THE PROTEIN SOCIETY

Combining protein sequence, structure, and dynamics: A novel approach for functional evolution analysis of PAS domain superfamily

Zheng Dong, Hongyu Zhou, and Peng Tao*

Department of Chemistry, Center for Drug Discovery, Design, and Delivery (CD4), Center for Scientific Computation, Southern Methodist University, Dallas, Texas, 75275

Received 11 April 2017; Accepted 15 October 2017
DOI: 10.1002/pro.3329
Published online 20 October 2017 proteinscience.org

Abstract: PAS domains are widespread in archaea, bacteria, and eukaryota, and play important roles in various functions. In this study, we aim to explore functional evolutionary relationship among proteins in the PAS domain superfamily in view of the sequence-structure-dynamics-function relationship. We collected protein sequences and crystal structure data from RCSB Protein Data Bank of the PAS domain superfamily belonging to three biological functions (nucleotide binding, photoreceptor activity, and transferase activity). Protein sequences were aligned and then used to select sequence-conserved residues and build phylogenetic tree. Three-dimensional structure alignment was also applied to obtain structure-conserved residues. The protein dynamics were analyzed using elastic network model (ENM) and validated by molecular dynamics (MD) simulation. The result showed that the proteins with same function could be grouped by sequence similarity, and proteins in different functional groups displayed statistically significant difference in their vibrational patterns. Interestingly, in all three functional groups, conserved amino acid residues identified by sequence and structure conservation analysis generally have a lower fluctuation than other residues. In addition, the fluctuation of conserved residues in each biological function group was strongly correlated with the corresponding biological function. This research suggested a direct connection in which the protein sequences were related to various functions through structural dynamics. This is a new attempt to delineate functional evolution of proteins using the integrated information of sequence, structure, and dynamics.

Keywords: domain superfamily; elastic network model (ENM); conservation analysis; sequence alignment; structure alignment

Abbreviations: ENM, elastic network model; MD, molecular dynamics; NMA, normal mode analysis

Additional Supporting Information may be found in the online version of this article.

Statement: How protein functions are evolved in a protein superfamily is an important and interesting question. In this study, we attempt to answer this question by integrating the information of protein sequence, structure and dynamics. Our research suggested a direct connection in which the protein sequences were related to various functions by changing structural dynamics. This could serve as a new approach to study functional diversity in protein superfamily using the integrated information of sequence, structure and dynamics.

Grant sponsor: Edward R. Biehl Graduate Fellowship (HZ), and Ralph E. Powe Junior Faculty Enhancement Award (PT). Computational time was provided by Southern Methodist University's Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin.

*Correspondence to: Peng Tao, Department of Chemistry, Southern Methodist University, 3215 Daniel Avenue, Dallas, TX 75275-0314. E-mail: ptao@smu.edu

Introduction

PAS domains can be found in all domain of life, including archaea, bacteria and eukaryota[1] and comprise of more than 6000 proteins based on multiple-sequence alignment analysis.[2] They were named by combining the first letter of the period clock protein in Drosophila (PER), aryl hydrocarbon receptor nuclear translocator in vertebrate (ARNT), and Single-minded protein in Drosophila (SIM), which are proteins in this family.[3,4] PAS domains normally contain around 100 residues and are conserved in three-dimensional (3D) structure.[5,6] PAS domains play important roles in many biological processes,[6] such as transducing regulatory signal in many cellular processes through binding events.[4,7] Protein photoactive yellow protein (PYP) and the light-oxygen-voltage sensing (LOV) domains are two common types of PAS domains and act as protein sensors with blue light.[7,8]

Recently, as an increasing number of PAS domains have been found,[9] more studies were carried to investigate the biological functions and mechanisms of PAS domains by analyzing their sequences and structures.[6,9,10] For instance, structure alignment for 63 PAS proteins showed that the structural relation among PAS domains could be explained by their location to cell membrane and their binding ligands.[6] However, the information from sequences and structures may be insufficient to fully understand the relationship between structure and function in this protein family. Dynamics as an important property of protein serve as critical connection between structures and functions.[11] It could also be a key factor contributing to the mechanisms of protein biological functions. Recently, the vibrational motion at secondary structure level was shown to be influential on the PAS domain functions through dynamical analysis.[12–15] In these studies, molecular dynamics (MD) simulations also revealed different stability/flexibility of different parts. For example, the central alignment part of six PAS domains (HERG, phy3, PYP, FixL, hPASK and HIF-2α) containing α-helix was revealed to be very flexible.[12] Therefore, it is necessary to bridge between structure and function using information from dynamics.[11]

Elastic network model (ENM), as a coarse-grained normal mode analysis (NMA),[16,17] is a powerful tool to characterize dynamics of biomolecules based on their crystallographic structures.[11,18] Recently, dynamics were shown through ENM analyses to be important to further understand protein functions including catalysis and allostery, as well as evolution of proteins.[18–22] ENM predicts the dynamics of protein using much fewer parameters and lower computational cost than all atomic force field models.[18] The low frequency normal modes from NMA are sufficient to show the intrinsic collective motions in proteins[11] and thus ENM is suitable to describe the functional dynamics of proteins.[23] Previous studies indicated that the protein motions calculated by ENM show a good agreement with the results from experimental observation of protein dynamics.[24] Therefore, ENM have been widely used to compare the dynamic patterns among multiple proteins. In the present study, ENM were employed to compare functional motions and construct the evolutionary relationship among proteins in PAS domain superfamily. Evolution analysis of protein sequences and structures has been widely applied as a powerful computational tool.[25] And protein dynamics have also been proven to play an important role in protein evolution.[19–22,26] Therefore, in this study sequence and structure alignment combining with ENM were used to delineate functional evolution of proteins by integrating sequence, structure and dynamics information.

