 doi:10.20944/preprints202209.0450.v1

Article

Intrinsic Disorder as a Natural Preservative: High Levels of Intrinsic Disorder in Proteins Found in the 2,600-Year-Old Human Brain

Aaron Mohammed [1] and Vladimir N. Uversky [1,2,*]

[1] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA; amohammed2@usf.edu (A.M.); vuversky@usf.edu (V.N.U.)
[2] USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
* Correspondence: Department of Molecular Medicine, University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, Florida 33612, USA, Phone: 1-813-974-5816; Fax: 1-813-974-7357; E-mail: vuversky@usf.edu

Simple Summary: Reported earlier, results of the proteomic analysis of the "Heslington brain" (which is at least 2,600-year-old brain tissue uncovered within the skull excavated in 2008 from a pit in Heslington, Yorkshire, England) revealed the preservation of many proteins. Five of these proteins, heavy, medium, and light neurofilament proteins, glial fibrillary acidic protein, and myelin basic protein are engaged in the formation non-amyloid protein aggregates, such as intermediate filaments and myelin sheath. Our analysis reported in this study revealed that these five proteins, their interactors, and many other proteins found in the Heslington brain are characterized by high level of intrinsic disorder, suggesting that intrinsic disorder might play a role in preserving brain tissue likely acting as a molecular as a molecular mortar or cement that glue together various brain proteins and rigidify the resulting assemblies thereby generating highly stable matter.

Abstract: Proteomic analysis revealed the preservation of many proteins in the "Heslington brain" (which is at least 2,600-year-old brain tissue uncovered within the skull excavated in 2008 from a pit in Heslington, Yorkshire, England). Five of these proteins ("main proteins"), heavy, medium, and light neurofilament proteins (NFH, NFM, and NFL), glial fibrillary acidic protein (GFAP), and myelin basic (MBP) protein are engaged in the formation non-amyloid protein aggregates, such as intermediate filaments and myelin sheath. We used a wide spectrum of bioinformatics tools to evaluate the prevalence of functional disorder in several related sets of proteins, such as "main proteins" and their 44 interactors, as well as all other protein identified in the Heslington brain. These analyses revealed that all five "main proteins", half of their interactors and almost one third of the Heslington brain proteins are expected to be mostly disordered. Furthermore, most of the remaining proteins are expected to contain sizable disordered regions. This is in contrary the expected substantial (if not complete) elimination of the disordered proteins from the Heslington brain. Therefore, it seems that the intrinsic disorder of NFH, NFM, NFL, GFAP, and MBP, their interactors and many other proteins might play a crucial role in preserving the Heslington brain by forming tightly folded brain protein aggregates, in which different parts are glued together via the disorder-to-order transitions.

Keywords: Heslington brain; intrinsically disordered protein; intrinsically disordered region; binding-induced folding; disorder-to-order transition

1. Introduction

In 2008, archaeological work in advance of construction at a site on the eastern edge of the Heslington village located about 1.9 miles south-east of York city center (Yorkshire, England) uncovered a skull buried in a pit. This skull belonged to an Iron Age man aged

 doi:10.20944/preprints202209.0450.v1

between 26–45 years old at the time of death, likely in his mid-30s. Radiocarbon dating found that the man had died sometime between 673-482 BC [1]. The most remarkable side of this discovery was the fact that several large fragments of brain were present inside the skull. Since this Heslington brain (which is at least 2,600-year-old) preserved many anatomical features despite being shrunk to about 20% of its original size [1], it opened a unique possibility to investigate the preservation of human brain proteins, and results of this analysis were recently reported [2]. Utilization of a broad spectrum of molecular techniques, ranging from careful exclusion of sample contamination to immunoelectron microscopy, to mass-spectroscopy, to quantification of brain-specific proteins by ELISA, and to gel electrophoresis and immunoblotting demonstrated the preservation of neurocytoarchitecture in the ancient brain and also revealed the exceptional preservation of some human brain proteins in the analyzed samples of the Heslington brain [2]. Among identified proteins with this extraordinary long-term stability and ability to survive for 2,600 years were proteins engaged in the formation non-amyloid protein aggregates, such as neurofilament proteins (NFs), glial fibrillary acidic protein (GFAP), and myelin basic protein (MBP) [2]. Based on these observations, the authors concluded that the preservation of brain proteins for millennia was driven by the formation of protein aggregates [2]. Curiously, our analysis reported in this study revealed that all the major proteins found in the Heslington brain are characterized by high level of intrinsic disorder.

The analysis of the micro-structure of the Heslington brain first began with immunoelectron microscopy in order to determine if there was a presence of specific axonal proteins. Neurofilament heavy (NFH), an NF subunit, was found to be present in the axons of the ancient tissue. The detection of an NF subunit indicated that other NF subunits, neurofilament middle (NFM) and neurofilament light (NFL), should be present as well. NFs are type IV intermediate filaments (IFs), which are cytoskeletal fibers found in eukaryotic cells. Neurofilaments are necessary for the radial growth of axons [3]. Next, sensitive immunoassays were carried out to screen for other brain proteins that may have been present. There were strong immune responses that indicated the presence of GFAP and MBP and weak immune responses that indicated the presence of NFs. GFAP is a type III IF that provides structural stability to astrocytic processes which results in the modulation of astrocyte motility and shape [4]. MBP is part of the myelin sheath surrounding nerve axons and is the second most abundant protein in the central nervous system [5]. Mass spectroscopy was then used to get idea of what other proteins may be present in the preserved brain. Peptide fragments were detected that matched 855 proteins, after accounting for possible contaminants. GFAP, MBP, and NFL, were matches for fragments that were detected, however, some of those fragments were also matches for other proteins. Only GFAP and NFL had a fragment was unique to them.

It was concluded by Petzold *et al*. that the preservation of the Heslington brain is likely due to the formation of protein aggregates [2]. The experiment that contributed most to this conclusion first involved taking samples from the white and grey matter of the preserved brain and a normal brain as the control in order to investigate the presence of protein aggregates. The samples were homogenized, centrifuged, and then the resulting liquid was subjected to gel electrophoresis. The first four bands were used for immune-blotting, which revealed the presence of hyperphosphorylated NFH with a molecular weight of 420 kDa, twice the amount to be expected. Hyperphosphorylation of NFs is thought to be one of the main triggers that lead to aggregate formation in NFs, so a large amount of hyperphosphorylated NFH is indicative that there are large protein aggregates within the preserved brain [6,7]. Another indication that the preserved brain contained large aggregates was the fact that they had a higher degree of resistance to urea than aggregates from normal brains.

Protein aggregates have a tendency to form from proteins that contain intrinsically disordered regions (IDRs), referred to as intrinsically disordered proteins (IDPs) [8]. IDPs lack the ability to fold and form rigid 3D structures; they make up one-third of all eu-

 doi:10.20944/preprints202209.0450.v1

karyotic proteins [9,10]. IDRs cannot fold because they contain a high amount of net charge and a low amount of hydrophobic residues, relative to ordered proteins. A high amount of net charge and low amount of hydrophobicity within a given region increases repulsive forces and reduces the drive for hydrophobic collapse respectively. IDRs allow IDPs to be very flexible and take on many different conformations. A high number of conformations allow IDPs to have numerous functions and binding partners. Although these traits are important for life, they are also the reason why IDPs are associated with a number of diseases, especially those that arise from misfolding due to certain conformational changes[8]. A higher number of possible conformations and binding partners increases the chances of misfolding taking place. Many proteins that play important roles in neurodegenerative pathways are IDPs e.g. amyloid-β, tau, and α-synuclein [11].

In this paper, intrinsic disorder and protein interactions of NFs, GFAP, and MBP are investigated. The proteins that were detected using mass spectroscopy are also analyzed to gauge the amount of intrinsic disorder that may be present within the preserved brain. Proteins that are found to be possible interactors with NFs, GFAP, and MBP are used to determine possible pathways that may have been responsible for the high degree of aggregate formation within the ancient brain.

2. Materials and Methods

2.1. Protein Datasets

ID of proteins analyzed in this study were retrieved from the rsif20190775_si_001 dataset posted by Axel Petzold, Ching-Hua Lu, Mike Groves, Johan Gobom, Henrik Zetterberg, Gerry Shaw, Sonia O'Connor on 16.12.2019, 05:17. This dataset includes proteins listed in a Supplementary Table with the mass spectrometry data from [2]. Sequences of these query proteins in FASTA format were gathered from the UniProt database [12,13] and are listed in Supplementary Materials. These retrieved sequences represent a "mass spectrometry-identified proteins" set. We conducted a comprehensive bioinformatics analysis of the neurofilament heavy, medium, and light chain proteins (NFs, UniProt IDs: P12036, P07197, P07196, respectively), glial fibrillary acidic protein (GFAP; UniProt ID: P14136), and myelin basic protein (MBP; UniProt ID: P02686), which were shown to be engaged in the formation the non-amyloid protein aggregates [2], and which are defined here as the "main proteins". Entire human proteome (20,317 manually curated proteins; proteome:UP000005640) was downloaded from the UniProt database [12,13], from which 10,611 brain proteins were further selected (proteome:UP000005640 AND brain).

2.2. Evaluation of the Intrinsic Disorder Predisposition

Predisposition for intrinsic disorder of all proteins were determined using a set of commonly used per-residue disorder predictors including PONDR® VLS2, PONDR® VL3, PONDR® VLXT, PONDR® FIT, IUPred-Long, IUPred-Short [14-19]. A web application called Rapid Intrinsic Disorder Analysis Online (RIDAO) was used to gather results from each predictor in bulk [20]. The percent of predicted intrinsically disorder residues (PPIDR) for each protein was used to classify each protein based on their level of disorder. A residue was considered to be disordered if it had a value of 0.5 or higher. Generally, a PPIDR value of less than 10% is taken to correspond to a highly ordered protein, PPIDR between 10% and 30% is ascribed to moderately disordered protein, and PPIDR greater than 30% corresponds to a highly disordered protein [21,22]. In addition to PPIDR, mean disorder score (MDS) was calculated for each query protein as a protein length-normalized sum of all the per-residue disorder scores. Here again, proteins are grouped based on their corresponding MDS values, being classified as highly ordered (MDS < 0.15), moderately disordered of flexible (MDS between 0.15 and 0.5) and highly disordered (MDS ≥ 0.5).

2.3. PPI Networks

 doi:10.20944/preprints202209.0450.v1

Protein-protein interaction (PPI) networks were generated using STRING (search tool for recurring instances of neighboring genes) platform (http://string-db.org/) [23]. In this study, STRING was used in two modes. First, a global interaction network for all 881 mass spectrometry-identified proteins was generated using a moderate confidence level of 0.7. Then, we generated networks of NFs, GFAP, and MBP using 0.7 confidence and 500 as the limit for possible number of interactors (max).

Agile Protein Interactomes DataServer (APID, http://apid.dep.usal.es) [24] was used to determine which mass spectrometry-identified proteins are involved in interactions with NFs, GFAP, and/or MBP. First, the IDs for the 5 main proteins were uploaded to APID. For each protein, the list of interactors with at least one evidence was saved, resulting in 5 lists of specific interactors. Each list was then compared to the list of mass spectrometry-identified proteins in order to determine if any of the interactors in the list were detected in the brain. The resulting lists were combined, and duplicates were removed. In total, 44 interactors were found. PPI networks including those interactors and the 5 main proteins were created using APID and STRING.

2.4. Disorder-Based Functional Annotations

The Database of Disordered Protein Predictions (D^2P^2) was used to determine binding sites based on the ANCHOR algorithm and sites of various posttranslational modifications (PTM) [25]. D^2P^2 also provides Structural Classification of Proteins (SCOP) domain predictions based on the SUPERFAMILY predictor and disorder predictions based on PONDR VLXT, PONDR VSL2b, PrDOS, PV2, ESpritz-DisProt, Espritz-XRay, Espritz-NMR, IUPred-Long, and IUPred-Short predictors [25].

2.5. CH-CDF Analysis

CH-CDF analysis was for gauging which proteins found through mass spectroscopy were disordered and was generated using RIDAO. This analysis combines results from charge-hydropathy (CH) and cumulative distribution function (CDF) plots, which are both binary predictors of disorder. For CH, net charge is plotted versus hydropathy for each protein [26]. Due to the observation that disordered proteins tend to have high net charge and low hydropathy, disordered and ordered proteins cluster two regions of the plot. A linear boundary separates disordered and ordered proteins [26,27]. Proteins that are disordered appear above the boundary while ordered proteins appear below [26,27]. In the CDF-plot predictor, PONDR scores for each residue of a single protein is plotted against their frequency within the sequence. If a CDF curve of a given protein is below the order-disorder boundary, this protein is considered to be disordered and ordered if the CDF curve is located above this boundary [27]. Data generated by CH- and CDF-plots are then combined to generate ΔCH-ΔCDF plot [28-30], which enables rapid discrimination between flavors of disorder [31]. To this end, for each query protein, ΔCH, the vertical distance of the corresponding point in CH plot from the boundary, is calculated, whereas ΔCDF is computed as the average distance between the order-disorder boundary and the CDF curve. Then, ΔCH is plotted against ΔCDF resulting in a CH-CDF plot. Proteins in the top-left quadrant are predicted to be disordered by both CH and CDF, the ones in the bottom-left are predicted to be ordered by CH and disordered by CDF, the ones in the top-right are predicted to be disordered by CH and ordered by CDF, and in the bottom-right quadrant predicted to be ordered by both [28-31].

2.6. Pathway Analysis

The gene IDs of the 44 proteins that were found to interact with NFs, GFAP, and/or MBP were inputted into a webserver called DAVID, which stands for Database for Annotation, Visualization, and Integrated Discovery (http://www.david.niaid.nih.gov) [32]. It was used to investigate if any of the 44 interactors that reside in the preserved brain are part of any pathways that involve protein aggregate formation. The 5 proteins were also uploaded to the KEGG (Kyoto encyclopedia of genes and genomes) [33] pathway database (https://www.genome.jp/kegg/pathway.html) to carry out their functional analysis.

 doi:10.20944/preprints202209.0450.v1

2.7. 3D Model Structures of Main Proteins

3D structural models of the main proteins (NFs, GFAP, and MBP) were generated by AlphaFold [34].

