Stockholm University

This is an accepted version of a paper published in *Biochimica et Biophysica Acta - Proteins and Proteomics*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Light, S., Sagit, R., Ekman, D., Elofsson, A. (2013)

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

Biochimica et Biophysica Acta - Proteins and Proteomics, 1834(5): 890-897

Access to the published version may require subscription.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-88549>

<http://su.diva-portal.org>

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,e,b}

^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*

^d*Medical Biochemistry and Biophysics, Karolinska Institute, SE-17177 Solna, Sweden*

^e*Swedish e-Science Research Center (SeRC)*

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

Email addresses: sara.light@scilifelab.se (Sara Light),
rauan.sagit@scilifelab.se (Rauan Sagit), diana.ekman@ki.se (Diana Ekman),
arne@bioinfo.se (Arne Elofsson)

¹Contributed equally

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.

Keywords: intrinsically disordered proteins, indels, protein evolution, protein structure, sequence alignment

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, 12, 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 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, 17], suggesting that disorder could be a component required for higher complexity [18], although it is possible that another reason for this finding is the compactness that characterizes prokaryotic genomes [19]. In-

trinsically disordered regions are in general fast evolving, but there are also examples of highly conserved intrinsically disordered regions [14, 20]. Further, many intrinsically disordered regions are important for binding [17, 13] and intrinsically disordered regions are a common feature of the hubs in protein-protein interaction network of *Saccharomyces cerevisiae* [21, 22].

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 [23] and Disopred [24], to analyze the evolutionary patterns of disordered residues in particular with respect to indels. There are many flavors of protein disorder [25, 13]. 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 [26], 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 *Caenorhabditis elegans* - *Drosophila 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 [27]. 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.

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 Figure 1.

First, we note that there is a tendency for indels to be longer at the termini, see Figure 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 Figure 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 Figure 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 [28].** 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 Figure 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 Figure 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 Figure 3, that indels at the C-termini are slightly more disordered than those at the N-terminus given the same length. This tendency is exemplified, for

instance, by the essential nucleolar protein KRR-1, one of the many proteins that has a clear cut expansion of disorder at one of the termini, in this case the C-terminus, see Figure 8. In all the fungal homologs of *S. cerevisiae* KRR-1 that were included in the study presented here, there is a lysine-rich disordered region at the C-termini that is substantially less conserved than the rest of the protein and, additionally, of variable length.

As can be seen in Figure 5 there **is significantly more disorder at indel positions compared to aligned positions and there is an increase in disorder among aligned position at the termini. In addition, the fraction of disorder is higher in internal indels compared with termini indels, in agreement with Figure 3.**

As we have shown, disorder is prevalent in indel positions, however it is **also** well known that indels are more frequent in loops than in regular secondary structures. To compare indels and disorder we calculated the Matthews Correlation Coefficient [29] between indels and disorder as well as with other predicted structural features, see Figure 7. It is clear from this figure that disorder, to a greater extent than loops, **overlaps** with indels.

2.3. Some disordered regions are evolutionarily conserved

Although, as we have shown here, disordered regions are quite overrepresented among indels, and indeed disordered regions tend to evolve faster than ordered regions [30], there are also several known exceptions to this rule [30]. Certainly, for a few cases, the protein family in question is evolving rapidly while the disordered region remains conserved [31, 32]. To examine such examples in our dataset, we compared the evolutionary events within the disordered and ordered regions of all proteins pairs, both with regard to

the number of indels in these regions and, additionally, sequence identity, see Figure 6. In general, disordered regions exhibit lower sequence identity and contain more indels but, again, there are a few notable exceptions.

We extracted the proteins where the disordered region is more conserved than the ordered region, retaining only those examples where Iupred30 and Disopred30 agree, thus gathering 14 pairs from *C. elegans* and *D. melanogaster* with this characteristic, see Table 2, and 94 such pairs from the fungal dataset, see Table 3. Utilizing, instead, the union of the protein pairs extracted using either Disopred30 or Iupred30 generated 118 protein pairs.