We aimed to explore the functional evolution of PAS domain superfamily using a five-step approach. (1) We collected protein sequences and crystal structure data of PAS domain superfamily from RCSB Protein Data Bank divided into three groups based on their functions (Group_1: nucleotide binding, Group_2: photoreceptor activity, and Group_3: transferase activity). (2) Protein sequences were aligned and then used to select sequence-conserved residues and to build phylogenetic tree showing the evolutionary relationship among different functional groups. (3) In each functional group, structure alignment was also applied to obtain structure-conserved residues. Both sequence- and structure-conserved residues are identified as conserved sites. (4) The fluctuation difference in conserved residues and other residues was illustrated through comparison of ENM results. (5) Finally, the ENM results were validated through comparison with fluctuation results via molecular dynamics (MD) simulation, and the correlation between conserved sites' fluctuation and biological functions is revealed. Using this approach, this study provides insight into the functional evolution of PAS domain superfamily.

Results

Proteins clustered together by their biological functions

The sequences of the selected proteins were used to determine evolutionary relationships among PAS domain proteins in different functional groups. The phylogenetic tree was constructed by maximum likelihood method and shown in Figure 1. The selected proteins from PAS domain superfamily were clustered together based on the sequence alignment. To further analyze the relation between protein sequences and evolution of PAS functions, the sequence-conserved residues were selected from alignment result (the number of the sequence-conserved residues: Group_1: 80, Group_2: 67, and Group_3: 64). In addition, structure-conserved sites were also

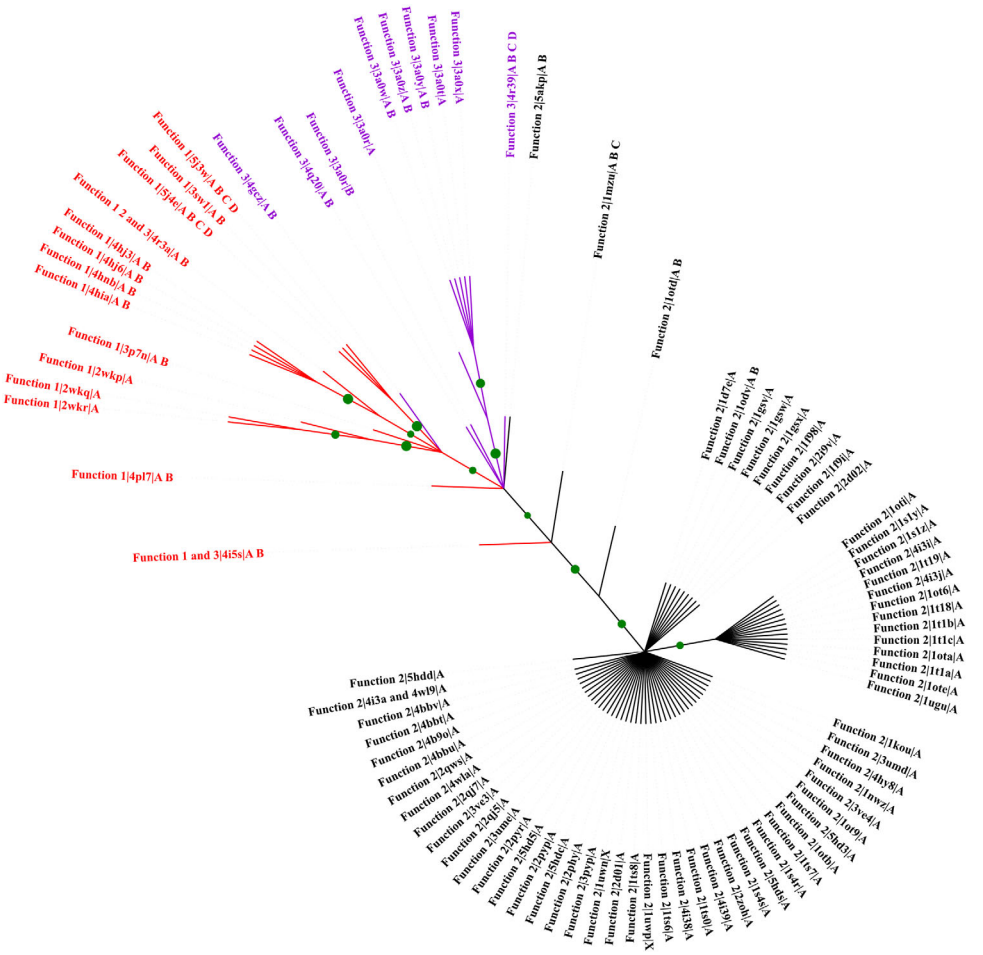

Figure 1. The phylogenetic tree for PAS domain superfamily. Function 1: nucleotide binding, Function 2: photoreceptor activity and Function 3: transferase activity.

identified based on the structure alignment (the number of the structure-conserved residues: Group_1: 34, Group_2: 65, and Group_3: 96). Residues that are both sequence-conserved and structure-conserved are referred as conserved site (the number of the conserved sites: Group_1: 14, Group_2: 33, and Group_3: 51) (see Supporting Information Table SII). All other residues are referred as normal sites. The conserved sites are highlighted in red in representative protein structures illustrated in Figure 2 and Supporting Information Figure S1. Most of the conserved sites are located in the beta-sheets of proteins with few in loops, especially in groups 1 and 2, suggesting that the special distribution of conserved sites in protein secondary structure might play a central role in evolutionary process of PAS domain superfamily.