3. Results

3.1. Functional intrinsic disorder in the "main proteins"

At the first stage of our study, we looked at the intrinsic disorder predisposition of five "main proteins", the neurofilament heavy, medium, and light chain proteins (NFH, NFM, and NFL, UniProt IDs: P12036, P07197, P07196, respectively), glial fibrillary acidic protein (GFAP; UniProt ID: P14136), and myelin basic protein (MBP; UniProt ID: P02686), which were shown to be engaged in the formation the non-amyloid protein aggregates, and which were considered as the major constituents defining preservation of the Heslington brain for at least 2,600 years [2]. Results of the multifactorial disorder analysis of these natural preservatives are shown in **Figures 1** through **5**.

For these five "main proteins", we looked at the per-residue disorder profiles generated by RIDAO, functional disorder profiles produced by the D2P2 platform, protein-protein interaction networks generated by STRING, and 3D structural models generated by AlphaFold.

3.1.1. NFH

Figure 1 represents results of functional disorder analysis of human NFH, which is a 1,026 residue-long intermediate filament (IF) protein that together with the NFM and NFL is involved in the maintenance of the neuronal caliber. This protein includes a head region (residues 1-200), an UF rod domain (residues 97-413) and a tail region (residues 414-1026) containing 30 × 6 amino acid repeats of K-S-P-[AEPV]-[EAK]-[AEVK]. Many serines of the K-S-P repeats of the NFH protein are phosphorylated in addition to Ser76, Ser124, Ser347, and Thr774. It is believed that the NFH phosphorylation within the repeat region leads to the formation of interfilament cross bridges that play important roles in the maintenance of axonal caliber [35], whereas phosphorylaron within the head and rod regions inhibits polymerization [35].

The PPIDR values generated by the PONDR® FIT, PONDR® VSL2, PONDR® VL3, PONDR® VLXT, IUPred Short, and IUPred Long were 56.34%, 90.45%, 85.77%, 50.19%, 58.87%, and 71.05% respectively. From the mean of these predictors, 83.82% of the NFH residues were predicted to be disordered (scores greater than or equal to 0.5). Based on these parameters, NFH can be classified as a highly disordered protein. The most disorder is found between residue 440 and the C-terminal tail (see **Figure 1A**). NFH exists as two isoforms generated by alternative splicing, with isoform 2 being different from the canonical form by missing residues 750-812 within the tail region.

According to the ANCHOR algorithm, NFH has 16 disorder-based binding sites: 86-92, 104-115, 264-269, 393-402, 412-421, 506-517, 520-528, 540-547, 614-623, 666-675, 680-688, 755-764, 777-787, 801-807, 912-917, and 1015-1020 (see **Figure 1B**). These indicates that NFH can be engaged a multiple disorder-driven interactions. This conclusion is supported by Figure 1C showing the PPI network generated by STRING using 0.7 as the confidence level. There are 18 interactors and 73 interactions (see **Figure 1C**). Since the expected number of interactions for a random set of proteins of the same size and degree distribution drawn from the genome is 18, and since the PPI enrichment p-value is $<10^{-16}$, the network centered at NFH has significantly more interactions than expected. Furthermore, the average node degree of this network is 8.11 (i.e., each protein in the network interact with at least 8 partners), and its average local clustering coefficient is 0.856.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

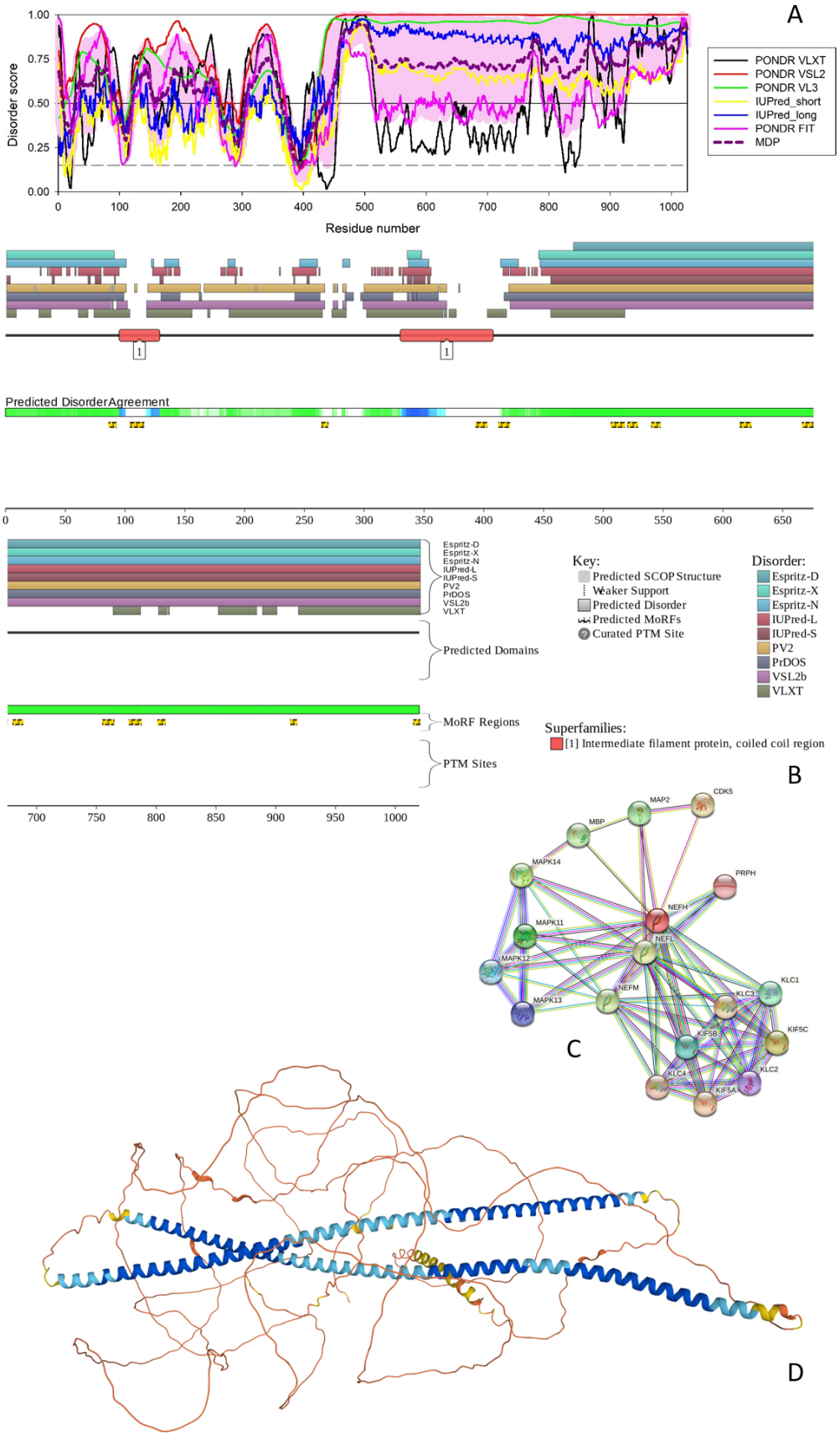


Figure 1. Functional disorder analysis of human NFH (UniProt IDs: P12036). **A**. Per-residue disorder profile generated by RIDAO. Solid and dashed vertical lines at disorder scores 0.5 and 0.15

 doi:10.20944/preprints202209.0450.v1

correspond to the disorder and flexibility thresholds. **B**. Functional disorder profile produced by the D²P² platform. Here, 9 colored bars represent the location of disordered regions as predicted by different disorder predictors. In the middle of the D²P² plots the blue-green-white bar shows the agreement between the outputs of nine disorder predictors (IUPred, PONDR® VLXT, PONDR® VSL2, PrDOS, PV2, and ESpritz), with blue and green parts corresponding to disordered regions by consensus. Above the disorder consensus bar are two lines with colored and numbered bars that show the positions of the (mostly structured) SCOP domains [36,37] predicted using the SUPERFAMILY predictor [38]. Yellow zigzagged bar shows the location of the predicted disorder-based binding sites (MoRF regions) identified by the ANCHOR algorithm [39], whereas differently colored circles at the bottom of the plot show location of various PTMs assigned using the outputs of the PhosphoSitePlus platform [40], which is a comprehensive resource of the experimentally determined post-translational modifications. **C**. STRING-generated PPI network centered at human NFH. In this network, the nodes correspond to proteins, whereas the edges show predicted or known functional associations. Seven types of evidence are used to build the corresponding network, where they are indicated by the differently colored lines: a green line represents neighborhood evidence; a red line – the presence of fusion evidence; a purple line – experimental evidence; a blue line – co-occurrence evidence; a light blue line – database evidence; a yellow line – text mining evidence; and a black line – co-expression evidence [23]. **D**. 3D structure modeled for NFH by AlphaFold. Structure is colored based on the pLDDT values, where segments predicted with very high (pLDDT > 90), high (90 > pLDDT > 70), low (70 > pLDDT > 50), and very low (pLDDT < 50) confidence are shown by blue, cyan, yellow and orange colors, respectively.

All these data indicate that the NFH-centered PPI network is highly connected. The most enriched biological processes of the proteins in this network (in terms of the Gene Ontology, GO) are Microtubule-based process (GO:0007017; $p = 2.29\times10^{-06}$), Axo-dendritic transport (GO:0008088; $p = 2.88\times10^{-06}$), Neurofilament bundle assembly (GO:0033693; $p = 5.28\times10^{-05}$), Microtubule-based movement (GO:0007018; $p = 5.28\times10^{-05}$), and Anterograde dendritic transport of neurotransmitter receptor complex (GO:0098971; $p = 5.85\times10^{-05}$). Among five most enriched molecular functions are Microtubule motor activity (GO:0003777; $p = 1.72\times10^{-09}$), Microtubule binding (GO:0008017; $p = 2.26\times10^{-06}$), MAP kinase activity (GO:0004707; $p = 4.25\times10^{-06}$), Cytoskeletal protein binding (GO:0008092; $p = 3.36\times10^{-05}$), and Kinesin binding (GO:0019894; $p = 4.02\times10^{-05}$). Finally, the most enriched cellular components related to the proteins in this network are Polymeric cytoskeletal fiber (GO:0099513; $p = 2.39\times10^{-12}$), Kinesin complex (GO:0005871; $p = 4.54\times10^{-11}$), Microtubule associated complex (GO:0005875; $p = 2.31\times10^{-10}$), Axon (GO:0030424; $p = 4.06\times10^{-10}$), and Microtubule (GO:0005874; $p = 7.04\times10^{-09}$.

Finally, **Figure 1D** shows that most of the NFH 3D structure is predicted by AlphaFold with low or very low per-residue confidence scores (70 > pLDDT > 50 and pLDDT < 50), indicating that most of this protein may be unstructured in isolation. Furthermore, it is very likely that the two long α-helices (residues 90-248 and 267-410), which are predicted by AlphaFold and overlap with the coil 1A/Bb (residues 101-132/146-244) and coil 2A/2B (residues 267-288/293-413) regions will be formed as a result of the coiled-coil-based assembly of the neuronal intermediate filaments, where "three polypeptides first make αhelical coiled-coil dimers, then form antisymmetric tetramers and finally assemble into 10–15 nm diameter filaments with long flexible side-arms extending 50–100 nm from the polymer core" [41]. It was also indicated that initial steps of the neurofilament assembly rely on the preferential NFH-NFL and NFM-NFL heterodimerization [42,43]. Since in the mature axons, the stoichiometry of NFL:NFM:NFH is 7:3:2, on average, one can find there 1 NFL:NFL homodimer, 2 NFH:NFL dimers, and 3 NFM:NFL dimers [41].

3.1.2. NFM

NFM (UniProt ID: P07197) is a 916-residue long protein. Similar to NFH, this protein has a head region (residues 2-104), IF rod domain (residues 101-412) containing coiled-coil regions 1A, 1B, 2A, and 2B (residues 105-136, 150-248, 266-287, and 292-412, respectively, and a tail region containing 6 × 13 AA approximate tandem repeats of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

K-S-P-V-[PS]-K-S-P-V-E-E-[KA]-[GAK]. The isoform 2 of NFM is different from the canonical form by missing head and most of the rod regions (residues 1-376).

Also similar to NFM, this intermediate filament protein is phosphorylated on multiple serines regions within the repeats of the tripeptide K-S-P and in the head and rod domains. In addition to containing multiple phosphoserines, NFM is phosphorylated on Thr571, contains two O-linked (GlcNAc) threonines at positions 47 and 431, and includes a Omega-N-methylarginine at position 42 (https://www.uniprot.org/uniprotkb/P07197/).

The PPIDR values according to each disorder predictor suggest that the NFM is a highly disordered protein. For PONDR® FIT, PONDR® VSL2, PONDR® VL3, PONDR® VLXT, IUPred Short, and IUPred Long, the PPIDR values were 58.52%, 88.32%, 84.83%, 56.44%, 45.20%, and 64.74% respectively. The PPIDR from the mean of these predictors was 77.62%. According to the mean, the disordered regions include residues 1-106, 160-213, 273-282, 303-364, 430-441, and 454 to the C-terminus (see **Figure 2A**). According to the ANCHOR algorithm, NFM has 16 disorder based binding sites: 1-6, 56-69, 110-122, 290-299, 390-401, 449-457, 469-486, 504-520, 543-551, 573-580, 594-609, 740-757, 798-806, 829-838, 853-858, and 869-877 (**Figure 2B**). The STRING-generated PPI network, shown in **Figure 2C**, includes 21 interactors, 77 interactions, an average node degree of 7.33, an average local clustering coefficient of 0.88.

