JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

String Kernels and High-Quality Data Set for Improved Prediction of Kinked Helices in α-Helical Membrane Proteins

Benny Kneissl,[*,†] Sabine C. Mueller,[‡,§] Christofer S. Tautermann,[∥] and Andreas Hildebrandt[†]

[†]Johannes Gutenberg-University of Mainz, 55128 Mainz, Germany

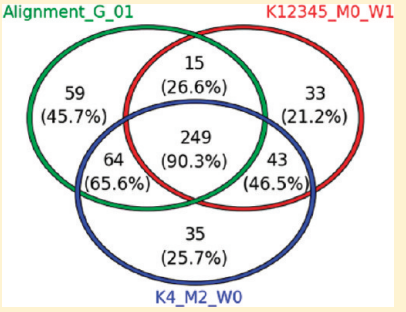
[‡]Intel Visual Computing Institute, Saarland University, 66123 Saarbrücken, Germany

[§]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany

[∥]Lead Identification and Optimization Support, Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany

Supporting Information

ABSTRACT: The reasons for distortions from optimal α-helical geometry are widely unknown, but their influences on structural changes of proteins are significant. Hence, their prediction is a crucial problem in structural bioinformatics. For the particular case of kink prediction, we generated a data set of 132 membrane proteins containing 1014 manually labeled helices and examined the environment of kinks. Our sequence analysis confirms the great relevance of proline and reveals disproportionately high occurrences of glycine and serine at kink positions. The structural analysis shows significantly different solvent accessible surface area mean values for kinked and nonkinked helices. More important, we used this data set to validate string kernels for support vector machines as a new kink prediction method. Applying the new predictor, about 80% of all helices could be correctly predicted as kinked or nonkinked even when focusing on small helical fragments. The results exceed recently reported accuracies of alternative approaches and are a consequence of both the method and the data set.



INTRODUCTION

The prediction of structural elements based on protein sequences is a major task in bioinformatics. Consequently, many algorithms dealing with secondary structure prediction have been developed, starting with very simple methods in the 1970s to more complex ones in the 1990s. Some of the previously reported procedures rely on templates,[1,2] applying machine learning techniques,[3−5] or combining both.[6] For a more detailed review, the interested reader is referred to Pirovana.[7]

However, it becomes increasingly apparent that distortions of perfect geometries in secondary structure elements are very important to create structural diversity from simple building blocks, e.g., in helix bundle membrane proteins.[8] The distortions can be divided into different types, e.g., wide turns or kinks in helices. Especially the latter one, which changes the helical axis noticeably and rather abruptly, yields a significant change of the structure. But even though the knowledge about kinks is crucial for successful modeling of new structures, the number of available algorithms for computational kink prediction is relatively small.

In 2003, a promising sequence pattern descriptors approach was developed by Rigoutsos et al.[9] Based on motifs extracted from 17 proteins, the authors created a search engine to discriminate not only between ideal helices and distortions but also between different distortion types from perfect α-helicity. However, the low prediction accuracy for new sequences not included in their data set (e.g., new GPCRs) as well as their very low false positive rate (0.03%) for nonmembrane spanning helical structures and nonhelical region indicates an overfitting of their descriptors. An alternative approach, due to Yohannan et al., is based on the so-called evolutionary hypothesis for kink generation.[8] Information derived from homologous protein sequences can then be used for kink prediction. In their study, the authors focused on kink patterns in different G-protein-coupled receptor (GPCR) classes and also on eight unrelated membrane structures. For these proteins, they predicted 36 of 39 proline and 14 of 17 nonproline kinks correctly without any false positives. While the former approaches work on the sequence level and are relatively fast, Hall et al. used a complex molecular dynamic simulation setup to reproduce kinks in 405 helices of which 44% have been kinked.[10] Approximately 79% of the 62 proline-induced kinks were predicted correctly with a very high specificity. Interestingly, the prediction accuracy of this structural approach decreased to 58% and 18% for vestigial proline and nonproline induced kinks, respectively. In 2010, Langelaan and co-workers provided a large data set of 842 TM helices. These helices have been automatically annotated using the MC-HELAN algorithm which found a kink in 64% of all cases. Thereafter, they applied support vector machines to predict kinks in a range of ±4 residues.[12] The approach achieved a maximal accuracy of 74% when predicting based on the presence

Received: June 17, 2011
Published: October 08, 2011

of proline. However, the F-scores of the prediction results were never above 0.6. This relatively weak score might indicate one of three things (or a combination thereof): either the sequence is only one (small) factor in producing kinks, the data set used was too error prone due to automated annotation, or the applied kernel functions are insufficient for predicting kinks.

Quite recently, Bowie et al. published an approach called TMKink.[11] They used a neural network with 5 hidden nodes and achieved the best performance for their data set (323 kinked and 567 nonkinked helices) for a window size of 9, resulting in a sensitivity and specificity of 0.7 and 0.89, respectively.

The approach put forth in the present work has been developed independently from the work of Langelaan et al. but is similar in spirit. We also started to collect a large data set, but instead of relying purely on automated kink annotation, we created a manually curated data set. We examined the sequential neighborhood of the annotated kinked residues and computed the solvent accessible surface area (SASA) per residue for kinked and nonkinked helices to quantify the influence of neighboring amino acids and helices.

In addition, to accentuate the need of our manually annotated data set, we used three alternative state-of-the-art methods for automatic kink annotation from the three-dimensional structure and compared their respective qualities. All four data sets were used to train string kernel-based support vector machines for predicting kinks from the protein sequence alone. Furthermore, we compared our performance to TMKink to show the significant better performance of string kernels for support vector machines.

In summary, our main goal in this work is to create a highly accurate data set for kink prediction, compare its quality to data sets automatically derived from the three-dimensional structure, and apply a statistical learning method to predict kinks from the primary sequence. This will yield insights into the state-of-the-art on structurally based kink detection, on the influence of annotation errors on statistical predictors, and on the influence of different sequence features on the likelihood of kink formation. In its current state, the method does not address the problem of determining the exact kink position in a distorted helix, but preliminary work in this direction will do.