The number of interaction partners for this set of proteins with conserved disordered regions does not differ from what is observed in the rest of the dataset, Materials and Methods. However, some functional classes of proteins appear to be overrepresented in this set. About half of the proteins are involved in information processing. The two sets of proteins, fungal and fly/worm, were analyzed with regard to functional classification according to EggNog [33]. Most notably, proteins involved in translation (i.e. ribosomal proteins) are clearly overrepresented in both datasets, see Table 4. There is also some overrepresentation of proteins involved in RNA processing, the cytoskeleton, functions pertaining to nuclear structure and intracellular trafficking.

3. Conclusion

Here, we have studied the homologous proteins from *C. elegans* and *D. melanogaster*, as well as homologous fungal proteins, with regard to the disorder content of indels. Due to the difficulty of aligning distantly related

proteins, even using state of the art HMM-HMM alignment methods, in particular disordered proteins, the results should be regarded with a measure of caution. However, given that the results remain essentially the same irrespective of disorder prediction method and dataset used, we are confident the results are robust.

Most strikingly we have found that:

1. Long indels contain more disorder than short indels
2. Terminal indels are longer than internal
3. Internal indels are more disordered than terminal
4. Some proteins involved in information processing contain highly conserved disorder regions.

This connects to previous observations of the flexibility of the length of disordered regions [34]. While most indels are, of course, quite short, this is particularly true for those in the center of the proteins, while terminal indels tend to be longer. Disordered indels are relatively more frequent internally than terminally, in particular those that exceed 10 residues. Further, indels at the termini are generally longer, regardless of whether the inserted/deleted region is disordered or not. Clearly, there is a tendency for structured units, whether protein domains or not, to be added at the protein termini. This is in line with the suggestion that structured domains or structural embellishments may be preferentially added at the termini since such additions do not disrupt the structure to the same extent as long internal insertions [35].

It is possible that there is some characteristic, such as tandem nuclear repeats, at the DNA level that could cause the observed increase of indels in disorder regions. However, in the absence of

proof of such correlations, it seems plausible that the dominant cause of an elevated indel rate in disordered regions is a lower selective pressures for indels. This is in agreement with the idea that the function of at least some disordered regions is performed by sequential motifs important for binding rather than by folded structures. Therefore, further studies are required to discern if disordered regions that fold into specific structures upon binding, may not follow the same trends as we have shown herein.

Finally, while the disordered regions of proteins are generally less conserved, both at the residue level and, as we have here shown, with regard to insertions and deletions, there are a number of proteins where the disordered region is at least as conserved as the ordered region. In accordance with a previous study by Bellay *et al* [36], we found that proteins that contain a highly conserved disordered region tend to be involved in information processes, such as RNA processing and translation.

4. Materials and Methods

4.1. Orthologous protein pairs

Orthologous protein pairs between *Caenorhabditis elegans* and *Drosophila melanogaster* were retrieved from pre-computed homology clusters from In-Paranoid (version 7) [37]. Additionally, an evolutionary distance filter was applied (Tree-Puzzle [38] distance ≤ 4) to avoid inclusion of non-homologs. In total, 3,736 protein pairs were included. Orthologous protein pairs between *Saccharomyces cerevisiae* and five other fungal species (*Candida albicans*, *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis* and

Yarrowia lipolytica) were retrieved from InParanoid (version 7) [37]. An evolutionary distance filter was used as described above. In total, 18,389 protein pairs were included.

4.2. HMM-HMM pairwise alignments

Every sequence in the protein pair set went through a PSI-BLAST [39] run to obtain one PSI-BLAST multiple sequence alignment per sequence. Every pair was aligned using HMM-HMM global pairwise alignment with their respective PSI-BLAST multiple sequence alignments as input to HHalign [40]. HHalign constructed one HMM profile per multiple sequence alignment and aligned the two HMM profiles. These terminal indels were then added in order to get full global pairwise alignments for the protein pairs.

