Genome-Wide Analysis of Domain-Swap Predicted Products in the Genome of Anti-Stress Medicinal Plant: *Ocimum tenuiflorum*

Bioinformatics and Biology Insights
Volume 13: 1–5
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1177932218821362
SAGE

Atul Kumar Upadhyay[1,2] and Ramanathan Sowdhamini[1]

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, India.
[2]Division of Bioinformatics, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, India.

ABSTRACT: Computational approaches to high-throughput data are gaining importance because of explosion of sequences in the post-genomic era. This explosion of sequence data creates a huge gap among the domains of sequence structure and function, since the experimental techniques to determine the structure and function are very expensive, time taking, and laborious in nature. Therefore, there is an urgent need to emphasize on the development of computational approaches in the field of biological systems. Engagement of proteins in quaternary arrangements, such as domain swapping, might be relevant for higher compatibility of such genes at stress conditions. In this study, the capacity to engage in domain swapping was predicted from mere sequence information in the whole genome of holy Basil (*Ocimum tenuiflorum*), which is well known to be an anti-stress agent. Approximately, one-fourth of the proteins of *O tenuiflorum* are predicted to undergo three-dimensional (3D)-domain swapping. Furthermore, function annotation was carried out on all the predicted domain-swap sequences from the *O tenuiflorum* and *Arabidopsis thaliana* for their distribution in different Pfam protein families and gene ontology (GO) terms. These domain-swapped protein sequences are associated with many Pfam protein families with a wide range of GO annotation terms. A comparative analysis of domain-swap-predicted sequences in *O tenuiflorum* with gene products in *A thaliana* reveals that around 26% (2522 sequences) are close homologues across the 2 genomes. Functional annotation of predicted domain-swapped sequences infers that predicted domain-swap sequences are involved in diverse molecular functions, such as in gene regulation of abiotic stress conditions and adaptation to different environmental niches. Finally, the positively predicted sequences of *A thaliana* and *O tenuiflorum* were also examined for their presence in stress regulome, as recorded in our STIFDB database, to check the involvement of these proteins in different abiotic stresses.

KEYWORDS: Machine-learning approaches, Random Forest, three-dimensional-domain swapping, protein sequences, genomes and proteome

RECEIVED: September 4, 2018. **ACCEPTED:** November 8, 2018.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ramanathan Sowdhamini, National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road, Bangalore 560065, India.
Email: mini@ncbs.res.in

Introduction

Studies suggest that around one-third of the proteins in a cell are in oligomeric state.[1] Protein oligomers have evolved because of their advantages over their monomers such as oligomers have more chances of allosteric control, form new active sites at subunit interfaces of oligomers, provide increase in local concentration of active sites as compared with monomers, have chances of retaining larger binding surfaces, and will also create economic ways to produce large protein interaction networks and molecular machines.

Three-dimensional (3D)-domain swapping is one of the mechanisms of protein oligomerisation. The 3D-domain-swapped oligomers form stronger interactions in comparison to side-by-side homologous oligomers.[2] The 3D-domain swapping is a mechanism of protein oligomer formation from monomeric units by exchanging their whole domains or small structural elements (Figure 1). The monomeric subunits first undergo partial unfolding to form open conformations, and these open conformations then undergo domain swapping at high concentrations. First time in 1962, 3D-domain swapping was proposed as the mechanism of dimerization of RNase A.[3] In 1990, crystal structures of β2-crystallin[4] were solved and analyzed by Tom Blundell and Christine Slingsby groups, as a dimer at 2.1 Å resolution to show that the linker region between 2 domains is extended.

This mechanism was first documented by Eisenberg and coworkers[5] in the structure of diphtheria toxin in 1994. There are reports of more than one domain swapping in a single protein, eg, RNase A dimers exhibit swapping in both the N- and C-terminal regions.[3] 3D-domain-swapped proteins have wide range of sizes and highly diverse sequences. Even the swapped domain can also be an entire tertiary domain consisting of hundreds of residues or small structural element like β-strand or α-helix. A flexible linker region, namely, "hinge," possesses intrinsic flexibility to facilitate 3D-domain-swapping process. The hinge region adopts different conformations in the monomer and in the domain-swapped oligomer. For example, the C-terminal β-strand of RNase A is exchanged in both the C-terminal-swapped dimer and the cyclic C-terminal-swapped trimer of RNase A. Hence, it is clear that the same hinge region adopts distinct conformations in the monomer, the C-terminal-swapped dimer, and the cyclic C-terminal-swapped trimer of RNase A, confirming the flexibility of hinge region.[6] The hinge region, which is not sufficiently long enough for the swapped domain to fold back to the core,[7] forms metastable and partially unfolded structures. These structures oligomerise to form domain-swapped oligomers.

