DISCOVERY OF AMINO ACID COMPOSITIONS AND MOTIFS RESPONSIBLE FOR TOPOLOGICAL TRANSITIONS IN PROTEIN COMPLEXES

by

ERHAN EKMEN

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of

the requirements for the degree of

Master of Science

Sabancı University

July 2021

DISCOVERY OF AMINO ACID COMPOSITIONS AND MOTIFS RESPONSIBLE FOR TOPOLOGICAL TRANSITIONS IN PROTEIN COMPLEXES

APPROVED BY:

[redacted] [redacted]

[redacted]

[redacted] [redacted]

[redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

DATE OF APPROVAL: 08/07/2021

ABSTRACT

DISCOVERY OF AMINO ACID COMPOSITIONS AND MOTIFS RESPONSIBLE FOR TOPOLOGICAL TRANSITIONS IN PROTEIN COMPLEXES

Erhan Ekmen

Molecular Biology, Genetics and Bioengineering, M.Sc. Thesis, July 2021

Thesis Supervisor: Canan Atılgan

Thesis Co-supervisor: Ali Rana Atılgan

Keywords: amino acid composition, secondary/quaternary structure, k-nearest neighbor, support vector machine, H/P model, dissipative particle dynamics

Prediction of structural classes of proteins has been pursued using various features of proteins such as amino acid composition (AAC), sequence information, structural motifs and amino acid coordinates. In some studies, it has been shown that using only AACs is enough to predict structural classes such as α, β, α+β, α/β and being monomer or dimer with high accuracy. These studies implicate that evolution has an impact on AAC for secondary and quaternary structure preferences of proteins. In this study, we use AACs to predict the topological preferences of protein complexes by applying several machine learning (ML) models. We used k-Nearest Neighbor (kNN) and Support Vector Machine (SVMs) algorithms utilizing AACs as the only feature for the prediction of secondary and quaternary structural classes of proteins. We successfully predicted the five secondary structural classes (α, β, α+β, α/β, s) of proteins with average F1-score of 0.65 with multiclass model. Different quaternary structural classes of complexes having four subunits have also shown that distinctive complexes which have higher symmetry can be predicted more robustly, up to an F1-score of 0.86, and proteins in two virus capsid structure classes with different symmetry can be predicted up to an F1-score of 0.89, proving how a simple feature of proteins is effective for quaternary structure of the protein complexes. To gain a physics-based understanding of these findings, we modeled the chains at the level of H/P (Hydrophobic/Polar) two-letter alphabet and detected unique 10-16 letter long sequences belonging to different quaternary topologies. We applied coarse-grained Dissipative Particle Dynamics (DPD) simulations on complexes which have repetitions of these sequences and found associations unique to the sequences. Thus, although the AACs are effective in the formation of quaternary structures, sequences creating special hydrophobic patches at the interface determine the topological details.

ÖZET

PROTEİN KOMPLEKSLERİNDE TOPOLOJİK GEÇİŞLERE SEBEP OLAN AMİNO ASİT YÜZDELERİNİN VE MOTİFLERİN KEŞFİ

Erhan Ekmen

Moleküler Biyoloji, Genetik ve Biyomühendislik, Yüksek Lisans Tezi, Temmuz 2021

Tez Danışmanı: Canan Atılgan

Tez İkinci Danışmanı: Ali Rana Atılgan

Anahtar Kelimeler: amino asit yüzdesi, ikincil/dördüncül yapı, k en yakın komşu, destek vektör makineleri, H/P model, dağılıcı parçacık dinamiği

Proteinlerin yapısal sınıflarının tahmini için amino asit yüzdeleri (AAY), sekansları, yapısal motifleri ve amino asit koordinatları gibi birçok özelliği kullanılmıştır. Bazı çalışmalarda, sadece AAY'sinin α, β, α+β, α/β veya bir proteinin monomer ya da dimer olması gibi yapısal sınıfların tahmininde yeterli olduğu gösterilmiştir. Bu çalışmalar AAY'sinin proteinlerin ikincil ve dördüncül yapı sınıflarına evrimsel etkisini açıkça ortaya koymaktadır. Bu çalışmada ise, sadece proteinlerin AAY'lerini kullanarak birçok makine öğrenmesi tekniği ile proteinlerin topolojik tercihleri tahmin edilmiştir. İkincil ve dördüncül yapı sınıflarının tahminlerinde AAY'leri kullanılarak, k en yakın komşu algoritması ve destek vektör makineleri ile tahminler yapılmıştır. 5 farklı ikincil yapı sınıfı (α, β, α+β, α/β, s) çoklu sınıf tahmini kullanılarak ortalama 0.65 F1 skoru ile doğru tahmin edilmiştir. Farklı dördüncül yapı sınıflarının tahmininde ise dört protein içeren komplekslerden simetrisi yüksek ve ayırt edilebilir olan komplekslerin F1 skoru 0.83'e kadar ulaşmıştır ve farklı simetrilere sahip iki virüs kapsit sınıfındaki proteinler 0.89 ortalama F1 skoru ile doğru tahmin edilmiştir. Bu durum AAY'si gibi basit bir özelliğin proteinlerin dördüncül yapısını ne kadar etkilediğini kanıtlamaktadır. Sonrasında fizik tabanlı bir anlayış elde edebilmek için, ikili alfabe modeli H/P (Hidrofobik/Polar) kullanılarak elde edilen zincirlerden komplekslere ait 10-16 harf uzunluğunda birbirinden farklı tekrarlayan motifler tespit edilmiştir. AAY'leri ve tespit edilen motiflerle oluşturulan zincirlere Dağılıcı Parçacık Dinamiği (DPD) benzetimleri uygulandığında oluşturulan bu zincirlerin birbirinden farklı özgün özellikleri gözlemlenmiştir. AAY'leri dördüncül yapıların oluşumunda önemli olsalar da sekansların oluşturduğu ve etkileşim yüzeylerinde bulunan hidrofobik kısımların topolojik detayları tanımladığı anlaşılmıştır.

To Özgür Gül,

in loving memory.

ACKNOWLEDGEMENTS

I would like to first thank my advisors Canan Atilgan and Ali Rana Atilgan for their great support. I was able to overcome any research related or personal problems with their genuine interest and care. The working place they created throughout the years has prepared me for my long academic career in future.

Next, I want to thank my jury members; Ogün Adebali, Öznur Taştan, and Deniz Eroğlu. The courses which were given by Ogün Adebali and Öznür Taştan really helped me understanding the theoretical and practical parts of this thesis.

Further, I would like to thank some of our group members. I met with Sofia during my internship. She was always right behind me (literally), for my any biological questions and she always had an answer. I also met with Ebru during my internship and even then, she was giving me scientific suggestions for my undergraduate thesis. After being a lab member, we spent most of our free times together in campus and exchanged valuable opinions mostly about science and life. One of our other group members, Kurt, made me always look another way around of scientific concepts with his never-ending questions. Oğuzhan really helped me at first while I was learning the new concepts even after he went to the US. Fürkan has always cheered up with his funny songs that come out of nowhere and have no meanings. In addition to that, he was always sincerely supporting me in the times when I was stressful and always made me look to the world from the bright side, even though, he considers himself as a dark person. Işık was the first person who I felt closer compared to others due to her Bilgi University history. We formed kind of an alliance right away and nothing could've stopped us after that point. Tamay was always there for me when I was down, and she encouraged me to take crucial steps in my academic career. I will never forget her ongoing optimist behavior against everything.

I also wanted to thank my roommate, Niang for useful late-night scientific talks which actually is the first thing I look for in a roommate.

Last but not least, İrem. I am very grateful to have a such friend who I spent most of my time with. I have no doubt that our friendship will last till the time when we've got our Nobel Prizes.

We also thank the Scientific and Technological Research Council of Turkey (Grant Number 117F389) for funding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAC	Amino Acid Composition
Å	Ångström
CG	Coarse-Grained
DPD	Dissipative Particle Dynamics
Fs	Femtosecond
H/P	Hydrophobic/Polar
K	Kelvin
kNN	k-Nearest Neighbors
MEME	Multiple Expectation Maximization Algorithm for Motif Elicitation
MD	Molecular Dynamics
ML	Machine Learning
Ns	Nanosecond
NPT	Isothermal–Isobaric Ensemble
PDB	Protein Data Bank
PPI	Protein-Protein Interaction
UniProtKB	Universal Protein Resource Knowledgebase
RDF	Radial Distribution Function
RoG	Radius of Gyration
QS	Quaternary Structure
SVD	Singular Value Decomposition
SVM	Support Vector Machine
VIPERdb	Virus Particle Explorer database

1. INTRODUCTION

In organisms, proteins mostly function in the form of complexes with other proteins; for example, 80% of the yeast proteins interact at least with one protein [1]. It is crucial to unravel the interactions between these protein complexes to understand which features of the proteins are involved in the process of evolution. Most recently, artificial intelligence approaches have been extremely successful in predicting the folded three-dimensional structures of monomeric proteins with unprecedented accuracy [2]. Nevertheless, deciphering the protein-protein interaction (PPI) surfaces remains an enigma.

Secondary, tertiary, and quaternary structures of proteins rely on the location and the proportion of amino acids, final poses are determined by nonbonded interactions acting within and between units. Using these features of the proteins, many prediction models were generated in the past; some of them rely on the coordinates of each atom to predict tertiary structures [3], while others use relatively simple features such as amino acid composition to predict secondary structural classes of proteins and protein-protein interactions [4-10].

Secondary structural classes of proteins are classified as α having a predominant amount of α-helices, β having a predominant amount of β-sheets, α+β consisting of separated out α and (mainly antiparallel) β regions, α/β consisting of alternating α and (mainly parallel) β regions, and small (s) proteins whose tertiary structures are formed by disulfide bridges, metal ligands, or cofactors [11]. For the classification of proteins into the set [α, β, α+β, α/β, s], there are several studies which reached high accuracy scores by using AACs and PPIs as features. First studies which use AACs date back to 1992 [9]. Based on 184 proteins, Chou and collaborators [5] have reached 100% prediction rate for α, β, and α+β classes and 96.7% for α/β proteins by using an algorithm which was later shown to be mathematically equivalent to the SVD method [4]. The latter study rationalized the success of the AACs by using simplified two-dimensional chain models to argue that the amino acid types impose constraints on the coordination number of the individual sites and affected the size and geometry of non-bonded clusters which in turn determine the fold. Other studies which use AACs as feature for their prediction models: for prediction of secondary structural classes, accuracy scores of 68% for α, 64% for β, and 92% for the cumulative α+β and α/β, based on 5796 proteins have been reached by combining AACs with evolutionary AA coupling information as features in machine learning (ML) algorithms [6] in a study. Another study has employed support vector machines (SVM) as a prediction method and achieved 74%, 82%, 88%, and 72% accuracy for α, β, α+β and α/β, respectively [12]. However, instead of multiclass prediction like the previous studies, they have used the "one-against others" method which we have also implemented [13, 14]. These studies demonstrate that even a simple feature such as AACs holds an important volume of information on the structural properties of the proteins. This raises the question of how proteins have evolved in terms of their AACs, and if its significance is also valid for quaternary structural classes of proteins complexes.

For tertiary structure prediction for instance, iTASSER [3] (Iterative Threading ASSEmbly Refinement) first identifies the template proteins that have a similar structure or structural motif as the query protein sequence. Then, the secondary structures are predicted based on these multiple template proteins. Later, these fragments are used to assemble structural conformations. It uses only C_α atoms and the center of mass of the sidechains to converge the models. Lastly, multiple Monte Carlo simulations with different temperatures are applied to the predicted structures and they are ranked. This algorithm mostly relies on structural motifs, proving their importance as a feature for predicting structures.

For prediction of quaternary structural classes such as being monomer, dimer, trimer, or any other oligomer, there has been only one study which use AACs as feature in literature. In this study, they used 3174 protein sequences for training set, and 332 protein sequences for testing set. In overall, among seven classes, they reached up to 90.5% success prediction rate by resubstitution test, however they did not use a validation test set, which might interfere the reliability of results and the identity information between sequences is not stated [15].