4.3. Intrinsic disorder predictions

The intrinsic disorder was predicted with Iupred [23] and Disopred [24]. For Iupred, the “long” setting was used. Additionally, two more predictions, “Iupred30” and “Disopred30”, were added. These were generated by using the corresponding predictor results and only accepting disordered regions of at least >30 as disordered and marking the rest as ordered.

4.4. Internal, N-terminal, and C-terminal indels

Every indel was classified as N-terminal, **C-terminal or internal based on the profile-profile alignment**. An N-terminal indel is defined as a residues aligned to a gap before the first position in a pairwise alignment, and a C-terminal indel is a gap found after the last matching position. An internal indel is a gap found anywhere between the first and last matching positions. The different types

of indels and the different types of positions (ordered, disordered and ambiguous) are shown in Figure 1. Within the *C. elegans* and *D. melanogaster* protein pairs, there are 44,236 internal, 2,631 N-terminal and 2,891 C-terminal indels. Within the Fungi protein pairs, there are 168,033 internal, 9,531 N-terminal and 11,170 C-terminal indels.

4.5. Ordered, disordered and ambiguous positions

Global pairwise alignments were created to compare orthologous protein pairs. Every position in a pairwise alignment was classified according to the terminology in Figure 1. A position is ordered when there are two ordered aligned residues or one ordered residue aligned with a gap. A position is disordered when there are two disordered aligned residues or one disordered residue aligned with a gap. A position is ambiguous when there is one ordered residue aligned with one disordered residue.

Within the *C. elegans* and *D. melanogaster* protein pairs, with intrinsic disorder predicted by Disopred30, there are 1,797,352 ordered, 484,394 disordered and 90,484 ambiguous positions. Within the Fungi protein pairs, with intrinsic disorder predicted by Disopred30, there are 7,744,350 ordered, 1,925,337 disordered and 344,531 ambiguous positions. Note that a position can be residue-residue or residue-gap. Except for ambiguous positions which are always residue-residue.

4.6. Conserved disordered regions

The conserved disordered regions are found by comparing the ordered and disordered regions of a pairwise alignment, see Figure 6. To include every long

disordered region predicted with Iupred30 and Disopred30, the ambiguous positions for these predictions were included as disordered positions.

For the assessment of the connectivity of the conserved disordered proteins, the protein interaction dataset was downloaded from BioGRID [41], version 3.1.84. The connectivity for the conserved disordered proteins was 3.0 and 6.2 for *C. elegans* and *D. melanogaster* respectively. This was not significantly different from the corresponding values for the entire dataset (3.8 and 7.1), according to a one sample t-test with confidence interval 95%. The functional categories associated with conserved disordered regions were collected from EggNog [33]. Z-scores were calculated by repeated (100,000 times) random sampling of 236 proteins from the set of *C. elegans* and *D. melanogaster* proteins, followed by application of the formula $(x - \mu)/\sigma$, where x is the observed value, μ is the average for the randomized population and σ is the standard deviation for the randomized population. The same procedure was performed on the fungal dataset consisting of 1096 proteins.

4.7. Matthews correlation coefficient

The Matthews correlation coefficient [29] (MCC) was calculated to measure the level of correlation between different properties. The MCC is calculated based on a confusion matrix that has four cells; true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). For example, when comparing Iupred30 and Disopred30, true positives are all positions that are disordered with both Iupred30 and Disopred30, false positives are ordered with Iupred30 but disordered with Disopred30, true negatives are ordered with both Iupred30 and Disopred30, and false negatives are disordered with Iupred30 but ordered with Disopred30. These four values are

then inserted in the following equation to calculate MCC.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

5. Acknowledgments

This work was supported by grants from the Swedish Research Council (VR-NT 2009-5072, VR-M 2010-3555), SSF, the Foundation for Strategic Research, Science for Life Laboratory the EU 6'th Framework Program is gratefully acknowledged for support to the GeneFun project, contract No: LSHG-CT-2004-503567 and the 7'th framework through the EDICT project, contract No: FP7-HEALTH-F4-2007-201924. Funding for SL was provided by BILS, Bioinformatics Infrastructure for Life Science.

