Improved Helix and Kink Characterization in Membrane Proteins Allows Evaluation of Kink Sequence Predictors

David N. Langelaan,[†] Michal Wieczorek,[†] Christian Blouin,[†,‡,§] and Jan K. Rainey*[,†,∥]

Department of Biochemistry & Molecular Biology and Centre for Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia B3H 1X5, Canada, Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 1W5, Canada, and Department of Chemistry, Dalhousie University, Halifax, Nova Scotia B3H 4J3, Canada

Received August 24, 2010

Although the α-helical secondary structure of proteins is well-defined, the exact causes and structures of helical kinks are not. This is especially important for transmembrane (TM) helices of integral membrane proteins, many of which contain kinks providing functional diversity despite predominantly helical structure. We have developed a Monte Carlo method based algorithm, MC-HELAN, to determine helical axes alongside positions and angles of helical kinks. Analysis of all nonredundant high-resolution α-helical membrane protein structures (842 TM helices from 205 polypeptide chains) revealed kinks in 64% of TM helices, demonstrating that a significantly greater proportion of TM helices are kinked than those indicated by previous analyses. The residue proline is over-represented by a factor >5 if it is two or three residues C-terminal to a bend. Prolines also cause kinks with larger kink angles than other residues. However, only 33% of TM kinks are in proximity to a proline. Machine learning techniques were used to test for sequence-based predictors of kinks. Although kinks are somewhat predicted by sequence, kink formation appears to be driven predominantly by other factors. This study provides an improved view of the prevalence and architecture of kinks in helical membrane proteins and highlights the fundamental inaccuracy of the typical topological depiction of helical membrane proteins as series of ideal helices.

INTRODUCTION

The α-helix is a common secondary structure in globular proteins and is often found spanning the lipid bilayer in membrane proteins. The characteristics of an ideal α-helix are well-defined, but the features of disruptions in helices are poorly understood. Even though it is well-recognized that membrane-spanning α-helices from mono- or polytopic membrane proteins may be kinked,[1−5] there is disagreement in the literature concerning the nature, amino acid composition, and underlying causes of helical kinks.[2,4] Compounding this issue, helical kinks are often neglected in the discussion and analysis of membrane protein structures. This neglect goes to the extent of regions of Protein Data Bank (PDB)[6] files for membrane proteins being annotated as helical even through regions of disruption. Kinks in otherwise α-helical transmembrane (TM) regions provide points that readily allow for conformational change and structural variability, making helical kinks often functionally important in proteins.[7−9]

Three analyses of kinking in data sets of α-helical membrane proteins have been carried out to date. In 2001, Riek et al. examined 119 TM helices from 11 membrane proteins, with annotation of noncanonical helical character (i.e., π-helices or 3_{10}-helices) as well as kinks.[2] This study identified 36 kinks in 31 helices, with ∼26% of helices being kinked and ∼16% of these kinked helices having two kinks. A 2004 study by Yohannan et al. examined a set of 39 kinks from 10 membrane protein structures.[4] This study had the unique aim of testing the hypothesis that kinks in TM helices can be traced back through evolution to ancestral proline residues and, as such, did not focus on kink detection and geometries. Finally, Hall et al. analyzed the positions of kinks in a larger subset of 405 TM helices,[1] determining that 44% of TM helices were kinked, with only 35% of these kinks being caused by prolines.

Here we present a new algorithm, implemented in a web-based Python application named MC-HELAN (Monte Carlo based HELix ANalysis). MC-HELAN uses heuristics to systematically detect and characterize helical kinks, regardless of whether the helix in question is an α-helix, π-helix, or 3_{10} helix. We classify kinks as either bends (a change in helix axis direction with all residues remaining helical) or disruptions (change in helix axis direction accompanied by a loss of helical character of the residues involved in the kink). Previous studies have used software to determine helical axes (ICM,[2,10] SIMULAID[11]) followed by analysis by in-house scripts or other software such as ProKink.[12] The HELANAL[13,14] program was also designed to locate kinks and classify helices. MC-HELAN is able to characterize and localize kinks using a Monte Carlo based algorithm.

Although MC-HELAN, HELANAL, and ProKink all have similar functions, the mechanism of kink detection and the output data from each program is unique. All three algorithms can be used to determine which residues are most likely to

* To whom correspondence should be addressed. E-mail: jan.rainey@dal.ca. Phone: (902) 494-4632. Fax: (902) 494-1355.
[†] Department of Biochemistry & Molecular Biology.
[‡] Faculty of Computer Science.
[§] Centre for Genomics and Evolutionary Bioinformatics.
[∥] Department of Chemistry.

10.1021/ci100324n
Published on Web 11/19/2010

be kinked as well as to get an estimate of the magnitude of the kink through different methods. HELANAL uses a local window of seven residues to calculate local bend angles, MC-HELAN determines its own helical axes through a Monte Carlo approach, and ProKink works in conjunction with SIMULAID[11] to determine helical axes.[15] In the output data, ProKink does not sort helices into categories and HELANAL sorts helices into kinked, curved, and linear categories, while MC-HELAN describes helices as either disrupted, bent, or linear. HELANAL and ProKink provide information on other helix parameters such as helix twist and rise, while MC-HELAN does not. Finally, MC-HELAN analyzes the TM helices and protein as a whole in the context of the plasma membrane, unlike either HELANAL or ProKink. Although MC-HELAN, HELANAL, and ProKink all allow analysis of helix kinks, these fundamental differences between the algorithms and output data make each program stand alone.

We have used MC-HELAN to examine all (as of March 2010) nonredundant α-helical membrane proteins contained in the PDB of Membrane Proteins[16] (PDB_TM). The data set was culled using PISCES.[17] To allow direct comparison of algorithm sensitivity, the same data set was analyzed using both HELANAL[13,14] and DSSP.[18,18] We present general features of bends and disruptions in membrane proteins and assess the predictive nature of sequence determinants in the vicinity of kinks.

