University of Groningen

Data-driven Modelling of Intrinsically Disordered Proteins

Tamiola, Kamil

DOI:
[10.33612/diss.96266373](https://doi.org/10.33612/diss.96266373)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Tamiola, K. (2019). *Data-driven Modelling of Intrinsically Disordered Proteins*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.96266373>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

university of
groningen

Data-driven Modelling of Intrinsically Disordered Proteins

Phd thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday 27 September 2019 at 11.00 hours

by

Kamil Tamiola

born on 27 December 1982
in Starachowice, Poland

Supervisor

Prof. D.B. Janssen

Co-supervisor

Prof. F.A.A. Mulder

Assessment Committee

Prof. S.J. Marrink

Prof. A. Bonvin

Prof. W. Vranken

Mojej Mamie, Babci i Bartkowi

Contents

1	Introduction	1
1.1	Intrinsically disordered proteins	1
1.1.1	Amino acid preferences of IDPs	2
1.1.2	Structural disorder comes in many 'flavors'	3
1.1.3	Functional spectrum of IDPs	14
1.1.4	Biophysical characterization of IDPs	15
1.2	Nuclear Magnetic Resonance Spectroscopy	18
1.2.1	Nuclear magnetism as a quantum phenomenon	18
1.2.2	The vector model	23
1.2.3	Chemical shielding	24
1.2.4	Secondary chemical shifts	26
1.3	Mathematical modelling	27
1.3.1	Influential observations in experimental data	28
1.3.2	Singular Value Decomposition (SVD)	31
1.3.3	Solving linear systems using SVD	35
1.4	Overview	38
2	ncIDP Database	43
2.1	Introduction	44
2.2	Materials and methods	44
2.2.1	Nomenclature and definitions	44
2.2.2	Chemical Shift Database	45
2.2.3	Determination of correction factors and random coil shifts	46
2.2.4	Self-consistent optimization protocol	48
2.3	Results and Conclusions	49
3	ncIDP-Assign	71
3.1	Introduction	71
3.2	Implementation	72
3.3	Results and Conclusion	73

4	ncSPC	77
4.1	Introduction	78
4.2	Materials and Methods	80
4.2.1	Sequence-specific 'random-coil' chemical shifts	80
4.2.2	Structural Propensity Calculations	81
4.3	Results	83
4.3.1	Example I: Structural characterization of the intrinsically disordered protein α -synuclein under native and membrane-mimicking conditions	84
4.3.2	Example II: Light activation of Photoactive Yellow Protein	86
4.4	Conclusion	87
5	pepKalc	91
5.1	Introduction	92
5.2	Approach	93
5.2.1	The general theory	93
5.2.2	The hybrid mean field method	99
5.3	Methods	101
5.3.1	Software implementation	101
5.3.2	Performance testing and deployment	101
5.4	Discussion	101
5.4.1	Numerical Convergence	102
5.4.2	Performance Gain Factor	103
5.4.3	Calculation Time Benchmarks	105
5.4.4	Strong Electrostatic Coupling	105
5.4.5	Site-specific protonation behavior of intrinsically disordered α -synuclein	107
5.5	Conclusion	109
6	Conclusions and future perspectives	113
6.1	Conclusions	113
6.2	Future perspectives	119
6.3	Nederlandse Samenvatting	121
7	Epilogue	125
	Bibliography	129
8	Acknowledgements	141
9	Publications	151

9.1	Scientific publications	152
9.2	Art and popular science publications	153

Introduction

Intrinsically disordered proteins, NMR and mathematical modelling

1.1 Intrinsically disordered proteins

Intrinsically disordered proteins are a class of polypeptides devoid of persistent secondary structures [1]. They constitute an elusive group of biomolecules, quite recently labelled as "dark proteome" [2]. The elusive character of IDPs is backed by the most recent statistical survey of DISPROT [3], a repository of experimentally confirmed IDPs which demonstrates that insofar only 803 proteins and 2167 regions were confirmed to be intrinsically disordered through experimental verification (statistics as of October 2018). When contrasted with the statistics on SWISS-PROT [4], a non-redundant and manually annotated repository of 558,590 proteins, the current number of annotated IDPs constitutes only a meagre 0.14% of all characterized proteins. Table 1.1 provides a taxonomic breakdown of the sources of known IDPs, suggesting the current knowledge on IDPs transpires mainly from biophysical studies of eukaryotic organisms. Although the actual number of experimentally confirmed intrinsically disordered proteins is very small, it has been shown by bioinformatic analyses that IDPs may constitute a significant (estimated 33%) part of the protein universe [5]. This claim has been supported by numerous computational predictions on the disorder penetrance at the proteome level [6, 7]. It has been shown that amino acid sequence preferences of IDPs are a key determinant for their problematic biophysical characterization [8]. On the other hand, sequence patterns within IDPs warrant ease of detection in genome-wide bioinformatic surveys [9–11].

Tab. 1.1: Taxonomic classification of experimentally confirmed IDPs in DISPROT [3].

Kingdom	Number of proteins
Metazoa	426
Proteobacteria	102
Fungi	62
Viridiplantae	51
ssRNA viruses	33
Firmicutes	29
dsDNA viruses, no RNA stage	23
Retro-transcribing viruses	16
Euryarchaeota	11
Alveolata	11
Actinobacteria	9
Euglenozoa	4
Deinococcus-Thermus	4
Crenarchaeota	4
Cyanobacteria	3
Thermotogae	3
Spirochaetes	3
Aquificae	2
Stramenopiles	1
dsRNA viruses	1
Amoebozoa	1
Parabasalia	1
ssDNA viruses	1
Deltavirus	1
environmental samples	1

1.1.1 Amino acid preferences of IDPs

A comparative analysis of 300+ natively folded and 100+ intrinsically disordered proteins revealed that the combination of low mean hydropathy and high net charge represents a prerequisite for the absence of canonical secondary structures in IDPs under physiological conditions [12]. High net charge yields electrostatic repulsion, whereas low hydropathy results in weak protein compaction in intrinsically disordered polypeptides [1]. Thus, the lack of stable secondary structures in IDPs seems to transpire from their amino acid sequence [13] just like stable structure is warranted by primary protein structure in natively folded peptides, in accordance to Anfinsen's dogma [14].

A more detailed analysis of IDP sequences revealed that disordered proteins are significantly depleted in so-called order-promoting residues [13] like bulky and hydrophobic: Ile, Leu, and Val, and aromatic amino acid residues: Trp, Tyr, and Phe, which would normally form the hydrophobic core of a folded globular protein [7, 9]. Besides that, IDPs were found to possess low content of Cys and Asn residues [8]. Importantly, IDPs were found to be substantially enriched in disorder-promoting residues: Arg, Gly, Gln, Ser, Glu, and Lys, and hydrophobic, but structure-breaking Pro and hydrophobic Ala [15].

Based on the ability of amino acids to promote order and disorder, a special amino acid scale (TOP-IDP) was introduced [16]. The scale enabled discrimination between ordered and intrinsically disordered proteins with a reasonably high accuracy [16]. The amino acids were ranked according to their capabilities to promote order or disorder resulting in the following arrangement (from the most order promoting to the most disorder promoting): W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P [16].

The combined use of charge-hydrophobicity ratio's (typically high for IDPs) [12] and the TOP-IDP scale [16] facilitated the discovery of numerous putative IDPs in large-scale genomic surveys, advancing state of the knowledge about disordered proteins and their cellular functions [17].

1.1.2 Structural disorder comes in many 'flavors'

The absence of persistent secondary structures in natively unfolded proteins is predominantly born out the electrostatic repulsion and low-compaction in their structural ensembles. Figure 1.1 demonstrates an exemplary structural ensemble for natively unfolded 140 residue-long human alpha-synuclein (aS), composed of 100 lowest-energy *in silico* generated conformers [18]. The depicted aS ensemble is complex and highly heterogeneous and cannot be approximated by a single or even a small number of distinct sets of coordinates. Thus, a robust proxy or a statistical descriptor is required to provide a concise classification of conformational features of aS ensemble.

Fig. 1.1: A superposition of 100 low-energy, computer-generated conformers of human alpha-synuclein, adapted from PED1AAD in Protein Ensemble Database [18].

Ramachandran plot

In their seminal work from 1963, Ramachandran and co-workers demonstrated that vast structural complexity of an atomistic 3D protein model can be effectively reduced by studying and plotting distributions of backbone torsion angles in polypeptide chain [19]. Figure 1.2 demonstrates a model of an arbitrary polypeptide backbone with cardinal torsion angles Φ and Ψ , which report on the rotation around N-C α , and C α -C, respectively. The Ramachandran diagram, which plots Φ versus Ψ backbone dihedral angles for each amino acid residue in a peptide chain provides an easy way to view the distribution of torsion angles in a protein structure. Figure 1.3 shows 3D models and their corresponding Φ and Ψ distribution plots for three seminal proteins, an all-helical human haemoglobin, a mixed α - and β -structure rich green fluorescent protein (GFP) and a single conformer taken from an aS

Fig. 1.2: Dihedral angles in arbitrary polypeptide chain.

ensemble, introduced in the previous subsection. The Φ , Ψ distribution in near all-helical human haemoglobin, visible in Figure 1.3b clusters around torsional angle values of -50° , -50° , whereas an inclusion of beta-sheet elements in GFP results in largely binominal dihedral angle distribution observed on Figure 1.3d. The Φ , Ψ angles in unstructured aS, visible on Figure 1.3f are very dispersed indicating significantly different conformational preferences in backbone conformation of alpha-synuclein from those observed for folded proteins. Importantly, the Φ , Ψ coordinate system can be used to outline the most energetically and statistically plausible torsion angle scenarios in protein structure validation, as depicted in Figure 1.3 with blue contour outlines. Since the Φ , Ψ values are not normally optimised in the X-ray model refinement process they can serve as sensitive indicators of local problem areas in model refinement [20].

With a simplified measure of local secondary structure warranted by residue-specific Φ , Ψ angle combinations, it has become possible to build the first semi-empirical models of 'random-coil' conformations and quantify the degree of structural disorder within proteins using experimental techniques. Three, core "flavors" of structurally disordered protein states were proposed: random-coils, pre-molten globules and molten globules. Their characterization was fundamentally important for complete understanding of structural heterogeneity of intrinsically disordered proteins.

Fig. 1.3: 3D protein models and their corresponding Ramachandran plots for (a,b) human haemoglobin (PDB 2DN1), (b,c) photo-activated form of GFP (PDB 3GJ2), and (e,f) singular conformation of human alpha-synuclein from ensemble (PED1AAD). Contours in Ramachandran plots correspond to Φ and Ψ distributions observed in 500 high definition protein structures analyzed by Lovell et al. [20]

Random-coil polymers

The very first random-coil polymer model assumed a freely joined chain of molecules in which consecutive residues were connected by bonds of fixed length and uncorrelated directions [21]. This simplistic approach was capa-

ble of predicting the mean radius of gyration of random-coil polymers with satisfactory accuracy. A major revision of previously introduced random-coil model came from Tanford in 1968, who suggested a polymer molecule is randomly coiled when internal rotation can take place at about every single bond within the said molecule with energy barriers that correspond to bond rotations observed in low-molecular weight residues containing the same kind of bonds. Tanford's model postulated a true random-coil had no preferred conformations with essentially featureless rotational free energy landscape [22]. Tanford's 'random-coil' model required a major revision in order to account for heterogeneity born out of variable amino acid composition of polypeptide chains. Eventually, it got replaced by the Flory random-coil [23] and worm-like chain models [24]. In the rotational isomeric approximation [23] to the Flory random-coil model, the conformational partition function for the polypeptide is written as a product of partition functions of independent interaction units, as illustrated in Figure 1.4. All interactions

Fig. 1.4: A Flory model of 3-amino acid long protein backbone, with Kuhn segments $Eb - Bc - Ab$.

between non-nearest neighbour units, or so-called Kuhn segments, are explicitly ignored although the intrinsic conformational preferences of individual units are captured in terms of weights for each of the possible rotational isomers, as shown in matrix Equation 1.1. The unit either spans the degrees of freedom of an individual residue or can take local effects into account to expand the unit to span multiple residues. In either case, each conformation for unit i is annotated by an intrinsic energy value that is calculated using an empirical potential function of one's choosing. The conformations are binned into rotational isomeric states based on the similarities of the backbone and side-chain dihedral angles, as shown in Figure 1.5. Residue i , might have m rotational isomers whereas residue j might have n rotational isomers. Thus for a given residue, each rotational isomer is assigned a weight that is calcu-

Fig. 1.5: Coarse graining of conformational space into discrete rotational isomers for an arbitrary residue (Kuhn segment) with known free energy landscape (red to blue). The tiles represent discrete rotational isomers, with distinct label and a statistical weight.

lated using the Boltzmann weights of energies of individual conformations, using Equation 1.1, that make up the rotational isomer.

$$\begin{bmatrix} AaAaAa & \dots & Aa \\ AaAaAa & \dots & Ab \\ DeBaFc & \dots & Bc \\ \vdots & \vdots & \vdots \\ GgGgGg & \dots & Gf \\ GgGgGg & \dots & Gg \end{bmatrix} \propto \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{M-1} \\ P_M \end{bmatrix} \quad (1.1)$$

Given an amino acid sequence of N residues, one can calculate, a priori, the probabilities associated with all combinations of rotational isomers from,

$$P_z = \prod_{i=1}^N W_m^i \quad (1.2)$$

For the sequence of interest the number of rotational isomers per residue, their statistical weights, and the sequence composition dictate the total number of conformational possibilities and the likelihood associated with each conformation. These likelihoods make up the predicted conformational distribution function and can be used to calculate a variety of conformational properties including the average end-to-end distance, the average radius

of gyration, the average hydrodynamic size, the average distance between residues i and j , and any observable that can be cast as a function of a moment of the conformational distribution function. Because of its simplicity and robustness, the Flory random-coil model has become a foundation for more elaborate semi-empirical models for IDPs.

Structural anisotropy in random-coil polypeptides

Steric hindrance and charge-charge interactions between amino acids in unfolded peptides lead to structural anisotropy and effectively restrict the accessible Φ , Ψ angles bonds which a polypeptide chain can sample [25]. The structural anisotropy in random-coil polypeptides may demonstrate itself as transient or persistent residual structure [26]. The notion that disordered proteins could contain important levels of residual secondary structure gained importance after preliminary NMR experiments with presumed random-coil protein samples, such as FK506 binding protein unfolded in urea and guanidine hydrochloride [27]. In their seminal paper, Shortle and Ackerman clearly demonstrated that a native-like topology can be found in denatured samples of staphylococcal nuclease challenging the 'featureless random coil' hypothesis and suggesting that secondary structure elements might form an inevitable part of the protein random-coil state [28]. Fitzkee and Rose produced a supporting body of evidence for the residual structure hypothesis by calculating two related measures of polypeptide compactness in the random-coil state: the radius of gyration R_G , and the end-to-end distance $\langle L^2 \rangle$ from computer-generated ensembles of artificially designed proteins that contained mainly (92%) short segments of defined structure (α -helices and β -strands) linked by much shorter fragments (8%) for which the backbone torsion angles were allowed to vary freely. The predicted R_G and $\langle L^2 \rangle$ values were very similar to compactness expected for completely random-coil peptides devoid of any secondary structures [29].

Unfolded polypeptides differ from statistical random-coils

The notion that natively unfolded proteins are not statistical random-coils but may contain regions with nascent secondary structures was further supported by the work of Schwalbe and Dobson who introduced a 'coil' model

born out of statistical analysis of amino acid-specific torsion angles found in experimental structures in the Protein Data Bank (PDB) [30, 31]. Deviations from the predictions of the 'coil' model could be interpreted as residual structuration or higher order resulting from long-range contacts within the unstructured ensembles [32]. An inclusion of nearest-neighbour effects on torsion angles enabled more accurate disorder inference, demonstrating that residual secondary structure is mostly absent with small local exceptions in purposefully denatured proteins [33]. The idea that artificially unfolded protein ensembles are largely devoid of local secondary structures was further explored by Wright, Dyson and co-workers who applied the rotational isomeric state theory of Flory [23] for the unstructured protein state, in which the chain is treated as a polymer of jointed statistical segments that are randomly oriented with respect to each other [34]. Such statistical segments comprise several amino acid residues (determined by small-angle X-ray scattering and NMR relaxation measurements to be five to seven amino acids), with a propensity towards extended backbone (β or polyproline II) conformations and that those segments are highly anisotropic in shape. The experimental results for unstructured proteins could be explained without having to invoke the presence of residual structuration as was suggested before. The authors of the study suggested that steric restrictions on rotations about the dihedral angles resulted in local stiffness leading to extended conformations [34]. The final layer of refinement in *in silico* approaches to 'coil' models was the flexible-meccano's Monte Carlo algorithm [35] for generating the backbone of the unfolded-state conformations. It used a subset of the database of amino acid-specific Φ - and Ψ -torsion angles obtained by exclusion of all residues in α -helices and β -sheets. The database had special cases for residues preceding a proline. In this case disordered state conformations were built by adding residues with a randomly selected pair of Φ - and Ψ -angles from the torsional subset database. If this introduced clashes, the angle pair was rejected and replaced by a randomly generated suggestion, thereby implicitly introducing the influence of the preceding neighbour. The flexible-meccano approach reinforced the notion that persistent secondary structure was mainly absent in unstructured protein ensembles and pointed towards the presence of extended (β and polyproline II) conformations.

Nearest-neighbour effects in disordered protein states

The experimental evidence born out previous examples suggested that conformational sampling in unstructured proteins might be determined by the residue type and the identity of its neighbours. To account for the variable impact of neighbouring effects in disordered proteins, the concept of 'persistence length' was introduced. It reported on an effective residue cutoff beyond which the remainder of the polypeptide chain could be considered of negligible effect on the residue of interest. Lippens and co-workers studied 'persistence length' in NMR spectroscopy experiments for small peptides of 13-17 amino acid residues in length, and demonstrated the spectral resonances of central residues with 2-residue cutoff on both sides in these peptides were identical as in the context of full length unstructured Tau protein [36].

Molten globule states

Upon an exposure to denaturing agents, globular proteins may transition to structurally labile conformations such as molten and pre-molten globular, which could be viewed as non-linear interpolations between ordered and completely disordered states [37–39].

Small-angle X-ray scattering experiments on proteins in chemically induced molten globule states showed that polypeptides in this intermediate state have globular structure typical of native globular proteins [40, 41]. Hydrogen-deuterium exchange 2D NMR spectroscopy experiments showed that protein molecules in molten globule states were characterised not only by the native-like secondary structure content, but also by the native-like folding pattern [42, 43]. A considerable increase in the accessibility of a protein molecule to proteases was noted as a specific property of the molten globule state [44]. Finally, it was established that the averaged value for the increase in the hydrodynamic radius in the molten globule state compared with the native form is no more than 15%, which corresponds to a volume increase of 50%. Interestingly, one of the seminal examples of well-characterized molten globule proteins is an insulin analogue, des-pentapeptide-(B26-B30)-insulin (DPI), whose model is given by Figure 1.6 [45, 46]. Structures of insulin in different crystal forms exhibit significant local and non-local differences, including correlated displacement of elements of secondary

Fig. 1.6: A superposition of 15 solution NMR structures (PDB 1HIS) for des-pentapeptide-(B26-B30)-insulin (DPI), (a) front and (b) back projection. Color code corresponds to the original secondary structure annotations found in the PDB file, (grey) turn, (magenta) α -helix, (cyan) β -turn and (yellow) β -strand.

structure [47]. As a monomeric insulin analogue, DPI exists in partially folded state formed by coalescence of distinct α -helix-associated micro-domains and confirmed in 2D nuclear Overhauser enhancement (NOE) spectroscopy (NOESY) experiments.

Pre-molten globule states

A thorough investigation of an anion-induced folding in Staphylococcal nuclease revealed multiple partially folded intermediates, which displayed distinct properties from previously described molten globule state [48]. The studied polypeptide was found to have 50% native secondary structure and roughly three times bigger hydrodynamic volume as compared to the native state. Ultimately, Uversky and co-workers demonstrated that Staphylococcal nuclease trapped in this intermediate state had no globular structure [48]. Figure 1.7 shows an OB-fold sub-domain of the above Staphylococcal nuclease, which gives NMR spectra characteristic of an unfolded protein, i.e. the wild-type nuclease sequence is insufficient to maintain a stable tertiary structure in

Fig. 1.7: A superposition of 10 solution NMR structures (PDB 2SOB) for OB-fold sub-domain of Staphylococcal nuclease, (a) front and (b) top projection. Color code corresponds to the original secondary structure annotations found in the PDB file, (grey) turn, (magenta) α -helix, (cyan) β -turn and (yellow) β -strand.

the absence of the C-terminal one-third of this single-domain protein. A large unfolded loop, composed of residues 1–103 can be clearly seen on the Figure 1.7. Interestingly, both hydrodynamic radius and NMR spectra for this polypeptide change upon Val66Leu and Gly88Val mutations, which appear to stabilize tertiary structure by consolidating the hydrophobic core of the nuclease OB-fold sub-domain and bringing it closer to more compact, molten globule state [49].

Consequently, a new term describing this intermediate but distinct structurally disordered form was introduced and named as pre-molten globule state. It is known now that protein molecules in pre-molten globule state are considerably less compact than in the molten globule or native states, but they are still more compact than random-coil (its hydrodynamic volume in the molten globule, the pre-molten globule, and the unfolded states, in comparison to that of the native state, increases 1.5, 3, and 12 times, respectively). Moreover, polypeptides in pre-molten globule states are devoid of rigid tertiary structures, yet may contain a considerable amount of secondary structure, although much less pronounced than that of the native or the molten globule protein states. Ultimately, it has been shown that the pre-molten globule is separated from the molten globule state by an all-or-none transition, which represents an intramolecular analogue of the first-order phase transition [50, 51].

Consensus view on disordered protein states

The evolution of 'coil' and disordered state models, disclosed in previous subsections, leads to a few fundamental conclusions about IDPs in general,

- IDPs are not amenable to descriptions by single or even small number of distinct coordinate sets [52]. Instead, statistical descriptors are required to provide a concise classification of conformational ensembles and this can be achieved in the language of polymer physics [53],
- residues in intrinsically disordered proteins rapidly sample Φ , Ψ angles according to their own properties [54],
- nearest neighbour-effects on Φ , Ψ sampling is particularly important for accurate modelling of disordered ensembles [54],
- molten and pre-molten globule states are not exclusive to globular proteins in denaturing conditions, but can be also detected in partially disordered proteins under native conditions,
- proteins function within a conformational continuum, shown in Figure 1.8, ranging from fully structured to completely disordered [55],
- the observed continuum of structure in proteins has led to a formulation of "protein quartet" paradigm [57], shown on Figure 1.9, as an alternative for the traditional protein structure–function paradigm.

1.1.3 Functional spectrum of IDPs

IDPs, although highly abundant in any given proteome [2] and effectively complementing the functional spectrum of ordered peptides [9, 58], are still considered an elusive and difficult to characterize part of the protein universe [2]. However, the prevalence of functional protein disorder demands reevaluation of the classical structure–function paradigm [59, 60], as biophysical features of IDPs and their protein interactions vary tremendously [61]. There may be no common mechanism that can explain the different binding modes observed experimentally [5]. The verified functions of IDPs include regulation of transcription and translation [62], cellular signal transduction [60,

Fig. 1.8: Schematic representation of the continuum model of protein structure. The color gradient represents a continuum of conformational states ranging from highly dynamic, expanded conformational ensembles (red) to compact, dynamically restricted, fully folded globular states (blue). Dynamically disordered states are represented by heavy lines, stably folded structures as cartoons. A characteristic of IDPs is that they rapidly interconvert between multiple states in the dynamic conformational ensemble. In the continuum model, the proteome would populate the entire spectrum of dynamics, disorder, and folded structure depicted. Adopted from Lee et al. [56].

63], protein phosphorylation [64], the storage of small molecules, and the regulation of the self-assembly of large multi-protein complexes such as the bacterial flagellum and the ribosome [61]. Interestingly, IDPs seem to be complementary to ordered proteins and protein domains by making combined use of features such as induced structure upon functional interactions and flexibility, depending on the individual system and the cellular context [65].

1.1.4 Biophysical characterization of IDPs

As explained in the seminal works of Dyson, Wright [66] and Dobson [58], naturally occurring protein disorder severely limits three-dimensional structure determination using X-ray crystallography. Missing electron density of backbone atoms in crystallographic structures has long been attributed to localized structural disorder [67]. One of the most prominent and demonstrative cases is a comprehensive analysis of Early E2A DNA-binding protein from human adenovirus type 5 [68]. This polypeptide is essential in the unwinding and replication of viral DNA. It also helps in virion assembly, host-range determination, direct control of transcription, and the regulation of the stability of

Fig. 1.9: The protein quartet model of protein conformational states. In accordance with this model, protein function arises from four types of conformations of the polypeptide chain (ordered forms, molten globules, pre-molten globules, and random-coils) and transitions between any of these states. Adopted from Lee et al. [56].

RNA. The E2A protein was found to contain 7 regions of intrinsic disorder, as gauged from the thorough analysis of X-ray scattering patterns and missing structural coordinates. A set of independent experimental studies, confirmed the initial findings of structural disorder in loop regions, serving as an illustrative case of disorder mapping using X-ray crystallography [68, 69]. A wide palette of experimental approaches have been developed to capture structural and spatio-temporal heterogeneity of intrinsically disordered proteins [6]. Table 1.2 provides a comprehensive overview of empirical approaches to intrinsic disorder identification employed in the latest edition of DISPROT (October 2018) [3]. It should be noted that besides the discussed "negative" X-ray crystallography, spectroscopic methods including NMR and variants of circular dichroism (CD) played a fundamental role in producing experimental evidence for structural disorder in polypeptides. A detailed description of the experimental techniques in IDP characterization is beyond the scope of this thesis and has been covered elsewhere [6].

Tab. 1.2: Experimental verification of structural disorder in DISPROT [3].

Experimental method	Number of proteins
X-ray crystallography	684
Nuclear magnetic resonance (NMR)	591
Circular dichroism (CD) spectroscopy, far-UV	352
Sensitivity to proteolysis	95
Proton-based NMR	69
Size exclusion/gel filtration chromatography	67
Circular dichroism (CD) spectroscopy, near-UV	39
Aberrant mobility on SDS-PAGE gel	36
Small-angle X-ray scattering (SAXS)	30
Fourier transform infrared spectroscopy (FTIR)	25
Dynamic light scattering (DLS)	22
Analytical ultracentrifugation	21
Hydrogen-deuterium exchange	20
Fluorescence, intrinsic	20
Differential scanning calorimetry	10
Stability at thermal extremes	10
Immunochemistry	9
Rotary shadowing electron microscopy	8
Fluorescence polarization/anisotropy	7
Site-directed spin-labeling EPR spectroscopy	6
Atomic force microscopy (AFM)	6
High relative B-factor	5
ESI-FTICR mass spectrometry	4
HDX-MS	4
Synchrotron radiation circular dichroism (SRCD)	4
Fluorescent probes (extrinsic fluorescence)	3
Fluorescent dynamic quenching	3
Raman optical activity	3
Fluorescence resonance energy transfer (FRET)	3
Viscometry	3
Stability at pH extremes	2
Small-angle neutron scattering (SANS)	2
Raman spectroscopy	1
Vibrational spectroscopy of cyanylated cysteines	1
Static light scattering	1
Radio-labelled enzyme activity assay	1

1.2 Nuclear Magnetic Resonance Spectroscopy

As evidenced by records in Table 1.2, multidimensional heteronuclear nuclear magnetic resonance (NMR) spectroscopy has proven to be a tremendously successful experimental technique in the detection of intrinsic protein disorder. Historically, the first unfolded protein characterized by NMR was the urea unfolded N-terminal domain of 434-repressor [70]. Since dynamic behaviour of chemically unfolded proteins, folding intermediates and natively disordered peptides usually differ a lot, numerous NMR techniques have been deployed to study unfolded and intermediate folding states of proteins and are described elsewhere [71, 72]. As this thesis concerns numerical approaches born out of NMR chemical shifts, the very fundamentals of NMR spectroscopy, which give rise to the chemical shift phenomenon are discussed in detail.

1.2.1 Nuclear magnetism as a quantum phenomenon

The theory of NMR spectroscopy describes the quantum mechanics of nuclear spin angular momentum. Both nuclear magnetism and nuclear magnetic resonance effect constitute empirical evidence for nuclear spin angular momentum, which physical origins are complex and far beyond the scope of this thesis. However, the spin angular momentum can be characterized by the nuclear spin quantum number, I , which displays systematic features: (1) nuclei with odd mass numbers have half-integral spin quantum numbers (e.g. $\frac{1}{2}$), (2) nuclei with an even mass number and an even atomic number have spin quantum numbers equal to zero, and (3) nuclei with an even mass number and an odd atomic number have integral spin quantum numbers.

Since the nuclear magnetic resonance phenomenon relies on the existence of nuclear spin, nuclei belonging to the second category are NMR inactive. Nuclei with spin quantum numbers greater than $\frac{1}{2}$ also possess electric quadrupole moments arising from nonspherical nuclear charge distributions.

