Bioinformatics, 32(16), 2016, 2427–2434
doi: 10.1093/bioinformatics/btw211
Advance Access Publication Date: 22 April 2016
Original Paper

Sequence analysis

Partitioning and correlating subgroup characteristics from Aligned Pattern Clusters

En-Shiun Annie Lee[1], Fiona J. Whelan[2], Dawn M. E. Bowdish[3] and Andrew K. C. Wong[1,*]

[1]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, [2]Department of Biochemistry and Biomedical Sciences and [3]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 26, 2015; revised on March 26, 2016; accepted on April 13, 2016

Abstract

Motivation: Evolutionarily conserved amino acids within proteins characterize functional or structural regions. Conversely, less conserved amino acids within these regions are generally areas of evolutionary divergence. A priori knowledge of biological function and species can help interpret the amino acid differences between sequences. However, this information is often erroneous or unavailable, hampering discovery with supervised algorithms. Also, most of the current unsupervised methods depend on full sequence similarity, which become inaccurate when proteins diverge (e.g. inversions, deletions, insertions). Due to these and other shortcomings, we developed a novel unsupervised algorithm which discovers highly conserved regions and uses two types of information measures: (i) data measures computed from input sequences; and (ii) class measures computed using a priori class groupings in order to reveal subgroups (i.e. classes) or functional characteristics.

Results: Using known and putative sequences of two proteins belonging to a relatively uncharacterized protein family we were able to group evolutionarily related sequences and identify conserved regions, which are strong homologous association patterns called Aligned Pattern Clusters, within individual proteins and across the members of this family. An initial synthetic demonstration and in silico results reveal that (i) the data measures are unbiased and (ii) our class measures can accurately rank the quality of the evolutionarily relevant groupings. Furthermore, combining our data and class measures allowed us to interpret the results by inferring regions of biological importance within the binding domain of these proteins. Compared to popular supervised methods, our algorithm has a superior runtime and comparable accuracy.

Availability and implementation: The dataset and results are available at www.pami.uwaterloo.ca/∼ealee/files/classification2015.

Contact: akcwong@uwaterloo.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As sequencing technologies continue to improve, the amount of biological information has become more difficult to effectively analyze. Biologists often need to identify conserved regions in DNA or protein sequences as these are often essential for biological function. Conversely, areas of low sequence homology are often useful in studying evolutionary divergence or novel function (Ng and Henikoff, 2006). Herein, we develop an unsupervised aligned pattern clustering tool to facilitate the identification of areas of conservation and divergence in protein sequences. Using two members of the class A scavenger receptors, whose evolutionary origin and relatedness has been previously described based on phylogenic approaches (Whelan *et al.*, 2012; Yap *et al.*, 2015), we demonstrate that our method is as effective as multiple sequence alignments for motif discovery. In addition, our algorithm is more effective in grouping the two proteins across a wide range of organisms, including the identification of conserved and mutated amino acids within each protein and across the family of proteins, thus identifying regions of potential functional significance amongst homologs and orthologs.

Protein families can consist of many thousands of members, and the distinction between these members becomes less clear with greater evolutionary distance. In supervised learning (also known as classification), the groups are predetermined as *a priori* knowledge, given as class labels, and the model is optimally learned to predict those labels (Leslie *et al.*, 2002). In unsupervised clustering, the samples (i.e. sequences) are grouped or partitioned based on their similarity (Perner and Rosenfeld, 2003) or common features (i.e. amino acid sequences/motifs with strong statistical association usually along the entire sequence) (Durston *et al.*, 2012). To show the uniqueness of our approach, we propose the following problem definition: Given a set of input sequences in pre-existing groups (i.e. with class labels), our algorithm discovers subsequences (i.e. motifs/regions) with strong statistical association which distinguishes subgroups within the data and in turn reveal other biologically relevant information. We can rank the quality of these discovered patterns, and group their inherent characteristics using our data measures. By then using our class measures, we can identify whether these discovered patterns and their amino acid mutations partitioned subgroups match the pre-existing groups as reflected by the class labels.

An example using the well-known cytochrome c protein (Fig. 1) is used to illustrate how local patterns can be related to species. Here the sequence patterns CAECH, CAWCH and CLQCH appear in the input sequences describing three class groups, namely *plant*, *mammal* and *fungi*. If these patterns were conserved amino acid sequences with presumed functional significance, the aligned patterns could reveal similarities and/or differences for associating with the species (i.e. classes). Once conserved regions and their associated patterns are localized, information measures are able to partition group or class characteristics from the aligned patterns. Such information can then be used to describe the classes as well as to justify and assist classification.

1.1 Motivations

We developed an unsupervised algorithm to discover regions of potential functional importance which is (i) flexible in pattern length and number of mutations and (ii) enriched with pattern and data representations for correlating subgroups. Our new unsupervised method, called WeMine-APC, first discovers non-redundant, statistically significant sequence patterns from a sequence family to create Aligned Pattern Clusters (APCs) (Lee and Wong, 2013; Wong *et al.*, 2012), which are clusters of sequence patterns that are aligned to localize the correlation of the amino acids within a region.

Therefore, we introduce the concept of pattern-data spaces (Wong and Li, 2008) (Fig. 2). The pattern space consists of the aligned patterns within the APC and the data space is formed by the occurrences of the subsequences containing the aligned patterns of the APC, i.e. the array of sequence data formed by all the subsequence occurrences or instances of the aligned patterns in the APC with sequence ID and location given (Wong *et al.*, 2012). First, from the pattern space, the patterns are discovered as statistical associations with ranking in the localized region. Next, from the data space, the information measures are computed to reveal partitions of the patterns and data as well as amino acid mutations within and between the aligned columns. Thus, all the probability used to define information measures are derived from the data space of the APC. Further role of the pattern-data space is summarized in supplementary materials.