Conserved sites have lower fluctuations

In the present study, we compare the fluctuation between conserved sites and normal sites in three different ways. One is the comparison of the averaged fluctuation of the conserved sites as well as the normal sites at residue level. The result showed that the conserved sites had significantly lower fluctuations

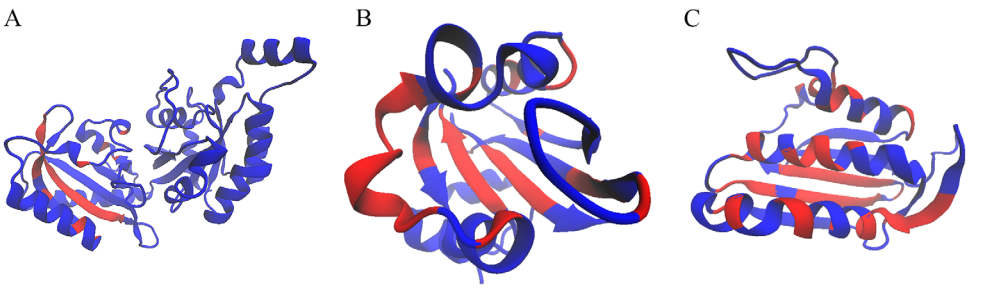

Figure 2. The conserved residues in the secondary structure of proteins. A, 2WKP in Group_1 (nucleotide binding); B, 1F9I in Group_2 (photoreceptor activity); and C, 3AOT in Group_3 (transferase activity), respectively.

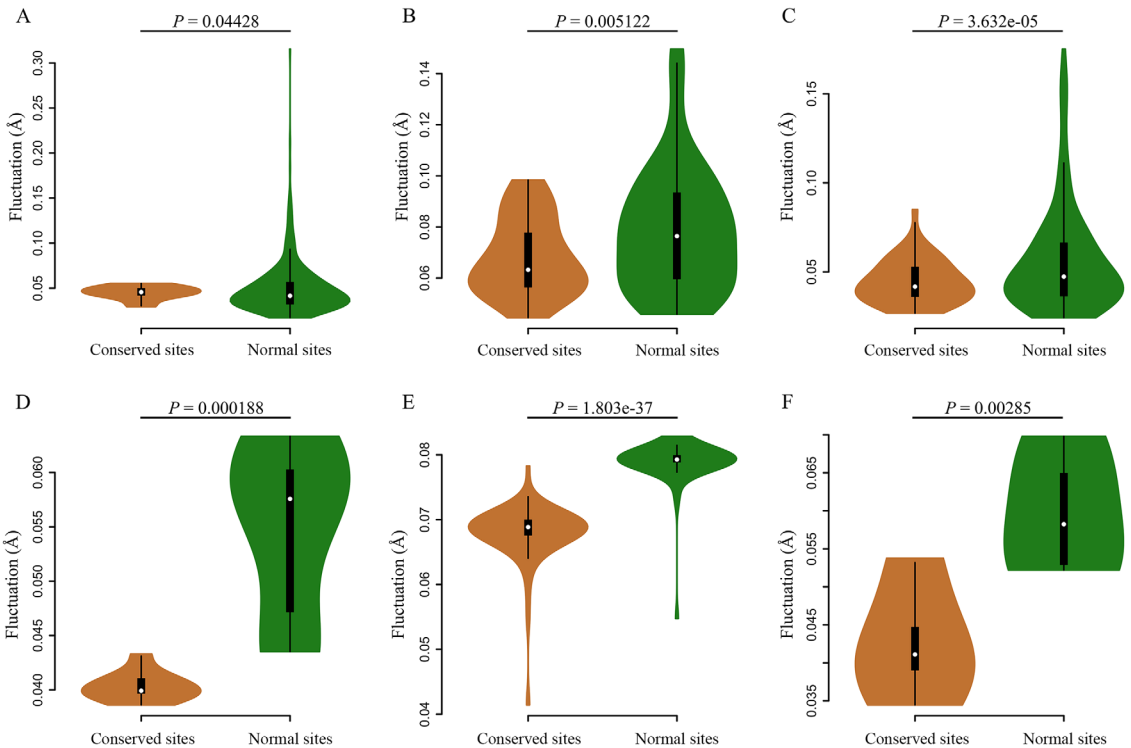

Figure 3. The comparison of fluctuation between conserved sites and normal sites by elastic network model. A, B, and C are the comparison in residue level among Groups 1 (nucleotide binding), 2 (photoreceptor activity) and 3 (transferase activity), respectively. D, E, and F are the comparisons in PDB structure level among Groups 1, 2 and 3, respectively. P value less than 0.05 was considered statistically significant.

than the normal sites in all three biological functional groups of PAS domain superfamily with *P* value of 0.04, 5.12E-03 and 3.63E-05, respectively [Fig. 3(A–C)]. Another is the comparison of the averaged fluctuation at individual protein structure level among various groups. Results also indicated significantly lower fluctuations in the conserved sites than in the normal sites ($P = 1.88E-04$, 1.80E-37 and 2.85E-03, respectively) [Fig. 3(D–F)]. Finally, to validate the above results, MD simulations were also carried out, and the results have a good agreement with the ENM results with *P* value of 2.22E-05, 1.05E-03 and 4.48E-05, respectively (Fig. 4). In summary, the conserved sites in PAS domain proteins are associated with the protein dynamics and have significantly lower fluctuations than the normal sites.

Fluctuation patterns in different biological functions

The fluctuation patterns along the aligned sequence of three selected biological functional groups in PAS domain superfamily are plotted in Figure 5. For each group, the averaged fluctuation was calculated for each aligned residue and used for the plot. In general, the conserved sites are shared by more proteins than the normal sites in each functional group.