Importantly, the lifetimes of the magnetic states for quadrupolar nuclei in solution normally are much shorter than are the lifetimes for nuclei with $I = \frac{1}{2}$ leading to significant line broadening in NMR spectra and detection difficulties. The most relevant properties of nuclei commonly utilized in biomolecular NMR studies are summarised in Table 1.3. The most important

Tab. 1.3: The properties of nuclei commonly used in biological NMR spectroscopy. ^a the nuclear spin angular momentum quantum number; ^b the magnetogyric ratio.

Nucleus	I^a	$\gamma^b (T\ s)^{-1}$	Natural abundance (%)
1H	1/2	$2.6752 \cdot 10^8$	99.99
2H	1	$4.107 \cdot 10^7$	0.012
^{13}C	1/2	$6.728 \cdot 10^7$	1.07
^{15}N	1/2	$-2.713 \cdot 10^7$	0.37
^{19}F	1/2	$2.518 \cdot 10^8$	100.00
^{31}P	1/2	$1.0839 \cdot 10^8$	100.00

nuclei in biomolecular NMR spectroscopy with $I = \frac{1}{2}$ are 1H , ^{13}C , ^{15}N , ^{19}F , and ^{31}P , whereas the most significant nucleus with $I = 1$ is the deuteron (2H). The nuclear spin angular momentum, $\mathbf{I}$, is a vector quantity with magnitude given by

$$I = [\mathbf{I} \cdot \mathbf{I}]^{1/2} = \hbar I[I + 1]^{1/2} \quad (1.3)$$

in which I is the nuclear spin angular momentum quantum number and $\hbar$ is Planck's constant divided by 2π . Due to the quantum mechanical uncertainty principle, only one of the three Cartesian components of $\mathbf{I}$ can be specified simultaneously with $I^2 = \mathbf{I} \cdot \mathbf{I}$. Conventionally, the value of the z-component (z-axis projection) of $\mathbf{I}$ is expressed as

$$I_z = \hbar m \quad (1.4)$$

in which m is the magnetic quantum number with values given according to $m = (-I, -I + 1, \dots, I - 1, I)$. Consequently, I_z can adopt $2I + 1$ possible values. The orientation of the spin angular momentum vector in space is quantized, as the magnitude of the vector is constant and the z-component has a set of discrete possible values. In the absence of external magnetic fields, the quantum states corresponding to the $2I + 1$ values of m have the same energy and the spin angular momentum vector $\mathbf{I}$ does not have a preferred orientation.

Nuclei that have nonzero spin angular momentum also possess nuclear magnetic moments. According to the Wigner—Eckart theorem [73], the nuclear magnetic moment, μ , is colinear with I and is defined by

$$\begin{aligned}\mu &= \gamma I \\ \mu_z &= \gamma I_z = \gamma \hbar m\end{aligned}\tag{1.5}$$

in which the magnetogyric ratio, γ , is a characteristic constant for a given nucleus (Table 1.3). The nuclear magnetic moment inherits quantization from angular momentum. Importantly, the magnitude of γ , in part, determines the magnetic receptivity of a nucleus in NMR spectroscopy. Thus, in the presence of an external magnetic field, the spin states of the nucleus have energies computed from

$$E = -\mu \cdot B\tag{1.6}$$

in which B is the magnetic field vector. An alignment of μ with B leads to the minimum energy state. However, since the magnitude of nuclear spin angular momentum I exceeds the value of its z-axis projection, that is $|I| > I_z$, μ cannot be completely colinear with B and the m spin states become quantized with energies proportional to their projection onto B . In an NMR spectrometer, the static external magnetic field is directed along the z-axis of the laboratory coordinate system. Under such conditions, Equation 1.6 can be reduced to

$$E_m = -\gamma I_z B_0 = -m \hbar \gamma B_0\tag{1.7}$$

in which B_0 is the static magnetic field strength. Hence, in the presence of a static magnetic field, the projections of the angular momentum of the nuclei onto the z-axis of the laboratory frame result in $2I + 1$ equally spaced energy levels, which are known as the Zeeman levels. The quantization of I_z and its geometrical representation are shown in Figure 1.10. In an unperturbed and equilibrated system, different energy states are unequally populated because lower energy orientations of the magnetic dipole vector are more probable. The relative population of a state under equilibrium conditions can be obtained from a Boltzmann distribution,

$$\frac{N_m}{N} = \frac{e^{\frac{-E_m}{k_B T}}}{\sum_{m=-I}^I e^{\frac{-E_m}{k_B T}}} \approx \frac{1}{2I + 1} \left(1 + \frac{m \hbar \gamma B_0}{k_B T} \right)\tag{1.8}$$

Fig. 1.10: A visualization of the allowed z -components, I_z of the angular momentum vectors, $\mathbf{I}$, for (a) a spin- $\frac{1}{2}$ particle and (b) a spin-1 particle. The location of $\mathbf{I}$ on the surface of the cone cannot be specified because of quantum mechanical uncertainties in the I_x and I_y components. Adopted from Cavanagh et al. [74]

in which N_m is the number of nuclei in the m -th state and N is the total number of spins, T is the absolute temperature, and k_B is the Boltzmann constant. The result of Equation 1.8 is obtained via the first order Taylor expansion of the exponential functions, as at temperatures relevant to solution NMR spectroscopy, $\frac{m\hbar\gamma B_0}{k_B T} \ll 1$. The populations of the states depend both on the nucleus type and on the applied field strength. An increase of the external field strength yields higher energy differences between the nuclear spin energy levels, which in turn translate into population differences between the states. Importantly, polarization of the spin system to generate a population difference between spin states is not an instantaneous phenomenon. Upon application of the magnetic field, the polarization, or magnetization, develop at a specific rate, commonly referred to as the spin—lattice relaxation rate constant. Magnetic properties at a macroscopic scale are given by the bulk magnetic moment, $\mathbf{M}$, and the bulk angular momentum, $\mathbf{M}$, which is a vector sum of the corresponding quantities for individual nuclei, $\boldsymbol{\mu}$ and $\mathbf{I}$. At thermal equilibrium, the transverse components (e.g., the orthogonal x - or y -components) of $\boldsymbol{\mu}$ and $\mathbf{I}$ for different nuclei in the sample are uncorrelated and thus their sum is zero. The small population differences between energy levels give rise to a bulk magnetization of the sample parallel (longitudinal) to the static magnetic field, $\boldsymbol{\mu} = M_0 \mathbf{k}$ in which $\mathbf{k}$ is the unit vector in the

z-direction. Thus using Equation 1.4, 1.5, and 1.8, M_0 can be represented as

$$\begin{aligned}
 M_0 &= \gamma \hbar \sum_{m=-I}^I m N_m \bigg|_{N_m} \\
 &= N \gamma \hbar \sum_{m=-I}^I m \frac{e^{\frac{m \hbar \gamma B_0}{k_B T}}}{\sum_{m=-I}^I e^{\frac{m \hbar \gamma B_0}{k_B T}}} \\
 &\approx N \gamma \hbar \sum_{m=-I}^I m \frac{\left(1 + \frac{m \hbar \gamma B_0}{k_B T}\right)}{\sum_{m=-I}^I \left(1 + \frac{m \hbar \gamma B_0}{k_B T}\right)} \quad (1.9) \\
 &\approx \left[N \gamma^2 \hbar^2 \frac{B_0}{k_B T (2I + 1)} \right] \sum_{m=-I}^I m^2 \\
 &\approx N \gamma^2 \hbar^2 \frac{B_0 I}{3 k_B T}
 \end{aligned}$$

Transitions between Zeeman levels can be facilitated by applying electromagnetic radiation, which is analogous to other forms of spectroscopy. Magnetic dipole transitions are selected according to $\Delta m = +/ - 1$ rule. Thus, the photon energy, ΔE , required to excite and facilitate a transition between the m and $m + 1$ Zeeman states can be expressed as

$$\Delta E = \hbar \gamma B_0 \quad (1.10)$$

which shows ΔE is directly proportional to the magnitude of the static magnetic field. As born out of the Planck's Law, the frequency of the electromagnetic radiation required to facilitate the state transition is given by

$$\omega = \frac{\Delta E}{\hbar} = \gamma B_0 \quad (1.11)$$

in units of s^{-1} , which can be also expressed in Hertz.

$$\nu = \frac{\omega}{2\pi} = \gamma \frac{B_0}{2\pi} \quad (1.12)$$

The population differences between Zeeman states define the sensitivity of NMR spectroscopy. Since the population difference can be only on the order of 1 in 10^5 e.g. for spins in an 11.7T magnetic field, NMR is a relatively insensitive spectroscopic technique compared to visible or ultraviolet spectroscopy.

1.2.2 The vector model

A simple vector formalism, commonly referred to as the Bloch model, can be used to describe the behaviour of a sample of non-interacting spin- $\frac{1}{2}$ nuclei in a static magnetic field [75]. In the presence of a magnetic field, which may include components in addition to the static field, $\mathbf{M}(t)$ experiences a torque that is equal to the time derivative of the angular momentum,

$$\frac{d\mathbf{J}(t)}{dt} = \mathbf{M}(t) \times \mathbf{B}(t) \quad (1.13)$$

which multiplied at both sides by γ yields

$$\frac{d\mathbf{M}(t)}{dt} = \mathbf{M}(t) \times \gamma \mathbf{B}(t) \quad (1.14)$$

A frame of reference rotating with respect to the fixed laboratory axes is introduced. The angular velocity of the rotating axes is represented by the vector ω . The two coordinate systems are assumed to be superposed initially. Vectors are represented identically in the two coordinate systems; however, time differentials are represented differently in the two coordinate systems. The equations of motion of $\mathbf{M}(t)$ in the laboratory and rotating frames are related by

$$\left[\frac{d\mathbf{M}(t)}{dt} \right]_{rot} = \left[\frac{d\mathbf{M}(t)}{dt} \right]_{lab} + \mathbf{M}(t) \times \omega = \mathbf{M}(t) \times [\gamma \mathbf{B}(t) + \omega] \quad (1.15)$$

The equation of motion for the magnetization in the rotating frame has the same form as in the laboratory frame, provided that the field $\mathbf{B}(t)$ is replaced by an effective field, $\mathbf{B}_{eff}$, given by

$$\mathbf{B}_{eff} = \mathbf{B}(t) + \frac{\omega}{\gamma} \quad (1.16)$$

For the choice $\omega = -\gamma \mathbf{B}(t)$, the effective field is zero, so that $\mathbf{M}(t)$ is time independent in the rotating frame. Consequently, as seen from the laboratory frame, $\mathbf{M}(t)$ precesses around $\mathbf{B}(t)$ with a frequency $\omega = -\gamma B$. For a

static field of strength B_0 , the precessional frequency, also known as Larmor frequency, is given by

$$\omega_0 = -\gamma B_0 \quad (1.17)$$

Thus, in the absence of other magnetic fields, the bulk magnetization precesses at the Larmor frequency around the main static field axis, defined as the z-direction. As described by Levitt [76], the Larmor frequency has different signs for spins with positive or negative gyro-magnetic ratios, e.g., ^1H and ^{15}N . The magnitude of the precessional frequency is identical to the frequency of electromagnetic radiation required to excite transitions between Zeeman levels, as brought by Equation 1.12. This identity is the reason that, within limits, a classical description of NMR spectroscopy is valid for systems of isolated spin- $\frac{1}{2}$ nuclei.

1.2.3 Chemical shielding

The observed resonance frequencies in NMR experiments depend on the local environments of individual nuclei. The deviations in resonance frequencies with respect to Equation 1.17 are referred to as chemical shifts and arise when identical nuclei are exposed to different chemical environments.

The chemical shift phenomenon arises due to the secondary magnetic fields induced by the motions of electrons in the presence of the external magnetic field. Thus the overall magnetic field at a specific location depends upon the static magnetic field and the local secondary fields. The effect of the secondary fields is referred to as nuclear shielding and can augment or diminish the effect of the main field. In general, the electronic charge distribution in a molecule is anisotropic and the effects of shielding on a particular nucleus are described by the second-rank nuclear shielding tensor, represented by a 3×3 matrix. In the principal coordinate system of the shielding tensor, the matrix representing the tensor is diagonal, with principal components σ_{11} , σ_{22} , and σ_{33} . If k th principal axis of the molecular orientation vector is aligned with the z-axis of the static field, the net magnetic field at the nucleus is given by

$$B = (1 - \sigma_{kk})B_0 \quad (1.18)$$

In isotropic liquid solution, rotational diffusion leads to the averaging of the shielding tensor. Under these circumstances the effects of shielding on a particular nucleus can be accounted for by modifying Equation 1.17 as

$$\omega = -\gamma(1 - \sigma)B_0 \quad (1.19)$$

in which σ is the average, isotropic shielding constant for the nucleus

$$\sigma = \frac{\sigma_{11} + \sigma_{22} + \sigma_{33}}{3} \quad (1.20)$$

The chemical shift anisotropy (CSA) is defined as

$$\Delta\sigma = \sigma_{11} - \frac{\sigma_{22} + \sigma_{33}}{2} \quad (1.21)$$

and the asymmetry of the shielding tensor is defined as

$$\eta = \frac{3(\sigma_{22} - \sigma_{33})}{2\Delta\sigma} \quad (1.22)$$

σ , $\Delta\sigma$, and η are effectively the principal components of the shielding tensor. Variations in σ due to different electronic environments translate into different resonance frequencies of the nuclei. Fluctuations in the local magnetic field as the molecule rotates results in the chemical shift anisotropy (CSA) relaxation mechanism.

Since resonance frequencies are directly proportional to the static field, B_0 , the difference in chemical shift between two resonance signals measured in frequency units increases with B_0 . In addition, the absolute value of the chemical shift of a resonance is difficult to determine in practice because B_0 must be measured very accurately. Therefore, chemical shifts are measured

in parts per million (ppm or δ) relative to a reference resonance signal from a standard molecule,

$$\delta = \frac{\Omega - \Omega_{ref}}{\omega_0} 10^6 = (\sigma_{ref} - \sigma) 10^6 \quad (1.23)$$

in which Ω and Ω_{ref} are the offset frequencies of the signal of interest and the reference signal, respectively. Chemical shift differences measured in parts per million (ppm) are independent of the static magnetic field strength so that, for example, chemical shifts reported from experiments on a 500MHz spectrometer will be the same as those determined on an 800MHz spectrometer.

1.2.4 Secondary chemical shifts

The deviation of a measured chemical shifts δ from their empirical, 'random coil' (r.c.) values δ_{rc} indicates the relative tendency of the polypeptide chain to adopt either helical or extended structures at that position (i) in the primary sequence,

$$\delta_S(i) = \delta(i) - \delta_{rc}(i) \quad (1.24)$$

Thus, by estimating the magnitude of δ_S and comparing it against known values, protein topology can be effectively computed from known experimental resonance assignments. This concept was utilized in one of the most prominent developments in chemical shift analysis, the chemical shift index (CSI) [77]. Furthermore, because chemical shifts are sensitive to structure, even structurally labile regions can be classified, and small propensities to transiently populate canonical types of secondary structure, such as α -helix or β -sheet, can be quantitatively determined. Marsh and co-workers utilized this concept and proposed the structural propensity score (SSP) as a measure of local preferences of a disordered chain to adopt canonical secondary structure, and demonstrated its application in structural characterization of an intrinsically disordered protein family of synucleins [78]. Ultimately, chemical shift information can be used for the assessment of protein flexibility and structural order, as elegantly demonstrated in the study of Berjanskii et al. [79]. Importantly, all of the methods listed above rely heavily on reference 'random-coil' chemical shift libraries, against which experimental data are compared. Multiple approaches have been proposed in order to

provide the most reliable and comprehensive set of 'random-coil' chemical shifts, for example by including nearest-neighbour effects on the backbone ^{15}N chemical shifts in short polypeptides [80], sequence-corrected backbone chemical shift libraries for the polypeptides AcGGXAGNH₂ in 1.0M urea and pH 5 [81], AcGGXGGNH₂ in 8.0M urea and pH 2.3 [82], as well as a more recent compilation for AcQQXQQNH₂ [83]. Alternative 'random coil' chemical shift libraries have been derived from the chemical shifts observed for protein regions which were found to be outside regular secondary structure elements and turns (i.e. assigned as 'coil') as gauged from their PDB structure [84, 85].

This thesis describes the development of chemical shift library compiled from experimental resonance assignments for IDPs [86], and demonstrates, how secondary chemical shifts computed from that library can be used to augment the resonance assignment process and structural characterization for intrinsically disordered proteins.

1.3 Mathematical modelling

Even in the absence of formal mathematical theories, numerical simulations based on experimental observations can be of fundamental importance for drawing inferences of how biochemical systems are organized, function, and are regulated. Previous subsections introduced the concept of random-coil, its mathematical representation and the notion that structural features of ordered and disordered peptides can be quantified through measurements of NMR chemical shifts.

In the forthcoming subsections of this chapter, the consecutive mathematical tools in the numerical modelling of random-coil chemical shifts of disordered proteins will be described. At first, an approach to robust identification of influential observations in experimental chemical shifts will be presented. Three orthogonal measures of data influence will be introduced in the framework of multiple regression modelling. Subsequently, the advantages of singular value decomposition (SVD) in the context numerical analysis will be discussed. A brief algebraic example of a full decomposition will be presented. Ultimately, an SVD-based approach to solving a linear system of equations will be presented.

1.3.1 Influential observations in experimental data

The numerical modelling in natural sciences owes its success to the possibility to relate quantitative hypotheses to experimental observations. However, the correctness, and robustness of even the most simplistic mathematical models depend heavily on the choice of input data. Consequently, a large number of statistical quantities have been proposed to efficiently identify major discrepant points and evaluate influence of individual observations in numerical modelling and analysis. In their seminal review, Chatterjee and Hadi [87], gave a detailed account of the influence measures in numerical analysis, evaluating influence indicators based on: residuals, the prediction matrix, the volume of confidence, influence functions, and partial influence. Although each measure was designed to detect specific phenomenon in the data, they were all closely related, being functions of the basic building blocks in model construction. The most significant finding of the comparative study was an identification of three, orthogonal influence measures: volume of confidence ellipsoids, Welsch-Kuh distance and Cook-Weisberg metrics. Simultaneous analysis of the aforementioned statistical quantities offered a comprehensive view on the influential observations in sparse but highly heterogeneous data sets. The numerical analysis of chemical shifts, and consecutive derivation of ncIDP database, described in detail in the **Chapter 2** of this dissertation, benefited greatly from recursive identification of outliers based on a combination of volume of confidence ellipsoids, Welsch-Kuh distance and Cook-Weisberg methods. A brief description of each method is given below in the context of multiple linear regression model applied in the formulation of ncIDP chemical shift database.

Multiple linear regression model

In order to evince the mathematics behind the orthogonal influence measures, let us consider an overdetermined multiple regression model,

$$Y = X\beta + \epsilon \quad (1.25)$$

where Y is an $N \times 1$ vector of the observables, X is an $N \times p$ full-column rank matrix of known predictors, β is a $p \times 1$ vector of unknown coefficients to be estimated, and ϵ is an $N \times 1$ vector of independent random variables each with zero mean and unknown variance σ^2 . The fitting of the multiple regression

model using the method of least squares can be represented analytically as,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (1.26)$$

where, $\hat{\beta}$ is the estimated parameter vector. The variance of the vector $\hat{\beta}$ can be gauged from,

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (1.27)$$

The vector of fitted values $\hat{Y}$ can be computed as,

$$\hat{Y} = X\hat{\beta} = PY \quad (1.28)$$

where P is the parameter prediction matrix, which relates to the input data X by,

$$P = X(X^T X)^{-1} X^T \quad (1.29)$$

The vector of regression residuals, which provides the most basic measure of model failures, can be computed from,

$$\epsilon = Y - \hat{Y} = (I - P)Y \quad (1.30)$$

where I is an unitary matrix. The variance of model residuals $\text{Var}(e)$ can be related to the parameter prediction matrix P with,

$$\text{Var}(e) = \sigma^2 (I - P) \quad (1.31)$$

Volume of Confidence Ellipsoids

A measure of the influence of the i th observation on the estimated regression coefficients can be based on the change in volume of confidence ellipsoids with or without the i th observation. As suggested by Cook and Weisberg [88], the logarithm of the ratio of the volume of the $(1 - \alpha)100\%$ confidence ellipsoids with and without the i th observation can be used as a measure of influence. The residue specific Cook-Weisberg parameter CW_i can be estimated from

$$CW_i = \frac{1}{2} \log(1 - p_i) + \frac{p}{2} \log \left(\frac{(N - p - 1)F_{\alpha, N-p}}{(N - p - t_i^2)F_{\alpha, N-p-1}} \right) \quad (1.32)$$

where p_i is the i th element of the prediction matrix P defined in Equation 1.29, N is the number of observables, p is a number of coefficients to be

estimated and F_α is the upper α -point of the F-distribution for $N - p$ and $N - p - 1$ degrees of freedom, respectively. The factor t_i^2 can be computed from,

$$t_i^2 = \frac{e_i^2}{\frac{e^T e}{N-p}(1 - p_i)} \quad (1.33)$$

where e_i is the i th element of the vector of residuals e . A large and positive value of Cook-Weisberg metrics indicates that deletion of the i th observation yields a substantial decrease in the volume of the confidence ellipsoids of the model solutions, allowing for a robust identification of influential observations, within the analysed dataset. Conversely, large and negative CW_i values, point to the influential data, which would increase the volume of the confidence ellipsoids.

Welsch-Kuh Distance

The impact of the i th observation on the i th predicted value can be measured by scaling the change in the prediction at position x_i , when the i th observation is omitted, that is,

$$\frac{|\hat{y}_i - \hat{y}_{(i)}|}{\sigma \sqrt{p_i}} = \frac{|x_i(\hat{\beta} - \hat{\beta}_{(i)})|}{\sigma \sqrt{p_i}} \quad (1.34)$$

Welsch and Kuh suggested using a derived $\hat{\sigma}_{(i)}^2$ as an estimate of σ^2 in Equation 1.34, yielding a derivative metric of impact, referred to as Welsch-Kuh Distance parameter [89–91],

$$WK_i = \frac{|x_i(\hat{\beta} - \hat{\beta}_{(i)})|}{\hat{\sigma}_{(i)} \sqrt{p_i}} \quad (1.35)$$

where $\hat{\beta}_{(i)}$ is the estimate of β when the i th observation is excluded from the analysis. The impact analysis in the context of WK_i parameter is performed by computing the aforementioned metrics and comparing it against the calibration point, $2\sqrt{p/N}$. Welsch-Kuh distance values bigger than 1 indicate data points of special attention.

Influence of an observation on a single coefficient

The influence measures discussed insofar assumed that all regression coefficients were of equal interest. However, an observation with a moderate influence on all regression coefficients may be judged more important than one with a large influence on one coefficient and negligible influence on all others. Cook and Weisberg proposed a statistical measure for the impact of the i th observation on a subset of β [88]. A special case of the Cook-Weisberg measure can be derived, which reflects the influence of the i th observation on the j th fitted coefficient.

$$D_{ij} = \frac{t_i^2(p_i - p_{i[j]})}{1 - p_i} \quad (1.36)$$

It was suggested that the values with $|D_{ij}| > 2/\sqrt{N}$ should be treated with special attention [88].

1.3.2 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) presents itself as an extremely valuable tool in the analysis and the solution of problems in mathematical engineering [92]. SVD has been successfully applied in all domains of modern numerical modelling, including genome-wide expression data processing and modelling [93], natural language analysis and latent semantic indexing [94], spectral data processing [95], signal analysis in biomedical applications [96], kinematic and dynamic characteristics of robotic manipulators [97].

As a numerical analysis tool, SVD provides an unifying framework, in which the conceptual formulation of the problem, the practical application, and a numerically robust solution, can be derived in a single algorithmic step. In the context of large-scale dataset analysis, SVD can be considered as a method for transforming correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items. Furthermore, SVD has been proven to excel at identifying and ordering the dimensions along which data points exhibit the most variation. Ultimately, the analytical advantage of singular value decomposition comes from the fact that once the most variable part of the data has been identified, it is possible to find the best approximation of the original data points using

fewer dimensions, thus perform a robust and meaningful data reduction. To illustrate the data-reduction capabilities of singular value decomposition, let us consider a two-dimensional dataset consisting of uniformly distributed random variables, plotted in Figure 1.11a. The regression line applied to the

Fig. 1.11: Graphical demonstration of dimensionality reduction for uniformly distributed random variables.

dataset can be interpreted as the best approximation of the original data. The plotted line minimizes the distance between each data point and its projection on the regression curve. Furthermore, if a perpendicular line would be drawn from each data point to the regression line, the intersection of those lines could be considered as an approximation of the original data. Consequently, a reduced representation of the original data with their intrinsic variability could be obtained. The second regression line, perpendicular to the first, can be drawn, as shown on Figure 1.11b. This line resembles the variability of the original data along the second dimension. With both regression lines drawn, it is possible to generate a set of uncorrelated data points that will expose subgroupings in the original data not necessarily visible at first glance. Hence, the SVD can be seen as a robust tool for transforming a high-dimensional and variable set of data points into a reduced representation that exposes the substructure of the original data more clearly, and orders it according to its variability.

The SVD has proven to be an instrumental tool in the recursive process of derivation of ncIDP chemical shift database described in detail in the **Chapter 2**. The following subsections explain the math behind the decomposition, introduce the concept of a rank of an arbitrary matrix and demonstrate how to use SVD to solve linear systems. A case-specific application of singular

value decomposition to our ncIDP chemical shift library is described in detail in **Chapter 2**.

The mathematics behind SVD

Singular value decomposition is based on a theorem from linear algebra which postulates that a rectangular $m \times n$ matrix A can be broken down into the product of three matrices - an orthogonal $m \times m$ matrix U , a diagonal $m \times n$ matrix S , and the transpose of an orthogonal $n \times n$ matrix V ,

$$A_{mn} = U_{mm} S_{mn} V_{nn}^T \quad (1.37)$$

where $U^T U = I$, $V^T V = I$; the columns of U are orthonormal eigenvectors of AA^T , the columns of V are orthonormal eigenvectors of $A^T A$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order.

One can think of A as a linear transformation taking a vector v_1 in its row space to a vector $u_1 = Av_1$ in its column space. The SVD arises from finding an orthogonal basis for the row space that gets transformed into an orthogonal basis for the column space, $Av_i = \sigma_i u_i$. An orthogonal basis for the row space can be found using the Gram–Schmidt process [98].

The crux of the problem is to find an orthonormal basis $v_1, v_2, \dots, v_r$ for the row space of A for which,

$$\begin{aligned} A \begin{bmatrix} v_1 & v_2 & \dots & v_r \end{bmatrix} &= \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \dots & \sigma_r u_r \end{bmatrix} \\ &= \begin{bmatrix} u_1 & u_2 & \dots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \end{aligned} \quad (1.38)$$

with $u_1, u_2, \dots, u_r$ being an orthonormal basis for the column space of A . Once we add in the nullspaces, equation 1.41 becomes,

$$AV = US \quad (1.39)$$

and the full SVD of matrix A can be achieved by solving for V , U and S in individual steps. Since V is orthogonal, we can multiply both sides of equation 1.39 by $V^{-1} = V^T$,

$$A = USV^T \quad (1.40)$$

Rather than solving equation 1.40 for U , V and S simultaneously, we can multiply both sides by $A^T = VS^T U^T$ to obtain,

$$\begin{aligned} A^T A &= VSU^{-1}USV^T \\ &= VS^2V^T \\ &= V \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} V^T \end{aligned} \quad (1.41)$$

and solve for V by diagonalizing the symmetric positive definite (or semi-definite) matrix $A^T A$. The columns of V are eigenvectors of $A^T A$ and the eigenvalues of $A^T A$ are the values σ^2 . (We choose σ_i to be the positive square root of λ_i .) The same approach can be applied in the solution for U , however this time utilizing multiplication by AA^T .

An end-to-end SVD algorithm will be demonstrated in an example of a small input matrix A ,

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} \quad (1.42)$$

The first step towards the singular value decomposition of matrix A is the computation of a product of the transpose of matrix A^T with A , that is $A^T A$,

$$A^T A = \begin{bmatrix} 4 & -3 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} = \begin{bmatrix} 25 & 7 \\ 7 & 25 \end{bmatrix} \quad (1.43)$$

The eigenvectors of $A^T A$ matrix will give us the vectors v_i , and the eigenvalues will yield the coefficients σ_i . Two orthogonal eigenvectors of $A^T A$ are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The orthonormal basis are obtained from Gram–Schmidt process and are $v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ with eigenvalues $\sigma_1^2 = 32$ and

$\sigma_2^2 = 18$. Consequently, the first parts of our singular value decomposition of matrix A can be written as,

$$\begin{bmatrix} 4 & -3 \\ 4 & 3 \end{bmatrix}^A = \begin{bmatrix} & \\ & \end{bmatrix}^U \begin{bmatrix} \sqrt{32} & 0 \\ 0 & \sqrt{18} \end{bmatrix}^S \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}^{V^T} \quad (1.44)$$

We can solve for the missing factor matrix U , by computing AA^T ,

$$AA^T = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} \begin{bmatrix} 4 & -3 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 32 & 0 \\ 0 & 18 \end{bmatrix} \quad (1.45)$$

whose orthogonal eigenvectors are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ -1 \end{bmatrix}$. Thus the complete SVD of our exemplary matrix A becomes,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \sqrt{32} & 0 \\ 0 & \sqrt{18} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \quad (1.46)$$

1.3.3 Solving linear systems using SVD

Let us consider a linear system,

$$A\vec{x} = \vec{b} \quad (1.47)$$

where A is the parameter matrix, $\vec{x}$ is a vector of independent variables and $\vec{b}$ represents a vector of dependent variables. The general solution to equation 1.47 could be obtained by minimizing

$$|A\vec{x} - \vec{b}| \quad (1.48)$$

but such a linear system may have no solution, or one solution, or even infinitely many solutions.

