2010 International Conference on Complex, Intelligent and Software Intensive Systems

A comprehensive system for identifying internal repeat substructures of proteins

Hua-Ying Kao, Tsang-Huang Shih, Tun-Wen Pai
Dept. of Computer Science and Engineering
National Taiwan Ocean University,
Keelung, Taiwan, R.O.C.
e-mail: twp@mail.ntou.edu.tw

Ming-Da Lu, Hui-Huang Hsu
Dept. of Computer Science and
Information Engineering,
Tamkang University, Taipei, Taiwan, R.O.C.
e-mail: hhsu@cs.tku.edu.tw

Abstract*—Repetitive substructures within a protein play an important role in understanding protein folding and stability, biological function, and genome evolution. About 25% of all proteins contain repeat structures for eukaryote species and most of them do not have the resolved structural information yet. Therefore, this study aimed to design a comprehensive system for identifying internal repeats either from a protein sequence or structural information. In this study, we have curated a set of internal repeat units as a benchmark dataset for performing both sequence and structural alignment with respect to the query sequence or structure. Except for the traditional BLAST algorithms on amino acid sequence or the optimal structural superposition approaches on structures, a novel method employing the predicted secondary structure element information for internal repeat identification was proposed. Sequences were firstly transformed into Length Encoded Secondary Structure (LESS) profiles and followed by autocorrelation analyses. From the primary experimental results, the developed Internal Repeat Identification System (IRIS) can successfully identify internal repeats from those known protein structures, and the web system is freely available at http://iris.cs.ntou.edu.tw/.

Keywords- internal repeat unit; secondary structure element; sequence alignment; structure alignment; Length Encoded Secondary Structure; solenoid

I. INTRODUCTION

Protein repeats were roughly classified into three different types according to the length of a repeat unit. The shortest repeats contain less than 4 residues within a repeat unit and form crystalline or fibrous structures. The second type of repeat is with unit length shorter than 45 residues and called as solenoid proteins which contain secondary structure elements within the repeat unit and the repeat units are coiled alone with a common axis or a specific direction sequentially in spatial domain. The third type represents a basic repeat unit possessing its length longer than 45 residues and forms a protein domain itself within a repeat structure[1]. Due to important biological features and particular construction frameworks of protein structures with internal repeats, increasing interest has recently been devoted to the study on detection of protein repeats. A repetitive substructure within a protein plays an important role in understanding protein folding and stability, biological function, and genome evolution [2] [3] [4]. These studies indicated that novel genes were evolved through duplications and transitions from existing genes within proteins possessing regular secondary structures and functional units[5], and the stability and repetition of structural unit directly reflected the structural and biophysical properties of proteins[6]. For example, different alleles of the fungus *Podospora anserine* possess different numbers of WD40 (WD or beta-transducin repeat) repeats[7]. The analysis of conserved cores of internal repeats often occur symmetric units on structures, such as the protein phophatase 2A PR65 (HEAT), a superhelix with repeats [8].

Over the last two decades, a bunch of tools were developed for repeat sequence and structure detection. Several implementations were designed at the DNA level, such as Reputer[9], CGSSR[10], Repseek[11], while Swelfe[12], REPRO[13], and REPETITA[14] performed at the structure level. Traditional approaches for identifying internal repeats were based on sequence alignment strategies, especially when coping with protein sequences without resolved protein structures. It is not easy to predict the internal repeats within a protein since the highly varied residue contents usually occurred for the identical substructures within a protein. However, it becomes relatively simple when a protein structure is known for quantitative analysis of repetitive composition.

The sequence alignment approaches were satisfied only confronting with sequences with high similarity and regularity. Such alignment tools combined with comprehensive genomic databases of various species can be efficiently applied to identify homologous sequences which possessing with known repetitive information. These well-known tools include PAM [15], BLAST[16], PSI-BLAST[17], and ClustalW[18]. However, if protein segments possess low sequence similarity, then sequence alignment based methods become invalid for internal repeat detection. Unfortunately, it has been verified that sequence contents of repeat structure units of most proteins with internal repeats are highly diverged among all various species. Hence, the structural information of secondary structure information provides an alternative way to analyze and predict the locations of repeat segments within a protein, since the secondary structures always possess highly

978-0-7695-3967-6/10 $26.00 © 2010 IEEE
DOI 10.1109/CISIS.2010.92

conserved and stable structural characteristics, even the residue contents are quite degenerate during evolution processes.

