Investigating the Conformational Ensembles of Intrinsically Disordered IA_3s Proteins

James R. Watt; Tianyan Li; Dr. Gail E. Fanucci, Ph.D.

Abstract

Intrinsically disordered proteins (IDPs) make up 52% of proteins in the human proteome and are linked to disorders such as Alzheimer's, Parkinson's, CVD, and some cancers. This work examines the structural properties of the intrinsically disordered IA_3 *Saccharomyces pastorianus* and *Saccharomyces arboricola* proteins with Circular Dichroism (CD) spectroscopy to establish a correlation between their sequence parameters and the CD-identified structure. CD spectra across varying concentrations of 2,2,2-Trifluoroethanol (TFE) revealed distinct, structurally linked spectral features. IA_3 *S. pastorianus* exhibited consistent α-helical structures across all tested TFE volumes, and IA_3 *S. arboricola* exhibited primarily α-helical structures with minor deviations at lower TFE concentrations. The BeStSel Single Spectrum Analysis tool was used to find the quantitative structural compositions from the CD data to highlight spectral characteristics representative of α-helical content. These findings provide insights into the structural composition of IA_3, laying the groundwork for subsequent studies into their functional and structural mechanisms and broader IDP implications.

1 Introduction

Intrinsically disordered proteins (IDPs) are a unique category of proteins that, under standard physiological conditions, inherently lack a well-defined secondary structure and never exhibit a tertiary structure[1,2,3]. This means that these proteins while being able to form some small structures, also known as conformations, along their polypeptide backbone, can never form a solid and stable 3^d structural assembly of these smaller structures. Additionally, they most often consist, either in part or entirely, of regions of total disorder, or a complete lack of structure[3]. Intrinsically disordered regions (IDRs) describe those strands of disordered residues along proteins that are known as partially ordered[4]. IDPs and IDRs are both rich in highly polar and charged amino acids: Arginine, Glutamic acid, Glutamine, Lysine, Proline, and Serine, which are known as the disorder-promoting residues[3]. Additionally, they contain relatively few hydrophobic residues: Cystine, Isoleucine, Leucine, Phenylalanine, Tryptophan, Tyrosine, and Val which are known as the order-premoting residues[3]. This study investigates boundaries of the correlation be-

tween these disorder-promoting residues, and the structural ensembles that a given protein exhibits.

In classical proteins, the structure is thought to be critically correlated to its function, similar to a lock and key mechanism. However, for IDPs and IDRs, this paradigm doesn't apply due to their inherent lack of structure. This gap in understanding has led to the development of a disorder-ensemble-function paradigm. In this, researchers are working to understand how the disorder within IDPs and IDRs influences their structure in order to eventually be able to connect disorder to function. The function of IDPs and IDRs is theorized to be structurally correlated but in a non-traditional way. For example, in some protein-receptor-enzyme processes, the unique flexibility of IDRs along some proteins is theorized to help facilitate the conformational transitions between the conformation required to bind to a receptor and the conformation required to bind to a modifying enzyme. To understand how this disorder is truly influencing function, researchers are first working to understand and classify the nature of the disorder itself and connect it to structure.

IDPs and IDRs are evolutionarily persistent[4], holding a range of functions across many biological domains. For example, in the human proteome, IDPs and IDRs make up 52% of all proteins[5]. This persistence is taken to mean that they likely hold some advantage over other proteins[4]. There are many theories for what makes them evolutionarily advantageous, however, this avenue is still widely uncertain and, for the most part, unknown. IDPs and IDRs have been linked with many human diseases such as Alzheimer's (AD), Parkinson's (PD), dementia, Huntinton's, amyloidosis, and some cancers[23]. Because of this, large amounts of research resources are currently being utilized to expand our understanding of IDPs and IDRs.

IDPs and proteins with IDRs aggregation and mutation characteristics have been the most significant link to neurodegenerative and cancerous diseases[27,30]. For AD, PD, and dementia, numerous previous studies have identified the aggregate lewy bodies of α-synuclien, an IDP, as a significant trait and potentially is the cause behind PD and dementias neurodegenerative effects[28,26,27]. Aggregation is theorized to be caused by the misfolding, Post-Translational modification, or aberrant protein-protein interactions of proteins[26,29]. Neurodegenerative diseases share the characteristic insoluble aggregate of misfolded or disordered proteins[26]. In cancer, the p53 tumor impression protein mutants are correlated with more than 50% of human cancers[27]. Understanding the structural mechanism of IDPs, especially those involved in these diseases, is a significant point of interest in intrinsic disorder research because their structure is seemingly directly involved in many human diseases[26].

Classical proteins are chains of amino acid residues characterized by their primary, secondary, tertiary, and quaternary structure[1]. The primary structure of a protein is the amino acid sequence that makes up the polypeptide chain and the secondary structure is the categorization of shapes, or conformations, on this backbone - commonly α-helicies and β-sheets[2]. Tertiary structure describes the arrangement of these conformational ensembles in space. that comprise the polypeptide chain[2]. The quaternary structure describes the spatial coordination of the amino acid subunits in proteins that are made up of more than one polypeptide chains[2].

IA_3 is a 68 amino acid IDP naturally found in the yeast *Saccharomyces* genus and in its biological role it acts as a yeast aspartic proteinase A (YPrA) inhibitor[11,16]. Because proteinases are a crucial part of many biological and pathological processes, there is great interest in understanding more about the pro-

teinase inhibitors to utilize them in biomedical and biotechnical applications[6,7,8].

IA_3 has the N-terminal domain (NTD) from residues 2-32 and C-terminal domain (CTD) from residues 33-68[616]. IA_3 has highly conserved NTD sequences, expect for amino acids 10 and 32 which can vary slightly, while their CTDs, sequence-wise, are very different (Figure 2.).

When IA_3 binds to YPrA, the NTD resolves into a α-helix while the C-terminal domain CTD stays unsolved(Figure 1.)[6,12]. The NTD α-helix binds with the YPrA and is responsible for the inhibitory behavior[612]. 2,2,2-Triflouroethanol can be used to induce the same structure in IA_3 that it takes when binding to YPrA[6,9]. This allows us to use approaches like Circular Dichroism (CD) to identify the structures present. IA_3 from *S. Cerevisiae* has been extensively studied in previous research and they solidified the existence of solely α-helical structures[6,14,16]. Previous studies also noted that TFE has its maximum induced structure when 23% of the sample is TFE by volume[6912]. In this study, we aimed to identify the structures present in TFE-structure induction using methodological approaches from the overexpressing and analyzing of IA_3 *Saccharomyces Cerevisiae* on *Saccharomyces pastorianus* and *Saccharomyces arboricola*.

