Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Tesi di Laurea Magistrale in Ingegneria Informatica

Automatic classification of repeat proteins with Raphael

Laureando

Gabriele Tiberti

Relatore

Prof. Carlo Ferrari

Correlatore

Prof. Silvio C.E. Tosatto

13 Ottobre 2014

Anno Accademico 2013/2014

Abstract

Tandem repeat proteins form a distinct class of structures of great relevance due to their connection to neurodegenerative diseases and their functions in human health. The rapid evolution of this proteins hampers the detection of periodicity at sequence level, being structure more evolutionarily conserved than sequence new method are developed to detect periodicity in protein structures. Starting from RAPHAEL, a solenoid detection tool developed by Biocomputing Lab at University of Padua, a new method is devised. The principal aim of this work is the upgrade of RAPHAEL to achieve a deeper level of classification of tandem repeats. The new method, RAPHAEL2.0, demonstrates to obtain good performances in detection and prediction of different classes of repeats. With the new tool thousands of repeats are now correctly detected and classified.

Contents

List of Figures

List of Tables

Introduction

Repeat proteins are a distinct class of structures of increasing importance. Several classes have been defined by dividing proteins into their periodic repeat length. Periodicity can be hidden at sequence level due to rapid evolution, for this reason recognition of repeats is difficult at the residue level but should be easier at the structural level due to structure having more conservation.
Detection and prediction of repeats is an interesting objective in bioinformatics, here a new method for classification of repeats is presented. Starting from RAPHAEL [1], a tool developed in 2012 by the BiocomputingUp lab at University of Padua, a refinement and adaption for new tasks has been created. The new method, RAPHAEL2.0, has good performances in accuracy of predictions and it allowed classification of thousands of unassigned repeats.
The structure of this thesis follows this schema:

- The first chapter introduces the biological problem with an explanation of repeat proteins and their classification.
- The second chapter describes RAPHAEL, a tool for solenoid detection developed at BiocomputingUp lab.
- The third chapter shows material and methods utilised for RAPHAEL2.0, the new method presented.
- The fourth chapter illustrates and discusses results obtained from RAPHAEL2.0 and a comparison with the previous version.
- Last chapter is a dedicated discussion, final classification on unassigned repeats, conclusions and future perspectives for this work.

Chapter 1

Tandem Repeats Proteins

Repeat Proteins are a broad class of proteins characterized by a repetition in both primary amino acid sequence and tertiary structure (3 dimensional fold). In eukaryote and prokaryote genomes there was identified large quantities of repeated DNA sequences (in turn proteins via translation) and it has been proven they are involved in at least five neurodegenerative diseases (Huntington's disease [2], Macado-Joseph disease [3], Spinal muscular atrophy [4], Spinocerebellar ataxia [5] and dentatorubral-pallidoluysian atrophy [6]) and estimated to occur in about one third human proteins [7, 8]. In the past few years relevance of repeats have increased especially due to their importance for health [9, 10] and protein engineering [11, 12].
From the evolutionary point of view it is interesting to examine some of the repeated proteins' characteristics; it is well known that repeated regions in protein sequences are due to errors in the duplication process (with higher probability than normal mutations) [13]. This fact could suggest a quicker evolution in repeats [13], for this reason periodicity in sequence could be rapidly hidden despite the structure remaining evolutionarily conserved and thus easier to determine.
The length of the repetition can vary from a single residue to large regions of 100 or more residues with heterogeneous function and structure.
Recent analysis on repeats have highlighted the presence of repeated proteins especially in eukaryotes [14], with low levels of similarity with ancient organisms' proteins. Repeats then seems to be a recent evolutionary mechanism.

1.1 Classification of Repeats

The increasing number of known protein structures containing repetitive structural elements necessitates their classification to facilitate further un-

derstanding of their sequence-structure-function relationships as well as the evolutionary mechanisms.

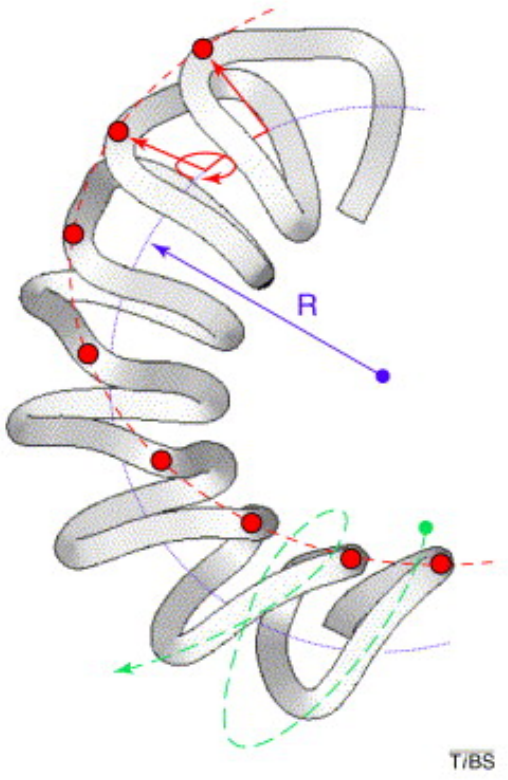


Figure 1.1: An example of repeat. Periodic distance is shown as the distance between red points. Curvature (magenta) is described by the radius (R). Handedness (green) describes the direction in which the polypeptide chain winds along the helical axis.

Figure 1.1 shows a typical repeat with period length indicated (red dashed line). An early classification of repeats was given by Kajava in 2001 [15] dividing repeats in four categories based on the repeat length or period; after ten years improvements of classification (with new 3D structures) a new classification was given [8].

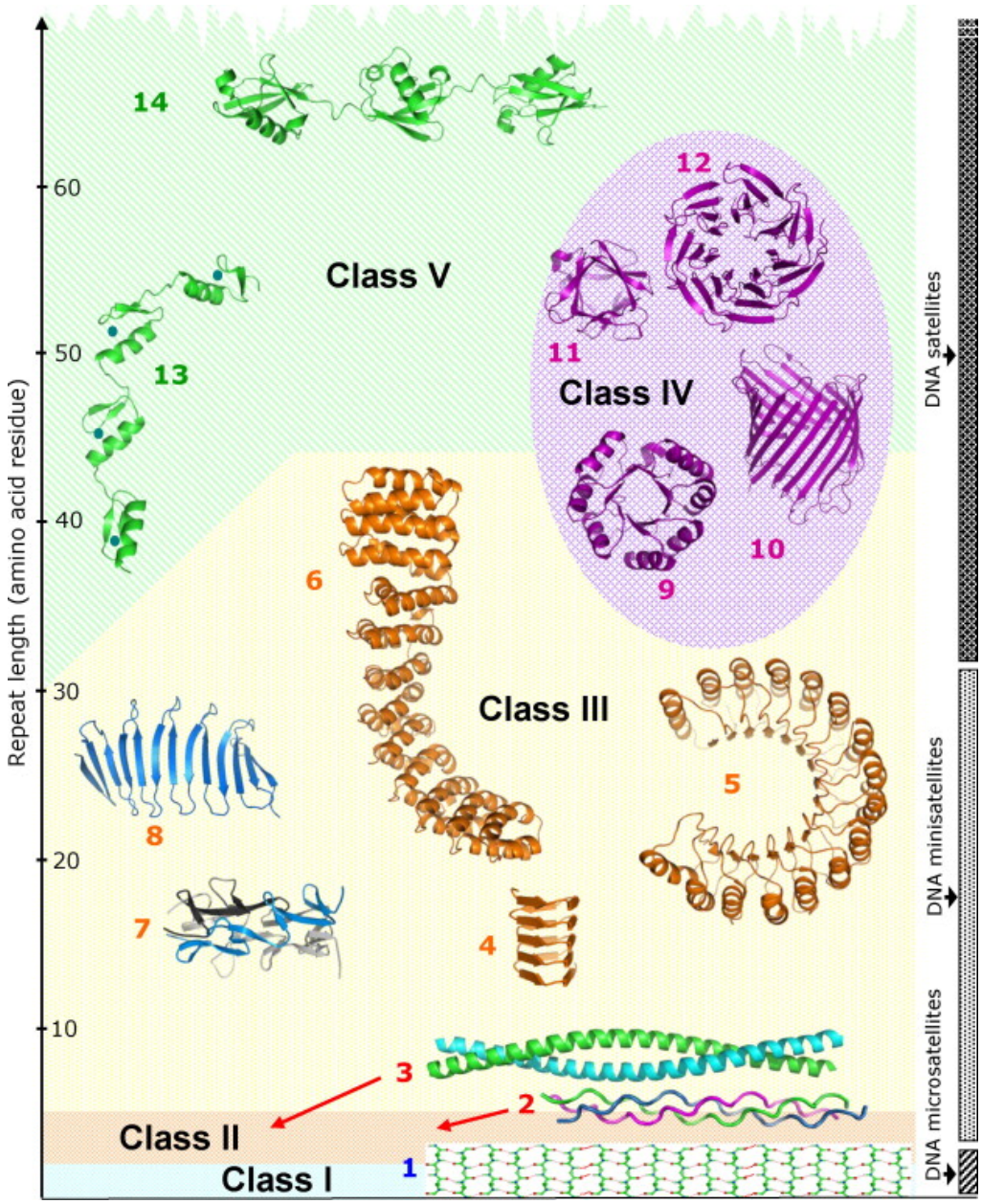


Figure 1.2: Structural classification of repeat proteins based on repeat length or period.

Repeats can be now classified into five classes (Figure 1.2):

- **Class I**: (Crystalline Aggregates of Unlimited Size)
  This class includes proteins and peptides with 1 or 2 residue-long re-

peats that form different types of crystallites of unlimited size which are harmful to living organisms.
In proteomes, regions with such repeats are predominantly hydrophilic and have high probability to be unfolded.Structures of these proteins are nowadays rare in the PDB. From experimental studies they normally form different type of crystallites which are composed of α-helices, the β-sheet structures, polyproline II helices or other regular conformations. Experimental evidence demonstrate these kind of proteins are linked to neurodegenerative disorders, including Huntington's disease [16].

- **Class II**: (Fibrous Structures stabilized by interchain interactions)
  This class includes two major fibrous structures that are collagens and α-helical coiled coils.
  Collagens have chians with extended polyproline II conformation and assemble into a triple helix.The α-helical coiled coils are characterized by several repeats of the same short structure ($\approx$ 7 residues) where apolar residues are spaced at intervals of 3-4 residues. Each chain folds into an α-helix wrapped around the axis of the coiled coil structure.
  It is possible to consider that class II structures have repeats ranging from 3 to 4 residues.

- **Class III**: (Elongated structures where repetitive units require one another to mantain structure)

  1. *Solenoid structures*:
     Structures with repeats of 5-40 residues are dominated by solenoid proteins; they are based on solenoidal windings of the polypetide chain. Solenoids tend to have elongated structures in contrast with the majority of globular proteins. The repeating structural unit of the solenoid proteins is an individual coil which consists of 12-45 amino acids, equivalent of one to four segments of secondary structure connected by loops. There are purely α-helical or β-structural solenoids in addition to units with a mixture of secondary structure elements.
     Most of the known solenoid structures have longer repeats of about 20-25 residues that correspond to a complete turn of the coil. Solenoids, as class II proteins, require one another to mantain the structure; in contrast with fibrous structures they can have a stable structure alone without forming an oligomer.
  2. *Non-Solenoid structures*: In the past few years new 3D structures different from solenoidal fold have emerged:

- **Trimer of β-spirals**:
  Characterized by long central β-strands that hold the trimer together through interchain hydrogen bonds and a short peripheral β-strands stabilizing the structure (Figure 1.3 A).
- **Single layer antiparallel β-structure**:
  Present in *Borrelia burgdorferi*; the repeat length of these structure is 23-26(Figure 1.3 B).
- **Antiparallel β-structure folded along long axis**:
  Conversely to the other antiparallel β-structure this protein is folded along the longest axis as a "burrito"-like shape filled by its amino acid side chains [8].Repeat length is in range 24-37 residues (Figure 1.3 C).
- **Spiral β-hairpin staircase**:
  This is a single-stranded β-fibrous fold with β-hairpins as 20-24 residue repetitive structural units(Figure 1.3 D).

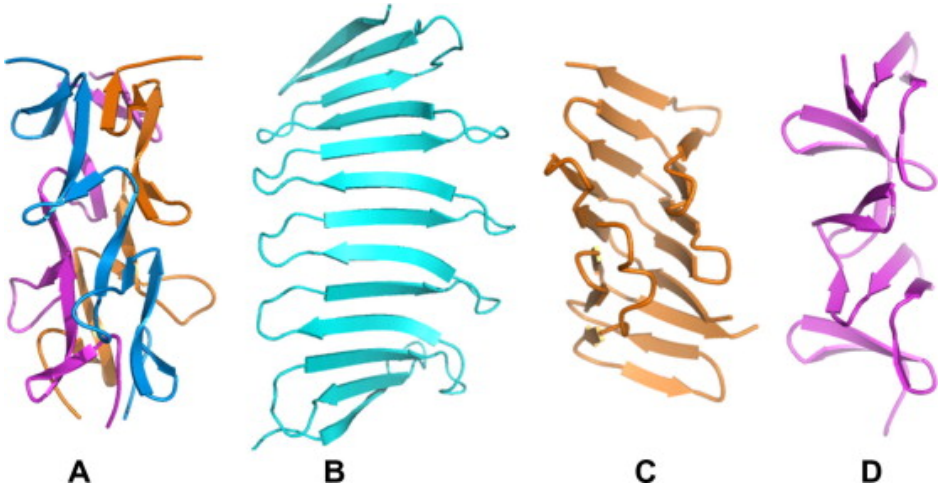


Figure 1.3: Non-solenoidal structures of Class III. (A) Trimer of β-spirals. (B) Single layer antiparallel β-structure. (C) Antiparallel β-structure folded along long axis. (D) Spiral β-hairpin staircase.

- **Class IV**: (Closed Structures)
  The structures mentioned above do not have any restrictions inherent to axial growth; in contrast, protein from this class have fixed number of repeats due to their "closed" structures. The repeat lengths of this class overlap with both class III and V (further described) structures. TIM-barrels, for instance, can be considered as "closed" α/β solenoids.

Most class IV structures, however, are not solenoids.
Due to numerous changes occurred during the evolution,numerous barrels do not have well-recognized repeats in their amino acid sequences.

- **Class V**: (Beads on a String)
  This class of structure includes repetitive units already able to fold independently into stable domains. Typical size is over 50-60 residues. The overall structure is mainly composed by globular domains ("beads") stabilized with either disulfide bonds or metal ion. A classic example (Fig.1.2) is the Zinc-finger domain, a common DNA-binding motif stabilized by zinc metal ions.
  Recently new types of class V structures has emerged; normally are elongated and semi-rigid proteins with tight connections between repetitive units. Spectrin-like repeats (Fig.1.4.C) composed by α-helical bundle with 3-5 helices aligned to the molecule axis (100-130 residues). Other semi-rigid examples from class V are several β-structural domains of circa 60 residues existent in a wide variety of complement and selectins (Fig.1.4.A-B).

