Systematic elicitation of sequence patterns associated with non-proline *cis* peptide bonds

Konstantinos P. Exarchos, Themis P. Exarchos, *Student Member, IEEE,* Costas Papaloukas,
Anastassios N. Troganis, Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*—Non-proline *cis* peptide bonds have been quite underrated for many years, due to the limited amount of structural information available. There is now significant evidence that non-proline *cis* peptide bonds occur more frequently than previously thought, and that they are often located at or near important sites of the protein molecule. In this work, we employ a combinatorial pattern discovery algorithm in order to identify simple and specific amino acid patterns, associated with the occurrence of non-proline *cis* peptide bonds in proteins. The derived patterns after careful validation help in gaining insight into the factors that influence the formation of non-proline *cis* peptide bonds.**

I. Introduction

Because of the partial double bond character of the peptide bond, two isomers are energetically preferred, *cis* and *trans* (Figure 1). In protein structures, the *trans* conformation is overwhelmingly preferred, whereas the *cis* conformation occurs rarely because of its higher intrinsic energy. A survey conducted by Weiss *et al.* [1] in a non-redundant set of 571 proteins, reported that 0.03% of the Xaa-nonPro (where Xaa denotes any of the 20 amino acids and nonPro is any amino acid except Proline) and 5.2% of the Xaa-Pro peptide bonds are in *cis* conformation. It is noteworthy that the resolution of the protein structure is indicative of the number of *cis* peptide bonds detected. This is due to the fact that the distance between two adjacent alpha carbons in *cis* conformation is nearly 1Å shorter than in the *trans* conformation [2]. The correlation between the resolution and the *cis* conformation content may have left many *cis* nonPro peptide bonds unrecognized, especially in experiments at medium or low resolutions.

Despite the low frequency of occurrence, *cis* nonPro conformation bears great importance in a variety of biological processes. It has been suggested that, *cis* nonPro formations occur frequently either at or near the active sites of protein molecules, and it is very likely that they have roles in the function of the protein [1, 3-5]. Such examples are carboxypeptidase A, dihydrofolate reductase and intein gyrA, where *cis* peptide bonds have been strongly proposed to bear a functional role [5, 6]. Moreover, several *cis* formations play a significant role in the final structure, as well as the folding and stability of many proteins [4, 7]. Furthermore, the occurrence of *cis* nonPro formations has been associated with steric strain in proteins and it has been speculated that these sites of strain comprise some kind of energy reservoir for the protein [7].

Certain factors have been proposed in the literature to affect the occurrence of *cis* peptide bonds. Nuclear Magnetic Resonance (NMR) experiments have reported that there is a strong connection between the primary amino acid sequence and the occurrence of *cis* conformations in proteins [8]. In addition, the physicochemical properties of the surrounding residues have been proven to influence the isomerization of peptide bonds [9]. Based on these facts, several approaches have been implemented in order to predict the conformation of the peptide bond either in Xaa-Pro amino acid pairs or between any two amino acids [10-12]. However, all these methods utilize a rather opaque architecture, whereby a classifier is employed to distinguish between the conformations, without providing insight regarding the nature and the interactions of the peptide bond isomerization.

In this work we perform a systematic attempt to discover non-random patterns that are associated with *cis* nonPro formations and accurately describe these bonds. For this purpose a combinatorial pattern discovery algorithm is employed which reports regular expression-type patterns that are overrepresented in a set of sequences. Thus, we circumvent the limitations of previously reported methods, and provide some hints about the physical causes of *cis* nonPro formations. Similar studies have emerged nonrandom patterns correlated with the secondary structure [13, 14] or regions of disorder in proteins [15]. Moreover, in our study several conservative substitutions are permitted among the amino acids, regarding their structural or chemical nature. The careful assessment of the derived patterns might further contribute to the biological interpretation of *cis* nonPro formations.

Manuscript received May 27, 2008.

K. P. Exarchos and T. P. Exarchos are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110, Ioannina, Greece, the Dept. of Medical Physics, Medical School, University of Ioannina, GR 45110, Ioannina, Greece, and the Institute of Biomedical Technology, CERETETH, GR 41222, Larissa, Greece (e-mail: kexarcho@cc.uoi.gr, exarchos@cc.uoi.gr).

