Towards 3D Modeling of Interacting TM Helix Pairs Based on Classification of Helix Pair Sequence

Witold Dyrka[1], Jean-Christophe Nebel[2], and Malgorzata Kotulska[1]

[1] Institute of Biomedical Engineering and Instrumentation, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
[2] Faculty of Computing, Information Systems and Mathematics, Kingston University, Penhryn Road, Kingston-upon-Thames, KT1 2EE, United Kingdom
{witold.dyrka,malgorzata.kotulska}@pwr.wroc.pl, j.nebel@kingston.ac.uk

Abstract. Spatial structures of transmembrane proteins are difficult to obtain either experimentally or by computational methods. Recognition of helix-helix contacts conformations, which provide structural skeleton of many transmembrane proteins, is essential in the modeling. Majority of helix-helix interactions in transmembrane proteins can be accurately clustered into a few classes on the basis of their 3D shape. We propose a Stochastic Context Free Grammars framework, combined with evolutionary algorithm, to represent sequence level features of these classes. The descriptors were tested using independent test sets and typically achieved the areas under ROC curves 0.60-0.70; some reached 0.77.

Keywords: stochastic context-free grammar, evolutionary algorithm, helix-helix interaction, transmembrane protein.

1 Introduction

It has been estimated that around 30% of proteins in human body are transmembrane (TM) proteins [1]. Moreover, since they are more accessible to drugs than intracellular proteins, they are prime targets for drug design. Unfortunately, the specific environment of cell membranes, their large size and dynamic behavior (e.g. ion channels) make them very difficult objects for current experimental techniques in structural biology: fewer than 2% of currently known protein structures are from TM proteins [2]. Thus, the lack of experimental structures cannot be compensated by template-based modeling, i.e. homology and threading, which would require availability of a large dataset of structures. The alternative is use of ab initio methods, which build protein 3D models directly from their sequences. However, these approaches have only been successful for small proteins up to 200 amino acids [3], mainly because computational power limits the size of the conformational phase space that can be searched. Moreover, the energy function is not accurate enough to guarantee the minimum at the native state [4]. Therefore, for larger proteins, such as protein channels, which typically contain

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 62–73, 2010.

1000's of amino acids, limitations of ab initio methods can only be overcome by integrating additional knowledge in the modeling process.

Contact maps have been shown to be promising constraints. It was estimated that as few as one contact in every eight residues would be sufficient to find the correct fold of a single domain protein [5]. Moreover, even the prediction of a few contacts is useful to constrain conformational searches in ab initio prediction [6]. Recent study also suggests that some contacts are structurally more significant than others [7]. Consequently, the prediction of intramolecular contacts has become an active field of research. According to [4] homologous template approaches achieve the highest accuracy (up to 50%). However, they are not suitable for TM proteins, since very few templates are available. As correlated mutations methods have the lowest accuracy (around 20%), machine learning methods seem to be the most appropriate.

Over 80% of known TM structures are classified as alpha-helical [2]. In these proteins, molecular contacts between helices are crucial as they provide a structural skeleton. A stable interaction between two helices requires that several residues from each helix are involved in the helix-helix contact. We call this structure a helix-helix (H-H) interface and define it more precisely later in the paper. A recent study by Walters and DeGrado [8] on helix packing motifs has revealed that 90% of known configurations of H-H interactions in TM proteins can be accurately represented using only a set of 8 3D templates (Fig. 2,3 in [8]). In their research, helix pairs were clustered according to the 3D similarity (RMSD $\leq$ 1.5 A) of their fragments involved in the H-H contact. Their study also highlighted position-specific sequence propensities of amino-acids and the occurrence of the well known [GAS]-X-X-X-[GAS] motif [9].

