Bioinformatics and Biology Insights

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Insights into Protein Sequence and Structure-Derived Features Mediating 3D Domain Swapping Mechanism using Support Vector Machine Based Approach

Khader Shameer[1], Ganesan Pugalenthi[2], Krishna Kumar Kandaswamy[3,4], Ponnuthurai N. Suganthan[2], Govindaraju Archunan[5] and Ramanathan Sowdhamini[1]

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road, Bangalore, 560065, India. [2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. [3]Institute for Neuro- and Bioinformatics, University of Lübeck, 23538 Lübeck, Germany. [4]Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany. [5]Department of Animal Science, Center for Pheromone Technology, Bharathidasan University, Tiruchirappalli, Tamilnadu, 620 024, India. Email: mini@ncbs.res.in

Abstract: 3-dimensional domain swapping is a mechanism where two or more protein molecules form higher order oligomers by exchanging identical or similar subunits. Recently, this phenomenon has received much attention in the context of prions and neurodegenerative diseases, due to its role in the functional regulation, formation of higher oligomers, protein misfolding, aggregation *etc*. While 3-dimensional domain swap mechanism can be detected from three-dimensional structures, it remains a formidable challenge to derive common sequence or structural patterns from proteins involved in swapping. We have developed a SVM-based classifier to predict domain swapping events using a set of features derived from sequence and structural data. The SVM classifier was trained on features derived from 150 proteins reported to be involved in 3D domain swapping and 150 proteins not known to be involved in swapped conformation or related to proteins involved in swapping phenomenon. The testing was performed using 63 proteins from the positive dataset and 63 proteins from the negative dataset. We obtained 76.33% accuracy from training and 73.81% accuracy from testing. Due to high diversity in the sequence, structure and functions of proteins involved in domain swapping, availability of such an algorithm to predict swapping events from sequence and structure-derived features will be an initial step towards identification of more putative proteins that may be involved in swapping or proteins involved in deposition disease. Further, the top features emerging in our feature selection method may be analysed further to understand their roles in the mechanism of domain swapping.

Keywords: 3D domain swapping, domain swap, machine learning, SVM, feature selection

Bioinformatics and Biology Insights 2010:4 33–42

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

Introduction

Many cellular functions rely on interactions between protein pairs and are mediated by proteins in oligomeric conformations. Although there are many possible mechanisms for oligomer formation, 3D domain swapping has been proposed as an important mechanism that explains the evolution from monomeric to oligomeric proteins.[1–4] 3D domain swapping can be defined as a mechanism for forming oligomeric proteins from their monomers by exchanging identical or similar subunits. The swapped region can be an entire domain or a helix or β-strand or loop regions.[5,6] Protein structures reported to be engaged in 3D domain swapping are distinct from the rest of the oligomers due to the signature-swapping phenomenon. Yet, they are extremely diverse based on their primary sequence and secondary structures and belong to different protein domain families and structural classes. Although domain swapping is an important mechanism for controlling multi-protein assembly, it has also been suggested as a possible mechanism for protein misfolding and aggregation.[5–8] Protein structures in swapped conformations are reported to initiate pathological conformations in prion proteins and human cystatin C. They are reported to aggregate same type of proteins to generate aberrant structures.[6,7,9–13] For example, amyloidogenic proteins like cystatin C and prion proteins have been shown to form dimers by exchange of subdomains of the monomeric proteins.[3,6,14] 3D domain swapping phenomenon is interesting not only due to its pathological conformation factor; it is also important due to a wide range of functions mediated by the proteins in swapped conformation.[7,12,13] It has been reported as a mechanism for dimer formation in odorant binding proteins[6,15,16] and has also been proposed as a possible mechanism for fibril formation.[7,14] Several well-studied examples for domain swapping events have been reported. For example, bovine seminal ribonuclease is a natural domain-swapped dimer that has special biological properties, such as cytotoxicity to tumor cells.[17] Barnase, a domain swapped trimer, is an enzyme that acquires enzymatic activity by cyclic domain swapping.[18] For example, Diptheria toxin, RNase, Cro (DNA repressor), Spectrin (cytoskeleton), antibody fragments, human prion protein (implicated in various types of transmissible neurodegenerative spongiform encephalopathy), human cystatin C (implicated in amyloidosis and Alzheimer's disease) and SH3 domains (important molecule in signal transduction) are shown to be having 3D domain swapped segments with crucial functional roles.[12] The functional diversity of proteins reported with 3-dimensional domain swapping is reflected in a diverse set of Gene Ontology (GO) annotations[19] obtained from PDB ID to GO annotation mapping. Table 1 is provided with the GO annotations (Molecular Function), SCOP fold and Pfam domain IDs of 10 different proteins reported with 3D domain swap mechanism along with their diverse function annotations. The study of 3D domain swapping events in proteins will be an important step towards understanding the molecular basis of the various factors that control this phenomenon and its crucial role in deposition diseases and evolution of swapping in oligomers. As 3D domain swapping is observed in different structures belong to different structural superfamilies (as an example, a set of 3 structures involved in 3D domain swapping is provided in Figure 1) with no common structural, sequence or functional patterns, identification of domain swapping events from features derived from combination of sequence and structural properties provides interesting insights into the patterns that could differentiate between the oligomers in swapped conformation and normal oligomers. In this manuscript, we report the details of a new Support Vector Machine (SVM) based classifier developed to differentiate between swapped oligomers or normal protein structures with a reliable accuracy of 73.81%. Further, the manuscript also discusses the top features emerging from the information-gain-based feature-selection method of the prediction model and its implication in large-scale analysis of 3D domain swapping in proteins.