C. Papaloukas and A. N. Troganis are with the Dept. of Biological Applications and Technology, University of Ioannina, GR 45110, Ioannina, Greece (e-mail: papalouk@cc.uoi.gr, atrogani@cc.uoi.gr).

D. I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110, Ioannina, Greece, and the Institute of Biomedical Technology, CERETETH, GR 41222, Larissa, Greece (phone: 0030-26510-98803; fax: 0030-26510-98889; e-mail: fotiadis@cs.uoi.gr).

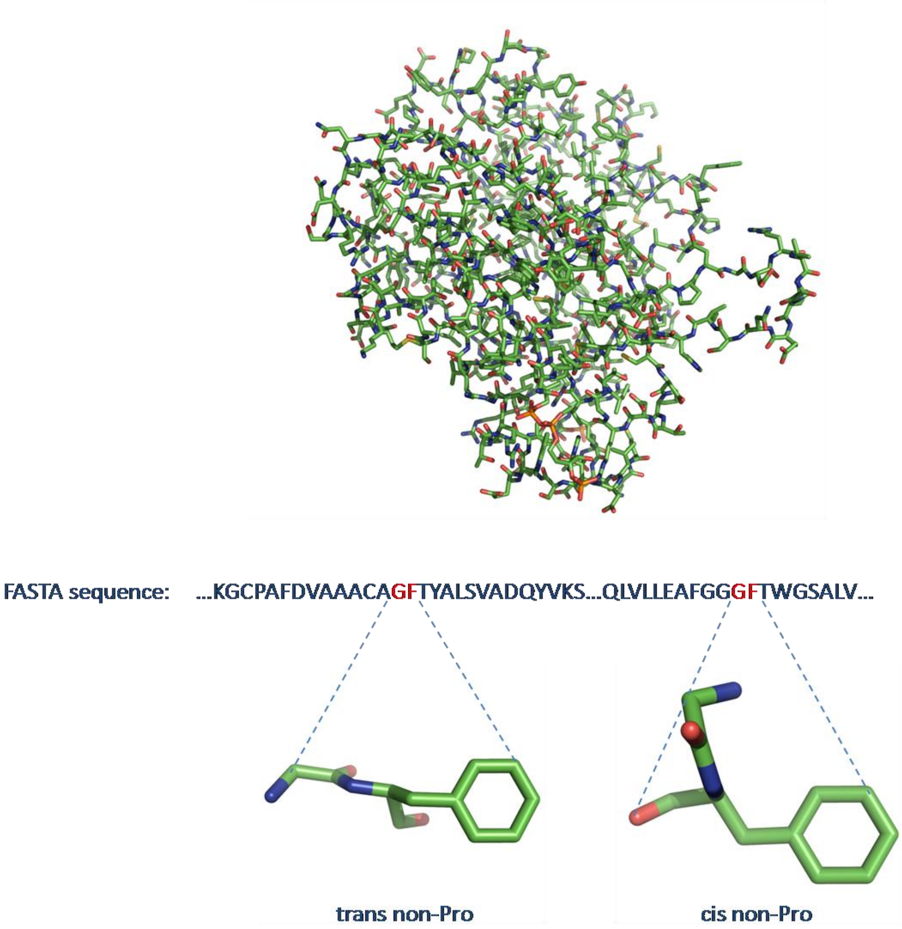

Fig. 1. Conformational isomers of a Glycine–Phenylalanine (G-F) peptide bond. The displayed structure is 1HNJ, acquired from the Protein Data Bank.

II. Materials and Methods

A. Dataset

The data necessary for our study consist of 3050 well resolved (resolution < 2.0Å) and refined protein sequences, extracted from the Protein Data Bank (PDB) [16]. The chosen proteins have been determined by X-ray crystallography and display less than 25% sequence identity; furthermore the R-factor is less than 0.25. The annotation of the dataset is performed using VADAR (Volume Area Dihedral Angle Reported) [17], which accepts PDB files and calculates the dihedral angle ω. Bonds with ω dihedral angle between -30° and +30° are considered as *cis*, whereas bonds with ω dihedral angle between 150° and 210° are assumed to be *trans*. For each nonPro peptide bond in the above non-redundant set of sequences, a region containing its immediate ±5 neighboring residues is assembled [10, 18]. In order to avoid interclass overlapping regions, the ±5 *trans* nonPro residues flanking a *cis* nonPro region are excluded from our study (Figure 2). Thus, two non-overlapping datasets are composed, the D_+ containing all *cis* nonPro regions and the D_- containing all *trans* nonPro regions. All extracted regions have a length of 11 residues.