The problem of H-H interaction prediction was addressed in [3] by creating sequence profiles from a library of helix pairs whose spatial configurations were known. In their method a helix pair in the query was compared to helix pairs in the library by calculating profile-profile scores between the pairs. While the overall accuracy of helix packing prediction was rather low, it was sufficient to constrain ab initio prediction of TM protein structures. Significantly, this approach does not model interactions between contacting residues from the two helices since this would require a more complex model than sequence profiles. Waldispuehl and Steyaert [10] proposed a multi-tape S-attributed grammar to represent helix bundles in TM proteins. In their model, a single pair of helices is described by a set of grammar rules of a non-probabilistic context-free language. At each stage of processing of a sequence, a value or attribute that reflects folding cost is calculated. The authors report that the predictive power gained from the ability to represent long range dependencies between contact residues allowed their method to outperform the best TM helix prediction software.

There are two main approaches for learning grammar rules: Maximum A Posteriori (MAP) Expectation-Maximization algorithms (EM) and evolutionary methods (Genetic Algorithms (GA) [11,12,13] or Genetic Programming (GP) [14]). Both EM and GP approaches managed to, respectively, learn probabilities of Stochastic Context-Free Grammars (SCFG) for RNA structure prediction

[15,16,17] and derive non-probabilistic CFGs for non-biological problems [18]. Successful applications of evolutionary algorithms to SCFG [19,20,21] include our earlier research on SCFGs for protein binding sites [22]. Since, unlike EM techniques, GA-based grammar inference allows introducing pressure towards more compact grammars (see Methods) and is less dependent on initial estimates of rule parameters [20], we choose this approach for learning grammar rules.

In this work we exploit the expressive power of Stochastic Context-Free Grammars to represent the subtle and complex sequence motifs underlying H-H interactions in TM proteins. The aim is to facilitate sequence based classification of helix pairs regarding their three-dimensional configuration. As a result, a class template can be assigned to a pair of helices with high accuracy. This would be extremely valuable to constrain ab initio protein structure predictions or for threading refinement.

2 Materials and Methods

2.1 Datasets

The first dataset was created on the basis of Walters and DeGrado (WDG) dataset [8]. It includes fragments of helix sequences that are in contact. We consider only the 4 most populous contact types (classes 1-4). Unlike the original set where lengths of fragments varied from 10 to 14, we kept only the 10 residues which provided the closest match with a class template. The second dataset is based on the non-redundant set of alpha-helical chains from PDBTM database [2] as of 30th November 2009. Then TM alpha helices with at least one contact residue according to Promotif3 [23] were extracted. RMSD to the representatives of the 4 WDG classes were calculated. A helix pair was assigned to a certain class if its RMSD was lower than the highest RMSD in the class of the original WDG set, i.e. 0.66, 0.93, 0.76 and 1.11A for classes 1 to 4 respectively. As a result, the PDBTM set comprises 641 helix pairs with a population of 174, 107, 64 and 69 assigned to classes 1 to 4, respectively. For training, each class used the 20 fragments which were the closest to their representative (PDBTM20). Finally, homologous sequences (40%) were removed using PAM250 matrix [24] from our combined training and test sets so that both sets were mutually independent. As result, the processed WDG test sets (WDGNR) contained 92, 49, 37 and 27 helix pair fragments for classes 1 to 4 respectively.

2.2 Principles and Formal Definitions

Amino-acid interactions between helices are subtle and complex in comparison to intra-helical interactions. Moreover, they display either parallel or anti-parallel topologies. Methods typically used for the purpose of protein pattern detection, Profile HMMs [25], cannot express these dependencies. Therefore, to classify the contact type class, we use a SCFG, which, not only, is capable of representing anti-parallel dependencies, but also can be induced automatically from a set of unrelated protein sequences which share common features [22]. The formal

definition of a context-free grammar G is the following [26]: $G =< V, T, P, S >$, where V is a finite set of non-terminal (NT) symbols, T is a finite set of terminal symbols, P is a finite set of production rules and S is a special start symbol ($S \in V$). The sets V and T are mutually exclusive. Each rule from the set P has the form: $A \rightarrow X$, where $A \in V$ and $X \in (V \cup T)*$. For a SCFG, probabilities are attributed to each rule. Usually, probabilities of all productions for one Left-Hand Side (LHS) symbol sum to one; the SCFG is then called proper.