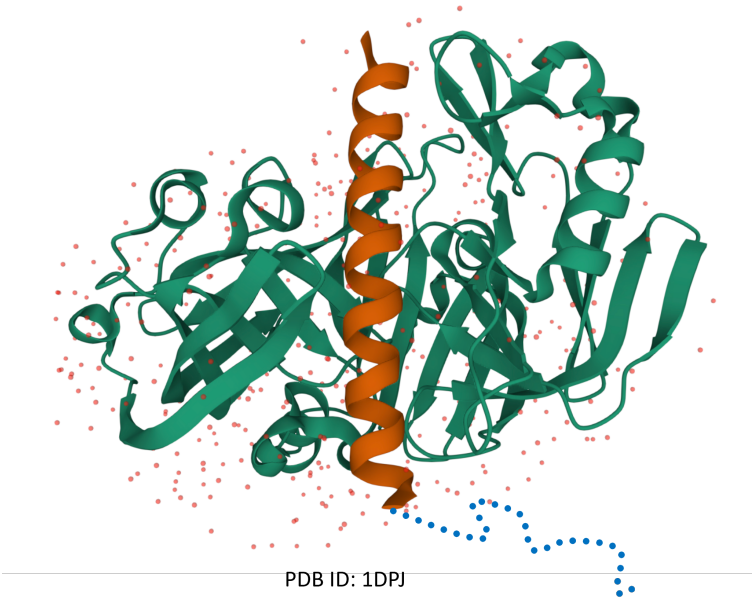


Figure 1: IA_3 Binding to YPrA with resolved N-terminus and unsolved C-terminus[5].

Species	1 — 32 — 68
S. pastorianus	MNTDQ QKVSE IFQSS KEKLQ GDAKV VSDAF KE MAN KDKNS DSNGD -SSHKP DYQEQ –YNK LKGAV KREE
S. arboricola	MNTDQ QKVSD IFQSS KEKLQ GDAKV VSDAF KE MAK KDKGG KADNI SDKD- KPDYQ EQYNK LKQSA HKKE

Figure 2: IA_3 amino acid sequences (red highlighting denotes the C terminus)

2 Methods

2.1 DNA Transformation

To produce the target proteins, IA_3 *S. pastorianus* and *S. arboricola*, we used the cellular machinery of Escherichia Coli (E. Coli) bacteria to transcribe and translate the DNA encoding them. To begin this process, the proteins of interest we first acquired commercially available DNA encoding our proteins and then inserted them into E. Coli bacteria cells. Commercial Pet22b(+) plasmids with the genes of interest *S. pastorianus* and *S. Arboriola* was purchased from Genscript. Then 6 μL of these valid encoded DNA plasmids were introduced to approximately 6μL of commercially bought BL21(DE3) competent Escherichia Coli(E. Coli) cells and then incubated in ice for 25 minutes. Then to insert the plasmids into the cells the plasmids and cells were heat shocked at 42°C for 45 seconds then immediately after placed in ice for 2 minutes. To allow the cells with our DNA in them to grow and replicate the samples were then added to 300μL of Luria Broth (LB) which is a mixture of nutrients designed to facilitate cell growth. Then to speed up cell growth in the LB media the samples were incubated for 1 hour at 37°C.

To allow the cells to build colonies agar plates infused with ampicillin (AMP) were streaked with 180μL of the cells. The AMP in the plates will kill all E. Coli cells lacking the DNA plasmids encoding the target proteins because those genes also code for AMP resistance. Then to allow these colonies to grow the plates were incubated at 37°C upside down for 12 hours. After the cell colonies had been incubated and grown, the cells were inoculated from the plates and transferred into 1 L of LB media containing 1:1000 AMP to grow further. The OD600 was then measured as an indicator of the maturity of the cell culture while being incubated at 37°C in a shaker to continue the cell growth until the OD600 reached 0.6.

2.2 Protein Overexpression

To get the E. Coli cells now containing the DNA encoding our target proteins to express this protein we adapted the overexpression optimization protocol developed by Dana Francis and Rebecca Page[24]. These methods were previously adapted and shown to be effective on overexpressing IA_3 *S. Cerevisiae*[6]. In this study, we applied these methods to IA_3 *S. pastorianus* and *S. arboricola*.

Once the cells reach an optical density of 0.6 or greater at 600nm, over-expression was triggered with the addition of 1 mL Isopropyl β-D-1-thiogalactopyranoside(IPTG) which utilizes an allolactose mimicking mechanism to begin the protein overexpression process. To allow the cells time to produce a significant amount of our protein, the samples were incubated at 16°C overnight.

With cells containing our target protein inside them, we then physically separated the cells from the solution and all other debris in the sample. To do this the samples were first centrifuged at 4°C for 15 minutes at 6,000 rpm to separate the cells containing the target protein from the supernatant which is discarded. This centrifugal separation was done twice more to separate more of the unwanted debris. For this, the pellets were then resuspended in a 30 mL, pH 8.0 buffer solution of 50 mM NaH_2PO_4 and 300 mM NaCl. The samples were then centrifuged again at 4,400 rpm for 20 minutes to separate our cells in a pellet from the waste supernatant. Then the cells were separated from excess debris and contained inside them the target proteins. The samples were then stored at -20°C until we were ready to continue.

To prevent the E. Coli's natural cleav-

ing enzymes from cleaving our target proteins once they're released from the E. Coli cells, the frozen samples were then resuspended in the 30 mL of the same buffer solution and 70 μL of a cocktail set VII-millipore protease inhibitor was added.

To break the cell membrane and release our target proteins along with all other cell contents, the samples were sonicated 3 times for 2 minutes each, cycling the sonicator on for 5 seconds and then off for 5 seconds throughout the 2-minute intervals. Following sonication, the samples were centrifuged at 4°C for 30 minutes at 12,000 rpm which separates our proteins into the supernatant from the debris which pellets and is discarded. The samples were then placed in a boiling water bath for 5 minutes to denature the very large, undesired proteins. After boiling, the samples were centrifuged again at 4°C for 30 minutes at 12,000 rpm to spin out the denatured proteins and some other leftover debris in the pellet which was discarded and the supernatant containing our target protein was kept.

2.3 Protein Purification

To purify our samples to contain only our target protein in the buffer we purified the proteins from any of the remaining debris with an ATKA PRIME Fast Protein Liquid Chromatography(FPLC) machine. We utilized the built-in UV spectrometer and conductivity meter in the ATKA PRIME to track the progression of our purification at 280 nm (Figure 3. & 4.). For purification, 3 buffers of 50 mM NaH_2PO_4 (phosphate) and 300 mM NaCl were made in 1 L nanopure water respectively. One buffer is pH corrected to pH 8.0, one is corrected to pH 6.3, and, lastly, one is corrected to pH 4.0 each by adding HCL or NaOH till the pH read by a pH meter is the desired pH. Before purification was started, the chelating column was loaded with Nickel Sulfate, which was used as it traps the 6xHis-Tag on our target proteins at pH 8.0 and pH 6.3, allowing other proteins and debris to be washed away in chromatography. With the column inserted into the FPLC, the pH 8.0 buffer is run through with no sample added to establish a UV and conductivity read-out baseline (Figure 6 & 7). Once the baseline was established, the sample was pumped into the FPLC which caused the UV and conductivity to spike. Once the UV and conductivity returned to baseline the buffer was switched to the pH 6.0 one and the process was repeated. Then the pH 4.0 buffer is added and the spectra are allowed to settle as chromatography completes. This is repeated for the other samples independently and the FPLC is cleaned in between cycles. This left a final sample of approximately pure target protein sample suspended in the pH 4.0 buffer solution.

