Bioinformatics and Biology Insights

ORIGINAL RESEARCH

Association of Putative Members to Family of Mosquito Odorant Binding Proteins: Scoring Scheme Using Fuzzy Functional Templates and Cys Residue Positions

Malini Manoharan*[,1–3], Kannan Sankar*[,2,4,5], Bernard Offmann[1,6] and Sowdhamini Ramanathan[2]

[1]Université de La Reunion, DSIMB, INSERM UMR-S 665, La Reunion, France. [2]National Centre for Biological Sciences, Tata Institute for Fundamental Research, GKVK campus, Bangalore, INDIA. [3]Manipal University, Madhav Nagar, Manipal, Karnataka, India. [4]Birla Institute of Technology, Pilani, Rajasthan, India. [5]Current address: Iowa State University, Ames, IA, USA. [6]Université de Nantes, UFIP CNRS FRE 3478, Nantes, France. *Both authors contributed equally.
Corresponding author email: mini@ncbs.res.in

Abstract: Proteins may be related to each other very specifically as homologous subfamilies. Proteins can also be related to diverse proteins at the super family level. It has become highly important to characterize the existing sequence databases by their signatures to facilitate the function annotation of newly added sequences. The algorithm described here uses a scheme for the classification of odorant binding proteins on the basis of functional residues and Cys-pairing. The cysteine-based scoring scheme not only helps in unambiguously identifying families like odorant binding proteins (OBPs), but also aids in their classification at the subfamily level with reliable accuracy. The algorithm was also applied to yet another cysteine-rich family, where similar accuracy was observed that ensures the application of the protocol to other families.

Keywords: cysteine-based scoring scheme, Classification of proteins, Functionally important residues, Ligand binding residues

Bioinformatics and Biology Insights 2013:7 231–251

doi: 10.4137/BBI.S11096

This article is available from http://www.la-press.com.

Introduction

The role of olfaction is the major source for host identification among mosquitoes. The molecular basis of this chemical signal recognition is systematically encoded by a series of proteins. Odorant binding proteins are thought to be the primary proteins involved in the transport of odorants and pheromones to the olfactory receptors.[1,2] Members of this protein family have been identified in a number of insect species, including four dipterian species *Drosophila melanogaster*.[3,4] *Anopheles gambiae*,[4,5] *Aedes aegypti*[6] and *Culex quinquefasciatus*.[7] Since their identification, this family of proteins has been of immense focus in the field of biology, as they could act as important target proteins. However, the sequence divergence of this family is very high in comparison to their function, which is to bind to a wide range of odorant molecules. It has been difficult to classify these proteins into different subfamilies for this reason. 3 major subfamilies have been defined previously in this family of proteins, which are *Classic, PlusC* and *Atypical* based on their cysteine conservation patterns.

In general, biological sequence data are accumulating rapidly as a result of advanced sequencing technology and concerted genome projects, at a greater rate than growth in computing efficiency.[8] The probability that a new protein can be classified as part of a sequence family is already near 30%.[9] Encouragingly, evolutionary constraints on protein sequences are imposed by requirements of 3-dimensional structure and biological function, which are main aspects employed for the classification of proteins. Generally, functional requirements are known to be more pronounced in terms of residue conservation, where an occurrence of completely conserved residues indicates a specific biological function. Many examples of such occurrences have been reported in protein sequences; 2 examples are the Ser-His-Asp triad of serine proteases[10] and the zinc finger motif of deoxyribonucleic acid (DNA)-binding proteins.[11] Mutation of such residues generally renders the protein inactive. Such residues can be either spread across the entire stretch of the protein or can be observed as conserved contiguous patterns termed "functional motifs". Such conservation status has been employed in annotating protein sequences by different methods reviewed by Ouzounis et al.[12] Though many of these methods fare well at assigning an unknown protein at a family level, the accuracy fails when a classification is required at a subfamily level. Several such function prediction algorithms require the availability of structural information, namely spatial interactions of residues of query sequences, in order to recognize preservation of geometry of functional residues. These include methods like Conserved functional group (CFG).[13,14] Whereas such methods could be quite applicable for proteins of unknown function, determined by structural genomics initiatives, structural information is either not available for most query sequences or the quality of models, derived by homology, could be limited. Residues near the active site might play an auxiliary role and are less easy to identify as part of "functional motifs". Sequence conservation of functional residues is therefore less obvious for residues that modulate the specificity of biological function. These residues change as a protein evolves to satisfy modified functional constraints, while the basic biochemical mechanism and the overall three-dimensional fold remain unaltered. In such cases, representative residues, associated with structural aspects of a protein, serve as better classifiers.

Cysteine, as a sulphur containing non-essential biogenic amino acid, plays critical roles in a number of metabolic processes. It is found as a part of a number of biological important proteins associated with important roles starting from folding to maintaining the integrity of structure to function. One of the most important roles of cysteines is the formation of disulphide bridges involved in the folding of proteins to form 3-dimensional structures. Disulphide bonds, which are formed by cysteines that may be sequentially apart but spatially proximate,[15] define the rigidity of large globular proteins. These disulphide bonds are generally conserved among related proteins[16–18] and the connectivity patterns can be used to identify proteins of similar 3-D structure.[19] The conservation of disulphide bond connectivity pattern enables the identification of remote homologues even when most of popular sequence search methods fail to do so. Such approaches, however, are complicated by observations of topologically equivalent disulphide bonds in non-homologues and also by non-equivalent numbers of disulphide bonds in close homologues.[20]

Owing to the fact that disulfide connectivity pattern formation in a protein is a directed (ie, non-random) process,[21] this property can be used to obtain

a structural classification of proteins. A large variety of connectivity patterns are found in disulphide-containing proteins.[21,22] In proteins with low sequence similarity, identical connectivity patterns can indicate high structural homology. Proteins that share a disulfide bonding pattern usually belong to the same structural family. Therefore, disulfide connectivity patterns provide a rapid and simple method for structural characterization of protein sequences and for examining structural properties, such as protein topologies.[21] entropic effects of cross-linkage,[22] structural superimposition of proteins by means of their disulfide bridge topology[20] and taxonomy of small disulfide-rich protein folds.[22] In addition, methods that classify proteins based on their connectivity patterns have also been established.[23] A systematic method for the classification of disulphide-rich proteins based on cysteine conservation is thus worth undertaking. Previous attempts on cysteine-based classification of proteins included approaches based on cysteine pairing,[23] identification of odorant binding proteins based on cysteine motifs,[4] conotoxin superfamily classification using pseudo amino acid composition and multi class support vector machines,[24] and classification of peroxiredoxins using regular expressions.[25]

An algorithm has been devised that can efficiently classify a new protein as an odorant binding protein belonging to a particular class by capturing specific information in terms of (1) functional residue conservation and (2) cysteine conservation and disulphide connectivity. The functional residue-based scoring scheme relies on the conservation of residues at functionally important sites (only sequence information) and a flexible distance-based scheme (also structural data). The functionally important sites were determined by the mapping of ligand binding residues on the structural alignment of the available structural members. The test sequences were aligned to the structural alignment and scores were assigned based on the residue conservation at these functional sites. The scoring of the distance-based scheme was based on a distance criterion between the residues at these positions. The distance criteria were established by observing the distances between the residues in the functional sites, including the 'fuzziness'; ie, the variation in distances observed among the crystal structures. The scores were calculated by a fit criterion after examining the distances within the models of unknown sequences. In our approach, for the queries whose structure is not yet available and homology modeling is unreliable due to relationship distance, a simple amino acid conservation-based scoring scheme is adopted that objectively measures the extent of conservation of functionally important residues (please see 'Scoring of query sequences' within the Methods section for details). Distances between such residues are not required or employed in this novel option. For the cysteine-based scheme, a "disulphide profile" of aligned sequences[19] has been employed of the various classes. The query sequences are aligned with these disulphide profiles followed by assigning a score based on the conservation of the cysteines in the query and further classifying them based on a composite classification scheme. These classification methods were primarily developed for the classification of odorant binding proteins in the mosquito genome. However, the functional residue-based classification was further extended to the serine protease family, where the classification of query sequences using the method into 3 subfamilies has been described. The cysteine-based classification was also implemented on the conotoxin family of proteins to extend the use of this method for the classification of disulphide-rich protein families at the subfamily level.

Methodology

Datasets

7 structural entries of odorant-binding proteins (OBPs; PDB ID: 1dqe, 2wcj, 2gte, 2erb, 3k1e, 3bfh, 1ow4), available then, were used for the construction of the structural alignment. The dataset used in this analysis is comprised of 116 conotoxin sequences[24] and 284 odorant binding proteins from mosquito genomes.[26] The conotoxins are classified into 7 classes. The odorant binding proteins are classified into 3 major classes including *Classic, PlusC* and *Atypical*; the *Atypical* are further divided into 4 subtypes (*MAtype1-4*). Representative sequences were chosen from the different classes for the construction of the training profile and the other sequences were used in the test set (Table 1).

Construction of profiles

A structural alignment constructed using COMPARER[27] was used as a profile for the functional

Table 1. Datasets used as training and test sets to build and assess scoring schemes for the identification of OBPs. (**A**) The OBP family dataset representing number of representative sequences used in constructing the profile (training dataset) and test set in the different classes respectively. (**B**) The conotoxin family dataset representing number of representative sequences used in constructing the profile (training dataset) and test set in the different classes respectively.

Protein subfamily	Training dataset	Test dataset
(A)		
Classic	18	104
Plus C	9	49
Minus C	18 (Classic OBPs)	17
Atypical 1	6	0
Atypical 2	6	26
Atypical 3	6	4
Atypical 4	6	33
(B)		
Class A	6	19
Class M	6	7
Class O	6	55
Class T	6	11

residue-based scoring scheme (Fig. 1). For the cysteine-based scoring scheme, representative sequences from each class, which have conserved cysteines at all the positions under consideration, were aligned separately using ClustalW.[28] This alignment of representative sequences was used as a training profile for the classification of query sequences. The number of sequences in the training profile and the number of cysteine positions under consideration vary for the different classes of the protein. Thus, a number of training profiles equal to the number of classes was generated.

