10

Pattern Discovery and Disentanglement for Aligned Pattern Cluster Analysis and Protein Binding Complexes Detection

Peiyuan Zhou[1] • En-Shiun Annie Lee[2] • Andrew K. C. Wong[1]

[1]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada;
[2]School of Continuing Studies, York University, Toronto, ON, Canada

Author for correspondence: Peiyuan Zhou, System Design Engineering, University of Waterloo, Waterloo, ON, Canada. Email: choupeiyuan@gmail.com

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch10

Abstract: Pattern discovery detects statistically significant associations among attribute values known as patterns. Traditional pattern discovery algorithms usually produce overwhelming numbers of overlapping/redundant patterns, weakening their interpretation and decision. Pattern Discovery and Disentanglement (PDD) is a new method that can decompose the entangled associations into groups related to specific factors to overcome this problem. Hence, the patterns discovered are much less in number, yet comprehensive and succinct for machine learning tasks and "explainability". PDD has a potential for proteomic research, drug discovery, and personalized genetic medicine by revealing subtle genetic/clinical patterns. This chapter provides an overview of the methodology of PDD and its two applications: association discovery on aligned pattern clusters (APCs) and residue-to-residue interactions (R2R-I) prediction. Discovery of patterns from APCs of cytochrome *c* and class A scavenger receptors are presented as example. Distinct subgroup characteristics of their functional domains and discovery of R2R-I patterns to enhance prediction of residue interactions between binding proteins are discussed.

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

Keywords: aligned pattern clusters; class A scavenger receptors; pattern discovery and disentanglement; proteomics; residue-to-residue interactions

INTRODUCTION

Proteins and their interactions govern the biological processes of a living organism. Proteins in the same family have similar functions, which can be revealed by the aligned sites and patterns of the biosequences in homologous functional regions. Machine learning and frequent pattern mining are common data analysis approaches for protein analysis. However, in protein analysis, the input (residue) and output (protein function/binding) relations are not often obvious, particularly when the correlation of residues in the sequence is governed by multiple factors. Pattern Discovery and Disentanglement (PDD) algorithm may help overcome the problem. In this chapter, we present two applications of PDD in protein analysis: association discovery on aligned pattern clusters (APCs) and residue-to-residue interactions (R2R-I) prediction.

As for the first application, discovering conserved sequence patterns (or associations) from a protein family is crucial for revealing region functionality. To identify subgroup characteristics in functional domains are of fundamental importance. Hence, we identify APCs (1, 2) from biosequences of protein families to locate and reveal conserved functional regions with variable width (Figure 1A), for example, rare substitution and frameshift mutations. The significance of APCs in biosequence analysis is due to their dual space representation—the pattern space (Figure 1B) and the data space (Figure 1C). The former displays the statistical patterns of the residue association. The latter allows the tracking of the discovered patterns in the sequence data, through the pattern addresses obtained in the pattern discovery phase. With this location-preserving information, it is easy to see how the aligned residues and their associations are entangled (Figure 1D) among different taxonomic classes (C_1, C_2, and C_3 in Figure 1C) within the conserved domain (Figure 1B). Pattern discovery (PD) (3), such as frequent pattern mining (4, 5), is the typical approach which provides succinct statistical support in predictive analytics (6). It has been used to discover patterns (for example, motifs) in biosequence data to reveal associations for interpretation and classification (7). Through the discovered patterns, knowledge can be revealed from data (8). However, traditional PD approaches usually produce an overwhelming number of overlapping/redundant patterns (9). These patterns/associations are hard to be clustered or summarized (9-11) to reveal precise "knowledge" inherent in the functioning environment that produced the associations, making interpretation difficult. Furthermore, due to the difficulties encountered in handling large volume of patterns effectively, along with too many redundant patterns, the accuracy of these methods also suffers. Hence the interpretability of these methods on relational data is still a challenge (12).