The uniqueness of the solution to equation 1.47 is dictated by the degree of linear dependence of row and column vectors in the parameter matrix A , which is collectively referred to as the rank of the matrix A .

The rank of a matrix

The rank of an arbitrary matrix A is the maximum number of linearly independent row and column vectors that the considered matrix has. To demonstrate the matrix rank concept, let's consider our exemplary matrix A from equation 1.42,

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} \quad (1.49)$$

$$\text{rank } A = 2 \quad (1.50)$$

Our exemplary matrix A has rank 2, because the first two row vectors are linearly independent, whereas all three column vectors are linearly dependent.

The effective rank of a matrix

An explicit estimation of rank for complex matrices is computationally expensive. Thus a stable measure of matrix rank is obtained by computing $A^T A$ or AA^T , which are symmetric but share the same rank as input matrix A . Based on the accuracy of the data, an arbitrary tolerance can be set, 10^{-6} for example, and singular values of $A^T A$ or AA^T counted above it. The number of singular values above the specified tolerance is the effective rank.

A unique least squares solution through a pseudoinverse

Let's transform the equation 1.47 by multiplying both sides by a transpose of matrix A ,

$$A^T A \vec{x} = A^T \vec{b} \quad (1.51)$$

If A has linearly independent columns and $A^T A$ is invertible then the equation 1.51 has an unique least squares solution given by,

$$\vec{x} = (A^T A)^{-1} A^T \vec{b} \quad (1.52)$$

However, if A does not have full rank, there might be many vectors $\vec{x}$ that minimize $|A\vec{x} - \vec{b}|$. Under such conditions, the optimal solution of the equation 1.47 is a minimum-length vector $\vec{x}^+$ computed from,

$$\vec{x}^+ = A^+ \vec{b} \quad (1.53)$$

where A^+ is a pseudoinverse of matrix A .

Pseudoinverse

The pseudoinverse of matrix A is computed from residual matrices U , S and V obtained in singular value decomposition of A , as introduced in equation 1.37,

$$A^+ = VS^+U^T \quad (1.54)$$

and yields a reciprocal singular values matrix S^+ . The singular values of A , $\sigma_1 \dots \sigma_r$ are on the diagonal of m by n matrix S , whereas the reciprocals of the singular values $\frac{1}{\sigma_1} \dots \frac{1}{\sigma_r}$ are on the diagonal of n by m matrix S^+ from equation 1.54.

As an example, the pseudoinverse A^+ of our matrix A from equation 1.42 is given by,

$$A^+ = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{32}} & 0 \\ 0 & \frac{1}{\sqrt{18}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (1.55)$$

Linear system solution via pseudoinverse

The minimum length solution to the equation 1.47 can be represented as,

$$x^+ = VS^+U^T \vec{b} \quad (1.56)$$

If a classical inverse of matrix A exists then $A^+ = A^{-1}$ and the considered general linear system from equation 1.47 has a unique least squares solution.

For the sake of demonstration let's solve a classical system defined by parameter matrix A from equation 1.42 and an arbitrary observables vector $b = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$,

$$\begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} x = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \quad (1.57)$$

which can be also represented in a standard notation as,

$$4x_1 + 4x_2 = 4 \quad (1.58)$$

$$3x_1 - 3x_2 = 4 \quad (1.59)$$

The solution to such linear problem through pseudoinverse of A can be computed from,

$$x = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{32}} & 0 \\ 0 & \frac{1}{\sqrt{18}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \end{bmatrix} \quad (1.60)$$

and yields $x = \begin{bmatrix} -0.167 \\ 1.167 \end{bmatrix}$, or $x_1 = -0.167$ and $x_2 = 1.167$ in classical notation.

1.4 Overview

Relationships between amino acid sequence and biophysical properties of intrinsically disordered proteins gauged from NMR experiments are investigated in this thesis.

The fundamental link between IDP sequences and NMR observables is introduced in **Chapter 2** where we describe the first inventory of IDP backbone $^1H^N$, ^{15}N , $^1H^\alpha$, ^{13}C , $^{13}C^\beta$, and $^{13}C^\alpha$ chemical shifts compiled from a set of 14 disordered proteins of unrelated sequence and function. We demonstrate that in proteins devoid of persistent secondary structures backbone chemical shifts can be approximated by a linear combination of the actual "random-coil" chemical shift and its correction factors dependent on the identity of its nearest neighbours. Our chemical shift prediction methodology is enhanced by inclusion of spectrum-specific reference offset corrections and automated outlier identification through a multiple linear regression procedure. We use singular value decomposition to find the pseudo-inverse of a parameter

matrix that represents 20 amino acid-specific random coil chemical shifts and 40 sequence-dependent left- and right-neighbour correction factors in a linear system of 6903 equations, which represent the measured ^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^{\alpha}$, ^{13}C , $^{13}\text{C}^{\beta}$, and $^{13}\text{C}^{\alpha}$ chemical shifts. The resulting data constitute the neighbour-corrected IDP chemical shift library (ncIDP). The random coil backbone chemical shifts computed from the primary sequences of 14 IDPs display root-mean-square deviations of 0.65, 0.14, 0.12, 0.50, 0.36, and 0.41 ppm from the experimentally measured values for the ^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^{\alpha}$, ^{13}C , $^{13}\text{C}^{\beta}$, and $^{13}\text{C}^{\alpha}$ chemical shifts, respectively. The prediction accuracy of our "random-coil" chemical shift library is significantly higher than that obtained with libraries for small peptides or "coil" regions of folded proteins suggesting that conformational sampling in these peptides is not representative of intrinsically disordered protein states.

A rapid and accurate prediction of "random-coil" backbone chemical shifts, born out of **Chapter 2**, can be of an immediate importance for the sequential resonance assignment in NMR spectra of disordered proteins, which often display a great level of resonance overlap due to poor signal dispersion. The predictive power of ncIDP library in the context of sequential NMR resonance assignment is utilised in **Chapter 3**. We introduce there an extension for the popular NMR assignment program SPARKY, which facilitates the effective matching of a set of connected resonances to the correct position in the protein sequence by predicting ^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^{\alpha}$, ^{13}C , $^{13}\text{C}^{\beta}$, and $^{13}\text{C}^{\alpha}$ ncIDP chemical shifts and calculating their most plausible resonance assignment scenarios. The relative gain in resonance assignment performance from the use of our extension is demonstrated for the 140-residue intrinsically disordered cytoplasmic domain of human neuroligin-3 (hNLG3cyt). Our approach achieves almost a two-fold improvement in the estimation of chemical shifts for hNLG3cyt with significantly higher detection sensitivity of chemical shift deviations from the sequence-specific "random-coil" values due to resonance misassignment.

Protein backbone chemical shifts can be used as powerful descriptors of local dynamic deviations from the "random-coil" state towards canonical types of secondary structure. These deviations were shown to describe very well functional or dysfunctional protein states, e.g. in adaptive molecular recognition and protein aggregation. As explained in the previous sections of this chapter, secondary chemical shifts are ideally suited for this task, given that appropriate reference data are available, and idiosyncratic sensitivity

of backbone chemical shifts to structural information is treated in a sensible manner. **Chapter 4** demonstrates how the ncIDP chemical shift library can be used to obtain accurate secondary shifts for ordered and intrinsically disordered proteins. The relative benefit of using our "random-coil" chemical shift library in secondary chemical shift calculations is demonstrated for a 140 residue long, natively unfolded human alpha-synuclein (aS). We show that an inclusion of neighbour corrections in the calculation of secondary shifts for aS vastly improves its secondary chemical shifts plots, and yields reliable and smooth structural propensity profiles (SPP) for this disordered polypeptide. Finally, we propose a refinement of an established method of computing SPP by an inclusion of ncIDP random-coil chemical shifts and chemical-shift specific weight parameters, which reflect their idiosyncratic sensitivity to structural information. The newly proposed approach, neighbour-corrected structural propensity calculator (ncPSC), is then applied to NMR data for two well-characterized systems, namely, the binding of human alpha-synuclein to micelles, and light activation of photoactive yellow protein to detect intricate structural changes. These examples spotlight the great power of NMR chemical shift analysis for the quantitative assessment of protein disorder at the atomic level, both in folded and natively unfolded peptides of very different biological roles.

The body of experimental evidence on structural features of IDPs, presented in **Chapter 2** and **4** allows us to formulate a simplistic model of an arbitrary intrinsically disordered protein, as a hetero-polymer in which distances between charged residues are sampled from a Gaussian-like distribution. Under such an assumption, all residual charge-charge interactions within considered system can be enumerated, their interaction strength computed from the linearized Debye-Hückel electrostatics model and collective protonation behaviour predicted in simulated *in silico* titration experiments. **Chapter 5** introduces a robust approach to the comprehensive evaluation of protein electrostatics in unfolded protein states. We demonstrate how the limitations of the previously reported Monte-Carlo approaches in the computation of protein electrostatics can be mitigated using a hybrid approach that effectively combines exact and mean-field electrostatic calculations to rapidly obtain accurate results. The theoretical foundation of our novel approach is discussed in detail. Finally, we demonstrate the relative benefit of our protein electrostatics approach using site-specific protonation data for natively unfolded human alpha-synuclein, obtained in residue-specific NMR pH-titration experiments. The results of our simulations are in excellent

agreement with experimental data demonstrating that our approach, named pepKalc, is capable of evaluating protonation behaviour for an arbitrary-size polypeptide in a sub-second time regime.

Ultimately, all the experimental work from this thesis is concluded with short summary and future directions, disclosed in **Chapter 6**.

ncIDP Database

Sequence-specific random coil chemical shifts of intrinsically disordered proteins

Modified from:

Tamiola K.; Acar B.; Mulder F.A.A.

J. Am. Chem. Soc. **2010**, 132, 18000–3.

DOI: 10.1021/ja105656t

Abstract

Although intrinsically disordered proteins (IDPs) are widespread in nature and play diverse and important roles in biology, they have to date been little characterized structurally. Auspiciously, intensified efforts using NMR spectroscopy have started to uncover the breadth of their conformational landscape. In particular, polypeptide backbone chemical shifts are emerging as powerful descriptors of local dynamic deviations from the "random coil" state toward canonical types of secondary structure. These digressions, in turn, can be connected to functional or dysfunctional protein states, for example, in adaptive molecular recognition and protein aggregation. Here we describe a first inventory of IDP backbone $^1H^N$, ^{15}N , $^1H^\alpha$, ^{13}C , $^{13}C^\beta$, and $^{13}C^\alpha$ chemical shifts using data obtained for a set of 14 proteins of unrelated sequence and function. Singular value decomposition was used to parametrize this database of 6903 measured shifts collectively in terms of 20 amino acid-specific random coil chemical shifts and 40 sequence-dependent left- and right-neighbor correction factors, affording the ncIDP library. For natively unfolded proteins, random coil backbone chemical shifts computed from the primary sequence displayed root-mean-square deviations of 0.65, 0.14, 0.12, 0.50, 0.36, and 0.41 ppm from the experimentally measured

values for the ^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^{\alpha}$, ^{13}C , $^{13}\text{C}^{\beta}$, and $^{13}\text{C}^{\alpha}$ chemical shifts, respectively. The ncIDP prediction accuracy is significantly higher than that obtained with libraries for small peptides or "coil" regions of folded proteins.

2.1 Introduction

In recent years, NMR spectroscopy has proven to be singular in its capacity to study intrinsically disordered proteins (IDPs) with atomic detail [55, 99–103]. Because of the lack of a unique three-dimensional structure, the conformational state of IDPs is described by extensive ensembles derived from a thorough analysis of various experimental data [103–106]. As an alternative to comprehensive structure determination, NMR chemical shifts are of significant value, since they reflect the conformational preferences of polypeptide chains with atomic resolution [77, 107, 108]. Flexible peptides and unfolded proteins display “random coil” chemical shifts, which in turn can be used as a hallmark of disorder. The deviation of a measured chemical shift from its random coil value indicates the relative tendency of the polypeptide chain to adopt either helical or extended conformations at that point in the primary sequence, thereby offering a sensitive and accurate proxy for changes in protein (dis)order and dynamics [78, 107, 109, 110].

Here we describe the first neighbor-corrected random coil chemical shift library for intrinsically disordered proteins, ncIDP, which enables the straightforward and accurate prediction of nuclear shielding constants for a polypeptide sequence.

2.2 Materials and methods

2.2.1 Nomenclature and definitions

Let x , a , and y each be any of the twenty naturally occurring protein amino acids. Then, for each tripeptide $x - a - y$ within the selected Intrinsically Disordered Protein (IDP) entry i , the observed chemical shift for nucleus n

in amino acid a , $\delta^n(x, a, y, i)$ can be expressed as a linear combination of the following terms:

$$\delta^n(x, a, y, i) = \delta_{RC}^n(a) + \Delta_{-1}^n(x) + \Delta_{+1}^n(y) + \epsilon^n(i), \quad (2.1)$$

where n represents: $^1H^N$, $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C^O$, ^{15}N ; $\Delta_{-1}^n(x)$ and $\Delta_{+1}^n(x)$ are the amino acid-specific correction factors for the preceding amino acid x and the subsequent residue y ; $\delta_{RC}^n(a)$ is the mean random-coil chemical shift for nucleus n in the residue a if the residue of interest would be in the *Gly* – a – *Gly* peptide i.e. $x = \text{Gly}$ and $y = \text{Gly}$; and $\epsilon^n(i)$ is the mean offset of all chemical shifts of type n within the IDP entry i to account for alternative referencing. Changes in chemical shifts due to residues beyond the immediate neighbors were not considered, since studies of unstructured peptides have established that these effects are minor [82].

2.2.2 Chemical Shift Database

The input chemical shift dataset was extracted from the BioMagResBank (<http://www.bmrb.wisc.edu>, September 2009) for intrinsically disordered proteins that met the following criteria:

- (1) the polypeptide is disordered under native conditions and studied in buffer without addition of denaturants,
- (2) at least two of the: $^1H^N$, $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C^O$, ^{15}N types of chemical shifts are available,
- (3) IUPAC 1H , ^{13}C , ^{15}N referencing was employed,
- (4) the data were acquired in the temperature range $4 - 30^\circ C$ and pH range $4.0 - 7.5$,
- (5) the protein sequences shared less than 5% sequence identity with other members of the database.

Two further proteins were assigned in-house following standard methods. The assignments of the first (N-terminal) and last (C-terminal) residues of each protein were excluded to avoid possible terminal and charge effects on chemical shifts. In total, 6903 ($^1H^N$ 1331, $^1H^\alpha$ 738, $^{13}C^\alpha$ 1306, $^{13}C^\beta$ 1107,

$^{13}\text{C}^{\text{O}}$ 1092, ^{15}N 1329) chemical shifts were selected for calculations. The complete list of manually selected NMR assignments with their properties is available in Table 2.1.

2.2.3 Determination of correction factors and random coil shifts

For each of the six nuclei $n \in \{^1\text{H}^{\text{N}}, ^1\text{H}^{\alpha}, ^{13}\text{C}^{\alpha}, ^{13}\text{C}^{\beta}, ^{13}\text{C}^{\text{O}}, ^{15}\text{N}\}$ with 20 mean random-coil chemical shifts $\delta_{\text{RC}}^n(a)$, and 40 nearest-neighbor corrections $\Delta_{-1}^n(x)$ and $\Delta_{+1}^n(y)$, we defined a system of linear equations,

$$\begin{bmatrix} \Delta_{-1}^n(A) + \delta_{\text{RC}}^n(H) + \Delta_{+1}^n(K) + \epsilon^n(1) \\ \Delta_{-1}^n(V) + \delta_{\text{RC}}^n(Y) + \Delta_{+1}^n(L) + \epsilon^n(2) \\ \vdots \end{bmatrix} = \begin{bmatrix} \delta^n(A, H, L, 1) \\ \delta^n(V, Y, L, 2) \\ \vdots \end{bmatrix}, \quad (2.2)$$

which was factorized and expressed as,

$$A\Lambda^n(a, i) = \Gamma^n(a, i), \quad (2.3)$$

where A is a binary coefficient matrix so that a direct product of A with the column vector Λ^n yields the left side of Equation 2.2. The correction coefficients and random-coil values for shift type- n are contained within the column vector Λ^n , expressed as,

$$\Lambda^n(a, i) = \begin{bmatrix} \Delta_{-1}^n(x) \\ \delta_{\text{RC}}^n(a) \\ \Delta_{+1}^n(y) \\ \epsilon^n(i) \end{bmatrix}, \quad (2.4)$$

Λ^n in Equation 2.3 is a vector of observed chemical shifts for nucleus type n for all available amino acids a , collected from i IDP entries. The effect of referencing errors in neighbor-correction determination is simultaneously minimized for all included IDP entries by inclusion of entry-specific $\epsilon^n(i)$ terms. The initial solution to Equation 2.3 is obtained by solving,

$$\Lambda^n(a, i) = A^{-1}\Gamma^n(a, i), \quad (2.5)$$

where A^{-1} is a pseudo inverse of a binary matrix A computed by means of SVD (Singular Value Decomposition) programmed in Matlab 2008b (Mathworks Inc, USA).

Tab. 2.1: An overview of manually selected Intrinsically Disordered Proteins used in the compilation of the ncIDP library.

ID	Entry Name	pH	T[°C]	$^1H^N$	$^1H^\alpha$	$^{13}C^\alpha$	$^{13}C^\beta$	$^{13}C^O$	^{15}N
4273	Candidacidal peptide domain of MUC7	4.5	30	11	11	0	0	0	0
4286	Snc1	4.5	15	88	84	0	0	0	88
6078	Aspartic proteinase inhibitor (IA3)	5.6	20	63	55	53	0	49	63
7244	Gamma-synuclein (monomer)	7.4	5	121	0	121	111	121	121
7279	Human high mobility group protein HMGA1	6.0	25	87	44	96	85	73	86
7358	Recombinant form (rmBG21) of the murine golli-myelin basic protein (MBP) isoform BG21	6.8	7	179	169	189	170	179	189
15131	Murine myelin basic protein (MBP)	6.5	4	143	117	161	137	0	143
15136	Endosulfine-alpha	7.5	10	93	0	108	97	91	93
15397	Activation domain of the nuclear hormone receptor coactivator (ACTR)	6.7	31	59	0	60	56	61	59
15409	Zeta-chain of T-cell receptor	6.7	15	102	0	109	97	109	102
15430	Gamma-subunit of cGMP phosphodiesterase (PDE)	4.0	25	73	69	80	70	73	73
15506	Translocation domain of Colicin N	6.8	15	82	0	86	63	82	82
NA1	Neuroigin 3	6.8	25	106	117	125	117	125	106
NA1	Alpha-synuclein	7.4	10	124	72	118	104	129	124
Total				1331	738	1306	1107	1092	1329

2.2.4 Self-consistent optimization protocol

Critical to the computation of $\Delta_{-1}^n(x)$, $\Delta_{+1}^n(x)$, and $\delta_{RC}^n(a)$ is a sufficient number of representative chemical shifts. In order to identify chemical shift outliers within the selected data sets, we designed a self-consistent optimization protocol:

(1) The parameters: $\Delta_{-1}^n(x)$, $\Delta_{+1}^n(x)$, and δ^n are computed as defined in Equation 2.5

(2) a Root Mean Square Deviation (RMSD), defined as,

$$RMSD(\delta_{obs}^n(a, i), \delta_{pred}^n(x, a, y, i)) = \sqrt{\frac{\sum_{i=1, a}^N (\delta_{obs}^n(a, i) - \delta_{pred}^n(x, a, y, i))^2}{N}}, \quad (2.6)$$

is computed, where $\delta_{obs}^n(a, i)$ is the observed chemical shift of nucleus n for amino acid a for an IDP entry i , and $\delta_{pred}^n(x, a, y, i)$ is the calculated chemical shift on the basis of Equation 2.1

(3) predicted chemical shifts $\delta_{pred}^n(x, a, y, i)$ are fitted against the experimental values $\delta_{obs}^n(a, i)$ with a multiple linear regression model [111],

$$Y = X\beta + \alpha, \quad (2.7)$$

where Y is a vector of computed chemical shifts $\delta_{pred}^n(x, a, y, i)$, X is a 2-column matrix composed of experimental chemical shifts $\delta_{obs}^n(a, i)$ and their corresponding predictors set to 1, β is the value of the linear regression coefficient that in the best fit scenario, which should be equal to 1.0, and α is a vector of independent random variables each with zero mean and unknown variance σ^2

(4) an assumption is being made that in a simple linear model defined in Equation 2.1, residuals, computed as

$$\xi^n(a, i) = \delta_{obs}^n(a, i) - \delta_{pred}^n(x, a, y, i), \quad (2.8)$$

follow a normal distribution with zero mean, but with different variances at different values of the predictors. A statistical outlier test, available as

regress in Matlab 2008b, is performed on the residuals. Values that do not follow normal distribution within a $p = 0.05$ confidence interval are marked as outliers and excluded from the analysis. In this way we excluded: 357 $^1H^N$ shifts out of 1331, 259 $^1H^\alpha$ out of 738, 629 $^{13}C^\alpha$ out of 1306, 372 $^{13}C^\beta$ out of 1107, 446 $^{13}C^O$ out of 1092, and 401 ^{15}N out of 1329, yielding a final set consisting of 4439 chemical shifts.

(5) after each outlier exclusion step, 10% of randomly selected chemical shifts are removed, followed by computation of $\Delta_{-1}^n(x)$, $\Delta_{+1}^n(x)$ and $\delta_{RC}^n(a)$. This procedure is repeated 10000 times. The resulting $\Delta_{-1}^n(x)$, $\Delta_{+1}^n(x)$ and $\delta_{RC}^n(a)$ are used to compute standard deviations of the correction factors and random coil chemical shifts in the randomly reduced data sets. The relative magnitude of calculated standard deviations with respect to initial $\Delta_{-1}^n(x)$, $\Delta_{+1}^n(x)$ and $\delta_{RC}^n(a)$ serves as an additional self-consistency check, in which deviations of the order $> 5\%$ point to inhomogeneities in the selected chemical shift data.

Iterative optimization is performed until no outliers are found within the chemical shift data set, and the decrease in computed $RMSD(\delta_{obs}^n(a, i), \delta_{pred}^n(x, a, y, i))$ is smaller than estimated standard deviations of the correction factors. Figure S1 summarizes the self-consistent optimization process, yielding the following RMSD values: $^1H^N$ 0.048 ppm, $^1H^\alpha$ 0.013 ppm, $^{13}C^\alpha$ 0.071 ppm, $^{13}C^\beta$ 0.078 ppm, $^{13}C^O$ 0.089 ppm, ^{15}N 0.294 ppm. Computed random coil chemical shifts $\delta_{RC}^n(a)$, neighbor correction factors $\Delta_{-1}^n(x)$, and $\Delta_{+1}^n(x)$ with their standard deviations and offsets $\epsilon(i)$, are available in Tables 2.2, 2.3, 2.4, and 2.5.

2.3 Results and Conclusions

Using a total of 6903 experimental nuclear shielding constants, we solved Equation 2.1, which states that for each protein entry i , the observed chemical shift of a nucleus $n \in \{^1H^N, ^1H^\alpha, ^{13}C^\alpha, ^{13}C^\beta, ^{13}C^O, ^{15}N\}$ in an amino acid a embedded in the tripeptide sequence $x - a - y$ consists of a random coil reference value δ_{RC}^n , a left-neighbor correction $\Delta_{-1}^n(x)$, and a right-neighbor correction $\Delta_{+1}^n(y)$. The fourth parameter, $\epsilon^n(i)$, is available to account for chemical shift offsets due to alternative referencing and also subsumes systematic deviations due to variations in pH or temperature. A single offset is included for chemical shifts of type n for each entry i . In a first round, the lin-

Tab. 2.2: Amino acid-specific chemical shift Δ_{-1}^n correction values (Corr) and their standard deviations (Stdev), given in ppm.

¹Abnormally high standard deviations for cysteine amino acid are directly related to very low occurrence of this residue within the collected data set yielding ill-conditioned input data matrix, which reflects generally small prevalence of CYS in the sequences of intrinsically unfolded proteins.

²High standard deviations are related to low natural occurrence of tryptophan residues in IDPs.

	$^1H^N$		$^1H^\alpha$		$^{13}C^\alpha$		$^{13}C^\beta$		$^{13}C^O$		^{15}N	
	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev
A	0.088	0.003	-0.046	0.001	0.068	0.006	0.161	0.004	-0.135	0.004	-0.639	0.015
R	0.251	0.003	-0.004	0.001	-0.246	0.008	0.029	0.005	-0.038	0.005	1.909	0.021
D	0.081	0.003	-0.024	0.002	0.059	0.007	0.290	0.005	-0.182	0.006	0.679	0.016
N	0.130	0.003	-0.004	0.003	-0.052	0.008	0.293	0.005	-0.205	0.006	0.699	0.018
C ¹	0.293	0.864	0.025	0.451	-0.417	17.517	-0.006	5.659	0.008	3.664	3.100	12.657
E	0.174	0.003	-0.048	0.002	-0.032	0.006	0.171	0.004	-0.119	0.005	1.226	0.014
Q	0.225	0.003	-0.009	0.002	-0.153	0.008	0.125	0.005	-0.064	0.005	1.700	0.019
G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H	0.141	0.004	-0.019	0.002	-0.243	0.010	0.069	0.006	-0.093	0.007	1.761	0.023
I	0.250	0.003	0.010	0.002	-0.045	0.012	-0.087	0.008	-0.111	0.008	4.205	0.024
L	0.105	0.003	-0.042	0.002	-0.115	0.007	0.083	0.005	-0.143	0.005	0.835	0.017
K	0.174	0.003	-0.031	0.002	-0.117	0.006	0.195	0.004	-0.102	0.004	1.311	0.015
M	0.205	0.005	-0.018	0.002	-0.118	0.009	0.028	0.009	-0.139	0.009	1.407	0.032
F	0.079	0.004	-0.040	0.002	-0.672	0.015	-0.163	0.006	0.059	0.007	2.238	0.025
P	0.292	0.003	-0.032	0.001	-0.040	0.007	0.021	0.005	-0.151	0.005	0.569	0.017
S	0.230	0.003	0.006	0.001	-0.128	0.007	0.155	0.004	-0.153	0.004	2.297	0.013
T	0.211	0.003	-0.005	0.002	-0.130	0.007	0.143	0.005	-0.072	0.004	2.680	0.016
W ²	-0.555	0.616	-0.168	0.042	-0.465	5.155	0.014	0.014	-0.046	0.399	0.795	0.585
Y	-0.052	0.005	-0.035	0.015	-0.633	0.012	-0.245	0.008	0.017	0.008	2.729	0.030
V	0.259	0.003	-0.007	0.002	-0.198	0.007	0.075	0.005	-0.147	0.005	4.507	0.019

Tab. 2.3: Amino acid-specific chemical shift δ_{RC}^n correction values (Corr) and their standard deviations (Stdev), given in ppm.

¹Abnormally high standard deviations for cysteine amino acid are directly related to very low occurrence of this residue within the collected data set yielding ill-conditioned input data matrix, which reflects generally small prevalence of CYS in the sequences of intrinsically unfolded proteins.

²High standard deviations are related to low natural occurrence of tryptophan residues in IDPs.

	$^1H^N$		$^1H^\alpha$		$^{13}C^\alpha$		δ_{RC}^n		$^{13}C^\beta$		$^{13}C^O$		^{15}N	
	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev
A	8.158	0.003	4.224	0.002	178.418	0.007	52.599	0.005	19.102	0.005	123.906	0.016		
R	8.232	0.003	4.239	0.002	176.821	0.008	56.088	0.005	30.691	0.005	121.288	0.019		
D	8.217	0.003	4.537	0.002	176.987	0.009	54.331	0.007	41.089	0.006	120.207	0.020		
N	8.366	0.004	4.632	0.003	175.825	0.009	53.231	0.006	38.790	0.006	118.668	0.022		
C ¹	8.410	0.938	4.447	0.477	174.927	17.588	58.327	6.265	28.085	1.165	119.068	11.986		
E	8.304	0.003	4.222	0.002	177.125	0.007	56.650	0.005	30.225	0.005	120.769	0.016		
Q	8.258	0.003	4.254	0.002	176.510	0.008	55.840	0.006	29.509	0.006	120.224	0.019		
G	8.307	0.003	3.980	0.007	174.630	0.007	45.236	0.005	-	-	108.783	0.017		
H	8.310	0.005	4.585	0.003	175.349	0.011	55.964	0.007	29.719	0.010	118.930	0.030		
I	7.963	0.004	4.076	0.002	176.897	0.010	61.247	0.009	38.563	0.006	120.512	0.024		
L	8.088	0.003	4.260	0.002	178.037	0.008	55.260	0.006	42.212	0.005	121.877	0.020		
K	8.221	0.003	4.237	0.002	177.224	0.007	56.412	0.005	32.921	0.005	121.353	0.018		
M	8.209	0.005	4.425	0.004	176.953	0.015	55.591	0.010	32.690	0.010	120.002	0.031		
F	8.107	0.004	4.573	0.002	176.368	0.011	57.934	0.007	39.660	0.009	120.138	0.029		
P	-	-	4.339	0.003	177.542	0.009	63.180	0.006	32.072	0.005	136.612	0.044		
S	8.215	0.003	4.392	0.002	175.236	0.007	58.352	0.005	63.766	0.005	115.935	0.017		
T	8.047	0.003	4.252	0.002	175.122	0.007	61.926	0.006	69.794	0.005	114.024	0.019		
W ²	7.725	0.711	4.567	0.047	174.549	5.717	57.500	0.668	29.380	0.531	120.733	0.431		
Y	8.026	0.005	4.504	0.003	176.284	0.009	57.761	0.008	38.750	0.008	120.228	0.028		
V	8.037	0.003	4.009	0.002	176.772	0.007	62.347	0.005	32.674	0.005	120.403	0.021		

Tab. 2.4: Amino acid-specific chemical shift Δ_{+1}^n correction values (Corr) and their standard deviations (Stdev), given in ppm.