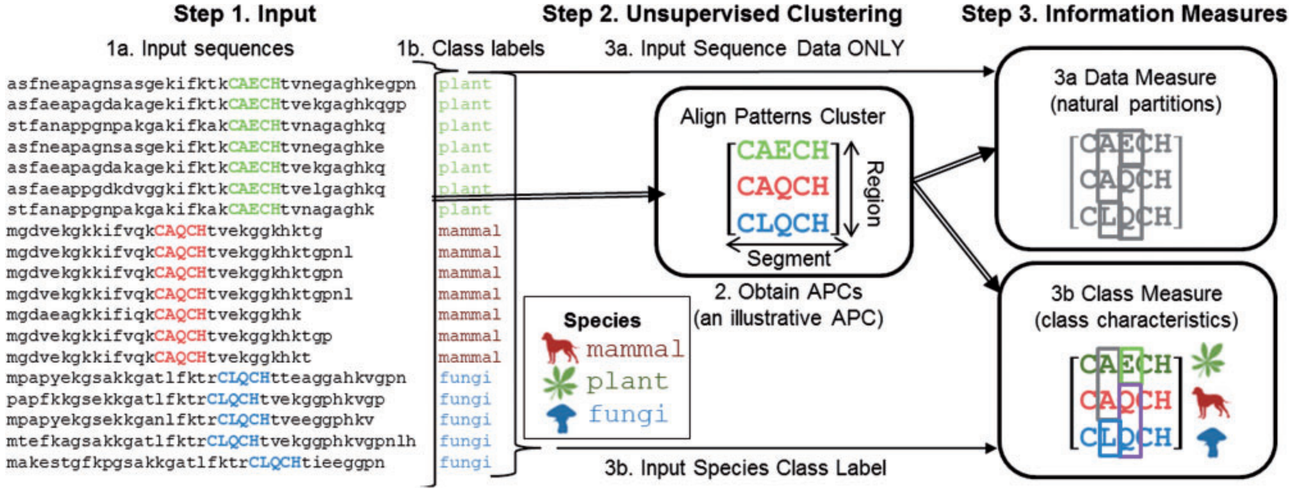
Because the data space contains a localized region correlating the sequences, distinct protein family subgroups can be revealed in a natural manner. Thus, both amino acid conservation and mutation within an APC can help partitioning the proteins data using simple information measures. This paper proposes two types of quantified information measures to assess the ability of our unsupervised clustering algorithm to obtain natural partitions within the data region (Fig. 1, Step 3). They are (i) *data measures* that assess natural partitions that reveal inherent characteristics of subgroups within the data induced by the APC; and (ii) *class measures* that assess the correlation of the constituent representations of APCs such as sequence patterns, amino acid mutation sites (columns) and specific amino acids that are identified from the data based upon the external class labels.

2 Methods

The proposed methodology (Fig. 1) is separated into two parts: (i) constructing APCs to reveal subgroups and/or classes; and (ii) using appropriate information measures to assess either the natural data partitions or the inherent subgroup class characteristics (Fig. 2). At the onset of unsupervised clustering, *a priori* class labels obtained externally are not incorporated into the pattern discovery process. Without using external class labels, APCs discovered with their informative sites of conservation and/or mutations are used to disclose natural partitions within the data. Once the natural partitions are found, externally collected class labels are then incorporated to compute the information measures so as to ascertain how closely the external class labels correspond to the natural partitions obtained. We use a biological example from the results (Fig. 3) to demonstrate the usefulness of our information measures.

2.1 Constructing APCs to reveal subgroups/classes

An APC is defined as a set of sequence patterns clustered together based on their aligned similarities. By allowing flexible mutations, APC maximizes the column-wise (vertical) similarity of the amino acids at each position while minimizing subgroup mutations between the sequence patterns. Since pattern alignment begins with statistically significant patterns, the AP clustering process (i) is faster due to its pattern-based clustering which takes advantage of the strong statistical constraints among the sequence patterns rather than considering all possible combinations of site mutations such as in most typical motif finding algorithms; (ii) is a flexible algorithm that relaxes the constraints of specific length, and number of mutations, as well as the amount of sequence coverage; (iii) renders a more practical ranking due to the alignment of multiple homologous local patterns as one cluster. When frameshifts occur, weak APCs allowing frameshift mutations could be adopted (Wong *et al.*, 2013) but will not be tackled in this paper. If frameshift mutations are conserved, they will be clustered into separate APCs. Hence, WeMine-APC significantly improves the discovery and interpretation of local functional subgroups for a large class of data.



1a. Input sequences	1b. Class labels
asfneapagnsasgekifktkCAECHtvnegaghkegpn	plant
asfaeapagdakagekifktkCAECHtvekgaghkqgp	plant
stfanappgnpakgakifkakCAECHtvnagaghkq	plant
asfneapagnsasgekifktkCAECHtvnegaghke	plant
asfaeapagdakagekifktkCAECHtvekgaghkq	plant
asfaeappgdkdvggkifktkCAECHtvelgaghkq	plant
stfanappgnpakgakifkakCAECHtvnagaghk	plant
mgdvekgkkifvqkCAQCHtvekggkhktg	mammal
mgdvekgkkifvqkCAQCHtvekggkhktgpnl	mammal
mgdvekgkkifvqkCAQCHtvekggkhktgpn	mammal
mgdvekgkkifvqkCAQCHtvekggkhktgpnl	mammal
mgdaeagkkifiqkCAQCHtvekggkhk	mammal
mgdvekgkkifvqkCAQCHtvekggkhktgp	mammal
mgdvekgkkifvqkCAQCHtvekggkhkt	mammal
mpapyekgsakkgatlfktrCLQCHtteaggahkvgpn	fungi
papfkkgsekkgatlfktrCLQCHtvekggphkvgp	fungi
mpapyekgsekkganlfktrCLQCHtveeggphkv	fungi
mtefkagsakkgatlfktrCLQCHtvekggphkvgpnlh	fungi
makestgfkpgsakkgatlfktrCLQCHtieeggpn	fungi