More complex features of proteins such as relative surface accessibilities of residues have been also used for prediction of PPIs: in 2005, among 6170 nonhomologous proteins, 62616 pairwise interactions were predicted and most of them were verified from various databases and literature [16]. Another study which mostly focuses on conserved residues on PPIs showed that hotspots discovered via experimental alanine-scanning within these interfaces have densely packed, preorganized, and conserved residues which contribute significantly to the stability of PPIs [17, 18]. These conserved residues have been found to be mostly hydrophobic, but also there more charged and polar residues on protein surfaces compared to protein cores, meaning they also have significant effect. In addition to these, structural properties of protein interfaces also have an important role between PPIs: a method which predicts PPIs first compares target proteins with known template protein-protein interfaces by considering them as rigid-body structures, then ranks the candidate structures based on energies calculated from docking simulations [19]. Moreover, in 2020, a method called Perturbation Response Scanning (PRS) [20] was used to study characteristics of residues which take role in disassociation of protein complexes. After perturbing these residues, unbound conformation of the complexes were more preferred and they demonstrated that this changes differ in different types of amino acids, suggesting the importance of amino acid type in interacting protein pairs [21].

These studies show that there are key factors such as sequence and structural information which differ PPIs with each other. Therefore, first we focus on a simple feature derived from sequences: AACs for our prediction model, then we focus on structure of protein complexes and their differentiative properties with a coarse-grained simulation technique: Dissipative Particle Dynamics.

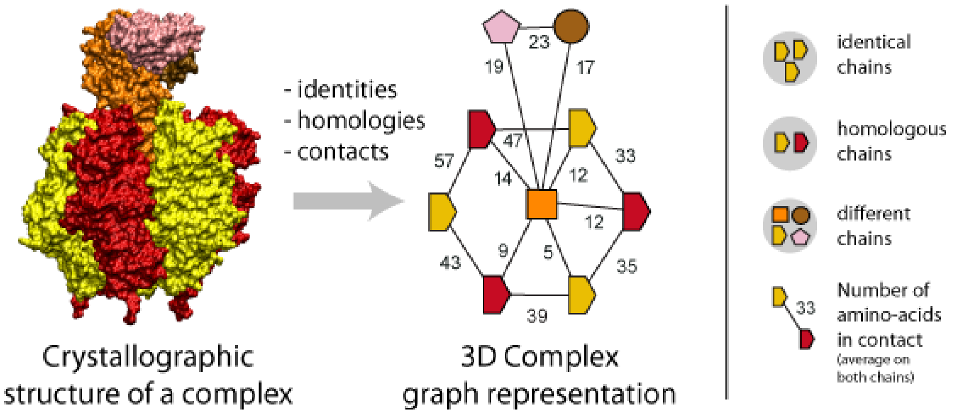


Figure 1 – A graph example of a complex in 3D Complex. These graphs provide topology of complexes; number of nodes (subunits), pattern of interfaces, number of residues at the interfaces and their symmetry [22].

In this thesis, we have not focused on classification of being monomer, dimer, trimer, or any other oligomer, instead, we focused on subclasses which have same number of subunits, however, are topologically totally different. Even though, this would make the prediction more complicated compared to previous study [15], it is a chance to unravel distinctive differences between these complexes. For this purpose, we use 3D complex database which classifies protein complexes based on their topologies [22]. Information such as identities between chains, homologies, and contacts can found in this database (Figure 1). The classification between complexes is based on chain domain architecture, the sequence, and the PPIs. Also, they added symmetry information calculated by rotating the complex around its center of mass for their classification. All this information is represented with a graph model (Figure 1). After creating the database, they found out that most of the protein complexes have four or less subunits and they have tendency to be homomeric and symmetric. Therefore, for quaternary structural class prediction and structural analysis of these complexes, we mostly focused protein complexes having four subunits and compare the symmetric ones with less symmetric ones.

For structural analysis, we first identified structural motifs and then sequences generated from these motifs were studied with coarse-grained dynamics simulations. There are many approaches to study dynamics of proteins, but the most common way is the all-atomic molecular dynamics (MD). However, since they are computationally costly, coarse-grained (CG) approaches which decrease the number of particles to be considered by many orders of magnitude have emerged. One of these CG approaches: DPD is used in literature to study proteins for many types of problems. In 2012, a method with DPD has been developed to mimic hydrogen bonding that stabilize secondary structure of proteins [23]. With this method, different folding characteristic of αSyn polypeptide in different pHs have been demonstrated successfully. Another study has showed that coating proteins of energy storage devices in different solvents shows different kinds of structural properties which can explain the experimental observations [24]. In 2020, folding properties of Chignolin and Superchignolin mini-proteins have been studied [25], and their characteristic hairpin structures have been observed by using DPD approach. These studies show that DPD is a robust tool to study protein structures, and therefore we use it to identify different characteristics of selected protein complexes which we predicted successfully.

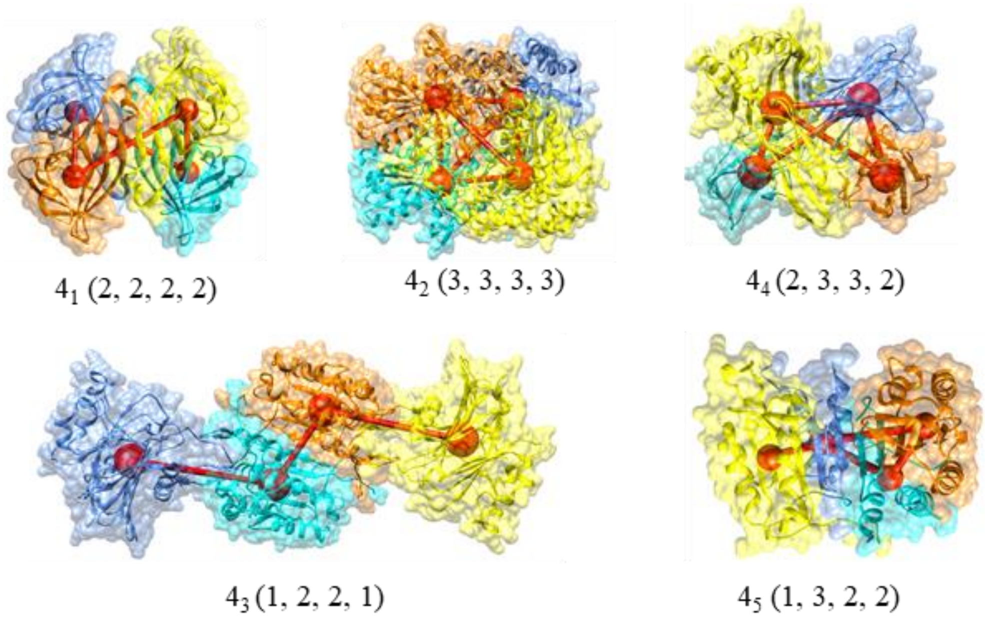


Figure 2 – Examples for complexes having four subunits used in the model. Labelling is based on the number of direct interactions between each protein in that complex. (PDB IDs; 4_1: 5N8E, 4_3: 1I4E, 4_2: 1NI4, 4_4: 1BQH, 4_5: 3P45)

In this thesis, we first use several ML techniques for the prediction of secondary structural classes of proteins using AACs as the only feature to obtain a parallelism between previous studies but using more than 30000 proteins now available in the SCOP database [11]. We then use the same models to predict quaternary structural classes of protein complexes obtained from the 3D Complex database [22] which classifies them in terms of their number of subunits, topological contacts, graph images, and symmetries. Then, we choose the protein complexes having four subunits (Figure 2) and make multiclass and binary classification between these groups to understand how the topology of protein-protein contacts affect the success rate of the model. We also obtained virus structures from Virus Particle Explorer database (VIPERdb) [26] and applied the same classification techniques on proteins complexes in virus capsids which have different symmetry. Lastly, to study dynamics and structure of these protein complexes, we used DPD simulations on chains constructed from motifs which are reoccurring within these protein complexes. The ultimate goal is to develop a physics-based understanding of the oligomerization preferences of proteins in terms of their AACs and motifs.

2. MATERIALS AND METHODS

2.1 Data preparation and feature extraction

Three different datasets were used in this study. Firstly, for secondary structural classes of proteins, the same dataset originally utilized by Chou was used as a benchmark to compare against previous findings [5]. Secondly, the SCOP database (build 1.0.6, data retrieved on 10th of November 2020) [11] was used. Thirdly, for quaternary structural classes of proteins, the 3D complex database (version 6.0, data retrieved on 20th of July 2020) [22] which classifies the protein complexes in terms of topological classes was used. Lastly, for virus capsid proteins VIPERdb (version 3.0, data retrieved on 24th of May 2021) which classify the viruses based on their family, genus, and T-numbers. SCOP, 3D complex, and VIPERdb data were randomized and split into three groups; 60% for training, 20% for testing, and 20% for validation by using scikit-learn ML library [27] and the same datasets were used for different classifiers.

The Chou set has 120 proteins for training, 63 proteins for testing. We note that many of the proteins in this dataset have been replaced by newer models; there are additional amino acids that have been resolved in the new structures and the predictions in this work are made on the updated information. Moreover, the protein coded 1PHY which was included in the training set for β proteins was later superseded by 2PHY and reclassified as α/β type due to misinterpretation of the original data as a β clam fold [28]. We have excluded this protein from the training set which has led to the differences between the results for the current SVD application against the original study as discussed under the Results section. We list the updated dataset with their Protein Data Bank (PDB) [29] identifiers in Table S1 for the training set and in Table S2 for the test set.

FASTA sequences for Chou's dataset [5] were obtained from the Protein Data Bank [29]. FASTA files for proteins in the SCOP, 3D complex and VIPER databases were downloaded from the respective websites. Since complexes in 3D complex and VIPERdb contain more than one chain, the repeated sequences in the homomers were discarded from our analyses. Also, we only selected the chains no higher than 70% identity with other chains by using a bioinformatic tool called t_coffee [30].

For a set on *N* proteins, the AAC of each amino acid, *i*, for a given protein *k* is calculated through,

$$x_k = \frac{x_{k,i}}{chain\ length_k} \quad (k = 1,2,\ldots,N), (i = 1,2,\ldots,19) \tag{1}$$

Thus, a 19-dimensional x_k vector was generated for each protein,

$$x_k = [x_{k,1} \quad x_{k,2} \quad \cdots \quad x_{k,19}] \tag{2}$$

2.2 Singular Value Decomposition (SVD) algorithm

SVD is a supervised learning model and was originally used for the classification of topics in a given text based on the frequency of words [31], and later for the prediction of secondary structural classes of proteins [4]. We followed the approach in ref. [5], which is mathematically identical to the method used in ref. [4]. Thus, we construct the matrices

$$\mathrm{x}_k = \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{k19} \end{bmatrix}, \bar{\mathbf{x}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_{19}} \end{bmatrix}, \mathbf{S} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,19} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,19} \\ \vdots & \vdots & \ddots & \vdots \\ S_{19,1} & S_{19,2} & \cdots & S_{19,19} \end{bmatrix} (k = 1,2,\ldots,N) \tag{3}$$

where

$$S_{i,j} = \sum_{k=1}^{N} [x_{ki} - \bar{x}_i]\left[x_{kj} - \bar{x}_j\right] (i,j = 1,2,\ldots,19) \tag{4}$$

$$\mathrm{D}_2(\mathrm{x}, \bar{\mathrm{x}}) = (\mathrm{x} - \bar{\mathrm{x}})^{\mathrm{T}} \mathrm{S}^{-1} (\mathrm{x} - \bar{\mathrm{x}}) \tag{5}$$

The AACs are calculated for each of the *N* proteins. Then, the mean of AACs for each class (α, β, α+β, α/β) is calculated to be used for constructing the 19x19 covariance matrix, **S** (Equation 3 & 4). The class that has the minimum Mahalanobis distance to the test data is assigned as the predicted class for each test data using Equation 5 (Table S2).

2.3 k-Nearest Neighbor (kNN) algorithm

This method is used for many classification problems in finance, handwriting detection, and image recognition [32]. Compared to other ML methods, the training phase of the algorithm is much faster; however, this slows down the testing phase. When the only parameter, *k*, is set low, the model will be more flexible with low bias and high variance, when it is high, the decision boundaries will be much smoother resulting in high bias and low variance. Normalization of the data is crucial for kNN, since it relies on distance, similar to SVD. Since the AACs are already normalized with the chain length for each protein, there is no need for normalization in this case. 10-fold cross-validation method was used with the range of *k* values of 1 to 50.

2.4 Support Vector Machine (SVM) algorithm

SVMs are also amongst the supervised learning models used for regression and classification [33]. Beside linear classification, SVMs can also perform non-linear classification by using different kernels such as polynomial, radial distribution function, and sigmoid. The algorithm finds a hyperplane in an *N* dimensional space between each class by maximizing the margin to increase the confidence of the model.

