Bioinformatics Advance Access published September 8, 2012

RAPHAEL: Recognition, periodicity and insertion assignment of solenoid protein structures.

Ian Walsh[1], Francesco G. Sirocco[1,2], Giovanni Minervini[1], Tomás Di Domenico[1], Carlo Ferrari[3] and Silvio C.E. Tosatto[1,*]

[1]Department of Biology, University of Padua, Viale G. Colombo 3, 35131 Padova, Italy.

[2]present address: Unit of Experimental Oncology 1, CRO National Cancer Institute, Via Franco Gallini 2, 33081 Aviano, Italy.

[3]Department of Information Engineering, University of Padua, Via Gradenigo 6, 35121 Padova, Italy.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Repeat proteins form a distinct class of structures where folding is greatly simplified. Several classes have been defined, with solenoid repeats of periodicity between ca. 5 and 40 being the most challenging to detect. Such proteins evolve quickly and their periodicity may be rapidly hidden at sequence level. From a structural point of view, finding solenoids may be complicated by the presence of insertions or multiple domains. To the best of our knowledge, no automated methods are available to characterize solenoid repeats from structure.

Results: Here we introduce RAPHAEL, a novel method for the detection of solenoids in protein structures. It reliably solves three problems of increasing difficulty: (i) recognition of solenoid domains, (ii) determination of their periodicity and (iii) assignment of insertions. RAPHAEL uses a geometric approach mimicking manual classification, producing several numeric parameters which are optimized for maximum performance. The resulting method is very accurate, with 89.5% of solenoid proteins and 97.2% of non-solenoid proteins correctly classified. RAPHAEL periodicities have a Spearman correlation coefficient of 0.877 against the manually established ones. A baseline algorithm for insertion detection in identified solenoids has a Q_2 value of 79.8%, suggesting room for further improvement. RAPHAEL finds 1,931 highly confident repeat structures not previously annotated as solenoids in the PDB records.

Availability: The RAPHAEL web server is available with additional data from the URL: http://protein.bio.unipd.it/raphael/

Contact: silvio.tosatto@unipd.it

1 INTRODUCTION

Protein repeats contain tandem arrays of smaller structural motifs where, unlike most globular domains, folding is reduced to simple coiling and long range interactions are greatly reduced (Andrade, et al., 2001; Kajava, 2011; Kobe and Kajava, 2000). Repetitive proteins evolve quicker due to the intrinsically error prone process connected with the formation of repeating sequences (Buard and Vergnaud, 1994). 14% of all known protein sequences are strictly periodic and it was hypothesized that repeating sequences occur more frequently in eukaryotic proteins (Marcotte, et al., 1999). Repeating sequences were estimated to occur in around one in three human proteins (Jorda and Kajava, 2010; Kajava, 2011).

Classification of repeating proteins is usually achieved in terms of repeat unit length (Kajava, 2001; Kajava, 2011). The length of the repeating unit can be as small as one or two residues for different types of crystallites of unlimited size. At the other extreme are repeating units of entire domains ("beads on a string") with a typical repeating unit of over 50 residues. The middle ground comprises solenoid repeats with units of 5-40 residues. These are elongated structures containing α-helices and/or β-strands with a large distance between the N and C termini (Kobe and Kajava, 2000). There has been increasing interest in solenoid proteins over the years, especially their relevance in health (de Wit, et al., 2011; Kajava, et al., 2006) and for protein engineering applications (Main, et al., 2005; Stefan, et al., 2011). Solenoid proteins have also been shown to fold sequentially, one unit at a time, suggesting that the sequence contains all necessary information to determine the local fold (Kajander, et al., 2005). Understanding solenoid function and evolution passes through their classification from sequence and structural information, which are two different problems. Solenoid sequences evolve quickly while maintaining their fold, thereby hampering detection (Andrade, et al., 2001). Several sequence-based methods predicting tandem repeats from self-alignments have been developed over the years, including RADAR (Heger and Holm, 2000), TRUST (Szklarczyk and Heringa, 2004) and HHrepID (Biegert and Soding, 2008). Our previous work REPETITA uses a fast Fourier transform to specifically detect solenoids (Marsella, et al., 2009). In all cases there is still room for improvement, with the best methods still missing out many solenoids, especially with insertions. Generally speaking, solenoid repeats tend to be easy to spot through visual inspection in a molecular viewer. However, the manual search of hundreds or thousands of structures to determine if they are solenoid repeats or not is extremely time consuming and inefficient. Moreover, the definition of repeat length, i.e. repeating

*To whom correspondence should be addressed.

blocks containing similar residue numbers, and detection of breaks in the periodicity require objective measures.

Available structural databases such as PDB (Berman, et al., 2007) and CATH (Pearl, et al., 2003) store solenoid structures but do not provide feasible means for extracting them. Tools for discriminating protein repeat structures from globular proteins are rare in the literature. DAVROS (Murray, et al., 2002; Murray, et al., 2004) is perhaps the first method developed for this purpose. Unfortunately, it is no longer maintained. ProSTRIP (Sabarinathan, et al., 2010) is designed to find all similar structural repeats. It requires the selection of the repeat length and alignments from a set of alternatives, making it impractical for large-scale analysis. The Propeat database was designed by extracting recurring protein sub-structures, including internal repeats, but most of the structures contain only 2 repeating units (Shih and Hwang, 2004). A similar self-alignment approach is used by Swelfe to detect internal repeats in structures (Abraham, et al., 2008). When developing REPETITA we had to manually derive a dataset of 105 solenoids (Marsella, et al., 2009). Other sequence repeat prediction methods had similar problems in defining the dataset, e.g. in HHrepID the authors resort to structural self-alignment due to the lack of available tools for unbiased detection of solenoid repeats from structure (Biegert and Soding, 2008).

