Establishing contact: detecting tertiary interactions in disordered proteins via coevolving mutations and conformational shifts

Connor Pitman

July 2022

List of Abbreviations

- **BDNF -** Brain-Derived Neurotropic Factor
- **IDP -** Intrinsically Disordered Protein
- **IDR -** Intrinsically Disordered Region
- **pLGIC -** Pentameric Ligand Gated Ion Channel
- **SNP -** Single Nucleotide Polymorphism
- **SNV -** Single Nucleotide Variant

Overview and Objectives

Intrinsically disordered proteins regulate a variety of biological processes, through specific interactions with DNA [1, 2, 3], RNA [1, 2, 3], and other proteins [1, 2, 3, 4, 5]. While IDPs lack stable tertiary structure, sequence-specific tertiary interaction networks have been detected in some IDPs [6] using computationally intensive Molecular Dynamics (MD) simulations. However, it is unknown if these type of interactions are widespread among IDPs. Furthermore, is unfeasible to perform MD on a significant number IDPs. Correlated mutations are one possible indicator of tertiary interactions, and can be feasibly studied using bioinformatics tools.

In most cases, correlated mutations establish contacts between two mutation sites. Conventionally, researchers have focused on contacts between oppositely charged residues [7, 8, 9]. However, this view overlooks specific interactions between neutral residues, which are critical for the complex folds of structured proteins [10]. Mutation of these residues also affects the conformation of IDPs [6], supporting a role for specific tertiary interactions. Examples of underappreciated residue-residue interactions of this nature include methionine-methionine (met-met), cysteine-cysteine (cys-cys), and aromatic-aromatic interactions. Correlated mutations which result in prolonged residue-residue contacts tend to be found in the hydrophobic cores of proteins [9]. Blobulation, an algorithm developed by the Brannigan Lab, groups residues into functional units based on contiguous hydrophobicity and has already proven useful in detecting enrichment in dSNPs [11], as well as coarse-graining IDPs [6]. Using blobulation, we can better identify the local sequence context in the absence of structure, predict whether groups will frequently establish contacts, and more accurately define and predict the effects of compensatory mutations on all proteins. To do this, we will follow these aims:

Aims

- **Aim 1: Test for correlated mutations involving neglected interaction pairs (met-met, cys-cys, aromatic-aromatic) in a previously studied bacterial protein dataset:** We will identify putative correlated mutations using multiple sequence alignment, generate a maximum likelihood phylogenetic tree, and use CoMap to predict coevolving positions. Next, we will refine the list of potential correlated mutations by testing for physical contacts using deposited structures (for structured proteins) or molecular dynamics simulations (for IDPs). Calculate enrichments of specialty hydrophobic (met-met, cys-cys,

aromatic-aromatic) correlated residue-residue interactions found in IDPs vs structured proteins, as well as neglected interaction pairs vs other charge-neutral substitutions in both IDPs and structured proteins. Compute the enrichment of neglected interaction pairs in h-blobs vs p-blobs.

- **Aim 2: Test the hypothesis that adaptive evolution tends to preserve blob topology and blob net charge.** Using the results from Aim 1, calculate the enrichments of mutations which preserve length, total number, and overall net charge of h-blobs vs those which cause changes in the bacterial dataset. For a set of pentameric Ligand Gated Ion Channels from bacteria, archaea, and unicellular eukaryotes, for which phylogeny has already been determined, calculate the enrichments of the same characteristics for proteins within the same branch.
- **Aim 3: Test the hypothesis that compensatory mutations preserve blob topology and blob contacts, in a dataset of vertebrates** Using the same pipeline as Aim 1, identify coevolving positions in an in-house *Anolis* Dataset. Test for enrichment in compensatory mutations in h-blobs vs p-blobs. Compute the enrichments of mutations which preserve length, total number, and overall net charge of h-blobs vs those which cause changes. Calculate the enrichment of correlated mutations which preserve the same characteristics for proteins found in the same branch of generated phylogenetic trees.

Background

Intrinsically Disordered Proteins

For almost a century, the function of a protein was associated with having a stable three-dimensional structure. This dogma is known as the "structure-function" paradigm, which states that "sequence determines structure determines function". This view holds for many proteins. Enzymes, for instance, have active sites that require a specific tertiary structure for substrate binding. However, an entire class of proteins does not have a well-defined tertiary structure and are vital to biological pathways: intrinsically disordered proteins (IDPs) [1]. Additionally, some structured proteins contain functional regions of intrinsic disorder (IDRs).

Intrinsic disorder is a sequence-based phenomenon. It arises from a high fraction of charged residues, coupled with a low fraction of hydrophobic residues [12, 13]. These proteins tend to remain unfolded in an aqueous environment. They are notable for their many functions within cells: from signaling [14, 15, 16], to facilitating intracellular molecular interactions [17], to acting as chaperones [18], to their medical relevance in forming disease-causing aggregates [19, 20]. They exist as a broad class of proteins and range from entirely unfolded to having the ability to fold upon binding with a second protein [21, 22].

Intrinsic disorder can be found in proteomes from bacteria to eukaryotes. Though many attempts to classify how prevalent intrinsic disorder is across different organisms, it is generally predicted that the percentages of amino acids found in intrinsically disordered regions in bacteria are somewhere between ~12% to ~24%. In eukaryotes, the estimates range from ~33% to ~50%. One example of this difference between prokaryotes and eukaryotes is that pLGICs found in eukaryotes typically contain a long intracellar IDRs [23] not found in prokaryotes.

Though IDPs lack tertiary structure, they can still sample a conformational ensemble of metastable states. Residual secondary structure can also be persistent in IDPs [12, 24].

Mutations and IDPs

Like structured proteins, mutations in IDPs can still have functional consequences [25]. Many researchers have investigated function-altering mutations in known IDPs - such as BDNF [26], MeCP2 [27], and p53 [28]. Mutations involve a change in charge are considered most likely to be highly disruptive.

The Brannigan lab has previously investigated the BDNF prodomain, an IDP. When the Val66Met mutation occurs in the protein, it causes it to bind to a new receptor (SorCS2) [29, 30], which produces disease-associated neurodegeneration. What was initially unexpected about this dSNP is that a hydrophobic to hydrophobic mutation can cause a significant shift in function.