Helix interface is defined as a set of residues which are in contact with residues from the other helix, i.e. distance between residues in contact cannot be greater than the sum of van der Waals radii of their atoms enlarged by 0.6A [27]. The residues of the inner or contact face of a helix are separated by either 1 or 2 residues of the outer face so that an average helix periodicity of 3.6 residue is preserved. Two helices are separated by a coil. In the anti-parallel configuration these can be described schematically by context-free grammar rules, such as [10]:

```
Interface -> InsideRes1 Outerface InsideRes2  | Turn
Outerface -> OutsideRes1 Interface OutsideRes2 | Turn
```

More specifically, we modified a non-probabilistic CFG proposed in [10] to obtain a grammar that imposes helix periodicity (3-4 residues) and is manageable within our probabilistic scheme (i.e. not extending ca. 200 rules):

```
Start -> [ Whatever OuterfaceP Whatever }
 | [ Whatever InterfaceP Whatever }
OuterfaceP -> TwoRes InterfaceP TwoRes | OneRes InterfaceL TwoRes
 | TwoRes InterfaceR OneRes | OneRes InterfaceB OneRes | Turn
OuterfaceL -> TwoRes InterfaceP TwoRes
 | TwoRes InterfaceR OneRes | Turn
OuterfaceR -> TwoRes InterfaceP TwoRes
 | OneRes InterfaceL TwoRes | Turn
OuterfaceB -> TwoRes InterfaceP TwoRes | Coil
InterfaceP -> TwoRes OuterfaceP TwoRes | OneRes OuterfaceL TwoRes
 | TwoRes OuterfaceR OneRes | OneRes OuterfaceB OneRes | Turn
InterfaceL -> TwoRes OuterfaceP TwoRes
 | TwoRes OuterfaceR OneRes | Turn
InterfaceR -> TwoRes OuterfaceP TwoRes
 | OneRes OuterfaceL TwoRes | Turn
InterfaceB -> TwoRes OuterfaceP TwoRes | Turn
Turn -> Whatever ] { Whatever
Whatever -> X Whatever | empty
TwoRes -> OneRes OneRes
```

where the symbols '[', ']', '{' and '}' refer to the beginning and end of helix 1 and helix 2 respectively. Four $Outer\text{-}face$ and $Interface$ NT symbols (marked with suffixed P, L, R, B) ensure that each complete helix turn is 3 or 4 amino-acids long, e.g. if $Outer\text{-}faceP$ is one-residue long, it can only be followed by $InterfaceB$ which is always two-residue long. Production rule

$Turn \rightarrow Whatever]\{Whatever$ imposes helix boundaries on parser by using] and { terminal symbols. Moreover, the *Whatever* non-terminal allows to deal with parts of the helix that are not involved in the contact and thus do not share contact pattern.

2.3 Representation of Amino-Acid Properties

OneRes symbol refers to one amino-acid in a sequence. However, instead of using the amino-acid identity, which would make the grammar induction intractable, information about the level of a physio-chemical property (described later in this section) of a residue is carried. More specifically, *OneRes* can be one of three NT symbols that represent low, medium and high level of the property of interest, e.g. van der Waals volume: $OneRes \equiv Low|Medium|High$. The rationale behind this representation is to integrate quantitative information about amino-acid properties into our stochastic framework. An important advantage of this method is that it reduces the number of possible combinations of the Right-Hand Side (RHS) symbols in production rules. Therefore, a number of rules, which is maintainable in the learning process, is kept without losing generality of the grammar in the beginning of induction. For each given property, our method relies on defining all the terminal rules in the form:

```
Low     ->  amino-acid identity 1..20
Medium  ->  amino-acid identity 1..20
High    ->  amino-acid identity 1..20
```

and associating them with proper probabilities which are calculated using the known quantitative values associated to the amino acid identities. Since all terminal rules are fixed with given probabilities, unlike probabilities of all other rules, they do not need to be induced during the learning process. Moreover, to avoid trivial solutions, non-terminals which are Left-Hand Side (LHS) symbols in the terminal rules are prohibited from being LHS non-terminals of the other rules. We use the 5 categories of amino-acids from AAindex [28] as suggested in [22]: beta propensity, alpha and turn propensity, composition, physio-chemical properties and hydrophobicity.