The present study aims to detect solenoid repeat structures using distance and periodic features extracted from the structural coordinates. The algorithm is efficient, has high discrimination power, can determine the repeat unit length and can find insertions which break the periodicity temporarily. The consequences of the algorithm are vast and here we tackle the large-scale extraction of repeats from CATH and the PDB. In addition to the novel data produced, a server is available (URL: http://protein.bio.unipd.it/raphael/) which can determine how periodic a structure is, the repeat length, periodicity and insertion plots.

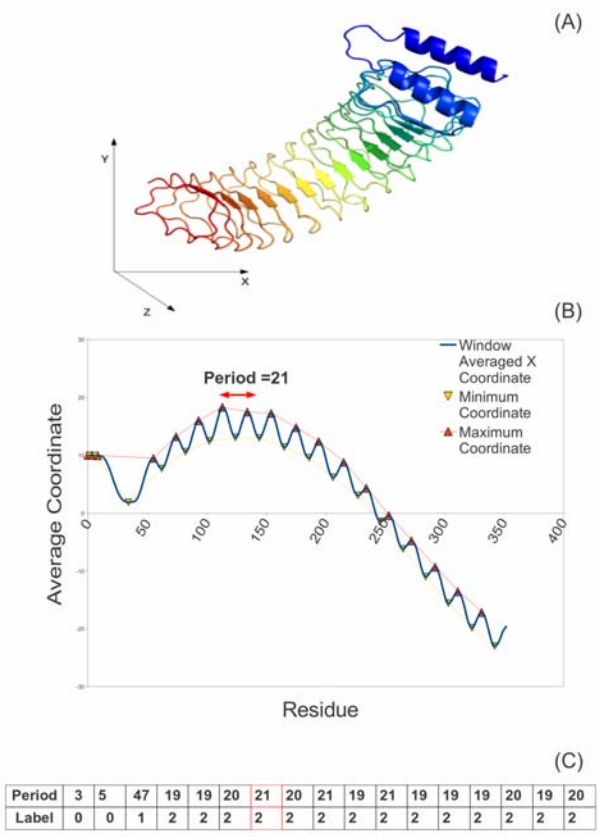
2 METHODS

Periodicity and distance measures are both important factors when considering a particular protein visually. The aim of our algorithm is to mimic the intuitive definition used by a manual curator, extracting these two factors from the three-dimensional coordinates of the structure. A set of parameters and filters are then derived to capture the essence of periodic spatial patterns. It should be noted that while signal processing methods such as fast Fourier transform can be used for repeat proteins, our previous experience suggests that they do not excel on biological data with intermittent insertions (Marsella, et al., 2009).

2.1 Periodicity

For each C-alpha coordinate (i.e. x, y and z) a profile/wave is generated, filtering by averaging the profile twice over a window for each coordinate profile. The first pass window size is 6 and the second pass window size is 3. Figure 1(b) shows an example of a coordinate profile derived from C-alpha coordinates. In order to avoid bias due to the initial orientation of the structure, the protein is anchored at a reference point by random translation and rotation. Anchoring is performed 200 times in order to build stable periodicity values, thus producing 3x200 profiles (i.e. 200 for each coordinate profile). A period is defined as the distance between consecutive local maxima on the profile curve (consecutive minima are also considered), see Figure 1(b). In order to score the periodicity, two observations are made: (i) frequent adjacent periods, termed window score, indicate solenoid proteins. (ii) frequent periods separated by rarely occurring periods, termed bridge score, indicate solenoid proteins.

Let $\theta_i = max_{i+1} - max_i$ $i=1,...,M-1$ be the period calculated between adjacent local maxima on the coordinate profile (similarly for minima) where M is the total number of local maxima. A labeled sequence is constructed from the sequence of periods $\theta_1, ..., \theta_{M-1}$ where $\theta_i \in \mathbb{N}$. A period θ_i is labeled with $k \in \mathbb{N}$ where k is the position of the first occurrence, i.e. $\theta_i \in [\theta_k - T, \theta_k + T]$, when scanning the periods from N- to C-terminus. T is the acceptable difference in residues between two periods that allows assignment of the same label. Otherwise a new label is attached. This labeling procedure results in a sequence of labels $L_1, ..., L_{M-1}$ representing periodicities found in the structure, which is the only information supplied to the window and bridge scoring functions described below. We found that $T=5$ produces optimal results, see Figure 1(c) for an example period sequence and the corresponding label sequence.



(C)

Period	3	5	47	19	19	20	21	20	21	19	21	19	19	19	20	19	20
Label	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Figure 1. Tagging the periods for the x coordinate of Leucine-rich effector protein YopM-a from *Yersinia pestis* (PDB code 1JL5). (a) The structure is shown colored from N-terminus (blue) to C-terminus (red). (b) the period as calculated from two consecutive local maxima on the averaged x coordinate profile. (c) the period sequence for the profile from (b) with the tagged label sequence below it. Notice how similar periods are assigned to the same tag. As this is clearly a solenoid protein, there are many identical tag labels adjacent to each other.

