Natural Selection of Intrinsic Disorder Characteristic of Proteins

Elfalem Alemu, Emily Jones, & Sridhar Hannenhalli
Center for Bioinformatics and Computational Biology, University of Maryland

Abstract

Intrinsically disordered proteins (IDP) lack stable tertiary structure and are implicated in cell signaling, transcription, and chromatin remodeling. They are a relatively recent in molecular evolution and are highly regulated due to their low affinity for their substrates. A previous study found that *in silico* evolution preserves secondary structure, but not intrinsic disorder (ID), suggesting that the disordered structure of proteins is selected for even when the sequence is not conserved. To test this, we developed a general framework for determining selection for protein properties not directly encoded in sequence. Our algorithm uses protein orthologs to create synthetic proteins using the same parameters as *in vivo* sequence evolution, then compares the difference in disorder conservation between the original orthologs to the difference between the originals and synthetics, measuring purifying selection for disorder by comparing selected mutations to nonselected ones. Our algorithm shows enrichment for natural selection in more disordered proteins and between less conserved sequences as expected. We plan to examine the enrichment for motifs and hub proteins within significantly selected disordered proteins and specific well-known disordered proteins as case studies of significant selection. We also intend to use our algorithm for measuring purifying selection of other protein properties not directly encoded in the sequence, such as overall charge or polarity.

Background

One-third of all eukaryotic proteins have an ID region at least 30 residues long. This disorder can be found in links between domains; proteins and can either become ordered upon binding or retain their disorder. IDPs have a larger interaction surface, most post-translational modifications, and have many substrates, which they bind with low affinity but high specificity. These properties lead them to be highly regulated through high rates of decay of ID RNA transcripts and low rates of synthesis and shorter half-lives of ID proteins. This low expression is correlated with faster rates of evolution (Brown et al, 2011).

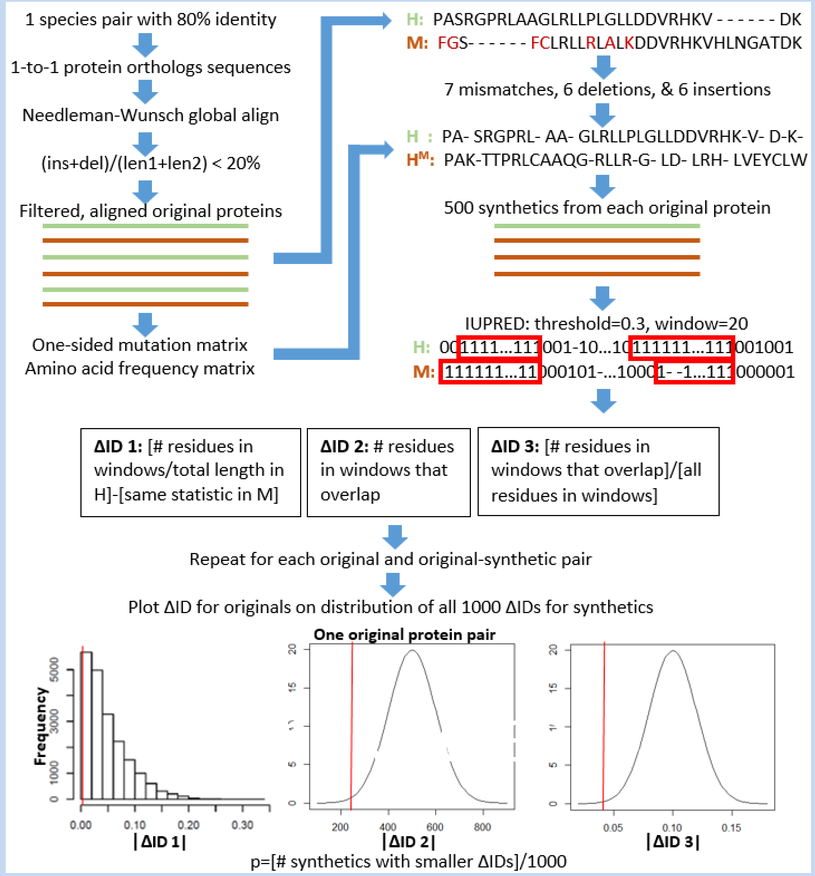
A previous study found that *in silico* evolution preserves secondary structure, but not intrinsic disorder (Schaefer et al, 2010). Interestingly, most IDPs evolve faster likely due to lack of structural constraints; thus, sequence conservation is not required to maintain dynamic behavior (Dosztányi et al, 2010). This suggests that disorder is selected for even when the sequence is not conserved. Sequence evolution can easily be measured by pairwise distances or alignment scores. In contrast, structural evolution requires study of interactions between residues and thus cannot be predicted based on sequence alone, so there are few models to measure structural selection.