There are two hyperparameters for SVMs. γ adjusts the influence of each training data; thus, when it is set to a high value, the classes will have low influence and vice versa. *C* acts as a regulation parameter for SVMs, trading off true classified training examples against maximization of the margin.

The validation method used in kNN was also used for the SVM model. For γ and C parameters, the sets [0.0002, 0.002, 0.02, 0.2, 1, 10] and [0.01, 0.1, 10, 100, 1000] are used respectively. The best performing parameters and kernel were chosen based on the F1 score of the fitted models. Radial basis function (RBF) kernel was the best performing kernel amongst other kernels for all models.

2.5 Creating H-P letter model and Determining the Motifs within complexes

In the literature there are many studies that utilize the DPD method to describe the dynamics of proteins using various degrees of detail in the modeling of the interactions [23-25, 34, 35]. In this study, we do not aim to model a particular protein, but rather to determine whether a stretch of selected two letter chain code is able to display distinguishing tetramerization profiles. For this purpose, first, chains which contain 20 letter alphabets of amino acids were transformed into 2 letter alphabets chains as H for aromatic and aliphatic residues and P for polar and charged residues. Classification of polar/nonpolar residues is based on a previous study [36]. Among 1800 chains in 4_1 complex and 577 chains in 4_5 complex, multiple HP motifs were identified by using MEME (Multiple Expectation Maximization Algorithm for Motif Elicitation) software. MEME can discover novel gapped motifs within the given sequences. Then, it statistically ranks the motifs based on their reoccurrences and widths [37].

To construct the sequences for DPD part of the project, one motif was chosen for each complex based on their significance and reoccurrence sites within chains (Table S4). We have selected motif six in 4_1 and motif five in 4_5 so that the repeats in the simulations would be similar. Then, 200 amino acid long chains were generated by making tandem repeats of these motifs (13 repeats for 4_1 and 12 repeats for 4_5) and considering AACs of these two complexes. The chains are capped with block copolymers of H and P units as listed in the Table S4.

2.6 Dissipative Particle Dynamics

DPD is a coarse-grained dynamics technique for mesoscopic systems such as different kids of lipid structures, polymers, complex fluids and so on and it is developed by Hoogerbrugge and Koelman in 1992 [38]. In this technique, each bead may represent one molecule or group of atoms and the motions of these beads are governed by Newton’s law of motion.

$$F_i = \sum_{j \neq i}(F_{ij}^C + F_{ij}^D + F_{ij}^R) \qquad (6)$$

There are 3 different forces acting on each bead: a conservative force, a dissipative force, and a random force (Equation 6). Conservative forces are soft repulsion that gives a chemical identity to each bead [39] and they are only effective within an adjusted cutoff radius. Dissipative forces and random forces act as a thermostat which keeps the temperature of the system constant. The main difference between DPD and Brownian dynamics is that each bead in DPD is under the effect of a random force independently [39].

In this study, each bead represents one amino acid molecule in the chain. According DPD theory, each bead should have similar radiuses to increase the confidence of the systems. To achieve this, we chose 5 water molecules to represent 1 solvent bead in the systems. Number of water molecules were selected by comparing average van der Waals volumes of amino acid molecules from their structure files and water molecules which were run 200 picoseconds of molecular dynamics simulation with NPT ensemble.

To calculate the interaction parameters between beads (H: hydrophobic, P: polar, S: water) for DPD simulations, first, solubility parameters (δ) should be calculated from MD simulations which were made with two different beads (H vs P, H vs S, P vs S).

$$\delta = (CED)^{1/2} \tag{7}$$

Cohesive Energy Density can be calculated from the relationship between solubility parameter (δ) which is from Hildebrand's definition (Equation 7). Then, ΔE mix is the mixing energy and it can be found as follows:

$$\Delta E_{mix} = \varphi_i (CED)_{ii} + \varphi_j (CED)_{jj} - (CED)_{ij} \tag{8}$$

φ_i and φ_j are the volume fractions of amino acid molecules and solvent in the mixture. Then, χ_{ij} which is the Flory-Huggins interaction parameter and the basis for DPD simulations can be calculated with following equation (V_{bead} is the average volume of the bead *i* and *j*. ΔE_{mix} is the mixing energy):

$$\chi_{ij} = \left(\frac{\Delta E_{mix}}{k_B T}\right) V_{bead} \tag{9}$$

To transform χ_{ij} parameter into DPD parameters, a relationship for a system which has 3 DPD units can be used (a_{ii} = 25) [40]:

$$a_{ij} \approx a_{ii} + 3.27\chi_{ij} \tag{10}$$

We previously calculated DPD parameters among amino acids and with water, however, since in our model we do not use each amino acid, we mostly focused on other studies which use H/P model or other hydrophobic and polar molecules, and used similar DPD parameter values for H-P, H-S, P-S interactions. Therefore, rather than individually parameterizing amino acids as two or three beads with individual parameters for the side chains [23-25, 35], we revert to a single bead representation for each amino acid fine-tuned for systems with number density $\rho = 3$ in DPD simulations [41-43].

Table 1 - DPD parameters for H, P, and water. Parameters were selected based on previous studies which also use H/P model for proteins or similar molecules. $a$ii = 25 is for neutral interaction. Above 25 is considered repulsive, belove 25 is considered attractive.

	H	**P**	**Water**
H	25.00		
P	30.00	25.00	
Water	35.00	28.00	25.00

For polar amino acid - water interactions, we use the value we had previously calculated for PIPOX units in water whereby the helical formations in water for these chains above their lower critical solution temperatures were well characterized [44].

For nonpolar amino acid - water interactions, we use a value typical for a hydrophobic monomer and a polar solvent, e.g. for styrene-DMF where $a_{\text{ij}} = 35$. [45]

Finally, for polar-nonpolar amino acid interactions, there are two competing interactions: All the backbones, irrespective of the identity of the side chains, have a tendency to hydrogen bond. On the other hand, the dissimilar side chains do not display a particular preference to interact and they tend to distance themselves from each other mainly due to the entropic cost of clustering. We therefore take a value intermediate to the above-mentioned two extremes and we set $a_{\text{ij}} = 30$. Similar DPD interaction parameters were calculated between hydrophobic and polar segments in polyurethanes [46, 47] and arylene ether sulphones [48].

After selecting the parameters between H, P, and water beads (Table 1), DPD chains were generated with identified motifs by also considering the AACs of H and P amino acids for 4_1 and 4_5 complexes. Then, 200 bead long four DPD chains were solvated in 200 x 200 x 200 Å^3 simulation box (800 beads for proteins, 52415 beads for water), and each system (5 replicas for each system listed) were run for 280 ns (1100000 DPD steps) with 254 fs time step at 298.0 K.

3. RESULTS AND DISCUSSION

3.1 Prediction of secondary structural classes of monomeric proteins.

We first set out to determine the extent to which kNN and SVM may predict the structural class of monomeric proteins. We list the individual Mahalanobis distances and assignments in the test set for the SVD method in Table S2; the comparison between the summarized results from all the methods are displayed in Table S3. For the original Chou dataset [5] these two algorithms display significantly better accuracy than SVD in predicting α/β class of proteins, whereas the reverse is true for the β class, although overall these models are weaker than SVD. Note that a single missing protein and the added residues in updated structures in this data set (see Methods for details) already significantly affects the predictions of the original SVD method for the accuracy of the α+β proteins.

Table 2 - Accuracy of SVD, kNN, and SVM for multiclass prediction of secondary structural classes of proteins with SCOP database. 18289 proteins for training, 6097 proteins for validation, 6097 proteins for testing were used for the models. (hyperparameters; k = 1 for kNN, γ = 0.2 and C = 10 for SVM)

Class	# of proteins in			Accuracy of		
	Training set	Validation set	Test set	SVD	kNN	SVM
α	3645	1221	1221	0.93	0.67	0.64
α+β	4970	1645	1645	0	0.55	0.49
α/β	4707	1630	1630	0	0.74	0.78
β	3929	1265	1265	0.11	0.60	0.63
s	1038	336	336	0	0.65	0.81
Total	**18289**	**6097**	**6097**			
F1-score (macro average)				**0.29**	**0.64**	**0.65**

To create a reliable ML model, we used SCOP database which has more than 30000 proteins. As the data in SCOP is well-balanced (Table 2), we have used accuracy for our classification metric; this also allows for comparing with previous studies which used accuracy over F1 scores. Finally, unlike the previous studies, the class of small proteins have also been included in the assignments, and the results are displayed in Table 2.

In this larger set, SVD was not able to distinguish between the structural classes since it assigned most proteins to α, decreasing the accuracy of the other classes (Table 2). This is because, the SVD model will always assign the class of the nearest protein for the class of the new test data due its dependence on the nearest Mahalanobis distance instead of selecting the cluster of that certain class. This choice increases the variance of the model while decreasing the bias, attested by the fact that when a new data point was added to our SVD model, the results changed dramatically. In contrast, modern ML classifiers such as kNN and SVM can be adjusted with hyperparameters to manipulate bias and variance rate. We find that for the protein structural class assignment, the best performing parameter for kNN is k = 1 in 10-fold cross validation. For SVM, the optimized hyperparameters are γ = 0.02 and C = 10. Both optimized models perform similarly well in terms of the average accuracy (Table 2; 0.64 for kNN, 0.65 for SVM).

Table 3 - Accuracies of SVM model for binary prediction of secondary structural classes of proteins. SVM was performed with the hyperparameters of γ = 0.2 and C = 10 for α vs. s and β vs. s; γ = 0.002 and C = 100 for α/β vs. s; γ = 0.2 and C = 1000 for α+β vs. s; γ = 0.2 and C = 1 for the rest.

			Accuracy of			**F1 score of**
	α	**β**	**α+β**	**α/β**	**s**	**One against others**
α	-					0.79
β	0.91	-				0.79
α+β	0.82	0.79	-			0.70
α/β	0.88	0.88	0.78	-		0.78
s	0.94	0.94	0.94	0.97	-	0.90

For the prediction of secondary structural classes of proteins, the best accuracy was obtained for the small proteins by SVM. The tertiary structures of these small proteins are formed by disulfide bridges, metal ligands, or cofactors which are highly residue specific for interaction; e.g., Cys for disulfide bridges; apparently, SVM is able to learn these better than kNN to get an accuracy of 0.81. On the other hand, the lowest accuracy attained is for α+β type proteins (0.55 by kNN and 0.49 by SVM), where they are frequently mis-assigned to α/β class. In contrast, α/β is amongst the best predicted classes.

Both ML techniques are adept in learning the features of amino acids leading to alternating patterns of α helices and β sheets, while they fall short of distinguishing a protein separated into all α and all β domains. We also carried the analysis with binary classification of each structural class with each other and one against others (Table 3). The worst performance was again for the classification of α+β and α/β with each other with the F1 score and accuracy of 0.78, and the best performance was again achieved with the small proteins with the accuracies of 0.96 for α, β, α+β proteins and 0.98 for α/β proteins. One against others method was also applied for each class (Table 3), because in multiclass prediction, the parameters for the SVM cannot be adjusted for each class individually. We then observed that some classes were indeed performing better with different γ and C values and this gives the information of how much these classes are distinguishable compared to other proteins with different classes in overall.

Overall, we find that even in the absence of sequence information, the amino acid composition of proteins already carries wealthy information on the structural class of a protein, with ca. 2/3 of proteins being correctly assigned using this information alone. Thus, regardless of having specific motifs, enrichment in certain amino acid types already signifies a certain class that will be selected in the three-dimensional structure. This is in alignment with the polymeric nature of proteins with the collapse to a given folded state being intimately related to the number of hydrophobic residues shielded from the solvent [49]. We next seek to determine the extent to which the oligomerization architectures of proteins are dictated by their AACs.

3.2 A Survey of AA types distributed in various architectures.

After predicting the five classes of secondary structural classes by only using AACs, we wanted to first see the significant differences between monomers, multimers and among different complexes. To do that, we have utilized the 3D complex database to make a survey of the distribution of AAs that display a tendency to form complexes as compared to the dataset of monomers which can be also found in 3D Complex database.