MATERIALS AND METHODS

Data Set Preparation. The PDB_TM extensible markup language (XML) file and transformed PDB files for all structures identified as helical membrane-spanning proteins in the PDB_TM[16,19] were obtained (all depositions to March 4, 2010). The XML file identifies transmembrane polypeptide chains and the predicted N- and C-terminal boundaries of each α-helical (or β-sheet) TM span for each PDB entry. The PDB_TM transformed PDB files have atomic coordinates that are rotated and translated such that the putative membrane is in the xy-plane and centered at $z = 0$ using the TMDET algorithm.[20] The Topology Data Bank of Transmembrane Proteins (TOPDB[21]) was used to determine the orientational topology (i.e., direction of cytoplasmic face) whenever the corresponding entry existed. The algorithm PISCES[17] was used to cull the set of PDB_TM entries, keeping only X-ray crystal structures with resolution better than 3.0 Å and an R-factor better than 0.75 (constraints are not placed on NMR-derived structures; only NMR and X-ray structures were used) and only the highest resolution polypeptide chain in cases of pairwise sequence identity ≥95%. Since the number of membrane protein structures available is still small, we used this high cutoff to remove direct duplicates but retain as many nonduplicate structures as possible. Of the structures analyzed, 25% of them had a pairwise sequence identity >40%. For PDB entries consisting of ensembles of NMR structures, we have limited our analysis to the first model (i.e., that contained in the PDB_TM) rather than examining all ensemble members. The biological assembly for a given PDB entry was determined through parsing of both the PDB_TM xml file and the original PDB file. MC-HELAN analysis (using a Python program, freely available upon request, rather than the interactive webpage) was carried out on the data.

Monte Carlo Method for Helical Axis Definition. A residue at position i is denoted as a helix seed if it is not already classified as part of a helix and the following three empirically derived conditions are met: (1) The dihedral angle pair (ϕ_i, ψ_i) lies within the 99.95% boundary of Lovell et al.'s statistically derived α- and 3_{10}-helical region.[22] (2) The angles denoted by the triplet of points $\Theta_x = (C^\alpha_{i+x}, C^\alpha_i, C^\alpha_{i+1})$ for $x = 2, 3, 4$ lie within ranges of 35−50° for Θ_2, 60−80° for Θ_3, and 45−65°for Θ_4. (3) The distances between C^α_i and C^α_{i+x} for $x = 2, 3, 4$ deviate from the distances observed in an ideal helix by at most 0.5 Å. Similar geometric criteria have been employed in other studies to locate helix kinks.[2,23] For each identified seed, a nascent helix including the backbone atoms of residues (i, $i + 1$, ..., $i + 4$) is used to approximate a helical axis. The axis is denoted by the tuple (c, $\dot{v}$), where c is a 3D point and $\dot{v}$ is a vector. The fit of the axis is minimized by a uniform random sampling in each of the components of c and $\dot{v}$. An empirically determined objective function (E_h) is calculated, and the new parameters are kept only when E_h is lower. For the first 800 iterations, E_h is defined using

$$E_h = \sum E_r \quad \text{and} \quad E_r = \begin{cases} (r-2)^2 & \text{if} \quad r \geq 20 \\ 0 & \text{if} \quad r < 2 \end{cases} \tag{1}$$

where r is the minimum distance between a given atom and the helix axis and the sum is carried out over all heavy backbone atoms of the helical segment. Next, 200 further iterations use the following modified E_r

$$E_r = |r - \bar{r}|^2 \tag{2}$$

where $\bar{r}$ is the average distance of all heavy backbone atoms of the atoms from the helical axis. This procedure was adequate for proper convergence of an accurate helical axis over as few as five residues, as is readily seen in examination of MC-HELAN output files. After definition of a helix axis, the nascent helix is extended to include a neighboring residue if the residue has helical ϕ and ψ dihedral angles and the backbone atoms of the residue are on average <3 Å away from the helix axis. The helix axis is then recomputed and the process repeated until no additional residues can be added. This procedure is applied from the N-terminus to the C-terminus of a peptide chain and then again from the C-terminus to the N-terminus in order to define the most complete helices possible.

Identification of disrupted TM helices is trivial, based on the definition given above. Bends are identified in all instances where two helices are adjacent or overlapping. The precise location of the bend is determined as residue i using a voting procedure according to the following criteria: (1) The residue with the largest deviation from ideal helical dihedral angles ($\phi = -62°$, $\psi = -41°$)[24] if the sum of the deviation is larger than 40°. (2) That residue i has helical Θ_x angles while $i - 1$ and $i - 2$ do not. (3) That the distance between C^α_i and C^α_{i+4} deviates from the distance expected in an ideal helix by more than 0.5 Å and either the C^α_{i-1} and C^α_{i+3} distance deviates by >1 Å or the distance between C^α_{i-1} and C^α_{i+3} and C^α_{i-2} and C^α_{i+2} both deviate by >0.5 Å from an ideal helix. For each criterion satisfied a suspected "bent" residue is given a point. The residue with the most points is

Table 1. Transmembrane Helix Characteristics ($n = 840$) As Classified by MC-HELAN, HELANAL, and DSSP for the Nonredundant Set of Currently Solved Protein Structures ($n = 205$) Containing at Least One Transmembrane Helix (N/A is not applicable)

	MC-HELAN[a]	HELANAL	DSSP
linear helices	293	68	N/A
curved helices	N/A	382	N/A
kinked helices	516	387	N/A
bent helices	462	N/A	N/A
disrupted helices	54	N/A	142[c]
bends	543	N/A	N/A
disruptions	53	N/A	153[c]
kinks used for pairwise comparison[b]	559	158	153

[a] The reported numbers are only for the converged helices and kinks found by MC-HELAN. [b] Only kinks detected more than two residues from the end of a helix. Multiple kinks within three residues are merged. [c] Value obtained by parsing the files as in the pairwise analysis.

taken as the location of the bend. In the case of ties in the voting process, the location of the bend was determined by identifying the residue *i* in the overlapping segments, which optimizes the fit of the helical axis for both incident helices.