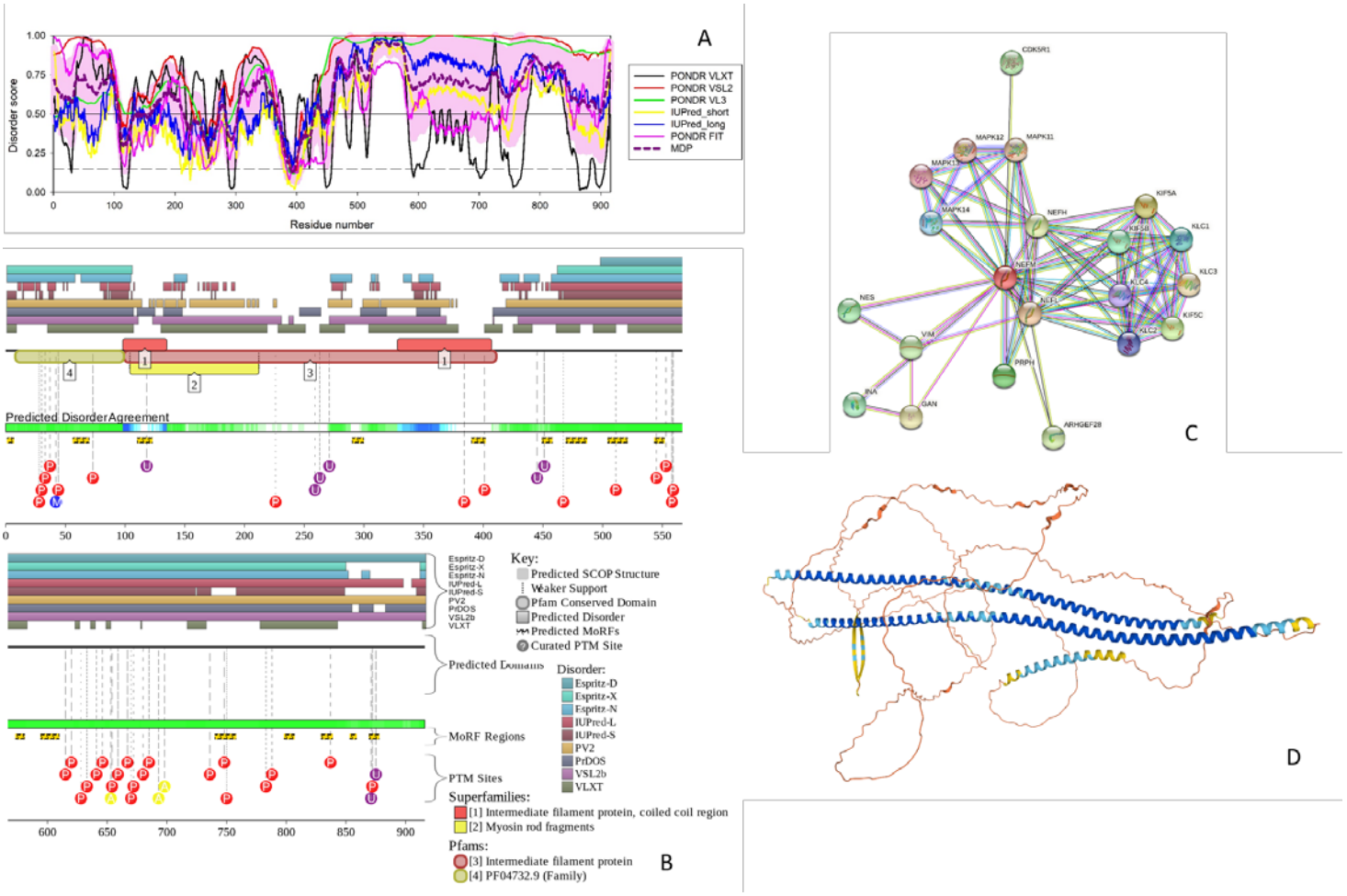


Figure 2. Functional disorder analysis of human NFM (UniProt IDs: P07197). **A**. Per-residue disorder profile generated by RIDAO. **B**. Functional disorder profile produced by the D2P2 platform. **C**. STRING-generated PPI network centered at NFM. **D**. 3D structure modeled for NFM by AlphaFold. Structure is colored based on the pLDDT values, where segments predicted with very high (pLDDT > 90), high (90 > pLDDT > 70), low (70 > pLDDT > 50), and very low (pLDDT < 50) confidence are shown by blue, cyan, yellow and orange colors, respectively.

 doi:10.20944/preprints202209.0450.v1

The expected number of edges for a random set of proteins of the same size and degree distribution drawn from the genome was 21 and the PPI enrichment p-value $<10^{-16}$, indicating that this NFM-centered network has significantly more interactions than expected. Among the most enriched biological processes related to proteins in this network are Intermediate filament cytoskeleton organization (GO:0045104; $p = 2.50\times10^{-7}$), Axo-dendritic transport (GO:0008088; $p = 4.52\times10^{-5}$), Neurofilament bundle assembly (GO:0033693; $p = 6.81\times10^{-5}$), Anterograde dendritic transport of neurotransmitter receptor complex (GO:0098971; $p = 6.81\times10^{-5}$), and Intermediate filament organization (GO:0045109; $p = 6.81\times10^{-5}$). The most enriched molecular functions of these proteins are Microtubule motor activity (GO:0003777; $p = 6.23\times10^{-9}$), MAP kinase activity (GO:0004707; $p = 1.10\times10^{-5}$), Structural constituent of postsynaptic intermediate filament cytoskeleton (GO:0099184; $p = 1.72\times10^{-5}$), Tubulin binding (GO:0015631; $p = 4.68\times10^{-5}$), and Structural constituent of cytoskeleton (GO:0005200; $p = 4.98\times10^{-5}$). These proteins were most abundantly found within the following cellular components: Polymeric cytoskeletal fiber (GO:0099513; $p = 1.11\times10^{-12}$), Supramolecular fiber (GO:0099512; $p = 1.11\times10^{-12}$), Kinesin complex (GO:0005871; $p = 6.73\times10^{-11}$), Cytoskeleton (GO:0005856; $p = 2.62\times10^{-9}$), and Neurofilament (GO:0005883; $p = 1.02\times10^{-7}$).

The high level of disorder in this protein is further evidenced by the **Figure 2D** showing its 3D structure modeled by AlphaFold, where one can see two very long and intertwined α-helices (residues 92-250 and 267-416), a shorter α-helix (residues 453-492) and a β-hairpin (residues 851-883), with the remaining protein being expected to be unstructured in isolation. Again, with a high probability, long helical segments are formed as a result of the coiled-coil-driven dimerization of NFM with NFL.

3.1.3. NFL

Human NFL (UniProt ID: P07196) is the shortest intermediate filament protein that contains 545 residues. Despite being noticeably shorter, this protein contains a head and IF rod domains (residues 2-92 and 90-400, respectively), and a tail region (residues 397-543) with the subdomains A and B (residues 397-443 and 4440543, respectively). Also, there are coils 1A (residues 93-124), 1B (residues 138-234), 2A (residues 253-271), and 2B (residues 281-396) in the rod domain. NFL is phosphorylated in the head and rod regions, contains O-linked (GlcNAc) threonine and O-linked (GlcNAc) serine at positions 21 and 30, respectively, and Omega-N-methylarginine at position 30. The Arg23 residue of this protein can be modified to Asymmetric dimethylarginine or Omega-N-methylarginine (https://www.uniprot.org/uniprotkb/P07196/entry#ptm_processing).

For this protein, the PPIDR values evaluated from the PONDR® FIT, PONDR® VSL2, PONDR® VL3, PONDR® VLXT, IUPred Short, and IUPred Long outputs were 58.38%, 82.69%, 60.22%, 67.40%, 20.81%, and 28.73% respectively, whereas the PPIDR from the mean disorder predictor was 56.72%, indicating that similar to other intermediate filament proteins, NFL is expected to be highly disordered (see **Figure 3A**). According to the mean disorder profile (see **Figure 3A**), the disordered regions of NFL include residues 1-7, 20-75, 147-213, 295-359, 416-424, and 441-545. **Figure 3B** shows that according to the ANCHOR algorithm, NFL has 4 disorder-based binding sites: 381-388, 422-465, 488-501, and 530-539 and contains multiple PTMs.

The NFL-centered PPI network shown in **Figure 3C** includes 46 interactors, 214 interaction, average node degree of 9.3, average local clustering coefficient of 0.77, and a PPI enrichment p-value of $<10^{-16}$. The expected number of edges was 63, so this network is highly enriched. Proteins in this network are enriched in the following biological processes: Regulation of NMDA receptor activity (GO:2000310; $p = 6.82\times10^{-16}$), Regulation of neurotransmitter receptor activity (GO:0099601; $p = 8.97\times10^{-15}$), Axo-dendritic transport (GO:0008088; $p = 6.23\times10^{-14}$), Modulation of chemical synaptic transmission (GO:0050804; $p = 3.94\times10^{-3}$), and Regulation of cation channel activity (GO:2001257; $p = 2.27\times10^{-11}$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

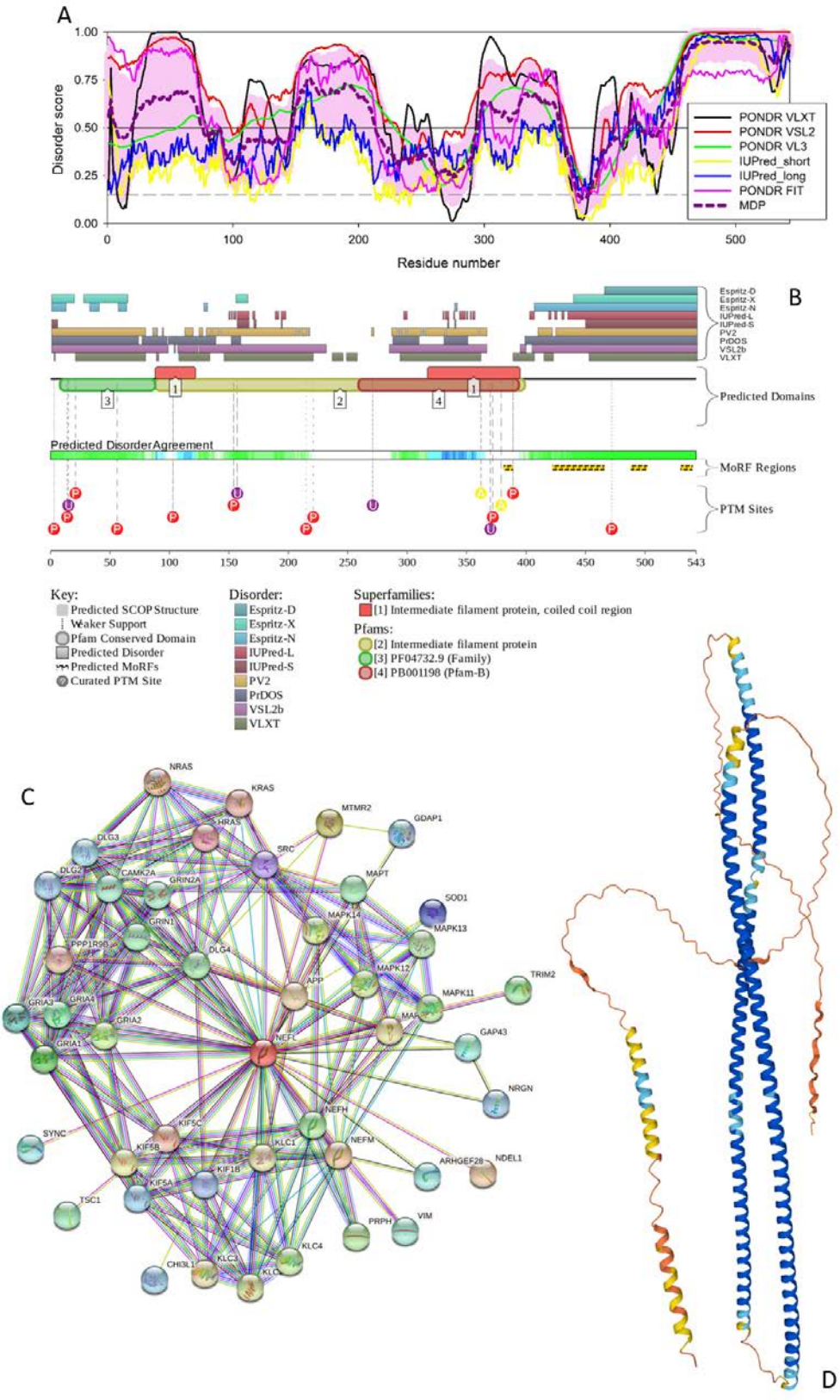


Figure 3. Functional disorder analysis of human NFL (UniProt IDs: P12036). **A**. Per-residue disorder profile generated by RIDAO. **B**. Functional disorder profile produced by the D2P2 platform. **C**. STRING-generated PPI network centered at NFL. **D**. 3D structure modeled for NFL by AlphaFold. Structure is colored based on the pLDDT values, where segments predicted with very high (pLDDT > 90), high (90 > pLDDT > 70), low (70 > pLDDT > 50), and very low (pLDDT < 50) confidence are shown by blue, cyan, yellow and orange colors, respectively.

 doi:10.20944/preprints202209.0450.v1

Most enriched molecular functions are Ionotropic glutamate receptor activity (GO:0004970; $p = 5.40\times10^{-8}$), Microtubule motor activity (GO:0003777; $p = 5.40\times10^{-8}$), Microtubule binding (GO:0008017; $p = 3.96\times10^{-7}$), AMPA glutamate receptor activity (GO:0004971; $p = 9.93\times10^{-7}$), and Transmitter-gated ion channel activity involved in regulation of postsynaptic membrane potential (GO:1904315; $p = 2.25\times10^{-6}$). Finally, among the most enriched cellular components, there are Neuron projection (GO:0043005; $p = 2.17\times10^{-23}$), Axon (GO:0030424; $p = 1.41\times10^{-20}$), Somatodendritic compartment (GO:0036477; $p = 5.13\times10^{-18}$), Postsynaptic density (GO:0014069; $p = 2.19\times10^{-17}$), and Dendrite (GO:0030425; $p = 2.19\times10^{-17}$).