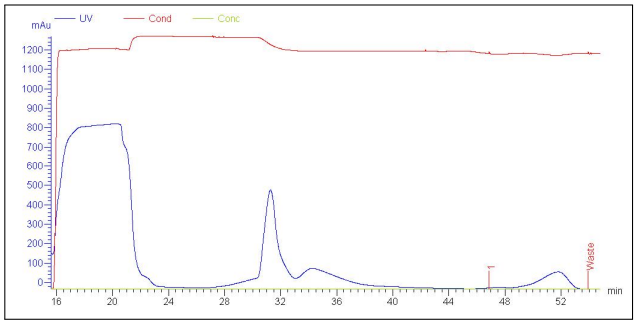


Figure 3: Chromatogram Produced during *S. pastorianus* purification

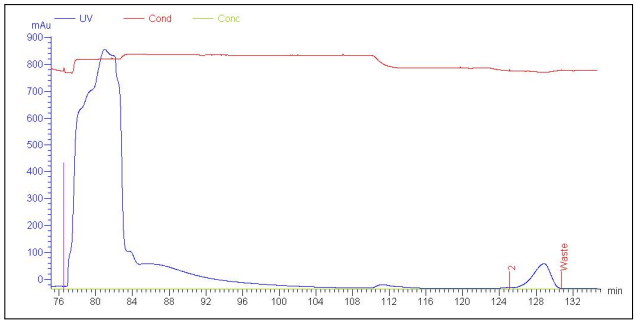


Figure 4: Chromatogram produced during *S. arboricola* purification

Following purification, dialysis was used to concentrate the samples so that we could use UV-vis spectroscopy to estimate their concentrations. The samples were placed in an Ultrafilter membrane and then placed in nanopure water and mixed with a stir bar for approximately 6 hours, changing the nanopure water every 1.5 hours to use the equilibrium pressure to replace much of the buffer and have the samples suspended in pH 7.0.

UV-vis spectroscopy was then used to measure the absorbance of the purified and concentrated samples to calculate the concentration of the samples by using Beer's law ($A = \varepsilon cl$) with known extinction coefficients (ε) calculated using BeStSel for 214nm and the measured path length of our cuvettes.

2.4 Preparing CD Samples

To survey the structures present in IA_3 *S. pastorianus* and *S. arboricola* Circular Dichroism (CD) spectroscopy was used. When surveying chiral molecules with the CD spectroscopy, such as the IA_3 protein, the different conformational ensembles produce different absorption, or more specifically ellipticity difference, spectra which can be analyzed based on their spectra shape to then identify these conformations[6,10,17,19]. To analyze this data, the BeStSel Single Spectrum Analysis tool was used to determine the relative amounts of those different conformations[17]. In this study, we performed CD Spectroscopy and BeStSel analysis on IA_3 *S. pastorianus* and *S. arboricola* in the 170 nm - 260 nm UV spectrum, to investigate the local conformational ensembles that make up the different activated IA_3 proteins.

For CD spectroscopy, we prepared 3 500 μL samples for each protein, each at a different 2,2,2-Triflouroethanol (TFE) volume. The 500 μL final samples comprise a 150 μL TFE system and a 350 μL protein system. The three samples for each protein are each at a different volumetric ratio of TFE: 0%, 15%, and 30%. This means for the TFE systems we made a 150 μL TFE - 0 μL NANOwater system (30% TFE), a 75 μL TFE - 75 μL NANOwater system (15% TFE), and a 0 μL TFE - 150 μL NANOwater system (0% TFE). Each of these was combined with our protein systems which are 350 μL solutions of 10 μM protein, 5mM phosphate (NaH_2PO_4), and 21.5 mM Sodium Chloride (NaCl) at pH 7.0.

2.5 Circular Dichroism and BeStSel Analysis

With our prepared samples, we ran CD spectroscopy at 22.5 °C on the samples 5 times each to return their average CD spectra in the 170nm to 260nm range. This data was then input into BeStSel's single-spectrum analysis tool to return the relative compositions of secondary structural elements.

2.6 CIDER and Intrinsic Parameters

With these secondary structures, we aimed to look for correlations between the intrinsic properties of the amino acid sequence and the structure or lack thereof. We cross-analyze our structural data with the charge patterning, the fraction of charged residues, the net charge per residue, and the Das-Pappu Phase.

CIDER is a tool being developed by the Pappu Lab which is used to calculate parameters of IDPs and IDRs[18]. This study uses CIDER to calculate the Charge Patterning (κ), Fraction of Charged Residues (FCR), and Net Charge Per Residue (NCPR). That allows for IA_3 *S. pastorianus* and *S. arboricola* to be classified on the Das-Pappu Phase Plot and to investigate how charge distribution impacts conformational ensembles[18].

Charge Patterning (κ) describes the segregation and grouping of like and unlike charges, a.k.a the square deviation, along a protein with the parameter κ which falls in the range $0 \leq \kappa \leq 1$[20]. For example, $\kappa = 1$ would indicate more segregation of unlike charges (i.e. $+ + + + + - - - - -$) and $\kappa = 0$ would indicate a homogeneous distribution of unlike charges(i.e. $+ - + - + - + - + -$)[18]. In highly charged IDPs and IDRs, when the charges are more segregated with a higher κ value, they are more likely to be folded because of the Coulombic attractions between the different regions of the protein or protein segment and vice-versa[21].

The Fraction of Charged Residues (FCR) describes the ratio of positively and negatively charged residues in a protein[18]. FCR is broken up into $f+$ and $f-$ which are the parameters for the ratio of positively charged residues to total residues and the of negatively charged residues to total residues respectively[18].

The Net Charge Per Residue (NCPR) is the net charge across the protein divided by the total number of residues. NCPR is related to FCR by the relationship, $NCPR = |(f+) - (f-)|$[18]. The NCPR parameter is yet another tool to categorize and indicate the structure of IDPs[18].

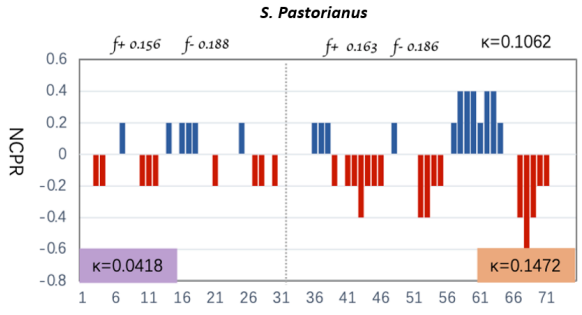


Figure 5: NCPR, FCR, and κ for*S. pastorianus*[22]

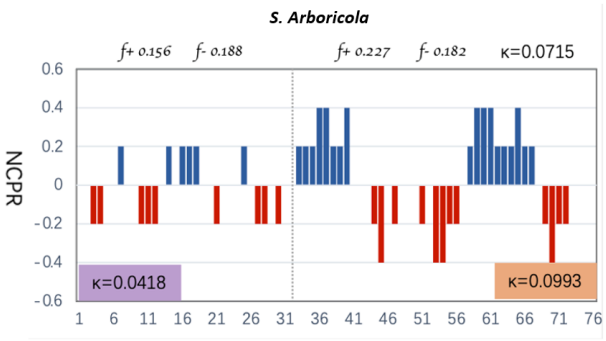


Figure 6: NCPR, FCR, and κ for *S.arboricola*[22]

localCIDER was used to calculate the Charge Patterning, FCR, and NCPR for IA_3 *S. pastorianus* and *S. arboricola*(Figure 5. & 6.)