Bend Prediction from Primary Sequence. The ability to identify bends from sequence information alone was explored using a machine learning approach. The identification of a bend position was formulated as the classification for a window of nine residues to have a bend site in position 5 or not. A data set of 471 converged sequences of length nine that contain a bend in the fifth (middle) position was compiled. Likewise, a set of 9067 sequences of equal length that do not contain a bend was compiled. In total, 10 classification replicates were performed using a 10-fold cross-validation on the 471 bent sequences combined with a random sample of negative examples of equal size. A feature vector for each character was defined with one or more numerical values from the set of possible features: hydropathy,[25] membrane helical propensity,[26] relative proportion of residue identity (Supporting Information, Table S1), and a binary encoding for six physicochemical properties (hydrophobic, polar uncharged, cationic, anionic, glycine, or proline). The final feature vector for one sequence instance is the concatenation of these features for each character in the input sequence. As a measure of baseline prediction considering only the presence of a proline residue, an alternative binary feature vector was used for each residue in the bend local sequence being assigned 1 for proline or 0 for nonproline. The performance of the classifiers was evaluated using precision, recall, and the *F*-score. The classifier used was the support vector machine (SVM)[27] using the radial basis function kernel. For each classification experiment, optimal C and γ parameters were determined using a grid search.

RESULTS AND DISCUSSION

Data Set Characteristics. After PISCES culling, the membrane protein data set consisted of 840 TM helices from 205 polypeptide chains in 137 PDB files derived by both X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (Table 1). Highlighting the underrepresentation of polytopic membrane protein structures in the PDB, of the 205 TM polypeptide chains, only 129 (63%) are polytopic (Figure 1A). The bitopic proteins were subdivided into monomeric vs oligomeric biological assemblies, with 57/74 being structured as oligomers.

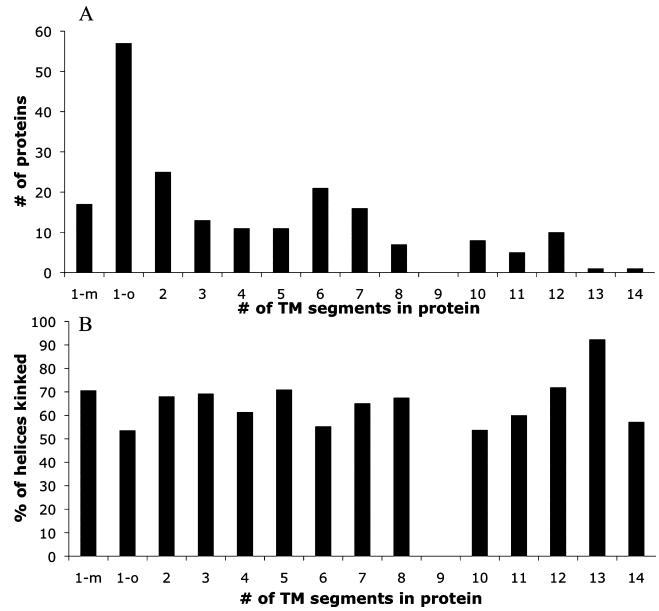

Figure 1. Composition of the culled helical membrane protein data set. (A) Number of polypeptide chains (205 total) with the indicated number of transmembrane (TM) helices. (B) The proportion of kinked (i.e., bent or disrupted) helical segments identified by MC-HELAN in TM proteins of each size. Only TM segments that converged by MC-HELAN analysis (>97%) are considered in part B. The bitopic proteins have been subdivided into those that are in a monomeric state (1-m) and those which are part of an oligomeric complex (1-o).

The MC-HELAN Server. For users wishing to analyze helical axes and kinks in a single protein (whether membrane-spanning or not), the Python program MC-HELAN is available in a web-mounted format (freely accessible at http://structbio.biochem.dal.ca/jrainey/MC-HELAN/). A sample MC-HELAN analysis of a protein, showing both the modified PDB file produced and the resulting membrane protein topology diagram, is provided in Figure 2. All reported statistics were derived using MC-HELAN in a batch script format (freely available upon request).

Convergence of MC-HELAN. Since the MC-HELAN algorithm relies on a Monte Carlo algorithm, identification of the same residue as being kinked in multiple analyses of the same data is not guaranteed. In order to assess the degree of convergence, the MC-HELAN analysis was completed 10 times for every protein to evaluate the standard deviation values in kink angles and the consistency of kink locations. In this work, a kink is retained for further analysis only if all 10 replicate predictions converged to a single solution. All converged kinks are enumerated in Table 1.

