proteomes

Article

Revealing Subtle Functional Subgroups in Class A Scavenger Receptors by Pattern Discovery and Disentanglement of Aligned Pattern Clusters

Pei-Yuan Zhou [1], En-Shiun Annie Lee [2], Antonio Sze-To [3] and Andrew K. C. Wong [3,*]

[1] VaryWave Technology Co., Ltd., 538A, Core Building 2, Hong Kong Science Park, Shatin, NT, Hong Kong; choupeiyuan@gmail.com

[2] VerticalScope Inc., 111 Peter Street, Suite 900, Toronto, ON M5V 2H1, Canada; alee@verticalscope.com

[3] Systems Design Engineering, 5th, 6th Floor, 200 University Avenue West, University of Waterloo, Waterloo, ON N2L 3G1, Canada; hy2szeto@uwaterloo.ca

* Correspondence: akcwong@uwaterloo.ca; Tel.: +1-519-888-4567 (ext. 35775)

Received: 24 November 2017; Accepted: 1 February 2018; Published: 8 February 2018

Abstract: A protein family has similar and diverse functions locally conserved as aligned sequence segments. Further discovering their association patterns could reveal subtle family subgroup characteristics. Since *aligned residues associations* (ARAs) in Aligned Pattern Clusters (APCs) are complex and intertwined due to entangled function, factors, and variance in the source environment, we have recently developed a novel method: Aligned Residue Association Discovery and Disentanglement (ARADD) to solve this problem. ARADD first obtains from an APC an ARA Frequency Matrix and converts it to an adjusted *statistical residual vector space* (SRV). It then disentangles the SRV into Principal Components (PCs) and Re-projects their vectors to a SRV to reveal succinct orthogonal AR groups. In this study, we applied ARADD to class A scavenger receptors (SR-A), a subclass of a diverse protein family binding to modified lipoproteins with diverse biological functionalities not explicitly known. Our experimental results demonstrated that ARADD can unveil subtle subgroups in sequence segments with diverse functionality and highly variable sequence lengths. We also demonstrated that the ARAs captured in a Position Weight Matrix or an APC were entangled in biological function and domain location but disentangled by ARADD to reveal different subclasses without knowing their actual occurrence positions.

Keywords: residue association; pattern discovery; disentanglement; protein; class A scavenger receptors

1. Introduction

Proteins from the same family have similar but also diverse functions [1]. Hence, discovering conserved yet varied sequence patterns with various subgroup characteristics is important for understanding the protein functionality of a protein family. However, existing broad and imprecise grouping definition and processes neglect the diversity of certain protein families [1]. One protein family may contain members that are strikingly variable in sequence length, three-dimensional structure, and hence biological function, where similar and diverse functional domains may reside in different sequence locations within the family. For example, class A scavenger receptors (SR-A) [1], biologically important for binding on modified lipoproteins, are complex particles composed of multiple proteins that transport all fat molecules (lipids) around the body within the water outside cells to promote macrophage differentiation into foam cells, leading to chronic conditions such as atherosclerosis [2]. SR-A is a diverse family of proteins classified based on their ability to bind modified lipoproteins [3]. Although the five members (Marco, Sra, Scara3, Scara4, Scara5) [1] of this family could bind modified lipoproteins, they are different in terms of their sequence patterns, locations, structures, and hence functions. For instance, within the same family, their protein length varies from 451 to 732 residues with the functional domains residing in different sequence locations. Thus SR-A is a protein family with conserved yet diverse function subgroups, ideal for exploration.

In order to reveal the functional subgroup characteristics of conserved sequence patterns corresponding to the diverse members of a protein family, we need mathematical transformations to disentangle the intriguing functionality related to conserved functional regions to reveal subgroups not explicitly manifested from the data. To this end, we have developed a novel method, known as Aligned Residue Association Discovery and Disentanglement (ARADD) [4], based on our previous work Attribute Value Association Discovery and Disentanglement (AVADD) [5], to discover and then disentangle the statistical representation of the Aligned Residue Associations (ARAs) derived from Aligned Pattern Clusters (APCs) for revealing their subgroups and subgroup characteristics.

In this study, we conducted experiments on SR-A, and found that the sequence patterns that are aligned and clustered into an APC via sequence similarity could be entangled in biological functions and domain location. Hence, we applied ARADD [4] to disentangle the statistics so as to reveal the ARs and ARA Clusters (Patterns) corresponding to their functionality and residing location in the family. We further showed that even the ARAs captured in a Position Weight Matrix (PWM) [6] discovered by the famous motif discovery MEME [7] could be entangled in biological functions and domain location. We then demonstrated the effectiveness of ARADD [4] in disentangling ARA patterns to reveal functional subgroups. The major contribution of our study is three-fold.

1. We were able to show that sequence patterns in the representation model, could be mixed or entangled in functionality and location through the study of the SR-A data.
2. We validated that ARADD [4] could reveal functional subgroups and subgroup characteristics of APCs and locate their residing domains through the case study on SR-A. Understanding its subgroup characteristics could also render new knowledge for gene therapy applications [2].
3. We demonstrated that the sequence patterns captured by the Position Weight Matrix (PWM) [6] could be entangled in biological functions and domain location, and they needed to be disentangled by an effective method such as ARADD [4].

The focus of this paper, apart from introducing ARADD [4] and emphasizing its novelty, is to produce a succinct analysis of subgroup characteristics obtained through pattern disentanglement, via the case study on SR-A. To our knowledge, no studies have reported similar experimental results.

The paper layout is as follows. Section 2 does a literature survey. Section 3 explains the methodology. Section 4 describes the materials and Section 5 illustrates the experimental results. Section 6 outlines the biological significance. Section 7 summarizes and concludes the research.

2. Related Work

Traditionally, computational sequence analysis methods have been developed to identify conserved sequence patterns from a protein family. Methods such as Multiple Sequence Alignment (MSA) [8] are only suitable for globally homologous sequences with a high level of sequence similarity [9], and motif discovery [10]; another method, is based on probabilistic model (such as position weight matrix [6]) which assumes independence between residue columns to represent the conserved sequence patterns. Such independence assumption is unrealistic in many cases, where correlation of residues along the sequence is commonly observed [11,12].

Pattern discovery is an essential element in predictive analytics [13,14] for knowledge discovery and analysis. Its essence is to discover patterns (motifs) occurring in the data to reveal association patterns for interpretation and classification [15]. Hence, we develop an algorithm to obtain APCs [16,17] which capture functional residue association and site conservation. Since APCs contain aligned residues in strong statistical association sequence patterns, this representation is more knowledge-rich [16,17] when compared with MSA and probabilistic models. Hence, APCs reveal locally conserved yet diverse function patterns of protein families. APCs can reveal biological function in conserved regions of protein families.

Association rule mining [18] is the most well-known methodology for mining item sets in relational dataset in the area of data mining. Algorithms such as Apriori [19] and FP-growth [20] can be applied for capturing associations from relational dataset. However, frequent patterns discovered by the above algorithms are extremely sensitive to threshold settings. Our new method, Aligned Residue Association Discovery and Disentanglement (ARADD), evolved from our AVADD [5] method, and is proposed to solve this problem. ARADD is able to reveal residue association patterns in different orthogonal PCs and *Re-projected SRVs* (RSRVs). ARADD is able to correlate different functionalities based only on the confidence intervals. As demonstrated in the results reported in this paper, ARADD achieves stable and succinct results in a simple fashion.

As observed in our recent paper [5], a challenging problem encountered when discovering association patterns is that the association could be masked or obscured in the data due to the entanglement of unknown factors in their source environment. To resolve this problem for general relational datasets, we developed a novel method known as AVADD in our previous work [5]. In this paper, we transformed the existing methodology to discover and disentangle ARAs from APCs. The reasons are as follows: (1) the aligned columns (sites) in an APC can be treated as attributes of relational dataset; (2) the residing residues on these sites can be treated as attribute values; (3) the residue associations in an APC can be treated as attribute value associations. The extended ARADD from AVADD [5] could discover and disentangle ARAs from APCs as if AVADD [5] could do that on *attribute value associations* (AVAs) from a relational dataset. This is the most game-changing part of ARADD in comparison with existing methods. Due to such capability, subtle entangled subgroup characteristics masked or conspicuous in APCs can be revealed. To the best of our knowledge, only ARADD could disentangle such ARA patterns in APCs while no other reported methods could.

In summary, compared to the above-mentioned algorithms, ARADD solved the most difficult problems in discovering and analyzing subgroup characteristics of APCs containing entangled associations and the variation among them in their aligned sequence patterns. We should conclude that: (1) local associations may occur in different sequence locations or functional domains; (2) subgroups with similar patterns (motifs) could have small differences in functionality; (3) similar functionality may occur in different function groups and domains; and (4) multiple functionalities may occur within a functional group dominated by a key function. We refer to such entwined phenomena as the results of entangled ARA patterns.

3. Methods

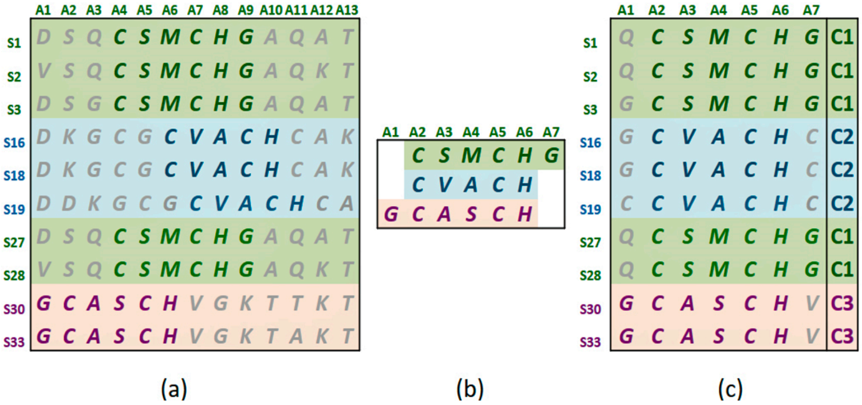
The method used in this paper is evolved from AVADD [5], which was developed by our team. It was used for discovering and disentangling AVAs from mixed-mode relational datasets (RDS) [5] very successfully. In this study, we extended AVADD to discover and disentangle the statistical representation SRV of the ARAs obtained from APCs to reveal their subgroups and subgroup characteristics as well as to locate their functional domains.

By Aligned Pattern Cluster (APC) [16,17], we mean an array of sequence segments containing a cluster of aligned statistically significant sequence patterns grouped together according to their similarity though alignment [16,17]. Figure 1a shows the statistically significant sequence patterns discovered from the sequence data. Figure 1b represents the pattern space of the APC when the discovered patterns are aligned. Figure 1c is referred to as the data space of the APC showing all the sequence data segments containing the patterns in the APC with their pattern sequence ID and head position registered. Through the data space of APCs, any ARs or ARAs of an APC identified in any PCs and RSRVs, respectively, can be located.

