Intrinsically Disordered Proteins

Maeve Andersen

December 4, 2023

Abstract

The structure and dynamics of intrinsically disordered proteins represent an active and exciting area of research in biophysics. Previous methods developed to describe protein structure and dynamics assume the protein to be ordered, in that the protein has a native conformation that is determined by its sequence. However, this is not the case with proteins that exhibit disorder. In this review, I will briefly overview how proteins with disorder are modeled today. Additionally, I will review experimental methods for measuring disorder and why these must be combined with theory in a hybrid approach to characterize proteins.

1 Introduction

In 1894, Fischer released a paper that presented the lock-and-key mechanism to explain enzyme function. The mechanism models a protein as a key of sorts, where the protein's unique shape determines its unique biological function (the lock). [9]. This way of thinking about functional proteins as rigid key-like objects would dominate the field of protein science for the next century.

Today however, proteins without a rigid 3D structure are being studied. These proteins without rigid structure are said to exhibit disorder, and modern models of proteins are beginning to account for this [21]. Thus, intrinsically disordered proteins represent an active area of research in protein science. Characterization of disorder in proteins is important as disordered proteins are involved in cellular signaling and regulation,[23] and are associated with human diseases, such as neurodegenerative disease, cardiovascular disease, amyloidoses, cancer, and diabetes.[22] Although challenging, modern methods of characterizing proteins can provide new insights to crucial protein function human biological mechanisms. [2].

2 Conformational Ensembles

In order to describe the structure of proteins with a high degree of heterogeneity, conformational ensembles are used [15],[6],[21]. When determining conformational ensembles, neither experiment nor computational model can be used in isolation [6], [15], [19]. The reason why experiment fails to solely determine structure is that data collected from experiment is sparse, meaning the number of unique observations is less than the degrees of freedom [19]. Computational methods such as Monte Carlo or molecular dynamics also fall short to determine protein structure for two reasons [6]: 1) The force fields used in even the most advanced methods are still approximations. This can lead to large errors in predicted properties of proteins. 2) Limited computational resources put an upper bound on the protein simulation time. This means that these methods may "run out of time" before a conformational space is fully explored.

The most promising method of determining protein structure is to combine data from experiment with computational methods is a promising method of protein structure determination [6] ,[21] ,[15]. When doing a combined approach, usually four components are incorporated [21]. 1) One or more experiments whose data gives information on protein strucure. 2) A method to sample protein conformations computationally. 3) A forward model, which calculates experimental observables from the conformational ensemble. 4) A refinement method, which refines the ensemble based on experimental data.

It is important to note that the term "ensemble" is often not carefully used within the field of structural biology. In statistical mechanics, a protein's ensemble, given some set of conditions, is

