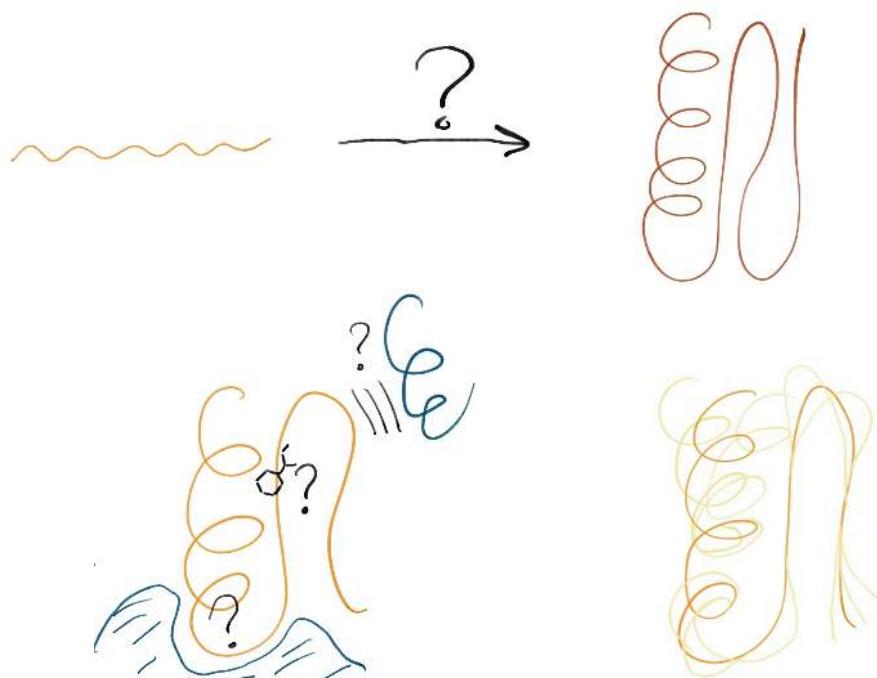


# Introduction to Protein Structural Bioinformatics

K. Anton Feenstra   Sanne Abeln

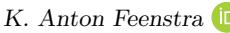


July 6, 2023



---

References . . . . .	14
<b>I Structure Comparison</b>	<b>15</b>
<b>1 Introduction to Protein Structure</b>	<b>17</b>
<i>Annika Jacobsen</i> <i>Erik van Dijk</i> <i>Halima Mouhib</i>	
<i>Bas Stringer</i> <i>Olga Ivanova</i> <i>Jose Gavaldá-García</i>	
<i>Laura Hoekstra*</i> <i>K. Anton Feenstra*</i> <i>Sanne Abeln*</i>	
1 Protein structure basics . . . . .	19
1.1 Primary structure . . . . .	19
2 Secondary structure . . . . .	24
2.1 Backbone hydrogen bonding . . . . .	24
2.2 $\alpha$ -helices . . . . .	26
2.3 $\beta$ -strands . . . . .	27
2.4 Loops . . . . .	29
2.5 Phi and psi angles . . . . .	31
2.6 Secondary structure assignment . . . . .	33
3 Tertiary structure . . . . .	33
3.1 Hydrophobic core . . . . .	35
3.2 Protein domains . . . . .	36
4 Quaternary structure . . . . .	36
5 Key points . . . . .	36
6 Further Reading . . . . .	36
References . . . . .	37
<b>2 Structure determination</b>	<b>38</b>
<i>Halima Mouhib</i> <i>Bas Stringer</i> <i>Hugo van Ingen</i>	
<i>Jose Gavaldá-García</i> <i>Katharina Waury</i> <i>Sanne Abeln</i>	
<i>K. Anton Feenstra</i>	
1 Introduction . . . . .	39
2 X-ray crystallography . . . . .	41
2.1 Crystallization . . . . .	41
2.2 Diffraction . . . . .	43
3 Nuclear magnetic resonance . . . . .	50
4 Cryo electron microscopy (cryo-EM) . . . . .	56
5 Other structure determination methods . . . . .	60
6 Dynamics and flexibility . . . . .	61
7 Key points . . . . .	63
8 Recommended further reading . . . . .	63
References . . . . .	64
<b>3 Structure Alignment</b>	<b>66</b>

Olga Ivanova  Jose Gavaldá-García  Dea Gogishvili 	
Isabel Houtkamp  Robbin Bouwmeester 	
K. Anton Feenstra*  Sanne Abeln* 	
<b>1 Comparing protein structures . . . . .</b>	<b>67</b>
1.1 Structure is more conserved than sequence . . . . .	67
<b>2 Structural superposition . . . . .</b>	<b>69</b>
2.1 PDB coordinates as a structure representation . . . . .	69
2.2 A score for comparing protein structures – RMSD . . . . .	69
2.3 Structural superposition and RMSD . . . . .	71
<b>3 Structural Alignment . . . . .</b>	<b>73</b>
3.1 The three key components of structural alignment . . . . .	75
3.2 Structure representation and contact maps . . . . .	75
3.3 Heuristic optimization algorithms . . . . .	76
3.4 Statistical scoring of structural alignments . . . . .	77
<b>4 Applications of structure comparison . . . . .</b>	<b>77</b>
<b>5 Key points . . . . .</b>	<b>78</b>
<b>6 Further reading . . . . .</b>	<b>78</b>
References . . . . .	79
<b>4 Data Resources for Structural Bioinformatics</b>	<b>80</b>
Jose Gavaldá-García  Bas Stringer  Olga Ivanova 	
Sanne Abeln*  K. Anton Feenstra*  and Halima Mouhib* 	
<b>1 Experimental protein structures . . . . .</b>	<b>81</b>
1.1 The Protein DataBank . . . . .	81
<b>2 Structure analysis and annotation . . . . .</b>	<b>86</b>
2.1 Structure validation . . . . .	86
2.2 Structural classification . . . . .	87
2.3 Protein sequences . . . . .	92
<b>3 Functional data resources . . . . .</b>	<b>94</b>
<b>4 Key points . . . . .</b>	<b>95</b>
<b>5 Further reading . . . . .</b>	<b>95</b>
References . . . . .	96
<b>5 Protein Function &amp; Interactions</b>	<b>98</b>
Annika Jacobsen  Qingzhen Hou  Bas Stringer  Sanne Abeln 	
K. Anton Feenstra 	
<b>1 Proteins are the machinery of the cell . . . . .</b>	<b>99</b>
1.1 Protein function . . . . .	100
1.2 Functional site . . . . .	102
<b>2 Protein-protein interactions &amp; complexes . . . . .</b>	<b>102</b>
2.1 Inferring protein interaction from experimental data . . . . .	103
2.2 Contact and desolvation . . . . .	106

3	Functional classes . . . . .	112
3.1	Protein Small Ligand Binding . . . . .	113
3.2	Protein-DNA interactions . . . . .	116
3.3	Transmembrane Proteins . . . . .	117
3.4	Functional motions . . . . .	119
3.5	Intrinsically disordered Proteins . . . . .	119
4	Protein function in the context of the living cell . . . . .	121
5	Key points . . . . .	123
6	Further Reading . . . . .	123
	References . . . . .	123
<b>II</b>	<b>Structure Prediction</b>	<b>127</b>
<b>6</b>	<b>Introduction to structure prediction</b>	<b>129</b>
Sanne Abeln  Jaap Heringa  K. Anton Feenstra 		
1	What is the protein structure prediction problem? . . . . .	131
1.1	Predicting the structure for a protein sequence . . . . .	131
1.2	Structure is more conserved than sequence . . . . .	132
1.3	Terminology in structure prediction . . . . .	132
1.4	Different classes of structure prediction methods . . . . .	132
1.5	Domains . . . . .	136
2	Assessing the quality of structure prediction methods . . . . .	136
2.1	Critical Assessment of protein Structure Prediction . . . . .	137
2.2	Comparing structures – RMSD and GDT_TS . . . . .	138
2.3	How difficult is it to predict? . . . . .	140
2.4	For which gene sequences can we predict a three-dimensional structure? . . . . .	141
2.5	How accurate do we need to be? . . . . .	142
3	Is there such a concept as a single native fold? . . . . .	142
3.1	Intrinsically disordered proteins . . . . .	143
3.2	Allostery and functional structural ensembles . . . . .	143
3.3	Amyloid fibrils . . . . .	143
4	Key Points . . . . .	144
5	Further Reading . . . . .	144
	References . . . . .	144
<b>7</b>	<b>Practical Guide to Model Generation</b>	<b>147</b>
Sanne Abeln  K. Anton Feenstra 		
1	Template based protein structure modelling . . . . .	149
1.1	Homology based Template Finding . . . . .	149
1.2	Fold recognition . . . . .	151
1.3	Generating the target-template alignment . . . . .	152

1.4 Generating a model . . . . .	153
1.5 Loop or missing substructure modelling . . . . .	153
2 Template-free protein structure modelling . . . . .	154
2.1 What if no suitable template exists? . . . . .	154
2.2 Generating models from structural fragments . . . . .	154
2.3 Fragment Assembly into decoys . . . . .	155
2.4 Constraints from co-evolution based contact prediction or experiments . . . . .	156
3 Selecting and refining models from structure prediction . . . . .	157
3.1 Model refinement . . . . .	157
3.2 Model quality assessment strategies . . . . .	158
3.3 Secondary Structure Prediction . . . . .	158
4 Key points . . . . .	159
5 Further Reading . . . . .	159
References . . . . .	160
<b>9 Structural Property Prediction</b>	<b>163</b>
Maurits Dijkstra  Katharina Waury  Dea Gogishvili 	
Punto Bawono  Juami H. M. van Gils  Jose Gavaldá-García 	
Mascha Okounev  Robbin Bouwmeester  Bas Stringer 	
Jaap Heringa  Sanne Abeln  K. Anton Feenstra 	
1 Introduction . . . . .	165
2 Structural property prediction as a machine learning problem . . . . .	166
2.1 Training and benchmarking structural property predictions	170
2.2 Sequence signatures . . . . .	172
2.3 Evolutionary information . . . . .	172
3 Secondary structural element (SSE) prediction . . . . .	173
3.1 Hydrophobicity patterns . . . . .	173
3.2 Intrinsic preference of amino acids for certain secondary structure types . . . . .	173
3.3 Locality of secondary structure . . . . .	174
3.4 Deriving Amino Acid Propensities . . . . .	175
3.5 Secondary structure prediction methods . . . . .	177
3.6 Special cases . . . . .	179
4 Other structural properties . . . . .	180
4.1 Surface accessibility prediction . . . . .	180
4.2 Disorder and flexibility prediction . . . . .	180
4.3 Transmembrane prediction . . . . .	181
4.4 Aggregation propensity prediction . . . . .	181
5 Practical advice . . . . .	182
6 Key Points . . . . .	183
7 Further Reading . . . . .	184
References . . . . .	184

---

**11 Function Prediction** 188

*Bas Stringer*  *Annika Jacobsen*  *Qingzhen Hou*   
*Hans de Ferrante*  *Olga Ivanova*  *Katharina Waury*   
*Jose Gavaldá-García*  *Sanne Abeln\**  *K. Anton Feenstra\** 

1	Introduction . . . . .	189
2	Different types of function prediction tasks . . . . .	189
2.1	Different function prediction methods . . . . .	190
3	Residue level function predictions . . . . .	191
3.1	Mutation impact analysis . . . . .	191
3.2	Active site prediction . . . . .	192
3.3	Structural annotation predictions . . . . .	193
4	Protein level function predictions . . . . .	193
4.1	Inferring function through homology . . . . .	194
4.2	Critical Assessment of Function Annotation . . . . .	194
5	Protein-protein interaction predictions . . . . .	195
5.1	Prediction of PPI from structure – docking method . . . . .	195
5.2	Prediction of PPI from sequence . . . . .	196
5.3	Protein interface prediction . . . . .	196
5.4	CAPRI . . . . .	197
6	Key points . . . . .	199
7	Further reading . . . . .	199
8	Author contributions . . . . .	199
	References . . . . .	199

**III Dynamics and Simulation** 203
**12 Introduction to Protein Folding** 205

*Juami H. M. van Gils\**  *Erik van Dijk*  *Ali May*   
*Halima Mouhib*  *Jochem Bijlard*  *Annika Jacobsen*   
*Isabel Houtkamp*  *K. Anton Feenstra\**  *Sanne Abeln\** 

1	Protein folding and restructuring . . . . .	207
1.1	Flexibility of protein chains & structural ensembles . . . . .	207
1.2	Defining the folded and unfolded states . . . . .	208
2	Folding and refolding . . . . .	210
2.1	Stability and probability . . . . .	210
2.2	Changing conditions . . . . .	210
3	Factors that (de)stabilize the native fold . . . . .	211
3.1	Hydrophobic effect . . . . .	212
3.2	Hydrogen bonds, salt-bridges and packing . . . . .	212
3.3	Backbone entropy . . . . .	212
4	Folding pathways . . . . .	213
4.1	Free Energy Landscapes . . . . .	214

5	Folding in the cell . . . . .	215
5.1	Chaperones . . . . .	215
5.2	Folded proteins are only marginally stable . . . . .	215
6	Alternative stable states of proteins . . . . .	216
6.1	Molten globules . . . . .	216
6.2	Natively disordered proteins . . . . .	217
6.3	Misfolding . . . . .	217
6.4	Aggregation and amyloid formation . . . . .	217
7	Key concepts . . . . .	218
8	Further reading . . . . .	219
	References . . . . .	219
<b>13</b>	<b>Thermodynamics of Protein Folding</b>	<b>221</b>
<i>Juami H. M. van Gils*</i>  <i>Halima Mouhib</i>  <i>Erik van Dijk</i> 		
<i>Maurits Dijkstra</i>  <i>Isabel Houtkamp</i>  <i>Arthur Goetzee</i> 		
<i>Sanne Abeln*</i>  <i>K. Anton Feenstra*</i> 		
1	Equilibrium and Dynamics . . . . .	223
2	Thermodynamic laws . . . . .	224
3	Entropy . . . . .	226
4	Enthalpy . . . . .	228
5	Free energy . . . . .	229
5.1	Temperature Dependence of Free Energy Landscapes . . .	233
6	From Microstates to Macrostates . . . . .	235
6.1	Order Parameters . . . . .	236
6.2	Ensemble Average . . . . .	236
7	Ensembles . . . . .	237
8	Conclusion . . . . .	239
	References . . . . .	240
<b>14</b>	<b>Molecular Dynamics</b>	<b>241</b>
<i>Halima Mouhib*</i>  <i>Juami H. M. van Gils</i>  <i>Jose Gavaldá-García</i> 		
<i>Qingzhen Hou</i>  <i>Ali May</i>  <i>Arriën Symon Rauh</i> 		
<i>Jocelyne Vreede</i>  <i>Sanne Abeln*</i>  <i>K. Anton Feenstra*</i> 		
1	Introduction to molecular dynamics . . . . .	243
1.1	Simulating a protein by classical physics . . . . .	243
2	Relevant time and length scales . . . . .	245
3	Forces & interactions . . . . .	248
3.1	Force fields . . . . .	248
3.2	Interactions . . . . .	249
3.3	Parameters . . . . .	251
4	Dynamics . . . . .	254
4.1	Integrating equations of motion . . . . .	256
4.2	Convergence of state properties . . . . .	263

---

4.3 Temperature dependence . . . . .	269
4.4 Homology model optimization . . . . .	270
5 Outlook and summary . . . . .	273
6 Key concepts . . . . .	273
7 Further reading . . . . .	274
References . . . . .	274
<b>15 Monte Carlo for Protein Structures</b>	<b>278</b>
<i>Juami H. M. van Gils*</i> <i>Maurits Dijkstra</i> <i>Halima Mouhib</i>	
<i>Arriën Symon Rauh</i> <i>Jocelyne Vreede</i>	
<i>K. Anton Feenstra*</i> <i>Sanne Abeln*</i>	
1 Introduction . . . . .	279
2 Proteins in equilibrium . . . . .	279
3 The Purpose of Simulations . . . . .	280
4 Comparison to experiments . . . . .	281
5 Monte Carlo Alogrithm . . . . .	282
5.1 Potential energies . . . . .	282
5.2 Sampling the partition function . . . . .	282
5.3 The Metropolis Monte Carlo algorithm . . . . .	283
6 Applications of Monte Carlo for proteins . . . . .	287
6.1 A simple protein lattice model . . . . .	287
6.2 Other applications in bioinformatics . . . . .	289
7 Enhanced sampling techniques . . . . .	291
7.1 Umbrella Sampling in MC . . . . .	292
8 Monte Carlo vs. Molecular Dynamics . . . . .	295
9 Key points . . . . .	296
10 Further reading . . . . .	297
References . . . . .	297

<b>References</b>	<b>301</b>
-------------------	------------

# Preface

Why did we write this book? Firstly, because all the authors involved are very enthusiastic about protein structures and scientific computation, and secondly because we have been setting up a successful MSc level Bioinformatics in Structural Bioinformatics course that is in immediate need for a good text book.

While many good textbooks are available on Protein Structure, Molecular Simulations, Thermodynamics and Bioinformatics methods in general, there is no good introductory level book for the field of Structural Bioinformatics. This book aims to give an introduction into Structural Bioinformatics, which is where the previous topics meet to explore three dimensional protein structures through computational analysis.