With localCIDER we could use the FCR parameter to plot the IDPs on the Das-Pappu Phase Plot which categorizes them based on their FCR and properties(Figure 7.)[18]. R1 contains proteins that have a relatively low ratio of charged residues to uncharged residues and present as weak polyampholytes and polyelectrolytes forming globule and tadpole structures[18]. R2 IDPs are categorized as Janus sequences, meaning they have contrasting regions and take either a globule or expanded structure[18]. R3 IDPs are strong polyelectrolytes forming coils, hairpins, and chimeras[18]. R4 IDPs are highly negative strong polyelectrolytes that form swollen coils[18]. R5 IDPs are highly positive strong polyelectrolytes that also form swollen coils[18]. R-categorization allows us to plot new less studied proteins based on their computationally calculated FCR and get an idea of their structure. *S. pastorianus* and *S. Arboriola* are characterized as either R2 or R3 and lying near the R2-R3 transition. *S. pastorianus* lies within R2 and *S. arboricola* lies in R3. It is then theorized that *S. pastorianus* and *S. Arboriola* all fall into the structural predictions of R2 or R3 respectively.

The BeStSel analyzed data we got from

the CD analysis was compared to these parameters in an attempt to build on the disorder-structure paradigm by correlating the CD data to these intrinsic parameters.

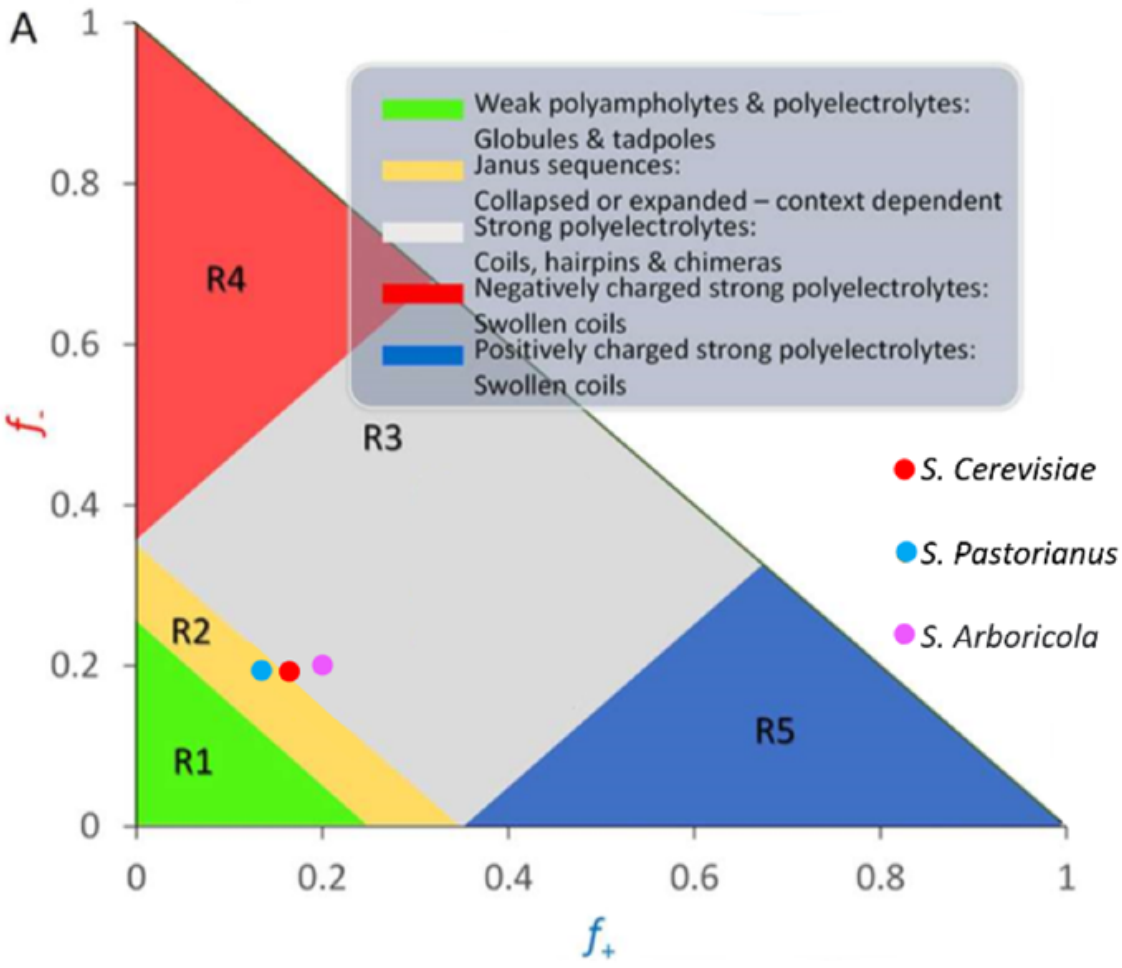


Figure 7: *S. Cerevisia*, *S. pastorianus*, and *S. Arboriola* plotted on the Das-Pappu Phase Plot[13,18,22]

3 Results

3.1 Circular Dichroism Spectra

Circular Dichroism was performed on IA_3 *S. pastorianus* and *S. arboricola* and returned the average CD spectra for each (Figure 9, Figure 10, Table 1, and Table 2). For all of the sample spectra, between ∼170 nm and ∼200 nm, the left-handed absorption is consistently greater than the right-handed absorption yielding the constant CD value. From ∼200 nm to ∼240 nm the right-handed ab-