References

- [1] M. Kalman, N. Ben-Tal, Quality assessment of protein model-structures using evolutionary conservation., *Bioinformatics* 26 (2010) 1299–1307.
- [2] G. Levinson, G. Gutman, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Mol. Biol. Evol.* 4 (1987) 203–221.
- [3] M. G. Giacomelli, A. S. Hancock, J. Masel, The conversion of 3' UTRs into coding regions., *Molecular biology and evolution* 24 (2007) 457–64.
- [4] A. K. Björklund, D. Ekman, A. Elofsson, Expansion of protein domain repeats, *PLoS Comp. Biol.* 2 (2006) e114.
- [5] W. J. Guo, J. Ling, S. P. Ye, Significant comparative characteristics between orphan and nonorphan genes in the rice (*oryza sativa* l.) genome, *Comp Funct Genomics* 21676 (2007).
- [6] R. Kim, J. T. Guo, Systematic analysis of short internal indels and their impact on protein folding, *BMC Struct. Biol.* 10 (2010) 24.
- [7] A. K. Björklund, D. Ekman, S. Light, J. Frey-Skott, A. Elofsson, Domain rearrangements in protein evolution, *J. Mol. Biol.* 353 (2005) 911–923.
- [8] A. K. Björklund, S. Light, R. Sagit, A. Elofsson, Nebulin: a study of protein repeat evolution, *J. Mol. Biol.* 402 (2010) 38–51.
- [9] A. Pascual-García, D. Abia, R. Mendez, G. S. Nido, U. Bastolla, Quantifying the evolutionary divergence of protein structures: the role of

- function change and function conservation., *Proteins: Structure, Function, and Bioinformatics* 78 (2010) 181–196.
- [10] D. Ekman, A. K. Björklund, J. Frey-Skott, A. Elofsson, Multi-domain proteins in the three kingdoms of like - orphan domains and other unassigned regions, *J. Mol. Biol.* 348 (2005) 231–243.
 - [11] P. E. Wright, H. J. Dyson, Intrinsically unstructured proteins: reassessing the protein structure-function paradigm., *J Mol Biol* 293 (1999) 321–331.
 - [12] V. N. Uversky, J. R. Gillespie, A. L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins* 41 (2000) 415–427.
 - [13] A. K. Dunker, I. Silman, V. N. Uversky, J. L. Sussman, Function and structure of inherently disordered proteins, *Curr. Opin. Struct. Biol.* 18 (2008) 756–764.
 - [14] P. Tompa, *From Protein Structure to Function with Bioinformatics*, Springer Science + Business Media B.V., Springer Science+Business Media UK Ltd. 236 Gray's Inn Road Floor 6 London WC1X 8HB United Kingdom tel: + 44 (0) 20 3192 2000, pp. 113–140.
 - [15] S. Müller-Späth, A. Soranno, V. Hirschefeld, H. Hofmann, S. Rüdiger, L. Reymond, D. Nettels, B. Schuler, From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins., *Proceedings of the National Academy of Sciences of the United States of America* 107 (2010) 14609–14.

- [16] J. Liu, H. Tan, B. Rost, Loopy proteins appear conserved in evolution., *J. Mol. Biol.* 322 (2002) 53–64.
- [17] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol.* 337 (2004) 635–645.
- [18] A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta, B. Rost, Protein disorder—a breakthrough invention of evolution?, *Curr Opin Struct Biol* 21 (2011) 412–418.
- [19] J. Zhang, Protein-length distributions for the three domains of life., *Trends Genet* 16 (2000) 107–109.
- [20] Z. Dosztanyi, B. Meszaros, I. Simon, Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, *Brief. Bioinformatics* 11 (2010) 225–243.
- [21] D. Ekman, S. Light, A. Björklund, A. Elofsson, What properties characterize the hub proteins of the protein-protein interaction network of *saccharomyces cerevisiae*?, *Genome Biol* 7 (2006) R45.
- [22] A. Dunker, M. Cortese, P. Romero, L. Iakoucheva, V. Uversky, Flexible nets. the roles of intrinsic disorder in protein interaction networks., *FEBS J.* 272 (2005) 5129–5148.
- [23] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, *J. Mol. Biol.* 347 (2005) 827–839.