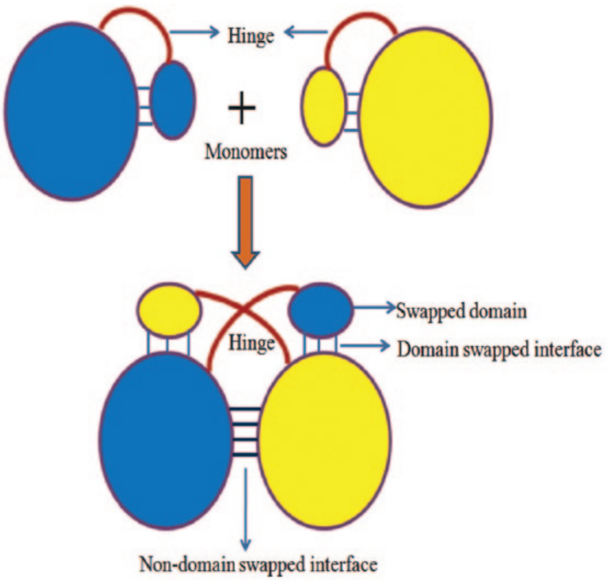

Figure 1. Cartoon representation of 3D-domain swapping.

There are many examples of 3D-domain-swapped proteins in plants, such as the light-harvesting complex of photosystem II,[8] bleomycin resistance,[9] and SET domain protein methyltransferase, involved in many diverse functions. Tulsi (*Ocimum tenuiflorum*, also referred as Holy basil) is a herbal plant that belongs to the genus *Ocimum L.* (Lamiaceae), which is also known as basil. This genus comprises 30 to 160 annual and perennial herbs and shrubs native to the tropical and subtropical regions of Asia, Africa, and America.[10] Holy basil is well known to be anti-stress agent.[11] There are many studies on the use of machine-learning approaches for the understanding of structure and function of these complex protein molecules. In this article, all the potential 3D-domain-swapped proteins are reported from the draft genome of Tulsi,[12] along with function elucidation of these proteins for their involvement in different stress tolerance. The positively predicted gene products are compared with Arabidopsis genome for homologues and in STIFDB database for their possible involvement in combating biotic and abiotic stresses in Arabidopsis.

Material and Methods

Prediction of 3D-domain-swapped sequences of Tulsi

Using 439 sequence features of known examples of 3D-domain swapping, a Random Forest model was generated. Physicochemical features were extracted from AAINDEX database.[13] WEKA was consulted for the selection of best features.[14] The detailed methodology of prediction method is already discussed in the earlier paper.[15] This model was applied for the prediction of 3D-domain swapping to 36 768 proteins of *O tenuiflorum* (Ote). Prediction model was also applied to the whole genome of reviewed proteins from UniProt of *Arabidopsis thaliana* (Ath), *Solanum tuberosum* (Stu), *Solanum lycopersicum* (Sly), and *Medicago truncatula* (Mtr). In these genomes, *Ocimum* has maximum number of proteins (36 768) and *M. truncatula* has minimum (186) sequences.

Genome-wide and comparative analyses of predicted 3D-domain-swapping sequences of Ocimum and Arabidopsis

Functional annotation of positively predicted sequences of *Ocimum* genome was performed by TRAPID,[16] an online tool for the functional and comparative analyses of high-throughput data. Functional annotation of all the positively predicted sequences of Tulsi genome was performed for their gene ontology (GO) term, protein family information, and protein domain associations.

A comparative analysis was also performed on the predicted sequences of *Ocimum* and *Arabidopsis*. Putative 3D-domain-swap predicted sequences (9419) from *Ocimum* were used for searching homologous sequences in different plant genomes. Those gene products, which were positively predicted for domain swapping in the draft genome of *O tenuiflorum*, were used as query against UniProt at *E*-value of 10 for assigning function to these sequences.

Domain-swap predicted sequences from *Arabidopsis* were checked for their involvement in abiotic stress, by using an in-house server STIFDB2.[17] Common protein sequences to domain swapping and stress were searched against the protein structure database. Some of the predicted sequences, which have homologues of known structure, were validated by manual visualization of these structures with the help of PyMol (https://www.pymol.org/), a molecular visualization tool.

