight backbone-sidechain collisions. The distance between colliding atoms is in the range of 2.2Å-2.6Å, where 2.6Å is defined as collision in PROCHECK. All these collisions correspond to atom-pairs that share H-bond contacts. In the most pathological case, conformations 10 and 11 along the pathways, there are nine colliding residues (one pair is at a distance of 2.2Å and all the others are at distances of 2.5Å or more). Furthermore, the model structure have four colliding residues (Tyr62-Pro63 and Pro83-Val84) placed in the selectivity-filter and thus these collisions are mutual to all the conformations that emerged from the model-structure tree by the RRT algorithm. Overall, the quality of the conformations along the pathway is similar to that of typical model-structures.
S2. Structural alignment between the selectivity filters of the closed- and open- (model) structures of KcsA

Structural-alignment between the selectivity-filters of the X-ray structure of KcsA (pdb 1k4c; red), considered to be the closed conformation, and the model-structure of KcsA (cyan) in an open-conformation. (Left) From a side-view it is evident that the selectivity filter of both structures superimposes very well (rmsd of 0.33Å between their Cα atoms). In contrast, the gate-region undergoes major conformational changes. (Right) An extracellular view of the structural alignment between the selectivity filters, the P-helices and four atoms of the C-terminus regions of TM1 superimposes very well (0.33Å). The molecular surface of KcsA in its closed-conformation is shown in the background.
S3. Algorithm for motion-paths generation

Planning a motion-pathway of a protein can be viewed as a search in conformational space, \( C_{\text{space}} \), where each point \( q \) in \( C_{\text{space}} \) represents a unique 3D conformation of the protein. A KcsA conformation is determined by the torsion-angles along its backbones and the rotameric state of the side-chains. \( C_{\text{space}} \) can be divided into feasible, \( C_{\text{feasible}} \), and forbidden, \( C_{\text{forbid}} \), regions. \( C_{\text{forbid}} \) is the union of all the conformations that involve steric clashes between atoms and conformations that reveal high energy. \( C_{\text{feasible}} \) is the complement region \( C_{\text{space}} \setminus C_{\text{forbid}} \). The definition of \( C_{\text{forbid}} \) implies that \( C_{\text{feasible}} \) is very much constrained, and comprises collision-free, albeit compact, conformations. This suggests applying the RRT algorithm to our problem, as it has been recognized as a very useful tool for designing paths in highly constrained, high-dimensional, spaces (1, 2).

The RRT algorithm attempts to find a collision-free path between initial and goal conformations using a greedy heuristic that biases the conformational exploration from the initial toward the goal conformation and vice versa, though at the same time avoids the pitfalls of local minima. Here the initial conformation corresponds to the native, closed conformation of KcsA, and the goal conformation is the model-structure of the open conformation of KcsA, which was derived by homology-modeling from the known KvAP structure.

In particular, the RRT algorithm grows two trees, \( T_{\text{init}} \) and \( T_{\text{goal}} \), rooted at the initial and goal conformations, congruently. Initially, these trees comprise a single-node related to the initial, \( q_{\text{init}} \), and the goal, \( q_{\text{goal}} \), conformations. The growth of these trees is interleaved, i.e., an expansion procedure is applied to both trees in an alternate manner. In the expansion procedure, a random conformation, \( q_{\text{rand}} \), is generated uniformly over the \( C_{\text{space}} \) dofs and the nearest node, \( q_{\text{near}} \), in the tree is expanded toward the random conformation using a local planner, i.e., moving by fixed incremental distances, resulting in an ordered sequence of close conformations towards \( q_{\text{rand}} \). The extension of \( q_{\text{near}} \) continues until it reaches \( q_{\text{rand}} \) or an unfeasible conformation (e.g., conformation that involves collision between atoms). If the expansion halts at the random-conformation, \( q_{\text{rand}} \) becomes a new vertex in the tree and an edge is added between \( q_{\text{near}} \) and \( q_{\text{rand}} \). Otherwise, the last feasible conformation towards \( q_{\text{rand}} \) is added to the tree (Fig. S3a), unless it is too close to the conformation \( q_{\text{near}} \). We denote the new conformation by \( q_{\text{new}} \), and add an edge between \( q_{\text{near}} \) and \( q_{\text{new}} \).

Both trees are incrementally-extended by the above procedure to efficiently-explore the feasible conformation-space. Following a cutoff number of iterations, an attempt is made to connect these two trees in order to find a feasible path, connecting the initial and goal conformations. Specifically, if \( q_{\text{new}} \) was added to \( T_{\text{init}} \) by the above procedure, then the nearest node in the target tree, \( q_{\text{near}} \), is expanded towards \( q_{\text{new}} \) with the aim to find a feasible local path that connects these two nodes, and merges the two trees (Fig. S3b). This procedure is applied to all pairs of nodes that emerge from different trees (i.e., one stored in \( T_{\text{init}} \) and the other in \( T_{\text{goal}} \)) that the distance between them is below a predefined threshold. Each path from \( q_{\text{init}} \) towards \( q_{\text{goal}} \) and vice versa is compatible with a feasible motion pathway of KcsA.

The conformations along the pathway were selected using the ROSETTA energy-function. Steric clashes are assigned high energy-values of several thousand units, and the low cutoff of zero that was used eliminated them. Formally, one cannot be sure that the pathway between pairs of these low-energy conformations is collision-free, and one should apply the local planner to examine this. In reality, we decided against the use of the local-planner in order to reduce the computational load. Instead, we used a small distance-cutoff of 1.6Å to connect conformations. The results indicated that the pathways are indeed collision-free.
Figure S3. The RRT algorithm expands a conformation tree in the feasible space (white); the forbidden space is marked in yellow. (a) In each iteration, a random conformation, \( q_{\text{rand}} \), is generated and the nearest node, \( q_{\text{near}} \), in \( T \) is expanded as close as possible towards \( q_{\text{rand}} \) up to the boundary of the forbidden space. The last feasible-conformation (\( q_{\text{new}} \)) in the process becomes a new vertex in \( T \). (b) Suppose that \( q_{\text{new}} \) was added in the previous procedure (a) to \( T_{\text{init}} \). An attempt is made to connect the two trees. To this end, the nearest node, \( q_{\text{near}} \), in \( T_{\text{goal}} \) is expanded towards \( q_{\text{new}} \) with the aim to find a feasible local path that connects these two nodes, and merges the two trees. Each path between \( q_{\text{init}} \) and \( q_{\text{goal}} \) defines a feasible motion pathway of KcsA.


The movie demonstrates the curved rotational-motion that leads to the opening of the gate region. The four monomers are shown as cartoons. The spheres mark the location of the alpha carbon of Val115. The suggested three phases, described from the gate perspective, are as follows: Following a Brownian motion near the close-native structure (V115 is in grey).

- A slight clockwise movement unlocks the conformation from its close-state (V115 is in cyan).
- Sliding of the TM2 helices over TM2 helices of neighboring-monomers and opening of the gate (V115 is in purple).
- A counter-clockwise movement of the helix-ends locks the channel in an open-state (V115 are in red).

A static picture of the motion is provided in Fig. 5 of the main text.

REFERENCES

Rapid Sampling of Molecular Motions with Prior Information Constraints

Barak Raveh¹²‡, Angela Enosh²‡, Ora Schueler-Furman ¹* and Dan Halperin ²*

¹Department of Molecular Genetics and Biotechnology, The Hebrew University, Jerusalem, Israel
²School of Computer Science, Tel-Aviv University, Tel Aviv, Israel

Abstract
Proteins are active, flexible machines that perform a range of different functions. Innovative experimental approaches may now provide limited partial information about conformational changes along motion pathways of proteins. There is therefore a need for computational approaches that can efficiently incorporate prior information into motion prediction schemes. In this paper, we present PathRover, a general setup designed for the integration of prior information into the motion planning algorithm of Rapidly-exploring Random Trees (RRT). Each suggested motion pathway comprises a sequence of low-energy clash-free conformations that satisfy an arbitrary number of prior information constraints. These constraints can be derived from experimental data or from expert intuition about the motion. In order to allow the use of state-of-the-art energy functions and conformational modeling, we have integrated this framework into Rosetta, an accurate protocol for diverse types of structural modeling. We have applied the overall PathRover framework to three systems: CesT type III secretion chaperone, Cyanovirin-N anti-viral protein and Ribose Binding Protein (RBP). We then investigated the effect of various types of prior information in these systems. Strikingly, scarce information from distant homologues is sufficient to predict domain-swapping motion at a high level of accuracy. We also show that the predicted energy during domain swapping motion correlated well with existing experimental data, and that intuitive prior assumptions can preserve the integrity of discontinuous domains of RBP during motion. As the incorporation of prior information is very straightforward and since it significantly narrows down the vast search in conformational space and subsequently the running time, the suggested framework can serve as an effective complementary tool for molecular dynamics, normal mode analysis and other common techniques for predicting motion in proteins.

Author Summary
Incorporating external knowledge into computational frameworks is a challenge of particular importance in many fields of biological research. In this study, we show how computational power can be harnessed to make use of limited external information, and to more effectively simulate the molecular motion of proteins. While experimentally solved protein structures restrict our knowledge to static molecular “snapshots”, a vast number of proteins are flexible entities that constantly transform shape. Protein motion is therefore an inherent property of protein function. State-of-the-art experimental approaches are still limited in the information that they provide about protein motion. Therefore, we suggest here a very general framework that can take into account diverse external constraints, and include experimental information or expert intuition. We explore in detail several biological systems of prime interest, including domain swapping and substrate binding, and show how limited partial information enhances the accuracy of predictions. Suggested motion pathways form detailed lab-testable hypotheses and can be of great interest to both experimentalists and theoreticians.

Introduction
Mechanistic understanding of protein motions intrigued structural biologists, bio-informaticians and physicists to explore molecular motions for the last four decades. In two seminal breakthroughs in 1960 [1,2], the structure of Haemoglobin and Myoglobin were solved and consequently, for the first time, mechanistic structural insights into the motion of a protein were deduced from its snap-shot image. This finding paved the way to a-by-now classical model for cooperativity in binding of allosteric proteins [3]. Nowadays, hundreds of proteins with known multiple conformations, together with their suggested molecular motion, are recorded in databases such as MolMovDB [4]. This number increases with the influx of solved structures from the Protein Data Bank [5]. An inherent flexibility
is characteristic of fundamental protein functions such as catalysis, signal transduction and allosteric regulation. Elucidation of motion embedded within protein structures is essential for understanding their function, and in particular, for understanding control mechanisms that prevent or allow protein motions. Understanding the relation between protein sequence and protein motion can allow de-novo design of dynamic proteins, enhance our knowledge about transition states and provide putative conformations for targeting drugs. Accurate prediction of protein motion can also help address other computational challenges. For instance, Normal Mode Analysis (NMA) motion predictions [6] can be used for efficient introduction of localized flexibility into docking procedures [7,8].