1.2 Identification of tandem repeats in protein sequence

The continuously growing amount of proteomic data and the important health and functional roles of repeats has led to increasing efforts in the development of methods for protein repeat recognition. Protein tandem repeats are frequently not perfect, containing a number of mutations (substitutions, deletions and insertions) triggered by evolution, and some of them cannot be easily identified. In order to solve this problem different algorithms and techniques have been developed relying mainly on protein sequence.
We can subdivide them into five general types of methods:

1. **Fast Fourier Analysis**:
   This approach finds periodic amino acid sequences using Fourier Transform Analysis, it does not rely on prior knowledge about the data representing an *ab initio* method for repeat recognition [17, 18]. It is mostly specialized in detection of long arrays of tandem repeats without insertions or deletions.

2. **Short String Extension Algorithms**:
   Specialized in detection of relatively short repeats ($\leq$ 15-20 residues).

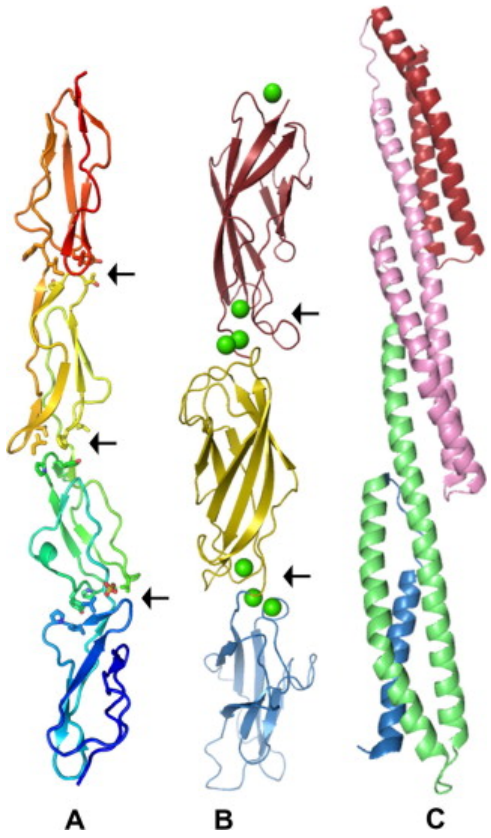


Figure 1.4: Examples of beads on a string structure. Like solenoids they are elongated but repetition has a much larger period (repeat unit length). (A) Four β-structural domains of Complement Control Protein modules. (B) An elongated structure of cadherin repeats. (C) Spectrin-like repeats representing an α-helical bundle.

They have $\mathcal{O}(n)$ complexity and, therefore, well suited for large-scale search of repeats (XSTREAM [19], T-REKS [20]).

3. **Sequence self-alignment**:
   Efficient for detection of arrays of long repeats (more than 15 units) but they frequently fail to identify short repeats(RADAR [21], TRUST [22]). The $\mathcal{O}(n^2)$ time complexity prevent the use of this kind of method for large scale analysis.

4. **Hidden Markov Models based on *a priori* generated repeats alignment or sequence profile**:
   The power of this method depends on the quality of sequences used

to create the HMMs or profiles. This approach is one of the best in detection of long and strongly imperfect tandem repeats, however, it requires an *a priori* generated alignment of repeats (Pfam [23], SMART [24]) which are currently scarce due to data unavailability.

5. **HMM-HMM or profile-profile comparisons**:
   An HMM is constructed from a multiple alignment of proteins that are homologous to the analyzed one for the sake of finding sub-optimal alignments of this HMM with itself (HHrepID [25]).
   Another approach concerns the comparison of sequence profile against Discrete Fourier Transform (REPETITA [26]) or stationary wavelet packet transform of sequences (WAVELET [27]). Also this kind of method turn out to be slow relatively slow and inappropriate for automated large scale analysis.

Identification of tandem repeats based on protein sequence tend to be a difficult problem because of their high divergence. As a result there is a lack of repeat sequence data to construct accurate algorithms for repeats detection. Focusing on structure should bring the problem to an easier level given that structure is more evolutionarily conserved. The hope being once an accurate structure based predictor is developed this will fill the gap in the sequence data scarcity. The remainder of this thesis is concerned about structural repeat detection giving improvements over existing algorithms.

Chapter 2

Identification of tandem repeats in protein structure

Periodicity and distance information are important to detect repeats visually (e.g. using a structure visualization tool) but this is slow and infeasible for detecting many. Algorithms should also benefit from these distance and periodic features. Currently repeat structure detection algorithms are very scarce, to the best of my knowledge only two exist: in [28] an algorithm based on distribution of suboptimal structural alignments of continuous fragments is developed, the second method is Console [29] based on modularity of contact map. Improvement in term of accuracy of detection and speed of these algorithms is one of the next challenges to the researchers given that repeat structure data is growing fast compared to power of computers [8].To enhance the previous structure detection methods, the BioComputing Up group developed RAPHAEL, a new approach for detection and recognition of repeated regions in proteins; this new method aims to find repeated structures using distance and periodic features extracted from the structural coordinates of proteins.

2.1 RAPHAEL

RAPHAEL [1] uses a geometric approach mimicking manual classification and producing several numeric parameters which are optimized for maximum performance.
It is created to solve three kinds of problem of increasing difficult:

- Recognition of solenoid domains.
- Determination of periodicity (i.e. its repeat unit length).

- Assignment of insertions (i.e. non-periodic parts).

RAPHAEL is very accurate and finds 1,931 repeat structures not previously annotated as solenoids in the PDB records.
Periodicity and distance measures are key factors when considering a particular protein visually; the algorithm basically extract these features allowing a machine learning approach in order to discriminate solenoidal proteins.
The next section will explain in detail the measures taken from protein structures.

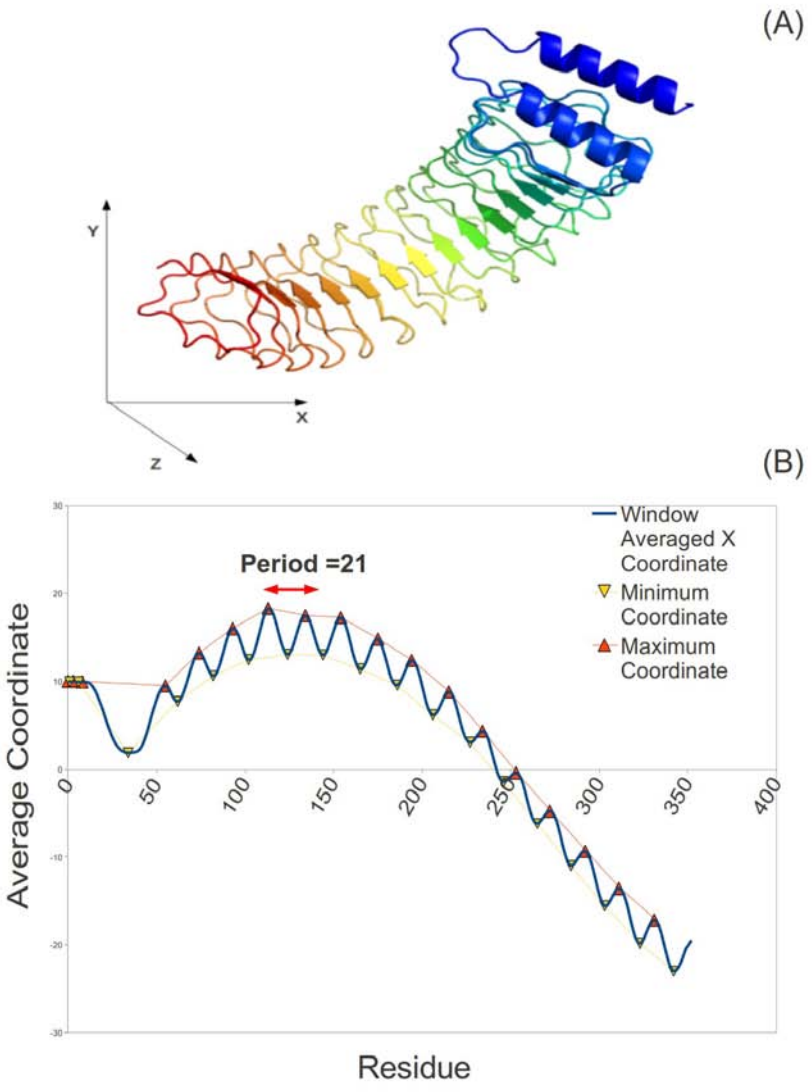
2.1.1 Periodicity features

Periodicity is the first characteristic to investigate in studying repeats.
Two observations should be made before beginning to study this factor:

- Frequent adjacent periods (taken as the distance between points on the protein) indicate repetition in structure.

- Frequent periods separated by rarely occurring periods(insertions) indicate repetition.

A complete description of the RAPHAEL periodic features is beyond the scope of this thesis, for the interested I suggest reading [1]. For simplicity here I describe the features using an example. Figure 2.1 shows the periodic profile (Figure 2.1 B) and its corresponding label sequence (Figure 2.1 C) for protein 1a9nC. The profile in figure 2.1 C is calculated from the x, y and z 3-dimensional coordinates and the distance between the maxima and minima on the profile are used to calculate the periods. The periods in turn are used to define the label sequence (see Figure 2.1 C). Let $\delta_i = max_{i+1} - max_i$ be the period, a labelling sequence is created comparing adjacent periods. A maximum threshold is chosen and the same label is attached if consecutive periods do not exceed this value, otherwise another label is added. This procedure produces a sequence of labels (Fig.2.1 C) the only information needed to score the periodicity. For example, in Figure 2.1 C, the period label "2" is the most frequent and it corresponds to a period of 20 $\pm$ 2.



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 2.1: Labelling function. PDB ID: 1a9nC. (A) Protein 1a9nC. (B) Profile wave calculated on the x coordinates. (C) Period sequence for the profile calculated in (B) with the tagged label sequence.

Let C(L_i) be the number of occurrences of L_i in the label sequence, two functions have been defined:

1. **Window Score**

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

   Window score is positive if identical labels are adjacent (i.e. $|i-j|$ =1)

2. **Bridge Score**

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

   Bridge score penalizes identical labels separated by an insertion of other labels

Once computed these values, the final periodic score is given by:

$$TotalScore = \frac{pW^* + (1-p)B^*}{N} \tag{2.1}$$

Where W^* is the final window score and B^* is the final bridge score calculated on the entire label sequence, N indicates the sequence length. An example of window and bridge scores calculation is shown in figure 2.2.

Next, another important feature coming from periodicity is caught measuring the variance among all the periods. In order to better discriminate this difference the Period Matrix *(PM)* is built. Let $P = \{\theta^x_{1j}, \theta^y_{1j}, \theta^z_{1j}, ..., \theta^x_{Rj}, \theta^y_{Rj}, \theta^z_{Rj}\}$ be the set of periods for residue j for all R rotations and translations along each coordinates x, y and z. F_{kj} is defined to be the frequency of period k in the set for the residue j. The PM is defined as a 2D matrix (60*N) with elements F_{kj}, $k = 0, ..., 60$ and $j = 0, ..., N-1$ where N is the length of the protein. The threshold of 60 is chosen considering that repeated units rarely exceeds 60 residues. Having this matrix it is possible to calculate the variation of periodicity on the entire protein using the standard deviation of the PM:

$$SD = \sum_{j=0}^{N-1} \sum_{k=0}^{60} (F_j^{avg} - F_{kj}) \tag{2.2}$$

where F_j^{avg} is the average frequency of the column j in the period matrix. The last periodicity feature considered is the average period calculated after filtering all the values in order to remove outliers (i.e. average of P after removing outliers).

(A)

Example label sequence:	Calculation	Window Score
0 0 0 1 0 1 1 1 2 0 2 0 1	2C(0)=12	12
0 **0 0** 1 0 1 1 1 2 0 2 0 1	2C(0)=12	24
0 0 **0 1** 0 1 1 1 2 0 2 0 1	0	0
0 0 0 **1 0** 1 1 1 2 0 2 0 1	0	0
0 0 0 1 **0 1** 1 1 2 0 2 0 1	0	0
0 0 0 1 0 **1 1** 1 2 0 2 0 1	2C(1)=10	34

(B)

Example label sequence:	Calculation	Bridge Score
0 0 0 **1** 2 0 **1** 1 1 2 0 2	2C(1)-C(2)-C(0)=10-3-5	12

Figure 2.2: Tagging and Scores. (A) shows an example of the window score. (B) shows a bridge score example.

2.1.2 Structural features

From the experimental process of classification of repeats some observation can be made:

- Most of the repeated proteins are elongated.
- Contacting residues should have low sequence separation.
- There should be regularity in sequence among the contacting residues.

Considering these informations new features are calculated. The elongation (MD) is measured considering the 3D euclidean distance between N-terminus and C-terminus. More robustness is given considering the minimum distance among the first and the last 40 residues.

$$MD = min[d(i,j)], \forall i \leq 40, \forall j \geq N - 40 \tag{2.3}$$

where $d()$ is the euclidean distance in 3D, i and j are the residue position. Next, the ratio of contacts at long sequence separation (NC) is calculated

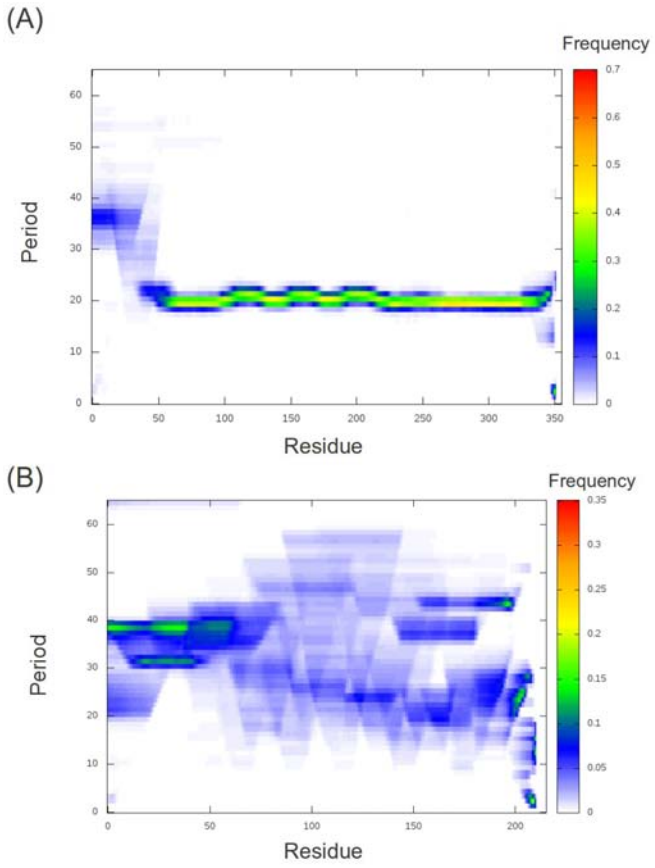


Figure 2.3: Examples of Period Matrix. (A) shows a repeat where the standard deviation is high at the N terminus. (B) shows the PM of a globular domain with high variance an thus SD will be high.

considering contacts occurring at sequence separation greater than 55 (repeats unit length rarely exceeds this value).

$$NC = \frac{\sum_{i=0}^{N-1} \sum_{i-55>j>i+55}^{N-1} C_{ij}}{N} \tag{2.4}$$

where $C_{ij} = 1$ if the distance between i and j is less than 6 Å (seemingly close to hydrogen bonds distance). Long Range contact are usually present in Non Repeats, so this score is particularly helpful in the discrimination process.