Fig. 1. An example using cytochrome *c* to illustrate the use of our unsupervised approach and information measures for partitioning and correlating subgroup characteristics for conserved regions. The input contains protein sequences with the discovered patterns for the proximal binding site (coloured). Each sequence belongs to a species group (i.e. class) of either mammal (red), plant (green) or fungi (blue). Overall, our approach involves two steps (Step 2 and Step 3). First in Step 2, an unsupervised algorithm clusters, aligns and partitions patterns simultaneously into an Aligned Pattern Cluster (APC). Our algorithm, WeMine-APC, marks, aligns and extracts a data block covering all the discovered patterns in the APC. We refer the data block as the data space that is constructed from the APC to represent a conserved region of the protein. Second, information measures are used for assessing the amino acid mutations in the natural data partitions (3a) and a priori class labels (3b), respectively, in the data space of the APC. Step 1a. At the onset of unsupervised clustering, a priori class labels obtained externally are not incorporated into the pattern discovery process. Step 2. Without using external class labels, an APC with specific conservations and mutations are discovered from sequence data in order to reveal natural partitions within the region. Step 3. Two types of information measures reveal partitions: data measures using only the sequence input; and class measures using the additional class labels. Step 3a. The natural partitions of an APC are measured based on data measure SR2 (a measure of interdependence of a column with all other columns in the APC) Step 3b. The a priori class labels are further input in Step 3b to associate with partitions of patterns (rows) and of mutations (columns) based on two class measures, Class Entropy and Information Gain (Color version of this figure is available at *Bioinformatics* online.)

Let the alphabet Σ be a collection of amino acids including the wildcard symbol, $*$, and the gap symbol, $-$, and let the pattern {p $= \sigma_1\sigma_2 \ldots\ \sigma_n$ be statistically significant ordered sequences containing symbols from Σ, where $\sigma_i \in \Sigma$. The APC is then a set of similar patterns of different lengths that is simultaneously assembled into an aligned set of patterns that are sequence segments of the same length by appropriately inserting gaps and wildcards.

Definition

Let an APC (Lee and Wong, 2013) be defined as:

$$C = \begin{pmatrix} \sigma_1^1 & \cdots & \sigma_n^1 \\ \vdots & \ddots & \vdots \\ \sigma_1^m & \cdots & \sigma_n^m \end{pmatrix}_{mxn} \qquad (17)$$

In the context of the above APC notation, the ith row vector (or segment) of C contains the aligned pattern p_i, for $i = 1,\ \ldots,\ m$, where each pattern is of length n. We use c_j to denote the jth column, for $j = 1,\ \ldots,\ n$.

2.2 Computing the information measures for quantifying subgroup partitions

There are two types of information measures to be considered. The first is the *data measure* that is unaffected by class label biases. It is derived from the natural partitions within the data, reflecting its inherent varied functional characteristics. The second is referred to as the *class measure* that quantifies the strength of a constituent representation of an APC associated with the externally acquired class labels, as well as their quality. It is for comparing and assessing the class partition against the natural partition inherent in the data.

Together, there are four measures as presented in Table 1. The two data measures are (i) data entropy, and (ii) normalized sum of mutual information redundancy (SR2) (Wang and Wong, 1978). The two class measures are (i) class entropy (H) for the APC and its sub-components, such as the APC itself, its patterns, and its distinct amino acids therein; and (ii) class information gain (IG) for aligned columns. These measures are based on the theoretic definitions of: (ii) Shannon's information entropy (Shannon, 2001; Strait and Dewey, 1996), (ii) mutual information and (iii) the change in information entropy (i.e. information gain). Detailed equations for the data and class measures can be found in supplementary materials Section 1.5 and 1.6, respectively.

2.3 A biologically relevant example using members of the class a scavenger receptor protein family

Within our results, we chose to focus on those which contain cysteine residues responsible for the disulfide bonds in the terminal domain of the cA-SRs. The 2nd highest scoring APC (Fig. 3b) displays five sequence patterns and thirteen aligned columns. The dataset contains five relevant patterns CRMLGYS, CRSLGY, VFCRMLG, SDATVFCRMLGYS and DATVFCRMLGYS. First considering class measures, the pattern $p_1 =$ CRMLGYS has a class entropy of $H_Y(p_1) = 0$ with $\vec{Y} = $((MARCO; 17), (SRAI; 0)) and the pattern p_2 =CRSLGYS has a class entropy of $H_Y(p_2) = 0$ with $\vec{Y} = $((MARCO; 0) and (SRAI; 15)). In the same table, the APC has a class entropy of $H_Y(\mathrm{C}) = 0.99$, which all the aligned columns also share. In aligned column 432, amino acid M has the class entropy 0 and amino acid S has class entropy 0, indicating that both are perfect partitions of the proteins MARCO and SRAI that lead to an information gain of 1. Next considering data measures, aligned column 432 has the second highest SR2 = 0.36; furthermore, aligned column 436 has the highest SR2 = 0.41.