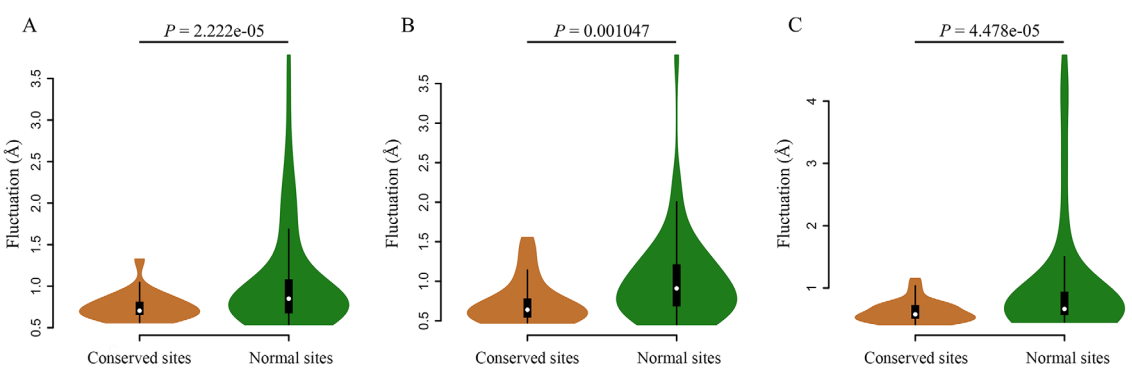

Figure 4. The comparison of fluctuation between conserved sites and normal sites by molecular dynamics simulations. A, B, and C were performed in Groups 1 (nucleotide binding), 2 (photoreceptor activity) and 3 (transferase activity), respectively. *P* value less than 0.05 was considered statistically significant.

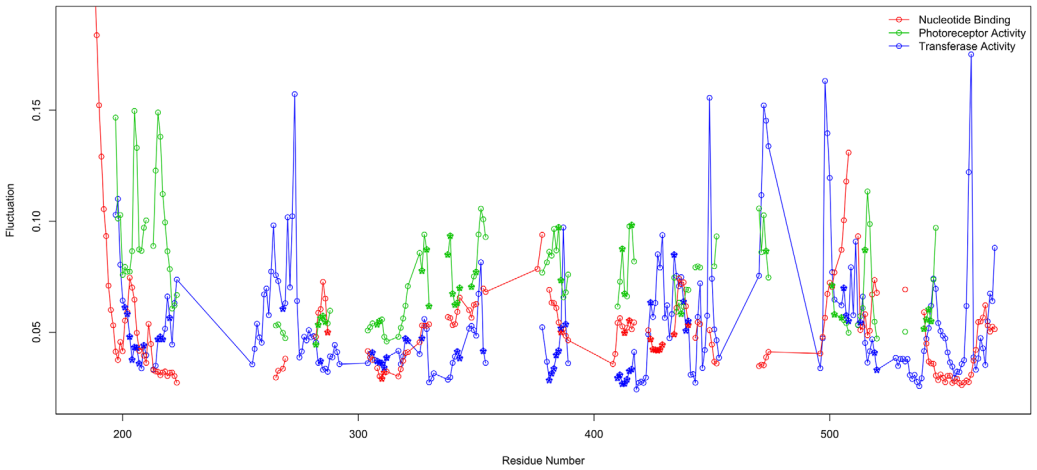

Figure 5. The fluctuation patterns in different biological functions. The conserved residues identified in this study are labeled using the symbol "*".

The results showed that the fluctuation patterns vary significantly in different biological function groups. Residues' fluctuation in function of photoreceptor activity (Group 2, green plot in Fig. 5) is the largest, and some of the transferase activity group conserved sites have the lowest fluctuation comparing to the conserved sites in other two groups (Fig. 5). The conserved sites in each group are marked by asterisk in Figure 5 and showed relative lower fluctuations than others. As a summary, each biological function in PAS domain superfamily is likely to adopt a specific residue fluctuation pattern.

The correlation between conserved sites fluctuation and biological function

The overall goal of this study is building connections among different aspects of PAS domain proteins: sequence, structure, dynamics, and functions. Therefore the correlation among these aspects is explored through comparison of fluctuation of conserved sites associated with each functional group. The comparison of conserved sites fluctuation was carried out between each pair of group and plotted in Figure 6. The computational details of these plots are described in the functional fluctuation analysis section in the Materials and Methods part of the article.

The plots in Figure 6 show that the proteins with the same function are clustered together based on the averaged fluctuation of function-specific conserved sites, and the different clusters are clearly separated. Regarding the comparison of Group_1 and Group_2 corresponding conserved residues, Group_1 proteins show lower fluctuations in the Group_1-specific conserved sites (averaged fluctuation: 0.044) than in the Group_2-specific conserved sites (average fluctuation: 0.056), while Group_2 proteins have similar averaged fluctuations of the Group_1- and Group_2- specific conserved sites (averaged fluctuation: 0.066 vs. 0.068) [Fig. 6(A)]. In the comparison of Group_1- and Group_3-specific conserved sites, proteins in Group_1 show lower fluctuations in the Group_1-specific conserved sites

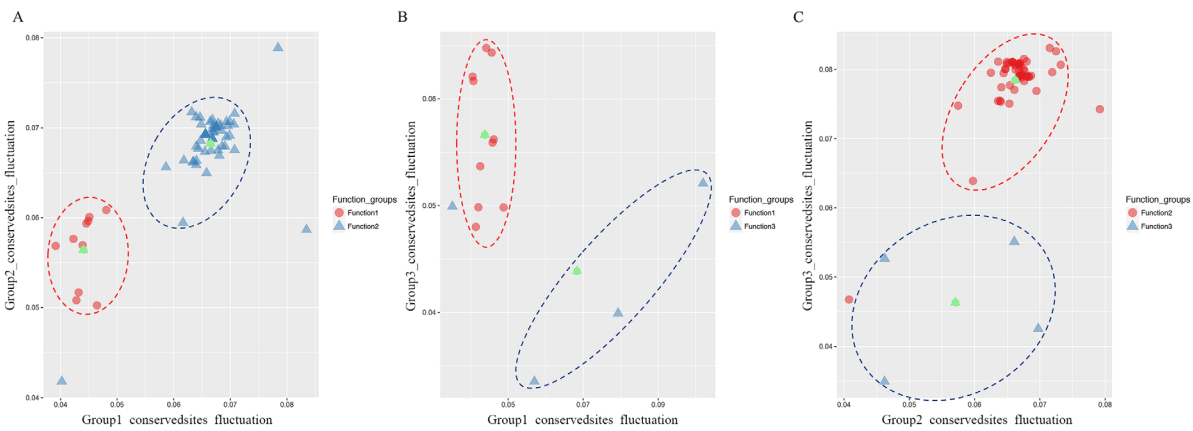

Figure 6. Functional clustering analysis of proteins based on conserved sites' fluctuations. A, B, and C were performed in Groups 1 versus 2, Groups 1 versus 3 and Groups 2 versus 3. Group 1: nucleotide binding; 2: photoreceptor activity; 3: transferase activity. The center of each cluster was calculated and labeled in a green color symbol.