Overall, 596 of the 614 kinks identified by MC-HELAN (>97%) converged to a unique solution. The presence of nonconverged kinks was not due to failure of the algorithm in defining an accurate helix axis. Failure to converge was attributable in general to either a small kink (~5°−10°) or to the location of a kink being assigned to pairs of neighboring residues. The defined kink angles were well-converged with the average value of the standard deviation of all converged kink angles being 0.043° and the largest standard deviation value being 4.2°. It should be noted that

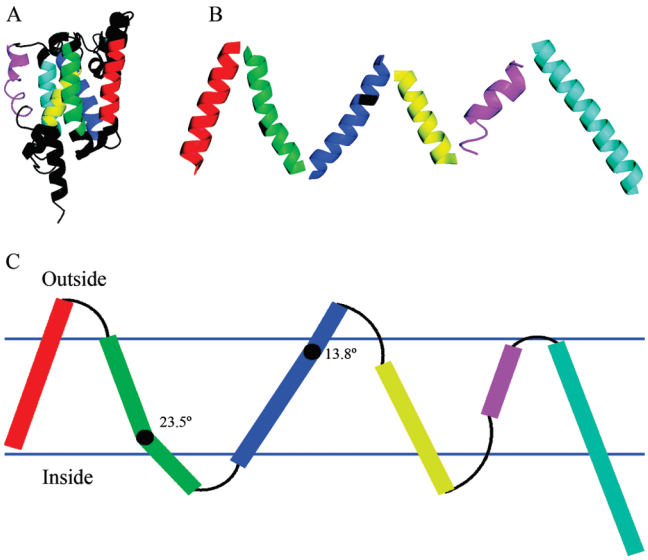

Figure 2. MC-HELAN analysis of a rhomboid peptidase from *Haemophilus influenzae* (PDB entry 2nr9 chain A[32]). (A) Cartoon representation of the entire transmembrane (TM) domain of 2nr9, with the residues identified by TMDET[20] as being membrane-spanning colored by TM helix. (B) MC-HELAN output illustrating the kinks detected (colored black) in each of the four TM helicies of 2nr9. (C) The 2nr9 TM domain topology diagram generated by MC-HELAN, with calculated bend or disruption angles indicated. The orientations and positions of each TM helix with respect to the putative membrane calculated by TMDET alongside kink characteristics determined by MC-HELAN are represented in the topology diagram. The cytoplasmic side of the membrane (inside) is indicated. Pymol (Delano Scientific, San Carlos, CA) was used for visualization of both parts A and B; the graphical topology diagram output from MC-HELAN (http://structbio.biochem.dal.ca/jrainey/MC-HELAN) is shown in part C.

there were only four kinks (<0.7%) that had a standard deviation of >1°. In these four cases, the kinks being analyzed had geometries that led to difficulty in exact helix axis definition by the MC-HELAN algorithm. Since both the convergence and the standard deviation for kink angle of a set of 10 predictions are a part of the standard output of MC-HELAN, lack of convergence or high standard deviation values, while generally very rare, should be employed by the user to highlight kinks where manual interpretation may be necessary.

Frequencies of Kinks. Out of 810 TM segments with 100% convergence when analyzed by MC-HELAN, 293 (36%) were an unperturbed helix, 462 (57%) contained a bend, 54 (7%) a disruption, and one segment had no helical sections. The proportion of kinks detected (64%) is significantly higher than those detected previously in TM helices using other methods (26%[2] and 44%[1]). MC-HELAN detects more kinks and, unlike HELANAL, does not consider "curved" helices as a separate class. In addition, the extension over dynamic regions of a helix from both N- and C-termini in the MC-HELAN algorithm improves sensitivity to small deviations in the helical axis that may be missed by algorithms that rely strictly upon local parameters over fewer residues. The MC-HELAN algorithm also avoids the issue that least-squares fitting derived helical axes are dependent upon the number of residues analyzed.[2] Table S3 of the Supporting Information contains information about all of the kinks found by MC-HELAN for comparison for future

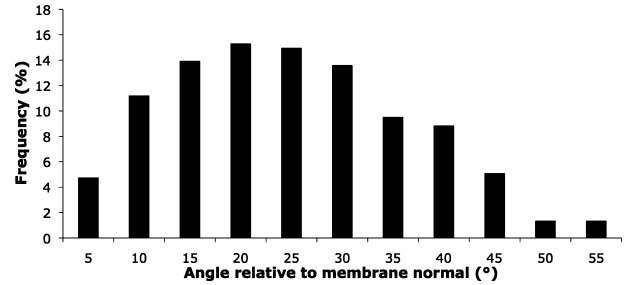

Figure 3. Observed distribution of the frequency of helix angles to the membrane normal for linear transmembrane helices (n = 293) identified by MC-HELAN with respect to the putative membrane normal defined by the TMDET algorithm.[20]

studies. In addition, MC-HELAN analysis files of the full data set are available for download at the MC-HELAN Web site.

The high prevalence of helical kinks in membrane proteins is likely reflective of both the degree of constraint and the uniqueness of the membrane environment. The physical presence of the lipid molecules and the anisotropy of a hydrophobic layer sandwiched on both sides by two potentially dramatically different hydrophilic environments limits the possible conformations of a protein. Due to this, the architectural toolkit of TM helices of membrane proteins may be hypothesized to require a high number of bends, disruptions, and orientational diversity to provide functional uniqueness.[1−4] This is upheld by the fact that even nonkinked TM helices very rarely cross the membrane at an angle perpendicular to the membrane surface. Rather, they are observed to cross the membrane at a range of angles, with 10°−30° relative to the membrane normal defined by TMDET being the most common (Figure 3).

Comparison of MC-HELAN to HELANAL and DSSP. The membrane protein data set was analyzed with both DSSP[18] and HELANAL[13] using the TM regions defined by PDB_TM as the boundaries for analysis (Table 1 and Figure 4). DSSP was included in the comparison as a baseline for detection of helical disruptions. Because both MC-HELAN and HELANAL determine helical axes and classify kinked vs straight helices, we directly compared the output of these programs and correlated this with the disruptions identified by DSSP.