MATERIALS AND METHODS

Data Set Generation. To create a reliable data set, we used the database MPtopo[13] to obtain all currently (January 2011) available α-helical transmembrane (TM) proteins, extracted their helices, and removed the ambiguous ones (pairwise sequence identity $\geq$95%) using the PISCES algorithm.[14] For a higher accuracy, the extraction of a single helix was manually curated by visual inspection using BALLView.[15] To select only those helices whose largest part is inside the membrane, we applied the online tool TMDet.[16] The final data set contains 132 proteins including 1014 helices (see Supporting Information). The largest helix has a length of 43 and the smallest of 12 residues, indicating at least 3 turns.

Kink Definition. Given this data set, the next task was to classify these helices as kinked or nonkinked. Previous work has relied on automated kink detection from three-dimensional structures.[12] From our own experiences, however, we expect the automated techniques to fail in some instances, which is also reflected in the fact that there exists no consistent definition of a kink. Moreover, in a canonical helix, the backbone torsion angles φ and ψ are fixed at $-57°$ and $-47°$, but according to the literature, real-world helices are commonly slightly curved toward the solvent, yielding torsions of about $-62°$ and $-41°$ with a higher variance, respectively.[17] Therefore, many helices are hard to classify using simple methods based on large local deviations of torsion angles; in our brief analysis, deviations for kinked, nonkinked, and curved helices look sometimes very similar. In addition, a residue causing merely a wide turn can lead to the same magnitude in angle deviation as one inducing a kink.

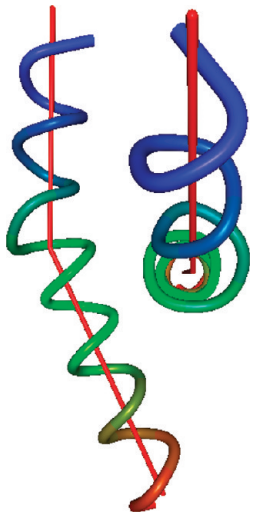


Figure 1. Side and top views of helix 1 of bovine rhodopsin (PDB ID: 1U19) with its computed helical axes using PCA (red lines).

To quantify the quality of automated structural kink detection and to understand the influence of annotation errors on statistical learning schemes, we decided to build a hand-curated data set first. To this end, we used visual inspection using BALLView to determine whether a given helix contains a kink. Exact criteria for the manual annotation of a kink are very hard to define. Each helix has to be examined from several perspectives to identify the residue producing a kink. Indeed, there are some arguable cases where different viewers will report different answers, but this problem is similar to the one of finding an appropriate cutoff in automated methods. Another advantage of manually annotated kinks is the possibility to have a look at local changes and their global effects at the same time. To reduce the bias of only one viewer, the helices were checked by two people. As a rule of thumb, a kink can be an abrupt change of the helical axis or a twisted residue causing for example a shifted axis. Small increases of the helix diameter for one turn, known as wide turns, were not annotated as kinks. While this method is also not entirely free of errors, it is more reliable than any automated criterion we tested. Our annotation leads to manually annotated data set (MDS) with 367 kinked, 461 nonkinked, and 196 curved helices (see Figure 2).

We then compared the data set to two automatically generated ones. First, we adapted principle component analysis (PCA) to compute the helical axis from the backbone atoms.[18] To this end, we split a helix into two segments, where the smallest one consisted of at least five consecutive residues, and computed the axis, i.e., the first principle component, for both parts. The axes had to be slightly adjusted—one end was set to the shortest distance point of the two vectors—to obtain a continuous axis (Figure 1). Afterward, we computed the minimal distance m over

all backbone atoms to the computed axis. This procedure was done for all possible pairs of two segments, and we chose the pair yielding the best fit to the original helix, which is the one with the maximal value m. Finally, we classified a helix as kinked if the angle was larger than a predefined cutoff and defined the residue closest in structure as the kink position. Supposing that our manually curated data set is indeed the most reliable one, we chose the cutoff such that the proportion of kinked and nonkinked helices was similar to the one of MDS. Therefore, an angle value of at least 11° was used, yielding to 364 kinked helices in the data set called PCA data set (PDS). Note that this method can only find global effects of disrupted helices and, therefore, works for at most one kink per helix.

In a second annotation approach, we applied the method HELANAL[19] using the python toolkit MDAnalysis.[20] In this method a helix was defined as kinked, if at least one local bending angle is larger than 20°. Although a few residues (mostly in a row) might fulfill this property, we set the kink to the position of the largest bending angle. In contrast to our PCA method analyzing global effects, this algorithm finds only local ones because just seven residues are used for the computation of an axis. Applying this method we obtained the data set HELANAL with 303 kinked helices.

Based on the choice of annotation algorithms for our manually extracted helices from α-helical membrane proteins, we can compare the performance of our string kernels on a data set created on both manual and automated methods, relying either on global or local effects.

Last but not least, we converted the data set available on the MC-HELAN Web site to our format. We suppose that the MC-HELAN method is currently the best available automated kink annotation method and, hence, an upper bound for our comparison between manually and automatically annotated kink data sets. Note, in contrast to our manually created data set (or to others published so far), the one created by Langelaan and coworkers has an inversed number of kinked and nonkinked helices.

When applying the TMKink webserver on all four data sets, we defined a helix as kinked if at least one kink was predicted, since we focus only on the classification of kinked and canonical helices. The cutoff was adjusted to $t \geq 0.7$ to obtain a higher balanced accuracy and, hence, in order to be able to compare our performance with the best possible results we can achieve using TMKink.