Construction of fuzzy functional template

For the functional residue-based scoring scheme based on functional residues, a fuzzy functional template was constructed. Ligand binding residues, for each of the ligand-bound forms of each of the structural entries of OBPs mentioned above, were identified using LIGPLOT. These residues were mapped on the structural alignment (Fig. 1). 12 residue positions were considered as functionally important as marked in Figure 1. C^{β}–C^{β} distances between residues at these positions for each of the structural entries were calculated and averaged. The upper and lower limit for the distances were set to ±2 standard deviations (SD) from the average distance and represented in the form of a matrix (Fig. 2). This logic of inscribing distance variation amongst functionally important residues is the same as that adopted by Skolnick's group in an earlier study.[29]

For the serine protease family, 11 structural entries from the thrombin subfamily (1ai8, 1avg, 1hao, 1mkx, 1ucy, 2hpp, 3hk3, 3k65, 3nxp, 3pma, 3qlp), 15 structural entries from the trypsin family (1aoj, 1aks, 1an1, 1fxy, 1hj8, 1jrs, 1pq7, 2a31, 2eek, 2f91, 2ra3, 3beu, 3fp7, 3mi4, 3p95) and 4 structural entries from the plasminogen activator (1a5h, 1a5i, 1bqy, 1rtf) subfamily were used for the construction of the structural alignment. The functional positions were adopted in a similar manner to the functional sites described by Skolnick et al[29] 125 annotated query sequences from all 3 subfamilies (derived from SWISSPROT) were aligned to each of the subfamily profiles and the scores were checked for every query sequence against each profile.

Scoring of query sequences

Functional residue based scoring scheme

Different scoring functions were defined for scoring the conservation of residues in the functional positions based on their occurrence, probability of occurrence and by consulting the Dayhoff matrix.

Majority-based scheme: In this scheme, a score of 1 is given to a position in the query sequence if it has the amino acid which occurs majority of times at that position in the structural alignment (from known observations) and finally these scores are averaged for all the 12 positions.

Probability-based scheme: A score is given to each amino acid at a position in the query sequence equal in magnitude to its probability of occurring at that position. In one scheme (PROB_1), the scores are finally averaged for all the 12 positions, and in the second scheme (PROB_2), the sum of scores is divided by the sum of the maximum probabilities of occurrence each position.

Dayhoff matrix-based scheme: For each position in the query sequence, the score is calculated as the product of probability of each amino acid occurring at that position in the template and the Dayhoff Matrix score for the amino acid substitution from that AA to the residue present in the query. Finally, the scores are averaged for all the

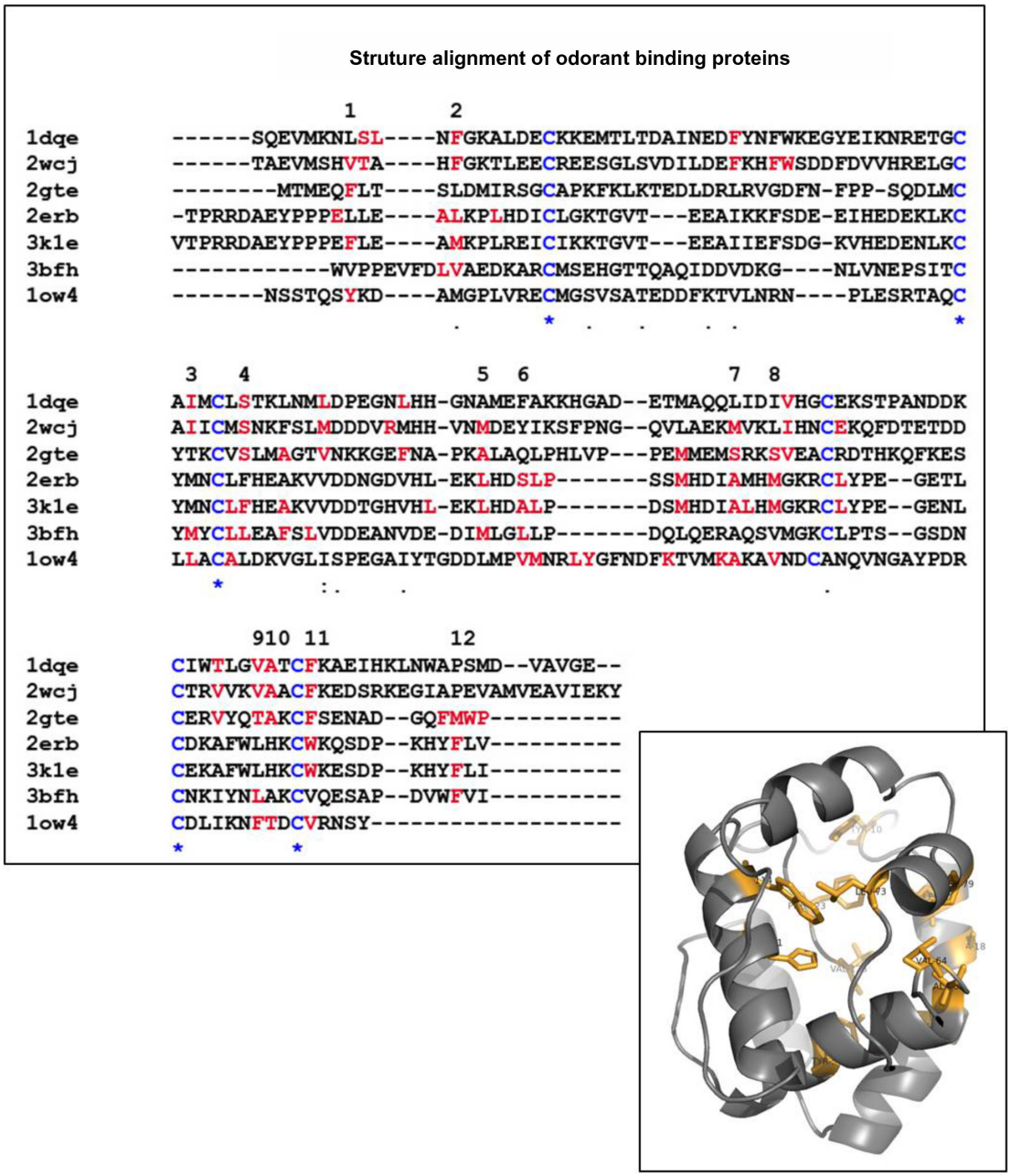

Figure 1. Alignment of available structures of odorant binding proteins using COMPARER.
Notes: The conserved cysteines are colored in blue ad functional residues are colored in red and the 12 positions used as functional sites for the scoring scheme are labeled respectively from 1–12 above the alignment. The functional residues are as shown on one example structure: 2erb.

12 positions. However, this matrix of amino acid exchanges are recorded and normalized as observed for large numbers of unrelated protein families and are also not position-specific in nature.

Given a query string Q with amino acid Qi at functional position i, where $0 \leq i \leq p$ and a training profile T which is an alignment with i functional positions.

The scores according to the different schemes are defined as follows:

Majority based score:

$$\frac{\sum_{i=1}^{p} Is\ Equal\,(P_i(Q_i), m_i)}{p} \times (Is\ Equal\,(P_i(Q_i), m_i) = 1 if\ P_i(Q_i) = m_i\ otherwise\ 0)$$

	6	10	48	52	69	72	85	88	109	110	113	122
6	0											
10	3.61–13.27	0										
48	13.4–19.88	7.25–23.77	0									
52	11.8–19.07	6.84–21.82	4.2–13.07	0								
69	8.43–20.81	8.41–24.37	11–19.13	5.7–14.49	0							
71	6.55–15.49	6.8–10.41	11–17.04	9–10.83	3.5–10.85	0						
85	8.66–17.62	12.3–19.2	13–25.44	9.7–22.68	4.8–16.64	6.88–18.3	0					
88	12.2–22.42	12.5–25.26	15–20.37	14–16.03	7.8–13.48	12.9–16.53	1.5–14.47	0				
109	14–24.37	11.3–26.84	6.4–18.3	6.6–14.24	9.1–12.57	13.9–17.69	9.1–22.86	7.4–14.85	0			
110	11.7–22.59	9.23–26.59	2.4–14.02	8–17.27	11–14.43	10.9–19.74	10–22.82	8.58–17.81	1.92–12.71	0		
113	12–21.81	11.3–23.82	10–18.69	11–14.35	9.6–13.59	13.5–16.82	4.7–16.89	4.87–5.7	3.83–11.93	3.05–15.98	0	
122	3.66–19.71	8.6–19.23	12–22.18	12–18.34	13–19.95	13.3–19.32	11–13.05	10.8–14.5	11.91–22.91	8.54–23.51	9.29–12.79	0

Figure 2. Fuzzy functional template investigated to score the dissimilarity between OBPs.
Notes: The matrix represents the distance criteria threshold between the 12 functional sites averaged over data from the available structural members. The distances between pairs which have an SD < 2 are colored yellow.

Probability_1 based score:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{n} M(T_{ij}, Q_i)}{p}$$

Probability_2 based score:

$$\frac{\sum_{i=1}^{p} P_i(Q_i)}{p}$$

Dayhoff Matrix based score:

$$\frac{\sum_{i=1}^{p} P_i(Q_i)}{\sum_{i=1}^{p} m_i}$$

where:
p = # of functional positions under consideration
n = # of sequences in the training profile (Structure alignment)
T_{ij} = Amino acid at position i in the sequence j of the training profile
Q_i = Amino acid at position i of the query sequence
m_i = Maximum probability of occurrence of any amino acid at position i
$M(A,B)$ = Entry in substitution matrix for amino acid A being substituted by B
$P_i(A)$ = Probability of amino acid A occurring at position i in the training profile.

Functional residue distance-based scoring scheme

C^{β}–C^{β} distances of the residues at the functional positions were calculated from the models of 131 classic OBP sequences (data not shown). The distances in the fuzzy functional template (FFT) residue pairs with SD < 2 were considered for the final scoring scheme. The query sequences were aligned to the structure alignment profile and the distances between residues corresponding to the functional position were calculated in their respective models. If the distance of the residue pairs fall within the upper and lower limits assigned for those residue pairs in FFT, a score of 1 was awarded (else score is 0) and averaged for the 12 functional positions.