2.2 Functions

Let $C(L_i)$ be the number of times L_i appears in the label sequence. The window score is defined as:

$$W(L_i, L_j) = \begin{cases} 2C(L_i) & f \mid i-j \mid = 1 \text{ and } L_i = L_j \\ 0 & otherwise \end{cases} \quad (1)$$

The window score is positive only for identical adjacent labels (i.e. $|i-j|=1$), see Figure 2(a). Assuming we have two identical labels separated by an insertion of other labels, the bridge score penalizes the periods between them as follows:

$$B(L_i, L_j) = \begin{cases} 2C(L_i) - \sum_{j>k}^{j} C(L_j) & \text{if } L_i = L_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Figure 2(b) shows an example of the bridge score labeled sequence. The total periodic score for one coordinate and one random rotation and translation is:

$$Totalscore = \frac{pW^* + (1-p)B^*}{N} \quad (3)$$

where W^* and B^* are the final window and bridge scores (respectively) when processing the entire labeled sequence and N is the sequence length. Using a linear grid search on the training set, $p=0.49$ was found to be the optimal balance parameter. The total score for the entire protein is the average of the 3 coordinate profiles and the 200 random rotations and translations.

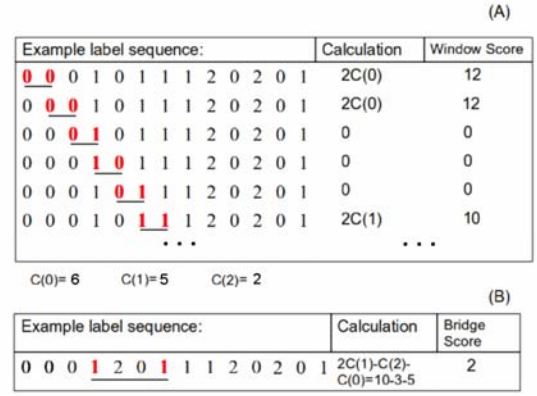


Figure 2. Example for the window and bridge scores. The positions being considered are shown bold faced in red and underlined. (a) The window score considers identical neighboring labels towards the total score. (b) The bridge score looks for identical labels separated by an insertion, here $i=4$ and $j=7$. See text for details.

2.3 Parameters and optimization

The variance among all the periods found within a structure should intuitively be another important factor for discriminating solenoids. Let $P = \{\theta_{1j}, \theta_{1j}, \theta_{1j}, ..., \theta_{Rj}, \theta_{Rj}, \theta_{Rj}\}$ be the set of periods for residue j for all R rotations and translations along each coordinate frame x, y and z. On this set, let F_{kj} be the frequency of period k found in P for residue j. We define the period matrix PM to be a 2D matrix of dimension $60*N$ with elements F_{kj}, $k=0,...,60$ and $j=0,...,N-1$, where N is the length of the protein and j is the index over residues. The cutoff was chosen to be the maximum allowed period since repeating units rarely exceed 60 residues for solenoid structures. Figure 2 shows the period matrix for a typical solenoid and non-solenoid protein. In order to measure the variation of periodicity within the entire protein, the standard deviation over all residues is calculated as:

$$SD = \sum_{j=0}^{N-1} \sum_{k=0}^{60} (F_j^{avg} - F_{kj}) \quad (4)$$

where F^{avg}_j is average frequency of column j in the period matrix. To complete the periodic information, we use the average period. Before calculating the average, set P is filtered by removing all outliers such that each period must be part of the interval $[P^{avg}-\sigma(P)/2, P^{avg}+\sigma(P)/2]$, where P^{avg} and $\sigma(P)$ are the average and standard deviation of all periods in P. This value is used to determine the solenoid periodicity length (termed P^* throughout the remaining sections).

Some observations about distance may be made through visual inspection of solenoid proteins: (i) solenoids, are usually elongated, (ii) contacting residues in solenoids should have low sequence separation relative to globular proteins and (iii) there should be regularity *in sequence* among the contacting residues (conversely there should be large variance for non-solenoids). Two residues are in contact if the distance between the C-alpha coordinates of both residues is less than a predefined threshold. To measure the distance in 3D space between the N- and C-terminus, the following distance is used:

$$MD = \min[d(i,j)] \quad \forall i \le 40,\ j \ge N-40 \quad (5)$$

where $d(i,j)$ is the distance between C-alpha atoms of residue i and j and N is the length of the protein. MD calculates the minimum distance between the first 40 residues and the last 40 residues. This value should give a good measure of protein elongation. Next the number of contacting residues at a sequence separation greater than 55 are calculated as follows:

$$NC = \frac{\sum_{i=0}^{N-1} \sum_{i-55>j>i+55}^{N-1} C_{ij}}{N} \quad (6)$$

Where $C_{ij} = 1$ if the distance between i and j is less than 6 Å, a value chosen because it closely resembles the hydrogen bond distance. The sequence separation cutoff at 55 was chosen since solenoid unit length rarely exceeds this value for solenoids and contacts between repeating units can therefore be counted by NC. In contrast, long range contacts are often present in globular protein structures (Kajander, et al., 2005; Main, et al., 2003).