STATISTICAL METHODS

Support vector machines (SVMs) using string kernels have been applied to all data sets using a nested five-fold cross validation setup.[21] The main idea of string kernels is to compare strings by means of the substrings they contain, while these substrings are not assumed to be contiguous.[22] While SVMs have previously been used for kink detection,[12] the use of string kernels as a measure of similarity is novel to this field. In addition, string kernel SVMs have been shown to produce clearly superior results to classical SVMs in fields, such as protein classification,[23−25] prediction of t-cell epitopes,[26] and other structural biology problems. Whereas many different string kernels have been proposed in the literature,[27] we concentrated our approach on the following:

Alignment Kernel. The alignment kernel is strongly motivated by the Needleman−Wunsch alignment score. We used the simple edit distance as scoring function to allow mismatches, insertions, and deletions of amino acids in the input sequences. Hence, the kernel value is the minimum number of operations to transform one sequence into the other.[28]

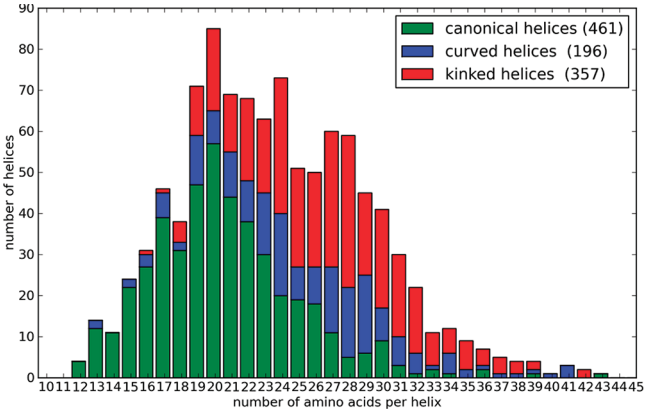


Figure 2. Length distribution of all helices in our data set. The ratio of kinked to canonical helices swaps if the length exceeds 24 amino acids.

K-mer Kernel. In general, a k-mer kernel measures sequence similarity by shared occurrences of fixed-length patterns in the data, allowing for mutations between patterns (mismatch kernel) and a weighting of pattern frequencies (spectrum kernel). In this work, we also allowed a combination of both as well as a combined spectrum kernel of different k-mer sizes.

For the sake of convenience, we introduce the following nomenclature: K for the k-mer length, M for the number of allowed mismatches, and $W1(0)$ to (not) weight the multiple occurrences of the k-mers. Thus, K123_M0_W1 takes all k-mers of size 1−3 where no mismatches are allowed and the occurrences are weighted. In the case of the alignment kernel, we tested different values for the parameter γ, e.g., Alignment_G0_01 represents the alignment kernel with a chosen γ of 0.01.

The SVM setup was realized using libsvm[29] with the precomputed kernel option and was integrated into the Biochemical Algorithms Library BALL.[30] The parameter C of the SVM was determined using five-fold nested cross-validation (from 10^{-6} to 10^{3}).[31]

At this point we want to stress that finding an optimal setup for training the SVM classifiers was out of scope of this work. Indeed, we assume that the prediction accuracy can be further increased with a more sophisticated setup such that our results can be seen as a lower bound for this method. This will be a focus of future work.

Statistical Performance Measures. To obtain an impression of the classifier's performance, we determined the so-called confusion matrix relating true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).[32] Based on this matrix, several performance measures, such as specificity, sensitivity, and accuracy, can be easily calculated. Due to the imbalance in our data sets, we will specify the balanced accuracy and the F-score.

The significance value for the solvent accessible surface area analysis was calculated applying the Welch two sample t test using R.[33] In this case, the null hypothesis for the two-sample t test was $\mu_1 = \mu_2$. Hence, the smaller the computed value, the more convincing the rejection of the null hypothesis.

Table 1. Amino Acid Distribution[a]

AA	Protein	Helix	Kink	AA	Protein	Helix	Kink
A	9.5	11.6	9.2	M	2.9	3.6	4.6
C	1.1	1.3	1.1	N	3.2	1.7	1.1
D	3.3	1.1	0	P	4.6	1.9	6.2
E	3.9	1.8	1.5	Q	2.7	1.8	2.4
F	6.4	8.1	8.1	R	3.9	2.4	1.1
G	8.5	8.2	10.5	S	5.7	4.8	6.5
H	2	1.6	1.1	T	5.5	5.5	4.3
I	7.6	10.3	7.8	V	8.3	10.4	11.6
K	3.4	1.8	1.9	W	2.4	2.8	2.4
L	11.6	15.8	15.1	Y	3.5	3.5	3.5

[a] Percentage distribution of all amino acids in the whole protein, the extracted helices and at the manually annotated kink position of MDS. The two interesting groups are highlighted in orange (low occurrences at kink position) and yellow (high occurrences at kink position).

RESULTS AND DISCUSSION

Data Set Analysis. As a first step of our analysis, we studied the manually annotated data set in detail. Figure 2 shows the length distribution of all helices: 461 helices were defined as nonkinked, with a maximal frequency at the length of 20 amino acids, equaling the number of residues needed to completely cross a typical cell membrane. In 357 cases, we found a residue obviously causing a change of the helical axis, thus introducing a kink. These helices are more uniformly distributed between 19 and 32 residues and, as expected, tend to be longer than canonical helices (the kink allows to fit a longer helix into the membrane).

Altogether, there are 196 helices featuring a distortion that could not be exactly assigned to one residue due to a curved structure. We removed these helices in the following to reduce the noise such that finally 357 kinked and 461 nonkinked helices are included in our manually created data set (MDS).

To obtain information about the data set and the amino acids it contains, we calculated their percentage distribution in the whole protein (see Supporting Information) and the helical regions as well as the determined kink positions (see Table 1).

Two interesting groups of amino acids are immediately apparent: the first and most important one for our work contains glycine, serine, and proline (marked yellow). Their percentage occurrence value is smallest when considering the helical region. The special role of proline and glycine is well established from many other studies,[34,35] and serine is also known as a potential helix breaker.[10] Thus, statistical analysis of the amino acid distribution at kink positions in the data set confirms our manual annotation of the helices.

Another interestingly distributed group, marked in orange, contains both acidic amino acids (aspartic and glutamic acids) and two basic ones (histidine and arginine) as well as asparagine. The members of this group appear only very rarely at kink positions (in total 4.8%), even though their general occurrences in proteins and helices are much higher (in total 16.3% and 8.6%, respectively). In addition, lysine, the third basic amino acid, also appears unfrequently at a kink position (1.9%). The rare occurrence of these amino acids at kink positions is due their sequence position in the helix. They appear mostly at the end of a helix (because of the membrane environment), where only a few kinks are determined (altogether 43 kinks in the first and last 30% of a helix). Hence, these findings do not allow to draw simple conclusions on the relevance of this group for kink formation.