As per AlphaFold, human NFL is expected to contain two long intertwined α-helices (residues 73-241 and 255-403) characteristic of the coiled-coil structures with a shorter C-terminal helix (residues 481-539) predicted with low confidence (see **Figure 3D**). This is a typical structural organization of the proteins involved in the formation of intracellular skeleton and extracellular matrix. It is likely that such long intertwined helices are formed as a result of the binding-induced folding, as extremely long helices cannot exists in isolation. In line with these considerations, it was noted that compared to the complete human proteome, the extracellular proteome is significantly enriched in proteins with high disorder content (> 50%) and includes many coiled-coil proteins [44].

Furthermore, it was indicated that coiled-coils are commonly predicted as unstructured regions [45]. Therefore, it is not surprising that human NFL is expected to show high helical content despite being predicted as mostly disordered protein. Curiously, NFL is the only IF protein that can form homodimers [46], being also crucial for the formation of heterodimers with NFM and NFH [42,43]. Therefore, intrinsically disordered nature of these three IF proteins and their capability to undergo binding-induced disorder-to-order transition leading to the formation of long α-helical segments are crucial the for the neurofilament assembly.

3.1.4. GFAP

Human glial fibrillary acidic protein (GFAP; UniProt ID: P14136) is a type III intermediate filament protein that serves as a highly specific marker for the central nervous system (CNS) astrocytes [47-49] and which is most abundantly expressed in the brain [50]. In addition to GFAP, a group of type III intermediate filament protein includes desmin, peripherin, and vimentin [51]. A very specific feature of GFAP is the presence of multiple isoforms generated by alternative splicing and alternative polyadenylation, where in human, in addition to the canonical isoform GFAPα, there are twelve splice-isoforms [50,52-54]. The domain structure of GFAP is typical for the intermediate filament proteins, and this protein contains a head domain (residues 1072), IF rode (residues 69-377) that includes coils 1A (residues 73-104), 1B (residues 116-214), 2A (residues 231-252), and 2B (residues 257-377), and the tail region (residues 378-432).

Human GFAP contains multiple PTMs, being phosphorylated at Thr7, Ser13, Ser38, Ser82, Thr110, Thr150, Ser323, Thr383, and Ser385 [55,56]. It also citrullinated on arginine residues at positions 30, 36, 270, 406, and 416 [57], and has an Omega-N-methylarginine at position 12.

Figure 4A shows that human GFAP is expected to contain high levels of intrinsic disorder. The PPIDR values derived for GFAP from PONDR® FIT, PONDR® VSL2, PONDR® VL3, PONDR® VLXT, IUPred Short, and IUPred Long were 48.84, 82.41%, 59.72%, 68.52%, 12.96%, and 22% respectively, whereas the PPIDR based on the mean disorder prediction was 55.09%, thereby classifying GFAP as a highly disordered protein. Looking at the mean disorder profile (see **Figure 4A**), one can find seven disordered regions in this protein, residues 1-51, 63-73, 136-217, 274-281, 291-343, 395-398, and 408-432. According to the ANCHOR algorithm, GFAP has 4 disorder-based binding sites at residues 52-59, 259-265, 289-298, and 352-366 (see **Figure 4B**). D^2P^2-based functional disorder profile also shows that the human GFAP is densely decorated by various PTMs (see **Figure 4B**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

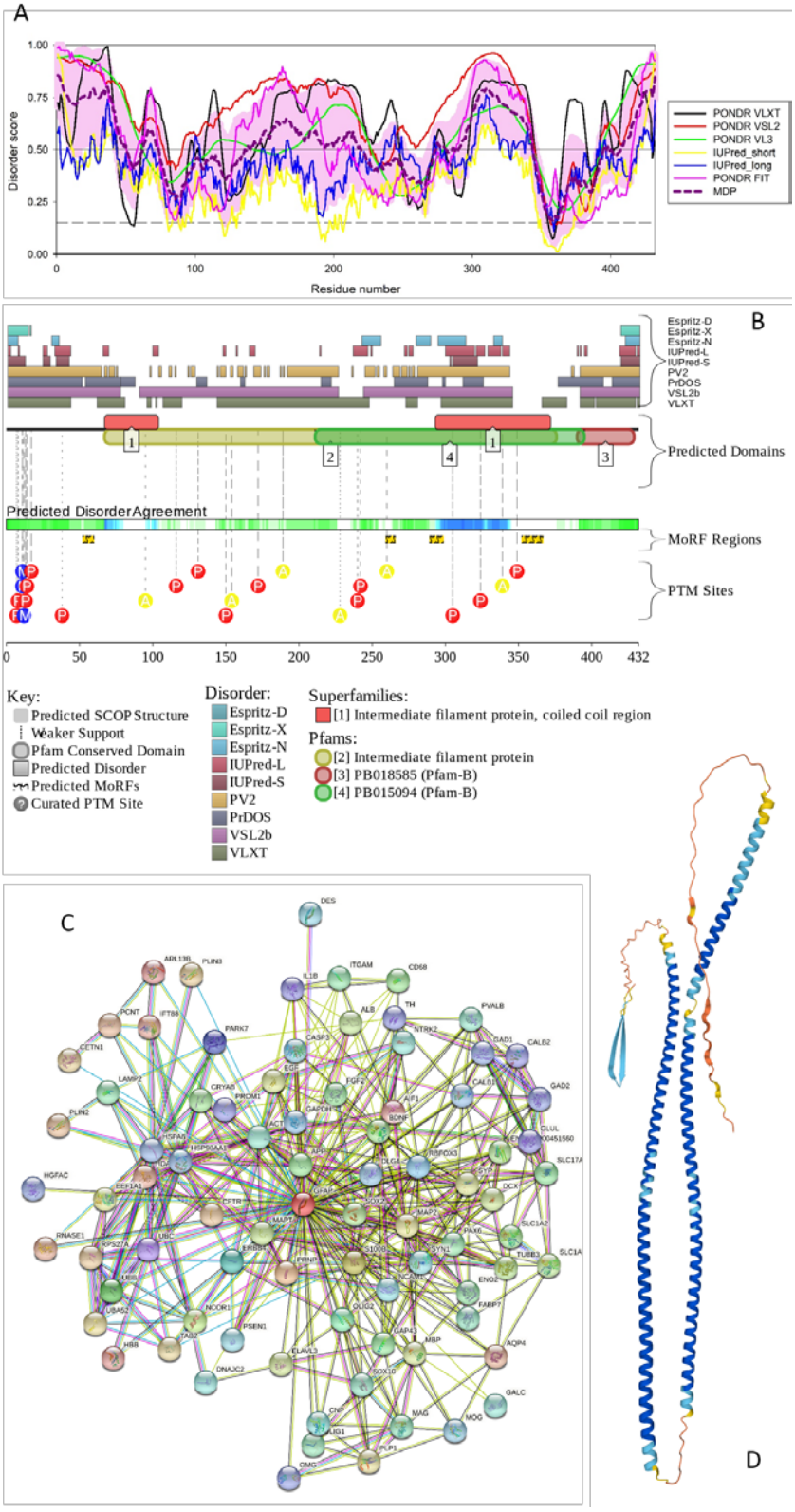


Figure 4. Functional disorder analysis of human glial fibrillary acidic protein (GFAP; UniProt ID: P14136). **A**. Per-residue disorder profile generated by RIDAO. **B**. Functional disorder profile produced by the D²P² platform. **C**. STRING-generated PPI network centered at GFAP. **D**. 3D structure modeled for GFAP by AlphaFold. Structure is colored based on the pLDDT values, where segments predicted with very high (pLDDT > 90), high (90 > pLDDT > 70), low (70 > pLDDT > 50), and very low (pLDDT < 50) confidence are shown by blue, cyan, yellow and orange colors, respectively.

 doi:10.20944/preprints202209.0450.v1

The STRING generated PPI network is displayed in **Figure 4C**. It has 81 interactors, 420 interactions, average node degree of 10.4, average local clustering coefficient of 0.646, and a PPI enrichment p-value of $<10^{-16}$. This network has significantly more interactions than expected, as the number of expected interactions is 159. Analysis of the functional enrichment of the proteins included into this network (in terms of the GO terms) revealed that five most enriched biological processes are Nervous system development (GO:0007399; $p = 1.50\times10^{-17}$), Neuron differentiation (GO:0030182; $p = 2.71\times10^{-15}$), Central nervous system development (GO:0007417; $p = 8.51\times10^{-15}$), Neuron development (GO:0048666; $p = 9.01\times10^{-14}$), and Generation of neurons (GO:0048699; $p = 1.97\times10^{-13}$). The most enriched molecular functions of these proteins are Protein binding (GO:0005515; $p = 3.76\times10^{-10}$), Binding (GO:0005488; $p = 3.96\times10^{-6}$), Identical protein binding (GO:0042802; $p = 0.00012$), Protein tag (GO:0031386; $p = 0.00014$), and Enzyme binding (GO:0019899; $p = 0.00014$). Among the most $3.07.40\times10^{-20}$), Somatodendritic compartment (GO:0036477; $p = 1.64\times10^{-17}$), Neuron projection (GO:0043005; $p = 1.64\times10^{-17}$), Plasma membrane bounded cell projection (GO:0120025; $p = 1.64\times10^{-17}$), and Cell projection (GO:0042995; $p = 1.64\times10^{-17}$).

Figure 4D represents the 3D structure modeled for GFAP by AlphaFold and shows that this protein is structurally similar to other IF proteins and contains two long inter-twined α-helices (residues 52-106/110-215 and 233-378). With a high probability, these structural elements are formed as a result of the binding-induced folding. In fact, similar to the neurofilament assembly discussed in sections dedicated to NFH, NFM and NFL, GFAP forms IF network, where "GFAP can form homodimers, but also heterodimers with vimentin. In the next step of IF network assembly, dimers bind in an antiparallel fashion to form tetramers, which then laterally associate into octamers forming structures called unit-length fragments (ULFs). Subsequent association of ULFs in a non-polar fashion leads to the formation of a filament, which then undergoes radial compaction leading to the final diameter of 10 nm" [50,58].

3.1.5. MBP

Human myelin basic protein (MBP; UniProt ID: P02686) is a 304-residue-long intrinsically disordered protein [59], which is highly abundant in the central nervous system (CNS) myelin sheath, a multilayered proteolipid membrane, which is crucial for the neural insulation and saltatory conduction of nerve impulses [60]. In fact, MBP is crucial for the maintenance of the structural integrity of the myelin sheath, where it holds together the apposing cytoplasmic leaflets of the oligodendrocyte mem-brane in a tight, spiral, multilamellar arrangement [61].

This protein is encoded by an *MBP* gene, which is included into the Golli (Genes of OLigodendrocyte Lineage) gene complex and which contains 11 exons in mice and 10 in humans, including the 7 exons translated into the classic MBP [62]. However, MBP is known to exist as several isoforms, which are noticeably different in size and charge (e.g., 21.5, 20.2, 18.5, 17.24, 17.22, and 14 kDa in the mouse and 21.5, 20.2, 18.5, and 17.2 kDa in humans) and which are generated by alternative splicing of a single mRNA [5,59,63-66]. Being characterized by a very basic pI (pI 9.79), the longest MBP isoform (Golli-MBP1 or HOG7; 304 residues, 33.1 kDa) contains 29 Arg, 19 Lys, 22 Asp, and 16 Glu residues. On the other hand, the alternative splicing-generated isoforms 2 (Golli-MBP2 or HOG5; 197 residues, 21.5 kDa), 3 (MBP1; 197 residues, 21.5 kDa), 4 (MBP2; 186 residues, 20.2 kDa), 5 (MBP3, 171 residues, 18.5 kDa), and 6 (MBP4, 160 residues, 17.2 kDa) are characterized by the pI values of 5.99, 11.35, 11.14, 11.38, and 11.14, respectively. Furthermore, MBP isolated from brains shows extensive PTMs, such as deimination, phosphorylation, deamidation, methylation, and N-terminal acylation [67,68]

Although no specific functional domains we assigned to MBP, this protein contains several regions with characteristic amino acid biases, such as regions enriched in basic and acidic residues (1-44 and 95-117) or polar residues (45-71 and 118-147) (https://www.uniprot.org/uniprotkb/P02686/entry).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

The intrinsically disordered nature of the MBP is known for more than 50 years [69], being confirmed over and over again in multiple comprehensive studies [61,70-72], including an impressive (but unsuccessful) attempt to crystallize MBP under 4600 different conditions that eventually resulted in an important conclusion that "18.5 kDa MBP and possibly also its isoforms will remain preeminent examples of proteins that cannot be crystallized" [73]. In agreement with these studies, our computational analysis revealed that the human MBP is an extremely disordered protein. In fact, the PPIDRs calculated using the outputs of PONDR® FIT, PONDR® VSL2, PONDR® VL3, PONDR® VLXT, IUPred Short, and IUPred Long were 82.57%, 100%, 100%, 58.22%, 82.57%, and 93.75% respectively (see **Figure 5**).