Quantifying tertiary contacts in structure-less proteins

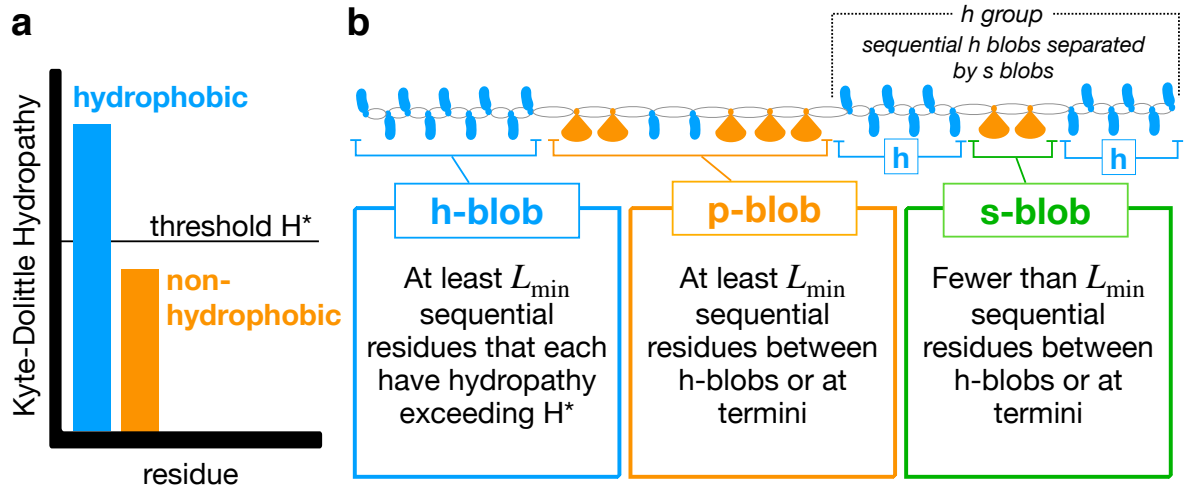

Figure 1: Figure and caption adapted from Lohia (2022) [11]. Whole-sequence blobulation algorithm for segmentation of proteins. (A) First, the sequence is digitized: Residues are classified as hydrophobic or nonhydrophobic depending on whether they have a Kyte–Dolittle [31] hydropathy falling above or below the user-provided threshold $\mathbf{H}^*$, respectively. (B) The clustering step acts on the digitized sequence, which is illustrated here as a cartoon: Residues above and below the $\mathbf{H}^*$ threshold are shown as blue ovals and orange fans, respectively. The clustering step scans the digitized sequence according to the indicated criteria, first detecting h-blobs, then p-blobs, and finally s-blobs. L_{min} is the user-provided minimum blob size. The blobulation outcome for this particular chain would be valid for $2<L_{min}<6$.

Initially, the study considered residue-residue contacts to track the shift in contacts in both the wild-type and mutated protein. However, this produced too many interactions, each of which had subtle changes, to detect statistically significant differences. To understand the origins of the conformational shifts of the BDNF prodomain, it became necessary to coarse-grain the protein into functional regions that do not include secondary structure. Here, contiguous hydrophobicity was utilized. Our lab developed blobulation algorithm. Blobulation not only groups similar residues into functional units (Figure 1) but also shows which groups will likely establish contacts. While having been identified as an individual amino acid property, hydrophobicity had not been considered in the case of contiguous hydrophobic residues forming “blobs”. It was found that the Val66Met mutation changes frequency of blob contacts, increasing between

blobs which contain methionine [6].

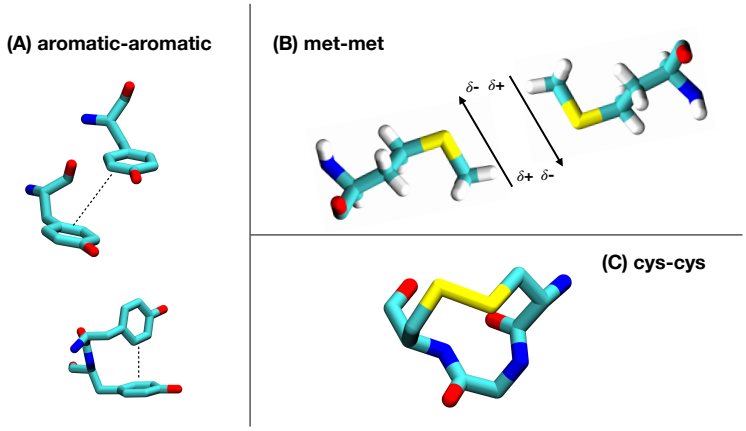

Figure 2: Examples of specialty residue-residue interaction pairs. Aromatic-aromatic (A), met-met (B), and cys-cys (C) example interactions are shown. (A) Tyrosines in two conformations: offset stacking (above), t-stacking (below). (B) Methionine-methionine dipole-dipole interaction occuring due to sulfur polarizability. (C) Two cysteines in a disulfide bond. Panels A and C were generated from a pLGIC (PDB accession: 6v4a)

Methionine

Methionine is one of two sulfur-containing amino acids, though it cannot form disulfide bridges like cysteine. In most contexts, methionine is classified as hydrophobic, but it also contains one polarizable sulfur. It has also recently been established that two methionines can undergo favorable interactions on the scale of -3.0 to -3.5 kcal/mol [32]. As two methionine residues approach each other, their dipoles cause the residues to interact such that the slightly negative sulfur on one residue comes into contact with the adjacent slightly positive methyl group (Figure 2). Methionine has been shown to play a significant role in IDPs [33].

Blobulation and dSNPs

Blobulation has also proven helpful in other contexts beyond IDP dynamics. Recently, a study from the Brannigan lab found that dSNPs are significantly enriched in certain blobs along a continuum: as blobs get longer and more hydrophobic, they tend to have a greater enrichment in dSNPs [11]. Additionally, mutations that cause changes to blobs, such as breaking or dissolving h blobs, are more likely to be dSNPs.

Correlated mutations

Not all mutations are deleterious. Sometimes they occur within the genome and allow an organism to adapt to its environment. Additionally, once a dSNP has occurred, there are many

ways nature can restore original functions to an organism. For instance, a second mutation can change the amino acid back to its original form. However, there are cases where a second mutation at an entirely different position compensates for the original deleterious one. This is referred to as a compensatory mutation. It has been demonstrated that compensatory mutations do not occur randomly and tend to be near the site of the original mutation [34]. It has also been shown that contact between mutated positions is essential in most cases [35].

At present, compensatory mutations have not been studied in IDPs specifically, as many studies use solved structures [9] or predicted secondary structure [36] to demonstrate intraprotein contacts between residues.

Our goal is to determine if tertiary contacts are persistant throughout the Intrinsically Disordered proteome. We seek to quantify the importance of these contacts through correlated mutations and conserved blob topology in bacteria, pLGICs, and *Anolis*.