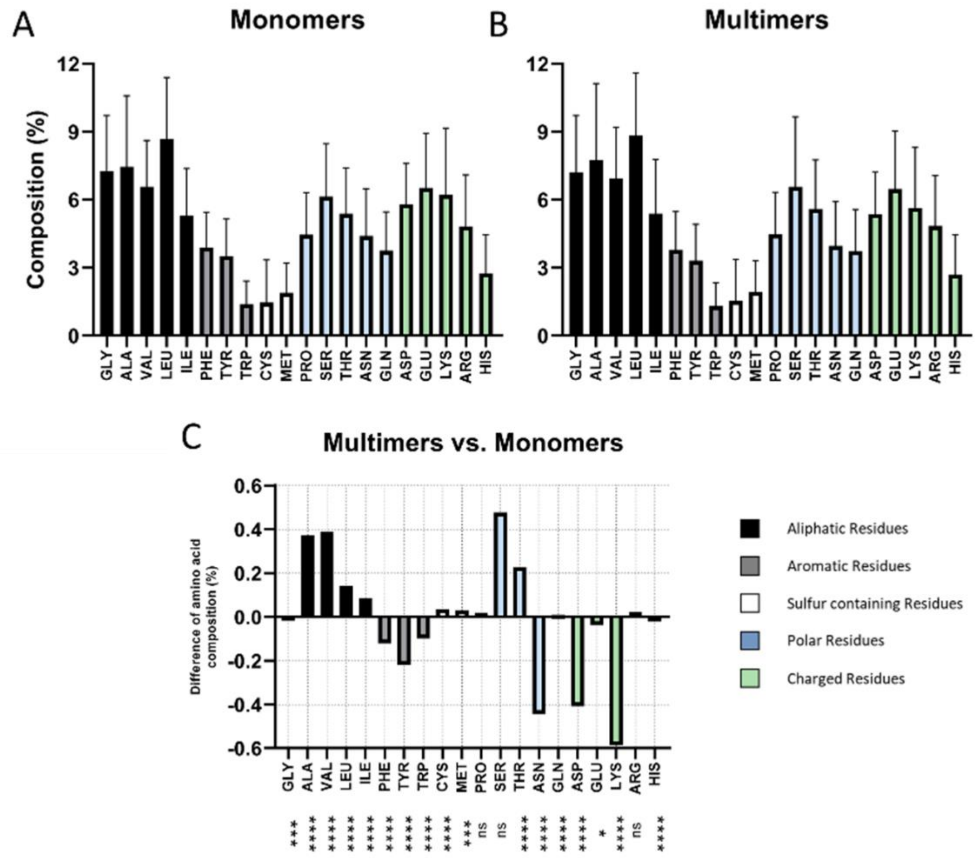
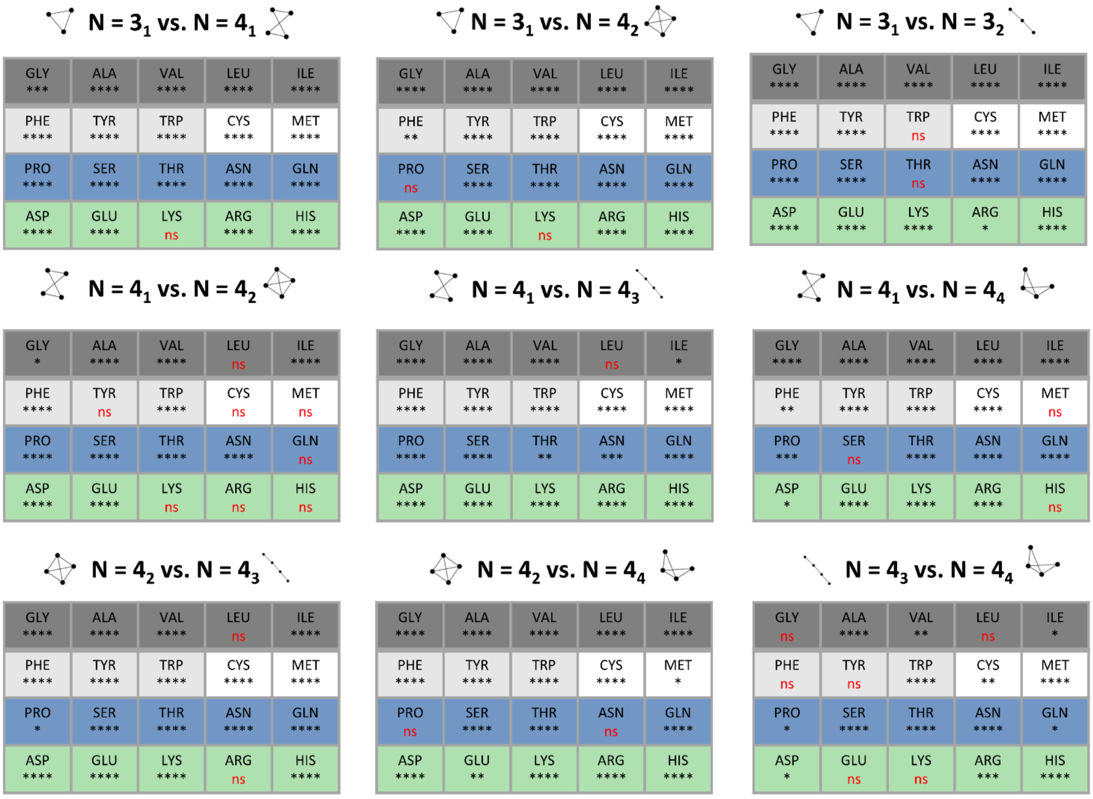


Figure 3 – AACs of proteins and their differences. (A) AACs of monomeric proteins in 3D Complex database, **(B)** AACs of multimeric proteins in 3D Complex database, **(C)** differences of AACs between multimeric and monomeric proteins. (p-values: ≤0.0001 = ****, ≤0.001 = ***, ≤0.01 = **, ≤0.05 = *, >0.05 = ns)

Figure 3 displays the AAC of the monomeric proteins (25437 proteins; Figure 3A) and multimeric complexes (48502 proteins; Figure 3B) in the 3D Complex database. The difference between each type of AA in monomeric vs. multimeric proteins are also displayed in Figure 3C. We find that there is a significant increase in the composition of aliphatic residues (e.g., Ala, Val, Leu, Ile) in the multimeric proteins with nonsignificant changes in polar residues (e.g., Pro, Ser). On the other hand, there is substantial decrease in the amount of charged residues (e.g., Asp, Glu, Lys, His). It was previously shown that intermolecular simple salt bridges (a non-bonded or hydrogen-bonded ion-paired interaction that joins a single pair of charged amino acid residues) are significantly less than intramolecular simple salt bridges [50]. In another study, it was shown that the charged residues were less conserved in the protein interaction interfaces compared to other amino acid groups [51], which may suggest the necessity to discard charged residues from the interface of multimeric proteins to form complexes. In fact, a recent study explains the complexation preferences of proteins by a hydrophobic ratchet model where neutral mutations have a tendency to drive interfaces to become more hydrophobic than water-exposed surfaces even when there is no apparent functional advantage [52]. This process is followed by purifying selection which entrenches the complex, simply because reverting to the monomeric form is destabilizing and opens the system to aggregation.



$N = 3_1$ vs. $N = 4_1$

GLY ***	ALA ****	VAL ****	LEU ****	ILE ****
PHE ****	TYR ****	TRP ****	CYS ****	MET ****
PRO ****	SER ****	THR ****	ASN ****	GLN ****
ASP ****	GLU ****	LYS ns	ARG ****	HIS ****

$N = 3_1$ vs. $N = 4_2$

GLY ****	ALA ****	VAL ****	LEU ****	ILE ****
PHE **	TYR ****	TRP ****	CYS ****	MET ****
PRO ns	SER ****	THR ****	ASN ****	GLN ****
ASP ****	GLU ****	LYS ns	ARG ****	HIS ****

$N = 3_1$ vs. $N = 3_2$

GLY ****	ALA ****	VAL ****	LEU ****	ILE ****
PHE ****	TYR ****	TRP ns	CYS ****	MET ****
PRO ****	SER ****	THR ns	ASN ****	GLN ****
ASP ****	GLU ****	LYS ****	ARG *	HIS ****

$N = 4_1$ vs. $N = 4_2$

GLY *	ALA ****	VAL ****	LEU ns	ILE ****
PHE ****	TYR ns	TRP ****	CYS ns	MET ns
PRO ****	SER ****	THR ****	ASN ****	GLN ns
ASP ****	GLU ****	LYS ns	ARG ns	HIS ns

$N = 4_1$ vs. $N = 4_3$

GLY ****	ALA ****	VAL ****	LEU ns	ILE *
PHE ****	TYR ****	TRP ****	CYS ****	MET ****
PRO ****	SER ****	THR **	ASN ***	GLN ****
ASP ****	GLU ****	LYS ****	ARG ****	HIS ****

$N = 4_1$ vs. $N = 4_4$

GLY ****	ALA ****	VAL ****	LEU ****	ILE ****
PHE **	TYR ****	TRP ****	CYS ****	MET ns
PRO ***	SER ns	THR ****	ASN ****	GLN ****
ASP *	GLU ****	LYS ****	ARG ****	HIS ns

$N = 4_2$ vs. $N = 4_3$

GLY ****	ALA ****	VAL ****	LEU ns	ILE ****
PHE ****	TYR ****	TRP ****	CYS ****	MET ****
PRO *	SER ****	THR ****	ASN ****	GLN ****
ASP ****	GLU ****	LYS ****	ARG ns	HIS ****

$N = 4_2$ vs. $N = 4_4$

GLY ****	ALA ****	VAL ****	LEU ****	ILE ****
PHE ****	TYR ****	TRP ****	CYS ****	MET *
PRO ns	SER ****	THR ****	ASN ns	GLN ****
ASP ****	GLU **	LYS ****	ARG ****	HIS ****

$N = 4_3$ vs. $N = 4_4$

GLY ns	ALA ****	VAL **	LEU ns	ILE *
PHE ns	TYR ns	TRP ****	CYS **	MET ****
PRO *	SER ****	THR ****	ASN ****	GLN *
ASP *	GLU ns	LYS ns	ARG ***	HIS ****

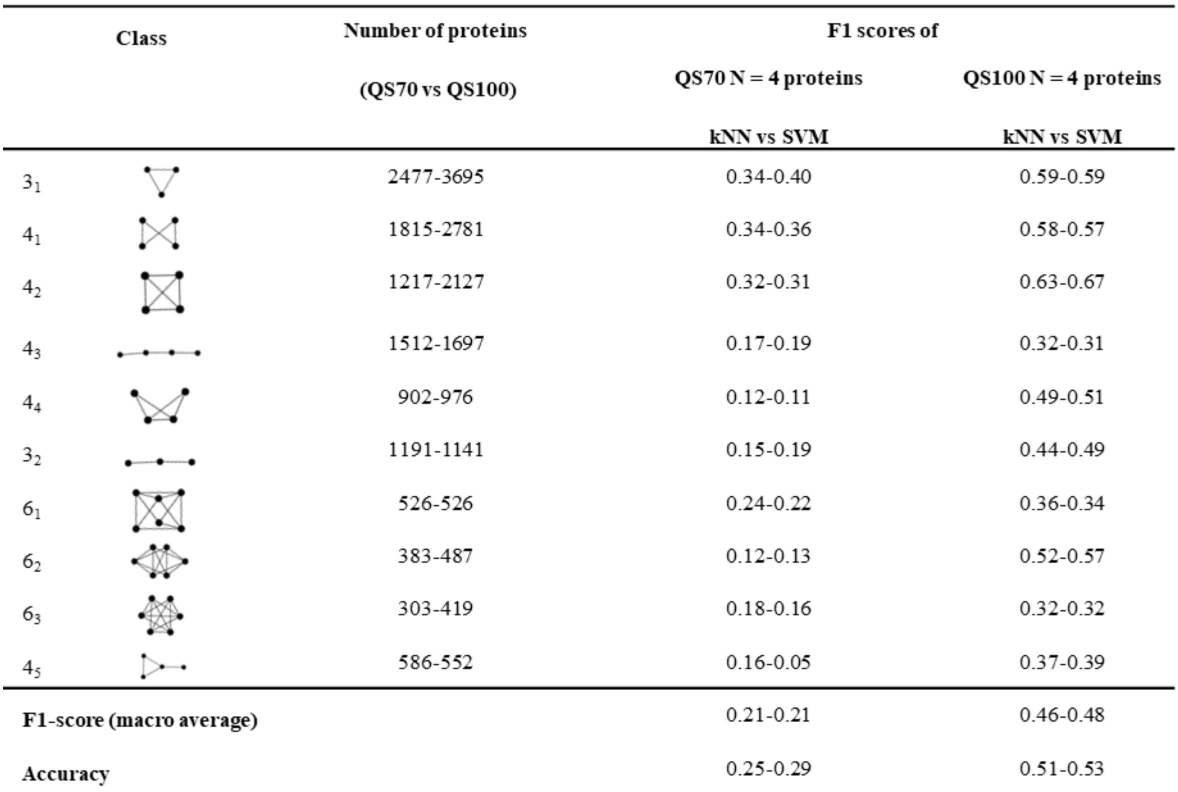
Figure 4 – Differences in AACs in different complexes. Differences in AAC between different subgroups in 3D Complex database. (dark gray = aliphatic residues, gray = aromatic residues, white = sulfur containing residues, blue = polar residues, green = charged residues) (p-values: ≤0.0001 = ****, ≤0.001 = ***, ≤0.01 = **, ≤0.05 = *, >0.05 = ns)

We also find differences in the AACs of various complexes of homomers of different numbers of subunits (Figure 4). Moreover, the interaction motifs between subgroups also contain variations, although some architectures (e.g., 4_1 vs. 4_2) have shown similar AACs for specific residues (Figure 4). Overall, we find the variations within the types of complexes distinctive enough to move on to more sophisticated classification methods.