**Experimental Limitations**
Experimental knowledge of macro-molecular motions has been discouragingly limited to this day by the fact that high-resolution structures solved by X-ray crystallography are merely the outmost stable conformations of proteins, in a sense a snap shot of a dynamic entity. Nonetheless, innovative breakthroughs in single molecule Förster resonance energy transfer (FRET), small-angle X-ray scattering (SAXS) and Laue method X-ray scattering, as well as advances in NMR spectroscopy such as residual dipolar coupling methods and paramagnetic relaxation enhancements [9,10] provide coarse experimental data on molecular motion, e.g., distance and angle constraints or measurements of rotational motion [11]. However, high resolution experimental data of molecular motion are still beyond reach.

**Computational Simulation of Motion**
In spite, and perhaps due to the limited amount of experimental information, computational techniques have been used extensively for the last three decades to simulate macro-molecular motion. Molecular dynamics (MD) simulations pioneered application of computationally intensive techniques into physics, chemistry and biology [12,13], with the vision of deriving realistic simulations of physical motion *ab initio*. Tremendous efforts of research were dedicated to devise force fields that approximate the quantum mechanic forces underlying atomic motions. Such approximations are based on Newton equations of motions, or even partial quantum mechanics calculations in QM/MM simulations [14]. However, standard MD simulations are computationally intensive and are limited to pico-to-nano second timescales of motion. This is prominently true for large molecules with many atoms, especially in slower motions such as protein folding and recognition, enzyme catalysis, signal transduction and channel gating. In these cases, MD simulations might remain trapped in repetitive cycles of Brownian motion at local minima throughout the simulation, and crossing an energy barrier into a conformation of interest might be a rare event that is not explored by the simulation. As researchers often possess intuition and explicit partial knowledge about the nature of a motion or target conformations, biasing techniques were devised in steered MD simulations [15]. Such methods incorporate prior knowledge or expert intuition about the system and compromise the intended purity of MD simulations as a physical simulation. Nonetheless, they still rely, to the most part, on an approximation of physical forces, and guarantee that some plausible assumptions are satisfied. Subsequent motion trajectories were shown useful for designing experiments and deriving mechanistic insights into protein motion. But since an immense number of step by step calculations of the physical forces are still explicitly required, complementary coarse-grained methods such as Normal-Mode Analysis and Gō models [6,16,17] (reviewed [9]) pose a faster alternative. They can provide a quick impression about protein conformational changes when given a native conformation, but do not aim at the very fine details of the motion.

**Sampling Based Approaches**
In recent years, a novel approach for sampling motion pathways, rooted in algorithmic robotics motion planning, has been applied to large-scale molecular motion prediction. This approach suggests an efficient alternative to slow step-by-step simulations of Newton equations. A sequence of clash-free conformations is generated by sampling the topology of the conformational space. This sequence is a fine discretization of continuous motion. In their original context, motion planning techniques like probabilistic road-maps (PRM) [18,19], Rapidly-exploring Random Trees (RRT) [20,21] and similar methods [22-24], all reviewed in [21,25], have been used to plan the
motion of objects\(^1\) with many degrees of freedom (dofs) among obstacles in a constrained environment [26]. Recently, motion planning techniques were successfully applied to predict motion pathways for molecules while considering a large numbers of dofs [27-35] For simplicity, we here collectively refer to this family of techniques as “motion planning” sampling techniques, as opposed to other sampling based methods used in molecular biology such as Monte-Carlo (MC) and Monte-Carlo with Minimization (MCM) [36]. The latter are popular techniques for sampling a single final low energy conformation, like in \textit{ab-initio} fold prediction, docking and protein design. In contrast to Monte-Carlo, motion planning techniques are optimized for finding complete motion pathways. They record the history and topology of the sampled search space in a tree or a graph data structure, the “road-map”. Molecular motions are extracted from paths or "roads" in the graph, where nodes stand for feasible (low-energy) conformations and edges connect close-by conformations. Therefore, paths in this graph are sequences of clash-free conformations. This also adds a whole new dimension of memory to the sampling process and the resulting search in conformation space is shown to be less prone to futile repetitive sampling\(^2\) [20].

But as the application of motion planning techniques to molecular motion is very new, in particular compared to established methods such as MD, further research is required in order to validate and calibrate its use. Motion planning techniques are very fast – they take between minutes and hours even for generating a full motion pathway of slow time-scales with dozens of dofs and hundreds of amino-acids [34,35] compared to weeks to months in MD simulations of motions with shorter time-scales. Nevertheless, as the dimensionality of the search space increases, it is advantageous to exploit prior information about the nature of the motion, in order to direct the search. A common practice in sampling methods of single conformations like MC is to bias the energy function itself towards known constraints [37]. Importance sampling and pseudo-measures were used to bias motion sampling in the context of algorithmic robotics motion planning, for instance for addressing narrow passages [21,38], but to the best of our knowledge, such ideas have not been applied in molecular motion planning hitherto.

\textbf{Our contribution}

Here, we present PathRover, a comprehensive and generalized framework for efficiently sampling and generating motion pathways that satisfy constraints of prior information with the RRT algorithm [20]. Our approach follows the notion that the combination of a number of partial constraints can significantly limit the number of feasible solutions. We rely on a generalized RRT formalism that allows efficient, flexible and straightforward integration of prior information into the basic RRT algorithm by truncating undesired pathways from the RRT tree (see Figure 1). To our knowledge, this is the first thorough generalized attempt to incorporate diverse prior biological information into the RRT algorithm. Such knowledge is often available in real biological contexts, e.g., when it is known or hypothesized that specific contacts between atoms are likely (see Table 1 for additional examples). For instance, progress in experimental techniques can yield samples of pair-distance constraints from “experimental rulers” such as single-molecule FRET and site-directed spin labeling experiments, or orientation constraints on secondary structures [11]. PathRover generates low-energy, clash-free motion pathways that are guaranteed to satisfy these constraints. This is in analogy to similar approaches for finding a single optimal structure (but not a motion pathway) under a set of experimental constraints [39,40].

PathRover is integrated into the Rosetta molecular modeling framework [41], a leading accurate protocol for a range of different structural modeling tasks (e.g. [42-45]). Thus, PathRover is equipped with state-of-the-art energy functions, sampling and optimization protocols.

\textbf{Studied Examples}

This study explores the effect of partial information constraints on generated pathways. We start with the effect of partial information constraints on RRT sampling and pathways on a simple toy example (see Figure 1). We then test various sources of partial information in detail on three biological examples.

\footnote{Referred to as “robots”, but can be any moving object such as digital avatars, manufactured parts, or molecules in the context of this study. (see Latombe. 1999).}

\footnote{See Discussion for comparison to constant memory extensions of Monte-Carlo like Tabu-Search (Chelouah \textit{et al.}, 2000).}
CesT and Cyanovirin-N – simulation of domain swapping using partial information

Many protein chains consist of several domains that are packed against each other, suggesting that these domains have an intrinsic property of complementing each other [46]. Domain swapping occurs when part of a domain from one chain packs against the complementary part of this domain in an identical chain. The subsequent pair is referred to as a pseudo-monomer, as the packing forms a globular unit that resembles the one in the non-swapped monomers (See Figure 2a, b). Domain swapping requires the unpacking of domains in the original chain, and the subsequent repacking to another chain. Investigation of domain unpacking and repacking may improve our understanding of the general mechanism of oligomerization [47]. Domain swapping can also lead to undesired effects of aggregation, such as the formation of amyloid fibrils [48]. We show that we can use very coarse information from a very distant homologue of the CesT protein to predict a low-energy pathway for a domain swapping motion, and also examine domain swapping motion in the Cyanovirin-N anti-viral protein.

Ribose Binding Protein – forcing the tertiary integrity of a domain throughout the motion pathway

Here we examine the conformational changes in an enzyme that consists of two domains. Each domain is composed of several discontinuous segments of the protein sequence (see Figure 4b), and standard single-tree RRT leads to partial disintegration of the domains throughout the motion. Domain disintegration during catalysis seems unlikely based on basic biological intuition, and we use this example to show how such sensible intuitions about the course of the motion can be easily translated and incorporated into the prediction of a motion pathway.

Since the use of RRT for predicting pathways of molecular motion is relatively recent, and as the possibility of alternative pathways cannot be completely excluded, suggested pathways should be regarded to an extent as a proof of concept for the feasibility of the method to incorporate constraints into molecular motion prediction. However, all generated motion pathways are guaranteed to form a sequence of clash-free low-energy conformations, and to satisfy the input constraints. Unlike coarse grained methods like ENM, Go models and others [6, 9, 17], we support a full-atom mode, in which the output conformations contain the coordinates of all side-chains and hydrogen atoms. As such, PathRover provides detailed lab-testable hypotheses that can be of great interest to both experimentalists and theoreticians.

Methods

Forbidden Space, Feasible Space and Energy Function

Rapidly-exploring Random Tree (RRT) is a general framework for rapid exploration of a conformation space in a highly constrained environment. It was first presented in algorithmic robotics, where it was used to plan the motion of moving objects among obstacles [20]. Here, we define the conformation space by the internal dofs of the protein, namely the torsion angles that are allowed to change throughout the motion pathway (see below). The conformation space is divided into forbidden and feasible regions (referred here as C-forbid and C-feasible, respectively). The forbidden regions correspond to all the conformations that involve high energy values, namely energy score above a threshold parameter, whereas the feasible regions comprise the low-energy conformations.

Rosetta Infrastructure

Energy Function: Full-Atom and Centroid Mode

We experiment with both the Rosetta full-atom energy function (Rosetta score12 [42]), which was shown useful for discerning native structures at atomic detail, and the coarser Rosetta centroid mode energy function, where side-chains are represented as centroid spheres (Rosetta score4 [41]). The latter allows rapid calculations at the expense of atomic detail, and has been used in a wide range of applications in Rosetta to speed up the conformational search by optimizing coarse features prior to atomic level optimization.