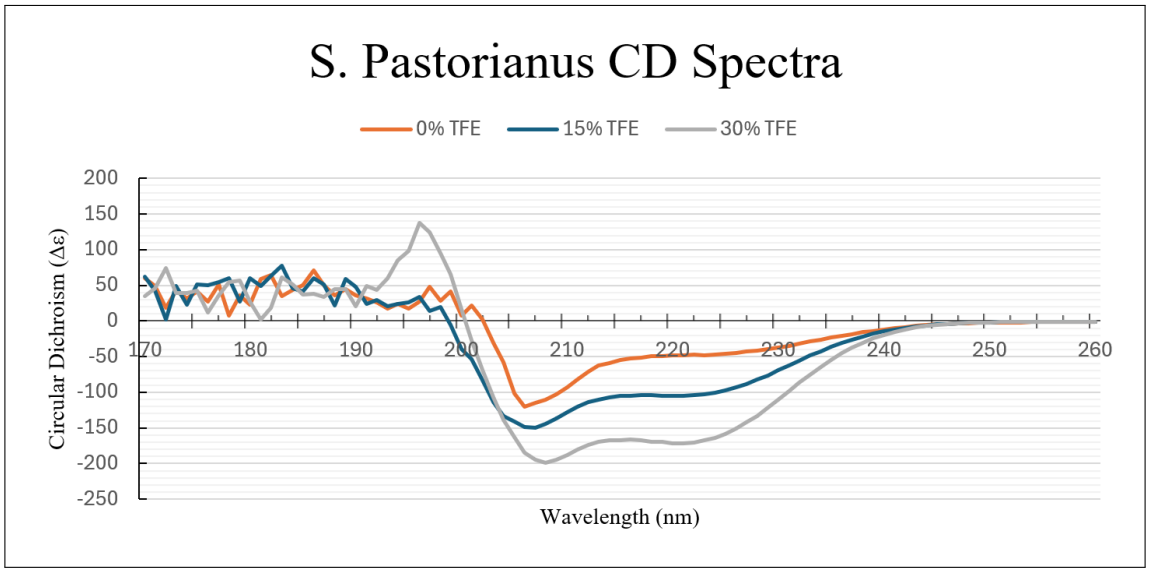


Figure 8: *S. pastorianus* Circular Dichroism Spectra at 0%,15%, and 30% TFE

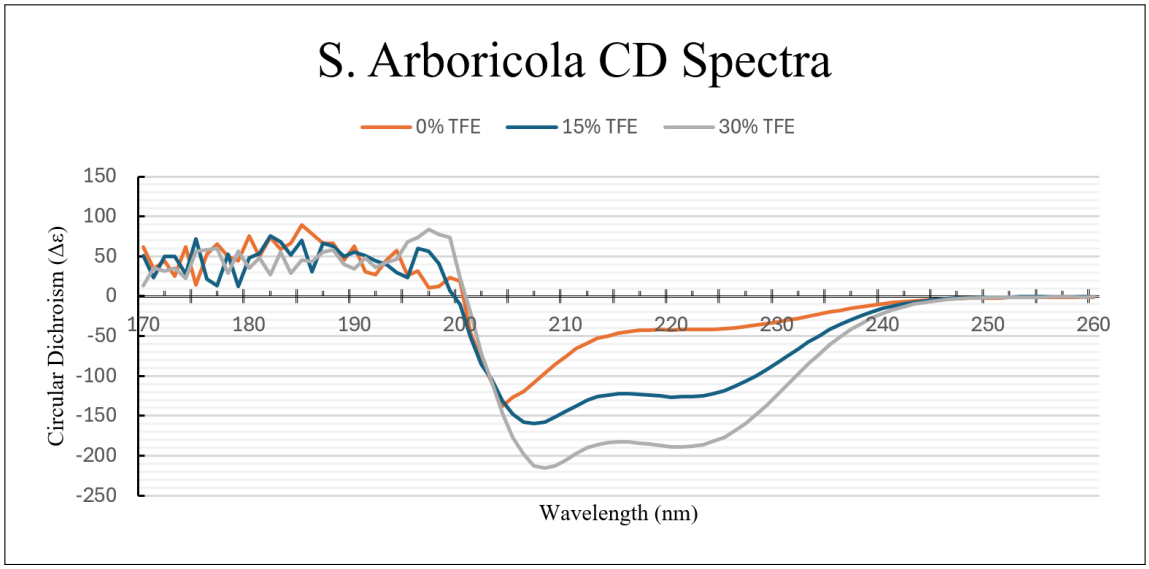


Figure 9: *S. arboricola* Circular Dichroism Spectra at 0%,15%, and 30% TFE

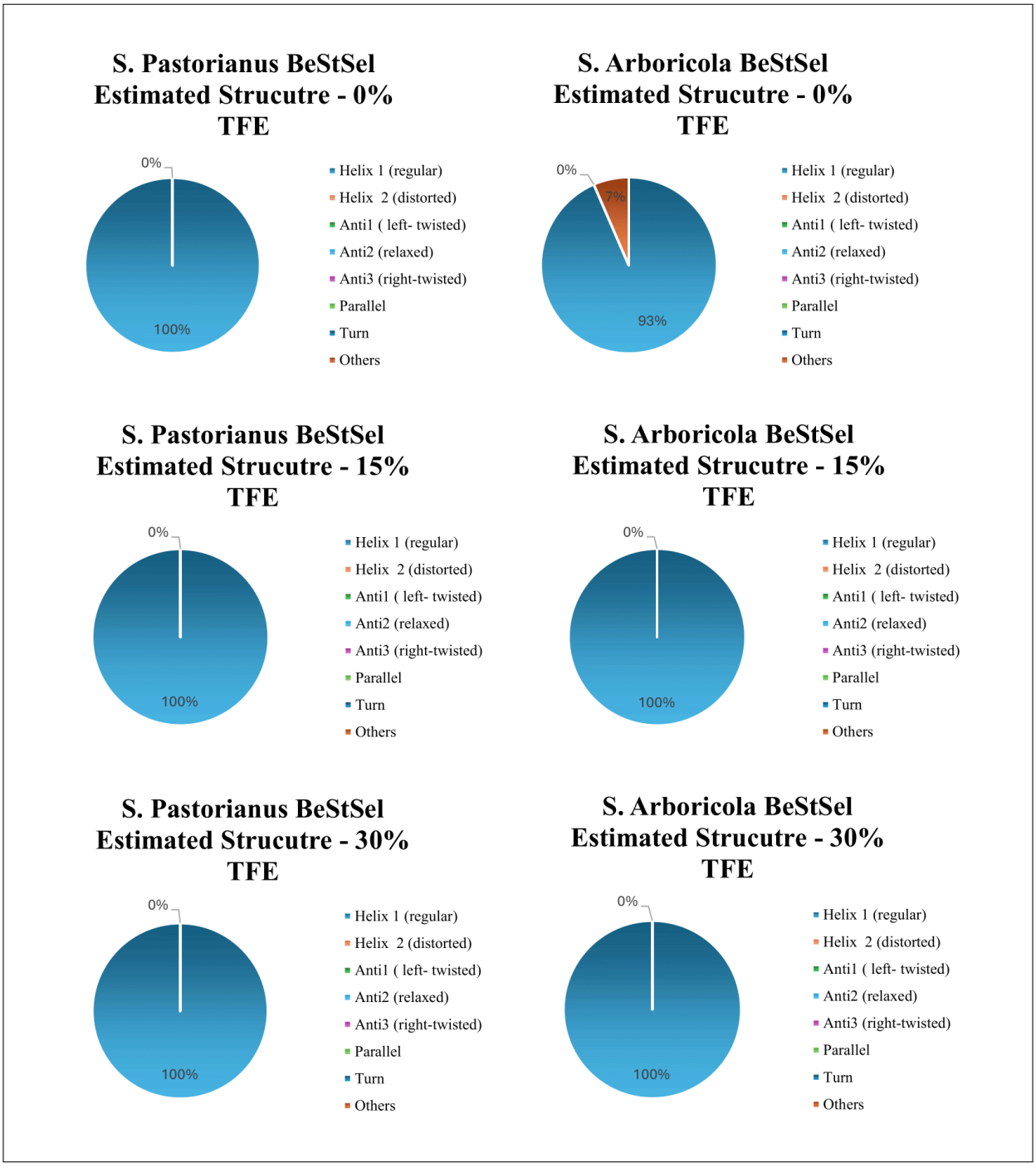


Figure 10: BeStSel structural analysis of CD spectra

S. arboricola							
Wavelength	0%	15%	30%	Wavelength	0%	15%	30%
170	61.49345	50.32565	13.35509	215	-46.6717	-122.583	-182.373
171	34.41792	22.94004	35.13133	216	-44.538	-122.245	-182.902
172	44.1761	49.97959	31.525	217	-42.8797	-122.815	-184.112
173	25.08519	49.9752	35.35534	218	-42.3979	-123.919	-185.654
174	61.902	28.16095	22.76739	219	-41.7968	-125.258	-187.482
175	14.03244	71.76833	56.65524	220	-42.4031	-126.426	-188.928
176	52.19134	21.21008	57.94	221	-42.0972	-126.014	-189.404
177	65.42251	12.85606	60.3482	222	-42.0473	-125.953	-188.521
178	49.68924	52.96245	28.51367	223	-41.9993	-124.634	-186.508
179	44.09528	12.74381	56.53915	224	-41.3754	-121.965	-182.105
180	75.564	48.32767	34.86061	225	-40.778	-118.149	-176.677
181	49.65286	53.37995	48.09665	226	-39.7554	-113.304	-169.112
182	73.6466	75.1987	27.19804	227	-38.4338	-106.755	-159.689
183	58.61146	67.7419	55.17342	228	-36.5876	-99.8889	-148.816
184	66.0496	51.58843	28.40266	229	-34.6576	-91.6476	-136.768
185	88.83938	69.75123	44.91164	230	-32.7186	-83.4259	-124.045
186	78.11276	30.26148	43.71554	231	-29.8638	-74.8824	-110.648
187	66.67144	66.463	55.4644	232	-27.5842	-66.3879	-97.6945
188	66.68754	63.0622	58.25376	233	-24.86	-57.5059	-84.4212
189	45.0988	49.86448	40.2028	234	-22.3088	-49.5963	-72.5847
190	62.53274	55.6432	33.8583	235	-19.8581	-42.0192	-61.1012
191	30.79874	50.38432	47.63378	236	-17.536	-35.1899	-50.9356
192	27.09058	44.1891	36.48838	237	-15.2614	-29.4266	-41.9054
193	45.41382	39.7603	42.20084	238	-13.1645	-23.9927	-34.02
194	56.90458	29.6391	46.11588	239	-11.3114	-19.4385	-27.3532
195	26.20781	23.66563	68.46828	240	-9.63147	-15.5029	-21.7385
196	31.8526	60.09318	73.5492	241	-8.19264	-12.2451	-17.0046
197	10.17668	56.5482	83.43248	242	-6.94072	-9.7465	-13.0626
198	12.41103	40.51468	77.6985	243	-5.72655	-7.48947	-10.0878
199	23.22818	7.186925	74.07614	244	-4.91092	-5.95789	-7.70855
200	18.90353	-10.9298	22.4931	245	-4.21289	-4.4176	-5.90778
201	-45.5409	-50.9009	-19.6998	246	-3.5737	-3.61788	-4.51205
202	-79.6649	-85.5298	-71.4313	247	-3.07102	-2.617	-3.47762
203	-107.186	-103.851	-106.279	248	-2.67297	-2.20228	-2.69757
204	-137.965	-131.228	-147.2	249	-2.32149	-1.75136	-2.14769
205	-127.028	-147.618	-176.816	250	-2.00754	-1.4288	-1.78341
206	-119.888	-157.92	-198.174	251	-1.94043	-1.12518	-1.41073
207	-108.593	-159.901	-212.609	252	-1.74374	-1.055	-1.33589
208	-96.2131	-158.068	-215.589	253	-1.71771	-0.9566	-1.159
209	-85.5253	-151.226	-213.023	254	-1.52632	-0.93124	-1.06051
210	-75.4337	-144.339	-205.7	255	-1.51906	-0.81598	-1.02515
211	-65.1453	-137.58	-197.499	256	-1.49245	-0.77062	-0.93676
212	-58.7833	-130.895	-190.086	257	-1.43731	-0.73176	-0.91973
213	-52.9894	-125.914	-185.902	258	-1.44231	-0.73973	-0.93037
214	-49.4798	-124.018	-183.358	259	-1.32156	-0.65443	-0.98226

Table 1: Averaged CD Data for IA_3 S. arboricola