Figure 2 shows the three major steps of ARADD. (1) ARADD converts an APC into an ARA Frequency Matrix (ARAFM) (Definition 1). It then converts the ARAFM into an Adjusted Statistical Residual Space (SRV) so that each frequency entry in ARAFM is transformed into a Statistical Residual

(SR) that accounts for the deviation of the frequency of occurrences of the ARA from the expected frequenting if the associations were random. In the SRV, each row represents a vector of an AR (referred to as an AR-vector or a-vector) whose coordinates represent the SRs of that AR associated with other ARs corresponding to the column AR-vectors; (2) Next, ARADD conducts the Principal Component Decomposition on the SRV to obtain the top Principal Components (PCs) ranked according to their variance, and then projects all the AR-vectors in the SRV onto the PC axis. The new SRV containing the vector projections on the PC is referred to as the Re-projected SRV. The new set of coordinates of these projections reflects the SRs of that AR associating with other ARs corresponding to the column vectors; (3) Finally, for each PC, ARADD identifies the distinct ARs and/or AR clusters with variance ≥1.0 from the center and then obtains the SR value of the ARA between the ARs within each AR clusters in the RSRV to reveal the association patterns.

Figure 3 shows the Graphical User Interface (GUI) of the ARADD software. After loading the CSV file of an APC; pushing the button labeled "Generate FM and SRV" will create both the ARAFM and SRV for the APC. The process in Step 2 will generate the set of top PCs and their corresponding RSRVs according to the number of PCs or percentage of the variance assigned in the box. The process in Step 3 will highlight the sub-cluster results according to the assigned confidence interval in the box.



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
S1	D	S	Q	**C**	**S**	**M**	**C**	**H**	**G**	A	Q	A	T
S2	V	S	Q	**C**	**S**	**M**	**C**	**H**	**G**	A	Q	K	T
S3	D	S	G	**C**	**S**	**M**	**C**	**H**	**G**	A	Q	A	T
S16	D	K	G	C	G	**C**	**V**	**A**	**C**	**H**	C	A	K
S18	D	K	G	C	G	**C**	**V**	**A**	**C**	**H**	C	A	K
S19	D	D	K	G	C	G	**C**	**V**	**A**	**C**	**H**	C	A
S27	D	S	Q	**C**	**S**	**M**	**C**	**H**	**G**	A	Q	A	T
S28	V	S	Q	**C**	**S**	**M**	**C**	**H**	**G**	A	Q	K	T
S30	**G**	**C**	**A**	**S**	**C**	**H**	V	G	K	T	T	K	T
S33	**G**	**C**	**A**	**S**	**C**	**H**	V	G	K	T	A	K	T

(a)

A1	A2	A3	A4	A5	A6	A7
	C	**S**	**M**	**C**	**H**	**G**
	C	**V**	**A**	**C**	**H**	
G	**C**	**A**	**S**	**C**	**H**	

(b)

	A1	A2	A3	A4	A5	A6	A7	
S1	Q	**C**	**S**	**M**	**C**	**H**	**G**	C1
S2	Q	**C**	**S**	**M**	**C**	**H**	**G**	C1
S3	G	**C**	**S**	**M**	**C**	**H**	**G**	C1
S16	G	**C**	**V**	**A**	**C**	**H**	C	C2
S18	G	**C**	**V**	**A**	**C**	**H**	C	C2
S19	C	**C**	**V**	**A**	**C**	**H**	C	C2
S27	Q	**C**	**S**	**M**	**C**	**H**	**G**	C1
S28	Q	**C**	**S**	**M**	**C**	**H**	**G**	C1
S30	**G**	**C**	**A**	**S**	**C**	**H**	V	C3
S33	**G**	**C**	**A**	**S**	**C**	**H**	V	C3

(c)

Figure 1. Protein Sequence and APC Data (**a**) Protein sequence dataset with high order association patterns (in bold) discovered by [11,12] with top row = aligned sites; fist column = sequence ID; (**b**) Aligned Pattern Cluster Pattern Space (*APC-P*) obtained using [16,17]; (**c**) APC Data Space (*APC-D*). The last column is the class labels.

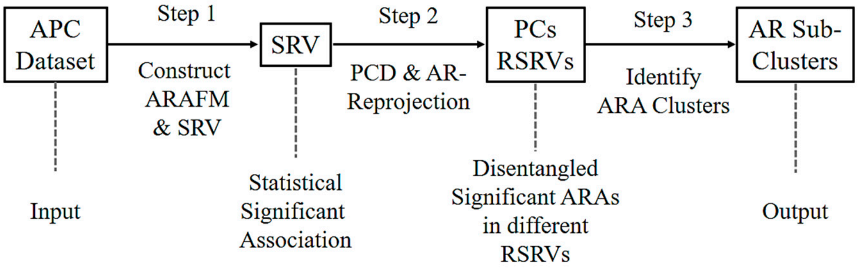
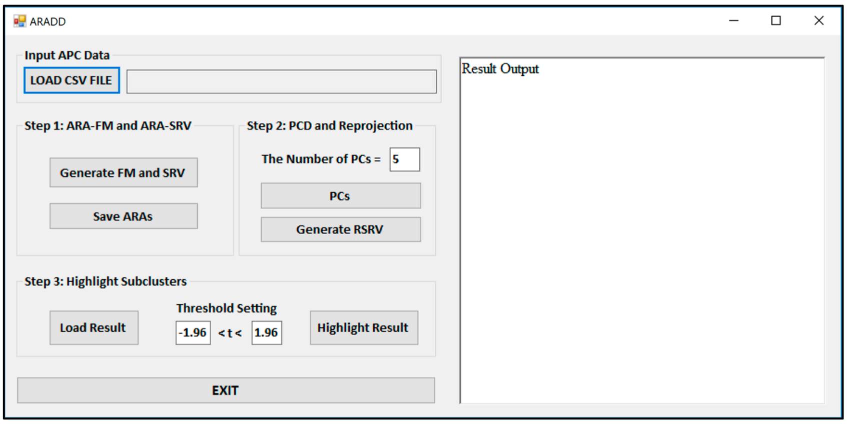


Figure 2. An overview of the proposed algorithm of Aligned Residue Association Discovery and Disentanglement (ARADD).

We can represent an APC dataset by M amino acids (residues) on N amino acid sites, denoted as $A = \{A_1, \ldots A_n \ldots A_N\}$. Each amino acid site A_n can assume a categorical value as a residue type. Thus A_n contains I_n values, denoted as $A_n = \left\{A_n^1, A_n^2, \ldots A_n^{I_n}\right\}$. Hence, $I = \sum_{n=1}^{N} I_n$ represents the total number of residue types (values) of all sites in an APC.



The Number of PCs = 5

-1.96 < t < 1.96

Figure 3. An example of the intermediate process and data produced by our Attribute Value Association Discovery and Disentanglement (AVADD) prototype via an Interactive Decision Support GUI. In the main ARADD GUI, the left-hand side shows the process of AVADD and the right-hand side shows the output result of each step but not displayed in this figure.

ARA represents the residue associations between residues on different aligned columns of the APC (i.e., the residue pairs). The ARA, $\left(A_n^i \leftrightarrow A_{n'}^j\right)$ represent the residue association between two aligned residues with the aligned site positions n and n' and the residue types i, j respectively.

Definition 1. *ARA Frequency Matrix. An ARAFM is a matrix of frequency counts of ARA between two aligned amino acids (residues) on two aligned sites, say* $\left(A_n^i \leftrightarrow A_{n'}^j\right)$*, within the same protein sequence. We denote the frequency by FM* $\left(A_n^i \leftrightarrow A_{n'}^j\right)$*, where* A_n^i *represents the residue of* i^{th} *type on the* n^{th} *aligned site in the APC, and* $A_{n'}^j$ *represents the residue of the* j^{th} *type on the* n'^{th} *aligned site* ($n \neq n'$)*. Hence, ARAFM is an* $I \times I$ *matrix.*

Now we shall describe the operations associated with each step in the ARADD GUI.

Step 1: Construct Adjusted Statistical Residual Vectors Space (SRV):

The button "Generate FM and SRV" in Figure 3 is used to construct the ARAFM first and then to convert the ARAFM into SRV (Definition 2).

Definition 2. *ARA Adjusted Statistical Residual Vector Space. An SRV is a vector space such that the* j^{th} *jjas coordinate of its* i^{th} *iithat row vector (a-vector or more precisely* A_n^i-*vector), corresponding to the* i^{th} *type residue on the* n^{th} *site in the APC. We denote the* n' *coordinate of the* A_n^i-*vector by* $SRV_{A_{n'}^j}$ *which is the SR obtained from* $FM\left(A_n^i \leftrightarrow A_{n'}^j\right)$*. Hence, an SRV can be expressed as a set of row vectors:* $SRV = < SRV_{A_1^1}, \ldots SRV_{A_1^{I_1}}, \ldots SRV_{A_n^{I_n}} \ldots, SRV_{A_N^{I_N}} >$*, where* $I = \sum_{n=1}^{N} I_n$ *is the total number of ARAs, and* I_1 *is the total number of unique SRs for the first aligned column* (A_1)*, and* I_n *is the number of unique amino acid for the* n^{th} *aligned column* (A_n)*. An a-vector is hence denoted as* $SRV_{A_n^i} = \{SR(A_n^i \leftrightarrow A_1^1), \ldots SR(A_n^i \leftrightarrow A_1^{I_1}), \ldots SR(A_n^i \leftrightarrow A_N^{I_N})\}$*, where* $SR(A_n^i \leftrightarrow A_{n'}^j)$ *represents the statistical residual for ARA* $(A_n^i \leftrightarrow A_{n'}^j)$*, and* $SR(A_n^i \leftrightarrow A_n^i) = 0$*.*

By replacing each ARA frequency with its *Adjusted Standard Statistical Residual* (SR), we can construct the SRV. We treat SRV as a vector space such that each row is taken as a row vector (AR-vector or a-vector) with its coordinates representing the SR of its ARAs associating with other ARs denoted by the column a-vectors.

To obtain statistically significant information from an APC, we transform an ARAFM into a SRV by converting each ARA frequency $FM\left(A_n^i \leftrightarrow A_{n'}^j\right)$ into an SR, denoted as $sr(A_n^i \leftrightarrow A_{n'}^j)$.

$$sr\left(A_n^i \leftrightarrow A_{n'}^j\right) = sr_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{1}$$

where o_{ij} represents the total number of occurrences when $A_{n'} = A_{n'}^j$ and $A_n = A_n^i$; $e_{ij} = \frac{\sum_{u=1}^{J} \mathrm{o}_{iu} \sum_{u=1}^{I} \mathrm{o}_{uj}}{M}$, where $\sum_{u=1}^{J} \mathrm{o}_{iu}$ represents the total number of counts when $A_{n'} = A_{n'}^j$ $\sum_{u=1}^{I} \mathrm{o}_{uj}$ represents the total number of counts when $A_n = A_n^i$ and M is the total number of records.

To reveal the statistical significance of an ARA, here, sr_{ij} measures the deviation of the observed frequency, o_{ij}, of the ARA from its default model e_{ij} assuming that the occurrence is a random association. Unlike the standardized residual, the Adjusted Statistical Residual considers the overall size of the sample and gives a more accurate indication of how far the observed count is from the expected count. So, in this paper we abbreviate Adjusted Statistical Residue as "*SR*", to be consistent with our previous paper [5]. We denote *SR* as $SR(A_n^i \leftrightarrow A_{n'}^j)$, which is the significance of ARAs. The value of $SR\left(A_n^i \leftrightarrow A_{n'}^j\right)$ is calculated using Equation (2).

$$SR\left(A_n^i \leftrightarrow A_{n'}^j\right) = \frac{sr_{ij}}{\sqrt{V_{ij}}} \tag{2}$$

Here, v_{ij} is the maximum likelihood estimate of the variance of SR_{ij}, and is defined as: $v_{ij} = (1 - \sum_{u=1}^{I_n} o_{uj}/M) \times (1 - \sum_{u=1}^{I_{n'}} o_{iu}/M)$. We should note that the respective statistical thresholds for significant and insignificant ARAs remain unchanged.