Rosetta Optimizations

In addition to the full-atom and centroid mode energy functions, the presented framework takes advantage of the elaborate infrastructure of Rosetta, including manipulation of molecular dofs, rapid side-chain optimization for fixed backbones (using the "rotamer trial" procedure described in [49]), and caching of energy calculations. The framework can also use the gamut of other Rosetta features, such as energy minimization protocols, closed loop sampling and sophisticated manipulation of backbone torsions, e.g. backbone fragment

---

1 In robotics, conformation space is usually referred to as the "configuration space"
libraries or backrub motions [50]. These features are out of the scope of the current study, and will be explored in future work.

**Algorithm**

In RRT, a tree of conformations is used to record the topology of the search space. Nodes stand for feasible (low-energy) conformations, edges connect close-by conformations, and paths are sequences of feasible conformations. Ideas for biasing the RRT motion prediction algorithms have been proposed in non-biological context, mainly for problems of narrow passages [21]. We suggest a different variant of the RRT algorithm that relies on truncation of non-favorable branches of exploration. We formulate diverse biological partial information to constrain the sampling of energetically favorable motion pathways of molecules (see Table 1 for examples). In Figure 1 we illustrate how this variant works compared to the basic algorithm on a simple toy example. We shall first review the basic RRT algorithm, and then present the branch truncating variant.

**Basic Single-Tree RRT Algorithm – only source conformation is known**

In the basic variant of the RRT algorithm, the tree is biased to grow towards unexplored regions of $C$-feasible in progressively increasing resolution [20] (see simulation of tree growth in Figure 1a). A single source conformation is the only input, and the tree is grown from this root conformation, where each node in the expanding tree corresponds to a valid conformation in feasible space $C$-feasible. The tree is grown towards random directions in the conformation space by very small incremental movements and hence each path in the RRT tree can be considered a fine discretization of a continuous motion pathway in the feasible conformation space. During the expansion, side-chains of generated conformations along the pathways are locally refined by the Rosetta rotamer-trial procedure to alleviate local steric clashes and optimize the bonding of side-chains with neighboring residues [49]. Upon satisfaction of a certain stop criterion, low-energy motion pathways can be extracted from paths in the tree.

**Single-Tree RRT with partial information about the target conformation**

Here we introduce diverse types of partial information in a generalized manner, which does not necessarily assume a target conformation even exists. Following a common practice, any type of partial information is formulated by a set of predicates that directs the search (see Table 1). Similar to the basic RRT algorithm, a single tree is grown by random sampling of new conformations. However, in a filtering step, branches that do not improve the given predicates, are truncated after a certain number of steps, even when they lead to energetically feasible conformations (Figure 1b). The aim of this filtering step is to avoid expensive energy calculations for undesired directions. Truncating branches that do not improve the predicates narrows down the search to relevant regions (Figure 1b, grey circle), but may impair the overall coverage of unexplored regions when using the basic variant (Figure 1a), and hence there is a tradeoff between the two variants. While in the toy example the partial information variant shows better performance (Figures 1c, 1d), we found that in some cases it is beneficial to disable the filtering step once in every $k$ iterations and allow the tree to grow in any direction. This helps to avoid local minima with respect to the biasing predicates and still benefit from the rapid sampling of unexplored regions in the basic algorithm.

**Types of partial information** The main aim of this study is to show the benefit in formulating and incorporating biologically motivated partial information. In Table 1 we outline several specific types of predicates that we implemented in this work. We chose to focus on partial information that can be motivated by experiments, comparative analysis and expert intuition. For example, comparative analysis over biological databases can provide partial information from homologue structures, or numerous alternative conformations of the native protein. Different types of partial information can be combined into a joint predicate as a weighted sum. Distance constraints and additional constraints have been previously used in Rosetta to direct Monte-Carlo with Minimization sampling, in quite a different algorithmic context (see Discussion).

**Choice of Dofs as a Form of Partial Information**

We define the conformation space in terms of internal coordinates, namely the backbone torsion angles uniquely define a conformation (but for each fixed backbone conformation, the side-chain torsion angles are optimized). However, as the combinatorial search space grows exponentially with the number of dofs, it is beneficial to restrict the choice of flexible torsion angles, again with the aid of expert intuition or alternative measures. An
automated accurate choice of mobile \textit{dofs} is a challenging aspect of motion prediction, and in this step, prior information can be most useful (see [34] for an attempt in this direction). A common practice is to assume that certain regions, such as secondary structures or whole domains, remain fixed throughout the motion. When a source conformation and a target conformation are available, we used the FlexProt flexible alignment tool [51] to extract fixed regions of the protein, and defined the \textit{dofs} by the regions in-between. However, one must consider the possibility that the above assumption does not hold, such that certain flexible regions are still identical among solved conformations. We combined information from homologue proteins, Normal Mode Analysis, comparison of torsion angles in alternative conformations and information from the literature to manually restrict the allowed \textit{dofs}. In addition, we fixed the bond angles and lengths, under the common assumption that changes in torsion angles can make up for the milder changes observed in bond angles and length. The impact of some of these choices is further elaborated in the Results section.

Results

Here we focus on three biological examples, two examples of domain swapping motion (\textit{CesT} and \textit{Cyanovirin-N} proteins), and one example of substrate binding-induced motion (\textit{Ribose Binding Protein}). In the first example we demonstrate the contribution of partial information to the final conformation of the motion. In the second we show how the motion pathway of domain swapping is altered in full-atom mode, where all side-chain atoms are considered. In the last example we show how partial information can enforce “quasi” loop-closure constraints during substrate binding. All runs were conducted on an AMD Opteron 275 2.2Ghz/1MB processor. The types of partial information predicates used throughout this study are listed in Table 1.

\textbf{Partial information on domain swapped target in the CesT protein}

The \textit{CesT} protein is a type III secretion chaperone in \textit{Enteropathogenic E. coli} that binds numerous effector proteins. Its structure and that of several of its distant homologues are known [52]. However, in contrast to its homologues, in \textit{CesT} the neighboring chains within the crystal lattice are domain swapped (Figure 2b) and domains from neighboring chains pack to form a monomer-like globular unit, the “\textit{pseudo-monomer}” [53] (Figure 2a). The \textit{pseudo-monomer} can be well aligned to monomers in the other homologues, suggesting that there is a monomeric form of \textit{CesT} that resembles the \textit{pseudo-monomer}. It is however not known whether the swapped conformation of \textit{CesT} is a crystallographic artifact or whether it is the physiologically active peptide-binding form [52]. Hence, we examine here whether we can model a hinge motion in which the domains of \textit{CesT} move from a swapped conformation to a monomeric conformation that resembles the distant homologues. We use this example to demonstrate different combinations of partial information predicates. Reassuringly, the partial information has led to accurate native-like conformations as described below, even though the energy calculations in this case have all been performed using the simplified centroid mode representation of the protein, where side-chain orientations are averaged by a single sphere (see Methods). Each reported run was repeated 50 times, and the best run with respect to the partial information predicate was used to analyze the motion pathway. We note however, that very similar results were obtained with a much smaller number of a few runs.

\textbf{Using the pseudo-monomer to model CesT domain swapping motion}

Packing of non-swapped monomers against each other is mostly identical to their packing in the \textit{pseudo-monomers}, as the sequence of domains is identical [47]. We therefore formulated a predicate to bias the two separate domains (see Figure 2a, 2b) to move towards their orientation in the \textit{pseudo-monomer}. We used the \textit{RMSD-Match} predicate (defined in Table 1), and biased the algorithm to reduce the RMSD between the current conformation and the target \textit{pseudo-monomer} conformation. Residues 34-37 separate the two domains and we allowed the torsion angles of these residues to change during the motion (Figure 2b, black). This is also in agreement with normal-mode analysis (NMA) that we conducted with HingeProt [51], which suggests a hinge motion around residue TYR37. In the subsequent runs (Movie S0), each motion pathway was generated within minutes on a single processor, and in the final conformation along the pathway the domains were packed exactly like in the \textit{pseudo-monomer} (0.8Å RMSD). This provides a proof of concept for the ability to generate a low-energy motion pathway between the source and the target-like conformation of the \textit{pseudo-monomer}.  
**Partial Information derived from SigE – a Homologue of CesT**

*SigE* (*Salmonella enterica* sigD specific chaperone, pdb-id 1K3S) is one of a few distant homologues of *CesT*, with 18% sequence identity [52]. The N-terminal and C-terminal domains of *SigE* are very similar to those of *CesT* (deviation of 1.7Å and 2.5Å respectively), and the pseudo-monomer of *CesT* can be aligned to the *SigE* monomer (Figure S1). However, due to very low sequence similarity, the *SigE* monomer deviates by 3.96Å from the pseudo-monomer (using the sequence alignment from [52]). We have used *SigE* as a platform for comparing the effect of different types of predicates derived from *SigE* on the motion pathways of *CesT*. We formulated different partial indicators on the relative orientation between the two domains as they appear in *SigE* and used them to guide the RRT algorithm. The results are summarized in Figure 2c and Table 2. Each run took only a few minutes on a single processor to produce a motion pathway from the domain-swapped source to the non-swapped monomer conformation in centroid mode representation. Below we discuss the different predicates that we extracted from *SigE* and show that partial information on a distant homologue is effective in guiding *CesT* to an accurate unswapped conformation.

**Atom Pair Distance** We first examined a simple constraint, requiring that the distance between atoms in two selected residues (in our example, PHE12 and LEU95) in the different domains of *CesT* will be approximately equal to the distance between atoms of the two corresponding residues in the homologue (PHE12 and GLU110 in *SigE*). This type of constraint can be motivated for instance from labeled spectroscopic experiments. The atoms were selected from helices of different domains (Figure 2c - i), hence this predicate is a rough indicator of the packing between the two domains. The final conformation of the generated motion pathway indeed satisfied the predicate (Table 2a), and during the motion, the pair-atom distance declined from 39.13Å to 6.06Å, which is indeed the distance between the atoms in the target homologue (Table 2a and Movie S1). However, the structural alignment between the final conformation of *CesT* and the pseudo-monomer is rather poor: helix H1 is aligned to the opposite side of helix H3 (see Fig 2c – i) compared to the pseudo-monomer. This demonstrates that the distance between a single pair of atoms should be combined with other partial information or atomic detail constraints, in order to derive a more reliable target conformation and motion pathway.