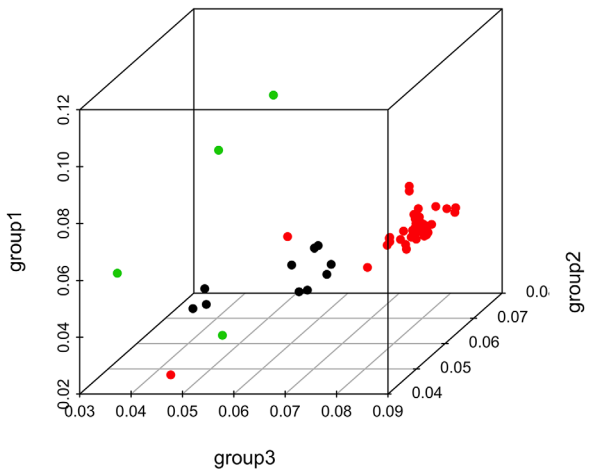

Figure 7. The three-dimensional cluster of proteins based on conserved sites' fluctuations. Group_1 (nucleotide binding); Group_2 (photoreceptor activity), and Group_3 (transferase activity), respectively.

than in the Group_3-specific conserved sites (averaged fluctuation: 0.044 vs. 0.056) and Group_3 proteins have lower fluctuations in the Group_3-specific conserved sites than in the Group_1-specific conserved sites (averaged fluctuation: 0.044 vs. 0.068) [Fig. 6(B)]. In the comparison of Group_2- and Group_3-specific conserved sites, proteins also show lower fluctuations in their corresponding group-specific sites (averaged fluctuation: Group_2, 0.066 vs. 0.078 and Group_3, 0.057 vs. 0.046) [Fig. 6(C)], which is consistent with the above results that the group-specific core sites have smaller motions than all other sites (Figs. 3 and 4). All three groups were also compared together, and a 3D plot shows distinct clusters based on the fluctuations of conserved sites (Fig. 7). These analyses suggest that the conserved site fluctuation is significantly correlated with the generation of PAS proteins' functions.

Discussion

PAS domain superfamily contains proteins that exist widely in living organisms[1] and contribute to many biological functions.[6] Many studies were carried out to analyze the biological functions and mechanisms of PAS domain superfamily using sequences and structures.[6,9,10] The information from protein sequences and structures is helpful but limited since protein dynamics play crucial roles to different protein functions.[11] ENM is an efficient computational tool to describe the functional dynamics of proteins[23] and helpful to bridge between structure and function using the protein dynamical information.[11] In addition, many studies suggested a strong correlation between sequence evolution and structural motions.[27,28] Therefore, it is important to explore the functional evolution of proteins with the integrated analyses of sequence, structure and dynamics.

In this study, we illustrated probable functional evolution of PAS domain proteins by integrating analyses of sequence, structure, and dynamics, suggesting a potential evolution correlation from protein sequence to function. The sequences and functions could be linked by two factors: conserved sites and protein dynamics. The correlation among protein sequence, conserved site, dynamics, and biological function for PAS domain superfamily are elucidated. In addition, we identified the relationship between conserved sites and dynamics using both ENM and MD simulations.

Although protein sequences in the PAS domain superfamily are highly diverse,[29] we constructed a phylogenetic tree to show the functional evolution of PAS domain superfamily (Fig. 1). Those proteins with the same biological function are clustered together and significantly distinguished from other functional proteins, suggesting that the same functional proteins share the same conserved residues belonging to the common evolutionary legacy. Therefore, those shared residues were selected in this study for dynamical analysis.

3D structure alignment for PAS domains was proven to be a powerful tool to study the function diversity.[6] Therefore, we carried out 3D structure alignment and combined these results with sequence alignment to obtain the candidate function-associated residues (called conserved sites in our study, see Supporting Information Table SII). Although one previous study showed that the buried sites are likely to be in hydrophobic core of proteins and have a higher conservation due to the local environment of protein core,[22] and some conserved sites identified in this study are hydrophobic core sites of proteins (54.54% in a blue light photoreceptor protein of Group_1 (PDB code: 5J3W); 32.35% in a photoactive yellow protein of Group_2 (PDB code: 1OTD) and 49.02% in a histidine kinase of Group_3 (PDB code: 3A0X)), many other conserved sites identified in this study not belonging to hydrophobic core still show high conservation in evolutionary process. This might be due to their potential role in controlling protein biological functions.

The conserved sites identified in this study, including hydrophobic core and hydrophilic residues, have a good agreement with the previous studies. Regarding the conserved sites related to nucleotide binding function, S98 and T100 affect the dimer interaction of LOV domain (PDB code: 3SW1) by modifying H-bonding interactions,[30] suggesting conserved sites might contribute to nucleotide binding function through the modification of dimer interaction. The key residues (R451, F494, N482 and N492 for photoactivatable Rac1, PDB code: 2WKP) interacting with mononucleotide substrates were also identified as conserved sites in the present study. The conserved sites associated with photoreceptor

are also important. For example, one mutation from arginine to glutamine in residue 52 (PDB code: R52Q) of photoactive yellow protein can elevate the chromophore pKa by one pH unit and lose two hydrogen bonds formed by R52 with T50 and Y98.[31] Residue R52 also plays a role as a lid on the binding pocket of chromophore and influences the exterior solvent accessibility.[31] Therefore, mutation at R52 are likely responsible for the regulation of pKa and thus might control photocycle process.[31] This is consistent with the result in the present study. Because the number of crystal structures of PAS proteins with transferase activity is limited in PDB, part of structure-conserved sites could come from false positive and lead to a relative larger number of conserved sites in Group_3 (51 residues). However, many conserved sites found in the study have been reported to affect their conformations and functions. For example, K710 might influence the side chain entering the pocket for nucleotide-binding and facilitate the electrostatic interaction between β-phosphate and E664 residue carboxylate.[32] Another conserved site T709 plays a key role in stabilizing the binding of ADP by hydrogen bond interaction.[32] Above all, the conserved sites identified in this study can influence their biological function through changing conformations and interactions.