Preliminary Work

Beta Amyloid

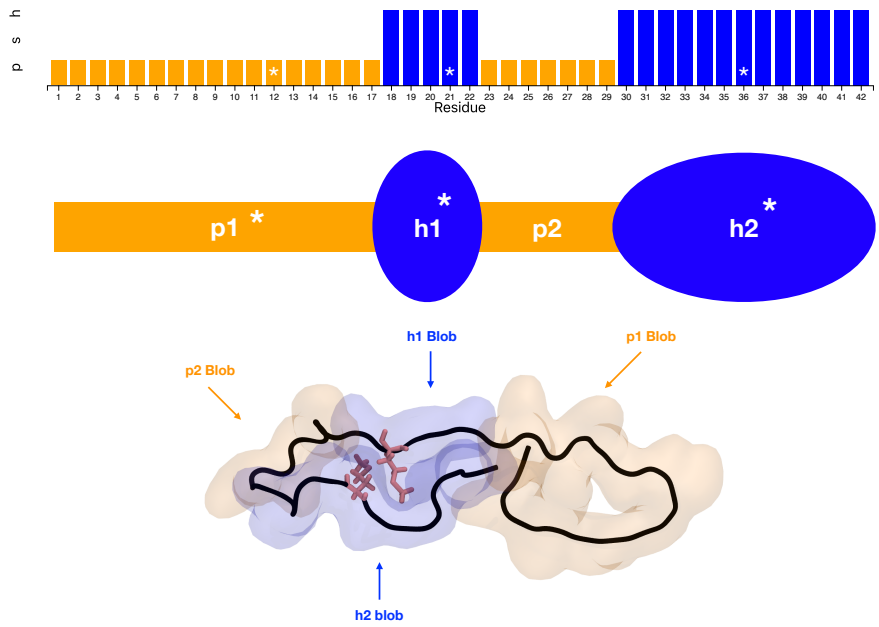

Figure 3: Beta amyloid protein after blobulation (top). "h-blobs" are shown in blue, while "p-blobs" are shown in orange. These blobs are also shown in a cartoon (middle), which is colored by blob type. Asterisks represent methionines in both the residue and blob panels. A21M mutant beta amyloid representation colored by blob in VMD (bottom). The backbone is shown in black, methionine in pink, and blob colored by type.

Beta amyloid peptides aggregate to form plaques in patients with Alzheimer's Disease [19] (reviewed in [37, 38]). Beta Amyloid is a 42 amino acid long IDP containing one native methionine. For this peptide, we hypothesized that introducing a second methionine to the sequence through mutation would cause an increased frquency in contacts between blobs containing methionines. To test this, we performed molecular dynamics simulations of the wild type peptide as well as two mutants which each contained an additional methionine at either the 12th (V12M) or 21st (A21M) positions.

We started by blobulating Beta Amyloid to classify functional segments (Figure 3). Beta Amyloid contains four blobs - termed h1, p1, h2, and p2. Then, we measured the frequency of contacts between blobs in the wildtype simulations (Figure 4). We found that h1 and h2 blobs have a contact frequency of 12% , and the p2 and h2 blobs have a contact frequency of 18% (Figure 4, left). Hydrophobic regions (h-blobs) and regions close in sequence space are expected to establish frequent contacts.

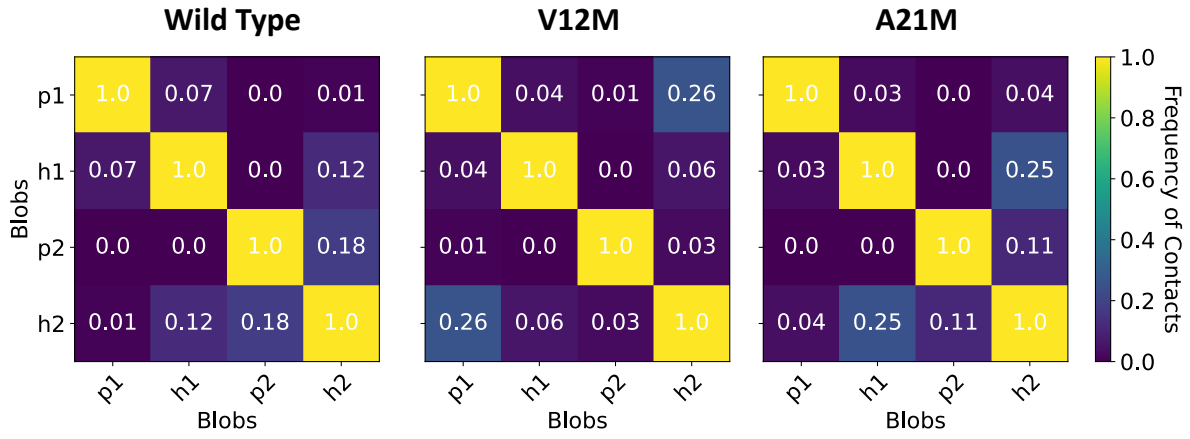

Figure 4: The frequency of contacts from MD simulations in the V12M and A21M mutants, as well as the wildtype peptide, between the defined blobs in the beta amyloid protein. Methionine is found in h2 in all proteins, and an additional methionine is found in p1 (V12M, middle) or h1 (A21M, right) in the mutants. The frequency of contacts is represented as a gradient from 0 (blue) to 1 (yellow).

To test whether the addition of a second methoinine changes the tertiary contact network of Beta Amyloid, we also measured the frequency of contacts in the mutant simulations. For both the V12M mutant (Figure 4, middle) and the A21M mutant (Figure 4, right), the frequency of contacts in blobs which contain methionine increase compared to the wildtype (a 25-fold and 13-fold increase, respectively). To quantify whether the contact shifts were influenced by met-met interactions, we measured the orientations of the methionines relative the established angles [32].

Figure 3 shows the frequencies of calculated P angles as the methionines get closer together for each mutant. Both show a peak around 90° (which is consistent with a favorable met-met interaction), and continues trending toward 90° as the residues get closer together.

This θ angle is the second metric we used to evaluate for met-met interactions. We calculated the frequencies of θ for both the V12M and A21M replicas. We see a minor peak around 30°, as well as 150° (Figure 6), and the methionines trend toward these expected values they get closer.