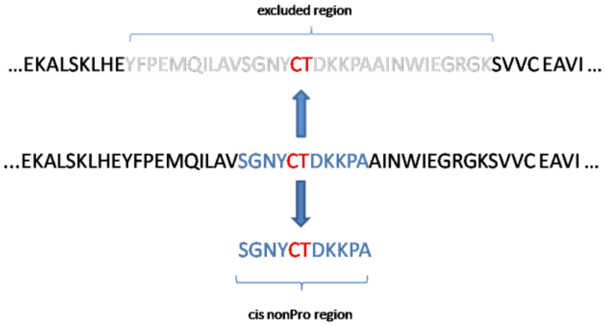
excluded region
...EKALSKLHEYFPEMQILAVSGNYCTDKKPAAINWIEGRGKSVVC EAVI ...
...EKALSKLHEYFPEMQILAVSGNYCTDKKPAAINWIEGRGKSVVC EAVI ...
SGNYCTDKKPA
cis nonPro region

Fig. 2. The above sequence segment has a *cis* nonPro peptide bond between C-T residues, in the center. The *cis* region taken into account, shown in the bottom of the figure, contains ±5 residues flanking the peptide bond. Residues in the excluded region, shown in the top of the figure, are eliminated from our study so that no overlapping regions exist both in D_+ and D_-.

B. Pattern discovery

All regions in D_+ are properly formatted and provided in the TEIRESIAS pattern discovery algorithm [19], which efficiently detects all maximal patterns present in the dataset, without enumerating the entire solution space. The algorithm operates in two phases: scanning and convolution. During the scanning phase, patterns exceeding a minimum support threshold are maintained; next these elementary patterns are progressively combined into larger patterns, until all existing maximal patterns are discovered. The algorithm detects all non-overlapping patterns of three to eleven residues (W=11), with minimum support K=2 [19, 20], requiring at least three constant literals (L=3). An initially low threshold is chosen regarding the support of the discovered patterns, since further validation is performed in the next stage of our analysis. The choice of L=3 is arbitrary but justified *a posteriori* by the observation that patterns with more specified residues are not frequent enough in our database, whereas patterns with fewer literals are rather uninformative. The pattern discovery process is carried out using three different types of analysis with biological insight: i) exact pattern discovery, ii) pattern discovery using a chemical equivalency set: [AG], [DE], [FY], [KR], [ILMV], [QN], [ST] and iii) pattern discovery using a structural equivalency set: [CS], [DLN], [EQ], [FHWY], [ITV], [KMR]. Residues in "[ ]" are allowed to substitute one another during the pattern discovery process, considering either their chemical behavior or structural nature. In the first type of pattern discovery, positions in the extracted patterns are occupied either by a specific residue or by a wildcard (denoted as a single dot "."); during the other two types of pattern discovery, positions may also be occupied by one of the specified character classes.

C. Pattern validation

Since the initial set of the discovered patterns, is derived by taking into account only D_+, the patterns themselves are not guaranteed to be highly specific. Thus, the contradiction of patterns against a negative control set (i.e. D_-) is necessary, in order to exclude patterns that are equally represented in both datasets. The pattern validation procedure is based on comparing proportionally the number of matching regions in D_+ and D_-. If P is a pattern and M(P) is the set of regions matching P, then we can define score as:

$$Score = \frac{|D_+ \cap M(P)|}{|D_+ \cap M(P)| + norm \times |D_- \cap M(P)|} \tag{1}$$