¹Abnormally high standard deviations for cysteine amino acid are directly related to very low occurrence of this residue within the collected data set yielding ill-conditioned input data matrix, which reflects generally small prevalence of CYS in the sequences of intrinsically unfolded proteins.

²High standard deviations are related to low natural occurrence of tryptophan residues in IDPs.

	$^1H^N$		$^1H^\alpha$		$^{13}C^\alpha$ Δ_{+1}^n		$^{13}C^\beta$		$^{13}C^O$		^{15}N	
	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev	Corr	Stdev
A	-0.056	0.002	-0.042	0.002	-0.816	0.007	-0.187	0.004	0.018	0.004	-0.004	0.014
R	-0.039	0.003	-0.017	0.001	-0.625	0.007	-0.258	0.005	0.046	0.005	-0.065	0.018
D	-0.003	0.003	-0.009	0.001	-0.853	0.007	-0.095	0.005	0.079	0.005	-0.373	0.018
N	-0.023	0.003	-0.033	0.002	-0.849	0.007	-0.067	0.005	-0.013	0.005	-0.219	0.020
C ¹	0.375	0.940	-0.008	0.425	-0.610	17.477	0.047	5.848	-0.170	3.547	0.785	12.764
E	-0.014	0.003	-0.005	0.001	-0.503	0.005	-0.071	0.004	0.065	0.004	-0.100	0.015
Q	-0.016	0.003	-0.039	0.002	-0.551	0.007	-0.098	0.005	0.004	0.005	-0.087	0.018
G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H	-0.136	0.004	-0.059	0.002	-0.677	0.009	-0.248	0.007	0.041	0.006	-0.225	0.026
I	-0.071	0.004	0.000	0.002	-0.559	0.012	-0.129	0.007	0.093	0.009	-0.167	0.023
L	-0.086	0.003	-0.021	0.002	-0.612	0.007	-0.184	0.005	-0.053	0.005	-0.397	0.018
K	-0.053	0.003	-0.011	0.002	-0.520	0.007	-0.142	0.004	0.029	0.004	0.068	0.015
M	-0.066	0.005	-0.018	0.002	-0.271	0.014	-0.061	0.010	-0.205	0.011	-0.145	0.031
F	-0.133	0.005	-0.052	0.003	-0.891	0.008	-0.247	0.007	0.023	0.008	-0.484	0.026
P	-0.028	0.003	0.276	0.002	-2.874	0.011	-2.423	0.006	-0.426	0.008	1.072	0.019
S	-0.002	0.003	0.030	0.001	-0.483	0.006	-0.151	0.004	0.087	0.004	-0.053	0.016
T	0.042	0.003	0.065	0.001	-0.327	0.006	-0.122	0.005	0.086	0.004	0.076	0.017
W ²	0.032	0.070	-0.082	0.136	-0.694	5.831	-0.201	0.440	0.069	0.377	-0.273	0.423
Y	-0.093	0.004	-0.044	0.003	-0.848	0.009	-0.148	0.008	0.019	0.007	-0.387	0.026
V	-0.052	0.003	0.021	0.002	-0.662	0.007	-0.214	0.004	0.044	0.005	-0.047	0.017

Tab. 2.5: Entry-specific offset values ϵ^n calculated during self-consistent SVD analysis. Offset values are given in ppm.
¹ ¹³C^O chemical shifts from BMRB 15506 were excluded from the analysis, due to severe mistakes in the resonance assignments.

ID	ϵ^n					
	¹ H ^N	¹ H ^α	¹³ C ^α	¹³ C ^β	¹³ C ^O	¹⁵ N
4273	0.094	0.125	-	-	-	-
4286	0.000	0.167	-	-	-	-0.436
6078	-	0.069	0.195	0.099	-	-0.394
	0.026					
7244	0.184	-	0.125	0.065	0.15	0.436
7279	-	0.066	-0.063	0.350	0.172	-0.241
	0.066					
7358	-	-0.102	-0.054	-0.121	-0.100	0.315
	0.132					
15131	0.062	-0.001	-	-0.176	-0.166	-0.083
15136	0.041	-	0.076	0.132	0.072	-0.136
15397	-	-	0.015	0.224	0.171	-0.619
	0.203					
15409	-	-	0.135	0.174	0.059	-0.406
	0.018					
15430	0.001	0.104	-0.071	0.040	0.153	-0.291
15506 ¹	-	-	-	0.042	0.017	0.007
	0.001					
NLG3	-	0.066	-0.063	0.000	0.088	-0.426
	0.064					
α-SYN	0.137	0.022	0.032	0.018	0.215	0.346

ear set of Equations 2.1 was solved for the 6903 experimental chemical shifts using singular value decomposition (SVD). The SVD algorithm effectively determined the ncIDP random coil chemical shift library, which comprises the reference chemical shift values of the 20 amino acids a when adjoined by glycine along with the 40 amino acid-specific corrections.

The presence of structure results in local changes in the (ensemble distribution of) bond angles, which are manifested through sequence-dependent deviations from the random coil chemical shifts. For example, the $^{13}\text{C}^\alpha$ chemical shift increases upon formation of α -helix and decreases in the context of a β -strand. On the basis of various types of experimental data, reports in the literature for the IDPs studied here indicate that these polypeptides do not form stable secondary or tertiary structures but sometimes display small segments that attain weakly populated, transient forms of organization. Thus, if accurate random coil chemical shifts are available, the distribution of secondary chemical shifts would be expected to consist of a sharp peak centered at zero for those nuclei present in random coil regions, augmented with broader features arising from segments that exhibit various levels of digression from the random coil state. Figures 2.1 and 2.2 demonstrates that this is indeed what was observed when the ncIDP library was used as a reference set.

Since a portion of the data contains conformational bias away from the random coil state, as gauged from the secondary chemical shifts, we devised a self-consistent optimization protocol based on multiple linear regression (described in the Section 2.2.4) to identify outliers in the experimental data and subsequently eliminate them prior to the derivation of a new, curated ncIDP library from the remaining data. Through this iterative procedure the tails of the secondary chemical shift distributions for all nuclei were eliminated (Figures 2.1 and 2.2). Only the data shown in yellow in Figures 2.1 and 2.2 were finally included in the derivation of the definitive ncIDP library. The values for δ_{RC}^n , $\Delta_{-1}^n(x)$, and $\Delta_{+1}^n(y)$ are given in Tables 2.2, 2.3, and 2.4 together with full details of their derivation. With the ncIDP library, knowledge of the primary sequence of a query protein alone is sufficient to predict its backbone proton, carbon, and nitrogen random coil chemical shifts.

Fig. 2.1: Distribution of secondary chemical shifts for 14 IDPs obtained using the ncIDP library as a reference set for the random coil chemical shift: (a) histogram of the distribution of $^1H^N$ secondary chemical shifts; (b) predicted vs experimentally observed $^1H^N$ chemical shifts. (c) histogram of the distribution of $^1H^\alpha$ secondary chemical shifts; (d) predicted vs experimentally observed $^1H^\alpha$ chemical shifts. (e) histogram of the distribution of $^{13}C^\alpha$ secondary chemical shifts; (f) predicted vs experimentally observed $^{13}C^\alpha$ chemical shifts. Outliers removed by the self-consistent optimization protocol are shown in blue, and the data retained in the curated set are shown in yellow.

In order to validate the statistical robustness of the new reference data set, the ncIDP library was derived several times, leaving out one protein entry at a time, and the eliminated chemical shift data were back-predicted. Some results of these predictions are displayed in Figure 2.11A,B (the predictions

Fig. 2.2: Distribution of secondary chemical shifts for 14 IDPs obtained using the ncIDP library as a reference set for the random coil chemical shift: (a) histogram of the distribution of $^{13}\text{C}^\beta$ secondary chemical shifts; (b) predicted vs experimentally observed $^{13}\text{C}^\beta$ chemical shifts. (c) histogram of the distribution of $^{13}\text{C}^{\text{O}}$ secondary chemical shifts; (d) predicted vs experimentally observed $^{13}\text{C}^{\text{O}}$ chemical shifts. (e) histogram of the distribution of ^{15}N secondary chemical shifts; (f) predicted vs experimentally observed ^{15}N chemical shifts. Outliers removed by the self-consistent optimization protocol are shown in blue, and the data retained in the curated set are shown in yellow.

for all of the chemical shifts for each of the 14 proteins and peptides are given in Figures 2.5–2.10). Figure 2.11(A,B) shows the agreement between the observed and predicted ^{15}N chemical shifts for the IDPs α -synuclein [78] (BMRB ID 7244) and endosulfon- α [112] (BMRB ID 15136), respectively.

The observed strong correlation between the measured and calculated data ($R^2 = 0.99$, $RMSD = 0.62$ ppm and $R^2 = 0.96$, $RMSD = 0.85$ ppm, respectively) demonstrates that accurate neighbor-corrected random coil ^{15}N chemical shifts can be calculated. Figure 2.11(C,D) illustrates this point further. The human prion protein [113] (hPrPC; BMRB ID 4402) is known to have an intrinsically disordered N-terminal domain (data shown in black in Figure 2.11C) while maintaining a well-structured C-terminal part (green data in Figure 2.11C). Indeed, the predicted chemical shifts are clearly at variance with complete disorder ($R^2 = 0.83$, $RMSD = 2.47$ ppm), and the two domains can be readily distinguished. Fully folded proteins display a pattern of even larger deviations, as shown in Figure 2.11D for the helical protein calbindin D_{9k} P43G (BMRB ID 16340) $R^2 = 0.48$, $RMSD = 4.35$ ppm). Evidently, the ncIDP database allows for the accurate detection of protein (dis)order: the root-mean-square difference (RMSD) between the observed and calculated ^{15}N chemical shifts can be used to define the level of structure in a protein (see Figures 2.5-2.10), with IDPs displaying RMSD values less than 1.0. Above this threshold, persistent structure is to be expected. The performance of ncIDP was subsequently benchmarked against alternative random coil chemical shift libraries available in the literature by calculating the RMSDs between the experimental and predicted chemical shifts for 14 unfolded proteins (Table 2.1), eliminating the query entry prior to building a database from the remaining protein entries. Figure 2.3 shows the results of these calculations, averaged over the 14 IDPs. (The individual comparisons are given in Figure 2.4). As a first reference set,

Fig. 2.3: Average chemical shift RMSDs for the following databases and libraries: RefDB [114] (black), Wishart et al. [81] (magenta), Schwarzsinger et al. [82] (green), Wang and Jardetzky [84] (blue), De Simone et al. [85] (orange), and this work (red). It should be noted that no rereferencing procedure was applied to the experimental input data.

we used the RefDB random coil database [114] (black bars), which amends

Fig. 2.4: Individual chemical shift RMSDs for NMR assignments (Table 2.1) used in ncIDP using the following databases and libraries: RefDB [114] (black), Wishart et al. [81] (magenta), Schwarzsinger et al. [82] (green), Wang and Jardetzky [84] (blue), De Simone et al. [85] (orange), and this work (red). It should be noted that no rereferencing procedure was applied to the experimental input data. (*) denotes missing resonance assignment data.

incorrectly referenced chemical shift data submitted to the BMRB [115]. In RefDB, chemical shift averages are reported for regions that do not classify as α -helix or β -sheet, and these are labeled as random coil. Since this database was not developed to predict random coil chemical shifts, it does not utilize the concept of neighbor corrections. Second, we tested the experimental libraries compiled for small synthetic peptides, which are used to mimic polypeptide random coil states. The first library we tested was compiled for Ac-Gly-Gly- α -Ala-Gly-Gly-NH₂ in 1 M urea (pH 5) by Wishart et al. [81] 2.3

Fig. 2.5: The evaluation of $^1\text{H}^N$ chemical shift back-computation using ncIDP for a set of intrinsically disordered proteins, using leave-one-out procedure. NA denotes entries for which no resonance assignments were available.

Fig. 2.6: The evaluation of $^1\text{H}^\alpha$ chemical shift back-computation using ncIDP for a set of intrinsically disordered proteins, using leave-one-out procedure. NA denotes entries for which no resonance assignments were available.

Fig. 2.7: The evaluation of $^{13}\text{C}^\alpha$ chemical shift back-computation using nCDP for a set of intrinsically disordered proteins, using leave-one-out procedure. NA denotes entries for which no resonance assignments were available.

Fig. 2.8: The evaluation of $^{13}\text{C}^\beta$ chemical shift back-computation using ncIDP for a set of intrinsically disordered proteins, using leave-one-out procedure. NA denotes entries for which no resonance assignments were available.

Fig. 2.9: The evaluation of ^{13}C chemical shift back-computation using nCIDP for a set of intrinsically disordered proteins, using leave-one-out procedure. NA denotes entries for which no resonance assignments were available.

Fig. 2.11: Correlation between experimental and computed ^{15}N chemical shifts for (a) α -synuclein, (b) endosulphin- α , (c) human prion protein (hPrPC), and (d) calcium-loaded calbindin D_{9k} P43G.

(magenta bars). For this reference set, it appears that ^1H and ^{13}C chemical shifts in particular are different from those for the other databases, which might result from a conformational bias in the case of Ala. For example, the $^{13}\text{C}^\alpha$ chemical shifts differ by 0.4 ppm on average from the experimental data for the peptides Ac-Gly-Gly-*a*-Gly-Gly-NH₂ in 8 M urea (pH 2.3) measured by Schwarzsinger et al. [82] (Asp and Glu not included). The use of amino acid-specific corrections in the Schwarzsinger database improves the correlation between the predicted and experimental data, as seen on Figure 2.3 (green bars), but the presence of urea and the low pH cause significant offsets for some nuclei, making this reference state less representative for the chemical shifts obtained under native conditions. As an alternative, neighbor-corrected random coil chemical shift databases have been derived from the nuclear shieldings observed for protein regions found to be outside regular secondary structure elements and turns (i.e., assigned as “coil”), as evaluated from the corresponding Protein Data Bank (PDB) structures. We also compared here the predicted chemical shifts utilizing the databases of Wang and Jardetzky [84] (blue bars) and De Simone et al. [85] (orange bars). The newly derived ncIDP library (red bars) clearly demonstrates that the prediction of chemical shifts for natively unfolded proteins benefits most from making use of the experimental data obtained for IDPs as a reference state. It also shows that neighbor corrections are significant for all backbone nuclei.

The detection of local structural propensity in a protein chain is crucially dependent on the availability of a reliable set of reference chemical shifts for the random coil state [77, 78, 103, 105–107, 116, 117]. Figure 2.12 buttresses this point by demonstrating the effect of different random coil libraries on the secondary chemical shifts for two intrinsically disordered protein domains. The first example, shown in Figure 2.12a, focuses on the aggregation-prone repeat regions of the human tau protein. Earlier studies presented conflicting results: one investigation [116] indicated significant α -helical propensity on the basis of the Wishart [81] random coil values, whereas another study [117] suggested that weakly populated β -sheet-like regions are predominant on the basis of a comparison with the random coil chemical shifts of Schwarzsinger et al. [82]. The present results do not identify α -helical regions in the K19 portion of human tau but are in good agreement with the presence of β -sheet propensity at the beginning of each of the repeat regions [117]. It is important to emphasize that this difference is not due to the experimental data obtained in the two studies but results entirely from the use of a different reference state for the calculation of secondary chemical shifts. The agreement with the analysis of Mukrasch et al. [117] holds true at a macroscopic level, but there are also differences observed at the residue level. Much of the residue-to-residue variation seen in the middle panel of the figure is strongly suppressed when using ncIDP (Figure 2.12 bottom). In addition, the ^{15}N secondary chemical shifts shown in Figure 2.12 demonstrate good agreement with those obtained from $^{13}\text{C}^\alpha$ (it should be noted that these secondary chemical shifts are strongly anti-correlated) and support the notion that the imperfect hexapeptide repeats contain a strong signal for aggregation into a cross- β structure.

The possibility of also detecting transitory helical structure is illustrated with a second example involving the kinase inducible domain (KID) of the rat CREB transcription factor. For CREB-KID, it was previously established that the two helices interacting with the KIX domain of the coactivator CBP are already differentially populated in the free form [118]. We present in Figure 2.12c the sequence-dependent secondary chemical shifts for the $^{13}\text{C}^\alpha$ and ^{15}N nuclei of CREB-KID (101-160) calculated using RefDB [114], the database of Schwarzsinger et al. [82], and ncIDP. The predictions made with ncIDP confirm that the first helix ($\text{h}\alpha$:120-129) is significantly populated in the absence of KIX, whereas the second helix ($\text{h}\beta$:134-144) in the complex is better described as a random coil in isolation. Predictions with the program

Fig. 2.12: (a) $^{13}\text{C}^\alpha$ secondary chemical shifts for the K19 construct of the human tau protein (repeat regions R1, R3, and R4) computed according to the protocols from Eliezer et al. [116] (magenta), Mukrasch et al. [117] (green), and this work (red). A -0.15 ppm referencing offset was computed using routines found in Marsh et al. [78]. The position of the hexapeptide essential for aggregation of tau protein is indicated. (b) ^{15}N secondary chemical shifts for human tau protein (repeat regions R1, R3, and R4) computed using ncIDP. (c) $^{13}\text{C}^\alpha$ and ^{15}N secondary chemical shifts for the kinase inducible domain of rat CREB (residues 101-160) computed using RefDB [114] (black), Schwarzinger et al. [82] (green), and this work (red).

AGADIR [119] yield levels of 30 and 1% for the two helices, respectively. The stark difference in helical propensity of the two regions is also consistent with the observation of “helical” amide ($i, i + 3$) NOE connectivities for $\text{h}\alpha$

but not for $h\beta$ [118].

The above examples show that the determination of a representative set of neighbor-corrected random coil chemical shifts for IDPs results in a more continuous pattern of secondary shifts, supporting the reliability of the method. Moreover, although the conformational sensitivity of ^{15}N chemical shifts is well-documented [107], deviations from random coil values appear not to be useful [78] in the absence of neighbor corrections. The improvement obtained in the prediction of random coil ^{15}N chemical shifts obtained here (Figure 2.3) clearly facilitates the use of secondary chemical shifts as a gauge for the conformational state of a protein (Figure 2.11) and the sequence-specific detection of weak signals of transient structure (Figure 2.12).

Finally, any method that utilizes a comparative analysis of protein chemical shifts with respect to a reference random coil state will greatly benefit from the accuracy and reliability offered by ncIDP. This new library can be readily interfaced with available protein chemical shift analysis tools, such as chemical shift index [77] (CSI), structural propensity score assessment for intrinsically disordered proteins [78], and protein structure modeling from chemical shift information [120] (CS-Rosetta).

ncIDP-assign

A SPARKY extension for the effective NMR assignment of intrinsically disordered proteins

Modified from:

Tamiola K.; Mulder F.A.A.

Bioinformatics. 2011, 27, 1039–40.

DOI: 10.1093/bioinformatics/btr054

Abstract

We describe here the ncIDP-assign extension for the popular NMR assignment program SPARKY, which aids in the sequence-specific resonance assignment of intrinsically disordered proteins (IDPs). The assignment plugin greatly facilitates the effective matching of a set of connected resonances to the correct position in the sequence by making use of IDP random coil chemical shifts.

3.1 Introduction

The structural characterization of intrinsically disordered proteins (IDPs) is a rapidly growing field in structural molecular biology. Over the past decade NMR spectroscopy has proven to be singular in its capacity to provide detailed structural characterization of these dynamic entities [106]. Although diverse experimental approaches have been developed [121, 122], the sequential resonance assignment of >10 kDa natively unfolded polypeptides is still not

a trivial task.

Here, we describe an enhanced version of the SPARKY [123] sequence repositioning extension, which assists in matching, connecting and assigning consecutive residues, and is specifically designed for intrinsically disordered proteins: **ncIDP-assign** (neighbor-corrected IDP chemical shift assignment). This tool makes use of a novel random-coil chemical shift library, enabling the accurate prediction of the chemical shifts of a queried protein on a basis of tripeptides [86]. Predicted sequence-specific chemical shifts are used as a template for re-assignment and validation of existing resonance assignments. The newly designed **ncIDP-assign** greatly accelerates the process of sequential resonance assignment of large intrinsically disordered proteins by drastically reducing the level of assignment ambiguities.

3.2 Implementation

The process of assignment validation and repositioning begins with the computation of the sequence-specific random-coil chemical shifts for the intrinsically disordered protein under study, based on its primary sequence. The random-coil chemical shift for a nucleus $n \in \{^1\text{H}^\alpha, ^1\text{H}, ^{13}\text{C}^\alpha, ^{13}\text{C}^\beta, ^{13}\text{C}^O, ^{15}\text{N}\}$ of amino acid residue a , within a tripeptide $x - a - y$, can be expressed as,

$$\delta_{pred}^n(x, a, y) = \Delta_{-1}^n(x) + \delta_{rc}^n(a) + \Delta_{+1}^n(y) \quad (3.1)$$

where $\delta_{rc}^n(a)$ is the "random-coil" chemical shift in the reference peptide $Gly - a - Gly$ and $\Delta_{-1}^n(x)$ and $\Delta_{+1}^n(y)$ are the neighbor corrections due to the preceding and the sequential residue, respectively. Subsequently, a collection of experimentally assigned resonances in a range $k \dots k + l$, specified by the user, is retrieved and compared against the predicted chemical shifts from Eq. 3.1. All plausible locations for the fragment along the sequence of a protein are considered. Each investigated position is characterized with a chemical shift deviation score S ,

$$S(k, l) = \sum_{i=k}^{k+l} \sum_n Z^n(i) \quad (3.2)$$

where $Z^n(i) = \left| \frac{\delta_{exp}^n(i) - \delta_{pred}^n(i)}{\sigma_{pred}^n(i)} \right|$, δ_{exp}^n and σ_{pred}^n are experimental chemical shifts and the standard deviations of expected chemical shieldings for $k \dots k + l$ sequence combination, respectively. A score of 1.0 implies that at the queried sequence position, the experimental shifts are one standard deviation from the computed values, on average. Hence, the "best-fit" solutions are characterized by the lowest S values.

3.3 Results and Conclusion

To assess the performance of **ncIDP-assign** we used the NMR chemical shift assignments for the 140-residue intrinsically disordered, cytoplasmic domain of human neuroligin-3 [124] (hNLG3cyt). The robustness of the assignment procedure was established using the data obtained in ^1H - ^{15}N HSQC [125], COCAHA [126] and HNCACB [127] experiments, providing access to backbone and $^{13}\text{C}^\beta$ chemical shifts. Table 3.1 displays the level of completeness of the resonance assignment for hNLG3cyt, and the accuracy of the chemical shift back-computation modules in the standard and ncIDP versions of the SPARKY sequence repositioning plugins. As borne out by Table 3.1,

Tab. 3.1: Comparative analysis of chemical shift back-computation for hNLG3cyt using the standard chemical shift computation module available in SPARKY, and **ncIDP-assign**. Root mean square difference (RMSD) values are given in ppm. *Chemical shift RMSD computed after removal of mean systematic offsets between the computed and experimental resonance assignments for hNLG3cyt in order to minimize chemical shift referencing errors.

Nucleus (n)	Standard*	ncIDP*	Assignments
$^1\text{H}^\alpha$	0.122	0.044	120
$^1\text{H}^N$	0.148	0.111	128
$^{13}\text{C}^\alpha$	0.810	0.324	128
$^{13}\text{C}^\beta$	0.410	0.212	120
$^{13}\text{C}^O$	0.717	0.393	131
^{15}N	1.314	0.664	129

ncIDP-assign offers an almost two-fold improvement in the estimation of chemical shifts for hNLG3cyt. The superior predictive power of **ncIDP-assign** translates into detection sensitivity of chemical shift deviations from the sequence-specific "random-coil" values due to resonance misassignment. This point is demonstrated by Fig. 3.1 where the performance of the two methods

Fig. 3.1: Comparative analysis of accuracy of standard and ncIDP SPARKY sequence repositioning extensions, as a function for fragment length, using resonance assignments in: ^1H - ^{15}N HSQC, COCAHA and HNCACB experiments, for the intrinsically disordered protein hNLG3cyt. **(A)** Normalized frequency of correct solutions as a function of fragment length in: standard (blue) and ncIDP-enhanced (green) repositioning, respectively. **(B)** Sensitivity as a function of sequence length for: standard (blue) and ncIDP (green) repositioning extensions

is compared against known resonance assignments for hNLG3cyt. **ncIDP-assign** identifies correct sequential assignment solutions much more readily, and at the level of dipeptides for the considered experiments (Fig. 3.1A). Further expansion of the length of a query fragment to tripeptide rapidly increases the probability of assigning the correct solution. Consequently, the information content contained within a combination of resonance frequencies in short peptides (length2) is unique enough to make the correct position guess in most cases.

Another critical parameter in the assignment process is the relative separation of the "best-fit" score S_{best} with respect to the second-best scoring suggestion $S_{second-best}$, expressed here as the sensitivity $S_{second-best}/S_{best}$ (Fig. 3.1B). Values for the sensitivity close to 1.0 indicate ambiguity. Given a sequentially assigned dipeptide, **ncIDP-assign** generates a list of solutions in which the "best-fit" scenario scores appreciably better than the next considered option. Already significant improvements are observed in the analysis of ^1H - ^{15}N HSQC resonance lists, clearly demonstrating that information content

of ^1H - ^{15}N resonance pairs in sequentially connected dipeptides can be effectively used in the assignment process of an intrinsically disordered protein.

In conclusion, we have shown here that ***ncIDP-assign*** is an effective tool to aid the sequential NMR resonance assignment of (intrinsically) disordered proteins. The source code of ***ncIDP-assign*** is available to download at <http://nmr.chem.rug.nl/wordpress/index.php/downloads/ncidp-assign/>.

Acknowledgement

Renee Otten (University of Groningen) and Thomas Goddard (University of California, San Francisco) are gratefully acknowledged for help with Python code development and valuable suggestions.

ncSPC

Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins

Modified from:

Tamiola K.; Mulder F.A.A.

Biochem. Soc. Trans. **2012**, 40, 1014–1020.

DOI: 10.1042/BST20120171

Abstract

NMR spectroscopy offers the unique possibility to relate the structural propensities of disordered proteins and loop segments of folded peptides to biological function and aggregation behaviour. Backbone chemical shifts are ideally suited for this task, given that appropriate reference data are available, and idiosyncratic sensitivity of backbone chemical shifts to structural information is treated in a sensible manner. We describe here methods to detect structural protein changes from chemical shifts, and present an online tool (ncSPC), which unites aspects of several current approaches. Examples of structural propensity calculations are given for two well-characterized systems, namely, the binding of alpha-synuclein to micelles, and light activation of photoactive yellow protein. These examples spotlight the great power of NMR chemical shift analysis for the quantitative assessment of protein disorder at the atomic level, and further our understanding of biologically important problems.

4.1 Introduction

The fact that intrinsically disordered proteins are widespread in Nature, and sustain various functions, advances a newfangled view on structural biochemistry, expanding far beyond the traditional 'structure-function' paradigm [59]. Intrinsically unstructured/disordered proteins (IUPs/IDPs) are known for the lack of a well-defined tertiary structure and exist as a multitude of conformations that dynamically inter-convert in time [58]. Their abundance puts forward the notion that protein dynamics has evolved for the adaptive benefit of higher organisms, and advances a 'systems view' on protein interaction networks [10]. Intrinsically disordered proteins control many important biological processes, effectively complementing the functional spectrum of ordered proteins [5]. The evaluation of the "flavours" of disorder observed in IDPs [9] demonstrates that structural lability can help these entities to fulfil their functions in various ways. Interestingly, disordered segments of large proteins often act as flexible linkers between folded subunits in multidomain proteins. Alternatively, many disordered proteins function by specific binding to a downstream partners, DNA or RNA. This phenomenon, referred to as coupled folding and binding involves a temporary excursion to a more ordered state with a structurally defined binding interface [128]. Abstracting from their fundamental biological role, numerous IDPs are implicated as causative agents in devastating human diseases, most notably in the area of neurodegeneration [129]. Thus, there is a strong need to develop our understanding of the functional role of labile protein conformations in cellular processes, as well as their involvement in protein aggregation, oligomerization and fibril formation.