Figure 2.4 shows the contact map C_{ij} of a repeat protein indicating very low number of long range contacts.

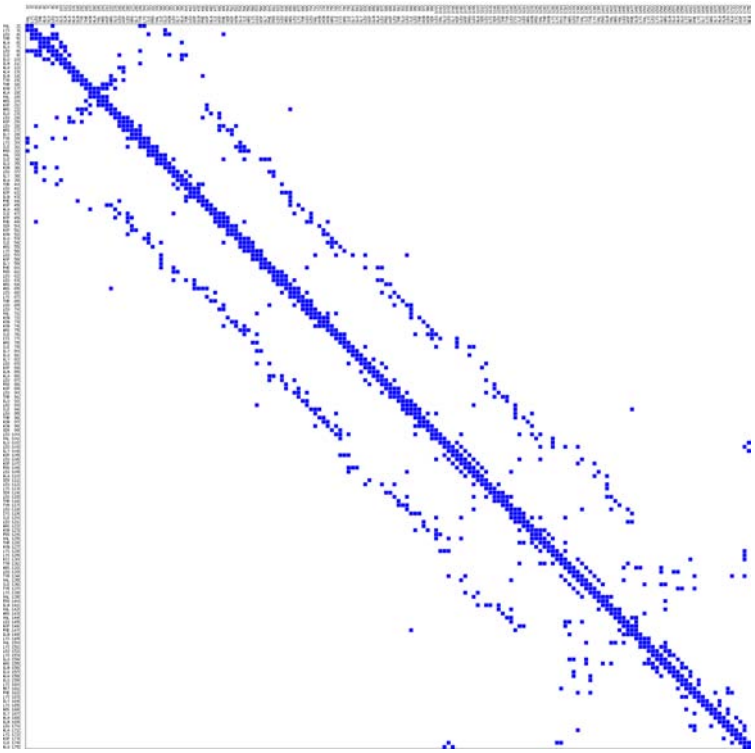


Figure 2.4: Example of Contact Map. PDB ID: 1a9nC. The element C_{ij} is coloured in blue if the distance from residue $i-th$ to residue $j-th$ is less than 15 Å .

The last feature considered is the regularity of contacting residues in the sequence; this measure is taken by calculating the variance of the Residue Wise Contact Order (RWCO), defined for the $i-th$ residue as:

$$RWCO_i = \frac{1}{N} \sum_{i-3>j>i+3}^{N-1} |i-j|\, C_{ij} \tag{2.5}$$

where $C_{ij} = 1$ if distance from $i-th$ to the $j-th$ residue is less than 15 Å . In this way it is possible to sum all the sequence separations starting from each one of the N residues to the respective contacting residues. Regularity of sequence separation for contacting residue should be given by the variance of this measure; let $RWCO^{avg}$ be the average and $\sigma(RWCO)$ the standard deviation of RWCO, the score used is defined by:

$$RWCO : RWCO_i \in \lfloor RWCO^{avg} - 0.6\sigma(RWCO), RWCO^{avg} + 0.6\sigma(RWCO) \rfloor \tag{2.6}$$

This range of values is taken in order to remove possible outliers (e.g. those produced by insertions).
These periodic and structural features are combined using Machine Learning approaches automating the discrimination between repeat and non repeat proteins.

2.2 Improvements to RAPHAEL

Despite the good results obtained by this algorithm improvements are needed to expand capabilities and enhance performances of this software; the aim of this thesis is to improve the tool for:

- Reduce false positive rate of prediction.
- Expand detection and classification at class level (Raphael was designed to detect only **Solenoids**).
- Adapt the tool for new data retrieved.
- Create a package to mine all the available data.

In the next chapter I will show the new method proposed to upgrade the performance of the currently available package.

Chapter 3

Materials and Methods

This chapter will present the materials used for design and implementation of the new tool. First of all a description of the dataset employed for the analysis is given, after that the new features for the software and the automatic approach will be presented.

3.1 RepeatsDB

RepeatsDB is a database of annotated tandem repeats protein structures [30]. Recognition and classification of repeats is a difficult problem; a structural approach has been devised and proposed by the Biocomputing Up group with RAPHAEL, based on the results obtained from that tool a new database for tandem repeats annotation has been created. It is the first large database of tandem repeats structures. The structures are diverse and range in their functional importance. The aim of RepeatsDB is to offer a central resource for classified and annotated repeat structures.

3.1.1 RepeatsDB description

The database was envisaged around the work of Kajava [8], repeats are divided into five classes (see Figure 3.1). A deeper level of classification is also proposed allowing a more precise subdivision of proteins in subclasses mainly based on repeat length and secondary/tertiary structures (Fig.3.1).

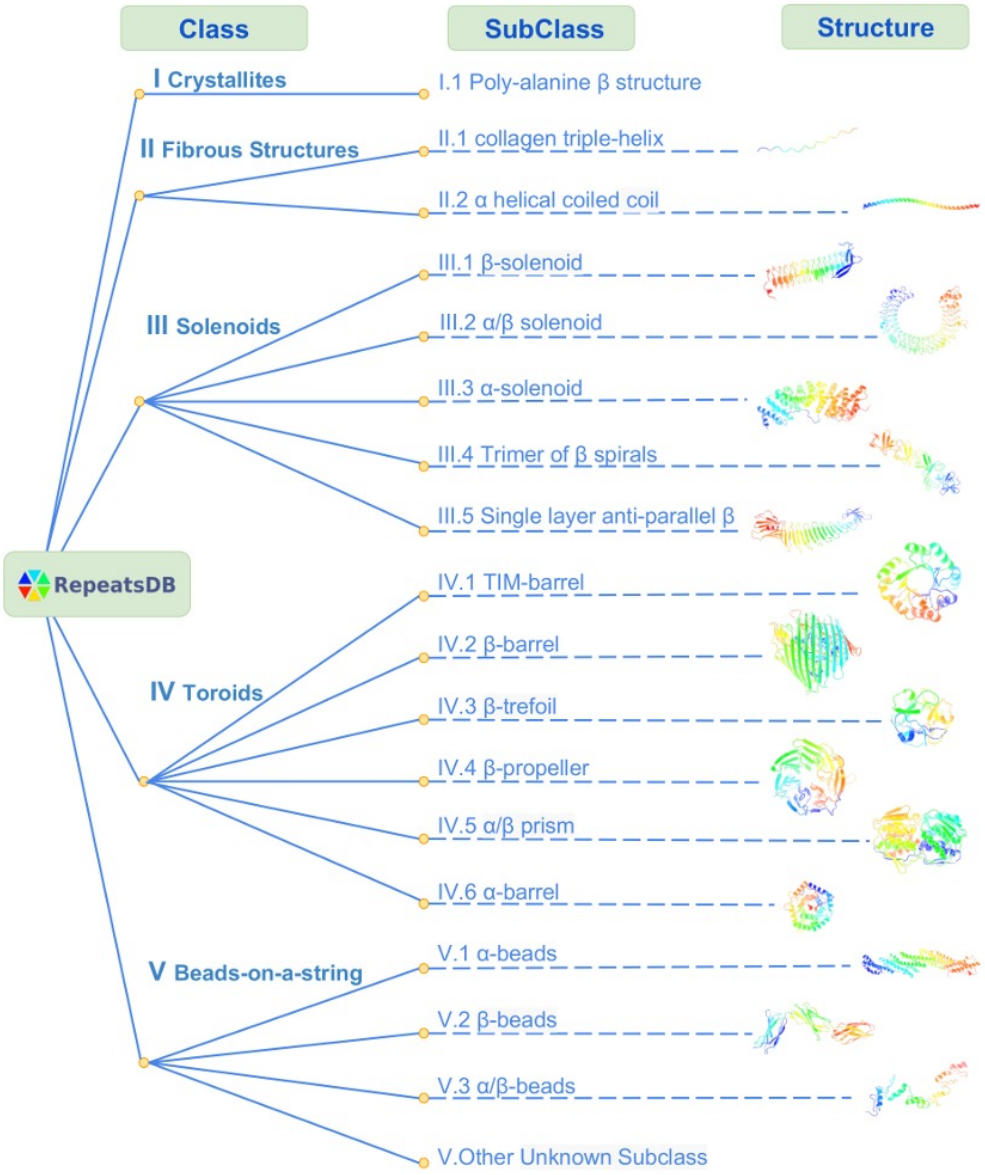


Figure 3.1: RepeatsDB structure. The subdivision in class and subclass is given by each column. The Structure field shows an example of each subclass.

Classification is conducted at different levels of accuracy; from RAPHAEL results an initial set of repeat candidates are stored as "predicted". After that, a two-level process of manual curation has been conducted; the first level ("manual classified") classifies a candidate into structural repeat class

and subclass using a group of manual curators in a similar vein to crowd classification of galaxies (see `www.galaxyzoo.org/`), the second level ("detailed") gives more precise informations about repeat units' lengths and positions, repeated regions and insertions.
The annotation process undergoing the manually classification is validated by consensus, more specifically for the first level annotation at least 75% of curators had to agree on the same decision, otherwise the entry is excluded and placed on a reserve list for future annotation,for the second level the threshold decrease to 65%. For controversial cases an expert decided the final annotation based on alternative proposals. Proteins at the second level of manual annotation ("detailed") have been used, using sequence similarities calculated with sequence alignment algorithm. In order to retrieve unclassified proteins annotated at a "classified by similarity" level (40% identity threshold and a minimum 80% coverage on the classified proteins).

3.1.2 Repeat Recognition

Using the previously mentioned tool (i.e. RAPHAEL) the entire Protein Data Bank has been tested by calculating the features mentioned in chapter 2. All the features have been combined and classified with an SVM model finding a set of $>$ 10,000 proteins. With the above-mentioned annotation framework predicted repeats are the starting point and each are then assigned into classes manually. This manual assignment is time consuming and one of the objectives of this thesis is to automate this class assignment. Figure 3.2 shows the number of repeat regions divided by class; different colours show the different annotation level for each class. As previously mentioned examples of Class I are not present in the PDB until now. Class III and IV are the most represented, they include the most studied types of repeat proteins.
Figure 3.3 shows the deeper classification at subclass level; it is important to notice that numerous subclasses are under-represented instead of others that have a good number of entries; for instance, α-solenoid (Cl.III.3) is the biggest and most present subclass in the database, Class IV.1 and IV.4 mostly cover the examples for closed structures. The distribution is quite skewed and this may be problematic using an automatic approach. However, we have the largest set of classified repeats to date and using this data as a Machine Learning training set should produce accurate algorithms at class level. The new method, presented in the next section, aims to create a new framework for repeat protein recognition at the class level shown in Figure 3.2.

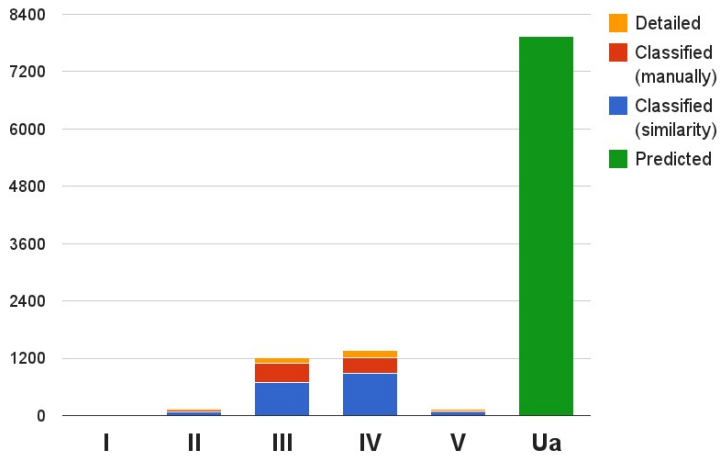


Figure 3.2: RepeatsDB, Entries by Class. Class I has yet to be filled. Class III and IV are abundant. For a description of the classes see Chapter 1.

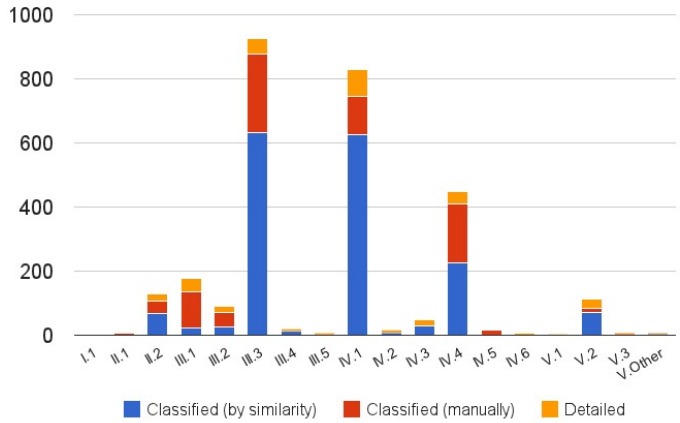


Figure 3.3: RepeatsDB, Entries by Subclass. Class III.3 (α-solenoid), class IV.1 (TIM-barrel) and class IV.4 (β-propeller) are the most represented in the database.

3.2 RAPHAEL2.0

The purpose of this work is to revise the existent RAPHAEL creating a new tool able to detect repeats, as expected, but also to assign for each protein the actual class. With the recent development of RepeatsDB this is now possible since for the first time we have a quality data source with different repeat types.
Looking at the various structures present in the five classes should be essential to identify structural features able to distinguish a single class from each other. This is the key idea to be effective in the evaluation and separation process between these different kind of repeats.
As previously mentioned, RAPHAEL was designed for solenoids' recognition, the most representative repeat in our class III; adapt the tool for a wider recognition is the main challenge of this work. As seen before, class I is not present in RepeatsDB so no mention will be given for this particular class.
Like the development of RAPHAEL in this work a visual examination of the different structures easily allows the extraction of some structural characteristics peculiar for each class:

- **Class II**: Entirely composed by coiled coil and long helix.

Figure 3.4: PDB ID: 1d7mA. Due to the simplicity and regularity of the motif is an attractive system to explore the principles of protein folding and stability.

- **Class III**: Mostly solenoidal structures, normally elongated.

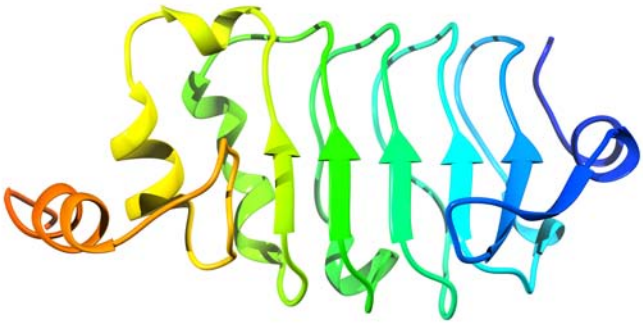


Figure 3.5: PDB ID: 1a9nC. Spliceosomal involved in pre-mRNA maturation.

- **Class IV**: Closed structures are defined by their N and C terminus being in close proximity.