This book will provide you an overview of existing computational techniques, to validate, simulate, predict and analyse protein structures. More importantly, it will aim to provide practical knowledge about how and when to use such techniques. We will consider proteins from three major vantage points, as illustrated in Figure 1:

- I) Protein structure & structure comparison;
- II) Protein structure prediction; and
- III) Protein simulation & dynamics.

*Part I Structure & structure comparison* deals with comparing one protein structure to another. This can either be in a general sense: a protein structure is compared to a large set of reference structures to validate the experimental reliability of the structure. Or the comparison may be to one or more specific protein structures, with the question how similar the protein folds are.

*Part II Structure prediction* deals with the question, how to predict the structure given a protein sequence. We start with a graceful introduction to protein structure basics (Abeln *et al.*, 2017a). We will see that there is a wide range of methods available, and that the reliability of such method varies strongly. It is important to understand how and when structure prediction will give you trustworthy outcomes (Abeln *et al.*, 2017b). Then we reach

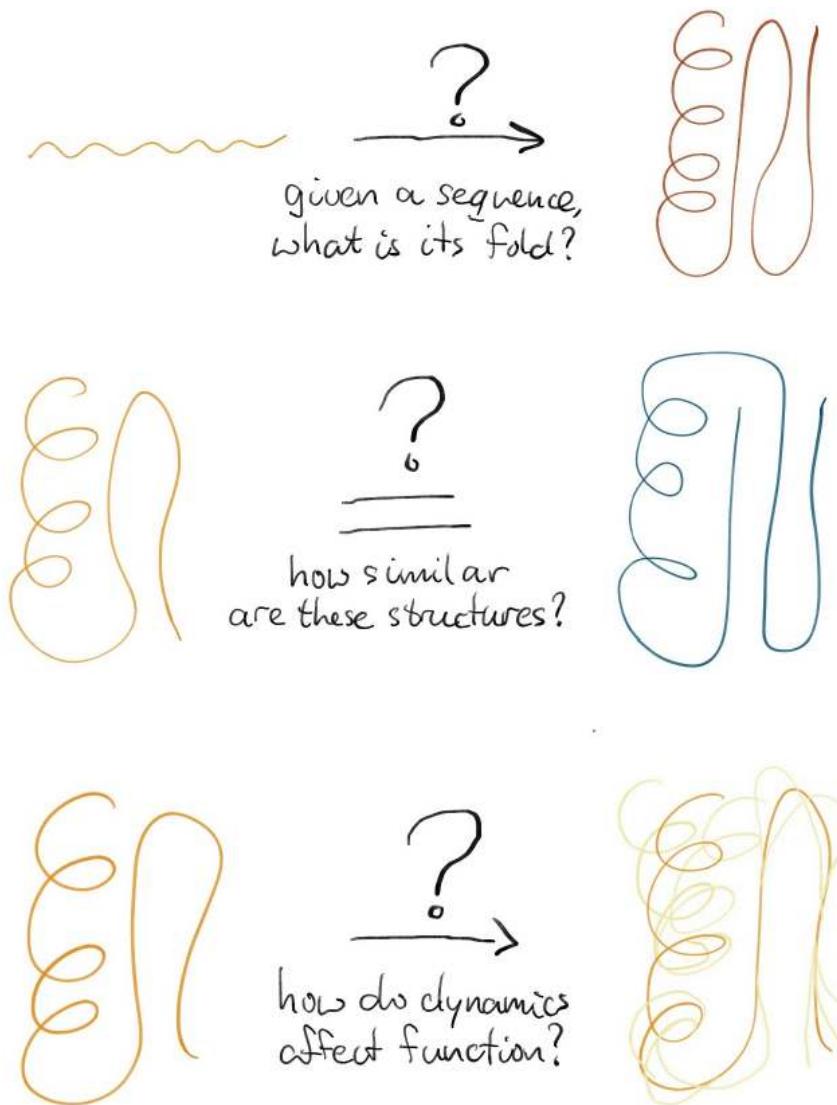


Figure 1: Within the field of Structural Bioinformatics three dimensional protein structures are investigated through computational analysis. Important problems that may be addressed computationally are shown in the form of cartoons. Firstly, how does the genomic sequence of a gene translate into the folded, functional protein structure? Secondly, when considering two proteins, how similar are their structures? And, last but not least, since we know proteins are not static entities, how do flexibility and dynamics play a role in the function of the protein?

perhaps the most salient question: from an available structure, what can we say about the possible function this protein could perform.

*Part III Simulation & dynamics* considers the dynamics or the ensemble of possible structures a protein may take up. Generally speaking however, it is not yet possible to allow proteins to fold within a simulation, nevertheless simulations are vital to understand functional mechanisms of structural ensembles of proteins, specifically in context of small metabolites, the cell membrane or other proteins.

Herein, we introduce relevant concepts such as structure basics, structure prediction, homology modeling, statistical thermodynamics, simulation. From here, the reader may be able to acquire further information through self study and further references. In particular, we will treat the protein not as a rigid structural entity, but as dynamic ensembles of structures which make the protein able to perform its function or functions.

Hence this book aims to provide a framework for all important concepts within the field of Structural Bioinformatics. However, we do not aim to give extensive reviews of all the methods available in each of these subjects; for these we will refer, where appropriate, to other books.

The primary audience of this book are master students in the area of bioinformatics, or a related discipline like biotechnology. Basic background knowledge is assumed on protein structure, bioinformatics, sequence analysis, dynamic programming algorithms, calculus and basic chemistry. A few good books we would like to mention here, upfront, as they may be important reference material for readers of this book, depending on their background:

- Protein Structure – Branden and Tooze (1998)
- Understanding Bioinformatics – Zvelebil and Baum (2008)
- Essential Bioinformatics – Xiong (2006) (very basic)
- An introduction to Thermal Physics – Schroeder (1999)
- Introduction to Python – Lutz (2013)

And some more advanced books, that go further on particular topics than this book:

- Structural Bioinformatics – Gu and Bourne (2009)
- Sequence Analysis – Durbin *et al.* (1998)
- Molecular Simulation – Frenkel and Smit (2002)
- Molecular Modelling – Leach (2001)

We hope to provide the reader with knowledge about the type of problems that are scientifically feasible to solve. For example, predicting a protein structure through homology modeling will generally give high quality and reliable results. In contrast, we also would like to point out which problems are still unlikely to yield good results, such as predicting protein-protein interactions, and which are still out of scope altogether for existing methods, such as correct prediction of protein folding from sequence information alone. Most importantly, we will introduce the fundamental concepts on

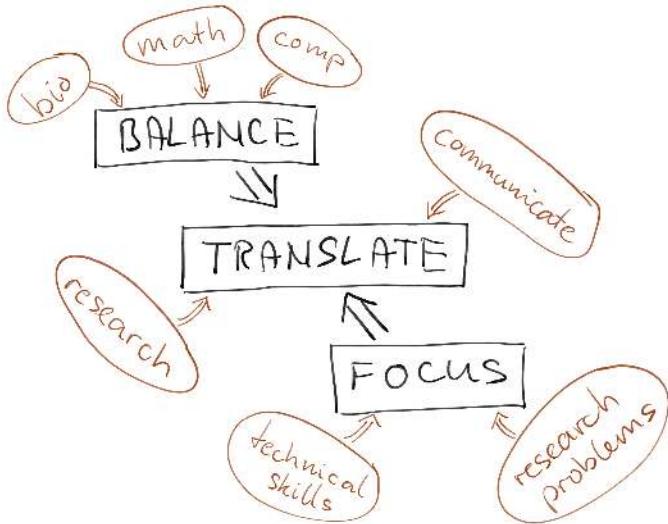


Figure 2: Conceptual organisation of our bioinformatics education programme along three key elements: Translate, Balance, and Focus.

which the method are based, and what assumptions there are for using such methods. Therefore it should become apparent which inherent limitations the techniques have and what the techniques can successfully be used for.

The setup of this books follows the same guiding principle that we apply throughout our masters' programme. We introduce a **focus** on open and challenging research problems to solve, and the technical skills to solve them; a **balance** between basic skills, such as biology, mathematics (modelling) and computational tools (programming); then, building on this focus and balance, **translate** between different adjoining disciplines, and between tools and methods on the one hand, and application and problems on the other (Abeln *et al.*, 2013; Feenstra *et al.*, 2018). This is summarized in Figure 2.

Here, we would like to thank all chapter authors: Annika Jacobsen , Ali May , Arriën Symon Rauh , Bas Stringer , Dea Gogishvili , Erik van Dijk , Hans de Ferrante , Hugo van Ingen , Halima Mouhib , Jochem Bijlard , Jose Gavaldá-García , Jaap Heringa , Juami H. M. van Gils , Jocelyne Vreede , Katharina Waury , Maurits Dijkstra , Mascha Okounev , Olga Ivanova , Punto Bawono , Qingzhen Hou , and Robbin Bouwmeester .

Lastly, we would like to thank all our students who have followed the MSc course Structural Bioinformatics at the VU University in Amsterdam for their enthusiasm in pointing out mistakes in our lectures and asking important additional questions; without this vital input it would have been absolutely impossible to write this book. In particular, we thank Nicola Bonzanni, Ashley Gallagher, Tim

Kwakman, Ting Liu, Arthur Goetzee, and Reza Haydarlou for insightful discussions and critical proofreading of early versions.

## References

- Abeln, S., Molenaar, D., Feenstra, K.A., Hoefsloot, H.C.J. et al (2013). Bioinformatics and Systems Biology: bridging the gap between heterogeneous student backgrounds. *Briefings in Bioinformatics*, **14**(5), 589–598.
- Abeln, S., Heringa, J. and Feenstra, K.A. (2017a). Introduction to Protein Structure Prediction. *arXiv*, **1712.00407**.
- Abeln, S., Heringa, J. and Feenstra, K.A. (2017b). Strategies for protein structure model generation. *arXiv*, **1712.00425**.
- Branden, C. and Tooze, J. (1998). *Introduction to protein structure*. garland publishing, New York.
- Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- Feenstra, K.A., Abeln, S., Westerhuis, J.A., Brancos dos Santos, F. et al (2018). Training for translation between disciplines : a philosophy for life and data sciences curricula. *Bioinformatics*, **34**(13), 1–9.
- Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Pr, San Diego, second edition.
- Gu, J. and Bourne, P.E. (2009). *Structural bioinformatics*. John Wiley & Sons,, Hoboken, 2nd ed. nv edition.
- Leach, A. (2001). *Molecular Modelling: Principles and Applications*. Pearson.
- Lutz, M. (2013). *Learning Python*. O'Reilly.
- Schroeder, D.V. (1999). *An Introduction to Thermal Physics*. Addison-Wesley Publishing Company, San Francisco, CA.
- Xiong, J. (2006). *Essential Bioinformatics*. Cambridge University Press.
- Zvelebil, M. and Baum, J. (2008). *Understanding Bioinformatics*. Garland Science, Taylor & Francis Group, New York – London.

**Part I**

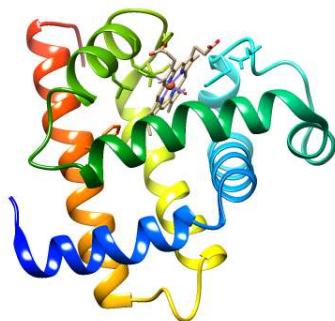
**Structure Comparison**



# Chapter 1

## Introduction to Protein Structure

Annika Jacobsen  Erik van Dijk  Halima Mouhib   
Bas Stringer  Olga Ivanova  Jose Gavaldá-García   
Laura Hoekstra\*  K. Anton Feenstra\*  Sanne Abeln\* 



\* editorial responsibility



Within the living cell, protein molecules perform specific functions, typically by interacting with other proteins, DNA, RNA or small molecules. They take on a specific three dimensional structure, or in some cases, an ensemble of three dimensional structures. It is this three dimensional structure that allows the protein to function within the cell. This structure is with high specificity encoded by its amino acid sequence; the precise amino acid sequence of a protein is in turn encoded by the genes of an organism. Hence, the understanding of a protein's function is tightly coupled to its three dimensional structure.

The current state of scientific understanding allows us to comprehend how the gene sequence encoded by the DNA is transcribed into RNA, and in its turn translated into amino acid sequence. However, experiments to determine protein structures and protein structural ensembles are difficult and laborious; we will come back to that in detail in Chapter 2. Recently, deep learning models like AlphaFold2, trained on existing protein structure data, have achieved success in predicting protein structures from sequences. However, simulating the transition from a protein sequence to its folded structure computationally is still challenging for moderately sized proteins. As a result, structural bioinformatics faces unresolved problems or, alternatively, presents exciting scientific challenges.

Before going into protein structure analysis and prediction, and protein folding and dynamics, we will first give a brief introduction into the basics of protein structures. This is deliberately kept short and shallow. The excellent book "Introduction to Protein Structure" by Branden and Tooze (1998) provides a much more in-depth introduction into this exciting field.

## 1 Protein structure basics

A protein structure may be described at four levels as depicted in Figure 1.1: **The primary structure** is simply the sequence of amino acids that make up the protein polypeptide chain.

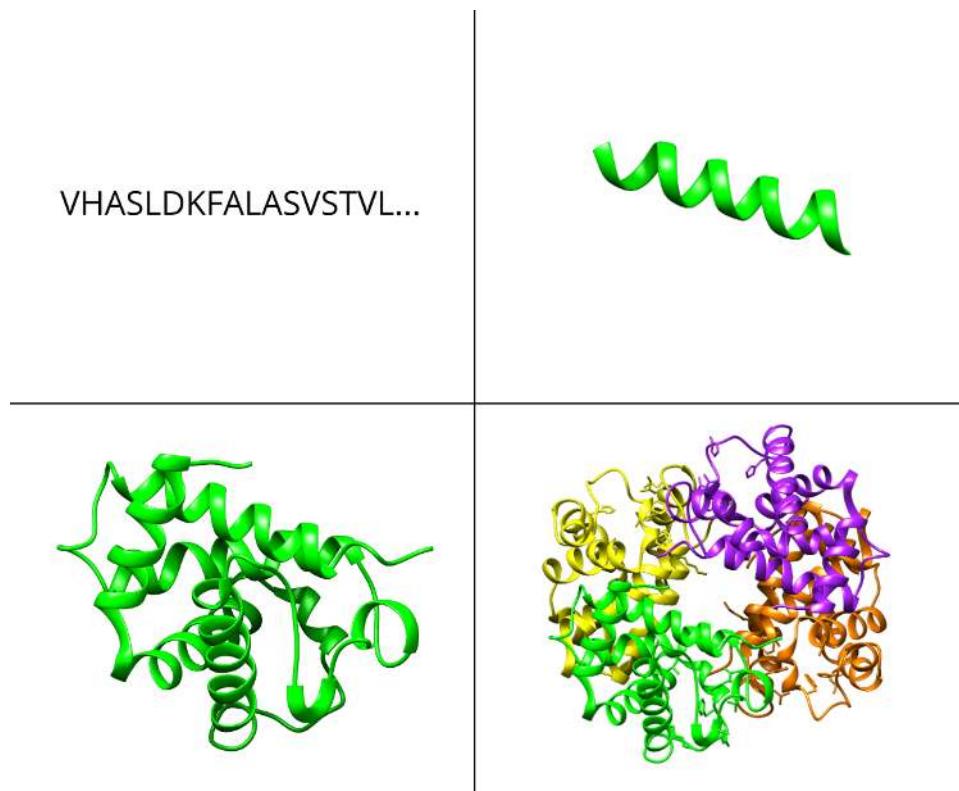
**Secondary structure** describes the organisation of this chain into regular  $\alpha$ -helices and  $\beta$ -strands and anything else, called 'coil' or loop.

**Tertiary structure** is the three dimensional arrangement – or topology – of the protein chain; it defines its overall shape.

**Quaternary structure** is the (three dimensional) organisation of the protein chain in context of the proteins and molecules it interacts with; i.e. the configurational ensemble multiple molecules adopt when binding to each other, forming macro-molecular complexes.

### 1.1 Primary structure

There are 20 naturally occurring amino acids that constitute the building blocks of proteins. Amino acids are linked together by a **peptide-bond**



**Figure 1.1: Levels of protein structure.** Top-left: Primary structure, given as polypeptide sequence in the one-letter code of amino acids. Top-right: Secondary structure, example of an alpha helix. Bottom-left: Tertiary structure, structure of one of the monomers of hemoglobin. Bottom-right: Full structure of Human hemoglobin, 4 chains make the whole structure (PDB:1BIJ). Ribbon representation obtained with UCSF-chimera (Pettersen et al., 2004).

between the carbonyl Carbon ( $\text{C}=\text{O}$ ) of the preceding residue and the amide Nitrogen ( $\text{NH}$ ) of the next residue in its primary sequence. This is why proteins are also referred to as “polypeptides”. Note that for each amino acid type, this part of the chemical structure is identical; it is also referred to as the *backbone*. The *sidechains* branch out from the central Carbon atom ( $\text{C}\alpha$ ) in the backbone. Unlike the backbone, sidechains are chemically different between the different amino acid types; see Panel “Amino acids, residues, and the peptide bond” for more detail. We can view the primary protein structure as a chain with 20 different colours of beads that all have different properties.