- [24] D. Jones, J. Ward, Prediction of disordered regions in proteins from position specific score matrices., *Proteins* 53 Suppl 6 (2003) 573–578.
- [25] S. Vucetic, C. J. Brown, A. K. Dunker, Z. Obradovic, Flavors of protein disorder., *Proteins* 52 (2003) 573–584.
- [26] B. Rost, C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure., *Proteins* 19 (1994) 55–72.
- [27] P. Radivojac, Z. Obradovic, C. J. Brown, A. K. Dunker, Improving sequence alignments for intrinsically disordered proteins., *Pac Symp Biocomput* (2002) 589–600.
- [28] G. Nido, R. Mendez, A. Pascual-Garcia, D. Abia, U. Bastolla, Protein disorder in the centrosome correlates with complexity in cell types number., *Mol Biosyst* 8 (2012) 353–367.
- [29] B. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme., *Biochim Biophys Acta* 405 (1975) 442–451.
- [30] C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, A. K. Dunker, Evolutionary rate heterogeneity in proteins with long disordered regions., *Journal of molecular evolution* 55 (2002) 104–10.
- [31] G. W. Daughdrill, P. Narayanaswami, S. H. Gilmore, A. Belczyk, C. J. Brown, Dynamic behavior of an intrinsically unstructured linker domain

- is conserved in the face of negligible amino acid sequence conservation., *J Mol Evol* 65 (2007) 277–288.
- [32] A. J. Ayme-Southgate, R. J. Southgate, R. A. Philipp, E. E. Sotka, C. Kramp, The myofibrillar protein, projectin, is highly conserved across insect evolution except for its pevk domain., *J Mol Evol* 67 (2008) 653–669.
 - [33] L. J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, P. Bork, eggno: automated construction and annotation of orthologous groups of genes., *Nucleic Acids Res* 36 (2008) D250–D254.
 - [34] J. Siltberg-Liberles, Evolution of structurally disordered proteins promotes neostructuralization., *Mol Biol Evol* 28 (2011) 59–62.
 - [35] G. A. Reeves, T. J. Dallman, O. C. Redfern, A. Akpor, C. A. Orengo, Structural diversity of domain superfamilies in the cath database, *J. Mol. Biol.* 360 (2006) 725–741.
 - [36] J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers, P. M. Kim, Bringing order to protein disorder through comparative genomics and genetic interactions., *Genome Biol* 12 (2011) R14.
 - [37] M. Remm, C. E. Storm, E. L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *J. Mol. Biol.* 314 (2001) 1041–1052.

- [38] H. A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing, *Bioinformatics* 18 (2002) 502–504.
- [39] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [40] J. Söding, Protein homology detection by HMM-HMM comparison, *Bioinformatics* 21 (2005) 951–960.
- [41] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, M. Tyers, The BioGRID Interaction Database: 2011 update., *Nucleic acids research* 39 (2011) D698–704.
- [42] D. Jones, Protein secondary structure prediction based on position-specific scoring matrices., *J. Mol. Biol.* 292 (1999) 195–202.
- [43] M. Clamp, J. Cuff, S. Searle, G. Barton, The jalview java alignment editor, *Bioinformatics* 20 (2004) 426.
- [44] R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput., *Nucleic acids research* 32 (2004) 1792–7.
- [45] J. Cherry, C. Adler, C. Ball, S. Chervitz, S. Dwight, E. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, D. Botstein, SGD: *Saccharomyces* genome database., *Nucleic Acids Res* 26 (1998) 73–79.

Figures

Figure 1: Terminology example. The image shows a schematic representation of two aligned proteins, represented by boxes. Thick boxes show disordered regions and thin boxes show ordered regions. Different types of aligned and unaligned positions are marked with ellipses.