This study aims to develop a comprehensive system which can hierarchically detect internal repeats within a protein either from its protein sequence contents and/or corresponding structural information. At the protein sequence level, the proposed system adopts multiple sequence alignment approaches and/or employs secondary structure prediction methodologies for identifying internal repeats. In the meanwhile, an internal repeat unit database was curately constructed for homologous sequence comparison. The BLAST algorithms were adopted here as the initial trial to efficiently search possible homologous sequences from the collected database. When dealing with sequences with low similarity, the system turned to focus on the predicted secondary structure information and apply autocorrelation analysis to identify the locations of internal repeat substructures. At the protein structure level, the query protein structure was aligned individually with respect to all the repeat units within the collected database under the criteria of desired RMSD values and minimum number of aligned residues. Based on the designed system flow, we have presented the Internal Repeat Identification System (IRIS) to identify the internal repeats either from a protein sequence or structure in a comprehensive mechanism.

II. Material and Method

A. Internal Repeat Unit Databse

The fundamental internal repeat unit (IRU) database was mainly extracted from the repeated structures collected in PROPEAT[19] which provides complete location information of each internal repeat unit within a protein. There are total 399 internal repeat structures from PROPEAT database found from RCSB [20] and all of the repeat units were manually verified and extracted for constructing the benchmark dataset in our study. Accordingly, 2230 repeat units with sequence contents and corresponding substructures were collected in our IRU database, and they were considered as the fundamental repeat units for primary analysis in our proposed system.

B. System Configuration

The proposed system including two different approaches: detection of internal repeats either from protein sequences or structures, and which system is depicted in Fig. 1. For detecting repeat units from sequences, there are three consecutive modules including autocorrelation analysis of primary sequences, BLAST with IRU database, and prediction of secondary structure elements for encoded profile analysis. The system flow chart for sequence detection is shown in Fig. 1. Users enter a query protein sequence into the system, an autocorrelation analysis on the input primary sequence was performed by shifting the sequence by variant lengths and aligning to its original sequence to discover the identical or homologous repeat segments within a protein. If there is no repeat sequence found in previous steps, the system applied the query sequence by Basic Local Alignment Search Tool (BLAST) algorithms with respect to our collected benchmark datasets. This study adopted BLASTP version to search local sequence segments which are highly similar to repeat unit segments in IRU database. If more than two repeat segments can be matched from the IRU database and these segments belong to a protein structure in IRU database, the query protein structure is considered as a protein with internal repeats. However, if the results of segment matching cannot be satisfied, the next module of SSE prediction and LESS profile transformation were employed for internal repeat analysis. In this module, an SSE prediction method based on SSPro4[21] was adopted here, and the predicted SSE information of the query sequence was transformed into a LESS profile which contained the length and type information of the predicted SSEs. With the encoded symbols in a LESS profile, an autocorrelation mechanism was applied again to observe the internal repeat relationship at the secondary structure level.

For handling the input at the structural level, either a PDB id or an uploaded PDB file is required. In this module, all fundamental IRUs will be structurally aligned to the query structure, and see if there is any IRU matched with the substructure of the query structure more than twice simultaneously and non-overlappedly. When such a condition is detected, the query structure possessing the same kind of IRU is identified. To guarantee the performance of structural alignment, three different types of SSEs including helices(H), strands(E), and coils(C) within the substructure should be identical with the target IRU. Furthermore, high percentages of alignment residue number and low values of root-mean-square-deviation between the IRU and the substructure of query protein should be strictly constrained.