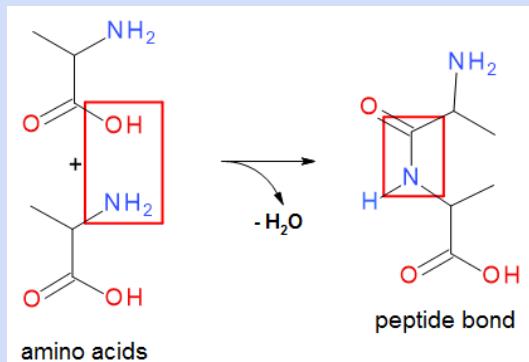
The amino acid type in the polypeptide chain is encoded by codons (sequences of three nucleotides) in the DNA sequence of a gene. After transcription, the translation from RNA codons into amino acids occurs at the ribosome. The exact sequence of the codon determines the amino

acid, or may indicate the start or end of a peptide chain. This codon table is universal across all species, although several microorganisms are known that use a (slightly) different table. The translation mechanism, including the codon table, the tRNA and the aminoacyl-transferases, is beyond the scope of this book.

Roughly speaking, there are three important classes of amino acids: *i*) hydrophobic, *ii*) charged, and *iii*) polar. These classes are based on their interaction properties with respect to water. Hydrophobic residues do not interact with water, whereas polar and charged residues do make contact with water favourably. Later in this chapter, we will see the importance of the difference between hydrophobic and polar amino acids for protein folding. The Panel “Amino acids, residues, and the peptide bond” gives more background detail on the chemical characteristics of those amino acids.

### Amino acids, residues, and the peptide bond

There are 20 naturally occurring amino acids that constitute the building blocks of proteins, shown in Panel “The 20 natural amino acid residues” below. (Chemically speaking many, many more types of amino acids are possible). To build up a protein, amino acids react under the loss of water to form an extremely stable peptide bond. The figure shows the formation of the peptide bond between two alanine amino acids:



Within proteins, amino acids differ in the sidechain part, the backbone of the protein, i.e., (NH–C<sub>α</sub>–C=O) is repetitive. Note that each amino acid can be referred to using a three or one letter code: here Ala or A for alanine. In Panel “The 20 natural amino acid residues”, amino acid residues are categorized into **charged** (positive/negative), **polar** (not charged), and **hydrophobic** (or aliphatic) amino acids.

**Hydrophobic/aliphatic/apolar** amino acids consist of only Carbon (and for Cysteine, Sulphur) atoms in the sidechain.

**Aromatic residues** all have a regular six- or five-sided ring (trypt-

tophan has both) consisting of mostly carbon atoms. Tyrosine is aromatic but due to a hydroxyl group also polar. Tryptophan does contain a nitrogen atom but is considered hydrophobic.

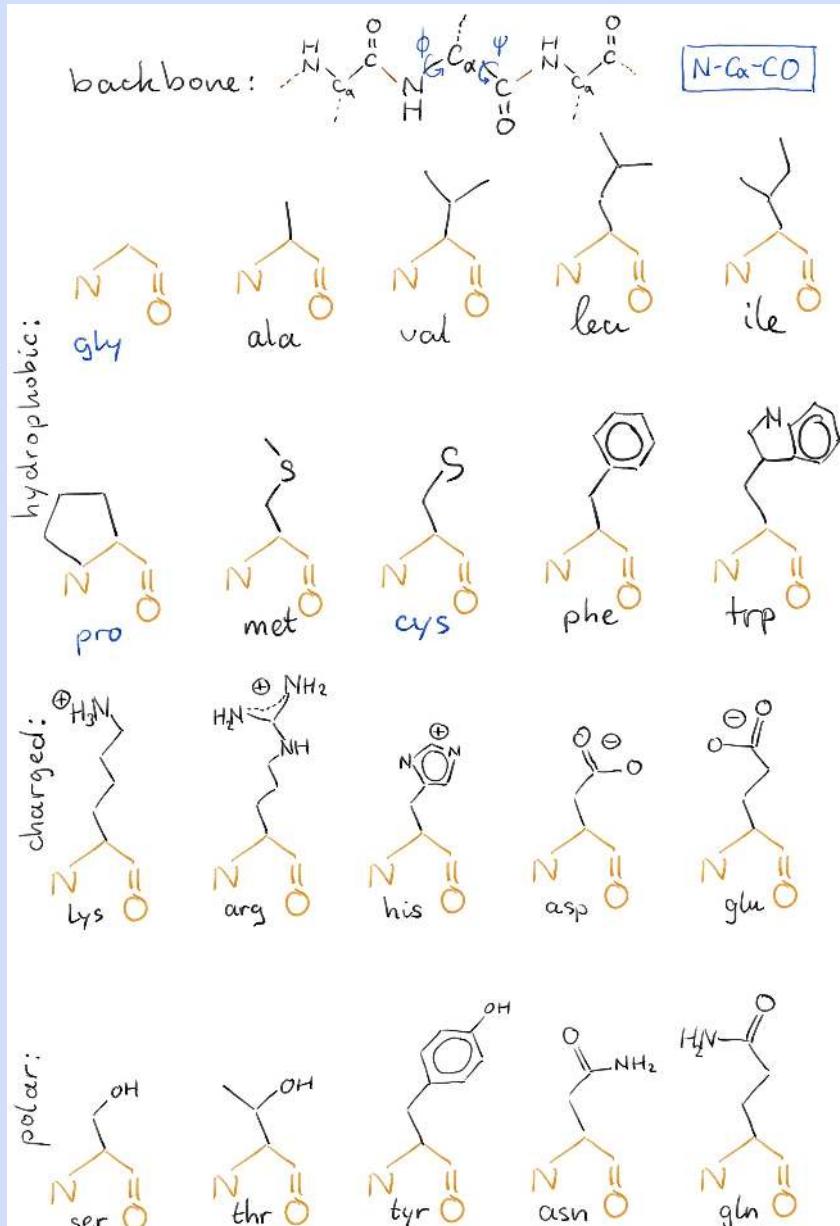
**Polar and charged** residues all have nitrogen and/or oxygen atoms in the sidechain. Charged residues are also considered polar, and both are hydrophilic.

**Small, medium, large:** A further subdivision of the hydrophobic amino acids can be made into small (glycine, alanine), medium (valine, leucine, isoleucine), large (methionine) and aromatic/ring (proline, which is also medium size, and phenylalanine and tryptophan which are also large). Of the charged amino acids, lysine and arginine are considered large, and among the polar ones, serine and threonine are small.

Finally, the backbone contains N and O and is therefore always polar. In the polar and charged residues, oxygen is always negative and nitrogen always positive.

Before moving on, it should be noted that biologists and bioinformaticians often use the terms ‘amino acid’ and ‘residue’ equivalently. However, ‘residue’ is more general and can also refer to e.g. a nucleotide in DNA or RNA. To be a bit more precise, for chemists, the amino acid is the free molecule, and it is called ‘amino acid residue’ only when part of a protein.

### The 20 natural amino acid residues



In addition to these broad categories of amino acids, there are a few special ones (labelled in blue in the figure):

**Cysteine** contains a sulphur atom. When two cysteines are close in the structure, the two sulphur atoms will form a covalent bond of similar strength to the other bonds within the amino acids. These are much stronger than hydrogen bonds (see next 1.1.2 for hydrogen bonds).

**Proline** contains a ring that loops from the C $\alpha$  back to the backbone nitrogen. This makes the backbone of the proline much less flexible than for other residues; proline often terminates helices or otherwise induces a kink, and proline is used to make a loop containing sharp turns. More about this below in the Panel “The omega torsion angle”.

**Glycine** has the smallest possible sidechain: only a single hydrogen atom (which is much smaller than a carbon). Due to this, there is less steric hindrance around the C $\alpha$  and more flexibility in its Phi/Psi angles (more detail about phi/psi angles in Section 2.5 below).

## 2 Secondary structure

All amino acids have a common part, the backbone, as discussed in the previous section. The backbone can make regular structures due to their chemical properties, see also Figure 1.3. These regular backbone structures are called secondary structure. Examples of  $\alpha$ -helices and  $\beta$ -strands are shown in Figure 1.4. Below we will introduce how backbone hydrogen bonding leads to secondary structure.

### 2.1 Backbone hydrogen bonding

Hydrogen bonds are a key part of the secondary and tertiary structure of a protein. Hydrogen bonding takes places between hydrogen atoms, with a

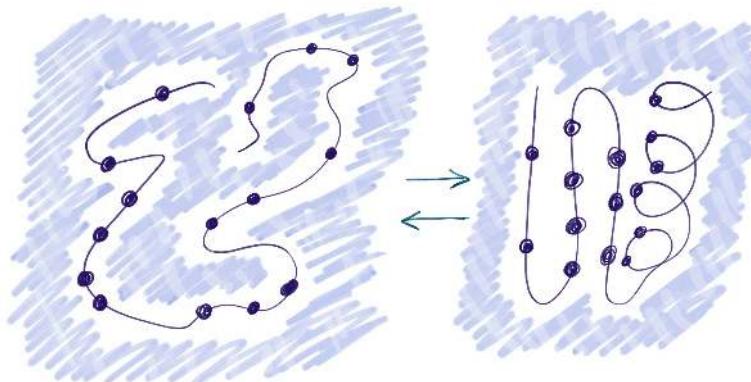


Figure 1.2: Hydrophobic collapse as the first step in a protein folding from its unfolded state (on the left) to a folded state (right). Hydrophobic residues, shown as black spheres, will tend to minimize contact with water and therefore end up in the interior of the protein. Hydrophilic (polar and charged) residues are not drawn explicitly here, they form the rest of the backbone, between the black spheres.

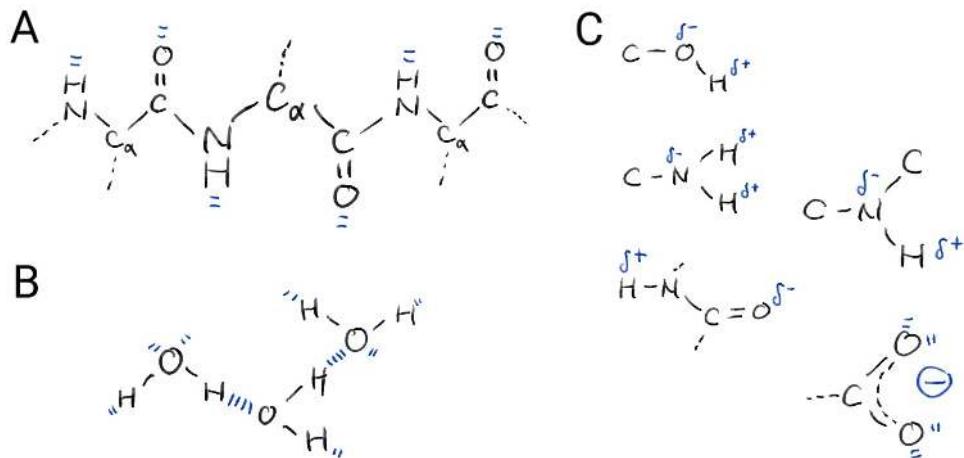


Figure 1.3: Hydrogen bonding in the backbone of the protein (A) and in water (B); hydrogen-bond forming groups are indicated with blue dashed lines. Hydrogen bonds are caused by atoms with slight negative charges ( $\delta^-$ ) being attracted to atoms with slightly positive charges ( $\delta^+$ ). In a protein, hydrogens on a nitrogen or oxygen are positive, oxygens and nitrogens themselves are negative (C).

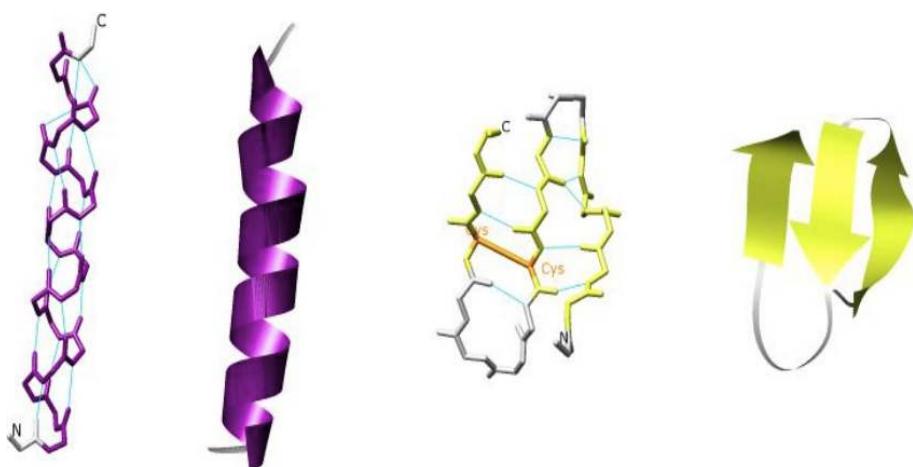


Figure 1.4: (Examples of  $\alpha$ -helical (left two) and  $\beta$ -strand (right two) structures made out of alanines (note the single atom in the sidechain) and a few cysteines (in the  $\beta$ -sheet). Both are shown in a ‘sticks’ (left) and a typical ‘cartoon’ (right) representation.

slight positive charge and nitrogen or oxygen atoms with a slight negative charge.

The attraction of opposite charges leads to strong H $\cdots$ O interactions called **hydrogen-bonds**. The N–H and C=O parts of the backbone, can also make these polar hydrogen-bond interactions (see Figure 1.3), and so can the polar and charged sidechains (O–H, N–H, S–H and C=O groups as shown in Panel “The 20 natural amino acid residues”). Energetically, hydrogen-bonds are very favourable, and most hydrogen bond donors and acceptors in the sequence of the protein will therefore also make a hydrogen bond, in any stable structure.

There are two main ways in which the amino acid chain in proteins are structured so that all backbone hydrogen bonds in the hydrophobic interior can become satisfied: *helix* or *strand*. In the helical structure, repeated local hydrogen bonding occur, as shown in Figure 1.4. This is called the  $\alpha$ -helix secondary structure type. An  $\alpha$ -helix only leaves unsatisfied hydrogen bonding capacity at the ends of a helix, these ends are thus usually found at the surface of a protein. In the strand structure, two stretches of sequence are adjacent in the structure, and hydrogen bonds occur ‘laterally’ between adjacent strands. This is the ‘ $\beta$ -sheet’ secondary structure. It leaves unsatisfied hydrogen bond capacity at the first and last strands of a sheet, called the ‘edge’ strands, which like helix ends are mostly found at the protein surface.

## 2.2 $\alpha$ -helices

Helical secondary structures are characterized by repeated, local hydrogen bonding between the backbone amide group of one residue, and the carbonyl group of a subsequent residue, as shown in Figure 1.4. The most common of these structures is the  $\alpha$ -helix, where the bond is formed between residues  $i$  and  $i + 4$ . Helices can be anywhere from 4 to 40 residues long, with an average length of  $\sim$ 10 residues, or about 3 turns.

This periodicity of  $\alpha$ -helices can typically also be observed in the sequence. Within a protein structure a helix typically has a solvent exposed and a buried side; this will lead to hydrophilic residues tending to point outside towards the solvent, and hydrophobic ones tending towards the inside of the protein. The side sticking into the core are typically tightly packed together, also referred to as helix packing. Helical structures are generally considered easier to predict from sequence due to this periodicity in the sequence, we will come back to this in Chapter 9. Please refer to the Panel “Helices” below for further details on helices.

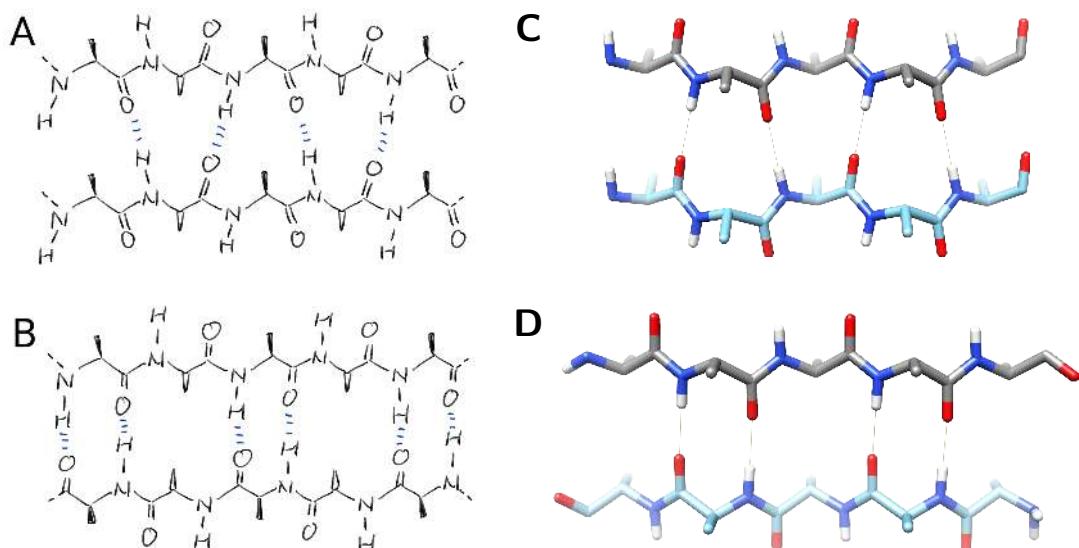


Figure 1.5: Two types of  $\beta$ -sheet, schematically: a) parallel, and b) anti-parallel; and in three-dimensions: c) parallel and d) anti-parallel (ideal geometries, generated using Chimera).