Results

Prediction of 3D-domain-swapped sequences of Tulsi

For this study, all the protein sequences (36 768) of *O tenuiflorum*[12] were used for the prediction of 3D-domain swapping by using the machine-learning approaches. For the sake of comparison, Random Forest model was applied on 5 plant genomes for the prediction of 3D-domain swapping. Prediction results range from a minimum of 25% in *O tenuiflorum* to a maximum of 64% in *A thaliana* for reviewed sequences from UniProt (Table 1). Plant genomes used in this study are *A thaliana, M truncatula, S tuberosum*, and *S lycopersicum*, where the percentage of genes predicted to be involved in domain swapping is 64, 48, 52, and 43, respectively. In *O tenuiflorum*, the lowest percentage of predicted sequences is observed among all the plant genomes considered in this study.

Genome-wide and comparative analysis of predicted 3D-domain sequences of O tenuiflorum and A thaliana

A total of 9346 (25%) protein sequences of Tulsi proteome were predicted to be engaged in 3D-domain swapping. The

Table 1. Prediction result of 3D-domain swapping by RF approach on different plant genomes.

S. NO.	GENOMES	TOTAL REVIEWED SEQUENCES	POSITIVE PREDICTION BY RF
1	*Arabidopsis thaliana*	12 033	7694 (64%)
2	*Medicago truncatula*	186	48 (26%)
3	*Solanum tuberosum*	400	208 (52%)
4	*Solanum lycopersicum*	423	183 (43%)
5	*Ocimum tenuiflorum*	36 768	9419 (25%)

Abbreviation: RF, Random Forest.

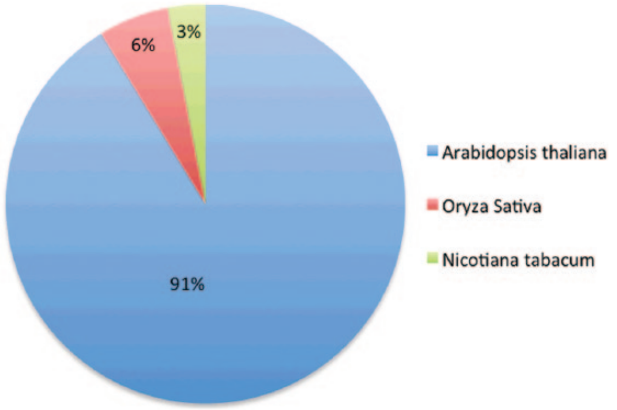

Figure 2. Distribution of homologues of predicted 3D-domain-swapped proteins of Tulsi in different plant genomes.

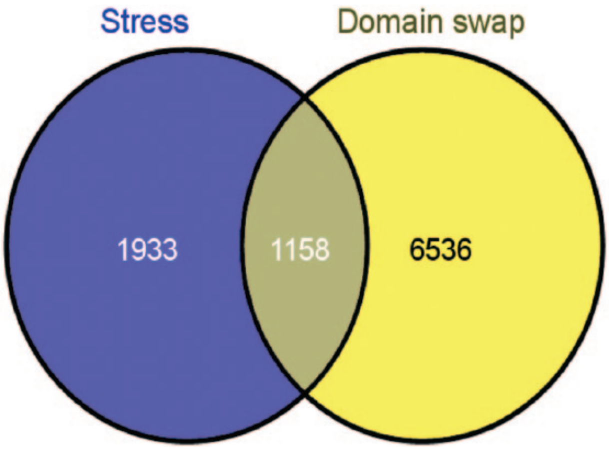

Figure 3. Venn diagram showing the number of common proteins to 3D-domain swapping and biotic and abiotic stresses in *A thaliana.*

average protein length of positively predicted sequences in this dataset is 520.2 bp. Out of 9346 positively predicted sequences of Tulsi genome, 9082 sequences have homologous sequence in other plant genomes with InterPro[18] gene family: 8027 of them retain GO term association and 8339 have protein domain association.

All the positively predicted 9346 sequences from Tulsi genome were searched against UniProt.[19] The distribution of homologous sequences in different organisms is plotted (Figure 2). Out of 9346 predicted sequences, 9265 sequences have homologous sequences in UniProt. Within 9265 homologous sequences, 3213 sequences have the best homologue in *Arabidopsis*, on the basis of identity percentage and bits score. Among the *Arabidopsis* sequence homologues, 2522 of them are also predicted to be involved in 3D-domain swapping.