As for the second application, predicting protein-protein interaction (PPI) and R2R-I are also important in proteomics. The current view is that the closer two residues are to each other, the stronger the physicochemical binding/interaction between them. However, this view is being challenged, as there might be other physicochemical factors that bring them close to each other (8). Three types of

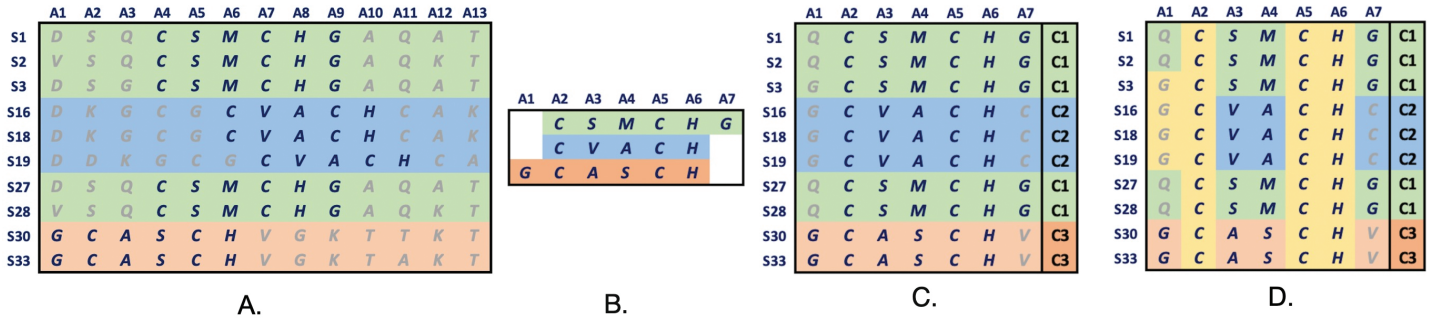


Figure 1. Pattern and Data Spaces of aligned pattern cluster (APC) and aligned residue associations (ARAs). A. A portion of protein sequence dataset with discovered high order patterns (in bold) with labels on the top row denoting the aligned sites, and those on the first column denoting the sequence ID. **B.** APC pattern space showing all the statistically significant patterns discovered in the domain. **C.** APC data space (C1, C2, C3 represents three classes) containing all the data with the embedded patterns in bold font. **D.** entangled ARAs (in S3 and S16, the ARAs A1G from C2 is entangled the A1G from C3. The entangled ARAs are highlighted in yellow).

computational R2R-I prediction methods have been developed. The first is computational docking (13), which simulates the interacting process based on physicochemical properties (14) of protein sequences such as shape complementarity, electrostatics, and biochemical information. This type of approach requires unbound structures of the target proteins that take special efforts to obtain. The second type (15, 16) is based on co-evolution conjecture which creates a Multiple Sequence Alignment (MSA) separately for both proteins and predicts statistically associated columns in spatial proximity. The prediction performance of this type requires homologous sequences of the given protein sequences to conduct MSA. It is easy to see that these applications are still limited since they often require additional data beyond the sequences given. Very few current methods can discover R2R-I sites using only information obtained directly from the sequence data. The third type uses machine learning. The methods first take a dataset of PPI complexes as input for both interacting pairs and non-interacting pairs, then derive a variety of features from the protein structures or MSA, and finally predict R2R-I using the same feature vectors for two input proteins. The structure-based methods require structures from the two input proteins (17, 18), while the sequence-based methods only require sequences from the two input proteins (18). The key drawback of the latter is that a large amount of time is needed to extract appropriate features and select optimum combinations of them.

To overcome these problems, we apply our recently developed PDD method. Realizing that the fundamental notions of these problems is related to the subtle associations of the items/events of the subject matter, we developed PDD to reveal more specific associations/functionality hidden in the acquired data.

PATTERN DISCOVERY AND DISENTANGLEMENT

PDD is a computational algorithmic process to discover a succinct set of statistically significant patterns from a relational dataset. Figure 2 provides an overview of PDD. It first obtains data and constructs a frequency matrix (FM). For the analysis of protein functional domain, the FM is constructed from the co-occurrence between pairs of residues in the APCs. For R2R-I prediction, the FM is constructed from the frequency counts of contact between residues obtained from the R2R-C data. The FM is then converted to a statistical residual vector space (SRV), accounting for the statistical significance. Next, the SRV is decomposed into principal components (PCs) and reprojected onto a new SRV, referred to as the reprojected SRV (RSRV), which reveals the association captured in the PC. We call a PC and its corresponding RSRV a disentangled space (DS).