## Helices

The  $\alpha$ -helix is the most common helix secondary structure type, where the bond is formed between residues  $i$  and  $i + 4$ . So-called 3/10- and  $\pi$ -helices, where the bond connects residues  $i \rightarrow i + 3$  and  $i \rightarrow i + 5$  respectively, are less common. In the helical structure, the carbonyl group ( $C=O$ ) is oriented along the direction of the sequence, while the amide group ( $NH$ ) points in the opposite direction (Figure 1.4a). This arrangement results in a tightly packed helix with minimal internal space.

## 2.3 $\beta$ -strands

A  $\beta$ -strand is a stretch of amino acids with the backbone in an extended configuration, typically 3 to 10 amino acids long. Two or more  $\beta$ -strands together make up a  $\beta$ -sheet. Hydrogen bonding patterns in the  $\beta$ -sheet are distinctly different from those in the  $\alpha$ -helix. In the  $\beta$ -sheet, hydrogen bonds occur ‘laterally’ between adjacent strands.

When two  $\beta$ -strands have the same direction, they are referred to as parallel  $\beta$ -sheets. Conversely, if the two  $\beta$ -strands have opposite directions, they are known as antiparallel  $\beta$ -sheets. These two types of  $\beta$ -sheets can be distinguished by the geometry of the hydrogen bonds. In parallel  $\beta$ -sheets,

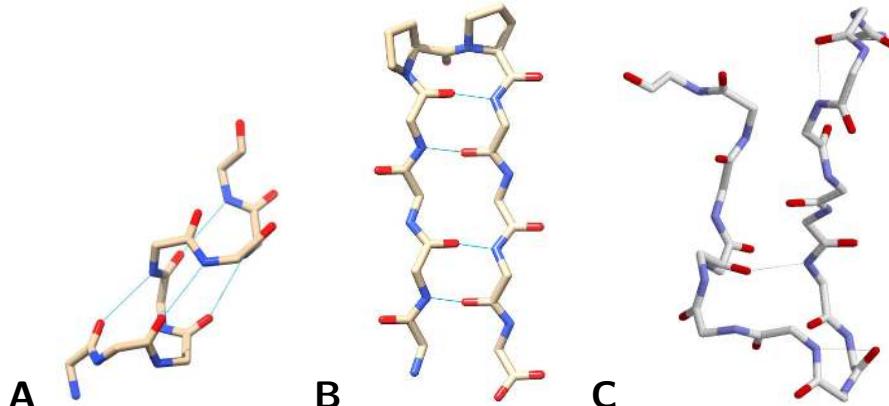


Figure 1.6: Details of  $\alpha$ -helix (A),  $\beta$ -strand (B) and coil (C). Note how regular patterns of hydrogen bonds (thin lines) stabilize  $\alpha$ -helix (A) and  $\beta$ -strand (B), but not coil (C). Also note how the hydrogen bonds in the  $\alpha$ -helix (A) point along the helix axis; the hydrogen bonds go from the hydrogen atom (which is not shown) on the nitrogen (blue) to the oxygen (red), pointing 'backwards' along the direction of the protein chain, which runs from bottom left to top right (A).

### Strands and sheets

Due to the chirality of amino acids,  $\beta$ -sheets are often twisted or pleated in a right-hand turn. Simple configurations of  $\beta$ -sheets include the commonly found hairpin and so-called psi-loop motif, whereas larger sheets can assume complex formations like  $\beta$ -barrels or  $\beta$ -propellers.

When the connecting sequence between two strands is a small loop (a  $\beta$ -loop- $\beta$  motif, often referred to as  $\beta$ -hairpin) Figure 1.6b, the sequence distance can be as low as 3. However, the sequence distance between two strands in a beta-sheet can be much larger. An extreme case occurs when a whole protein domain is in between the two strands, which might be hundreds of residues (more on protein domains later in Chapter 4 “Data Resources for Structural Bioinformatics”). Therefore, no clear relation (such as ‘i–i+4’ for  $\alpha$ -helix) occurs between hydrogen-bonded residues in a  $\beta$ -sheet. Thus  $\alpha$ -helices can be considered “local” compared to  $\beta$ -sheets, as the hydrogen-bonding in  $\alpha$ -helices is formed between nearby residues in the sequence, while in  $\beta$ -sheets the hydrogen-bonded residues may be far away in the sequence.

the hydrogen bonds are formed diagonally between the carbonyl group of one amino acid residue and the amide group of the neighboring residue on the adjacent strand, resulting in a slanted hydrogen bonding pattern. On the other hand, in antiparallel  $\beta$ -sheets, the hydrogen bonds are formed directly between the carbonyl group of one amino acid residue and the amide group of the neighboring residue on the opposite strand, creating a linear hydrogen bonding pattern Figure 1.5. Please refer to the Panel “Strands and sheets” below for further details on beta-sheets.

## 2.4 Loops

In loop regions, see also Figure 1.6c, there is no regular pattern of hydrogen bonding. Nevertheless, the hydrogen bond donors and acceptors, as present in the backbone, do need to form hydrogen bonds. In loop structures, the backbone atoms may make hydrogen bonds with the solvent, with the sidechains of polar amino acids, or even with backbone atoms – but not in a regular pattern.

The configuration of loops are much less regular, or ordered compared the helices and  $\beta$ -sheets. Generally, loops lie on the surface of a protein, and are much more solvent exposed. Often loop regions can be flexible, and can change conformation in the functional state of the protein, even when the protein is fully folded. Loop regions are therefore also much more likely to loose (deletion) or gain (insertion) small parts during evolution. In a (multiple) sequence alignment, loop regions typically contain many gaps compared to helical or sheet regions. Very long loops ( $> 20$  residues) are also called *disordered regions*. Such regions will not take up a rigid three-dimensional structure in their folded state, see also Panel “Unusual secondary structures”.

### Unusual secondary structures

In addition to ‘typical’ proteins secondary structures, there are three main classes of ‘a-typical’ cases: amyloid fibrils, coiled-coils, and disordered proteins. Examples are shown in Figure 1.7.

**Amyloid fibrils** ( $\beta$ -fibrils) are a particular case of  $\beta$ -sheets. Here the  $\beta$ -strands are also formed between the chains of different protein molecules, and such structures can become infinitely long. The resulting fibrils may form larger aggregate fibers, that may disrupt the cell functioning or even kill cells. Initially, the ability to form these fibrils was thought to be a particular property of specific proteins and associated with particular pathologies, like the prion protein in scrapie (sheep), mad cow disease (cattle) or Creutzfeldt-Jacob’s disease (human). It has now become clear

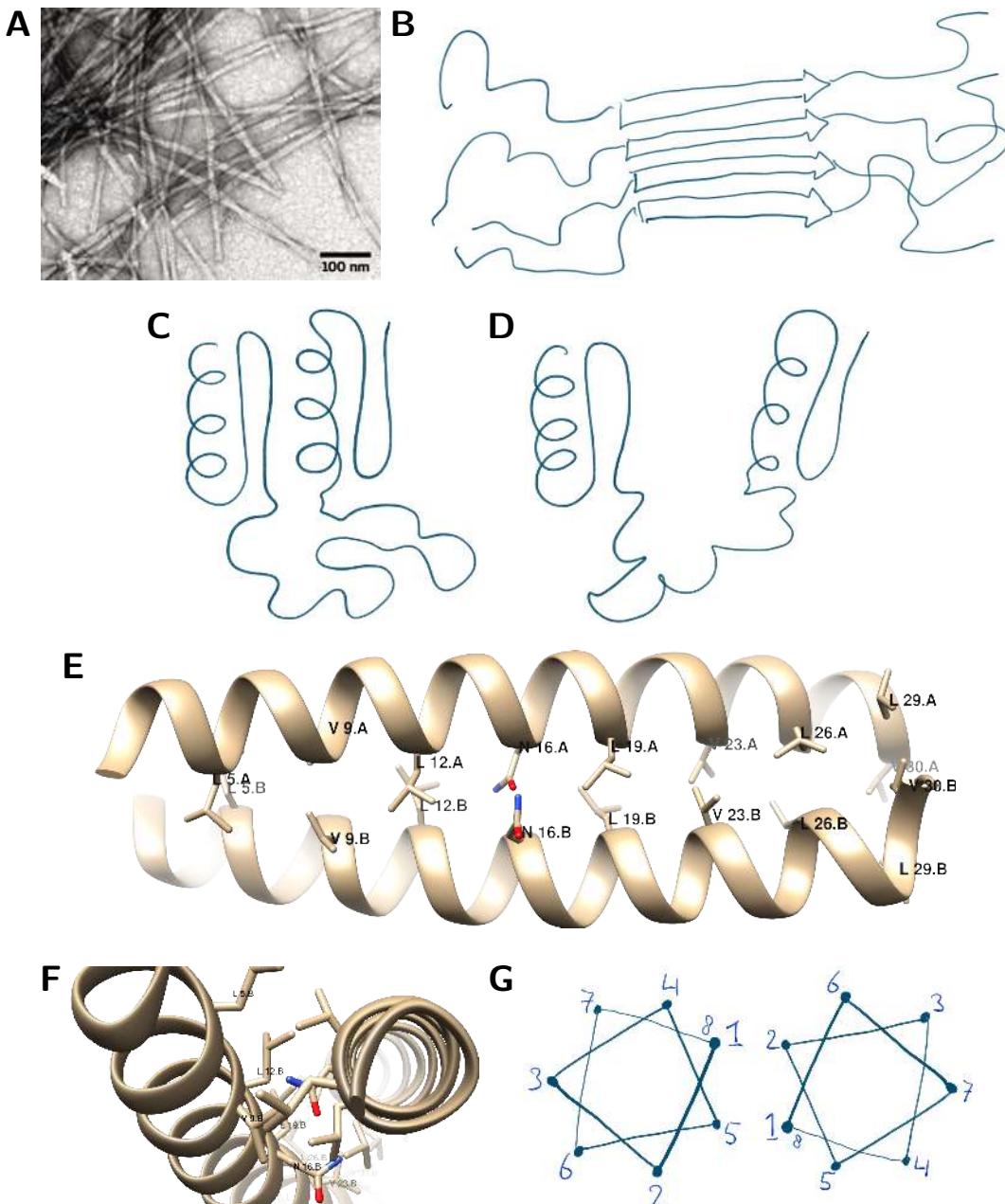


Figure 1.7: Three main classes of atypical protein structures: amyloid-fibrils or  $\beta$ -fibrils (A), as an example of bound ordered structure flanked by disordered loops or termini (B). Disordered proteins or regions; shown are schematically a disordered loop within a protein domain (C) and a disordered linker between two protein domains (D). (E+F)  $\alpha$ -helical coiled coils (PDB:2ZTA) which are characterized by the repetition of a Leucine every 7th residue, hence also referred to as ‘leucine zippers’. Shown length-wise (E), from the top (F) and schematically (G).

that the ability of proteins to form amyloidic structures tends to be generic (Dobson, 2003). Several other diseases have now been associated with the formation of ‘amyloid plaques’, which are large-scale deposits of  $\beta$ -fibrils that can be highly disruptive to tissue. A well-known example is some forms of Parkinson’s disease. However, it is not yet clear in most cases if these plaques are involved in causing the disease or merely a result of the disease process.

The **coiled-coil** is a twisted rod formed by a pair of  $\alpha$ -helices, this is shown in Figure 1.7b. It resembles a pair of tweezers, with one end slightly open, and both helix ends binding on either side of the DNA double helix in certain DNA binding proteins. The coiled-coil has a repetitive element of 7 residues where both helices are in direct contact. Typically, every 7th residue is a leucine, and valines or isoleucines are found in between. This creates a pattern like Lxx[VI][VI]xxL ([VI] means either V or I at that position). These structures are also referred to as “leucine zippers” and “leucine-rich repeats” because of the repeating leucine every 7th residue.

**Disordered protein regions:** some proteins never fold in a fixed three-dimensional structure, and are referred to as “**disordered proteins**”. These lack a folded structure, but display a highly flexible, random-coil-like conformation under physiological conditions. They will be further discussed in Chapter 9 “Structural Property Prediction” will briefly go into prediction of disordered proteins and regions based on sequence patterns. Many proteins contain large disordered segments (33% of eukaryotic, 2% for archaea, and 4.2% in bacteria) (Ward *et al.*, 2004).

## 2.5 Phi and psi angles

The backbone of a peptide consists of two flexible chemical bonds: NH–C $\alpha$  and C $\alpha$ –CO. These bonds can rotate around their axes, and they are referred to as torsion angles or dihedral angles. In this context, we will use the term torsion angles. The NH–C $\alpha$  torsion angle is denoted as  $\Phi$  or phi and is located at the beginning of each residue. The C $\alpha$ –CO torsion angle is denoted as  $\Psi$  or psi and is found at the end of each residue. Figure 1.8 illustrates this arrangement. The omega torsion angle is covered in the Panel “The omega torsion angle”.

Specific combinations of phi and psi angles allow the formation of favourable amide (NH) to carbonyl (C=O) hydrogen bonding patterns in the backbone. These combinations facilitate the formation of either  $\alpha$ -helix or  $\beta$ -sheet structures. However, certain phi and psi angle combinations can result in clashes within the backbone or between adjacent sidechains in the

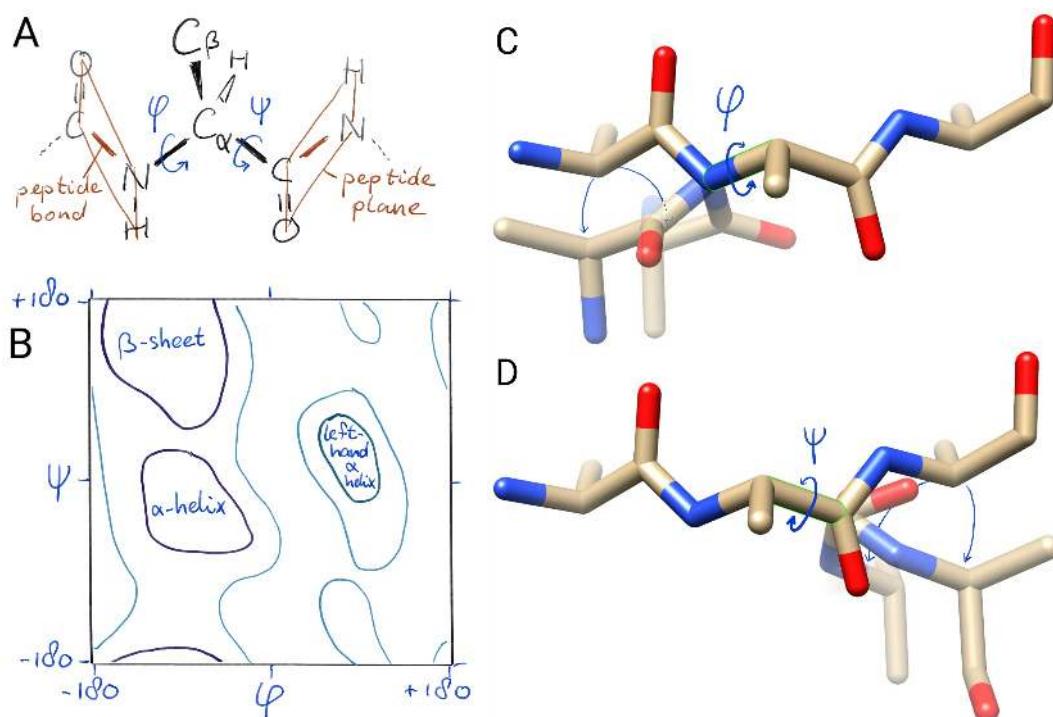


Figure 1.8: (a)  $\text{Phi}$  or  $\Phi$  and  $\text{psi}$  or  $\Psi$  angles defined in the backbone of an Alanine. (b)  $\text{Phi}$  and  $\text{psi}$  angles analyzed in a ramachandran plot. The contoured areas indicated allowed (light) and preferred (dark) combinations of  $\text{phi}$  and  $\text{psi}$  angles, which coincides with the two major secondary structure elements:  $\beta$ -sheets and  $\alpha$ -helices; in addition the smaller area of left-handed  $\alpha$ -helices can be seen in the positive quadrant (Chen et al., 2010). (c+d) Backbone re-arrangements in a tri-Alanine peptide, when adjusting the  $\Phi$  backbone dihedral angle of the central residue (c) or the  $\Psi$  angle (d). The backbone nitrogen ( $\text{N}$ ) is in blue, the carboxyl ( $\text{C}=\text{O}$ ) oxygen in red, and the carbon atoms in tan. The  $\text{C}\alpha$  atom is the one without (red) oxygen bound, and with the  $\text{C}\beta$  (tan) branching off of the backbone. These structural illustrations were created using Chimera.

protein chain, particularly if the sidechains are large. In some cases, clashes can even occur between the larger sidechains and the backbone of neighboring residues. Steric hindrance caused by both the backbone atoms and sidechains restrict the occurrence of certain combinations of phi and psi angles. Consequently, the number of potential conformations that may be adopted by the polypeptide is reduced. Sequence and propensity patterns that arise from this are exploited in secondary structure prediction, to which we will turn in Chapter 9.