Materials and Methods

Curation of the datasets

We have performed extensive database and literature curation to collect sequence and structural data for proteins with the structural features of domain swapping. We have collected a set of PDB structures from Protein Data Bank (PDB)[20] using a combination of integrative database searches and extensive literature curation of the existence and extent of 3D domain swapping. These entries were further manually analyzed using combination

Table 1. List of 10 structures with GO annotation, SCOP fold and Pfam domain ID.

PDB ID	GO annotation (Molecular function)	SCOP fold	Pfam domain ID
1A64[57]	antigen binding, protein binding, protein homodimerization activity, protein self-association	Immunoglobulin-like beta-sandwich	V-set
1OQF[58]	catalytic activity, lyase activity, methylisocitrate lyase activity	TIM beta/alpha-barrel	ICL
1K6 W[59]	cytosine deaminase activity, iron ion binding, hydrolase activity, hydrolase activity, acting on carbon-nitrogen (but not peptide) bond, metal ion binding	Composite domain of metallo-dependent hydrolases	Amidohydro_3
11BA[60]	nucleic acid binding, nuclease activity, endonuclease activity, pancreatic ribonuclease activity, hydrolase activity	RNase A-like	Rnase A
1EK1[61]	magnesium ion binding, catalytic activity, epoxide hydrolase activity, hydrolase activity, metal ion binding	alpha/beta-Hydrolases, HAD-like	Abhydrolase_1, Hydrolase
1I21[62]	glucosamine 6-phosphate N-acetyltransferase activity, N-acetyltransferase activity, acyltransferase activity, transferase activity	Acyl-CoA N-acyltransferases (Nat)	Acetyltransf_1
1M5M[63]	sugar binding	Cyanovirin-N	CVNH
1FRO[64]	lactoylglutathione lyase activity, zinc ion binding, lyase activity, metal ion binding	Glyoxalase/Bleomycin resistance protein/ Dihydroxybiphenyl dioxygenase	Glyoxalase
1DDT[65]	transferase activity, transferase activity, transferring glycosyl groups, NAD^+-diphthamide ADP-ribosyltransferase activity	Common fold of diphtheria toxin/transcription factors/cytochrome f	Diphtheria_R, Diphtheria_T, Diphtheria_C
1LSS[66]	catalytic activity, binding, cation transmembrane transporter activity, potassium ion binding	NAD(P)-binding Rossmann-fold domains	TrkA_N

of macromolecular visualization tools PyMol,[21] Rasmol[22] and literature reports. The structural entries were further processed using Domain Identification ALgorithm (DIAL) server[23] to identify probable swapped segments from the structural data. PDB ID to PubMed ID mapping and PDBSum database[24] were used to obtain primary literature reports. Since many structures are not available in quaternary state from the PDB, Protein Quaternary Structure server (PQS)[25] was consulted to obtain the quaternary assembly of the structures. From the extensive curation, 3Dswap: Knowledge-base of 3D domain swapping in Proteins, unpublished data, 315 PDB entries with 344 chains were obtained for the positive dataset. These chains were further mapped to their respective SCOP[26] folds. To curate the negative dataset, we scanned different databases (PDB, PQS, and PDBSum) for dimers or higher order oligomers that are not included in positive dataset. PDB was scanned for

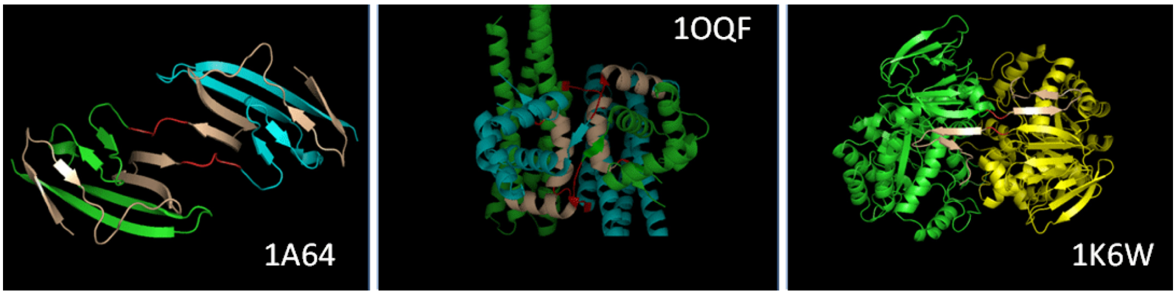

Figure 1. Structures of three different proteins involved in 3D domain swapping (PDB IDs: 1A64,[57] 1OQF,[58] 1K6W[59]). Hinge region is colored in red and swapped segment is in coffee brown.