The influence of local sequence context on met-met interactions

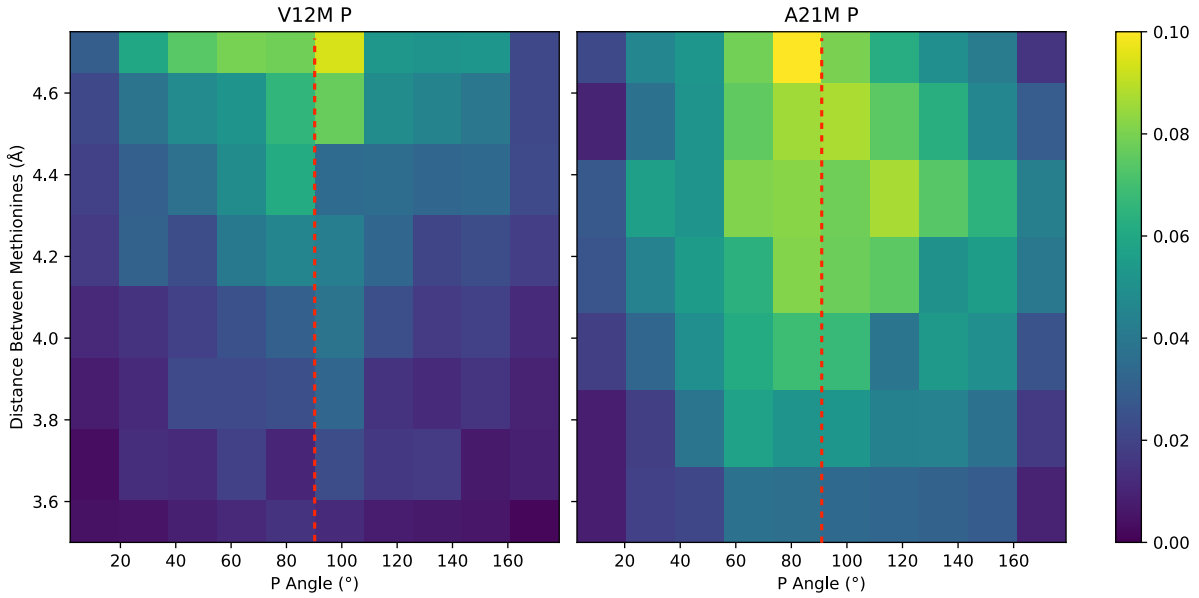

Figure 5: Frequency of P angle calculations binned by angle (x axis) and distance between methionines (y axis) across all V12M and A21M replicas. Vertical dashed lines represent the expected angles as defined by Gomez-Tamayo et al [32].

We also tested whether local sequence context and distance between methionines affected the frequency of met-met interactions. We hypothesized that less met-met interactions would occur if more residues separate the pair of methoinines. Additionally, we hypothesized that sequence content would have a stronger influence on met-met contacts at longer distances, with less met-met contacts occurring in a polar peptide.

To test this hypothesis, we simulated met-polya and met-polyk peptides (composed of 12 alanine or lysine and 2 methionines, respectively), as well as control polyalanine and polylysine. We varied the number of residues between the two methionines (2, 4, 6, and 8). Polyalanine is considered an h-blob, and polylysine is considered a p-blob.

After simulating the peptides for at least 2μs each and removing the equilibration portions of the trajectories, we measured the radius of gyration to quantify differences in overall shape between peptides (shown in Figure 7). For the met-polylysine peptide, the more space between the methionines, the smaller the radius of gyration tends to be. The same trend holds for the met-polyalanine simulations.

Finally, we tested for met-met interactions in both sets of met-polya and met-polyk peptides (Figures 8 and 9, respectively). For peptides where methionine residues are close together,

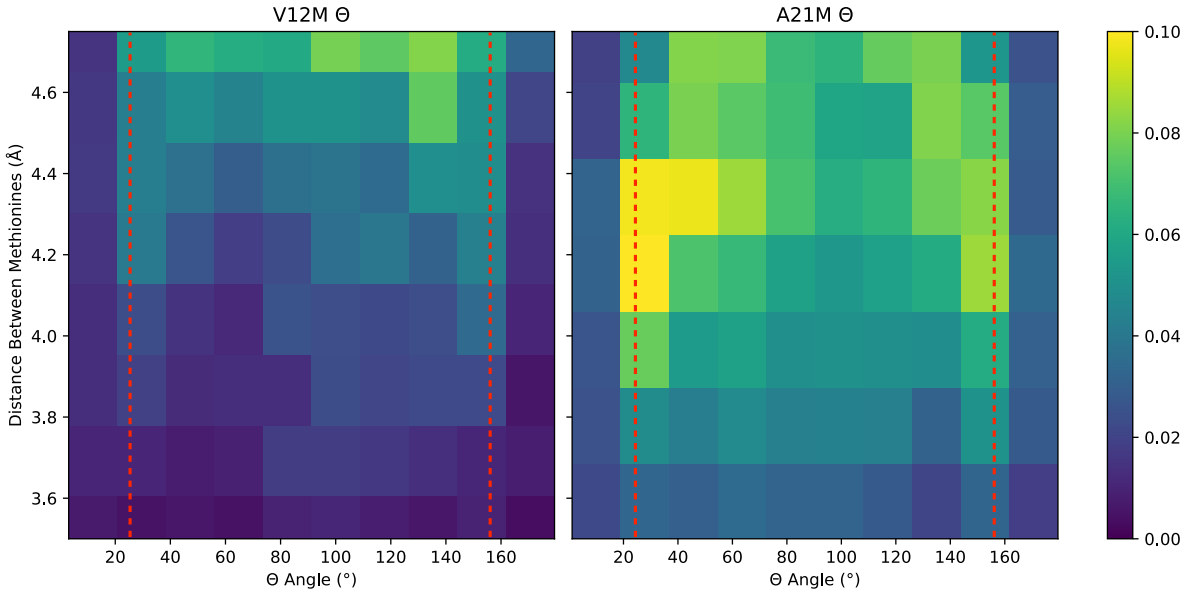

Figure 6: Frequency of θ angle calculations binned by angle (x axis) and distance between methionines (y axis) across all V12M and A21M replicas. Vertical dashed lines represent the expected angles as defined by Gomez-Tamayo et al [32].

the methoinines approach orientations consistent with met-met interactions. As the number of residues between methionines increases, these contacts tend to diminish, especially in the met-polyk peptides (met-met interactions occurred once in met-polyk if methionines are separated by 6 residues, and did not occur when they were separated by 8). These results indicate that both blob-type and sequence separation has an influence on met-met interactions.

In summary, we have shown that the addition of one methoinine to Beta Amyloid changes the protein's blob interaction network via met-met interactions. Additionally, we have quantified the importance of blob-type as well as placement in sequence for met-met interactions. This demonstrates that specific interactions between neutral residues can induce conformational shifts in IDPs, and blob properties can influence whether or not these interactions occur.