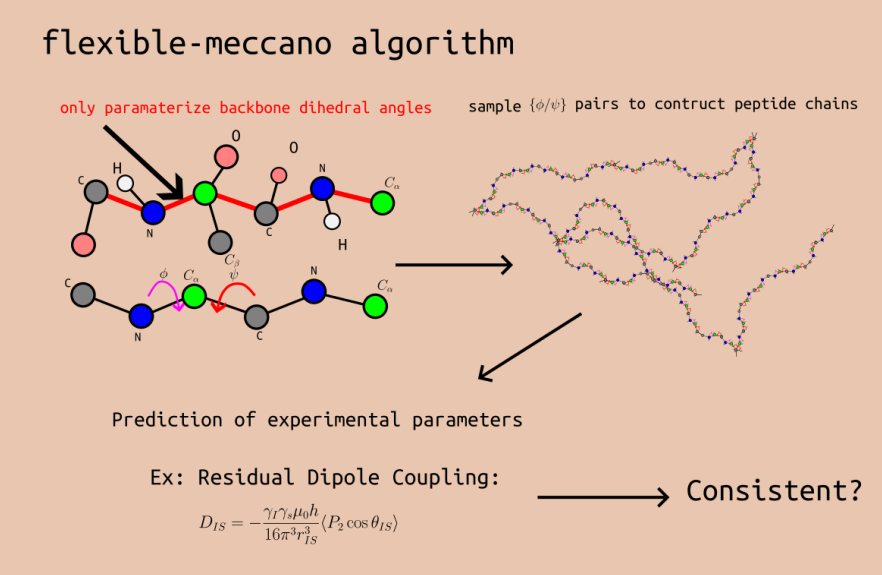

Figure 1: Flexible-meccano, an algorithm where dihedral angles are randomly sampled to construct a set of conformations. The resulting set of conformations is considered the structural ensemble. Experimental parameters are then calculated for each conformation in the ensemble. The calculated results are averaged and compared to the observed experiment. If the agreement is good, then the set of conformations may be close to the "true" ensemble [17].

a probability distribution over all possible conformations of that protein. In structural biology, the term is overloaded with several meanings.[12] There is of course a thermodynamic ensemble, which refers to the distribution of a protein's conformations under thermal equilibrium. Additionally there are uncertainty ensembles, which are collections of conformations that are degenerate due to sparse, ambiguous, or noisy data. Uncertainty ensembles may be further broken down depending on their source. For example, they may come from an arbitrary algorithm, or they may come from a well defined distribution.[12] In the following sub-section I will describe an explicit construction of one of these enembles to give a sense of how conformational ensemble are used and discussed in structural biology and biophysics.

2.1 Structural Ensemble Representation

A structural ensemble can be represented as a set of structures $s_1, ..., s_n$ and a set of population weights $\vec{w} = w_1, ..., w_n$, where $\sum_i w_i = 1$. Our goal is to find a set of "true" weights $\vec{w}^T = w_1^T, ..., w_n^T$ for protein. If one knew the relative free energy of the states, $\vec{w}^T$ could be calculated. Finding the relative free energy function is unfortunately not feasible [11]. Thus what is usually done is to find a set of weights that satisfy experimental observations.

A simple example of this is the flexible-meccano algorithm (Figure 1) [17]. This is a simple sampling approach, in which polypeptides conformations are constructed by randomly selecting dihedral angles from a set of amino-acid specific potentials. There is a potential for each of the 20 different amino-acid types as well as 3 special case potentials for prolines and glycines adjacencies. The result is a list of conformations that make up the "ensemble". Finally, observables are calculated for each of the generated conformations. The observables are then averaged over all conformations and then compared to actual experimental data. If the deviation of the observables is less than some threshold, then the conformation is considered to be of good fit.

The major problem with the above approach is that there may exist more than one set of weights whose deviation passes the threshold of experimental observables. In this case, given the deviation of $\vec{w}$ from some experimental observable M, $\zeta M(\vec{w})$, there exists N unique weight vectors, $\vec{w}_1, ...\vec{w}_N$ such that $\zeta_M(\vec{w}_l)$ is less than the threshold for each l. This degeneracy can only be lifted by making additional assumptions. One method suited to combat this degeneracy is the Bayesian weighting (BW) approach [10].

2.2 Bayesian Weighting (BW) Approach

The BW approach, [10] aims to generate a probability distribution for each of the conformation population weights. The resulting distribution is called the posterior probability distribution and is given by:

$$f_{\vec{W}|\vec{M}}\left(\vec{w}|\vec{m}\right) = \frac{f_{\vec{M}|\vec{W}}\left(\vec{m}|\vec{w}\right) f_{\vec{W}}\left(\vec{w}\right)}{\int d\vec{w}\, d_{\vec{M}|\vec{W}}\left(\vec{m}|\vec{w}\right) f_{\vec{W}}\left(\vec{w}\right)} \tag{1}$$

where $\vec{m}$ is a vector of experimental measurements, $f_{\vec{W}}(\vec{w})$ is the prior distribution, and $f_{\vec{M}|\vec{W}}\left(\vec{m}|\vec{w}\right)$ is the likelihood function.