oligomers that are not reported to be involved in domain swapping. The negative dataset was generated after excluding the SCOP folds reported in the positive dataset. To add diversity to the negative dataset, members from a single SCOP fold was represented only once in the negative dataset. The redundant entries were removed by considering their sequence identity. Sequences extracted from structures that have >70% sequence identity were removed using the CD-HIT program.[27] We retained 213 domain swap sequences for the positive dataset. Equal number of negative data was obtained from the Protein Data Bank. The training dataset was constructed using 150 domain swapping and 150 non-domain swapping sequences. Remaining 63 domain swap sequences and 63 non-domain swapping sequences were employed for testing. Schematic representation of data curation steps, followed to generate positive and negative data, are given in Figure 2.

Features

The SVM model is generated using a combination of features derived from sequence, structure and physico-chemical properties. Initially, each sequence is represented by a set of 66 features. Further, a set of features that contribute to the prediction model is identified using the feature-selection approach explained in 'Feature selection' section. The features sets used in the prediction can be classified into three groups as sequence-derived, structure-derived and physico-chemical features.

Sequence features

Sequence features are derived exclusively from sequence of proteins in the positive and negative datasets. The frequencies of 20 amino acids were calculated from the total number of each amino acid in a given sequence divided by protein length as explained previously in Pugalenthi and coworkers.[28] In addition, the amino acids are grouped into hydrophobic,

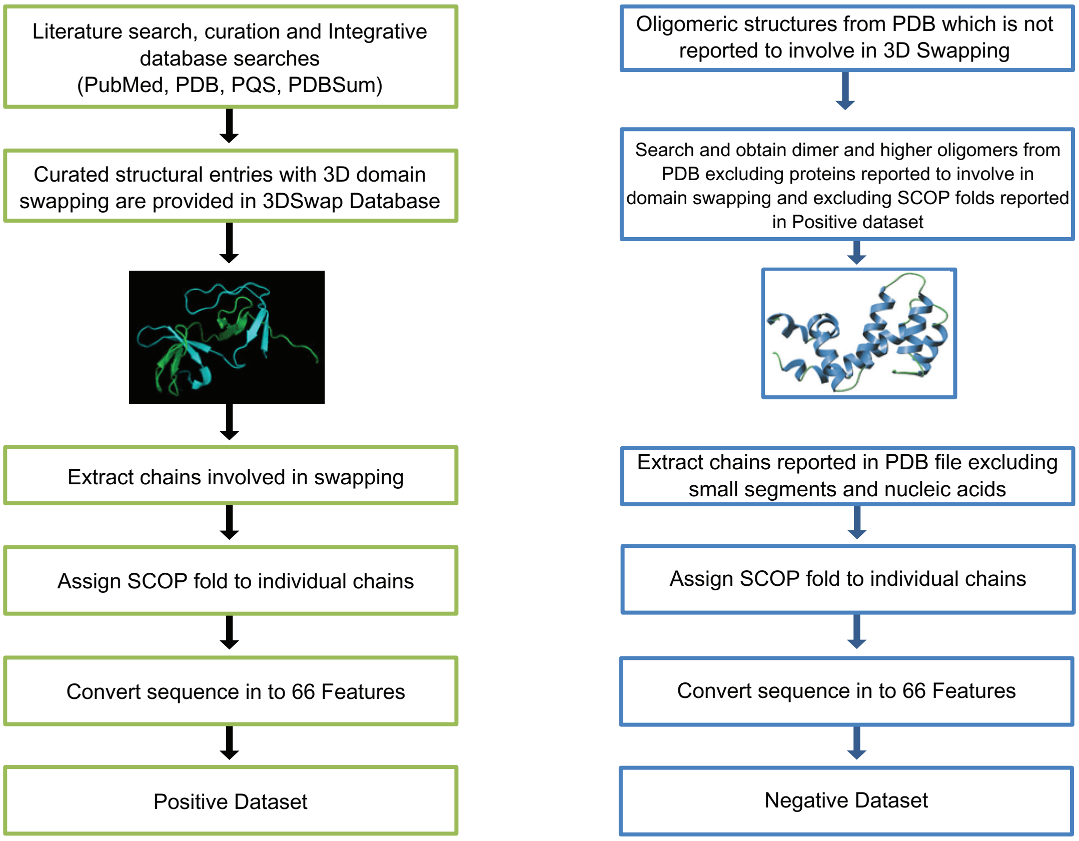
Literature search, curation and Integrative database searches (PubMed, PDB, PQS, PDBSum)
Curated structural entries with 3D domain swapping are provided in 3DSwap Database
Extract chains involved in swapping
Assign SCOP fold to individual chains
Convert sequence in to 66 Features
Positive Dataset
Oligomeric structures from PDB which is not reported to involve in 3D Swapping
Search and obtain dimer and higher oligomers from PDB excluding proteins reported to involve in domain swapping and excluding SCOP folds reported in Positive dataset
Extract chains reported in PDB file excluding small segments and nucleic acids
Assign SCOP fold to individual chains
Convert sequence in to 66 Features
Negative Dataset