In the *Arabidopsis* genome, a total of 7694 (64%) protein sequences were predicted as 3D-domain swapping, out of 12 033 reviewed sequences from UniProt and 3029 (25.2%) proteins are associated with different stress conditions. Proteins, which are common to domain swapping and stress are 1158 (9.6%) in number (Figure 3; Supplemental Table 1). Out of 1158 protein sequences, predicted to be involved in 3D-domain swapping and also observed in abiotic stress, only 69 of them are observed to retain homologues of known structure.

Few of the important examples of well-known cases of 3D-domain swapping in plants (Table 2), and also found as predicted in Tulsi genome, are Bleomycin resistance protein (1BYL), serine/threonine phosphatase activity (3NMT), glyoxalase/bleomycin resistance protein (1TG5), dimerization domain protein binding (3C6N), and glyceraldehyde-3-phosphate dehydrogenase chloroplastic (3RVD).

Bleomycin resistance protein is from Glyoxalase/Bleomycin resistance protein/dihydroxybiphenyl dioxygenase superfamily. It belongs to alpha and beta proteins class ($\alpha + \beta$) of SCOP,[20] which has a β-α-β type of fold. There are 10 families in this superfamily. Three Pfam[21] protein domain families are associated with this superfamily: these are PF13669 (Glyoxalase_4), PF00903 (Glyoxalase), and PF06983 (3-dmu-9_3-mt). These entire protein domain families are from the same clan, ie, Glyoxalase clan (CL0104). This clan is known to have enzymes which catalyze isomerization, epimerization, oxidative cleavage of C–C bond, and nucleophilic substitution reactions. Bleomycin-resistant protein is of 14 kDa and consists of compact dimer with a hydrophobic interface involved in mutual chain exchange. The hinge region is of 3 residues (residue numbers 8V, 9P, and 10V). The hinge region and swapped domain are marked in red and green color, respectively. Another example of predicted protein engaged in 3D-domain swapping from Tulsi genome and also reported earlier to be involved in this phenomenon among plant homologues is glyceraldehyde-3-phosphate dehydrogenase (GAPA). In plants, there are 3

Table 2. Some of the plant protein crystal structures with 3D-domain swapping.

S. NO.	MONOMER	DOMAIN SWAP	DESCRIPTION
1	1GNU	1WZ3	Ubiquitin-like
2	1G6J	1GJZ	Ubiquitin-like
3	1KMZ	1XY7	Glyoxalase/bleomycin resistance
4	1GQ9	1W77	Nucleotide-diphospho-sugar transferases
5	1KL7	1E5X	Threonine synthatase
6	1X91	1X8Z	Plant invertase/pectin methylesterase
7	–	1Z84	Galactose-1-phosphate uridyltransferase-like
8	–	1MLV	Ribulose-1,5 bisphosphate
9	2A5V	1EKJ	Beta-carbonic anhydrase
10	–	1L3A	Plant transcriptional regulator pbf-2
11	–	3A8R	Respiratory burst NADPH oxidase
12	–	1Z7W	Cysteine synthase
13	–	2Q48	Protein AT5G48480
14	–	2NTX	EMB|CAB41934.1
15	–	2AAO	Calcium-dependent protein kinase
16	–	2Q4H	Nucleocapsid protein
17	–	2066	PII protein
18	–	1Z7Y	Cysteine synthase
19	–	2PC5	DUTP pyrophosphate-like protein
20	–	2P90	DUTP pyrophosphate-like protein
21	–	1MLV	Serine/threonine-protein kinase 10
22	–	2BHW	Chlorophyll a-b binding protein AB80

types of GAPA, namely, GAPA1, 2, and 3. GAPA1 helps in the reduction of 1,3-diphosphateglycerate by nicotinamide adenine dinucleotide phosphate (NADPH). The GAPA1 is of 396 residues long with a molecular mass of 42.4 kDa. These proteins belong to 2 Pfam families, ie, PF02800 (glyceraldehyde-3-phosphate dehydrogenase, C-terminal) and PF00044 (glyceraldehyde-3-phosphate dehydrogenase, NAD-binding domain).

