Biochimica et Biophysica Acta 1834 (2013) 890–897

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Long indels are disordered: A study of disorder and indels in homologous eukaryotic proteins☆

Sara Light [a,b,c,1], Rauan Sagit [a,b,1], Diana Ekman [d], Arne Elofsson [a,b,e,*]

[a] Science for Life Laboratory, Stockholm University SE-171 21 Solna, Sweden
[b] Department of Biochemistry and Biophysics, Stockholm University, SE-171 21 Solna, Sweden
[c] Bioinformatics Infrastructure for Life Sciences, Sweden
[d] Medical Biochemistry and Biophysics, Karolinska Institute, SE-17177 Solna, Sweden
[e] Swedish e-Science Research Center (SeRC), Sweden

ARTICLE INFO

Article history:
Received 1 November 2012
Received in revised form 30 December 2012
Accepted 3 January 2013
Available online 17 January 2013

Keywords:
Intrinsically disordered protein
Indel
Protein evolution protein structure
Sequence alignment

ABSTRACT

Proteins evolve through point mutations as well as by insertions and deletions (indels). During the last decade it has become apparent that protein regions that do not fold into three-dimensional structures, i.e. intrinsically disordered regions, are quite common. Here, we have studied the relationship between protein disorder and indels using HMM–HMM pairwise alignments in two sets of orthologous eukaryotic protein pairs. First, we show that disordered residues are much more frequent among indel residues than among aligned residues and, also are more prevalent among indels than in coils. Second, we observed that disordered residues are particularly common in longer indels. Disordered indels of short-to-medium size are prevalent in the non-terminal regions of proteins while the longest indels, ordered and disordered alike, occur toward the termini of the proteins where new structural units are comparatively well tolerated. Finally, while disordered regions often evolve faster than ordered regions and disorder is common in indels, there are some previously recognized protein families where the disordered region is more conserved than the ordered region. We find that these rare proteins are often involved in information processes, such as RNA processing and translation. This article is part of a Special Issue entitled: The emerging dynamic view of proteins: Protein plasticity in allostery, evolution and self-assembly.

1. Introduction

A number of different genetic mechanisms cause mutations in coding genes, ranging in size from point mutations, through insertions and deletions (indels) of a few residues, to rearrangements of protein domains and fusion of entire genes. In general, mutations occur at random but are under selective pressure. One general result of this is that residues in the core of a protein are more likely to be maintained through evolution compared to those on the surface of the protein [1]. Further, short indel events are more likely to occur in loops than in secondary structures.

Short indels occur by, for instance, DNA replication slippage during replication or repair [2]. Longer extensions can occur through the conversion of 3′ UTRs into coding regions [3] and through cassette duplications of protein domain repeats, a feature that is particularly common in higher eukaryotes [4]. Novel coding regions may also be created through tandem repetitions of short nucleotide sequences (microsatellites) within the coding region [5].

As some regions of proteins are less crucial to the functionality to the protein than others it is safe to assume that indels within some regions are less likely to be deleterious than indels in other regions. Short indels that become fixed in the population preferentially occur in solvent accessible loop regions [6]. Longer indel events involve the insertion or deletion of entire protein domains, primarily at the N- and C-termini of proteins [7] but also, when it comes to repeated domains, within the central parts of a protein [7]. The selective pressure acting on these longer indel events is less well understood. However, in the case of repeated proteins it is clear that the duplication of particular domain combinations are strongly favored [8]. The large length variation caused by indels of several protein repeat domains affects binding properties of the proteins, i.e. longer indels events are often associated with functional changes [9].

During the last decade it has become evident that while most proteins contain folded domains, and indeed most proteins contain more than one domain [10], some proteins are partially or even fully disordered [11–13]. These sequences are characterized by two primary features; (i) a low level of hydrophobicity which precludes the formation of a stable globular core; (ii) a high net charge which favors an

☆ This article is part of a Special Issue entitled: The emerging dynamic view of proteins: Protein plasticity in allostery, evolution and self-assembly.

* Corresponding author at: Science for Life Laboratory, Stockholm University SE-171 21 Solna, Sweden. Tel.: +46 706951045.
E-mail addresses: sara.light@scilifelab.se (S. Light), rauan.sagit@scilifelab.se (R. Sagit), diana.ekman@ki.se (D. Ekman), arne@bioinfo.se (A. Elofsson).

[1] Contributed equally.

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.01.002

extended structural state due to electrostatic repulsion [14]. These properties lead to that intrinsically disordered proteins are, in general, more expanded in native conditions than foldable proteins [15].

One important observation concerning intrinsically disordered regions is the fact that they are not at all as common in prokaryotes as in eukaryotes [16], suggesting that disorder could be a component required for higher complexity [17], although it is possible that another reason for this finding is the compactness that characterizes prokaryotic genomes [18]. Intrinsically disordered regions are in general fast evolving, but there are also examples of highly conserved intrinsically disordered regions [14,19]. Further, many intrinsically disordered regions are important for binding [13] and intrinsically disordered regions are a common feature of the hubs in protein–protein interaction network of *Saccharomyces cerevisiae* [20,21].