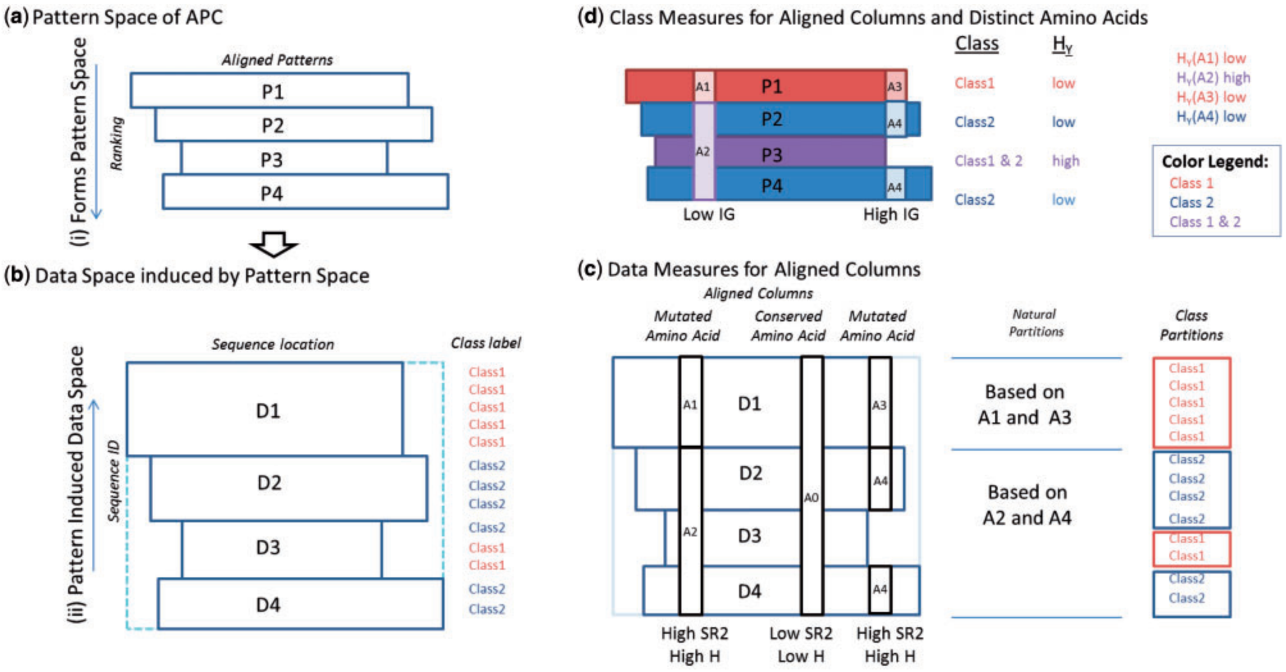


Fig. 2. The pattern and data spaces of apcs and the use of information measures for class and natural data partitioning. A segment is a sub-sequence of a full sequence. A region is considered as an area in the set of input sequences consisting of several related segments. The interdependent association of amino acids that comprise the segment itself and vertically accounted by amino acids conserved or mutated on a site (or column) are defined as (**a**) pattern Space of APCs: the set of aligned patterns in an APC with statistical ranking, statistical residuals and probability distribution; (**b**) data space of an APC (or pattern induced data space of an APC): a rectangular array of sequence data covering all the aligned patterns in the APC with sequence id and pattern location given in each row of the block. (**c**) Natural data partitions subdivide the input sample sequences by the feature in the APC Data Space, such as amino acid mutations or aligned columns. It is considered natural because it is based on the data feature itself and not obtained from external a priori knowledge such as class labels. Data Measures H and SR2 are obtained from the data space to reveal the natural data partitions. In c, columns with low H contains specific amino acid (such as A0). Columns with High SR2 (those with A3 and A4) are those with strong interdependence with other columns (those with A1 and A2) within the APC. Class partitions subdivide the samples by the class labels that were given as a priori knowledge. (**d**) Class measures such as class entropy and information gain relate APC representation (such as patterns, columns and distinct amino acids in specific columns) with the class labels. Low entropy and high information gain suggest that their strong association with a specific class, the reverse suggests the association with more classes. Amino acid partitions further subdivide the samples by their aligned column using the different amino acid mutations in that column. Distinct amino acids like A3 reveal a correlation to Class 1, whereas A4 correlates to Class 2 (Color version of this figure is available at *Bioinformatics* online.)