Cysteine-based scoring scheme

Each query sequence was aligned separately with each of the training profiles using the sequence to profile alignment method in ClustalW[28] and checked for the conservation of cysteines. If a cysteine was found at a position, a score of '1' was given; otherwise a score of '0' was given. In this study, a cysteine in the query is assumed to be 'strictly conserved' if it aligns perfectly with the cysteine position in the training profile. However, according to the 'relaxed criterion', an arbitrary shift of 2 residues on either side of the cysteine positions in the training profile is allowed for uncertainties in the sequence alignment. In addition to the scores for cysteine conservation, an extra score of '1' is added for the conservation of each cysteine pair involved in disulphide bond formation. Such position-scores are normalized for all the positions within that class and an average score is obtained for each class for each query sequence (Supplementary Fig. 1). Thus, score of a query with the training profile of each class is a measure of its likelihood of belonging to that class.

Composite classification scheme

A composite classification scheme was devised for the classification of OBPs and conotoxins based on the scores for each class, the length of the query and the distance between the cysteines involved in disulphide formation (loop spacing; Supplementary Figs. 2 and 3). Thus, if it is an 'N'-class problem, then for each query, there will be 'N' score parameters (one for each class), a length parameter and a variable number of loop spacing (depending upon the classes). The loop spacing (number of amino acids along the sequence between the 2 cysteines involved in disulphide bonding) parameter would be extremely useful to distinguish between classes with the same cysteine motif but different disulphide connectivity patterns. This flexibility was introduced since it is expected that the loop spacing is more or less conserved throughout the members of a family, even if other inter-cysteine distances are not.

Re-substitution test of the cysteine based classification scheme

The re-substitution test is one of the important methods of evaluating predictive accuracy. In this test, the training set used to generate the classifier is itself used to test the classification model. In other words, the test set is the same as the training set. The re-substitution test is extremely important because it reflects the self-consistency of an identification scheme, and most importantly, the algorithm.

Results and Discussion

Functional sites and fuzzy functional template

Functional residues of proteins involved in ligand binding are generally conserved through the evolution of proteins and generally considered as good classifiers of protein families and for function annotation.[13] The ligand-binding residues from the bound complexes of the available PDB entries were mapped to the structural alignment generated by COMPARER.[27]

For the family of insect odorant binding proteins, the positions of the alignment, which had ligand-binding entries in at least 4 of the 7 PDB entries, were considered to be significant functional residue positions. 12 such positions were considered to be components of the functional template (Fig. 1). The C^{β}–C^{β} distance between these 12 residues were calculated and averaged in the form of a matrix called the 'fuzzy functional template' (FFT). The distance limits were set by indicating the average ±2 SDs, since the distances between the residues pairs were quite variable. The distances in the matrix that were less than 2 SDs from the mean were considered for the calculation of the scores. 12 such distances were identified involving 12 residue pairs in the matrix (Fig. 2). These distances were used for the scoring function.

Structure-based scoring scheme

The structure-based scoring scheme shows a good range of scores (0.3–1.0). However, there were low scoring sequences observed in the test cases. The scores were independent of the sequence identity to its template (Fig. 3). However, a limitation of this method is the fact that the test set consisted of models derived from members of the training set used as templates. This method could be applied only to proteins that have a structural entity or for query sequences for which a homology model could be derived, and thus the method was applied only on the classic odorant-binding proteins.

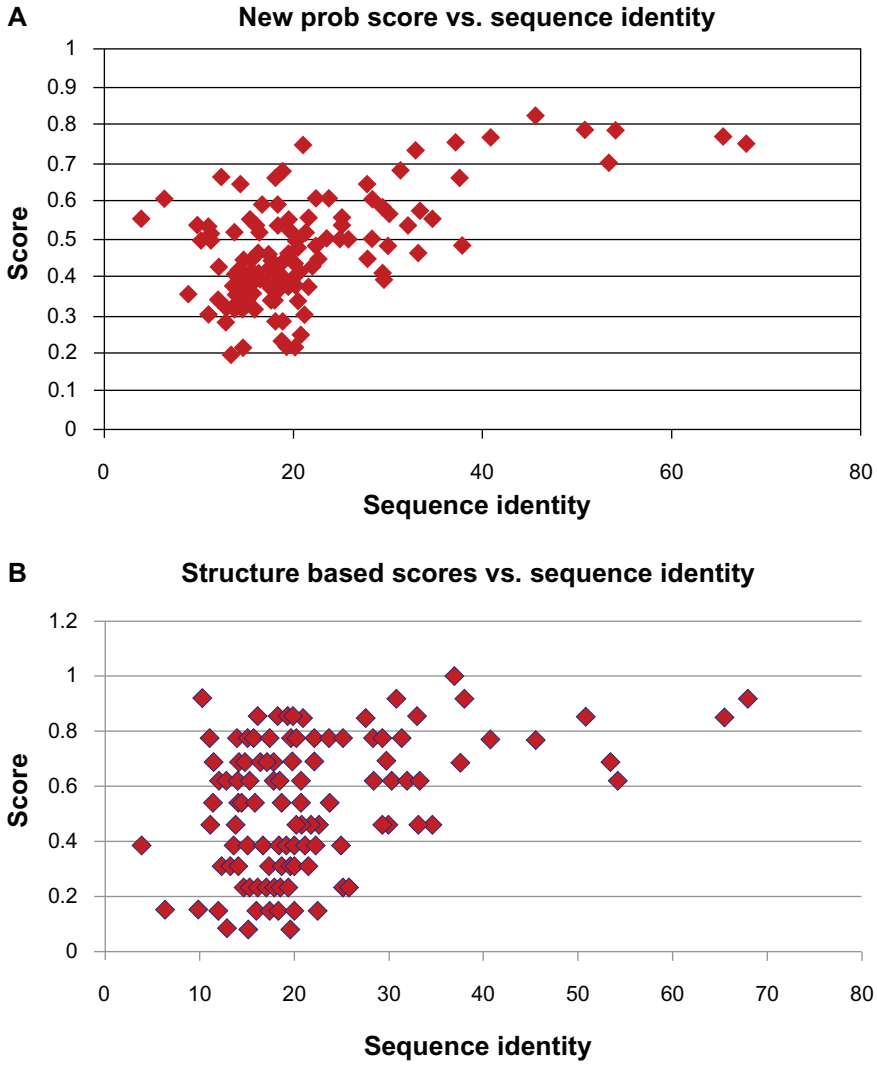

Figure 3. Scatter plot representing the effect of sequence identity on the sequence-based scores with sequence identity on the X-axis and scores on the Y-axis. (**A**) Effect of sequence identity on sequence-based scoring scheme. (**B**) Effect of sequence identity on structure based scoring scheme.

Functional residue-based scoring scheme

The 'PROB_2' scoring scheme, with the addition of homologues, achieves the best range and correlation. The scores were based on the occurrence, probability of occurrence and Dayhoff matrix as described in the Methods section. For the family of insect odorant-binding proteins, different training datasets were analyzed that include (1) a 7-member training set, which is the initial structure alignment, (2) a 25-member dataset where the 7-member dataset was populated (to include evolutionary data) with one additional close homologue from each of the mosquito genomes to every member in the 7-member dataset, (3) a 5-member dataset where the 2 mosquito crystal structures 2erb and 3k1e were removed to avoid potential bias in scoring the models (since these 2 structures served as templates for modeling) and (4) an 18-member dataset from which the 2 mosquito crystal structures and their homologues were excluded. The range of scores for each of the methods on every training set were analyzed and it was observed that the probability score PROB_2 achieved the best range, followed by the majority-based scores (Table 2A), and that they also achieved the best correlation when compared to other 2 methods (Table 2B). It was also observed that addition of homologues to the initial dataset significantly improved the range and correlation.

All 12 positions in the scoring scheme are equivalent in importance

It was important to analyze whether certain functional site positions contributed more to the scores in order

Table 2. Correlation and distribution of scores by the different schemes. (**A**) Distribution of the scores obtained from each of the different schemes based on each training set showing that the Prob_2 scheme achieves the highest range among the 4. (**B**) Correlation between the scores of the different schemes tested on various training sets showing that the Probability based scores have higher correlation with the other 2 types of scores.

	7 member training set	25 member training set	5 member training set	18 member training set
(A) Scoring scheme				
Majority	0.08–0.75	0.0–0.92	0.0–0.33	0.0–0.67
Prob_1	0.01–0.35	0.03–0.42	0.02–0.27	0.05–0.25
Prob_2	0.03–0.88	0.08–0.98	0.02–0.59	0.2–0.95
Dayhoff	0.3–0.75	0.3–0.54	0.33–0.44	0.28–0.41
(B) Score				
Prob. vs. Maj.	0.87	0.96	0.76	0.81
Day vs. Maj.	0.84	0.86	0.63	0.66
Day vs. Prob.	0.72	0.81	0.46	0.6

to provide different weights on the positions. This was done by jack-knifing each of the 12 individual positions and recalculating the scores for the initial 7-member dataset. The Pearson correlation coefficient between the scores were calculated after removing each of the 12 residue positions (Table 3) and it was observed that the removal of any one position from the scoring scheme did not significantly alter the scores.

The scores are independent of the sequence identity of the query sequence with the template

Since the scoring scheme is based on the probability of occurrence of an amino acid, the effect of sequence identify on the scores had to be considered carefully. A histogram of the number of sequences versus the sequence identity of the protein with the closest structural template in the dataset was plotted (Fig. 3). The distribution of the graph indicated that the scores are indeed independent of the sequence identity. A histogram of the number of sequences versus the percentage sequence identity of the query sequence with the template was plotted and the consistently high-scoring and low-scoring sequences were marked on it (Fig. 4). It was observed that the distribution of the low scoring and high scoring queries was independent of sequence identity.