where $norm = |D_+|/|D_-|$, which is used to phase out the imbalance in our datasets, since it is reasonable to expect $|D_+ \cap M(P)|$ and $|D_- \cap M(P)|$ to be roughly proportional to the database size. Score ranges in [0,1], with score=0 in the least favorable case and score=1 in the most favorable one. If a pattern is equally represented in the *cis* and *trans* regions, then score $\simeq$ 0.5 and values greater than that should indicate a propensity towards *cis* nonPro regions. In our case a much more rigorous threshold is chosen (score $\geq$ 0.9) in order to maintain only highly selective associations, that capture the nature of *cis* nonPro formations. A further constraint is imposed on the derived patterns:

$$|D_+ \cap M(P)| \geq 4 \tag{2}$$

This ensures that every pattern should match the *cis* regions at least four times; thus excluding patterns that match *trans* regions zero times and *cis* regions. This constraint also ensures that patterns with zero matches in D_-, yielding score=1, are not essentially descriptive and should not be maintained unless they are adequately represented in D_+.

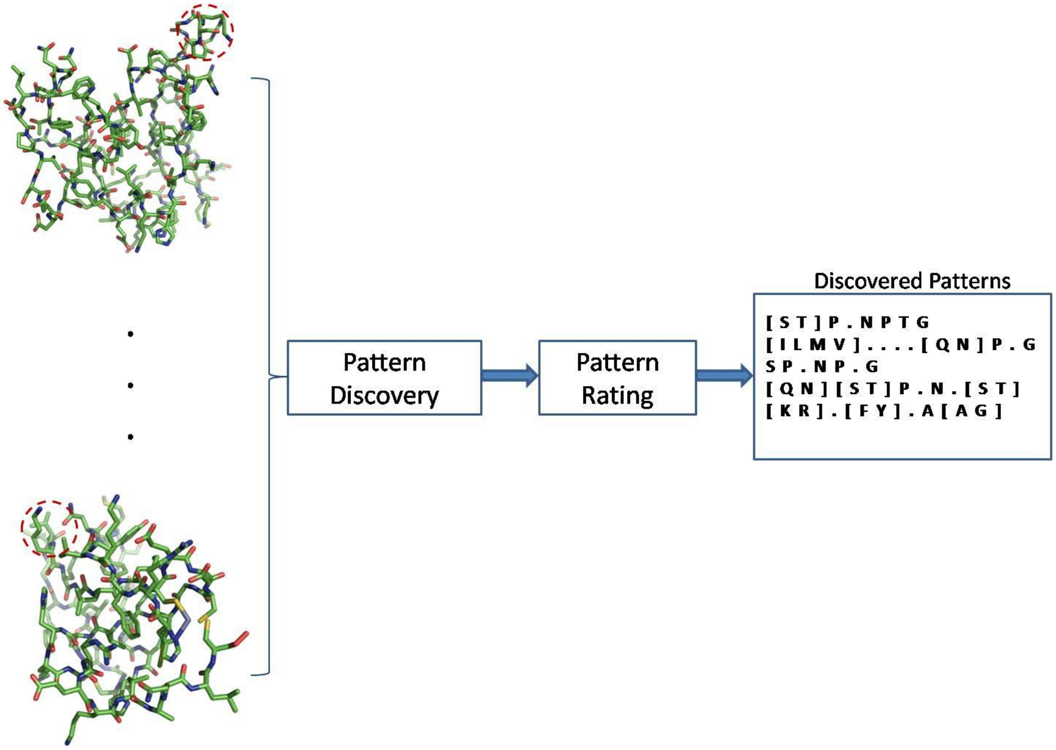

Fig. 3. The steps of the proposed methodological analysis.

Moreover, the reliability of the derived patterns is assessed; besides the score, a measure of significance is computed and attached to every pattern. This measure is estimated using the Bayes theorem in conjunction with a second order Markov chain and represents the probability that the pattern under consideration is found by chance in a very large biological database. The steps of the proposed methodological analysis are shown in Figure 3.