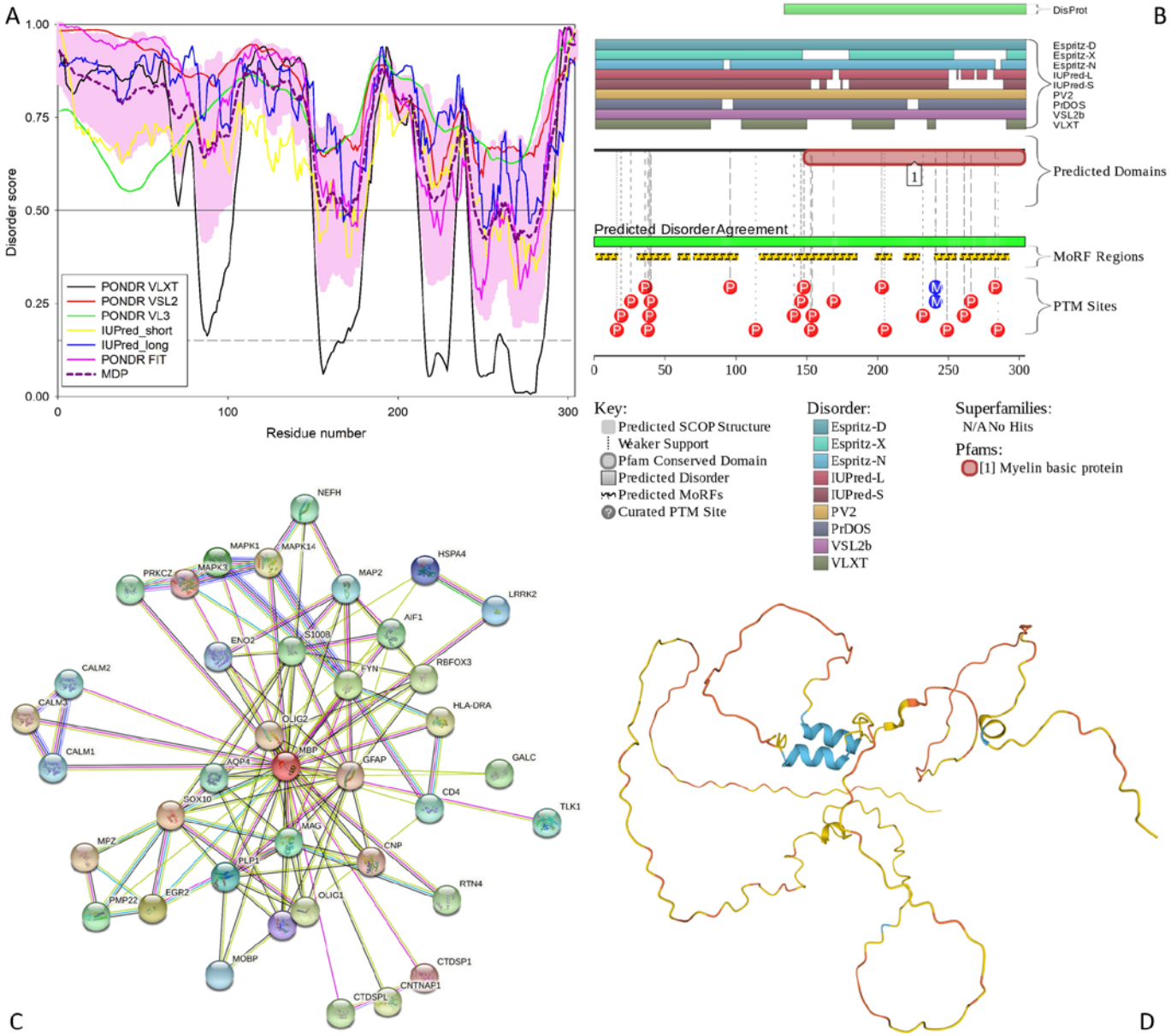


Figure 5. Functional disorder analysis of the myelin basic protein (MBP; UniProt ID: P02686). **A**. Per-residue disorder profile generated by RIDAO. **B**. Functional disorder profile produced by the D²P² platform. **C**. STRING-generated PPI network centered at MBP. **D**. 3D structure modeled for MBP by AlphaFold. Structure is colored based on the pLDDT values, where segments predicted with very high (pLDDT > 90), high (90 > pLDDT > 70), low (70 > pLDDT > 50), and very low (pLDDT < 50) confidence are shown by blue, cyan, yellow and orange colors, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

The PPIDR evaluated from the mean disorder profile was 86.51%. Therefore, out of the 5 "main proteins" analyzed in this study, MBP displayed the most disorder (see **Figure 5A**). The disordered regions include residues 1-154, 157-168, 173-245, 261-264, and 285-304.

Importantly, this highly disordered nature seems to be crucial for the MBP functionality, as this protein is heavily decorated by diverse PTMs, such as multiple phosphorylation, citrullination, and methylation sites (see **Figure 5B**), many of which are isoform-specific [5]. Furthermore, MBP is predicted by the ANCHOR to contain 10 MoRFs: residues 1-16, 30-53, 59-67, 70-101, 116-139, 141-185, 198-209, 218-229, 240-255, and 258-292 (see **Figure 5B**). Therefore, 67.4% of the MBP residues are potentially involved in disorder-based interactions with binding partners.

This is reflected in the well-developed PPI network generated for human MBP by STRING (see **Figure 5C**), which contains 38 interactors and 112 interactions and is characterized by the average node degree of 5.89 and average local clustering coefficient of 0.791. The MBP-centered network has significantly more (PPI enrichment value of 3.33×10^{-16}) interactions than 46 interactions expected for a random set of proteins of the same size and degree distribution drawn from the genome. Proteins involved in the formation of this network are commonly participate in the following biological processes: Nervous system development (GO:0007399; $p = 2.08\times10^{-15}$), Central nervous system development (GO:0007417; $p = 5.74\times10^{-14}$), Myelination (GO:0042552; $p = 7.53\times10^{-11}$), Gliogenesis (GO:0042063; $p = 7.53\times10^{-11}$), and System development (GO:0048731; $p = 7.53\times10^{-11}$). The most enriched molecular functions of these proteins are N-terminal myristoylation domain binding (GO:0031997; p = 0.00028), MAP kinase kinase activity (GO:0004708; p = 0.00028), Adenylate cyclase activator activity (GO:0010856; p = 0.00040); Disordered domain specific binding (GO:0097718; p = 0.00068), and Structural constituent of myelin sheath (GO:0019911; p = 0.0011. Finally, these proteins are most commonly found in the following cellular components: Myelin sheath (GO:0043209; $p = 1.50\times10^{-10}$), Cell body (GO:0044297; $p = 9.55\times10^{-6}$), Synapse (GO:0045202; $p = 9.55\times10^{-6}$), Neuronal cell body (GO:0043025; $p = 1.88\times10^{-5}$), and Somatodendritic compartment (GO:0036477; p = 0.00015).

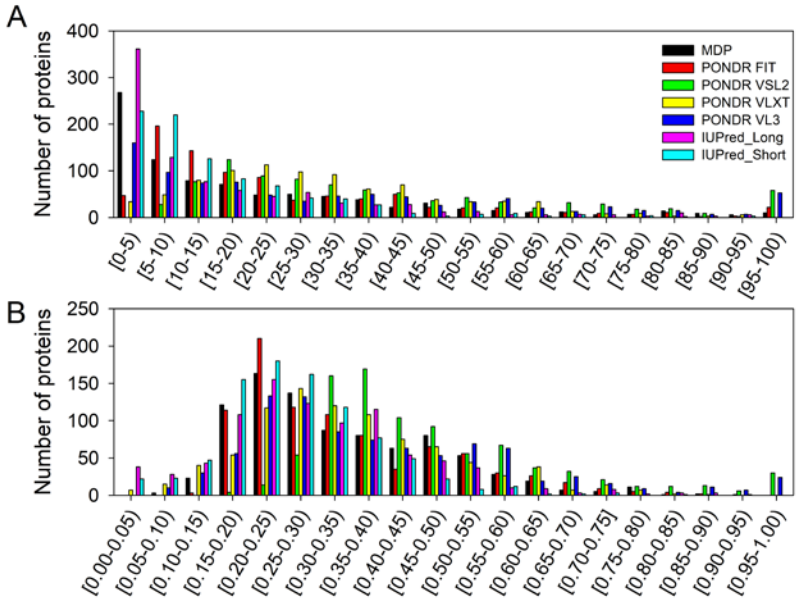


Figure 6. Distribution of the "Heslington brain proteins" based on their PPIDR (**A**) and MDS (**B**) values evaluated by various per-residue disorder predictors utilized in this study.

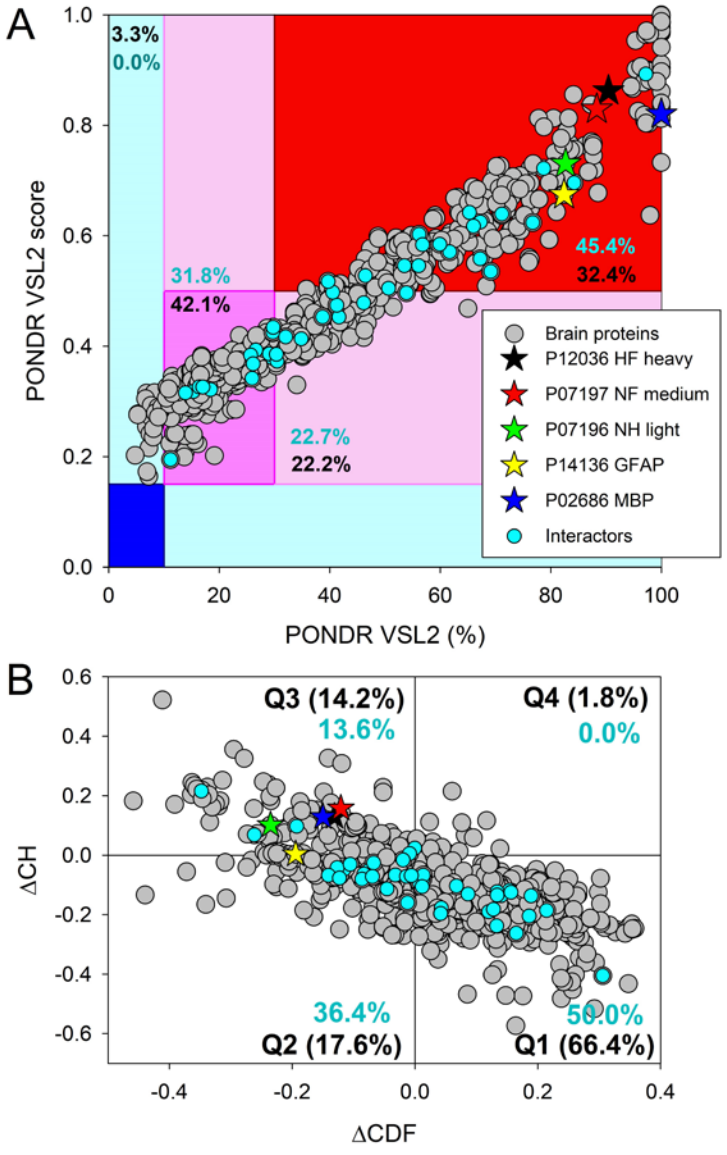
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

In line with the outputs of our bioinformatics analysis and in line with the available experimental data, **Figure 5D** shows that human MBP is predicted by the AlphaFold as a highly disordered protein.

3.2. Global intrinsic disorder analysis of the Heslington brain proteins

In total, the mass spectroscopy analysis of the Heslington brain samples identified 881 proteins, which, in this study, we subjected to the global intrinsic disorder analysis. Results of this analysis are summarized in **Figures 6** and **7**. **Figure 6** shows distribution of these proteins based on their PPIDR and MDS values and indicates that proteins in this dataset are characterized by noticeable levels of predicted intrinsic disorder, as evaluated by at least several of the per-residue disorder predictors used in this study.

Figure 7 indicates that the "Heslington brain proteins" are characterized by the noticeable levels of intrinsic disorder. **Figure 7A** represents the results of the classification of the disorder status of these proteins based on the outputs of the per-residue disorder predictor PONDR® VSL2. This classification is based the accepted in the field practice to group proteins based on their PPIDR values [21], where proteins with PPIDR < 10% are considered as ordered or mostly ordered; proteins with 10% ≤ PPIDR < 30% are considered as moderately disordered; whereas proteins with the PPIDR ≥ 30% are considered as highly disordered [21].



 doi:10.20944/preprints202209.0450.v1

Figure 7. Evaluation of the global disorder status of 881 human proteins found in the Heslington brain samples (gray circles). Corresponding data for five "main proteins" are shown by colored stars, whereas data for their interactors are shown by small cyan circles. **A**. PONDR® VSL2 output. PONDR® VSL2 score is the mean disorder score (MDS) for a query protein. PONDR® VSL2 (%) is a percent of the predicted disordered residues (PPDR) in a query protein; i.e., the percent of residues with the disorder scores above 0.5. Color blocks indicate regions in which proteins are mostly ordered (blue and light blue), moderately disordered (pink and light pink), or mostly disordered (red), as per accepted classification (see the text). If the two parameters agree, the corresponding part of background is dark (blue or pink), whereas light blue and light pink reflect areas in which only one of these criteria applies. **B**. Charge-hydropathy and cumulative distribution function (CH-CDF) plot. The Y-coordinate is calculated as the distance of the corresponding protein from the boundary in the CH plot, whereas the X-coordinate is calculated as the average distance of the corresponding protein's CDF curve from the CDF boundary. The quadrant that the protein is located determines its classification. Q1, protein predicted to be ordered by CH-plot and CDF. Q2, protein predicted to be ordered to by CH-plot and disordered by CDF-plot. Q3, protein predicted to be disordered by CH-plot and CDF. Q4, protein predicted to be disordered by CH-plot and ordered by CDF.