Comparison of the sequence-based scoring scheme with sequence searches and phylogenetic analyses

We find that our simple sequence-based objective scoring scheme works better than domain-based subfamily association or phylogeny-based associations; for example, in the case of odorant binding proteins, which fall into three major subfamilies the *Classic PlusC* and *Atypical* as described earlier in the manuscript. When each of these members are searched against the conserved domain database it is observed that in many cases cross-talk is seen with respect to subfamily (Supplementary Table 1). For example, most of the Plus C Obps are never identified to carry the PBP_GOBP domain, and atypical OBPs,

Table 3. Pearson correlation co-efficient between the scores using all 12 functional positions and on jack-knifing each position from the 7-member dataset to analyze the contribution of individual function positions on the score.

Score	W/O 1	W/O 2	W/O 3	W/O 4	W/O 5	W/O 6	W/O 7	W/O 8	W/O 9	W/O 10	W/O 11	W/O 12
Maj.	0.95	0.96	0.98	0.99	0.97	0.95	0.95	0.96	0.98	0.99	0.97	0.95
Prob.	0.98	0.98	0.98	0.98	0.97	1.00	0.98	0.98	0.98	0.98	0.97	1.00
Day.	0.97	0.95	0.98	0.99	0.98	0.99	0.97	0.95	0.98	0.99	0.98	0.99

Note: All the scores are very similar after jack-knifing any of the positions, which leads to the conclusion that all the 12 positions in the profile are equivalent.

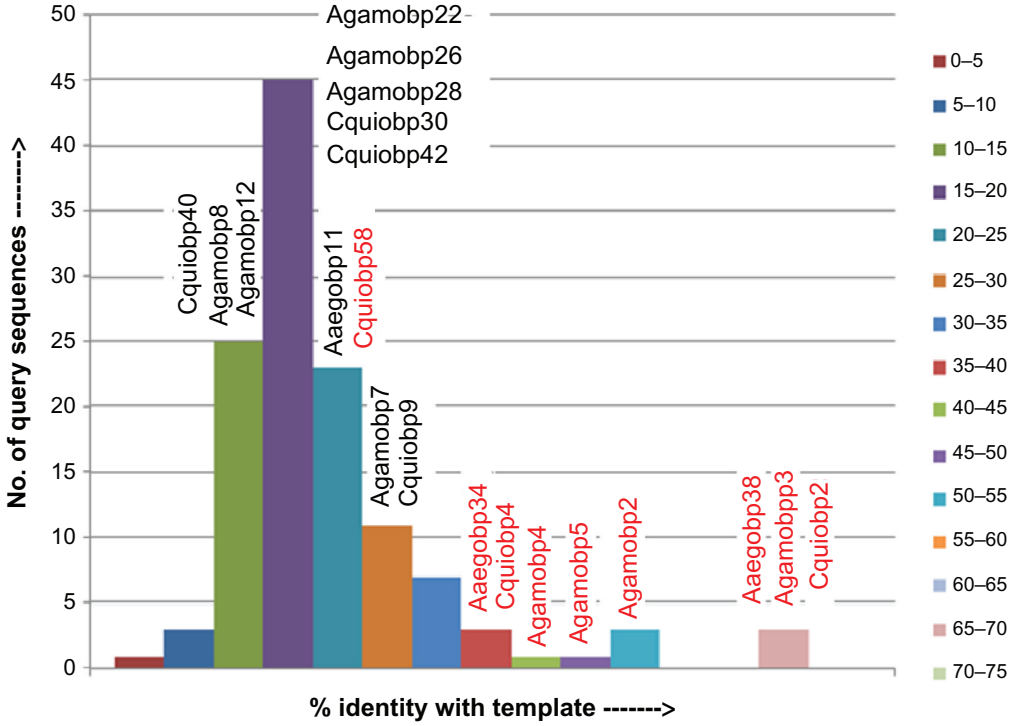

Figure 4. Histogram of the number of sequences versus the % identity of the query sequence with the template.
Note: The sequences labeled in red are high scoring while those labeled in black are low scoring.

which should be predicted to have two PBP_GOBP domains, are predicted to have only 1 PBP_GOBP domain. In contrast, the current method is able to exactly classify these proteins to their respective subfamilies.

It is also difficult to infer sequence associations from phylogenetic trees to provide a meaningful classification of the different subfamilies in the case of the odorant binding proteins. The phylogenetic trees were inferred separately for odorant binding proteins from each of the mosquito genomes using the neighbor-joining method in MEGA 4.0 26 (Supplementary Fig. 4A–C). In the phylogenetic trees of OBPs from *Anopheles gambiae*, *Aedes aegpti* and *Culex quinquifasciatus,* the different subfamilies were not clustered together with significant bootstrap support due to the high sequence divergence that is observed.

Application of sequence-based scoring scheme on serine protease subfamilies

Serine proteases are one of the largest groups of proteolytic enzymes with a nucleophilic serine residue at the active site and are believed to constitute nearly 1/3 of all the known proteolytic enzymes. They include exopeptidases and endopeptidases belonging to different protein families grouped into clans. They function as part of diverse biological processes such as digestion, blood clotting, fertilization, development, complement activation, pathogenesis, apoptosis, immune response, secondary metabolism, with imbalances causing diseases like arthritis and tumors. The current method was applied to 3 families to see if the method can classify the sequences into these 3 subfamilies: Trypsin, Thrombin and Plasminogen Activator. The method was tested on 125 serine protease sequences from the three subfamilies (Supplementary Table 2). It was observed that the method could classify the proteins into their respective subfamilies effectively.

Cysteine-based scoring scheme

Cysteine positions in protein sequences are the other evolutionarily conserved sites in disulphide-rich protein families. They can be used as effective regular expressions in protein sequences, even among distantly-related proteins, whose classification based on other methods would be quite challenging.

However, a sequence-to-sequence alignment algorithm, using one representative sequence for a family, would not provide sufficient accuracy in terms of accounting for the insertions and deletions observed in diverse sequences. A disulphide profile, derived from representative sequences, is more suitable for compensating the occurrences of insertions and deletions.[18] The cysteine-based scoring scheme was found to be a more direct way for the identification of OBPs in insects and was used previously in the use of identification of OBPs.[4] In this work, however, the scheme has been further extended to classify the OBPs in the mosquito genome. Hence, practically, the algorithm not only predicts the chance of a query sequence to be a putative OBP protein, but also facilities its classification into 1 of the different classes of OBPs that are described below. The OBPs are classified into 4 major classes (i) *Classic*, which carry 6 conserved cysteine motifs, (ii) *PlusC* OBPs, which carry an additional 3 conserved cysteines, (iii) *Dimer* OBPs or *Atypical* OBPs, which carry 2 Classic OBP domains and hence 12 conserved cysteines and (iv) *Minus-C* OBPs, which lack 2 Cys residues in comparison with *Classic* OBPs. The dimer OBPs can be further classified as *MAtype1-4*; all of them hold 12 conserved cysteines except *MAtype2*. From the alignments used in the construction of phylogenetic trees, it was observed that the cysteine conservation patterns and spacing could play an important role in the classification of OBPs. This was analyzed by observing the cysteine conservation patterns of sequences in the test datasets when aligned to profiles that were constructed using a training set of each of the classes described above.

A training set for the 7 different classes of OBPs (disulphide profiles) was prepared (as summarized in Table 1A). A representative sequence was identified from a phylogeny of odorant binding proteins of each class. For the *Minus-C* class, the same profile for Classic OBPs was used, but only the 1st, 3rd, 4th and 6th cysteine positions were considered. A composite classification scheme was devised for the family of OBPs incorporating the 7 different scores and the length of sequence as attributes. The protocol was applied to a dataset of 284 mosquito OBP sequences (from *Anopheles gambiae, Aedes aegypti* and *Culex quinquefasciatus*) and the class predictions were compared with the predictions of class association independently made from phylogenetic analysis. The 'confusion matrix' of the classes predicted by the cysteine based classification scheme versus the phylogeny-based classification is given in Figure 5A. The scheme provides an accuracy of 90.14% when compared with the phylogeny-based classification for the test set sequences. The effect of different classes to this was tested using a re-substitution test.

The re-substitution test on the training set gave accuracies of 100%, 100%, 0%, 100%, 66.66% and 100% for *Classic, PlusC, Atypical1, Atypical2, Atypical3 and Atypical4* classes, respectively. The sequences in *Atypical1*, however, form a small group of 6 sequences and do not follow a strict conservation of cysteines as the other classes of OBPs. Hence it was difficult to classify these members by our scheme

A

	C	P	M	A1	A2	A3	A4
C	97	1	3	0	1	0	2
P	3	45	0	0	1	0	0
M	2	0	15	0	0	0	0
A1	0	0	0	0	0	0	0
A2	0	1	2	0	21	2	0
A3	0	0	0	0	1	0	0
A4	0	0	4	1	3	3	25

B

	A	M	O	T
A	18	1	0	3
M	0	7	0	0
O	0	1	53	1
T	2	0	0	9

Figure 5. Confusion Matrix of Classification. (**A**) Confusion matrix between the phylogeny based classification of odorant binding proteins and the cysteine scoring based classification scheme. (**B**) Confusion matrix between the classification of conotoxins and the cysteine scoring based classification scheme.

explaining the poor performance of the re-substitution test for the *Atypical1* class.

Application of cysteine-based scoring schemes on well-known superfamily of conotoxins

Since the accuracy of the classification scheme needed further convincing, the algorithm was extended to the well-known cysteine-rich superfamily of conotoxins. Conotoxins are small neurotoxic peptides found in the venom of the predatory cone snails of the genus *Conus* that act primarily by modulating the activity of specific ion channels. The mature conotoxins are characterized by the presence of multiple disulphide bonds and have been classified into 7 families including A, M, O, I, P, T and S, again on the basis of a highly conserved N-terminal precursor sequence, disulphide connectivity and mode of action.[24] Each family is characterized by the presence of 1 or 2 characteristic patterns of disulphide cross-links.[30] The prominent disulphide connectivity patterns in the 4 major families of conotoxins are shown in Supplementary Figure 5, and these alone were used for scoring purposes.

A classification scheme was developed for conotoxins as shown in Supplementary Figure 3, incorporating the 4 scores corresponding to each of the 4 major families. The classifier (constructed using the training set as shown in Table 2) was tested on a dataset of 116 conotoxin sequences obtained from Mondal et al[24] and the predictions made by the scheme were compared with the known classes of the sequences in the study by Mondal et al.[24] The scheme gave an accuracy of 93.1% for the test set and the confusion matrix is presented in Figure 5B. The re-substitution test on the training set provided an accuracy of 100% for all 4 families.