PAS domain superfamily conserved sites have lower fluctuation than other sites (Figs. 3 and 4). This is consistent with the results that the most conserved residues have decreased mobility in other proteins.[26,33] One reason behind this could be that the low fluctuation residues increase the thermostability of protein.[33] The other one could be that the conserved residues are critical for proteins to maintain low fluctuations in their active states. For example, a previous study based on Gaussian network model showed that proteins had lower structural fluctuations in active states than in inactive states.[34] Another study about PAS domains stated that β-strands had lower mobility than the inter-strand loops.[12] Most of the conserved sites identified in this study are in β-sheet with very few in loops, agreeing with the observation of low fluctuation in conserved residues. In the present study, we analyzed the relationship between conserved sites and dynamics using elastic network model (Fig. 3) with validation based on molecular dynamics simulations (Fig. 4). Our overall results indicate that the conserved sites in PAS domain superfamily have a lower fluctuation than other residues. This could be utilized as a new way to identify function-associated residues in proteins. Furthermore, the function-associated residues contributed to the diverse fluctuation patterns in different biological functions (Fig. 5).

By building the relations among protein sequence, structure, dynamics, and biological function for PAS domain superfamily, we aimed to construct a theoretical framework integrating protein sequences, structures, and dynamics to biological functions in this study. The results showing a specific cluster of proteins with the same function (Figs. 6 and 7) are in a good agreement with the previous result from sequence evolution analysis. The outliers for specific clusters are one form of photoactive yellow protein (PDB code: 2D02) [Fig. 6(A,C)] and a blue-light activated histidine kinase (PDB code: 4R39) [Fig. 6(B)]. A critical functional mutant R52Q in the photoactive yellow protein might contribute to the difference from other proteins. The selected histidine kinase is an isolated DHp/CA catalytic construction and not the full-length protein, which might also affect its structural movements.

Overall, this study suggests an important role of conserved sites dynamics as the connection between sequence and the biological function generation in PAS domain superfamily. It also implies the importance of integrating the information of sequence, structure, and dynamics for protein functional and evolutionary analysis.

Conclusion

Our study is a new approach to delineate functional evolution of protein superfamily using the integrated information of sequence, structure and dynamics. This study systematically revealed the key role of dynamics on the potential path of PAS domain superfamily from sequence to biological function, and suggested a direct connection from protein sequences and structure to dynamics and functions. In addition, the present research also revealed that the conserved residues in evolution and function showed a lower fluctuation than other residues. Our study sheds light on the understanding of PAS domain superfamily and provides a new insight to study the functional evolution of protein superfamily.

Materials and Methods

Data collection

The crystallographic structures of PAS domain superfamily were searched in the Protein Data Bank (PDB) in Europe (http://www.ebi.ac.uk/pdbe/). A total of 344 proteins were found. Both sequences and structure data were downloaded. Based on the selection of protein functions, 89 proteins belonging to three biological function groups (Group_1: nucleotide binding, Group_2: photoreceptor activity and Group_3: transferase activity) were selected in our study. Other proteins with different biological functions were not considered due to the limited number of proteins in those functional groups. The detailed information of the candidate protein including data quality, resolution, gene name, gene ontology id and super kingdom, is presented in Supporting Information Table SI.

Sequence alignment, phylogenetic analyses, and conservation analysis

All coding sequences of proteins in three biological function groups were aligned by using MUSCLE

program in MEGA 7.0.[35,36] The default parameters in MUSCLE were used for the analysis. The aligned sequences were applied to construct a maximum likelihood (ML) tree in Poisson model.[35,37] ML trees were reconstructed by bootstrap analysis with 1000 replication.[35,37] Finally, The phylogenetic tree was visualized by iTOL (http://itol.embl.de).[38]

Furthermore, the aligned sequences in each biological function group were used to do residue conservation analysis through the Protein Residue Conservation Prediction website (http://compbio.cs.princeton.edu/conservation/score.html) with Shannon entropy scoring method.[39] After scoring, residues with conservation scores among top 10%, which are measured according to the evolutionary relations of proteins and the substitution rate of one residue to another residue in amino acid substitution table[40] (averaged conservation score of all sites: Group_1, 0.239; Group_2, 0.460; and Group_3, 0.212; averaged conservation score of selected sites: Group_1, 0.507; Group_2, 0.730; and Group_3, 0.403), and with fewer than 3 gaps in aligned sequences were considered as sequence-conserved sites.

Structure alignment

Multiple structural alignments of protein structures in each biological function group were performed by MultiProt,[41] which is a powerful tool for multiple protein structure alignments.[41] In MultiProt method, multiple structural superposition is carried out for input proteins to maximize the number of aligned residues among these proteins. The alignment with the largest number of aligned residues was used for further analysis. These structurally aligned residues are referred as structure-conserved residues. Residues belonging to both of sequence-conserved residues and structure-conserved residue are grouped together and referred as conserved sites. All other residues are referred as normal sites.