The goal of this project is two-fold: to develop a general framework for determining selection for protein properties not directly encoded in sequence and to measure properties associated with proteins for which disorder is significantly selected for.

Methods

To measure selection, we here mimic sequence evolution using the same parameters as *in vivo* sequence evolution. First, we identify species pairs that shared 80% identity across the protein coding genes: human and mouse, *Drosophila melanogaster* and *D. pesudoobscura*, and *Sacchromyces cerevisea* and *S. bayanus*. We then obtained the protein sequences for all 1-to-1 protein orthologs between each pair of species: human and mouse from Ensembl, fly from FlyBase with sequences from Batch Entrez, and yeast from Kellis et al, 2003 with sequences from the Yeast Genome Order Browser. Next, we globally aligned each protein pair using an open-source Needleman-Wunsch tool with a BLOSUM62 matrix. From this alignment, we generated a one-sided substitution rate matrix for all amino acids from the entire proteome; then, between each orthologous protein pair, we counted the number of deletions, insertions, and mismatches. From this, we would generate synthetic proteins.

After creating synthetics, we calculated the disorder score of each protein through IUPRED, which assigns a disorder score to each amino acid based on the pairwise energy score. We defined a disordered residue as one in a series of residues of a minimum length which all have scores above a threshold of 0.3, then calculated differential disorder in three ways: first by subtracting the number of disordered residues in one protein in a pair from the number in another, second by aligning sequences and counting the overlapping disordered residues, and third by dividing the second score by the total number of disordered residues. For each protein pair, we calculated a p-value based on where the differential ID score of the real-real pair falls on the score distribution of the real-synthetic pairs, putting the structural selection in the context of sequence conservation. A significant p-value indicates significant purifying selection. Thus, our algorithm asks: to what extent the structure would change if evolution preserved sequence, but not structure?



Results

P-value Enrichment [E=(% sig)/.05]

pid \ meanidp	[0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
[0,0.5]	0.2667	0	0.8333	1.5385	1.5094
	75	43	24	26	53
(0.5,0.65]	0.4717	0.1802	0.4211	0.5769	2.963
	212	111	95	104	108
(0.65,0.8]	0.4199	0.7595	0.7792	1.81	1.8357
	762	316	231	221	207
(0.8,0.9]	0.2539	0.793	0.8581	1.4035	1.2931
	1024	454	303	228	232
(0.9,1]	0.135	0.53	0.542	0.9589	1.0256
	1185	566	369	292	234

Table 1: P-value enrichment (the number of significant proteins divided by the number expected by chance) for bins divided up by percent identity and mean ID 1 of the two original human and mouse proteins in a pair. Bins with ID 1 above 0.6 and percent identity below 90% are enriched for significance; thus, disorder is more selected for in more disordered proteins, and our tool can measure significant disorder selection even with high sequence conservation.

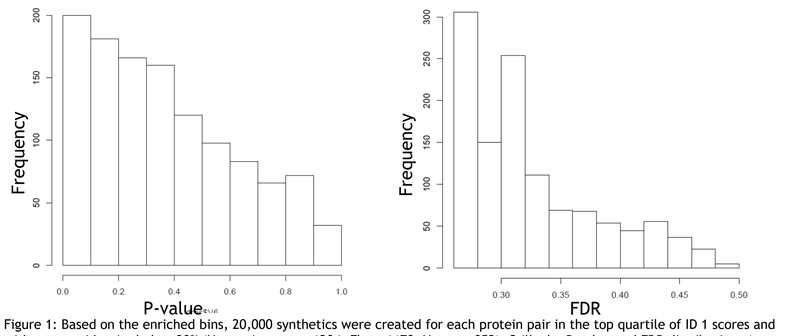


Figure 1: Based on the enriched bins, 20,000 synthetics were created for each protein pair in the top quartile of ID 1 scores and with percent identity below 90% (Human/mouse n=1284, Fly n=1178, Yeast n=952). Still, the P-value and FDR distributions (as shown here for fly protein pairs) were too high.