Conclusion

Simple domain-finding techniques such as association to Pfam families, can be helpful only to relate mosquito OBPs to the broad family of 'odorant binding proteins' (PF01395), but cannot be distinguished as *Classic, PlusC* and *Atypical* odorant binding proteins. These subfamilies differ in their sequence features, even though they carry the basic PBP/GOBP domain. In the case of families where the sequence divergence in very high, it is important that family-specific classification methods are derived to obtain a more meaningful functional classification of the family. Evolutionarily-constrained functional and structural entities/signatures, combined with family-specific profile-based scoring, improve the function annotation quality and can also be further extended to a subfamily level classification. Fuzzy functional template-based objective methods, encoded in our structure-based scoring scheme, provide a clear representation of the extent of spatial preservation of known functionally important residues. Such scoring schemes provide an early indication of family members with deviations from the parent family in biological function or the lack of function. Such structure-based scoring schemes could be convenient to rapidly validate a large number of gene products whose high-quality homology models can be automatically obtained.

Most popular function prediction methods reported in the literature require structural information or models of query sequences for scoring and recognizing functionally important residues which are only applicable for SGI targets or those sequences where homology models can be obtained reliably. In our approach, there is a novel option to employ only sequence information to score the preservation of functionally important residues. Our pure sequence-based approach is different from other methods that use sequence alignments (like the functional-residue-clustering (FRC) method)[31] that lead to abstract data by defining amino acid alphabets and require a joint alignment including subfamily members.

The above-described algorithms are shown to work efficiently for protein families such as odorant-binding proteins, serine proteases and conotoxins. We demonstrate that it is possible to apply this approach using large-scale annotation and classification by applying it to new odorant-binding proteins, which are indeed a diverse family of proteins and pose a lot of challenges for regular identification and classification algorithms.[32] This could be extended to other diverse families of proteins. However, an in-depth analysis of every superfamily for family-specific signatures and the construction of a composite classification scheme at the subfamily level is required.

Acknowledgements

Malini Manoharan was supported by an international PhD fellowship from Conseil Regional de La Reunion in the framework of the joint dual-

studentship program between Manipal University and University of La Réunion. This work was in part supported by a grant from Conseil Régional de La Réunion, French Ministry of Research and European Union in the framework of the GRI Phase III project. SK thanks BITS Pilani Practice School and RS thanks Universite de La ReUnion for a Visiting Professorship. We thank NCBS (TIFR) for infrastructural facilities.

Funding

We would like to thank Conseil Régional de La Réunion, French Ministry of Research and European Union for the funding and fellowship.

Author Contributions

Conceived and designed the experiments: SR, MM, KS, BO. Analyzed the data: KS. MM. Wrote the first draft of the manuscript: MM. Contributed to the writing of the manuscript: SR, KS, BO. Agree with manuscript results and conclusions: SR, MM, KS, BO. Jointly developed the structure and arguments for the paper: SR, MM. Made critical revisions and approved final version: SR. All authors reviewed and approved of the final manuscript.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Pelosi P, Maida R. Odorant-binding proteins in insects. *Comp Biochem Physiol B Biochem Mol Biol*. 1995;111(3):503–14.
2. Vogt RG, Callahan FE, Rogers ME, Dickens JC. Odorant binding protein diversity and distribution among the insect orders, as indicated by LAP, an OBP-related protein of the true bug Lygus lineolaris (Hemiptera, Heteroptera). *Chemical Senses*. 1999;24(5):481–95.
3. Hekmat-Scafe DS, Scafe CR, McKinney AJ, Tanouye MA. Genome-wide analysis of the odorant-binding protein gene family in Drosophila melanogaster. *Genome Res*. 2002;12(9):1357–69.
4. Zhou JJ, Huang W, Zhang GA, Pickett JA, Field LM. "Plus-C" odorant-binding protein genes in two Drosophila species and the malaria mosquito Anopheles gambiae. *Gene*. 2004;327(1):117–29.
5. Xu PX, Zwiebel LJ, Smith DP. Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, Anopheles gambiae. *Insect Mol Biol*. 2003;12(6):549–60.
6. Zhou JJ, He XL, Pickett JA, Field LM. Identification of odorant-binding proteins of the yellow fever mosquito Aedes aegypti: genome annotation and comparative analyses. *Insect Mol Biol*. 2008;17:147–63.
7. Pelletier J, Leal WS. Genome analysis and expression patterns of odorant-binding proteins from the Southern House mosquito Culex pipiens quinquefasciatus. *PloS One*. 2009;4:e6237–7.
8. Butte AJ. Challenges in bioinformatics: infrastructure, models and analytics. *Trends Biotechnol*. 2001;19(5):159–60.
9. Chothia C. One thousand families for the molecular biologist. *Nature*. 1992; 357:543–4.
10. Kraut J. Serine proteases: structure and mechanism of catalysis. *Annu Rev Biochem*. 1977;46:331–58.
11. Miller J, McLachlan AD, Klug A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J*. 1985;4(6):1609–14.
12. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB. Classification schemes for protein structure and function. *Nat Rev Genet*. 2003;4(7):508–19.
13. Innis CA, Anand AP, Sowdhamini R. Prediction of functional sites in proteins using conserved functional group analysis. *J Mol Biol*. 2004;337(4): 1053–68.
14. Wangiker PP, Tendulkar AV, Ramya S, Deepali MN, Sarawagi S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*. 2003;326(3):955–78.
15. Thornton JM. Disulphide bridges in globular proteins. *J Mol Biol*. 1981;151(2):261–87.
16. Richardson JS. The anatomy and taxonomy of protein structure. *Adv Protein Chem*. 1981;34:167–339.
17. Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P. Conformations of disulfide bridges in proteins. *Int J Pept Protein Res*. 1990;36(2):147–55.
18. Johnson MS, Overington JP. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol*. 1993;233(4):716–38.
19. Thangudu R, Manoharan M, Srinivasan N, Cadet F, Sowdhamini R, Offmann B. Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families. *BMC Struct Biol*. 2008;8:55.
20. Mas JM, Aloy P, Marti-Renom MA, et al. Classification of protein disulphide-bridge topologies. *J Comput Aided Mol Des*. 2001;15:477–87.
21. Benham CJ, Jafri MS. Disulfide bonding patterns and protein topologies. *Protein Sci*. 1993;2(1):41–54.
22. Harrison PM, Sternberg MJ. The disulphide beta-cross: From cystine geometry and clustering to classification of small disulphide-rich protein folds. *J Mol Biol*. 1996;264(3):603–23.
23. Lenffer J, Lai P, El Mejaber W, et al. CysView: protein classification based on cysteine pairing patterns. *Nucleic Acids Res*. 2004;32(Web Server issue): W350–5.
24. Mondal S, Bhavna R, Mohan Babu R, Ramakumar S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J Theor Biol*. 2006;243(2):252–60.
25. Chon JK, Choi J, Kim SS, Shin W. Classification of peroxiredoxin subfamilies using regular expressions. *Genomics Informatics*. 2005;3:55–60.
26. Manoharan M, Ng Fuk Chong M, Vaïtinadapoule A, Frumence E, Sowdhamini R, Offmann B. Comparative genomics of odorant binding proteins in Anopheles gambiae, Aedes aegypti and Culex quinquefasciatus. *Genome Biol Evol*. 2013;5(1):163–80.
27. Sali A, Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*. 1990;212(2):403–28.

28. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997;25(24): 4876–82.
29. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol*. 1998;281(5):949–68.
30. Olivera BM. Conus venom peptides: Reflections from the biology of clades and species. *Annu Rev Ecol Syst*. 2002;33:25–47.
31. Shah PK, Tripathi LP, Jensen LJ, et al. Enhanced function annotations for Drosophila serine proteases: a case study for systematic annotation of multi-member gene families. *Gene*. 2008;407(1–2):199–215.
32. Manoharan M. Genomic, structural and functional characterization of odorant binding proteins in olfaction of mosquitoes involved in infectious disease transmission. Ph.D. Thesis, Manipal University and Universite de la ReUnion. 2011.

Supplementary data

Figure S1. Schematic representation of the cysteine based scoring scheme.

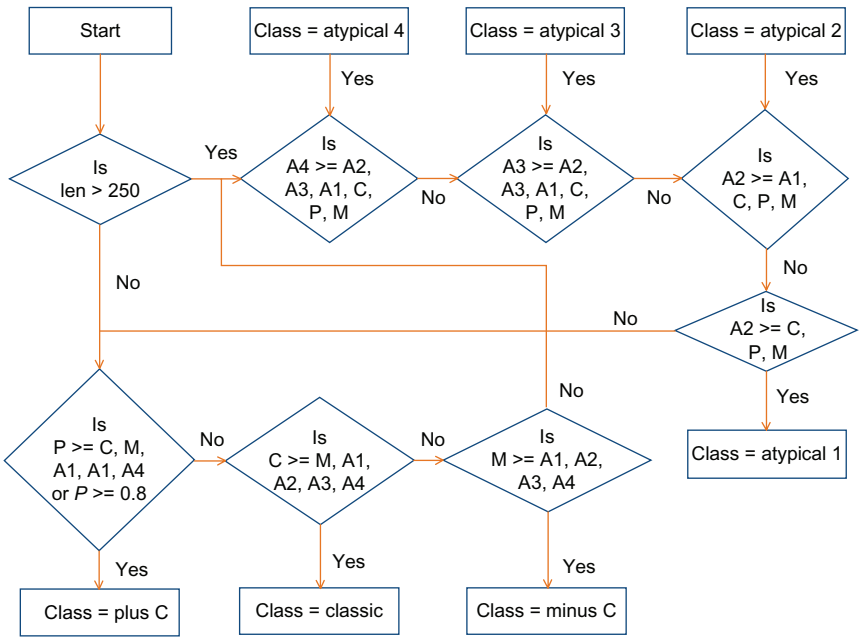

Figure S2. Flowchart of the logistics used in the composite classification scheme of OBPs.