**Relative Orientation of Secondary Structures** We now formulate a predicate on the relative orientation of various secondary structures in the two domains of *CesT*, as partial indicators about the packing of these domains. The α-helices H1 and H3 each lie in a different domain of the protein (Figure 2a). We used least mean square line fitting (LMSLs, Table 1) to approximate the main axes of these α-helices. We then combined axis angle and axis distance with centers of mass distance (Table 2 - ii and Figure 2c - ii). The combined measure defines the coarse orientation between these helices in *SigE* and in *CesT*, but it does not depend on the residue by residue alignment. Hence, it is less sensitive to alignment shifts that are characteristic for α-helices. During the run, the line angle and line distance predicates were perfectly matched (Table 2a – ii and Movie S2). Intriguingly however, the center of mass distance between the helices remained far from that in *SigE*. In fact, the center of mass distance in the final *CesT* conformation (11.1Å) is much closer to that in the pseudo-monomer (12.46Å) than that in *SigE* (15Å). Moreover, the RMSD of the final structure is only 1.87Å from the pseudo-monomer residues, compared to 5.34Å in *SigE* (Table 2b – ii), although the only predicate that guided the motion is the orientation of helices in *SigE*, and even though the run is in centroid-mode. This shows that the native packing of the domains could be recovered with partial information predicate on a distant homologue.

Three additional predicate formulations over secondary structures orientations were evaluated. For each, the relative orientation was measured by an explicit residue-by-residue matching - **Aligned RMSD-Match** (Table 1). In case (iii), the RMSD was measured over helices H1 and H3, as before; in case (iv), a similar measure was used to align β-strand B0 next to β-strand B1 (Figure 2a), using residues 32-34 from the B0 strand; In case (v), we tried to match both the β-strands and the α-helices using the same criterion. In all cases, the final conformation was consistently closer to the pseudo-monomer conformation (see Table 2b – iii / iv / v, Figure 2c – iii / iv / v, and movies S3, S4 and S5).

**β-sheet formation** It is interesting to see how β-sheet formation affects α-helix orientation and vice versa. In cases (ii) and (iii) in Figure 2c, the predicates focus on orienting helices H1 and H3, whereas the predicate in case (v) combines helix and sheet formation. The final conformation is consistently similar to the pseudo-monomer form, even though the information is from the homologue (Table 2). It is interesting to note that when the motion was guided by partial information on the orientation of the α-helices alone (helices H1 and H3 in Figure 2a), the
β-strands B0 and B1 still came close together and the final conformation exhibited similar structure to the *pseudo-monomer*. In contrast, the helices did not move to the correct orientation when the only partial information provided was β-sheet formation. This may imply that the motion is, in principle, guided by the motion of H1 towards H3, and only when the helices contact each other, the β-sheet is created. However, further analysis - e.g. that investigates additional *dofs* - might shed more light on the order of secondary structure formation.

**Cyanovirin-N domain swapping**

*Cyanovirin-N* is an anti-viral fusion inhibitor protein that binds to viral sugars, and is trialed for preventing sexual transmission of HIV. The domain swapped dimer (Figure 3a) has higher anti-viral affinity than the monomer [54], and it was shown that the two forms can exist in solution, with a high energy barrier for transition between them. In addition, it has been reported that certain mutations can affect the energy barrier and stabilization of the alternative conformations [55]. We examined here how two repeat domains of a single chain can unpack from the tightly-intertwined monomeric conformation (Figure 3b, *pdb-id 2EZM*) to an extended domain-swapped conformation (Figure 3a, *pdb-id 1L5E*), first in the simplified centroid mode, then in detailed full atom representation.

**Allowed Torsion Degrees of Freedom**

We relied on structural conservation patterns and allowed torsion flexibility in the stretch of residues 48-55 (Figure 3a - loop L) that connects the two repeat domains of 30% sequence identity at residues 4-48 and 55-100 (Figure 3a, D1 and D2). Each domain consists of two β-sheets that are connected by a small helical loop. Torsion angles in the L region differ significantly in the two structures, whereas the two repeat domains deviate by 0.5Å RMSD only, and are conserved in alternative structures. The conformational transition during swapping is substantial, as the swapped conformations deviate by 14Å RMSD.

**The Cyanovirin-N Monomer is sterically locked by the side-chains**

In a first step, we checked the feasibility of generating a motion pathway in the simplified Rosetta centroid mode representation (see Methods). We were able to generate low-energy motion pathways from the packed monomer to the swapped dimer conformation within minutes by a Single-Tree RRT with an *RMSD-Match Align* (Table 1) predicate on residues 10-90 for biasing the growth towards the domain-swapped conformation. However, considering the experimentally determined high energy barrier for this motion, it is rather surprising that such a pathway could indeed be easily created, and we thought this might be an artifact of the simplified centroid mode representation of the structure: the barrier might be apparent only at a higher resolution level. We therefore proceeded to a full-atom representation, where all side-chains atoms and hydrogen atoms are modeled explicitly.

It was now impossible to unlock the intertwined monomer using the same degrees of freedoms (*dofs*), unless the energy threshold was substantially raised to 10^7 Rosetta score12 units, allowing for extreme steric clashes. In fact, the monomer remained packed even when a very large RRT was generated (consisting of 300,000 conformations), and moved by no more than 1.5Å from the initial conformation, thus remaining 13Å away from the swapped target conformation. Movies S6 and S7 demonstrate how the side-chains of one domain are tightly locked within the other domain, and restrict the motion to Brownian perturbations within a “cage”.

**Adding Secondary Hinges**

We postulated that introduction of more *dofs* might allow the monomer to unwind. We therefore inspected the structure and located additional potential flexible regions, namely two symmetry-related loops at residues 36-40 and at residues 87-91 that connect the two β-sheets in each domain (Figure 3a and 3b, cyan and yellow loops), which look like “weak links” in the protein chain. Unlike the prime hinge, these regions are structurally conserved between structures, but might move during the motion and allow unlocking of the monomer. Coarse-grained normal mode analysis of monomer 2EZM (performed with HingeProt [51]) suggested hinges at near-by residues 34 and 85. As the variation in phi / psi torsion angles between conformations is larger in residues 36-40, we allowed flexibility in these residues and in the symmetric region at residues 87-91. We note that the NMA did not recover a domain-swapping motion trajectory and did not detect the primary hinge at residues 48-55, which clearly moves during the motion to accommodate the big difference in phi / psi torsion angles between conformations.
Additional Hinges Allow Cyanovirin-N Domains to Unlock

The addition of flexibility in the two loop regions into the algorithm results in a motion pathway of the conformational change in full atomic representation that includes the side-chains (Movie S8). First, the two domains slightly disintegrate in a coordinated motion of the three flexible loops. The central flexible loop (residues 48-55), which was previously wrapped around one domain, unwinds in a circular motion. The domains are then brought to the vicinity of the target swapped conformation. We suggest that the small dual-stranded hairpins adjacent to the loops (residues 40-49 and residues 91-100) might trap the side-chains of the monomers and form an atomic “gate”. Flexibility in these regions may serve to open the “gate” and allows the domains to unpack.

Cyanovirin-N is Free to Move after Unlocking

Analysis of the energy plot throughout the motion demonstrates the sensitivity of the Van der-Waals repulsive term of Rosetta score to atom clashes in cluttered environments: high-energy peaks are observed during domain dissociation and unwinding (steps 1-35 in Figure 3d). This stands in contrast to the smoother landscape (Figure 3d, steps 36-68 on the x-axis) that is seen once the two domains have completed to separate. Now that the two domains are unlocked, they are free to sample many conformations without significant clashes. This finding correlates well with the fact that various additional domain swapped conformation have been solved, and all are related by movement in the primary hinge at L’ (e.g. pdb-id 1L5B, see [55]).

Ribose Binding Protein

In the last inspected system, we focus on how intuitive constraints about the nature of the motion can be incorporated within the motion prediction framework. The Ribose Binding Protein (RBP) belongs to a family of ligand-binding proteins that consist of two domains connected by a hinge: upon binding of the ligand in a cleft between the two domains, the domains approach each other to close the cleft (Figure 4a). However, unlike Cyanovirin-N for example, in RBP each domain is discontinuous with respect to the sequence. One domain consists of residues 1-101 and residues 236-261, and the other consists of residues 104-234 and residues 262-271 (Figures 4b, 4c). The hinge that connects the two domains is made of three separate stretches of sequence (Figure 4b). Consequently, the hinge torsion angles must change in a coordinated fashion, as otherwise the two domains would disintegrate. Indeed, when we simulated the motion of RBP without any partial information constraints, the domains disintegrated during the motion (results not shown).

The problem of fixing remote structural segments that are connected by a flexible loop is known in the literature as the protein loop closure problem [56]. Previous attempts have been made for ad-hoc solutions to this problem during RRT simulations [57,58], as well as in the broader context of structural modeling [56,59,60]. Here, instead of explicitly forcing loop closure constraints, we added a simple partial information predicate that punishes disintegration of the domains in terms of RMSD to the intact domain. In the resulting pathway (Movie S9), the two domains are kept in one piece throughout the motion, and only a small β-strand at the C-terminal of the protein (residues 266-269) deforms from its β-sheet during the motion (bottom domain in the supplementary movie). The bulk of the two domains abide loop closure constraints, and the simulation of the motion in centroid mode is performed within a few minutes time. This demonstrates the flexibility of the partial information framework to efficiently address diverse settings, without the need of explicit ad-hoc calculations.

Discussion

In this work we showed how different types of partial information can be incorporated into the Rapidly-exploring Random Tree (RRT) algorithm. We present PathRover as a comprehensive framework, implemented within the Rosetta modeling infrastructure. In non-biological contexts, the idea of biasing the growth of the tree towards predicates has been mainly restricted to specific problems such as narrow passages [21]. In structural biology, partial information constraints are widely used in single conformation prediction [39,40] and in MD simulations. The novelty in this work is the systematic introduction of the integration of partial information to sampling-based motion prediction. In this sense, sampling based methods like RRT pose a natural framework for integrating prior biological information. We incorporated partial information into simulations of three different systems: CesT type III secretion chaperone, Cyanovirin-N anti-viral protein and Ribose Binding Protein (RBP). Partial information
limits the search in the vast space of possible motion pathways using sensible expert assumptions or partial experimental results. In turn, it allows for further subsequent validation.

