Biochemical and Biophysical Research Communications 494 (2017) 602–607

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Sequence conservation of protein binding segments in intrinsically disordered regions

Haruki Ota, Satoshi Fukuchi*

Faculty of Engineering, Maebashi Institute of Technology, 460-1 Kamisadori, Maebashi 371-0816, Japan

ARTICLE INFO

Article history:
Received 12 October 2017
Accepted 18 October 2017
Available online 21 October 2017

Keywords:
Intrinsically disordered protein
Protein–protein interaction
Peptide motif
Sequence conservation

ABSTRACT

Intrinsically disordered proteins are proteins with intrinsically disordered regions (IDRs) that do not adopt a globular structure in their free state. IDRs have unique regions having protein-binding segments of which play pivotal roles in many biological processes. The binding sites in IDRs are heterogeneous in terms of their sequence conservation: some are conserved and others are not. We have been running a database of intrinsically disordered proteins, IDEAL, and collecting such binding segments, which are called protean segments (ProSs), within it. In this study, we compared the sequence conservation of ProSs, structural domains (SDs), and IDRs other than ProSs (non-ProSs) and found that i) functionally constrained residues in ProSs tend to be conserved, ii) the distribution of conservation scores of ProSs is similar to those of SDs but not non-ProSs, and iii) ProSs found in human proteins are mostly conserved only in vertebrates. These results indicate that the conservation patterns in ProSs principally follow the general rules found in SDs. However, we need to consider evolutionary distance when comparing IDR sequences because ProSs can readily emerge and disappear over the course of protein evolution. Moreover, many ProSs may remain to be identified, which may account for the heterogeneity of the sequence conservation of IDRs.

1. Introduction

Intrinsically disordered proteins (IDPs) are proteins that do not adopt unique three-dimensional (3D) structures under physiological conditions [1–4]. They are fully or partially disordered, abundant among eukaryotic proteins, and localize preferentially in the nucleus [5–7]. One of the unique features of IDPs is their ability to bind upon binding partners. The regions performing such binding can adopt local 2D structures in association with this binding, which has been referred to as the coupled folding and binding mechanism. Through this mechanism, IDRs play crucial roles in many biological processes, such as signal transduction and transcriptional regulation [1–3,8].

We have developed an IDP database, IDEAL (http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/) [9,10], and been collecting protein binding regions in IDRs, which are called ProSs (protean segments), within it. During the collection of ProSs, we have identified heterogeneous patterns of sequence conservation of ProSs, in which some ProSs are well conserved, some moderately conserved, and others show no conservation. Although short sequence motifs (peptide motifs) have been collected [11], not all ProSs are necessarily similar enough that their sequence motifs can be defined, and no motifs have been defined for both completely conserved and much less conserved regions. In fact, there are two conflicting conclusions regarding sequence conservation in IDRs: that IDRs are conserved as befits their functional significance and that they are much less conserved than folded regions. For example, the compositions of amino acid sequences of IPDs are conserved [12], specific motifs for PTM and binding show sequence conservation [13], and specific highly conserved segments of particular lengths have been reported in IDRs [14]. In contrast, there have also been reports on less conservation of IDRs compared with folded regions [15,16] and an increase of insertions and deletions due to high mutation rates in IDRs [17].

To elucidate the pattern of sequence conservation of IDRs, we compared amino acid sequences derived from the IDEAL database. Although peptide motifs are mostly found in IDRs, some of them reside outside of IDRs, and motifs for post-translational modification sites are also included in peptide motifs. On the other hand, ProS regions have evidence of being disordered in their free state,

* Corresponding author.
E-mail address: sfukuchi@maebashi-it.ac.jp (S. Fukuchi).

https://doi.org/10.1016/j.bbrc.2017.10.099

and have structures with its binding partners deposited in Protein Data Bank (PDB) [9,10]. Then, ProSs are part of IDRs, and the structures binding upon partner proteins are available. Thus, we can analyze sequence conservation patterns of ProSs, considering their structures in binding states.

2. Materials and methods

We used the XML file of the IDEAL database version 12/Jun/2015. We obtained amino acid sequences from the XML file and divided them into three categories, ProSs, SDs, and non-ProSs, where ProSs are the interaction segments in IDRs, SDs are the structural regions found in PDB, and non-ProSs are the IDRs excluding ProSs. We selected proteins from human or other mammalian model organisms such as mouse and rat, while excluding those from other resources. We also excluded ProSs that are found in loop regions in SDs. We used SDs deposited as monomers in PDB and discarded short non-ProS regions shorter than 20 residues.

To estimate the degree of conservation, we selected orthologs by the bidirectional Blast best hits method. We conducted a Blast search by using one of the amino acid sequences obtained as described above as a query against the proteome dataset derived from UniProt [18]. Then, using the top hit of the search as a query, we conducted the next Blast search. When the query of the first search was also the top hit in the second search, these proteins were referred to as an orthologous pair. The proteomes selected were from the following organisms: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Macaca mulatta* (rhesus macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Xenopus tropicalis* (Western clawed frog), *Danio rerio* (zebrafish), *Takifugu rubripes* (Japanese pufferfish), *Gallus gallus* (chicken), *Caenorhabditis elegans*, *Arabidopsis thaliana* (mouse-ear cress), *Oryza sativa* (rice), *Aspergillus nidulans*, and *Saccharomyces cerevisiae* (baker's yeast). The orthologs were aligned with the ClustalW program [19] and the conservation score of each column of the alignments was estimated by the method of Pei and Grishin [20]. In the case of the alignments including invertebrates, conservation was estimated by the method by Valdar and Thornton [21].