Given a set of unique structures $s_i ... s_n$, one could use a Boltzman distribution as the estimate for the weights. Here, the Boltzman distribution for a set if unique structures is [10]:

$$w_i^P = \frac{e^{-U(s_i)/k_B T}}{\sum_j e^{-U(s_i)/k_B T}}. \tag{2}$$

In this case, the prior distribution would rely on the energy of the structures, but one could use other sources priori information besides energy. The simplest prior distribution would be a Gaussian, however in practice one does a coordinate transform since the domain of integration is an n-dimensional simplex [10]. Still, the beauty of this approach is that one can additionally choose a variance to reflect the accuracy of the underlying potential. One can even treat the variance itself as a random variable if finding uncertainties in the priori information is not feasible [10].

3 Experimental Observables

There are several experimental techniques available to measure protein structure and dynamics. In this section I will go over a select few, however more complete descriptions are available [21].

3.1 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) spectroscopy is a well established tool in the study of protein structure. An important note is that NMR measurements are all time and ensemble averaged [16], which is especially relevant when determining protein dynamics. Additionally, NMR measurements can only take place after determination of resonances, which for disordered proteins requires special treatment [16].

One NMR measurable is the paramagnetic relaxation effect (PRE). When studying disordered proteins, you can look for evidence of secondary structure or compaction by subtracting out the expected PRE of a statistical random coil [7]. Additionally, the PRE can amplify information on the existence of lowly populated states if the exchange rate between a highly populated state is fast (pico seconds to nano seconds range) [7].

PRE rates are given by the Solomen Bloemburgen Equations. These describe the longitudinal (Γ_1) and transverse ($Gamma_2$) paramagnetic relaxation effect (PRE) rates [1],[20] :

$$\Gamma_1 = \frac{2}{5}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_1^2 g^2 \mu_B^2 S(S+1) J_{SB}(\omega_I) \tag{3}$$

$$\Gamma_2 = \frac{1}{15}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_1^2 g^2 \mu_B^2 S(S+1) \tag{4}$$

where g is the electron g-factor, γ_1 is the proton gyromagnetic ratio, $\omega_I/2\pi$ is the Lamor frequency of the proton, and $J_{SB}(\omega)$ is the spectral density of the reduced correlation function:

$$J_{SB}(\omega) = r^{-6}\frac{\tau_c}{1+(\omega\tau_c)^2} \tag{5}$$

The correlation time τ_c is $\tau_c = \left(\tau_r^{-1} + \tau_s^{-1}\right)^{-1}$. Here, τ_r is the rotational correlation time of the macromolecule and t_s is the effective electron relation time. A key note is the scaling r^{-6}. This allows one to study long range PRE effects (36 Angstroms away from a paramagnetic center) [7]. This is in

contrast to the nuclear Overhauser effect (nOe), which only has a range of a few Angstroms [7]. Other parameters provided my NMR are scalar J-couplings, which provide bond connectivity and dihedral angles of proteins, and residual dipolar couplings (RDCs) which give directions of bond vectors with respect to a global alignment tensor [16]. One can also use a pre-developed framework to determine dihedral angles of a protein using combinations of RDCs and measured chemical shifts, where the combinations of the two measurements are such that the Ramachandram space of the protein is well determined [18]. Finally, using just measured chemical shifts, one can also find dihedral angles. That being said, the dihedral angles may not be uniquely determined [14].

3.2 Cryogenic Electron Microscopy

Cryogenic Electron Microscopy (Cryo-EM) can be combined with NMR observables, or other experimental data. When combining with NMR observables, one must first estimate the kinetic properties of the ensemble generated by the cryo-EM data. Then NMR observables can be incorporated. This is because NMR observables are ensemble averages, which depend on the transition rates/time scales of interconversion between states [3]. There exist several methods that determine structural ensembles using Cryo-EM. BioEM takes a set of single particle cryoEM images and a structure. Then uses a Bayesian approach to find the consistency between the two [8] .

Mosaics-EM [24] uses natural move monte carlo and simulated annealing to refine protein structure [13]. Natural Move Monte Carlo is used to model the macromolecule segments connected by flexible loops, on multiple scales. This method depends on 2D class averages measured with Cryo-EM to refine conformations. This method was applied to study Methonococcus maripaludis chaperonin, where using one class averaged projection, it refined an initial closed conformation to the open conformation. When using heterogeneous projection images, Natural Moves were used to iteratively refine in a mix of open and closed states.