To test for differences in regional vs localized kink detection, pairwise comparison between methods for kinks detected by each method at a given location was performed. For this comparison, additional parsing of output data was required. In the case of MC-HELAN, only converged kinks were considered valid for comparison. In the HELANAL algorithm, helices are classified as linear, curved, and kinked. Because HELANAL does not determine the location of kinks, any residue that had a local helix axis bend angle >20° (the cutoff used by HELANAL to categorize a helix as kinked) was considered kinked. Furthermore, it should be noted that three residues are effectively lost at both the N- and C-termini by HELANAL due to the requirements of its helix axis calculation algorithm. For the DSSP analysis, any residue that did not have helical character was considered kinked. For any of the three methods, an identified kink was only counted if there were at least two helical residues between the kink and the N- or C-terminus of the helix and there were at least three helical residues between it and the

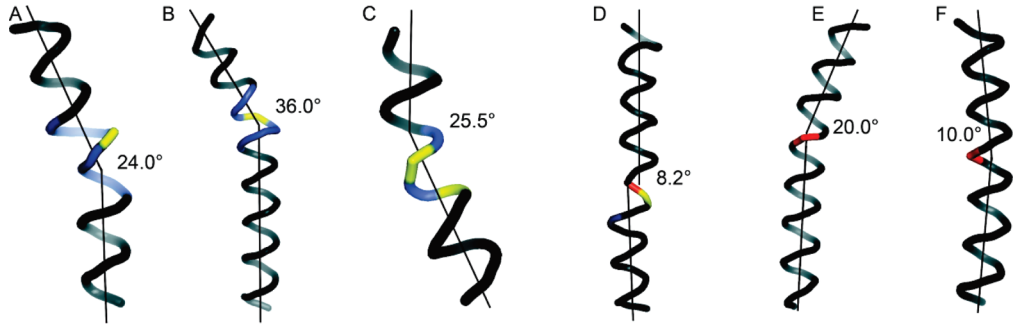

Figure 4. Examples of transmembrane helices classified as kinked. Colors indicate the kinked residues detected by HELANAL[13] or DSSP[18] (blue), by MC-HELAN (red), or by MC-HELAN and either HELANAL or DSSP (yellow). Kinks are shown which were (A, B) detected by all three algorithms, (C, D) detected by HELANAL and DSSP but did not have a converged position over 10 replicates of MC-HELAN analysis, and (E, F) detected by MC-HELAN only. In the illustrated cases, an overlap window of ±4 residues was used to define a kink between two different algorithms as being overlapped. In (C, D) the kinks shown by MC-HELAN were not converged, causing the kink to be "missed" by MC-HELAN.

next nearest kink. Putative kinks located less than three residues apart were annotated as a single kink. Pairwise comparison was performed over windows of ±0−4 residues. Increasing the window size beyond ±1 residue led to only a small increase in the number of overlapping kinks observed (Supporting Information, Figure S1). For consistency, all results are reported with a window size of ±4.

MC-HELAN was able to locate ~87% of kinks detected by either HELANAL or DSSP (Supporting Information, Figure S1), as well as many additional kinks, demonstrating the effectiveness of the algorithm. Notably, although there are approximately the same number of regions of helical disruption located by DSSP and kinks located within TM helices by HELANAL (Table 1), they are largely nonoverlapping, with only ~37% of kinks being detected by both methods. Finally, two kinks were detected by HELANAL with a lack of helical character as defined by DSSP that were detected at a level of <100% convergence by MC-HELAN (e.g., Figure 4).

It is important to note that while MC-HELAN did locate more kinks than HELANAL, many of these kinked helices would have been classified as curved by HELANAL. In fact, out of the 840 helices analyzed, HELANAL only considers 68 helices to be linear, while MC-HELAN identifies 293 converged linear helices. This large discrepancy arises from fundamental differences in the algorithms used to define helix axes, as HELANAL will fit some classes of kinked helices to curves, while MC-HELAN will instead fit a kinked helix to multiple straight axes.

Architecture of Helical Kinks. Examination of the frequency of kinks as a function of depth within the bilayer region (defined here relative to the center of the lipid membrane as determined by TMDET) demonstrates that kinks associated with the membrane surface are not overly prevalent, since bend and disruption depths are highly variable and occur throughout the membrane (Supporting Information, Figure S2). Helical disruptions appear equally probable at any depth, with a surprisingly high frequency of bends located near the center of the putative membrane bilayer region, where the environment should be most uniform, versus near one of the membrane surfaces. When considering bends that occur in the central 70% of the membrane, thus removing the headgroup associated bends, TM helix bends are distributed over a range of kink angles (Figure 5), with 10°−20° being the most frequently observed.

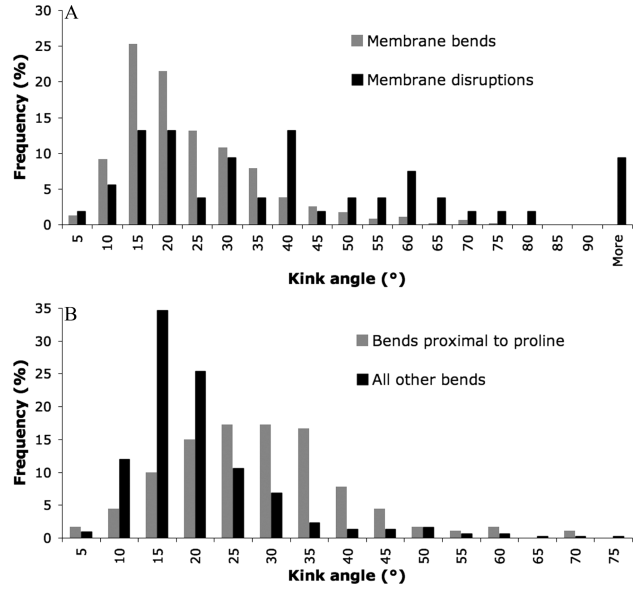

Figure 5. (A) The frequency distribution of transmembrane helix kink angles observed in bends ($n = 471$) and disruptions ($n = 53$) located in the center 70% by depth of the putative membrane. (B) The bend angles in part A subdivided into two distinct sets of bends with or without a proline within a range of ±4 residues.