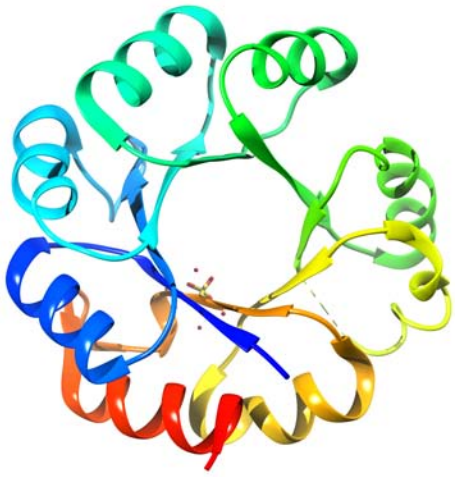


Figure 3.6: PDB ID: 1dl3A. Phosphoribosylanthranilate isomerase (PRAI) connected with activation and trasformation of the basis composing the RNA.

- **Class V**: Probably to most difficult structure to define and recognize, several repeated regions linked by flexible regions.

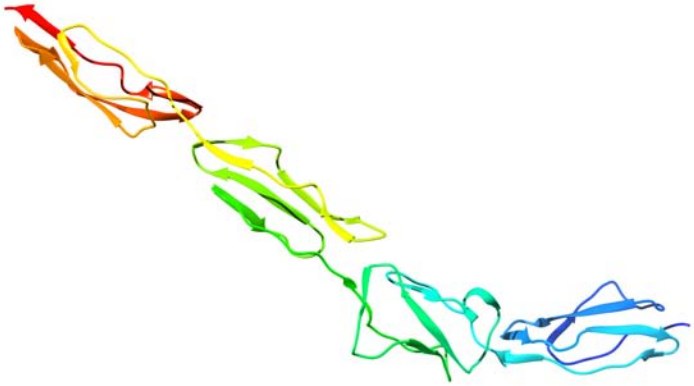


Figure 3.7: PDB ID: 1ojvB. Human regulatory complement associated with the immunological response in human.

Recalling the feature described in the previous chapter it is easy to notice that the first version of RAPHAEL already has some specific feature suitable for this new kind of analysis.
Elongation is measured as in formula 2.3, this may be helpful to discriminate closed structures between all the others. Class III should also be well recognized being the software designed for this purpose. In order to improve detection and prediction for structures classification new features are added.

3.2.1 Periodicity Features

RAPHAEL already predict the average period for each protein, starting from this feature it is possible to measure the 3D distance between residues separated by an average period. This value should give a good measure about internal elongation (Class III and Class V).
As previously mentioned the Period Matrix is computed calculating (eqn. 2.2), the matrix could also be used to calculate the frequency of every occurring period; regular repeats tend to have rare high-recurring periods (Fig.3.8), irregular repeats show different numerous low-recurring periods (Fig.3.9).

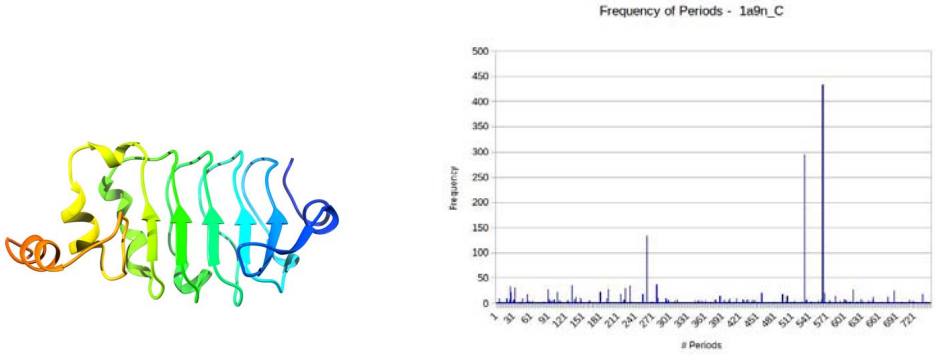


Figure 3.8: PDB ID: 1a9nC. Protein image and frequency of periods.

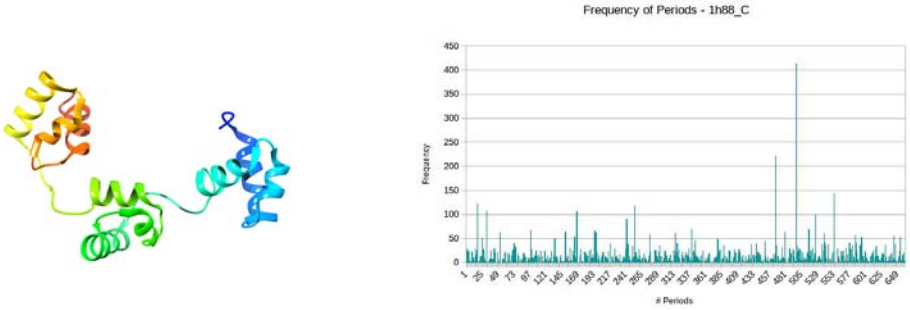


Figure 3.9: PDB ID: 1h88C. Protein image and frequency of periods.

Considering this frequency distribution, a new score is defined; let $f(P_i)$ be the frequency of the $i-th$ period and f^* be the maximum frequency amongst the period matrix:

$$PS = \sum_{i \in Periods} \frac{f(P_i)}{f^*} \tag{3.1}$$

This score should give good information about regularity of proteins.In particular helping to discriminate classes, for example in recognition of class V containing irregular periodicity.

3.2.2 Structural Features

Several structural features were calculated in RAPHAEL. More specific features are introduced here with the idea to calculate short range contacts present in a sphere having radius equal to the average periodic distance

(Fig.3.10).
For each residue the number of short contacting amino acids are counted. The standard deviation of this measure is utilised as another feature.

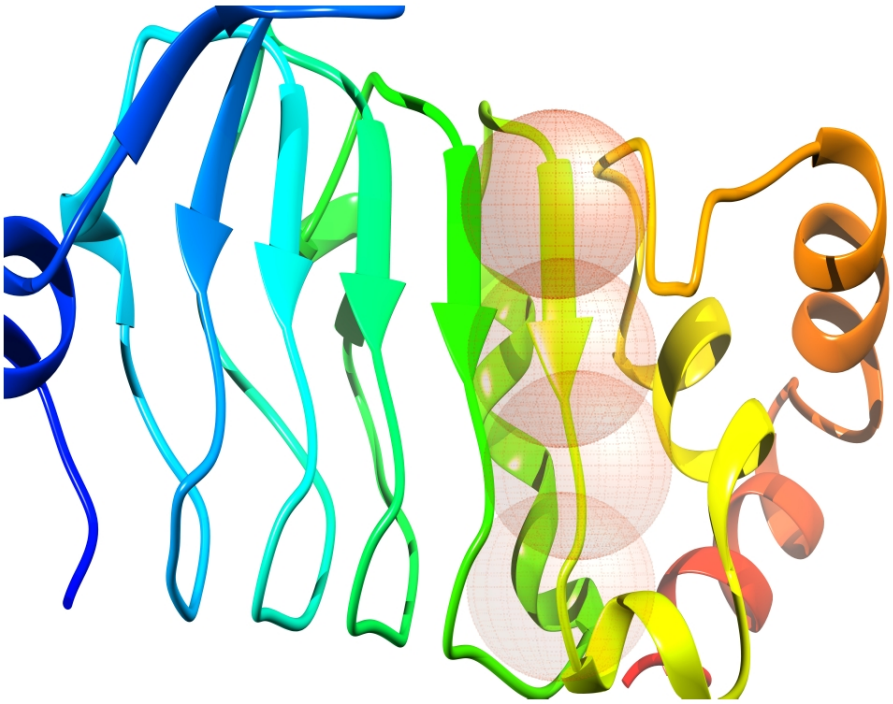


Figure 3.10: PDB ID: 1a9nC. Short Contacts measured in a sphere. The variation of residues lying in this space should give a good information about periodicity of the protein.

Solenoids, collagens and closed structures should present approximately the same number of residues lying in a 3 dimensional space surrounding the residue considered; beads on a string (Class V) shows a stronger variation in this sense, links connecting the "beads" give more variation to this measure.

3.2.3 Secondary Structure

Aiming to classify different types of proteins has strong support provided by analysing the secondary structure (*SS*). From three dimensional structure it is possible to calculate the secondary structure sequence to collect new features. The most popular method to calculate *SS* is **DSSP** [31], this algorithm operates by pattern-recognition of hydrogen bonds and geometrical

features extracted from x-ray coordinates. A sequence of secondary structure can be calculated with each amino acid having one of the three classes: α-helix, β-strand or coil. From the *SS* sequence we measure the periodicity of α-helices and β-strands in terms of average and standard deviation in this case, periodicity represents the distance between middle points of two adjacent identical elements of secondary structure, either helices or sheets. Figure 3.11 shows a simple example of the *SS* periodicity.

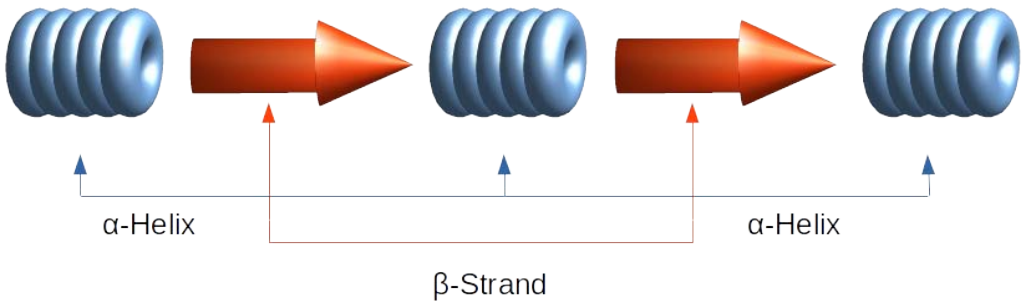


Figure 3.11: Periodicity in Secondary Structure. Periodic distances are taken between middle points of two adjacent identical elements of secondary structure.

Regular periods of secondary structure elements should be ideally found for proteins belonging to class II, III and IV. Class V proteins, as previously seen, have numerous periodicity breaks due to the "beads" having non-repeating nature.
Another important feature derived from the 3D coordinates is Surface-to-Volume Ratio (SVR). SVR is defined as the ratio between the accessible surface area and the volume of the region [32]; it measures the compactness of a protein and the tendency to be exposed to the solvent. For example, proteins of class II should exhibit higher scores (Fig.3.12) and different classes should display different solvent exposure character (Class III example in Fig.3.13). The figures show proteins coloured by solvent accessibility score of each residue where red means highly accessible, blue means low solvent accessibility.

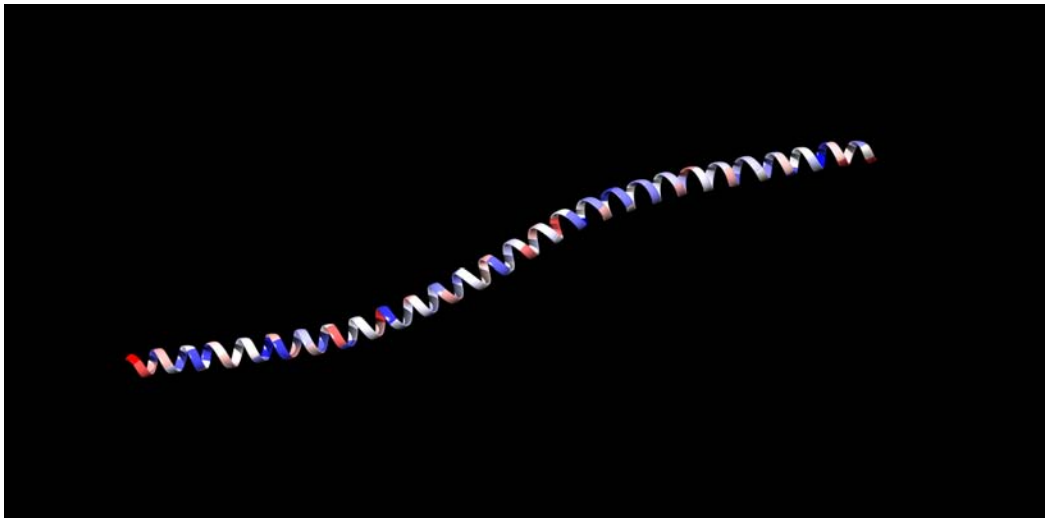


Figure 3.12: PDB ID:1d7mB. The protein is coloured by solvent accessibility score (Red means high accessibility, blue low accessibility).

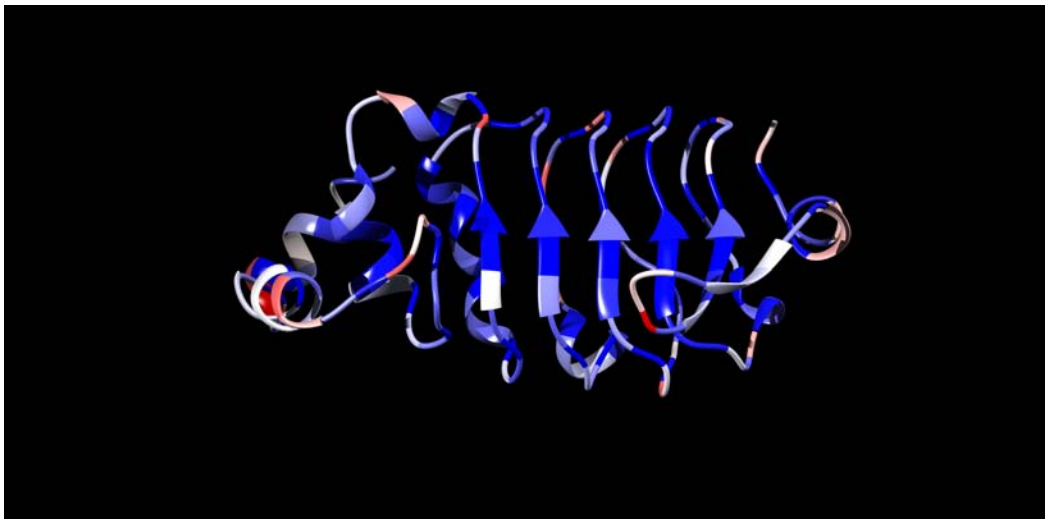


Figure 3.13: PDB ID:1a9nC. The protein is coloured by solvent accessibility score (Red means high accessibility, blue low accessibility).

SVR, calculated as the average of the solvent accessibility on all residues present in the *PDB* file, is also calculated with the **DSSP** algorithm. In 2011 a new version has been developed by Hekkelman [33], this version is the one used for this work.

3.3 Summary of the features

After having explained the features created for the new method we can briefly summarize all the features involved in RAPHAEL2.0:

- **RAPHAEL**:
  - Periodicity Score
  - Variance of periods in Period Matrix
  - Average Period
  - Elongation
  - Long range contacts
  - Regularity of contacting residues
- **RAPHAEL2.0**:
  - Average Periodic Distance
  - Periodic Score
  - Short Range Contacts
  - Secondary Structure Periodicity
  - Surface-to-Volume Ratio

All these features will be combined with the automatic approach explained in the next section.