Since the MDS of a given protein is not directly related to its PPIDR value (e.g. theoretically, a protein with the PPIDR of 100% might have the MDS ranging from 0.5 to 1.0; whereas a protein with the PPIDR of 0% might have any MDS < 0.5), proteins can also be classified based on their corresponding MDS values, being annotated as highly ordered (MDS < 0.15), moderately disordered or flexible (MDS between 0.15 and 0.5) and highly disordered (MDS ≥ 0.5). Based on these classification criteria, none of the "Heslington brain proteins" was predicted as ordered by both MDS and PPIDR, and only 3.3% of these were predicted as mostly ordered based on their MDS values, with remaining proteins being either moderately or highly disordered. In fact, **Figure 7A** shows that 42.1% of these proteins were predicted as moderately disordered/flexible/containing noticeable IDRs based on both their MDS and PPIDR values. Additional 22.2% of the dataset were expected to be moderately disordered based on their MDS values, whereas 32.14% of the "Heslington brain proteins" are expected to be highly disordered, with 259 proteins (~29.3%) being predicted to have PPIDR ≥ 50% and MDS ≥ 0.5.

The global disorder status of proteins can be further evaluated using the ΔCH-ΔCDF plot (a tool that combines the outputs of two binary predictors (i.e., tools classifying proteins as mostly ordered or mostly disordered), charge-hydropathy (CH) plot and cumulative distribution function (CDF) plot, see *Materials and Methods* section). Due to the principle difference between the CH and CDF plot analyses, this approach allows classification of a query protein as mostly ordered, molten globule-like or hybrid, or highly disordered based on their position within the CH-CDF phase space. **Figure 7B** shows that 66.4% of the "Heslington brain proteins" are located within the quadrant Q1 (bottom right corner) that contains proteins predicted to be ordered by both predictors. On the other hand, 17.6% of these proteins are positioned within the quadrant Q2 (bottom left corner) that includes proteins predicted to be ordered/compact by the CH-plot and disordered by the CDF analysis (i.e., it contains either molten globular proteins, which are compact, but do not have unique 3D structures, or hybrid proteins containing high levels of ordered and disordered residues). Quadrant Q3 (top left corner) includes 14.2% of the "Heslington brain proteins", which are predicted as disordered by both predictors and therefore, are expected to be highly disordered and behave as native coils or native pre-molten globules in their unbound states. Finally, 1.8% of the Heslington brain proteins are found in quadrant Q4 (top right corner), being predicted as disordered by CH-plot and ordered by CDF analysis. Therefore, 33.6% of the "Heslington brain proteins" are located outside the quadrant Q1 and can be are considered as proteins with high disorder levels.

There is a reasonably good agreement between the outputs of the tools used here for the global disorder evaluation. For example, most of the 14.2% proteins found in the quadrant Q3 of the ΔCH-ΔCDF plot can be found within the set of ~12.5% proteins that

 doi:10.20944/preprints202209.0450.v1

have PPIDR ≥ 50% and MDS ≥ 0.5 and therefore predicted as very disordered. Similarly, the "red segment" in **Figure 7A** contains 32.14% of the analyzed dataset, and 33.6% of the "Heslington brain proteins" are located outside the quadrant Q1 in **Figure 7B**. Also, comparison of data shown in **Figure 7** suggest that many protein predicted as mostly ordered by the CH-CDF analysis in fact might contain noticeable levels of disordered residues.

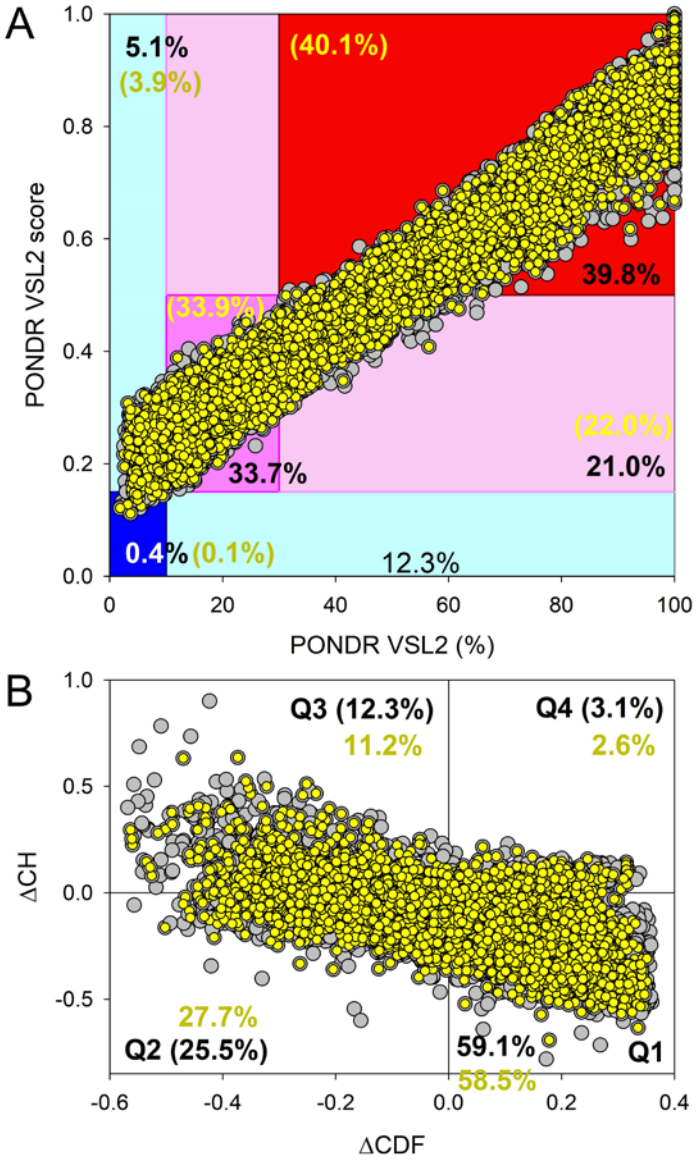


Figure 8. Evaluation of the global disorder status of 20,317 human proteins (gray circles) and 10,611 brain proteins (small yellow circles). **A**. PONDR® VSL2 output. **B**. Charge-hydropathy and cumulative distribution function (CH-CDF) plot. For explanations, see legend to **Figure 7**.

To place these data for the "Heslington brain proteins" in perspective, **Figure 8** represents the results of similar analyses of the entire human proteome (20,317 manually curated proteins). Comparison of the results shown in **Figure 7A** and **8A** suggest that the whole human proteome contains a bit more of highly disordered proteins as evidenced by the contents of the corresponding red segments, where one can find 32.4% and 39.8% of the "Heslington brain" and whole proteome proteins, respectively. Furthermore, 7,381 human proteins (i.e., 36.3% of the whole set) are predicted to have PPIDR ≥ 50% and MDS ≥ 0.5, which is slightly higher than the proportion of highly disordered proteins with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

PPIDR ≥ 50% and MDS ≥ 0.5 among the "Heslington brain proteins" (~29.3%). On the other hand, the whole human proteome contains a bit less of moderately disordered proteins and noticeably more mostly ordered proteins than the set of the "Heslington brain proteins". Comparison of **Figures 7B** and **8B** shows a rather different picture, where a set of the "Heslington brain proteins" contains a bit more of Q3 proteins than the whole human proteome (14.2% *vs*. 12.3%) and Q1 proteins (66.4% *vs*. 59.1%) and noticeably less Q2 proteins (17.6% *vs*. 25.5%).

Clearly, not all 20,317 human proteins are found in the brain. Therefore, at the next stage, we analyzed 10,611 manually curated human proteins, which contain the term "brain" either in their name or in the annotation. Results of these analyses are added to **Figure 8**. In comparison with brain-related proteins from the whole human proteome, the "Heslington brain proteins" shows 1.27- and 1.13-time increase in the content of quadrants Q3 and Q1, respectively and 1.57-time decrease in the proteins populating quadrant Q2.

3.3. Functional intrinsic disorder of proteins potentially interacting with "main proteins"

As explained in the Materials and Methods section, interactors for the 5 "main proteins" (NFH, NFM, NFL, GFAP, and MBP) were determined using APID. A list of the interactors that were also detected in the preserved Heslington brain using mass spectroscopy was then made. This list included 44 interactors with at least one of the 5 "main proteins". Results of the global intrinsic disorder analysis in these interactors are summarized in **Figure 7**. **Figure 7A** shows that 20 interactors (45.5%) are located within the red area, indicating that they are predicted as highly disordered proteins. Remaining 25 proteins are moderately disordered, being spread between the dark pink and light pink areas that contain 14 (31.8%) and 10 (22.7%) proteins, respectively. Since blue and cyan areas do not have any points, none of these proteins is predicted to be mostly ordered. As per CH-CDF analysis (**Figure 7B**), interactors are found in quadrants Q1 (22 proteins, 50%), Q2 (16 proteins, 36.4%), and Q3 (6 proteins 13.6%), supporting the idea that, in general, the set of interactors is noticeably more disordered than the proteins identified in the preserved Heslington brain (see **Figure 7**), or the whole human proteome and human brain proteins (see **Figure 8**).

Figure 9A represents the STRING-generated PPI network between the 5 main proteins and their 44 interactors. This network was generated using medium confidence level of 0.4 to insure maximal inclusion of interactors. The resulting network contains 48 proteins connected via 163 interactions. This network is characterized by the average node degree of 6.79 and the average local clustering coefficient of 0.489. Since the expected number of interactions for a random set of proteins of the same size and degree distribution drawn from the genome is 44, this network has significantly more interactions than expected (p-value $<10^{-16}$). **Figure 9B** shows PPI network centered at these proteins. Here, network includes 548 proteins linked by 15,492 interactions.

The PPI network shown in **Figure 9B** includes significantly (p-value $<10^{-16}$) more interactions than expected (8171) and is characterized by the average node degree of 56.5 and the average local clustering coefficient of 0.54. Protein included to this network preferentially participate in the following biological processes: Negative regulation of biological process (GO:0048519; $p = 7.02\times10^{-79}$), Symbiotic process (GO:0044403; $p = 3.76\times10^{-76}$), Viral process (GO:0016032; $p = 3.75\times10^{-76}$), Regulation of metabolic process (GO:0019222; $p = 4.05\times10^{-74}$), and Response to organic substance (GO:0010033; $p = 2.26\times10^{-73}$). The most enriched molecular functions of these proteins are Protein binding (GO:0005515; $p = 2.33\times10^{-60}$), Enzyme binding (GO:0019899; $p = 1.71\times10^{-50}$), Binding (GO:0005488; $p = 1.09\times10^{-39}$), RNA binding (GO:0003723; $p = 1.74\times10^{-38}$), and Beta-catenin binding (GO:0008013; $p = 5.07\times10^{-38}$), and they serve as cellular components in Protein-containing complex (GO:0032991; $p = 2.97\times10^{-95}$), Cytosol (GO:0005829; $p = 8.38\times10^{-88}$), Vesicle (GO:0031982; $p = 4.93\times10^{-63}$), Intracellular (organelle GO:0043229; $p = 5.81\times10^{-61}$), and Extracellular exosome (GO:0070062; $p = 6.55\times10^{-61}$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

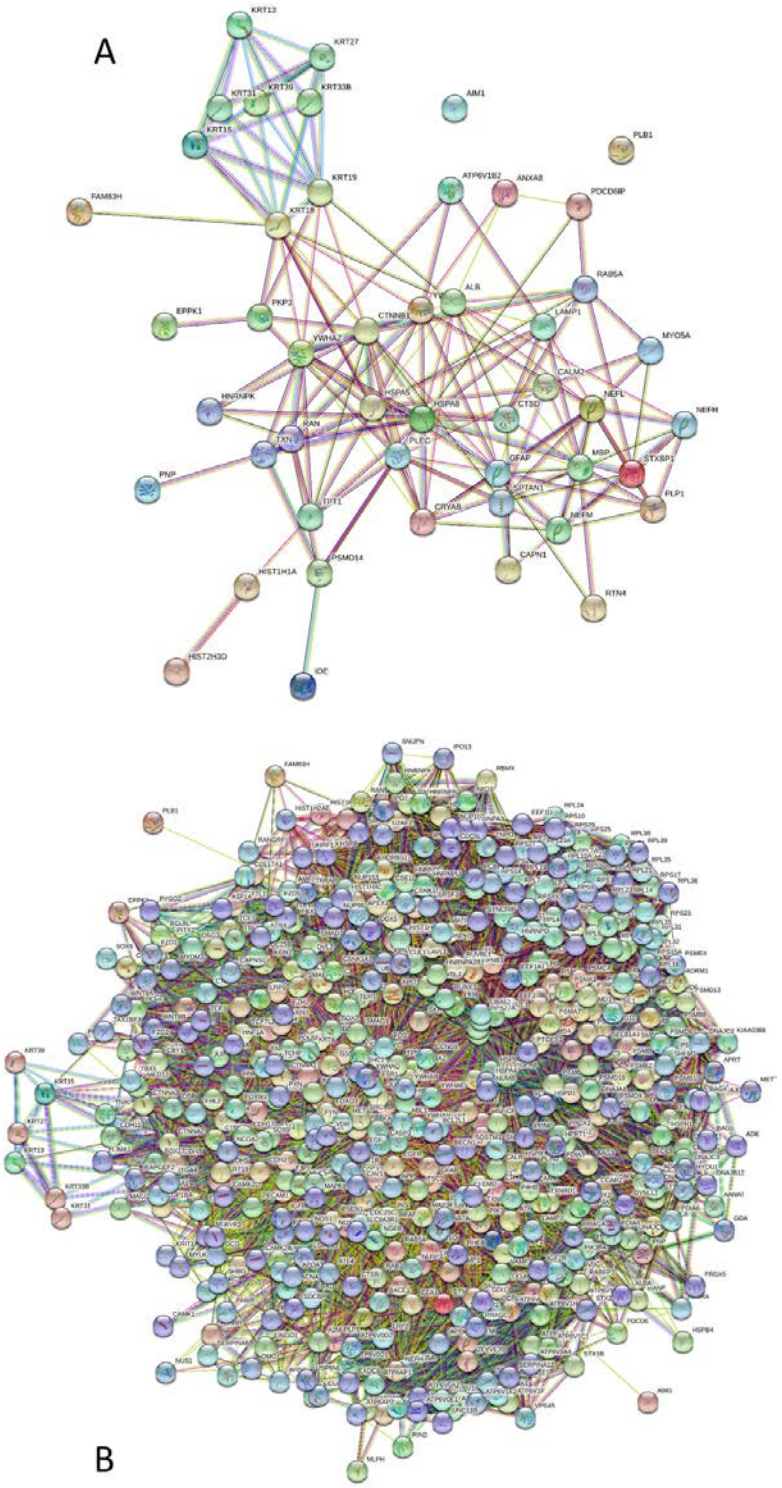


Figure 9. STRING-generated PPI network between NFs, GFAP, MBP and their 44 interactors found in the Heslington brain sample using APID (**A**), and a network that also includes a 1st shell of interactors of these proteins.