Based on known protein structures, we can derive empirical distributions of phi and psi angle combinations. This distribution is visualised in a so-called Ramachandran plot of phi (horizontal) vs. psi (vertical), as shown in Figure 1.8. Firstly, we can observe that only some combinations of phi

and psi angles are allowed, e.g. light and dark outlined areas, whilst others are very uncommon (outside areas). The allowed regions are the secondary structure elements, such as the  $\alpha$ -helix (with negative phi and psi angles) and  $\beta$ -sheet (with negative phi and positive psi angles). Additionally, there is a smaller area that corresponds to the left-handed helix, which is observed but less frequently encountered. The areas without data points, the disallowed regions, indicate combinations of phi and psi angles that result in steric hindrance among the backbone atoms and are therefore not observed.

### The omega torsion angle

The peptide bond (between C=O and N–H) chemically connects two amino acid residues together. The consecutive peptide bonds form the backbone of the protein. Strictly speaking the peptide bond is a torsion angle like the phi and psi angles, however this bond is different. Due to the physicochemical properties of this bond it cannot rotate freely, this is related to the fact that it is in between a C=O and N–H group. This bond angle is called  $\omega$  or omega. In proline residues, it can switch between two possible angles in a process called ‘proline isomerisation’.

## 2.6 Secondary structure assignment

Secondary structure *assignment* involves determining the secondary structure class for each residue in a protein based on its structure. It is a structure-based definition for secondary structure. Protein structures are typically stored as a set of coordinates for each atom in the structure, see also Chapter 4. Various features such as **phi and psi angles or hydrogen bonding patterns**, can be used to assign secondary structure.

The most commonly used method to assign secondary structure is the Dictionary of Secondary Structure of Proteins, or **DSSP** (Kabsch and Sander, 1983), but several others exist such as **Stride** (Heinig and Frishman, 2004). DSSP first assigns hydrogen bonds to pairs of atoms, and uses these pairs to infer the secondary structure. For example, if several consecutive residues have hydrogen bonds that are four places ahead in the sequence, these residues are designated to be part of an  **$\alpha$ -helix** by DSSP. **Minimum lengths** of secondary structure elements are also considered by these methods, to avoid assigning a single residue as an  $\alpha$ -helix.

## 3 Tertiary structure

The tertiary structure of the protein, which represents its complete structure, consists of secondary structure elements, or motifs (see also Panel

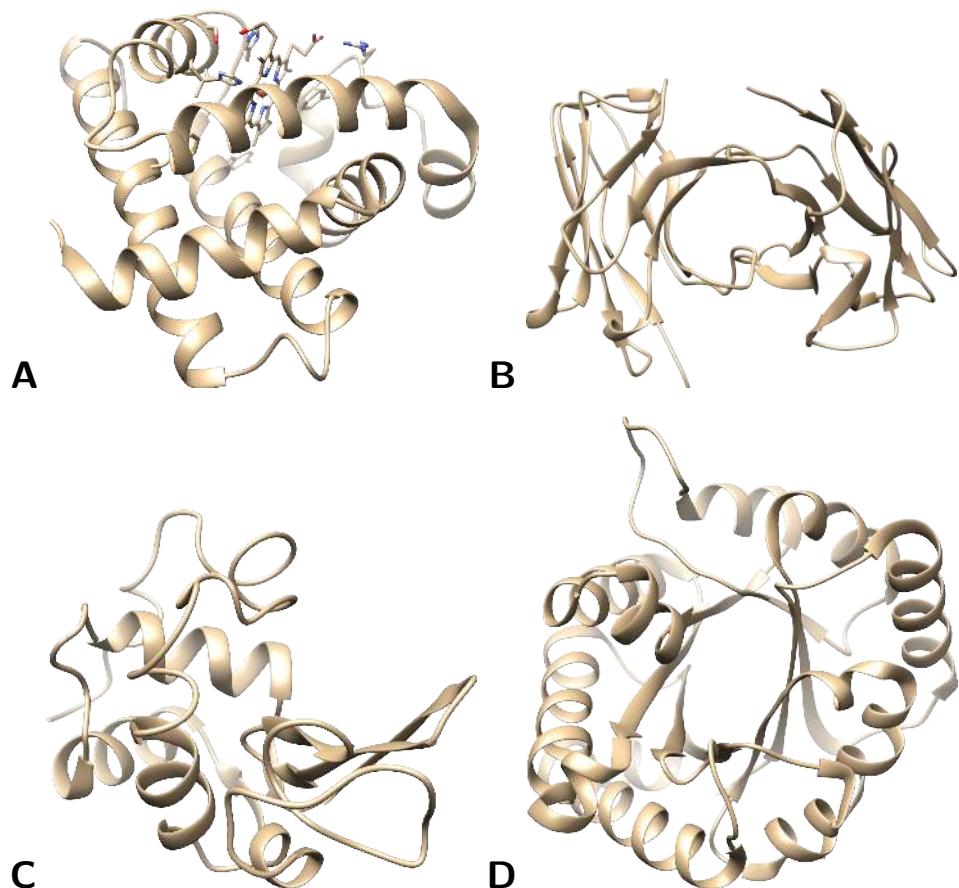


Figure 1.9: The four main protein fold classes, here showing a more or less famous example for each of them. (A) all- $\alpha$ : myoglobin, solved in 1960 by Sir John Kendrew, for which he received the Nobel prize (PDB:1mbn, Kendrew et al., 1960), consisting of only alpha-helices. (B) all- $\beta$ : Immunoglobulin domain, consisting of only beta-strands (PDB:1igt, Harris et al., 1997). (C)  $\alpha/\beta$  lysozyme by DC Phillips, where one domain is helical, and another strands (here on the bottom right) (PDB:1lyz, Diamond, 1974). (D)  $\alpha+\beta$  triose phosphate isomerase, where helical and strand regions intermingle (PDB:1tim, Banner et al., 1976).

“Secondary structure motifs”). The arrangement of secondary structure elements along the protein chain, and their folding to establish contacts in the three-dimensional structure, is called the *protein fold*. This is also referred to as ‘protein topology’. Four main fold categories are distinguished: all- $\alpha$ , all- $\beta$ ,  $\alpha/\beta$  and  $\alpha + \beta$  as shown in Figure 1.9.

### Secondary structure motifs

Secondary structure elements often occur in the protein structure in particular combinations, called ‘secondary structure motifs’. Some of these can be very informative, e.g. for assessing protein function, which is why they form the basis of structural classification schemes. As an example, Figure 1.9 shows the four main fold classes from one of the protein databases, SCOP (Andreeva *et al.*, 2008). For a comprehensive treatise on secondary structure motifs, we refer to Branden and Tooze (1998). Also, you can find more on this in Chapter 4 “Data Resources for Structural Bioinformatics”.

## 3.1 Hydrophobic core

Proteins in a cell are typically surrounded by water (except for transmembrane proteins which are located inside of cell membranes). Water is a polar molecule: with the oxygen atom slightly negative and the hydrogen slightly positive, see also Figure 1.3. Hydrophobic sidechains cannot make hydrogen bonds with the water molecules, therefore the solvent avoids to make contact with hydrophobic residues. Oil is also a hydrophobic substance, and as you may well know: oil and water do not mix well.

In a protein, the hydrophobic residues want to be shielded from the water, in any stable configuration. The result is that the protein will adopt a conformation in which the exposure of the hydrophobic sidechains to the water is minimized; hydrophobic residues will tend to become buried in the interior of the protein. This effect is known as the ‘hydrophobic effect’ and is the main driving force for protein folding. In the folding process, this leads to what is called the ‘**hydrophobic collapse**’. Figure 1.2 sketches the role of hydrophobic residues in the folded and unfolded states of a protein. We will come back to the folding process and the **thermodynamics and driving forces** behind in much more depth in Chapter 12 “Introduction to Protein Folding”.

For now, it is important to realize that the interior of the protein will consist mostly of residues with hydrophobic sidechains, but the backbone is polar. It is **not possible** to keep **all the polar backbone parts of the buried hydrophobic residues at the surface of the protein**. At the surface, the polar sidechains as well as the backbone, form hydrogen bonds with the water, but the backbone of the buried hydrophobic residues cannot do this. This creates a problem, as it is very unfavourable if these backbone hydrogen bonding capacity remains unsatisfied. The solution to this problem is the formation of **secondary structure**, as covered in the previous section. **Regular secondary structures** will therefore usually make up the core of the protein.

### 3.2 Protein domains

Protein domains are conserved regions that will be mentioned often in this book. More generally, we could define them as self-folding, evolutionary conserved subunits of structure. Domains typically have a specific molecular function, and may recurrently appear in different proteins. Most eukaryotic proteins have multiple domains, which may be linked together by a small linker, or large disordered regions. There are several distinct ways in which domains may be described, each of which will be explained in further detail elsewhere in this book. In Chapter 4 we will see how structural domains can be defined; in Chapter 6 it becomes clear that structure prediction is most effective at the domain level.

## 4 Quaternary structure

Protein-protein interactions (PPIs) involve the binding of two or more proteins, leading to the formation of a protein complex known as the quaternary protein structure. This structure represents a natural extension of the primary, secondary, and tertiary structures. It is worth noting that protein function often emerges at the level of the quaternary structure, as it determines the specific function performed by the protein complex. We will come back to protein function, and the role of interactions, in Chapter 11.

## 5 Key points

- Proteins fold from an unstructured polypeptide coming from the ribosome into their functional native conformations.
- Structure Basics:
  - primary, secondary, tertiary, quaternary
  - phi/psi angles
  - hydrogen-bonds
- Loops tend to be more flexible
- Hydrogen bonds may be satisfied by backbone, sidechain or water
- PDB & Structural genomics: bias in data
- Protein structure may be predicted from sequence
- Function may be derived from structure

## 6 Further Reading

- “Sequence Analysis” (Durbin *et al.*, 1998)
- “Introduction to Protein Structure” (Branden and Tooze, 1998)

## Author contributions

Wrote the text:	AJ, EvD, BS, HM, KAF, SA
Created figures:	HM, JG, KAF, SA
Review of current literature:	HM, KAF, SA
Critical proofreading:	BS, KAF, SA
Editorial responsibility:	SA, KAF
The authors thank Reza Haydarlou	 and Nicola Bonzanni 
	for non-expert feedback.

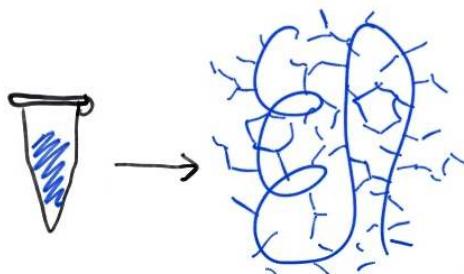
## References

- Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E. et al (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, **36**(SUPPL. 1), 419–425.
- Banner, D., Bloomer, A., Petsko, G., Phillips, D. and Wilson, I. (1976). Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochemical and Biophysical Research Communications*, **72**(1), 146–155.
- Branden, C. and Tooze, J. (1998). *Introduction to protein structure*. garland publishing, New York.
- Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A. et al (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(1), 12–21.
- Diamond, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *Journal of Molecular Biology*, **82**(3), 371–391.
- Dobson, C.M. (2003). Protein folding and misfolding. *Nature*, **426**(6968), 884–890.
- Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- Harris, L.J., Larson, S.B., Hasel, K.W. and McPherson, A. (1997). Refined Structure of an Intact IgG2a Monoclonal Antibody.
- Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, **32**(Web Server), W500–W502.
- Kabsch, W. and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577–2637.
- Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G. et al (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, **185**(4711), 422–427.
- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S. et al (2004). UCSF Chimera – A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13), 1605–1612.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, **337**, 635645.

# Chapter 2

## Structure determination

Halima Mouhib [ID](#) Bas Stringer [ID](#) Hugo van Ingen [ID](#)  
Jose Gavaldá-García [ID](#) Katharina Waury [ID](#) Sanne Abeln [ID](#)  
K. Anton Feenstra [ID](#)



\* editorial responsibility

## 1 Introduction

The main emphasis of this work is to provide a background on experimental techniques for protein structure determination. The focus is set on X-ray crystallography and Nuclear Magnetic Resonance spectroscopy (NMR), which are by far the main methods used to determine the structure of soluble proteins. We will also introduce cryogenic Electron Microscopy (cryo-EM) and electron diffraction which are more suited to analyze **membrane proteins and larger protein complexes**. At the end, more qualitative techniques are summarized that are used to obtain insight on the overall structure and dynamics of proteins. Note that this introduction to protein structure determination aims at familiarizing the reader to different experimental techniques, their benefits and bottlenecks, but that a thorough mathematical and technical description of the concept is beyond the scope of this work. For the interested reader, Section 8 provides selected works that go deeper into the details.

Generally speaking, structure determination is an immediate result of understanding the interaction between **light (radiation) and matter**. Figure 2.1 shows an overview of the electromagnetic radiation (e.g. light) and the corresponding wavelengths used by the different experimental techniques. A corresponding overview of the methods including the used wavelengths, their reachable **resolutions, benefits, and limitations** is given in Table 1. Depending on the sample of interest, different techniques are applied to resolve the structure.

X-ray Crystallography and NMR have traditionally been the two main methods for protein structure determination, as they can obtain the highest '**atomic**' resolutions – in the range of **1-2 Å**. X-ray crystallography uses

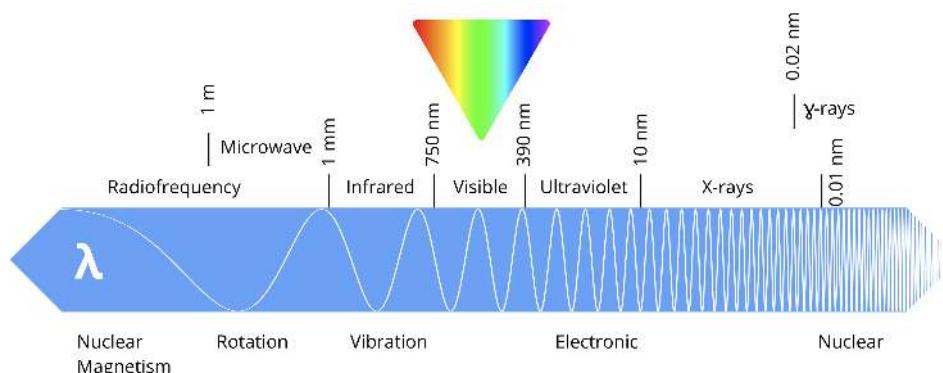


Figure 2.1: Electromagnetic waves spectrum and their applications. In principle all (or most) wavelengths of the electromagnetic spectrum can be used to obtain information from molecules. Different kinds of electromagnetic waves (top) are used to obtain diverse information on molecular systems (bottom).

Table 1: Overview of different Methods used for protein structure determination.

Method	Resolution	Wavelengths	Strengths	Limitations
X-ray	1 – 2 Å	nm-pm	atomic resolution	purity & crystal static
NMR	1 – 2 Å	m	can be in solution dynamics direct interactions	only small proteins
EM	mm-μm	mm-μm*	direct imaging unpurified	very low resolution static
Cryo-EM	~ 2 – 10 Å	μm-nm*	measure phases large complexes	not atomic resolution static
IR	–	mm-μm	global structure dynamics	no atomic assignments
CD	–	10 – 100 nm	global structure dynamics	no atomic assignments

X-ray: protein X-ray crystallography; NMR: Nuclear Magnetic Resonance spectroscopy; EM: Electron microscopy; IR: Infrared spectroscopy; CD: Circular Dichroism; \* electron radiation.

Note: Atomic resolution lies in the range of 1-2 Å: the atomic (Van der Waals) radius of carbon is about 1.5 Å, that of nitrogen and oxygen 1.1 Å. Bond lengths between carbon, nitrogen or oxygen are in that same range.

short wavelength X-ray radiation; for X-ray, the wavelengths used limit the resolution of the diffraction data obtained. **Shorter wavelengths give higher resolution information.**

NMR uses the long wavelength radio waves as these frequencies correspond to the energy levels of **spin-state** transitions in the nuclei of atoms, which are sensitive to the **local (atomic) environment**. This local environment yields information about the relative positions of atoms, from which the overall protein structure is constructed.

Over the past decades, cryo-EM has been steadily pushing down on the resolution limit, going down from 4 Å in 2008 (Yu *et al.*, 2008; Zhang, 2008), over “near-atomic” resolutions of 3-3.5 Å (Li *et al.*, 2013; Earl *et al.*, 2017) to atomic resolution close to 1 Å (Herzik, 2020). This technique is particularly interesting for **large assemblies** such as viruses (Jiang and Tang, 2017; Ward and Wilson, 2017) and flagella (Egelman, 2017).

All successfully resolved and published structures, usually obtained from X-ray, NMR or cryo-EM, are accessible via the Protein Data Bank (PDB; [www.pdb.org](http://www.pdb.org), see Chapter 4 for more detail). Besides X-ray, NMR and cryo-EM, which yield direct information on atomic coordinates, other spectroscopic techniques (visible, UV, IR) are sensitive to electronic and molecular vibrations. These measurements can be used to probe various properties

of molecular systems and obtain more **global and qualitative** information (e.g. the amount of secondary structure elements). We will introduce these methods in some more detail throughout the chapter to provide a general overview of available techniques in structure determination as well as sufficient references to dig in deeper yourselves.