3.3 Prediction of quaternary structural classes of multimeric proteins.

Survey have shown us that monomers, multimers and different protein complexes have significantly different AACs when compared. These differences differ one complex to another, suggesting that they can be used as feature for more sophisticated classification methods such as kNN and SVM.

Table 4 - F1-scores of kNN and SVM for multiclass prediction of quaternary structural classes of QS70 and QS100 complexes. Some classes were predicted with higher precision and recall resulting in higher F1-scores. (hyperparameters; *k* = 1 for kNN, γ = 0.2 and *C* = 10 for SVM).



Class		Number of proteins	F1 scores of	
		(QS70 vs QS100)	QS70 N = 4 proteins	QS100 N = 4 proteins
			kNN vs SVM	kNN vs SVM
3_1		2477-3695	0.34-0.40	0.59-0.59
4_1		1815-2781	0.34-0.36	0.58-0.57
4_2		1217-2127	0.32-0.31	0.63-0.67
4_3		1512-1697	0.17-0.19	0.32-0.31
4_4		902-976	0.12-0.11	0.49-0.51
3_2		1191-1141	0.15-0.19	0.44-0.49
6_1		526-526	0.24-0.22	0.36-0.34
6_2		383-487	0.12-0.13	0.52-0.57
6_3		303-419	0.18-0.16	0.32-0.32
4_5		586-552	0.16-0.05	0.37-0.39
F1-score (macro average)			0.21-0.21	0.46-0.48
Accuracy			0.25-0.29	0.51-0.53

The data collected from 3D complex database is unbalanced with few types of complexes having many representatives and most complexes having few representatives. Also, complexes having even numbers of subunits are favored. In our prediction scheme we have therefore focused on the types of complexes made of three or more proteins and having more than 300 representatives at the level of 100% identity (labelled QS100 in 3Dcomplex). This leads to 10 types of complexes on which predictions are performed (Table 4). Since QS100 data is highly biased, we also used QS70 data for our model. We note that the bias of QS100 comes from the nature itself, meaning it is likely that the sequence and structure of proteins which belong same type of quaternary structural class would be similar even though they have different functions. We refer this as “resampling of nature”.

Due to the unbalanced nature of the data, we have used the classification metric of F1-score which considers the proportion of each class in the model. We have optimized and performed both kNN and SVM. SVM slightly outperformed kNN with an average F1 score of 0.48 vs. 0.46 for QS100 data. We both present detailed results for kNN and SVM in what follows, for which the optimized parameters are $k = 1$ for kNN, and $\gamma = 0.2$ and $C = 10$ for SVM. Both models have similar prediction individual scores for the 10 most populated classes of multimeric proteins studied in this work.

The class 4_3, 6_1, 6_3 and 4_5 which have lower F1 scores (below 0.4) compared to other classes (up to a maximum of 0.67) were mostly classified as class 4_1, decreasing the average F1 score of the overall model. We rationalize this observation by realizing that since class 4_1 proteins interact at least with three other proteins, this may increase the AAC variation at the interfaces between the available surfaces in the training set. This extra variability is expected to make the 4_1 complexes in the test set prone to misclassification.

3.4 Tetrameric complexes in detail.

Since most of the multimers in nature formed by four or less subunits [22], we decided to focus on their prediction and structures. This would narrow down our perspective on importance of AACs on quaternary structure of proteins, so that, a better understanding can be achieved and would lead us to study their structures more specifically. We therefore selected the complexes having four subunits from 3D Complex and used QS70 and QS100 proteins for prediction with kNN and SVM. We first made a multiclass prediction of quaternary classes.

Table 5 - F1-scores and accuracy of SVM for multiclass prediction of quaternary structural classes of all, homomeric, and heteromeric tetrameric complexes. Some classes were predicted with higher precision and recall resulting higher F1-scores. (hyperparameters; $\gamma = 0.2$ and $C = 10$)

Class	Number of proteins (QS70 vs QS100)	F1 score of QS70 N = 4 proteins kNN vs SVM	F1 score of QS100 N = 4 proteins kNN vs SVM
4_1	1815 – 3117	0.38 - 0.47	0.70 - 0.70
4_2	1217 – 2391	0.39 - 0.43	0.72 - 0.71
4_3	1512 – 1920	0.29 - 0.42	0.49 - 0.39
4_4	902 – 1288	0.17 - 0.06	0.42 - 0.45
4_5	586 - 770	0.25 - 0.27	0.52 - 0.42
F1-score		**0.30 - 0.33**	**0.57 - 0.62**
Accuracy		0.32 - 0.40	0.62 - 0.62

We also found out that all possible architectures for tetrameric complexes (Figure 2) have high representation in the 3D complex database and are included in the sample set analyzed for multiclass classification (Table 4). We therefore further study this subset to understand the extent to which AACs are sufficient to classify them to the correct architecture. For this subset of multimers that contains five of the 10 classes listed in Table 4, the F1 score increases to 0.62 (Table 5).

To clarify, the number of proteins in common complexes in Table 4 and Table 5 is different due to removal of sequences which are present in multiple classes to remove any possible bias from our model.

Table 6 - F1-scores of SVM for binary prediction of quaternary structural classes of all, homomeric, and heteromeric tetrameric complexes. Some protein complexes were predicted with higher precision and recall when cross prediction was applied. For these results, we used five complexes having four subunits. γ and C hyperparameters differ in each classification.

	F1 score of					
	4_1 (2, 2, 2, 2)	4_2 (3, 3, 3, 3)	4_3 (1, 2, 2, 1)	4_4 (2, 3, 3, 2)	4_5 (1, 3, 2, 2)	One against others
4_1 (2, 2, 2, 2)	-					0.60
4_2 (3, 3, 3, 3)	0.59	-				0.66
4_3 (1, 2, 2, 1)	0.65	0.73	-			0.54
4_4 (2, 3, 3, 2)	0.54	0.67	0.52	-		0.52
4_5 (1, 3, 2, 2)	0.77	0.83	0.54	0.67	-	0.57

For binary prediction, we chose QS70 data due to its low bias and high variance over QS100. We think that the QS70 data would give us more statistically significant sequences for our structural analysis in the next part.

To do that, we tested the capacity of our model to make binary predictions on QS70 data, i.e. to select between pairs of possible tetrameric arrangements. We find some complexes to be more distinguishable by our model, reaching up to 0.83 F1-score (Table 6).

Interestingly, the capacity of our model to differentiate between two data sets of symmetric topology complexes that have the most type of amino acids with insignificant ACC differences (4_1 and 4_2 in Figure 4 where eight of the 20 amino acids have the same composition), is amongst the best performing. In fact, for prediction of quaternary structural classes of proteins, we have observed that protein complexes which have higher symmetry can be predicted with higher precision and recall. This was also observed for the binary classification between these complexes. We suggest that the complexes which have low symmetry, for example the 4_5 complex with the (1,3,2,2) topology, does not necessarily have to have similar sequences with other proteins within that complex. Because one protein only interacts with one protein, two proteins with two others, another one with all the other three, the ACC variation within these proteins is increased, adversely affecting the performance of the model. To further study these complexes, we chose 4_1 and 4_5 complexes for structural analysis because of their evident different symmetry with same total amount of interactions (total of 8) with other proteins.

3.5 Tetramer formation is observed in both complexes, but they have different structural properties.

After proving that AACs carry information for quaternary structures of proteins and seeing that this information is more pronounced between specific protein complexes, we identified the motifs which are reoccurring within these two complexes (4_1 and 4_5). Then, pseudo sequences (Table 7) were generated with the repetition of these motifs, and DPD simulations were applied (See methods for details).

Table 7 – 200 bead long sequences which were used for DPD simulations. Generated with tandem repeats of significantly reoccurring motifs within complexes. Tails are added to reach 200 bead length. Numbers 13 and 12 represent repetition number of motifs which are shown bolded. Other identified motifs can be found in Table S4.

Complex	Sequences
4_1	[PPPHHHHHH[**HPPHHPHHHPPHPP**]$_{13}$HHHHHHPPP]
$4_{1\ \text{reverse loop}}$	[HHHHHHPPP[**HPPHHPHHHPPHPP**]$_{13}$PPPHHHHHH]
4_5	[PPPPPPPPPPPHHHHH[**HPHHPPPHHHPPPH**]$_{12}$HHHHHPPPPPPPPPPP]
$4_{5\ \text{reverse loop}}$	[HHHHHPPPPPPPPPPP[**HPHHPPPHHHPPPH**]$_{12}$PPPPPPPPPPPHHHHH]

The sequences generated with tandem repeats of the identified motifs within complexes have total of 200 beads, thus the length of the chain would not make a structural difference. To complete the sequences into 200 bead long, we added H/P loops into both sides. In nature, loops of proteins usually made of polar residues, therefore, our main sequences are generated with polar heads on both termini [53].

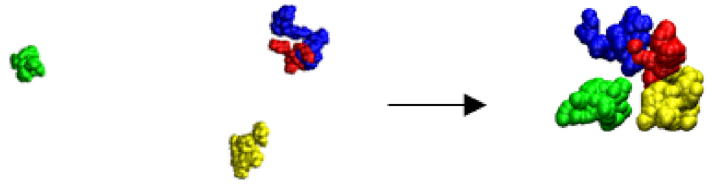


Figure 5 – Example of Tetramer formation in 4_1 complex. Tetramer formation occurred in all of the complexes. Tetramer formation is completed after 30th ns of simulation time.

As a result of DPD simulations, tetramer formation was observed in all systems (Figure 5). Assembly time of the chains differ from one to another with no significant differences, however, chains usually assemble one by one, and sometimes all together depending on the initial location of each chain. Tetramer formation is completed before 30th ns in all of the systems, and systems reach an equilibrium when tetramer is formed, and chains never disassemble.

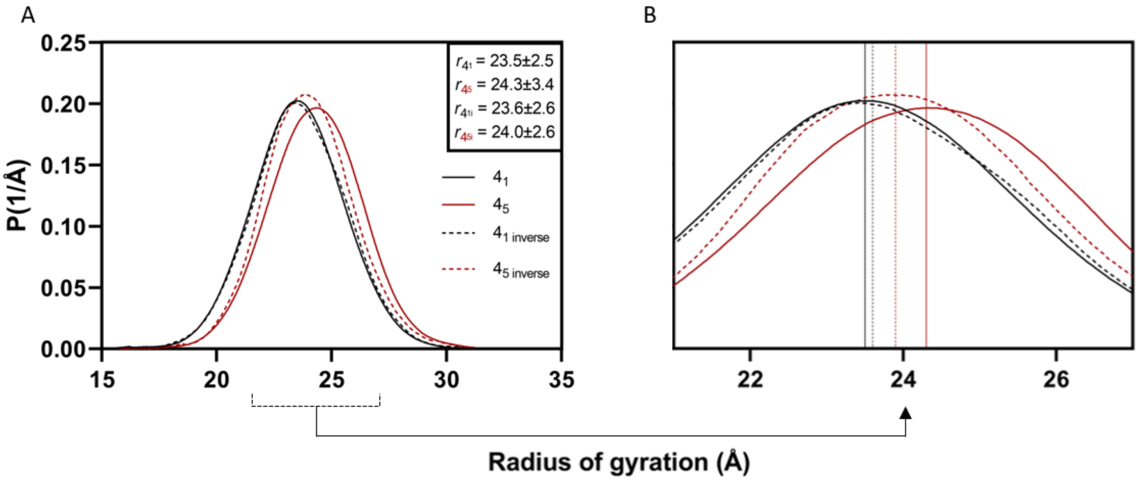


Figure 6 – Average Radius of Gyration (RoG) of both complexes and their inverted loop versions. (A) Average RoGs and radiuses of both complexes. **(B)** Peaks are zoomed in.