III. Results and Discussion

A summary of the most representative patterns (top-20 highest scoring patterns), obtained from our study, is shown in Table 1; for each pattern, the score it obtained, as well as, the significance measure attached to it, are given. Our methodological analysis has rediscovered some known facts about the nature of *cis* nonPro bonds, while it has also detected certain unknown ones. From the careful assessment of the highest scoring patterns in Table 1, several interesting conclusions can be drawn. The most obvious remark is that *cis* nonPro regions are replete with Glycines (G), which are found in 50 out of the 60 patterns presented in Table 1. Glycine is found with very high frequency either as part of the peptide bond or in its immediate neighborhood [3]. Some representative patterns are “H...G..GL”, “K.G.G...P”, “F..G.G.R”, “[AG].[QN].[ILMV]K.V[ST]”, “[KR].G.G.R.P”, “GT[ILMV]..QL”, “[KMR].G.G.R.P”, “[FHWY][KMR]..KG..K”, “K.[DLN]GT....L” and their variations, as well as many others. The tiny Glycine residue raises the probability of acquiring a *cis* conformation, probably due to lack of steric hindrance. Besides Glycine, Alanine (A) is also frequently observed in the neighborhood of cis nonPro formations, due to its confined size [3]; It is observed in many patterns, such as “A.SG.Y”, “GA.D.A”, “[AG].[QN].[ILMV]K.V[ST]”, “A[DLN]..[DLN][KMR][DLN]V” and “A.[CS]..YG[DLN]”. Moreover, residues Leucine (L), Lysine (K), Threonine (T) and Serine (S) have relatively high propensities for occurrence near *cis* nonPro bonds, as we can see in many patterns shown in Table 1 (e.g. “H...G..GL”, “K..GT....L”, “LN.LK”, “E.K[KMR][KMR].L”, “[ST]T.E..A[ILMV]”, “K.[DLN]GT....L”, “S[AG].[FY]GL”). Especially Leucine and Lysine are more frequent than the other two residues. Regarding Leucine, Serine and Threonine similar observations have been reported in the literature, whereas Lysine has been previously found to occur scarcely near *cis* nonPro formations [3]. This can be attributed to the limited amount of solved structures available at that time. Aromatic residues (Phenylalanine (F), Tryptophan (W), Tyrosine (Y) and Histidine (H)) are also found in many of the highest scoring patterns, such as: “F..G.G.R”, “G.F.W”, “S[AG].[FY]GL”, “[FHWY][KMR]..KG..K” and “H...G..GL”, to name a few.

Table 1: The patterns are grouped according to the type of pattern discovery. Important details such as the score and significance of the derived patterns are also provided. The reported patterns are sorted first by score and then by significance.

Exact pattern discovery			**Chemical equivalency set**			**Structural equivalency set**		
Pattern	**Score**	**Significance**	**Pattern**	**Score**	**Significance**	**Pattern**	**Score**	**Significance**
KPGKGRRK	1	-37.673	KPGKGRRK	1	-37.673	KPGKGRRK	1	-37.673
H...G..GL	0.998	-14.556	[AG].[QN].[ILMV]K.V[ST]	0.999	-22.208	A[DLN]..[DLN][KMR][DLN]V	0.999	-21.106
K.G.G...P	0.997	-14.595	[KR].G.G.R.P	0.999	-19.440	[KMR].G.G.R.P	0.999	-19.228
G.G.R.P	0.996	-14.201	GT[ILMV]..QL	0.999	-18.060	[FHWY][KMR]..KG..K	0.999	-18.591
K..GT....L	0.996	-13.646	[DE][ST]G.YG	0.999	-18.941	K.[DLN]GT....L	0.999	-18.390
GT...QL	0.995	-13.957	LN.LK[ILMV]	0.998	-18.015	A.[CS]..YG[DLN]	0.999	-18.902
F..G.G.R	0.994	-14.369	G..[DE].K..S[ILMV]	0.998	-17.942	H...G..GL	0.998	-14.556
GT.E..L	0.992	-13.719	H...G..GL	0.998	-14.556	[DLN]G.[KMR].PL	0.998	-17.778
A.SG.Y	0.991	-14.401	[ST]T.E..A[ILMV]	0.998	-17.590	D[KMR].[EQ]...[ITV]L	0.997	-17.185
FKPG	0.990	-14.695	[AG]...[ILMV]K[ILMV][ILMV][ST]	0.997	-20.226	E.K[KMR][KMR].L	0.997	-17.626
GA.D.A	0.988	-18.877	S[AG].[FY]GL	0.997	-18.287	K.G.G...P	0.997	-14.595
LN.LK	0.987	-13.369	[AG][ST].EP.[ILMV]	0.997	-17.609	L..V[ITV].Q[DLN]	0.997	-17.560
G...W...D	0.980	-9.985	[DE][DE][AG]T[KR]..[ILMV]	0.997	-21.966	LG..[ITV]N.[DLN]	0.995	-17.440
G.F.W	0.977	-10.267	GT[ILMV]I.[QN]	0.997	-17.827	GT...QL	0.995	-13.957
G..G...W	0.972	-9.414	K.G.G...P	0.997	-14.595	F..G.G.R	0.994	-14.369
G....N...S	0.969	-8.639	[ILMV].G.[AG]..AT	0.997	-17.157	[KMR].G.G...P	0.994	-13.515
H.E.....L	0.963	-8.835	K..GT....L	0.996	-13.646	L..L.[DLN][ITV]T	0.994	-16.699
F....G..K	0.960	-8.888	F..G.G..[KR]	0.996	-13.761	[KMR]..KG..K	0.994	-13.462
P..G..K	0.958	-8.932	L[AG].[ILMV][ILMV]N.[ILMV]	0.996	-20.443	F..G.G..[KMR]	0.994	-13.565
H......GL	0.956	-8.798	E[DE][AG]....[ILMV]L	0.995	-16.296	[KMR]EP[DLN][DLN]	0.993	-17.265