The input data can be of various types, such as a relational table (for example, APCs) or sequence data (for example, a collection of sequences for predictor training and testing or two protein sequences for R2R-I prediction). An APC can be represented as a relational table described by an $N \times M$ matrix for N amino acid sites and M protein sequences. For R2R-I prediction, the R2R-C in PPI 3D configurations are acquired from protein data bank (PDB) (19) (Table 1). The notations, definitions and terminologies are tabulated in Table 1.

Since we attempt to apply PDD to APC and R2R-I at the same time, we use the superscript ^ and * to reference them, respectively, in the description below.

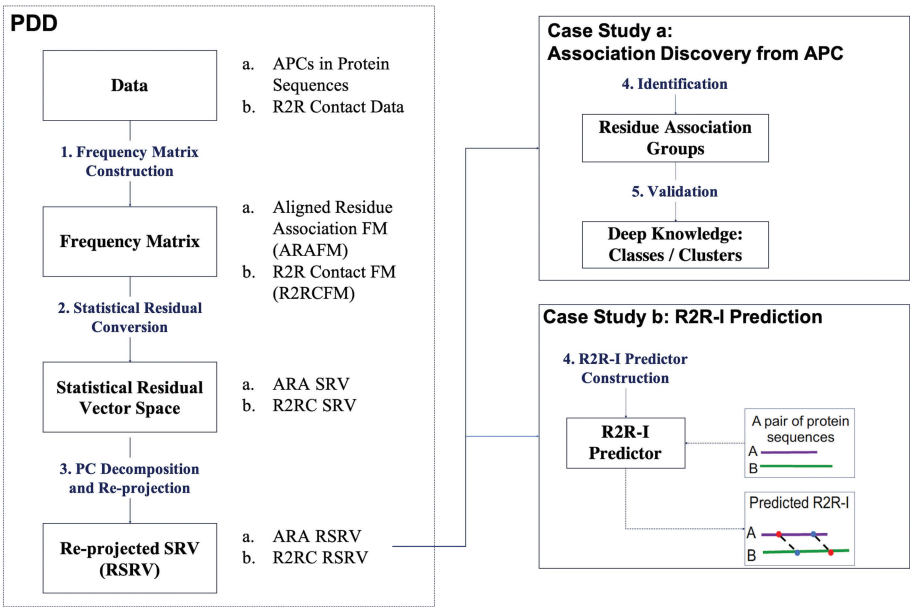


Figure 2. Overview of the methodology of PDD. This overview describes the methodology walkthrough for two applications **(a)** discovering the association patterns in a protein aligned pattern cluster and **(b)** building a predictor to predict the interacting residue and the sites between two interacting proteins.

TABLE 1 Terminologies used in this chapter and their definitions

Terminology	Definition
R2R pair	A pair of residues residing on two different protein sequences in PPI complexes were obtained from PDB.
Residue-Residue Contact (R2R-C)	An event where two residues are considered to be in close contact in the 3D coordinate space when the closet Euclidean distance between their C-Beta atoms is less than 6Å (17) (18). A R2R-C pair is referred to as a pair of residues (r_i, r_j) in close contact in the 3D coordinate space.
Aligned Residue Association (ARA)	An APC is denoted as A = {A_1, A_2,… A_N} (N is the number of amino acid sites), and for each site of amino acid, the values are denoted as $A_n = \{A_n^j \mid j = 1,2,\ldots I_n$ (I_n represent the total number of values on the n*th* site). Then the ARA refers to a pair of aligned residues, denoted as ($A_n^i \leftrightarrow A_{n'}^j$)

APC, aligned pattern cluster; PDB, protein data bank; PPI, protein-protein interaction; R2R-C, residue-residue close-contact

In step 1, the frequency counts of aligned residue associations (ARA)^ and R2R-C* are obtained to construct an FM. The item of ARAFM^ is denoted as $FM(A_n^i \leftrightarrow A_{n'}^j)$. Similarly, the item of R2RCFM* is denoted as $FM(r_i, r_j)$. The dimension of FM is T × T, where, in the APC case, T represents the sum of the number aligned residue pairs for all amino acid sites in the APC, while on the R2R-I case, it represents the number of different residues in R2R-C*.