Finally, Metainference Approach/Metainference for cryo-EM (EMMI) [2],[4] is a recently developed method that extends the metainference approach ([5]) to model conformational ensembles using cryo-EM. The metainference approach thrives at determining conformal heterogeneity in the low resolution areas of high resolution maps [5].

References

[1] N. Bloembergen and L. O. Morgan. "Proton Relaxation Times in Paramagnetic Solutions. Effects of Electron Spin Relaxation". In: *The Journal of Chemical Physics* 34.3 (Mar. 1, 1961), pp. 842–850. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1731684. URL: https://pubs.aip.org/jcp/article/34/3/842/79898/Proton-Relaxation-Times-in-Paramagnetic-Solutions (visited on 12/02/2023).

[2] Massimiliano Bonomi, Riccardo Pellarin, and Michele Vendruscolo. "Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy". In: *Biophysical Journal* 114.7 (Apr. 2018), pp. 1604–1613. ISSN: 00063495. DOI: 10.1016/j.bpj.2018.02.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006349518302881 (visited on 11/28/2023).

[3] Massimiliano Bonomi and Michele Vendruscolo. "Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy". In: *Current Opinion in Structural Biology* 56 (June 2019), pp. 37–45. ISSN: 0959440X. DOI: 10.1016/j.sbi.2018.10.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X18301180 (visited on 11/28/2023).

[4] Massimiliano Bonomi et al. "Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling". In: *Structure* 27.1 (Jan. 2019), 175–188.e6. ISSN: 09692126. DOI: 10.1016/j.str.2018.09.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S096921261830337X (visited on 12/02/2023).

[5] Massimiliano Bonomi et al. "Metainference: A Bayesian Inference Method for Heterogeneous Systems". In: *Science Advances* 2.1 (Jan. 2016), e1501177. ISSN: 2375-2548. DOI: 10.1126/sciadv.1501177. URL: https://www.science.org/doi/10.1126/sciadv.1501177 (visited on 12/02/2023).

[6] Massimiliano Bonomi et al. "Principles of Protein Structural Ensemble Determination". In: *Current Opinion in Structural Biology* 42 (Feb. 2017), pp. 106–116. ISSN: 0959440X. DOI: 10.1016/j.sbi.2016.12.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X16302330 (visited on 11/28/2023).

[7] G. Marius Clore and Junji Iwahara. "Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes". In: *Chemical Reviews* 109.9 (Sept. 9, 2009), pp. 4108–4139. ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr900033p. URL: https://pubs.acs.org/doi/10.1021/cr900033p (visited on 11/02/2023).

[8] Pilar Cossio and Gerhard Hummer. "Bayesian Analysis of Individual Electron Microscopy Images: Towards Structures of Dynamic and Heterogeneous Biomolecular Assemblies". In: *Journal of Structural Biology* 184.3 (Dec. 2013), pp. 427–437. ISSN: 10478477. DOI: 10.1016/j.jsb.2013.10.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S1047847713002712 (visited on 12/01/2023).

[9] Emil Fischer. "Einfluss der Configuration auf die Wirkung der Enzyme". In: *Berichte der deutschen chemischen Gesellschaft* 27.3 (Oct. 1894), pp. 2985–2993. ISSN: 0365-9496. DOI: 10.1002/cber.18940270364. URL: https://onlinelibrary.wiley.com/doi/10.1002/cber.18940270364 (visited on 04/15/2023).

[10] Charles K. Fisher, Austin Huang, and Collin M. Stultz. "Modeling Intrinsically Disordered Proteins with Bayesian Statistics". In: *Journal of the American Chemical Society* 132.42 (Oct. 27, 2010), pp. 14919–14927. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja105832g. URL: https://pubs.acs.org/doi/10.1021/ja105832g (visited on 11/23/2023).

[11] Peter L. Freddolino et al. "Force Field Bias in Protein Folding Simulations". In: *Biophysical Journal* 96.9 (May 2009), pp. 3772–3780. ISSN: 00063495. DOI: 10.1016/j.bpj.2009.02.033. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006349509005888 (visited on 12/02/2023).