When trajectories were analyzed, even though these two complexes have similar amino acid compositions, it was observed that they have unique structural properties, suggesting motifs are also a significant factor for transitions in protein complexes. We first checked the compactness of the complexes with radius of gyration analysis. 4_1 complexes having 23.5±2.5 Å radius are significantly more compact when we compare them with 4_5 complex which have 24.3±3.4 (Figure 6), and this can be explained by 4_1's more symmetric structure. However, inreversing the loops into P to H beads make significant changes for 4_5 complex; shifting the structure to similar radius with 4_1 complex. This phenomenon can be explained with the longer polar ends of 4_5 complex, when not reversed, increases the radius of the complex due to free polar loops which make interactions with water. This is not observed when the loops of 4_5 is reversed: instead of making interaction with water, loops are more likely to fold into itself, which makes the overall structure more compact, however, it still does not reach the compactness of 4_1 complex, showing the effect of motifs instead of loops.

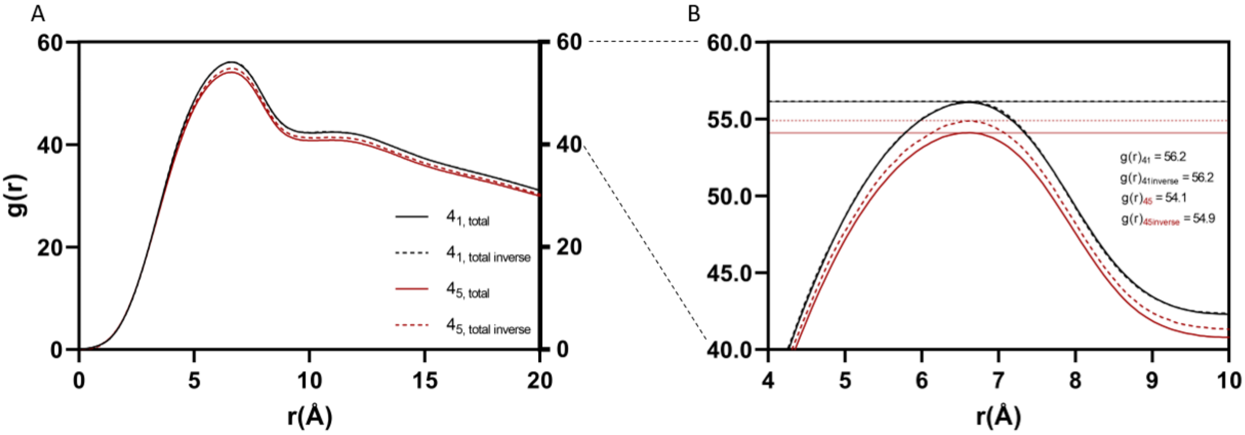


Figure 7 – Average Radial Distribution Function of both complexes. (A) Average RDFs of both complexes. **(B)** Peaks are zoomed in and g(r) values are shown.

To get an interaction-based information, radial distribution function (RDF) analyses were applied on all systems. RDF, *g(r)* is a function that describes the average density of particles around a particle, in our case bead, or some other reference point such as center of mass, within varying radiuses at given cutoff. When *g(r)* is high, we can estimate that, within that radius, the particles are closer to each other, forming a cluster.

Intramolecular and intermolecular RDFs have been checked together and separately. Also, we later checked interactions between H and P beads to get further information from the systems (Figure S1).

When intramolecular and intermolecular RDFs checked together, results have shown that 4_1 complexes have higher total amount of interactions between beads, even when the loops are reversed (Figure 7). Even though the total interaction increases when the loops of 4_5 complexes are reversed, they again cannot reach the amount of interactions of 4_1 complex.

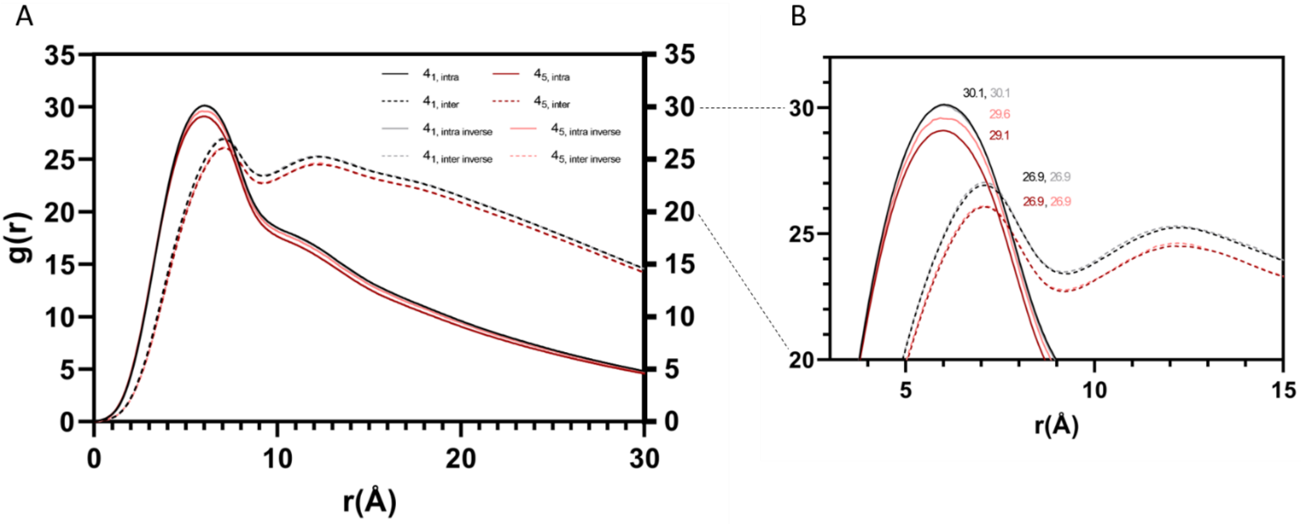


Figure 8 - Average Intramolecular and Intermolecular Radial Distribution Function of both complexes. Average intramolecular and intermolecular RDFs of both complexes. **(B)** Peaks are zoomed in and g(r) values are shown.

Next, we checked intramolecular and intermolecular RDFs separately (Figure 8). Overall intramolecular interaction higher in both complexes, proving that chains are more likely to fold into itself instead of making interaction with other chains. Comparison of two complexes shows that, again, both intermolecular and intramolecular interactions are higher in 4_1 complexes, and there is exception to this phenomenon when the loops are reversed.

Note that the symmetric structure of 4_1 complex can be again a proof to this result. In 4_1, all proteins make interactions with other two proteins within that complex, however, in

4_5, number of interactions for each protein is different, which lowers the symmetry of this complex, and the interactions between proteins (Figure 2).

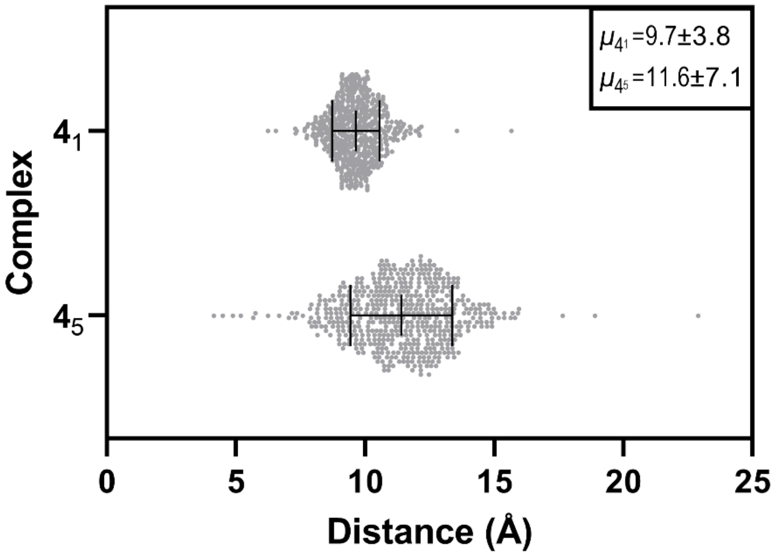


Figure 9 - Distance calculation between center of mass of each chain and center of mass of complex. The distance measurement was done throughout the trajectory, after tetramer is formed. (after 30th ns)

To check whether these tetramer formations have resemblances with real topologies on 3D complex, we measured the distance between center of mass of each chain and center of max of tetramer throughout the trajectory after the tetramer is formed. From our calculation, we saw that the distances between these two reference points are significantly smaller in 4_1 complex compared to 4_5 complex (p-value < 0.0001). Average distance for 4_1 complex is 9.7 ± 3.8 Å, and 11.6 ± 7.1 Å for 4_5 (Figure 9). This was expected due to more symmetric topology of 4_1 complex compared to 4_5. Less symmetric topology of 4_5 would have longer distances between these reference points, because the chain which makes interaction with only one protein would be in a much further distance away from center of mass of the complex (see Figure 2).

3.6 Virus capsids which follow different symmetry are also distinctive.

The results we obtained from tetramer complexes showed that along with AACs, motifs were also important for specific multimer formation of protein complexes. Thus, we wanted to go further with our analysis and see whether virus capsid structures can be also predictable with their AACs which might be more pronounced within their complexes due to their repetitive oligomer formation. In addition to that, they have also higher surface area of protein-protein interfaces which highly affect their AACs.

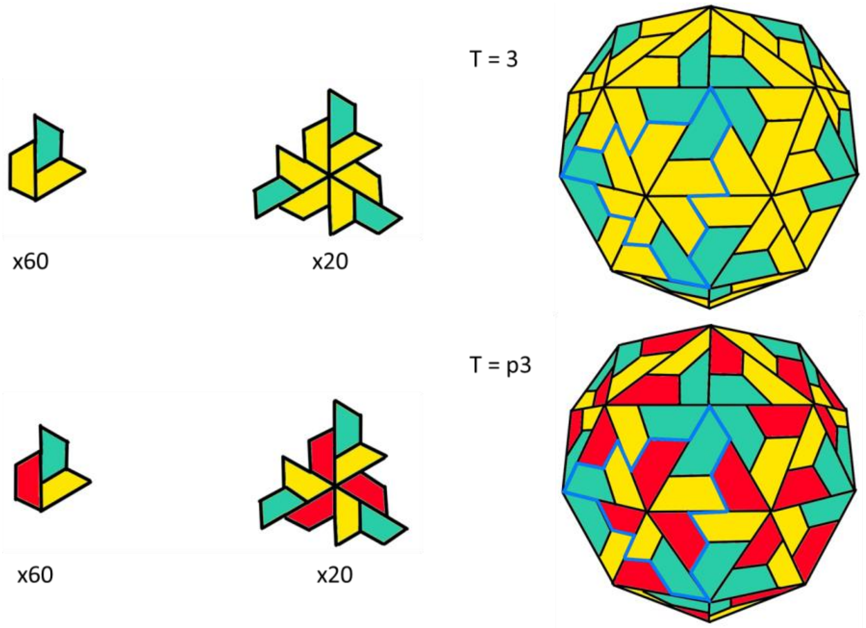


Figure 10 - Two different icosahedral capsid structures classified based on their T numbers. T = 3 and T = p3 icosahedral capsid structures. The only difference between them is number of different proteins in their basic unit.

Virus capsids are made of repeated protein complexes that act as a shell by enclosing the genetic material. In our previous results we stated that the number of amino acids interacting with other proteins should be less than the amino acids which form secondary structure of proteins. This highly affects the prediction of quaternary structural classes of proteins using AACs. However, since virus capsids are made of repeating protein complexes, we considered that the evolutionary information of AAC might be carried more efficiently in these complexes. Therefore, we also made the prediction of distinguishing from each other two virus capsid structure proteins which were classified based on their T number to prove our point.