Conclusions

Nearly 25% of the protein sequences in Tulsi genome are predicted to be involved in the mechanism of 3D-domain swapping of protein oligomerization by Random Forest method. Nearly 98% of these predicted sequences have homologues in different plants, with majority in *A thaliana*. Functional annotations for almost 90% of the sequences could be assigned. The GO term associations could be performed for 8027 (85%) of Tulsi sequences predicted to be involved in domain swapping, and protein domain association could be observed for 8339 (89%) of them. Those sequences (85%), which have GO term associations, also have protein domain family associations from Pfam database. Some 1158 (12% of domain-swapped predicted) protein sequences are involved in both the phenomenon of 3D-domain swapping and in abiotic stress.

These sequences are involved in many diverse biological functions such as

1. *Defense mechanism of plants against biotic stress*. Proteins involved in this mechanism help in the production of secondary metabolites, known as allelochemicals, which acts as plant defense against herbivores by changing the behavior, growth, or survival of herbivores.
2. *Against abiotic stress like heat-shock, cold*. In case of abiotic stresses, there is production of a group of proteins called heat-shock proteins (Hsps) or stress-induced proteins. On the basis of molecular weight of these proteins, they are grouped into 5 classes in plants: (1) Hsp100, (2) Hsp90, (3) Hsp70, (4) Hsp60, and (5) small heat-shock proteins (sHsps). It is reported that because of this high

diversification of these proteins in plants, they have an adaptation to tolerate the different stresses such as heat and cold stress.

3. *Involved as transcription factors to regulate the function of many genes*. Transcription factors are proteins that regulate gene expression. These proteins bind to specific sites of gene and works as regulatory elements such as enhancer. Plants alter transcription and gene expression levels during development and in response to environmental conditions.

Some of the examples of the protein sequences involved in 3D-domain swapping and also involved in defense mechanism against biotic stress and transcription factors, as concluded in the above section, are listed in Supplemental Table 2.

Author Contributions

RS conceived the project ideas. AU carried out all the work and analyses. AU wrote first draft of the manuscript and RS critically read to improve it.

ORCID iD

Ramanathan Sowdhamini https://orcid.org/0000-0002-6642-2367

REFERENCES

1. Goodsell DS, Olson AJ. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*. 2000;29:105-153.
2. Nagradova NK. Three-dimensional domain swapping in homooligomeric proteins and its functional significance. *Biochem*. 2002;67:839-849.
3. Crestfield AM, Stein WH, Moor S. On the aggregation of bovine pancreatic ribonuclease. *Arch Biochem Biophys*. 1962;Suppl 1:217-222.
4. Bax B, Lapatto R, Nalini V, et al. X-ray analysis of beta B2-crystallin and evolution of oligomeric lens proteins. *Nature*. 1990;347:776-780.
5. Bennett MJ, Choe S, Eisenberg D. Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci U S A*. 1994;91:3127-3131.
6. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. *Protein Sci*. 2002;11:1285-1299.
7. Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci*. 1995;4:2455-2468.
8. Standfuss J, Terwisscha van Scheltinga AC, Lamborghini M, Kühlbrandt W. Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 A resolution. *EMBO J*. 2005;24:919-928.
9. Trievel RC, Beach BM, Dirk LMA, Houtz RL, Hurley JH. Structure and catalytic mechanism of a SET domain protein methyltransferase. 2002;111:91-103.
10. Vieira R, Grayer R, Paton A, Simon J. Genetic diversity of Ocimum gratissimum L. based on volatile oil constituents, flavonoids and RAPD markers. *Biochem Syst Ecol*. 2001;29:287-304.
11. Cohen M. Tulsi - Ocimum sanctum: a herb for all reasons. *J Ayurveda Integr Med*. 2014;5:251.
12. Upadhyay AK, Chacko AR, Gandhimathi A, et al. Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties. *BMC Plant Biol*. 2015;15:212.
13. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res*. 2000;28:374.
14. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using WEKA. *Bioinformatics*. 2004;20:2479-2481.
15. Upadhyay AK, Sowdhamini R. Genome-wide prediction and analysis of 3D-domain swapped proteins in the human genome from sequence information. *PLoS ONE*. 2016;11:e0159627.
16. Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biol*. 2013;14:R134.
17. Naika M, Shameer K, Mathew OK, Gowda R, Sowdhamini R. STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in arabidopsis and rice. *Plant Cell Physiol*. 2013;54:e8.
18. Mulder NJ, Apweiler R, Attwood TK, et al. InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*. 2002;3:225-235.
19. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32: D115-D119.
20. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247:536-540.
21. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2009;38:D211-D222.