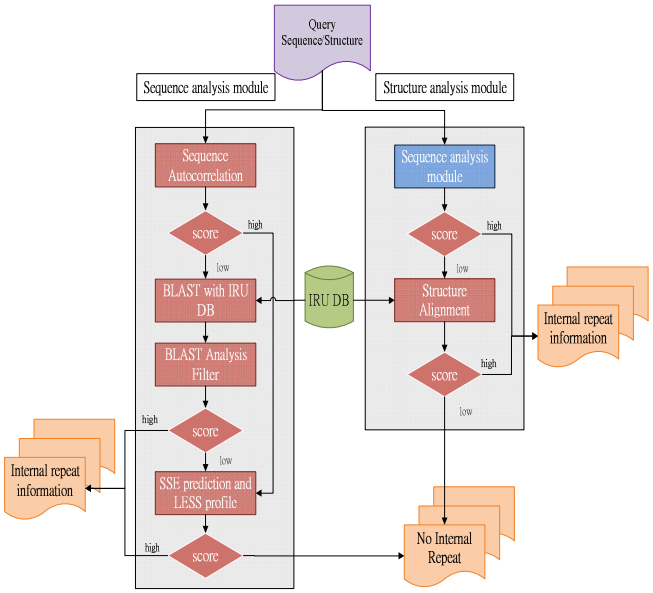


Fig. 1 System flow chart for internal repeat detection from a sequence or a structure.

C. *Sequence Similarity by Autocorrelation Analysis*

The internal repeat detection based on intra-sequence evaluation is approached by an autocorrelation analysis. The core algorithms detected repeat locations by shifting the query sequence to right with N residues and matching with its original primary sequence. If a high score of sequence similarity of local segments can be obtained, then a repeat condition exists. However, it may not be easy to discover the repeat segments at the residue level, since the contents of a protein sequence are generally degenerated to a certain extent during the evolution processes. To avoid the difficulties, the proposed system applied the same mechanism at the secondary structure level. The query sequence was represented by its predicted secondary structure format and the autocorrelation analysis was then applied on the transformed profiles. The details for a novel encoded SSE profile are described in the next section.

D. *BLAST with IRU Database*

Except for applying autocorrelation analysis on primary sequences or secondary structure profiles, the developed system provides another approach for discovering repeat segments in a protein. We have collected most of known internal repeat segments in advance and stored them in our IRU database for BLAST searching. In these matching processes, the query sequence will be locally aligned with all 2230 repeat segments in our curated IRU database. These various segments were extracted from 399 protein structures with verified internal repeat characteristics. If the homologus segments can be found non-overlappedly within the query sequence from IRU database, the system will verify the selected IRU segments and see if they possess similar types of substructures. To search the homologous segments from IRU database, BLASTP version was applied with default parameter settings of BLOSUM 62 matrix, initial gap penalty of 11, gap extension penalty of 1, minimal number of aligned residues of 5, and less than E-value of 0.1. Under these criteria, only top ranked 30 candidates were selected for further IRU analysis, these matched fundamental repeat units from IRU database were finally recognized as a repeat unit of the query sequence when there were at least two IRU segments were originated from an identical protein structure.

E. *SSE Prediction and Eencoded LESS profile*

When the input primary sequence itself cannot provide valuable information for internal repeat detection, the proposed system turns to focus on its predicted structural information. The structural information of the query sequence was obtained by SSE prediction methods to allocate the possible locations of alpha-helix and beta-strand segments within a protein. In this study, we adopted one of the best SSE predictors, SSPOR4[21], which was verified with an accuracy of 78.7% and was tested by the independent assessor EVA. With the predicted SSE information, the system executed a length encoded secondary structure (LESS) transformation to encode the predicted SSE sequence into a profile which is composed of three secondary structure codes including helix(H), strand(E), coil (C) types. According to the length distribution, each type possesses six different groups and categories as ***H0~H5***, ***E0~E5***, and ***C0~C5*** respectively. Based on the statistical analysis of equally distributed intervals, a look up table for LESS profile transformation is shown in table 1. The LESS transformation generates a corresponding LESS profile from predicted secondary structure information, and the profile contains the SSE information but in a shorter representation. To reduce the noise generated from prediction processes, a smoothing filter was performed to tolerate the variation of LESS profile. The last step is to apply the autocorrelation procedures for allocating the repeat segments as described in the previous section. If the query sequence indeed possesses repeat segments within it and the predicted SSE elements holds satisfied prediction rates, the transformed LESS profile certainly reflect the conditions in its spatial representation.