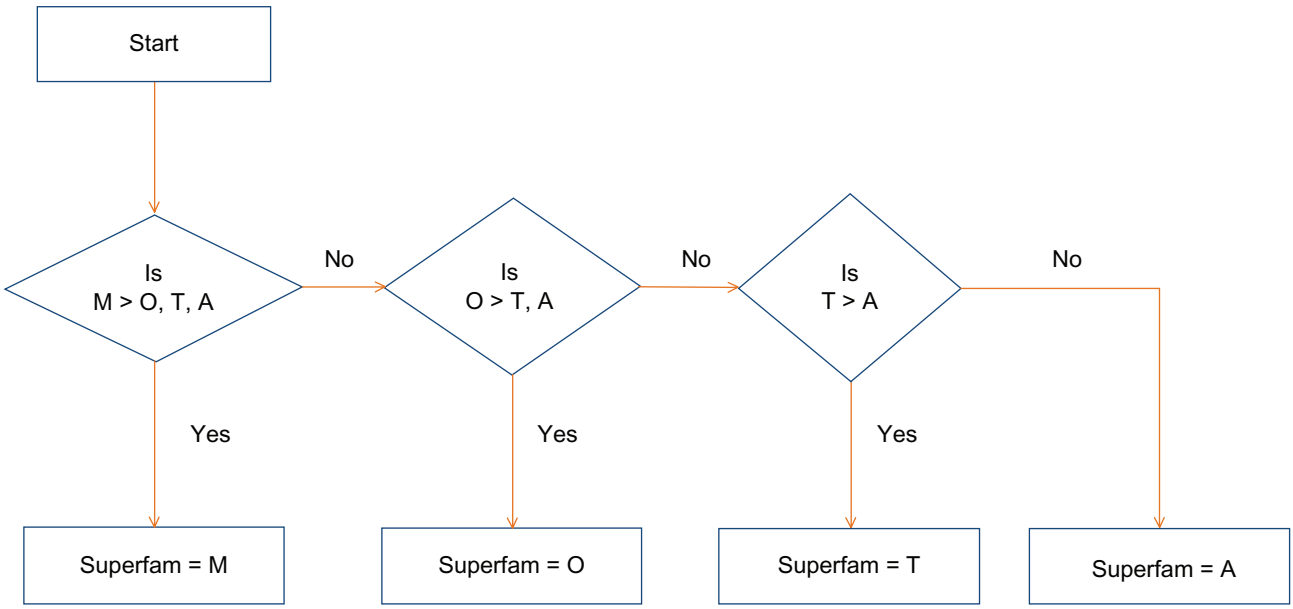

Figure S3. Flowchart of the logistics used in the composite classification scheme of the conotoxin family.

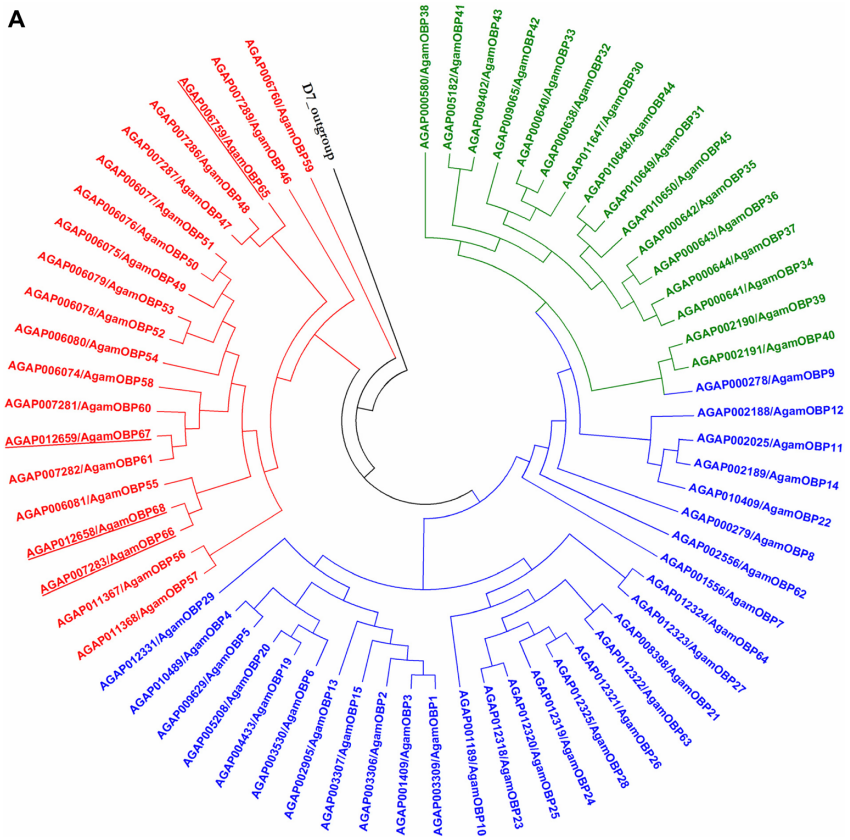

Figure S4. (*Continued*)

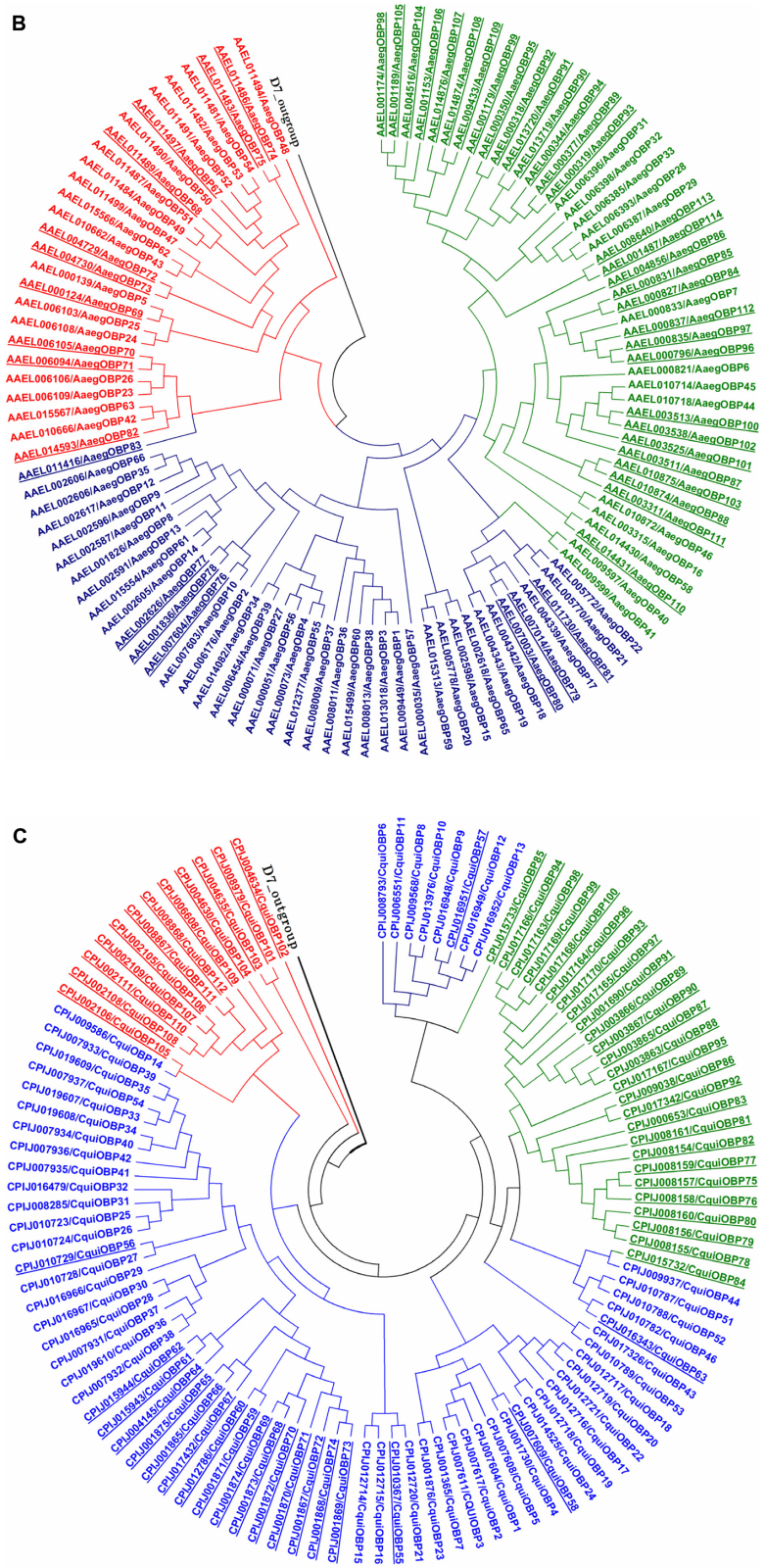

Figure S4. (**A**) Rooted phylogenetic tree of the odorant binding proteins in the *Anopheles gambiae* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red. (**B**) Rooted phylogenetic tree of the odorant binding proteins in the *Aedes aegypti* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red. (**C**) Rooted phylogenetic tree of the odorant binding proteins in the *Culex quinquifasciatus* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red.

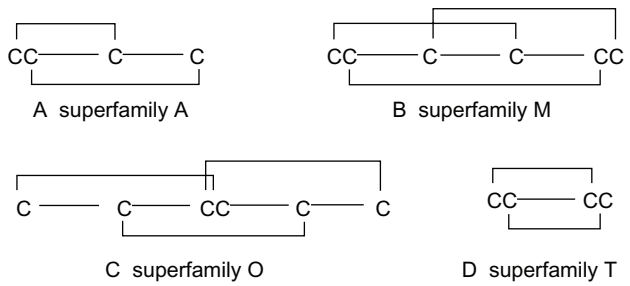

Figure S5. Cysteine connectivity patterns in the four major superfamilies of conotoxins, namely superfamily A (**A**), superfamily M (**B**), superfamily O (**C**) and superfamily T (**D**).

Table S1. Sequences mispredicted by domain based methods and correctly predicted by the current method.

