Identification of structurally conserved residues of proteins at low sequence identity: A structural/sequence comparison

Goyal Amit, Uwitonze Hosanna, Kim Sejung, Lee Sun Gu, Hwang Kyu Suk*
Department of Chemical Engineering, Pusan National University, Korea
(kshwang@pusan.ac.kr)

Abstract

Identification of the conserved residues among the proteins sharing the similar SSS at low sequence identity has been a big challenge for biologist. Many conventional approaches has been used traditionally but, proved unable to identify this quest with good efficiency. In the present study, we utilize the combination of the protein sequence, structure and bonding information to identify the conserved residues which satisfy the conservedness criteria for the important evolutionary factors. Overlapped conserved residues (OCR) based approach was described and applied on GFP-like protein to identify the conserved residues. Fold scan analysis using the sequence pattern identified using OCR-based approach is proved to be highly efficient to identify the structural homologues from the structure and sequence database.

Introduction

For more than a decade, protein sequence-structure relationship has been the base of identification of the structurally and / or functionally conserved residues and protein structure prediction. The relationship defines two proteins to fold into similar structure if they share a sequence identity of greater than 30%. Several exceptions are available where protein does not follow the sequence-structure relationship, e.g. GFP-like protein, Sandwich like protein, SH3 like protein, etc. Identification of the conserved residues among the structurally similar protein sequences, with the sequence identity in mid-night zone, i.e. 10-20%, has been of keen interest for molecular biologist. The aim of this research is to characterize the crucial residues responsible for protein structure determination, especially in the case of protein family of high diversity.

Last two decades are devoted to bioinformatics research, which provides us various statistical comparison and genetic algorithm based approaches to identify the conserved residues. Traditional methods mainly utilize the protein sequence, structure and bonding information separately, to identify the conserved residues. The efficiency of conventional alignment methods has also been low and lacks the criteria to eliminate the residues which do not involved in structure conservation directly. In the present study, we attempt to identify the conserved residues that satisfy the conservedness criteria of all the three methods, i.e. every conserved residue must be identified as the conserved position by each method. Conserved residues identified by this approach are important to preserve the protein structure at each level of evolution, i.e. sequence evolution, structural evolution and crucial conserved bond. It significantly eliminates the supporting residues which fail to satisfy any of the evolutionary criteria. OCR-based approach identifies the conserved residues crucial foe the structural determination and the number of identified residues are relevant to the sequence identity conserved among the diverse members of the protein family. High efficiency of the fold scan analysis, using the sequence pattern obtained from the OCR-based approach, was obtained to detect the similar structural fold. OCR-based approach was successfully tested on GFP like protein were validated using the SPs and Cyanins like protein.

Result

Determination of the efficiency of three methods to identify conserved residues among non-homologous protein sequences

To determine the efficiency of the identification of conserved residues using each approach, multiple sequence alignment of the GFP-like protein sequences was obtained using MSA, SBA and SSS-based method. The alignment of the 10 distantly related sequences of GFP-like proteins was done independently using the multiple sequence alignment (MSA) tool ClustalW, structure based alignment (SBA) tool Dali server and supersecondary structure (SSS) based sequence alignment algorithm. In this research, hydrophobicity and hydrophilicity of residues was chosen as the criterion of residue similarity over conventional scoring matrices for three methods. Multiple sequence alignment using SSS-based method revealed the 46 conserved positions, i.e. 19% of the total amino acid residues form the SSS-determining pattern. Similarly, in the case of alignment using ClustalW and Dali server, we obtained a total of 55 and 44 conserved residues respectively.

Alignment Method / Protein Sequence pattern	SSS	MSA	SBA
Protein Sequence Pattern scan against PDB (A total of 273 GFP like structure)			
Conserved Residues Pattern	268	9	12
Protein Sequence Pattern scan against SwissProtKB database (A total of 534 GFP like sequence)			
Conserved Residues Pattern	463	135	148
Conserved Residues Pattern with a single Mismatch	521	177	215

Table1. Specificity and sensitivity of protein sequence pattern obtained via three methods. Protein sequence/structure database scan using SSS-determining pattern reveals the high specificity and sensitivity of the SSS-based method.

A conserved pattern of protein sequence is considered to be highly specific and sensitive if it can detect all or almost all the protein with the given structural folds and no or very few false-positive results. In the present research, ExPASy ScanProsite tool was used for the detection of the similar folds against the protein sequence/structure database such as PDB and SwissProtKB. SSS-determining pattern identifies almost 98% of sequences with the similar SSS in PDB. Protein sequence pattern using the multiple sequence alignment tool ClustalW scanned just 9 similar fold. Similarly, sequence pattern obtained using structural based alignment tool Dali scan 12 GFP like fold. Thus, the SSS-determining pattern is highly specific and sensitive for the SSS. Similar kinds of results were obtained via the fold scan against sequence database.