Although the SRs can reveal the significance of an ARA, subtle associations could still be entangled and masked. Hence, by treating the ARA *SR* matrix as a *SR* vector space (SRV), we have developed a novel method to disentangle the SRV through Principal Component Decomposition (PCD) onto a set of PCs and reproject the projections of the AR-vectors on each PC back to a new SRV referred to as its corresponding RSRV.

Step 2: Conduct PCD on SRV and Obtain for Each PC Its Corresponding RSRV.

The button "PC" in Figure 3 is used to initiate the application of the PCD (Definition 3) on the SRV and obtain a number of top PCs according to the number (or the threshold of the total variance) set in the window. In the meantime, in order to disentangle the discovered AVAs from SRV, the a-vector projections on each PC is re-projected onto a new SRV, referred to as the Re-projected SRV (Definition 4). The new transformed a-vector positions in the RSRV correspond to a new set of ARA SRs for each AR with other ARs in the RSRV. These new positions of a-vectors reflect the ARAs captured in the corresponding PC.

Definition 3. *Principal Components. In PCD, PCs are a set of k PCs, denoted as PC = {* $PC_1, PC_2, \ldots PC_k$*}, where* PC_k *is a set of projections of the a-vectors from SRV, denoted as* PC_k *= {* $PC_k\left(A_n^i\right) \middle| n = 1, 2, \ldots N,\ i = 1, \ldots I_n$ *}, where N is the total number of all aligned columns in APC and* I_n *is the total number of distinct amino acids on the column* A_n*.*

Definition 4. *Re-projected SRV (RSRV). RSRV is the SRV containing the a-vector projections on a PC. The coordinates of an a-vector projection on the PC in the RSRV represent the ARA SRs of the AR of that a-vector associating with other ARs corresponding to the column vectors as captured by the PC (Equation (3)).*

$$RSRV_k = SRV \cdot PC_k \cdot {PC_k}^T \tag{3}$$

After PCD, we first identify the distinct projection(s) of the a-vectors and their furthest clusters from the mean (center) of the PC-Axis. The a-vector projections in PCs with large eigenvalue should have strong association or strong presence captured by the orthogonal PCs.

If the class labels are included in the APCs, its position in the PCs appears just as a virtual AR (essentially the centroid of the ARs within an AR cluster pertaining to that class). In the illustrative example in Figure 4, the colored dots enclosed in the square boxes are centroids of the AR clusters associated to specific classes like mammal, plant, etc.

Hence, the new positions of the projections of the a-vectors on the PC, when transformed to the SRV, represent the a-vectors with a new set of coordinates in the RSRV. We mark the correspondence by attaching to them the same subscript, i.e., k in PC_k and $RSRV_{k.}$

Figure 5 shows how the SRV with AVAs, related to the taxonomical subgroups and obtained from APC-6382 (described later in the Results Section), is disentangled into different RSRVs (such as, RSRV1 and RSRV2) and thus revealing ARA subgroups corresponding to different functionality pertaining to the taxonomical classes in the region. Note that the ARAs entangled in SRV (Figure 5a) are disentangled into succinct sub-groups corresponding to difference taxonomical classes in RSRV1 and RSRV2 (Figure 5b,c, respectively).

To summarize this step, PCD uses an orthogonal transformation to transform a set of possible correlated variables into a set of linear uncorrelated variables known as PCs. The first PC, PC_1, has the largest possible variance, which accounts for ARs with the highest ARAs with other ARs. PCs with less variance follow.

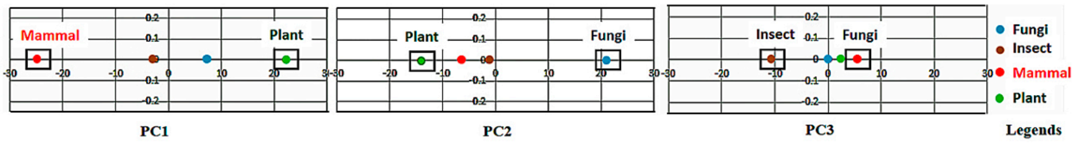
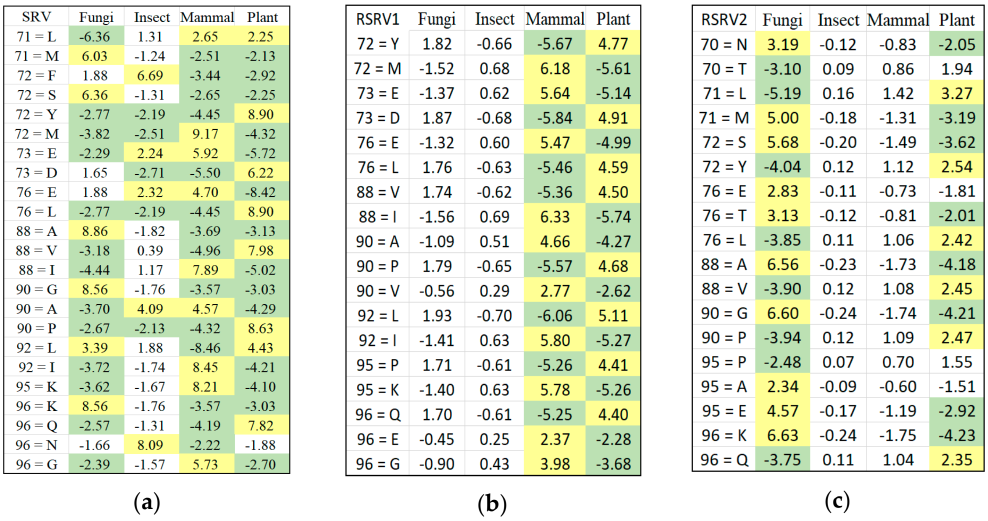


Figure 4. PC1, PC2 and PC3 plots for APC-6382 used as an illustrative example in Section 5. Here, only the points representing the class labels are displayed. A full plot with or without class labels is given in Figure 6.



SRV	Fungi	Insect	Mammal	Plant
71 = L	-6.36	1.31	2.65	2.25
71 = M	6.03	-1.24	-2.51	-2.13
72 = F	1.88	6.69	-3.44	-2.92
72 = S	6.36	-1.31	-2.65	-2.25
72 = Y	-2.77	-2.19	-4.45	8.90
72 = M	-3.82	-2.51	9.17	-4.32
73 = E	-2.29	2.24	5.92	-5.72
73 = D	1.65	-2.71	-5.50	6.22
76 = E	1.88	2.32	4.70	-8.42
76 = L	-2.77	-2.19	-4.45	8.90
88 = A	8.86	-1.82	-3.69	-3.13
88 = V	-3.18	0.39	-4.96	7.98
88 = I	-4.44	1.17	7.89	-5.02
90 = G	8.56	-1.76	-3.57	-3.03
90 = A	-3.70	4.09	4.57	-4.29
90 = P	-2.67	-2.13	-4.32	8.63
92 = L	3.39	1.88	-8.46	4.43
92 = I	-3.72	-1.74	8.45	-4.21
95 = K	-3.62	-1.67	8.21	-4.10
96 = K	8.56	-1.76	-3.57	-3.03
96 = Q	-2.57	-1.31	-4.19	7.82
96 = N	-1.66	8.09	-2.22	-1.88
96 = G	-2.39	-1.57	5.73	-2.70

(**a**)

RSRV1	Fungi	Insect	Mammal	Plant
72 = Y	1.82	-0.66	-5.67	4.77
72 = M	-1.52	0.68	6.18	-5.61
73 = E	-1.37	0.62	5.64	-5.14
73 = D	1.87	-0.68	-5.84	4.91
76 = E	-1.32	0.60	5.47	-4.99
76 = L	1.76	-0.63	-5.46	4.59
88 = V	1.74	-0.62	-5.36	4.50
88 = I	-1.56	0.69	6.33	-5.74
90 = A	-1.09	0.51	4.66	-4.27
90 = P	1.79	-0.65	-5.57	4.68
90 = V	-0.56	0.29	2.77	-2.62
92 = L	1.93	-0.70	-6.06	5.11
92 = I	-1.41	0.63	5.80	-5.27
95 = P	1.71	-0.61	-5.26	4.41
95 = K	-1.40	0.63	5.78	-5.26
96 = Q	1.70	-0.61	-5.25	4.40
96 = E	-0.45	0.25	2.37	-2.28
96 = G	-0.90	0.43	3.98	-3.68

(**b**)

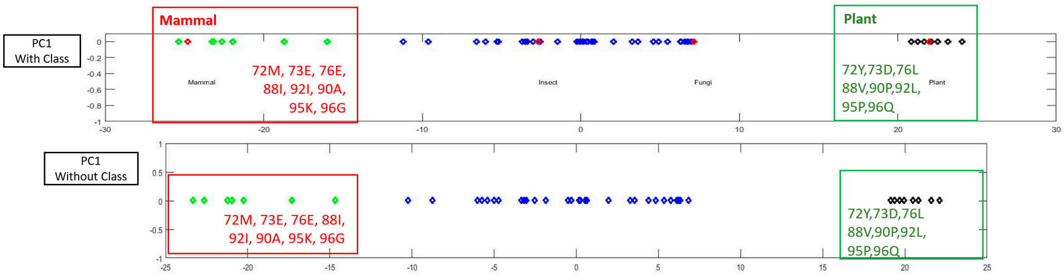
RSRV2	Fungi	Insect	Mammal	Plant
70 = N	3.19	-0.12	-0.83	-2.05
70 = T	-3.10	0.09	0.86	1.94
71 = L	-5.19	0.16	1.42	3.27
71 = M	5.00	-0.18	-1.31	-3.19
72 = S	5.68	-0.20	-1.49	-3.62
72 = Y	-4.04	0.12	1.12	2.54
76 = E	2.83	-0.11	-0.73	-1.81
76 = T	3.13	-0.12	-0.81	-2.01
76 = L	-3.85	0.11	1.06	2.42
88 = A	6.56	-0.23	-1.73	-4.18
88 = V	-3.90	0.12	1.08	2.45
90 = G	6.60	-0.24	-1.74	-4.21
90 = P	-3.94	0.12	1.09	2.47
95 = P	-2.48	0.07	0.70	1.55
95 = A	2.34	-0.09	-0.60	-1.51
95 = E	4.57	-0.17	-1.19	-2.92
96 = K	6.63	-0.24	-1.75	-4.23
96 = Q	-3.75	0.11	1.04	2.35

(**c**)

Figure 5. Part of SRV and RSRVs for APC-6382. (**a**) SRV shows that there is no clear partition between classes; (**b**) In RSRV1, plant and mammal are succinctly separated; (**c**) In RSRV2, plant and fungi are also distinct. Such disentanglements are captured in PC1 and PC2 as shown in Figures 4 and 6.

Step 3: Identify Distinct ARs and AR-Clusters Through Their Statistical Significant ARA SRs Between ARs.