Figure 2. Schematic representation of data curation steps.

hydrophilic and neutral amino acids (see Pugalenthi and coworkers[29]) and the frequency was obtained for each sequence in the datasets.

Structure-derived features

Structure-derived features refer to a set of features derived from the PDB coordinates of the positive and negative datasets. Structure-based features such as solvent accessibility, secondary structures, hydrogen bonds and residue compactness were computed from the individual protein structures using JOY package.[30] Basic structure-based features used in the prediction model are overall composition of helix, overall composition of strand and overall composition of coil. Along with the generic structure-based features, we have also used 'structure-derived fusion-features' like hydrogen bonds in helix, hydrogen bonds in strand, and hydrogen bonds in coil where the frequency of hydrogen bonds in a given structure is coupled with secondary structure of residues that mediate the hydrogen bonds. The frequency of solvent inaccessible residues in the secondary structure classes like helix, strand and coil was also computed. Another set of structure-derived fusion-features includes the number of cysteine residues in helix, the number of cysteine residues in strand and the number of cysteine residues in coil regions. Hydrogen bonds were calculated using HBOND routine available from the JOY package. Secondary structure information was inferred using the SSTRUC program available from the JOY package. Solvent accessibility was calculated using the routine available in the PSA routine in JOY package to compute the Ooi number. Composition of secondary structural elements and frequency of hydrogen bonds mediated by residues in secondary structural elements were calculated using custom Perl scripts.

Physicochemical features

We obtained 18 physico-chemical properties from AAINDEX[31] and its derivative UMBC AAINDEX database.[32] The computed physico-chemical properties include molecular weight, hydrophobicity, hydrophilicity, refractivity, average accessible surface area, flexibility, melting point, side chain volume, side chain hydrophobicity, polarity, heat capacity, isoelectric points and normalized frequency of α-helix, β-sheet and coil. Physico-chemical features were derived from the protein sequence of proteins from positive and negative datasets using custom Perl scripts.

Support vector machine

SVM, rigorously based on Vapnik's statistical learning theory[33,34] possesses excellent generalization capability. Due to its excellent generalization capabilities, it is widely used in bioinformatics applications.[28,29,35–37] When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest in each class. Details of the formulation and solution methodology of SVM for binary classification task can be found elsewhere.[34] We provide here only final form of the decision function and the type of kernel function employed in our study.

Let x_i Є RN, $i = 1, 2 \ldots, N$ be input feature vectors and y_i Є {+1, −1} be its corresponding class label, where, N be the total number of proteins in training database. To assign a class label for a query sequence x, the trained SVM model applies the following function form:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x_j) + b \qquad (1)$$

In this equation, where, m is the number of support vectors, a subset of training dataset, m < N having non-zero positive values of the Lagrange multipliers, α_i which are obtained by solving a quadratic optimization problem and b is the bias term. We have conducted our study with Radial Basis Function (RBF) kernel function defined by Equation 2.

$$K(x_i, x_j) = exp\left(-\sigma \cdot \frac{\left(x_i - x_j\right)^2}{2} \right) \qquad (2)$$

$K\ (x_i,x_j)$ represents Radial Basis Function (RBF) kernel. Parameter σ in Equation (2) decides the width of the Radial Basis Function kernel function.[33,34] Simulations were performed using LIBSVM version 2.81 (C.C. Chang, 2001). SVM training was carried out by optimization of the value of regularization parameter and the value of RBF kernel parameter.

5 fold cross validation experiment was carried out to evaluate performance of SVM model.

Feature selection

To identify the important features that distinguish positive and negative classes, we used Information Gain algorithm with the ranker method for the feature selection. This method was implemented using Weka 3.5.[38] The information gain for each feature was calculated and the features were ranked according to this measure.