3. Results

First, we illustrate heterogeneous patterns of sequence conservation in ProSs. Fig. 1a presents an alignment of a ProS region found in human Adenomatous polyposis coil protein (APC), which is IDEAL entry IID00035. This short segment shown against a black background provides the binding region with β-Catenin, which is important in the Wnt signaling pathway, and has also been reported to be disordered when it is isolated [22]. This is an example in which all of the residues are completely identical within vertebrate orthologs in the ProS region. The asterisks shown at the bottom of the alignment represent the complete conservation at each column. Fig. 1b shows an example of the identification of a motif-like structure. This ProS found in Human Axin-1 (IID00007) binds to glycogen synthase kinase-3β(GSK3β) [23–26], also in the Wnt signaling pathway. Here, the asterisks are more sparsely distributed than in panel a), forming a motif-like structure. Fig. 1c presents an alignment of a ProS found in human X-ray repair cross-complementing protein 6 (XRCC6, IID00023). This region is one of the nuclear localization signals, binds to importin [27], and is disordered without importin [28]. The lack of asterisks at the bottom of the alignment in this example indicates the lack of significant conservation in the ProS region. Note that the conservation patterns vary between the alignments, despite the proteome set of the aligned sequences being identical in the three alignments.

Generally, sequence motifs can be found when there is variation

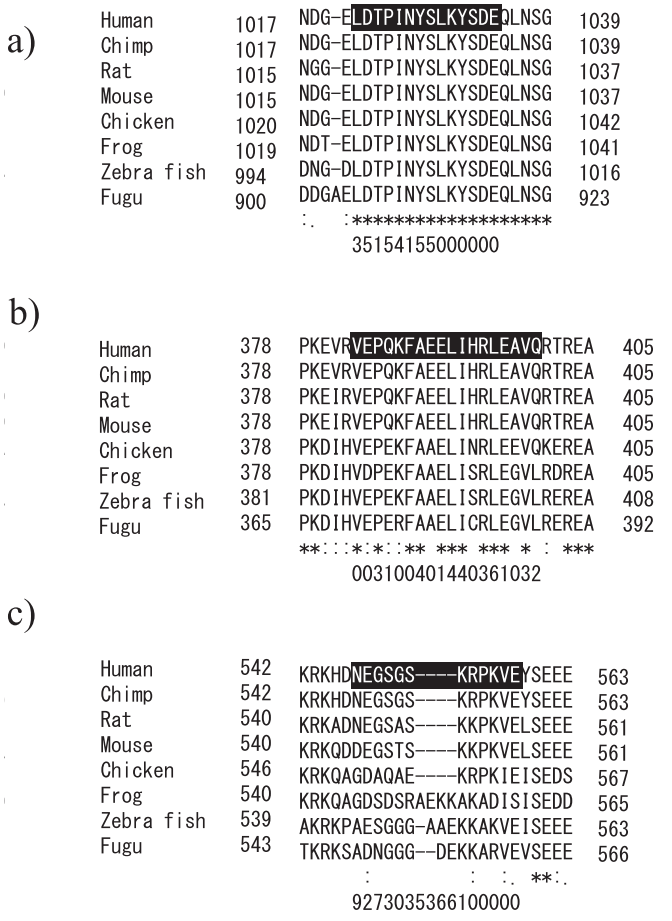

Fig. 1. Examples of the heterogeneous conservation of ProSs.
The alignments of some of the ProSs are shown. The ProSs are shown against a black background in the alignments. The conserved residues are shown by asterisks at the bottom of the alignments. The numbers under the alignments represent the numbers of contacts with binding partners. a) Adenomatous polyposis coil protein (IDEAL, IID00035), b) Axin-1 (IID00007), and c) X-ray repair cross-complementing protein 6 (IID00023).

in the functional constraint on each residue in a particular length of amino acid sequences. We first considered the functional constraints on ProSs. ProSs are protein binding segments found in IDRs, which are defined by evident disordered states and binding structures in PDB [9,10]. Given the nature of ProSs, when considering their functional constraints, it is natural to consider their residues that contact their binding partners. Because a PDB structure of the state when bound with a binding partner is available for each ProS, the number of contacts of each of the residues can be estimated. The residues located within a Cβ–Cβ distance of 8 Å were defined as those in contact. IDRs can sometimes not be aligned with homologs of closely related organisms or protein families, so we limited the orthologs to those within the vertebrates for alignment in this analysis.

Fig. 2 shows the relationship between the numbers of contacts of each residue in ProSs and the conservation scores. The line with the dots shows the mean conservation scores for each of the bins. This figure shows that residues with a large number of contacts tend to be highly conserved. Although the deviations of the gray dots are large, they are distributed in the upper left section of the chart, suggesting that less conserved residues do not have many contacts. Empirically, functionally important residues tend to be conserved in structural domains, and this result confirms that this