Then, in step 2, the FM is transformed into an SRV to summarize the significance of respective associations. Those ARAs^ or R2R-C* with *SR* ≥1.96 or ≤ −1.96 corresponding to the confidence level = 95%, are considered as positive and negative significant associations, respectively. In step 3, PC decomposition (20) is applied to the SRV to create RSRV. RSRVs can disentangle and filter the entangled associations to reveal the distinct functional associations often masked in the SRV. When setting the same threshold, 1.96 corresponding to the confidence level as 95%, strong residue associations or interactions are revealed in the RSRVs.

In each RSRV, groups of associations are generated automatically if they share strong residue associations (21). For R2R-I, both positive and negative R2R-Is of a candidate pair and its six neighboring pairs are used to construct a *126-dimension* feature vector (8) for training the predictor as well as for prediction.

CASE STUDIES

PDD can discover more succinct patterns from the conserved region of proteins (for example, APC) and enable more efficient interpretability that is absent in existing sequence alignment methods in revealing the domain functionality of proteins. In addition, in the case of R2R-I prediction, the feature vectors construction using disentangled R2R-I patterns can dramatically improve the binding residue and sites prediction.

Application I: Association discovery on class A scavenger receptors APC

The dataset APC (1, 22) used in the first application was class A scavenger receptors (SR-A) (23) with 95 sequences (24). They are crucial for binding modified lipoproteins. Then an APC containing 12 aligned residue patterns from the following five subclasses were obtained: *Marco* (macrophage receptor with collagenous domain), *Sra* (scavenger receptor class A), *Scara3* (scavenger receptor class A, member 3), *Scara4, Scara5*. All five subclasses have the following domains: cytoplasmic, collagenous, transmembrane, α-helical and coiled-coil motifs. The classes *Marco*, *Sra*, and *Scara5* contain the collagenous domain. Only *Sra* contains the Scavenger Receptor Cysteine-Rich (SRCR) domain. So, the dimension of the APC dataset is 13(12 residues and 1 class label) × 95 (protein sequences).

PDD, unlike traditional association mining methods, is able to discover distinct patterns or associations related to different classes and domains from the original data. Figure 3 shows the statistical residual (SR) of the ARAs obtained in SRV (Figure 3A) and in the three RSRVs (Figures 3B, C) respectively. The yellow and green shaded cells in the figures denote the associations with positively and negatively significant SR, according to the 1.96 confidence level. In Figure 3A, we observed that before the disentanglement, the ARAs in the SRV correlated with classes were entangled. For example, both residues C and R in the position of 234 and 235 (234 = C and 235 = R) in Figure 3A are associated with the three classes *Marco*, *Scara5* and *Sra* when only SRs are used to measure ARAs. So, in the SRV result, ARA associated with *Marco* is entangled with those discovered in *Scara5*

	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

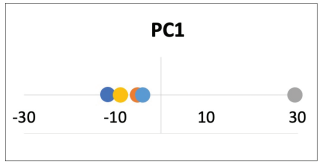
B.

	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.



E.

Seq ID \ APC Column Position	Class	1 234	2 235	3 236	4 237	5 238	6 239	7 240	8 241	9 242	10 243	11 244	12 245	Sequence Position
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

F.