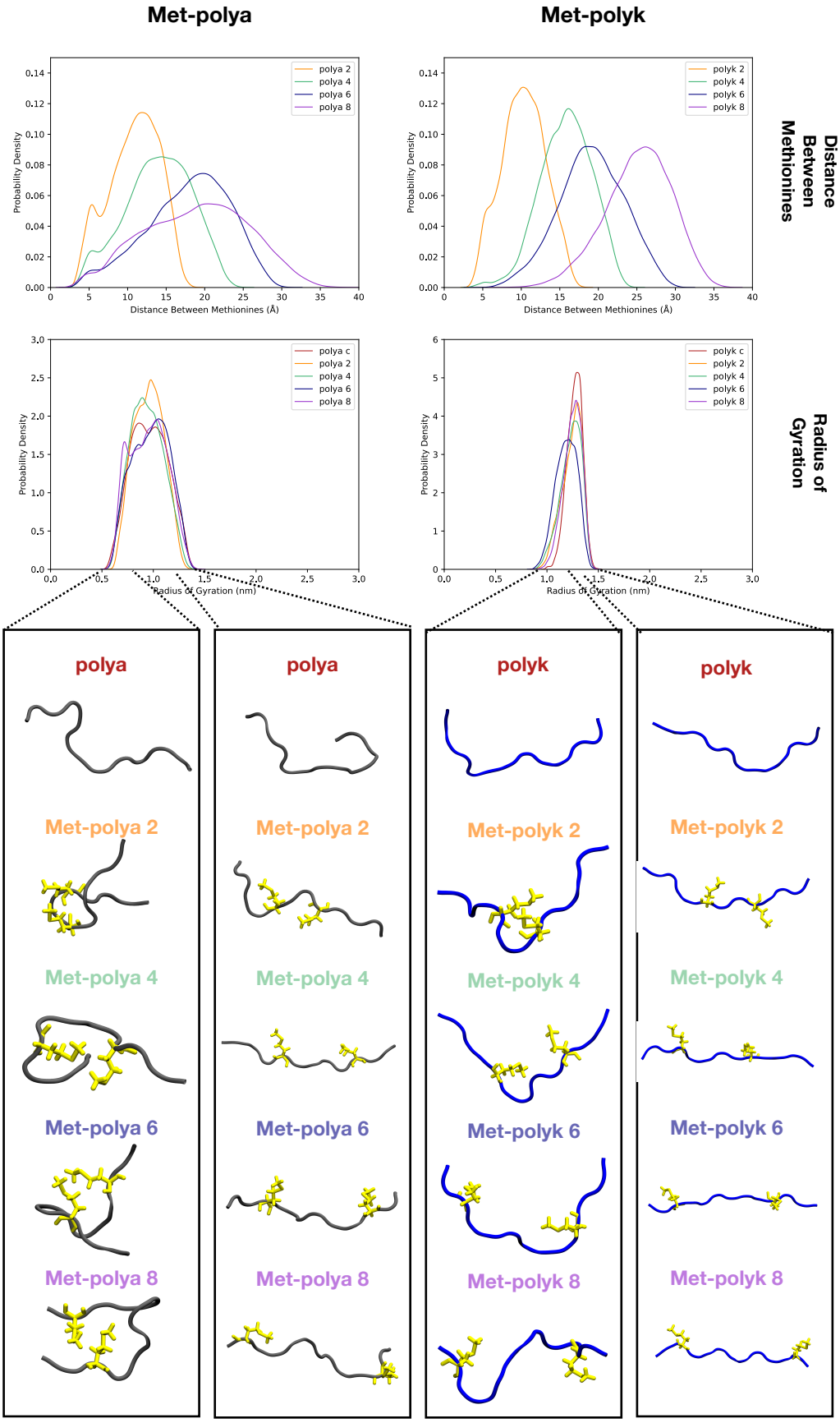

Figure 7: Met-polya and met-polyk peptide conformational trends. The probability distributions of distance between methoinines (top), and radii of gyration (middle) are shown in four panels for all peptides. A selection of images of peptides with high or low radii of gyration is also shown for all met-polya and met-polyk peptides, as well as the control peptides (red) (bottom).

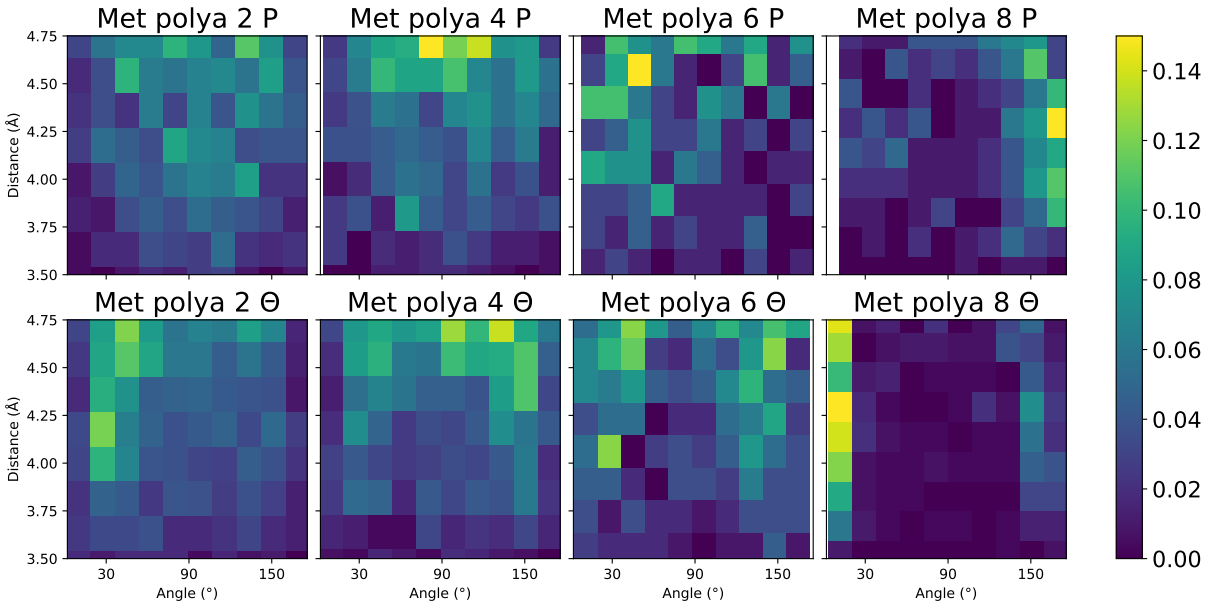

Figure 8: Frequency of P (top row) and θ (bottom row) angle calculations binned by angle (x axis) and distance between methionines (y axis) across all polyalanine peptides. From left to right, the number of residues between methionines (starting at 2) increases by 2.

