Identifying key amino acid types that distinguish paralogous proteins using Shapley value based feature subset selection

Pranav Machingal[1](✉), Rakesh Busi[2], Nandyala Hemachandra[1], and Petety V. Balaji[2]

[1] Department of Industrial Engineering and Operations Research, Indian Insitute of Technology Bombay, Mumbai, India
pranav.mac@iitb.ac.in

[2] Department of Biosciences and Bioengineering, Indian Insitute of Technology Bombay, Mumbai, India

Abstract. We view a protein as the composite of the standard 20 amino acids, ignoring their order in the protein sequence. With this view, we try to identify the important amino acid types that distinguish pairs of paralogous proteins, thereby playing a role in their functional difference. Using only the amino acid composition (AAC) as features and a linear classifier, we find that many pairs of paralogous protein families can be classified accurately. Next, we use an existing Shapley value-based feature subset selection algorithm, SVEA, to identify the important amino acid types that distinguish a pair of paralogous proteins. The SVEA algorithm assigns a score, Shapley value, to each feature, amino acid type, based on its contribution to the classifier's training error. We identify the important distinguishing amino acid types as those whose Shapley value exceeds a data-driven threshold. We refer to these as the amino acid feature subset (AFS). We find that many paralog pairs can still be accurately classified using only the AFS composition. We partition AFS based on the classifier weights to infer class-wise amino acid importance. We verify whether the identified AFS amino acids indeed play a role in the functional difference of the paralog pairs using various methods - multiple sequence alignment, 3D structure analysis, and supporting evidence from biology literature. We also discuss some consistencies observed in the Shapley value based ranking and the AFS when comparing the AFS of two different but related paralog pairs. We demonstrate the results for 15 pairs of paralogous proteins.

1 Introduction

Proteins form the fundamental machinery in living systems, having several vital functions such as DNA replication, catalysis, transport, environmental interaction, etc. Advancements in sequencing technologies have resulted in exponential growth of protein sequence databases [25]. However, the number of experimentally verified annotations constitute a tiny fraction; for instance, out of around 250 million sequences in UniProtKB [25] only about 0.57 million sequences have manually reviewed annotations. Experimental methods for determining biological process level functions (transcription, DNA repair, etc.) are high-throughput whereas methods for molecular function (catalysis, ligand specificity, etc.) are low-throughput and hence are not scalable. The relationship between sequence and function is subtle and has not been fully decoded yet.

Paralogs are proteins that have a common ancestor but have diverged functionally. The difference in functionality in two paralogous proteins is considered to arise due to specific evolutionary changes in the sequences [9]. The acquired mutations could lead to differences in the composition of certain amino acid types between two paralogous proteins. Figuring out these amino acid types can be useful to identify the residues that determine the functional specificity of paralogous proteins. Towards this, we view a protein as the composite of its constituent standard 20 amino acids, ignoring the ordering of the amino acids in the protein sequence. We use amino acid composition (AAC) features, Shapley value [21] based feature ranking and subset selection algorithm (Shapley Value based Error Apportioning, SVEA) [26], and support vector machine (SVM) based linear classifier [24] as tools to identify key amino acid types that can distinguish one paralogous protein family from another. The proposed method yields quick results based on which biologists can conduct detailed experiments which are resource-intensive (time, cost, trained manpower, etc.).

Following are our key results from experiments on 15 datasets (pairs of paralogous proteins):

- AAC alone suffices as input feature set to classify several paralogous protein pairs using a linear SVM.
- The SVEA algorithm identifies a subset of amino acid types (referred to as AFS) that are adequate for distinguishing two paralogous proteins. The number of amino acids in AFS is typically less than or equal to 10. A linear SVM classifies accurately using only AFS composition features.
- The significance of the amino acids identified in the AFS was validated for 9 datasets using various methods like multiple sequence alignment (MSA), structural analysis and supporting evidence from literature that report the conservation of these amino acids at structurally/functionally essential positions.
- For a paralog pair dataset, say families P and Q, the computed AFS is partitioned into $AFS(P)$ and $AFS(Q)$ based on the sign of the linear classifier weights. In the multiple sequence alignments (MSA) of 7 datasets, we find that the amino acids in $AFS(P)$ are relatively conserved at more positions in family P sequences than in family Q and similarly, $AFS(Q)$ amino acids are relatively conserved at more positions in family Q sequences than in P.
- We observe logical consistencies in the Shapley value and the AFS when comparing the AFS of two or more different but related paralog pairs (See globins, Section 3.6, and GPCRs, Section 3.7).
  - For families P vs Q and P vs R, with their respective AFS as AFS_1 and AFS_2, we find common amino acids selected in $AFS_1(P)$ and $AFS_2(P)$.
  - For families P vs Q, Q vs R and R vs P; with their respective AFS as AFS_1, AFS_2 and AFS_3, we find that: amino acids in $AFS_i \cap AFS_j$ are either excluded from AFS_k or have a low rank / Shapley value in AFS_k, AFS_i, or AFS_j.

Use of deep learning methods trained on large datasets is becoming commonplace in Biology; for example, prediction of molecular function via EC number or GO annotation [3, 20], identifying regions in an input sequence relevant to its prediction based on instance-level feature attribution [28] and learning sequence-function mapping from deep mutational scanning experiment data [23]. Use of large datasets for training makes this approach highly resource intensive. The approach we present herein needs much smaller datasets and consequently, (i) is computationally cheap and (ii) has far wider applicability because the amount of labelled data validated by wet lab experiments is very limited. Furthermore, our method ignores the sequence order of the amino acids in a protein, which is an essential input to deep learning methods as well as traditional sequence alignment based methods like profile-HMM [8] and BLAST [1].

In the next section, we describe our methodology in detail. In Section 3 we present the results of our computational experiments. Our conclusions and further discussions are provided in Section 4.

2 Methodology

The datasets and three main components of our methodology are discussed in subsections below.

2.1 Datasets of 15 paralog pairs

We apply our method for identifying amino acid types that distinguish paralogous proteins using the datasets described in Table 1. We have considered paralogous proteins wherein functional differences vary from fine-grained (e.g., trypsin/chymotrypsin) to coarse-grained (e.g., GPCR). This diversity is considered for robust evaluation of the method. All datasets are taken from publicly available data abases (UniProt [25] and GPCR-PEnDB [2]). Well-known pairs of paralogous proteins were curated from millions of sequences from UniProt considering the number of sequences and manually reviewed labels available for them.

Table 1: The number of sequences in the train and test sets of the protein families considered in computational experiments.

Family	Train (Swiss-Prot)	Test (TrEMBL)
Lysozyme-like		
α-Lactalbumin	22	53
Lysozyme C	74	14
Trypsin-like		
Trypsin	66	3813
Chymotrypsin	17	281
Tubulin		
α	117	190
β	191	347
Histone		
H2A	180	16599
H2B	177	7599
Interleukin-1		
α	16	12
β	25	194

Family	Train (Swiss-Prot)	Test (TrEMBL)
Cytochrome P450		
CYP3	32	818
CYP51	32	601
Globins		
Myoglobin	107	479
Hemoglobin-α	303	525
Hemoglobin-β	285	261
	(GPCR-PEnDB)	
	Train (80%)	Test (20%)
GPCR families		
Rhodopsin-like	181	45
↳ Lipid receptors	113	28
Peptide receptors	367	92
Aminergic receptors	186	47
Glutamate-like	89	23
Secretin-like	90	23

For all datasets, except GPCR, we use manually curated Swiss-Prot sequences for training and electronically annotated TrEMBL sequences for testing. These proteins have very specific functions. In contrast, GPCRs are a large and diverse group of transmembrane proteins that mediate cellular responses to extracellular signals. We chose to use an already curated dataset in this case. For each of the GPCR families considered (Table 1), the sequences are randomly split as 80%-train/20%-test. Use of GPCR-PEnDB data is to illustrate the effectiveness of our method with random slicing which is inevitable when additional curated data are not available. If one or many UniProt entries in a dataset had identical sequences, then only one of them was retained and the remaining were deleted. More details regarding the method/criteria for collecting the data along with the data files (with accession numbers) are provided in the Supplementary Section S1.

2.2 AAC features

Consider a paralogous pair of proteins, families P and Q. We first curate a set of sequences, say P_{train} and Q_{train}, from standard protein sequence databases with n_P and n_Q number of sequences each from families P and Q respectively.

For a protein sequence $\mathbf{p} = (p_1, p_2, \ldots, p_L)$ of length L with $p_k \in \{1, 2, \cdots, 20\}$ corresponding to the standard 20 amino acids, the AAC feature $\mathbf{x}^{AAC} \in [0,1]^{20}$ for $\mathbf{p}$ is computed as follows,

$$x_i^{AAC} = \frac{1}{L} \sum_{k=1}^{L} \mathbf{1}_{\{p_k = i\}}, \; \forall i \in [20]$$

Here feature x_i^{AAC} for a protein sequence is the normalised count of the standard amino acid i, $i \in \{1, 2, \cdots, 20\}$.

2.3 Feature subset selection

Given a set of N features from the protein sequences of P and Q, we try to find the features $S \subseteq N$ which contribute the most to the classification of P and Q sequences. With AAC features, we have $N = \{1, 2, \ldots, 20\}$ corresponding to each of the standard 20 amino acid types.