F. *Structural alignment*

When the query possessing structural information, the proposed system will extract the sequence contents from the query PDB file first and run through all of the sequence analysis modules. If there is no confident results confirmed by the system, the structural alignment between the IRU elements and the query structure will be performed. Any structure alignment tool can be adopted here for searching possible repeat units within the query protein structure. Here we employed our own developed multiple structural alignment algorithms which is based on the geometrical characteristics of secondary structure elements, dynamic programming on residue level approaches, and iterative refinement procedures. The common measurement of structural similarity is evaluated by two features simultaneously, which are the number of equivalent residues and the root-mean square deviation (RMSD) among equivalent residues. Each IRU element is superposed on all possible corresponding substructures. When the number of aligned residue is greater than 70% of an IRU element and the RMSD value is less than 2.5Å, the matched IRU element will be considered as one of its internal repeat candidates.

III. RESULT AND DISCUSSION

A. *The IRIS system and IRU database*

We have developed an internal repeat identification system for a protein sequence or a structure. The system is named as Internal Repeat Identification System (IRIS), and it is freely accessible at http://iris.cs.ntou.edu.tw/. In this web system, users can learn a protein structure with internal repeats, verify the repeat unit substructure within a structure, and compare their similarity among repeat units. The dataset

were classified into 6 categories according SCOP classification rules. Each repeat unit within a repetitive structure was extracted individually and aligned together for comparison. It is clearly found the type and quality of repeat units in a protein according the RMSD value. Up to now, there are initially 399 fundamental units were collected here for homologous matching, and the number of primitive structure units will be increased when a novel and distinct substructure is found. Fig. 2 shows a snap shot of the IRU database. The upper structure is an example of repeat protein structure (1a04:a) and the lower part depicts the aligned repeat units of 1a04:a with RMSD value of 0.565. The sequence contents were also aligned according to it geometrically aligned coordinates. Those vertically aligned residues represented spatially matched condition, while the gaps showed those not aligned resides in spatial domain. For all of collected protein repeats in IRU database, an interesting statistical information is shown in Fig. 3 which shows the length distribution of each IRU segments and the number of repeats within a repeat protein. For example, most of repeat proteins possess only two repeat units within it, and higher number of repeats coexists with shorter length of a unit. More specifically, the length of repeat unit shorter than 50 residues occupies more than 86% of our collected IRU database. In other words, most of the repeat proteins collected in IRU database are compatible with definition of solenoid repeats, and the rest of large proteins belong to domain repeats.

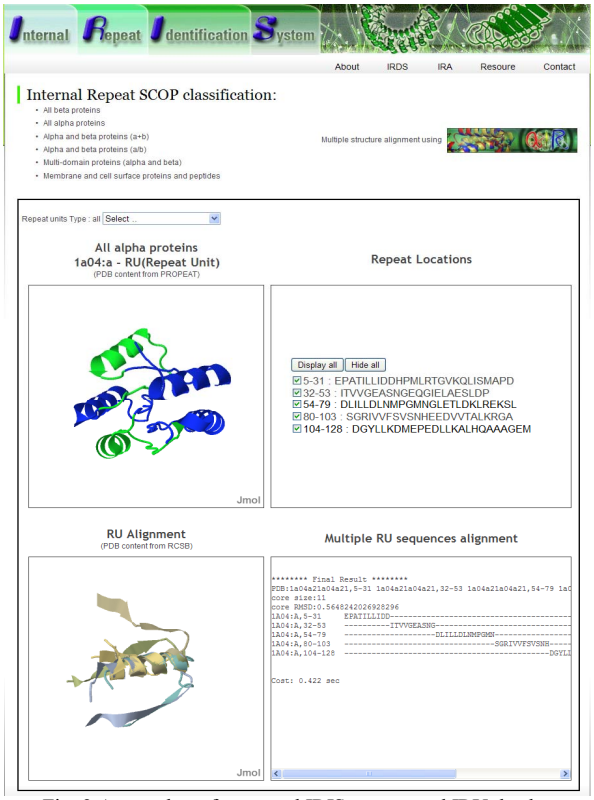


Fig. 2 A snapshot of proposed IRIS system and IRU database.