Disruptions show much greater variability in kink angle (Figure 5A) and a reduced preference for small kink angles in comparison to bends. In total, five disruptions were observed with kink angles >90°. Two of these disruptions occur in the same TM helix, the others in three other proteins. Interestingly, all of these kinks are in proteins that are part of a larger complex, with three of the four proteins being some form of transporter (Supporting Information, Figure S3). Two of these kinks, which occur in cytochrome *b* and the formate transporter FocA, are quite near the surface of the membrane and may be thought of as reentrant loops. The other three kinks occur closer to the membrane center and are found on the pore lining faces of transporters.

A reasonable a priori hypothesis would be that helical kinks in membrane proteins are more likely in situations with decreased constraint due to tertiary structuring. Examination of the likelihood of bends or disruptions as a function of the number of TM spans, however, clearly demonstrates that the frequency of helical kinking is completely independent of the number of TM spans in a membrane protein or oligomer (Figure 1B). In other words, regardless of the topology of

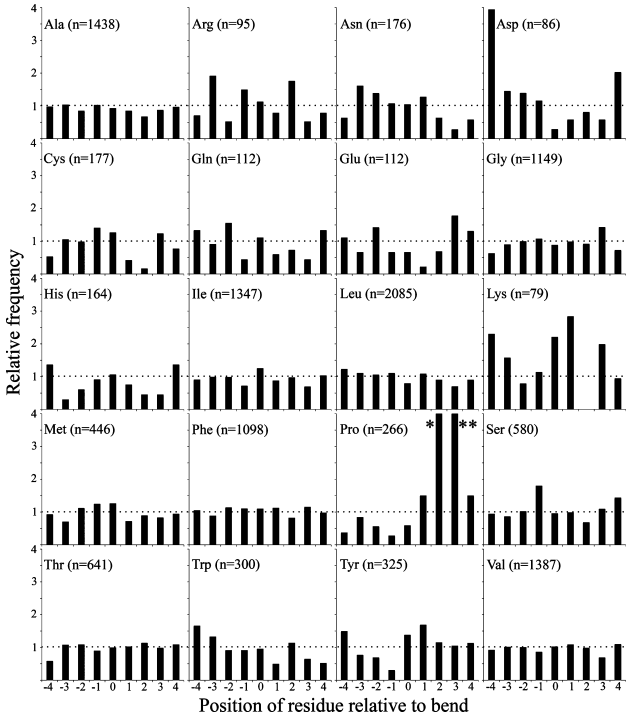

Figure 6. The relative frequencies of each amino acid type at or within four residues of a bend with respect to the frequency of the given residue in the central 70% of the putative membrane by depth in all TM helices. Note that for proline the relative frequency values are beyond the scale of the plot (* = 6.5 and ** = 6.7)

an α-helical membrane protein, ca. two-thirds of its TM helices should be expected to be kinked and/or significantly disrupted.

Amino Acid Prevalence at Helical Kinks. The prevalence of amino acids at and in the region of helical bends and disruptions is a matter of significant literature debate. Prolines are well-recognized as inducing kinks in helices due to the disruption in hydrogen-bonding arising from the lack of a backbone amide proton.[1,2,4,5] In the work of Yohannan et al., this was taken to the extreme of hypothesizing that proline residues are at the evolutionary root of all kinks in TM helices, and this hypothesis was tested using a relatively small set of TM helices.[4] Notably, the recent investigation of a much larger set of membrane proteins by Hall et al.,[1] employing proKink,[12] found statistics somewhat at odds to those of Yohannan et al., with 19% of kinks being due to proline and 16% being associated with vestigial proline. Beyond proline, glycine is also typically assumed to be helix-destabilizing,[28] although it should be noted that this may not necessarily be a valid assumption in nonpolar environments.[26] Since the MC-HELAN algorithm is entirely independent of amino acid sequence, it provides an ideal tool to investigate the induction of helical kinks by particular residues or motifs.

Examining the residues identified as being at the bend and within one helical turn of the 471 converged bends detected in the center 70% of TM helices, with direct comparison to the likelihood of finding the same amino acid in a TM helix in general (Figure 6A and Supporting Information, Table S1), several trends become apparent. First, proline is over 500% more frequent at both the $i + 2$ and $i + 3$ positions C-terminal to a bend (where the bend occurs at position i) than anywhere else in a helix (Figure 6A). However, proline

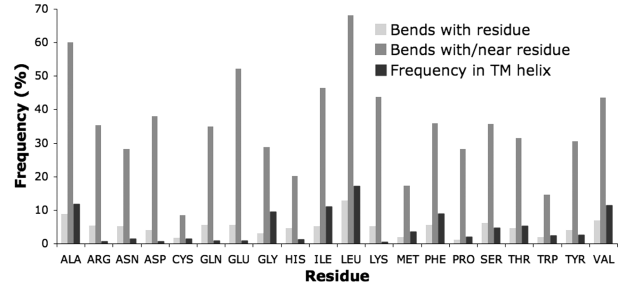

Figure 7. The proportion of TM helix bends ($n = 471$) located within the central 70% by depth of the putative membrane that are found within one helical turn of each residue type. Multiple occurrences of a given residue at a given bend are only counted once, and the relative frequencies of each residue in the central membrane region are shown.