3.4 Automatic Approach

Machine Learning algorithms are widely used in bioinformatics due to their capabilities to discover and learn hidden patterns, sometimes even difficult for expert humans experts to explain. These approaches seem also to be very robust to noisy and missing data, a main characteristic of biological data. Machine Learning is used in this thesis combining the so far described features using a Support Vector Machine.
An SVM classifier is a machine learning approach to learn separation of different classes by a maximum margin hyperplane (Fig.3.14).

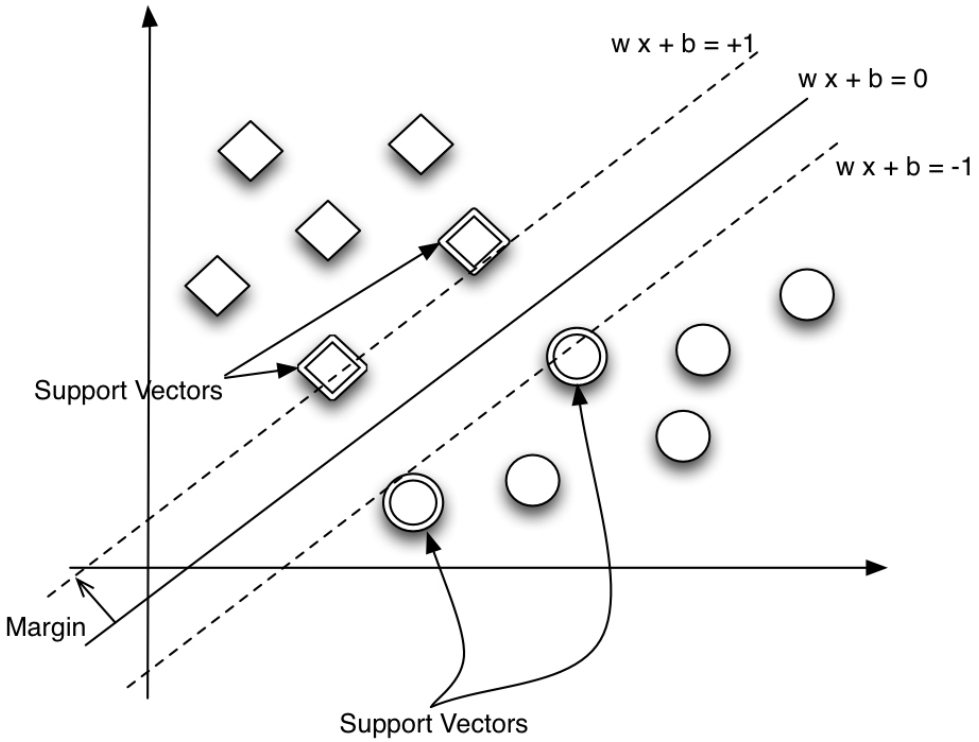


Figure 3.14: SVM example, Support vectors and hyperplanes are highlighted.

This margin is defined by the algorithm support vectors [34]. SVM is mainly designed to directly perform binary classification. As previously mentioned the aim of this work is to discriminate four different classes. In order to achieve this objective we split the multiclass classification problem into several binary classification experiments. The approach selected is the *One versus All* where for each training experiment one class is labelled as positive and all the others as negatives. This approach leads to the creation of four different models used for the classification of new examples. The SVM^{light} library was chosen. SVM^{light} is an implementation of Support Vector Machine developed in C by Thorsten Joachims [35].The library consists of two modules `svm_learn` and `svm_classify`; `svm_learn` reads the training set, learns the separation hyperplane and writes the related classification model. `Svm_classify` classifies new examples according to the models learned. For each class a model has been created considering the manually curated ("detailed" and "manually classified") proteins present in RepeatsDB. In the next section the performance of the models are assessed.

Chapter 4

Results and Discussion

The new features devised for the second version of RAPHAEL were added to the existing tool with the hope of reaching a finer level of classification. The first task is to refine and improve the performances of RAPHAEL, a new model for repeat protein discrimination is assessed. After that, a more precise explanation about all the other tests on the individual class are given. For every test we chose to compute a 5-fold Cross Validation in order to avoid possible overfitting and to assess the performance fairly.

4.1 Repeats versus Non Repeats

In 2012 RAPHAEL was tested on a small set of 242 solenoidal domains. Conversely a negative set of 342 globular was taken to build a classification model for solenoid proteins.
Improvements to data quality and quantity via RepeatsDB allowed us to expand the learning dataset to 1,081 repeated domains (the new positive dataset). Proteins with non repeated regions chosen as the lowest SVM scores from RAPHAEL were taken as negative examples. This dataset was initially composed by 162,416 proteins (i.e. *pdb* chains).
In order to enforce some diversity on the negatives and reduce the size of the dataset the CD-HIT Algorithm [36, 37] has been used. CD-HIT is a clustering algorithm working on FASTA sequences. Setting a maximum identity threshold it clusters the data giving a compact representation of relevant sequences. For this experiment we picked an identity threshold of 40% reducing from 162,416 initial sequences to 13,096 centroids, the new negative set. Table 4.1 shows the distribution of the positive repeats split into the four classes used in this work.

Type	Class	# Examples
Repeats	Class II	64
	Class III	516
	Class IV	449
	Class V	52
Non Repeats		13,096

Table 4.1: Repeats Distribution.

4.1.1 Results

The first test is conducted by comparing the accuracy of the two methods to discriminate repeats and the efficacy of the new features introduced for the new method. All the detailed and manually classified proteins from each different class are labelled as positive examples with the class distribution in Table 4.1. An example of Non Repeat protein is given in figure 4.1.

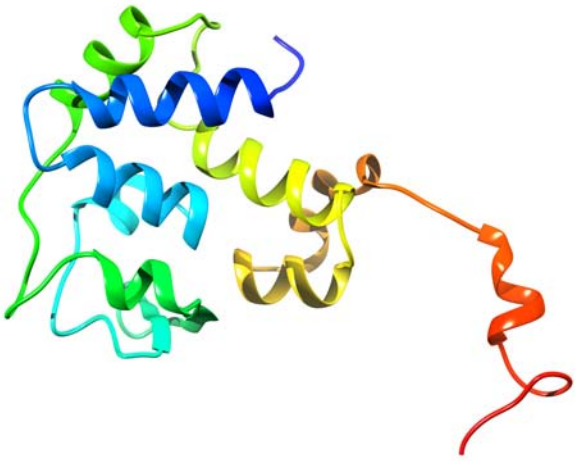


Figure 4.1: Example of Non Repeat protein showing very little periodicity among its structure.

The previous version, RAPHAEL, appears to be more precise in repeats classification having slightly better results especially in terms of Precision-

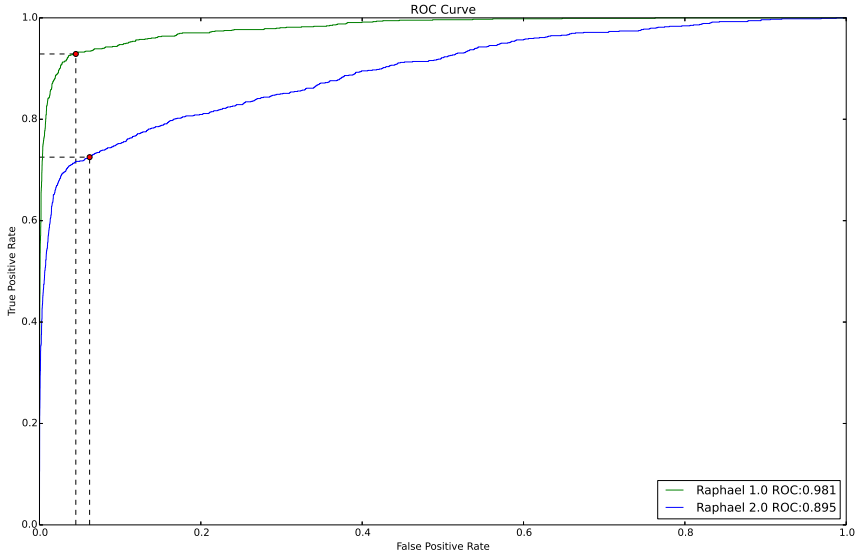


Figure 4.2: Comparison of the two methods. ROC Curve.

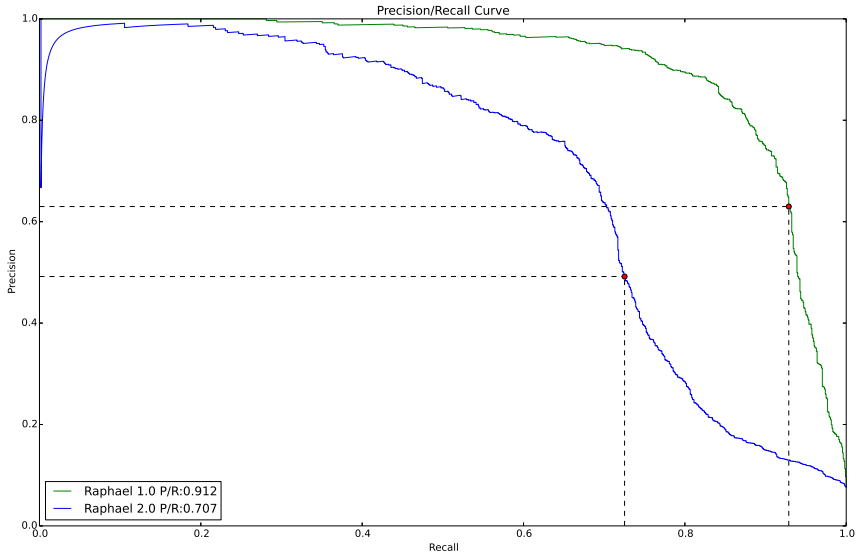


Figure 4.3: Comparison of the two methods. Precision-Recall Curve.

Recall and Receiver Operating Characteristic (ROC) AUC. From this test it is possible to hypothesize that secondary structure and advanced geometrical features are not helpful in discrimination of repeats from globular ones. For this reason RAPHAEL should be chosen for this task (See section 5.1). New features are taken in order to correctly discriminate between different classes so it is reasonable the new method is not able to handle different kind of proteins as a unique positive set. Considering the data distribution it is also possible to notice that solenoids are roughly the 50% of the positive examples favouring better performances of the first method. These results could also be biased by the initial data gathering, in fact repeats and non repeats were classified based on the SVM score predictions from RAPHAEL features. This test still produce a more accurate and general model for repeat detection compared with the one produced in 2012.

4.2 Class Test

Class detection is the main target of this thesis, this task will introduce a new level of repeat recognition not yet achieved. To this end several binary classification problems are defined selecting positive examples by proteins from a unique class and using all the examples from other classes as negative training set. A more detailed view on the data distribution of each class will be given in order to deeply understand the strong class imbalance caused by the nature of the data. In this section we compare the performances of the two methods inspecting if an improvement over RAPHAEL is obtained. For each class model a 5-fold Cross Validation test were conducted, from the ROC curves we chose the 5% FPR thresold (marked as a red point in each graph) to assure low detection error.

4.2.1 Class II

Class II Subclass	Subclass Name	# Examples
II.1	collagen triple-helix	5
II.2	α-helical coiled coil	59

Table 4.2: Class II Distribution.

Class II is basically composed by two types of structures where almost the entire positive set consist of α-helical coiled coil (Table 4.2), collagens are

rarely present in our dataset. Having a sufficient number of examples and considering the substantial difference with all the other structures we expect to perform well in this task. ROC and P-R curves compare the performance of the models obtained.

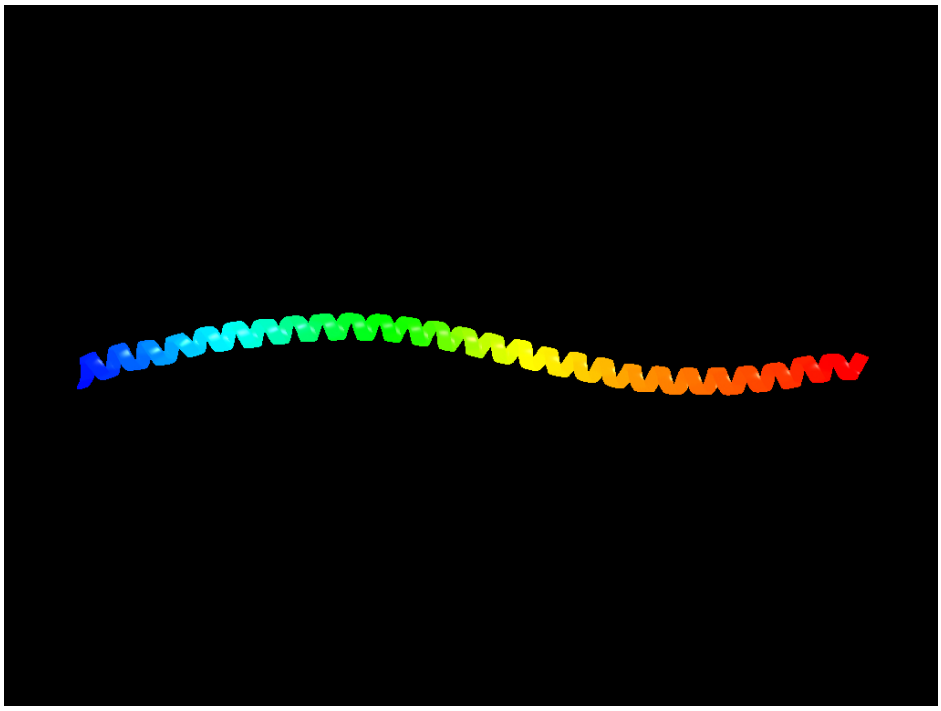


Figure 4.4: Class II example.

The graphs 4.5 and 4.6 show the new method has great results, especially concerning precision and recall. The new features seem to be strongly discriminating for this separation task. In particular, Surface Volume Ratio and Secondary Structure periodicity are powerful features for class II detection.

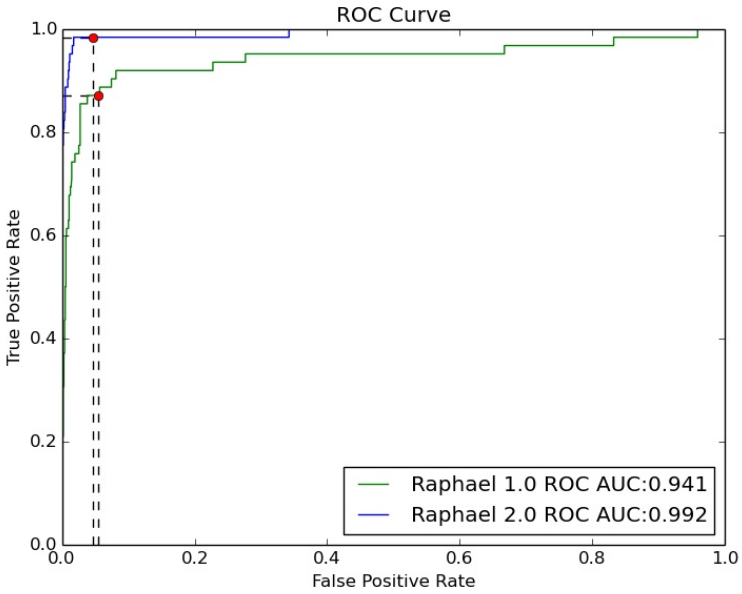


Figure 4.5: Class II ROC Curve.