Prediction assessment

The prediction system is evaluated using sensitivity, specificity, accuracy, positive prediction value (PPV), negative prediction value (NPV) and Mathew's Correlation Coefficient (MCC). These measurements are expressed in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP). The measurements are defined as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

$$NPV = \frac{TN}{TN + FN} \tag{7}$$

$$MCC(X) = \frac{(TPTN - FPFN)}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}} \tag{8}$$

The MCC ranges from $-1 \leq MCC \leq 1$. A value of MCC = 1 indicates the best possible prediction while MCC = −1 indicates the worst possible prediction (or anti-correlation). Finally, MCC = 0 would be expected for a random prediction scheme (Matthews, 1975). Five-fold cross-validation method is also used to evaluate the performance of the model with respect to different sub-sets of the data. Results of the prediction assessment using five-fold cross validation on training dataset (Table 2) and independent validation dataset (Table 3) are provided.

Results and Discussion

We have developed a new SVM model to differentiate structures in swapped conformation from normal oligomers or normal structures. The model was trained on a training dataset containing 150 proteins from the positive dataset and 150 proteins from the negative dataset. The performance of the model was evaluated using the five-fold cross-validation method. As shown in Table 2, overall prediction accuracy of 76.33% was obtained by five-fold cross validation. In order to identify the prominent features, feature selection (information gain with ranker method) was performed on this dataset. We selected five feature subsets by decreasing the number of features and the performance of each feature subset was evaluated using five-fold cross-validation. As seen in Table 2, feature selection generally does not deteriorate the classification performance much until the number of features decreases to 10. Using 10 features, our model obtained 71.67% accuracy that is comparable to accuracy obtained using all features. Similar performance was observed using 25 and 50 feature subsets. This result suggests that our feature reduction approach selected useful features by eliminating the uncorrelated and noisy features. In order to examine the performance of the newly developed model, we tested our training model on the test dataset consisting of 63 proteins from the positive dataset and 63 proteins from the negative dataset. As shown in Table 3, our model achieved 73.81% accuracy with 73.02% sensitivity and 74.60% specificity using all features and 76.19% accuracy with 73.02% sensitivity and 79.37% sensitivity using 50 features. We

Table 2. Performance evaluation on training data (150 proteins from positive dataset and 150 proteins from negative dataset).

Feature subset	5 fold cross validation (%)
10 features	71.67
25 features	75.33
50 features	76.33
All features (66)	76.33

Table 3. Test with independent validation dataset (63 proteins from positive dataset and 63 proteins from negative dataset).

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)	PPV (%)	NPV (%)
10 features	69.84	66.67	0.37	68.25	67.69	68.85
25 features	73.02	65.08	0.38	69.05	67.65	70.69
50 features	73.02	79.37	0.52	76.19	77.97	74.63
All features (66)	73.02	74.60	0.48	73.81	74.19	73.44

Abbreviations: MCC, Matthews Correlation Coefficient; PPV, Positive prediction value; NPV, Negative prediction value; AROC, Asymptotic receiver operating characteristic.

investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Fig. 3) derived from the sensitivity (true positive rate) and specificity (false positive rate) values for the classifiers using all the features and the 10 best performing features (Table 4), respectively.

The list of top 10 features clearly indicates that features with higher classification strength are a mix of sequence, structural and physicochemical derived features. This feature distribution in both sequence and structural classes also asserts that swapping can be detected from combination of features from sequence and structural information. The 10 best performing features emerged from the feature selection using information gain algorithm offers interesting leads into the mechanism that mediate domain swapping. As no generic sequence or structure based common pattern is reported to be a hallmark of structures with domain swap mechanism, the set of top 10 features could be considered further for detailed analysis. A generic sequence or structure analysis approach could have likely missed the identification of these features, but the combination of features and machine learning based approach used in the current work enables the identification of the specific patterns between the positive and the negative datasets. Top 10 features (Table 4) identified by the feature selection method can be classified into three categories based on the mode of feature derivation. Top 10 features include four sequence-derived features (frequency of neutral amino acids, valine, tyrosine and tryptophan), one physico-chemical feature derived from sequence (refractivity), one structure-derived feature (composition of coil) and four structure- derived fusion-features (solvent inaccessible residues in coil, frequency of residues that form hydrogen bond to main chain CO in helix, number of cysteine residues in strand and number of cysteine residues in helix).