Finally, the regularity of contacting residues in the sequence is measured by the variance of the Residue Wise Contact Order (RWCO) (Kinjo and Nishikawa, 2005), which for residue i is defined as:

$$RWCO_i = \frac{1}{N} \sum_{i-3>j>i+3}^{N-1} |i-j| C_{ij} \quad (7)$$

where $C_{ij} = 1$ if the distance between i and j is less than 15 Å. This cutoff was chosen to relax the distance strength and thus allow a sufficient count at all sequence separations. $RWCO_i$ is the sum of sequence separations between the i-th residue $\forall i=0,...,N-1$ and all contacting residues. The variance of this property will give a measure of how regular the sequence separation is for contacting residues. Let $RWCO^{avg}$ be the average and $\sigma(RWCO)$ be the standard deviation of $RWCO$. The final value used for discrimination of solenoids is the standard deviation of the set defined by:

$$\{RWCO: \quad RWCO_i \in [RWCO^{avg} - 0.6\sigma(RWCO), RWCO^{avg} + 0.6\sigma(RWCO)]\} \quad (8)$$

This gives a measure of the variance of the sequence separation between the contacts while ignoring extreme outliers.

The previously described periodic and distance features were combined using a Support Vector machine (SVM). The SVM C parameter was set to 0.02 and a simple linear kernel was used. The SVM produces a real number score with positive values indicating predicted solenoids and negative values indicating non-solenoids. The more positive the SVM score the more solenoid the protein should be.

2.4 Finding insertions

A simple baseline method is used to discriminate non-periodic residues or insertions in a structure from the core solenoid repeat. The main source of data is the variation of distances between residue j and $j \pm P^*$ where P^* is the calculated period. For each residue j we define the minimum periodic distance towards the N and C termini:

$$PD_j^N = \min[d(j, j-P^* \pm \Delta)] \quad \forall \Delta = 1,..,w \quad (9)$$
$$PD_j^C = \min[d(j, j+P^* \pm \Delta)]$$

$d(.,.)$ is the Euclidean distance between C-alpha atoms on residue pairs. PD^N_j and PD^C_j are used as a double pointed probe on the structure at residue j. First, it is important to determine the representative distance of a given period since proteins with the similar period do not necessarily repeat at the same distance. Given a protein of length L the raw set of periodic

distances $D = \{ PD_1^C, PD_1^N, ..., PD_L^C, PD_L^N \}$ is reduced to the subset $Df \subseteq D$ using the following conditions:

$$PD_j^{N/C} < T$$
$$PD_j^{N/C} \in \left[D^{avg} - \delta(D)/2, D^{avg} + \delta(D)/2\right] \quad (10)$$

where D^{avg} and $\sigma(D)$ are the average and standard deviation of D respectively. These conditions ensure the removal of extreme outliers and large non-meaningful distances (i.e. non chemical bonds). Let mDf denote the median of the set Df. It is in fact the variation from mDf which will measure the potential for non-periodicity. This variance profile is defined as follows:

$$VP = \sum_{j=1}^{L} \begin{cases} 1 & if \quad PD_j^{N/C} < mDf \pm \lambda \\ 0 & otherwise \end{cases} \quad (11)$$

when calculating distances boundary conditions, $|j+P^*\pm\Delta| \leq L$ and $|j+P^*\pm\Delta| \geq 1$, were implemented. The parameters w, T and λ were determined using a grid search on the training folds of the leave one out procedure. Values for the parameters were found to range $w \in [9,10]$, $\lambda \in [1.5, 2.0]$ and $T \in [12,15]$ depending on the training fold. Intuitively, the idea is to capture the maximum deviation of each residue from the median periodicity. This is a simple algorithm which may be further improved with more parameters and machine learning but should nevertheless provide a valid baseline for detecting insertions and repeat boundaries. Throughout the paper we will refer to insertions as non-repeated residues surrounded by solenoid repeats. Only the final experiment is shown in the manuscript, with results for the two partial optimizations shown in the Supplementary Material. All thresholds were found by maximizing Q_2 on the training sets.

2.5 Datasets

The training and test sets are based on publicly available data from the REPETITA paper (Marsella, et al., 2009). Briefly put, an initial set of 32 solenoid repeat proteins was taken from a previous review (Kobe and Kajava, 2000) and expanded using TESE (Sirocco and Tosatto, 2008) to find more protein domains in CATH (Pearl, et al., 2003) belonging to the same solenoid folds as the initial set. Choosing representatives with at most 35% pairwise sequence identity (i.e. CATH 'S' level) yielded a set of 105 solenoid domains. The set of non-solenoid protein domains was generated with TESE by randomly choosing x-ray structures with different topologies and no detectable sequence similarity (i.e. CATH 'T' level), for a total of 247 domains. The sets of solenoid and non-solenoid protein domains was randomly split into training and test sets, with the constraint that solenoid structures of low similarity fall in the same partition. It is worth mentioning that closed repeating structures such as beta-barrels or propellers are not included in the set and our algorithm does not consider these toroidal structures, but may still find their periodic signal.

In addition to the training and test sets, RAPHAEL was also benchmarked on CATH and PDB. The 'S' and 'O' level classifications, with a maximum sequence identity of 35% and 60%, were downloaded from the CATH website for the current version (v3.4). The PDB was downloaded as of July 1st 2011. DNA, RNA and protein chains with length less than 30 amino acids were removed. Each structure was separated into chains and reduced to 40% sequence identity using CD-HIT (Li and Godzik, 2006) with options -c 0.4 –n 2, creating a diverse set of 16,226 unique chains.