## 2 X-ray crystallography

A simplified work-flow used in **X-ray crystallography**, which to this date is still the method that provided most of the available protein structures, is shown in Figure 2.2. First, the proteins need to form a regular **crystal** structure, such that their orientation is regular and very densely packed against each other. The fixed orientation of the crystals allows the **X-ray diffraction**, in the next step, to be recorded as a regular pattern. In a third step the **electron density** may be acquired from the diffraction pattern, but first the **phase problem** needs to be solved. As a result, a **3D structure** may be **fitted to the density**. The different steps and relevant concepts are explained in the following sections. Note that obtaining good crystals of the protein, and the derivation of the electron density from the diffraction data are the more challenging parts (highlighted with red lightning bolts in Figure 2.2).

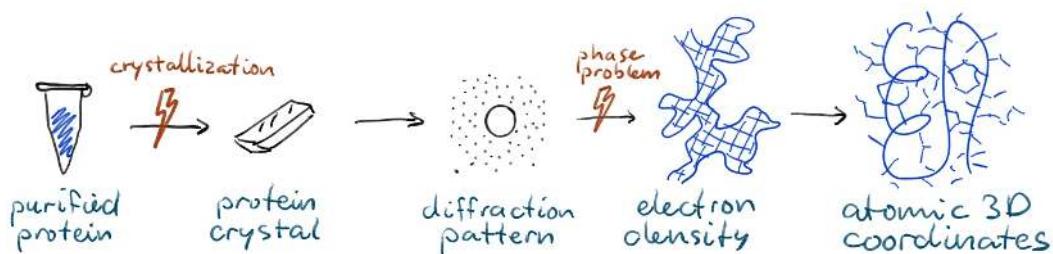


Figure 2.2: Simplified work-flow of protein structure determination through X-ray crystallography. Crystallization and the phase problem are the main bottlenecks.

### 2.1 Crystallization

When you try to imagine a crystal, you may first think about something like salt or sugar grains, or some rock or gem. While protein crystals have a similar appearance, inside, they contain a surprisingly **large amount of water**: between 20% and up to as much as 80% by volume. Figure 2.3 shows examples of a particularly dense (little water) and an open (much water) crystal elementary cell. You should realize that, for example, a regular packing of spheres (think of a box of marbles) also contains about 25% empty space (see also Atkins and De Paula (2014) on crystal packing); however 80%

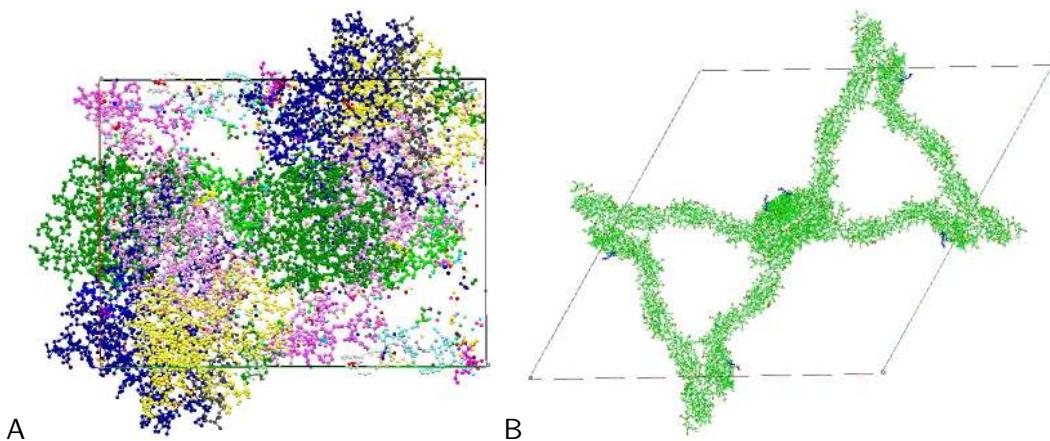


Figure 2.3: A) A typical protein crystal of human deoxyhemoglobin (PDB:4hhb) with a relatively low fraction of water. (B) The crystal packing of Myelin-associated glycoprotein (PDB:5lf5), which contains an exceptional amount of water. The elementary cell (repetitive units in the crystal) are outlined with rectangles.

water is more like a **gel** than a crystal. This is not so different compared to the typical cytosol (inside a living cell), which is about 70% water (Luby-Phelps, 1999) with most of the rest (20-30%) taken up by proteins (Ellis, 2001). As one might expect, anything that makes a protein **flexible** can interfere with the crystallization process. This problem is two-fold: not all proteins will be in the **same conformation**, making it difficult to obtain the regular packing required in the crystal. (Although in a few cases, a single crystal may contain different conformations of the same protein – this happens already naturally in some virus capsids.) The second part of the problem is related to the **loss of entropy** upon crystallization, since the protein molecules become more ordered as they occupy their fixed positions on the crystal lattice (we will return to entropic effects in Chapter 13).

### Challenging structures

**Membrane proteins** such as G protein coupled receptors (GPCRs) in particular pose a challenge. These protein represent about **60% of the drug targets** and are thus extremely important for drug design. They need to be embedded within the membrane to be stable, but the membrane consists of many small (lipid) molecules and is very flexible, which makes it almost impossible to fit into a crystal. Even though a whole array of tricks has been invented (like using simple detergents instead of lipids; or even inducing two-dimensional crystallization inside a membrane), this is still the largest bottleneck in protein crystallography.

phy. Until 2005 the only available crystal structures of GPCRs were of rhodopsin. Also, until today, it is still not possible to crystallize olfactory GPCRs which are responsible for detecting odorants in the nose.

**Glycoproteins** are another example that pose a double challenge for crystallographers: first, the **glycan (sugar)** groups are very **flexible**, which like flexible linkers, loops or termini, interferes with crystallization. But the (even) greater problem is **heterogeneity**. The glycan groups are added after translation, i.e. so-called post-translational modifications. This is done by enzymes which have specific affinities for attaching certain glycans in given places on the protein. But the placing, number and types attached may vary from protein molecule to protein molecule. This means the crystal must now accommodate protein molecules which have **slightly different shapes**. You can imagine that this will not work very well. Moreover, **without** the glycans attached, many of these proteins adopt **different conformations**, or even remain **largely disordered**. And, finally, the enzyme machinery for attaching the glycans can vary between species, and only eucaryotes have them. This makes **production** of these proteins in the right form experimentally challenging as well.

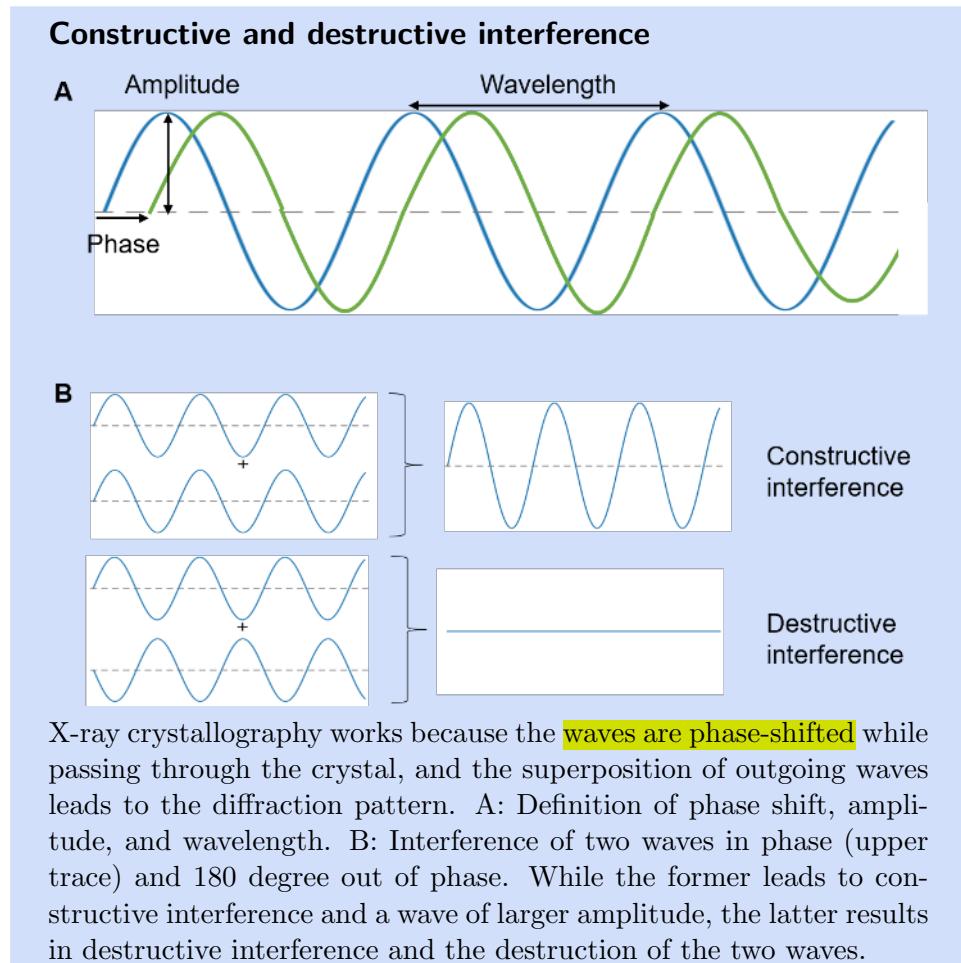
New developments in cryo-electron microscopy (cryo-EM) will allow to obtain more information on these kind of challenging systems. Since the determination of the first EM structure of an activated GPCR at approximately 4 Å resolution in 2017 (Zhang *et al.*, 2017), cryo-EM has very recently moved towards atomic resolution close to 1 Å and is even expected to reach resolutions below 1 Å in the years to come (Herzik, 2020). We will come back to cryo-EM in Section 4.

## 2.2 Diffraction

In an X-ray diffraction experiment, a well-defined and narrow X-ray beam is directed at the crystal. Usually some tungsten based X-ray generator is used as the radiation source to provide light of a given wavelength between 0.1-10 nm (see also Figure 2.1). During the data acquisition, the crystal is often cooled by so-called ‘**cryogen**’: a stream of liquid nitrogen or helium . This has two reasons. First, the X-ray radiation hitting the crystal will **heat** it up and eventually cause damage. Second, to obtain highest resolutions, **atomic motions have to be reduced**.

The radiation that is **diffracted** by the crystal is then captured and recorded by a detector (all radiation that goes straight through the crystal on the other hand is stopped by a small slab of metal called the “beam stop”). The diffracted radiation makes up a **diffraction pattern** which contains the information that we need to derive the coordinates of the atoms

in our protein.



It is the easiest to envision the process for a single protein atom at a time: the incoming radiation will strike the atoms in each of the protein molecules in the crystal. Because the atoms are in different positions, the radiation for each atom will travel its own specific distance (“path length”) from the source to the detector. Each radiation wave is defined by two properties: its **amplitude and its phase** (see Panel “Constructive and destructive interference” A). For different directions, the waves will be effectively randomized, and most energy gets **averaged away or even canceled out** (destructive interference). However, in some specific directions the waves are in phase, so the waves will be **amplified**, and the intensities add up (constructive interference) to produce a spot on the detector. Both types of interference are shown in Panel “Constructive and destructive interference” B. Each different atom in the protein produces **multiple** such spots, but many different atoms may also contribute to the **same** spot. The relation to describe the

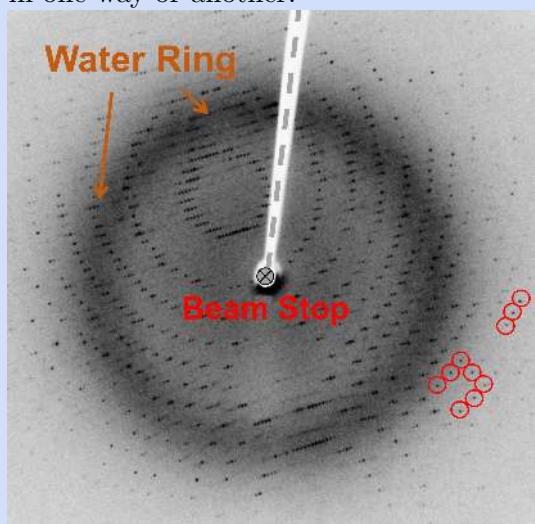
path length differences as a multiple of the wavelength,  $n\lambda$ , is known as Bragg's law in crystallography:

$$2d \cdot \sin\theta = n\lambda \quad (1)$$

Hereby,  $d$  represents the distance between two planes in the crystal,  $\theta$  (theta) is the glancing angle,  $\lambda$  the wavelength, and  $n$  the diffraction order. Figure 2.4 shows Bragg's law in context of protein molecules inside a crystal. Considering basic trigonometric rules, you can easily derive the relations between the angles and lengths to determine the distance  $d$  between the planes of the crystal. Note that constructive interference will only be achieved if Bragg's law is fulfilled.

### Diffraction Pattern

A typical diffraction pattern is shown below. The amount of detail, or resolution, of the data increases with distance from the centre. Thus, the crystallographer can immediately say what the maximum resolution could be, by looking at the furthest observed 'reflections' – the black dots scattered in patterns across the image (highlighted in red below). It is important to note that the diffraction pattern is the (only) primary data that an X-ray experiment produces. All the rest (densities, atomic coordinates, B-factors), are modelled onto the primary data in one way or another.



X-ray diffraction pattern of crystallized 3Clpro, a SARS protease (2.1 Å resolution)<sup>a</sup>.

The diffraction pattern has several striking features that do not carry information about the protein structure. In the middle is a blank area, caused by the beam stop preventing any ('direct beam')

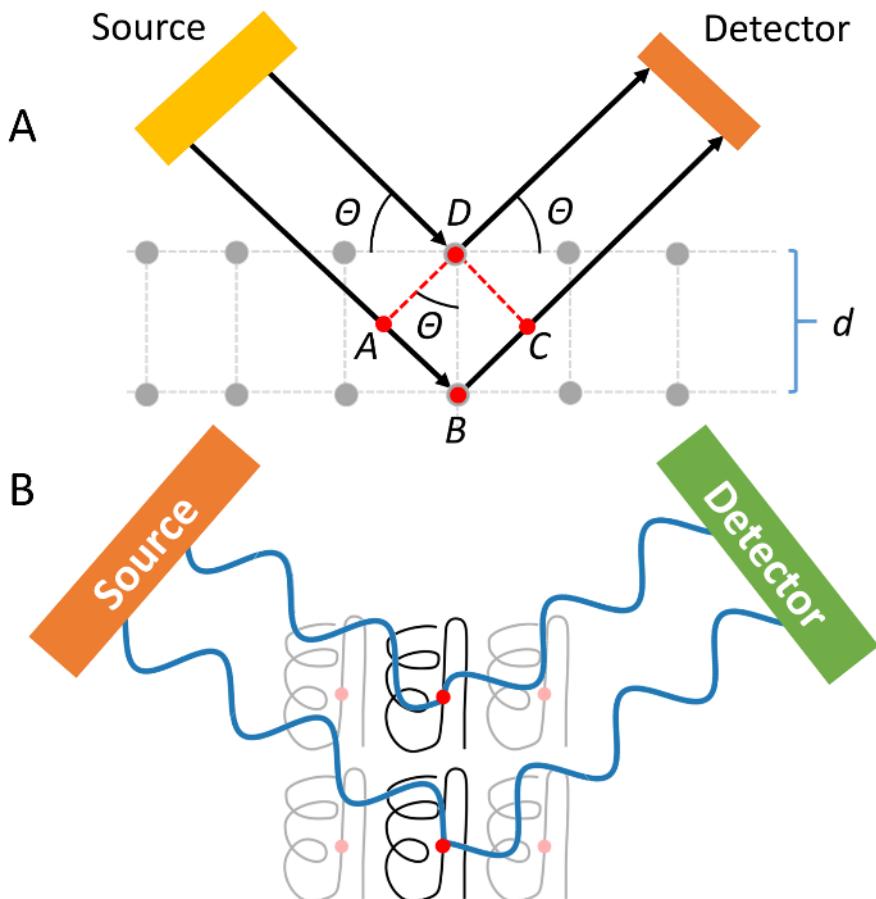


Figure 2.4: Bragg's law. A: Simplified scheme to define the distance  $d$  between two planes of the crystal lattice, the glancing angle  $\theta$ , wavelength  $\lambda$ , and the diffraction order  $n$  using 4 atoms (A, B, C, D) in a crystal lattice. B: Put into context of atoms inside the protein molecules of the crystal. Incident radiation is drawn to come from the left. For simplification, we show two atoms in two different lattice layers, that scatter the radiation in a specific angle onto the detector.

radiation from reaching the detector. Unfortunately, protein crystals only **diffract a (small) fraction of the incoming radiation**, so not using a beam stop would completely overwhelm the detector (akin to pointing a camera at the sun). There will be a ring caused by **diffraction** on the randomly oriented **water** in the crystal, known as the '**water ring**'. The intensity of this ring will depend on the fraction of water present. Also the **loop** or, in this case, the rod that holds the crystal will scatter some of the radiation (not diffraction, just bouncing off the surface).

Finally, we see many small dots known as reflections. You can see

they lie in a pattern, which will vary depending on the type (symmetry) of the arrangements (packing) of the protein molecules in the crystal. In this lattice pattern, each point has a set of three indices, relating to the angles of the diffraction. The actual data used are the intensities of each (observed) spot at each possible lattice (index) position.