Sequential and Structural Kink Environment. Besides the kink position itself, the kink environment may play an important role. Therefore, similar to Langelaan et al.,[12] we computed the occurrence probability for each amino acid around an identified kink (Figure 3). There are four amino acids with an under-representation of at least 5% in this region: arginine, glutamine, glutamic acid, and lysine, which supports the results drawn from Table 1. Over 50% of all prolines in all helical sequences occur in a range of ±5 residues around a kink, in particular, between 0 and +5 amino acids next to a kink, which corroborates its great relevance. For this reason, we defined all kinks as proline induced if a proline occurs in this range. This is slightly different to the annotation results of the MC-HELAN algorithm,[12] where especially positions 2 and 3 after an identified kink are over-represented by proline. Glycine and serine do not reveal such a distribution, because their ratio between helix and kink occurrences is much lower compared to proline. Some amino acids were not found at specific positions, e.g., neither was a histidine present one or two residues after a kink nor an aspartic acid found at residue position −1, 0, or 4. Another aspect is the small increase of aspartic acid one turn before a kink, which fits very well to the data of Langelaan.

In our last data set analysis, we compared the amino acid composition of kinked and canonical helices at specific positions. To this end, we superimposed the kink position of the kinked helices on the center of nonkinked helices, computed the occurrences of each amino acid in a range of ±4 residues, and calculated the ratio of their frequencies. As illustrated in Figure 4, several amino acids are over- or under-represented at different positions. Here, we want to mention only three: the five times over-represented tryptophan (it has two rings), the several times over-represented arginine and lysine (long and charged side chain), and the charged aspartic acid, which is over-represented at the first three positions but under-represented at positions four, six, and nine. To determine the exact kink position in later projects or with regard to a better understanding of kink formation, this analysis might be very helpful.

To explore the influence of neighboring helices we calculated the SASA per residue for each helix (see Figure 5). We applied a two-tailed t test to assess whether the SASA means are statistically different for kinked vs ideal helices, for kinked vs curved helices, and for ideal vs curved helices. For the first comparison (kinked against the nonkinked helices), we obtained a p-value of 0.0031, while the p-value for the second case (kinked against curved) was 0.0511. The first value in particular indicates a large difference in the environment of the compared helix types. The environment for nonkinked and curved helices, however, seems to be very similar (p-value: 0.8354). Due to the significantly smaller SASA mean value for kinked helices, we conclude that the resulting larger amount of potential tertiary interactions can help to enforce and stabilize these kinks.

Evaluation of the Automated Detection Methods. Figure 6 shows a Venn diagram for the kinked helices. Only 59% of all manually annotated kinked helices have been identified by both automated methods, but in total, only 29 kinks have been manually, but not automatically, detected, and 125 kinks were annotated automatically by either PCA or HELANAL, but only 4

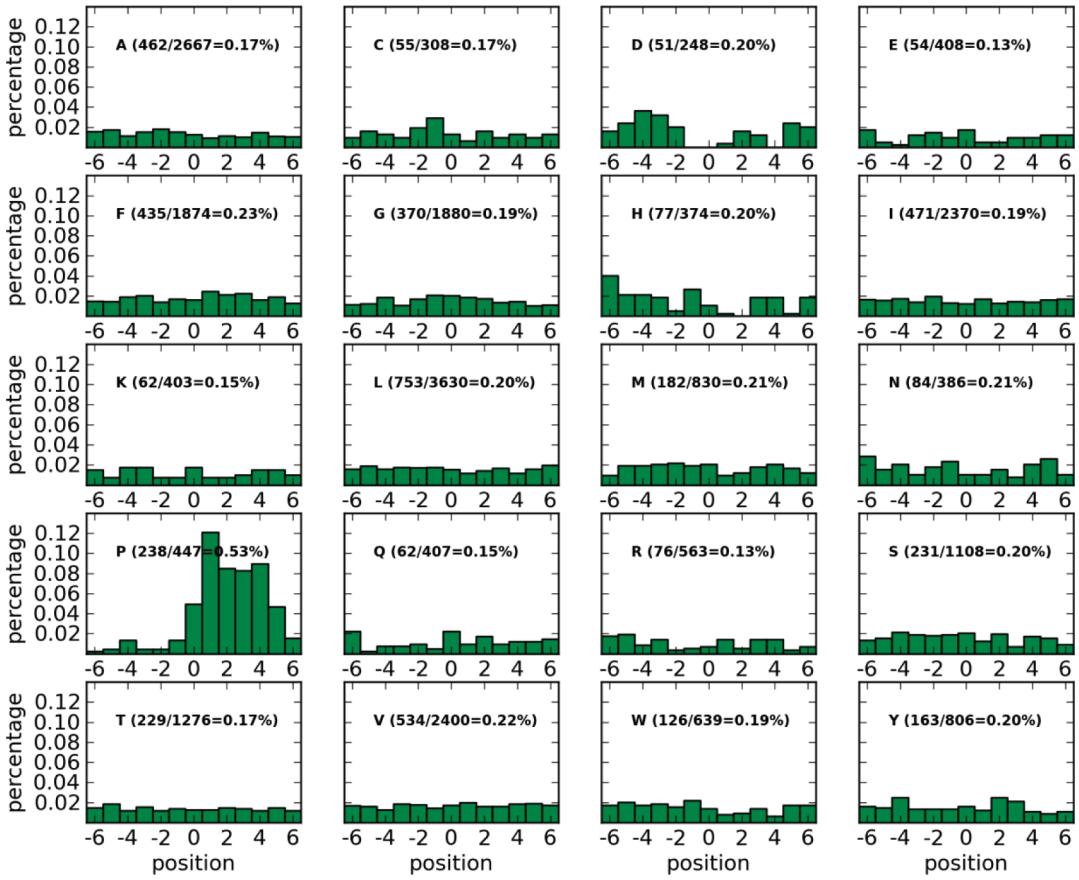


Figure 3. Neighborhood of kinks. For every amino acid type, we computed the probability to be found at a specific position (± 6 residues) next to a kink (position 0). We divided the number of each amino acid found in this region by the number found in the complete helix. The expected value is 20.5%.