Identification of conserved residues that satisfy the conservedness criteria of all the three methods

Identification of the conserved residues crucial for the determination of the protein structure requires a specific selection of the residues from the identified conserved residues via the three methods. In the present study, elimination of the supporting residues is performed via deriving the conserved residues that do not satisfy the conservedness criteria of the three methods. A comparison of the distribution of the conserved residues among the three sequence pattern describes the pattern of conservedness. We divided

the conserved residues by two groups; “Overlapped Conserved Residues” (OCR) and “Critical Conserved Residues” (CCR). “Overlapped Conserved Residues” are the common residues among the identified residues via the three methods, i.e. OCR satisfy the conservedness criteria of all the three methods and are supposed to be important for specificity of the sequence pattern. “Critical Conserved Residues” are the distinct residues among three methods and affects the sensitivity of the protein sequence pattern.

Effect of OCR and CCR on the sensitivity and specificity of fold detection

To understand the effect of OCR and CCR on the sensitivity and specificity of the fold detection, we perform the fold scan for four different cases as per shown is table 2. The four cases include the sequence pattern made from either OCR or CCR from any of the alignment method. The OCR are common for the three alignment method. Critical conserved residues for SSS-method make it highly specific and sensitive for the fold detection.

Sr. No.	Conserved Sequence Pattern	PDB		SwissProtKB		SwissProtKB^{SM}	
		Structural hits	% GFP fold	Structural hits	% GFP fold	Structural hits	% GFP fold
1.	OCR	271	99%	508	95%	522	97%
2.	CCR from SSS	327	-	623	-	636	-
3.	CCR from MSA	28	10%	275	51%	312	58%
4.	CCR from SBA	47	17%	362	67%	426	79%

Table2. Protein fold scan using the protein sequence pattern made of either OCR or CCR from the three different sequence alignment methods.

Here SwissProtKB^{SM} represents the database scan allowing a single mismatch in the conserved sequence pattern.

As shown is first case, sequence pattern consist of only OCR is found to be more sensitive as compared to the SSS-determining pattern as it scan 3 additional GFP-like fold against the PDB. Similarly, fold scan against the SwissProtKB detects 1 additional fold when using the OCR as the sequence pattern. Present results describe that OCR-based sequence pattern that is comprised of the relatively low number of conserved residues, i.e. ~10% of the sequence length, is highly efficient to detect the structure fold. Here, OCR-based approach can be used for the identification of the crucial residues for the structural determination which has been a daunting task via using the conventional approaches.

Effect of the conserved residues in strands or loops on the efficiency of fold detection

Conservation of the sequence is rarely homogenous along the entire sequence length, but it is considered to be localized to specific regions. These specific secondary structure elements (SSEs) are conserved during the evolution. To understand the effect of the conserved residues in strands or loops region, we perform the fold scan for the four different cases. As shown in table 3,

fold scan analysis using the sequence patterns comprise of the conserved residues either from strand or loop region, each for SSS-based and OCR-based sequence pattern, were performed.

Sr. No.	Conserved Sequence Pattern	PDB		SwissProtKB		SwissProtKB^{SM}	
		Structural hits	% GFP fold	Structural hits	% GFP fold	Structural hits	% GFP fold
1.	SSS^{S}	270	99	494	92	518	97
2.	SSS^{L}	297	-	568	-	586	-
3.	OCR^{S}	271	99	508	95	528	98
4.	OCR^{L}	5782	-	37639	-	43562	-

Table3. Protein fold scan using the protein sequence pattern comprised of the conserved residues either from strand or loop region, each for SSS-based and OCR-based sequence pattern

Fold scan results, in second and fourth case, indicate that in the absence of the conserved residues for strand, specificity of the sequence pattern reduced and as a result we found false positives in the fold scan result. In the case of sequence pattern made only from conserved residues for strand, high specificity and sensitivity of the fold scan is obtained. The results obtained shows that the conserved residues in strands increases the efficiency of the fold detection for the group of Green Fluorescent-like protein (GFP) and no significant effect of the conserved residues from loop is observed for fold scan. Fold scan result in third case shows the high efficiency where the sequence pattern consist of only 17 conserved residues, which represent only 7% of the protein sequence length. The results provides an important conclusion that a relatively small number of the amino acid residues represent the protein family with diverse sequence and these residues are conserved during the protein evolution.

Conclusion

Present study emphasizes the use of protein sequence, structure and bonding information during the evolution process to perform the alignment of the amino acid residues for identification of the conserved residues. OCR-based approach detects around 10% of the sequence residues as the conserved location that provide crucial role in structural determination.

References

1. Oliver C Redfern, Benoit Dessailly, Christine A Orengo (2008) Exploring the structure and function paradigm. Curr Opin Struct Biol 18(3): 394–402.
2. Alexander E. Kister, Israel Gelfand (2009) “Finding of residues crucial for supersecondary structure formation”, Proc Natl Acad Sci USA, 106(45): 18996–19000.
3. Michael Hopf, Walter Gohring, Albert Ries, Rupert Timpl, Erhard Hohenester (2001) “Crystal structure and mutational analysis of a perlecan-binding fragment of nidogen-1”, Nature Struct Biol, 8, 634 - 640.
4. Chiang Y-S, Gelfand TI, Kister AE, Gelfand IM (2007), “New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage” *Proteins* 68:915–921.