As explained in the seminal works of Dyson and Wright [66] and Dobson [58], dynamic disorder severely limits protein three-dimensional (3D) structure determination using X-ray crystallography, rendering our knowledge of the conformational state of intrinsically disordered proteins rudimentary. However, among the available experimental approaches, NMR spectroscopy has proven to be singular in its capacity to study disordered proteins with atomic detail, both in the solid and solution state [6]. Due to the lack of a defined 3D structure, the conformational state of IDPs is described by extensive ensembles, derived from a thoroughgoing analysis of various experimental data [102]. As an alternative to comprehensive NMR structure determina-

tion, which for large ($> 25\text{kDa}$) proteins may be a tedious task, it would be of significant value to identify a simple, sensitive and accurate proxy for changes in protein (dis)order and dynamics, and relate this to functional consequences. For example, what is the immediate structural consequence of a protein's contact with a membrane, or a mutation? NMR chemical shifts are perfectly suited to help answering such questions, since they reflect the conformational preferences of polypeptide chains with atomic resolution [107], and display exquisite sensitivity to dynamics [79, 109]. Furthermore, chemical shifts are easy to measure experimentally, and can be efficiently assigned to individual atoms in the protein molecule.

The advancement of experimental NMR spectroscopy of proteins stimulated the development of computational tools that utilize chemical shifts in the secondary structure assignment. Since the deviation of a measured chemical shift from its 'random coil' (r.c.) value indicates the relative tendency of the polypeptide chain to adopt either helical or extended structures at that point in the primary sequence, protein topology can be effectively computed from known experimental resonance assignments. One of the most prominent developments in chemical shift analysis is the chemical shift index (CSI) [77]. Furthermore, because chemical shifts are so sensitive to structure, even structurally labile regions can be classified, and small propensities to transiently populate canonical types of secondary structure, such as α -helix or β -sheet, can be quantitatively determined. Marsh and co-workers utilized this concept and proposed the structural propensity score (SSP) as a measure of local preferences of a disordered chain to adopt canonical secondary structure, and demonstrated its application in structural characterization of an intrinsically disordered protein family of synucleins [78]. Ultimately, chemical shift information can be used for the assessment of protein flexibility and structural order, as elegantly demonstrated in the study of Berjanskii et al. [79]. Importantly, all of the methods listed above rely heavily on 'random-coil' chemical shift libraries, against which experimental data are compared. Multiple approaches have been proposed in order to provide the most reliable and comprehensive set of 'random-coil' chemical shifts, for example by including nearest-neighbor effects on the backbone ^{15}N chemical shifts in short polypeptides [80], sequence-corrected backbone chemical shift libraries for the polypeptides AcGGXAGNH₂ in 1.0M urea and pH 5 [81], AcGGXGGNH₂ in 8.0M urea and pH 2.3 [82], as well as a more recent compilation for AcQQXQQNH₂ [83]. Alternative 'random coil' chemical shift

libraries have been derived from the chemical shifts observed for protein regions which were found to be outside regular secondary structure elements and turns (i.e. assigned as 'coil') as gauged from their PDB structure [84, 85]. Our recent study of a novel chemical shift library compiled of experimental resonance assignments for IDPs, ncIDP, demonstrates that secondary chemical shift computation will benefit the most from the use of r.c. shieldings derived from disordered proteins under native conditions [86].

4.2 Materials and Methods

With better predictive power of the chemical shifts of the 'random coil' state, it is now possible to more accurately detect local propensities for structure formation both for natively unfolded polypeptides, as well as for ordered proteins with disordered regions. We use here experimental resonance assignments for the well-characterized IDP human α -synuclein [78, 130–134] to demonstrate this fact.

4.2.1 Sequence-specific 'random-coil' chemical shifts

For the reference 'random-coil' state, we use here the ncIDP chemical shift library, compiled specifically for intrinsically disordered proteins [86], in which the random-coil chemical shift for a nucleus $n \in \{^1\text{H}^N, ^1\text{H}^\alpha, ^{13}\text{C}^\alpha, ^{13}\text{C}^\beta, ^{13}\text{C}^O, ^{15}\text{N}\}$ of amino acid residue a , within a tripeptide $x - a - y$, is expressed as

$$\delta_{calc}^n(x, a, y) = \Delta_p^n(x) + \delta^n(a) + \Delta_n^n(y) \quad (4.1)$$

where $\delta^n(a)$ is the 'random-coil' chemical shift in the $G - a - G$ reference sequence, and $\Delta_p^n(x)$ and $\Delta_n^n(y)$, are the neighbour corrections due to preceding ('p') and next ('n') residue, respectively.

The sequence-dependent deviations of experimental resonance assignments from the 'random coil' chemical shifts, are known as secondary chemical shifts, and report on the increased sampling of compact (e.g. α -helix) or extended (e.g. β -strand) structures. Figure 4.1 displays the $^{13}\text{C}^\alpha$ and ^{15}N secondary chemical shifts of human α -synuclein, using the RefDB [114] random coil

Fig. 4.1: Secondary chemical shifts of α -synuclein. (left) computed using RefDB, and (right) IDP-based 'random-coil' chemical shift library (**ncIDP**).

chemical shifts as utilized by Marsh et al. [78] (left), and those calculated from the ncIDP 'random coil' database (right) [86]. The differences in Figure 4.1 mainly arise from the absence of neighbour corrections in RefDB, which are particularly large in the case of ^{15}N [80, 86], and are also responsible for the significant residue-to-residue variation of the $^{13}\text{C}^\alpha$ chemical shifts. Using the more accurate IDP 'random coil' database correctly identifies most of the protein to be a 'random coil' (right), whereas the neglect of sequence sensitivity of chemical shifts yields less reliable, or even biased estimates of the conformational state of a disordered protein (left).

As borne out by Figure 4.1, the success of assigning local conformational preferences is predicated on the fact that backbone and side chain $^{13}\text{C}^\beta$ chemical shifts demonstrate a strong and idiosyncratic sensitivity to secondary structure. However, chemical shifts depend on structure in different ways. For example, the reliability to discriminate α -helix from 'coil' decreases in the order $^{13}\text{C}^\alpha > ^{13}\text{C}^\circ > ^1\text{H}^\alpha > ^{13}\text{C}^\beta > ^{15}\text{N} > ^1\text{H}^\text{N}$, whereas the sensitivity to identify β -sheet structure follows the order $^1\text{H}^\alpha > ^{13}\text{C}^\beta > ^1\text{H}^\text{N}, ^{15}\text{N}, ^{13}\text{C}^\alpha, ^{13}\text{C}^\circ$ [84]. Consequently, the most reliable predictions can be made using a properly weighted combination of secondary chemical shifts.

4.2.2 Structural Propensity Calculations

The Secondary Structure Propensity (SSP) algorithm was proposed by Marsh et al. [78], and used to explain the fibrillation propensities of human α - and γ -synuclein. In this procedure, the calculated secondary chemical shift for a particular nucleus is divided by that expected in the case of fully formed secondary structure, and yields a residue-specific score between -1.0 (β -sheet) and $+1.0$ (α -helix). In their analysis of synucleins $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, and $^1\text{H}^\alpha$

appeared to be most useful, whereas the sensitivity of ^{15}N chemical shifts to peptide conformation may have been missed due to lack of sequence consideration.

In this paper we refine the secondary structure propensity score of Marsh et. al [78]. As in the original implementation, the contributions of different chemical shifts to the total residue-specific SSP score $\Psi(k)$ are weighted by their uneven sensitivity to α - or β -structure, but, in addition, weights are defined that reflect their relative sensitivity to backbone conformation (*vide infra*). Eq. 4.2 demonstrates how the SSP score for a residue k in the sequence of amino acids $x - a - y$, is then computed:

$$\Psi(k, w) = \frac{\sum_n \sum_{j=k-w}^{k+w} \text{sign} \times \theta_{SS}^n \times \frac{\Omega_{obs}^n(j, x-a-y)}{\sigma_{SS}^n(j, a)}}{\sum_n \sum_{j=k-w}^{k+w} \theta_{SS}^n \times \frac{\Omega_{SS}^n(j, a)}{\sigma_{SS}^n(j, a)}} \quad (4.2)$$

$\Omega_{obs}^n(j, x - a - y)$ is the neighbour corrected secondary chemical shift for amino acid a at the position j in the sequence, computed as a difference between the experimental and the predicted shift from the Eq. 4.1, $\Omega_{SS}^n(j, a)$ is the tabulated average secondary chemical shift of fully formed secondary structure (α or β) for amino acid a , taken from RefDB [114]. Both terms are weighted by a standard deviation $\sigma_{SS}^n(j, a)$ of $\Omega_{SS}^n(j, a)$ as defined in RefDB. The parameter θ_{SS}^n reflects the relative sensitivity of the chemical shift n to secondary structure of type $SS = \{\alpha, \beta\}$. Normalized values of θ_{SS}^n are given

Tab. 4.1: Normalized weight parameters θ_{SS}^n , reflecting relative sensitivity of chemical shifts to the canonical secondary structures. θ_{SS}^n are given in arbitrary units.

Nucleus (n)	α -helix	β -sheet
$^1\text{H}^N$	0.15	0.30
$^1\text{H}^\alpha$	1.00	1.00
$^{13}\text{C}^O$	0.5	0.25
$^{13}\text{C}^\alpha$	1.00	1.00
$^{13}\text{C}^\beta$	1.00	1.00
^{15}N	0.125	0.250

in Table 4.1 and are based on a principal component analysis of the various contributions to backbone and $^{13}\text{C}^\beta$ chemical shift, performed by Neal et al. [107]. The SSP score is calculated from a comprehensive set of $^1\text{H}^N$, $^1\text{H}^\alpha$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}^O$, and ^{15}N chemical shifts. Eq. 4.2 may include a user-defined weighted average over $2w+1$ residues, where w is the size of a sliding window,

to identify consistent trends in the residual secondary structure around the investigated position k , as often used for the CSI. Secondary structure type discrimination in Eq. 4.2 is performed using a test in which the sign of the product of $\Omega_{obs}^n(j, x - a - y)$ and $\Omega_{SS}^n(a)$ is calculated, as in Eq. 4.3.

$$\Omega_{SS}^n(j, a) = \begin{cases} \Omega_{\alpha}^n(j, a) \wedge sign = 1 & \text{if } \Omega_{obs}^n(j, x - a - y) \times \Omega_{\alpha}^n(j, a) > 0 \\ \Omega_{\beta}^n(j, a) \wedge sign = -1 & \text{if } \Omega_{obs}^n(j, x - a - y) \times \Omega_{\beta}^n(j, a) > 0 \end{cases} \quad (4.3)$$

Consequently, residual SSP scores of 1.0 and -1.0 reflect fully formed α - or β -structure, respectively, whereas a score of 0.5 suggests that 50% of the conformers in the ensemble are helical at that position.

4.3 Results

Being aware of the indispensable role of the 'reference' random-coil state in secondary structure assignment, along with the idiosyncratic sensitivity of the chemical shifts to structural features of protein backbone, we implemented the rationale developed above into a neighbour-corrected Structural Propensity Calculator, **ncSPC**, which uses NMR chemical shift data as sole input to determine the molecular conformation of proteins, both in disordered and ordered state. The ncSPC algorithm is expected to provide reliable propensity calculations, as (1) the calculation of 'random coil' chemical shifts using the IDP 'random coil' database naturally includes these sequence corrections, and greatly improves the accuracy and completeness of secondary chemical shift calculations and (2) numerically optimized weight factors for the secondary chemical shifts of the various nuclei afford the comprehensive and accurate calculation of structural propensities. Thus, **ncSPC** can detect and classify areas of disorder more reliably than currently available methods because it can better predict 'random coil' chemical shifts of disordered proteins, and weigh the contributions of the different nuclei to the propensity score by the relative sensitivity of chemical shifts to secondary structure effects.

The use of chemical shifts as proxies for protein conformational state is now illustrated using two examples, for which structural rearrangements can be readily linked with their biological function: (1) detection of structural propensities in the intrinsically disordered human protein α -synuclein under

native and membrane mimicking conditions; (2) the structural transition of the bacterial photoactive yellow protein (PYP) upon light excitation.

4.3.1 Example I: Structural characterization of the intrinsically disordered protein α -synuclein under native and membrane-mimicking conditions

Human α -synuclein is a small, intrinsically disordered protein of 140 amino acids that has attracted great interest in recent years [130, 131, 133, 135, 136]. Within neurons, α -synuclein is particularly localized in the nucleus and in the pre-synaptic terminals [137] where it is loosely associated with the distal pool of synaptic vesicles [138] and known to interact with the lipid rafts of the plasma membrane [139]. The propensity of α -synuclein to interact with neuron membranes *in vivo* was confirmed by extensive *in vitro* studies, including binding to membrane mimicking agents [140, 141]. At synapses, however, the distribution of the protein is highly dynamic and directly related to neuronal activity, with detectable, rapid exchanges taking place among neighbouring synapses [142]. α -Synuclein is directly implicated in human neuropathology. In the brain of patients with neurodegenerative disorders, termed α -synucleinopathies, the protein is known to undergo conformational transitions yielding deposits of proteinaceous material [129, 143]. Consequently, at the onset of neurodegenerative processes, malfunction of synapses develops.

Much effort has been put in understanding the fundamental causes of α -synuclein aggregation and the abrupt loss of functional membrane interactions. Several biophysical methods, including NMR [78, 132, 134, 144–146], EPR [147, 148], and fluorescence spectroscopy [149], have provided valuable insights into the structural features of natively disordered and membrane-bound α -synuclein. Figure 4.2A shows the result of a structural propensity calculation from NMR chemical shift data, using the ncSPC web-tool. The program detects enhanced β -propensity only in the C-terminal region, residues 105–140. The extended conformation of the C-terminus was confirmed from NMR residual dipolar couplings in an independent study and shown to result from electrostatic forces [144]. In contrast, as shown in Figure 4.2B–C, in membrane-mimic states of the physiological situation at nerve termini, residues 1–97 form a broken, loosely ordered helical structure

Fig. 4.2: Structural interpretation of Neighbour-Corrected Structural Propensity Score (ncSPC) for pure and micelle-bound α -synuclein. (A) Propensity plot for pure α -synuclein (upper panel), and α -synuclein in the presence of 50 mM SDS (lower panel). (B) Structural interpretation of **ncSPC** score in the context of micelle-bound NMR model of α -synuclein (PDB ID: 1XQ8). The positions of structural "kinks" in the helices are annotated with green circles. (C) Schematic representation of partially folded α -synuclein upon the binding to SDS micelle (grey). The cartoon representation of the protein is color-coded according to the **ncSPC** score. (calculations performed using: $^1\text{H}^\alpha$, $^1\text{H}^N$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, and ^{15}N chemical shifts from in-house assigned α -synuclein and BMRB 5744).

while maintaining disorder in the C-terminal domain [148]. Moreover, the regions of decreased α -helical propensity can be reliably assigned to the pres-

ence of structural "breaks" in well-defined α -helices upon the contact with SDS micelles, as gauged from RDC measurement by Ulmer and co-workers [150]. The **ncSPC** assignment of structural propensity for α -synuclein is in line with available structural data, offering the reliable detection of small but meaningful structural features of an intrinsically disordered protein.

4.3.2 Example II: Light activation of Photoactive Yellow Protein

The second example concerns the blue light photoreceptor Photoactive Yellow Protein (PYP), involved in bacterial phototaxis, and a model for the action of Per/Arnt/Sim (PAS) signal transduction domains [151]. PYP is a 14 kDa water-soluble protein, which contains a thioester linked p-coumaric acid cofactor [152]. Upon light excitation, PYP undergoes a significant conformational change, leading to the dissolution of native structure around the chromophore [153]. Although the 3D structure of the photocycle ground state was solved using NMR spectroscopy, the light-activated 'signaling' state has been elusive, showing ubiquitous line broadening as a result of slow (ms) protein conformational interconversion [153]. We analyze here the truncation mutant PYP Δ 25 which displays a photocycle that is very similar to wild type, and for which near-complete heteronuclear NMR resonance assignments are available [154]. Using the chemical shift data alone, we are able to detect and interpret the structural consequence of blue light irradiation on PYP, as shown in Figure 4.3: the helix at the PAS core domain, comprised of residues *AAEGDIT*, dissolves and forms a region of extended structure. At the same time an adjacent region with β -sheet propensity becomes further consolidated. Our analysis demonstrates that the long-lived intermediate of the PYP Δ 25 photocycle in solution exhibits structural and dynamic disorder. Consequently, exposure of a new surface isotope can mediate signal transduction through protein-protein interaction with a downstream partner. The structural change described above then marks the start of the response of the bacterium to evade harmful blue light.

4.4 Conclusion

Structural changes in adaptable regions of proteins are widespread, and have large significance in biology. However, they often remain elusive to the 'standard' tools of structural biochemistry. The above examples demonstrate that, given a judiciously compiled reference set of 'random coil' chemical shifts and a simple propensity calculation algorithm, NMR spectroscopy is able to detect meaningful conformational shifts in disordered proteins or labile regions of folded proteins. These conformational rearrangements harness biological function, or distinguish benign from malign cellular protein interactions. The development of a web tool, that can instantly perform such analyses on any query protein for which assigned chemical shift data are available, will facilitate its use in the structural biology community. The neighbour-corrected structural propensity calculator (**ncSPC**) is available as a WW3-compliant application and can be accessed under: <https://www.protein-nmr.org>. The structural propensity computations are divided into three stages. The first phase is data upload. At this stage experimental resonance assignments can be retrieved from the BioMagResBank or uploaded from a local workstation as a STAR 2.x/3.x file. Since the correct chemical shift referencing is substantial in the analysis, manual or ^{13}C automatic reference offset corrections, according to method described by Marsh and co-workers, can be performed upon upload [78]. The presence of deuterium isotope shifts for uniformly deuterated proteins [155] can be accounted for. Subsequently, a choice of 'random-coil' chemical shift library, with 5 commonly available 'random-coil' tables implemented with their neighbour-corrections is made [80, 82, 84, 86]. The last step transfers the user to a selection of features divided into three major groups: predictive, statistical, and structural propensity computation. The calculations are concluded with short summary of selected options and a list of downloadable ASCII and graphic files. The data and graphics presented in this paper were prepared with **ncSPC**.

Fig. 4.3: Investigation of secondary structure propensity changes in ground and photoactivated states of PYP Δ25 (A) Propensity plot for ground state PYP Δ25 (upper panel), and PYPΔ25 after photoactivation (lower panel). (B) Structural interpretation of ncSPC score in the context of NMR model of ground state PYP Δ25 (PDB code 1XFN). The position of structural rearrangement is indicated in yellow. The position of p-coumaric acid is marked with a green circle. Calculations performed using $^1\text{H}^\alpha$, $^1\text{H}^N$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and ^{15}N chemical shifts from BMRB 6321 and 6322 respectively.

pepKalc

Scalable and comprehensive calculation of electrostatic interactions in random coil polypeptides

Modified from:

Tamiola K.; Scheek R.M.; van der Meulen P.; Mulder F.A.A.

Bioinformatics **2018**, *bty033*

DOI: [10.1093/bioinformatics/bty033](https://doi.org/10.1093/bioinformatics/bty033)

Abstract

Polypeptide sequence length is the single dominant factor hampering the effectiveness of currently available software tools for *de novo* calculation of amino acid-specific protonation constants in disordered polypeptides. Thus, we have developed **pepKalc**, a robust simulation software for the comprehensive evaluation of protein electrostatics in unfolded states. Our software completely removes the limitations of the previously reported Monte-Carlo approaches in the computation of protein electrostatics by using a hybrid approach that effectively combines exact and mean-field calculations to rapidly obtain accurate results. Paired with a modern architecture GPU, **pepKalc** is capable of evaluating protonation behavior for an arbitrary-size polypeptide in a sub-second time regime.

5.1 Introduction

Protonation is a ubiquitous and important process in biology. Protein folding, ligand recognition, enzyme catalysis, membrane potentials, and the energetics of cells depend on ionization and proton transfer [156]. Thus, great theoretical and experimental effort have been devoted to the elucidation of the intricate workings of protonation-related electrostatics at the molecular level [157–159].

In their seminal work, Tanford and Kirkwood introduced a theoretical framework for the treatment of protein electrostatics in solution [160]. The "hard-sphere" model of protein electrostatics, proposed by Tanford and Kirkwood, was further improved by an inclusion of a mean-field approximation of pair-wise interactions in a polypeptide chain [161]. Furthermore, the developments in X-ray crystallography and advancements in computational structural biology opened a path to a new class of protein electrostatics models. Das et al. showed that a continuum electrostatics with distance-distributions derived from the structural ensembles of proteins could be used to compute the site-specific protonation behavior in folded polypeptides [162]. The continuum model got further improved by an inclusion of protein flexibility and local dynamics gauged from molecular dynamics sampling [163]. Although significant progress in the theoretical treatment of the protonation behavior of folded polypeptides has been made, a complete characterization of electrostatic interactions in folded proteins still remains a formidable task [164]. Incomplete treatment of protein dielectric properties and rudimentary conformational sampling were implicated as crucial factors hampering the elucidation of protonation behavior at the amino acid level. The experimental study of RNase T1 by Bombarda and Ullmann elegantly demonstrates the global challenges in predicting and explaining the intricate protonation behavior of amino acid residues in folded proteins [164]. Their study suggested that the interactions between protonatable residues in proteins can help to maintain the energy required to protonate a site in the protein nearly constant over a wide pH range, suggesting an existence of coupled networks of residue-residue interactions.

In addition, numerous successful protonation-state prediction studies have been reported for purposefully unfolded and intrinsically disordered proteins (IDPs) [165–169]. It was demonstrated that elucidation of electrostatic in-

teractions in polypeptides devoid of canonical structural features benefits greatly from the adoption of a Gaussian-chain, as a model for residue-to-residue distances in a structurally disordered state [166]. Unfortunately, the Monte-Carlo based free energy integration, presented by Zhou was successful only at reproducing experimental data for small-size polypeptides [167]. The predictions of the protonation behavior for proteins with more than twenty residues were unattainable given the need for representative sampling of all possible configurations [166]. Motivated by the simplicity and the effectiveness of the Gaussian-chain model, we decided to extend its applicability to an arbitrary peptide of any sequence length and complexity.

We present here **pepKalc**, a robust software for the calculation of protonation constants in an unfolded polypeptide of arbitrary length. Combining the simplicity of the Gaussian-chain model for the disordered state with the robustness and speed of a hybrid mean-field approximation treatment [170, 171], all the relevant pair-wise electrostatic interactions can be calculated within seconds. Finally, we demonstrate how **pepKalc** can help rationalize experimental pKa shifts in an intrinsically disordered protein.

5.2 Approach

5.2.1 The general theory

The association reaction to form an unfolded polypeptide chain in a specific protonation state x at position i can be schematically represented as,

where $X_i^{q_0}$ is the fully deprotonated form of a amino acid residue at the position i , with the unit charge of q_0 . H represents the protons, and $HX_i^{q_0+x_i}$ is the site i of polypeptide in protonation state x . Correspondingly, the collective protonation state of an arbitrary unfolded polypeptide with N titratable sites can be expressed in terms of a vector x , whose elements,

Tab. 5.1: Model compound pK_0 values [159].

Residue Type	Model pK_*
N-Terminus	8.0
C-Terminus	3.6
D	4.0
E	4.4
H	6.3
C	8.3
Y	9.6
K	10.4
R	13.5

x_i , take on values of 0 or 1 according to whether the corresponding site i is unprotonated or protonated, respectively.

$$x = \{x_1, x_2, x_3 \dots x_N\} \quad x_i = \begin{cases} 0 & \text{if deprotonated} \\ 1 & \text{if protonated} \end{cases} \quad (5.2)$$

Residues within the polypeptide chain can be classified based on their residual charge in the deprotonated form into two distinct groups, acidic $A \in \langle C - terminus, C, D, E, Y \rangle$ with $q_0 = -1$, and basic $B \in \langle N - terminus, H, K, R \rangle$ and $q_0 = 0$ with the intrinsic and amino-acid specific pK_a values listed in Table 5.1. The explicit treatment of protonation behavior of an arbitrary unfolded protein with multiple electrostatically interacting sites, can be demonstrated on an example of a fully deprotonated polypeptide, composed of basic residues BBB . In the presence of a single proton, the protonation reaction of BBB at its first available site,

yields a partially protonated polypeptide B^+BB , whose change in free energy of protonation relies solely on the intrinsic dissociation constant K_1 .

$$\Delta G = -RT \ln(K_1) \quad (5.4)$$

Under such conditions, the concentration of protonated polypeptide B^+BB can be expressed as,

$$[B^+BB] = [BBB] \frac{[H^+]}{K_1} = [BBB] 10^{pK_1 - pH} \quad (5.5)$$

where pK_1 is the intrinsic pK_a value of the first residue in the polypeptide. An inclusion of a second proton,

causes a significant change in the free energy of the protonation reaction due to charge-charge interactions between the protonated sites 1 and 2,

$$\Delta G = \Delta G_1 + \Delta G_2 + G_{12} \quad (5.7)$$

where ΔG_1 and ΔG_2 are the self-association free energy changes, and G_{12} denotes an electrostatic contribution to the overall free energy change due to charge-charge interactions. In the presented approach, the electrostatic charge-charge interactions are assumed to follow the linearized Debye-Hückel equation [161], adopted for a solvated polymer chain in a solution of known dielectric constant ϵ , and ionic strength I . Following the approach of Zhou, the distance between the charged residues i and j , that is $r_{i,j}$, is assumed to be sampled from the distribution for a Gaussian Chain $p(r)$,

$$p(r_{i,j}) = 4\pi r_{i,j}^2 \left(\frac{3}{2\pi d_{i,j}^2} \right)^{\frac{3}{2}} \exp \left(-\frac{3r_{i,j}^2}{2d_{i,j}^2} \right) \quad (5.8)$$

where $d_{i,j}$ is the root-mean-square (RMS) distance between the interacting sites, approximated by the empirical relation,

$$d_{i,j} = b \sqrt{l_{i,j}} + s \quad (5.9)$$

in which, $l_{i,j}$ is the number of peptide bonds separating the interacting i, j residues, b is the effective residue separation, and s represents the shift distance due to side chain. These model values are adjustable parameters, and are taken to be $b = 7.5\text{\AA}$ and $s = 5\text{\AA}$ [168]. An assumption is made that $p(r)$ is independent of temperature, ionic strength and . The mean electrostatic interaction energy between sites i, j , that is G_{ij} , can be computed from,

$$G_{ij} = 332 \sqrt{6/\pi} \frac{1 - \sqrt{\pi} \beta e^{\beta^2} (\beta)}{\epsilon d_{i,j}} \quad (5.10)$$

where $\beta = \frac{\kappa d_{i,j}}{\sqrt{6}}$, $\kappa = \sqrt{\frac{8\pi I e^2}{\epsilon k_B T}}$, I is ionic strength, T is temperature, ϵ is the dielectric permittivity of a solvent and (β) is the complementary error function [168]. Consequently, all possible configurations of the model polypeptide

BBB arising from the complete protonation, can be expressed in terms of $[BBB]$ and the concentrations of the free proton ligand, $[H^+]$, as follows,

$$\begin{aligned}
[BBB^+] &= [BBB] \frac{[H^+]}{K_3} \\
[BB^+B] &= [BBB] \frac{[H^+]}{K_2} \\
[BB^+B^+] &= [BBB] \frac{[H^+]^2}{K_2 K_3 p_{23}} \\
[B^+BB] &= [BBB] \frac{[H^+]}{K_1} \\
[B^+BB^+] &= [BBB] \frac{[H^+]^2}{K_1 K_3 p_{13}} \\
[B^+B^+B] &= [BBB] \frac{[H^+]^2}{K_1 K_2 p_{12}} \\
[B^+B^+B^+] &= [BBB] \frac{[H^+]^3}{K_1 K_2 K_3 p_{12} p_{13} p_{23}}
\end{aligned} \tag{5.11}$$

where p_{12} , p_{13} , p_{23} , represent an increase in the free energy of a particular configuration, due to the electrostatic interactions between the like (positive) charges at sites 1-2, 1-3, and 2-3, respectively. The pair-wise electrostatic interaction parameters p_{ij} are dimensionless and computed from

$$\log(p_{ij}) = \frac{4184.0}{RT \ln(10)} G_{ij} \tag{5.12}$$

where G_{ij} represents electrostatic repulsion introduced in Equation 5.10. Consequently, the electrostatic interaction Hamiltonian of a simulated system can be represented as W_{ij} .

$$W_{ij} = \begin{bmatrix} 0 & \log(p_{12}) & \cdots & \log(p_{1N}) \\ \log(p_{21}) & 0 & \cdots & \log(p_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ \log(p_{N1}) & \log(p_{N2}) & \cdots & 0 \end{bmatrix} \tag{5.13}$$

Furthermore, the factor $\frac{[H^+]^2}{K_i K_j p_{ij}}$, can be expanded to $10^{pK_i + pK_j - 2pH - \log(p_{ij})}$ in our symbolical framework.

For reasons that will become clear below we shall express the concentrations of all configurations relative to that of the fully deprotonated configuration.