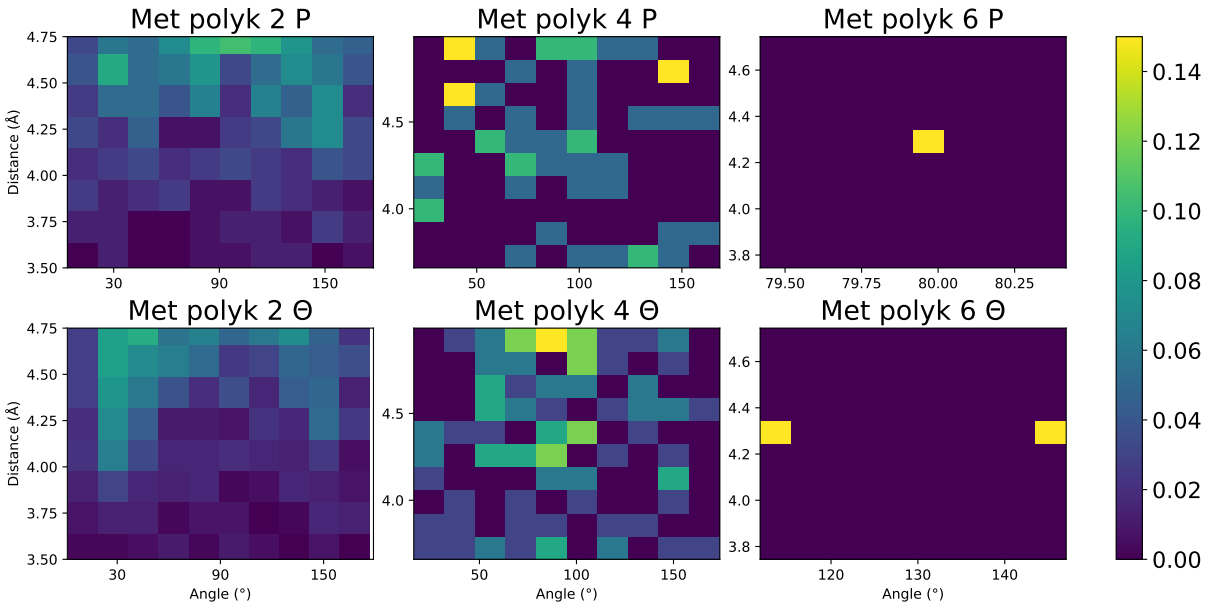

Figure 9: Frequency of P (top row) and θ (bottom row) angle calculations binned by angle (x axis) and distance between methionines (y axis) across all polyalanine peptides. From left to right, the number of residues between methionines (starting at 2) increases by 2. Note that Met polyk 8 is not shown because the methionines never come within 5Å of each other in the equilibrated simulation, and are therefore not considered to come into contact.

Research Plan

Aim 1: Test for correlated mutations involving neglected interaction pairs (met-met, cys-cys, aromatic-aromatic) in a bacterial protein dataset.

Introduction

Correlated mutations have thus far been considered mainly in proteins for which well-defined structures have been determined. Our *objective* in this Aim is to test the hypothesis that correlated mutations are found in contiguously hydrophobic regions of proteins, and specialty residue interactions play an essential role in some of these mutations (particularly in IDPs). Our *approach* will be using an established bioinformatics pipeline, an algorithm which groups residues into functional units, and utilizing molecular dynamics simulations to detect residue-residue interactions in proteins which lack a solved structure. Our *rationale* is that these interaction pairs have already been shown to be influential in determining the conformational ensemble of IDPs in previous research [6], and in our preliminary data in h-blobs. Our expected *outcome* that these neglected interaction pairs will be enriched in h-blobs (particularly in IDPs).

Research design

It has been a challenge in recent decades to determine with a high degree of certainty if two residue positions are truly correlated. To this end, we will utilize a recent pipeline created by Chaurisa et al. [9] to align sequences collected from the HOGENOM [39] database. The identification of correlated mutations requires two things: aligned sequences and predicted phylogenetic trees. However, to generate high quality sequence alignments, high quality predicted phylogenetic trees are necessary. The reverse is also true, to maximize the quality of predicted phylogenetic trees, high quality sequence alignments are necessary. As a solution to this problem, we will do an initial alignment using Clustal Omega [40, 41], and follow this with an initial phylogenetic tree construction using FastTree [42]. Both methods are relatively computationally inexpensive, and we will use the results to filter out low-confidence sequences. Then, with the filtered dataset, we will do a high-quality alignment using Muscle [43] and Clustal Omega. To generate a consensus alignment from these two outputs, we will use a sum-of-pairs score for each alignment site, masking any site deemed inefficient by the bppAInScore from Bio++. Then we will generate a high-quality phylogenetic tree using bppPhySamp from Bio++ [44]. Both the aligned sequences and phylogenetic trees will be used to predict coevolving residue

positions using CoMap [45]. Then, we will blobulate each of the protein sequences which contain predicted coevolving positions. Additionally, for protiens which have a solved structure deposited in the Protein Databank, we will use the distance between correlated positions to determine if they are interacting. For a selection of proteins which do not have a solved structure, we will perform molecular dynamics simulations to determine if sets of coevolving residues are indeed interacting. To achieve this, we propose GROMACS T-REMD [46, 47] simulations carried out in explicit TIP4P-D water [48] using the a99sb*-ildn-q force field [49], using GROMACS 5.1.2. To determine which IDPs are intrinsically disordered, we will use PONDR [50] and Espritz [51]. Any protein region identified by either will be considered as disordered, and proteins with > 50% disordered region will be considered IDPs, those with less will be considered proteins with IDRs. Finally, we will calculate the enrichments of neglected types of residue-residue interactions in IDPs vs structured proteins, as well as neglected interaction pairs vs other charge-neutral substitutions in both IDPs vs structured proteins and h-blobs vs p-blobs.

Expected outcomes

We will produce a set of predicted correlated mutations which includes IDPs, which we will utilize in Aim 2. Additionally, we will produce a list of distances between residue positions from these contacts using the protein databank. We will also generate a library of simulation data for a subset of IDPs, which will then be used to calculate their radii of gyration and generate contact probabilities and a list of geometric measurements for detecting residue-residue interactions. We hope to investigate tertiary contacts in these bacterial proteins, particularly those without well-defined structures, which will give further insight into tertiary contacts within IDPs and IDRs.

Potential problems and alternative approaches

One potential challenge we may face in Aim 1 is the alignment of IDPs, which is sometimes difficult due to their low sequence conservation. If the alignment methods in our pipeline are insufficient, we can use KMAD (a tool to aid in IDP alignment) [52], as well as blobulation to generate consensus alignments. Additionally, achieving convergence on the proteins we hope to simulate may exceed our computational resources. In this case, we could use implicit solvent to simulate for faster and longer.

Aim 2: Test the hypothesis that adaptive evolution tends to preserve blob topology and blob net charge.

Introduction

While recent studies have identified that compensatory mutations and other evolutionary mechanisms tend to preserve the overall structural motifs of proteins, we hypothesize that the underlying foundations of these motifs is contiguous hydrophobicity and mean net charge. To this end, our *objective* is to study evolutionary mechanisms that cause changes in two protein datasets: the correlated mutations in prokaryotes from Aim 1, and pentameric ligand gated ion channels in a second dataset. For the pLGICs, we will be able to additionally investigate intrinsic disorder in the context of these proteins, as many eukaryotes contain intracellular IDRs where prokaryotes do not. Our *approach* will be utilizing the blobulation algorithm to determine the blobs for all proteins in each dataset, and then calculating the enrichments of mutations or changes to blobs in proteins within the same phylogenetic branch which preserve length, total number, and overall net charge of h-blobs vs those that cause changes to these attributes. Our *rationale* is that it is ultimately hydrophobicity and blob net charge which determines protein contacts, which in turn inform structure. The natural end to this reasoning is that disruptions to these basic properties have the potential to disrupt a protein's overall structure. Our expected *outcome* is that we will find that these properties are largely conserved by evolutionary mechanisms, and they will tend to be preserved at the coarse-grained level.