3 Results

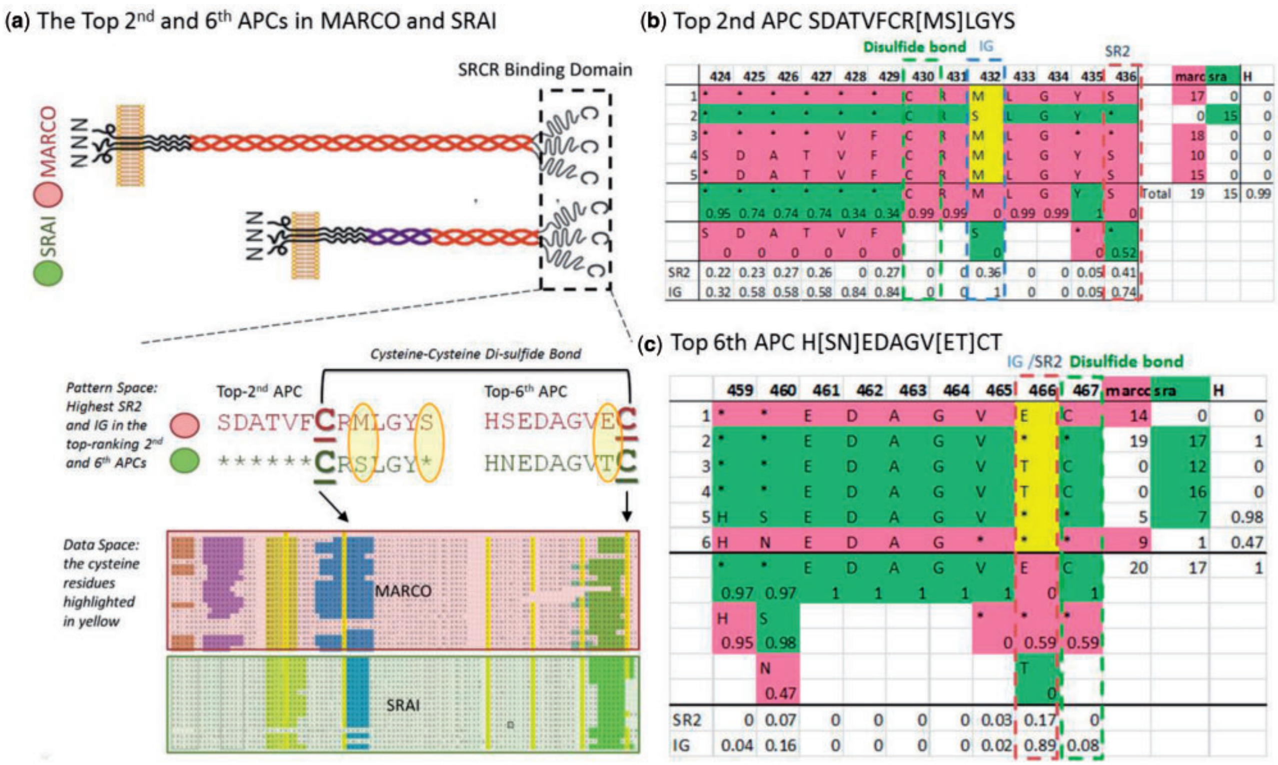
3.1 In silico case study of the cA-SRs

Although synthetic experiments (Supplementary Materials Section 1.8) are important in developing and testing our unsupervised approach, it is also critical to ensure that WeMine-APC is applicable to biologically relevant protein sequences in an unsupervised manner. Here, we applied WeMine-APC to two proteins of a known family of innate immune receptors called the class A Scavenger Receptors (cA-SRs). This family of receptors consists of five members, which are hypothesized to have originated via gene duplication events; however, these orthologs have diverged greatly from one another over time. Of particular interest are two proteins, MARCO and SRAI. Although, these proteins share much sequence similarity, they have both evolved to have very unique functions which are likely to be expressed within their protein sequences. Thus, these two proteins are excellent test cases for the WeMine-APC algorithm since members of this family possess protein sequences which share domains of high conservation as well as areas of great dissimilarity. As described in Whelan *et al.* (2012), protein sequences of cA-SRs were gathered from available genomes of vertebrate organisms in NCBI, EBI and Ensembl databases. The dataset for the unsupervised clustering algorithm contained two classes, MARCO and SRAI; there are 21 MARCO sequences and 16 and SRAI sequences included from vertebrate species. We discovered a set of APCs which contain amino acid mutations that can be ranked by data and class measures and are believed to hint at the functional differences between these two proteins.

3.2.1 Measuring class label quality by class information gain

To establish the previous observation that IG can indeed assess the quality of class partitions, two different sets of class labels were used for one input protein sequence dataset. One dataset corresponds to *good class label partitions*, i.e. the gene class labels MARCO and SRAI, which actually denote how the protein sequences are naturally partitioned. The other corresponds to *poor class labels*, i.e. species taxonomy which has no direct relationship with these sequences. Thus, the APC with all its patterns discovered in the data are the same, but the two different class labels lead to two different values with regards to revealing the capability of these class measures.

The maximum IGs of each discovered APC are presented as a bar graph (Fig. 4). When the class label is correct (Fig. 4, blue bars), the value of IG is near 1, indicating class partitions that correlate perfectly with the natural data partitions. Good class labels also yield substantial values of IG close to 0, which indicates no partitions within the input sequences that correspond to the accompanying class labels. When the class label is incorrect (Fig. 4, red bars), the values of IG are low, specifically more uniformly low for all columns, indicating that the class partitions do not correlate with the


(a) The Top 2nd and 6th APCs in MARCO and SRAI
SRCR Binding Domain
MARCO
SRAI
N
C
Cysteine-Cysteine Di-sulfide Bond
Pattern Space: Highest SR2 and IG in the top-ranking 2nd and 6th APCs
Top-2nd APC
SDATVFCRMLGYS
******CRSLGY*
Top-6th APC
HSEDAGVEC
HNEDAGVTC
Data Space: the cysteine residues highlighted in yellow
MARCO
SRAI
(b) Top 2nd APC SDATVFCR[MS]LGYS
Disulfide bond
IG
SR2
(c) Top 6th APC H[SN]EDAGV[ET]CT
IG /SR2
Disulfide bond

	424	425	426	427	428	429	430	431	432	433	434	435	436		marc	sra	H
1	*	*	*	*	*	*	C	R	M	L	G	Y	S		17	0	0
2	*	*	*	*	*	*	C	R	S	L	G	Y	*		0	15	0
3	*	*	*	*	V	F	C	R	M	L	G	*	*		18	0	0
4	S	D	A	T	V	F	C	R	M	L	G	Y	S		10	0	0
5	*	D	A	T	V	F	C	R	M	L	G	Y	S		15	0	0
	*	*	*	*	*	*	C	R	M	L	G	Y	S	Total	19	15	0.99
	0.95	0.74	0.74	0.74	0.34	0.34	0.99	0.99	0	0.99	0.99	1	0				
	S	D	A	T	V	F			S			*	*				
	0	0	0	0	0	0			0			0	0.52				
SR2	0.22	0.23	0.27	0.26	0	0.27	0	0	0.36	0	0	0.05	0.41				
IG	0.32	0.58	0.58	0.58	0.84	0.84	0	0	1	0	0	0.05	0.74				