S. pastorianus							
Wavelength	0%	15%	30%	Wavelength	0%	15%	30%
170	59.52946	62.37358	34.35	215	-54.896	-105.332	-166.823
171	48.72755	42.368	45.55304	216	-52.4843	-104.715	-166.45
172	18.41556	2.232752	73.6031	217	-50.9449	-104.114	-167.124
173	42.14774	48.51212	39.00892	218	-49.3862	-104.283	-169.051
174	31.85276	22.8161	39.0041	219	-49.1278	-104.615	-169.916
175	40.86275	51.4313	42.88442	220	-47.935	-105.145	-171.242
176	26.8812	50.41617	12.2937	221	-47.6633	-104.981	-171.11
177	50.77857	54.54426	36.18352	222	-47.23	-104.325	-170.12
178	7.87979	60.10604	54.00715	223	-47.7338	-102.68	-167.571
179	35.35821	26.88362	56.34011	224	-46.8476	-100.389	-163.971
180	22.91802	60.43914	27.14138	225	-45.715	-97.3024	-158.42
181	59.36964	49.0879	1.750818	226	-45.2333	-93.3857	-151.237
182	64.80026	62.81238	18.9541	227	-43.2465	-88.226	-142.608
183	35.15067	77.1204	60.66974	228	-41.59	-82.4157	-133.166
184	43.68768	46.17622	51.54286	229	-39.7262	-76.108	-121.878
185	49.62516	41.02192	36.80798	230	-37.2751	-69.1959	-110.58
186	70.78828	59.82564	37.81466	231	-34.6576	-62.1641	-98.5269
187	50.12031	51.72262	33.57308	232	-32.0527	-55.5737	-86.9489
188	36.38588	21.96498	44.33724	233	-29.049	-48.6707	-75.1836
189	44.18394	59.19058	44.69314	234	-26.4846	-42.3231	-64.741
190	35.46368	47.50286	20.3847	235	-23.6017	-36.2648	-54.4659
191	31.68716	23.40146	49.32938	236	-20.805	-30.8638	-45.439
192	26.20394	29.68878	43.53032	237	-18.434	-25.9824	-37.4301
193	17.61982	20.96017	60.44746	238	-15.8929	-21.4792	-30.38
194	23.73522	23.34461	85.28682	239	-13.8457	-17.7264	-24.3487
195	17.65605	25.86005	97.783	240	-11.6648	-14.4555	-19.3104
196	26.94206	33.35362	137.4462	241	-9.99234	-11.6992	-15.2088
197	48.3104	14.27558	123.9784	242	-8.40472	-9.45607	-11.9125
198	28.13216	19.07155	95.27688	243	-7.18365	-7.61161	-9.16606
199	41.50514	-5.79247	65.08266	244	-6.04587	-6.15993	-7.15196
200	7.512994	-39.0182	16.75842	245	-5.03784	-4.99914	-5.55897
201	21.25012	-53.4178	-27.0596	246	-4.37546	-4.03784	-4.21888
202	2.465412	-83.296	-69.3121	247	-3.58544	-3.28405	-3.34353
203	-31.0719	-113.884	-107.362	248	-3.17254	-2.69363	-2.61473
204	-58.302	-133.036	-138.358	249	-2.84434	-2.29119	-2.01114
205	-101.479	-140.554	-162.987	250	-2.50618	-2.00046	-1.83954
206	-119.801	-148.937	-184.875	251	-2.17299	-1.71661	-1.52543
207	-114.8	-149.5	-194.73	252	-2.09808	-1.51659	-1.24341
208	-110.47	-144.161	-198.657	253	-1.94703	-1.45051	-1.22776
209	-102.457	-137.171	-195.105	254	-1.70124	-1.49664	-1.1293
210	-93.4337	-127.998	-187.727	255	-1.5496	-1.30097	-1.06163
211	-82.2586	-120.138	-180.777	256	-1.67158	-1.29298	-1.00361
212	-71.4493	-113.889	-173.758	257	-1.54468	-1.24331	-0.95327
213	-62.7608	-110.454	-169.795	258	-1.4893	-1.21312	-0.80763
214	-58.9026	-106.972	-167.214	259	-1.459	-1.24035	-0.95696

Table 2: Averaged CD Data for IA_3 S. pastorianus

sorption grew to be greater than left-handed absorption which is shown in the negative CD values. Between ∼200 nm and ∼240 nm the intensity of the difference between the right-handed and left-handed absorption grew with greater TFE present. Then from ∼240 nm to ∼260 nm the directional absorption splitting stopped.