We utilise the Shapley value based feature ranking and subset selection algorithm, SVEA [26], to identify the most important $S \subseteq N$. Shapley value is a well known solution concept from cooperative game theory [15] for distributing the total worth of a coalition of players fairly among each of them by quantifying each player's effective marginal contribution. The SVEA algorithm considers the binary classification task as a cooperative game among the features and apportions the total training error among the features using Shapley values. The Shapley value $\phi(i)$ for a feature $i \in N$ is computed as follows,

$$\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Where $v(S) = tr_er(\varnothing) - tr_er(S)$, $tr_er(S)$ is the minimum training hinge loss using S as features and $\varnothing$ is the empty set. The term $(v(S \cup \{i\}) - v(S))$ is the marginal contribution of feature i to the feature subset S. Thus, $\phi(i)$ is a weighted sum of the marginal contribution of feature i to all the possible feature subsets that do not contain i. We use a class-balanced hinge loss for computing $tr_er(S)$ for which $tr_er(\varnothing) = 1$. Shapley values are unique solution concepts satisfying the axioms - efficiency, symmetry and marginality [27]. With efficiency implying $\sum_{i \in N} \phi(i) = v(N) = 1 - tr_er(N)$.

The higher the $\phi(i)$, lower is the contribution of feature i to the classifier training error and consequentially higher the importance of feature i to the classification. We set $\phi_{cutoff} = \frac{1}{20} \sum_{i=1}^{20} \phi(i)$ for selecting the feature subset, $AFS = \{i : \phi(i) > \phi_{cutoff}\}$. Each of the features in AFS uniquely corresponds to $d \leqslant 20$ amino acids from the standard 20. Exact Shapley value computations are known to be exponential time, hence, the Shapley value ϕ_i for a feature i is computed using Monte Carlo based linear time (in number of features) approximation in the SVEA algorithm. As the number of features is small (20), good approximations can be computed fast via larger sampling. More details of the SVEA algorithm are given in Supplementary Section S2.

2.4 Identifying class-wise partitions of AFS

We train a linear SVM, to classify P vs Q, using the composition of the amino acids in AFS as the features, i.e. using $\mathbf{x}^{AFS} \in [0, 1]^d$, with $x_j^{AFS} = x_i^{AAC}$ and each $j \in \{1, 2, \cdots, d\}$ uniquely maps to a $i \in AFS$. We use these linear SVM weights $\mathbf{w} \in \mathbb{R}^d$ to divide the set AFS into disjoint sets $AFS(P)$ and $AFS(Q)$ based on the sign of the weights. Since $x_j^{AFS} \geqslant 0 \; \forall j \in [d]$, the sign of the linear classifier weight w_j indicates which class is relatively prominent in the amino acid corresponding to j. So if the $+1$ class is P, then we divide AFS class-wise as $AFS(P) = \{j \in [d] : w_j > 0\}$ and similarly $AFS(Q) = \{j \in [d] : w_j < 0\}$.

We use 5-fold cross-validation to tune the SVM regularisation hyperparameter C from $\{0.1, 1, 10, 100, 1000\}$ that gives the best average classification score for the 5 folds. C is inversely proportional to the strength of regularisation. In general, we find that there is an imbalance in the number of sequences that we find for the two paralogous proteins, i.e. say $n_P >> n_Q$. It is known that accuracy is not a well-suited performance measure of the classifier in class imbalance settings. Therefore we use the arithmetic mean of sensitivity and specificity (AM) to measure the performance of the classifier [4]. Further, we use a class-balanced version of hinge loss for training the SVM as suggested in [13] for statistical consistency with the AM score. We report the train and test scores of the trained linear SVM with AFS features on the protein family datasets considered in our computational experiments in Section 3.

A flowchart summarizing the sequence of steps for computing $AFS(P)$ and $AFS(Q)$ is shown in the Figure 1. For different paralogous protein pairs we provide supporting evidence from the literature for the significance of amino acids in $AFS(P)$ and $AFS(Q)$ (in Section 3). We generate a multiple sequence alignment (MSA) of a set of randomly selected sequences from P_{train} and Q_{train} and analyze the conservation

of $AFS(P)$ and $AFS(Q)$ amino acids within and across the respective families (Figure 3). When classifying families P vs Q and P vs R, we compare the amino acids in the two $AFS(P)$ sets corresponding to family P that we get from the two classifications (in Section 3).

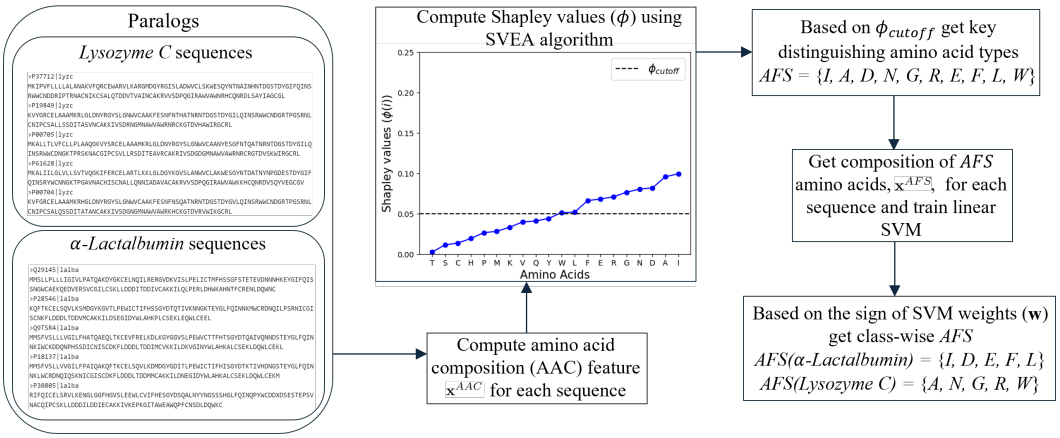


Fig. 1: Flowchart summarizing the sequence of steps used in our method to compute the key amino acid types (AFS: amino acid feature subset) that distinguish paralogous proteins, using amino acid composition (AAC) features, Shapley value based SVEA algorithm for feature subset selection and class-wise feature subsets using linear SVM. Lysozyme C and α-Lactalbumin are used here as representative examples of paralogous proteins. AFS is identified for other pairs of paralogous proteins using the same approach, and are discussed in Section 3.

3 Results

***AAC* as well as *AFS* features are adequate to classify many paralogous protein pairs**: We train linear SVMs with AAC and AFS features respectively for all datasets (15 paralog pairs) using class-balanced hinge loss [13], the classification scores for which are reported in Tables 2. We get at least 84% test AM score when classifying a paralog pair using AAC features. For 10 of 15 paralog pairs we get greater than 90% test AM scores with AAC. Using AFS features we get at least 70% test AM score. For 13 of 15 paralog pairs we get greater than 83% test AM scores using AFS features. For 8 of 15 paralog pairs we get greater than 90% test AM scores using AFS features.

Role of the amino acids identified in *AFS* in the functional difference of paralog pairs: For 14 paralog pairs, we discuss the significance of the amino acids identified in the respective AFS (Table 3). We use various methods - multiple sequence alignment, 3D structure analysis, and supporting evidence from biology literature. We also discuss some logical consistencies that we observe in the identified AFS and the Shapley value scores when comparing the AFS of two or more different but related paralog pairs

3.1 Lysozyme C and α-Lactalbumin

Lysozyme C and α-Lactalbumin are sequence and structure homologs with mutually exclusive functions and high fold conservation. Based on phylogenetic analysis, they are considered to have diverged from a common ancestor millions of years ago [19].

We observe in the MSA of the sequences, in Figure 3a, that the identified feature sets $AFS(\alpha\text{-Lactalbumin})$ and $AFS(\text{Lysozyme C})$ (Table 3) show significant conservation in respective families resulting in a pattern which distinguishes the sequences in one family from the other. The amino acids D and E which are part of $AFS(\alpha\text{-Lactalbumin})$ are known to be found in the Ca^{2+} and Zn^{2+} binding sites respectively of α-lactalbumin [16, 17]. All α-lactalbumins studied so far are known to bind Ca^{2+} and Zn^{2+} whereas several (but not all) lysozymes do not bind Ca^{2+}.

Table 2: Classification scores for different pairs of paralogous proteins using the train/test datasets described in Table 1, using AAC and AFS features. The AFS amino acids computed for each pair are given in Table 3. The train score is the mean (± 1 standard deviation) 5-fold cross-validation score. **AM.** is the arithmetic mean of specificity and sensitivity. **Acc.** is the accuracy.