Figure 2: Length distribution of indels located at different distances from the protein termini. N; N-terminal indels. C; C-terminal indels. Internal; internal indels. **See materials and methods for how the indels are defined.** The inset figure shows the average length of indels of the three different location types, including standard errors. A two sample t-test shows that the average length of internal indels differs to the termini indels with a P-value $< 10^{-100}$. While the N-terminus and C-terminus average indel lengths differ with a P-value of 8×10^{-10} .

Figure 3: Length and intrinsic disorder for internal and terminal indels. The **inserted** text in the graphs shows which intrinsic disorder prediction that is used (Disopred30 in the left panel and IUPred30 in the right panel). Standard errors for every tenth x-value are included in the graphs.

Figure 4: Length of ordered and intrinsically disordered indels with respect to whether they are terminal or internal. Ordered indels are gaps that align against residues that have $<25\%$ disorder. Disordered indels are those that have $>75\%$ disorder. The vast majority of all indels fall into one of these two categories and the rest (around 5% of all indels) are ignored. Internal and terminal indels are defined as described in Materials and Methods. The inserted text shows which intrinsic disorder prediction that was used.

Figure 5: Intrinsic disorder at aligned regions and indels for internal and terminal alignment positions. Standard errors are shown for every bar. The N-terminal, C-terminal and internal indels are defined as in Materials and Methods. The aligned regions are defined in a different manner. The N-terminal aligned region is the first 50 matching alignment positions. The C-terminal aligned region is the last 50 matching positions. The internal aligned region is all matching positions. The bars show the average disorder per alignment. The inset text shows the disorder prediction. Two sample t-tests between aligned regions and indels for Disopred30 predictions all have highly significant P-values $< 3 \times 10^{-12}$. The same is valid for Iupred30 predictions. Two sample t-tests between indels for Disopred30 are all highly significant except for N vs. C = 3×10^{-2} . The corresponding P-values for Iupred30 are N vs. Internal = 1×10^{-2} , N vs. C = 7×10^{-1} and C vs. Internal = 5×10^{-2} .

Figure 6: Sequence identity (SeqID) and gaps per position (GPP) between the ordered and disordered regions of every pairwise alignment. Only alignments with 30 or more ordered and 30 or more disordered matching positions were included. SeqID was calculated based on matching positions only. The two upper panels compare the SeqID [%] while the lower two panels compare the GPP [%] of the ordered and disordered regions.

Figure 7: Matthews correlation coefficients between indels and different predicted properties. The secondary structure elements coils, helices and strands were predicted using Psipred V2.3 [42].

Tables

	Positions [%]	GPP [%]	Positions [%]	GPP [%]
Disopred30	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

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.

CEL Protein	DME Protein	CEL Function	DME Function	ID Ord [%]	ID Dis [%]
CE03253	FBpp0091091	histone H3	histone H3	97	98
CE02131	Fbpp0079168	ribosomal	ribosomal	79	89
CE26774	FBpp0077142	ribosomal	ribosomal	57	85
CE04691	Fbpp0081845	ribosomal	ribosomal	72	79
CE16127	FBpp0083973	Ca-signalling	synaptic vesicle	67	71
CE29829	FBpp0074298	muscle protein	muscle protein	44	62
CE09008	FBpp0081552	chaperone	protein folding	42	50
CE01104	FBpp0077214	protein-bind.	unknown	34	36
CE41557	FBpp0083319	RNA-bind.	RNA-bind.	30	31
CE27796	FBpp0077645	unknown	phospholipase act.	22	30
CE23858	FBpp0087063	cell division	nucleocytoplasmic	25	27
CE29953	FBpp0078766	lipid metabolism	unknown	25	27
CE10508	FBpp0083356	telomeric	transcription	18	19
CE39656	FBpp0082093	regulation	golgi function	16	19

Table 2: Homologous protein pairs where the disordered region is more conserved than the ordered region for *C. elegans* (CEL) and *D. melanogaster* (DME). Functional annotations were retrieved from UniProt and disorder content was assigned using Disopred30 [24]. Only pairs where Disopred30 [24] and Iupred30 [23] both show higher conservation for the disordered region are shown in this table.