2.6 Performance measures

Throughout the manuscript, TP, FP, TN and FN are used for true positives, false positives, true negatives and false negatives respectively. Sensitivity and precision values are calculated for both periodic (P, positive class) and non-periodic residues/structures (N, negative class). The following measures are used: *sensitivity (P) = TP/(TP+FN)*, *precision (P) = TP/(TP+FP)*, *sensitivity (N) = TN/(TN+FP)*, *precision (N) = TN(/TN+FN)*. Accuracy is used as synonymous to sensitivity and Q_2 is the fraction of correctly predicted residues, i.e. *(TP+TN)/(TP+FP+TN+FN)*. The receiver operator characteristic (ROC) curve describing the overall performance at variable thresholds is plotted as TP rate vs. FP rate.

To compare RAPHAEL to existing methods, we chose the structure-based method Swelfe (Abraham, et al., 2008) and three sequence-based methods: REPETITA (Marsella, et al., 2009), TRUST (Szklarczyk and Heringa, 2004) and RADAR (Heger and Holm, 2000). Since Swelfe returns several alternative predictions, the best was considered in order to over- rather than underestimate its performance. Results for the sequence-based methods are taken from our previous publication (Marsella, et al., 2009). The comparison should be considered a baseline only, given that all of these tools (except REPETITA) are not explicitly designed for solenoid detection.

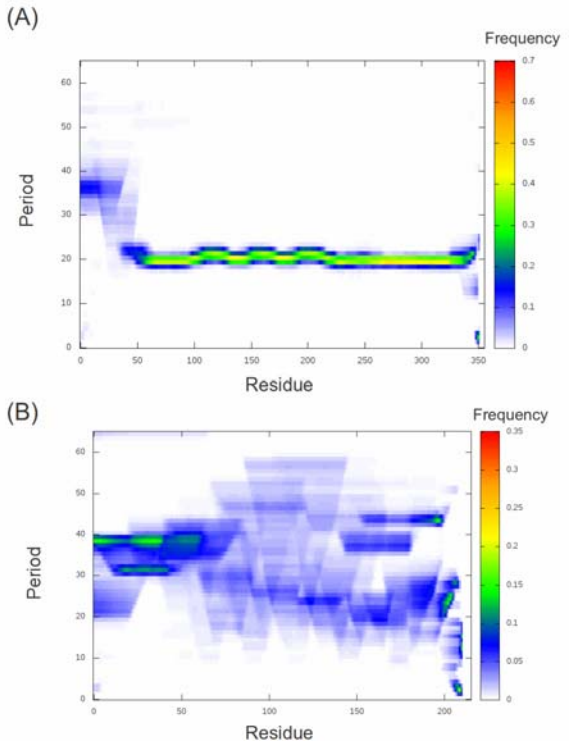


Figure 3. The period matrix for (a) Solenoid protein YopM-a Leucine-rich Effector protein from *Yersinia pestis* (PDB code 1JL5, as in Figure 1) and (b) sulfhydryl protease from the latex of the papaya fruit (PDB code 9PAP). Notice the variation of period frequency for 9PAP while 1JL5 periodicity appears regular.

3 RESULTS

3.1 Solenoid identification

In order to identify possible solenoids, RAPHAEL transforms the coordinates of the protein structure into a period matrix. An example for the transformation of a solenoid and a clearly non-repetitive structure can be seen in Figure 3. The solenoid structure produces a compressed signal of higher intensity, which can be used for detection. Several parameters were derived to take advantage of this information (see Methods). The performance at discriminating solenoids with the combined SVM score on the training set is shown in Table 1, while the individual parameters are reported in Supplementary Table S1. While the window

function is the most discriminating feature, the SVM combination improves performance by ca. 4% for solenoids and ca. 7% for non-solenoids, suggesting that different information is captured. Due to the limited number of training data and to be more statistically robust, we also tested the performance of a leave one out cross-validation. Here, training is performed with N-1 protein chains and testing with the remaining chain, while counting the results for all the testing examples (n=351). This produces results somewhere between the training and test sets, with only 7 false solenoids and 11 false non-solenoids for the entire dataset. Table 1 also shows how a stricter SVM threshold of 1.0 produces just 1 false solenoid, at the expense of losing 14 solenoids, thereby increasing positive precision to 98.9% compared to 93.1% for an SVM score of 0. In other words, an SVM threshold of 1.0 corresponds to very confident solenoid assignments.

Table 1. Accuracy on the training set and test set combining all six features through an SVM.

	TP	FP	TN	FN	Solenoids	Non-solenoids
Training	49	2	117	1	98.0	98.3
Testing	48	6	122	7	87.3	95.3
Leave one out > 0	94	7	240	11	89.5	97.2
Leave one out > 1	91	1	246	14	86.7	99.6

Results are shown for the method optimized on the training set (first two rows) and on the leave one out split (last two row) respectively. The latter are further reported at an SVM threshold of 0 and 1.