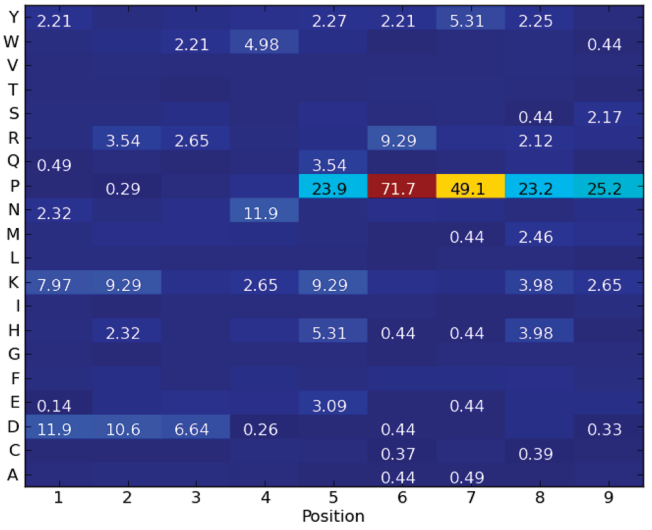


Figure 4. Amino acid composition: Ratio of frequencies in kinked and canonical helices is shown. Numbers are given if a residue is at least two-fold over-represented (>2.0) or under-represented (<0.5) in kinked helices compared to nonkinked helices. Position 5 refers to the annotated kink and the center of the helix in kinked and canonical helices, respectively.

of these were detected by both methods. As mentioned above, these methods focus on different aspects of structural changes in kinked helices and are insufficient to create a reliable data set on their own.

Table 2 shows if the manually and automated methods defined a helix as kinked, these positions do not differ much. Over 90% are in between 1 helical turn (≤4 residues). This means that automated methods work well, but our further results will demonstrate that the exact kink annotation is necessary for predicting kinked helices with a high accuracy.

Application of SVMs. Table 3 gives an overview of the results for the nine best string kernels for SVMs and the neural network of TMKink. Annotating the extracted helices in our data set manually yields a much higher balanced accuracy and *F*-score compared to both automated methods, HELANAL and PCA. Applying our method to the data set from the MC-HELAN Web site, we achieve the same sensitivity but a much lower specificity and, hence, a lower balanced accuracy—even compared to HELANAL. The higher *F*-score is mainly due to the larger number of kinked helices in their data set. This supports our assumption that automated kink detection from structural information is still too error prone and noisy to be useful for training statistical classifiers.

Interestingly, the *F*-score we achieve significantly exceeds the one reported by Langelaan et al. and demonstrates the usefulness of string kernel-based SVMs, even more with respect to the neural networks results of TMKink, where the balanced accuracy and *F*-score decreases dramatically. But again, MDS performs best, while the data set of Langelaan and co-workers has the highest *F*-score.

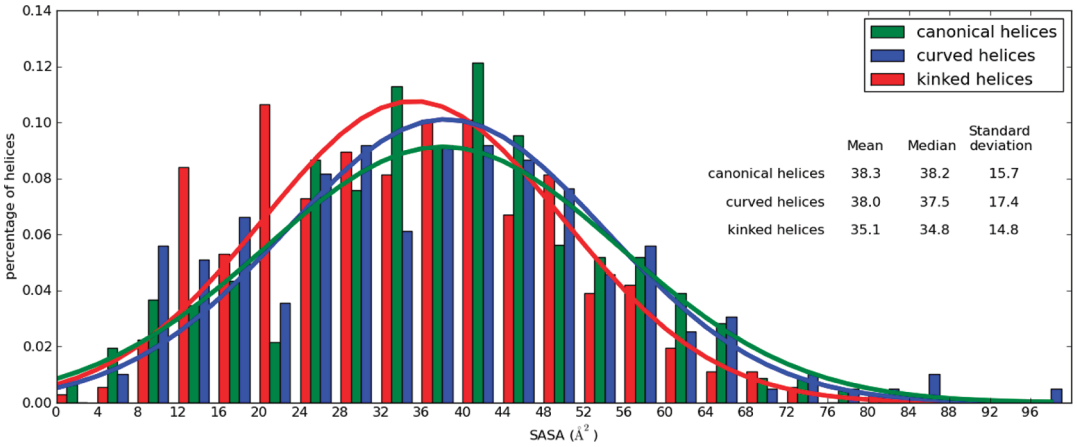


Figure 5. Histogram of the SASA for all identified helices. In addition, the corresponding probability distributions are shown.

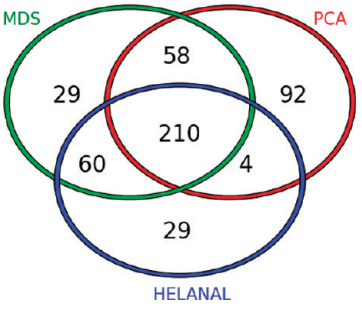


Figure 6. Venn diagram for kinked helices of our 3 data sets: 210 helices are defined as kinked by all the methods, and 29 have been identified by the two viewers as kinked but neither by PCA nor by the HELANAL method, while 125 sequences are identified by at least 1 automated method, they have only 4 of these in common.

The three alignment kernels (with different parameter γ) are ranked in the top four. This seems to indicate that the order of the amino acids might play a decisive role, information that is lost when using k-mer kernels. Whereas a combination of mismatch and spectrum kernel slightly decreases the performance in most cases, we were able to improve the results remarkably using a combined spectrum kernel with different k-mer sizes. For example, K12345_M0_W1 (76%) is better than the non-weighted version K12345_M0_W0 (73%) and compared to the best single spectrum kernel K3_M0_W1 (62%). Furthermore, mismatches yield better results than weighting the occurrences: especially using k-mers of size 4 and allowing 2 mismatches without weighting the occurrences (K4_M2_W0) seems to be a good choice for kink prediction.