Figure 3. Discovered ARAs and ARA groups. A. ARAs discovered in SRV show that their associations with class (in yellow shade) are entangled (shared with several classes). **B.** RSRV1 shows that ARAs associated with *Marco* and Scar4 are disentangled (no more sharing the same AR). **C.** RSRV3 shows the subset of ARAs positively associated with *Scara3* and negatively with *Scara5*. **D.** RSRV4 shows the ARAs positively associated with *Marco* and negatively with *Scara5*. **E.** disentanglement of functional groups in the three PCs reveals ARs associated with different classes as displayed by dots signifying their association with different class labels, note that in certain DS, their associations are close and in other are different (on the opposite side of the PC), quite different from traditional clustering results not at the level of patterns or subset of AVAs. **F.** experimental results of alighted residue groups (referred as AR groups) associating with different Scavenger Receptor-A 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 highlights the ARs in an AR-vector with strong SR with other ARs. The last column tabulates the range of the sequence position (of the sequences listed in the first column) on which each AR group resides.

and *Sra* (Figure 3A). While, after disentangling, in $RSRV_1$ (Figure 3B), the pattern of *Marco* is disentangled with that of *Scara5* and *Sra* among the amino acids residing on the aligned sites 234, 235 and 236. In Figures 3C and D, the patterns in *Scara3* and *Marco* respectively are disentangled from each other, and the entangled patterns are manifested as distinct groups from other classes.

Figure 3E gives a succinct view of how the statistical association strength of residues of different classes are disentangled and displayed in the PC. The color dots represent different classes. Their spatial closeness in the PC signifies the characteristic closeness of the classes. Hence, in different PCs, different classes and their closeness in correlation are revealed. To be more specific, the difference and similarity of the disentangled patterns pertaining to different classes are exemplified in Figure 3F. For example, the alighted residue groups (referred as AR groups) for *Scara5* and *Sra* are very similar to each other with only a single difference in their significant ARs (236 = M in *Scara5* and 236 = S in *Sra*) (Figure 3F). Their similar ARAs are also revealed in $RSRV_2$ among the negative class-AR associations (Figure 3C). This can also be seen by the closeness of the orange (*Scara5*) and the blue (*Sra*) dots in PC_2 and both deviate significantly from Scara3 (Figure 3E).

Compared to the traditional pattern discovery approaches, the advantage of PDD is that the tabulated results of PDD (Figure 3F) provide important scientific support to the significance of ARA disentanglement in proteomic research. The result reveals the crucial information of the "what" and "where" of the ARs in the primary structure of a protein family. In addition, it reveals the AR groups discovered in the APC pattern space and explicitly displays them in the APC data space. The residues with statistically significant ARAs with other ARs are plotted in bold colored fonts in Figure 3F. The first and the last columns tabulate the sequence IDs and the range of the positions in the protein sequences, respectively. We also note that the AR pattern for *Scara5* is CRM****G***V which is similar to CR***Y*G***V in Sara. However, they are mapped on two distant domains. It is difficult to distinguish and locate them using traditional approaches based only on statistical measurements, since these class patterns are entangled and scattered in the sequences of the scavenger receptor family.

Application II: Binding site (R2R-I) prediction

When two proteins A and B are given, binding site prediction aims to find out which residues and sites in protein A interact with which residues and sites in protein B, assuming proteins A and B can interact (17, 18). In this case, the proposed PDD is applied to the protein-protein docking benchmark dataset version 4.0 (abbreviated as DBD 4.0 (25)) containing 176 non-redundant PPI complexes. To evaluate the performance of PDD, we compared the PDD prediction results with those of PPiPP (Protein-Protein Interacting Pair Prediction) (17), an existing sequence-based R2R-I prediction software using feature engineering over external knowledge of protein sequences. The dataset is divided into a training dataset consisting of 124 non-redundant PPI complexes equivalent to protein-protein docking benchmark dataset version 3.0 (DBD3.0) (26), and a testing dataset consisting of the remaining 52 non-redundant PPI complexes in DBD 4.0 (25).