**Sampling Based Methods vs. Molecular Dynamics and NMA**

The focus of this work is the integration of partial data into the RRT algorithm within a biological context. While their biological application is quite recent, sampling based algorithms such as RRT have contributed to our understanding of molecular kinetics in applications such as energy landscape exploration, protein and nucleic acids folding pathways and ligand binding [27,29-31]. They were shown to provide clash-free motion pathways of slow-scale, and their performance has been compared to molecular dynamics [31] and integrated with Normal Mode Analysis [34]. In several cases, they capture known conformational intermediates and other experimental indicators [28,33-35]. An elaborate comparison of our method to MD and NMA is out of the scope of this work, as a serious comparison to MD would require generations of many MD pathways for different systems. We do however emphasize that RRT is capable to capture motion pathways of much slower time scales than MD, using far less computational effort. Hence, in contrast to MD simulations, sampling based methods are fast enough to generate a very large number of alternative pathways, whereas in an MD simulation it is often hard to decide if the pathway is representative or just the outcome of specific random start conditions. In addition, MD simulations simulate molecular motion *ab-initio*, whereas sampling based methods aim to give a global view of the energetic search space. Experience in sampling based methods has not yet accumulated to the many years of research and calibration of molecular dynamics simulations, and perhaps an integrative view can result in more effective motion prediction algorithms.

Normal mode analysis (NMA), in full atom or in coarse representation, was often shown to provide a fair approximation of atomic motions around local minima, and to provide prediction of hinge regions [6]. As such, it is complementary to RRT, where the full scope of a motion pathway is examined at once. Indeed elastic network models have been recently integrated into RRT and examined on a database of several proteins [34]. Recently, deformable elastic networks potential have been used to incorporate distance constraints from low-resolution density maps and restrain sampling around approximate structural models, though the objective in this case is the prediction of a single equilibrium structure rather than the simulation of molecular motion [61]. We note that compared to NMA, the incorporation of any prior information on the motion in the framework presented herein is particularly straightforward, since it is simple to formulate predicates that bias the motion according to general experimental or intuitive constraints.

**CesT – recovery of native conformation from a distant homologue**

In CesT, backbone atoms play a major role in the motion, as β-strands switch partners. Strikingly, diverse types of partial information about the conformational changes were shown to lead to the same final conformation. Structural hints from a very distant homologue (SigE) resulted in a native-like conformation (Table 2), even though the simulation was made in simplified centroid mode (defined in Methods). At such level of predictive precision, many side-chains can be already modeled quite accurately. We note that the original swapped CesT chain is very far from either conformation and the motion is large in scale. The fact that one term in the predicate reached a value similar to that in the *pseudo-monomer* (Figure 2c) is particularly reassuring, as it demonstrates how energetic considerations might prevent over-biasing when using partial-information predicates. Although our aim is motion prediction and not homology modeling, it is promising that the near-native conformation is recovered using very simple predicates.

**Cyanovirin-N – unlocking intertwined sidechains**

In the case of Cyanovirin-N, we analyzed the domain swapping motion itself and not only the final conformation. We suggested a possible motion pathway of its two domains out of a tightly-interlocked position, where only three small flexible regions allow this motion. These flexible regions can be good candidates for mutational analysis experiments based on our predictions.

We note that in this case, the centroid-mode representation, where side-chains positions are averaged, did not capture how constrained this movement is. This raises a more general question of how well a simplified representation of a molecule can capture the details of molecular motion. While in CesT, centroid mode recovered
accurate conformations, in the case of Cyanovirin-N, the side-chains seem to be responsible for securing the tight packing of the two domains. This might relate to the fact that CesT packing is mediated by secondary structure formation at the backbone level. The fact that the energy landscape is smoother towards the end of the motion (Figure 3d) suggests that the two domains reach a free unlocked form at this stage where they can sample many conformations without significant clashes or energy loss. An interesting support comes from the fact that additional solved structures of the domain swapped conformation (e.g. pdb-id 1L5B) are all related to 1L5B by modest hinge rotations, yet their overall conformation is substantially different [55]. Thus, once the domains are unpacked from each other, the chain is free to sample diverse conformations. The difference between Cyanovirin-N and CesT shows how domain swapping can take different forms and how examination of a larger set of domain swapping motions via computational tools can allow us to shed light on different classes of domain swapping mechanisms.

Ribose Binding Protein – Incorporating Expert Intuition about Motion

The last inspected system mainly serves to show how intuitive constraints effectively dictate the nature of the motion pathway itself and not just the end conformation. The domains of the Ribose Binding Protein are not continuous with respect to the sequence (Figures 4b, 4c). Maintaining the domains integrity during hinge motion is therefore not a trivial task, since three remote stretches of torsion angles are manipulated. This might require complex loop closure calculations or interpolating changes in torsion angles from Normal-Mode Analysis [34,58]. We were able to maintain the integrity of domains using simple and general predicates of partial information that do not require ad-hoc calculations or external methods.

Generally speaking, quite often an expert might have a rough intuition about certain aspects of a protein motion. Such intuition should, of course, be incorporated with great care, in order to avoid erroneous over-constraining of the motion, but can still result in invaluable working hypotheses. In this respect, the weight of the constraints can be softened, so that domain integrity can be somewhat compromised, and energetic considerations would still dominate the motion.

Incorporation into Rosetta Modeling Framework

An important aspect of PathRover is its full embedding into the Rosetta modeling framework. Rosetta has repeatedly demonstrated an exceptional ability to produce high-quality results for a variety of different modeling tasks in the field of protein modeling, docking, protein design and other modeling challenges at atomic level detail (e.g. [42-45]). This setup allows us to address questions of algorithmic efficiency within an established frame that includes (1) well-calibrated energy functions (both at the centroid and full-atom level) (2) efficient energy calculations, and finally (3) a battery of established conformational sampling protocols. In particular, we use the Rosetta centroid and full-atom energy functions to direct the search and maintain energetic plausibility of the pathways. Rosetta side-chain optimization schemes and options for caching energy calculations are used to improve the efficiency and accuracy of PathRover, and in principle, sophisticated local moves from other Rosetta protocols can be integrated into the RRT sampling process. The presented framework is designed for easy incorporation of a wide range of different predicates (see Table 1). The next step will involve the straightforward incorporation of specified Rosetta conformational sampling protocols, as well as the extension to additional predicates used previously, such as NMR coupling measurements and docking interface constraints. These have been used to guide and filter Rosetta Monte-Carlo searches, and will here be incorporated into RRT-based motion prediction.

A Full-atom Energy Function and a Simplified Centroid Mode Representation

Previous applications of the RRT algorithm have mainly been based on geometric considerations of clash avoidance or Van-der-Waals terms of established force fields. In some cases, more sophisticated terms were employed [27,29]. Here we introduce, for the first time, the established Rosetta full-atom energy function into sampling based methods. Hence, we are able to rapidly generate motion pathways for complex movements that are at the same time energetically favorable and that abide by possibly known constraints about the motion. The full-atom energy function of Rosetta (we used here score12 [42]) includes physical terms such as hydrogen bonding and solvation terms, as well as statistical knowledge based terms like the Ramachandran score, rotamer
likelihood, and a simplified electrostatic score [41]. In some cases we observed that the repulsive energy term dominates the motion pathway: in a cluttered environment, clash avoidance is indeed probably the main contribution. Naturally, however, additional energy terms will affect the details of the motion pathways, such as solvation effects and electrostatic interactions [62]. We experimented with the coarser but rapid centroid mode of Rosetta, in which side-chains are represented as an average spheric entity and energy calculations are directed by the backbone with a coarse approximation of the contacts between the side-chains. As shown previously for other applications, the coarse-grained energy function can rapidly provide insights into the nature of the motion, and some rough estimates over its feasibility. It can thus complement full-atom predictions, where the energy function is more susceptible to fine side-chain rearrangements. A useful and common practice in other Rosetta protocols is to integrate the full-atoms and the centroid modes, and we currently are working on integrating such an approach within the RRT framework to allow faster search of motion pathways in atomic detail. As experimental knowledge of intermediate states progresses, it should be particularly useful to optimize energy scoring functions for transient conformations, and not only for static low-energy states.

Future applications
Analyzing Multiple Motion Pathways

One of the advantages of RRT-based techniques is their speed, and thus a large body of motion pathways can be quickly created at atomic level that includes side-chain atom positions. A large number of pathways provide further insights about the connectivity of the conformational space under a wide range of settings. In contrast, it is difficult to generate a large number of pathways using, e.g., MD simulations, due to slower running times. We recently proposed a method to compare, cluster and merge multiple motion pathways from independent runs of the RRT algorithm. The merged pathways have lower energy than all input pathway [35]. It would be interesting to examine the clusters of pathways that are generated with different types of partial information, also in terms of robustness to variations in input partial information.

Applications to Other Types of Molecular Motion

In this work PathRover was applied to motions of domain swapping and substrate binding. However, different types of molecular motions might have different characteristics with respect to the number of torsion angles that are involved in the motion, the scale of the motion, the role of side-chains, etc. One challenging class of molecular motions involves allosteric protein motions [63]. In this case, a large number of torsion angles are often involved in the motion, but each of them changes by rather small increments, and partial information might constrain the overall nature of the motion. Another interesting type of motion involves more than one molecule, such as docking of a protein or a flexible peptide onto another protein. Of particular interest within this framework are cases where partial information can provide details about the approximate location of the interface, and conformational backbone flexibility of the monomer needs to be modeled efficiently [60,64].

Experimental and Computational Validation

Predictions that were made in this study, and in related studies by us and others [33-35], have been shown to correlate with existing experimental and theoretical observations. More importantly, our predictions can be readily subjected to further experimental validation so that biological insights deduced from these trajectories are verifiable. Within the framework of PathRover, it will be interesting to examine the accuracy of RRT for variable time-scales, degrees of freedom and aggressiveness of sampling. While full-atom experimental motion pathways of high resolution are still not in sight, more partial information on pathway arises quite naturally in molecular biology research. Knowledge about alternative conformations, or even homologues, can be incorporated in order to direct the search for plausible motion pathways. As a natural extension of the present study, we plan to analyze more predicates, e.g. NMR constraints. Such constraints have contributed to computational sampling of single conformations in the past, but have not been examined in sampling of full conformational motions. In this sense, our vision is that innovative experimental measurements of limited scope, such as single molecule FRET, advanced NMR measurements and CD spectroscopy can help focus and enhance computational techniques, and will effectively allow researchers to generate realistic and energetically favorable motion pathways that incorporate as much external information as possible within the currently suggested framework of PathRover.
Particularly, this can allow for the design of experiments that target specific states within a motion pathway based on *in silico* predictions of large-scale motions.