Elastic network models analysis

The elastic network model (ENM) implemented in CHARMM[42,43] was employed to approximate protein dynamics. ENM treats macromolecule as a network of beads connected by Hookean springs.[44] Each amino acid residue is represented as a bead centering at its Cα atom. The twenty lowest frequency modes corresponding to large-scale protein motions were selected to calculate the averaged fluctuation value for each residue using the following equation

$$Flu_{\mathrm{res}} = \frac{\sum_{m=1}^{20} Flu_m \frac{1}{Eig_m}}{\sum_{m=1}^{20} \frac{1}{Eig_m}}.$$

Flu_{res} and Flu_m are the averaged fluctuation for each residue and the amplitude for each mode, respectively. Eig_m is the mth lowest eigenvalue among all vibrational normal modes using ENM. 67 out of 89 selected protein structures from PDB were subjected to ENM analysis. The remaining 22 structures were not considered due to the incompleteness of the structure. The comparison of the averaged fluctuation between conserved sites and normal sites was calculated in two ways: residue-based and protein structure-based. The residue-based fluctuation is the averaged fluctuation of either conserved or normal sites across different proteins in one specific functional group. The protein structure-based fluctuation is the averaged fluctuation of conserved or normal sites in each protein structure in one specific functional group. These two ways of fluctuation calculation were designed to explore the correlation of residues across different proteins in one group and of residues in each individual protein.

MD simulations and root-mean-square fluctuation (RMSF) calculation

MD simulations were conducted to validate the result of our analysis using ENM. One protein was randomly selected for each group for validation purpose: a blue-light photoreceptor protein (PDB code 5J3W)[45] in Group_1, a histidine kinase ThkA (PDB code 3A0X)[32] in Group_2, a photoactive yellow protein (PDB code 1OTD)[46] in Group_3. The initial protein structures were processed by adding missing hydrogen atoms and solvated using explicit water model (TIP3P).[47] Sodium cations and chlorine anions were added to systems to balance the overall charge and maintain ionic concentration of simulation box around 100 mM. Afterwards, those simulation systems were subjected to energy minimization with 200 steepest descent steps, and sufficient adopted-basis Newton-Raphson (ABNR) minimization steps, which yielded a total gradient of less than 0.03 kcal/(mol•Å). Each simulation box was subjected to 12 picoseconds equilibrium with temperature raising from 100K to 300K, followed by 12 nanoseconds isothermal-isobaric (NPT) MD simulation. The 12 nanoseconds simulation trajectory was used for RMSF analysis, which is the measurement of averaged atomic fluctuation during MD simulation.[48] Three independent MD simulations were conducted for each protein. The RMSF value of each residue was calculated based on these MD simulations.

The cubic periodic boundary conditions were used in all three simulations. Particle mesh Ewald method was applied to calculate electrostatic interactions.[49] CHARMM version c40b1 was used to carry out all the simulations with CHARMM force field of version 27.[43]

Functional fluctuation analysis

To characterize conserved sites specific to each group for the comparison, any shared conserved sites are excluded for the analysis. For example, Group_1 has

14 conserved sites. Two of these conserved sites are shared with Group_2. Therefore, the remaining 12 conserved sites specific to Group_1 (with reference to Group_2) were used for residue fluctuation analysis (see Supporting Information Table SIII). Similar to Group_2, after excluding the two shared conserved sites with Group_1, the remaining 31 conserved sites specific to Group_2 were used for residue fluctuation analysis. The similar process was carried out for Group_1:Group_3 and Group_2:-Group_3 comparison, respectively. The residues used for this analysis are referred as function-specific conserved sites. When comparing Group_1 and Group_2, the averaged fluctuation of function-specific conserved sites for Group_1 (total of 12) and the averaged fluctuation of function-specific conserved sites for Group_2 (total of 31) are calculated for each protein and plotted in Figure 6(A). Therefore, for Group_1 proteins plots [red round plots in Fig. 6(A)], 12 function-specific conserved sites for Group_1 are actually conserved sites. But 31 function-specific conserved sites for Group_2 are not conserved for the proteins in Group_1. It should be noted that not all proteins in Group_1 contain residues corresponding to all 31 function-specific conserved sites for Group_2. Therefore, for each protein in Group_1, the averaged fluctuation was calculated for the only existing residues corresponding to the function-specific conserved sites for Group_2. It is similar for Group_2 proteins plots [blue triangle plots in Fig. 6(A)]. 12 function-specific conserved sites for Group_1 are not conserved sites for Group_2 proteins. But 31 function-specific conserved sites for Group_2 are conserved. It is a reminder that these residues are identified through sequence comparison among all selected proteins.

Statistical analysis

Data on residue fluctuations are illustrated as violin plots. The differences in fluctuation between conserved sites and normal sites were calculated by Student's *t*-test. *P* value below 0.05 was considered statistically significant. All statistical analyses in this study were performed using R script (Version 3.3.0: www.r-project.org/)

Conflict of Interest Statement

The authors have declared that no competing interests exist.