(a) Lysozyme C vs α-Lactalbumin

	AAC	**AFS**
Train AM.	1.0	0.993 ($\pm$0.013)
Test AM.	0.896	0.898
Train Acc.	1.0	0.99 ($\pm$0.02)
Test Acc.	0.836	0.881

(b) Trypsin vs chymotrypsin

	AAC	**AFS**
Train AM.	0.992 ($\pm$0.015)	0.977 ($\pm$0.031)
Test AM.	0.873	0.835
Train Acc.	0.988 ($\pm$0.024)	0.965 ($\pm$0.047)
Test Acc.	0.844	0.756

(c) Tubulin-α vs Tubulin-β

	AAC	**AFS**
Train AM.	0.996 ($\pm$0.009)	0.997 ($\pm$0.006)
Test AM.	0.992	0.992
Train Acc.	0.997 ($\pm$0.006)	0.994 ($\pm$0.008)
Test Acc.	0.991	0.994

(d) Histone H2A vs Histone H2B

	AAC	**AFS**
Train AM.	0.983 ($\pm$0.016)	0.983 ($\pm$0.01)
Test AM.	0.91	0.934
Train Acc.	0.983 ($\pm$0.016)	0.983 ($\pm$0.01)
Test Acc.	0.889	0.922

(e) Globins

Dataset		**AAC**	**AFS**
Myoglobin vs Hemoglobin-α	**Train AM.**	0.998 ($\pm$0.003)	0.994 ($\pm$0.009)
	Test AM.	0.968	0.97
	Train Acc.	0.998 ($\pm$0.005)	0.995 ($\pm$0.006)
	Test Acc.	0.969	0.971
Myoglobin vs Hemoglobin-β	**Train AM.**	1.0 ($\pm$0.0)	1.0 ($\pm$0.0)
	Test AM.	0.957	0.936
	Train Acc.	1.0 ($\pm$0.0)	1.0 ($\pm$0.0)
	Test Acc.	0.949	0.919
Hemoglobin-α vs Hemoglobin-β	**Train AM.**	0.983 ($\pm$0.008)	0.976 ($\pm$0.007)
	Test AM.	0.961	0.935
	Train Acc.	0.983 ($\pm$0.008)	0.976 ($\pm$0.006)
	Test Acc.	0.966	0.947

(f) GPCRs

Dataset		**AAC**	**AFS**
Secretin-like vs Glutamate-like	**Train AM.**	0.933 ($\pm$0.042)	0.95 ($\pm$0.032)
	Test AM.	0.888	0.845
	Train Acc.	0.933 ($\pm$0.042)	0.95 ($\pm$0.032)
	Test Acc.	0.889	0.844
Rhodopsin-like vs Glutamate-like	**Train AM.**	0.884 ($\pm$0.042)	0.85 ($\pm$0.045)
	Test AM.	0.967	0.934
	Train Acc.	0.867 ($\pm$0.038)	0.837 ($\pm$0.032)
	Test Acc.	0.956	0.926
Rhodopsin-like vs Secretin-like	**Train AM.**	0.917 ($\pm$0.051)	0.878 ($\pm$0.065)
	Test AM.	0.934	0.846
	Train Acc.	0.908 ($\pm$0.06)	0.863 ($\pm$0.073)
	Test Acc.	0.941	0.853
Aminergic vs Lipid receptors	**Train AM.**	0.949 ($\pm$0.014)	0.943 ($\pm$0.005)
	Test AM.	0.922	0.843
	Train Acc.	0.943 ($\pm$0.017)	0.94 ($\pm$0.008)
	Test Acc.	0.92	0.84
Aminergic vs Peptide receptors	**Train AM.**	0.835 ($\pm$0.06)	0.818 ($\pm$0.053)
	Test AM.	0.844	0.79
	Train Acc.	0.83 ($\pm$0.06)	0.819 ($\pm$0.051)
	Test Acc.	0.827	0.784
Lipid vs Peptide receptors	**Train AM.**	0.829 ($\pm$0.022)	0.76 ($\pm$0.035)
	Test AM.	0.845	0.709
	Train Acc.	0.838 ($\pm$0.018)	0.75 ($\pm$0.032)
	Test Acc.	0.858	0.725

(g) Interleukin-1 α vs Interleukin-1 β

	AAC	**AFS**
Train AM.	0.98 ($\pm$0.04)	0.98 ($\pm$0.04)
Test AM.	0.979	0.985
Train Acc.	0.975 ($\pm$0.05)	0.975 ($\pm$0.05)
Test Acc.	0.961	0.971

(h) Cytochrome P450 CYP3 vs Cytochrome P450 CYP51

	AAC	**AFS**
Train AM.	0.967 ($\pm$0.041)	0.933 ($\pm$0.062)
Test AM.	0.902	0.92
Train Acc.	0.969 ($\pm$0.038)	0.936 ($\pm$0.062)
Test Acc.	0.894	0.908

Table 3: AFS and its class-wise partition computed for 15 paralog pairs. AFS amino acids are written in decreasing Shapley values from left to right for each paralog pair. Figures 2,4 and 5 show the Shapley value of the amino acids for each paralog pair.

Paralog pair	AFS	**Class-wise AFS parition**
Lysozyme C and α-Lactalbumin	$\{I, A, D, N, G, R, E, F, L, W\}$	$AFS(\alpha\text{-Lactalbumin}) = \{I, D, E, F, L\}$ $AFS(\text{Lysozyme C}) = \{A, N, G, R, W\}$
Trypsin and Chymotrypsin	$\{Y, W, T, A, V, K, P\}$	$AFS(\text{Trypsin}) = \{Y, A\}$ $AFS(\text{Chymotrypsin}) = \{W, T, V, K, P\}$
Tubulin-α and Tubulin-β	$\{M, Q, K, N, F, I, H, A, C, Y\}$	$AFS(\text{Tubulin-}\alpha) = \{K, I, H, C, Y\}$ $AFS(\text{Tubulin-}\beta) = \{M, Q, N, F, A\}$
Histone H2A and Histone H2B	$\{L, G, S, M, K, N, T, Y, F\}$	$AFS(\text{Histone H2A}) = \{L, G, N\}$ $AFS(\text{Histone H2B}) = \{S, M, K, T, Y, F\}$
Interleukin-1 α and Interleukin-1 β	$\{C, G, T, S, V, Q, A, N, P\}$	$AFS(\text{Interleukin-1 }\alpha) = \{T, S, A, N\}$ $AFS(\text{Interleukin-1 }\beta) = \{C, G, V, Q, P\}$
Cytochrome P450 CYP3 and CYP51	$\{H, F, G, K, A, P, N\}$	$AFS(\text{CYP3}) = \{F, K, P, N\}$ $AFS(\text{CYP51}) = \{H, G, A\}$
Globins		
Myoglobin and Hemoglobin-α	$AFS_1 = \{E, S, Y, V, K, P, I, G, C, W\}$	$AFS_1(\text{Myoglobin}) = \{E, K, I, G, W\}$ $AFS_1(\text{Hemoglobin-}\alpha) = \{S, Y, V, P, C\}$
Myoglobin and Hemoglobin-β	$AFS_2 = \{K, V, C, E, W, N, F, M, Y, I\}$	$AFS_2(\text{Myoglobin}) = \{K, E, M, I\}$ $AFS_2(\text{Hemoglobin-}\beta) = \{V, C, W, N, F, Y\}$
Hemoglobin-α and Hemoglobin-β	$AFS_3 = \{W, P, N, S, G\}$	$AFS_3(\text{Hemoglobin-}\alpha) = \{P, S\}$ $AFS_3(\text{Hemoglobin-}\beta) = \{W, N, G\}$
GPCRs		
Rhodopsin-like and Glutamate-like	$AFS_1 = \{D, Q, E, G, M, L\}$	$AFS_1(\text{Rhodopsin}) = \{M, L\}$ $AFS_1(\text{Glutamate}) = \{D, Q, E, G\}$
Secretin-like and Glutamate-like	$AFS_2 = \{W, H, Y, V, D\}$	$AFS_2(\text{Secretin}) = \{W, H, Y\}$ $AFS_2(\text{Glutamate}) = \{V, D\}$
Rhodopsin-like and Secretin-like	$AFS_3 = \{W, E, M, S, V, H, Q, A\}$	$AFS_3(\text{Rhodopsin}) = \{M, S, V, A\}$ $AFS_3(\text{Secretin}) = \{W, E, H, Q\}$
Rhodopsin-like GPCRs		
Aminergic receptors and Lipid receptors	$AFS_1 = \{L, P, E, W, F, M, D\}$	$AFS_1(\text{Aminergic receptors}) = \{P, E, W, D\}$ $AFS_1(\text{Lipid receptors}) = \{L, F, M\}$
Aminergic receptors and Peptide receptors	$AFS_2 = \{L, F, E, M, K, D, V, R\}$	$AFS_2(\text{Aminergic receptors}) = \{E, K, D, R\}$ $AFS_2(\text{Peptide receptors}) = \{L, F, M, V\}$
Lipid receptors and Peptide receptors	$AFS_3 = \{P, R, G, I, W, S, V\}$	$AFS_3(\text{Lipid receptors}) = \{\text{R, G, S}\}$ $AFS_3(\text{Peptide receptors}) = \{P, I, W, V\}$

3.2 Trypsin and Chymotrypsin

The amino acids Y and W get the highest $\phi(\cdot)$ score in $AFS(\text{Trypsin})$ and $AFS(\text{Chymotrypsin})$ respectively (Table 3 and Figure 2b). In experiments to convert trypsin to chymotrypsin [10, 11] it has been shown that Y to W conversion in loop-3 of trypsin leads to significant increase in chymotrypsin activity. Also, we find that the catalytic triad amino acids, S, H and D, which are common to both families, are not part of the AFS

3.3 Tubulin-α and Tubulin-β

We find in the MSA of the tubulin sequences in Supplementary Figure S4 that the amino acids identified in $AFS(\text{Tubulin-}\alpha)$ and $AFS(\text{Tubulin-}\beta)$ show significant conservation in respective families resulting in a pattern which distinguishes the sequences in one family from the other.