**Conclusions**

This study proposes PathRover as a general and flexible setup where novel molecular systems can be explored, and constraints can be incorporated in a general and straightforward manner. Partial information can improve the performance of sampling based algorithms, by narrowing down the search in the vast conformational space of proteins. This is demonstrated in the present study on a number of molecular motions of specific interest. Future work will concentrate on refining protocols for additional systems and types of motions. Beneficial crosstalk between experimental procedures and *in silico* simulations will ultimately optimize the wide integration of partial information into fast sampling-based algorithms – and forward our general understanding of protein motion and function.

**Acknowledgements**

We thank the many scientists who have participated in the development of the Rosetta software suite. We thank Dan Reshef, Nir London and Nir Ben-Tal for critical reading of the manuscript and useful advice. All structural images have been produced using the PyMol software [65].
References


<table>
<thead>
<tr>
<th>Name of predicate</th>
<th>Predicate formal definition</th>
<th>Motivating examples of relevant partial information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSD-Match &amp; Aligned RMSD-Match</strong></td>
<td>RMSD between subsets of Cα atoms (or all residues) in two conformations, with or without alignment</td>
<td>The target structure of a native protein or a homologue is given; only the structure of an active site region is known; known relative orientation of two secondary structures</td>
</tr>
<tr>
<td><strong>Line-Distance &amp; Line-Angle</strong></td>
<td>The distance, or angle, between two subsets of Cα atoms, fitted by a least mean square line (LMSL)</td>
<td>Pairing of two beta strands, relative orientation between a helix and a sheet, etc. This method is less susceptible to residue shift than RMSD-Match</td>
</tr>
<tr>
<td><strong>Pair-Distance</strong></td>
<td>The distance between a pair of residues</td>
<td>Experimental distance constraints (Spin-Label NMR, Single-Molecule FRET, Cross-Linking); known salt bridge between charged residues</td>
</tr>
<tr>
<td><strong>Cent-Mass Distance</strong></td>
<td>The distance between the center of mass of two subsets of Cα atoms</td>
<td>A coarse Cryo-electron Microscopy image indicates the center of mass distance between two domains</td>
</tr>
<tr>
<td><strong>Form-H-Bond</strong></td>
<td>The extent of H-bond formation in the protein</td>
<td>Circular Dichroism (CD) spectroscopy indicates the creation of helices / sheets, but provides no indication regarding their specific location within the protein sequence</td>
</tr>
</tbody>
</table>

**Table 1**: Examples for predicates of partial information in PathRover (motivated by experimental techniques) and comparative methods for analysis of the resulting paths.
Table 2 Moving CesT swapped chain towards SigE homologue with five different types of partial information, see legend box and Figure 2. All values are given in Angstroms, except where indicated. (a) Comparison of predicate evaluation on Start (CesT), Target (SigE) and Final conformation in the motion pathway. (b) RMSD between the final conformation and either SigE, or the CesT pseudo-monomer conformation (see Figure 2b for comparison). Although SigE was used to direct the motion, the final conformation is, in general, more similar to the CesT pseudo-monomer, except in one type of partial information (first column).

### Table 2

<table>
<thead>
<tr>
<th>Partial Data Evaluated on</th>
<th>(i) LLD</th>
<th>(ii)† LLA††</th>
<th>(iii) CMD</th>
<th>(iv)</th>
<th>(v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Conformation (CesT)</td>
<td>39.1</td>
<td>25.1</td>
<td>91°</td>
<td>47.2</td>
<td>13.8</td>
</tr>
<tr>
<td>Target Conformation (SigE)</td>
<td>6.1</td>
<td>8.8</td>
<td>279°</td>
<td>15.0</td>
<td>0</td>
</tr>
<tr>
<td>Final Conformation</td>
<td>6.1</td>
<td>8.6</td>
<td>275°</td>
<td>11.1</td>
<td>1.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSD to Final Conf.</th>
<th>SigE</th>
<th>pseudo-monomer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>6.7</strong></td>
<td><strong>8.9</strong></td>
</tr>
</tbody>
</table>

† LLD / LLA = Least Mean Square Line Distance / Angle ; CMD = Center of Mass Distance
†† LLA measured in angles

Legend for types of partial information: The technical formulation of predicates is listed in Table 1. (i) Pair-Distance predicate between atoms (PHE12;LEU95) in CesT and (PHE12;GLU110) in SigE. (ii) Line-Distance, Line-Angle and Cent-Mass Distance predicates between helices in different domains of CesT and SigE (H1 and H3 in Figure 2a). (iii) / (iv) / (v) Aligned RMSD-Match predicate between pairs of secondary structures (α-helices: H1 and H3, β-strands:B0’ and B1; see Figure 2a) in different domains of CesT and SigE: (iii) α-helices (iv) β-strands (v) α-helices + β-strands.
Figure Legends

Figure 1
Comparison of pathway motion predictions in a 2D toy example. Here, we aim to find collision-free paths for a point robot in 2D-space starting from a source configuration. (a) The basic Single-RRT algorithm provides fast but rough coverage of unexplored regions, and the target is often missed (red star, top left). During the run, the tree grows in feasible space (white) among obstacles (orange rectangles). In biological examples, these obstacles are high-energy conformations. Each point stands for a two dimensional conformation, and the tree grows from a source conformation (violet star, middle of figure), towards random directions (see Methods). (b) In the Partial-RRT variant, we use partial information to truncate branches that do not grow towards the target (like the truncated branch in the grey ellipse, compare to the branch in the magenta ellipse). The search is more confined to relevant regions, at the expense of overall coverage of the search space. (c, d) Comparison between the basic Single-RRT algorithm and RRT with partial information (Partial-RRT), for the toy example in a and b. The partial information used here is the distance to the target. In SingleRRT-t50 and PartialRRT-t50 the target is also used as an explicit direction of growth once in 50 iterations, in case the tree reaches the proximity of the target but not its exact location. This test follows a common assumption that RRT running time is dominated by the number of collision tests. We compare the euclidean distance of the RRT node that is closest to the target (y-axis) as a function of the number of collision tests (x-axis) throughout the run. Results are the average distance (in c) or the minimum distance (in d) over 50 independent runs. PartialRRT performed better than SingleRRT, especially for a lower number of collision checks. Better performance is achieved in less running time. As the number of collision checks grows to infinity, all methods converge. Note that this is only a toy example for illustrative purposes; In many biological examples, the target conformation might not be given at all, and the dimensionality is in general much higher, so the number of collision tests or energy calculations is low compared to the complexity of the problem.

Figure 2
Partial information over CesT domain swapping (a) In the swapped CesT crystal, the pseudo-monomer comprises two domains from different chains. The C-terminal (blue) is from a different chain than the N-terminal domain (red). The N-terminal includes helix H1 and strand B0. Monomers of distant homologues can be aligned to the pseudo-monomer (b) In the same chain of the swapped conformation, the N-terminal domain (H1 and B0, magenta circle) is not packed to the C-terminal domain (blue). They are connected by a small hinge (black). When the domains unswap, the N-terminal domain repacks against the C-terminal, probably in the same orientation as the pseudo-monomer of 2a. (c) The final conformation along the motion pathway of CesT (cyan) is shown for five different examples of predicates (see Results). The orientation of the N-terminal domain of the pseudo-monomer is shown in red for comparison. The predicates (also see Table 1 and 2) are based on homologue SigE. (i) Pair-Distance constraint on C-alpha atoms of residues PHE12 and LEU95 (spheres). PHE12 is also shown for the N-terminal domain of the pseudo-monomer (red). Note that the C-terminal domain is fixed, so only PHE12 moves during the motion. For clarity, this structure is not from the same viewpoint as the other structures (ii) Helix alignment by Line-Distance, Line-Angle and Centroid-of-mass Distance predicates between helices H1 and H3. The lines fit the helix axes. (iii) Helix alignment by Aligned-RMSD-Match predicate. (iv) Strand alignment (strands B0 and B1) by Aligned-RMSD-Match predicate. (v) Joint helix and strand alignment by Aligned-RMSD-Match predicate.

Figure 3
Swapping and unlocking of Cyanovirin-N domains (a) Cyanovirin chain in a swapped orientation (pdb-id 1L5E). N-Terminal (D1, pink, residues 4-48) and C-terminal (D2, blue, residues 55-100) repeat domains are connected via a central loop (L’, magenta). Blue and yellow show possibly flexible linkers within each domain (see Results). (b) Cyanovirin in locked monomer orientation (pdb-id 2EZM, side-chains not shown for clarity) (c) The phi/psi values of the residues along the central loop L’ throughout the motion are plotted over a Ramachandran plot. PRO51 (cyan) lies in a characteristic Proline region. Neighboring SS52 lies in a constrained region of the plot. The SS25 mutation was shown to destabilize the monomeric conformation [55]. (d) Energy plot of Cyanovirin monomer unpacking. Once in an open conformation (step 36), the monomer shows a transition to a “free” energy landscape with only mild energy barriers, in agreement with the diverse conformations observed for domain swapped Cyanovirin.

Figure 4
Ribose Binding Protein (RBP) architecture (a) The RBP structure in its open and closed form (pdb-ids 1URP and 2DRI respectively). Domains are in cyan and blue. (b) The architecture of RBP – each domain consists of discontinuous segments of residues. The two domains are connected by three hinges that must move in a coordinated fashion in order to maintain
domain integrity. Domain boundaries are rough estimates (c) The sequence of RBP showing the discontinuous domains and the secondary structures. Domain A in cyan box, Domain B in blue box, hinge regions indicated by red vertical line. This Illustration was taken from 1urp Protein Data Bank entry at http://www.rcsb.org/pdb/ [5]; and domain assignments are from [66].
Discussion

The prediction of TM protein structures is an ongoing important challenge due to the difficulty of determination of these structures experimentally. During the last decade, there was an increase in the pace of structure determination of TM proteins (Fleishman et al., 2006). However, the rate of structure determination of TM proteins lags behind that of soluble proteins. The main objective of my work was to develop computational tools for assistance in the prediction of structure and motion in TM proteins of the alpha-helix bundle type.