ID	CD-search	Current method
AAEL000139	No family	PlusC subfamily
AAEL006109	No family	PlusC subfamily
AAEL006108	No family	PlusC subfamily
AAEL006103	No family	PlusC subfamily
AAEL010666	No family	PlusC subfamily
AAEL010662	No family	PlusC subfamily
AAEL011494	No family	PlusC subfamily
AAEL011499	No family	PlusC subfamily
AAEL011484	No family	PlusC subfamily
AAEL011490	No family	PlusC subfamily
AAEL011487	No family	PlusC subfamily
AAEL011491	No family	PlusC subfamily
AAEL011482	No family	PlusC subfamily
AAEL011481	No family	PlusC subfamily
AAEL015566	No family	PlusC subfamily
AAEL015567	No family	PlusC subfamily
AAEL011497	No family	PlusC subfamily
AAEL011489	No family	PlusC subfamily
AAEL006105	No family	PlusC subfamily
AAEL006904	No family	PlusC subfamily
AAEL004729	No family	PlusC subfamily
AAEL004730	No family	PlusC subfamily
AAEL011486	No family	PlusC subfamily
AAEL011483	No family	PlusC subfamily
AAEL014593	No family	PlusC subfamily
AAEL000139	Classic	Atypical
AGAP007287	No family	PlusC subfamily
AGAP006065	No family	PlusC subfamily
AGAP006076	No family	PlusC subfamily
AGAP006077	No family	PlusC subfamily
AGAP006078	No family	PlusC subfamily
AGAP006079	No family	PlusC subfamily
AGAP006080	No family	PlusC subfamily
AGAP006081	No family	PlusC subfamily
AGAP011367	No family	PlusC subfamily
AGAP011368	No family	PlusC subfamily
AGAP006074	No family	PlusC subfamily
AGAP006760	No family	PlusC subfamily
AGAP007281	No family	PlusC subfamily
AGAP007282	No family	PlusC subfamily
AGAP006759	No family	PlusC subfamily
AGAP007283	No family	PlusC subfamily
AGAP012659	No family	PlusC subfamily
AGAP008793	No family	Classic
AGAP008979	No family	PlusC subfamily
CPIJ004634	No family	PlusC subfamily
CPIJ004635	No family	PlusC subfamily
CPIJ004630	No family	PlusC subfamily
CPIJ002105	No family	PlusC subfamily
CPIJ002109	No family	PlusC subfamily
CPIJ002108	No family	PlusC subfamily
CPIJ006608	No family	PlusC subfamily
CPIJ002111	No family	PlusC subfamily
CPIJ008867	No family	PlusC subfamily
CPIJ008868	No family	PlusC subfamily
CPIJ017524	No family	PlusC subfamily
CPIJ007337	No family	PlusC subfamily
CPIJ017168	Classic	Atypical
CPIJ017169	Classic	Atypical
CPIJ017163	Classic	Atypical
CPIJ017165	Classic	Atypical
CPIJ017164	Classic	Atypical
CPIJ017166	Classic	Atypical
CPIJ017170	Classic	Atypical
CPIJ001690	Classic	Atypical
CPIJ003867	Classic	Atypical
CPIJ003863	Classic	Atypical
CPIJ003865	Classic	Atypical
CPIJ000653	Classic	Atypical
CPIJ008154	Classic	Atypical
CPIJ008160	Classic	Atypical
AGAP000641	Classic	Atypical
AGAP000642	Classic	Atypical
AGAP000643	Classic	Atypical
AGAP000644	Classic	Atypical
AGAP011647	Classic	Atypical
AAEL001487	No family	PlusC subfamily
AAEL000837	Classic	Atypical
AAEL001153	Classic	Atypical
AAEL001189	Classic	Atypical
AAEL004516	Classic	Atypical
AAEL010875	Classic	Atypical

Table S2. Scores for the different subfamilies of serine proteases obtained from the sequence based scoring schemes.

Query	Trained using trypin dataset			Trained using Thrombin dataset			Trained using Plasminogen dataset		
	Majority	Probability	Dayhoff	Majority	Probability	Dayhoff	Majority	Probability	Dayhoff
PLAS_P00750	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
PLAS_P11214	0.571429	0.618868	0.60017	0.714286	0.717489	0.640649	0.857143	0.88	0.662976
PLAS_P15638	0.619048	0.671698	0.617857	0.666667	0.668161	0.614502	0.857143	0.906667	0.68131
PLAS_P19637	0.571429	0.618868	0.60017	0.666667	0.668161	0.629697	0.857143	0.88	0.666786
PLAS_P49150	0.619048	0.671698	0.617857	0.666667	0.668161	0.614502	0.857143	0.906667	0.68131
PLAS_P98119	0.619048	0.671698	0.617857	0.666667	0.668161	0.614502	0.857143	0.906667	0.68131
PLAS_P98121	0.619048	0.671698	0.617857	0.666667	0.668161	0.614502	0.857143	0.906667	0.68131
PLAS_Q28198	0.571429	0.618868	0.597755	0.666667	0.668161	0.618745	0.952381	0.946667	0.688571
PLAS_Q5R8J0	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
PLAS_Q8SQ23	0.619048	0.671698	0.614932	0.666667	0.668161	0.607792	0.904762	0.92	0.683095
PLAS_B4DN26	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
PLAS_B4DNJ1	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
PLAS_B4DRD3	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
PLAS_B4DV92	0.571429	0.618868	0.60017	0.666667	0.668161	0.620649	0.952381	0.946667	0.688571
THR_119760	0.666667	0.732075	0.614456	0.714286	0.70852	0.619784	0.666667	0.706667	0.591548
THR_122144690	0.619048	0.679245	0.599388	1	1	0.715628	0.619048	0.68	0.574524
THR_135806	0.619048	0.679245	0.58966	1	1	0.715628	0.619048	0.68	0.574524
THR_135807	0.619048	0.679245	0.600102	1	1	0.715628	0.619048	0.68	0.574524
THR_135808	0.619048	0.679245	0.597993	1	1	0.715628	0.619048	0.68	0.574524
THR_135809	0.619048	0.679245	0.597993	1	1	0.715628	0.619048	0.68	0.574524
THR_338817876	0.619048	0.671698	0.602007	0.761905	0.753363	0.651818	0.619048	0.666667	0.584881
THR_48427854	0.666667	0.728302	0.631701	0.714286	0.70852	0.61645	0.666667	0.706667	0.579762
THR_51701719	0.809524	0.849057	0.661939	0.619048	0.623318	0.59342	0.619048	0.64	0.589405
THR_51704215	0.761905	0.807547	0.650544	0.571429	0.573991	0.560563	0.619048	0.64	0.576071
THR_62511155	0.619048	0.679245	0.600102	1	1	0.715628	0.619048	0.68	0.574524
TRY_A1L3H8	0.761905	0.807547	0.650986	0.714286	0.717489	0.611948	0.619048	0.666667	0.594643
TRY_A1Z7M7	0.714286	0.735849	0.591871	0.571429	0.58296	0.525931	0.52381	0.573333	0.530238
TRY_A1Z8J7	0.571429	0.607547	0.558707	0.52381	0.524664	0.544978	0.47619	0.52	0.527143
TRY_A5CG75	0.619048	0.675472	0.602755	0.666667	0.668161	0.589913	0.619048	0.68	0.584286
TRY_A7UNZ4	0.619048	0.671698	0.56	0.571429	0.578475	0.56355	0.619048	0.68	0.60631
TRY_A9WQU1	0.428571	0.471698	0.428231	0.428571	0.426009	0.413636	0.428571	0.48	0.415357
TRY_A9WQW1	0.428571	0.45283	0.437687	0.333333	0.336323	0.3829	0.428571	0.453333	0.436667
TRY_B5DZ08	0.619048	0.656604	0.617857	0.619048	0.618834	0.583506	0.571429	0.613333	0.597381
TRY_B7P4W6	0.333333	0.343396	0.382619	0.095238	0.09417	0.342641	0.095238	0.093333	0.344286
TRY_B7P5I0	0.238095	0.25283	0.320986	0.095238	0.121076	0.247273	0.142857	0.146667	0.219167
TRY_B7P8G5	0.666667	0.720755	0.617857	0.666667	0.668161	0.609437	0.571429	0.64	0.575952
TRY_B7P9I9	0.571429	0.603774	0.573197	0.52381	0.533632	0.564892	0.428571	0.466667	0.52131
TRY_B7PAU0	0.619048	0.649057	0.593197	0.571429	0.58296	0.573117	0.428571	0.48	0.538571
TRY_B7PC21	0.666667	0.716981	0.629286	0.666667	0.668161	0.589177	0.714286	0.746667	0.629881
TRY_B7PDD5	0.380952	0.411321	0.521803	0.380952	0.38565	0.496407	0.333333	0.36	0.509286
TRY_B7PF42	0.571429	0.618868	0.557993	0.52381	0.529148	0.55671	0.428571	0.52	0.535595

(*Continued*)

Table S2. (*Continued*)