Table 8 - F1-scores and accuracy of kNN and SVM for prediction of two different virus capsid proteins. For kNN, *k* = 3 for QS70, *k* = 2 for QS100 were used. For SVM *γ* = 0.02 and *C* = 10 were used.

Class	**Number of proteins**		**F1 score of QS70 proteins kNN vs SVM**	**F1 score of QS100 proteins kNN vs SVM**
	QS70	**QS100**		
T = 3	90	199	0.84 - 0.75	0.86 - 0.81
T = p3	178	580	0.94 - 0.89	0.94 - 0.95
F1-score (macro average)			**0.89 - 0.82**	**0.91 - 0.88**
Accuracy			0.91 - 0.85	0. 93 - 0.91

The triangulation (T) number defines the size and complexity of the capsids. In this project, we focused on two different virus capsid proteins: T = 3, and T = p3 (Figure 9). The reason we chose these two classes is because the size and the complexity of these two capsids are same, however, T = 3 capsid are made of two different chains, while T = p3 capsids are made of three different chains. This is the reason why they are called pseudo T = 3 capsids. After extracting AACs of these proteins, we applied kNN and SVM to see how distinguishable they are. As we expected, the model was performing better compared to the binary prediction of protein complexes having four proteins, up to F1-score of 0.89 with QS70 dataset (Table 8), proving the information of AACs which defines the quaternary structure of protein complexes are more pronounced in virus capsids compared to regular protein complexes.

4. CONLUSION AND FUTURE WORK

In conclusion, we first showed that AAC of proteins are also significant for determining the quaternary structure of proteins as it is for secondary structures. The reason we have obtained worse results for quaternary structural class prediction compared to secondary structural class prediction is because AACs carry information for all residues which contribute all secondary structures of proteins; however, quaternary structures mostly depend on protein-protein interface residues. Yet, we can classify quaternary structural classes with more than 80% F1 score (Table 6), and this suggests AACs residing at the protein interaction surfaces are also effective to form different complexes. Moreover, the symmetry of these protein complexes is highly related to the proportion of their amino acids, which also makes them more predictable with ML models (Table 5). In addition, virus capsid proteins which are highly symmetric with different structural classes can be distinguished with even better performance with ML models (Table 8). We think that this is due to their nature of need to assemble in a host cell, increasing the importance of AACs within capsid proteins. We also note that more sophisticated prediction tools such convolutional neural networks can be also used to get higher performance. Thus, weight of each amino acids in a learning model can be adjusted differently from each other. We know that protein-protein interfaces mostly contain nonpolar residues, meaning they should have higher weights in a learning model.

Along with AACs, we discovered that motifs which are reoccurring within same quaternary structural class are also significant for transitions of protein complexes. Motifs affect the compactness, symmetry, and interaction preferences of complexes, making them again distinguishable in structural level (Figure 6-8). Also, we were able to show similar topologies for tetramers with our distance calculation which supports this point (Figure 9). These discovered motifs can be also used as features for ML models, so that, higher performance might be achieved.

We were able to identify H/P motifs within these virus capsid proteins in QS100 data (Table S5), however, we did not want to use them for DPD simulations due to presence of homologous proteins in that data. No significant H/P motifs were identified using QS70 data of virus capsid proteins. We think that the small number of proteins in QS70 data was the reason for this.

Next, we want to add more virus capsid classes to our data to identify motifs and run DPD simulation with them. We are aiming to end up with sphere structures with hollow inside in the end like real capsid structures.

REFERENCES

1. Alberts, B., *The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists.* Cell, 1998. **92**(3): p. 291-294.
2. John Jumper, R.E., Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Martin Steinegger, Michalina Pacholska, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis., *High Accuracy Protein Structure Prediction Using Deep Learning.* 2020.
3. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nat Protoc, 2010. **5**(4): p. 725-38.
4. Bahar, I., et al., *Understanding the recognition of protein structural classes by amino acid composition.* Proteins: Structure, Function, and Genetics, 1997. **29**(2): p. 172-185.
5. Chou, K.-C., *Does the folding type of a protein depend on its amino acid composition?* FEBS Letters, 1995. **363**(1-2): p. 127-131.
6. Lee, S., B.C. Lee, and D. Kim, *Prediction of protein secondary structure content using amino acid composition and evolutionary information.* Proteins, 2006. **62**(4): p. 1107-14.
7. Rizzetto, S., et al., *Context-dependent prediction of protein complexes by SiComPre.* NPJ Syst Biol Appl, 2018. **4**: p. 37.
8. Roy, S., et al., *Exploiting amino acid composition for predicting protein-protein interactions.* PLoS One, 2009. **4**(11): p. e7813.
9. Zhang, C.T. and K.C. Chou, *An optimization approach to predicting protein structural class from amino acid composition.* Protein Sci, 1992. **1**(3): p. 401-8.
10. Zhang, Q.C., et al., *Structure-based prediction of protein-protein interactions on a genome-wide scale.* Nature, 2012. **490**(7421): p. 556-60.
11. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
12. Cai, Y.D., et al., *Support vector machines for predicting protein structural class.* BMC Bioinformatics, 2001. **2**: p. 3.
13. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.
14. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks.* Bioinformatics, 2001. **17**(4): p. 349-58.
15. Chou, K.C. and Y.D. Cai, *Predicting protein quaternary structure by pseudo amino acid composition.* Proteins, 2003. **53**(2): p. 282-9.
16. Aytuna, A.S., A. Gursoy, and O. Keskin, *Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.* Bioinformatics, 2005. **21**(12): p. 2850-5.
17. Keskin, O., et al., *Principles of protein-protein interactions: what are the preferred ways for proteins to interact?* Chem Rev, 2008. **108**(4): p. 1225-44.
18. Keskin, O., B. Ma, and R. Nussinov, *Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues.* J Mol Biol, 2005. **345**(5): p. 1281-94.

19. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM.* Nat Protoc, 2011. **6**(9): p. 1341-54.
20. Atilgan, C. and A.R. Atilgan, *Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein.* PLoS Comput Biol, 2009. **5**(10): p. e1000544.
21. Abdizadeh, H., et al., *A Coarse-Grained Methodology Identifies Intrinsic Mechanisms That Dissociate Interacting Protein Pairs.* Front Mol Biosci, 2020. **7**: p. 210.
22. Levy, E.D., et al., *3D complex: a structural classification of protein complexes.* PLoS Comput Biol, 2006. **2**(11): p. e155.
23. Vishnyakov, A., D.S. Talaga, and A.V. Neimark, *DPD Simulation of Protein Conformations: From alpha-Helices to beta-Structures.* J Phys Chem Lett, 2012. **3**(21): p. 3081-7.
24. Li, C., et al., *Dissipative Particle Dynamics Simulations of a Protein-Directed Self-Assembly of Nanoparticles.* ACS Omega, 2019. **4**(6): p. 10216-10224.
25. Okuwaki, K., et al., *Folding simulation of small proteins by dissipative particle dynamics (DPD) with non-empirical interaction parameters based on fragment molecular orbital calculations.* Applied Physics Express, 2020. **13**(1).
26. Montiel-Garcia, D., et al., *VIPERdb v3.0: a structure-based data analytics platform for viral capsids.* Nucleic Acids Res, 2021. **49**(D1): p. D809-D816.
27. Fabian Pedregosa, G.V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2012.
28. Borgstahl, G.E., D.R. Williams, and E.D. Getzoff, *1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore.* Biochemistry, 1995. **34**(19): p. 6278-87.
29. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
30. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
31. Berry, M.W., S.T. Dumais, and G.W. O’Brien, *Using Linear Algebra for Intelligent Information Retrieval.* SIAM Review, 1995. **37**(4): p. 573-595.
32. Altman, N.S., *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression.* The American Statistician, 1992. **46**(3): p. 175-185.
33. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.
34. Okuwaki, K., et al., *Theoretical analyses on water cluster structures in polymer electrolyte membrane by using dissipative particle dynamics simulations with fragment molecular orbital based effective parameters.* RSC Advances, 2018. **8**(60): p. 34582-34595.
35. Peter, E.K., K. Lykov, and I.V. Pivkin, *A polarizable coarse-grained protein model for dissipative particle dynamics.* Phys Chem Chem Phys, 2015. **17**(37): p. 24452-61.
36. Frihart, C.R. and M.J. Birkeland, *Soy Properties and Soy Wood Adhesives*, in *Soy-Based Chemicals and Materials*. 2014. p. 167-192.

37. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.

38. Hoogerbrugge, P.J. and J.M.V.A. Koelman, *Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics.* Europhysics Letters (EPL), 1992. **19**(3): p. 155-160.

39. Tan, H., et al., *A dissipative particle dynamics simulation study on phase diagrams for the self-assembly of amphiphilic hyperbranched multiarm copolymers in various solvents.* Soft Matter, 2017. **13**(36): p. 6178-6188.

40. Groot, R.D. and P.B. Warren, *Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation.* The Journal of Chemical Physics, 1997. **107**(11): p. 4423-4435.

41. Can, H., G. Kacar, and C. Atilgan, *Surfactant formation efficiency of fluorocarbon-hydrocarbon oligomers in supercritical CO2.* J Chem Phys, 2009. **131**(12): p. 124701.

42. Kacar, G., C. Atilgan, and A.S. Özen, *Mapping and Reverse-Mapping of the Morphologies for a Molecular Understanding of the Self-Assembly of Fluorinated Block Copolymers.* The Journal of Physical Chemistry C, 2009. **114**(1): p. 370-382.

43. Ozen, A.S., U. Sen, and C. Atilgan, *Complete mapping of the morphologies of some linear and graft fluorinated co-oligomers in an aprotic solvent by dissipative particle dynamics.* J Chem Phys, 2006. **124**(6): p. 64905.

44. Furuncuoğlu Özaltın, T., et al., *Multiscale modeling of poly(2-isopropyl-2-oxazoline) chains in aqueous solution.* European Polymer Journal, 2017. **88**: p. 594-604.

45. Ozden-Yenigun, E., et al., *Molecular basis for solvent dependent morphologies observed on electrosprayed surfaces.* Phys Chem Chem Phys, 2013. **15**(41): p. 17862-72.

46. Avaz, S., et al., *Soft segment length controls morphology of poly(ethylene oxide) based segmented poly(urethane-urea) copolymers in a binary solvent.* Computational Materials Science, 2017. **138**: p. 58-69.

47. Avaz Seven, S., et al., *Tuning Interaction Parameters of Thermoplastic Polyurethanes in a Binary Solvent To Achieve Precise Control over Microphase Separation.* J Chem Inf Model, 2019. **59**(5): p. 1946-1956.

48. Sevinis Ozbulut, E.B., et al., *Blends of highly branched and linear poly(arylene ether sulfone)s: Multiscale effect of the degree of branching on the morphology and mechanical properties.* Polymer, 2020. **188**.

49. Munson, M., et al., *What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties.* Protein Sci, 1996. **5**(8): p. 1584-93.

50. Musafia, B., V. Buchner, and D. Arad, *Complex salt bridges in proteins: statistical analysis of structure and function.* J Mol Biol, 1995. **254**(4): p. 761-70.

51. Guharoy, M. and P. Chakrabarti, *Conserved residue clusters at protein-protein interfaces and their use in binding site identification.* BMC Bioinformatics, 2010. **11**: p. 286.

52. Hochberg, G.K.A., et al., *A hydrophobic ratchet entrenches molecular complexes.* Nature, 2020. **588**(7838): p. 503-508.

53. Jacob, E. and R. Unger, *A tale of two tails: why are terminal residues of proteins exposed?* Bioinformatics, 2007. **23**(2): p. e225-30.