Research design

Even when a protein's structure is known, it is challenging for researchers to fully understand the conformational shifts induced by sequence changes in proteins, due to the sheer amount of data generated at the residue level. To reduce the potential number of, we will utilize blobulation as a coarse-graining method for these proteins. Then, using our correlated mutation dataset, we seek to quantify changes to blobs within the prokaryotic proteins. Because our goal with the pLGIC dataset in particular is to consider the impact of evolutionary mechanisms on functional segments (blobs) throughout the tree of life, we will organize the aligned protein sequences by their branch on their phylogenetic tree. Within both of the sets of prokaryotic correlated mutations and eukaryotic pLGICs on the same phylogenetic branch, we aim to calculate the enrichment of change to blob length, total number, and net charge of proteins in each dataset. We will test for enrichment in each data subset for the aforementioned properties.

Expected outcomes

We plan to translate our initial set of alignments from Aim 1 into a library of blobulated proteins, which we will use to measure conservation at the tertiary level in both the bacteria and pLGIC proteins. We will calculate the enrichments of various properties across both structured and unstructured proteins and mechanisms of evolution, ultimately producing a list of enrichments. By the end of this Aim, we hope to quantify how conserved blob properties tend to be by evolutionary mechanisms, particularly in IDPs and IDRs, and ultimately gain insight into the units which drive tertiary contacts.

Potential problems and alternative approaches

A major potential challenge in Aim 2 is alignment of sequences from the pLGIC dataset. While we have tools to aid in this, (McBASC [53], GRAPES [54], and CAPS [55]), we can also utilize blobulation. Should proteins differ in number and placement of blobs, we can use other properties of the proteins to aid in alignment, such as patterns in the overall net charge and predicted disordered regions of the proteins. Both of these properties are outputted by the blobulator.

Aim 3: Test the hypothesis that compensatory mutations preserve blob topology and blob contacts, in a dataset of vertebrates

Introduction So far, we have considered correlated mutations in prokaryotes. Our next *objective* would be to research correlated mutations in a eukaryotic dataset. To do this, we will collaborate with the Geneva Lab at CCIB to study correlated mutations and their effects on structure, and elucidate the effects of neglected interaction pairs. Our *approach* will be to begin by identifying coevolving residue positions in proteins using the same pipeline referenced in Aim 1, blobulating each protein using the blobulation algorithm, and utilizing either deposited PDB structures or MD simulation to determine if positions are interacting. Our *rationale* is that the same principles from Aims 1 and 2 will still apply to the *Anolis* dataset. Our expected *outcome* is that we will find that correlated mutations in this dataset will follow the same patterns we hope to find in Aims 1 and 2, and demonstrate that they hold in vertebrates as well as prokaryotes.

Research design

We will use aligned genomes for ~400 species of *Anolis*. Though it has been a challenge to determine if two residue positions are truly correlated, we hope to utilize the same bioinformatics pipeline followed in Aim 1. This will involve generating high quality alignments and

phylogenetic trees, with the goal of detecting predicted correlated residues. Once we have the alignments and trees, we will perform the same set of analyses followed in Aims 1 and 2, calculating the same enrichments in this novel dataset. To determine which IDPs are intrinsically disordered, we will use PONDR [50] and Espritz [51]. Any protein region identified by either will be considered as disordered, and proteins with > 50% disordered region will be considered IDPs, those with less will be considered proteins with IDRs.

Expected outcomes

We expect to produce a list of potential correlated mutations and blobulated proteins. We will also generate a list of categorized proteins based on whether they are predicted to be IDPs, structured proteins with IDRs, or structured proteins. We plan to compute enrichments for both types of correlated mutations and tertiary conservation for the identified protein types. We expect to have confirmed or refuted the hypothesis that tertiary interactions are prevalent in *Anolis* IDPs, which will hopefully precede further investigation into tertiary contact networks in IDPs in general.

Potential problems and alternative approaches

Due to the challengingnature of curating a dataset of correlated mutations ab initio, we expect that our greatest potential challenge will be generating high confidence alignments and phylogenetic trees. If needed, we may augment our pipeline with additional analyses. If we have difficulty aligning IDPs, we may use KMAD [52]. If we can use an additional phylogenetic tree, such as RAxML [56] find a consensus. To refine correlated residue positions, we can use McBASC [53], GRAPES [54], CAPS [55], to filter out false positives by only considering pairs identified by at least two softwares. Additionally, we can use blobulation to group correlated mutation pairs by they occur in h-blobs, p-blobs, or one in each.

Bibliography

[1] P.E. et al Wright. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol., 1999.

[2] A. Keith Dunker, Christopher J. Oldfield, Jingwei Meng, Pedro Romero, Jack Y Yang, Jessica Walton Chen, Vladimir Vacic, Zoran Obradovic, and Vladimir N Uversky. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 2008.

[3] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, apr 2014.

[4] A. Keith Dunker, Celeste J. Brown, J. David Lawson, Lilia M. Iakoucheva, and Zoran Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 2002.

[5] Vladimir N. Uversky. Multitude of binding modes attainable by intrinsically disordered-proteins: a portrait gallery of disorder-based complexes. *Chemical Society Reviews*, 2011.

[6] Ruchi Lohia, Reza Salari, and Grace Brannigan. Sequence specificity despite intrinsic disorder: How a disease-associated val/met polymorphism rearranges tertiary interactions in a long disordered protein. *PLOS Computational Biology*, 15(10):e1007390, oct 2019.

[7] David D. Pollock, William R. Taylor, and Nick Goldman. Coevolving protein residues: Maximum likelihood identification and relationship to structure. Journal of Molecular Biology, 1999.

[8] Pierre Tuffery and Pierre Darlu. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Molecular Biology and Evolution, 2000.

[9] Shilpi Chaurasia and Julien Y. Dutheil. The structural determinants of intra-protein compensatory substitutions. Molecular Biology and Evolution, March 2022.

[10] Amit Kessel and Nir Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. CRC Press, 2018.

[11] Ruchi Lohia, Grace Brannigan, and Matthew Hansen. Contiguously hydrophobic sequences are functionally significant throughout the human exome, 2022.

[12] Vladimir Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, 2002.

[13] Rahul K. Das and Rohit V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions ofoppositely charged residues. PNAS, 2013.

[14] Vladimir N. Uversky. Multitude of binding modes attainable by intrinsically disordered-proteins: a portrait gallery of disorder-based complexes. *Chemical Society Reviews*, 2010.

[15] Vladimir N. Uversky. Intrinsic disorder-based protein interactions and their modulators. *Current Pharmaceutical Design*, 2013.

[16] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation.

[17] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews*, 2014.

[18] Peter Tompa and Peter Csermely. The role of structural disorder in the function of RNA and protein chaperones. *The FASEB Journal*, 18(11):1169–1175, aug 2004.

[19] George G. Glenner and Caine W. Wong. Alzheimer's disease: Initial report of the purification and charactization characterization of a novel cerebovascular amyloid protein. *Biochemical and Biophysical Research Communications*, 1984.

[20] Zhe-Yu Chen, Kevin Bath, Bruce McEwen, Barbara Hempstead, and Francis Lee. Nih public access author manuscript novartis found symp. author manuscript; available in pmc 2009 september. *Novartis Found Symp.*, 2008.