3.2 BeStSel Structural Analysis

The CD data was then analyzed using the Single-Spectrum analysis tool developed by BeStSel (Figure 11.). This resulted in IA_3 *S. pastorianus* only exhibiting α-helicies across all TFE volumes and IA_3 *S. arboricola*, at 0% TFE exhibiting 93.5% α-helix and 6.5% other structures, and at 15% and 30% TFE exhibiting entirely α-helix.

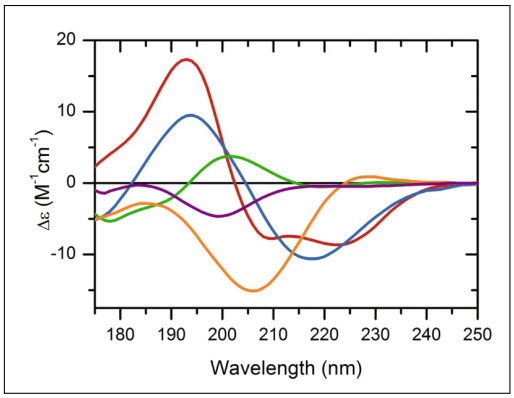
4 Discussion

With solely the BeStSel-analyzed CD data, we were not able to identify the correlation or lack thereof between the entirely α-helical structure observed in the TFE activated IA_3 *S. pastorianus* or *S. arboricola* and the charge patterning, FCR, NCPR, or the Das-Pappu phase. With the nature of BeStSel-analyzed CD data, we can only observe the estimated amounts of each structure present across the entire protein. Without further investigations into the positions of these structures across the protein and how the structures are arranged in 3^d space, it is difficult to make any conclusions about how the charge patterning, the FCR, the NCPR, and the Das-Pappu phase impact that structure.

For IA_3 *S. pastorianus* we observed an entirely α-helical across the different levels of TFE activation. However, in IA_3 *S. arboricola* when no TFE was present, only 93.5% of the structures were α-helicies, and 6.5% were identified as other structures. This is an unexpected result for an IA_3 protein and a subsequent study would likely need to be done to determine whether this is a result of an error or if IA_3 *S. arboricola* takes a not entirely α-helical structure when no TFE is present. Some possible points of error could've been low CD resolution, BeStSel misinterpretation, sample contamination, etc.

In the CD spectra for both IA_3 *S. pastorianus* and *S. arboricola* across the TFE volumes, they all matched the spectra characteristics of α-helical structures. At 200 nm across all CD spectra, we can observe the α-helix characteristic drop in CD value indicating the right-handed ellipticity grows larger than the left-handed ellipticity. Additionally, there are visible local minima at 207 nm and 225 nm which is expected in the CD spectra for α-helicies. Finally, we can observe the termination of the left-hand and right-hand splitting indicated by the CD value returning to zero.

Between the difference TFE volumes, 0%, 15%, and 30%, we observed the increasing amplitude of the CD spectra. This was because when more TFE is present, up until 23% at which all proteins are solved, more proteins solve into the α-helix structure, hence more proteins contribute to the CD effect thus being seen as the larger amplitude.



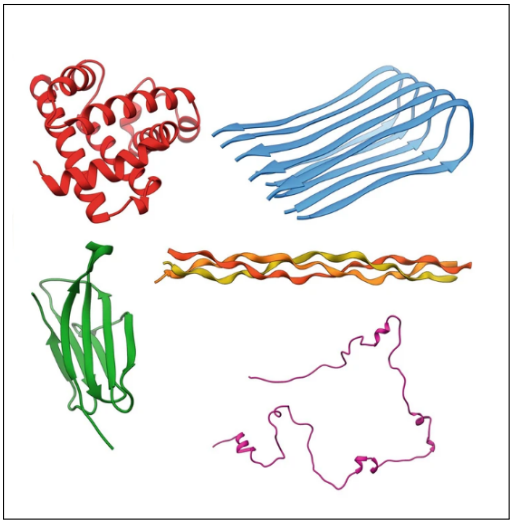


Figure 11: *Circular Dichroism secondary structure spectra*[17]

To correlate the intrinsic parameters of *S. Pastorius* and *S. arboricola* more data is needed. A possible next step is to perform sit-directed mutagenesis on the differing amino acids in the NTD of each protein to redo CD spectroscopy and analyze how the bound CTD impacts the conformational composition and CD spectra with identical NTDs. Additionally, IA_3 *S. pastorianus* and *S. arboricola* could be cleaved at the linkage between the NTD and CTD to perform CD spectroscopy on them independently. This would allow for the structure of each to be identified which could indicate a correlation between the linking of the NTD and CTD and the structure that they present. Additionally, with site-directed spin labeling, CW EPR could be used to identify the local tumbling volumes along the proteins which could reveal the locations of the CD-identified structures along the protein. Additionally, by studying more IDPs or proteins with IDRs we would amass a larger set of data in which correlations between intrinsic parameters and structure may be more easily identifiable.

5 Conclusions

From this study, we have seen that the methods adapted from *S. Cerevisiae* IA_3 transformation, overexpression, and purification can be applied to IA_3 *S. pastorianus* and *S. arboricola* to successfully transform, overexpress, and purify them. This is an advantageous application because till now IA_3 *S. pastorianus* and *S. arboricola* had entirely unknown structures and undeveloped methods, thus having the optimized methods for *S. Cerevisiae* will streamline further research on IA_3 *S. pastorianus* and *S. arboricola* IA_3.

Additionally, we have shown that IA_3 *S. pastorianus* exhibits only α-helices across the tested TFE volumed and that IA_3 *S. arboricola* exhibits almost entirely α-helical structure across TFE volumes, though more research is required to support these results for IA_3 *S. arboricola*.

In closing, this study contributes the overall structural compositions of IA_3 *S. pastorianus* and *S. arboricola* under TFE semi-

saturated and saturated conditions, allowing for more in-depth studies to be done in the future that may get us closer to understanding and preventing human diseases in which IDPs have been linked.

6 Acknowledgements

I would like to extend my sincerest thanks to Dr. Gail Fanucci and Tianyan Li for their unwavering support and mentorship throughout this research process.

I'd also like to give a huge thanks to Danae Perry, Ayan Desai, Siddarth Ranganath, Rachel Pinskey, Mily Gandhi, and the rest of the SSTP and CPET staff who made this research possible.

---

7 References

1. What are proteins and what do they do? Medline Plus. *The National Library of Medicine*. **2021** Mar. https://medlineplus.gov

2. Sun, P. D.; Foster, C. E.; Boyington, J. C. Overview of Protein Structural and Functional Folds. *National Library of Medicine.* **2004** Feb; DOI: 10.1002/0471140864.ps1701s35.

3. Ferreon, J. C.; Ferreon, A. C. Intrinsically Disordered Proteins: An Overview. *National Library of Medicine.* **2022** Nov; 23(22): 14050. DOI: 10.3390/ijms232214050.

4. Liu Z, Huang Y. Advantages of proteins being disordered. *Protein Sci.* **2014** May;23(5):539-50. doi: 10.1002/pro.2443.