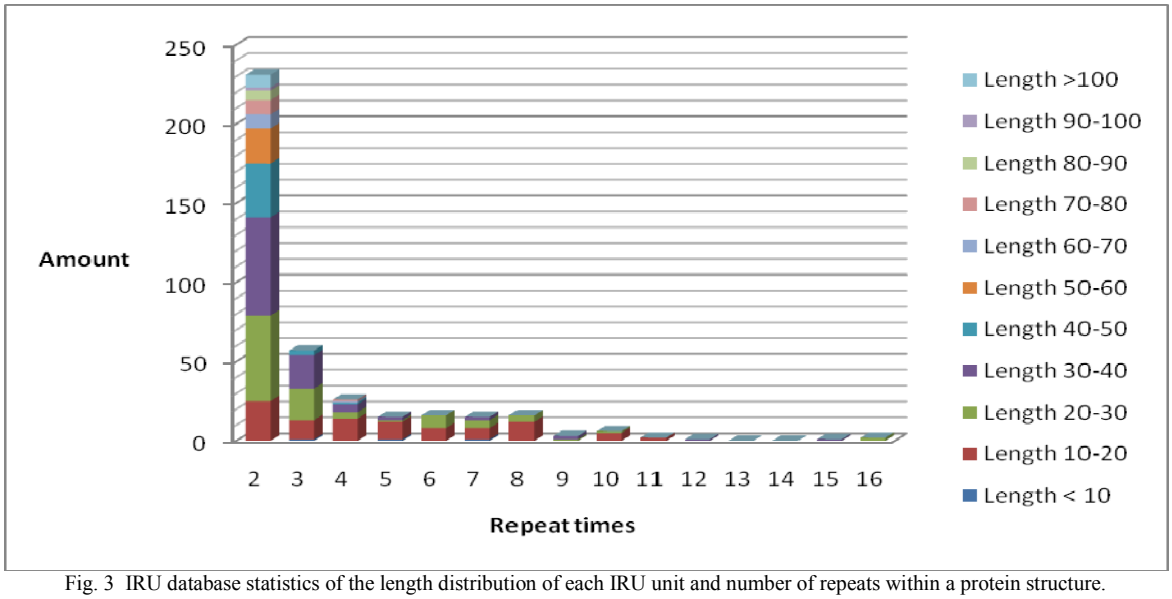
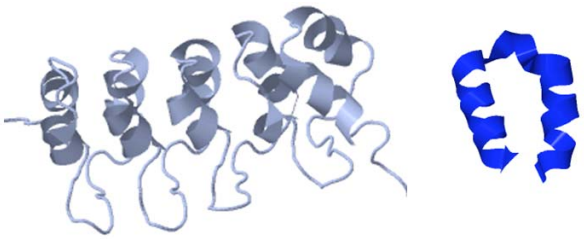


Fig. 3 IRU database statistics of the length distribution of each IRU unit and number of repeats within a protein structure.

B. *Verification of IRIS system*

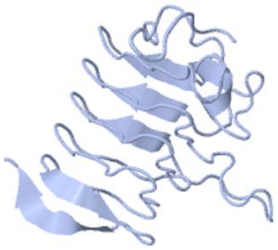
To verify the performance of IRIS system, we have selected another repeat protein dataset from REPETITA [14]. There were 50 repeat proteins selected as a testing dataset. Through the evaluation by automatic identification processes in IRIS, the results have shown that more than 84% of testing repeat proteins can be successfully identified by their corresponding repeat units. For example, the query protein 1AP7 (PDBID) was correctly recognized by the

repeat unit from IRU dataset (PDBID: 1B89), and the query protein structure and matched fundamental unit were shown in Fig. 4(a). In addition, those misrecognized repeat proteins can be considered as extra candidates for additional collection of repeat units in the IRU database. For another example, the repeat protein of 1XAT (PDBID) shown in Fig. 4(b) could not be identified by IRIS, since there was no repeat unit from IRU database perfectly matched with the substructure of 1XAT (PDBID). Therefore, the fundamental repeat unit of 1XAT (PDBID) can be manually added into the IRU dataset for its integrity. In conclusion, the developed IRIS system and collected IRU database provide an efficient and effective approach for identifying internal repeat structures either from verified repeat unit segments of known repetitive proteins or from information of predicted secondary structure elements.