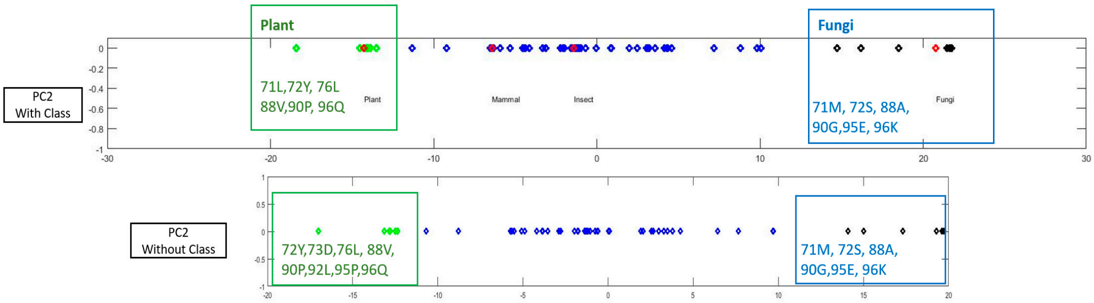
The button "Highlight Result" in Figure 3 is implemented for highlighting the results from the sub-clusters. When a cluster of ARs forms an association pattern, they should share strong ARAs. By setting the threshold to 1.96, the strong ARAs can be grouped together to represent an AR sub-cluster in the RSRVs. A careful comparison of the PCs (Figure 4) and their corresponding RSRVs (Figure 5) shows that the distinct AR clusters captured in PCs are reflected by the statistical significant ARAs in their corresponding rows (a-vectors) in the RSRVs (yellow cells). More succinct representations are found in Figures 6 and 7.



	Mammal	Plant
92 = L	-6.06	5.11
73 = D	-5.84	4.91
72 = Y	-5.67	4.77
90 = P	-5.57	4.68
76 = L	-5.46	4.59
88 = V	-5.36	4.50
95 = P	-5.26	4.41
96 = Q	-5.25	4.40
96 = E	2.37	-2.28
90 = V	2.77	-2.62
96 = G	3.98	-3.68
90 = A	4.66	-4.27
76 = E	5.47	-4.99
73 = E	5.64	-5.14
95 = K	5.78	-5.26
92 = I	5.80	-5.27
72 = M	6.18	-5.61
88 = I	6.33	-5.74

	88 = I	72 = M	92 = I	95 = K	73 = E	76 = E	90 = A	96 = G	90 = V	96 = E	96 = Q	95 = P	88 = V	76 = L	90 = P	72 = Y	73 = D	92 = L
92 = L	-6.07	-5.93	-5.56	-5.54	-5.41	-5.25	-4.47	-3.81	-2.66	-2.28	5.04	5.05	5.15	5.24	5.35	5.44	5.60	5.82
73 = D	-5.96	-5.83	-5.47	-5.46	-5.33	-5.17	-4.42	-3.79	-2.68	-2.31	4.73	4.74	4.83	4.92	5.03	5.11	5.27	5.47
72 = Y	-5.87	-5.74	-5.39	-5.38	-5.25	-5.10	-4.37	-3.76	-2.68	-2.32	4.52	4.52	4.61	4.70	4.80	4.89	5.04	5.24
90 = P	-5.78	-5.66	-5.32	-5.30	-5.18	-5.03	-4.31	-3.71	-2.65	-2.30	4.42	4.43	4.52	Plant		4.79	4.94	5.13
76 = L	-5.67	-5.55	-5.21	-5.19	-5.08	-4.93	-4.23	-3.64	-2.60	-2.26	4.33	4.34	4.43	[illegible]	[illegible]	4.69	4.84	5.03
88 = V	-5.54	-5.42	-5.09	-5.08	-4.96	-4.82	-4.12	-3.55	-2.53	-2.19	4.28	4.28	4.37	4.45	4.55	4.63	4.77	4.96
96 = Q	-5.42	-5.30	-4.98	-4.96	-4.85	-4.71	-4.03	-3.47	-2.47	-2.14	4.19	4.20	4.28	4.37	4.46	4.54	4.68	4.86
95 = P	-5.33	-5.21	-4.89	-4.87	-4.76	-4.62	-3.94	-3.37	-2.37	-2.04	4.30	4.31	4.39	4.48	4.57	4.65	4.79	4.97
96 = E	2.48	2.43	2.28	2.27	2.22	2.15	1.83	1.56	1.09	0.94	-2.04	-2.04	-2.08	-2.12	-2.17	-2.20	-2.27	-2.36
90 = V	2.82	2.75	2.58	2.57	2.51	2.43	2.06	1.75	1.20	1.02	-2.45	-2.45	-2.50	-2.54	-2.59	-2.64	-2.71	-2.81
96 = G	4.05	3.96	3.71	3.69	3.61	3.50	2.97	2.53	1.75	1.49	-3.45	-3.45	-3.52	-3.58	-3.66	-3.72	-3.83	-3.97
90 = A	4.86	4.75	4.46	4.44	4.34	4.21	3.60	3.08	2.17	1.87	-3.89	-3.90	-3.98	-4.05	-4.13	-4.21	-4.33	-4.50
76 = E	5.79	5.67	5.32	Mammal		5.04	4.31	3.71	2.64	2.29	-4.47	-4.47	-4.57	-4.66	-4.75	-4.84	-4.99	-5.18
73 = E	5.86	5.73	5.38	5.36	5.24	5.08	4.34	3.72	2.61	2.25	-4.72	-4.72	-4.82	-4.91	-5.01	-5.10	-5.25	-5.45
95 = K	5.91	5.78	5.41	5.40	5.27	5.11	4.35	3.71	2.58	2.21	-4.93	-4.93	-5.03	-5.12	-5.23	-5.31	-5.47	-5.68
92 = I	5.92	5.78	5.42	5.40	5.27	5.11	4.35	3.71	2.58	2.20	-4.96	-4.96	-5.06	-5.15	-5.26	-5.35	-5.50	-5.71
72 = M	6.34	6.20	5.81	5.79	5.66	5.49	4.67	3.99	2.78	2.38	-5.25	-5.25	-5.36	-5.46	-5.57	-5.66	-5.83	-6.05
88 = I	6.46	6.31	5.92	5.90	5.76	5.58	4.75	4.05	2.82	2.41	-5.40	-5.40	-5.51	-5.61	-5.73	-5.82	-5.99	-6.22

(**a**)

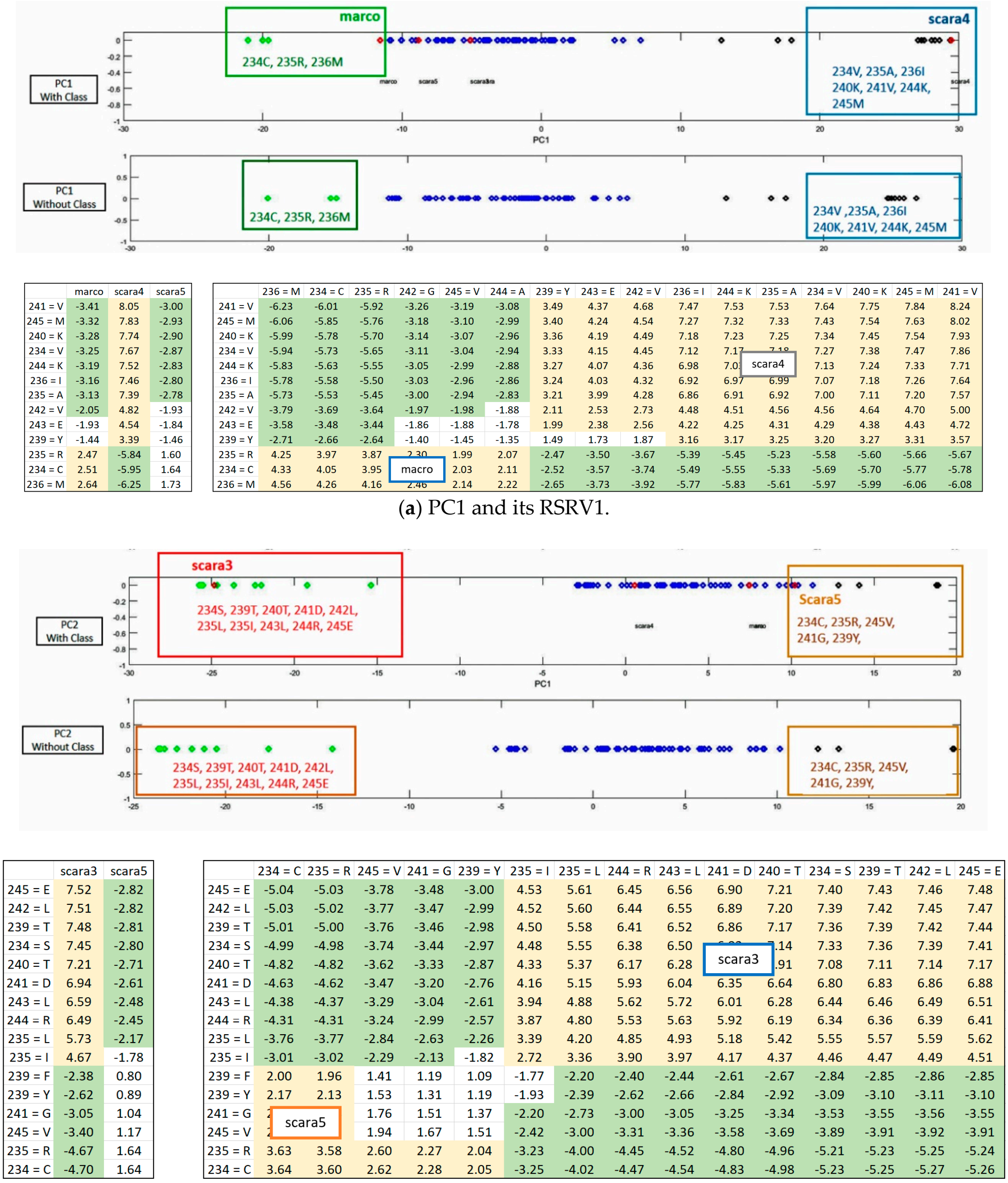


	Fungi	Plant
71 = L	-5.19	3.27
72 = Y	-4.04	2.54
90 = P	-3.94	2.47
88 = V	-3.90	2.45
76 = L	-3.85	2.42
96 = Q	-3.75	2.35
70 = T	-3.10	1.94
95 = P	-2.48	1.55
95 = A	2.34	-1.51
76 = E	2.83	-1.81
76 = T	3.13	-2.01
70 = N	3.19	-2.05
95 = E	4.57	-2.92
71 = M	5.00	-3.19
72 = S	5.68	-3.62
88 = A	6.56	-4.18
90 = G	6.60	-4.21
96 = K	6.63	-4.23