Structure Prediction. I have developed a novel method for assigning TM segments in the sequence of a transmembrane protein to helices in intermediate-resolution structures (M1: Enosh et al., 2004), which is a first step toward modeling of TM proteins. My method eliminates many putative helix assignments based on the (estimated) maximal lengths of each of the loops in the protein. Moreover, the feasible assignments were ranked by a novel score function based on the capability of loops to connect pairs of helices. I ranked feasible assignments of TM-protein chains of known structures taken from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/), and according to the results, the native-state assignments rank high.

The results show that my method can eliminate many infeasible assignments when the extramembrane segments are of short length (i.e., 7 amino-acids or less). For example, in the case of lactose permease, only 12 out of 479,000,000 putative assignments were found to be feasible due to the short lengths of its extramembrane segments. This 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 program can be extended to the problem of oligomeric proteins. This extension is quite straightforward since the assignment complexity depends on the loop lengths rather than the number of helices in the units. When parts of the structure are related by symmetry or quasi-symmetry (M2: Fleishman et al., 2006) the possible assignments can be further reduced. In addition, one can add to the current program other geometric constraints that can be derived empirically. For example, consider secondary-structure elements that may be part of loops comprising the extramembrane segments; obviously, these elements impose additional constraints on the loop lengths.

The suggested score function for ranking the capability of a loop to connect pairs of helices has shown success for short loops (i.e., loops comprising 7 amino-acids or less) regardless of their amino-acid types. This score function is a knowledge-based function that was derived based on a data-set of helix-loop-helix motifs of soluble proteins. Formally, if we fix the N-terminus side of the loop in a specific position in three-dimensional space, the score function stores the reachable space of the C-terminus end of the loop. Due to the relatively small number of helix-loop-helix motifs in the database where the loops in the motif comprise more than seven amino-acids, the reachable space of long loops cannot be deduced by our score function. Thus, it would be interesting to explore the reachable space for longer loops with other means as geometric constrains rather than relying on a data-set. In other words, similar to the Ramachandran plot that divides the two-dimensional torsion angle space to allowed and restricted regions for one amino-acid, the reachable space of loops of length K can be divided on a 2K-dimensional torsion-angles space with K amino-acids. Moreover, in the current implementation, all the loops were analyzed merely by their lengths and therefore it would be appealing to develop a more reliable score function that considers the amino-acids types along the loops, which in turn can lead to a more profound understating of loops’ conformational space.

Motion Prediction. The tracing of motion between conformations of TM proteins is crucial for understanding their mechanisms. However, due to technical obstacles, in most cases in which
structures of TM proteins have been solved, the structure of only one conformation is known. In the thesis I have developed new computational tools to predict other feasible conformations and to suggest pathways between them. These tools are based on advanced geometric algorithms combined with energy-based criteria, and enable the computation of conformational changes at rates much faster than those attainable by current simulation methods.

In a preliminary work (M3: Enosh et al., 2007), I have focused on simple systems comprised pairs of alpha-helices. More recently, I extended this method to handle the pore-forming domain of the potassium channel from Streptomyces lividans, KcsA (M4: Enosh et al., 2008). Potassium channels are the most common type of ion channels. They regulate cellular processes, such as neuronal signaling and secretion of hormones (see a minireview on potassium channels by MacKinnon, 2003). The pore-forming region in the potassium channel is composed of a bundle of eight helices from four identical monomers, through which ions are conducted. The addition of more helices obviously increases the number of dofs. However, it also reduces the feasible conformation space due to self-avoidance effects. The pathway that was found to be the most plausible is in agreement with previous studies (Shrivastava and Bahar, 2006; Shimizu et al., 2008) in spite of the use of very different search methodology and energy function, which may suggest that the motion is governed predominantly by geometric restraints.

The flexible nature of the algorithm enables testing the effects of changes in the energy function on the results. It is straightforward to integrate the algorithm with various molecular simulation packages that use different forcefields. In the analysis of the gating process in KcsA, I have applied two different forcefields to examine the conformations emerged from the RRT, namely, the Lennard-Jones (LJ) potential as implemented in the CHARMM (Neria et al., 1996) and in GROMOS (Oostenbrink et al., 2004) forcefields. As part of a joint work with Barak Raveh (M5:Raveh* et al., 2008), we have integrated the RRT framework within the Rosetta software package (Rohl et al., 2004) and thus examined the gating process with a version of Rosetta score12 energy function that includes attractive and repulsive VDW energy, hydrogen-bonding terms, and a statistical bias for prevalent Ramachandran torsion angles and side-chain rotamers. All the emerging pathways lead to the opening of the gate. However, when strictly LJ-potential was considered, the resulting pathways were absent the rotational mechanism that was observed by previous scientists (Shrivastava and Bahar, 2006; Shimizu et al., 2008). Using score12, we indeed obtained a rotational mechanism that we referred to as the three-phase safe-lock mechanism (M4: Enosh et al., 2008). This observation demonstrates the importance of the choice of the forcefield and it might suggest that in order to gain a better understating of motion pathways, various forcefields should be considered with care. Although an energy function for TM-proteins has been recently introduced into Rosetta (Barth et al., 2007), it was not made publicly available. Thus, it would be interesting to examine the suggested motion pathway with this particular energy function when it will be available.

In (M5:Raveh* et al., 2008), we have extended the RRT framework to various different applications in biology, e.g., to the study of dynamics in soluble proteins and specifically to the investigation of domain swap mechanisms in CesT (a type III secretion chaperone in Enteropathogenic E. coli) and Cyanovirin-N (an anti-viral fusion inhibitor protein that binds to viral sugars). Moreover, the RRT algorithm was extended to find motion pathways that follow specific constrains (e.g., loop closure constraints) during the pathway or motion pathways towards vague targets that are defined by partial biological information that is known for the protein in study (by, e.g., NMR, spin labeling, homologous structures, etc.).

RRT can be applied to various other problems in the field of bioinformatics, such as, model refinement applications and docking refinement. In a preliminary work, I have extended the RRT to handle these two problems. However, I have examined only one case for each of these problem, therefore a further investigation is required. A model structure of a TM protein derived from cryo-EM data includes the approximate locations of the Cα atoms along the helices embedded within the membrane. Given the Cα template model, canonical helices can be built to
fit the C\(^{\alpha}\) model template. Then, my RRT-based algorithm may be used to conduct an exhaustive search for collision-free conformations in the vicinity of the template helices. In my current implementation, the search includes rigid body movements of the helices, minor changes of the dihedral angles of the backbone (up to \(\pm 10^\circ\) from the \(\alpha\)-helix values) and changes in the rotameric states of the side-chains. The collision-free conformations obtained this way are assigned energy scores depending on the selected forcefield, and are clustered based on mutual rmsd. The rank of these clusters based on the energies of their conformations may suggest the most favorable conformations that should be further examined by experimental data.

In a similar way, the RRT-framework can be used to refine rigid-body docking solutions (as alternative approach to FireDock; Andrusier et al., 2007). Candidates of the complex structure, generated, e.g., by PatchDock (Duhovny et al., 2002), may be refined by generating an RRT that includes various conformations in the vicinity of these candidates by small changes in the orientation between the unbounded structures and changes in the rotameric states of side-chains in their interface. Due to the rapid running time of my algorithm, many complex structures can be generated and analyzed further by means of their energy using various forcefields.

As stated before, the RRT framework generates energetically feasible motion pathways. However, due to the random nature of the RRT, different runs can lead to different pathways. This emphasizes the need for a computational tool that compares and clusters motion pathways. To this end, I have developed a novel framework for the comparison and hybridization of pathways. This framework is modular and may be used in various other applications, e.g., within the context of molecular dynamics and Monte Carlo simulations or for comparing motion paths in Robotics. In the current implementation, the energy (defined by the forcefield in use) was used to rank the different conformations along the pathways and a preferable pathway was hybridized for each cluster based on the other generated pathways. Yet, many other scoring functions may be integrated in a much broader context such as in motion planning in Robotics. In this case, the forcefield can be replaced by different scoring functions depending on the problem at hand.

The use of the RRT for molecular pathway prediction is relatively new, and suggested pathways should be regarded as a proof of concept for their feasibility and further examined by experimental techniques. For highly constrained conformation spaces, where only several motion pathways are feasible, I believe that the RRT can provide insight into their mechanism of action. In order to investigate the quality of the motion pathways suggested by the RRT, it is essential to compare these pathways to more reliable pathways emerging from, e.g., MD simulations or to pathways generated by other common methods such as MC simulations. This comparison is vital for the approval of the RRT framework as a reliable tool for mechanism elucidation in biology.

* Angela Enoosh and Bark Raveh contributed equally to this work.
REFERENCES


RRT (LaValle and Kuffner, 2001; LaValle, 2006) is a method for finding a collision-free path in a configuration space. It consists of a probabilistic roadmap (PRM) technique. In order to find a collision-free path, a number of possible paths are generated from the source configuration to the target configuration. Each path is then evaluated based on its potential to reach the target configuration. The evaluation is based on a heuristic function that considers the distance between the two configurations. The path with the lowest cost is chosen as the final path.

The RRT algorithm was developed to solve the problem of finding a collision-free path in a configuration space. It is particularly useful when the configuration space is large and complex. The algorithm works by randomly sampling points in the configuration space and connecting them to form a roadmap. The roadmap is then used to find a path from the source configuration to the target configuration. The process is repeated until a collision-free path is found.

The RRT algorithm has been applied in various fields, including robotics, computer vision, and evolutionary biology. In robotics, RRTs are used to plan collision-free paths for robots in environments with obstacles. In computer vision, RRTs are used to plan paths for robots to navigate through cluttered environments. In evolutionary biology, RRTs are used to simulate the evolution of complex systems, such as protein structures and gene regulatory networks.
A probabilistic roadmap (PRM) algorithm is one of the first and most well-known methods for sampling-based motion planning.


The PRM algorithm first generates a large number of random points in the configuration space and then connects neighboring points to form a roadmap. The basic steps of the PRM algorithm are as follows:

1. **Sampling:** Draw a large number of random points in the configuration space.
2. **Nearest Neighbor:** For each new point, find its nearest neighbor in the roadmap.
3. **Connect:** If the new point is sufficiently close to the nearest neighbor, add a connection between them.
4. **Avoid Collisions:** Ensure that the connection does not intersect any obstacles.
5. **Optimize:** Optionally, refine the roadmap by optimizing the connections for efficiency.