However, each deprotonated acidic residue in a polypeptide introduces a negative charge in this reference state, making this case different from the all-basic case above. We shall illustrate our proposed approach with the tetrapeptide $[BBBA^-]$. Two pathways connect the doubly protonated configuration $[BBB^+A]$ with the fully deprotonated state $[BBBA^-]$,

and,

resulting in the overall equation,

$$[BBB^+A] = [BBBA^-] \frac{[H^+]}{K_3} \frac{[H^+]}{K_4} \quad (5.16)$$

The second pathway, Equation 5.15, suggests an alternative point of view, similar to the all-basic case above,

$$[BBB^+A] = [BBBA^-] \frac{[H^+]}{\frac{K_3}{p_{34}}} \frac{[H^+]}{K_4 p_{34}} \quad (5.17)$$

Substitution of $\frac{K_3}{p_{34}}$ by $K'_3 = \frac{K_3}{p_{34}}$ yields,

$$[BBB^+A] = [BBBA^-] \frac{[H^+]}{K'_3} \frac{[H^+]}{K_4} \frac{1}{p_{34}} \quad (5.18)$$

This view implies a correction of the intrinsic pK_a of residue 3 ($K_3 \rightarrow K'_3$) for the negative charge at position 4. Thus we achieve our goal, namely that in the case of double protonation of a basic and an acidic residue one repulsive interaction term ($\frac{1}{p_{34}}$) is introduced, like in the case of protonation of two basic residues, expressed in Equation 5.11. Applying this reasoning to the other configurations allows us to represent all 16 configurations of

this tetrapeptide relative to the concentration of the fully deprotonated state $[BBBA^-]$ in a homogeneous manner,

$$\begin{aligned}
[BBB^+A^-] &= [BBBA^-] \frac{[H^+]}{K'_3} \\
[BB^+BA^-] &= [BBBA^-] \frac{[H^+]}{K'_2} \\
[BB^+B^+A^-] &= [BBBA^-] \frac{[H^+]^2}{K'_2 K'_3 p_{23}} \\
[B^+BBA^-] &= [BBBA^-] \frac{[H^+]}{K'_1} \\
[B^+BB^+A^-] &= [BBBA^-] \frac{[H^+]^2}{K'_1 K'_3 p_{13}} \\
[B^+B^+BA^-] &= [BBBA^-] \frac{[H^+]^2}{K'_1 K'_2 p_{12}} \\
[B^+B^+B^+A^-] &= [BBBA^-] \frac{[H^+]^3}{K'_1 K'_2 K'_3 p_{12} p_{13} p_{23}} \\
[BBBA] &= [BBBA^-] \frac{[H^+]}{K_4} \\
[BBB^+A] &= [BBBA^-] \frac{[H^+]^2}{K'_3 K_4 p_{34}} \\
[BB^+BA] &= [BBBA^-] \frac{[H^+]^2}{K'_2 K_4 p_{24}} \\
[BB^+B^+A] &= [BBBA^-] \frac{[H^+]^3}{K'_2 K'_3 K_4 p_{24} p_{34} p_{23}} \\
[B^+BBA] &= [BBBA^-] \frac{[H^+]^2}{K'_1 K_4 p_{14}} \\
[B^+B^+BA] &= [BBBA^-] \frac{[H^+]^3}{K'_1 K'_2 K_4 p_{14} p_{24} p_{12}} \\
[B^+BB^+A] &= [BBBA^-] \frac{[H^+]^3}{K'_1 K'_3 K_4 p_{14} p_{34} p_{13}} \\
[B^+B^+B^+A] &= [BBBA^-] \frac{[H^+]^4}{K'_1 K'_2 K'_3 K_4 p_{14} p_{24} p_{34} p_{12} p_{13} p_{23}}
\end{aligned} \tag{5.19}$$

In the light of the explicit formalism presented in Equation 5.19, the individual fractional protonation states $\langle x_i \rangle$ at each , can be calculated from

$$\langle x_i \rangle (pH) = \frac{\sum [XH_i]}{\sum [XH]} \tag{5.20}$$

where $\sum [XH]$ denotes the summation over all 2^N possible protonation states a model polypeptide with N electrostatically interacting sites would have

and the index i limits the sum to states with site i protonated. Ultimately, the pK_a values of titratable residues are estimated by fitting the dependence of $\langle x_i \rangle$ to the Hill equation,

$$\log \frac{\langle x_i \rangle}{1 - \langle x_i \rangle} = n_i(pK_a -) \quad (5.21)$$

where n_i is the Hill parameter, under the assumption that all simulated sites adapt their equilibrium protonation states during the titration event [164].

5.2.2 The hybrid mean field method

An explicit numerical integration over all possible protonation-state combinations is required to solve Equation 5.20, yielding a typical $O(2^n)$ scaling problem. To mitigate the scalability issue, a hybrid mean field approach [170, 171] was developed. Algorithm Table 1 provides an overview of our approach. In the proposed methodology, given a polypeptide with N sites an explicit treatment of all plausible protonation states is applied only to a window of interacting residues w , defined as,

$$w = 2\phi + 1 \quad (5.22)$$

where ϕ is the interaction cutoff parameter. The calculation window w is recursively iterated α_{max} times over $i \in \langle \phi + 1, N - \phi \rangle$. In each cycle α corrected $pK_a(s)$ are calculated for residue $i \in \langle 1, N \rangle$, following the explicit procedure given by Equations 5.20 and 5.21, but now implying only residues within the range $i - \phi, i + \phi$. In addition to the explicit corrections for negative (unit) charges on acidic residues within that window (see above), also pK_a -corrections for charges q_j outside the window are taken into account, according to,

$$W_i = \sum_j q_j W_{ij} \quad (5.23)$$

assuming that averaged (fractional) charges can be used for these. These averaged charges q_j follow directly from the fractional protonations $\langle x_j \rangle$ calculated in the previous cycle. (In the first cycle they are calculated from tabulated pK_a values (Table 5.1) assuming Hill parameters of $n = 1.0$). This is repeated for all pH values in the chosen range before moving to the next residue (and the next window w). A new cycle (α) is started if the new $pK_a(s)$ differ too much from those calculated in the previous cycle.

Algorithm 1 Hybrid Mean Field Approximation

Require:

protein sequence with N interacting sites,
residue cutoff ϕ ,
number of supercycles α : α_{max}

Ensure:

acidity constants pK_a ,

Hill parameters n ,

$W_{ij} \leftarrow \log(p_{ij}) \leftarrow G_{ij}$

▷ position-dependent interaction energies

for α **do**

for $i \in \langle 1, N \rangle$ **do**

for pH **do**

if $\alpha = 1$ **then**

▷ the first cycle

$q_j \leftarrow (pK_a, n_i = 1.0)$

▷ use intrinsic $pK_a(s)$

else

$q_j \leftarrow \langle x_j \rangle$

▷ use values from previous iteration

end if

if $j < i - \phi \wedge j > i + \phi$ **then**

▷ residues outside w

$W_i \leftarrow \sum_j q_j W_{ij}$

▷ according to Equation 5.23

end if

$X_i, X \leftarrow W_{ij}, W_i$

$\langle x_i \rangle \leftarrow \frac{X_i}{X}$

▷ Equation 5.20

end for

pK_a, n for residue $i \leftarrow \langle x_i \rangle$

▷ Equation 5.21

end for

if $\alpha = \alpha_{max}$ **then**

 Stop

end if

end for

5.3 Methods

5.3.1 Software implementation

pepKalc software is available as a stand-alone, interactive HTML5 application at <http://protein-nmr.org> and command-line utility at <https://github.com/PeptoneInc/pepkalc>. The tool was developed using Python 2.7 [172]. The self-consistent numerical integration routines were implemented using high performance nVidia CUDA libraries, according to single-data-multiple-instruction (SIMD) paradigm. **pepKalc** supports deployment in virtual machine environments (cloud servers) with concurrent thread technology and transparent execution on nVidia graphical processing units (GPU). The command-line utility version of **pepKalc**, can be readily containerized using Docker engine [173] and deployed as a ultra-scalable application with an integrated support for remote-procedure-call (RPC) servers.

5.3.2 Performance testing and deployment

All performance benchmarks were performed on Apple Mac Pro computer (Mid 2012 model), equipped with two 2.4GHz 6-Core Intel Xenon E5645 processors, 16GB 1333 MHz DDR3 ECC memory, nVidia GeForce GTX680 4GB GPU, operating under the control of OSX 10.11.4 64-bit system with nVidia CUDA v7.5 architecture drivers. Multi-GPU scalability benchmarks were performed on nVidia DGX-1 supercomputing node.

5.4 Discussion

The performance of **pepKalc** was assessed. We have benchmarked the numerical convergence of the tool as a function of input parameters and the time required for the completion of benchmark simulations. Subsequently, we have evaluated the predictive power of **pepKalc** for the strong electrostatic-coupling case. Ultimately, we have compared experimental protonation constants for intrinsically disordered human alpha-synuclein against **pepKalc** predictions.

Fig. 5.1: The assessment of numerical convergence of **pepKalc**. The standard deviation of the residuals of the pK_a values for a D_{40} polypeptide with $N = 40$. The standard deviation σ is plotted on a logarithmic scale against the number of the integration super-cycles α , for four window sizes: $w = 3$ (dotted line $\cdots$), $w = 5$ (dashed line $--$), $w = 7$ (semi-dashed line $- \cdot -$), and $w = 9$ (continuous line $-$). The threshold line $\sigma = 0.05$ is marked on the plot with gray. The remaining simulation parameters were: $T = 298.15K$, $I = 0.1M$ and $\epsilon = 78.5$.

5.4.1 Numerical Convergence

One of the most important issues with the self-consistent numerical integration routines is an estimation of the minimal number of iterations required for the method to reach a satisfactory level of numerical convergence [174]. A protocol was devised to determine the minimum, acceptable number of iterations for **pepKalc** tool that lead to numerical convergence. The test consisted of a computation of pK_a values for a $N = 40$ residue polypeptide, composed solely of aspartate residues (D_{40}). The residue-specific pK_a values, predicted every iteration j , were stored in a form of a row-ordered matrix K . The numerical convergence criterion was defined as a standard deviation σ of a residual difference between the computed pK_a values in the consecutive iterations j and $j - 1$. The following steps were taken in the estimation of the numerical convergence of **pepKalc** tool:

(1) a computation of the residuals vector $\Delta pK_{a|(j)}^{\rightarrow}$, for every iteration j ,

$$\Delta pK_{a|(j)}^{\rightarrow} = \vec{K}_j - K_{j-1}^{\rightarrow} \Leftrightarrow j \in \langle 2, \alpha \rangle \quad (5.24)$$

where $\vec{K}_j$ is the row-vector of predicted pK_a values after iteration j .

(2) an estimation of the standard deviation $\sigma(j)$, for every iteration j according to,

$$\sigma(j) = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\Delta pK_{a|j}^{\rightarrow}(i) - \mu \right)^2} \quad (5.25)$$

where,

$$\mu = \frac{1}{N} \sum_{i=1}^N \Delta pK_{a|(j)}^{\rightarrow}(i) \wedge j \in \langle 2, \alpha \rangle \quad (5.26)$$

Figure 5.1 summarizes the outcome of the convergence estimation procedure. Four, ten super-cycle ($\alpha = 10$) simulations were performed with variable interaction window sizes w . It was found that calculations reached numerical convergence below a threshold of $\sigma = 0.05$ already after $\alpha = 3$ cycles, irrespective of the computational window w size. An increase in the number of computational iterations ($\alpha \geq 6$) yielded even lower standard deviation levels of computed residuals, surpassing the value of 0.0001 for all computed scenarios.

5.4.2 Performance Gain Factor

The relative performance of the hybrid mean field method in **pepKalc** software with respect to an exact result from an exhaustive calculation variant was assessed. A performance gain factor R was introduced. The gain is defined here as a ratio of the number of steps required for a numerically convergent simulation in an explicit approach, to the overall number of iterations in the hybrid mean field method algorithm,

$$R = \frac{2^{N-2\phi-1}}{\alpha N} \quad (5.27)$$

where N is the number of interacting sites, ϕ is the size of the hybrid mean field method cutoff, and α is the number of self-consistency integration super-cycles. For the sake of an example, let us consider a protonation state prediction for a 40-site polypeptide. An $N = 40$ calculation in an explicit approach would require approximately 4TB of RAM memory in order to ac-

Fig. 5.2: The theoretical performance gain of hybrid mean field approach over an explicit calculation, computed from Equation 5.27. The performance advantage is expressed on a logarithmic scale $\log(R)$, as a function of the number of interacting sites N . The hybrid mean field method calculation was performed with $\alpha = 3$ super-cycles. The performance curves are computed for four cutoff parameter ϕ values: $\phi = 1$ (dotted line $\cdots$), $\phi = 2$ (dashed line $--$), $\phi = 3$ (semi-dashed line $- \cdot -$), and $\phi = 4$ (continuous line $-$). The dashed gray line corresponds to a performance threshold ($\log(R) = 1$), above which computational gain is expected when using **pepKalc** tool.

commodate the protonation state matrix, with 2^{40} state combinations, stored as 32-bit precision floats. Subsequently, $2^{40} = 1099511627776$ integration cycles would be required to compute a numerically convergent solution to Equation 5.20. The very same computational problem, given $\alpha = 3$ integration super-cycles, with a cutoff parameter of $\phi = 2$, requires approximately, 480KB of RAM memory, to store 2^5 protonation-state combinations, and $3 \cdot 40 \cdot 2^{2 \cdot 2 + 1} = 7680$ integration cycles, when computed with **pepKalc** software. The expected performance gain R due to switching from an explicit computation to reduced site-approximation is $\frac{2^{35}}{(3 \cdot 40)} \approx 285000000$. Figure 5.2 demonstrates the relative gain in the computational performance, expressed as a $\log(R)$, against the number of interacting sites N .

Tab. 5.2: Simulation execution time benchmarks in seconds [s], reported for D_{40} polypeptide $N = 40$ with variable cutoff ϕ size.

*Using nVidia GeForce 680 GTX card with 4GB DDR and 1536 computational threads.

ϕ	1 CPU	2 CPUs	4 CPUs	GPU*
2	3.4	1.8	0.9	0.04
3	4.6	2.6	1.3	0.06
4	8.4	4.4	2.3	0.12
5	28.3	14.9	7.9	0.44
6	125	69.4	32.9	2.05

A meaningful, that is $\log(R) \gg 1$, performance enhancement can only be achieved if the number of the interacting sites N significantly exceeds the size of the computational cutoff (ϕ). In the case of $\phi = 1$ computational performance improvements are expected for proteins with $N > 7$, reaching 100-fold speedup for a polypeptide with fourteen interacting sites ($N = 14$). A clear performance shift can be observed in the theoretical performance curves for $\phi = 2, 3, 5$ and 10, advising the use of small computational window sizes in most cases of practical interest.

5.4.3 Calculation Time Benchmarks

A series of **pepKalc** execution-time benchmarks was performed. The multi-processor scalability and the numerical performance of the GPU implementation were assessed. Table 5.2 summarizes the outcome of the testing procedure for D_{40} , $N = 40$ polypeptide with three $\alpha = 3$ integration super-cycles, but variable calculation cutoff ϕ . Each benchmark simulation was executed fifty times and concluded with the calculation of the mean simulation time. As evidenced by the benchmark data, **pepKalc** retains near-linear scalability across the tested range of CPU cores. However, a GPU-accelerated implementation of **pepKalc** offers a remarkable two-orders of magnitude increase in the computational efficiency with respect to the CPU code, achieving 60x speedup for the most complex $\phi = 6$ computational scenario.

5.4.4 Strong Electrostatic Coupling

The predictive power of **pepKalc** in the strong-electrostatic coupling scenario was assessed. The theoretical protonation curves for DDD tripeptide were

Fig. 5.3: (A) Protonation curves for D_1 and D_3 , (B) D_2 residues of DDD polypeptide simulated with variable dielectric constant ϵ . The remaining simulation parameters are, $\alpha = 3$, $T = 298.15K$, and $I = 0.1M$

simulated. Five predictions were made with variable dielectric constant ϵ values. Figure 5.3 depicts the outcome of the calculations.

Low dielectric constant simulation $\epsilon = 2$ reveals very interesting protonation behavior, elegantly demonstrating the power of **pepKalc** in predicting a convoluted electrostatic coupling scenario. In the case of very low dielectric screening, a multi-site, correlated deprotonation of terminal D_1 , and D_3 residues takes place whilst D_2 happens to retain its neutral character and reverts to its nearly neutral form at $pH = 10$. Doubling of a dielectric constant, that is $\epsilon = 4$, yields complete deprotonation of D_2 at $pH = 12$, smoothing out the multi-site character of curves for terminal residues D_1 and D_3 . A typical, sigmoidal deprotonation behavior can be observed for simulations with $\epsilon > 20$, whereas the case of $\epsilon = 80$ resembles deprotonation events expected from Equation 5.21, although the Hill parameter does not have a unique interpretation for this particular case of weak-coupling [164, 175].

5.4.5 Site-specific protonation behavior of intrinsically disordered alpha-synuclein

Fig. 5.4: Comparative analysis of experimental (black) and predicted (grey) ΔpK_a values for intrinsically disordered human α -synuclein. Calculation parameters were derived from Croke et al., $T = 283.15K$, $\epsilon = 83.83$, and $I = 0.150M$.

Human alpha-synuclein is a 14.5 kDa intrinsically disordered protein expressed predominantly at the presynaptic terminals of brain neurons, [177]. It is known, that the misfolding of alpha-synuclein leads to the formation of fibrillar cytoplasmic aggregates [178], often referred to as Lewy bodies, which are defining characteristic of Parkinson's disease, [179]. Because of

its pivotal role in the etiology of neurodegeneration, great effort has been devoted to its biophysical characterization. Human alpha-synuclein contains 140 amino acids, of which 46 have the capacity to act as acids or bases: 15 Lys, 1 His, 4 Tyr, 6 Asp, 18 Glu, and the N- and C-termini. Croke et al. reported an experimental site-specific characterization of electrostatic interactions in alpha-synuclein using solution-state NMR spectroscopy. The pK_a values for all 26 sites that ionize below pH 7 were characterized using 2D ^1H - ^{15}N HSQC and 3D $C(CO)NH$ NMR experiments.

We have used **pepKalc** to predict the pK_a constants of ionizable residues in human alpha-synuclein adopting the experimental conditions found in the work of Croke et al. as model input parameters [176]. The calculation was performed for 46 independently interacting sites over the full pH range; thereby accurately treating pH-dependent protonation in the full-length polypeptide. It should be noted, the prediction of the electrostatic interactions with a similar level of numerical complexity is virtually impossible using the Monte-Carlo approach published by Zhou [168]. We assumed in our calculations the experimental salt-free samples could be approximated by low ionic strength ($I = 0.001M$). Figure 5.4 depicts a comparative analysis of experimentally and computationally derived differences between the observed pK_a values and the reference constants pK_0 listed in Table 5.1. As demonstrated in Figure 5.4 the pK_a values predicted by **pepKalc** follow the general trend found in the experimental study by Croke et al. [176]. The root mean square deviation (RMSD) of the computed pK_a constants with respect to the experimental values was used to assess the predictions. The simulations for the low ionic strength solution $I = 0.001M$ yielded an RMSD of 0.65, whereas calculations for $I = 0.150M$, resulted in much better agreement with the experimental values reflected by RMSD of 0.15.

The observed discrepancy in the estimation of ΔpK_a for human alpha-synuclein is expected, based on a massive body of experimental work done for this particular polypeptide. Growing experimental evidence from small angle X-ray scattering (SAXS) [180], solution state NMR spectroscopy [72, 181] and empirical modeling [182] studies, suggest intrinsically disordered proteins cannot be represented as homopolymers devoid of canonical structures, commonly referred to as random-coil [11, 183]. Thus, the Gaussian-chain model which was proven to successfully simulate the intra-molecular distance distributions in polypeptides exposed to highly concentrated chaotropic agents [166–168], may be providing a baseline to capture the intricate tendency of

intrinsically disordered proteins to adopt partially folded, non-canonical structures. A detailed structural propensity analysis of human alpha-synuclein clearly demonstrates the both the N- and C-terminal parts of the protein populate a non-canonical structures, which deviate from the "random-coil" behavior [184]. Consequently, the pair-wise distance distributions of charged moieties in intrinsically disordered alpha-synuclein are expected not to follow behavior predicted by Gaussian-chain model implemented in **pepKalc**. Thus, our computations can identify sites where locally persistent interactions may shift a pK_a constant to a value beyond that expected for a featureless chain, and could hence provide novel structural information.

In the current implementation, **pepKalc** and all previous published methods, which rely on simplified Debye-Hückel electrostatics, do not account for the electrostatic effects caused by highly concentrated protein in solution. Protein-protein and protein-solvent interactions are known to cause a significant and detectable effect on the ion binding affinities [185], as well as have a profound effect on the overall dielectric properties of their solutions [186, 187]. Thus, in a current form, given the approximate nature of Debye-Hückel interactions in **pepKalc**, overestimation of intra-molecular charge-charge interaction energies is expected at very low ionic strength. The generality of the method described here remains, however, and trivially allows Equation 5.10 to be adapted to obtain more accurate calculations. In the future, rather than the Gaussian Chain model for unfolded proteins, three-dimensional ensembles of structures could be used. In addition, the pH-dependent structural deformations that take place in the case of particular charge distributions [188] could be considered. These refinements would all help to model protein unfolded states in a more realistic manner, and thereby improve the accuracy of the prediction.

Finally, the method has the innate flexibility to consider any type of acid or base in the polypeptide sequence, such that non-natural amino acids and post-translational modifications can also be included. These extensions are also offered by **pepKalc**.

5.5 Conclusion

We have developed **pepKalc**, a robust simulation software for the comprehensive evaluation of protein electrostatics in disordered polypeptides. Our

software completely removes the limitations of previously reported Monte-Carlo approaches in the computation of protein electrostatics, by using a hybrid methodology that effectively combines exact and mean-field calculations to rapidly obtain accurate results. Paired with a modern architecture GPU, **pepKalc** is capable of evaluating protonation behavior for a typical protein in seconds. Our approach can be combined with other polymer models, including explicit ensembles. Thus the deviations from the protonation behavior obtained in high accuracy **pepKalc** simulations might be helpful in structure validation or as ensemble restraints for intrinsically disordered proteins with variable levels of secondary structure propensity.

Acknowledgements

Authors thank Alison Lowndes and Carlo Ruiz, (NVIDIA Corporation) for facilitating collaboration and access to DGX-1 supercomputing node.

Conclusions and future perspectives

Abstract

This thesis describes the studies of relationships between amino acid sequence and biophysical properties of intrinsically disordered proteins gauged from NMR spectroscopy experiments. A database (ncIDP) of amino acid-specific backbone chemical shifts with their nearest-neighbour corrections from 14 natively unfolded proteins is presented. The prediction accuracy of backbone chemical shifts exceeds previously reported approaches to 'random-coil' chemical shift calculations. Subsequently, the ncIDP database is shown to accelerate the process of sequential resonance assignment of IDPs, through accurate detection of resonance misassignment. The secondary-chemical shifts computed from ncIDP facilitate accurate detection of residual structural propensity in folded and natively unfolded proteins and constitute a novel tool for secondary structure detection (ncSPC). Finally, this thesis describes a theoretical model of a disordered peptide chain, which permits accurate and scalable prediction of protonation behaviour in disordered peptides. The robustness of the model is compared against residue-specific pH titrations obtained in high resolution NMR spectroscopy experiments. Key conclusions are summarized followed by an outlook for the underlying themes covered in this thesis.

6.1 Conclusions

Chemical shifts of intrinsically disordered proteins can be predicted from their amino acid sequences

Chapter 2 describes the first inventory of IDP backbone $^1H^N$, ^{15}N , $^1H^\alpha$, ^{13}C , $^{13}C^\beta$, and $^{13}C^\alpha$ chemical shifts (ncIDP), which is derived from NMR data for 14 disordered proteins of unrelated sequence and function. We demonstrate that backbone chemical shifts in IDPs can be accurately predicted

from a linear relationship, given by Equation 2.1, which includes sequence-dependent nearest neighbour corrections. Our ncIDP library permits the prediction of chemical shifts for 14 IDPs with root-mean-square deviations of 0.65, 0.14, 0.12, 0.50, 0.36, and 0.41 ppm from the experimentally measured values for the ^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^{\alpha}$, ^{13}C , $^{13}\text{C}^{\beta}$, and $^{13}\text{C}^{\alpha}$ chemical shifts, respectively. Ultimately, the analysis of the magnitude of chemical shift corrections in **Chapter 2** show that neighbour corrections are significant for all backbone nuclei.

The accuracy of our prediction model is hampered by structural anisotropy of IDPs related to experimental conditions

As explained in Section 1.1.2 of **Chapter 1**, natively unfolded proteins are not statistical random-coils but may contain regions with nascent secondary structures. This structural anisotropy of IDP polypeptide chains is expected to yield variable backbone chemical shifts, an effect described in detail in Section 1.2.3 of **Chapter 1**. To account for the effects of nascent secondary structures in IDPs, we introduce a self-consistent optimization protocol with an automated outlier selection, described in Section 2.2 of **Chapter 2**. The detected and excluded structural anisotropy within 14 proteins is evident from Figures 2.1 and 2.2. Furthermore, we introduce a protein-specific error parameter ϵ in Equation 2.1, to capture and separate systematic effects of resonance offset corrections and an influence of temperature and pH on our chemical shifts. However, an analysis of a large set of IDPs revealed a significant tendency of IDPs to form α -helical secondary structure elements and to preferentially fold into more compact structures under acidic conditions [189]. Our self-consistent optimization protocol is not capable of discerning pH- or temperature- induced structural anisotropy from a systematic deviation caused by sample conditions.

Very recently, Nielsen and Mulder proposed a new method for random-coil chemical shift calculations, POTENCI, which is parameterized using a large curated database of chemical shifts for protein segments with validated disorder [190]. It takes pH and temperature explicitly into account, and includes sequence-dependent nearest and next-nearest neighbour corrections as well as second-order corrections. The predictions obtained with POTENCI show root-mean-square values that are lower by 25–78%, with respect to ncIDP approach with the largest improvements observed for $^1\text{H}^{\alpha}$, and ^{13}C .

Libraries compiled from chemically unfolded and short-length polypeptides are not representative of intrinsically disordered states

The comparative assessment of chemical shift prediction accuracy presented in **Chapter 2** clearly demonstrates that the prediction of chemical shifts for natively unfolded proteins benefits most from making use of the experimental data obtained for IDPs as a reference state. It also shows that neighbour corrections are significant for all backbone nuclei. This result is further reinforced by a comparative analysis of 'random-coil' chemical shift predictions in the derivation of PROTENCI database [190]. The outcomes obtained from IDPs are always at great variance with respect to other chemical shift libraries, irrespective of their origin. Evidently, the structural anisotropy in IDP ensembles differs greatly from conformational sampling in chemically unfolded polypeptides or fragments of folded proteins with non-canonical secondary structures. Thus, our results hint at the correctness of "protein quartet" model of protein conformational states, which recognizes native disorder as a completely separate thermodynamic state of a polypeptide chain.

The prediction of IDP chemical shifts can aid the assignment of NMR spectra of intrinsically disordered proteins

The relative gain in NMR resonance assignment performance from the use of ncIDP chemical shift library is presented in **Chapter 3**. We introduce there ncIDP-Assign, an extension for popular NMR resonance assignment programme SPARKY and show that a two-fold improvement in the estimation of chemical shifts for 140-residue intrinsically disordered cytoplasmic domain of human neuroligin-3 (hNLG3cyt). Our approach has significantly higher detection sensitivity of chemical shift deviations from the sequence-specific 'random-coil' values due to resonance misassignment. Although the ncIDP-Assign extension was developed specifically for SPARKY, the approach born out of resonance matching to predicted chemical shifts could be also applied in other popular assignment programmes including MARS [191] and TSAR [192].

Accurate detection of structural anisotropy in IDPs and folded proteins benefits from chemical shifts of disordered proteins

Structural anisotropy in intrinsically disordered proteins results in detectable differences in chemical shift patterns. Section 1.2.4 of **Chapter 1** introduces the concept of secondary chemical shifts, as a measure of residual structure in ordered and disordered polypeptides. **Chapter 4** demonstrates how ncIDP chemical shift library can be used to obtain accurate secondary shifts for a 140 residue long, natively unfolded human alpha-synuclein (aS). An inclusion of neighbour corrections in the calculation of secondary shifts for aS vastly improves its secondary chemical shifts plots, as seen in Figure 4.1. The importance of neighbour effects in chemical shifts goes in concert with the early observations of nearest-residue effects on J-couplings and RDCs discussed in Section 1.1.2 of **Chapter 1**. We show that neighbour corrections are significant for all backbone nuclei and that their inclusion along with chemical-shift specific weight parameters (which reflect their idiosyncratic sensitivity to structural information), permits accurate structural propensity calculations. The applicability of our structural propensity calculator (ncSPC) is demonstrated for two well-characterized systems, namely, the binding of human alpha-synuclein to micelles, and light activation of photoactive yellow protein. To date, more than 30 different studies have been reported to benefit from ncSPC calculations, including state of the art examples, such as the study of the human telomere repeat binding factor 1 (hTRF1) in complex with *Escherichia coli* Hsp70 (DnaK), which clearly demonstrates structural propensity analysis can be used to study conformational changes in large macromolecular complexes [193].