Our current prediction model has its limitations due to smaller sample size of the positive dataset. Depending upon the availability of more crystal structures with swapped conformation, the method could be improved by re-training the model using larger

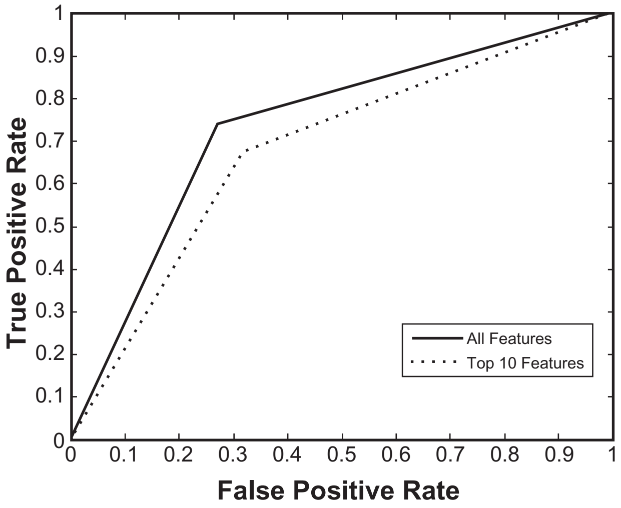

Figure 3. ROC curves plotted utilizing the fractions of true positives and false positives values derived using top 10 features and all features.

Table 4. List of top 10 selected features.

No	Features
1	Solvent inaccessible residues in coil
2	Frequency of residues (that form hydrogen bond to main chain CO) in helix
3	Number of cysteines in strand
4	Physico chemical properties (Refractivity)
5	Number of cysteines in helix
6	Frequency of neutral amino acids (THSQ)
7	Frequency of valine
8	Frequency of tyrosine
9	Frequency of tryptophan
10	Composition of coil

Table 5. Example results using the prediction model.

PDB ID	Protein name	Result
1YVS	Barnase	Domain swap
2NZ7	Caspase-recruitment domain	Domain swap
2OQR	Response regulator RegX3	Domain swap
2VTY	Novel Bcl-2-Like domain swapped dimer	Domain swap
2B9I	GITRL	Domain swap
3EXM	Cyanovirin-N	Domain swap
2V4N	Sur E	Non swap
2PQM	Cysteine synthase	Non swap

datasets. Due to unavailability of other methods or classifiers for the prediction of swapping events from sequence or structure data, the current method is not compared with any of the existing methods. To show the results of the prediction model, a set of example input PDB files and their respective results obtained using the current prediction model is provided in Table 5.

Conclusion

Domain swapping mechanism is essential for the formation of higher protein oligomers from their monomer, protein misfolding, protein aggregation *etc*. Several experimental[39–49] and computational studies[50–56] are performed to understand various aspects of domain swapping. We have attempted to predict the phenomenon of domain swapping from the sequence and structure-derived features of a protein using machine-learning approach based on support vector machines. Identification of common sequence or structure-based features from the structures that show this phenomenon is a challenging task. We developed SVM-based classifier to predict domain swapping event using sequence and structure-derived features. This method obtained 76.33% accuracy from training and 73.81% accuracy from testing. This method could be extremely useful for the identification of domain swap phenomenon from protein structure data based on features derived from protein sequence data and structural co-ordinates. The set of features identified using our feature-selection method is providing new insights to understand a common pattern behind domain swapping and need to be explored further. The method can be improved by considering exclusive sequence based features, so that a classifier could be designed which can perform prediction using (3Dswap-pred—prediction of 3D domain swapping from protein sequence, unpublished data). Such a method could be applied at the whole genome level to scan and identify putative proteins showing domain swapping.

Acknowledgements

R.S. and K.S. acknowledge National Centre for Biological Sciences (TIFR) for infrastructural and financial support. R.S. was a Senior Research Fellow of the Wellcome Trust, U.K. R.S. and G.A. thank Department of Biotechnology, Government of India for financial support. G.P. and P.N.S acknowledge the financial support offered by the A*Star (Agency for Science, Technology and Research, Singapore) under the grant # 052 101 0020. KKK acknowledges Prof. Thomas Martinetz and Dr. Stefen Moller, Institute for Neuro- and Bioinformatics, University of Lübeck, Germany for their support.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Almassy RJ, Janson CA, Hamlin R, Xuong NH, Eisenberg D. Novel subunit-subunit interactions in the structure of glutamine synthetase. *Nature*. 1986;323:304–9.
2. Anfinsen CB. The formation and stabilization of protein structure. *Biochem J*. 1972;128: 737–49.
3. Bennett MJ, Choe S, Eisenberg D. Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci U S A*. 1994;91:3127–31.
4. Parge HE, Arvai AS, Murtari DJ, Reed SI. Tainer JA. Human CksHs2 atomic structure: a role for its hexameric assembly in cell cycle control. *Science*. 1993;262:387–95.
5. Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci*. 1995;4:2455–68.
6. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. *Protein Sci*. 2002;11:1285–99.
7. Bennett MJ, Sawaya MR, Eisenberg D. Deposition diseases and 3D domain swapping. *Structure*. 2006;14:811–24.
8. Khare SD, Dokholyan NV. Molecular mechanisms of polypeptide aggregation in human diseases. *Curr Protein Pept Sci*. 2007;8:573–9.
9. Bennett MJ, Eisenberg D. The evolving role of 3D domain swapping in proteins. *Structure*. 2004;12:1339–41.
10. Gronenborn AM. Protein acrobatics in pairs—dimerization via domain swapping. *Curr Opin Struct Biol*. 2009;19:39–49.