5. Deiana A, Forcelloni S, Porrello A, Giansanti A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One*. **2019** Aug 19. DOI: 10.1371/journal.pone.0217889

6. Dunleavy, K. M.; Li, T.;Milshteyn E.; Jaufer, A. M.; Walker, S. A.; Fanucci, G. E. Charge Distribution Patterns of IA_3 Impact Conformational Expansion and Hydration Diffusivity of the Disordered Ensemble. *The Journal of Physical Chemistry B* **2023** 127 (45), 9734-9746 DOI: 10.1021/acs.jpcb.3c06170.

7. Bobrovs*, R.; Jaudzems, K.; and Jirgensons, A. Exploiting Structural Dynamics To Design Open-Flap Inhibitors of Malarial Aspartic Proteases. *J. Med. Chem. 2019, 62, 20, 8931–8950*. May 7, **2019**. DOI: 10.1021/acs.jmedchem.9b00184

8. Bobrovs, R.; Basens, E.E.; Drunka, L.; Kanepe, I.; Matisone, S.; Velins, K.K.; Victor Andrianov, Gundars Leitis, Diana Zelencova-Gopejenko, Dace Rasina, Aigars Jirgensons, and Kristaps Jaudzems. Exploring Aspartic Protease Inhibitor Binding to Design Selective Antimalarials. *Journal of Chemical Information and Modeling* **2022** 62 (13), 3263-3273 DOI: 10.1021/acs.jcim.2c00422

9. Danilo Roccatano* , Giorgio Colombo , Marco Fioroni , and Alan E. Mark. Mechanism by which 2,2,2-trifluoroethanol/watermixtures stabilize secondary-structure formationin peptides: A molecular dynamics study. *Proceedings of the National Academy of SciencesVol. 99, No. 19*. September 17, **2002** DOI: 10.1073/pnas.182199699

10. Pirman, N. L.; Milshteyn, E.; Galiano, L.; Hewlett, J. C.; Fanucci, G. E. Characterization of the Disordered-to-A-Helical Transition of IA3 by SDSL-EPR Spectroscopy. *Protein Sci.* **2011**, 20, 150– 159, DOI: 10.1002/pro.547

11. Green, T. B.; Ganesh, O.; Perry, K.; Smith, L.; Phylip, L. H.; Logan, T. M.; Hagen, S. J.; Dunn, B. M.; Edison, A. S. IA3, an Aspartic Proteinase Inhibitor from Saccharomyces Cerevisiae, Is Intrinsically Unstructured in Solution. *Biochemistry* **2004**, 43, 4071– 4081, DOI: 10.1021/bi034823n

12. Ganesh, O. K.; Green, T. B.; Edison, A. S.; Hagen, S. J. Characterizing the Residue Level Folding of the Intrinsically Unstructured IA3. *Biochemistry* **2006**, 45, 13585– 13596, DOI: 10.1021/bi061358w

13. Dunleavy, K. M.; Milshteyn, E.; Sorrentino, Z.; Pirman, N. L.; Liu, Z.; Chandler, M. B.; D'Amore, P. W.; Fanucci, G. E. Spin-Label Scanning Reveals Conformational Sensitivity of the Bound Helical Interfaces of IA3. *Aims Biophys.* **2018**, 5, 166– 181, DOI: 10.3934/biophy.2018.3.166

14. Casey, T. M.; Liu, Z. L.; Esquiaqui, J. M.; Pirman, N. L.; Milshteyn, E.; Fanucci, G. E. Continuous Wave W- and D-Band Epr Spectroscopy Offers "Sweet-Spots" for characterizing Conformational Changes and Dynamics in Intrinsically Disordered Proteins. *Biochem. Biophys. Res. Com.* **2014**, 450, 723- 728, DOI: 10.1016/j.bbrc.2014.06.045

15. Song, L. K.; Liu, Z. L.; Kaur, P.; Esquiaqui, J. M.; Hunter, R. I.; Hill, S.; Smith, G. M.; Fanucci, G. E. Toward Increased Concentration Sensitivity for Continuous Wave Epr Investigations of Spin-Labeled Biological Macromolecules at High Fields. *J. Magn. Reson.* **2016**, 265, 188– 196, DOI: 10.1016/j.jmr.2016.02.007

16. Green, T.; Perry, K.; Smith, L.; Phylip, L. H.; Logan, T. M.; Hagen, S. J.; Dunn, B. M.; Edison, A. IA3, a Yeast Proteinase a Inhibitor, Is Intrinsically Unstructured in Solution. *Sci. World J.* **2002**, 2, 99– 101, DOI: 10.1100/tsw.2002.47

17. Micsonai, A., Bulyáki, É., Kardos, J. BeStSel: From Secondary Structure Analysis to Protein Fold Prediction by Circular Dichroism Spectroscopy. In *Structural Genomics.* Methods in Molecular Biology, vol 2199. **2021**; pp 181-195. DOI: 10.1007/978-1-0716-0892-0.

18. Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O.G. Richardson, Rohit V. Pappu. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal Vol 112*. **2017**. DOI: 10.1016/j.bpj.2016.11.3200

19. Fasman GD. Circular dichroism and the conformational analysis of biomolecules. *Plenum Press, New York*. **1996**. DOI: 10.1007/978-1-4757-2508-7

20. Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J. Phys. Chem. Lett.* **2020**, 11, 9, 3408–3415. DOI: 10.1021/acs.jpclett.0c00288.

21. Bianchi G, Mangiagalli M, Barbiroli A, Longhi S, Grandori R, Santambrogio C, Brocca S. Distribution of Charged Residues Affects the Average Size and Shape of Intrinsically Disordered Proteins. *Biomolecules*. **2022** Apr 9;12(4):561. DOI: 10.3390/biom12040561

22. Li, T. Investigating the Impact of Charge Distributions on the Unfolded and Folded Conformations of IA_3. *University of Florida.* **2023** Mar.

23. Rodger, A. Circular Dichroism and Chirality. *Encyclopedia of Biophysics; Springer, Berlin, Heidelberg*. **2013**. DOI: 10.1007/978-3-642-16712-6_646

24. Francis DM, Page R. Strategies to optimize protein expression in E. coli. *Curr Protoc Protein Sci*. **2010** Aug; DOI: 10.1002/0471140864.ps0524s61

25. Santofimia-Castaño P, Rizzuti B, Xia Y, Abian O, Peng L, Velázquez-Campoy A, Neira JL, Iovanna J. Targeting intrinsically disordered proteins involved in cancer. *Cell Mol Life Sci*. **2020**. DOI: 10.1007/s00018-019-03347-3

26. Ayyadevara S, Ganne A, Balasubramaniam M, Shmookler Reis RJ. Intrinsically disordered proteins identified in the aggregate proteome serve as biomarkers of neurodegeneration. *Metab Brain Dis*. **2022** Jan. DOI: 10.1007/s11011-021-00791-8

27. Uversky VN. Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the D(2) concept. *Expert Rev Proteomics*. **2010** Aug. doi: 10.1586/epr.10.36

28. Spillantini, M., Schmidt, M., Lee, VY. et al. α-Synuclein in Lewy bodies. *Nature 388*, **1997**. DOI.org/10.1038/42166

29. Bobalova, J., Strouhalova, D., and Bobal, P. *Journal of Agricultural and Food Chemistry*, **2023**. DOI: 10.1021/acs.jafc.3c00886

30. Kim, D. and Han, K. Chapter Six - Target-binding behavior of IDPs via pre-structured motifs. *Progress in Molecular Biology and Translational Science*. **2021**. DOI: 10.1016/bs.pmbts.2021.07.031