Charge–charge distance can be approximated with Gaussian-chain model in IDPs and used to accurately compute protonation profiles for acidic and basic residues

The notion that structural features of natively unfolded proteins can be approximated using mathematical modelling has been explored thoroughly in **Chapter 1**. However, relatively little space was devoted to unravelling the nature of electrostatic interactions within the IDPs. The work presented in **Chapter 5** demonstrates that an arbitrary IDP can be modelled as a heteropolymer in which distances between charged residues are sampled from a

Gaussian-like distribution. Under such an assumption, all residual charge–charge interactions within the considered system can be enumerated, their interaction strength computed from the linearized Debye–Hückel electrostatics model and collective protonation behaviour predicted in simulated *in silico* titration experiments. We demonstrate how the limitations of the previously reported Monte–Carlo approaches in the computation of protein electrostatics can be mitigated using a hybrid approach that effectively combines exact and mean–field electrostatic calculations to rapidly obtain accurate results, named pepKalc. Our pepKalc approach is convergent, scalable and permits simulations of complex electrostatic–coupling scenarios, such as the case of *DDD* tripeptide, shown on Figure 5.3. Ultimately, we compare the results of site–specific simulations against residual protonation data for natively unfolded human alpha–synuclein, obtained in residue–specific NMR pH–titration experiments. Our predictions, seen on Figure 5.4 achieve high accuracy, with RMSD values of 0.65 for low, 0.15 high ionic strength scenarios, respectively. The agreement between our pepKalc–generated electrostatics and experimental data is very good, yet few predictions in the low ionic strength regime seem not to follow experimental trends.

pH–induced structural changes in IDPs may be detectable through NMR pH–titration measurements

The deviations of simulated pH–profiles from residual pH–titration data in low ionic strength regime ($I = 0.001\text{mM}$) can be explained by pH–induced structural anisotropy reported in different experimental studies for human alpha–synuclein [189]. Thus, our pepKalc approach can be used to generate baseline electrostatic–response profiles, which could be in turn compared against residual pH–titration profile from NMR experiments to identify deviating protein segments. An underlying hypothesis here is, pH–driven compaction should yield charge–charge distance distributions, which differ from pepKalc’s Gaussian–chain approach.

The ‘random–coil’ term does not reflect the conformational preferences of peptide chains and should be used with great care

Almost every chapter of this thesis concludes with the notion that IDPs display varying but detectable level of structural anisotropy. Thus, in the light of

the experimental evidence disclosed in this dissertation and the last 25 years of research on IDPs, it is fair to state that **behaviour of natively unfolded proteins is anything but random**. It is fuzzy, complex, and difficult to quantify with simple physical models, yet not random. Thus, the 'random-coil' term should be permanently removed from the structural biochemistry theory of proteins, and replaced by the operational definition of structural disorder, given by the "protein quartet" model of protein conformational states.

6.2 Future perspectives

A global scientific community made great progress in unravelling of the features of a relatively small set of medically important natively unfolded proteins, such as human alpha-synuclein or Tau. However, much work still needs to be done to characterize functional spectrum of other unfolded proteins. Kulkarni and Uversky summarize the state of our knowledge on IDPs with the powerful concept of "dark proteome", that is a good portion of the 'protein universe' that comprises proteins not amenable to experimental structure determination by existing means and inaccessible to homology modelling [194]. Should the state of our knowledge about IDPs advance, protein science will need to adopt:

- the advent of powerful and data-driven computational methods to screen protein sequences, and even the sequences of entire genomes, for intrinsic disorder. This will undoubtedly reveal many more proteins that belong to this elusive class. Machine learning and artificial intelligence will play the central role in such studies [195], as they were shown to surpass classical protein-modelling approaches by significant margins, e.g. in CASP protein modelling challenges [196];
- parallel progress in functional genomics to advance our understanding of the functions of these disordered regions of proteins. The role of sequences of low complexity that are between structured domains is just beginning to be addressed. In addition, the degree of polypeptide mobility that is consistent with (or required for) protein function is still unclear;
- the mindset that is probable that the present concepts of proteins and their functions will evolve to encompass a continuum; from fully unstructured proteins that fold on binding their target, through strings of protein domains that assemble on binding their target, to relatively rigid proteins with mobile functional regions.

Our concept of a functional protein must therefore evolve from a static picture to a highly dynamic one, in which several conformations that are

consistent with various aspects of function are represented. This is why a major renovation of protein–dynamics will be required and should include:

- tools adapted from the theories of chaos and information, which will bring systems biology to the brink of automatically discovering hidden patterns, quantifying their complexity and advancing our understanding of the molecular processes that define life [197];
- a possibility that the behaviour of an IDP, being chaotic, complex, and extremely sensitive to the environmental conditions [8], can be described in terms of the strange (Lorentz) attractor, where system will neither converge to a steady state nor diverge to infinity, but will stay in a bounded but chaotically defined region [198] being a subject to well known and understood forces that define protein folding [199–201];
- exploration of new forms of computing, such as gate–operated quantum computing and annealing, which hold a promise of computational supremacy over classical, silica–based computing solutions, and thus open up a completely new world of possibilities such as explicit density functional theory (DFT) calculations, as opposed to severely limited molecular dynamics (MD) simulations of proteins [202].

6.3 Nederlandse Samenvatting

De relatie tussen de aminozuurvolgorde en biofysische eigenschappen van intrinsiek ongevouwen eiwitten (“intrinsically disordered proteins”, IDPs) wordt in dit proefschrift onderzocht. Voor de experimentele studie van dit soort eiwitten maken we hier gebruik van kernspinresonantie (in het Engels nuclear magnetic resonance, afgekort NMR) met speciale aandacht voor informatie die de chemische verschuiving (“chemical shift”) bevat. **Hoofdstuk 2** introduceert de fundamentele link tussen aminozuurvolgordes van IDPs en grootheden, zoals chemical shifts, die met NMR gemeten kunnen worden. Er volgt een eerste inventarisatie van de chemical shifts van kernen in de hoofdketens ($^1H^N$, ^{15}N , $^1H^\alpha$, ^{13}C , $^{13}C^\beta$, en $^{13}C^\alpha$) van 14 IDPs met verschillende aminozuurvolgordes en functies. We tonen aan dat in eiwitten zonder een stabiele secundaire en tertiare structuur, de chemical shifts van kernen in de hoofdketen kunnen worden benaderd door een lineaire combinatie van aminozuur-specifieke “random-coil” chemical shifts en een correctiefactor die afhankelijk is van de identiteit van de direct naburige aminozuren. Onze methodiek voor het voorspellen van de chemical shifts bevat tevens opties om automatisch problemen op te lossen die verband houden met het feit dat in de literatuur verschillende referenties worden gebruikt voor het rapporteren van chemical shifts, en om met behulp van lineaire regressie methodes uitbijters te detecteren die mogelijk naar toekenningfouten te herleiden zijn. De 6903 gemeten chemical shifts van deze 14 eiwitten kunnen wiskundig beschreven worden als een stelsel van 6903 lineaire vergelijkingen in termen van 20 aminozuur-specifieke “random-coil” waarden (voor elk van de 6 kernen in de hoofdketen) en 40 correctie-factoren voor elk aminozuur als linker- of rechterbuur. “Singular-value decomposition” (SVD) is toegepast om van deze 60 parameters de waarden te berekenen die de gemeten chemical shifts van de betreffende kern het best reproduceren. Dit heeft geleid tot onze “neighbor-corrected IDP chemical shift” (ncIDP) bibliotheek, die dus waarden bevat voor de “random-coil” chemical shifts van alle 20 verschillende aminozuren en daarnaast 40 correctie-factoren voor elk aminozuur als linker- of rechterbuur. De chemical shifts van kernen in de hoofdketens van de 14 IDPs die hiermee konden worden voorspeld, vertoonden een “root mean square deviation” ten opzichte van de experimenteel gemeten waarden van 0.65, 0.14, 0.12, 0.50, 0.36 en 0.41 ppm voor ^{15}N , $^1H^N$, $^1H^\alpha$, ^{13}C , $^{13}C^\beta$,

en $^{13}\text{C}^\alpha$, respectievelijk. De nauwkeurigheid van de voorspelling met behulp van onze bibliotheek is aanzienlijk beter dan die van methodes die gebruik maken van korte peptides of alleen de “ongevouwen” delen van eiwitten met een verder goed gedefinieerde drie-dimensionale structuur. Deze waarneming suggereert dat de conformationele heterogeniteit in dergelijke korte peptides niet representatief is voor de situatie in IDPs.

Een snelle en nauwkeurige voorspelling van “random-coil” chemical shifts van kernen in de hoofdketen, zoals beschreven in **Hoofdstuk 2**, kan van direct belang zijn voor de sequentiële toekenning van signalen in NMR spectra van ongevouwen eiwitten, die vaak veel overlap vertonen. In **Hoofdstuk 3** introduceren we een nieuwe module voor het populaire programma SPARKY, die profiteert van de betrouwbaarheid van onze voorspellingen van de chemical shifts en die specifiek gericht is op het analyseren van NMR spectra van ongevouwen eiwitten. Op basis van de berekende chemical shifts van kernen in de hoofdketen (^{15}N , $^1\text{H}^{\text{N}}$, $^1\text{H}^\alpha$, ^{13}C , $^{13}\text{C}^\beta$, en $^{13}\text{C}^\alpha$) wordt de meest waarschijnlijke residu-specifieke toekenning (oftewel de correcte positie in de aminozuurvolgorde van het eiwit) bepaald. We laten het gebruik van onze methodologie zien voor de toekenning van het 140 residuen tellende cytoplasmatische deel van het humane eiwit neurologine-3 (hNLG3cyt). Onze aanpak bereikt bijna een tweevoudige verbetering in de schatting van chemical shifts voor hNLG3cyt en is als gevolg daarvan aanzienlijk beter in het correct bepalen van de juiste toekenning en in het detecteren van eventuele fouten door de gebruiker.

Chemical shifts van kernen in de hoofdketen bevatten ook belangrijke informatie over de lokale secundaire structuur (bijvoorbeeld alfa-helix of beta-sheet) van een eiwit. De verschillen tussen de experimentele data en de “random-coil” waardes (de zogenaamde “secundaire chemical shifts”) zijn uitermate bruikbaar voor het karakteriseren van (dis)functionele conformaties die van belang zijn in eiwitaggregatie of het herkennen van bindingspartners. Het moge duidelijk zijn dat een goede referentieset voor deze “random-coil” data van groot belang is en in **Hoofdstuk 4** laten we zien dat onze “ncIDP” bibliotheek, in het bijzonder vanwege de correctie voor naburige aminozuren, zeer geschikt is voor het verkrijgen van betrouwbare “secundaire chemical shifts” in zowel gevouwen als ongevouwen eiwitten. Als voorbeeld gebruiken we het eiwit alpha-synucleine (aS), in de natieve toestand een ongevouwen polypeptide van 140 aminozuren. We tonen aan dat onze methodologie resulteert in meer betrouwbare waardes voor

secundaire chemical shifts en daarmee tot een betere beschrijving van de neiging tot het vormen van een bepaalde secundaire structuur (“structural propensity profiles”, SSP) voor dit ongevouwen eiwit. Tot slot verfijnen we de veelgebruikte SSP methode door gebruik te maken van onze nieuwe ncIDP referentiedata en we passen onze methodiek (genaamd “neighbor-corrected structural propensity calculator”, ncSPC) toe op twee veel bestudeerde biologische systemen: de binding van alpha-synucleine aan micellen en de activatie met behulp van licht van photoactive yellow protein. De resultaten onderstrepen de gevoeligheid van NMR chemical shifts voor structurele kenmerken op atomair niveau van zowel gevouwen als ongevouwen eiwitten met verschillende biologische functies.

Aangemoedigd door de experimentele resultaten betreffende structurele kenmerken van IDPs, beschreven in **Hoofdstuk 2** en **Hoofdstuk 4**, gebruiken we in **Hoofdstuk 5** een simpel model voor een willekeurig ongevouwen eiwit. We beschrijven zo’n eiwit als een heteropolymeer waarin de afstanden tussen de geladen aminozuren een Gaussian-achtige verdeling vertonen. Onder deze aanname kunnen we een Debye-Hückel electrostatisch model gebruiken om de interacties te berekenen en zo het collectieve protoneringsgedrag te voorspellen met in silico titratieexperimenten. **Hoofdstuk 5** introduceert een robuuste methode voor het evalueren van eiwitelektrostatica in ongevouwen eiwitten. We laten zien dat de beperkingen van eerder beschreven Monte-Carlo methodes omzeild kunnen worden door gebruik te maken van een tweeledige aanpak die “exacte” en “mean-field” electrostaticaberekeningen combineert voor snelle doch accurate resultaten. De theoretische onderbouwing van onze nieuwe methodiek, genaamd pepKalc, wordt in dit hoofdstuk tot in detail beschreven. We demonstreren de algemene toepasbaarheid van pepKalc in een studie aan alpha-synucleine en demonstreren dat de resultaten van onze simulaties goed overeenkomen met experimentele data, verkregen uit NMR pH-titratie experimenten aan specifieke residuen van alpha-synucleine.

Ter afsluiting bevat **Hoofdstuk 6** een korte samenvatting van het werk beschreven in dit proefschrift en beschrijft mogelijke richtingen voor vervolgonderzoek.

Epilogue

Even in the absence of formal mathematical theories, numerical simulations based on experimental observations can be of fundamental importance for drawing inferences of how biochemical systems are organized, function, and are regulated. This is why in the course of the past twenty years computational biology has soared from being an auxiliary discipline to the position of an integral element for progress in practically all aspects of modern biological sciences. Michael Levitt, one of the most prominent contributors to the field of computational biophysics and Nobel Prize laureate, explained in an elegant way the synergy between the computations and biology [203]:

... computers were made for biology: biology would never have advanced as it did without the dramatic increase in computer power and availability. One day we would like to be able to simulate complicated biological processes, perhaps even going from the genomic sequence to a full simulation of the organism's phenotype. In thinking about how to do this, it is interesting to compare Nature with simulated biology. Some things that are very difficult in Nature are trivial for computers: consider how much cellular machinery is needed to transcribe DNA sequence to RNA sequence — in the computer all one needs to do is change T to U. Translating RNA sequence to protein sequence is even more difficult in the cell, but in a computer one just applies the genetic code table. Other things that appear very easy for Nature are almost impossibly hard for computers: once synthesized a protein sequence spontaneously folds into the native structure, whereas simulating even a part of this process is still well beyond our computational capabilities. Computational structural biology will remain very challenging well into the 21st century ...

Although many challenges lie ahead, it is fair to say huge leaps in biological computing were made [204]. We now have at our disposal large information-rich resources, and we are increasingly able to integrate and understand the vast quantities of data that they encompass. The most remarkable progress, however, has been made in the area of multiscale biological modelling. Analysis of massive gene expression and proteomic data permitted the construction of comprehensive and predictive models for cellular pathways, as well as development of software for inferring interaction networks, and steps toward modelling of whole cells [205–207]. Genes susceptible to disease have been identified and, on a different level, the electrical behaviour of neurons has been modelled [208]. Molecules have been imaged in action and networks that regulate cell functions untangled [209]. A rapid emergence of multiscale computing in structural biology owes its success to the **data-driven computational methods** which span across many experimental sub-disciplines and data sources. Hybrid methods in structural biology integrate data from, for example, serial crystallography and time-resolved wide-angle X-ray scattering, micro- and nano-crystals for (future) free-electron lasers, electron microscopy, fluorescence resonance energy transfer (FRET), cross-linking data, small-angle X-ray scattering, crystallography, and nuclear magnetic resonance spectroscopy (NMR).

Most importantly, computational biophysics is not any longer a domain of a finely tailored group of theoreticians, but a rapidly evolving field of natural sciences advanced by vast collective of experimentalists with an open access to computational resources [210]. The future of biophysics is with **data-driven approaches**, which will propel our understanding of spatio-temporal nature of proteins towards a general and unifying theory that explains the process of protein folding and structural disorder.

Bibliography

- (1) Dunker, A.; Lawson, J.; Brown, C. J., et al. *J. Mol. Graphics Modell.* **2001**, *19*, 26–59.
- (2) Uversky, V. N. *Proteomics* **2018**, 1800352.
- (3) Piovesan, D.; Tabaro, F.; Mičetić, I., et al. *Nucleic Acids Res.* **2017**, *45*, D219–D227.
- (4) Bairoch, A.; Boeckmann, B. *Nucleic Acids Res.* **1991**, *19*, 2247–2249.
- (5) Xie, H.; Vucetic, S.; Iakoucheva, L. M., et al. *J. Proteome Res.* **2007**, *6*, 1882–1898.
- (6) Uversky, V.; Longhi, S., *Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation*; Wiley: 2010.
- (7) He, B.; Wang, K.; Liu, Y., et al. *Cell Res.* **2009**, *19*, 929–949.
- (8) Uversky, V. N. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2013**, *1834*, The emerging dynamic view of proteins:Protein plasticity in allostery, evolution and self-assembly, 932 –951.
- (9) Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 573–584.
- (10) Ward, J.; Sodhi, J.; McGuffin, L.; Buxton, B.; Jones, D. *J. Mol. Biol.* **2004**, *337*, 635–645.
- (11) Das, R. K.; Ruff, K. M.; Pappu, R. V. *Curr. Opin. Struc. Biol.* **2015**, *32*, 102–112.
- (12) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
- (13) Romero, P.; Obradovic, Z.; Li, X., et al. *Proteins: Structure, Function, and Genetics* **2000**, *42*, 38–48.
- (14) Anfinsen, C. B. *Science* **1973**, *181*, 223–230.
- (15) Shirota, M.; Kinoshita, K. *Protein Sci.* **2013**, *22*, 725–33.
- (16) Campen, A.; Williams, R. M.; Brown, C. J., et al. *Protein Pept. Lett.* **2008**, *15*, 956–963.

- (17) Muller-Spath, S.; Soranno, A.; Hirschfeld, V., et al. *Proceedings of the National Academy of Sciences* **2010**, *107*, 14609–14614.
- (18) Varadi, M.; Tompa, P. In *Adv. Exp. Med. Biol.* 2015; Vol. 870, pp 335–349.
- (19) Ramachandran, G. N.; Ramakrishnan, C; Sasisekharan, V *J. Mol. Biol.* **1963**, *7*, 95–9.
- (20) Lovell, S. C.; Davis, I. W.; Arendall 3rd, W. B., et al. *Proteins* **2003**, *50*, 437–50.
- (21) Kuhn, W.; Kuhn, H. *J. Colloid Sci.* **1948**, *3*, 11 –32.
- (22) Tanford, C *Adv. Protein Chem.* **1968**, *23*, 121–282.
- (23) Flory, P. J., *Statistical mechanics of chain molecules*; Interscience Publishers: New York, 1969.
- (24) Shimada, J.; Yamakawa, H. *The Journal of Chemical Physics* **1977**, *67*, 344–352.
- (25) Ackerman, M. S.; Shortle, D. *Biochemistry* **2002**, *41*, 13791–7.
- (26) Dill, K. A.; Shortle, D *Annu. Rev. Biochem.* **1991**, *60*, 795–825.
- (27) Logan, T. M.; Thériault, Y; Fesik, S. W. *J. Mol. Biol.* **1994**, *236*, 637–48.
- (28) Shortle, D; Ackerman, M. S. *Science* **2001**, *293*, 487–9.
- (29) Fitzkee, N. C.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12497–502.
- (30) Smith, L. J.; Bolin, K. A.; Schwalbe, H, et al. *J. Mol. Biol.* **1996**, *255*, 494–506.
- (31) Smith, L. J.; Fiebig, K. M.; Schwalbe, H; Dobson, C. M. *Fold. Des.* **1996**, *1*, R95–106.
- (32) Schwalbe, H; Fiebig, K. M.; Buck, M, et al. *Biochemistry* **1997**, *36*, 8977–91.
- (33) Penkett, C. J.; Redfield, C; Dodd, I, et al. *J. Mol. Biol.* **1997**, *274*, 152–9.
- (34) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–42.
- (35) Bernadó, P.; Blanchard, L.; Timmins, P., et al. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–7.
- (36) Smet, C.; Leroy, A.; Sillen, A., et al. *Chembiochem* **2004**, *5*, 1639–46.
- (37) NYu, P.; Kireeva, M. L.; Murzina, N. V., et al. *Protein Eng.* **1994**, *7*, 1373–7.
- (38) Uversky, V. N.; Ptitsyn, O. B. *Biochemistry* **1994**, *33*, 2782–91.
- (39) Ptitsyn, O. B.; Uversky, V. N. *FEBS Lett.* **1994**, *341*, 15–8.
- (40) Fink, A. L.; Calciano, L. J.; Goto, Y; Nishimura, M; Swedberg, S. A. *Protein Sci.* **1993**, *2*, 1155–60.

- (41) Hamada, D; Hoshino, M; Kataoka, M; Fink, A. L.; Goto, Y *Biochemistry* **1993**, 32, 10351–8.
- (42) Eliezer, D; Yao, J; Dyson, H. J.; Wright, P. E. *Nat. Struct. Biol.* **1998**, 5, 148–55.
- (43) Wu, L. C.; Laub, P. B.; Elöve, G. A.; Carey, J; Roder, H *Biochemistry* **1993**, 32, 10271–6.
- (44) Merrill, A. R.; Cohen, F. S.; Cramer, W. A. *Biochemistry* **1990**, 29, 5829–36.
- (45) Hua, Q. X.; Ladbury, J. E.; Weiss, M. A. *Biochemistry* **1993**, 32, 1433–42.
- (46) Hua, Q. X.; Kochoyan, M; Weiss, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, 89, 2379–83.
- (47) Mirmira, R. G.; Tager, H. S. *Biochemistry* **1991**, 30, 8222–9.
- (48) Uversky, V. N.; Karnoup, A. S.; Segel, D. J., et al. *J. Mol. Biol.* **1998**, 278, 879–94.
- (49) Alexandrescu, A. T.; Gittis, A. G.; Abeygunawardana, C; Shortle, D *J. Mol. Biol.* **1995**, 250, 134–43.
- (50) Ptitsyn, O. B. *Trends Biochem. Sci.* **1995**, 20, 376–9.
- (51) Ptitsyn, O. B. *Adv. Protein Chem.* **1995**, 47, 83–229.
- (52) Ozenne, V.; Bauer, F.; Salmon, L., et al. *Bioinformatics* **2012**, 28, 1463–70.
- (53) Mao, A. H.; Lyle, N.; Pappu, R. V. *Biochem. J.* **2013**, 449, 307–18.
- (54) Cragnell, C.; Rieloff, E.; Skepö, M. *J. Mol. Biol.* **2018**, 430, 2478–2492.
- (55) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, 6, 197–208.
- (56) Van der Lee, R.; Buljan, M.; Lang, B., et al. *Chem Rev* **2014**, 114, 6589–631.
- (57) Uversky, V. N. *Protein Sci.* **2002**, 11, 739–56.
- (58) Dobson, C. M. *Nature* **2003**, 426, 884–890.
- (59) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, 293, 321–331.
- (60) Wright, P. E.; Dyson, H. J. *Nat. Rev. Mol. Cell Biol.* **2014**, 16, 18–29.
- (61) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, 6, 197–208.
- (62) Krois, A. S.; Ferreon, J. C.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, 113, E1853–62.
- (63) Wright, P. E.; Dyson, H. J. *Nat. Rev. Mol. Cell Biol.* **2015**, 16, 18–29.
- (64) Iakoucheva, L. M. *Nucleic Acids Res.* **2004**, 32, 1037–1049.
- (65) Mollica, L.; Bessa, L. M.; Hanouille, X., et al. *Front. Mol. Biosci.* **2016**, 3, 52.
- (66) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, 104, 3607–3622.

- (67) Tsernoglou, D; Tsugita, A; Tucker, A. D.; van der Vliet, P. C. *FEBS Lett.* **1985**, 188, 248–52.
- (68) Tucker, P. A.; Tsernoglou, D; Tucker, A. D., et al. *EMBO J.* **1994**, 13, 2994–3002.
- (69) Dekker, J; Kanellopoulos, P. N.; van Oosterhout, J. A., et al. *J. Mol. Biol.* **1998**, 277, 825–38.
- (70) Neri, D.; Billeter, M.; Wider, G.; Wuthrich, K. *Science* **1992**, 257, 1559–1563.
- (71) Salmon, L.; Nodet, G.; Ozenne, V., et al. *J. Am. Chem. Soc.* **2010**, 132, 8407–8418.
- (72) *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*; Felli, I. C., Pierattelli, R., Eds.; Advances in Experimental Medicine and Biology, Vol. 870; Springer International Publishing: Switzerland, 2015.
- (73) Merzbacher, E., *Quantum mechanics*, 2d ed; J. Wiley: New York, 1970.
- (74) Cavanagh, J., *Protein NMR spectroscopy: principles and practice*, 2nd ed; Academic Press: Amsterdam, 2007.
- (75) Bloch, F. *Phys. Rev.* **1946**, 70, 460–474.
- (76) Levitt, M. H. *J. Magn. Reson.* **1997**, 126, 164–182.
- (77) Wishart, D. S.; Sykes, B. D.; Richards, F. M. *Biochemistry* **1992**, 31, 1647–1651.
- (78) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. *Protein Sci.* **2006**, 15, 2795–2804.
- (79) Berjanskii, M. V.; Wishart, D. S. *Nucleic Acids Res.* **2007**, 35, W531–7.
- (80) Braun, D.; Wider, G.; Wuthrich, K. *J. Am. Chem. Soc.* **1994**, 116, 8466–8469.
- (81) Wishart, D. S.; Bigam, C. G.; Holm, A.; Hodges, R. S.; Sykes, B. D. *J. Biomol. NMR* **1995**, 5, 67–81.
- (82) Schwarzing, S; Kroon, G. J.; Foss, T. R., et al. *J. Am. Chem. Soc.* **2001**, 123, 2970–2978.
- (83) Kjaergaard, M.; Poulsen, F. M. *J. Biomol. NMR* **2011**, 50, 157–165.
- (84) Wang, Y.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, 124, 14075–14084.
- (85) De Simone, A.; Cavalli, A.; Hsu, S.-T. D.; Vranken, W.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, 131, 16332–16333.
- (86) Tamiola, K.; Acar, B.; Mulder, F. A. A. *J. Am. Chem. Soc.* **2010**, 132, 18000–18003.
- (87) Chatterjee, S; Hadi, A. S. *Stat. Sci.* **1986**, 1, 415–416.

- (88) Cook, R. D.; Weisberg, S. *Technometrics* **1980**, 22, 495–508.
- (89) Welsch, R. E.; Kuh, E. *Natl. Bur. Econ. Res. Work. Pap. Ser.* **1977**, 173, 923–977.
- (90) Welsch, R. E.; Peters, S. C. *Proc. of the Elev. interface Symp. Comput. Sci. Stat.* **1978**, 240–244.
- (91) Belsley, D. A.; Kuh, E.; Welsch, R. E. *J. Oper. Res. Soc.* **1981**, 32, 157–158.
- (92) Klema, V.; Laub, A. J. *IEEE Trans. Automat. Contr.* **1980**, 25, 164–176.
- (93) Alter, O.; Brown, P. O.; Botstein, D. *Proc. Natl. Acad. Sci.* **2000**, 97, 10101–10106.
- (94) Furnas, G. W.; Deerwester, S.; Dumais, S. T. In *Proc. 11th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* 1988, pp 465–480.
- (95) Boardman, J. W. In *Geosci. Remote Sens. Symp. 1989. IGARSS'89. 12th Can. Symp. Remote Sens.* 1989, pp 2069–2072.
- (96) Kanjilal, P. P.; Palit, S.; Saha, G. *IEEE Trans. Biomed. Eng.* **1997**, 44, 51–59.
- (97) Maciejewski, A. A.; Klein, C. A. *Int. J. Rob. Res.* **1989**, 8, 63–79.
- (98) Björck, Å. In *Linear Algebra Appl.* 1994, pp 297–316.
- (99) Tompa, P. *Trends Biochem. Sci.* **2002**, 27, 527–533.
- (100) Bartlett, A. I.; Radford, S. E. *Nat. Struct. Mol. Biol.* **2009**, 16, 582–588.
- (101) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, 37, 215–246.
- (102) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, 128, 52204.
- (103) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, 17, 3–14.
- (104) Shortle, D. R. *Curr. Opin. Struct. Biol.* **1996**, 6, 24–30.
- (105) Eliezer, D. *Methods Mol. Biol.* **2007**, 350, 49–67.
- (106) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, 19, 23–30.
- (107) Wishart, D. S.; Case, D. A. *Methods Enzymol.* **2001**, 338, 3–34.
- (108) Mulder, F. A.; Filatov, M. *Chem. Soc. Rev.* **2010**, 39, 578–590.
- (109) Berjanskii, M.; Wishart, D. S. *Nat. Protoc.* **2006**, 1, 683–688.
- (110) Berjanskii, M. V.; Wishart, D. S. *J. Biomol. NMR* **2008**, 40, 31–48.
- (111) Chatterjee, S.; Hadi, A. S., *Regression Analysis by Example*; John Wiley & Sons: 2013.
- (112) Boettcher, J. M.; Hartman, K. L.; Lador, D. T., et al. *Biomol. NMR Assign.* **2007**, 1, 167–169.