Furthermore, b-branched amino acids (Valine (V), Isoleucine (I) and Threonine (T)) can also be observed in some of the patterns shown in Table 1 (e.g. “[AG].[QN].[ILMV]K.V[ST]”, “[ST]T.E..A[ILMV]”, “D[KMR].[EQ]...[ITV]L”).

It is noteworthy that some patterns are observed in all three types of pattern discovery, either completely unaltered or with slight variations. Some of these predominant patterns are “H...G..GL”, “K.G.G...P” and “F..G.G.R” (or its variations “F..G.G..[KR]” and “F..G.G..[KMR]”). In Figure 4 we can see an approximation of the three dimensional structure of these patterns. In should be noted that in all three patterns an underlying “pocket” like motif exists. Regarding the pattern “KPGKGRRK”, which yielded very high scores in all three types of pattern discovery, all its instances are found in the same protein sequence (2GECA), which has four successive *cis* nonPro bonds; thus, no conclusions about the nature of *cis* nonPro formations, in general, can be drawn from this pattern only.

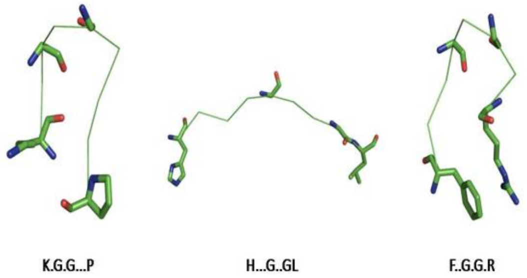

Fig. 4. Approximation of the three dimensional structure of the predominant patterns “K.G.G...P”, “H…G..GL” and “F..G.G.R”.

Overall, we can see that the significance measure for all retained patterns is very low, especially in the cases of pattern discovery using a chemical or a structural equivalency set. Such low values of significance ensure that the discovered patterns are quite unlikely to appear by chance, even in a very large biological database. Hence, the retained associations formulate highly specific and reliable descriptors of *cis* nonPro formations.

IV. Conclusions

Cis nonPro peptide bonds have been underrated for several years, due to their scarce occurrence. However, it is now clear that these formations are not irregularities of the respective structure that they are found, but actually play a significant role in the structure and function of the protein molecule. Although the limited amount of solved structures has impeded the study of *cis* nonPro bonds, with more three dimensional structures at hand today, more systematic approaches have become feasible.

Our analysis has yielded several descriptive patterns regarding the nature of *cis* nonPro bonds, most of which are consistent with previous findings, but it has also discovered some previously unknown associations. More important though, is that the surrounding and the interactions of *cis* nonPro conformations have been formulated in a list of simple and understandable patterns. Furthermore, the extracted patterns can be contradicted with available patterns concerning important sites in proteins (e.g. active sites) and, thus, enhance the functional prevalence of *cis* nonPro formations.