	459	460	461	462	463	464	465	466	467	marcc	sra	H
1	*	*	E	D	A	G	V	E	C	14	0	0
2	*	*	E	D	A	G	V	*	*	19	17	1
3	*	*	E	D	A	G	V	T	C	0	12	0
4	*	*	E	D	A	G	V	T	C	0	16	0
5	H	S	E	D	A	G	V	*	*	5	7	0.98
6	H	N	E	D	A	G	*	*	*	9	1	0.47
	*	*	E	D	A	G	V	E	C	20	17	1
	0.97	0.97	1	1	1	1	1	0	1			
	H	S					*	*	*			
	0.95	0.98					0	0.59	0.59			
		N						T				
		0.47						0				
SR2	0	0.07	0	0	0	0	0.03	0.17	0			
IG	0.04	0.16	0	0	0	0	0.02	0.89	0.08			

Fig. 3. Top Ranking APCs containing cysteine residues from the SRCR binding domain. MARCO and SRAI proteins have both evolved to have very unique functions which are likely to be expressed within their protein sequences. Members of this family possess protein sequences which share domains of high conservation as well as areas of dissimilarity. (**a**) From the discovered APCs that are in the SRCR binding domain, amino acid mutations from two top-ranking APCs (2nd APC: SDATVFCR[MS]LGYS; 6th APC: H[SN]EDAGV[ET]CT) have relatively high SR2 and IG (yellow ellipses) and separate the data space of these two APCs into the MARCO and SRAI classes. The cysteine-cysteine disulfide bond is between the 460 aligned column of the second APC, and 518 aligned column of the sixth APC. The data space of the SRCR domain and the cysteine residues highlighted in yellow, the coloured regions are the APCs clustered in this particular co-occurrence graph (not shown here) (Lee et al., 2014). In particular, these two APCs highly co-occur across protein sequences. (**b**) The second APC SDATVFCR[MS]LGYS. The APC is represented in table form. Each numbered row associates with a statistically significant pattern and each numbered column is an aligned column of amino acids, which are either conserved or mutated. The instance counts of the patterns for MARCO and SRAI are listed on the right hand columns labelled with MARCO in pink and SRAI in green. The Class Entropy of each pattern and the APC is displayed in the last column with the heading H and each amino acid is displayed below the patterns; additionally, the SR2 and IG is summarized at the bottom. Note that the aligned column 432 [MS] is a strong class partition due to its perfect IG of 1 where M corresponds to MARCO and S to SRAI. Next, we evaluate the top SR2 class measure for each APC to help interpret the data partition. The aligned column 432 with high SR2 is a [S] amino acid mutation. (**c**) The sixth APC H[SN]EDAGV[ET]CT. The APC is read similarly to (b), except that the top ranking aligned column for class partition measure (IG) is the same as the data partition measure (SR2). The aligned column 466 [ET] with the highest SR2 is also a strong data partition column with E corresponding to MARCO and T to SRAI in addition to being the strong class partition with highest (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Summary of measures for assessing aligned columns

		Dependent on only data or also with class	
		Data measures (dependent on data)	Class measures* (dependent on class)
Property measured	Stability Partitions	Data entropy $H(c_j)$ Sum of mutual information $SR2(c_i)$	Class entropy $H_Y(c_i)$ Class information gain $\Delta\ H_Y(c_j)$

*Y is defined as class distribution, thus H_Y denotes class measures, meaning that the measure uses class labels for its computation.

natural partitions in the data. When the class labels of the amino acid mutations are correct, the value of IG is close to 1, indicating that the natural data partitions resulting from the amino acid information strongly correlates with the class labels. However, when the natural data partitions only weakly correlate with the class labels, the value of IG is low. The natural data partitions are based on composition of the amino acid in the input sequence and class partitions are based on the *a priori* class labels. IG measures how close these two different partitions agree with one another.

3.2.2 Ranking amino acid partition

In Table 2, we observe that the sequence columns containing mutated amino acids follow the same order when ranked by data measures (SR2) as well as with class measures (IG). This indicates that both of these partition measures are able to qualify the natural partitions in the dataset. However, when the representation contains only one type of class label, the value of IG is always zero while the value of SR2 is still non-zero. Therefore, SR2 is a more robust measure than IG due to its ability to rank single class amino acid mutations. So when the class labels are not computable by IG, SR2 can still assess the data partitions, which likely correspond to the natural subclasses in the data.

3.2.3 Disulfide bonds in the top ranking aligned columns

The SRCR domain in the cA-SR contains 6 cysteine residues paired into 3 disulfide bonds that are important for its three-dimensional structure (Yap et al. 2015). We observe that five top-ranking APCs overlap or are adjacent to a conserved cysteine residue (Table 2). We discover all six cysteines in the disulfide bonds: where three are directly in a good-partitioning APC, two in a poor-partitioning APC and one is directly adjacent to an APC. Furthermore, two cysteines (C bold and underscored) in the top APCs (top 2nd APC and top 6th APC, Table 2) directly bind with one another in a disulfide bond (Fig. 3a). Furthermore, each APC contains amino acid mutations that correlate precisely with the class labels (Fig. 3b).