Figure 7a,b gives a Venn diagram of the prediction results for the best alignment, mismatch, and spectrum kernel. About 90% of the helices predicted as kinked by all three string kernels are indeed kinked. For canonical helices, this number is even higher. Compared to the number of kinked and nonkinked helices in MDS, nonkinked helices are predicted with a balanced accuracy of 63.0% (64.4%) by all three kernel methods. Taking the majority vote of these three kernels, we can predict nonkinked helices with a balanced accuracy of 82.1% (82.6%). These results strongly indicate that string kernel-based SVMs yield a very stable and adequate method for kink prediction.

Table 2. Comparison of the Annotated Kink Position[a]

distance, AA	HELANAL (270 helices)	PDS (268 helices)
0	84 (31%)	51 (19%)
1	90 (33%)	106 (40%)
2	48 (18%)	56 (21%)
3	15 (6%)	20 (8%)
4	11 (4%)	17 (6%)
>4	22 (8%)	18 (6%)

[a] Absolute distance in amino acids (AA) to our manually created data set for all helices labeled automatically as kinked. The most ($\geq$90%) are in between 1 helical turn.

Another question of interest in kink prediction is the sensitivity and specificity for proline and nonproline kinks. As mentioned above, we defined a proline-induced kink if proline occurs in the range of $\pm$5 residues away from a specified kink position. Table 4 shows that the alignment kernel is significantly better than the k-mer kernels in predicting proline kinks due to the very high specificity. We suppose the exact position of proline to play an important role, which is confirmed by the neighborhood analysis (Figure 3), where proline occurs mainly 0−5 residues after a kink. Nonproline kinks have been detected with a high and balanced sensitivity and specificity. These results are very promising, although we are not focusing on the exact kink position in this work. In particular, our approach reveals SVMs to be capable to find other general features besides the occurrence of proline in the sequence.

Kink Neighborhood. Today, the main influences for kink formation are still unknown. In fact, it is even unclear whether kinks are a very global (influences over different helices and nonhelical parts), a mostly local (influences only inside the same helix), or a very localized (influences only from a few residues around the kink) effect. To decide whether a few residues around the kink are enough to classify into kinked and nonkinked, we created further data sets containing only the so-called core subsequence (CDSX) of each helix, where X denotes the length of the subsequences. In cases of kinked helices, the kinked residue corresponds to the center of the considered subsequence. For nonkinked helices, we decided to set the center of the complete helix to the center of the subsequence, because this part is mostly in the membrane center and usually more important

Table 3. Prediction Results for the Four Data Sets[a]

Ranking	MDS	HELANAL	PDS	MC-HELAN
1	A_G0_1 0.820 (0.801)	K4_M2_W0 0.766 (0.707)	K4_M2_W0 0.630 (0.635)	K4_M2_W0 0.742 (0.811)
2	K4_M2_W0 0.811 (0.791)	A_G0_05 0.759 (0.699)	A_G0_1 0.616 (0.604)	A_G0_1 0.719 (0.802)
3	A_G0_05 0.808 (0.788)	A_G0_1 0.755 (0.693)	K4_M2_W1 0.606 (0.619)	K3_M1_W0 0.710 (0.779)
4	A_G0_01 0.786 (0.770)	A_G0_01 0.751 (0.692)	K5_M2_W1 0.605 (0.548)	K3_M1_W1 0.690 (0.765)
5	K4_M2_W1 0.778 (0.748)	K1234_M0_W1 0.742 (0.679)	K123_M0_W0 0.604 (0.606)	K1234_M0_W1 0.690 (0.786)
6	K12345_M0_W1 0.767 (0.734)	K4_M2_W1 0.737 (0.669)	K3_M1_W0 0.601 (0.611)	K4_M2_W1 0.688 (0.770)
7	K1234_M0_W1 0.765 (0.731)	K12345_M0_W1 0.734 (0.668)	K1234_M0_W1 0.600 (0.542)	K12345_M0_W0 0.686 (0.792)
8	K3_M1_W0 0.752 (0.730)	K3_M1_W0 0.733 (0.667)	K12345_M0_W0 0.600 (0.578)	K1234_M0_W0 0.679 (0.779)
9	K5_M2_W1 0.748 (0.703)	K123_M0_W1 0.733 (0.667)	K5_M2_W0 0.597 (0.529)	A_G0_05 0.673 (0.753)
-	TMKink 0.714 (0.707)	TMKink 0.691 (0.641)	TMKink 0.618 (0.621)	TMKink 0.630 (0.724)

[a] Kernel name as well as the corresponding balanced accuracy and *F*-score (in brackets) are given. The colors are due to the balanced prediction accuracy from yellow (low) to green (high). The last row shows the results of the TMKink method.

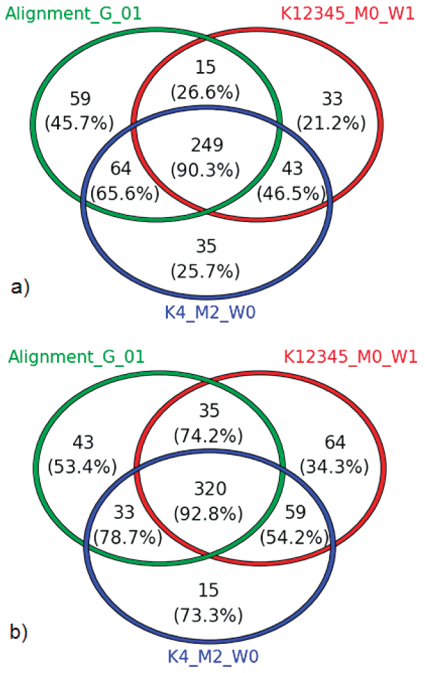


Figure 7. Venn diagram of the prediction results for kinked (a) and nonkinked (b) helices. The number in brackets is the percentage rate of correctly identified kinked (nonkinked) helices.

and reliable for the stability of the helical structure than regions at the end of a helix. Compared to our first results, we obtain a lower balanced accuracy and *F*-score but in many cases still over 75% with a maximal *F*-score of 0.74 (see Table 5). These results indicate that a large part, but not all, of the effects behind kink formation seems to be very localized. In addition, we classified each subsequence by TMKink and got clearly lower prediction results. Because Bowie and co-workers predicted kinks in a range of ±4 residues and also Langelaan et al. reported their results with this window size, we suppose that their results are comparable to the one of CDS9.