References

[1] M. S. Weiss, A. Jabs, and R. Hilgenfeld, "Peptide bonds revisited," *Nat Struct Biol,* vol. 5, p. 676, Aug 1998.

[2] D. E. Stewart, A. Sarkar, and J. E. Wampler, "Occurrence and role of cis peptide bonds in protein structures," *J Mol Biol,* vol. 214, pp. 253-60, Jul 5 1990.

[3] D. Pal and P. Chakrabarti, "Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations," *J Mol Biol,* vol. 294, pp. 271-88, Nov 19 1999.

[4] A. Jabs, M. S. Weiss, and R. Hilgenfeld, "Non-proline cis peptide bonds in proteins," *J Mol Biol,* vol. 286, pp. 291-304, Feb 12 1999.

[5] B. L. Stoddard and S. Pietrokovski, "Breaking up is hard to do," *Nat Struct Biol,* vol. 5, pp. 3-5, Jan 1998.

[6] T. Klabunde, S. Sharma, A. Telenti, W. R. Jacobs Jr, and J. C. Sacchettini, "Crystal structure of GyrA intein from Mycobacterium xenopi reveals structural basis of protein splicing," *Nat Struct Biol,* vol. 5, pp. 31-36, 1998.

[7] G. Fischer and T. Aumuller, "Regulation of peptide bond cis/trans isomerization by enzyme catalysis and its implication in physiological processes," *Rev Physiol Biochem Pharmacol,* vol. 148, pp. 105-50, 2003.

[8] C. Grathwohl and K. Wuethrich, "NMR studies of the rates of proline cis-trans isomerization in oligopeptides," *Biopolymers,* vol. 20, pp. 2623-2633, 1981.

[9] C. Frommel and R. Preissner, "Prediction of prolyl residues in cis-conformation in protein structures on the basis of the amino acid sequence," *FEBS Lett,* vol. 277, pp. 159-63, Dec 17 1990.

[10] J. Song, K. Burrage, Z. Yuan, and T. Huber, "Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information," *BMC Bioinformatics,* vol. 7, p. 124, 2006.

[11] M. L. Wang, W. J. Li, M. L. Wang, and W. B. Xu, "Support vector machines for prediction of peptidyl prolyl cis/trans isomerization," *J Pept Res,* vol. 63, pp. 23-8, Jan 2004.

[12] D. Pahlke, D. Leitner, U. Wiedemann, and D. Labudde, "COPS--cis/trans peptide bond conformation prediction of amino acids on the basis of secondary structure information," *Bioinformatics,* vol. 21, pp. 685-6, Mar 1 2005.

[13] M. J. Rooman, J. Rodriguez, and S. J. Wodak, "Relations between protein sequence and structure and their significance," *J Mol Biol,* vol. 213, pp. 337-50, 1990.

[14] M. J. Rooman and S. J. Wodak, "Weak Correlation Between Predictive Power Of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins," *Proteins: Structure, Function, and Genetics,* vol. 9, pp. 69-78, 1991.

[15] S. Lise and D. T. Jones, "Sequence patterns associated with disordered regions in proteins," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 58, pp. 144-150, 2005.

[16] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res,* vol. 28, pp. 235-242, 2000.

[17] L. Willard, A. Ranjan, H. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, and D. S. Wishart, "VADAR: a web server for quantitative evaluation of protein structure quality," *Nucleic Acids Res,* vol. 31, pp. 3316-9, Jul 1 2003.

[18] K. P. Exarchos, T. P. Exarchos, C. Papaloukas, A. N. Troganis, and D. I. Fotiadis, "Prediction of cis/trans isomerization using

feature selection and support vector machines," *J Biomed Inform*, 2008.

[19] I. Rigoutsos and A. Floratos, "Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm," *Bioinformatics,* vol. 14, pp. 55-67, 1998.

[20] I. Rigoutsos, A. Floratos, C. Ouzounis, Y. Gao, and L. Parida, "Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins," *Proteins,* vol. 37, pp. 264-77, Nov 1 1999.