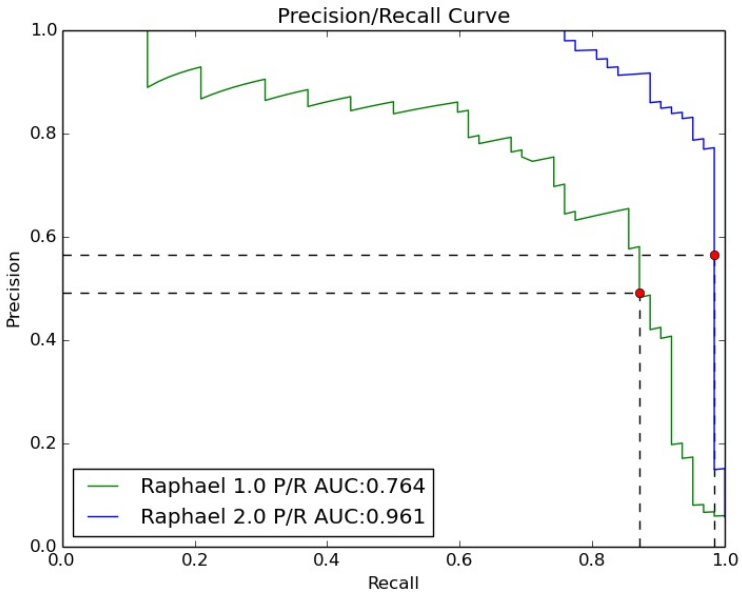


Figure 4.6: Class II Precision-Recall Curve.

4.2.2 Class III

Class III Subclass	Subclass Name	# Examples
III.1	β-solenoid	149
III.2	α/β-solenoid	62
III.3	α-solenoid	292
III.4	trimer of β spirals	7
III.5	single layer β	7

Table 4.3: Class III Distribution.

Class III collects all the examples of solenoidal-like structures. Looking at table 4.3, α and β solenoids are the most present subclasses for this class. From the previous work features for solenoid detection were already chosen. This could be the most stressful test for the new method because RAPHAEL was specifically designed for such detection. ROC and P-R curves compare the performance of the models obtained.

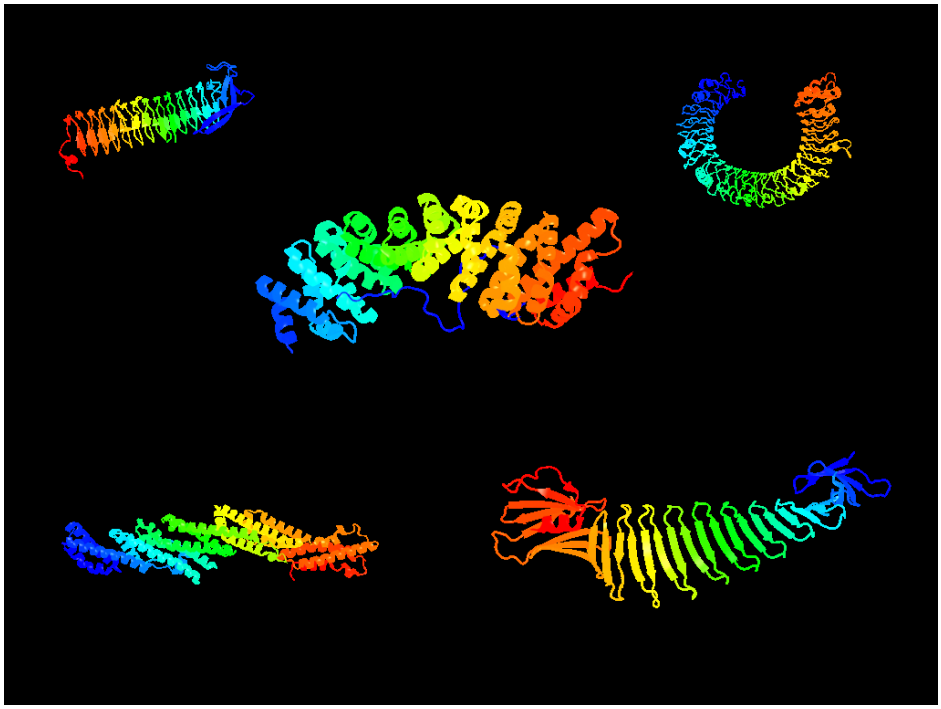


Figure 4.7: Class III examples.

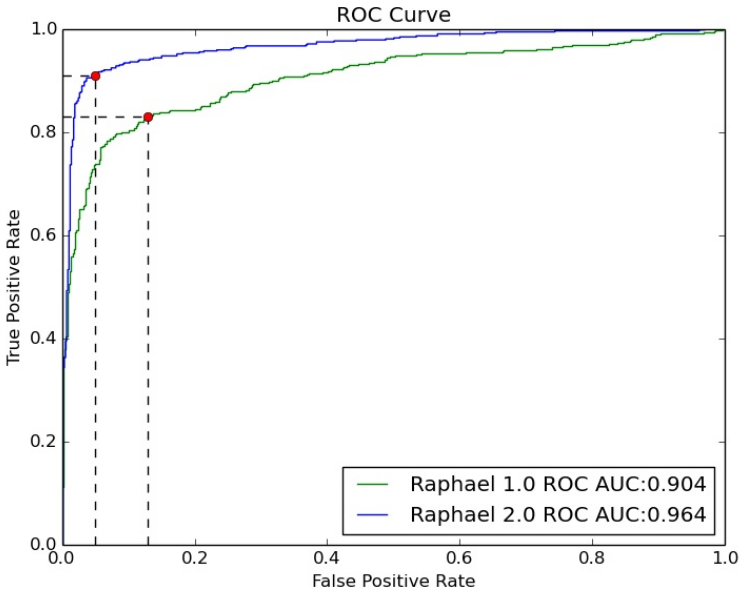


Figure 4.8: Class III ROC Curve.

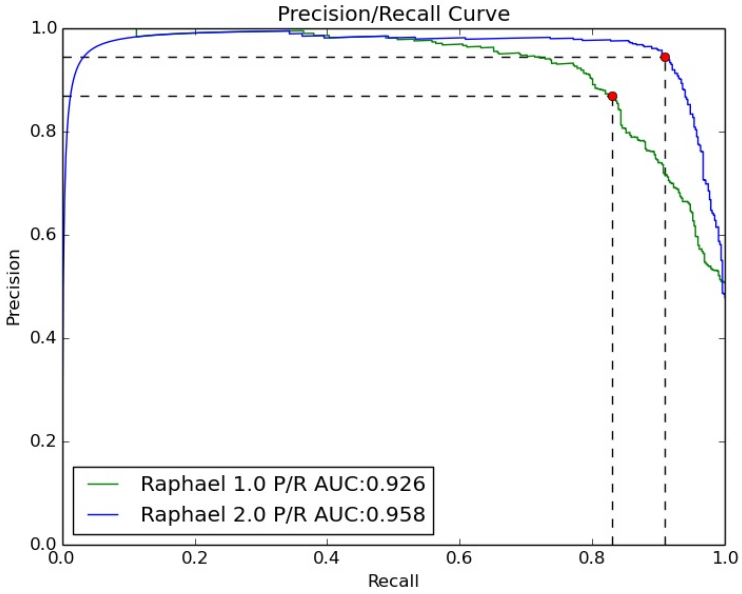


Figure 4.9: Class III Precision-Recall Curve.

Results demonstrate an improvement over the previous method. This is a significative result, specially concerning the quality of prediction (Figure 4.9).
Being proteins of class III the most regular type of repeats such enhancement could be motivated by the addition of secondary structure periodicity and structural features (short range contacts).

4.2.3 Class IV

Class IV Subclass	Subclass Name	# Examples
IV.1	TIM-barrel	201
IV.2	β-barrel	9
IV.3	β-trefoil	15
IV.4	β-propeller	206
IV.5	α/β prism	14
IV.6	α-barrel	5

Table 4.4: Class IV Distribution.

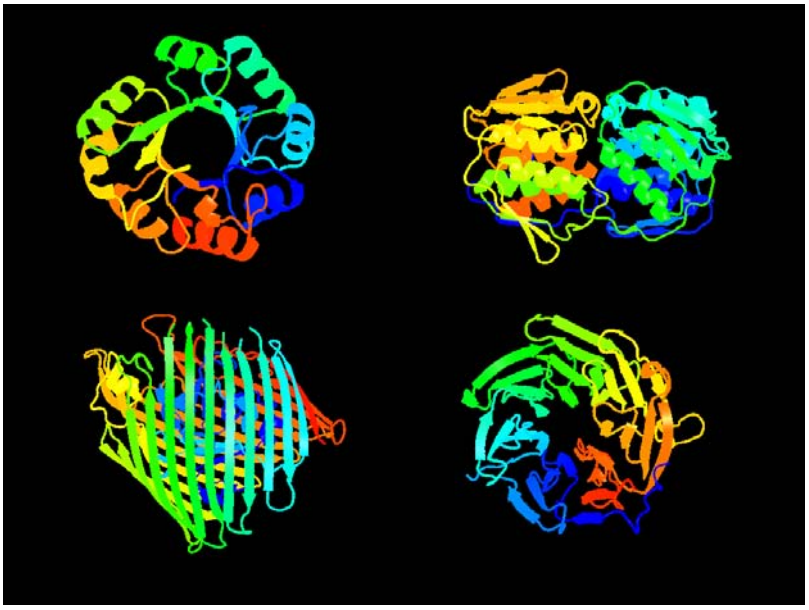


Figure 4.10: Class IV examples.

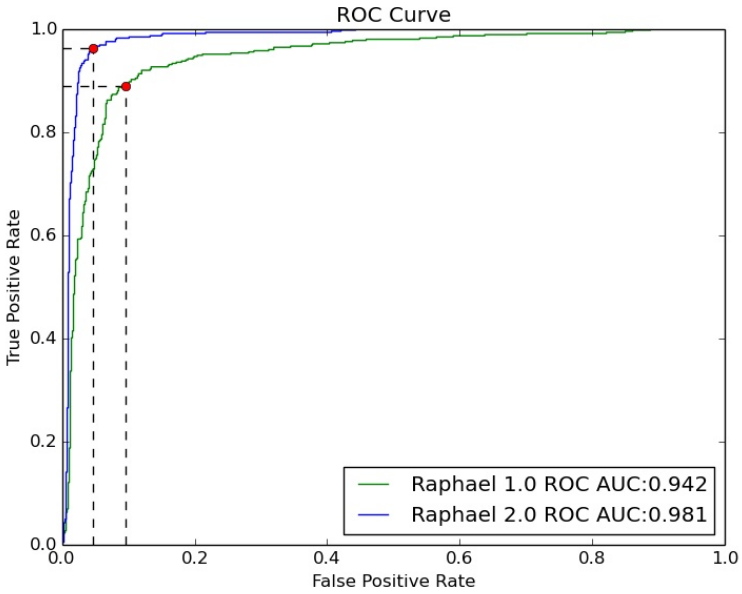


Figure 4.11: Class IV ROC Curve.

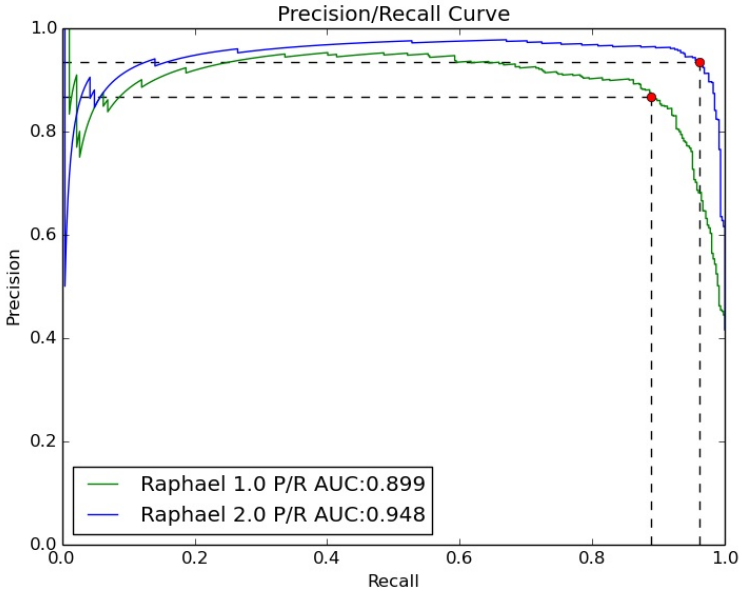


Figure 4.12: Class IV Precision-Recall Curve.

Class IV contains all the examples of closed structures (Fig.4.10). This is one of the most represented class in RepeatsDB and, given the significantly different structure compared to all the others, we expect to have good results in detection of these proteins. ROC and P-R curves compare the performance of the models obtained. Both the models have good performances for class IV detection. A strong feature for this task is given by elongation measure (eqn. 2.3), already present in RAPHAEL. It is possible to hypothesize that improvement in RAPHAEL2.0 may be caused by average periodic distance that seems to have a good discrimination power for closed structures detection.

4.2.4 Class V

Class V Subclass	Subclass Name	# Examples
V.1	α-beads	3
V.2	β-beads	41
V.3	α/β-beads	6
V.Other	Unknown subclass	2

Table 4.5: Class V Distribution.

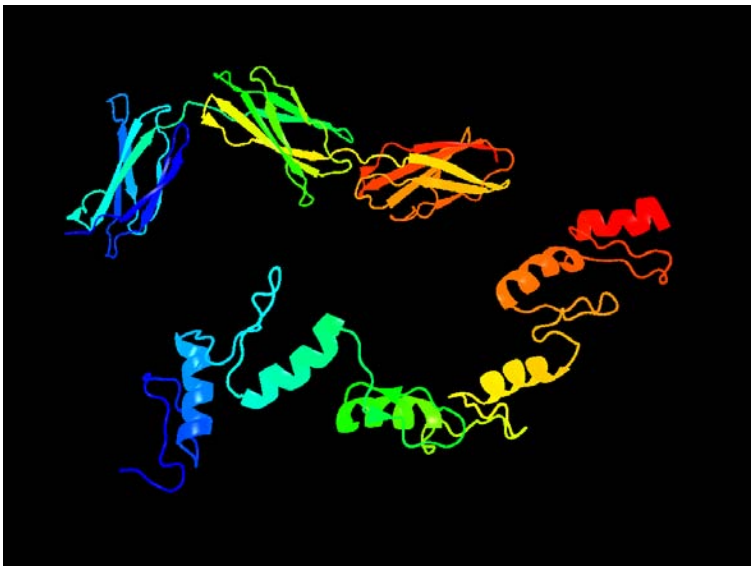


Figure 4.13: Class V examples.