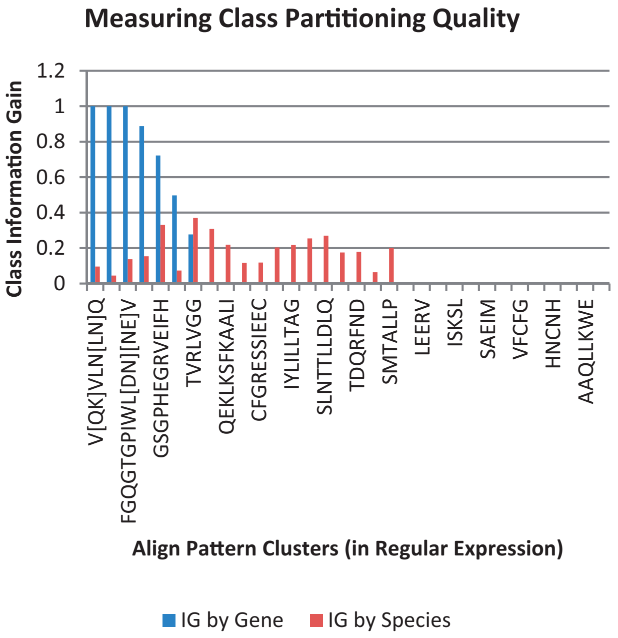


Fig. 4. Class label quality greatly affects the acquired class information gain (IG). Class IG can be used to measure the quality of the external class labels, i.e. how well class partitions correlate with data partitions. The aligned column with maximum IG from each APC is presented as a bar graph: the blue bar is class IG computed using the good gene class labels and the red bar is the class IG computed with the poor species labels. Some APCs have IG of 0 for both gene class labels and species class labels because they occur within only one class (Color version of this figure is available at *Bioinformatics* online.)

To demonstrate the biological interpretability of our information measures, the second and sixth APCs were examined in full detail (Fig. 3b, c, respectively). First, we evaluate the top IG class measure for each APC to help interpret the class partitions. Next, we evaluate the top SR2 class measure for each APC to interpret the data partition as stated in the caption. Note that the top IG and SR2 are the same in the top second APC but different for the top sixth APC.

4 Discussion

In this study (Table 2), we observed 7 out of the top 14 APCs with positive IG scores that were able to partition the cA-SRs. Three of these APCs have amino acid mutations with an IG value of 1, indicating that they partitioned the protein classes precisely. These APCs contain amino acid mutations that are natural data partitions that match the class partitions. Some of these partitions have been observed previously and are thought to explain some of the functional differences between these proteins. For example, both MARCO and SRAI contain a ligand binding, Scavenger Receptor Cysteine Rich (SRCR) domain. MARCO's SRCR domain contains a motif (RGRAEVYYSGT) which is not shared by SRAI and is thought to explain some of the differences in the ligand binding capabilities of these two receptors (Brännström *et al.*, 2002). WeMine-APC correctly identifies this motif as in the 11th ranking APC, NRGRAEVYY. Consistent with previous results, this motif is only found within the MARCO class and not within any SRAI sequence . Additionally, the ninth ranking APC, WGT[IV]CDDRW, was also recently identified by Yap *et al.* (2015) via FastML and is hypothesized to be important structurally given its inclusion of a conserved cysteine residue. This demonstrates that SR2 can reveal subclass dependency even when class labels are unreliable or IG is 0. In summary, both of these patterns have been previously identified with methods other than WeMine-APC to represent motifs of known functional importance.

Table 2. The top 14 APCs ranked by partition measures SR2 and IG

	APC	SR2	IG	SRCR domain	Disulfide bond
1	V[QK]VLN[LN]Q	0.44	1.00	No	No
2	SDATVF**C**R[**MS**]LGYS	0.41	1.00	Yes	Yes
3	FGQGTGPIWL[**DN**][**NE**]V	0.36	1.00	Yes	Adjacent
4	VLNNITNDLRLKDWEHSQTL	0.36	0.00	No	No
5	QEKLKSFKAALI	0.30	0.00	No	No
6	H[SN]EDAGV[**ET**]**C**T	0.25	0.89	Yes	Yes
7	GSGPHEGRVEIFH	0.25	0.72	No	No
8	EHFQNFS	0.11	0.00	No	
9	WGT[**IV**]**C**DDRW	0.09	0.50	Yes	Yes
10	**C**FGRESSIEE**C**	0.07	0.00	Yes	Yes
11	NRGRAEVYY	0.06	0.00	Yes	
12	IYLILLTAG	0.05	0.00	No	
13	FKQQEE	0.02	0.00	No	
14	TVRLVGG	0.01	0.50	No	No

Therefore, WeMine-APC is able to assess the impact of an amino acid mutation by partitioning the protein sequences by related APC patterns and their amino acid mutations to the regions in the protein and their relation with the class labels.

Table 3. Runtime comparisons to supervised methods

Method	Training time	Testing time	Total time
HMM	4.0302*	8.59950	12.63052
SVM	1.7594	0.01217	1.77157
WeMine-APC	0.1092	0.01745	0.12665

*Not including time to build profile HMM from online server for .hmm model (depends on load and usage of the server).

In addition to these previously described patterns, another 4 of the top 14 APCs were identified between the SRCR domains of MARCO and SRAI; these novel motifs represent patterns of possible future interest to the cA-SR field (Table 2). In particular, the second APC, SDATTVFCR[MS]LGYS, is a perfect partition between MARCO and SRAI and also contains a conserved cysteine necessary for disulfide bonding. The sixth APC, H[SN]EDAGV[ET]CT, also contains a conserved cysteine residue.

4.1 Comparison to supervised methods

To study the predictive power of WeMine-APC, supervised prediction was implemented to compare against other currently available supervised methods. The dataset described in the previous section was separated into two parts: 70% training set and 30% testing set. The training and testing sets are mutually exclusive and were randomly selected. Both SVM and HMM are supervised learning methods; however, our method is unsupervised. As such, the WeMine-APC algorithm was modified to predict class labels of an unknown sequence (Supplementary Materials Section 1.9). The goal of our modified supervised algorithm was to acquire a prediction for each input sequence with unknown class labels based on the set of discovered APCs that were generated from the training set (Supplementary Materials Section 1.8).

4.1.1 Runtime comparisons

To determine the runtime advantage of our method, the training and testing time of our method and other supervised methods is measured. Our results show that the runtime for WeMine-APC is much faster than HMM and SVM, due to the faster training time (Table 3). The testing time is about the same as SVM because both methods match patterns to make their predictions; however, the testing time is much slower for HMM due to full sequence alignment. The training time for WeMine-APC is 16 times faster than SVM and 37 times faster than HMM. This faster training runtime is because an APC aligns statistically significant patterns instead of matching all possible combinatorial mutations such as in SVM. Training in WeMine-APC is substantially faster than HMM because full sequence alignment is not required.