<sup>a</sup>Source: Jeff Dahl [https://commons.wikimedia.org/wiki/File:X-ray\\_diffraction\\_pattern\\_3clpro.jpg](https://commons.wikimedia.org/wiki/File:X-ray_diffraction_pattern_3clpro.jpg)

Mathematically, since the atoms are regularly arranged in the crystal, the observed pattern now corresponds to a (three dimensional) Fourier transform of the positions of the atoms. So, in principle we would just need to do a reverse Fourier transform to obtain the positions from the diffraction pattern, which can be found by:

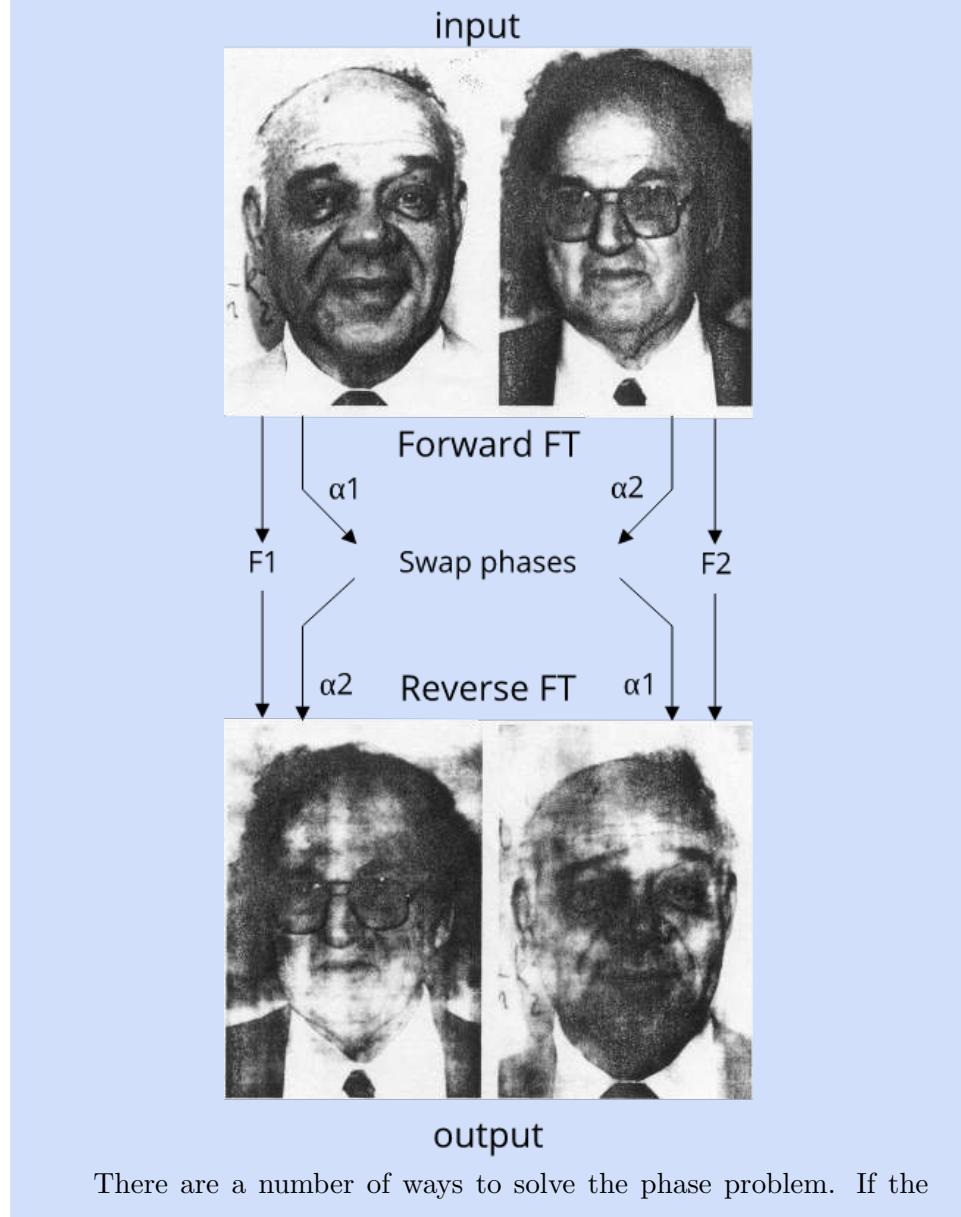
$$\rho(x, y, z) = \frac{1}{V} \sum_{h,k,l} |F(h, k, l)| \exp(i\alpha_{hkl}) \exp(-2\pi i(hx, ky, zl)) \quad (2)$$

The outcome of this reverse Fourier transform is  $\rho$ , the (electron) density that we want to observe. The three integers  $h$ ,  $k$ ,  $l$  are known as Miller indices and are an established notation system in crystallography to describe the planes of the different crystal lattices. Input is the structure factor  $F$ , i.e. the amplitudes that we get from the reflections (spots) in the diffraction data. Through this Fourier relationship, only the spots of the diffraction pattern are required. This means that experimentally, the amplitudes of the diffraction pattern (strengths of the spots arising from constructive interferences) are directly accessible. Unfortunately, the experiment does not provide any information on the associated phase  $\alpha$  of the diffraction (contained in the imaginary component  $\exp(i\alpha_{hkl})$ ). Without the phase information, it is not possible to reconstruct the electron density of the crystal cell. This is known as the “Phase Problem” in X-ray crystallography. For a more detailed description of the Fourier synthesis and the phase problem see the recommended further readings and Cowtan (2003). The Panel “There is a lot of information in the phases!” illustrates this problem using photographs of the two pioneers in (protein) crystallography.

### There is a lot of information in the phases!

The examples in the figure below show how important the phases,  $\phi$ , are for the reverse Fourier transform (The pictures show two pioneers of X-ray crystallography: Karle on the left and Hauptman on the right). The pictures were forward Fourier transformed (data not

shown) which results in the phase and amplitude, and then subsequently reverse Fourier transformed. However, for the reverse transformation, the **phases** between the two datasets are swapped, i.e. we get Karle with Hauptman's phases, and Hauptman with Karle's phases. This is a rather extreme example, and when the model and real structure are closer (see below) the effect is a lot less severe. Still, clearly, a lot of the information is contained in the phase and not in the amplitudes. This effect is known as "the phase problem" in crystallography (Cowtan, 2003).



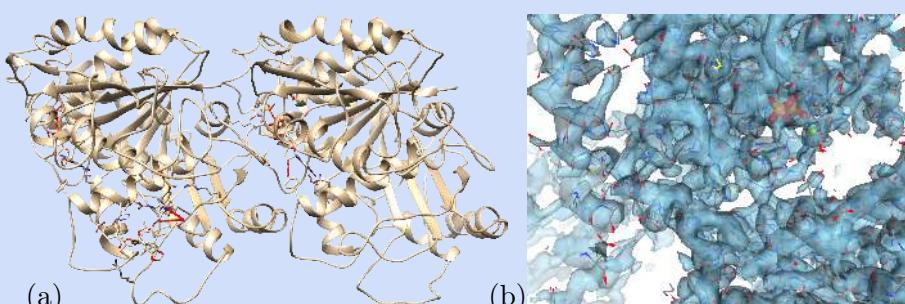
There are a number of ways to solve the phase problem. If the

resolution of the data is very high (better than 1 Å) and the protein is (very) small, there may be enough information in the amplitudes. Otherwise (in almost all cases for proteins), one can get some of the phase information by giving some of the radiation an “offset”; the amplitude changes caused by the offset are a measure of the phase. Incorporating heavy atoms into the protein structure does just that, but this requires chemical modification of the protein and carefully replication of the X-ray data collection. Something similar can be done by using not a single X-ray wavelength (which gives cleaner data), but multiple wavelengths. The differences in diffraction of different wavelengths also yield some phase information when compared.

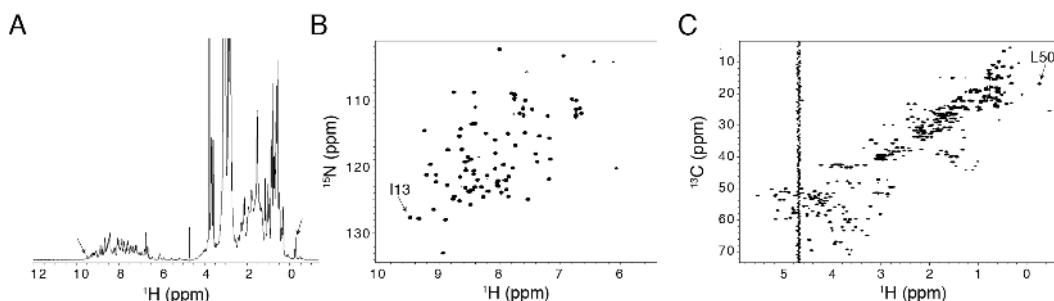
A very different solution makes use of the fact that once you have a reasonable estimate of the structure of the protein of interest, one can calculate the phases from a (forward) Fourier transform of the electron densities derived from the estimated structure. These calculated phases can then be used in the reverse Fourier transform, yielding electron densities. From these electron densities one obtains a new (usually better) set of coordinates for the structure of the protein, which is then used to calculate update phases. This process is usually iterated till convergence. However, convergence is not strictly guaranteed. This approach is sometimes called ‘molecular replacement’.

*Images used with permission from Randy J. Read (Read, 1997)*

### Electron Diffraction



(a) Near atomic resolution structure by electron crystallography and diffraction of the  $\alpha\beta$  tubulin dimer by electron diffraction from PDB:1tub (Nogales *et al.*, 1998). Great advantage of electron diffraction over X-ray diffraction, is that electron detectors do allow the direct measurement of phases. But resolution is limited to ‘near atomic’ (3.7 Å), just good enough to identify all secondary structure elements. Most importantly, because of the measurement of phases, the densities are well enough defined to accurately thread the protein chain through



*Figure 2.5: NMR spectra of ubiquitin (76 residues): (a) 1-dimensional hydrogen spectrum. Arrows correspond to the proton signals of the labelled peaks in (b, c). Intense peaks between 3 and 4 ppm are from the buffer. (b) 2-dimensional hydrogen-nitrogen (HN) spectrum. The backbone NH signal of Ile13 is labelled. (c) 2-dimensional hydrogen-carbon (HC) spectrum. One of the methyl  $\text{CH}_3$  signals of Leu50 is labelled. The vertical ridge is from the water signal. The spectrum axes (horizontal in a, and both in b,c) are expressed in parts-per-million ('ppm') deviation of the frequency with respect to a standard reference. Due to two spectrum dimensions being used in the 2D experiments, most of the overlapping peaks that appear in the 1D spectrum are resolved. The HN spectrum shows signals of the backbone amide NH groups and signals from NH group in side chains of some amino acids. This spectrum is very sensitive to changes in protein conformation, see panel 'NMR-based modelling of protein complexes'. The HC spectrum shows signals of the - $\text{CH}$ , - $\text{CH}_2$  and - $\text{CH}_3$  groups in aliphatic side chains as well as the backbone CH group at the alpha-position.*

them. (b) For comparison the density map of beta-tubulin at 3 Å from PDB:5yls (Yang *et al.*, 2018).

### 3 Nuclear magnetic resonance

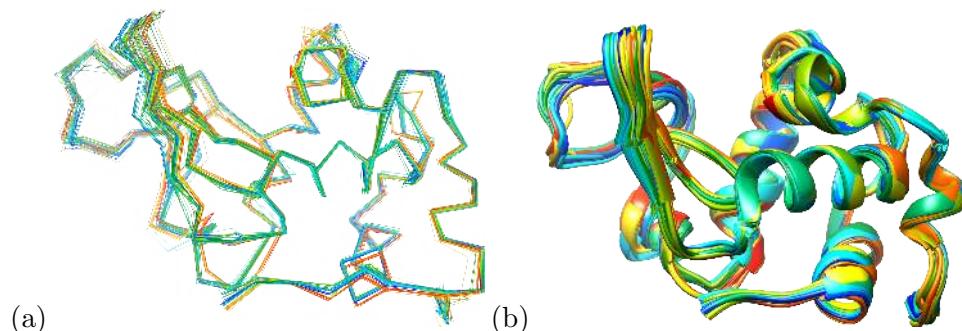
Nuclear Magnetic Resonance (NMR) relies on a property of some atomic nuclei known as '**nuclear spin**' to characterize the structure and dynamics of molecules. The nuclear spin can be thought of as a tiny bar magnet. Fortunately for (bio)molecular scientists, the hydrogen nucleus is magnetic. To reveal its magnetism a very strong magnetic field is needed. When placed in the external magnetic field, the nuclear spin will tend to align with this field. This so-called '**up**' orientation is the energetically most favourable state. The spin can also transition to a higher energy state, the '**down**' state, with exactly opposite orientation. The **energy difference** between these two states corresponds to wavelengths in the radio frequency range and is measured in an NMR experiment. Next to hydrogen, some isotopes of other atoms found in proteins, such as  $^{15}\text{N}$ ,  $^{13}\text{C}$ , are also NMR active and can be used to obtain additional information on the structure.

The frequency associated with transitions between the two levels not only depends on the type of atom (say hydrogen vs. phosphorus) but also on the

local chemical environment of the atom (say a hydrogen atom in a methyl-group vs. in an aromatic ring). The sensitivity of the transition frequency to the local environment of the spin is the basis for the application of NMR in chemistry. To stress this importance, the transition frequency is usually called the '*chemical shift*'. Chemical shifts of a given nucleus are very small and therefore expressed in parts per million (ppm), based on the relative change in transition frequency compared to a standard compound (usually tetramethylsilane, which is added as a reference in the experiment). The NMR spectrum of a protein will typically contain hundreds to thousands distinct signals, because each hydrogen nuclear spin will have a slightly different chemical environment. This results in a very crowded spectrum, as can be seen in Figure 2.5A, showing the NMR spectrum of ubiquitin. To use NMR in an intelligible manner two things have to be accomplished: first, the different signals need to be resolved; second, structural information about the relative position of the nuclear spins needs to be encoded in the NMR signal.

To resolve the overlapping signals, one usually performs a two-dimensional or three-dimensional experiment. Here, the chemical shifts of two or three different nuclear spins are measured simultaneously along the different dimensions of the spectrum. Typically, the additional dimensions are used to measure the chemical shift of the heavier isotope of nitrogen and carbon ( $^{15}\text{N}$  and  $^{13}\text{C}$ ) which also have spin. Figure 2.5B and C show the 2D H-N and H-C spectra of ubiquitin, illustrating that the addition of another chemical shift dimension allows to resolve nearly all signals. The Panel "NMR two-dimensional spectrum" shows some more detail in part of a 2D spectrum of a small peptide. The principle for generating 2D spectra can be extended to include multiple dimensions. In practice, 3D NMR, e.g. H-N-C, spectra are common in protein structure determination and are crucial to find out which peak corresponds to which atom (see the panel "NMR chemical shift assignments and structure determination").

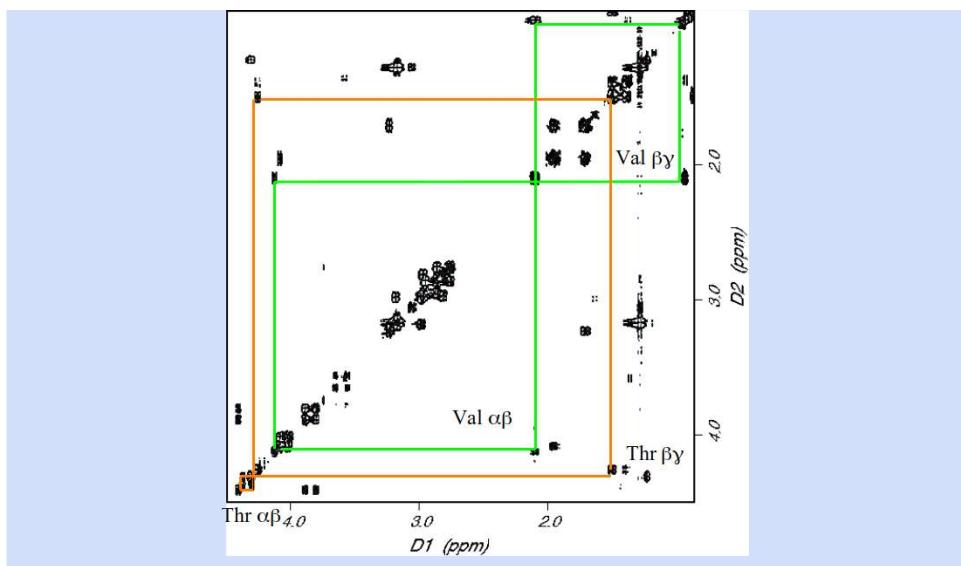
To encode structural information in the NMR signal, again multi-dimensional experiments are used. The spectra in Figure 2.5B and C already encode information about the secondary structure, as the chemical shifts of the backbone nuclei are sensitive to the backbone dihedral angles. To get data on the 3D fold of the protein, one makes use of the fact that the nuclear spin can 'sense' the presence of other nearby nuclear spins through their mutual magnetic interaction. This allows one to transfer the magnetic energy of one nuclear spin to another spin, this is known as the nuclear Overhauser effect (NOE). In a dedicated 2D experiment one then measures the chemical shifts of the two spins involved. Since the magnetic interaction between spins is distance dependent, the energy transfer and thus signal intensity ('NOE intensity') is also distance dependent. In this way, the distances between nuclei can be measured. For more explanation, please refer to Panel "NMR chemical shift assignments and structure determination".