SCE Protein	OTH Protein	SCE Function	ID Ord [%]	ID Dis [%]
YBR010W	DEHA2D12694g	Nucleosome	98	98
YBR010W	CAL0001203	Nucleosome	95	95
YGL103W	CAGL0J10384g	60S ribosomal	90	90
YLR249W	CAGL0B03487g	Elongation factor	88	88
YLR249W	KLLA0F12210g	Elongation factor	86	86
YGL103W	DEHA2A12782g	60S ribosomal	81	81
YER165W	CAGL0L11792g	mRNA-binding	79	79
YLR249W	DEHA2F24398g	Elongation factor	77	77
YGL103W	YALI0B21076g	60S ribosomal	77	77
YPL090C	DEHA2B02904g	40S ribosomal	77	77
YGL103W	CAF0006966	60S ribosomal	77	77
YGR285C	KLLA0F03124g	Ribosome associated	76	76
YLR249W	CAL0005746	Elongation factor	76	76
YOR091W	CAGL0J08679g	Unknown	72	72
YLR249W	YALI0E13277g	Elongation factor	70	70
YOR294W	DEHA2F09834g	Ribosome biogenesis	69	69
YJL122W	CAGL0L05500g	Ribosome assembly	62	62
YBR155W	CAGL0H07425g	Hsp90 protein binding	61	61
YPL106C	CAL0001523	Protein folding	60	60
YNL059C	KLLA0F03421g	Chromatin remodeling	60	60
YHR088W	YALI0F21483g	Aminoacyl-Trna ligase	60	60
YBR269C	CAGL0I09394g	Unknown	60	60
YPL106C	DEHA2D15840g	Protein folding	59	59
YOR051C	CAGL0L10164g	Translational termination	57	57
YMR309C	KLLA0F03014g	Translation initiation	57	57
YBR269C	KLLA0F01639g	Unknown	57	57
YER019C-A	CAGL0A04191g	Protein targeting to ER	55	55
YGR285C	YALI0A00594g	Ribosome associated	53	53
YBR269C	DEHA2A04862g	Unknown	53	53
YNL151C	CAGL0J05698g	RNA polymerase III	52	52
YGL151W	CAGL0E05060g	RNA polymerase related	50	50
YDL115C	CAGL0H10472g	General transcription	49	49
YJL122W	KLLA0E15093g	Ribosome assembly	49	49
YOR281C	YALI0D13090g	Regulation of transcription	48	48
YNR021W	CAGL0M10615g	Unknown	48	48

Continued on next page

Table 3 – continued from previous page

SCE Protein	OTH Protein	SCE Function	ID Ord [%]	ID Dis [%]
YLR439W	KLLA0B05181g	54S ribosomal protein L4	47	47
YLR295C	KLLA0A04543g	ATP synthase related	47	47
YNL059C	CAL0000047	Chromatin remodeling	46	46
YNL059C	DEHA2F21076g	Chromatin remodeling	45	45
YEL055C	CAGL0B03553g	Dna-Directed DNA polymerase	45	45
YKL117W	DEHA2E19866g	Regulation of telomere maintenance	42	42
YNL059C	YALI0B10538g	Chromatin remodeling	42	42
YPR137W	DEHA2E09658g	Snorna binding protein	39	39
YBL004W	CAL0002346	rRNA processing	39	39
YOR158W	CAGL0L04334g	Mitochondrial ribosomal	39	39
YMR014W	CAGL0M07678g	Ribosomal small subunit biogenesis	36	36
YBR055C	CAGL0H10098g	Mrna-Splicing factor	36	36
YDR432W	DEHA2D04400g	mRNA export from nucleus	35	35
YLR095C	KLLA0F19514g	ATP synthase related	35	35
YBR269C	CAL0005646	Unknown	35	35
YNL050C	CAGL0G01276g	Unknown	35	35
YDR432W	CAL0003843	mRNA export from nucleus	33	33
YBR055C	KLLA0F04191g	mRNA-Splicing factor	33	33
YGR119C	YALI0A20702g	RNA transport	31	31
YMR014W	KLLA0D11220g	Ribosomal small subunit biogenesis	30	30
YOR276W	DEHA2F10868g	Translation	30	30
YNL050C	KLLA0F03707g	Unknown	30	30
YBL014C	CAGL0E02893g	Transcription	29	29
YNL119W	YALI0B10263g	tRNA-related	29	29
YJL073W	KLLA0D06402g	Heat shock protein binding	29	29
YLR437C	KLLA0B05423g	Import facilitator	29	29
YLR295C	DEHA2F02024g	ATP synthase related	29	29
YOR051C	YALI0D22132g	Translational termination	28	28
YKL079W	CAGL0L05104g	Microtubule motor protein	28	28
YDL218W	KLLA0D08646g	Unknown	27	27
YHR052W	CAL0003033	Ribosomal biogenesis	27	27
YKR025W	YALI0D19448g	RNA polymerase	26	26
YJL157C	KLLA0B07469g	Response to pheromone	26	26
YMR211W	YALI0E10285g	Protein polymerization	26	26