is less likely to be at the position of the bend itself or N-terminal to it than it is to be in the rest of the helix. These results mirror those of Hall et al, who find proline to be concentrated two residues away from a kink, although the trend detected by MC-HELAN is much more apparent. Even though proline is more frequent near a kink, only 33% of TM bends actually have a proline within one helical turn (Figure 7). Also, only 71% TM helix proline residues are found within one turn of a bend or disruption, although many others appear at the termini of helices. This is a similar trend to that found by Hall et al., who found 35% of bends to be due to proline,[1] lower than the ∼50% of helical disruptions to be due to proline estimated by Riek et al. using a smaller set of helices.[2] The structural and functional significance of membrane protein kinks, both with and without proline residues, has been reviewed previously.[29] Interestingly, the presence of a proline residue clearly shifts the distribution of bend angles toward higher values (Figure 5B). However, previous studies indicated that proline-induced kinks in transmembrane helices[2] and globular protein helices[2,24] have a consistent angle of ∼23°−26°, which was not upheld in this study.

Beyond proline, other residues also show increased frequency in the vicinity of TM helical bends. Most notably, aspartic acid is >3 times more frequent at the $i - 4$ position and lysine is over twice as frequent at the $i + 1$ position. Decreased frequencies are also clear for a number of residues (Asn, Asp, Cys, Glu, His, Lys, and Tyr) at specific positions surrounding a bend in a TM helix (Figure 6B). Considering helical propensities in a nonpolar environment,[26] the observed trends of increased and decreased prevalence do not show a straightforward, direct relationship. Disruptions in TM helices show a similar trend in amino acid frequencies to those at bends (Figure 8; not plotted as a function of position due to the relatively low number of residues involved), with aspartic acid, asparagine, glutamic acid, and proline all being significantly more frequent in a membrane disruption. All of these residues, notably, have relatively low helical propensities in a nonpolar environment.[26] Finally, mirroring its relatively good helical propensity in nonpolar environment,[26] glycine is no more likely to be found in the vicinity of a TM helical kink than elsewhere in the helix. Overall, the polar or charged amino acids that show greatly increased frequency near kinks (Figures 7 and 8) may be capable of forming hydrogen bonds to the backbone of the helix.[30] This would break the typical hydrogen-bonding pattern of a α-helix, destabilizing it and allow a kink to form.

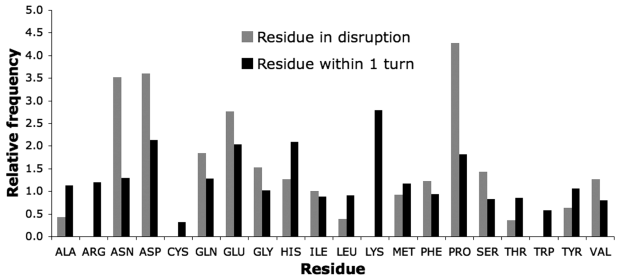

Figure 8. The relative frequencies of residues, with respect to the frequency of the given residue within all examined transmembrane helices, found both in and within ca. one helical turn (i.e., four residues) of transmembrane helix disruptions ($n = 53$) located in the central 70% by depth of the putative membrane.

It is important to note that the charged residues that are in proximity to kinks are themselves quite rare (Figure 6 and Supporting Information, Table S1), meaning that the observed discrepancies in frequencies may in some part arise from the limited data set.

Prediction of Kinks from Sequence Information. Although proline has been advanced as being essential for kink formation,[4] this study and others[1,2,29] find instances of proline residues that do not induce a kink and many kinks in the absence of a proline. With this in mind, we used machine learning algorithms to attempt to predict kinks in TM helices from primary sequence. Through iteration, it was determined that the relative frequency of a residue in the vicinity of a bend (Supporting Information, Table S1) for residues $i + 1$ to $i + 4$ was on its own the best feature to train a SVM classifier (F-score 0.59), versus training sets including side-chain identity or physicochemical properties (Supporting Information, Table S2). Predicting from proline residues alone, in a position-sensitive manner, achieves a maximum F-score of 0.51. Predicting kinks solely on the basis of the presence of proline provides the highest accuracy at 74%, but consequently, the recall of bend prediction suffers greatly (38%) since most experimentally observed kinks do not involve a proline at all. For comparison, using the full range (residues $i - 4$ to $i + 4$) during training and prediction led to slightly worse bend prediction (Supporting Information, Table S1).

Since the classifier based on machine learning is able to accurately predict bends to some extent, primary sequence definitely plays some role in creating kinks. However, even the best F-scores over all training sets and window sizes are relatively low (<0.6), indicating either that sequence information in the vicinity of a kink is only one factor in inducing bends or that the currently available data set is still much too small. This conclusion is also supported by the sequence logo[31] of the 471 converged bends showing <1 bit of information content in the sequences surrounding bends. This is true, despite proline being frequent at the $i + 2$ and $i + 3$ positions relative to a bend, as there are many bends that do not have nearby proline residues. In other words, although many proline residues are near a bend, a much lower proportion of bends are near a proline. Other factors, such as tertiary structure, may therefore play large roles in kink formation. Recall, however, that bitopic and polytopic TM proteins show an equal prevalence of kinks (Figure 1B), implying that tertiary structure is not exclusively responsible.

CONCLUSIONS

At this point in time, the exact factors causing helices to kink in TM helices remain elusive. Sequence determinants—including proline C-terminal to a kink—certainly play some role but are not, based on the currently available set of membrane protein structures, sufficiently well-defined to accurately predict the presence of a kink. What is clear, however, is that the standard topology diagram inference that a helical membrane is made up of a bundle of canonical helices (which has unfortunately pervaded to the level of misleading annotation in PDB entries) is far from the truth. In reality, topologies should be expected in which ca. two-thirds of TM helices are kinked and in which the membrane-traversal angles of the segments are highly variable (e.g., Figure 2C). This level of nonideality in TM helices, arising from the improved sensitivity of the MC-HELAN algorithm to helical kinks and supported by analysis of a data set made up of over double the number of TM helices of any previous work, is ~50% more extreme than would be expected from previous studies.