[12] Kari Gaalswyk, Mir Ishruna Muniyat, and Justin L MacCallum. "The Emerging Role of Physical Modeling in the Future of Structure Determination". In: ().

[13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. "Optimization by Simulated Annealing". In: *Science* 220.4598 (May 13, 1983), pp. 671–680. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.220.4598.671. URL: https://www.science.org/doi/10.1126/science.220.4598.671 (visited on 12/02/2023).

[14] Jaka Kragelj et al. "Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts". In: *ChemPhysChem* 14.13 (Sept. 16, 2013), pp. 3034–3045. ISSN: 1439-4235, 1439-7641. DOI: 10.1002/cphc.201300387. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.201300387 (visited on 12/02/2023).

[15] Kresten Lindorff-Larsen et al. "Simultaneous Determination of Protein Structure and Dynamics". In: 433 (2005).

[16] Dominique Marion. "An Introduction to Biological NMR Spectroscopy". In: *Molecular & Cellular Proteomics* 12.11 (Nov. 2013), pp. 3006–3025. ISSN: 15359476. DOI: 10.1074/mcp.O113.030239. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535947620345850 (visited on 12/02/2023).

[17] Valéry Ozenne et al. "*Flexible-Meccano:* A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and Their Associated Experimental Observables". In: *Bioinformatics* 28.11 (June 1, 2012), pp. 1463–1470. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/bts172. URL: https://academic.oup.com/bioinformatics/article/28/11/1463/266803 (visited on 11/29/2023).

[18] Valéry Ozenne et al. "Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution". In: *Journal of the American Chemical Society* 134.36 (Sept. 12, 2012), pp. 15138–15148. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja306905s. URL: https://pubs.acs.org/doi/10.1021/ja306905s (visited on 12/02/2023).

[19] Dina Schneidman-Duhovny, Riccardo Pellarin, and Andrej Sali. "Uncertainty in Integrative Structural Modeling". In: *Current Opinion in Structural Biology* 28 (Oct. 2014), pp. 96–104. ISSN: 0959440X. DOI: 10.1016/j.sbi.2014.08.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X14000943 (visited on 12/02/2023).

[20] I. Solomon. "Relaxation Processes in a System of Two Spins". In: *Physical Review* 99.2 (July 15, 1955), pp. 559–565. ISSN: 0031-899X. DOI: 10.1103/PhysRev.99.559. URL: https://link.aps.org/doi/10.1103/PhysRev.99.559 (visited on 11/24/2023).

[21] F. Emil Thomasen and Kresten Lindorff-Larsen. "Conformational Ensembles of Intrinsically Disordered Proteins and Flexible Multidomain Proteins". In: *Biochemical Society Transactions* 50.1 (Feb. 28, 2022), pp. 541–554. ISSN: 0300-5127, 1470-8752. DOI: 10.1042/BST20210499. URL: https://portlandpress.com/biochemsoctrans/article/50/1/541/230748/Conformational-ensembles-of-intrinsically (visited on 11/28/2023).

[22] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. "Intrinsically Disordered Proteins in Human Diseases: Introducing the D ^{2} Concept". In: *Annual Review of Biophysics* 37.1 (June 1, 2008), pp. 215–246. ISSN: 1936-122X, 1936-1238. DOI: 10.1146/annurev.biophys.37.032807.125924. URL: https://www.annualreviews.org/doi/10.1146/annurev.biophys.37.032807.125924 (visited on 11/28/2023).

[23] Peter E. Wright and H. Jane Dyson. "Intrinsically Disordered Proteins in Cellular Signalling and Regulation". In: *Nature Reviews Molecular Cell Biology* 16.1 (Jan. 2015), pp. 18–29. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm3920. URL: https://www.nature.com/articles/nrm3920 (visited on 11/28/2023).

[24] Junjie Zhang, Peter Minary, and Michael Levitt. "Multiscale Natural Moves Refine Macromolecules Using Single-Particle Electron Microscopy Projection Images". In: *Proceedings of the National Academy of Sciences* 109.25 (June 19, 2012), pp. 9845–9850. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1205945109. URL: https://pnas.org/doi/full/10.1073/pnas.1205945109 (visited on 12/02/2023).