First, as mentioned in the methodology section, to extract R2R-C information from the dataset, we marked all R2R-C pairs with contact distance <6Å as positive (+ve) R2R-I, and those ≥6Å as negative (–ve) R2R-I. Figure 4A gives an example of a R2R pair (A-V) with three neighbors on both sides. After constructing statistical matrices and the disentanglement process, a 126-dimensions feature vector (FV) for each R2R pair is constructed as shown in Figure 4B as directed by the knowledge discovered in the DSs. The first 84 features are the statistical interaction, which would be the top 6 PC projections of the 14 residue alphabets (from P to I), reflecting their R2R-I statistical strength on the PC (that is their distance from the center of the PC). The following 42 (7R2R pairs × 6 top RSRVs) features are the average SR of each of the 7 R2R-I pairs obtained from the top 6 RSRVs (8). All these features account for the statistical strength of the R2R-I of the center pair and their neighbors. Then, the FVs for all training residue pairs are put into an extra decision tree classifier (Random Forest) with 1,000 trees, and the testing

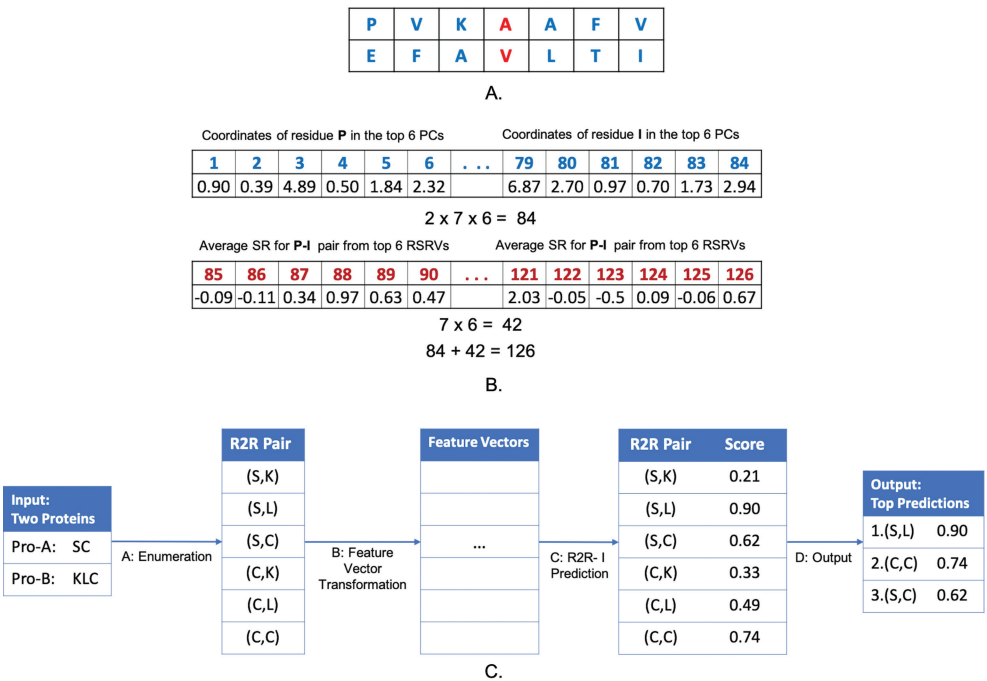


Figure 4. An overview of a feature vector (FV) creation for R2R-I predictor construction and prediction. A FV is created for an R2R-I site with a R2R-I pairs between two protein sequences. **A.** An example to show how the FV is created for an R2R pair (A-V) in red with its six neighboring pairs. **B.** A 126-dimensional FV created for the A-V pairs given in **A** based on the statistical strength of the central pairs and that of 6 neighboring pairs obtained from the training/test sequence data. The first 84 (2 x 7 x 6) features are the absolute values of the coordinates of all the 14 residues (from **P** to **I**) obtained from the 6 top PCs, reflecting their R2R-I statistical strength (their distance from the center of the principal component (PC)). The following 42 (7 x 6) features are the average SR of each of the 7 R2R-I pairs obtained from the top 6 RSRVs. Hence, there are 126 of them in total. **C.** an illustrative example of the R2R-I prediction process. This should be the first example in machine learning to use the discovered deep knowledge to train the classifier with data rather than using the statistics obtained directly from the data/ground-truth which could be masked/entangled.