It is noticeable that only in one case, the alignment kernel achieves the 10th best result. Hence, we suppose this kernel to be influenced by the length of a sequence. CDS9−CDS13 work very similar, which means that we have to consider 4−6 amino acids to the left and to the right of a specified kink. This observation correlates with the kink environment plot in Figure 3. Kernel K4_M2_W0 is again a good choice. The resulting confusion matrix for CDS11 for proline and nonproline kinks is illustrated in Table 6. While proline kinks using the mismatch string kernel K4_M2_W0 are predicted with a slightly higher balanced accuracy, we found a decrease in all other cases. The reason is, in particular, the very low sensitivity for nonproline kinks indicating that nonproline kinks are not local ones and that a larger sequence range has to be taken into account.

To confirm this supposition, we trained a statistical model on a data set containing only the nonproline sequences of MDS and CDS11 (see Table 7). These data sets are called NP_MDS and NP_CDS11, respectively. Applying all kernels, we obtained a significantly better performance including the whole helical sequence. In this case, more than 55% of nonproline kinks were predicted correctly. Focusing on just 5 neighboring residues this value decreased to just below 40%. The lower sensitivity and higher specificity compared to the usage of the complete data sets might be a result of the extremely unbalanced data set.

In addition, we compared the helix length of proline and nonproline kinked helices, finding an average length of 26.4 and 27.1, which results in a nonsignificant difference (*p*-value: 0.15). Hence, the helix length itself does not influence the result, implying that for nonproline kinks, the whole sequence seems to be very important for a better prediction.

Preliminary Investigation of Detecting the Exact Kink Position. In principle, SVMs can also be used for detecting the exact kink position in the helical sequence in addition to the binary kink/nonkink classification. The last results show that a SVM is able to predict also smaller sequences with a high balanced accuracy correctly. Focusing on the results of CDS9, we can predict kinks in a range of ±4 residues with a balanced accuracy of more than 75%, which is higher than reported accuracies in former studies. However, in further studies we want to be even more precise. One idea is to use a window of a specific size to create all subsequences of a helix while iterating over it. The subsequences will be labeled as kinked if and only if it contains the kinked residue. After applying SVMs to this modified data set, we can retranslate the prediction result to the whole helix. Our first tests indicated some signals, but there is still a large degree of noise due to the very short sequences, which have in some cases different labels in various helices. But combined with the information from the data set analysis, we are optimistic that this is a promising approach to detect the most probable or even the exact kink position.

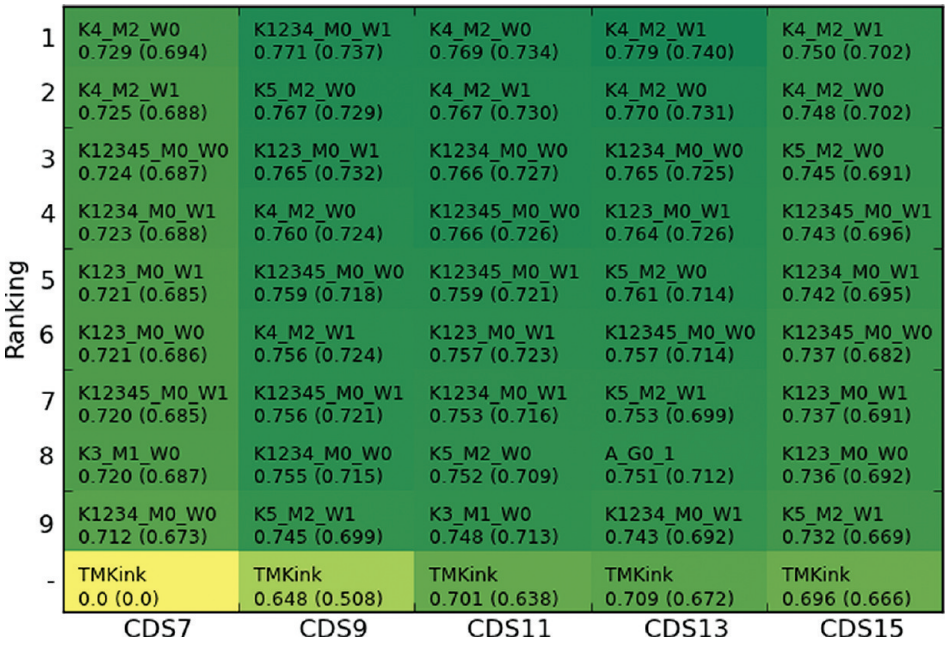
CONCLUSION

Our work gives new insights on kinks in α-helical membrane proteins. First, our data set analysis affirms the great importance of proline for distortions but reveals also a disproportionately high occurrence of glycine and serine. Moreover, the data set analysis can help to assess and improve homology

Table 4. Confusion Matrix of MDS for Proline and Nonproline Kinks

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 proline	192	14	19	18	93.2	51.4	72.3
AlignmentG0_1 proline	181	25	33	4	87.9	89.2	88.5
K4_M2_W0 nonproline	104	47	347	77	68.9	81.8	75.4
AlignmentG0_1 nonproline	117	34	339	85	77.5	80	78.7

Table 5. Prediction Results for CDSX[a]



[a] Kernel name as well as the balanced accuracy (and *F*-score) are given. The last row represents the results of the TMKink method.

Table 6. Confusion Matrix of CDS11 for Proline and Nonproline Kinks

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 proline	188	9	23	20	95.4	53.5	74.5
AlignmentG0_1 proline	182	15	31	12	92.4	72.1	82.2
K4_M2_W0 nonproline	60	80	344	70	42.9	83.1	63.0
AlignmentG0_1 nonproline	61	79	306	108	43.6	73.9	58.7

Table 7. Confusion Matrix of NP_MDS and NP_CDS11

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 NP_MDS	67	54	316	41	55.4	88.5	71.9
AlignmentG0_1 NP_MDS	73	48	328	29	60.3	91.9	76.1
K4_M2_W0 NP_CDS11	45	68	315	38	39.8	89.2	64.5
AlignmentG0_1 NP_CDS11	43	70	281	72	38.1	79.6	58.8

models by incorporating the gained information. Some of these results are already confirmed by the findings of the related work of Langelaan et al, e.g., the high occurrence of proline a few positions after a kink.