Structures in class V are the most heterogeneous and difficult to detect and predict. Figure 4.13 shows examples of repeats belonging to this class and table 4.5 shows the subclass distribution. Structurally they present several repeated units linked by flexible regions; as example the most representative class V subclass in RepeatsDB is β-beads, in this case repeated units resemble to solenoidal-like structures. Large efforts were made in order to create effective features for this class. ROC and P-R curves compare the performance of the models obtained. The graphs show this model perform generally worse compared to all the other models. As explained, the geometrical aspect of these repeats is hardly detectable considering that every single unit is likely to be recognized in different manner (for example as soleinodal-like structure) misleading the discrimination. Periodic score feature (eqn. 3.1), short range contacts and SS Periodicity were specifically designed for recognition of such repeats, despite that results are not completely satisfying. More precisely, most of class V examples tend to be mistakenly not detected and predicted with scores similar to class III repeats.

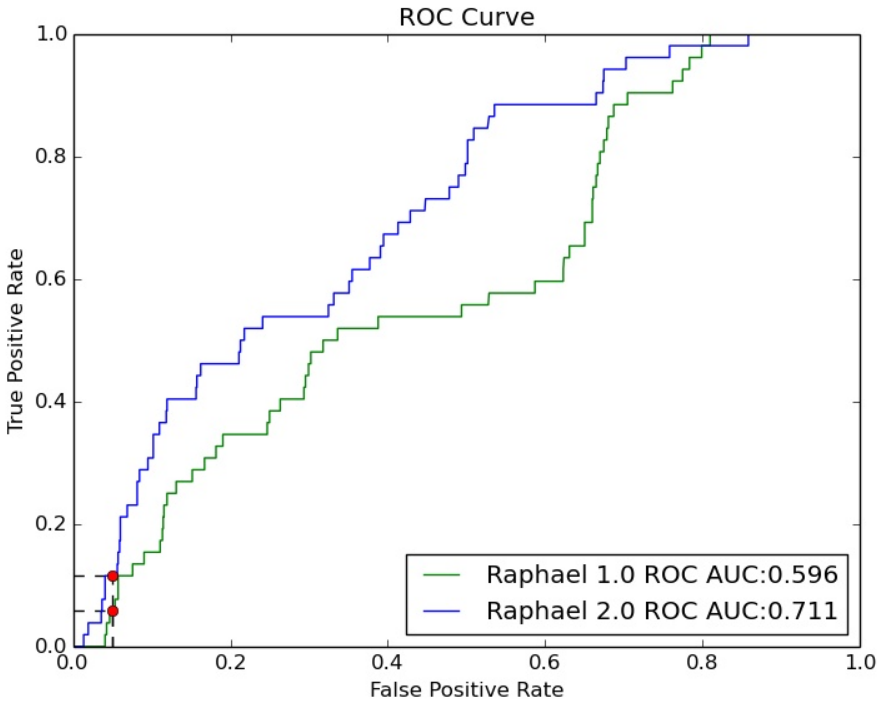


Figure 4.14: Raphael2.0. Class V ROC Curve.

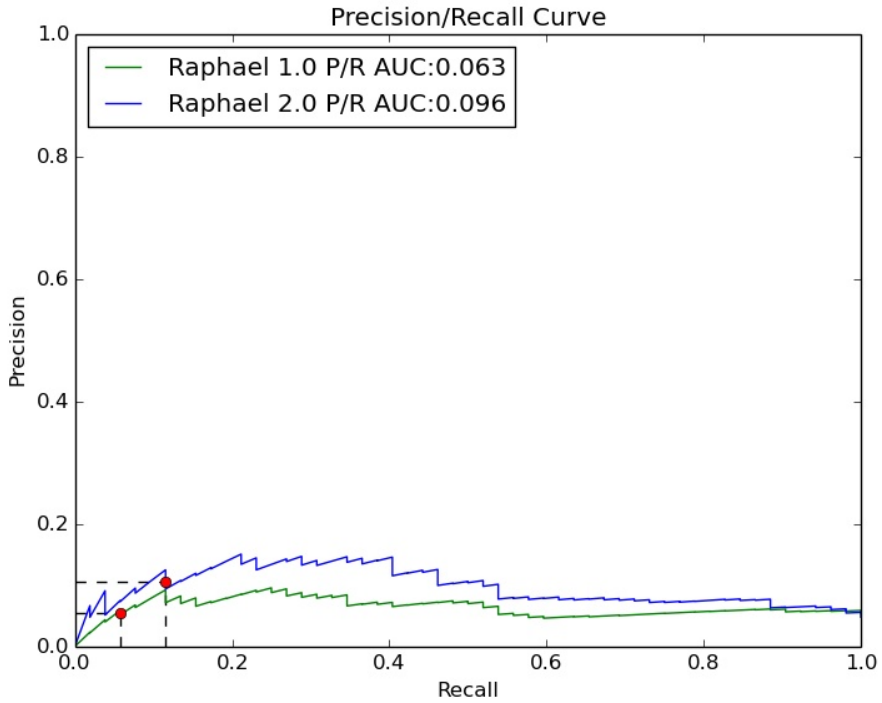


Figure 4.15: Class V Precision-Recall Curve.

4.3 Discussion

The main concern of this thesis is to discriminate repeats at a finer class level. The previous version of RAPHAEL was specifically devised to recognize solenoid repeats (class III subclasses). Here I expanded the method in order to retrieve new class-specific features making it able to detect and classify repeats belonging to different classes.
The results show that RAPHAEL2.0 is much improved over RAPHAEL when detecting the four classes. This specificity is achieved collecting features able to significantly discriminate a class when compared to the others. This new method demonstrates to have optimal performance except for class V. As previously described class V detection is the trickiest problem in this study and a possible explanation the size of the data for class V and its similarity to solenoid class III at least in terms of the designed features. In addition, the structure of such repeats is usually elongated "beads on a string" with beads having globular structure possibly disrupting the periodicity. Although

beads on a string are probably the simplest repeats to distinguish visually they still represent a delicate problem in such type of automatic recognition. Another big factor that must be taken into account is the class size; as seen in table 4.5 we have only 52 examples divided in four subclasses, considering the complexity of this kind of structures it is unlikely that a robust model for this discrimination task could be achieved at this time for class V. It is reasonable to hypothesize that increasing the data size will assure better performance specially for accuracy of prediction.
Besides this high global accuracy in classification, the algorithm developed is also fast. For this reason it is possible to automatically mine thousands of structures present in the Protein Data Bank. Analysing these predictions may give insights into their function and evolution. The next section is concerned with the description of the new package with the mining pipeline explained.

4.4 RAPHAEL2.0 Pipeline

Starting from RAPHAEL (`http://protein.bio.unipd.it/raphael/`) a new pipeline for class recognition was devised. Figure 4.16 shows the flowchart of the new package. Considering the results obtained we chose to employ the first method for the initial discrimination step, so the renewed model created using RAPHAEL is used to detect repeats. For all the other comparisons, models obtained with the new method have been used.
All the thresholds selected for the models were chosen optimizing the False Positive Rate at 5%.
Given a protein structure in *pdb* format a test for repeat detection is initially conducted; with a positive repeat detection the new features of RAPHAEL2.0 are retrieved in order to examine the SVM scores coming from the class models thus assigning it to a particular class.
If one or more threshold is overtaken the analysed protein is labelled with the respective class tag. Ambiguity exists since either a protein could have more than one type of repeated region.
The pipeline just discussed is tested in the classification of the unassigned repeats presents in RepeatsDB. The following describes the execution of the pipeline on currently unassigned proteins. The quantities and assignments are shown in detail.

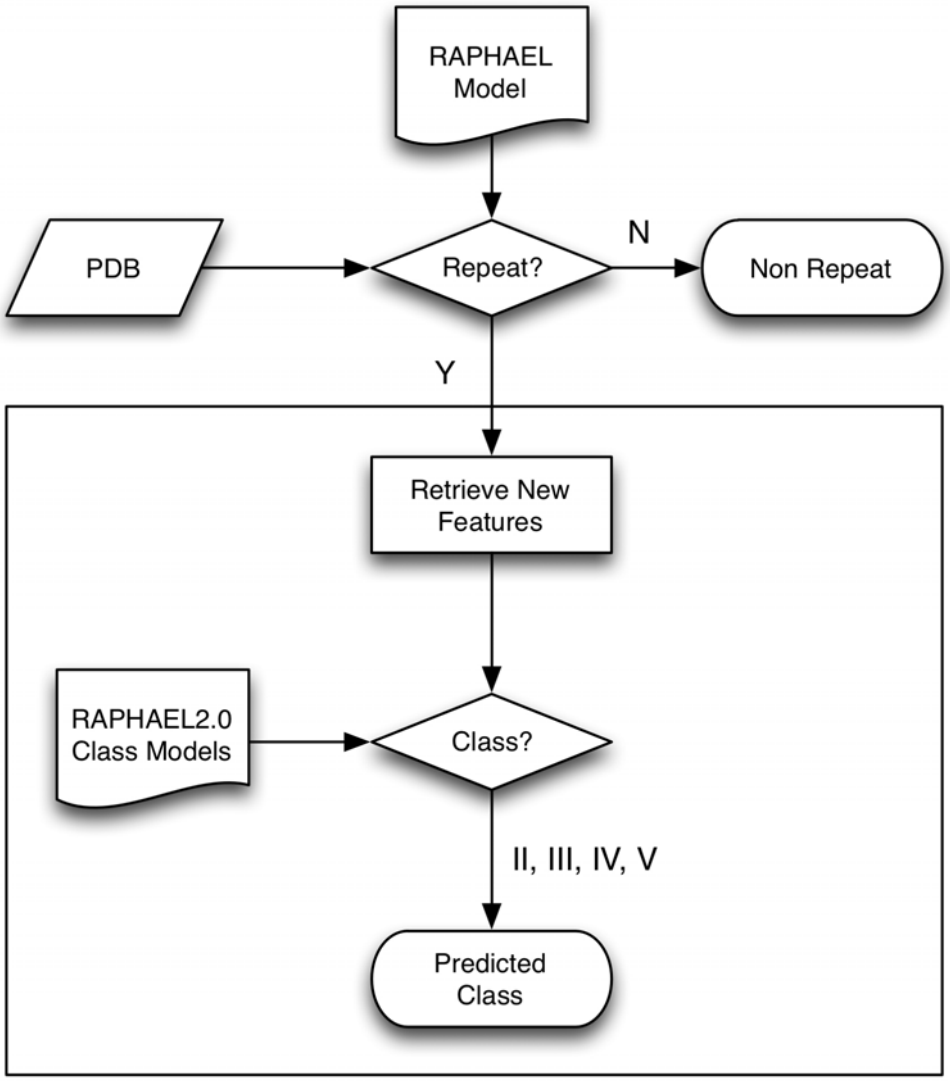


Figure 4.16: RAPHAEL2.0 Pipeline. RAPHAEL is used to discriminate repeats from globular proteins. The main focus of this thesis was the construction of this pipeline and the development of class prediction highlighted in the black box.

4.5 Prediction on Unassigned Proteins in RepeatsDB

The new package was tested on a set of 7,984 chains which were predicted repeats from RAPHAEL without a specific class assignment. The goal of this test is to assign each chain to the related class and identify wrong predictions. Each protein was so processed and the output file was analysed.
We are specifically interested in how many proteins are predicted with a single, double or triple (if any) class label, the number of unassigned repeats, the erroneous predictions coming from RAPHAEL and the variation of the dataset exploited after the test. Data distribution is shown in figure 4.17.

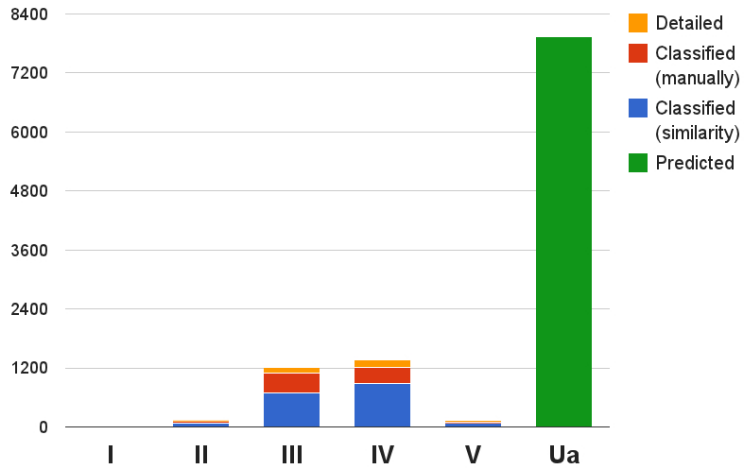


Figure 4.17: RepeatsDB data distribution before running Raphael2.0. The green column highlights the unassigned proteins to be predicted with the new pipeline.

4.5.1 Results

Predictions given by RAPHAEL2.0 allow us to refine the data distirbution in RepeatsDB, figure 4.18 shows the new subdivision. As visible from the figure 4.18 there is a linear growth for class III and IV remaining the most repre-

Type	Class	# Predicted
Repeats	Class II	1,341
	Class III	2,213
	Class IV	2,489
	Class V	7
	Unassigned	1,114
Non Repeats		1,007

Table 4.6: Raphael2.0 predictions.

sented in this database, a surprisingly growth comes from class II meanwhile class V still remain the smallest represented. Table 4.6 shows some statistics about the prediction task.

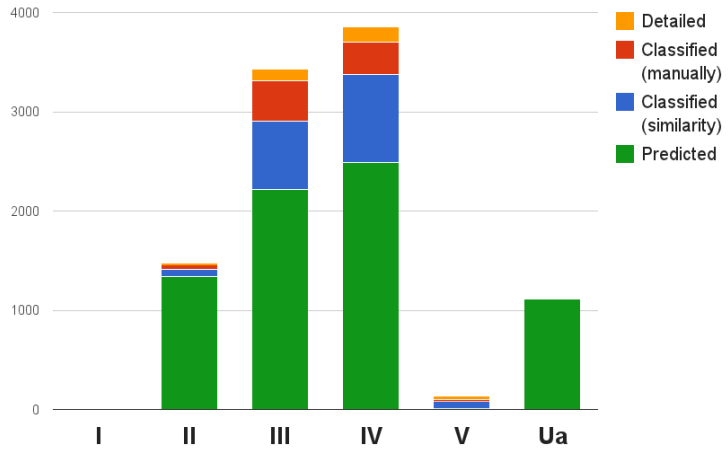


Figure 4.18: RepeatsDB data distribution after running Raphael2.0. Class III and IV are the most increased in terms of number of new proteins predicted. Proportionally class II is the most increased class. Class V has a small variation.

Type	Classes	# Predicted
Double	Class II , III	221
	Class II , IV	9
	Class II , V	3
	Class III , IV	22
	Class III , V	1
	Class IV , V	0
Triple	Class II , III , V	2

Table 4.7: Multiple Predictions.

In addition to this we have also investigated multiple assignments in order to understand if predictions are truly made. Table 4.7 shows the results of this study. Table 4.7 shows that in rare cases double assignments are given. This is entirely possible since a multi complex protein can indeed contain multiple repeat classes. Figure 4.19 shows an example. Multiple class assignments happens for 256 cases but these are not errors.