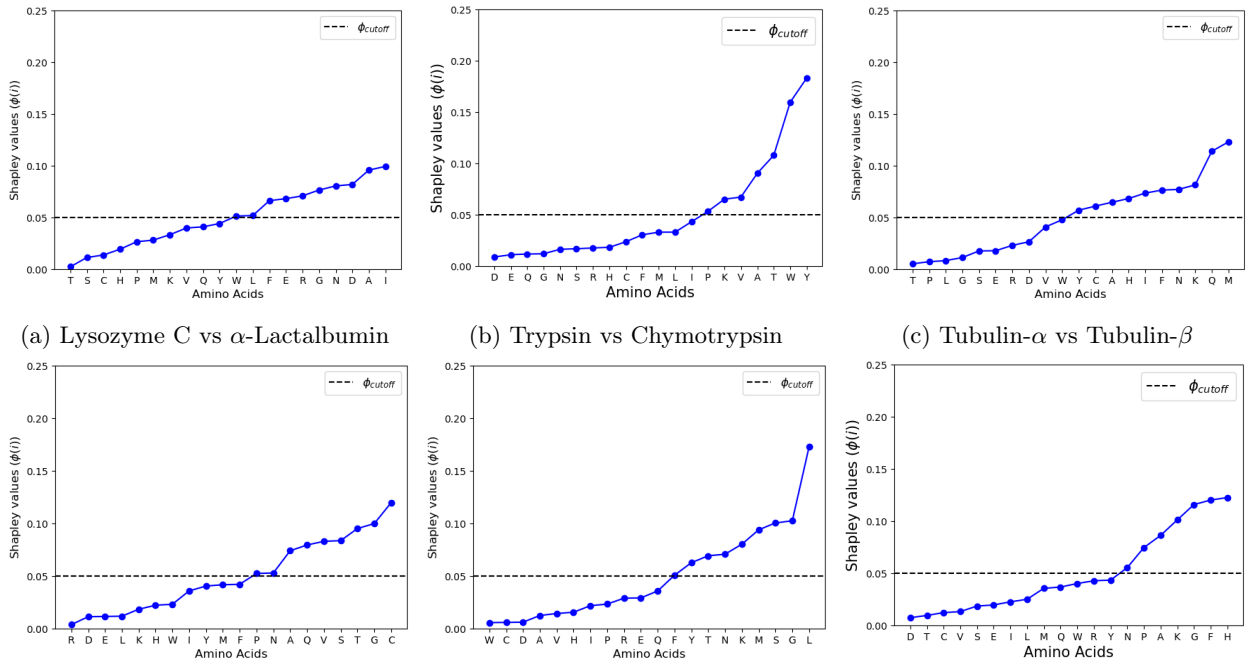
Structural analysis of AFS: Tubulin α and β heterodimers form the building blocks of microtubules, which are important components of the eukaryotic cytoskeleton [14]. We looked at the contact residues of a tubulin-α chain and tubulin-β chain in the 3D structure of tubulin-α/β heterodimer (PDB IDs: 3JAR,

5N5N). We find that the contact points of the tubulin-α chain in the heterodimer structure have more AFS(Tubulin-α) amino acids than AFS(Tubulin-β). Similarly, the contact points of the tubulin-β chain in the heterodimer structure have more AFS(Tubulin-β) amino acids than AFS(Tubulin-α). Thus, the amino acids identified in AFS can be considered to be significant towards the quaternary structure of tubulin-α/β heterodimer. See Supplementary Section S3.2 for more details.

3.4 Histone H2A and Histone H2B

In the MSA of the sequences in Supplementary Figure S5, we find that the amino acids identified in AFS(Histone H2A) and AFS(Histone H2B) show significant conservation in respective families resulting in a pattern which distinguishes the sequences in one family from the other.

Structural analysis of AFS: Histones are known to have a heterooctameric structure comprised of two H2A/H2B dimers and one H3/H4 tetramer [7]. We looked at the contact residues of an H2A chain and H2B chain in the heteroocatmer structure of histone (PDB IDs: 3KWQ, 1AOI). We find that the contact points of H2A chain in the heterooctomeric structure has more AFS(Histone H2A) amino acids than AFS(Histone H2B). This is interesting since, AFS(Histone H2A) has only three amino acids, while AFS(Histone H2B) has six amino acids. Similarly, the contact points of H2B chain in the heterooctomeric structure has more AFS(Histone H2B) amino acids than AFS(Histone H2A). Thus, the amino acids identified in AFS can be considered to be significant towards the quaternary structure of the histone heterooctamer. See Supplementary Section S3.3 for more details.



(a) Lysozyme C vs α-Lactalbumin (b) Trypsin vs Chymotrypsin (c) Tubulin-α vs Tubulin-β

(d) Interleukin-1 α vs Interleukin-1 β (e) Histone H2A vs Histone H2B (f) Cytochrome P450 CYP3 vs CYP51

Fig. 2: Shapley value ($\phi(i)$) for AAC features computed using SVEA.

3.5 Interleukin-1 α and Interleukin-1 β

We observe in the MSA of the sequences, in Supplementary Figure S6, that the amino acids identified in AFS(Interleukin-1 α) and AFS(Interleukin-1 β) show significant conservation in respective families resulting in a pattern which distinguishes the sequences in one family from the other.

(a) Lysozyme C (left) vs α-Lactalbumin (right)

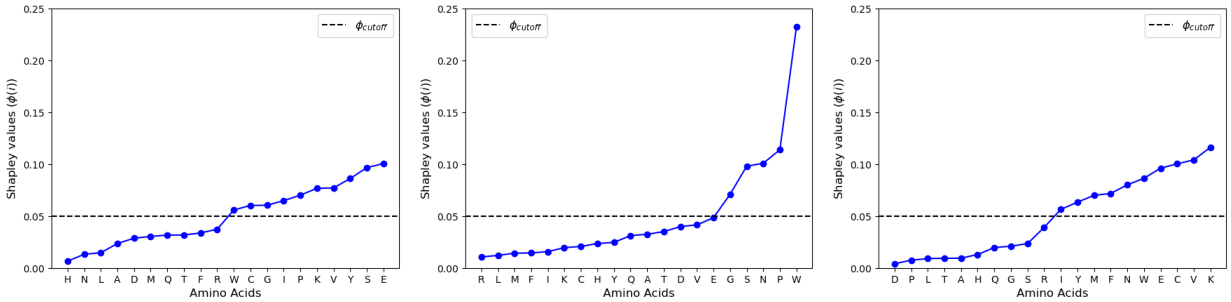
(b) Hemoglobin-α (left) vs Myoglobin (right)

(c) Hemoglobin-β (left) vs Hemoglobin-α (right)

Fig. 3: Multiple sequence alignment of sequences from the respective families in (a), (b) and (c). Within each alignment, 15 sequences on the left are from one family, and those on the right are from the other family in each of (a), (b) and (c). The sequences are randomly selected from the train set of the families. For each aligned sequence in (a) AFS(α-Lactalbumin) amino acids are in green and AFS(Lysozyme C) are in red, in (b) the amino acids in AFS_1(Myoglobin) are in green and AFS_1(Hemoglobin-α) are in red, and in (c) the amino acids in AFS_2(Hemoglobin-α) are in green and AFS_2(Hemoglobin-β) are in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (Figure 2).

3.6 Globins

For the three globin paralog pairs (Table 3), we observe in the MSA of the sequences the conservation of the class-wise partition of AFS in the respective families (Figures 3b,3c and Supplementary Figure S2).



(a) Myoglobin vs Hemoglobin-α (b) Hemoglobin-α vs Hemoglobin-β (c) Myoglobin vs Hemoglobin-β

Fig. 4: Shapley value ($\phi(i)$) for AAC features computed using SVEA.

The amino acid W receives a significantly high Shapley value $\phi(W)$ (Figure 4b) and is present in AFS_3(Hemoglobin-β) $= \{W, N, G\}$. It is highly conserved at position 40 in the MSA (Figure 3c) in hemoglobin-β sequences as compared to hemoglobin-α sequences. This W residue at position 40 has been determined to be present in hemoglobin-β at one of its contact position to hemoglobin-α in the tetrameric structure [22] and is, therefore, a structurally and functionally significant residue.

Structural analysis of AFS: Myoglobin is a monomer, while α and β chains together constitute hemoglobin, a tetramer of composition $\alpha_2\beta_2$ [6]. We superimposed the 3D structures of myoglobin, hemoglobin-α and hemoglobin-β (PDB IDs: 3RGK, 1HHO) and mapped the α, β contact residues (based on [22]) of hemoglobin tetramer to that of myoglobin. We find that the amino acids K, E, I, which are common in AFS_1(Myoglobin) and AFS_2(Myoglobin), are less in number at the contact residues of hemoglobin tetramer and more in number at the corresponding locations in myoglobin, which is a monomer (see Supplementary Figure S1).

The amino acids which are common in AFS_1(Hemoglobin-α) and AFS_2(Hemoglobin-β), V, Y, C, when classifying against myoglobin, are not found in AFS_3 in α vs β hemoglobin classification. Furthermore, the amino acid C has been shown to play an important role in the tetrameric structure of hemoglobin [12]. Hence, it can be expected that V, Y, C would not be *key* in AAC for distinguishing α vs β hemoglobin.