Furthermore, we have developed and validated a new kink prediction method using string kernels for SVM and our manually annotated data set. The very high consensus of all applied string kernels demonstrates that there is much information about kinks coded in the amino acid sequence of a helix. Most importantly, using string kernels allows us to detect also nonproline kinks with a high accuracy, where the most of the previously published methods more or less failed. The basis of these considerably improved results is our manually created data set and the usage of string kernels, which is demonstrated in the comparison between both manually and automatically annotated data sets as well as different methods. Nevertheless, we agree with Langelaan and

co-workers that the helical sequence is only one factor and that tertiary interactions or the spatial environment (membrane) cannot be neglected, which is confirmed by the SASA analysis of the different helix types.

Finally, we provide a large data set for further studies. This, for example, can be used to develop and evaluate future algorithms for determining kinks from three-dimensional structures automatically. The complete source code and data sets will be published in the next release of the biochemical algorithm library BALL.

■ ASSOCIATED CONTENT

Supporting Information. Table includes all helices with manually annotated kinked residues. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: b.kneissl@uni-mainz.de.

■ ACKNOWLEDGMENT

B.K. and S.M. thank Boehringer Ingelheim Pharma GmbH & Co. KG for financial support.

■ REFERENCES

(1) Yi, T. M.; Lander, E. S. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* **1993**, *232*, 1117–1129.
(2) Salamov, A. A.; Solovyev, V. V. Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **1997**, *268*, 31–36.
(3) Qian, N.; Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **1988**, *202*, 865–884.
(4) Tusnády, G. E.; Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **2001**, *17*, 849–850.
(5) Kim, H.; Park, H. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.* **2003**, *16*, 553–560.
(6) Bondugula, R.; Xu, D. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins* **2007**, *66*, 664–670.
(7) Pirovano, W.; Heringa, J. Protein secondary structure prediction. *Meth. Mol. Biol.* **2010**, *609*, 327–348.
(8) 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.
(9) Rigoutsos, I.; Riek, P.; Graham, R. M.; Novotny, J. Structural details (kinks and non-alpha conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. *Nucleic Acids Res.* **2003**, *31*, 4625–4631.
(10) Hall, S. E.; Roberts, K.; Vaidehi, N. Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *J. Mol. Graph. Modell.* **2009**, *27*, 944–950.
(11) Meruelo, A. D.; Samish, I.; Bowie, J. U. TMKink: A method to predict transmembrane helix kinks. *Protein Sci.* **2011**, *20*, 1256–1264.
(12) Langelaan, D. N.; Wieczorek, M.; Blouin, C.; Rainey, J. K. Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J. Chem. Inf. Model.* **2010**, *50*, 2213–2220.
(13) Jayasinghe, S.; Hristova, K.; White, S. H. MPtopo: A database of membrane protein topology. *Protein Sci.* **2001**, *10*, 455–458.
(14) Wang, G.; Dunbrack, R. L. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **2005**, *33*, 94–98.
(15) Moll, A.; Hildebrandt, A.; Lenhof, H. P.; Kohlbacher, O. BALLView: An object-oriented molecular visualization and modeling framework. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 791–800.
(16) Tusnády, G. E.; Dosztányi, Z.; Simon, I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **2005**, *21*, 1276–1277.
(17) 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.
(18) Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2* (6), 559–572.
(19) Bansal, M.; Kumar, S.; Velavan, R. HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.* **2000**, *17*, 811–819.
(20) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *10*, 2319–2327.
(21) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
(22) Zhang, Y.; DeVries, M. E.; Skolnick, J. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput. Biol.* **2006**, *2*, 88–99.
(23) Gubbi, J.; Shilton, A.; Parker, M.; Palaniswami, M. Protein topology classification using two-stage support vector machines. *Genome Inf.* **2006**, *17*, 259–269.
(24) Leslie, C.; Eskin, E.; Noble, W. S. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.* **2002**, 564–575.
(25) Leslie, C. S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004**, *20*, 467–476.
(26) Zhao, Y.; Pinilla, C.; Valmori, D.; Martin, R.; Simon, R. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics* **2003**, *19*, 1978–1984.
(27) Shi, F.; Huang, J. Prediction of T-cell Epitopes Using Support Vector Machine and Similarity Kernel. *CIS* **2005**, *1*, 604–608.
(28) Saigo, H.; Vert, J. P.; Ueda, N.; Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics* **2004**, *20*, 1682–1689.
(29) Chang, C.-c.; Lin, C.-J. LIBSVM: a Library for Support Vector Machines. *ACM Trans. Sys. Tech.* **2011**, *27*, 1–27.
(30) Hildebrandt, A.; Dehof, A. K.; Rurainski, A.; Bertsch, A.; Schumann, M.; Toussaint, N. C.; Moll, A.; Stöckel, D.; Nickels, S.; Mueller, S. C.; Lenhof, H. P.; Kohlbacher, O. BALL—biochemical algorithms library 1.3. *BMC Bioinformatics* **2010**, *11*, 531.
(31) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2003.
(32) Kohavi, R.; Provost, F. Glossary of Terms. In *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, Kluwer Academic Publishers: New York, 1998; 30, 271—274.
(33) Gentleman, R.; Ihaka, R. et al. *R: A Language and Environment for Statistical Computing*; Institute for Statistics and Mathematics, Vienna University of Economics and Business: Vienna, Austria, 2011; http://www.R-project.org, (accessed 2008).
(34) von Heijne, G. Proline kinks in transmembrane alpha-helices. *J. Mol. Biol.* **1991**, *218*, 499–503.
(35) Geetha, V. Distortions in protein helices. *Int. J. Biol. Macromol.* **1996**, *19*, 81–89.