[21] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. Structural Biology, 2002.

[22] Steffen P. Graether. Disorder and function: a review of the dehydrin protein family. Frontiers in Plant Science, 2014.

[23] Rebecca J. Howard. Elephants in the dark: Insights and incongruities in pentameric ligand-gated ion channel models. *Journal of Molecular Biology*, 433(17):167128, aug 2021.

[24] Orkid Coskuner-Weber and Vladimir N. Uversky. Insights into the molecular mechanisms of alzheimer's and parkinson's diseases with molecular simulations: Understanding the roles of artificial and pathological missense mutations in intrinsically disordered proteins related to pathology. International Journal of Molecular Sciences, 2018.

[25] Vladimir Vacic, Phineus R. L. Markwick, Christopher J. Oldfield, Xiaoyue Zhao, Chad Haynes, Vladimir N. Uversky, and Lilia M. Iakoucheva. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Computational Biology*, 8(10):e1002709, oct 2012.

[26] Fatima Soliman, Charles E. Glatt, Kevin G. Bath, Liat Levita, Rebecca M. Jones, Siobhan S. Pattwell, Deqiang Jing, Nim Tottenham, Dima Amso, Leah H. Somerville, Henning U. Voss, Gary Glover, Douglas J. Ballon, Conor Liston, Theresa Teslovich, Tracey Van Kempen, Francis S. Lee, and B. J. Casey. A genetic variant BDNF polymorphism alters extinction learning in both mouse and human. *Science*, 327(5967):863–866, feb 2010.

[27] Juan AusiÃ³. Mecp2: The genetic driver of rett syndrome epigenetics, January 2021.

[28] Claus Preudhomme, Pierre Fenaux, Service des Maladies du Sang, and C.H.U. Lille. The clinical significance of mutations of the p52 tumour suppressor gene in haematological malignancies. British Journal of Haematology, 1997.

[29] Agustin Anastasia, Katrin Deinhardt, Moses V. Chao, Nathan E. Will, Krithi Irmady, Francis S. Lee, Barbara L. Hempstead, and Clay Bracken. Val66met polymorphism of BDNF alters prodomain structure to induce neuronal growth cone retraction. *Nature Communications*, 4(1), sep 2013.

[30] Joanna I. Giza, Jihye Kim, Heidi C. Meyer, Agustin Anastasia, Iva Dincheva, Crystal I. Zheng, Katherine Lopez, Henrietta Bains, Jianmin Yang, Clay Bracken, Conor Liston, Deqiang Jing, Barbara L. Hempstead, and Francis S. Lee. The BDNF val66met prodomain disassembles dendritic spines altering fear extinction circuitry and behavior. *Neuron*, 99(1):163–178.e6, jul 2018.

[31] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 1982.

[32] José C. Gómez-Tamayo, Arnau Cordomí, Mireia Olivella, Eduardo Mayol, Daniel Fourmy, and Leonardo Pardo. Analysis of the interactions of sulfur-containing amino acids in membrane proteins. *Protein Science*, 25(8):1517–1524, jun 2016.

[33] Jung Mi Lim, Geumsoo Kim, and Rodney L. Levine. Methionine in proteins: It's not just for protein initiation anymore. *Neurochemical Research*, 44(1):247–257, jan 2018.

[34] Brad H. Davis, Art F. Y. Poon, and Michael C. Whitlock. Compensatory mutations are repeatable and clustered within proteins. Proceedings of the Royal Society, 2009.

[35] Amrita Bhattacherjee, Saurav Mallik, and Sudip Kundu. Compensatory mutations occur within the electrostatic interaction range of deleterious mutations in protein structure. *Journal of Molecular Evolution*, 80(1):10–12, nov 2014.

[36] Vladimir Vacic, Phineus R. L. Markwick, Christopher J. Oldfield, Xiaoyue Zhao, Chad Haynes, Vladimir N. Uversky, and Lilia M. Iakoucheva. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. PLOS Computational Biology, 2012.

[37] John Hardy and Dennis J. Selkoe. The amyloid hypothesis of alzheimer's disease:progress and problems on the road to therapeutics. *Science's Compass*, 2002.

[38] Dennis J. Selkoe. Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews*, 2001.

[39] Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, and Perriere G. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 2009.

[40] Dineen DG Gibson TJ Karplus K Li W Lopez R McWilliam H Remmert M Söding J Thompson JD Higgins D Sievers F, Wilm A. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 2011.

[41] Li W Valentin F Squizzato S Paern J Lopez R Goujon M, McWilliam H. A new bioinformatics analysis tools framework at embl-ebi. *Nucleic acids research*, 2010.

[42] M.N. Price, Dehal P.S., and A.P Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 2010.

[43] Robert C. Edgar. Muscle: multiple sequence alignment with highaccuracy and high throughput. Nucleic Acids Research, 2004.

[44] Laurent Gueguen, Sylvain Gaillard, Bastien Boussau, Manolo Gouy, Mathieu Groussin, Nicolas C. Rochette, Thomas Bigot, David Fournier, Fanny Pouyet, Vincent Cahais, Aure ́lien Bernard, Ce ́line Scornavacca, Beno ̂ıt Nabholz, Annabelle Haudry, Lo ıc Dachary, Nicolas Galtier, Khalid Belkhir, and Julien Y. Dutheil. Bio++: Efficient extensible libraries and tools forcomputational molecular evolution. Molecular Biology and Evolution, 2013.

[45] Hobolth A. Tataru, P. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time markov chains. *BMC Bioinformatics*, 2011.

[46] H. Bekker, H.J.C. Berendsen, E.J. Dijkstra, S. Achterop, R. Van Drunen, D. van derSpoel, A. Sijbers, H. Keegstra, and et. al. Gromacs: A and parallel computer for molecular dynamics simulations. *Physics Computing*, 1993.

[47] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 1999.

[48] Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The Journal of Physical Chemistry B*, 119(16):5113–5123, apr 2015.

[49] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, mar 2010.

[50] Kissinger C Villafranca JE Dunker AK Romero P, Obradovic Z. Identifying disordered regions in proteins from amino acid sequence. *Proc IEEE Int Conf Neural Networks*, 1997.

[51] Di Domenico T Tosatto SC. Walsh I, Martin AJ. Espritz: accurate and fast prediction of protein disorder. *Bioinformatics*, 2012.

[52] Joanna Lange, Lucjan S. Wyrwicz, and Gert Vriend. Kmad: knowledge-based multiple sequence alignment for intrinsically disordered proteins. *Bioinformatics*, 2016.

[53] Anthony A. Fodor and Richard W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. PROTEINS: Structure, Function, and Bioinformatics, 2004.

[54] Nicolas Galtier. Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics*, 12(1):e1005774, jan 2016.

[55] Mario A Fares and Simon A A Travers. A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. *Genetics*, 173(1):9–23, may 2006.

[56] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 2014.