2.4 Parsing

We use an implementation of the stochastic Earley parser [29]. In our framework Baum-Welch style Earley algorithm, where a probability for a certain node is calculated as a sum of probabilities of all sub trees, is used for training during grammar induction. This helps avoiding rapid convergence to trivial local minima in the absence of a negative training set. On the other hand, Viterbi style Earley algorithm is used for scanning, where a probability for any node in the parse tree is calculated as a maximal probability from all sub trees. According to our previous experiments, the Viterbi algorithm produces better discrimination between positive and negative samples and therefore it is more appropriate for scanning. Moreover [15,22] suggest that for a correctly induced grammar, the

most likely parse tree could reflect structural features of a molecule. The output of the stochastic parser is the log probability of the couple of residues involved in a long range helix contact of a certain type, so it is a similarity measure, which estimates how the sequence of interest matches the rules associated to the interaction class.

2.5 Learning Method for Stochastic Context-Free Grammars

In order to generate interface specific descriptors using the rules described in the previous section, a training set composed of positive examples of sequence fragments containing the interface is used to infer rule weights. The general principle behind our framework is to start the learning process with the complete set of rules expressing prior knowledge of the intra-helix interaction. Then, during training, rule probabilities are inferred to express contact type specific dependencies. Although this approach leads to quite large sets of rules even for moderate alphabets, it avoids bias which would be introduced by additional constraints. In this work, induction is performed by a genetic algorithm.

Similarly to [22] in this work a single individual in GA represents a whole grammar. The genotype is coded with real numbers ($< 0, 1 >$) linked to rule probabilities. The original population of size 200 is initialized randomly and then iteratively subjected to evaluation, reproduction, genomic operators and finally succession. The objective function of the GA is defined as an arithmetic average of logs of probabilities returned by the parsing algorithm for all positive training samples. The reproduction step of the GA uses the tournament method with 2 competitors [30], which ensures that the selective pressure is held at the same level during the whole induction process. In addition, the diversity pressure is kept by using a sharing function that decreases fitness score of individuals on the basis of their similarity to other individuals in the population. The distance between individuals takes into account that probability of a rule depends not only on its own gene but also on all genes referring to rules with the same LHS non-terminal [22]. In each GA epoch (generation of individuals), only the poorer 50% of the population is substituted by new individuals to ensure the stability of the GA algorithm. Offspring are produced by averaging genetic information of two individuals with some random distortion in order to enhance exploratory capabilities of the algorithm. Subsequently, a classical one point mutation operator is used to mutate randomly chosen genes. The probabilities of crossover and mutation are 0.9 and 0.01 respectively. The algorithm stops when there is no further significant improvement in the best scores (ratio 1.001 over 100 iterations). The implementation of our grammar induction algorithm is based on M. Wall's GAlib library which provides a set of C++ genetic algorithm objects [22].

A new genotype to phenotype function $f2 = phene(gene(W \rightarrow XYZ))$ was designed to facilitate rapid convergence and enhance exploring capabilities of the genetic algorithm. Let $A \rightarrow BCD$ is a context-free rule with LHS non-terminal A, $gene(A \rightarrow BCD)$ is a real number from range 0 to 1 linked with $A \rightarrow BCD$ rule and $geneavg(A)$ is a mean value of all genes associated with rules that

start with LHS non-terminal A. Then $tmpval(A \rightarrow BCD)$ is calculated in the following way:

```
if  gene(A->BCD)>2*geneavg(A)
then    tmpval(A->BCD)=gene(A->BCD)
else    tmpval(A->BCD)=gene(A->BCD)^10/(2*geneavg(A))^9.
```

Finally, normalization is carried out to obtain proper probabilities for each rule:

```
phene(A->BCD) = tmpval(A->BCD)/sum(XYZ){tmpval(A->XYZ)}.
```

Thus, $phene(A \rightarrow BCD)$ is the proper probability of the rule $A \rightarrow BCD$. The function assures that for a certain range of gene values, even small variations lead to significant changes in the phenotype. It reduces the number of active rules, since many of them have a near zero probability from the beginning of the induction. Thus, it speeds up the processing of each individual. The definition of the $f2$ function is consistent with a natural trend during grammar evolution where probabilities of unnecessary rules are reduced. This is an inherent property of proper stochastic grammars: distributions of probabilities with a small number of rules, which express well the pattern of interest, give better scores than even distributions of probabilities for all possible rules. After grammar induction, the final set of rules can be pruned to omit those which have a limited impact on the overall score of a scanned sequence.