Here, we present an investigation into insertions and deletions within disordered regions. We show that indels, here defined as regions that are aligned against gaps, contain much more disordered residues than aligned positions. Further, the longer the indel, the more likely that it is disordered. Finally, among the proteins where the disordered region is at least as conserved as the ordered region, we find an overrepresentation of proteins that are involved in processes related to translation.

2. Results and discussion

We have applied two disorder predictors, Iupred [22] and Disopred, to analyze the evolutionary patterns of disordered residues in particular with respect to indels. There are many flavors of protein disorder [13,23]. For instance, short and long disordered regions appear to perform different functional roles, where the short disordered regions often serve as loops in otherwise structurally ordered proteins [16]. Such regions are less conserved than their structured surroundings [24], whereas long disordered regions often are more conserved than the surroundings [16].

The identification of disordered regions is to a large extent performed by using different predictors. What these predictors detect depend on their accuracy as well as what they been trained to identify. The training is often based on missing, or high B-factor, residues in crystal structures. However, these properties also characterize flexible loops. Therefore, there are not exact rules to distinguish between flexible loops and short disordered regions. With this in mind, we have focused our analysis on long disordered regions, by applying a filter, where all predicted regions shorter than 31 residues are removed. Predictions performed using Iupred are based on single sequences and identifies primarily long disordered regions. Disopred on the other hand makes use of multiple sequence alignments and is considered superior for detection of short disordered regions. It is worth noting that even after filtering Disopred predicts about twice as many disordered residues as Iupred, see Table 1, showing that the exact amount of disordered residues is somewhat ambiguous.

The first dataset used herein consists of 3,736 pairs of homologous proteins from two well-annotated eukaryotic genomes — *Caenorhabditis elegans* and *Drosophila melanogaster*. Additionally, in order to avoid possible artifacts that arise as a result of different splice forms, and incorrect gene predictions, present in higher eukaryotes, we have performed the same study on a similar set of 18,389 fungal protein pairs. These results are primarily located in the supplementary material and only discussed briefly in the main text where notable differences compared to the *C. elegans–D. melanogaster* dataset are found.

The correct alignment of distantly related proteins is a genuinely difficult problem and, since sequence alignment methods were developed for structured proteins, it may be particularly troublesome to align distantly related disordered proteins [25]. Here, we have used HMM–HMM alignments methods to obtain the best possible alignments of all protein pairs in our dataset. Although using state of the art methods should minimize the problem, the exact details of the alignments in fast evolving proteins are sometimes difficult to conclusively establish. Nevertheless, the general trends noted here are, to the best of our knowledge, independent of the choice of alignment method.

Table 1
Average proportion of positions and average proportion of gaps per position (GPP) per alignment for the different position types (ordered, disordered and ambiguous) as described in Materials and methods.

Disopred30	Positions [%]	GPP [%]	Positions [%]	GPP [%]
	Invertebrates		Fungi	
Ordered	80	12	80	8
Disordered	17	36	16	23
Ambiguous	3	0	4	0
Iupred30				
Ordered	87	14	88	9
Disordered	9	34	8	21
Ambiguous	4	0	4	0

All indels were classified to be ordered (<25% disordered), disordered (>75% disordered) or mixed (25–75% disordered). The mixed category was small, 5% regardless of disorder prediction method, and was therefore not included below.

2.1. Long indels are disordered

In our analysis, each residue in an alignment is first classified as either ordered or disordered, according to a disorder prediction method, and, further, established as either aligned with a gap, and shall herein be referred to as an indel residue, or with another residue, here referred to as an aligned residue, see Fig. 1.

First, we note that there is a tendency for indels to be longer at the termini, see Fig. 2. It is well known that protein domains often are added at the protein termini [7], which in part explains our observation. But it is also clear that indels shorter than a complete domain are more common at the termini. We find that indels starting at the N-terminal are the longest, spanning on average around 47 residues, compared to 33 residues at the C-terminal while non-terminal indels consist of about six residues. The fungal set shows a similar trend, but the terminal indels are shorter. Naturally, the number of short indels far exceeds the number of long ones, see Fig. 2, both at the termini and internally. However, relatively short indels are less common toward the termini.

Second, for longer indels, the fraction of disordered residues grows as the length of the region increases, see Fig. 3. This suggests that, while small changes may affect ordered regions, disordered regions tend to accept larger changes. This is in agreement with earlier observations of evolution of disordered regions in the centrosomes [26]. Further, given indels of the same length, those located at the internal regions contain more disordered residues than terminal indels.

2.2. Terminal indels are often disordered

Next, we studied the length distribution of ordered and disordered indels. From Fig. 4 it is clear that the shortest indels are the ordered internal indels, followed by disordered internal indels, ordered terminal indels and disordered terminal indels. Actually the length distribution of disordered terminal indels is quite flat. The rapid increase of average disorder in internal indels observed in Fig. 3 can be explained by the fact that ordered internal indels are on average very short (average length 4 residues).

In comparison the frequency of long internal disordered indels is much higher (average length 8 residues). The slower increase of average disorder with length in terminal indels is a consequence of relatively longer ordered indels at the termini (average length 21 residues).

Also, for the larger indels that occur at the termini it is clear, judging by Fig. 3, that indels at the C-termini are slightly more disordered