SUPPLEMENTARY INFORMATION

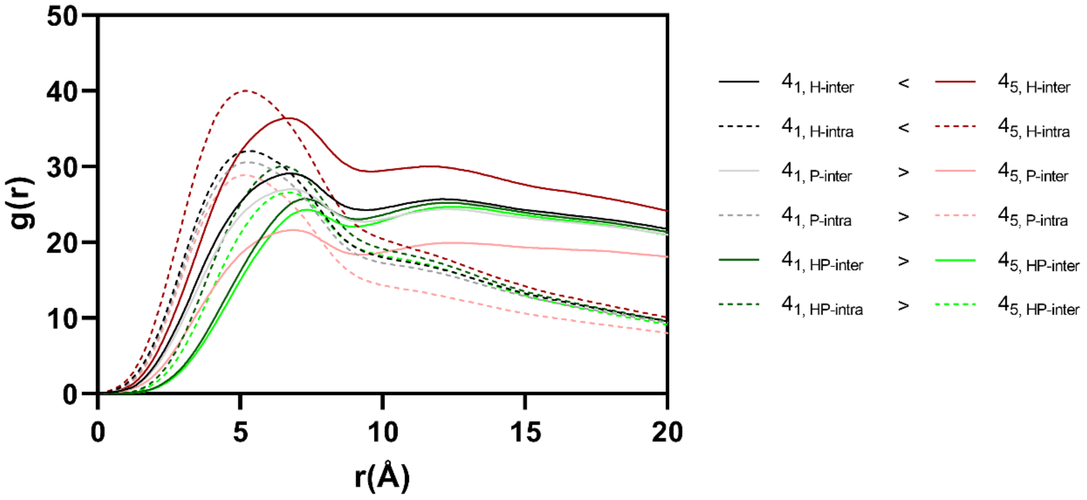


Figure S1 - Average Intra and Intermolecular Radial Distribution Function of H and P beads

Table S1 – List of proteins used as the training set in ref. [5] and replacements for obsolete structures; * indicates unchanged PDB id.

[a] reclassified as α/β type in ref. [28]; excluded from current analyses.

α		β		α+β		α/β	
ORIGINAL PDB	**UPDATED PDB**	ORIGINAL PDB	**UPDATED PDB**	ORIGINAL PDB	**UPDATED PDB**	ORIGINAL PDB	**UPDATED PDB**
1AVHA	*	1ACX-	*	1AAK-	2AAK-	1ABA-	*
1BABB	*	1AYH-	2AYH-	1CTF-	*	1CIS-	*
1BRD-	*	1CD8-	*	1DNKA	*	1CSEI	*
1C5A-	*	1CDTA	*	1EAF-	*	1CTC-	*
1CPCA	*	1CID-	*	1HSBA	*	1DHR-	*
1CPCL	*	1DFNA	*	1LTSA	*	1DRI-	2DRI-
1ECO-	*	1HILA	*	1LTSD	*	1ETU-	*
1FCS-	*	1HIVA	*	1NRCA	1OIAA	1FX1-	*
1FHA-	*	1HLEB	*	1OVB-	*	1GPB-	*
1FIAB	*	1MAMH	*	1POC-	*	1OFV-	*
1HBG-	*	1MONA	*	1PPN-	*	1PAZ-	*
1HDDC	*	1OMF-	2OMF-	1PRF-	2PRF-	1PFKA	*
1HIGA	*	1PHY-	2PHY-[a]	1RND-	*	1PGD	2PGD
1LE4-	*	1REIA	*	1SNC-	*	1Q21	*
1LIG-	2LIG-	1TEN-	*	1TFG-	*	1S01-	*
1LTSC	*	1TLK-	*	1TGSI	*	1SBP-	*
1MBC-	*	1VAAB	2VAAB	2ACHA	*	1SBT-	*
1MBS-	*	2ALP-	*	2ACT-	*	1TIMA	*
1RPRA	*	2AVIA	*	2BPA1	*	1TMD-	2TMD-
1TROA	*	2BPA2	*	2SNS-	*	1TREA	*
1UTG-	*	2HHRC	3HHRC	2SSI-	3SSI-	1ULA-	*
256BA	*	2ILA-	*	3IL8-	*	1WSYB	*
2CCYA	*	2LALA	*	3RUBS	*	2HAD-	*
2LH1-	*	2SNV-	*	3SGBI	*	2LIV-	*
2LHB-	*	3CD4A	*	3SICI	*	3GBP-	*
2MHBA	*	4GCR-	*	4BLMA	*	4FXN-	2FOX-
2MHBB	*	7APIB	*	4TMS-	*	5CPA-	*
2ZTAA	*	8I1B-	*	8CATA	*	5P21-	*
4MBA-	*	8FABA	*	9RNT-	*	8ABP-	*
4MBN-	*	8FABB	*	9RSAA	*	8ATCA	*

Table S2 – List of proteins used as the testing set in ref. [5] and their Mahalanobis distances calculated with SVD method.

PDB code (original PDB)	Mahalanobis Distances				Observed	Predicted
	$D^2(X,\bar{X}_\alpha)$	$D^2(X,\bar{X}_\beta)$	$D^2(X,\bar{X}_{\alpha+\beta})$	$D^2(X,\bar{X}_{\alpha/\beta})$		
1BBL-	1.60	3.09	3.72	10.07	α	**α**
1HBBA	0.86	4.18	2.02	6.85	α	**α**
1IFA-	2.45	2.67	3.83	3.61	α	**α**
1MRRA	0.93	0.54	0.91	0.91	α	β
2PDE- (1PDE-)	3.07	4.00	5.05	5.59	α	**α**
1PRCM	4.62	4.99	5.23	9.98	α	**α**
2SAS- (1SAS-)	3.37	3.93	2.98	1.91	α	α/β
2TMVP	2.91	1.51	2.45	12.17	α	β
4CPV-	1.73	7.99	5.15	8.94	α	**α**
2AAIB (1AAIB)	6.81	2.99	2.31	16.15	β	α + β
1ATX-	25.77	5.20	9.65	68.26	β	**β**
1COBA	7.02	3.28	4.86	4.17	β	**β**
1EGF-	18.90	2.79	4.90	38.13	β	**β**
1EST	7.38	1.46	3.94	9.63	β	**β**
1GPS-	18.22	5.73	15.45	128.14	β	**β**
1HCC-	9.49	2.62	4.46	12.30	β	**β**
1IXA-	18.21	7.17	11.32	73.72	β	**β**
1MDAA	4.16	2.30	2.61	5.76	β	**β**
1PPFE	4.44	1.99	7.14	14.50	β	**β**
1R1A2	3.75	1.22	2.22	2.41	β	**β**
1SHFA	5.70	0.65	2.55	3.84	β	**β**
1TIE-	2.77	0.70	1.77	3.67	β	**β**
1TNFA	2.79	1.08	1.17	5.32	β	**β**
1ACHB (2ACHB)	8.73	2.58	6.08	29.47	β	**β**
2CTX-	12.61	3.30	5.99	109.80	β	**β**
2MEV1	2.43	0.56	2.93	4.71	β	**β**
2PLV1	1.78	0.39	2.21	1.85	β	**β**
2SODO	7.19	3.89	5.17	4.67	β	**β**
3RP2A	1.63	0.61	1.41	3.22	β	**β**
4SGBI	8.99	4.22	6.63	109.74	β	**β**
5NN9-	10.35	1.49	1.46	11.83	β	α + β
2ABH- (2ABH-)	2.07	2.00	1.11	1.26	α + β	**α + β**
1BBPA	6.77	1.75	2.62	7.48	α + β	β
1BW4-	8.28	4.35	1.80	15.53	α + β	**α + β**
3COX- (1COX-)	3.31	0.79	0.81	1.42	α + β	β
1DNKA	1.02	0.83	0.86	1.57	α + β	β
1GLAG	4.08	0.97	1.01	0.94	α + β	α/β
2MS2A (1MS2A)	3.16	1.77	0.92	6.52	α + β	**α + β**
1OVOA	3.90	3.83	1.34	52.77	α + β	**α + β**
1POC-	7.18	2.85	0.69	13.38	α + β	**α + β**
2PPBA	2.62	1.54	1.75	5.32	α + β	β
1SHAA	1.04	2.59	1.73	6.87	α + β	α
1THO-	2.28	3.44	0.89	4.18	α + β	**α + β**
1XOB (1TRX-)	2.28	2.84	1.17	5.31	α + β	**α + β**
2AAA-	4.11	2.22	2.01	5.86	α + β	**α + β**
2PIA-	2.61	0.77	0.79	6.66	α + β	β
2SN3-	10.75	5.81	3.26	62.67	α + β	**α + β**
2TAAA	2.67	0.72	0.55	3.90	α + β	**α + β**
3B5C-	9.23	2.03	2.81	17.62	α + β	β
3SC2A	4.07	0.58	0.42	1.81	α + β	**α + β**
3SC2B	6.11	1.30	1.32	6.46	α + β	β
8TLN-	3.68	0.45	0.64	2.18	α + β	β
4ENL-	0.50	1.51	0.49	1.15	α + β	**α + β**
4INSB	6.03	19.73	4.36	13.27	α + β	**α + β**
4RCRH	2.71	1.12	0.83	0.94	α + β	**α + β**
1GPB-	0.97	0.62	1.02	0.50	α/β	**α/β**
2MINA (1MINA)	2.85	1.07	1.20	0.69	α/β	**α/β**
1NIPB	1.49	1.24	6.49	0.98	α/β	**α/β**
1SBP-	1.81	1.96	0.75	0.52	α/β	**α/β**
1BKS (1WSYA)	4.16	1.56	1.00	2.98	α/β	α + β
4ICD	1.36	1.12	1.73	0.82	α/β	**α/β**
7AATA	0.94	1.34	0.65	0.76	α/β	α + β
9RUBB	2.44	1.28	0.84	0.88	α/β	α + β
1GD1O	2.68	1.10	2.90	0.79	α/β	**α/β**

Table S3 - Accuracy of SVD, kNN, and SVM for prediction of secondary structural classes of proteins with Chou's data. 119 proteins for training, 64 proteins for testing were used for the models. Some proteins structures and sequences had been updated throughout the years causing small changes on results for SVD. kNN and SVM display similar overall accuracy.

	Accuracy			
	reported for SVD (Bahar et al., 1997)	**SVD** (this work)	**kNN**	**SVM**
α	0.67	0.67	0.67	0.67
α+β	0.81	0.58	0.54	0.63
α/β	0.67	0.67	0.89	1
β	0.90	0.91	0.67	0.50
Average	**0.81**	**0.72**	**0.63**	**0.64**

Table S4 – Motifs identified for each tetramer complex with their width, sites and p-values. Nonhomologous QS70 data was used.

4_1 (2,2,2,2)	**Sequence**	**Width**	**Site**	**p-value**
Motif 1	PHHHHHHPPPHHPPHHPPHHPHPHHHHHHH	30	158	9.2e-004
Motif 2	HHHPPHHHPHHPPHH	15	223	1.1e-003
Motif 3	HPHPPHHPHHHHHHPPPHPHHPHPHHPHP	29	159	2.1e-003
Motif 4	HPHHHPPHHPHHHHPPHPH	19	166	1.5e-002
Motif 5	HPPPPPPPPPPHHHHPHPPHHPHPPH	26	67	1.6e-002
Motif 6	PPHPPHHPPH	10	517	2.5e-002
Motif 7	PPPPPPPPHHPHHPPPHPHPH	21	51	3.1e-002
Motif 8	HPPHHPHHHPPHPP	14	326	4.3e-002
4_5 (1,3,2,2)	**Sequence**	**Width**	**Site**	**p-value**
Motif 1	PPHHHHPHHHHHHPPHPHPP	20	54	9.8e-004
Motif 2	HHPPHPPPPPHPPPPPPPPPHPHPP	25	68	2.0e-003
Motif 3	HPPHPHPHPHPPHPHHHPPHPPPP	24	92	3.9e-003
Motif 4	HPHHPHHPHHPHPPHPHPHHHPH	23	80	5.9e-003
Motif 5	HPHHPPPHHHPPPH	14	134	5.9e-003
Motif 6	PHHPPHHPHPHHHPH	15	35	7.8e-003
Motif 7	PHHPPPHPHPHPHHP	15	33	7.8e-003
Motif 8	PHPPPHPPPHPHPHHPPPPPHPPPP	25	52	7.8e-003

Table S5 – Motifs identified for each virus capsid proteins with their width, sites and p-values. QS100 data was used.

T=3	Sequence	Width	Sites	p-value
Motif 1	PHPPPPPPHPHPPPPPPPPPHPPHPHPHPPPPPPPPHPPPP	41	219	3.2e-171
Motif 2	PPPPPPHPPPPPHPHPPPPPPPPPPPHPP	29	13	2.8e-080
Motif 3	PHHPPPPPPPHPPHPPPHPPPPHPPPPPP	29	2	4.1e-105
T=p3	**Sequence**	**Width**	**Sites**	**p-value**
Motif 1	PHPHPHPHPHPHPHPPPHPHH	21	133	1.9e-004
Motif 2	PHPHPHPHPPPPHHPHPHHPP	21	178	7.6e-076
Motif 3	HPPHPHPPPHPPHPPHHHHHH	21	318	1.2e-073