	71 = L	72 = Y	90 = P	88 = V	76 = L	96 = Q	70 = T	95 = P	95 = A	76 = E	76 = T	70 = N	95 = E	71 = M	72 = S	88 = A	90 = G	96 = K
90 = G	-5.46	-4.27	-4.16	-4.13	-4.07	-3.96	-3.28	-2.64	2.39	2.89	3.20	3.27	4.70	5.15	5.86	6.78	6.83	6.85
96 = K	-5.42	-4.23	-4.12	-4.08	-4.02	-3.92	-3.24	-2.60	2.44	2.94	3.26	3.33	4.77	5.21	5.92	6.84	6.89	6.92
88 = A	-5.37	-4.18	-4.08	-4.04	-3.98	-3.88	-3.21	-2.57	2.41	2.91	3.22	3.29	4.71	5.16	5.85	6.77	6.81	6.84
72 = S	-4.81	-3.78	-3.68	-3.66	-3.60	-3.52	-2.93	-2.38	1.95	2.39	2.66	2.71	3.95	4.34	4.94	5.74	5.78	5.80
71 = M	-4.20	-3.29	-3.21	-3.19	-3.14	-3.06	-2.55	-2.06	1.74	2.12	2.36	2.41	3.50	Plant		5.07	5.11	5.13
95 = E	-3.75	-2.93	-2.85	-2.83	-2.79	-2.72	-2.25	-1.81	1.66	2.01	2.22	2.27	3.26			4.69	4.72	4.74
76 = T	-2.62	-2.05	-2.00	-1.99	-1.96	-1.91	-1.59	-1.29	1.09	1.32	1.47	1.50	2.18	2.39	2.72	3.16	3.18	3.19
70 = N	-2.62	-2.04	-1.99	-1.97	-1.95	-1.90	-1.57	-1.26	1.15	1.39	1.54	1.57	2.26	2.48	2.82	3.26	3.28	3.29
76 = E	-2.23	-1.72	-1.68	-1.66	-1.64	-1.60	-1.31	-1.04	1.09	1.30	1.43	1.46	2.07	2.26	2.56	2.95	2.97	2.98
95 = A	-2.00	-1.58	-1.54	-1.53	-1.51	-1.47	-1.23	-1.00	0.78	0.95	1.06	1.09	1.60	1.75	2.00	2.33	2.35	2.35
95 = P	2.13	1.67	1.63	1.62	1.60	1.56	1.29	1.05	-0.89	-1.08	-1.20	-1.23	-1.78	-1.95	-2.22	-2.58	-2.59	-2.61
70 = T	2.62	2.06	2.00	1.99	1.96	1.91	1.59	1.28	-1.12	-1.36	-1.51	-1.54	-2.22	-2.44	-2.77	-3.21	-3.24	-3.25
96 = Q	3.11	2.42	2.36	2.34	2.31	2.25	1.86	1.49	-1.41	-1.70	-1.88	-1.92	-2.74	-3.00	-3.41	-3.94	-3.96	-3.98
76 = L	3.15	2.45	[illegible]	[illegible]	2.33	2.27	1.87	1.49	-1.47	-1.77	-1.96	-2.00	-2.84	-3.10	-3.52	-4.06	-4.09	-4.11
90 = P	3.23	2.51	Fungi		2.38	2.32	1.91	1.53	-1.51	-1.81	-2.00	-2.04	-2.91	-3.17	-3.60	-4.15	-4.18	-4.20
88 = V	3.23	2.52	2.45	2.43	2.40	2.33	1.93	1.54	-1.47	-1.77	-1.96	-2.00	-2.86	-3.13	-3.55	-4.10	-4.13	-4.15
72 = Y	3.33	2.59	2.53	2.50	2.47	2.40	1.98	1.59	-1.53	-1.84	-2.04	-2.08	-2.97	-3.24	-3.68	-4.25	-4.28	-4.30
71 = L	4.39	3.44	3.35	3.32	3.28	3.20	2.66	2.15	-1.86	-2.26	-2.51	-2.56	-3.70	-4.06	-4.62	-5.35	-5.39	-5.41

(**b**)

Figure 6. AR Clusters captured in PCs and their corresponding ARAs reflected in RSRVs for APC-6382. (**a**) RSRV1 ARA Clustering Result with and without Class Labels; (**b**) RSRV2 ARA Clustering Result with and without Class Labels. In both (**a**,**b**) cases, we observed that the AR clusters on the right-handed side RSRV plots obtained from SRV without class labels correspond closely to the class association respectively as indicated on the left-handed side plots taken from the RSRVs obtained from SRV when class labels are included. Hence, this shows that ARADD is able to obtain closely corresponding ARA patterns with or without class labels.



	marco	scara4	scara5
241 = V	-3.41	8.05	-3.00
245 = M	-3.32	7.83	-2.93
240 = K	-3.28	7.74	-2.90
234 = V	-3.25	7.67	-2.87
244 = K	-3.19	7.52	-2.83
236 = I	-3.16	7.46	-2.80
235 = A	-3.13	7.39	-2.78
242 = V	-2.05	4.82	-1.93
243 = E	-1.93	4.54	-1.84
239 = Y	-1.44	3.39	-1.46
235 = R	2.47	-5.84	1.60
234 = C	2.51	-5.95	1.64
236 = M	2.64	-6.25	1.73

	236 = M	234 = C	235 = R	242 = G	245 = V	244 = A	239 = Y	243 = E	242 = V	236 = I	244 = K	235 = A	234 = V	240 = K	245 = M	241 = V
241 = V	-6.23	-6.01	-5.92	-3.26	-3.19	-3.08	3.49	4.37	4.68	7.47	7.53	7.53	7.64	7.75	7.84	8.24
245 = M	-6.06	-5.85	-5.76	-3.18	-3.10	-2.99	3.40	4.24	4.54	7.27	7.32	7.33	7.43	7.54	7.63	8.02
240 = K	-5.99	-5.78	-5.70	-3.14	-3.07	-2.96	3.36	4.19	4.49	7.18	7.23	7.25	7.34	7.45	7.54	7.93
234 = V	-5.94	-5.73	-5.65	-3.11	-3.04	-2.94	3.33	4.15	4.45	7.12	7.17	7.18	7.27	7.38	7.47	7.86
244 = K	-5.83	-5.63	-5.55	-3.05	-2.99	-2.88	3.27	4.07	4.36	6.98	7.0[illegible]	scara4	7.13	7.24	7.33	7.71
236 = I	-5.78	-5.58	-5.50	-3.03	-2.96	-2.86	3.24	4.03	4.32	6.92	6.97	6.99	7.07	7.18	7.26	7.64
235 = A	-5.73	-5.53	-5.45	-3.00	-2.94	-2.83	3.21	3.99	4.28	6.86	6.91	6.92	7.00	7.11	7.20	7.57
242 = V	-3.79	-3.69	-3.64	-1.97	-1.98	-1.88	2.11	2.53	2.73	4.48	4.51	4.56	4.56	4.64	4.70	5.00
243 = E	-3.58	-3.48	-3.44	-1.86	-1.88	-1.78	1.99	2.38	2.56	4.22	4.25	4.31	4.29	4.38	4.43	4.72
239 = Y	-2.71	-2.66	-2.64	-1.40	-1.45	-1.35	1.49	1.73	1.87	3.16	3.17	3.25	3.20	3.27	3.31	3.57
235 = R	4.25	3.97	3.87	2.30	1.99	2.07	-2.47	-3.50	-3.67	-5.39	-5.45	-5.23	-5.58	-5.60	-5.66	-5.67
234 = C	4.33	4.05	3.95	macro	2.03	2.11	-2.52	-3.57	-3.74	-5.49	-5.55	-5.33	-5.69	-5.70	-5.77	-5.78
236 = M	4.56	4.26	4.16	2.46	2.14	2.22	-2.65	-3.73	-3.92	-5.77	-5.83	-5.61	-5.97	-5.99	-6.06	-6.08

(**a**) PC1 and its RSRV1.

	scara3	scara5
245 = E	7.52	-2.82
242 = L	7.51	-2.82
239 = T	7.48	-2.81
234 = S	7.45	-2.80
240 = T	7.21	-2.71
241 = D	6.94	-2.61
243 = L	6.59	-2.48
244 = R	6.49	-2.45
235 = L	5.73	-2.17
235 = I	4.67	-1.78
239 = F	-2.38	0.80
239 = Y	-2.62	0.89
241 = G	-3.05	1.04
245 = V	-3.40	1.17
235 = R	-4.67	1.64
234 = C	-4.70	1.64

	234 = C	235 = R	245 = V	241 = G	239 = Y	235 = I	235 = L	244 = R	243 = L	241 = D	240 = T	234 = S	239 = T	242 = L	245 = E
245 = E	-5.04	-5.03	-3.78	-3.48	-3.00	4.53	5.61	6.45	6.56	6.90	7.21	7.40	7.43	7.46	7.48
242 = L	-5.03	-5.02	-3.77	-3.47	-2.99	4.52	5.60	6.44	6.55	6.89	7.20	7.39	7.42	7.45	7.47
239 = T	-5.01	-5.00	-3.76	-3.46	-2.98	4.50	5.58	6.41	6.52	6.86	7.17	7.36	7.39	7.42	7.44
234 = S	-4.99	-4.98	-3.74	-3.44	-2.97	4.48	5.55	6.38	6.50	scara3	7.14	7.33	7.36	7.39	7.41
240 = T	-4.82	-4.82	-3.62	-3.33	-2.87	4.33	5.37	6.17	6.28	[illegible]	[illegible].91	7.08	7.11	7.14	7.17
241 = D	-4.63	-4.62	-3.47	-3.20	-2.76	4.16	5.15	5.93	6.04	6.35	6.64	6.80	6.83	6.86	6.88
243 = L	-4.38	-4.37	-3.29	-3.04	-2.61	3.94	4.88	5.62	5.72	6.01	6.28	6.44	6.46	6.49	6.51
244 = R	-4.31	-4.31	-3.24	-2.99	-2.57	3.87	4.80	5.53	5.63	5.92	6.19	6.34	6.36	6.39	6.41
235 = L	-3.76	-3.77	-2.84	-2.63	-2.26	3.39	4.20	4.85	4.93	5.18	5.42	5.55	5.57	5.59	5.62
235 = I	-3.01	-3.02	-2.29	-2.13	-1.82	2.72	3.36	3.90	3.97	4.17	4.37	4.46	4.47	4.49	4.51
239 = F	2.00	1.96	1.41	1.19	1.09	-1.77	-2.20	-2.40	-2.44	-2.61	-2.67	-2.84	-2.85	-2.86	-2.85
239 = Y	2.17	2.13	1.53	1.31	1.19	-1.93	-2.39	-2.62	-2.66	-2.84	-2.92	-3.09	-3.10	-3.11	-3.10
241 = G	2[illegible]	scara5	1.76	1.51	1.37	-2.20	-2.73	-3.00	-3.05	-3.25	-3.34	-3.53	-3.55	-3.56	-3.55
245 = V	2[illegible]	[illegible]	1.94	1.67	1.51	-2.42	-3.00	-3.31	-3.36	-3.58	-3.69	-3.89	-3.91	-3.92	-3.91
235 = R	3.63	3.58	2.60	2.27	2.04	-3.23	-4.00	-4.45	-4.52	-4.80	-4.96	-5.21	-5.23	-5.25	-5.24
234 = C	3.64	3.60	2.62	2.28	2.05	-3.25	-4.02	-4.47	-4.54	-4.83	-4.98	-5.23	-5.25	-5.27	-5.26

(**b**) PC2 and RSRV2.

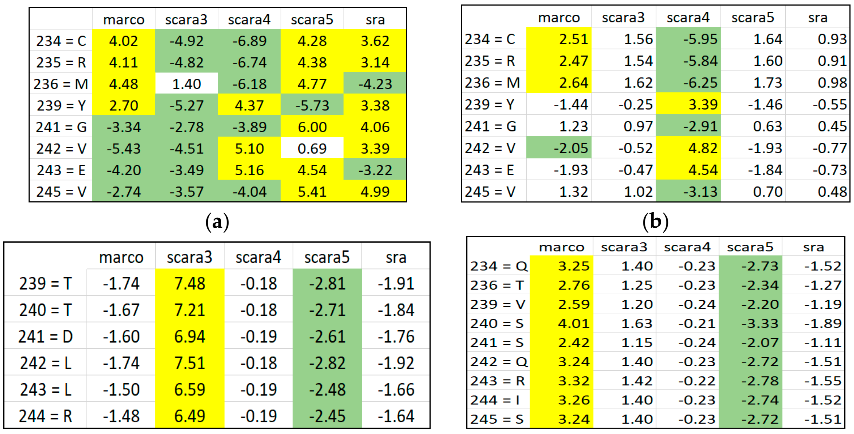
Figure 7. PC and RSRV results for AP2859 obtained from protein sequences of SR-A family. (**a**) PC1 and its RSRV1; (**b**) PC2 and RSRV2. Note the consistency between the AR clusters in the PCs and the a-vector groups in the RSRVs. Note also in both PC pairs, distinct AR groups associating with SR-A classes are succinctly discovered with/without the inclusion of the class labels in the APCs.