Logical consistencies in AFS (refer to Table 3 (Globins) for AFS_1, AFS_2, AFS_3):

- $AFS_1 \cap AFS_2 = \{E, Y, V, K, I, C, W\}$. Except for W the remaining of these amino acids are excluded from AFS_3, while W receives the least score in AFS_1 (Figure 4a).
- $AFS_2 \cap AFS_3 = \{W, N\}$. Except for W the remaining of these amino acids are excluded from AFS_1, while W receives the least score in AFS_1 (Figure 4a).
- Similarly, $AFS_3 \cap AFS_1 = \{W, P, S, G\}$. Except for W the remaining of these amino acids are excluded from AFS_2, while W receives the least score in AFS_1 (Figure 4a).

The Shapley value for W is very close to the cut-off in AFS_1 (Figure 4a). If it is dropped from AFS_1, then the exclusion principle illustrated above would be more prominent as in GPCRs (Section 3.7).

3.7 G-protein coupled receptors (GPCRs)

We consider pair-wise three GPCR families, rhodopsin-like, secretin-like and glutamate-like, identified in GRAFS [2] system. Further, we consider pair-wise three subfamilies of rhodopsin-like GPCR proteins - aminergic receptors, lipid receptors and peptide receptors.

Analysing AFS of secretin-like GPCRs W which receives the highest $\phi(\cdot)$ score in AFS_2(Secretin) and AFS_3(Secretin) (Table 3 and Figure 5), has been reported by [5] to be well conserved at multiple positions with structural and functional importance in secretin-like GPCR sequences.

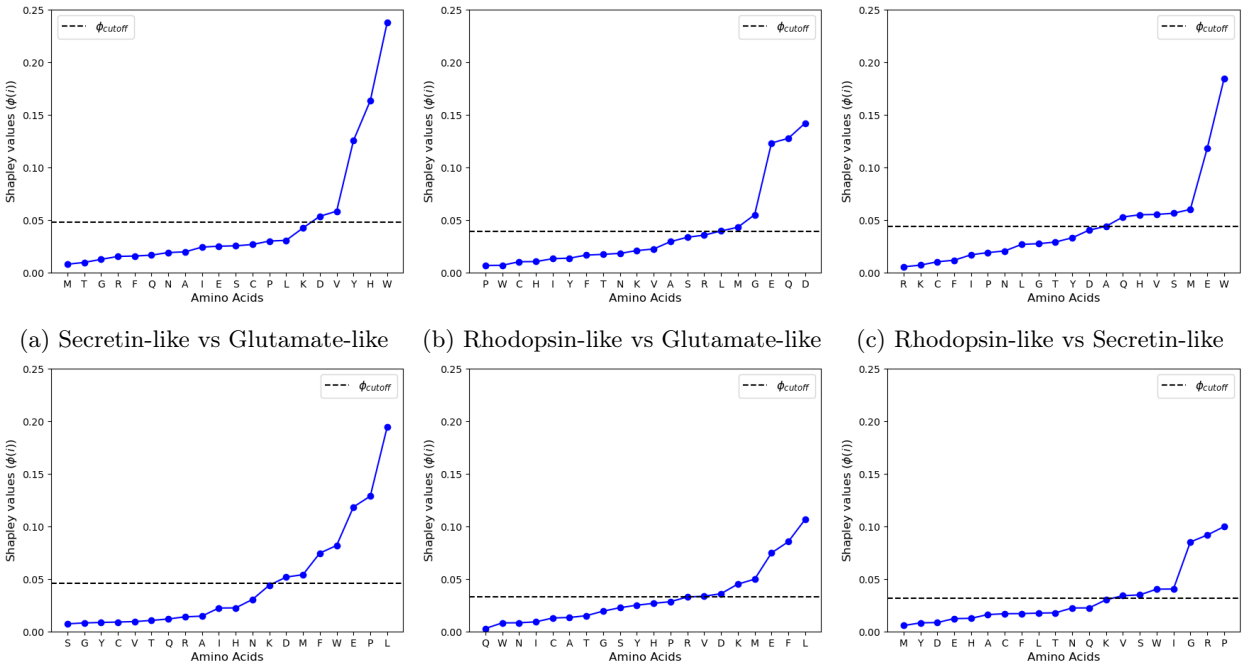
Logical consistencies in AFS **of GPCRs** (refer to Table 3 (GPCRs) for AFS_1, AFS_2, AFS_3):

- $AFS_1 \cap AFS_2 = \{D\}$. This is excluded from AFS_3.

– $AFS_2 \cap AFS_3 = \{W, H, V\}$. These amino acids are excluded from AFS_1.
– Similarly, $AFS_3 \cap AFS_1 = \{Q, E, M\}$. These amino acids are excluded from AFS_2.

Logical consistencies in AFS of Rhodopsin-like GPCR subfamilies (refer to Table 3 (Rhodospin-like GPCRs) for AFS_1, AFS_2, AFS_3):

– $AFS_1 \cap AFS_2 = \{L, E, F, M, D\}$. This is excluded from AFS_3.
– $AFS_2 \cap AFS_3 = \{R, V\}$. These amino acids are excluded from AFS_1.
– Similarly, $AFS_3 \cap AFS_1 = \{P, W\}$. These amino acids are excluded from AFS_2.



(a) Secretin-like vs Glutamate-like (b) Rhodopsin-like vs Glutamate-like (c) Rhodopsin-like vs Secretin-like

(d) Aminergic vs Lipid receptors (e) Aminergic vs Peptide receptors (f) Lipid vs Peptide receptors

Fig. 5: Shapley value ($\phi(i)$) for AAC features computed SVEA.

4 Discussion

Gene duplication, followed by functional differentiation, is a recurring theme in the evolution of proteins. The observation that the number of residues conserved across all cytochrome c proteins is <10 [18] illustrates that a protein may tolerate mutations at a large number of sequence positions conserving function. Thus, we see sequence divergence both among orthologs (same function) and paralogs (different functions). Thus when a protein of unknown function shows sequence similarity to an already characterized protein, the question that arises is whether transfer of function annotation is justified (i.e., the two proteins are orthologs) or not (i.e., the two are paralogs). In this background, it will be useful to have computationally inexpensive (time + resources) methods that can distinguish paralogous pairs of proteins. The work presented herein is one such approach. The robustness of this approach as demonstrated by considering a diverse set of paralogous protein pairs illustrates its wider applicability. Notably, amino acids in the AFS typically occur >1 in the sequence but our method is silent on the specific sequence positions where the amino acid has a functionally distinguishing role. This lacuna may be addressed by engineering features that incorporate sequence order information from the protein. However, these features can be very high-dimensional, for example, 20^k-dimensional for k-mer features. Shapley value computation from high-dimensional features can be expensive. This is because the Monte Carlo based approximation algorithm would require exponentially more sampling (in the number of features) for good approximations of Shapley value.