- (113) Zahn, R.; Liu, A.; Luhrs, T., et al. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 145–150.
- (114) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173–195.
- (115) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F., et al. *Nucleic Acids Res.* **2008**, *36*, 402–408.
- (116) Eliezer, D.; Barré, P.; Kobaslija, M., et al. *Biochemistry* **2005**, *44*, 1026–1036.
- (117) Mukrasch, M. D.; Bibow, S.; Korukottu, J., et al. *PLoS Biol.* **2009**, *7*, e34.
- (118) Radhakrishnan, I.; Pérez-Alvarado, G. C.; Dyson, H. J.; Wright, P. E. *FEBS Lett.* **1998**, *430*, 317–322.
- (119) Lacroix, E.; Viguera, A. R.; Serrano, L. *J. Mol. Biol.* **1998**, *284*, 173–191.
- (120) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *J. Biomol. NMR* **2009**, *43*, 63–78.
- (121) Bermel, W.; Bertini, I.; Felli, I. C., et al. *J. Am. Chem. Soc.* **2006**, *128*, 3918–3919.
- (122) Mäntylahti, S.; Aitio, O.; Hellman, M.; Permi, P. *J. Biomol. NMR* **2010**, *47*, 171–181.
- (123) Goddard, T.; Kneller, D. Sparky 3., 2008.
- (124) Paz, A.; Zeev-Ben-Mordehai, T.; Lundqvist, M., et al. *Biophys. J.* **2008**, *95*, 1928–1944.
- (125) Bodenhausen, G.; Ruben, D. J. *Chem. Phys. Lett.* **1980**, *69*, 185–189.
- (126) Dijkstra, K.; Kroon, G. J.; van Nuland, N. A.; Scheek, R. M. *J. Magn. Reson. Ser. A* **1994**, *107*, 102–105.
- (127) Wittekind, M.; Mueller, L. *J. Magn. Reson. Ser. B* **1993**, *101*, 201–205.
- (128) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
- (129) Uversky, V. N. *Curr. Protein Pept. Sci.* **2008**, *9*, 507–40.
- (130) Uversky, V. N. *J. Biomol. Struct. Dyn.* **2003**, *21*, 211–234.
- (131) Weinreb, P. H.; Zhen, W.; Poon, A. W.; Conway, K. A.; Lansbury, P. T. *Biochemistry* **1996**, *35*, 13709–13715.
- (132) Cho, M.-K.; Nodet, G.; Kim, H.-Y., et al. *Protein Sci.* **2009**, *18*, 1840–1846.
- (133) Bellani, S.; Sousa, V. L.; Ronzitti, G., et al. *Commun. Integr. Biol.* **2010**, *3*, 106–9.
- (134) Rao, J. N.; Jao, C. C.; Hegde, B. G.; Langen, R.; Ulmer, T. S. *J. Am. Chem. Soc.* **2010**, *132*, 8657–68.
- (135) Cahill, C. M.; Lahiri, D. K.; Huang, X.; Rogers, J. T. *Biochim. Biophys. Acta - Gen. Subj.* **2009**, *1790*, 615–628.

- (136) McCormack, A. L.; Di Monte, D. *Curr. Protein Pept. Sci.* **2009**, *10*, 476–482.
- (137) Jakes, R; Spillantini, M. G.; Goedert, M *FEBS Lett.* **1994**, *345*, 27–32.
- (138) Kahle, P. J.; Neumann, M; Ozmen, L, et al. *J. Neurosci.* **2000**, *20*, 6365–73.
- (139) Fortin, D. L.; Troyer, M. D.; Nakamura, K., et al. *J. Neurosci.* **2004**, *24*, 6715–6723.
- (140) Ruipérez, V.; Darios, F.; Davletov, B. *Prog. Lipid Res.* **2010**, *49*, 420–428.
- (141) Chandra, S.; Chen, X.; Rizo, J.; Jahn, R.; Südhof, T. C. *J. Biol. Chem.* **2003**, *278*, 15313–15318.
- (142) Fortin, D. L.; Nemani, V. M.; Voglmaier, S. M., et al. *J. Neurosci.* **2005**, *25*, 10913–10921.
- (143) Spillantini, M. G.; Crowther, R. A.; Jakes, R, et al. *Neurosci. Lett.* **1998**, *251*, 205–8.
- (144) Bertoncini, C. W.; Jung, Y.-S.; Fernandez, C. O., et al. *Proc. Natl. Acad. Sci.* **2005**, *102*, 1430–1435.
- (145) Wu, K.-P.; Baum, J. *J. Am. Chem. Soc.* **2010**, *132*, 5546–5547.
- (146) De Genst, E. J.; Guilliams, T.; Wellens, J., et al. *J. Mol. Biol.* **2010**, *402*, 326–343.
- (147) Jao, C. C. *Proc. Natl. Acad. Sci.* **2004**, *101*, 8331–8336.
- (148) Bortolus, M.; Tombolato, F.; Tessari, I., et al. *J. Am. Chem. Soc.* **2008**, *130*, 6690–6691.
- (149) Rhoades, E.; Ramlall, T. F.; Webb, W. W.; Eliezer, D. *Biophys. J.* **2006**, *90*, 4692–4700.
- (150) Ulmer, T. S.; Bax, A.; Cole, N. B.; Nussbaum, R. L. *J. Biol. Chem.* **2005**, *280*, 9595–9603.
- (151) Philip, A. F.; Kumauchi, M.; Hoff, W. D. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17986–91.
- (152) Hoff, W. D.; Dux, P.; Hard, K., et al. *Biochemistry* **1994**, *33*, 13959–13962.
- (153) Rubinstenn, G.; Vuister, G. W.; Mulder, F. A. A., et al. *Nat. Struct. Biol.* **1998**, *5*, 568–570.
- (154) Bernard, C.; Houben, K.; Derix, N. M., et al. *Structure* **2005**, *13*, 953–962.
- (155) LeMaster, D. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **1994**, *26*, 371–419.
- (156) Kumar, S.; Nussinov, R. *Chembiochem* **2002**, *3*, 604.
- (157) Alexov, E.; Mehler, E. L.; Baker, N., et al. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 3260–3275.

- (158) Wallerstein, J.; Weininger, U.; Khan, M. A. I.; Linse, S.; Akke, M. *J. Am. Chem. Soc.* **2015**, *137*, 3093–3101.
- (159) Hass, M. A.; Mulder, F. A. *Annu. Rev. Biophys.* **2015**, *44*, 53–75.
- (160) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.
- (161) Tanford, C. *Physical chemistry of macromolecules*; Wiley: New York, 1961.
- (162) Das, R. K.; Pappu, R. V.; Bhowmick, A., et al. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *3*, 10219–10225.
- (163) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721–33.
- (164) Bombarda, E.; Ullmann, G. M. *J. Phys. Chem. B* **2010**, *114*, 1994–2003.
- (165) Elcock, A. H. *J. Mol. Biol.* **1999**, *294*, 1051–1062.
- (166) Zhou, H.-X. *J. Phys. Chem. B* **2001**, *105*, 6763–6766.
- (167) Zhou, H.-X. *Biophys. J.* **2002**, *83*, 2981–2986.
- (168) Zhou, H.-X. *Proc. Natl. Acad. Sci.* **2002**, *99*, 3569–3574.
- (169) Geist, L.; Henen, M. A.; Haiderer, S., et al. *Protein Sci.* **2013**, *22*, 1196–1205.
- (170) Gilson, M. K. *Proteins Struct. Funct. Genet.* **1993**, *15*, 266–282.
- (171) Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins Struct. Funct. Genet.* **1993**, *15*, 252–265.
- (172) Van Rossum, G.; van Rossum, G. *Python tutorial*; tech. rep.; Amsterdam: Centrum voor Wiskunde en Informatica (CWI), 1995.
- (173) Hykes, S. Docker., 2013.
- (174) Gear, C. W.; Butcher, J. C.; Dahlquist, G., et al. *Math. Comput.* **1967**, *21*, 146.
- (175) Lindman, S.; Linse, S.; Mulder, F. A. A.; André, I. *Biochemistry* **2006**, *45*, 13993–14002.
- (176) Croke, R. L.; Patil, S. M.; Quevreaux, J.; Kendall, D. A.; Alexandrescu, A. T. *Protein Sci.* **2011**, *20*, 256–269.
- (177) George, J. M. *Genome Biol.* **2002**, *3*, 3002–3010.
- (178) Wood, S. J.; Wypych, J.; Steavenson, S., et al. *J. Biol. Chem.* **1999**, *274*, 19509–19512.
- (179) Cookson, M. R. *Annu. Rev. Biochem.* **2005**, *74*, 29–52.
- (180) Receveur-Brechot, V.; Durand, D. *Curr. Protein Pept. Sci.* **2012**, *13*, 55–75.
- (181) Nielsen, J. T.; Mulder, F. A. A. *Front. Mol. Biosci.* **2016**, *3*, 4.
- (182) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13099–13104.

- (183) Bhowmick, A.; Brookes, D. H.; Yost, S. R., et al. *J. Am. Chem. Soc.* **2016**, *138*, 9730–9742.
- (184) Tamiola, K.; Mulder, F. A. a. a. *Biochem. Soc. Trans.* **2012**, *40*, 1014–1020.
- (185) Linse, S.; Jönsson, B.; Chazin, W. J. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 4748–4752.
- (186) Moser, P.; Squire, P. G.; O’Konski, C. T. *J. Phys. Chem. A* **1966**, *70*, 744–756.
- (187) Bibi, F.; Villain, M.; Guillaume, C.; Sorli, B.; Gontard, N. *Sensors* **2016**, *16*, 1232.
- (188) Das, R. K.; Pappu, R. V. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 13392–13397.
- (189) Geist, L.; Henen, M. A.; Haiderer, S., et al. *Protein Sci.* **2013**, *22*, 1196–205.
- (190) Nielsen, J. T.; Mulder, F. A. A. *J. Biomol. NMR* **2018**, *70*, 141–165.
- (191) Jung, Y.-S.; Zweckstetter, M. *J. Biomol. NMR* **2004**, *30*, 11–23.
- (192) Zawadzka-Kazimierczuk, A.; Koźmiński, W.; Billeter, M. *J. Biomol. NMR* **2012**, *54*, 81–95.
- (193) Sekhar, A.; Rosenzweig, R.; Bouvignies, G.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 10395–400.
- (194) Kulkarni, P.; Uversky, V. N. *Proteomics* **2018**, *18*, e1800061.
- (195) Wei, L.; Zou, Q. *Int. J. Mol. Sci.* **2016**, *17*, DOI: 10.3390/ijms17122118.
- (196) Service, R. *Google’s DeepMind aces protein folding*; tech. rep.; Science Magazine Blog, 2018.
- (197) Ravasz, E. *Science* **2002**, *297*, 1551–1555.
- (198) Uversky, V. N. *J. Biol. Chem.* **2016**, *291*, 6681–6688.
- (199) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404–406.
- (200) Dobson, C. M.; Karplus, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92–101.
- (201) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6679–6685.
- (202) Perdomo-Ortiz, A.; Dickson, N.; Drew-Brook, M.; Rose, G.; Aspuru-Guzik, A. *Sci. Rep.* **2012**, *2*, 571.
- (203) Levitt, M. *Nat. Struct. Biol.* **2001**, *8*, 392–393.
- (204) Nussinov, R. *PLoS Comput. Biol.* **2015**, *11*, e1004053.
- (205) Tomita, M. *Trends Biotechnol.* **2001**, *19*, 205–210.
- (206) Loew, L. M.; Schaff, J. C. *Trends Biotechnol.* **2001**, *19*, 401–406.
- (207) Takahashi, K.; Kaizu, K.; Hu, B.; Tomita, M. *Bioinformatics* **2004**, *20*, 538–546.

- (208) Brette, R.; Rudolph, M.; Carnevale, T., et al. *J. Comput. Neurosci.* **2007**, *23*, 349–398.
- (209) Caspi, R.; Altman, T.; Dale, J. M., et al. *Nucleic Acids Res.* **2010**, *38*, D459–D471.
- (210) Muth, T.; Peters, J.; Blackburn, J.; Rapp, E.; Martens, L. *J. Proteomics* **2013**, *88*, 104–108.

Acknowledgements

There are quite likely N -number of ways I could start this final (and likely the most important) chapter of my thesis, where N is a cosmological number that expresses my gratitude for all the help and support I have received during my years at University of Groningen, both whilst the Top–Master Biomolecular Sciences Programme, as well as my Doctoral Research.

Let me, however, open this chapter with the words of my supervisor, **Prof. Frans A. A. Mulder** who did conclude one of his excellent lectures during the Dutch NMR Society meeting in Veeldhoven, NL on the 6th of December 2010 in the following way,

"This work is not an achievement of one person, but an effort of many."

Indeed, the work presented in this thesis, as well as anything I have built around that time, is most definitely not just my achievement, but an effort of very many brilliant and unbelievably patient people. They have helped me to shape my scientific interests, but also have had a profound impact on my life and the understanding of the surrounding world.

Before I start personal acknowledgements, I would like to thank the authorities of *University of Groningen*, more specifically *Faculteit Wiskunde en Natuurwetenschappen* and *Groningen Biomolecular Sciences Institute (GGB)* for accepting me to the Top–Master Biomolecular Sciences programme, and subsequently absolutely frictionless transition to doctoral studies. Whether in sun, rain or horribly windy weather, my daily bicycle commute from the very heart of Groningen to Zernike Complex, and our research facilities at *Nijenborgh 4* has always put a smile on my face, and it does remain like this, till now.

I would also like to extend my thanks to members of my PhD Assessment Committee, Prof. Siewert–Jan Marrink, Prof. Alexandere Bonvin and Prof. Wim Vranken for their time and effort put into reading my work.

Prof. Dick B. Janssen Dear Dick, words cannot describe how grateful I am for your help, patience and support over all those years. Although your laboratory is well known for impeccable experimental research, you have entrusted me with a Top Master project, which de facto had very little to do with experimentation, and focused solely on numerical modelling. After that I have started my work on Dh1A, which once again got derailed and (quite incredibly) evolved into my 'obsession' with disordered proteins. My doctoral research project had many ups and downs, yet I knew I could always count on your help and constant support. I just want you to know that you have made a very significant mark on my life, and whenever I pick up a pen and start jotting down a scientific idea, I ask myself: how would Dick Janssen do it? Working with you was a great honour and I am looking forward to catching up and endlessly talking about great climbs in Mont Blanc Massif.

Prof. Frans A.A. Mulder Dear Frans, there are no words that can describe my gratitude for all the effort and time you have invested in teaching me the very principles of NMR spectroscopy and (truth to be told) helping me to become an independent researcher. The said 'investment' came with massive risk profile born from my immaturity, stubbornness and plain stupidity. Only now, when I am working on a small scientific venture of my own, I can comprehend the magnitude of responsibility and stress that you must have felt whilst leading our NMR group. The fact is I should have known better and this thesis should have been written sooner. I want you to know I have sincerely admired your intellect, incredibly amusing and sarcastic comments and impeccable scientific integrity. I feel incredibly proud about the work I was able to complete under your supervision and I am hoping we could 'pull' some nice papers in the future.

Prof. Gerrit Poelarends Dear Gerrit, I wanted to cordially thank you for accepting and putting up with me, as an aspiring wet-lab experimentalist. The experience born from wet-lab work was invaluable in making my career-defining decision, that is devoting myself entirely to NMR spectroscopy. I was always profoundly impressed by our scientific intuition and incredible determination.

Dr. Renee Otten Dear René, I am eternally grateful for your amazing company, help and guidance. I did enjoy our conversations and scientific exchange tremendously, and I must admit I was always profoundly impressed by your knowledge and honesty. It was absolutely amazing to meet you after all those years in Cambridge, MA and have a decent round of beers, 'Pintelier-style'. I wish you and your lovely wife the very best.

Dr. Ruud Scheek My dear Ruud (also known as Rude Mean and Square, as we used to joke), I wanted to thank you from the bottom of my heart, for all your time invested in explaining the very principles of NMR spectroscopy. Every coffee and tea break in our cafeteria has always brought a smile on my face and made me seriously concerned about my religious beliefs. These are very precious moments, I will always treasure. Finally, only now, after spending some time in the Silicon Valley, I understand your love for California and Berkeley.

Ing. Klaas Dijkstra Dear Klaas, I wanted to thank you for all the hours put into our experiments and an amazing opportunity to see the very technical aspects of NMR. I simply loved observing your work in our little 'technical corner', especially when you were disassembling our NMR probes.

Ing. Pieter van der Meulen Dear Pieter, I wanted to thank you for countless hours spent on classical NMR experiments and our little, and I dare say, crazy electrostatic-field alignment idea. I am so happy we haven't inflicted any permanent damage in our NMR facility. I have always admired your incredible precision and passion for experimentation. I do miss your peculiar but spot-on sense of humour.

Dr. Alex de Vries My dearest Alex, I want you to know the time we have shared an office at Nijeborgh 4 building were simply the most intellectually- and artistically- stimulating I ever had. I do miss profoundly our conversations about force fields, molecular dynamics, teaching physics and classical music. I am very grateful for an amazingly inclusive atmosphere you have created over so many years for me.

Dr. Xavier Periole Monsieur Periole, thank you so much for your friendship, guidance and support through all these years. Many conversations we had about life and science still resound in my mind. I have always felt your

support throughout my academic career, and I want you to know I am profoundly grateful for your friendship and mentorship.

Dr. Francesco Maddalena Mio caro Francesco, ti ringrazio per la tua amicizia, per le fantastiche conversazioni e per le avventure altrettanto divertenti intercorse durante i miei primi giorni da studente a Groningen per il Top Master. Voglio che tu sappia che sono sempre stato profondamente colpito dalla tua intelligenza ed espressività. Auguro a te e alla tua cara moglie il meglio in Corea del Sud e non vedo l'ora di incontrarvi in Friuli.

Dr. Antonio Careta Mio caro Bibi, sono eternamente grato per la tua amicizia, il tuo umorismo, la tua compagnia e le nostre meravigliose arrampicate. Non dimenticherò mai tutto l'aiuto e il supporto che da te ho ricevuto nei momenti più impegnativi della mia vita. Spero condivideremo presto nuove avventure, sia scientificamente parlando sia per quanto riguarda l'arrampicata. Inoltre non vedo l'ora di incontrare Mr Sebastiano Careta.

Dr. Davide Fracasso Mio caro Ragazzo Milano, ti ringrazio per la fantastica atmosfera, la compagnia e l'ottimo caffè preparato sulla tua piccola e bizzarra piastra elettrica.

Dr. Djuure de Jong My dear Djuure, I wanted to thank you for your friendship and an amazing company. Only recently I have opened the books from the NMR Masterclass and Summer School we have attended at University of Nijmegen. I can vividly remember the hardcore quantum physics lectures, interwoven with amazingly warm summer evenings out in Nijmegen and the final karaoke night. Good ol' times!

Prof. Siewiert J. Marrink I want to cordially thank you Siewiert for creating an amazingly welcoming and family-like atmosphere in Molecular Dynamics laboratory. I have been dreaming of working in this lab from the very beginning of my academic curriculum at University of Groningen, and the actual experience surpassed my expectations, greatly! Both scientifically, but even more on the personal level!

Dr. Durba Sengupta My dearest Durba, your smile, laugh and positivity were always highly contagious. Did I mention amazing alcohol-spiced brown-

ies? I am very grateful for those funny conversations we had whilst coffee breaks and group outings.

Dr. Jelger Risselada Jelger, beyond any doubt you are one of the most intelligent and socially awkward people I have ever met. Our conversations and scientific divagations were anything but normal, but I want you to know I have enjoyed every single one of them, and I am very grateful for your companionship.

Dr. Tsjerk Wassenaar Tsjerk, I did enjoy very much our conversations about molecular dynamics, statistic and Friesland. I wanted to thank you for all the time you have invested in showing me how to use GROMACS in the most cost-effective way. I especially liked the conversations about periodic boundary conditions. I want you to know I have always been inspired by your visualization and data presentation skills. A true rarity, yet an asset that many people take for granted. Dank U wel!

Dr. Nur Alia Oktaviani Alia, thank you for your kindness, sense of humour and politeness. I do hope I was a sufferable fellow doctoral student, despite my childish practical jokes (especially scarring the hell out of you in the NMR laboratory).

Prof. Herman Berendsen Professor Berendsen was always able to find a time after our Friday Molecular Dynamics Group meetings and answer all my questions. Not to mention the fact he was instrumental in the development and testing of the wicked idea behind the electrostatic-field induced alignment of proteins in NMR.

Prof. Roelf Boelens Dear Roelf, I do want to thank you for all the inspirational conversations we had whilst our short meetings during Benelux NMR Discussion Group meetings and NMR conferences. I did enjoy them very much. I can also vividly remember the smirk on your face, when together with Ing. Peter van der Meulen, we have appeared at your doorstep at Utrecht University NMR centre and asked if we can pick up power amplifiers for our 'wicked' oscillating electrostatic-field alignment experiments. The chance of success with these was slim, not to mention we were likely to damage this equipment. Yet, there we were and you helped us without any hesitation. I will always remember that. Dank U wel!

Dr. Magdalena Siwko–Rampioni Dear Magda, it is you who has introduced me to University of Groningen and the Molecular Dynamics research group. Besides that you have been always very supportive and helpful. I want to thank you cordially for your friendship and warmth whilst our time at University of Groningen.

Dr. Engel Vrieling Dear Engel, you were very likely one of the first RuG staff members who has welcomed me to the academic world at University of Groningen. I wanted to cordially thank you for all your help, effort and goodwill. I am finally happy to report, my journey at University of Groningen has neared to its end and the most desired outcome, my doctoral degree. Dank U wel!

Theo Jurriens Dear Theo, thank you so much for all your help concerning the very first days of my stay in Groningen and The Netherlands, in general. I have very much enjoyed my stay at Rabenhauptstraat 32, the studio apartment you have found for me whilst my Top–Master programme. I have made many Dutch friends there, and we still stay in touch. Dank U wel!

Geert van der Veen Dear Geert, thank you so much for welcoming me in Bjoeks, letting me practice and photograph on Excalibur and an excellent and friendly atmosphere. You have always had a great interest in my scientific work and PhD, and here we are. It is done!

Marco Tamponi Caro Marco, sono molto grato per tutti i fantastici servizi fotografici, le nostre avventure e arrampicate sul Massiccio del Monte Bianco. Mi aspetto molte altre nuove avventure e altrettanti bei racconti riguardo il genepi fatto in casa.

Santiago Katz Dear Santi, I want you to know I count the time we have spent living together on Herestraat in the very heart of Groningen as the most amazing and exhilarating moments of my life. I do look forward to catching up soon in London or Texas, USA.

Navid Saleki Dear Navidir, I am eternally grateful for your friendship, support and countless hours spent talking about life, adventures and big plans. I am very happy you have landed in Italy and I wish you the very best in your life. You will always be my best friend, no matter where we are.

Dr. Matthew M. Heberling My dear Blind Sparrow, there are no words that can describe how grateful I am for your friendship and all the energy you have put in our little venture. You have always been the 'good guy' in our 'Peptone duo', something that I never take for granted. I can't wait to host you and Raquel here in the Alps and have a well deserved rest for both of our families.

La Famiglia Perosa Cari Donatella, Cristiano, Marco, Paolo e Bruno, grazie di cuore per aver reso ogni soggiorno a Palmanova e Gorizia un vero piacere. Le mie parole non saranno mai abbastanza per esprimere quanto il vostro supporto, i vostri consigli e la straordinaria compagnia mi abbiano aiutato nel trovare la motivazione per finire questo lavoro. Non vedo davvero l'ora di condividere con voi molti nuovi meravigliosi momenti e spero di potervi ripagare con il mio affetto e con il mio sostegno.

Moja rodzina Moi drodzy, dziękuję wam bardzo za zainteresowanie moją pracą. Zawsze czułem wasze wsparcie. Prawdą jest fakt, że ta dysertacja powinna zostać napisana wcześniej, jednakże do niektórych rzeczy trzeba dorosnąć i nie będę ukrywał, że to "dorastanie" zajęło trochę czasu. Nie mogę się doczekać kiedy was wszystkich zobaczę i wręczę kopię tej małej książki. Dziękuję wam bardzo za wszystko!

Babcia, Mama i Bartek Moja kochana Mamo, Babciu i Bartku. Prawdą jest fakt, że bez waszej bezwzględnej miłości, wsparcia, zaangażowania i ciężko zarobionych pieniędzy mój doktorat nigdy nie doszedłby do skutku. Zdaje sobie doskonale sprawę, że zawdzięczam wam, waszej pracy i poświęceniu praktycznie wszystko. Spłatenie długu wdzięczności jaki wobec was mam, nie będzie możliwe, bo nie istnieje cena za takie poświęcenie i zaufanie. Obiecać jedynie mogę, że to wszystko co od was dostałem, "przekuje" w miłość i poświęcenie dla mojego syna i podążając waszym przykładem zawsze będę go wspierał. Kocham was bardzo i choć bardzo lubiłem moje życie w Holandii, nie było dnia żebym nie tęsknił za naszym starachowickim domem i rodziną.

Francesca Perosa My dearest princess, there are no words, which could describe how grateful I am for your love, support and help. When we first met, I knew you were not an ordinary lady. Nothing however, could have prepared me for unconditional love and support I have received from you

during all those years. I am desperately hoping I will be able to offer some of this help and support back, while we expect our son, Leonardo. I love you with all my heart.

Publications

Science and Art

Fig. 9.1: On assignment with Phase One and their film crew 3400m above Courmayeur Mont-Blanc, Valle d'Aosta, Italy. Source: Phase One

Besides an academic curriculum, Kamil Tamiola has worked as a commercial outdoor photographer specializing in adventure and mountain sports. His work has been featured in *The Atlantic*, *New York Times*, *National Geographic*, *Outdoor Photography Magazine*, and many other popular outlets. His work has been used in commercial campaigns by *Apple*, *G-Technology*, *Nikon*, *Phase One*, *Jeep*.

Kamil is a registered member and artist in *Nederlands Instituut voor Kunstgeschiedenis (KDR)* (Record: 446160).

9.1 Scientific publications

"Preparation of D-amino acids by enzymatic kinetic resolution using a mutant of penicillin-G acylase from *E. coli*." Carboni C.; Kierkels H.G.T.; Gardossi L.; **Tamiola K.**; Janssen D.B.; Quaedflieg P.J.L.M. *Tetrahedron: Asymmetry*, **2006**, 17(2), 245–251.

DOI: 10.1016/j.tetasy.2005.12.023

"Deeply branching c6-like cytochromes of cyanobacteria." Bialek W.; Nelson M.; **Tamiola K.**; Kallas T.; Szczepaniak A. *Biochemistry* **2008**, 47(20), 5515–22.

DOI: 10.1021/bi701973g

"Probing the free energy landscape of the FBP28WW domain using multiple techniques." Periole X.; Allen L.R.; **Tamiola K.**; Mark A.E.; Paci E. *J. Comput. Chem.* **2009**, 30(7), 1059–68.

DOI: 10.1002/jcc.21128

"Sequence-specific random coil chemical shifts of intrinsically disordered proteins." **Tamiola K.**; Acar B.; Mulder F.A.A. *J. Am. Chem. Soc.* **2010**, 132, 18000–3.

DOI: 10.1021/ja105656t

"ncIDP-assign: a SPARKY extension for the effective NMR assignment of intrinsically disordered proteins." **Tamiola K.**; Mulder F.A.A. *Bioinformatics* **2011**, 27, 1039–40.

DOI: 10.1093/bioinformatics/btr054

"Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins." **Tamiola K.**; Mulder F.A.A. *Biochem. Soc. Trans.* **2012**, 40, 1014–1020.

DOI: 10.1042/BST20120171

"Controlled long-exposure imaging of a celestial object." **Tamiola K.**; Toma L. *USA Patent Office* **2016**, 15/034, 488

Source: US Patent App. 15/034,488

"Structural propensity database of proteins" **Tamiola K.**; Heberling M.M.; Domanski J.; *bioArxiv*, **2017**

DOI: <https://doi.org/10.1101/144840>

"pepKalc: scalable and comprehensive calculation of electrostatic interactions in random coil polypeptides." **Tamiola. K.**; Scheek R.M.; van der Meulen P.; Mulder F.A.A. *Bioinformatics* **2018**, *bty033*

DOI: [10.1093/bioinformatics/bty033](https://doi.org/10.1093/bioinformatics/bty033)

9.2 Art and popular science publications

"Spectacular photos of National Geographic Traveler Photo Contest 2012" *The Atlantic*, **2012**

Source: The Atlantic Online

"Naar de Top." *Teylers Magazijn*, **2012**, 31(115), 8–9

Source: Teylers Museum

"Cave new world: Photographer captures breathtaking and ever-changing ice caverns of the French Alps." *Daily Mail*, **2012**

Source: Daily Mail Online

"Wie is fotograaf Kamil Tamiola?" *Hoogtelijn*, **2013**

Source: Hoogtelijn Online

"Imagine yourself and night photography." *TedX Talks*, **2014**

Source: TedX

"De Melkweg boven de Alpen." *National Geographic Traveler*, **2015**

Source: National Geographic Traveler Online

"Luck Favours The Prepared. Kamil Tamiola brings the scientific method to alpine adventure photography." *Outdoor Photographer Magazine*, **2015**

Source: Outdoor Photographer Magazine Online

"Underwater Pixelguide, presented by Kamil Tamiola." *Amazon Prime*,
2016
Source: Amazon Prime