4.1.2 Accuracy comparisons

To evaluate the accuracy of our method, class mislabeling was injected into the training set to evaluate the robustness of accuracy (see algorithm in Supplementary Materials Section 1.9). Our accuracy (Fig. 5) decreased after SVM but before HMM, implying that HMM is robust to mislabeling due to the computationally expensive pairwise full sequence alignment. Since APC is faster than both SVM and HMM (Table 3), we recommend WeMine-APC to be used for divergent sequences or sequences with mislabels, thus allowing the discovery of conserved patterns, the measurement of label quality and the interpretation of the amino acid mutation.

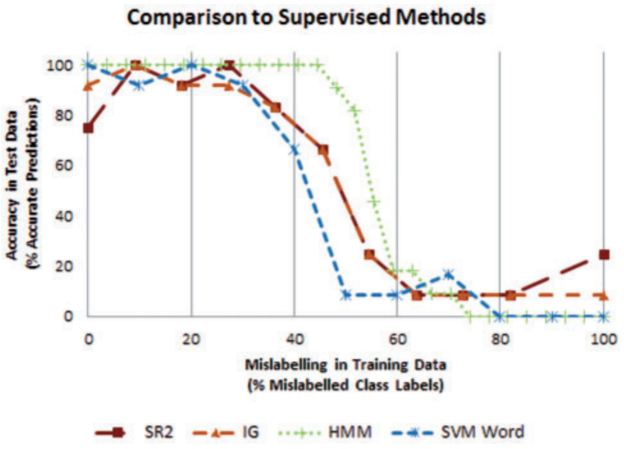


Fig. 5. Supervised classification accuracy. The accuracy of SR2 and IG is compared against HMM and SVM with four different kernels. APC is able to capture subtle mutational difference for subgroup partition between the classes, thus preserving a higher minimal accuracy of at least 10% and also gradual slower decrease in accuracy than other methods

5 Conclusions

In this study, we demonstrated that unsupervised methods on local functional regions are preferred over supervised algorithms because they explore, analyze and organize the patterns discovered from the data without relying on externally acquired class labels. However, any unsupervised method should be assessed using data with reliable external class labels in order to determine the accuracy of the method in identifying class members within families. Hence, while we utilized an unsupervised method to obtain the natural partitions within the data, we also assessed the quality of the results using two information measures to be justified by the externally collected class labels, even if those class labels are not included or used in the clustering phase.

APC uses two types of information measures: (i) data measures which are unbiased by external class labels and (ii) class measures which identify mislabeling errors. Our experiments demonstrate that labeling quality of cA-SRs can be measured quantitatively and the APC can make biological interpretations of disulfide bonds in the SRCR binding domain. Compared to supervised methods, APC has superior runtime and comparable accuracy. In conclusion, the contributions of the data and class information measures are (i) unbiased to class labels, (ii) interpretability of the mutations and the data partition and (iii) APC classification being comparable to supervised learning.

Conflict of Interest: none declared.

References

Brännström,A. *et al.* (2002) Arginine residues in domain V have a central role for bacteria-binding activity of macrophage scavenger receptor MARCO. *Biochem. Biophys. Res. Commun.*, **290**, 1462–1469.

Durston,K.K. *et al.* (2012) Statistical discovery of site inter-dependencies in sub-molecular hierarchical protein structuring. *EURASIP J. Bioinf. Syst. Biol.*, **2012**, 8.

Lee,E.A. and Wong,A.K.C. (2013) Ranking and compacting binding segments of protein families using aligned pattern clusters. *Proteome Science*, 11, 1–23.

Lee,E.S. *et al.* (2014) Discovering co-occurring patterns and their biological significance in protein families. *BMC Bioinformatics*, **15**, S2.

Leslie,C. *et al.* (2002) Mismatch string kernels for SVM protein classification. In: Advances in neural information processing systems. pp. 1417–1424.

Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.

Perner,P. and Rosenfeld,A. (2003) In: Perner,P. and Rosenfeld, A. (eds.) *Machine Learning and Data Mining in Pattern Recognition*. vol. **2734**. Springer, Berlin, Heidelberg.

Strait,B.J. and Dewey,T.G. (1996) The Shannon information entropy of protein sequences. *Biophys. J.*, **71**, 148–155.

Wang,D.C.C.W.D.C.C. and Wong,A.K.C.W.A.K.C. (1978). Classification of discrete data with feature space transformation. In: *1978 IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive*.

Whelan,F.J. *et al*. (2012) The evolution of the class A scavenger receptors. *BMC Evol. Biol.*, **12**, 227.

Wong,A.K. and Lee,E.S.A. (2014) Aligning and clustering patterns to reveal the protein functionality of sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), **11**, 548–560.

Wong,A.K.C. *et al*. (2012) Discovery of delta closed patterns and noninduced patterns from sequences. *IEEE Trans. Knowl. Data Eng.*, **24**, 1408–1421.

Wong,A.K.C. and Li,G.C.L. (2008) Simultaneous pattern and data clustering for pattern cluster analysis. *IEEE Trans. Knowl. Data Eng.*, **20**, 911–923.

Yap,N.V.L. *et al*. (2015) The evolution of the scavenger receptor cysteine-rich domain of the class A scavenger receptors. *Front. Immunol.*, **6**, 1–9.

Zhuang,D.E.H. *et al*. (2014) Discovery of temporal associations in multivariate time series. Knowledge and Data Engineering, IEEE Transactions on, **26**, 2969–2982.