11. Jaskolski M. 3D domain swapping, protein oligomerization, and amyloid formation. *Acta Biochim Pol*. 2001;48:807–27.
12. Nagradova NK. Three-dimensional domain swapping in homooligomeric proteins and its functional significance. *Biochemistry (Mosc)*. 2002;67: 839–49.
13. Newcomer ME. Protein folding and three-dimensional domain swapping: a strained relationship? *Curr Opin Struct Biol*. 2002;12:48–53.
14. Guo Z, Eisenberg D. Runaway domain swapping in amyloid-like fibrils of T7 endonuclease I. *Proc Natl Acad Sci U S A*. 2006;103:8042–7.
15. Pelosi P. Odorant-binding proteins. *Crit Rev Biochem Mol Biol*. 1994;29: 199–228.
16. Ramoni R, et al. Control of domain swapping in bovine odorant-binding protein. *Biochem J*. 2002;365:739–48.
17. Liu Y, Hart PJ, Schlunegger MP, Eisenberg D. The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-A resolution. *Proc Natl Acad Sci U S A*. 1998;95:3437–42.
18. Zegers I, Deswarte J, Wyns L. Trimeric domain-swapped barnase. *Proc Natl Acad Sci U S A*. 1999;96:818–22.
19. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25:25–9.
20. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*. 2007;35:D301–3.
21. DeLano WL. The PyMOL Molecular Graphics System. 2002.
22. Bernstein HJ. Recent changes to RasMol, recombining the variants. *Trends Biochem Sci*. 2000;25:453–5.
23. Pugalenthi G, Archunan G, Sowdhamini R. DIAL: a web-based server for the automatic identification of structural domains in proteins. *Nucleic Acids Res*. 2005;33:W130–2.
24. Laskowski RA. Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics*. 2007;23;1824–7.
25. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci*. 1998;23:358–61.
26. Andreeva A, et al. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*. 2008;36:D419–25.
27. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–9.
28. Pugalenthi G, Tang K, Suganthan PN, Archunan G, Sowdhamini R. A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. *BMC Bioinformatics*. 2007; 8:351.
29. Pugalenthi G, Kumar KK, Suganthan PN, Gangal R. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochem Biophys Res Commun*. 2008;367;630–4.
30. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. JOY: protein sequence-structure representation and analysis. *Bioinformatics*. 1998;14:617–23.
31. Kawashima S, et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 2008;36:D202–5.
32. Bulka B, desJardins M, Freeland SJ. An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices. *BMC Bioinformatics*. 2006;7:329.
33. Vapnik V. *The Nature of Statistical Learning Theory*, (Springer, NY, 1995).
34. Muller KR. An introduction to kernel-based learning algorithms. *IEEE Transactions in Neural Network*. 2001;2:181–201.
35. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem*. 2002;277:45765–9.
36. Ding YS, Zhang TL, Chou KC. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein Pept Lett*. 2007;14:811–5.
37. Du QS, Wei YT, Pang ZW, Chou KC, Huang RB. Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction. *Protein Eng Des Sel*. 2007;20: 417–23.
38. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. *Bioinformatics*. 2004;20:2479–81.
39. Ahuja U, Rozhkova A, Glockshuber R, Thony-Meyer L, Einsle O. Helix swapping leads to dimerization of the N-terminal domain of the c-type cytochrome maturation protein CcmH from Escherichia coli. *FEBS Lett*. 2008;582:2779–86.
40. Alcantara EH, Kim DH, Do SI, Lee SS. Bi-functional activities of chimeric lysozymes constructed by domain swapping between bacteriophage T7 and K11 lysozymes. *J Biochem Mol Biol*. 2007;40:539–46.
41. Andjelkovic M, Maira SM, Cron P, Parker PJ, Hemmings BA. Domain swapping used to investigate the mechanism of protein kinase B regulation by 3-phosphoinositide-dependent protein kinase 1 and Ser473 kinase. *Mol Cell Biol*. 1999;19:5061–72.
42. Aravind P, Suman SK, Mishra A, Sharma Y, Sankaranarayanan R. Three-dimensional domain swapping in nitrollin, a single-domain betagamma-crystallin from Nitrosospira multiformis, controls protein conformation and stability but not dimerization. *J Mol Biol*. 2009;385:163–77.
43. Back K, Chappell J. Identifying functional domains within terpene cyclases using a domain-swapping strategy. *Proc Natl Acad Sci U S A*. 1996;93: 6841–5.
44. Back K, et al. Cloning of a sesquiterpene cyclase and its functional expression by domain swapping strategy. *Mol Cells*. 2000;10:220–5.
45. Bakker RA, et al. Domain swapping in the human histamine H1 receptor. *J Pharmacol Exp Ther*. 2004;311:131–8.
46. Balciunas D, Ronne H. Evidence of domain swapping within the jumonji family of transcription factors. *Trends Biochem Sci*. 2000;25:274–6.
47. Chan YH, Cheng CH, Chan KM. Study of goldfish (Carassius auratus) growth hormone structure-function relationship by domain swapping. *Comp Biochem Physiol B Biochem Mol Biol*. 2007;146:384–94.
48. Chintakayala K, et al. Domain swapping reveals that the C- and N-terminal domains of DnaG and DnaB, respectively, are functional homologues. *Mol Microbiol*. 2007;63:1629–39.
49. Cho SS, Levy Y, Onuchic JN, Wolynes PG. Overcoming residual frustration in domain-swapping: the roles of disulfide bonds in dimerization and aggregation. *Phys Biol*. 2005;2:S44–S55.
50. Alonso DO, Alm E, Daggett V. Characterization of the unfolding pathway of the cell-cycle protein p13 suc1 by molecular dynamics simulations: implications for domain swapping. *Structure*. 2000;8:101–10.
51. Esposito L, Daggett V. Insight into ribonuclease A domain swapping by molecular dynamics unfolding simulations. *Biochemistry*. 2005;44: 3358–68.
52. Lin YM, et al. Molecular dynamics simulations to investigate the domain swapping mechanism of human cystatin C. *Biotechnol Prog*. 2007;23: 577–84.
53. Liu HL, et al. Molecular dynamics simulations of human cystatin C and its L68Q varient to investigate the domain swapping mechanism. *J Biomol Struct Dyn*. 2007;25:135–44.
54. Chahine J, Cheung MS. Computational studies of the reversible domain swapping of p13 suc1. *Biophys J*. 2005;89:2693–700.
55. Cozza G, Moro S, Gotte G. Elucidation of the ribonuclease A aggregation process mediated by 3D domain swapping: a computational approach reveals possible new multimeric structures. *Biopolymers*. 2008;89:26–39.
56. Gouldson PR, et al. Dimerization and domain swapping in G-protein-coupled receptors: a computational study. *Neuropsychopharmacology*. 2000;23:S60–S77.
57. Murray AJ, Head JG, Barker JJ, Brady RL. Engineering an intertwined form of CD2 for stability and assembly. *Nat Struct Biol*. 1998;5:778–82.
58. Liu S, et al. Crystal structures of 2-methylisocitrate lyase in complex with product and with isocitrate inhibitor provide insight into lyase substrate specificity, catalysis and evolution. *Biochemistry*. 2005;44: 2949–62.
59. Ireton GC, McDermott G, Black ME, Stoddard BL. The structure of Escherichia coli cytosine deaminase. *J Mol Biol*. 2002;315:687–97.
60. Vitagliano L, et al. Binding of a substrate analog to a domain swapping protein: X-ray structure of the complex of bovine seminal ribonuclease with uridylyl(2',5')adenosine. *Protein Sci*. 1998;7:1691–9.