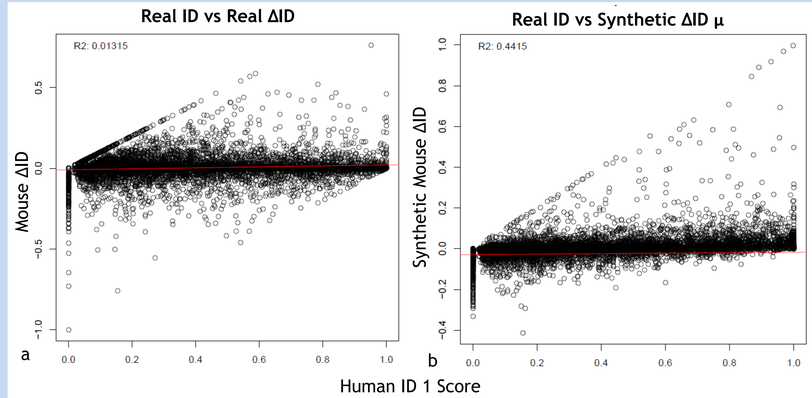


Figure 2: Scatterplot of original human ID 1 score vs. original mouse ΔID 1 score (a) or mean synthetic mouse ΔID 1 score (b). There is no correlation in either plot, thus the disorder of any protein does not bias its ΔID. Furthermore, this plot suggests that disorder is more easily lost than gained with random mutations, as less disordered (lower ID) human proteins can gain disorder (more negative ΔID) when compared to their mouse orthologs, but not when compared to synthetics (lowest ΔID is -0.4). The same trend was seen in fly and yeast populations.

Examples

Below are two examples of human proteins we found to be disordered (IDP 1 > 0.6) and significantly selected for ($p<0.05$) that are involved in human pathogenesis.

PRPF40B: PRP40 pre-mRNA processing factor 40 homolog B (ENSP00000369634)
Homologous to ENSMUSP00000119556
Human IDP 1 score: 0.798, Mouse IDP 1 score: 0.867, ΔIDP 1: 0.069
Synthetic IDP 1 μ: 0.550, Synthetic IDP 1 σ: 0.197
P-value: 0.005
Interacts with Huntingtin and MeCP2. Truncation of the WW-domain is involved in Huntington and Rett Syndrome pathogenesis.

ZNF469: Zinc Finger Protein 469 (ENSP00000402343)
Homologous to ENSMUSP00000057897
Human IDP 1 score: 0.958, Mouse IDP 1 score: 0.838, Δ IDP 1: 0.120
Synthetic IDP 1 μ: 0.534, Synthetic IDP 1 σ : 0.178
P-value: ~0
A zinc-finger protein which may act as a transcription factor for collagen fiber synthesis. Mutations cause brittle cornea syndrome.

Discussion

Our general algorithm for analyzing selection is working: p-value enrichment is as expected and conservation of sequences does not necessarily conserve structure as predicted. However, our current methods return very few significant protein pairs: Human/mouse: 245/7475, Drosophila: 181/5075, Saccharomyces: 154/4347.

Future Directions

- Repeat disorder calculation with RONN
- Look for well-known conserved disordered proteins as case studies of significant selection
- Compare enrichment of significant selection in hub vs non-hub disordered proteins
- Examine motif (MoRF and ELM) enrichment in proteins with significant selection
- Look for greater protein interaction between significant proteins across species by computing the shortest distance pair on a protein interaction network between significant disordered proteins in one species pair and another
- Use the pipeline to measure selection in specific disease-causing proteins and in other protein properties, such as overall charge or polarity

References
Brown, C.J., Johnson, A.K., Dunker, A.K, and Daughdrill, G.W. (2011). Evolution and Disorder. Curr Opin Struct Biol. 21(3):441-446.
Dosztányi, Z., Mészáros, B., and Simon, I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform. 11 (2): 225-243.
Schaefer, C., Schlessinger, A., and Rost, B. (2010). Protein secondary structure appears to be robust under in silico

Funding
Research reported in this poster was supported by the National Institutes of Health under award number R01GM100335. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.