Although genetic algorithms converge whatever their initial population [30], they may not find the global optimal solution. Therefore, for each grammar generation, we produced several grammars and selected the best one. Time needed for producing a grammar could take up to ca. 20 hours using Intel Xeon 2.4GHz quad-core processor systems at Wroclaw Centre for Networking and Supercomputing. The scanning took approximately one minute for parsing the whole test set by one grammar.

2.6 Protocol for Evaluation of Transmembrane H-H Interaction Prediction

For each of the four H-H interaction classes, 3 grammars were generated using PDBTM20 training set for each of the 6 selected amino-acid properties. The sequences of helix pair fragments from the WDGNR dataset were parsed for the four classes using all grammars. As a result, logs of probability that a sequence could have been generated by a given grammar were assigned to each H-H contact. The scores for positive and negative validation sets were analyzed by means of Receiver Operator Characteristics (ROC) methodology. The Area Under ROC Curve (AUC ROC) was used for general assessment of classifier quality and selection of the best grammar. In addition, Specificity and Sensitivity measures were calculated. Although for many applications it is desirable to maintain high Specificity or Sensitivity, we assume that the highest value of their product marks the optimal threshold for the parse score. For this threshold, Accuracy is provided.

3 Results and Discussion

3.1 Performance of Classifiers on Independent Test Set

The performance of grammar descriptors was assessed in a series of class-by-class classifications using WDGNR independent test set. On the basis of AUC ROC results for each class against other 3 classes, the properties, which lead to best scoring grammars, were selected. These were accessibility for class 1, van der Waals (vdW) volume for classes 2 and 3 and beta/turn propensity for class 4 (Tab. 1). The overall quality of classifiers measured by the Area under ROC curve

Table 1. H-H contact fragments classification performance using independent test set

Trained for	Using property	Tested against	AUCROC	Sensitivity	Specificity	Accuracy
c1	accessibility	c2	0.61	0.65	0.55	0.62
		c3	0.63	0.51	0.73	0.57
		c4	0.55	0.62	0.56	0.61
		c2+c3+c4	0.60	0.67	0.52	0.59
c2	van der Waals volume	c1	0.70	0.78	0.63	0.68
		c3	0.59	0.73	0.51	0.64
		c4	0.77	0.58	0.74	0.76
		c1+c3+c4	0.68	0.78	0.61	0.65
c3	van der Waals volume	c1	0.71	0.62	0.78	0.74
		c2	0.59	0.49	0.76	0.64
		c4	0.73	0.54	0.89	0.69
		c1+c2+c4	0.68	0.54	0.79	0.75
c4	beta-sheet propensity	c1	0.56	0.67	0.48	0.52
		c2	0.52	0.56	0.51	0.53
		c3	0.73	0.63	0.81	0.73
		c1+c2+c3	0.59	0.67	0.50	0.52

Table 2. Properties used by best class-by-class classifiers. Class-by-class classification of helix-helix pair contact fragments performance measured by Area and ROC curve using independent test set.

c1		c2		c3		c4	
c1		accessibility	0.61	accessibility	0.63	frequency	0.57
c2	VdW volume	0.70		frequency	0.64	vdW volume	0.77
c3	vdW volume	0.71	vdW volume	0.59		vdW volume	0.73
c4	beta prop.	0.56	accessibility	0.59	beta prop.	0.73	

varied from 0.59 for c4 to 0.68 for c2 and c3. The optimal thresholds for scores yielded in different balances between Sensitivity and Specificity. More precise evaluation of the classifiers is possible by analysis of their ROC curves (Fig. 1). There is a shift towards Sensitivity for c2 and a shift towards Specificity c3 vdW volume grammars. Typically, the relatively worst performance was obtained in classification of c1 vs. c4 or c2 vs. c3 classes. This is, however, consistent with