Output: RSRVs and ARA Sub-Clusters

1. *Disentangled ARAs by distributing and grouping them in different RSRVs.* After PCD, it may not be obvious why an a-vector is significant. However, when the a-vectors are examined in the RSRVs, it can be observed that their high SR coordinate(s) contributed to their high variance on their corresponding PCs. In general, PCD is sensitive to the relative scaling of the original aligned columns. By unifying the scaling of ARA measures through SR, our results showed that SRV is rather stable and can reveal statistically significant associations between ARs with statistical strength reasonably well. However, when ARAs become entangled, the SRV disentanglement

is crucial for yielding highly distinct, stable, and specific results as manifested in the RSRVs obtained from both datasets (Figures 4–8).

2. *ARs Sub-clusters.* Although the disentangled PCs can already reveal significant ARs/AR-Clusters on a one-dimensional space (Figures 4, 6 and 8), the statistical strength (SR) of the ARAs can further reveal the significance of the ARAs and the AR-Clusters by identifying their RSRVs through the SRs from the row and column a-vectors (RSRV1 and RSRV2 on bottom half of Figure 6a,b and Figure 7a,b. By disentanglement, we can obtain different ARs sub-clusters in different orthogonal PC spaces as shown in both figures. These AR subgroups may have functional meaning. They may help to build classification or subgroup partitioning models with less features but can achieve comparably superior performance to the classification models based on statistical significant ARAs derived from SRV. Furthermore, it is much easier to select discriminative features from RSRVs after disentanglement than from SRV and from the distinct ARs and AR clusters in the PCs.



	marco	scara3	scara4	scara5	sra
234 = C	4.02	-4.92	-6.89	4.28	3.62
235 = R	4.11	-4.82	-6.74	4.38	3.14
236 = M	4.48	1.40	-6.18	4.77	-4.23
239 = Y	2.70	-5.27	4.37	-5.73	3.38
241 = G	-3.34	-2.78	-3.89	6.00	4.06
242 = V	-5.43	-4.51	5.10	0.69	3.39
243 = E	-4.20	-3.49	5.16	4.54	-3.22
245 = V	-2.74	-3.57	-4.04	5.41	4.99

(**a**)

	marco	scara3	scara4	scara5	sra
234 = C	2.51	1.56	-5.95	1.64	0.93
235 = R	2.47	1.54	-5.84	1.60	0.91
236 = M	2.64	1.62	-6.25	1.73	0.98
239 = Y	-1.44	-0.25	3.39	-1.46	-0.55
241 = G	1.23	0.97	-2.91	0.63	0.45
242 = V	-2.05	-0.52	4.82	-1.93	-0.77
243 = E	-1.93	-0.47	4.54	-1.84	-0.73
245 = V	1.32	1.02	-3.13	0.70	0.48

(**b**)

	marco	scara3	scara4	scara5	sra
239 = T	-1.74	7.48	-0.18	-2.81	-1.91
240 = T	-1.67	7.21	-0.18	-2.71	-1.84
241 = D	-1.60	6.94	-0.19	-2.61	-1.76
242 = L	-1.74	7.51	-0.18	-2.82	-1.92
243 = L	-1.50	6.59	-0.19	-2.48	-1.66
244 = R	-1.48	6.49	-0.19	-2.45	-1.64

(**c**)

	marco	scara3	scara4	scara5	sra
234 = Q	3.25	1.40	-0.23	-2.73	-1.52
236 = T	2.76	1.25	-0.23	-2.34	-1.27
239 = V	2.59	1.20	-0.24	-2.20	-1.19
240 = S	4.01	1.63	-0.21	-3.33	-1.89
241 = S	2.42	1.15	-0.24	-2.07	-1.11
242 = Q	3.24	1.40	-0.23	-2.72	-1.51
243 = R	3.32	1.42	-0.22	-2.78	-1.55
244 = I	3.26	1.40	-0.23	-2.74	-1.52
245 = S	3.24	1.40	-0.23	-2.72	-1.51

(**d**)

Figure 8. Discovered ARAs based on SRV and RSRVs by ARADD for APC-2859. (**a**) SRV where ARAs associating with class entangled; (**b**) RSRV1 with ARAs associating with Marco and Scar4 disentangled; (**c**) RSRV2 with ARAs associating with Scar3; (**d**) RSRV3 with ARAs associating with Scara5.

4. Materials

Dataset 1-APC-6382 is a cluster of aligned patterns (width: 17) covering in total 85 protein sequences consisting of 4 subclasses (Fungi, Insect, Mammal and Plant). It was one of the APCs with highest coverage obtained by applying Aligned Pattern Clustering (APCn) algorithm [16,17] on a set of 93 protein sequences of Cytochrome c obtained from [16] after preprocessing on more than 300 protein sequences. For simplicity, we selected the protein sequences belonging to Fungi, Insect, Mammal, Mammal-Primate, Mammal-Rodent, Plant, Chlorophyta and Cryptophyta. We then regrouped Mammal, Mammal-Primate, Mammal-Rodent as the same subclass Mammal, and, Plant, Chlorophyta, Cryptophyta as the same subclass Plant. This dataset is mainly for illustrating our methodology since we found that while its taxonomical classes are succinct, their aligned patterns are still somehow entangled due to their entwining common and different evolutionary functions.

Dataset 2-APC-2859 is a cluster of aligned patterns (width: 12) obtained from APCn [16,17] covering in total 95 protein sequences coming from 5 subclasses (Marco, Sra, Scara3, Scara4, Scara5) of class A scavenger receptors originally taken from a dataset with 106 sequences used in [21]. Among all APCs, APC-2859 is the one with the highest coverage. All five subclasses of proteins contain domains: Cytoplasmic, Collagenous, Transmembrane, a-helical and coiled-coil motifs. Marco, Sra, and Scara5 contain the Collagenous domain. Only Sra contains the SRCR domain.

Here we first report our experimental results on Dataset 1-APC-6382 from Cytochrome c as an illustrative example of the proposed methodology. We then report our experimental results on Dataset 2-APC-2859 from class A scavenger receptors to reveal ARADD capability of discovering, disentangling and locating much more subtle subgroup characteristics entangled in the source environment.

5. Experimental Results

5.1. Experimental Results on Dataset 1-APC-6382 of Sequences from Cytochrome C Family

We used APC-6382 obtained from the Cytochrome c protein family with taxonomic class labels to illustrate how ARADD works. We first constructed SRVs between each pair of ARs in the APC dataset. The $SR(A_n^i \leftrightarrow A_n^i)$ with zero value corresponds to the co-occurrence of the same AR and therefore has no meaning. The value of SRs which is above or below ±1.96 represent respectively positive or negative significant ARAs and are shaded respectively in yellow and in green. Figure 5a shows the SRV result for the association of ARs with class labels if they are included in the data space of the APC. We found that in many cases, the entangled ARs cannot distinguish different classes. For example, as a case of ARA entanglement from different classes, "AR71 = L" is associated with both Mammal and Plant but is highlighted as the disentangled result given in the SRV in Figure 5.

We then applied PCD on the SRV obtained from APCs with class labels attached and obtained PCs ranked after their variance. We projected the a-vector projections on the PCs to RSRVs to obtain a new set of coordinates that are the SRs of each ARA between ARs corresponding to the row and column a-vectors. Figure 5 shows the disentanglement of the class labels. Figure 6a,b show the ARs made up of distinct clusters in the PCs and their corresponding ARAs from their a-vectors in the RSRVs. We conclude that the entangled ARAs in SRV are disentangled in the RSRVs.

From the above experimental result, we conclude that when an APC with class labels is given as input, we could detect different groups of ARs related with class labels on different PCs and their ARAs on their corresponding RSRVs. The RSRVs results could succinctly reveal ARAs sharing with or discriminating from different subgroups conditioned (or governed) by certain biological functions on that spot. In order to show that such functional associations are intrinsic even if the class labels are not included, we conduct separate experiments on APCs with and without class labels included.

On each PC pair (with class labels or without) on top of Figure 6a,b respectively, we observed that ARs that made up of the AR clusters corresponding to taxonomical classes remain essentially the same, though their positions have changed a little due to the change in statistics, resulting from the presence or absence of the class labels. The results in both Cytochrome c (Figure 6) and SR-A (Figure 7) support such claim. Note that on the plot with class label included, the red diamonds represent the class label when treated as an attribute with its values representing different classes. This shows that ARADD is robust in discovering and disentangling ARAs with or without the explicit reliance on class labels. Hence, it is effective for supervised classification and unsupervised subgroup analysis without explicit reliance on prior knowledge, a significant advantage in mining large volume of omitted data.

5.2. Experimental Results on Dataset 2-APC-2859 from Class A Scavenger Receptor Family Sequences

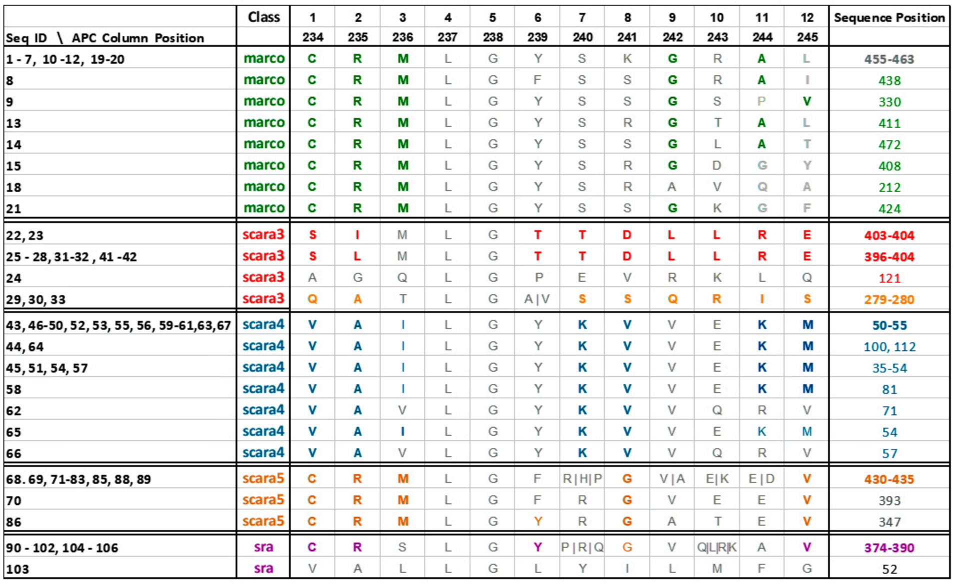
To meet the challenge of a well-recognized difficult proteomic problem, we applied ARADD on APC-2859 obtained from sequences of SR-A, the dataset 2 as described in Section 4. First, we compared the discovered ARAs obtained in RSRVs with those using only the adjusted statistical residual (SR) in SRV [1] with the same threshold SR > 1.96 all through (Figure 8).