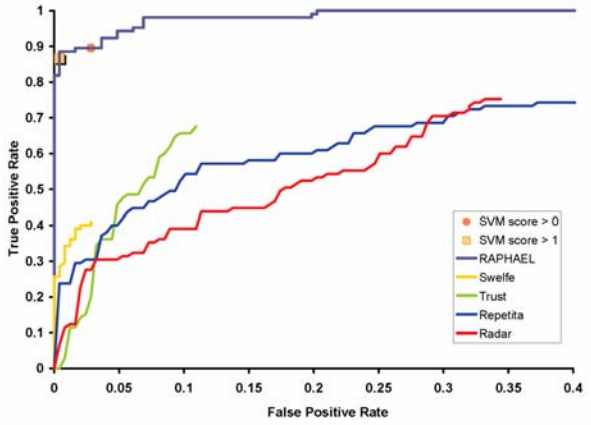


Figure 4. ROC curve on the combined training and test set. RAPHAEL trained using the leave one out split is compared to four other methods. The curve ends when a method does not produce further output, i.e. believes to have found all solenoids. Two SVM score thresholds are shown at 0 (orange circle) and 1 (yellow square) respectively.

A full ROC curve for the leave one out cross-validation is shown in Figure 4, also comparing to Swelfe and three sequence-based methods. Swelfe is not specifically designed for solenoids, but rather tries to detect internal repeats in proteins. It should also be emphasized that solenoid detection from sequence is more difficult and hence such methods can be expected to perform less well. The difference in ROC curve is nevertheless remarkable, with RAPHAEL detecting three times more solenoids than the other methods at low FP rates and the most difficult solenoid at an FP rate of ca. 20%. Table 2 shows the distribution of correct and incorrect classifications for leave one out training split in terms of CATH class. Interestingly it is the alpha-beta class which produces the most errors on solenoids (i.e. 7), suggesting that it may be somewhat more difficult to find solenoids when they have an alpha-beta mix. The datasets does not take into account class 4 (few secondary structures) as either negative or positive examples.

Table 2. Precision as a function of CATH class.

Class	TP	FP	TN	FN	Solenoids	Non-solenoids
Mainly α	40	0	59	2	100.0	96.7
Mainly β	31	0	16	9	100.0	64.0
Mixed α–β	23	7	165	0	76.7	100.0

Results calculated on the leave one out split. Precision results on solenoids and non-solenoids.

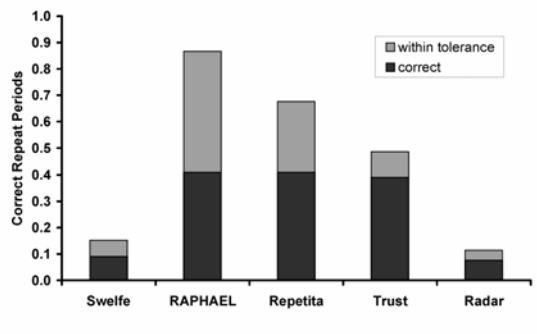


Figure 5. Detection of repeat periodicity for RAPHAEL and four other methods. See main text for details on the thresholds used to define the two levels of correctness.

3.2 Periodicity estimation

Once the presence of a solenoid has been established, it is important to define its periodicity, i.e. the length of the repeating unit. Supplementary Figure S1 shows a comparison of the periods determined from the period matrix (see methods) to a manual derivation from our previous work (Marsella, et al., 2009). The relationship is clearly linear, with an overall Spearman correlation coefficient of 0.877 indicating a strong relationship between RAPHAEL and the manually extracted repeat lengths. Upon inspection the small number of outliers exhibit period matrices which are highly variable and contain insertions and/or deletions. As expected it is difficult to determine the repeat length when insertions or deletions are present in the structure. Looking in more detail at the difficulty level of the solenoids, the hard (i.e. solenoids containing many insertions) cases have a Spearman correlation coefficient of 0.753 compared to 0.934 for the easy ones (i.e. solenoids with few or no insertions). Figure 5 shows a comparison of RAPHAEL to Swelfe and three sequence-based methods in terms of detecting the correct periodicity. Since the exact period in solenoids with insertions can be somewhat arbitrary, we allow two distinct levels of correctness. In analogy to our previous work (Marsella, et al., 2009), we consider one residue around the manually curated periodicity correct for all predictions. For sequence-based methods we also consider half or double the structural repeat as correct within tolerance. As structure-based methods (RAPHAEL and Swelfe) may be sensitive to insertions,

we allow five residues around the exact period as correct within tolerance. The effect of the window size on RAPHAEL predictions is shown in Supplementary Figure S2. As can be seen in Figure 5, RAPHAEL and the more accurate sequence-based methods have similar performances in recognizing correct periods. This is somewhat unexpected, but likely due to correct classification of solenoids without insertions where a clear sequence signal corresponds to the structural unit.

Table 3. Performance of simple insertion finding algorithm on leave one out cross validation.

Measure	All	Easy	Hard
Q2	79.8	83.4	74.1
Sensitivity (P)	95.5	95.7	95.1
Precision (P)	79.5	84.9	69.6
Sensitivity (N)	44.2	40.3	47.4
Precision (N)	81.2	72.7	88.5

The Q2, sensitivity and precision measures are shown after leave one out optimization for maximum Q2.

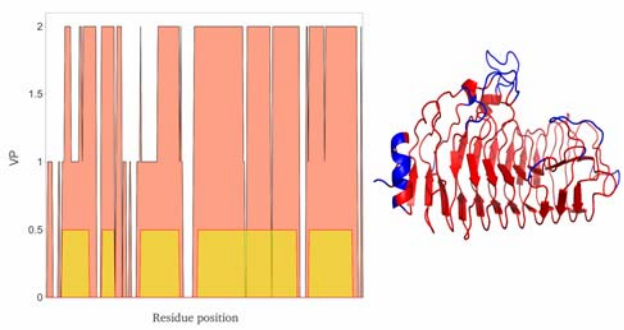


Figure 6. Example of insertions found for a β-solenoid. The variance plot (left) shows the score VP used to determine the location of insertions for Endopolygalacturonase (PDB code 1HG8 chain A). The yellow area indicates the true positions of the periodic residues (periodicity at 0.5, insertion 0).The same structure (right) is colored in red for residues assumed to be repeated and in blue for insertions. Notice how the algorithm identifies the core solenoid domain, while mispredicting some C-terminal residues.