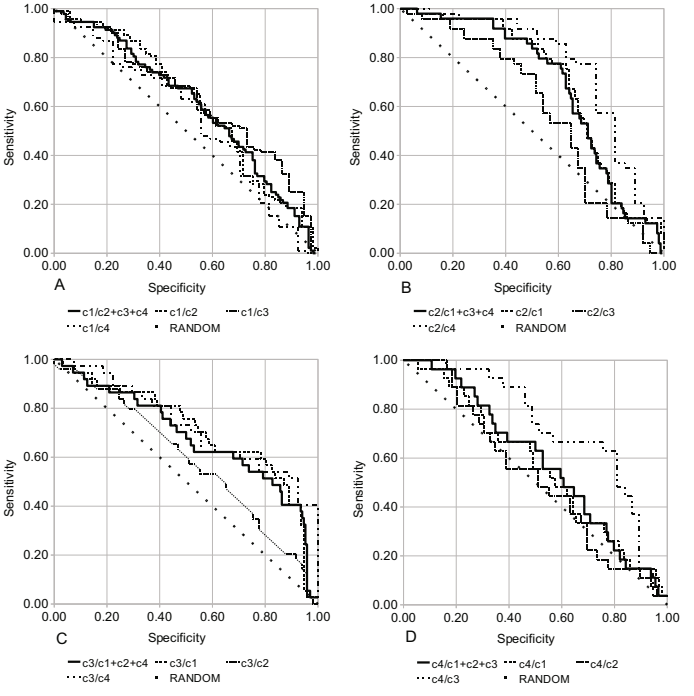

Fig. 1. ROC curves for H-H contact fragment classifiers: c1 accessibility-based (A), c2 vdW volume-based (B), c3 vdW volume-based (C) and c4 beta propensity-based (D)

the fact that these pairs of classes are similar in terms of RMSD. They differ in relative direction of helices (anti-parallel for c1 and c2, parallel for c3 and c4).

Representatives of 5 categories of amino-acid properties were utilized for grammar training resulting in varying robustness for different class-by-class comparisons. The properties that were used in best scoring grammars are presented in Table 2. In general, area under ROC curve values of the best grammars, for each class-by-class classification, were in the range from 0.56 to 0.77. Accessibility and vdW volume were most useful for distinguishing between classes unrelated in terms of their 3D shape. Frequency and beta-sheet propensity were the properties that allow for classification between anti-parallel and parallel versions of classes that share similar spatial configurations.

3.2 Analysis of Classifiers Features

Our analysis details the features of the SCFG classifiers, which contribute to the overall performance of the method. Our findings suggest that the difference in sequence composition, in terms of the property underlying the grammar, is the main factor. However, in a few cases descriptors that performed better than expected, according to sequence composition comparison, were obtained. Such examples include classifications between: c1 and c2 using grammar based on accessibility, c3 and c1+c4 using grammar based on van der Waals volume and c4