Uploading these main proteins and their interactors to DAVID and the use of the KEGG pathway database revealed that out of the 49 that were uploaded, only 39 had hits in the KEGG pathway database, and only the NFs had pathway hits out of the 5 main proteins. **Table 1** summarizes these results and shows that 6 pathways that involve the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2022 doi:10.20944/preprints202209.0450.v1

formation of aggregates appear to be associated with the interactors, with some of those also including NFs. These are neurodegeneration, Alzheimer's disease, ALS, fluid shear stress atherosclerosis, Parkinson's disease, and lipid atherosclerosis pathways.

Table 1. Potential pathways, according to the KEGG database, involving interactors detected in the brain and the main proteins. 49 proteins were uploaded to DAVID and only 39 had KEGG pathway hits.

Term	Genes	PValue	Count
hsa04915:Estrogen signaling pathway	KRT33B, HSPA8, KRT19, KRT18, KRT39, KRT27, KRT15, KRT13, CALM1, KRT31, CALM2, CTSD	1.78E-11	12
hsa05150:Staphylococcus aureus infection	KRT33B, KRT19, KRT18, KRT39, KRT27, KRT15, KRT13, KRT31	2.35E-07	8
hsa05022:Pathways of neurodegeneration - multiple diseases	PSMD14, HSPA5, NEFL, CTNNB1, NEFM, CAPN1, CALM1, CALM2, NEFH, RAB5A	2.59E-04	10
hsa05010:Alzheimer disease	PSMD14, CTNNB1, CAPN1, IDE, CALM1, CALM2, RTN4	0.008044994	7
hsa05152:Tuberculosis	LAMP1, CALM1, CALM2, CTSD, RAB5A	0.009433096	5
hsa04114:Oocyte meiosis	CALM1, CALM2, YWHAZ, YWHAG	0.022691054	4
hsa05014:Amyotrophic lateral sclerosis	PSMD14, HSPA5, NEFL, NEFM, NEFH, RAB5A	0.025753752	6
hsa05418:Fluid shear stress and atherosclerosis	CTNNB1, TXN, CALM1, CALM2	0.026464114	4
hsa05012:Parkinson disease	PSMD14, HSPA5, TXN, CALM1, CALM2	0.034238029	5
hsa04141:Protein processing in endoplasmic reticulum	HSPA8, HSPA5, CAPN1, CRYAB	0.044718031	4
hsa05417:Lipid and atherosclerosis	HSPA8, HSPA5, CALM1, CALM2	0.077697703	4
hsa04916:Melanogenesis	CTNNB1, CALM1, CALM2	0.080063685	3

We are reporting here the results of the comprehensive bioinformatics analysis of the prevalence of functional intrinsic disorder in five proteins (neurofilament proteins (neurofilament heavy (NFH), neurofilament middle (NFM), and neurofilament light (NFL)), glial fibrillary acidic protein (GFAP), and myelin basic protein (MBP)) that were established to play an important role in the preservation of the Heslington brain (which is at least 2,600-year-old). This analysis revealed that these "main" proteins are extensively disordered and four of them NFH, NFM, NFL, and GFAP are likely to undergo noticeable folding induced by the assembly of the neurofilaments. On the other hand, MBP, which is crucial for the maintaining the structural integrity of the myelin sheath, is found "in the major dense lines, electron-dense lamellae formed by the tight apposition of the cytoplasmic leaflets of the oligodendrocyte membrane" [61]. The ability of MBP to cause tight adhesion of the two lipid bilayers and also to act as a scaffold protein that binds many other proteins to the membrane is determined by the peculiarities of its amino acid sequence, where there are no specific membrane-binding domains and the basic residues distributed over its entire length rather than in a cluster [61]. Therefore, it seems that these disorder-based functional features of the "main protein" allows them to serve as long-lasting preservatives via formation of tightly packed intramolecular complexes. It is likely that the formation of these highly intertwined and densely packed assemblies makes these protein highly resistant to degradation and helps them and their partner to sustain for a long time (at least 2,600 years).

Comparison of these five proteins with the remaining proteins from the Heslington brain (as well as the whole human proteome or human brain proteins) revealed that NFH, NFM, NFL, GFAP, and MBP are positioned among the top 10% of the most disordered proteins in these datasets. Furthermore, we also established that many of the proteins known to interact with NFH, NFM, NFL, GFAP, and MBP are predicted to be highly disordered as well. Furthermore, our analysis revealed that the Heslington brain, which was preserved for more than 2,600 years, contains surprisingly high levels of highly disordered proteins. This finding is rather counterintuitive, as due to the well-known high accessibility of IDPs and IDRs to the proteolytic cleavage [9,74-78], one would expect almost complete elimination of the proteins with the noticeable disorder levels from the Heslington brain. The only reasonable explanation for these observations is the capability of IDRs/IDPs to act as a molecular mortar or cement that glue together various brain proteins and rigidify the resulting assemblies thereby generating highly stable matter. In other words, high disorder content in the Heslington brain proteins serves as another il-

lustration of one of the disorder-related paradoxes of protein universe, namely stability of instability, where "sturdiness of intrinsic disorder and its capability to "ignore" harsh conditions provides some interesting and important advantages to its carriers, at the molecular (e.g., the cell wall-anchored accumulation-associated protein playing a crucial role in intercellular adhesion within the biofilm of *Staphylococcus epidermidis*), supramolecular (e.g., protein complexes, biologic liquid-liquid phase transitions, and proteinaceous membrane-less organelles), and organismal levels (e.g., the … case of the microscopic animals, tardigrades, or water bears, that use intrinsically disordered proteins to survive desiccation)" [79].

Finally, although preservation of the brain tissue for more than 2,600 years seems to be a miracle, the Heslington brain is not the only ancient human brain tissue uncovered by the archeologists. In fact, excavations at Kanaljorden in Sweden uncovered ~8,000-year-old brain material inside human skulls that had received an underwater burial [80]. Although no proteomic analysis was conducted on the samples of the "Kanaljorden brain" as of yet, it is tempting to hypothesize that the noticeable fraction of the Kanaljorden brain proteins would have high levels of predicted intrinsic disorder as well. Future research is needed to verify this exciting hypothesis.

4. Conclusions

- Five main proteins that are assumed to be responsible for the preservation of the Heslington brain (which is at least 2.600-year-old), NFH, NFM, NFL, GFAP, and MBP are predicted to be highly disordered.
- 44 proteins from the Heslington brain, which are expected to interact with these five main proteins, are also predicted to have high levels of intrinsic disorder.
- Contrarily to the expected substantial (if not complete) elimination of the disordered proteins from the brain found inside a skull buried in a pit in Heslington, Yorkshire, England, many proteins in this Heslington brain are predicted to be highly disordered, with most of these proteins being expected to contain noticeable levels of intrinsic disorder.
- Intrinsic disorder of NFH, NFM, NFL, GFAP, and MBP and their interactors (in combination with other factors, such as the way in which the person was buried) might play a crucial role in preserving the Heslington brain by forming tightly folded brain protein aggregates.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Amino acid sequences of the Heslington brain proteins analyzed in this study.

Author Contributions: Conceptualization, V.N.U.; methodology, V.N.U.; validation, A.M. and V.N.U.; formal analysis, A.M. and V.N.U.; investigation, A.M. and V.N.U.; writing—original draft preparation, A.M. and V.N.U.; writing—review and editing, A.M. and V.N.U.; visualization, A.M. and V.N.U.; supervision, V.N.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article or supplementary material.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. O'Connor, S.; Ali, E.; Al-Sabah, S.; Anwar, D.; Bergström, E.; Brown, K.A.; Buckberry, J.; Buckley, S.; Collins, M.; Denton, J., et al. Exceptional preservation of a prehistoric human brain from Heslington, Yorkshire, UK. *J Archaeol Sci* **2011**, *38*,

1641-1654, doi:https://doi.org/10.1016/j.jas.2011.02.030.

2. Petzold, A.; Lu, C.H.; Groves, M.; Gobom, J.; Zetterberg, H.; Shaw, G.; O'Connor, S. Protein aggregate formation permits millennium-old brain preservation. *J R Soc Interface* **2020**, *17*, 20190775, doi:10.1098/rsif.2019.0775.
3. Yuan, A.; Rao, M.V.; Veeranna; Nixon, R.A. Neurofilaments and Neurofilament Proteins in Health and Disease. *Cold Spring Harb Perspect Biol* **2017**, *9*, doi:10.1101/cshperspect.a018309.
4. Eng, L.F.; Ghirnikar, R.S.; Lee, Y.L. Glial fibrillary acidic protein: GFAP-thirty-one years (1969-2000). *Neurochem Res* **2000**, *25*, 1439-1451, doi:10.1023/a:1007677003387.
5. Boggs, J.M. Myelin basic protein: a multifunctional protein. *Cell Mol Life Sci* **2006**, *63*, 1945-1961, doi:10.1007/s00018-006-6094-7.
6. Didonna, A.; Opal, P. The role of neurofilament aggregation in neurodegeneration: lessons from rare inherited neurological disorders. *Mol Neurodegener* **2019**, *14*, 19, doi:10.1186/s13024-019-0318-4.
7. Sihag, R.K.; Inagaki, M.; Yamaguchi, T.; Shea, T.B.; Pant, H.C. Role of phosphorylation on the structural dynamics and function of types III and IV intermediate filaments. *Exp Cell Res* **2007**, *313*, 2098-2109, doi:10.1016/j.yexcr.2007.04.010.
8. Uversky, V.N.; Oldfield, C.J.; Midic, U.; Xie, H.; Xue, B.; Vucetic, S.; Iakoucheva, L.M.; Obradovic, Z.; Dunker, A.K. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* **2009**, *10 Suppl 1*, S7, doi:10.1186/1471-2164-10-S1-S7.
9. Uversky, V.N.; Dunker, A.K. Understanding protein non-folding. *Biochim Biophys Acta* **2010**, *1804*, 1231-1264, doi:10.1016/j.bbapap.2010.01.017.
10. Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **2004**, *337*, 635-645, doi:10.1016/j.jmb.2004.02.002.
11. Uversky, V.N. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci* **2013**, *22*, 693-724, doi:10.1002/pro.2261.
12. UniProt Consortium, T. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **2021**, *49*, D480-D489, doi:10.1093/nar/gkaa1100.
13. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2018**, *46*, 2699, doi:10.1093/nar/gky092.
14. Meszaros, B.; Erdos, G.; Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **2018**, *46*, W329-W337, doi:10.1093/nar/gky384.
15. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **2005**, *61 Suppl 7*, 176-182, doi:10.1002/prot.20735.
16. Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A.K.; Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **2006**, *7*, 208, doi:10.1186/1471-2105-7-208.
17. Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C.J.; Dunker, A.K.; Obradovic, Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* **2005**, *3*, 35-60, doi:10.1142/s0219720005000886.
18. Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence complexity of disordered protein. *Proteins* **2001**, *42*, 38-48, doi:10.1002/1097-0134(20010101)42:1<38::aid-prot50>3.0.co;2-3.
19. Xue, B.; Dunbrack, R.L.; Williams, R.W.; Dunker, A.K.; Uversky, V.N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **2010**, *1804*, 996-1010, doi:10.1016/j.bbapap.2010.01.011.
20. Dayhoff, G.W.I.; Uversky, V.N. Rapid prediction and analysis of protein intrinsic disorder. *Protein Science* **2022**, In press.
21. Rajagopalan, K.; Mooney, S.M.; Parekh, N.; Getzenberg, R.H.; Kulkarni, P. A majority of the cancer/testis antigens are intrinsically disordered proteins. *J Cell Biochem* **2011**, *112*, 3256-3267, doi:10.1002/jcb.23252.