Code and data availability. Available at `https://github.com/Pranav-Machingal/AFS_AAC_SVM.git`.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990). https://doi.org/10.1016/S0022-2836(05)80360-2, https://www.sciencedirect.com/science/article/pii/S0022283605803602
2. Begum, K., Mohl, J.E., Ayivor, F., Perez, E.E., Leung, M.Y.: GPCR-PEnDB: a database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors. Database **2020** (11 2020). https://doi.org/10.1093/database/baaa087, https://doi.org/10.1093/database/baaa087
3. Bileschi, M.L., Belanger, D., Bryant, D.H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M.A., Colwell, L.J.: Using deep learning to annotate the protein universe. Nature biotechnology **40**(6), 932—937 (June 2022). https://doi.org/10.1038/s41587-021-01179-w, https://doi.org/10.1038/s41587-021-01179-w
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th ICPR. pp. 3121–3124 (2010). https://doi.org/10.1109/ICPR.2010.764
5. Cary, B.P., Zhang, X., Cao, J., Johnson, R.M., Piper, S.J., Gerrard, E.J., Wootten, D., Sexton, P.M.: New Insights into the Structure and Function of Class B1 GPCRs. Endocrine Reviews **44**(3), 492–517 (12 2022). https://doi.org/10.1210/endrev/bnac033, https://doi.org/10.1210/endrev/bnac033
6. Dill, K., Jernigan, R., Bahar, I.: Protein Actions: Principles and Modeling. CRC Press (2017), https://books.google.co.in/books?id=NHs2DwAAQBAJ
7. Dutta, S., Akey, I.V., Dingwall, C., Hartman, K.L., Laue, T., Nolte, R.T., Head, J.F., Akey, C.W.: The crystal structure of nucleoplasmin-core: Implications for histone binding and nucleosome assembly. Molecular Cell **8**(4), 841–853 (2001). https://doi.org/10.1016/S1097-2765(01)00354-9, https://www.sciencedirect.com/science/article/pii/S1097276501003549
8. Eddy, S.R.: Profile hidden Markov models. Bioinformatics **14**(9), 755–763 (10 1998). https://doi.org/10.1093/bioinformatics/14.9.755, https://doi.org/10.1093/bioinformatics/14.9.755
9. Fitch, W.M.: Homology: a personal view on some of the problems. Trends in Genetics **16**(5), 227–231 (2000). https://doi.org/10.1016/S0168-9525(00)02005-9, https://www.sciencedirect.com/science/article/pii/S0168952500020059
10. Hedstrom, L.: Serine protease mechanism and specificity. Chemical Reviews **102**(12), 4501–4524 (2002). https://doi.org/10.1021/cr000033x, https://doi.org/10.1021/cr000033x, pMID: 12475199
11. Hedstrom, L., Perona, J.J., Rutter, W.J.: Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. Biochemistry **33 29**, 8757–63 (1994)
12. Kan, H.I., Chen, I.Y., Zulfajri, M., Wang, C.C.: Subunit disassembly pathway of human hemoglobin revealing the site-specific role of its cysteine residues. The Journal of Physical Chemistry B **117**(34), 9831–9839 (2013). https://doi.org/10.1021/jp402292b, https://doi.org/10.1021/jp402292b, pMID: 23902424
13. Menon, A.K., Narasimhan, H., Agarwal, S., Chawla, S.: On the statistical consistency of algorithms for binary classification under class imbalance. In: Proceedings of the 30th ICML - Volume 28. p. III–603–III–611. ICML'13, JMLR.org (2013)
14. Mühlethaler, T., Gioia, D., Prota, A.E., Sharpe, M.E., Cavalli, A., Steinmetz, M.O.: Comprehensive analysis of binding sites in tubulin. Angewandte Chemie International Edition **60**(24), 13331–13342 (2021). https://doi.org/10.1002/anie.202100273, https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202100273
15. Narahari, Y.: Game Theory and Mechanism Design. WORLD SCIENTIFIC / INDIAN INST OF SCIENCE, INDIA (2014). https://doi.org/10.1142/8902, https://www.worldscientific.com/doi/abs/10.1142/8902
16. Permyakov, E.A.: α-actalbumin, Amazing Calcium-Binding Protein. Biomolecules **10**(9), 1210 (Aug 2020). https://doi.org/10.3390/biom10091210, http://dx.doi.org/10.3390/biom10091210
17. Permyakov, E.A., Berliner, L.J.: α-Lactalbumin: structure and function. FEBS Letters **473**(3), 269–274 (2000). https://doi.org/10.1016/S0014-5793(00)01546-5, https://www.sciencedirect.com/science/article/pii/S0014579300015465
18. Ptitsyn, O.B.: Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? Journal of Molecular Biology **278**(3), 655–666 (1998). https://doi.org/10.1006/jmbi.1997.1620, https://www.sciencedirect.com/science/article/pii/S002228369791620X
19. Qasba, P.K., Kumar, S., Brew, D.K.: Molecular divergence of lysozymes and α-lactalbumin. Critical Reviews in Biochemistry and Molecular Biology **32**(4), 255–306 (1997). https://doi.org/10.3109/10409239709082574, https://doi.org/10.3109/10409239709082574
20. Sanderson, T., Bileschi, M.L., Belanger, D., Colwell, L.J.: Proteinfer, deep neural networks for protein functional inference. eLife **12**, e80942 (feb 2023). https://doi.org/10.7554/eLife.80942, https://doi.org/10.7554/eLife.80942
21. Shapley, L.S.: 17. A Value for n-Person Games, pp. 307–318. Princeton University Press, Princeton (1953). https://doi.org/doi:10.1515/9781400881970-018, https://doi.org/10.1515/9781400881970-018
22. Shionyu, M., Takahashi, K., Gō, M.: Variable subunit contact and cooperativity of hemoglobins. J. Mol. Evol. **53**(4-5), 416–429 (Oct 2001)

23. Song, H., Bremer, B.J., Hinds, E.C., Raskutti, G., Romero, P.A.: Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. Cell Systems **12**(1), 92–101.e8 (2021). https://doi.org/10.1016/j.cels.2020.10.007, https://www.sciencedirect.com/science/article/pii/S2405471220304142
24. Steinwart, I., Christmann, A.: Support Vector Machines. Springer Publishing Company, Incorporated, 1st edn. (2008)
25. The UniProt Consortium: UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research **49**(D1), D480–D489 (11 2020). https://doi.org/10.1093/nar/gkaa1100, https://doi.org/10.1093/nar/gkaa1100
26. Tripathi, S., Hemachandra, N., Trivedi, P.: Interpretable feature subset selection: A Shapley value based approach. In: 2020 IEEE BigData. pp. 5463–5472 (2020). https://doi.org/10.1109/BigData50022.2020.9378102
27. Young, H.P.: Monotonic solutions of cooperative games. Int. J. Game Theory **14**(2), 65–72 (jun 1985). https://doi.org/10.1007/BF01769885, https://doi.org/10.1007/BF01769885
28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of IEEE CVPR (June 2016)

Supplementary Material for Identifying key amino acid types that distinguish paralogous proteins using Shapley value based feature subset selection

S1 Data collection and code

The following queries were used for collecting data from UniProt [5],

- **lysozyme C**: `(protein_name:"lysozyme C") AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200]) AND (ec:3.2.1.17) AND (xref:cazy-GH22) AND (reviewed:true)`
- **α-lactalbumin**: `(protein_name:"alpha lactalbumin") AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200]) AND (reviewed:true)`
- **myoglobin**: `(protein_name:"myoglobin") AND (xref:interpro-IPR002335) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **hemoglobin-α**: `(protein_name:"hemoglobin alpha") AND (xref:interpro-IPR002338) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **hemoglobin-β**: `(protein_name:"hemoglobin beta") AND (xref:interpro-IPR002337) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **trypsin**: `(protein_name:trypsin) AND (fragment:false) AND (ec:3.4.21.4) NOT (existence:5)`
- **chymotrypsin**: `(protein_name:chymotrypsin) AND (fragment:false) AND (ec:3.4.21.1) NOT (existence:5)`
- **tubulin-α**: `(protein_name:"tubulin alpha") AND (family:"tubulin family") AND (length:[300 TO 600]) AND (fragment:false) NOT (annotation_score:1) NOT (annotation_score:2)`
- **tubulin-β**: `(protein_name:"tubulin beta") AND (family:"tubulin family") AND (length:[300 TO 600]) AND (fragment:false) NOT (annotation_score:1) NOT (annotation_score:2)`
- **interleukin-1** α `(protein_name:"interleukin-1 alpha") AND (family:il-1) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[200 TO 400]) NOT (annotation_score:1)`
- **interleukin-1** β: `(protein_name:"interleukin-1 beta") AND (family:il-1) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[200 TO 400]) NOT (annotation_score:1)`

- **Histone H2A**: `(protein_name:"histone h2a") AND (family:histone) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200])`
- **Histone H2B**: `(protein_name:"histone h2b") AND (family:histone) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200])`
- **Cytochrome P450 CYP3**: `(family:"Cytochrome P450") AND ((gene:cyp3) OR (gene:cyp3A*)) AND (fragment:false) NOT (existence:4) NOT (existence:5) NOT (annotation_score:1)`
- **Cytochrome P450 CYP51**: `(family:"Cytochrome P450") AND ((gene:cyp51) OR (gene:cyp51A*) OR (gene:cyp51B*) OR (gene:cyp51C*)) AND (fragment:false) NOT (existence:4) NOT (existence:5) NOT (annotation_score:1)`

The GPCR sequences were collected from the GPCR-PEn database (URL: `https://gpcr.utep.edu/`) [1]. The rhodopsin-like family sequences were reduced using CD-hit [2] with 30% sequence similarity cutoff.

Code

The code to reproduce the computational experiments is available at `https://github.com/Pranav-Machingal/AFS_AAC_SVM.git`. The protein sequences, along with their Uniprot-IDs, which are used in the computational experiments, are provided in the `datasets` folder as `.csv` files for each family.

S2 The SVEA algorithm for AFS

Algorithm 1 ϕ_i Monte-carlo approximation algorithm as suggested in [6]

Input: Feature set $N = \{1, 2, \ldots, n\}$, Number of sample permutations $samPerm$, Number of examples m, Set of coalitions $Sam_co_set = [()]$

Initialise: $v(()) = 0$, $\hat{\phi}_j := 0 \forall j \in N$

Append N to Sam_co_set.
for $s = 1, 2, \ldots, samPerm$ **do**
  Take $\pi \in PermSet(N)$ with probability $\frac{1}{n!}$.
  for $j = 1, 2, \ldots, n$ **do**
    Compute $Pred^j(\pi) = \{\pi(1), \pi(2), \ldots \pi(k-1) | j = \pi(k)\}$
    if $Pred^j(\pi)$ not in Sam_co_set **then**
      Compute $v(Pred^j(\pi)) = 1 - tr_er(Pred^j(\pi))$.
      Append $Pred^j(\pi)$ to Sam_co_set.
    end if
    if $Pred^j(\pi) \cup j$ not in Sam_co_set **then**
      Compute $v(Pred^j(\pi) \cup \{j\}) = 1 - tr_er(Pred^j(\pi) \cup \{j\})$.
      Append $Pred^j(\pi) \cup \{j\}$ to Sam_co_set.
    end if
    $\hat{\phi}_j = \hat{\phi}_j + v(Pred^j(\pi) \cup \{j\}) - v(Pred^j(\pi)))$
  end for
end for
$\hat{\phi}_j = \frac{\hat{\phi}_j}{samPerm}, \forall j \in N$

For n_1 samples of class-1, $\{x_i\}_{i=1}^{n_1}$ and n_2 samples of class-2, $\{x_i\}_{i=n_1+1}^{n_2}$, the class balanced hinge loss based training error with linear classifier using features $S \subseteq N$ is defined as,

$$
\begin{aligned}
tr_er(S) = & \min_{w,\{\xi_i\}_{i=1}^{n_1+n_2}} \frac{1}{2n_1}\sum_{i=1}^{n_1}\xi_i + \frac{1}{2n_2}\sum_{i=n_1+1}^{n_2}\xi_i \\
\text{s.t. } & y_i\left(\sum_{j\in S} w_j x_{ij} + b\right) \geq 1-\xi_i,\ \forall i \in [n_1+n_2] \\
& \xi_i \geq 0,\ \forall i \in [n_1+n_2]
\end{aligned} \tag{1}
$$

The Shapley value is computed using the characteristic function $v(S) = tr_er(\emptyset) - tr_er(S)$. Here $tr_er(\emptyset) = 1$, therefore, $v(S) = 1 - tr_er(S)$. Algorithm 1 describes the Monte-Carlo based approximation algorithm for computing the Shapley value-based feature importance score.