Figure 8a shows the discovered ARAs only using SRV result for APC-2859. In this dataset, ARAs in different classes discovered are all included in the APC due to their similarity. However, we observed that their ARAs and class relationship are entangled, implying that the patterns in the APC are also entangled. When we checked the discovered ARAs obtained through using ARADD in different RSRVs, we observed the disentangled results as shown succinctly in Figure 8b–d.

In Figure 8b, if we select those entangled residue values in SRV from RSRV1, we found that in SRV, Marco is entangled with Scara5 and Sra, while in RSRV1, Marco is disentangled from Scara5 and Sra among residues in aligned sites 234, 235 and 236. In Figure 8c,d, Scara5 and Marco are disentangled and manifested as distinct groups from other classes.

Table 1 shows the AR subgroups discovered in the APC sequence pattern space plotted on the APC data space. The ARs with statistical significant ARAs with other ARs are in bold colored fonts. The first and the last column tabulate the sequence IDs and sequence positions range of the a-vectors respectively. Note that the AR pattern for Scara5 is CRM****G***V and that for Sra is CR***Y*G***V, which are similar but mapped onto two distant domains. Hence, ARADD not only can disentangle functional association in the pattern space but also disentangle their sequence location that is related to different domains of the family.

Table 1. Experimental results of AR groups associating with different SR-A classes. The first column tabulates the sequence IDs of the AR groups in the data space. The second column tabulates the SR-A Class with which each AR group is associating. The AR in bold on the third column indicates that it is the ARs in an a-vector with strong SR with other ARs. The last column tabulates the range of the sequence position on which each AR groups resides.



	Class	1	2	3	4	5	6	7	8	9	10	11	12	Sequence Position
Seq ID \ APC Column Position		234	235	236	237	238	239	240	241	242	243	244	245	
1 - 7, 10 -12, 19-20	marco	**C**	**R**	**M**	L	G	Y	S	K	**G**	R	**A**	L	**455-463**
8	marco	**C**	**R**	**M**	L	G	F	S	S	**G**	R	**A**	I	438
9	marco	**C**	**R**	**M**	L	G	Y	S	S	**G**	S	P	**V**	330
13	marco	**C**	**R**	**M**	L	G	Y	S	R	**G**	T	**A**	L	411
14	marco	**C**	**R**	**M**	L	G	Y	S	S	**G**	L	**A**	T	472
15	marco	**C**	**R**	**M**	L	G	Y	S	R	**G**	D	G	Y	408
18	marco	**C**	**R**	**M**	L	G	Y	S	R	A	V	Q	A	212
21	marco	**C**	**R**	**M**	L	G	Y	S	S	**G**	K	G	F	424
22, 23	scara3	**S**	**I**	M	L	G	**T**	**T**	**D**	**L**	**L**	**R**	**E**	**403-404**
25 - 28, 31-32 , 41 -42	scara3	**S**	**L**	M	L	G	**T**	**T**	**D**	**L**	**L**	**R**	**E**	**396-404**
24	scara3	A	G	Q	L	G	P	E	V	R	K	L	Q	121
29, 30, 33	scara3	**Q**	**A**	T	L	G	A | V	**S**	**S**	**Q**	**R**	**I**	**S**	**279-280**
43, 46-50, 52, 53, 55, 56, 59-61,63,67	scara4	**V**	**A**	I	L	G	Y	**K**	**V**	V	E	**K**	**M**	**50-55**
44, 64	scara4	**V**	**A**	I	L	G	Y	**K**	**V**	V	E	**K**	**M**	100, 112
45, 51, 54, 57	scara4	**V**	**A**	I	L	G	Y	**K**	**V**	V	E	**K**	**M**	35-54
58	scara4	**V**	**A**	I	L	G	Y	**K**	**V**	V	E	**K**	**M**	81
62	scara4	**V**	**A**	V	L	G	Y	**K**	**V**	V	Q	R	V	71
65	scara4	**V**	**A**	**I**	L	G	Y	**K**	**V**	V	E	K	M	54
66	scara4	**V**	**A**	V	L	G	Y	**K**	**V**	V	Q	R	V	57
68. 69, 71-83, 85, 88, 89	scara5	**C**	**R**	**M**	L	G	F	R | H | P	**G**	V | A	E | K	E | D	**V**	**430-435**
70	scara5	**C**	**R**	**M**	L	G	F	R	**G**	V	E	E	**V**	393
86	scara5	**C**	**R**	**M**	L	G	Y	R	**G**	A	T	E	**V**	347
90 - 102, 104 - 106	sra	**C**	**R**	S	L	G	**Y**	P | R | Q	G	V	Q|L|R|K	A	**V**	**374-390**
103	sra	V	A	L	L	G	L	Y	I	L	M	F	G	52

From the experimental results, we found that ARADD not only can discover the statistically significant ARAs, though entangled in the SRV, but much more novel and crucial is that it can discover significant ARs and/or AR Clusters (ARCs) captured in orthogonal PCs to bring out their separability as shown in Figure 8a,b respectively. We note that class Scara4 stands out in RSRV1; and Scara3 and Scara5 are two distinct subgroups, one as an opposite in RSRV2; and Scara3 and Scara4 are separated in RSRV3.

To show that the disentangled results remain intact with minor changes in AR position on APC when class labels are absent, we apply ARADD to the same set of data with class labels removed.

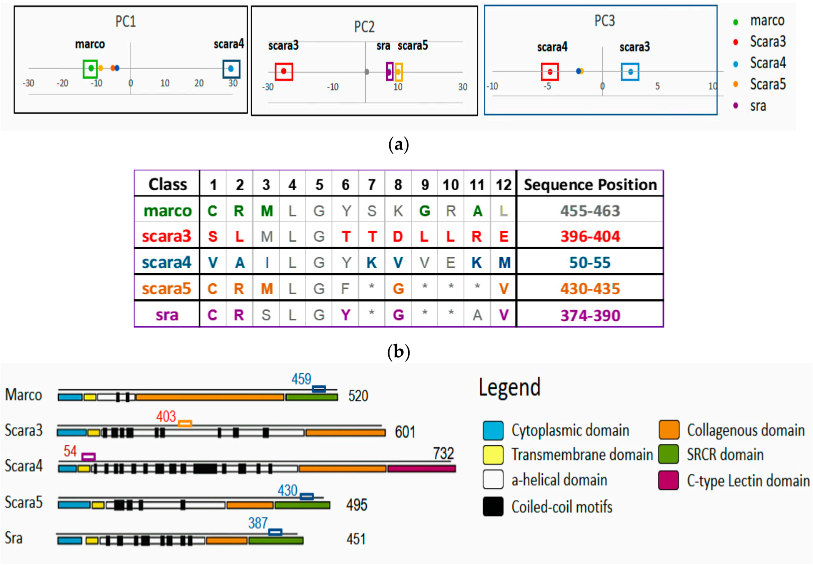
Figure 8 also shows from the PCs the closeness of the ARs found in each succinct cluster with and without class label in respective PC1 and PC2 (Figure 8a,b). We observed that that the AR cluster configurations have little change in their respective PC spaces. When class labels are included they appear in the PCs as a projection denoted by the red diamonds.

6. Discussion

The tabulated results in Table 1 give strong scientific support to the significance of ARA disentanglement in proteomic research, revealing the crucial information of “what” and “where” in a protein family. Figure 9 gives a succinct view of the discovered results both in pattern and data space.

Figure 9a shows that the class labels associating with ARs of that class are discovered within their associating clusters in the one-dimensional disentangled PC space. As revealed in Table 1, the AR groups for Scara5 and Sra are very close with only a single difference in their significant ARs. Their closeness is also revealed in RSRV2. Both deviate significantly from Scara3. Hence, from the PCs (Figure 9a) and the plots of the significant AR clusters (pattern space) we have at a glance of their similarity and differences with statistical backing. From the APC data space in Table 1, we observed that both the sequence ID and sequence position of each of the AR pattern were revealed. They are surprisingly closely correlated with the domain regions annotated the legends of a figure taken from [1]. Table 2 summarizes the sequence class and position information of the discovered patterns we obtained through ARADD and details how they correspond to domain regions obtained from biological experiment as reported in [1].

The results for SR-A are more profound than what we perceive on the surface. The experimental result we present in this paper provides significant evidence to support our conjecture that the discovered subtle deep knowledge entangled in the source environment and masked on the surface of the observed data can be discovered by ARADD without explicit reliance on prior knowledge. In a nut shell, ARADD [4] is able to reveal and locate the significant AR clusters, the "what" and "where" of subtle functional groups. The former is revealed through the disentangled PCs and RSRVs while the latter through the address table assembled during the significant ARA discovery process. Hence, we first introduce the notion on the ARA pattern space and the data space of an APC [16,17] before we dive into the experiments and the results. The former expresses which ARs, AR clusters, ARAs and ARA clusters are functionally and statistically significance through their association strength in the PCs and RSRVs. The latter displays where they are in the family sequences and relating them to the family domains found or validated through biology experiments. Hence, two functional groups with similar ARAs (like Scara5 and Sra, with the only difference in sites 236 and 239 discovered by ARADD represent different functional groups occurring in different regions (sequence position) of the family (shown in last column of Table 1 and Figure 9c).



(**a**)

Class	1	2	3	4	5	6	7	8	9	10	11	12	Sequence Position
marco	C	R	M	L	G	Y	S	K	G	R	A	L	455-463
scara3	S	L	M	L	G	T	T	D	L	L	R	E	396-404
scara4	V	A	I	L	G	Y	K	V	V	E	K	M	50-55
scara5	C	R	M	L	G	F	*	G	*	*	*	V	430-435
sra	C	R	S	L	G	Y	*	G	*	*	A	V	374-390

(**b**)

(**c**)

Figure 9. ARA disentanglement in pattern space and sequence location. (**a**) Disentanglement of functional groups corresponding to class displayed by their class labels; (**b**) Typical AR Groups in distinct bold colored ARs corresponding to different classes with sequence ID and the range of positions listed; (**c**) A mapping of the 5 patterns (AR groups) onto the protein sequences with domain regions annotated described in the legends [1] and class labels associating with the patterns. It shows the sequence mean position of the patterns (in small colored boxes and sequence position indices) and the sequence lengths in black digits at the end the sequences. Though patterns were clustered in the same APC due to sequence similarity, the ARA patterns reside in different domain corresponding to different biological function. Patterns found by ARADD fall into the range of each domain.

Table 2. A summary of the occurrences of the discovered patterns and their position in different class sequences and their correspondence with the biology domains reported in [1].

Covering Sequences Subclasses	Average Pattern Occurrence (or Address) Position	Domain
Marco	431.5	SRCR
Sra	367.8	SRCR
Scara5	423.7	SRCR
Scara4	58.0	Transmembrane/alpha-helical with coil-coiled motifs
Scara3	355.9	alpha-helical with coil-coiled motifs

Expanded Discussion of Protein Sequence Analysis: A Comparison with MEME

To investigate if the sequence patterns discovered and clustered by other models have been trapped in the entanglement of biological functions of similar ARAs due to unknown factors or carrying different functionality in different domain locations, we conducted an experiment referred to as Experiment 6A. We used a popular motif discovery algorithm, MEME [7], to obtain motifs on Dataset 2 as a case study. Later in Experiment 6B, we showed that the exhibition of AR clusters for different subclasses, with their positions located in the family sequence domains, can be obtained via disentanglement through ARADD.

In Experiment 6A, we first applied MEME [7] on Dataset 2, covering in total 95 protein sequences coming from five subclasses (Marco, Sra, Scara3, Scara4, Scara5) of class A scavenger receptors. The details for Dataset 2 could be referred to Section 4 on Materials. The parameter setting was as follows: minimum width = 10, maximum width = 15, number of motifs = 30. These parameters were set to find motifs with comparison to APC-2859.