3.3 Insertions

Given the performance in detecting solenoid proteins, the next question becomes whether the method is able to detect insertions for these proteins. To test this, every residue in each solenoid structure was annotated as either repeated or not. The Q_2, sensitivity and precision measures for the dataset are shown in Table 3. It should be emphasized that we are proposing a simple baseline algorithm with a few caveats. First of all, several solenoid structures are rather degenerate, prompting a somewhat arbitrary distinction between approximately repeated and inserted residues. Second, RAPHAEL tends to find clear insertions but finds it difficult to determine less obvious cases, as clear insertions disrupt the regular spatial pattern at the basis of our algorithm. Hence, smaller insertions can be underpredicted, while longer insertions are found but often reported as more disruptive than necessary. An example can be seen in Figure 6. To the best of our knowledge, this is the first time that an automatic classification of structural repeat insertions is attempted in the literature. It certainly also expands our view on the previously released REPETITA dataset (Marsella, et al., 2009).

3.4 Large-scale extraction of periodicity data

In order to test RAPHAEL, we decided to process large sets such as the PDB and CATH to generate datasets for future use. For this large scale search we trained the SVM on the combined data sets (105 solenoids and 247 non-solenoid domains).

RAPHAEL was used to detect solenoids on the entire CATH database at the S(35) and O(60) levels corresponding respectively to 35% and 60% maximum sequence identity. Supplementary Figure S3 shows the SVM score for all domains at S(35). Choosing this identity cutoff guarantees that solenoid domains are diverse at least at the sequence level but it can also be assumed to be true at the structural level. In total the algorithm considered 748 domains to be solenoids at this sequence identity cutoff (see Table 5 and Supplementary Figure S4). Obviously the higher the score the more expressed the periodicity should become. Upon visual inspection the better solenoids are represented by an SVM score greater than 1 (221 domains, see inset in Supplementary Figure S4). In order to find more solenoid domains which may be useful we also processed CATH with no sequence pair sharing 60% sequence identity. Using this less stringent cut-off the algorithm detected 1,156 CATH domains, with the distribution of SVM scores shown in Supplementary Figure S5. A list of CATH domains ranked by the solenoid score produced by the SVM can be found on the RAPHAEL website.

Table 4. Solenoid frequency in CATH.

	S(35)		O(60)	
Class	%	n	%	n
Mainly α	7.3	141	6.6	200
Mainly β	15.1	301	15.2	492
Mixed α–β	6.4	302	5.6	456
Few sec. struct.	5.2	4	7.3	8

The frequency (%) and absolute number (n) of solenoids found in each CATH class is shown for the S and O levels at 35% and 60% maximum sequence identity respectively.

Of course the extracted CATH domains will intersect with the set used for algorithm construction. Using a 50% sequence identity cutoff, we identify 696 proteins on S(35) and 1,089 on O(60) which are not homologous to our training data. At the more stringent SVM score greater than 1, the number of newly mined domains is 172 for S(35) and 245 for O(60). This has to be compared to the currently available list of 105 solenoid repeats (Marsella, et al., 2009).

In addition to CATH domains, we also processed PDB chains with RAPHAEL, finding 1,131 chains to be considered solenoids at 40% maximum sequence identity. A more confident set of 551 solenoid chains with SVM score greater than 1 was also generated. These numbers increase to 5,419 and 2,478 for the full PDB (see Table 5). It is interesting to note how the PDB analysis contains a comparatively higher number of confidently predicted solenoid structures than CATH. This might suggest the existence of solenoid structures outside the already known CATH superfamilies, although further analysis will have to be carried out to verify this

hypothesis.

To validate the results and verify the extent to which RAPHAEL detects previously unknown solenoid proteins, we have calculated the overlap of our predictions with PDB entries of proteins having the "repeat" (or "repeats") keyword in their respective header records. The results are drawn as a Venn diagram in Figure 7. It should be noted that PDB entries with the "repeat" keyword contains proteins that are not true solenoid repeats, e.g. the repeated spectrin or fibronectin domains. Nevertheless, RAPHAEL overlaps well with the PDB annotation but provides an even greater amount of novel automatic annotations. These can be useful for the automatic annotation of proteins by structural genomics consortia or the PDB itself.

Table 5. Number of solenoids found in CATH and the PDB.

	Structures	SVM > 0	SVM > 1
CATH S(35)	11,330	748	221
CATH O(60)	15,778	1,156	308
PDB 40	16,226	1,131	551
PDB full	74,020	5,419	2,478

The number of structures found with an SVM score greater than 0 and 1 is shown for the CATH S and O levels (i.e. 35% and 60% maximum sequence identity) as well as for the PDB dataset made non-redundant at 40% sequence identity and the full PDB.

