Intrinsically disorder in proteins (IDP) is the lack of stable tertiary structure. Intrinsic disorder (ID) is enriched in proteins implicated in cell signaling, transcription, differentiation, and chromatin remodeling and depleted in integral membrane and catalytic proteins. Intrinsic disorder is a relatively recent evolutionary phenomena, and one-third of all eukaryotic proteins have an intrinsic disorder 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).

Little is known about the evolution of intrinsic disorder. A previous study found that *in silico* evolution preserves secondary structure, but not intrinsic disorder (Schaefer et al, 2010). Interestingly, most IDPs evolve faster 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. This project is an extension of my current research; we have developed the algorithm in the following paragraph, but all other work is novel.

To measure selection, we would mimic sequence evolution using the same parameters as *in vivo* sequence evolution. First, we would identify species pairs that shared 80% identity across the entire genome, such as human and mouse. Then, we would obtain the protein sequences for all 1-to-1 protein orthologs and align each pair. From this alignment, we would generate a one-sided mutation matrix for all amino acids from the entire proteome; then, between each pair, we would count the number of deletions, insertions, and mismatches. From this, we would generate synthetic proteins; for instance, given that a human protein aligned with a mouse protein had 2 insertions and 1 mismatch, we would create synthetic mouse proteins from the human sequence by randomly inserting two amino acids and mutating one based on the generated mutation matrix. This algorithm would be more true to *in vivo* evolution than most *in silico* evolution studies because the mutation matrix would be true to the input proteome and insertions and deletions would be considered.

After creating synthetics, we would calculate the disorder score of each protein through two prediction tools: IUPRED, which assigns a disorder score to each amino acid based on the pairwise energy score, and RONN, which aligns input sequences to those with experimentally verified disordered segments. We would define a disordered residue as one in a series of residues of a minimum length which all have scores above a threshold, then calculate disorder in two ways: first by subtracting the number of disordered residues in one protein in a pair from the number in another, and second by aligning sequences and counting the overlapping disordered residues. For each protein pair, we could calculate a p-value based on where the ID score of the real-real pair falls on the ID score distribution of the real-synthetic pairs, putting the structural selection in the context of sequence conservation. A significant p-value would indicate significant purifying selection. Thus, our algorithm asks: what would occur if evolution preserved sequence, but not structure?

IDPs have many unique biologically relevant characteristics; to test whether selection for disorder is correlated with these properties, we would examine enrichment of significant p-values in IDPs that hold these properties. First, since hub proteins commonly use disordered regions to bind multiple partners, hub proteins should have greater selected disorder. This could be measured by comparing significant p-value enrichment in hub vs non-hub IDPs using protein-protein interaction

networks. Second, IDPs are enriched in two types of special motifs: molecular recognition features (MoRFs), short sequences that undergo disorder to order transitions upon binding, and eukaryotic linear motifs (ELMs), binding motifs in a linker or tail which are conserved within a larger region of nonconserved sequence. Third, disordered regions are enriched for post-translational modifications which enable binding and prevent folding; this may allow conserved disordered without conserved sequence. Finally, it has been predicted that 80% of all alternative splicing events occur in disordered regions; this also would help explain the low sequence conservation (Dunker, 2008). Thus, there should be enrichment for motifs, post-translational modification, and alternative splicing events in disordered portions of significantly selected IDPs.

The pipeline could also be modified to map disordered regions first, then create synthetics for the ordered and disordered segments separately and recombine to create the final protein. Although disordered regions are known to have higher evolutionary rates, this may be because common mutation matrices (such as PAM or BLOSUM) are based on proteins whose structures are ordered. In fact, disorder-promoting amino acids are more conserved in disordered regions and may only appear to have faster mutation rates due to the fact that charged and hydrophilic residues, which make up most of disordered regions, have a higher conversion rate to other charged and hydrophilic amino acids than do aromatic or hydrophobic residues which make up ordered regions (Brown et al, 2011). By creating separate alignments and mutation matrices for disordered and ordered regions, this would allow us to emulate how disordered and ordered protein regions evolve differently, which has yet to be modeled.

This pipeline could be used as a tool to both measure and teach Darwinian evolution of protein structure. Previous tools have only measured conservation of properties directly encoded in the sequence. Researchers could use this algorithm to measure whether other protein properties, such as overall charge, polarity, and post-translational modifications, are selected for. Teachers could also use this pipeline as a teaching tool to demonstrate how natural selection works beyond the sequence.

This research would also contribute to the field of knowledge about ID, which is of particular importance because malfunctioning IDPs are implicated in many diseases. For instance, p53 is disordered and implicated in cancer; its disordered region allows binding promiscuity in malfunction. Tau is also disordered and implicated in dementia; due to the high exposure of its disordered residues, it easily forms protein aggregates. The disordered nature of pathogenic IDPs make them poor targets for current drug designs, which usually target catalytic functions. If the nature of this disorder were better investigated through this research, better drugs could be developed to target the protein characteristics that have been selected for in pathogenic disordered proteins, such as binding motifs.

This research could be reasonably completed as a thesis project at Stanford University. The resources needed are access to public databases and knowledge of bioinformatic analysis tools. I have been working on similar research for several months and have training in developing algorithms and coding in scripting languages as would be necessary. At Stanford, Dr. Aaron Gitler works on the prion property of proteins, and at our meeting this summer, he expressed interest in using this pipeline to measure selection for prions. Thus, this research could be completed by me at Stanford, where I would present my findings at a molecular evolution conference, make my selection pipeline available for public use, and submit my findings for publication in a journal such as PLOS Computational Biology.

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.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics. 9 (Suppl 2): S1.

Schaefer, C., Schlessinger, A., and Rost, B. (2010). Protein secondary structure appears to be robust under in silico