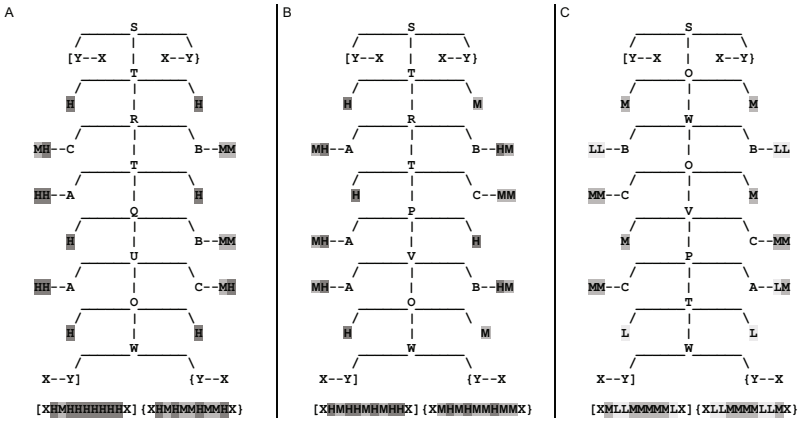

Fig. 2. Parse trees that would give maximum scores for (A) c1 accessibility grammar, (B) c4 accessibility grammar and (C) c3 vdW volume grammar. H, M, L are property level NTs, which refer to high, medium, low level of a given property. X is any amino-acid (probability of 1/20 to each amino-acid type). S is a start symbol. Subsets of NTs T,U,V,W and O,P,Q,R are designated to model Inter- and Outer-face of the helix pair (order of subsets is arbitrary). The sans-serif font for property level NTs for (B) indicates a modified method of assignment of probabilities to the rules started with those symbols (in text).

and c1 using grammar based on accessibility. The last was obtained in a scheme that included modified training and test sets. Moreover, property levels were related to the average property level in a training set, instead of the average over 20 amino-acids as utilized in the basic scheme. In Fig. 2, example of parse trees that would give maximum scores for these grammars are shown. Although they would not necessarily result in maximal parse scores for individual sequences, their structure is very likely to be found in real parses. It would be difficult at this stage of study to induce relations between parse tree structures and biological features of helix pairs, especially for classes 3 and 4, which are parallel. However, the analysis of the parse trees suggests that grammar classifiers can benefit from representation of dependencies between helices. For example, in (C) the most probable rules typically require that amino-acids from two helices have similar size at each stage of derivation. These results confirm the value of a strategy which uses amino-acid properties instead of amino-acid identities for modeling non-homologous helix pair sequences. However, the exact assignment of amino-acid to property levels remains an issue. We noticed that non-terminals related to property levels underrepresented in H-H bundles were rarely used in induced grammars, which hampered the capability of representing class defining patterns.

4 Conclusions

Our SCFG framework produced sequence-based descriptors, which represent classes of transmembrane helix-helix interaction configurations. The grammar

descriptors were tested using independent test sets. Amino-acid properties most relevant to each class-by-class classification were selected. Areas under ROC curves obtained for best classifiers were typically between 0.60 to 0.70 and in some cases higher. This shows that amino-acid sequence based descriptors can be used for prediction of H-H interaction structural class, for a pair of H-H sequences. Thus, they can be used to constrain the search space of an ab initio prediction method for transmembrane proteins. Another strategy could be use of predicted conformations of H-H interactions to deprive sets of structures modeled in the process of ab initio prediction of low quality items.

At this stage of research, the predictive power of the classifiers is mainly grounded in differences in amino-acid composition of H-H pairs in terms of the amino-acid properties. However, some grammar descriptors perform above expected level, based on sequence composition. This suggests that capability of CFG to represent higher level (anti-parallel) dependencies between interacting helices can contribute to the classification. Currently, we investigate the influence of several factors, including choice of the class representatives and the training sets, definition of the amino-acid property levels and design of the initial grammar structure. We also research the hypothesis that there are subclasses within WDG classes of H-H sequences more prone to structural description than others.

The other factor, important for the procedure of training, is the selection of the training set. According to recent publications [3,8], the optimal length of a helix fragment is from 10 to 14 residues. However the position of cutting of fragments could potentially have an impact on the quality of prediction. Finally, the clustering of H-H interfaces is still an open problem. The numbers of PDBTM sequences assigned to each WDG class representative were linearly correlated to the cut-off levels. This suggests, that the level of RMSD around 1.50 A prohibits the classes from overlapping but only conveys a limited biological meaning.

Acknowledgments. This work was partially supported by Ministry of Science and Higher Education of Poland (N N519 401537), British Council Young Scientists Programme (WAR/324/108) and MLODA KADRA Programme.