22. Uversky, V.N. Analyzing IDPs in interactomes. In *Intrinsically Disordered Proteins*, Kragelund, B.B., Skriver, K., Eds. Humana New York, NY, 2020; Vol. Methods in Molecular Biology, pp. 895-945.
23. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P., et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **2011**, *39*, D561-568, doi:10.1093/nar/gkq973.
24. Alonso-Lopez, D.; Campos-Laborie, F.J.; Gutierrez, M.A.; Lambourne, L.; Calderwood, M.A.; Vidal, M.; De Las Rivas, J. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (Oxford)* **2019**, *2019*, doi:10.1093/database/baz005.
25. Oates, M.E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M.J.; Xue, B.; Dosztanyi, Z.; Uversky, V.N.; Obradovic, Z.; Kurgan, L., et al. D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* **2013**, *41*, D508-516, doi:10.1093/nar/gks1226.
26. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415-427, doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7 [pii].
27. Oldfield, C.J.; Cheng, Y.; Cortese, M.S.; Brown, C.J.; Uversky, V.N.; Dunker, A.K. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **2005**, *44*, 1989-2000, doi:10.1021/bi047993o.
28. Mohan, A.; Sullivan, W.J., Jr.; Radivojac, P.; Dunker, A.K.; Uversky, V.N. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol Biosyst* **2008**, *4*, 328-340, doi:10.1039/b719168e.
29. Sun, X.; Xue, B.; Jones, W.T.; Rikkerink, E.; Dunker, A.K.; Uversky, V.N. A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development. *Plant Mol Biol* **2011**, *77*, 205-223, doi:10.1007/s11103-011-9803-z.
30. Xue, B.; Oldfield, C.J.; Van, Y.Y.; Dunker, A.K.; Uversky, V.N. Protein intrinsic disorder and induced pluripotent stem cells. *Mol Biosyst* **2012**, *8*, 134-150, doi:10.1039/c1mb05163f.
31. Huang, F.; Oldfield, C.; Meng, J.; Hsu, W.L.; Xue, B.; Uversky, V.N.; Romero, P.; Dunker, A.K. Subclassifying disordered proteins by the CH-CDF plot method. *Pac Symp Biocomput* **2012**, 128-139.
32. Dennis, G., Jr.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **2003**, *4*, P3.
33. Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**, *28*, 27-30, doi:10.1093/nar/28.1.27.
34. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589, doi:10.1038/s41586-021-03819-2.
35. Mukai, H.; Toshimori, M.; Shibata, H.; Kitagawa, M.; Shimakawa, M.; Miyahara, M.; Sunakawa, H.; Ono, Y. PKN associates and phosphorylates the head-rod domain of neurofilament protein. *J Biol Chem* **1996**, *271*, 9816-9822, doi:10.1074/jbc.271.16.9816.
36. Andreeva, A.; Howorth, D.; Brenner, S.E.; Hubbard, T.J.; Chothia, C.; Murzin, A.G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **2004**, *32*, D226-229, doi:10.1093/nar/gkh039.
37. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **1995**, *247*, 536-540, doi:10.1006/jmbi.1995.0159.
38. de Lima Morais, D.A.; Fang, H.; Rackham, O.J.; Wilson, D.; Pethica, R.; Chothia, C.; Gough, J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res* **2011**, *39*, D427-434, doi:10.1093/nar/gkq1130.

39. Meszaros, B.; Simon, I.; Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* **2009**, *5*, e1000376, doi:10.1371/journal.pcbi.1000376.
40. Hornbeck, P.V.; Kornhauser, J.M.; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **2012**, *40*, D261-270, doi:10.1093/nar/gkr1122.
41. Janmey, P.A.; Leterrier, J.-F.; Herrmann, H. Assembly and structure of neurofilaments. *Current opinion in colloid & interface science* **2003**, *8*, 40-47.
42. Carter, J.; Gragerov, A.; Konvicka, K.; Elder, G.; Weinstein, H.; Lazzarini, R.A. Neurofilament (NF) assembly; divergent characteristics of human and rodent NF-L subunits. *J Biol Chem* **1998**, *273*, 5101-5108, doi:10.1074/jbc.273.9.5101.
43. Carpenter, D.A.; Ip, W. Neurofilament triplet protein interactions: evidence for the preferred formation of NF-L-containing dimers and a putative function for the end domains. *J Cell Sci* **1996**, *109 (Pt 10)*, 2493-2498, doi:10.1242/jcs.109.10.2493.
44. Peysselon, F.; Xue, B.; Uversky, V.N.; Ricard-Blum, S. Intrinsic disorder of the extracellular matrix. *Mol Biosyst* **2011**, *7*, 3353-3365, doi:10.1039/c1mb05316g.
45. Szappanos, B.; Suveges, D.; Nyitray, L.; Perczel, A.; Gaspari, Z. Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett* **2010**, *584*, 1623-1627, doi:10.1016/j.febslet.2010.03.026.
46. Heins, S.; Wong, P.C.; Muller, S.; Goldie, K.; Cleveland, D.W.; Aebi, U. The rod domain of NF-L determines neurofilament architecture, whereas the end domains specify filament assembly and network formation. *J Cell Biol* **1993**, *123*, 1517-1533, doi:10.1083/jcb.123.6.1517.
47. Uceda-Castro, R.; van Asperen, J.V.; Vennin, C.; Sluijs, J.A.; van Bodegraven, E.J.; Margarido, A.S.; Robe, P.A.J.; van Rheenen, J.; Hol, E.M. GFAP splice variants fine-tune glioma cell invasion and tumour dynamics by modulating migration persistence. *Sci Rep* **2022**, *12*, 424, doi:10.1038/s41598-021-04127-5.
48. Bignami, A.; Dahl, D. Astrocyte-specific protein and neuroglial differentiation. An immunofluorescence study with antibodies to the glial fibrillary acidic protein. *J Comp Neurol* **1974**, *153*, 27-38, doi:10.1002/cne.901530104.
49. Bignami, A.; Eng, L.F.; Dahl, D.; Uyeda, C.T. Localization of the glial fibrillary acidic protein in astrocytes by immunofluorescence. *Brain Res* **1972**, *43*, 429-435, doi:10.1016/0006-8993(72)90398-8.
50. van Asperen, J.V.; Robe, P.; Hol, E.M. GFAP Alternative Splicing and the Relevance for Disease - A Focus on Diffuse Gliomas. *ASN Neuro* **2022**, *14*, 17590914221102065, doi:10.1177/17590914221102065.
51. Geisler, N.; Weber, K. Amino acid sequence data on glial fibrillary acidic protein (GFA); implications for the subdivision of intermediate filaments into epithelial and non-epithelial members. *EMBO J* **1983**, *2*, 2059-2063, doi:10.1002/j.1460-2075.1983.tb01700.x.
52. Middeldorp, J.; Boer, K.; Sluijs, J.A.; De Filippis, L.; Encha-Razavi, F.; Vescovi, A.L.; Swaab, D.F.; Aronica, E.; Hol, E.M. GFAPdelta in radial glia and subventricular zone progenitors in the developing human cortex. *Development* **2010**, *137*, 313-321, doi:10.1242/dev.041632.
53. van Bodegraven, E.J.; van Asperen, J.V.; Robe, P.A.J.; Hol, E.M. Importance of GFAP isoform-specific analyses in astrocytoma. *Glia* **2019**, *67*, 1417-1433, doi:10.1002/glia.23594.
54. van Bodegraven, E.J.; van Asperen, J.V.; Sluijs, J.A.; van Deursen, C.B.J.; van Strien, M.E.; Stassen, O.; Robe, P.A.J.; Hol, E.M. GFAP alternative splicing regulates glioma cell-ECM interaction in a DUSP4-dependent manner. *FASEB J* **2019**, *33*, 12941-12959, doi:10.1096/fj.201900916R.
55. Kawajiri, A.; Yasui, Y.; Goto, H.; Tatsuka, M.; Takahashi, M.; Nagata, K.; Inagaki, M. Functional significance of the specific sites phosphorylated in desmin at cleavage furrow: Aurora-B may phosphorylate and regulate type III intermediate

filaments during cytokinesis coordinatedly with Rho-kinase. *Mol Biol Cell* **2003**, *14*, 1489-1500, doi:10.1091/mbc.e02-09-0612.

56. Kosako, H.; Amano, M.; Yanagida, M.; Tanabe, K.; Nishi, Y.; Kaibuchi, K.; Inagaki, M. Phosphorylation of glial fibrillary acidic protein at the same sites by cleavage furrow kinase and Rho-associated kinase. *J Biol Chem* **1997**, *272*, 10333-10336, doi:10.1074/jbc.272.16.10333.
57. Jin, Z.; Fu, Z.; Yang, J.; Troncosco, J.; Everett, A.D.; Van Eyk, J.E. Identification and characterization of citrulline-modified brain proteins by combining HCD and CID fragmentation. *Proteomics* **2013**, *13*, 2682-2691, doi:10.1002/pmic.201300064.
58. Etienne-Manneville, S. Cytoplasmic Intermediate Filaments in Cell Biology. *Annu Rev Cell Dev Biol* **2018**, *34*, 1-28, doi:10.1146/annurev-cellbio-100617-062534.
59. Harauz, G.; Ishiyama, N.; Hill, C.M.; Bates, I.R.; Libich, D.S.; Fares, C. Myelin basic protein-diverse conformational states of an intrinsically unstructured protein and its roles in myelin assembly and multiple sclerosis. *Micron* **2004**, *35*, 503-542, doi:10.1016/j.micron.2004.04.005.
60. Martinsen, V.; Kursula, P. Multiple sclerosis and myelin basic protein: insights into protein disorder and disease. *Amino Acids* **2022**, *54*, 99-109, doi:10.1007/s00726-021-03111-7.
61. Harauz, G.; Ladizhansky, V.; Boggs, J.M. Structural polymorphism and multifunctionality of myelin basic protein. *Biochemistry* **2009**, *48*, 8094-8104, doi:10.1021/bi901005f.
62. Givogri, M.I.; Bongarzone, E.R.; Schonmann, V.; Campagnoni, A.T. Expression and regulation of golli products of myelin basic protein gene during in vitro development of oligodendrocytes. *J Neurosci Res* **2001**, *66*, 679-690, doi:10.1002/jnr.10031.
63. Vassall, K.A.; Bamm, V.V.; Harauz, G. MyelStones: the executive roles of myelin basic protein in myelin assembly and destabilization in multiple sclerosis. *Biochem J* **2015**, *472*, 17-32, doi:10.1042/BJ20150710.
64. de Ferra, F.; Engh, H.; Hudson, L.; Kamholz, J.; Puckett, C.; Molineaux, S.; Lazzarini, R.A. Alternative splicing accounts for the four forms of myelin basic protein. *Cell* **1985**, *43*, 721-727, doi:10.1016/0092-8674(85)90245-4.
65. Kamholz, J.; Toffenetti, J.; Lazzarini, R.A. Organization and expression of the human myelin basic protein gene. *J Neurosci Res* **1988**, *21*, 62-70, doi:10.1002/jnr.490210110.
66. Barbarese, E.; Carson, J.H.; Braun, P.E. Accumulation of the four myelin basic proteins in mouse brain during development. *J Neurochem* **1978**, *31*, 779-782, doi:10.1111/j.1471-4159.1978.tb00110.x.
67. Kim, J.; Mastronardi, F.; Wood, D.; Lubman, D.; Zand, R.; Moscarello, M. Multiple sclerosis: an important role for post-translational modifications of myelin basic protein in pathogenesis. *Molecular & Cellular Proteomics* **2003**, *2*, 453-462.
68. Moscarello, M.A. Myelin Basic Protein, The "Executive" Molecule of the Myelin Membrane. In *Cell Biology and Pathology of Myelin: Evolving Biological Concepts and Therapeutic Approaches*, Juurlink, B.H.J., Devon, R.M., Doucette, J.R., Nazarali, A.J., Schreyer, D.J., Verge, V.M.K., Eds. Plenum Press: New York, 1997; pp. 13-25.
69. Chao, L.P.; Einstein, E.R. Physical properties of the bovine encephalitogenic protein; molecular weight and conformation. *J Neurochem* **1970**, *17*, 1121-1132, doi:10.1111/j.1471-4159.1970.tb03360.x.
70. Majava, V.; Wang, C.; Myllykoski, M.; Kangas, S.M.; Kang, S.U.; Hayashi, N.; Baumgartel, P.; Heape, A.M.; Lubec, G.; Kursula, P. Structural analysis of the complex between calmodulin and full-length myelin basic protein, an intrinsically disordered molecule. *Amino Acids* **2010**, *39*, 59-71, doi:10.1007/s00726-009-0364-2.
71. Krigbaum, W.R.; Hsu, T.S. Molecular conformation of bovine A1 basic protein, a coiling macromolecule in aqueous solution. *Biochemistry* **1975**, *14*, 2542-2546, doi:10.1021/bi00682a038.
72. Libich, D.S.; Ahmed, M.A.; Zhong, L.; Bamm, V.V.; Ladizhansky, V.; Harauz, G. Fuzzy complexes of myelin basic protein: NMR spectroscopic investigations of a polymorphic organizational linker of the central nervous system. *Biochem Cell Biol* **2010**, *88*, 143-155, doi:10.1139/o09-123.
73. Sedzik, J.; Kirschner, D.A. Is myelin basic protein crystallizable? *Neurochem Res* **1992**, *17*, 157-166, doi:10.1007/BF00966794.

74. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W., et al. Intrinsically disordered protein. *J Mol Graph Model* **2001**, *19*, 26-59.
75. Iakoucheva, L.M.; Kimzey, A.L.; Masselon, C.D.; Bruce, J.E.; Garner, E.C.; Brown, C.J.; Dunker, A.K.; Smith, R.D.; Ackerman, E.J. Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci* **2001**, *10*, 560-571.
76. Fontana, A.; Zambonin, M.; Polverino de Laureto, P.; De Filippis, V.; Clementi, A.; Scaramella, E. Probing the conformational state of apomyoglobin by limited proteolysis. *J Mol Biol* **1997**, *266*, 223-230.
77. Fontana, A.; Polverino de Laureto, P.; De Phillips, V. Molecular aspects of proteolysis of globular proteins. In *Protein Stability and Stabilization*, van den Tweel, W., Harder, A., Buitelear, M., Eds. Elsevier Science: Amsterdam, The Netherlands, 1993; pp. 101-110.
78. Fontana, A.; Fassina, G.; Vita, C.; Dalzoppo, D.; Zamai, M.; Zambonin, M. Correlation between sites of limited proteolysis and segmental mobility in thermolysin. *Biochemistry* **1986**, *25*, 1847-1851.
79. Uversky, V.N. Paradoxes and wonders of intrinsic disorder: Stability of instability. *Intrinsically Disord Proteins* **2017**, *5*, e1327757, doi:10.1080/21690707.2017.1327757.
80. Gummesson, S.; Hallgren, F.; Kjellström, A. Keep your head high: skulls on stakes and cranial trauma in Mesolithic Sweden. *antiquity* **2018**, *92*, 74-90.