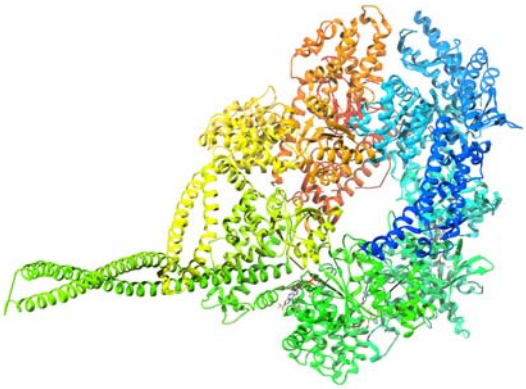


Figure 4.19: PDB ID: 3vkgB. This is an example of protein with multiple repeated regions. The protein contains two helices (left side) like a class II example and a closed structure from class IV (right side).

In particular, repeats with multiple classes need further algorithmic development to extract each class separately and assign them to their appropriate class. This will involve finding the boundaries between class and splitting them. Analysing some of the multiple predicted repeats it is possible to hypothesize that the method is divert by the size of the structure, huge proteins tend to have more than one recognizable class preventing the assignment of a single label.
The accuracy of the methods were already proven so the new distribution can become part of the RepeatsDB update adding additional quality to the database.

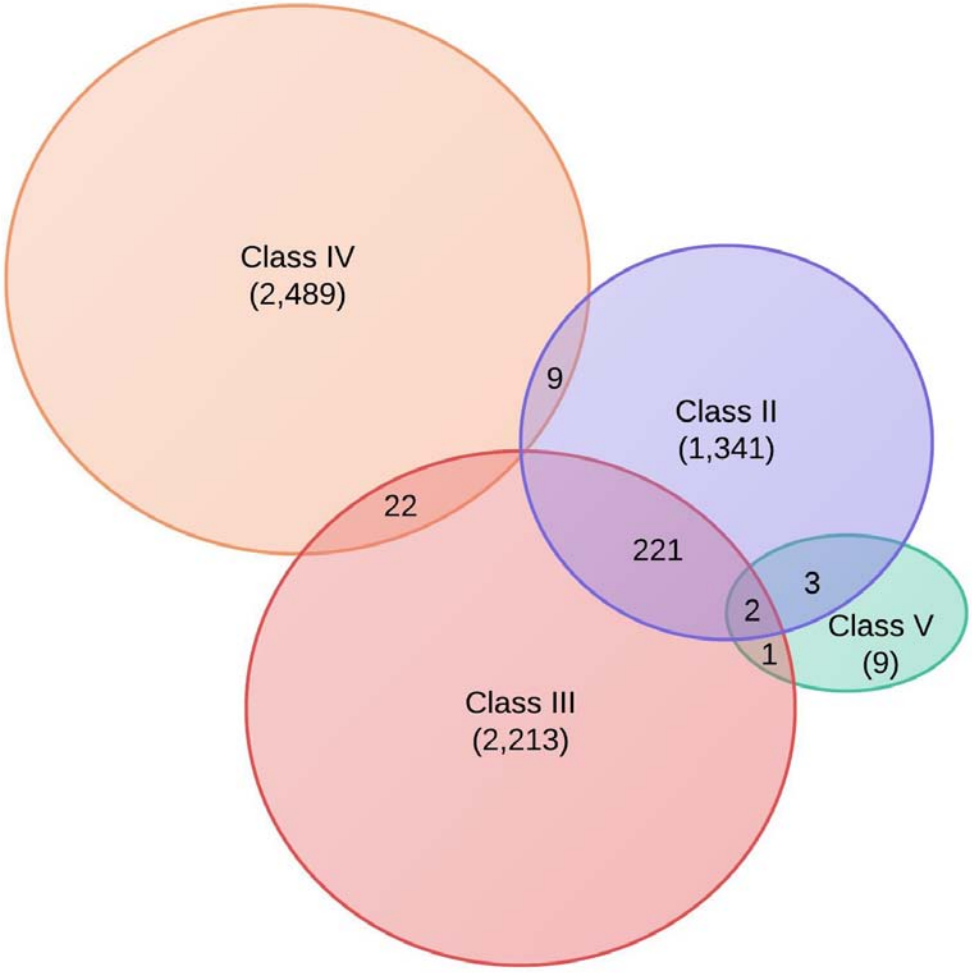


Figure 4.20: Venn diagram representing RAPHAEL2.0 predictions. Intersections between sets represent the multiple predictions made by RAPHAEL2.0.

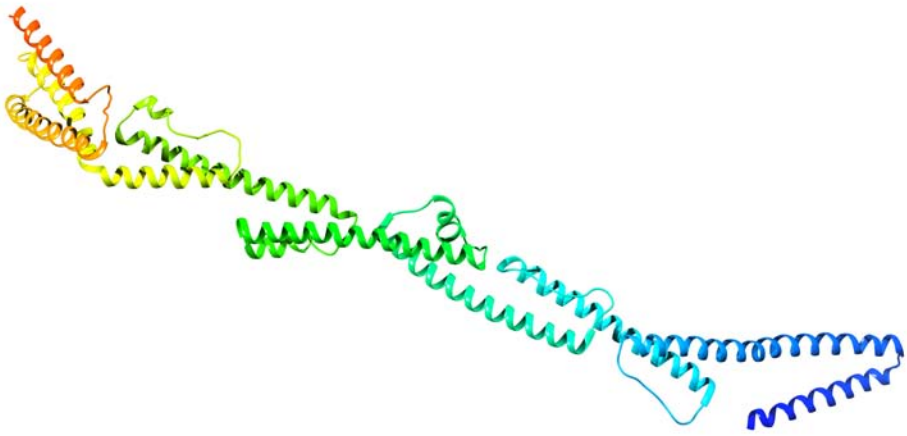


Figure 4.21: PDB ID: 4l3iA. This spectrin-like protein (class V) is a typical example of non correct prediction. RAPHAEL2.0 predicts the protein as class II and III due to its particular structure.

Figure 4.20 shows the intersection of multiple class assignments. The first observation is the occurrence of this phenomenon is rare. Secondly class II and III have the most frequent combination with the rest significantly less. Figure 4.21 shows an example of a class II and III mixture. This spectrin-like protein is normally classified as "beads on a string" (class V), due to its particular structure it is recognized as a mixture of class II (completely α) and class III (helices are disposed at a regular distance, for this reason recognized as solenoidal).

Chapter 5

Conclusions

In 2012 with RAPHAEL a new category of structure based repeat detection algorithm has been devised. During these years the increasing amount of data and the relative quality allowed to refine and improve the existing tool. In this sense, the development of RepeatsDB provides the most qualitative source of repeats data at the moment permitting to build solid models with a machine learning approach.
RAPHAEL2.0 aims to enhance the accuracy of repeat detection and extend the capabilities of the tool making prediction at the deeper class level. Describing RAPHAEL we underlined some major points to be upgraded such as repeat mining and classification. Analysing the results obtained with the new data we improved the first version of the tool giving more accuracy in repeats recognition. A new tool is so created able to correctly classify repeats in four classes with good precision and low false positive rate.
The accuracy of the models comes directly from the datasets so, validation of the proteins classified with this new tool will permit to create more accurate models strengthen the ones already built. Retrieving new data will also strongly improve performances for the class V model creating the most accurate repeats class predictor at this time.
With the models built from the new dataset a new package has been released, which allows mining multiple data in turn detecting each protein and the actual class. Mining the entire Protein Data Bank (more than 100,000 structures at the moment) may be the first step in order to discover new informations on repeats. This mining task will also permit to easily classify repeat proteins in subclasses eventually creating the opportunity to develop an algorithm able to automatically classify repeats at subclass level.

Bibliography

[1] I. Walsh, F. G. Sirocco, G. Minervini, T. Di Domenico, C. Ferrari, and S. C. E. Tosatto, "RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures." *Bioinformatics*, vol. 28, no. 24, pp. 3257–64, Dec. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22962341

[2] M. a. Andrade, C. Perez-Iratxeta, and C. P. Ponting, "Protein repeats: structures, functions, and evolution." *J. Struct. Biol.*, vol. 134, no. 2-3, pp. 117–31, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11551174

[3] P. Maciel, C. Gaspar, A. L. DeStefano, I. Silveira, P. Coutinho, J. Radvany, D. M. Dawson, L. Sudarsky, J. Guimarães, and J. E. Loureiro, "Correlation between CAG repeat length and clinical features in Machado-Joseph disease." *Am. J. Hum. Genet.*, vol. 57, no. 1, pp. 54–61, 1995.

[4] J. Kalita, U. K. Misra, D. K. Mishra, K. Thangaraj, R. D. Mittal, and B. R. Mittal, "Nonprogressive juvenile-onset spinal muscular atrophy: A clinico-radiological and CAG repeat study of androgen receptor gene," *J. Neurol. Sci.*, vol. 252, no. 1, pp. 24–28, 2007.

[5] H. T. Orr, M. Y. Chung, S. Banfi, T. J. Kwiatkowski, A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. Ranum, and H. Y. Zoghbi, "Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1." *Nat. Genet.*, vol. 4, no. 3, pp. 221–226, 1993.

[6] R. Koide, T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi, K. Endo, H. Takahashi, R. Kondo, A. Ishikawa, and T. Hayashi, "Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA)." *Nat. Genet.*, vol. 6, no. 1, pp. 9–13, 1994.

[7] J. Jorda and A. V. Kajava, "Protein homorepeats: Sequences, structures, evolution, and functions," *Adv. Protein Chem. Struct. Biol.*, vol. 79, no. C, pp. 59–88, 2010.

[8] A. V. Kajava, "Tandem repeats in proteins: from sequence to structure." *J. Struct. Biol.*, vol. 179, no. 3, pp. 279–88, Sep. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21884799

[9] J. de Wit, W. Hong, L. Luo, and A. Ghosh, "Role of Leucine-Rich Repeat Proteins in the Development and Function of Neural Circuits," pp. 697–729, 2011.

[10] A. V. Kajava, J. M. Squire, and D. A. D. Parry, "??-Structures in Fibrous Proteins," pp. 1–15, 2006.

[11] E. R. G. Main, A. R. Lowe, S. G. J. Mochrie, S. E. Jackson, and L. Regan, "A recurring theme in protein engineering: The design, stability and folding of repeat proteins," pp. 464–471, 2005.

[12] N. Stefan, P. Martin-Killias, S. Wyss-Stoeckle, A. Honegger, U. Zangemeister-Wittke, and A. Plückthun, "DARPins recognizing the tumor-associated antigen EpCAM selected by phage and ribosome display and engineered for multivalency," *J. Mol. Biol.*, vol. 413, no. 4, pp. 826–843, 2011.

[13] J. Buard and G. Vergnaud, "Complex recombination events at the hypermutable minisatellite CEB1 (D2S90)." *EMBO J.*, vol. 13, no. 13, pp. 3203–3210, 1994.

[14] E. M. Marcotte, M. Pellegrini, T. O. Yeates, and D. Eisenberg, "A census of protein repeats." *J. Mol. Biol.*, vol. 293, no. 1, pp. 151–160, 1999.

[15] A. V. Kajava, "Review: proteins with repeated sequence–structural prediction and modeling." *J. Struct. Biol.*, vol. 134, no. 2-3, pp. 132–44, Jan. 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047847700943284

[16] H. Y. Zoghbi, "Trinucleotide repeat disorders," in *Princ. Mol. Med.* Humana Press, 2006, pp. 1114–1122.

[17] E. Coward and F. Drablø s, "Detecting periodic patterns in biological sequences," *Bioinformatics*, vol. 14, no. 6, pp. 498–507, 1998.

[18] M. Gruber, J. Söding, and A. N. Lupas, "REPPER - Repeats and their periodicities in fibrous proteins," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, 2005.

[19] A. M. Newman and J. B. Cooper, "XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences." *BMC Bioinformatics*, vol. 8, p. 382, 2007.

[20] J. Jorda and A. V. Kajava, "T-REKS: Identification of Tandem REpeats in sequences with a K-meanS based algorithm," *Bioinformatics*, vol. 25, no. 20, pp. 2632–2638, 2009.

[21] A. Heger and L. Holm, "Rapid automatic detection and alignment of repeats in protein sequences," *Proteins Struct. Funct. Genet.*, vol. 41, no. 2, pp. 224–237, 2000.

[22] R. Szklarczyk and J. Heringa, "Tracking repeats using significance and transitivity," in *Bioinformatics*, vol. 20, no. SUPPL. 1, 2004.

[23] E. L. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, "Pfam: Multiple sequence alignments and HMM-profiles of protein domains," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 320–322, 1998.

[24] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting, "SMART, a simple modular architecture research tool: identification of signaling domains." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 11, pp. 5857–5864, 1998.

[25] A. Biegert and J. Söding, "De novo identification of highly diverged protein repeats by probabilistic consistency," *Bioinformatics*, vol. 24, no. 6, pp. 807–814, 2008.

[26] L. Marsella, F. Sirocco, A. Trovato, F. Seno, and S. C. E. Tosatto, "REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform." *Bioinformatics*, vol. 25, no. 12, pp. i289–95, Jun. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2687986&tool=pmcentrez&rendertype=abstract

[27] A. Vo, N. Nguyen, and H. Huang, "Solenoid and non-solenoid protein recognition using stationary wavelet packet transform," *Bioinformatics*, vol. 26, no. 18, 2010.

[28] R. G. Parra, R. Espada, I. E. Sánchez, M. J. Sippl, and D. U. Ferreiro, "Detecting Repetitions and Periodicities in Proteins by Tiling the

Structural Space," *J. Phys. Chem. B*, vol. 117, no. 42, pp. 12 887–12 897, Jun. 2013. [Online]. Available: http://arxiv.org/abs/1306.2852

[29] T. Hrabe and A. Godzik, "ConSole: using modularity of Contact maps to locate Solenoid domains in protein structures." *BMC Bioinformatics*, vol. 15, no. 1, p. 119, 2014. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4021314&tool=pmcentrez&rendertype=abstract

[30] T. Di Domenico, E. Potenza, I. Walsh, R. G. Parra, M. Giollo, G. Minervini, D. Piovesan, A. Ihsan, C. Ferrari, A. V. Kajava, and S. C. E. Tosatto, "RepeatsDB: a database of tandem repeat protein structures." *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D352–7, Jan. 2014. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3964956&tool=pmcentrez&rendertype=abstract

[31] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers*, vol. 22, no. 12, pp. 2577–637, Dec. 1983. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6667333

[32] M. Shirota, T. Ishida, and K. Kinoshita, "Effects of surface-to-volume ratio of proteins on hydrophilic residues: decrease in occurrence and increase in buried fraction." *Protein Sci.*, vol. 17, no. 9, pp. 1596–1602, 2008.

[33] R. P. Joosten, T. a. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend, "A series of PDB related databases for everyday needs." *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D411–9, Jan. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013697&tool=pmcentrez&rendertype=abstract

[34] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1995, vol. 8. [Online]. Available: http://portal.acm.org/citation.cfm?id=211359

[35] T. Joachims, U. Dortmund, and T. Joachimscsuni-dortmundde, "Making Large-Scale SVM Learning Practical," in *Adv. Kernel Methods - Support Vector Learn.*, B. Scholkopf, C. J. C. Burges, and A. J. Smola, Eds. MIT-Press, 1999, ch. 11, pp. 41–56. [Online]. Available: http://svmlight.joachims.org/$\delimiter"026E30F$nhttps://eldorado.uni-dortmund.de/handle/2003/2596

[36] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[37] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.