Continued on next page

Table 3 – continued from previous page

SCE Protein	OTH Protein	SCE Function	ID Ord [%]	ID Dis [%]
YGL172W	YALI0C23166g	RNA transport	26	26
YNL206C	YALI0A08107g	Histones H3 and H4 chaperones	26	26
YIL154C	DEHA2E08668g	Transcription activator	26	26
YJL203W	DEHA2B02926g	Spliceosome	26	26
YIL154C	CAL0002294	Transcription activator	26	26
YPL068C	CAGL0H05247g	Unknown	26	26
YNL127W	YALI0C09779g	Re-entry into mitotic cell cycle	25	25
YPL210C	KLLA0B02288g	in protein targeting to ER	25	25
YNR021W	YALI0F10703g	Unknown	24	24
YPL083C	YALI0A18293g	Spliceosome	24	24
YER083C	DEHA2B10626g	posttranslational	23	23
YEL009C	YALI0E27742g	Protein dimerization	23	23
YLL003W	KLLA0F12958g	spindle pole body duplication	22	22
YHR195W	CAGL0M07205g	Nuclear envelope	22	22
YGR184C	YALI0B03718g	Ubiquitin-Protein ligase	20	20
YER083C	CAL0005366	Posttranslational	20	20
YPL210C	YALI0E34309g	Protein targeting to ER	20	20
YJL073W	CAL0006020	Heat shock protein binding	20	20
YLL003W	CAGL0D05764g	Spindle pole body duplication	19	19
YKR006C	YALI0D21934g	Mitochondrial ribosomal	19	19
YHR182W	KLLA0A09317g	Unknown	19	19
YBL014C	CAL0001777	Transcription	17	17
YMR198W	CAGL0L10626g	Microtubule motor protein	16	16
YHR197W	YALI0F20350g	Pre-mRNA processing	16	16
YDR210C-C	KLLA0E25180g	Transposable element protein	15	15

Table 3: Homologous protein pairs where the disordered region is more conserved than the ordered region for the fungal dataset. Functional annotations for the *Saccharomyces cerevisiae* proteins were retrieved from EggNog [33] and SGD [45] and disorder content was assigned using Disopred30 [24]. Only pairs where Disopred30 [24] and Iupred30 [23] both show higher conservation for the disordered region are shown in this table.

Function	CEL/DME	Fungal
Cytoskeleton	1.5 (0.9)	2.2 (2.7)
Intracell. trafficking	3.3 (6.8)	1.4 (0.3)
Nuclear structure	7.4 (4.5)	3.3 (2.7)
RNA processing	1.6 (1.3)	1.7 (1.6)
Translation	3.7 (7.4)	2.7 (8.6)

Table 4: Functional categories overrepresented among the proteins where the disordered region is at least as conserved as the ordered region. Functional categories were extracted using EggNog [33]. The numbers show the ratio between the observed number of proteins for each category divided by the number of expected proteins from the same category given the corresponding proteomes. Z-scores (in parenthesis) were calculated using 100,000 random samples.