ACKNOWLEDGMENT

Thanks to Drs. Brian Sykes and Larry Fliegel of the University of Alberta, Edmonton, Canada, for discussions inspiring this study. This work was supported by startup funding from Dalhousie University (to J.K.R.) and by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (RGPIN 342034-2007 to J.K.R. and DGP 298397-2010 to C.B.). D.N.L. is the recipient of a doctoral level Canada Graduate Scholarship from NSERC.

Supporting Information Available: A PDF file with Figures S1–S3 and Tables S1–S3. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hall, S. E.; Roberts, K.; Vaidehi, N. Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *J. Mol. Graphics Modell.* **2009**, *27*, 944–950.
(2) Riek, R. P.; Rigoutsos, I.; Novotny, J.; Graham, R. M. Non-alpha-helical elements modulate polytopic membrane protein architecture. *J. Mol. Biol.* **2001**, *306*, 349–362.
(3) von Heijne, G. Membrane–protein topology. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7*, 909–918.
(4) Yohannan, S.; Faham, S.; Yang, D.; Whitelegge, J. P.; Bowie, J. U. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 959–963.
(5) von Heijne, G. Proline kinks in transmembrane alpha-helices. *J. Mol. Biol.* **1991**, *218*, 499–503.
(6) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(7) Harris, T.; Graber, A. R.; Covarrubias, M. Allosteric modulation of a neuronal K+ channel by 1-alkanols is linked to a key residue in the activation gate. *Am. J. Physiol. Cell Physiol.* **2003**, *285*, C788–96.
(8) Reddy, T.; Ding, J.; Li, X.; Sykes, B. D.; Rainey, J. K.; Fliegel, L. Structural and functional characterization of transmembrane segment IX of the NHE1 isoform of the Na+/H+ exchanger. *J. Biol. Chem.* **2008**, *283*, 22018–22030.
(9) Singh, R.; Hurst, D. P.; Barnett-Norris, J.; Lynch, D. L.; Reggio, P. H.; Guarnieri, F. Activation of the cannabinoid CB1 receptor may involve a W6 48/F3 36 rotamer toggle switch. *J. Pept. Res.* **2002**, *60*, 357–370.

(10) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.
(11) Mezei, M. Simulaid: A simulation facilitator and analysis program. *J. Comput. Chem.* **2010**, *31*, 2658–2668.
(12) Visiers, I.; Braunheim, B. B.; Weinstein, H. Prokink: A protocol for numerical evaluation of helix distortions by proline. *Protein Eng.* **2000**, *13*, 603–606.
(13) Kumar, S.; Bansal, M. Structural and sequence characteristics of long alpha helices in globular proteins. *Biophys. J.* **1996**, *71*, 1574–1586.
(14) Kumar, S.; Bansal, M. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys. J.* **1998**, *75*, 1935–1944.
(15) Kahn, P. C. Defining the axis of a helix. *Comput. Chem.* **1989**, *13*, 185–189.
(16) Tusnady, G. E.; Dosztanyi, Z.; Simon, I. PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* **2005**, *33*, D275−8.
(17) Wang, G.; Dunbrack, R. L. J. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591.
(18) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
(19) Tusnady, G. E.; Dosztanyi, Z.; Simon, I. Transmembrane proteins in the Protein Data Bank: Identification and classification. *Bioinformatics* **2004**, *20*, 2964–2972.
(20) Tusnady, G. E.; Dosztanyi, Z.; Simon, I. TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **2005**, *21*, 1276–1277.
(21) Tusnady, G. E.; Kalmar, L.; Simon, I. TOPDB: Topology data bank of transmembrane proteins. *Nucleic Acids Res.* **2008**, *36*, D234−9.
(22) Lovell, S. C.; Davis, I. W.; Arendall, W. B., 3rd; de Bakker, P. I.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure validation by Calpha geometry: Phi, psi and Cbeta deviation. *Proteins* **2003**, *50*, 437–450.
(23) Mohapatra, P. K.; Khamari, A.; Raval, M. K. A method for structural analysis of alpha-helices of membrane proteins. *J. Mol. Model.* **2004**, *10*, 393–398.
(24) Barlow, D. J.; Thornton, J. M. Helix geometry in proteins. *J. Mol. Biol.* **1988**, *201*, 601–619.
(25) Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **1984**, *53*, 595–623.
(26) Liu, L. P.; Deber, C. M. Uncoupling hydrophobicity and helicity in transmembrane segments. Alpha-helical propensities of the amino acids in non-polar environments. *J. Biol. Chem.* **1998**, *273*, 23645–23648.
(27) Chang C.-C.; Lin C.-J., LIBSVM: A library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm
(28) O'Neil, K. T.; DeGrado, W. F. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **1990**, *250*, 646–651.
(29) Sansom, M. S.; Weinstein, H. Hinges, swivels and switches: The role of prolines in signalling via transmembrane alpha-helices. *Trends Pharmacol. Sci.* **2000**, *21*, 445–451.
(30) Creighton, T. E. *Proteins: Structure and Molecular Properties*; W. H. Freeman and Co.: New York, 1996.
(31) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190.
(32) Lemieux, M. J.; Fischer, S. J.; Cherney, M. M.; Bateman, K. S.; James, M. N. The crystal structure of the rhomboid peptidase from *Haemophilus influenzae* provides insight into intramembrane proteolysis. *Proc. Natl. Acad. Sci U.S.A.* **2007**, *104*, 750–754.

CI100324N