TABLE 2 **Comparison of Average AUC for PDD, Random Predictor and PPiPP on 52 PPI complexes newly introduced in DBD4.0**

Predictors	Average AUC
Random Predictor	0.500 ± 0.000
PPiPP (17)	0.501 ± 0.003
PDD	0.643 ± 0.042

AUC, area under curve; PDD, pattern discovery and disentanglement (the proposed algorithm).

data are also transformed into FVs but without class labels. An example is given in Figures 4A and B. The constructed predictor is used for predicting the R2R-I without knowing whether they are binding or not. Figure 4C shows the prediction process.

The AUC (area under curve) is used for evaluating the binary classification, which refers to the area under a receiver operating characteristic (ROC) curve (27). The higher the AUC value, the better the prediction performance. As shown in the comparison result (Table 2), prediction using a PDD approach achieved a higher average AUC (0.643 ± 0.042) than that of a random predictor (0.50000 ± 0.00000), and, also, PPiPP with AUC (0.50112 ± 0.00257) (17). The improvement of the predictor is attributed to the use of the rectified ground truth by replacing R2RCFM with the top 6 DSs (PCs and RSRVs) and the proper use of the discriminative information in the DSs in the construction of the FVs for training and prediction.

CONCLUSION

By applying PDD to datasets acquired from complex source environments affected by entangled underlying factors, this study has shown that PDD can discover succinct patterns from the disentangled spaces. Both applications (patterns discovery in the APC of class A scavenger receptors and R2R-I prediction), show that PDD is able to discover deeper knowledge of association patterns masked at the data level due to the subtle entangled factors in the source environment. In the case of APC data analysis, PDD can discover residue associations correlated to different functional subgroups, regions, and domains in the class A scavenger receptor family, and succinctly locate and plot them in different statistically disentangled spaces. The explicit displayable result shows the efficacy of the pattern revealing and interpretability ability of PDD that is absent in existing sequence alignment methods. In the case of R2R-I prediction, PDD used more succinct and precise statistical measures to analyze R2R-I data. It disentangled the statistics from R2R-C data acquired from measurement such as residue closeness entangled in the three-dimensional physicochemical interaction environment, unveiled and extracted more specific deep knowledge on R2R-I between binding proteins, and used it to extract unbiased features to construct feature vectors for

training and classification, rather than conducting time-consuming feature engineering as in current machine learning. In AUC evaluation, the AUC of PDD achieved a higher average AUC (0.643 ± 0.042) than its contemporary PPiPP (0.501 ± 0.003) (17). This is 22% better with statistical significance (two-tailed paired student's *t*-test *p*-value: 1.9E-08 < 0.05). The result strongly indicates that the deep knowledge discovered from R2R-C data is effective for R2R-I prediction if disentangled by PDD. In summary, PDD not only discovers hidden deep knowledge, but also explains that knowledge and using it to predict unknown data using to achieve higher accuracy. Hence, PDD represents a pioneering work in deep knowledge discovery and explainable AI.

Acknowledgment: Publication costs were funded by NSERC Discovery Grant (xxxxx 50503-10275 500).