The PRM algorithm is particularly useful in high-dimensional spaces, where other algorithms might struggle.

Rapidly-exploring Random Trees (RRTs) are similar to PRMs but explore the configuration space in a more directed manner. They start from a random point and try to reach another random point. If this fails, the algorithm tries to reach an existing point in the roadmap. This process repeats, gradually filling in the space until a connection is found between the start and goal configurations.

Expansive Spaces Trees (ESTs) are another variant of PRMs that can handle higher-dimensional and more complex spaces.
Monte-Carlo (MC) Simulations

Exploring complex and non-trivial conformational spaces requires sophisticated computational methods. Steered Molecular Dynamics (SMD) and Targeted Molecular Dynamics (TMD) are two such methods. SMD, for instance, allows for the calculation of large conformational changes in timescales not accessible to standard MD, by applying external forces to the system (Izrailev et al., 1998). TMD, on the other hand, can be used to sample and control the movement of molecules by closely mimicking the behavior of two conformations (Schlitter et al., 1994).

Normal Mode Analysis (NMA)

Normal Mode Analysis is a technique widely used in computational chemistry. It is based on the concept of normal modes, which are the principal eigenvectors of the Hessian matrix of a system. The Hessian matrix contains second derivatives of the potential energy function with respect to the atomic coordinates. The eigenvalues and eigenvectors of this matrix provide information about the vibrational modes of the molecule, with the lowest frequencies corresponding to the largest normal modes.

For example, in proteins, the largest normal modes are often associated with global motions, such as domain movements or overall protein folding. This method has been applied extensively to study protein dynamics and has been particularly useful in understanding the mechanisms of protein-protein interactions and protein-ligand binding.

Source: Haywood, 2001; Krebs et al., 2002; Cui et al., 2004; Trakhanov et al., 2005.

References:
- Izrailev et al., 1998
- Schlitter et al., 1994
- Doyle et al., 1998
- Lotan et al., 2004
- Binder and Heerman, 1992
- Becker, 2001
- Holm and Sander, 1992
- Li and Scheraga, 1987
- Shimada et al., 2001
- Apaydin et al., 2002
- Apaydin et al., 2003
- Tikhonov and Zhorov, 2004
- Haywood, 2001
- Krebs et al., 2002
- Cui et al., 2004
- Trakhanov et al., 2005
- Ma and Karplus, 1998
- Krebs et al., 2002
- Cui et al., 2004
- Trakhanov et al., 2005.
In my research, I developed a new method to find the alignment between transmembrane segments that were determined from cryo-EM images. I used non-geometric constraints in M1: Enosh et al., 2004 and M2: Fleishman et al., 2006.

We proposed a mechanism for substrate binding and presented results of simulations with a partner in a collaborative work with EmrE. (M2: Fleishman et al., 2006).

Predicting sequence trajectories is like solving a puzzle, where the movements of a protein are composed of various types of movements: side-chains, conformational changes, slow and large movements in domains.

Large movements also depend on the type and nature of the movements and are related to functional movements of the protein. Therefore, it is important to consider small movements that are critical for tracking the movements of two protein conformations in a transmembrane segment. It is difficult to establish these proteins, and thus, the structure of the conformation is often determined due to technical difficulties. The mechanism of the protein, the prediction of additional conformations, and the equipment I developed are based on geometric algorithms. Among these, simulations based on force fields are essential for success in any computational approach to study dynamics in molecules. The quality of the potential functions is crucial for such evaluations. These functions are intended to estimate the flexibility of the system. However, these methods are limited to small molecules. They are based on quantum mechanics. The functions were developed empirically based on classical physics, which is a realistic way to calculate the energy of these functions. These functions are dependent on the types of atoms in the system.

A cluster of parameters is often used in simulations of movements in order to rank the conformations according to their energy. However, there is a wide variety of force fields, and some of them are used in computational studies. There are also comparisons between these force fields, for example, in Whitlow and Teeter, 1986; Kini and Eveans, 1992; Hobza et al., 1997. The same force fields are used in simulations of the same molecule, but their nature depends on the system being studied.

Molecular Dynamics (MD) Simulations

Molecular Dynamics (MD) simulations were first performed at the atomic resolution in 1977 in simulations by McCammon et al. Using force fields such as CHARMM (Chemistry at Harvard Molecular Mechanics; Brooks et al., 1983; MacKerell et al., 1998), AMBER (Assisted Model Building with Energy Refinement; Pearlman et al., 1991; Pearlman et al., 1995; Case, et al., 2002), GROMOS (Groningen Molecular Simulation; Berendsen, et al., 1995; Gunsteren et al., 1996), and many others.

This field has expanded significantly over time, and different force fields have been developed.
החלבונים חוצי למברנה הם חלק מרכזי של מembrana כל התהליכים במגוון רחב של מערכות ביולוגיות. בקובנה שלこれら המחלקים הם נמצאים מבני ומאגרים של יונים, חלקם גם מגדירים את המיקום של אנרגיות ויציבות של התהליכים שנעשים בתהליך זה. הם גם מהווים מטרות רופיות וידע נוסף, כאשר מחקר ופיתוחם מתנהלים על ידי מחקר תזתות וסיבות שונות של חלבונים חוצי למברנה.

במהלך הדוקטורט שלי פיתחתי כלים חישוביים לניבוי מבנה ותנועה בחלבונים. זה יכול לעזור בתכנון תרופותizona הificant את הידע שלנו בהקשר של חלבונים חוצי למברנה. מבני והם מתנהלים על ידי מחקר תזתות וסיבות שונות של חלבונים חוצי למברנה.

החלבונים חוצי למברנה הם מבני ומפתחים ובו של לוח וإنتاج חלבונים חוצי למברנה. מבני והם מתנהלים על ידי מחקר תזתות וסיבות שונות של חלבונים חוצי למברנה.

כדי למברנה מבנים של מספר חלבונים חוצי למברנה נקבעו ברזולוציה של cryo-EM (cryo-electron microscopy)._Currently, we have cryo-EM data of several membrane-bound helicases.cryo-EM data of several membrane-bound helicases.cryo-EM data of several membrane-bound helicases.cryo-EM data of several membrane-bound helicases.


למשל, ניתן למצוא ערכות של חלבונים חוצי למברנה שלSIDE CHAINS

(1) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(2) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(3) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(4) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(5) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(6) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(7) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(8) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(9) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(10) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(11) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(12) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(13) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(14) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(15) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(16) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(17) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(18) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(19) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(20) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(21) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(22) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(23) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(24) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(25) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(26) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(27) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(28) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(29) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(30) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(31) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(32) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; سאמavored, 2001)

(33) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(34) שעון, קינון ויסודות, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(35) שעון, קינון ויסedores, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(36) שעון, קינון ויסedores, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(37) שעון, קינון ויסedores, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(38) שעון, קינון ויסedores, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(39) שעון, קינון ויס.RunWith, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(40) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(41) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(42) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(43) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(44) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(45) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(46) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(47) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(48) שעון, קינון ויס.codigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(49) שעון, קינון ויסodigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(50) שעון, קינון ויסodigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(51) שעון, קינון ויסodigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(52) שעון, קינון ויסodigo, 1997; נאנגר, נל, 1999; סאמavored, 2001)

(53) שעון, קינון ויסodigo, 1997; נאנגר, נל, 1999; סאמavor
מציתת חציו הלבונים - בין אותות של העברה הכוללים, רביםתאיים בתהליכים תפקידים בעליهم מבנה הממברנה. הם משמשים לפיכך. התא ארגון הקשורים ותהליכים הממברנה דרך מולקולות יונים של ההעבר, תאים חוצי חלבונים. תרופות אטרקטיביות כמטרות כמהוים ממברנה 20–30%. בגןום המקודדים מהחלבונים בשיטות, לדוגמא )גבוה ברזולוציה נקבע אלו חלבונים של המבנים קטן החלק רק, הרבה חשיבותם אף על פי חלבונים של גנום שובייה ניקוי, כי ההפקה בתהליך הקשורות טכניות בעיות בגלל. אלו חלבונים של אלה פיתח את המכניזם המבנה לניתוח חישובי יטיות בפיתוח ממשי יצחו םחלבוניה מבנה שלם המיקו לגבי הערכה - בתלת הממברנה - תיבתיו את התאמה אתמצא הוא והאתגר, ימדמ באופניShockأن חישובי כלי פיתחתי כך לצורך. אלו קסיםיההל לבין ברצף ממברנה חציו מקטעים ביןemmפות שנגזרו ההליקס ואת Jaime. חציו בחלבונים תנועהל - חציו חלבונים. ביולוגיות תופעות במגוון חשוב תפקיד יש ממברנה - ממברנה. הם, בנוסף, הממברנה דרך ומד👏גולים יונים של להעברה שאחראיות ותעלות משאבות יוצריםחלבון של מבנה נקבע כאשר, לפיכך. ואقيادة הורמונים, מולקולות יהיוילז האחראיים כרצפטורים מנוגנים והוצע כאו עד. שלו הפעולה למנגנון הקשורות שאלות עולות, גבוהה ברזולוציה ממברנה חוצה חציו קסיםילהה של םהמיקו לגבי הערכה - בתלת הממברנה - תיבתיו את התאמה אתמצא הוא והאתגר, ימדמ באופניShockanas מרופד של מבנה קול的方式来 לחקור - Jaime. הקולטן של הממברנה receptor (ErbB2 האשלגן תעלתשל וKcsA. מסלולי המקרים שניב. נמוכה אנרגיהנות ב קונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה ניסיוניות המדווחות בספרות._RESULTS ל שמצאתי מתאימיםהתנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלولي המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים ביותר הקצרים הם שהצעתי התנועה מסלולי המקרים שניב. נמוכה אנרגיהנהקונפורמציות הכוללים steheneren.
טוקולעה לפי מדענים מדעיים ע"שريمונד וברלי סאקרל

בית הספר למדעי המחשב

הפקולטה למדעים מדויקים ע"שريمונד וברלי סאקרל

השיטה האלגוריתמית לפייר מדבקת והנוגה בשלבונים

ורジー-يمنברה

הרבר לשים קבלת תואר "דוקטור בפילוסופיה"

מאת אנ'ל אברג

העבורה בוצעו בהנחייתו של

פרופ' דוד הלפרין ופרופ' ניר בן-מיל

נוהג לטסת של אנכי-ברשת תל-אביב

אוגוסט 2008