S3 More details for computational experiments

S3.1 Globin Family

The 3D structures of hemoglobin-α/β (PDB ID:1HHO) were aligned with myoglobin (PDB ID:3RGK) using the online pairwise structure alignment tool available at `https://www.rcsb.org/alignment`, with the default parameter settings (`algorithm: jFATCAT(rigid)` — `RMSD Cutoff: 3` — `AFP Distance Cutoff: 1600` — `Fragment Length: 8`).

```
>3RGK.A (Myoglobin)

GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDE
MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHPGD
FGADAQGAMNKALELFRKDMASNYK

>1HHO.A (Hemoglobin-α)

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDL------
SHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAE
FTPAVHASLDKFLASVSTVLTSKYR

---------------------------------------------------------------------

>3RGK.A (Myoglobin)

GLSDGEWQLVLNVWGKVEA
DIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIK
PLAQSHATKHKIPVKYLEFISEAIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK

>1HHO.B (Hemoglobin-β)

HLTPEEKSAVTALWGKV—
NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Figure S1: The highlighted AMINO ACIDS in myoglobin chain correspond to (after structure alignment) the positions which are hemologlobin-α/β tetramer contact points (as identified in Table 3 and Table 4 of [4]). We find that the amino acids K, E, I, which are common in AFS_1(Myoglobin) and AFS_2(Myoglobin), are less in number at the contact residues of hemoglobin tetramer and more in number at the corresponding locations in myoglobin, which is a monomer.

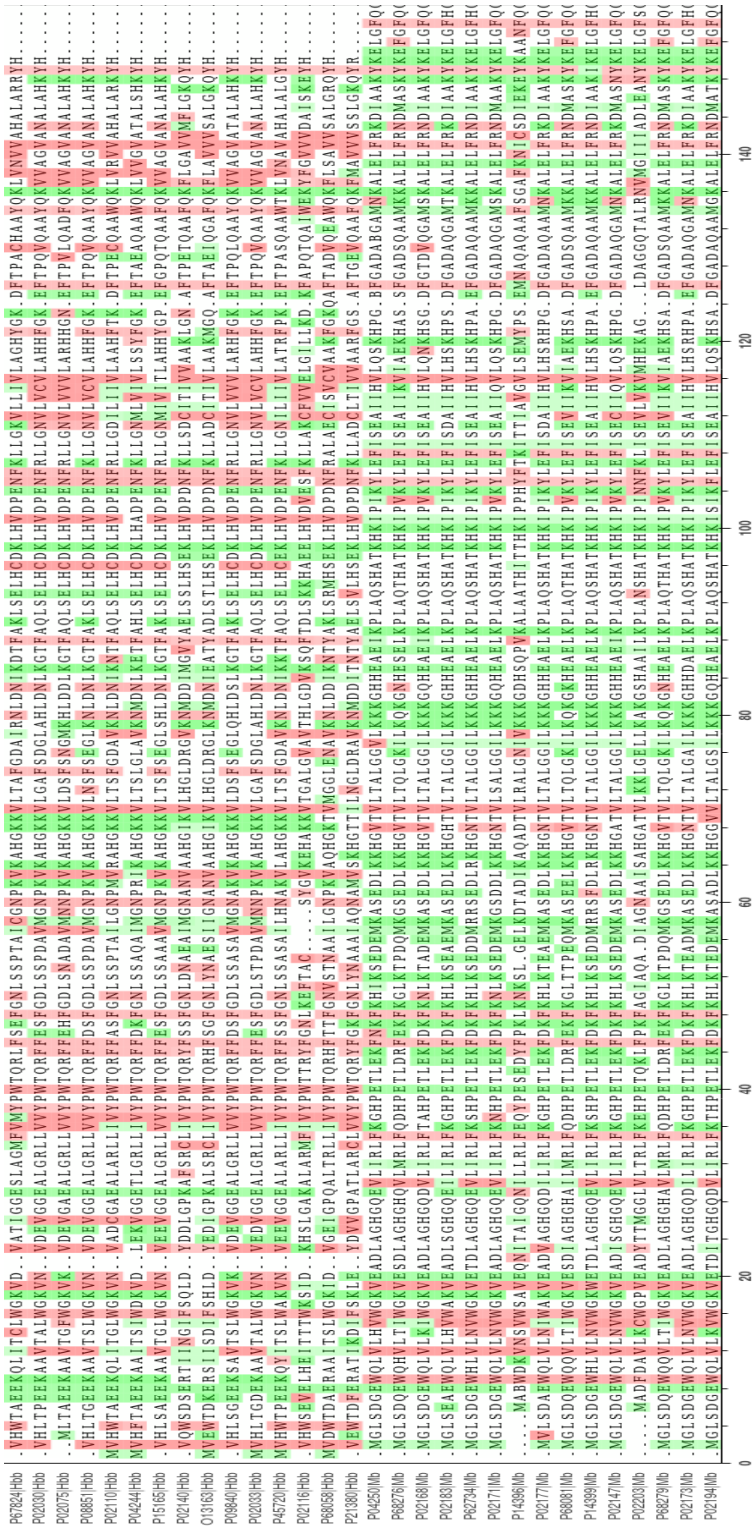


Figure S2: Multiple sequence alignment of hemoglobin-β and myoglobin sequences. 15 sequences on the left are from hemoglobin-β and on the right are from myoglobin. The sequences are randomly selected from the train set of the protein families. AFS(Myoglobin) amino acids are in green and AFS(Hemoglobin-β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 4c)

S3.2 Tubulin

The inter-chain contact residues from the tubulin-α/β heterodimer were identified using ChimeraX 1.4 [3]. The *Contacts* tool available in *Tools* $\rightarrow$ *Structure Analysis* was used with settings as shown in Figure S3. For PDB ID:3JAR we count the residues of chain-A (tubulin-α) and chain-B (tubulin-β) which are in contact with the residues of other tubulin chains. Similarly, for PDB ID:5N5N we count the residues of chain-G (tubulin-α) and chain-B (tubulin-β) which are in contact with the residues of other tubulin chains. The code for counting the *AFS* residues at the identified contact points of the respective chains is available at `https://github.com/Pranav-Machingal/AFS_AAC_SVM.git`.

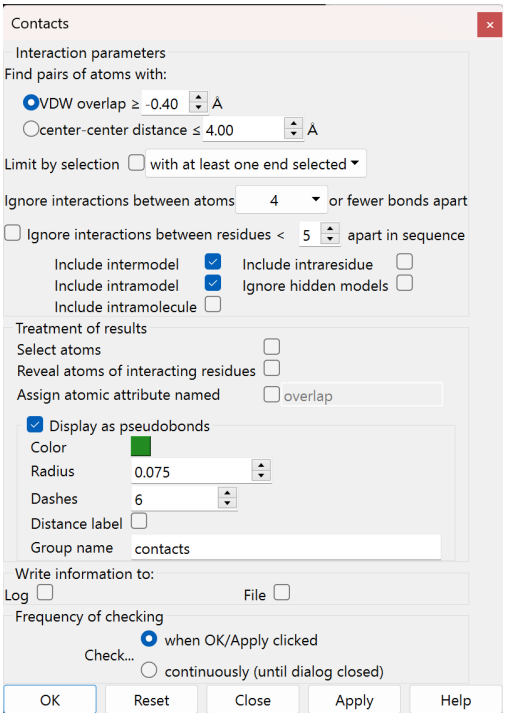


Figure S3: ChimeraX 1.4 [3] settings for identifying inter-chain contact points from the tubulin-α/β heterodimer and from the histone heterooctamer