61. Argiriadi MA, et al. Binding of alkylurea inhibitors to epoxide hydrolase implicates active site tyrosines in substrate activation. *J Biol Chem*. 2000;275:15265–70.
62. Peneff C, Mengin-Lecreulx D, Bourne Y. The crystal structures of Apo and complexed Saccharomyces cerevisiae GNA1 shed light on the catalytic mechanism of an amino-sugar N-acetyltransferase. *J Biol Chem*. 2001; 276:16328–34.
63. Botos I, et al. Structures of the complexes of a potent anti-HIV protein cyanovirin-N and high mannose oligosaccharides. *J Biol Chem*. 2002;277:34336–42.
64. Cameron AD, Olin B, Ridderstrom M, Mannervik B, Jones TA. Crystal structure of human glyoxalase I—evidence for gene duplication and 3D domain swapping. *EMBO J*. 1997;16:3386–95.
65. Bennett MJ, Choe S, Eisenberg D. Refined structure of dimeric diphtheria toxin at 2.0 A resolution. *Protein Sci*. 1994;3:1444–63.
66. Roosild TP, Miller S, Booth IR, Choe S. A mechanism of regulating transmembrane potassium flux through a ligand-mediated conformational switch. *Cell*. 2002;109:781–91.

Supplementary Data

List of 66 features, positive and negative datasets (training and testing) and features derived from positive and negative dataset (training and testing) are provided as supplementary material. The supplementary data can be accessed from the following URL: http://caps.ncbs.res.in/download/3dswap_seq_struc_svm