Query	Trained using trypin dataset			Trained using Thrombin dataset			Trained using Plasminogen dataset		
	Majority	Probability	Dayhoff	Majority	Probability	Dayhoff	Majority	Probability	Dayhoff
TRY_B7PF43	0.380952	0.392453	0.384524	0.380952	0.381166	0.399221	0.285714	0.32	0.35131
TRY_B7PFF7	0.571429	0.615094	0.578605	0.571429	0.573991	0.561255	0.571429	0.64	0.586786
TRY_B7PKT9	0.714286	0.743396	0.660102	0.666667	0.659193	0.61013	0.619048	0.666667	0.597381
TRY_B7PS16	0.380952	0.411321	0.425068	0.380952	0.381166	0.470043	0.333333	0.36	0.383571
TRY_B7PTD2	0.52381	0.573585	0.461054	0.571429	0.573991	0.488528	0.428571	0.48	0.454762
TRY_B7Q2U2	0.571429	0.618868	0.458503	0.571429	0.565022	0.479697	0.47619	0.52	0.480357
TRY_B7Q613	0.619048	0.660377	0.585	0.619048	0.623318	0.578398	0.571429	0.626667	0.56619
TRY_B7QBB9	0.52381	0.569811	0.542619	0.47619	0.475336	0.499351	0.52381	0.573333	0.536905
TRY_B7QCX1	0.47619	0.520755	0.43602	0.571429	0.573991	0.540173	0.47619	0.506667	0.432619
TRY_B7QGT2	0.52381	0.566038	0.479116	0.52381	0.524664	0.437359	0.428571	0.453333	0.38
TRY_B7QH62	0.285714	0.301887	0.351395	0.238095	0.264574	0.288182	0.095238	0.106667	0.189286
TRY_B7QKP8	0.571429	0.618868	0.549796	0.666667	0.668161	0.606234	0.619048	0.68	0.584048
TRY_B7QLM5	0.666667	0.716981	0.593503	0.571429	0.578475	0.551732	0.47619	0.533333	0.506905
TRY_B7QNG9	0.428571	0.441509	0.49085	0.380952	0.394619	0.465022	0.333333	0.373333	0.465357
TRY_C6WB29	0.619048	0.671698	0.61102	0.619048	0.61435	0.567792	0.619048	0.653333	0.577857
TRY_D0D5G3	0.666667	0.720755	0.617687	0.666667	0.668161	0.612857	0.619048	0.666667	0.594524
TRY_D2Y5C3	0.666667	0.69434	0.611837	0.47619	0.475336	0.51052	0.619048	0.613333	0.560357
TRY_D2YGB8	0.714286	0.762264	0.634558	0.666667	0.672646	0.608095	0.571429	0.626667	0.590238
TRY_D3PK18	0.857143	0.890566	0.677585	0.619048	0.623318	0.593939	0.619048	0.68	0.591071
TRY_E0V917	0.52381	0.569811	0.55398	0.47619	0.484305	0.554372	0.571429	0.626667	0.574167
TRY_E0V971	0.428571	0.467925	0.549898	0.380952	0.38565	0.506494	0.428571	0.466667	0.5125
TRY_E0VCQ1	0.619048	0.667925	0.535884	0.571429	0.578475	0.555758	0.52381	0.573333	0.53119
TRY_E0VDU6	0.571429	0.6	0.589388	0.571429	0.573991	0.565195	0.47619	0.546667	0.586429
TRY_E0VFA8	0.666667	0.713208	0.604558	0.571429	0.578475	0.580216	0.619048	0.64	0.588095
TRY_E0VFA9	0.619048	0.660377	0.587483	0.571429	0.573991	0.595195	0.571429	0.613333	0.5875
TRY_E0VJE5	0.761905	0.796226	0.661701	0.619048	0.623318	0.585065	0.571429	0.626667	0.58619
TRY_E0VN67	0.666667	0.709434	0.610238	0.666667	0.668161	0.583723	0.619048	0.666667	0.589762
TRY_E0VPJ3	0.714286	0.754717	0.635918	0.571429	0.573991	0.563853	0.571429	0.613333	0.595
TRY_E0VQ98	0.619048	0.656604	0.611803	0.619048	0.618834	0.58303	0.571429	0.613333	0.600833
TRY_E0VQA9	0.571429	0.615094	0.550374	0.52381	0.529148	0.563463	0.428571	0.48	0.542738
TRY_E0VQK2	0.666667	0.713208	0.611599	0.666667	0.672646	0.584502	0.52381	0.586667	0.576548
TRY_E0VW09	0.714286	0.769811	0.645068	0.714286	0.70852	0.61645	0.714286	0.76	0.631905
TRY_E0VW10	0.666667	0.724528	0.625238	0.666667	0.668161	0.622251	0.666667	0.72	0.6125
TRY_E0VW14	0.761905	0.811321	0.652109	0.714286	0.713004	0.622078	0.619048	0.68	0.603929
TRY_E0W074	0.571429	0.603774	0.543469	0.52381	0.529148	0.532078	0.47619	0.52	0.529881
TRY_E0W1D0	0.285714	0.309434	0.450374	0.190476	0.192825	0.359567	0.142857	0.173333	0.431429
TRY_E3X3A6	0.428571	0.45283	0.549252	0.47619	0.484305	0.512208	0.428571	0.466667	0.532381
TRY_E8UA10	0.571429	0.618868	0.567721	0.571429	0.578475	0.550433	0.47619	0.533333	0.54881
TRY_E9GB32	0.714286	0.762264	0.630034	0.666667	0.668161	0.602035	0.619048	0.666667	0.594286
TRY_E9GI06	0.714286	0.766038	0.624388	0.619048	0.623318	0.580476	0.571429	0.626667	0.583929
TRY_E9GYN1	0.761905	0.807547	0.662585	0.666667	0.672646	0.603593	0.571429	0.626667	0.592619
TRY_E9HT87	0.714286	0.769811	0.651905	0.619048	0.61435	0.565411	0.619048	0.68	0.606667
TRY_F0RPS3	0.571429	0.618868	0.565986	0.571429	0.578475	0.546623	0.47619	0.546667	0.575833

TRY_F6Y1Q1	0.666667	0.720755	0.590306	0.571429	0.573991	0.564459	0.571429	0.626667	0.581309
TRY_G0K2W4	0.761905	0.815094	0.636395	0.714286	0.713004	0.616364	0.619048	0.666667	0.604167
TRY_G8S323	0.619048	0.675472	0.564286	0.619048	0.623318	0.551688	0.619048	0.68	0.568452
TRY_H2B653	0.285714	0.320755	0.427279	0.190476	0.188341	0.398745	0.047619	0.053333	0.328214
TRY_H2K281	0.714286	0.766038	0.634728	0.666667	0.668161	0.595758	0.619048	0.666667	0.588095
TRY_H2XPX5	0.52381	0.539623	0.467041	0.47619	0.475336	0.407922	0.380952	0.426667	0.373095
TRY_O97399	0.761905	0.784906	0.643809	0.619048	0.623318	0.59697	0.52381	0.586667	0.572024
TRY_P00765	0.809524	0.837736	0.670782	0.666667	0.672646	0.608701	0.571429	0.626667	0.584881
TRY_P04814	0.761905	0.781132	0.642891	0.666667	0.668161	0.587835	0.666667	0.706667	0.597381
TRY_P07146	0.857143	0.875472	0.656905	0.666667	0.672646	0.604892	0.571429	0.626667	0.577738
TRY_P07477	0.809524	0.856604	0.657313	0.666667	0.672646	0.612641	0.571429	0.626667	0.570833
TRY_P07478	0.857143	0.875472	0.668197	0.666667	0.672646	0.604892	0.571429	0.626667	0.577738
TRY_P08426	0.857143	0.875472	0.660238	0.619048	0.623318	0.595844	0.571429	0.626667	0.586429
TRY_P24664	0.619048	0.671698	0.614354	0.619048	0.61435	0.598918	0.571429	0.626667	0.566548
TRY_P35004	0.809524	0.830189	0.652143	0.619048	0.623318	0.581082	0.619048	0.666667	0.591071
TRY_P35005	0.809524	0.826415	0.666973	0.666667	0.668161	0.602597	0.666667	0.72	0.622976
TRY_P35030	0.761905	0.807547	0.642619	0.619048	0.623318	0.58697	0.52381	0.6	0.568095
TRY_P35033	0.857143	0.875472	0.658605	0.666667	0.672646	0.604892	0.571429	0.626667	0.577738
TRY_P35036	0.761905	0.792453	0.632449	0.666667	0.668161	0.595455	0.666667	0.706667	0.596905
TRY_P35038	0.666667	0.716981	0.635884	0.666667	0.672646	0.601602	0.619048	0.666667	0.603929
TRY_P35048	0.714286	0.762264	0.613741	0.571429	0.573991	0.555844	0.619048	0.68	0.591905
TRY_P35049	0.666667	0.735849	0.626905	0.619048	0.623318	0.56	0.571429	0.64	0.56119
TRY_P35051	0.190476	0.184906	0.165952	0.047619	0.058296	0.122121	0	0	0.025714
TRY_P42278	0.761905	0.784906	0.652381	0.666667	0.668161	0.619264	0.619048	0.666667	0.599643
TRY_P51588	0.761905	0.807547	0.65932	0.666667	0.668161	0.604805	0.666667	0.706667	0.6075
TRY_P52905	0.666667	0.686792	0.585136	0.714286	0.717489	0.59342	0.571429	0.626667	0.565
TRY_Q1D1D2	0.666667	0.701887	0.622143	0.52381	0.529148	0.556883	0.52381	0.573333	0.557619
TRY_Q28EV7	0.714286	0.769811	0.644388	0.619048	0.623318	0.603593	0.571429	0.626667	0.583929
TRY_Q4VSI2	0.619048	0.664151	0.59949	0.609048	0.608834	0.568788	0.52381	0.573333	0.55631
TRY_Q54179	0.619048	0.683019	0.56102	0.47619	0.475336	0.529481	0.619048	0.693333	0.590238
TRY_Q6MQB3	0.333333	0.362264	0.435238	0.285714	0.286996	0.460303	0.190476	0.226667	0.424286
TRY_Q6QX59	0.809524	0.837736	0.677517	0.714286	0.717489	0.635931	0.619048	0.666667	0.595833
TRY_Q6QX60	0.857143	0.901887	0.670816	0.619048	0.623318	0.5929	0.571429	0.626667	0.584643
TRY_Q7JPN9	0.714286	0.769811	0.615646	0.714286	0.717489	0.607359	0.619048	0.666667	0.572738
TRY_Q8IYP2	0.333333	0.366038	0.468095	0.428571	0.434978	0.539827	0.333333	0.373333	0.484524
TRY_Q8MS52	0.714286	0.762264	0.623639	0.666667	0.672646	0.601948	0.571429	0.626667	0.602857
TRY_Q8ZSE3	0.238095	0.267925	0.417415	0.285714	0.282511	0.4071	0.142857	0.173333	0.402024
TRY_Q9VBY4	0.714286	0.762264	0.618741	0.619048	0.623318	0.594892	0.666667	0.733333	0.625833
TRY_Q9VUG2	0.666667	0.724528	0.621633	0.619048	0.623318	0.61	0.619048	0.693333	0.625357

Notes: The scores for every query sequence with respect to subfamily specific training set is shown. The highest score for each of the sequences have been highlighted and it can be seen that in majority of the cases a correct classification was obtained for the query sequences.