S3.3 Histone

The inter-chain contact residues of histone H2A and H2B were identified from its heterooctameric structure comprising of two H2A/H2B dimers and one H3/H4 tetramer, using ChimeraX 1.4 [3]. The *Contacts* tool available in *Tools* $\rightarrow$ *Structure Analysis* was used with settings as shown in Figure S3. For PDB ID: 1AOI and 3KWQ, we count the residues of an H2A and an H2B chain, which are in contact with other histone chains in the heterooctameric structure. The code for counting the *AFS* residues at the identified contact points of the respective chains is available at `https://github.com/Pranav-Machingal/AFS_AAC_SVM.git`.

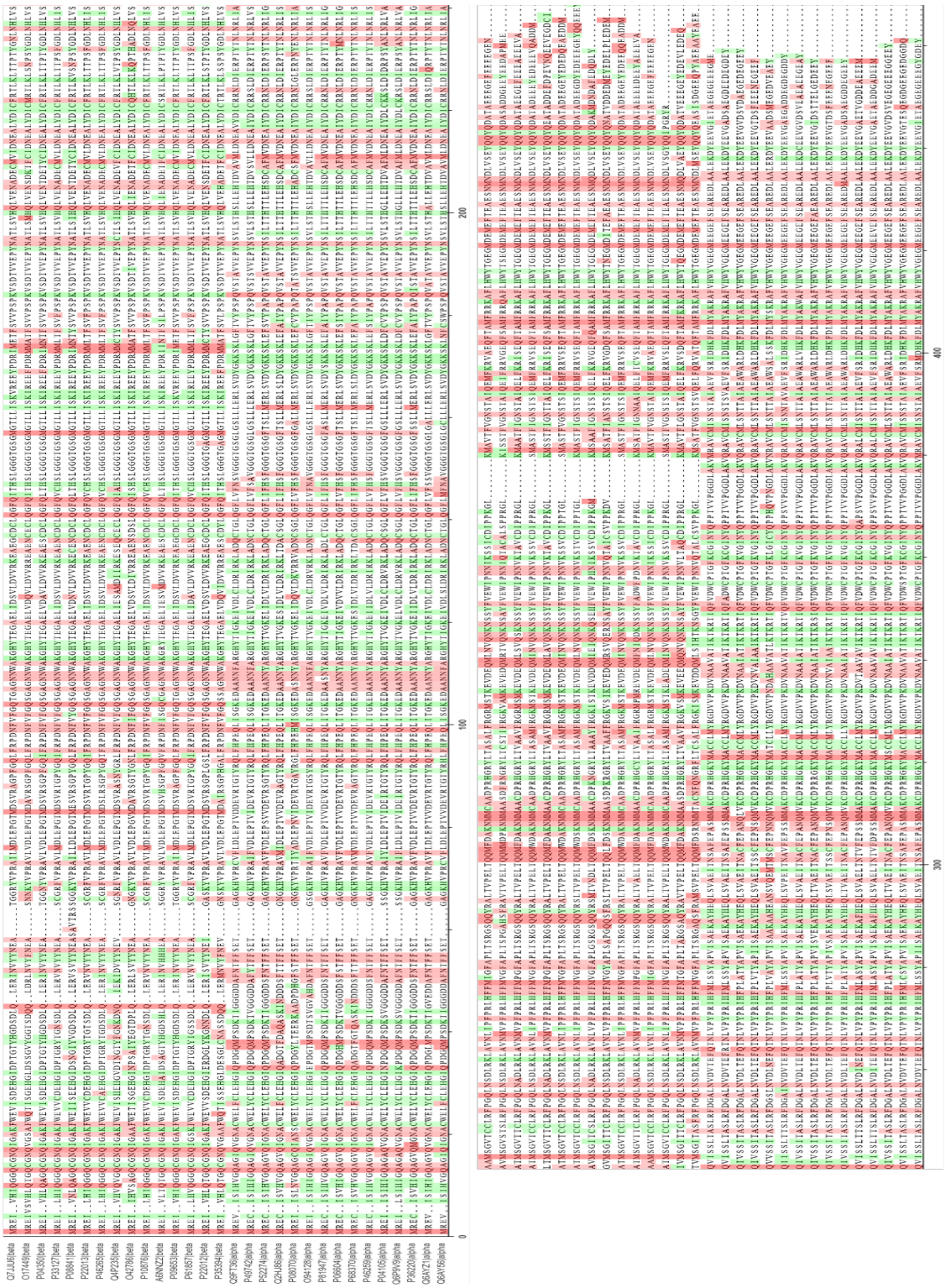


Figure S4: Multiple sequence alignment of tubulin-α and tubulin-β sequences. 15 sequences on the left are from tubulin-β and on the right are from tubulin-α. The sequences are randomly selected from the train set of the protein families. AFS(Tubulin-α) amino acids are in green and AFS(Tubulin-β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 2c)

Figure S5: Multiple sequence alignment of histone H2A and histone H2B sequences. 15 sequences on the left are from histone H2B and on the right are from histone H2B. The sequences are randomly selected from the train set of the protein families. AFS(Histone H2A) amino acids are in green and AFS(Histone H2B) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 2e)

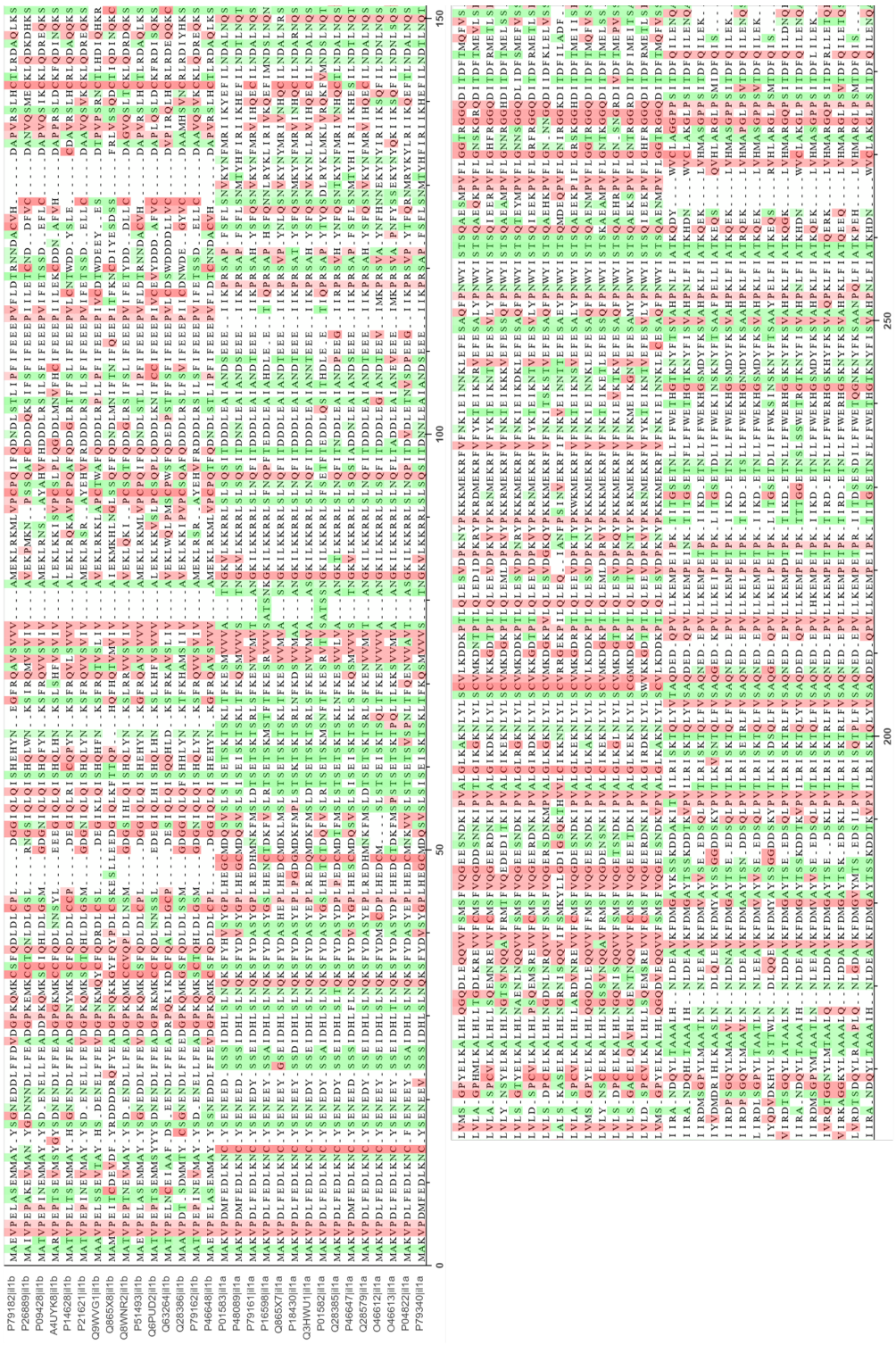


Figure S6: Multiple sequence alignment of interleukin-1 α and interleukin-1 β sequences. 15 sequences on the left are from interleukin-1 β and on the right are from interleukin-1 α. The sequences are randomly selected from the train set of the protein families. AFS(Interleukin-1 α) amino acids are in green and AFS(Interleukin-1 β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 2d)

References

[1] K. Begum, J. E. Mohl, F. Ayivor, E. E. Perez, and M.-Y. Leung. GPCR-PEnDB: a database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors. *Database*, 2020, 11 2020. ISSN 1758-0463. doi: 10.1093/database/baaa087. URL https://doi.org/10.1093/database/baaa087.

[2] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

[3] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin. Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021. doi: 10.1002/pro.3943. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3943.

[4] M. Shionyu, K. Takahashi, and M. Gō. Variable subunit contact and cooperativity of hemoglobins. *J. Mol. Evol.*, 53(4-5):416–429, Oct. 2001.

[5] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.

[6] S. Tripathi, N. Hemachandra, and P. Trivedi. Interpretable feature subset selection: A Shapley value based approach. In *2020 IEEE BigData*, pages 5463–5472, 2020. doi: 10.1109/BigData50022.2020.9378102.