PDBID:1AP7 IRU PDBID:1B89

Fig. 4 (a) Correctly identified internal repeat (the query protein: 1AP7; protein from IRU: 1B89)



PDBID: 1XAT

Fig. 4 (b) Misrecognized repeat protein and the repeat unit will be collected in our IRU database. (the query protein: 1XAT)

IV. ACKNOWLEDGMENT

This work is supported by the Center for Marine Bioscience and Biotechnology (CMBB) in National Taiwan Ocean University, Keelung, Taiwan, and the National Science Council in Taiwan, R. O. C. (NSC98-2627-B-019-003 and NSC98-2221-E-019-031-MY2 to Tun-Wen Pai).

REFERENCES

[1] Kobe, B. and A. V. Kajava (2000). "When protein folding is simplified to protein coiling: the continuum of solenoid protein structures." Trends Biochem Sci **25**(10): 509-515.

[2] Heger, A. and L. Holm. "Rapid automatic detection and alignment of repeats in protein sequences." Proteins 41(2): 224-37, 2000.

[3] Forrer, P., H. K. Binz, et al. (2004). "Consensus design of repeat proteins." Chembiochem **5**(2): 183-189.

[4] Andrade, M. A., C. P. Ponting, et al. (2000). "Homology-based method for identification of protein repeats using statistical significance estimates." J Mol Biol **298**(3): 521-537.

[5] Marcotte, E. M., M. Pellegrini, et al. "A census of protein repeats." Journal of Molecular Biology 293(1): 151-160, 1999.

[6] Andrade, M. A., C. Perez-Iratxeta, et al. "Protein repeats: structures, functions, and evolution." J Struct Biol 134(2-3): 117-31, 2001.

[7] Saupe, S., B. Turcq, et al. (1995). "A gene responsible for vegetative incompatibility in the fungus Podospora anserina encodes a protein with a GTP-binding motif and G beta homologous domain." Gene **162**(1): 135-139.

[8] Groves, M. R., N. Hanlon, et al. (1999). "The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs." Cell **96**(1): 99-110.

[9] Kurtz, S. and C. Schleiermacher, *REPuter: fast computation of maximal repeats in complete genomes.* Bioinformatics, 1999. **15**(5): p. 426-7.

[10] Pai, T.-W., Chen, C.-M., Hsiao, M.-C, Chen R., Tzou, W.-S., and Hu C.-H., "An online conserved ssr discovery through cross-species comparison". Computational Biology and Chemistry: Advances and Applications 2: pp. 23-35, 2009.

[11] Achaz, G., F. Boyer, et al. "Repseek, a tool to retrieve approximate repeats from large DNA sequences." Bioinformatics 23(1): 119-21, 2007.

[12] Abraham, A. L., E. P. C. Rocha, et al. "Swelfe: a detector of internal repeats in sequences and structures." Bioinformatics 24(13): 1536-1537, 2008.

[13] George, R. A. and J. Heringa. "The REPRO server: finding protein internal sequence repeats through the Web." Trends Biochem Sci 25(10): 515-7, 2000.

[14] Marsella, L., F. Sirocco, et al. (2009). "REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform." Bioinformatics **25**(12): i289-295.

[15] Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-10919.

[16] Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

[17] Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

[18] Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

[19] PROPEAT web server. (200910) URL is http://gln.ibms.sinica.edu.tw/product/repeat/index.html

[20] RCSB protein data bank (200910) URL is http://gln.ibms.sinica.edu.tw/product/repeat/index.html

[21] Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W72-6.