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1. Wong AKC, Lee AES. Aligning and clustering patterns to reveal the protein functionality of sequences. IEEE/ACM Trans Comput Biol Bioinform. 2014;11(3):548–560. https://doi.org/10.1109/TCBB.2014.2306840
2. Szeto AHY, Wong AKC. Discovering Patterns from Sequences Using Pattern-Directed Aligned Pattern Clustering. IEEE Trans Nanobioscience. 2018;14(8). https://doi.org/10.1109/TNB.2018.2845741
3. Wong AKC, Wang Y. High-Order Pattern Discovery from Discrete-Valued Data. IEEE Trans Knowl Syst. 1997;9(6):877-893. https://doi.org/10.1109/69.649314
4. Naulaerts S, Bittremieux W, Vu TN, Vanden Berghe W, Goethals B, Laukens K. A Primer to frequent itemset mining for bioinformatics. Brief Bioinform. 2015;16(2):216–231. https://doi.org/10.1093/bib/bbt074
5. Aggarwal CC, Han J. Frequent pattern mining: Springer; 2014. https://doi.org/10.1007/978-3-319-07821-2
6. Wang Y, Wong AKC. From association to classification: Inference using weight of evidence. IEEE Trans Knowl Data Eng. 2003;15(3):764–767. https://doi.org/10.1109/TKDE.2003.1198405
7. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. USA, Elsevier; 2011. 560p.
8. Wong AKC, Sze-To AH, Johanning GL. Pattern to Knowledge: Deep Knowledge-Directed Machine Learning for Residue-Residue Interaction Prediction. Sci Rep. 2018; 8:14841. https://doi.org/10.1038/s41598-018-32834-z
9. Wong AKC, Li GC. Simultaneous pattern and data clustering for pattern cluster analysis. IEEE Trans Knowl Data Eng. 2008; 20(7):911–923. https://doi.org/10.1109/TKDE.2008.38
10. Zhou P, Li GC, Wong AKC. An Effective Pattern Pruning and Summarization Method Retaining High Quality Patterns With High Area Coverage in Relational Datasets. IEEE Access. 2016;4:7847-7858. https://doi.org/10.1109/ACCESS.2016.2624418
11. Cheng J, Ke Y, Ng W. \delta-Tolerance Closed Frequent Itemsets. In Sixth International Conference on Data Mining (ICDM'06). IEEE Xplore. 2006;139–148. https://doi.org/10.1109/ICDM.2006.1

12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Med. 2019; 25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
13. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics. 2014; 30(12): 1771–1773. https://doi.org/10.1093/bioinformatics/btu097
14. Ezkurdia l, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. Progress and challenges in predicting protein--protein interaction sites. Brief Bioinform. 2009; 10(3):233–246. https://doi.org/10.1093/bib/bbp021
15. Hamer R, Luo Q, Armitage JP, Reinert G, Deane CM. i-Patch: Interprotein contact prediction using local network information. Proteins: Structure, Function, and Bioinformatics. 2010; 78(13): 2781–2797. https://doi.org/10.1002/prot.22792
16. Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014; 3:e03430. https://doi.org/10.7554/eLife.03430
17. Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS One. 2011; 6(12):e29104. https://doi.org/10.1371/journal.pone.0029104
18. Afsar M, Geiss FuA, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. Proteins: Structure, Function, and Bioinformatics. 2014; 82(7): 1142–1155. https://doi.org/10.1002/prot.24479
19. Berman Wea. The Protein Data Bank (PDB). Nucleic Acids Res. 2000;(28): p. 235–242. https://doi.org/10.1093/nar/28.1.235
20. A tutorial on principal component analysis. https://arxiv.org/abs/1404.1100v1 [access date: 3 Apr 2014].
21. Zhou P, Sze-To A, Wong AKC. Discovery and disentanglement of aligned residue associations from aligned pattern clusters to reveal subgroup characteristics. BMC Med Genomics. 2018;11(5):103. https://doi.org/10.1186/s12920-018-0417-z
22. Lee AES, Wong AKC. Ranking and compacting binding segments of protein families using aligned pattern clusters. Proteome Sci. 2013;11:S8. https://doi.org/10.1186/1477-5956-11-S1-S8
23. Whelan FJ, Meehan CJ, Golding GB, McConkey BJ, Bowdish DM. The evolution of the class A scavenger receptors. BMC Evol Biol. 2012;12(1):1–11. https://doi.org/10.1186/1471-2148-12-227
24. Lee ESA, Whelan FJ, Bowdish DM, Wong AK. Partitioning and correlating subgroup characteristics from Aligned Pattern Clusters. Bioinformatics. 2016; 32(16):2427–2434. https://doi.org/10.1093/bioinformatics/btw211
25. Hwang H, Vreven T, Janin, J, Weng Z. Protein--protein docking benchmark version 4.0. Proteins: Structure, Function, and Bioinformatics. 2010; 78(15):3111–3114. https://doi.org/10.1002/prot.22830
26. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein--protein docking benchmark version 3.0. Proteins: Structure, Function, and Bioinformatics. 2008; 73(3):705–709. https://doi.org/10.1002/prot.22106
27. Powers DMW. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. J Mach Learn Technol. 2011;2(1):37–63.