After running MEME on Dataset 2, we obtained 30 motifs in the form of PWM's [6]. We found that the fifth-ranked motif was most similar to APC-2859. This motif covering all 95 sequences, with an E-value of 1.7 × 10^{-690}, is depicted in Figure 10. All sequence patterns that compose the motif (Figure 10) are summarized in the Supplementary Table S1.

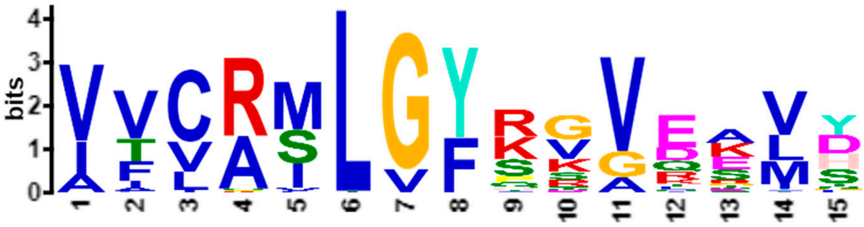


Figure 10. The fifth-ranked motif obtained by running MEME on Dataset 2. The motif covers all 95 sequences, with an E-value of 1.7 × 10^{-690}.

By averaging the starting position of the sequence patterns (Supplementary Table S1) that compose the motif depicted in Figure 10, we report that the domains where the average starting positions of these patterns reside correspond to the five subclasses (Marco, Sra, Scara3, Scara4, Scara5) of class A scavenger receptors domains as shown in Table 3.

Table 3. A summary of the occurrences of the discovered patterns by MEME [7] and their position in different class sequences and their correspondence with the biology domains reported in [1]. The problem here is that these sequence patterns are clustered in the same motif, but they actually occur in different biological domains.

Covering Sequences Subclasses	Average Pattern Occurrence (or Address) Position	Domain
Marco	425.1	SRCR
Sra	367.6	SRCR
Scara5	422.7	SRCR
Scara4	55.6	Transmembrane/alpha-helical with coil-coiled motifs
Scara3	69.6	alpha-helical with coil-coiled motifs

As shown in Table 3, we observed that the sequence patterns discovered and clustered by MEME [7] clearly show the entanglement phenomena of biological functions as similar ARAs are occurring in different domain location. In other words, although the sequence patterns are discovered and clustered in the same motif by MEME [7], they actually reside in different biological domains and thus may have a little different biological functions.

In Experiment 6B, we applied ARADD to the APCs of sequence patterns as summarized in Supplementary Table S1 and obtained the PC projection plots as depicted in Figures 11 and 12. We found that, in the PC projection plots, the domain information corresponding to the subclasses was reflected exactly as reported in [1], while neither the pattern occurrence position nor the domain information were inputted into the ARADD algorithm. We observed that the subclasses Scara4 and Scara3 are distinctly revealed in PC1 (Figure 11) and PC2 (Figure 12) respectively. This is consistent to the biological domain distinction shown in Table 3. In other words, ARADD algorithm has successfully disentangled the sequence patterns composing the motif (Figure 10) not relying on the knowledge of the pattern occurrence positions.

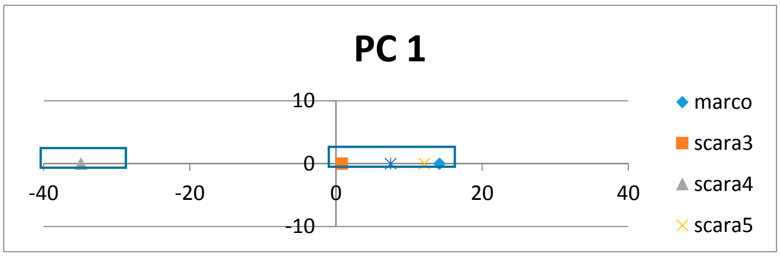


Figure 11. The 1st PC projection plot obtained by applying ARADD algorithm on the APCs of sequence patterns as summarized in Supplementary Table S1. It should be noted that the pattern occurrence position (indicated by the column "start" in Table S1) was not inputted to ARADD algorithm.

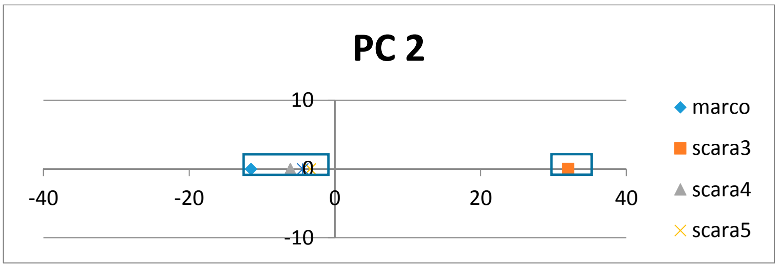


Figure 12. The 2nd PC projection plot obtained by applying ARADD algorithm on the APCs of sequence patterns as summarized Supplementary Table S1. It should be noted that the pattern occurrence position (indicated by the column "start" in Table S1) was not inputted to ARADD algorithm.

The above results pointed out that not only in APCs but also in motifs obtained by MEME [7], the sequence patterns that were discovered and clustered could be entangled in biological functions and domain location. These results show that in both scenarios [4,7], ARADD is able to disentangle the sequence patterns to reflect the domain information corresponding to different subclasses without knowing the actual pattern occurrence positions.

7. Conclusions

By applying the ARADD algorithm [4] to an entangled APC obtained from class A scavenger receptor, this study has shown that AR clusters (patterns), associating with different functional subgroups, regions and domains of the family obtained from an APC, could be succinctly plotted and statistically separated in different PCs and RSRVs as well as in different location through their sequence ID and sequence position of the family.

The most significant finding of this study is that the aligned patterns could be disentangled into ARA subgroups associating with different classes or subgroups, residing in different functional regions or domains of the family, even within an APC where conserved functional segments from different sequences are aligned. Such findings are absent in existing sequence alignment methods, but clearly revealed in our study on the class A scavenger receptor family.

Biologically, entangled patterns in class A scavenger receptor that are aligned in a sequence conservation model, such as in an APC, reveal biological functional patterns pertaining to similar or different classes. Entangled patterns can be disentangled into subgroups pertaining to different functionality, such as class A scavenger receptor classes into ARA subgroups mapped onto different PCs and RSRVs and located in different functional domains of the class A scavenger receptor family. Therefore, the successful application of ARADD algorithm to class A scavenger receptor demonstrates its capability to open a new way for analyzing conserved regions and their distribution, with potential to reveal new knowledge for gene therapy applications [2].

Supplementary Materials: The following are available online http://www.mdpi.com/2227-7382/6/1/10/s1, Supplementary document 1: The software executable used in this study and a set of test data; Supplementary Table S1: All sequence patterns that compose the motif (Figure 10).

Acknowledgments: This research is supported by NSERC Discovery Grant (2450 500 105 0503 116642 xxxx 0000 000) and Waterloo/China Graduate Scholarship.

Author Contributions: A.W. originated the fundamental concept and coordinated the development; P.Z., A.S. and A.W conceived and designed the experiments; E.L contributed the experimental data; P.Z. implemented the methodology. P.Z. and A.W. performed the experiments; E.L, A.S. and A.W. analyzed the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Whelan, F.J.; Meehan, C.J.; Golding, G.B.; McConkey, B.J.; Bowdish, D.M.E. The evolution of the class A scavenger receptors. *BMC Evol. Biol.* **2012**, *12*, 227. [CrossRef] [PubMed]
2. Zani, I.A.; Stephen, S.L.; Mughal, N.A.; Russell, D.; Homer-Vanniasinkam, S.; Wheatcroft, S.B.; Ponnambalam, S. Scavenger receptor structure and function in health and disease. *Cells* **2015**, *4*, 178–201. [CrossRef] [PubMed]
3. Plüddemann, A.; Mukhopadhyay, S.; Sankala, M.; Savino, S.; Pizza, M.; Rappuoli, R.; Tryggvason, K.; Gordon, S. SR-A, MARCO and TLRs differentially recognise selected surface proteins from neisseria meningitidis: An example of fine specificity in microbial ligand recognition by innate immune receptors. *J. Innate Immun.* **2009**, *1*, 153–163. [CrossRef] [PubMed]
4. Zhou, P.; Wong, A.K.C.; Sze-To, A. Discovery and Disentanglement of Protein Aligned Pattern Clusters to Reveal Subtle Functional Subgroups. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2017), Kansas City, MO, USA, 13–16 November 2017.
5. Wong, A.K.C.; Zhou, P.; Sze-To, A. Discovering Deep Knowledge from Relational Data by Attribute-Value Association. In Proceedings of the 13th International Conference on Data Mining (DMIN'17), Las Vegas, NV, USA, 17–20 July 2017; pp. 51–57.
6. Xia, X. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. *Scientifica (Cairo)* **2012**, *2012*, 917540. [CrossRef] [PubMed]
7. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*. [CrossRef] [PubMed]
8. Edgar, R.C.; Batzoglou, S. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **2006**, *16*, 368–373. [CrossRef] [PubMed]
9. Thompson, J.D.; Linard, B.; Lecompte, O.; Poch, O. A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS ONE* **2011**, *6*. [CrossRef] [PubMed]
10. D'haeseleer, P. How does DNA sequence motif discovery work? *Nat. Biotechnol.* **2006**, *24*, 959–961. [CrossRef] [PubMed]
11. Altschuh, D.; Lesk, A.M.; Bloomer, A.C.; Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **1987**, *193*, 693–707. [CrossRef]
12. Kass, I.; Horovitz, A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins Struct. Funct. Genet.* **2002**, *48*, 611–617. [CrossRef] [PubMed]
13. Chau, T.; Wong, A.K.C. Pattern discovery by residual analysis and recursive partitioning. *IEEE Trans. Knowl. Data Eng.* **1999**, *11*, 833–852. [CrossRef]
14. Wang, Y.; Wong, A.K.C. From association to classification: Inference using weight of evidence. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 764–767. [CrossRef]
15. Jiawei, H.; Kamber, M.; Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2012; ISBN 978-0-12-381479-1.
16. Lee, E.-S.; Wong, A.K. Ranking and compacting binding segments of protein families using aligned pattern clusters. *Proteome Sci.* **2013**, *11*, S8. [CrossRef] [PubMed]
17. Wong, A.K.C.; Lee, E.S.A. Aligning and clustering patterns to reveal the protein functionality of sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 548–560. [CrossRef] [PubMed]
18. Naulaerts, S.; Meysman, P.; Bittremieux, W.; Vu, T.N.; Vanden Berghe, W.; Goethals, B.; Laukens, K. A primer to frequent itemset mining for bioinformatics. *Brief. Bioinform.* **2015**, *16*, 216–231. [CrossRef] [PubMed]

19. Agrawal, R.; Imielinski, T.; Swami, A. Mining Association in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 25–28 May 1993; pp. 207–216. [CrossRef]
20. Han, J.; Pei, J.; Yin, Y.; Mao, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.* **2004**, *8*, 53–87. [CrossRef]
21. Lee, E.-S.A.; Whelan, F.J.; Bowdish, D.M.E.; Wong, A.K.C. Partitioning and correlating subgroup characteristics from Aligned Pattern Clusters. *Bioinformatics* **2016**, *32*, 2427–2434. [CrossRef] [PubMed]