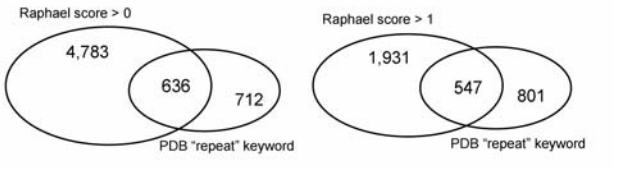


Figure 7. Venn diagram of RAPHAEL predictions and PDB repeat annotations. The predictions are shown at the SVM score cutoffs of 0 (left) and 1 (right) on the entire PDB. The PDB headers were scanned for the "repeat" keyword.

4 CONCLUSIONS

In this paper we have presented a novel method, RAPHAEL, for the accurate determination of solenoid repeats from PDB structures. The method quantifies repeat structures by mimicking visual interpretation by experts through various parameters. Combination in a SVM provides exceptionally accurate predictions, as tested on a previously published dataset. To the best of our knowledge, we show for the first time that our method is also able to broadly recognize insertions and repeat boundaries. Scanning the entire CATH and PDB databases provides hundreds or thousands of additional solenoid repeats, with automatic annotation for repeat regions. RAPHAEL was implemented in a new web-server based application for automatic repeat protein recognition. Due the importance of repeat proteins in both design and human diseases, we plan to use this method for systematic large scale analysis of protein structures, in order to improve our understanding of these peculiar proteins and their impact on organism evolution.

ACKNOWLEDGEMENTS

The authors are grateful to members of the BioComputing UP lab for insightful discussions.

Funding: University of Padova grant CPDA098382 and FIRB Futuro in Ricerca grant RBFR08ZSXY to S.T. G.M. is an AIRC research fellow.

REFERENCES

Abraham, A.L., Rocha, E.P. and Pothier, J. (2008) Swelfe: a detector of internal repeats in sequences and structures, *Bioinformatics*, **24**, 1536-1537.

Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) Protein repeats: structures, functions, and evolution, *J Struct Biol*, **134**, 117-131.

Berman, H., *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data, *Nucleic Acids Res*, **35**, D301-303.

Biegert, A. and Soding, J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency, *Bioinformatics*, **24**, 807-814.

Buard, J. and Vergnaud, G. (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90), *Embo J*, **13**, 3203-3210.

de Wit, J., *et al.* (2011) Role of leucine-rich repeat proteins in the development and function of neural circuits, *Annu Rev Cell Dev Biol*, **27**, 697-729.

Heger, A. and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences, *Proteins*, **41**, 224-237.

Jorda, J. and Kajava, A.V. (2010) Protein homorepeats sequences, structures, evolution, and functions, *Adv Protein Chem Struct Biol*, **79**, 59-88.

Kajander, T., *et al.* (2005) A new folding paradigm for repeat proteins, *J Am Chem Soc*, **127**, 10188-10190.

Kajava, A.V. (2001) Review: proteins with repeated sequence--structural prediction and modeling, *J Struct Biol*, **134**, 132-144.

Kajava, A.V. (2011) Tandem repeats in proteins: From sequence to structure, *J Struct Biol*, **179**, 279-288.

Kajava, A.V., Squire, J.M. and Parry, D.A. (2006) Beta-structures in fibrous proteins, *Adv Protein Chem*, **73**, 1-15.

Kinjo, A.R. and Nishikawa, K. (2005) Recoverable one-dimensional encoding of three-dimensional protein structures, *Bioinformatics*, **21**, 2167-2170.

Kobe, B. and Kajava, A.V. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures, *Trends Biochem Sci*, **25**, 509-515.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.

Main, E.R., Jackson, S.E. and Regan, L. (2003) The folding and design of repeat proteins: reaching a consensus, *Curr Opin Struct Biol*, **13**, 482-489.

Main, E.R., *et al.* (2005) A recurring theme in protein engineering: the design, stability and folding of repeat proteins, *Curr Opin Struct Biol*, **15**, 464-471.

Marcotte, E.M., *et al.* (1999) A census of protein repeats, *J Mol Biol*, **293**, 151-160.

Marsella, L., *et al.* (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform, *Bioinformatics*, **25**, i289-295.

Murray, K.B., Gorse, D. and Thornton, J.M. (2002) Wavelet transforms for the characterization and detection of repeating motifs, *J Mol Biol*, **316**, 341-363.

Murray, K.B., Taylor, W.R. and Thornton, J.M. (2004) Toward the detection and validation of repeats in protein structure, *Proteins*, **57**, 365-380.

Pearl, F.M., *et al.* (2003) The CATH database: an extended protein family resource for structural and functional genomics, *Nucleic Acids Res*, **31**, 452-455.

Sabarinathan, R., Basu, R. and Sekar, K. (2010) ProSTRIP: A method to find similar structural repeats in three-dimensional protein structures, *Comput Biol Chem*, **34**, 126-130.

Shih, E.S. and Hwang, M.J. (2004) Alternative alignments from comparison of protein structures, *Proteins*, **56**, 519-527.

Sirocco, F. and Tosatto, S.C. (2008) TESE: generating specific protein structure test set ensembles, *Bioinformatics*, **24**, 2632-2633.

Stefan, N., *et al.* (2011) DARPins recognizing the tumor-associated antigen EpCAM selected by phage and ribosome display and engineered for multivalency, *J Mol Biol*, **413**, 826-843.

Szklarczyk, R. and Heringa, J. (2004) Tracking repeats using significance and transitivity, *Bioinformatics*, **20 Suppl 1**, I311-I317.