References

1. Repik A, Rebbapragada A, Johnson MS, Haznedar JÖ, Zhulin IB, Taylor BL (2000) PAS domain residues involved in signal transduction by the Aer redox sensor of *Escherichia coli*. Mol Microbiol 36:806–816.
2. Gilles-Gonzalez M-A, Gonzalez G (2004) Signal transduction by heme-containing PAS-domain proteins. J Appl Physiol 96:774–783.
3. Gu Y-Z, Hogenesch JB, Bradfield CA (2000) The PAS superfamily: Sensors of environmental and developmental signals. Annu Rev Pharmacol Toxicol 40:519–561.
4. Taylor BL, Zhulin IB (1999) PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 63:479–506.
5. Dunham CM, Dioum EM, Tuckerman JR, Gonzalez G, Scott WG, Gilles-Gonzalez MA (2003) A distal arginine in oxygen-sensing heme-PAS domains is essential to ligand binding, signal transduction, and structure. Biochemistry 42:7701–7708.
6. Henry JT, Crosson S (2011) Ligand binding PAS domains in a genomic, cellular, and structural context. Annu Rev Microbiol 65:261–286.
7. Zoltowski BD, Schwerdtfeger C, Widom J, Loros JJ, Bilwes AM, Dunlap JC, Crane BR (2007) Conformational switching in the fungal light sensor vivid. Science 316:1054–1057.
8. Suetsugu N, Wada M (2013) Evolution of three LOV blue light receptor families in green plants and photosynthetic stramenopiles: Phototropin, ZTL/FKF1/LKP2 and aureochrome. Plant Cell Physiol 54:8–23.
9. Pellequer J-L, Wager-Smith KA, Kay SA, Getzoff ED (1998) Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. Comput Biomol Sci 95:5884–5890.
10. Imamoto Y, Kataoka M (2007) Structure and photoreaction of photoactive yellow protein, a structural prototype of the PAS domain superfamily. Photochem Photobiol 83:40–49.
11. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 15:586–592.
12. Pandini A, Bonati L (2005) Conservation and specialization in PAS domain dynamics. Protein Eng Des Sel 18:127–137.
13. Zhao JM, Lee H, Nome RA, Majid S, Scherer NF, Hoff WD (2006) Single-molecule detection of structural changes during Per-Arnt-Sim (PAS) domain activation. Proc Natl Acad Sci USA 103:11561–11566.
14. van Aalten DM, Hoff WD, Findlay JB, Crielaard W, Hellingwerf KJ (1998) Concerted motions in the photoactive yellow protein. Protein Eng 11:873–879.
15. Zayner JP, Antoniou C, Sosnick TR (2012) The amino-terminal helix modulates light-activated conformational changes in AsLOV2. J Mol Biol 419:61–74.
16. Na H, Jernigan RL, Song G (2015) Bridging between NMA and elastic network models: Preserving all-atom accuracy in coarse-grained models. PLoS Comput Biol 11:e1004542.
17. Moritsugu K, Kurkal-Siebert V, Smith JC (2009) REACH coarse-grained normal mode analysis of protein dimer interaction dynamics. Biophys J 97:1158–1167.
18. Bastolla U (2014) Computing protein dynamics from protein structure with elastic network models. Wiley Interdiscip Rev Mol Sci 4:488–503.
19. Zheng W, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci USA 103:7664–7669.
20. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB (2015) Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. Mol Biol Evol 32:132–143.
21. Kim H, Zou T, Modi C, Dörner K, Grunkemeyer TJ, Chen L, Fromme R, Matz MV, Ozkan SB, Wachter RM (2015) A hinge migration mechanism unlocks the

evolution of green-to-red photoconversion in GFP-like proteins. Structure 23:34–43.
22. Echave J, Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. Nat Rev Genet 17:109–121.
23. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Elastic network model for protein structural dynamics. Biophys J 80:505–515.
24. Wang Y, Rader AJ, Bahar I, Jernigan RL (2004) Global ribosome motions revealed with elastic network model. J Struct Biol 147:302–314.
25. Levasseur A, Pontarotti P, Poch O, Thompson JD (2008) Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evol Bioinformat 2008:121–137.
26. Nevin Gerek Z, Kumar S, Banu Ozkan S (2013) Structural dynamics flexibility informs function and evolution at a proteome scale. Evol Appl 6:423–433.
27. Liu Y, Bahar I (2012) Sequence evolution correlates with structural dynamics. Mol Biol Evol 29:2253–2263.
28. Maguid S, Fernandez-Alberti S, Echave J (2008) Evolutionary conservation of protein vibrational dynamics. Gene 422:7–13.
29. Narikawa R, Okamoto S, Ikeuchi M, Ohmori M (2004) Molecular evolution of PAS domain-containing proteins of filamentous cyanobacteria through domain shuffling and domain duplication. DNA Res 11:69–81.
30. Circolone F, Granzin J, Jentzsch K, Drepper T, Jaeger KE, Willbold D, Krauss U, Batra-Safferling R (2012) Structural basis for the slow dark recovery of a full-length LOV protein from Pseudomonas putida. J Mol Biol 417:362–374.
31. Changenet-Barret P, Plaza P, Martin MM, Chosrowjan H, Taniguchi S, Mataga N, Imamoto Y, Kataoka M (2007) Role of arginine 52 on the primary photoinduced events in the PYP photocycle. Chem Phys Lett 434:320–325.
32. Yamada S, Sugimoto H, Kobayashi M, Ohno A, Nakamura H, Shiro Y (2009) Structure of PAS-linked histidine kinase and the response regulator complex. Structure 17:1333–1344.
33. Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA (2012) Structure fluctuations and conformational changes in protein binding. J Bioinform Comput Biol 10:1241002.
34. Kalaivani R, Srinivasan N (2015) A Gaussian network model study suggests that structural fluctuations are higher for inactive states than active states of protein kinases. Mol Biosyst 11:1079–1095.
35. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874.
36. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.
37. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306.
38. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128.
39. Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875–1882.
40. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302.
41. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56:143–156.
42. Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77:1905–1908.
43. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: The biomolecular simulation program. J Comput Chem 30:1545–1614.
44. Fuglebakk E, Tiwari SP, Reuter N (2015) Comparing the intrinsic dynamics of multiple protein structures using elastic network models. Biochim Biophys Acta 1850:911–922.
45. Röllen K, Granzin J, Panwalkar V, Arinkin V, Rani R, Hartmann R, Krauss U, Jaeger K-E, Willbold D, Batra-Safferling R (2016) Signaling states of a short blue-light photoreceptor protein PpSB1-LOV revealed from crystal structures and solution NMR spectroscopy. J Mol Biol 428:3721–3736.
46. Anderson S, Crosson S, Moffat K (2004) Short hydrogen bonds in photoactive yellow protein. Acta Cryst 60: 1008–1016.
47. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79: 926.
48. Frenkel D, Smit B (2002) Understanding molecular simulation: From algorithms to applications. Academic Press.
49. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593.