References

1. Yarov-Yarovoy, V., Schonbrun, J., Baker, D.: Multipass Membrane Protein Structure Prediction Using Rosetta. Proteins 62, 1010–1025 (2006)
2. Tusnady, G.E., Dosztányi, Z., Simon, I.: PDB_TM: selection and membrane localization of transmembrane proteins in the PDB. Nucleic Acids Res. 33, D275–D278 (2005)
3. Barth, P., Wallner, B., Baker, D.: Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. 106, 1409–1414 (2009)
4. Wu, S., Zhang, Y.: A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24, 924–931 (2008)
5. Li, W., et al.: Application of sparse NMR restraints to large-scale protein structure prediction. Biophys. J. 87, 1241–1248 (2004)
6. Izarzugaza, J.M.G., Grana, O., Tress, M.L., Valencia, A., Clarke, N.D.: Assessment of intramolecular contact predictions for CASP7. Proteins 69(suppl. 8), 152–158 (2007)

7. Sathyapriya, R., Duarte, J.M., Stehr, H., Filippis, I., Lappe, M.: Defining an Essence of Structure Determining Residue Contacts in Proteins. PLoS Comput. Biol. 5, e1000584 (2009)
8. Walters, R.F.S., De Grado, W.F.: Helix-packing motifs in membrane proteins. Proc. Natl. Acad. Sci. 103, 13658–13663
9. Russ, W.P., Engelman, D.M.: The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. 296(3), 911–919 (2000)
10. Waldispühl, J., Steyaert, J.-M.: Modeling and predicting all-transmembrane proteins including helix-helix pairing. Theoretical Computer Science 335, 67–92 (2005)
11. Holland, J.H.: Adaptation in Natural and Artificial Systems. Univ. Michigan (1975)
12. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning Reading. Addison-Wesley, Reading (1989)
13. O'Neill, M., Ryan, C.: Grammatical Evolution. IEEE Trans. Evol. Comput. 5, 349–358 (2001)
14. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
15. Sakakibara, Y., Brown, M., Underwood, R.C., Mian, I.S.: Stochastic Context-Free Grammars for Modeling RNA. In: Procs 27th Hawaii Int. Conf. System Sciences (1993)
16. Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R., Haussler, D.: Stochastic Context-Free Grammars for tRNA. Nucleic Acids Res 22, 5112–5120 (1994)
17. Knudsen, B., Hein, J.: RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 15, 446–454 (1999)
18. Mernik, M., Crepinsek, M., Gerlic, G., Zumer, V., Viljem, Z., Bryant, B.R., Sprague, A.: Learning CFG using an Evolutionary Approach. Technical report (2003)
19. Sakakibara, Y.: Learning context-free grammars using tabular representations. Pattern Recognition 38, 1372–1383 (2005)
20. Keller, B., Lutz, R.: Evolutionary induction of stochastic context free grammars. Pattern Recognition 38, 1393–1406 (2005)
21. Cielecki, L., Unold, O.: Real-valued GCS classifier system. Int. J. Appl. Math. Comput. Sci. 17, 539–547 (2007)
22. Dyrka, W., Nebel, J.-C.: A Stochastic Context Free Grammar based Framework for Analysis of Protein Sequences. BMC Bioinformatics 10, 323 (2009)
23. Hutchinson, E.G., Thornton, J.M.: PROMOTIF - A program to identify structural motifs in proteins. Protein Science 5, 212–220 (1996)
24. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345–352 (1978)
25. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D.: Hidden Markov models in computational biology: Applications to protein modeling. J. Mol. Biol. 235, 1501–1531 (1994)
26. Revesz, G.E.: Introduction to Formal Languages. McGraw-Hill, New York (1983)
27. Gimpelev, M., Forrest, L.R., Murray, D., Honig, B.: Helical Packing Patterns in Membrane and Soluble Proteins. Biophysical J. 87, 4075–4086 (2004)
28. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: AAindex: amino acid index database. Nucleic Acids Res. 36, D202–D205 (2008)
29. Stolcke, A.: An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities. Computational Linguistics 21(2), 165–201 (1995)
30. Arabas, J.: Wyklady z algorytmow ewolucyjnych Warsaw: WNT (2004)
31. Wall, M.: GAlib library documentation (version 2.4.4). MIT, Cambridge (1999)