*Figure 2.6: Representation of the ensemble of NMR solution structures by (a) an explicit ensemble of backbone traces, and (b) an overlay of ‘cartoon’ renderings. One can clearly see that variability between conformations is different in different places of the protein. PDB:1e8l (Schwalbe et al., 2001)*

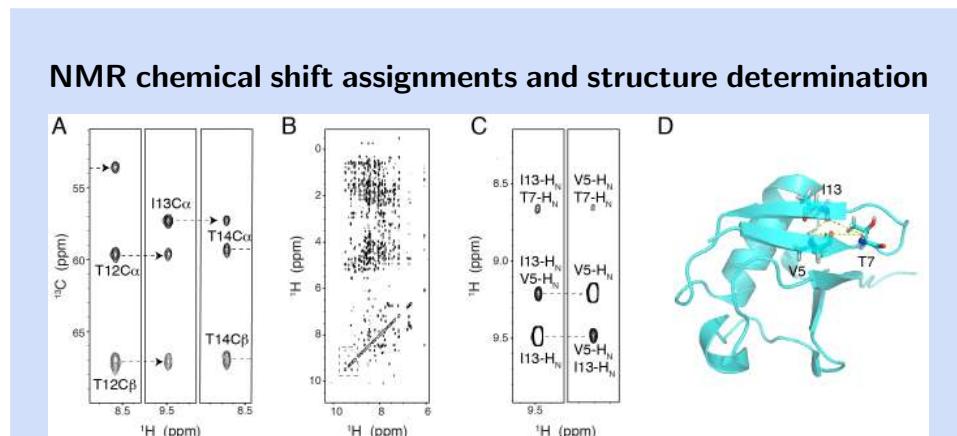
### NMR two-dimensional spectrum

In the Correlation Spectroscopy (COSY) experiment, the setup is such that energy transfer is predominant through chemical bonds. The figure shows an example for a small peptide (Feenstra *et al.*, 2002). Diagonal peaks are where the absorbing hydrogen nucleus also emits, and cross peaks correspond to energy transfer; note that transfer here is induced via chemical bonds. The cross-peak connections between the  $\alpha$  and  $\beta$  hydrogens and between  $\beta$  and  $\gamma$  hydrogens for the Valine and Threonine are traced out. This allows us to identify frequencies with unique individual protons in the molecule, which is rather important, because we cannot know beforehand which hydrogen will respond to which frequency. Since we know the sequence of our molecule, using the 2D-COSY spectrum we can trace out which hydrogen atom is where in the spectrum. The process is called ‘chemical shift assignment’: the assignment of which specific frequencies correspond to which atoms.

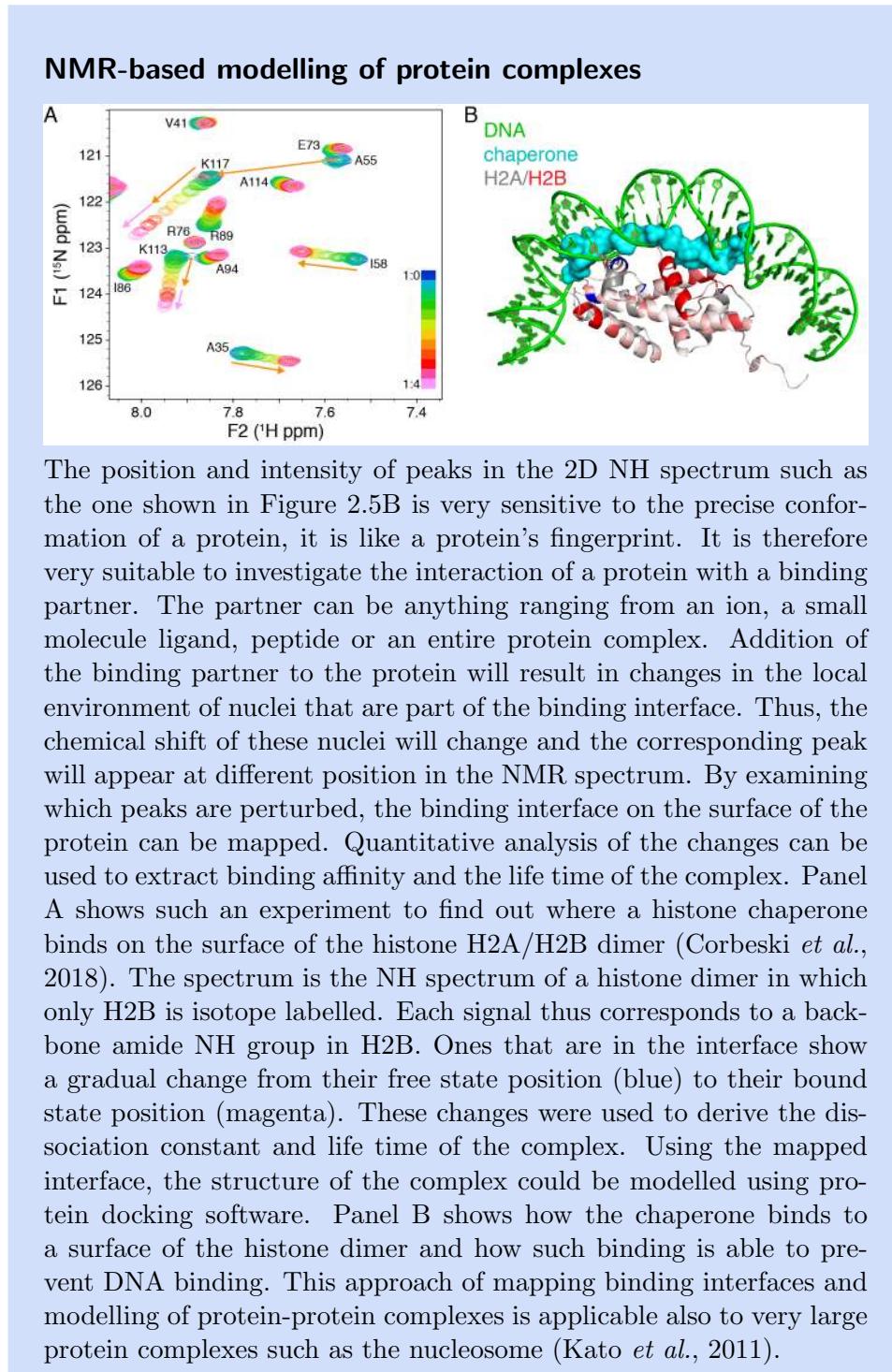


The distances and dihedral angles that are measured from NMR spectra are combined with a priori knowledge of the amino acid sequence of the protein and the structures of the amino acids to derive the 3D protein structure. It should be noted that NMR distance and angle data do not give direct access to the atomic coordinates, but these rather serve as constraints in the structure calculation process. The structure is fitted to these constraints in a calculation much like a molecular dynamics simulation or Monte Carlo sampling (see Chapter 14 and Chapter 15 for more on those techniques). Usually multiple, slightly different solutions are possible, which results in the typical bundle appearance of NMR structures, see Figure 2.6. Some of the variability in the models may show real structural fluctuations in solution, where other variation may indicate a lack of data or accuracy.

This procedure for structure determination by NMR works well for protein structures up to about 300 residues. Importantly, proteins need to be isotope-labelled with  $^{15}\text{N}$  and  $^{13}\text{C}$ , which works best if the protein can be expressed and purified from *E. coli*. Samples are typically solutions of the protein of interest, but can also be semi-solid samples of membrane proteins embedded in native membranes or suitable membrane-mimics. In both cases the crystallisation step is not required. Next to structure determination, important applications of NMR are the study of intrinsically disordered proteins, the study of protein dynamics and the study of protein-protein interactions (see the Panel “NMR-based modelling of protein complexes”)

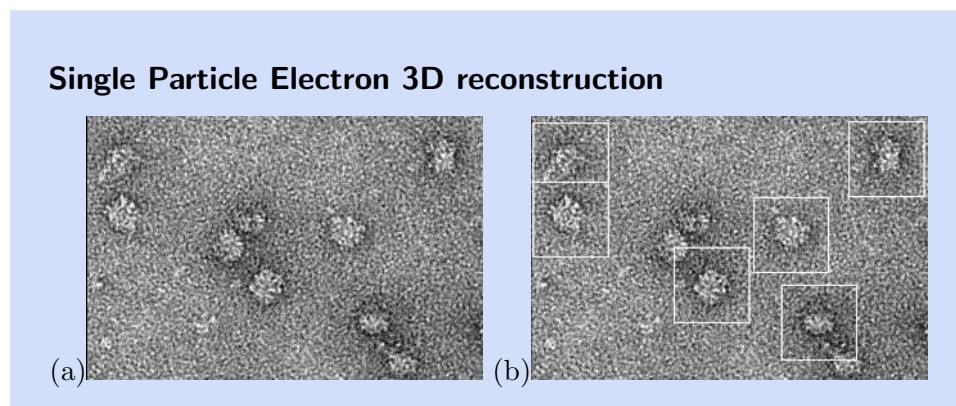


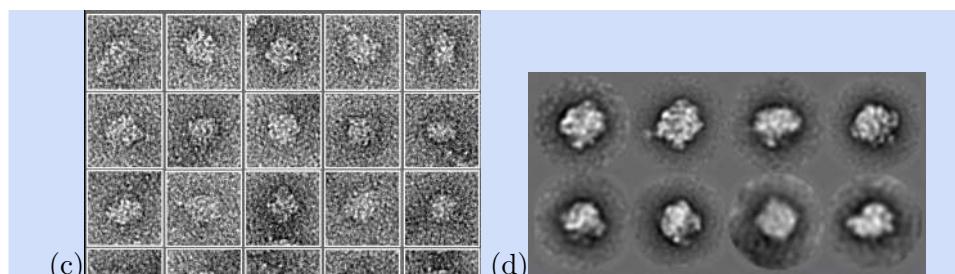
One of the main tasks in protein studies by NMR is the assignment of the signals for the backbone nuclei, in particular of the backbone amide NH groups. In this process, the transition frequency or chemical shift of a certain signal is assigned to a particular atom in the protein. This is achieved through the analysis of a dedicated set of so-called triple resonance NMR experiments. In these experiments, the chemical shifts of the amide NH and backbone  $^{13}\text{C}$  of each residue are measured. For each residue also the chemical shifts of the previous residue are measured. The resulting puzzle is to trace back the amino acid sequence by finding the matching connections from one residue to the next. This is illustrated in panel A for residue I13. Panel B shows a segment from a so-called NOESY experiment that is used to measure the distance between atoms. Each signal corresponds to an energy-transfer between nearby spins. The signal intensity can be used to derive the interatomic distance, where intense peaks mean that the atoms are close ( $< 3.5 \text{ \AA}$ ) and weak peaks that the atoms are within  $5 \text{ \AA}$ . When atoms are further than  $5$  to  $6 \text{ \AA}$  apart, the magnetic interaction between spins is too weak to result in transfer and no peak will be observed. The expanded plot in panel C shows NOE peaks observed for I13. Each peak corresponds to two chemical shifts, one along the horizontal and one along the vertical dimension, each corresponds in turn to a particular atom, as shown in the labels. The intense peak labelled “I13-HN V5-HN” is the result of transfer between amide protons of I13 and V5, residues that are far apart in the primary sequence. This means that these two are in close proximity in the 3D fold of the protein, as shown in panel D. Such information is particularly valuable when determining structures.



## 4 Cryo electron microscopy (cryo-EM)

Next to X-ray and NMR, cryo electron microscopy (cryo-EM) has become more and more popular in the past decade to obtain structural information on protein and biomolecular systems. With cryo-EM, as it is a microscopy technique, we can directly observe the objects of interest. The fundamental limitation of this technique is the wavelength of the electrons. In principle they can be tuned to any desired wavelength, however shorter wavelengths are progressively higher in energy. At some point the energy input into the protein will quickly destroy the sample. Atomic resolution is thus only obtainable in certain specific conditions, including cooling to extremely low temperatures. The method allows to image two-dimensional crystals that can be obtained from membrane-bound proteins, which are typically hard to crystallize into ‘normal’ 3D crystals for X-ray diffraction (see also Panel “Challenging structures”). Especially large-scale structures such as kinesin, tubulin,  $\beta$ -amyloid fibres, and virus capsids have become routine work for the cryo-electron microscopist. Unfortunately, unlike optical microscopy, EM does not work on ‘live’ samples. Because of the wavelength/energy problem mentioned above, it is hard to get sufficient contrast at high resolutions without damaging the sample. In cryo-EM, the sample is cooled with a cryogen, liquid nitrogen or helium, allowing the electron density of the protein molecules to be observed without the ‘blurring’ that is caused by atomic motions at room temperature. To obtain atomic-level resolution, cryo-EM density maps are often combined with X-ray structures in what can be thought of as an ‘X-ray jigsaw’ solution.





Electron microscopy allows single particles to be imaged, but to obtain more detail (higher resolution), shorter wavelengths are needed which means a higher energy beam. This greatly increases the damage done during data acquisition. To solve this, an average over many such particles can be used. Importantly, these particles do not need to be in a crystal, and that can be a huge advantage.

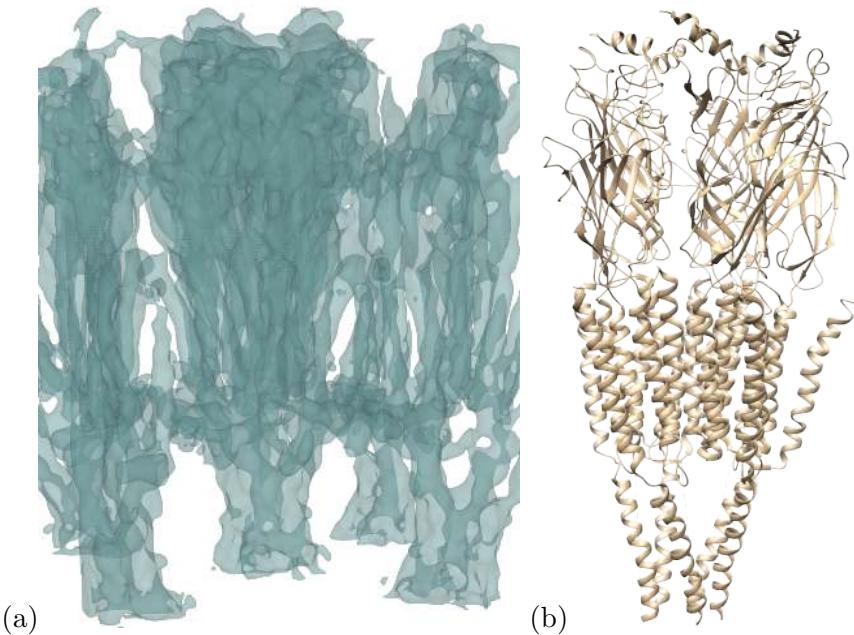
Images are taken of individual particles, such as virus capsid structures, lying on a grid (a). To minimize radiation damage the exposure is very brief and the contrast between the image and the background is minimal. In this method particles are selected (b) and sorted into different classes (c), which correspond (approximately) to the orientations in which the particle lies on the grid.

Each class (orientation) is then averaged to improve the signal to noise (d), which are then used to reconstruct a 3D image (not shown here). The method depends on the molecule falling on the grids in a fairly random way. Molecules that preferentially fall in only a few orientations need to be treated specially by tilting the grid.

Images courtesy of Peter Shen and Janet Iwasa (University of Utah) <https://cryoem101.org/chapter-1/>.

Figure 2.7 shows the membrane bound receptor for which the first density map was measured in 1993, at low resolution ( $9\text{\AA}$ ) by Unwin (1993). Those maps did not allow the modelling of atomic coordinates. Given the challenges of crystallography for these type of proteins, even this low resolution can give valuable insights. Later, the same protein was solved in the same lab at much improved resolution ( $4\text{\AA}$ ), where secondary structure elements may be clearly resolved, but not all atomic details become available (Unwin, 2005; Unwin and Fujiyoshi, 2012).

Methods for single molecule cryo-EM structure determination have been developed since the early nineteen nineties, e.g. on a skeletal muscle calcium channel (Radermacher *et al.*, 1994). See Panel “Single Particle Electron 3D reconstruction” for a more detailed description of an application to viral envelope structures. Many methodological advances have since pushed the achievable resolution to near atomic (Van Heel *et al.*, 2000; Frank, 2002; Li *et al.*, 2013), and applications of large macromolecular machines, e.g. Baumeister and Steven (2000). Some landmarks include the calcium channel already mentioned (Radermacher *et al.*, 1994), the E. coli ribosome (Mal-



*Figure 2.7: The first low resolution density map of a membrane bound receptor, the nicotinic acetylcholine receptor, was created at 9 Å (Unwin, 1993). Later, greatly improved resolution of the cryo-EM experiments yielded a maps at 4 Å, allowing atomic models to be constructed (Unwin, 2005; Unwin and Fujiyoshi, 2012). (a) Overview of the density map PDB:4aq9. (c) Full details of the protein structure PDB:2bg9. Images generated by LiteMol (Sehnal et al., 2017).*

hotra *et al.*, 1998), icosahedral viruses (Baker *et al.*, 1999), and the plant photosystem II complex (Barber *et al.*, 2000). A particularly impressive example of ‘cryo-EM reconstruction’, or X-ray/EM jigsaw solution of the structure of a large complex is the structure of the tail of bacteriophage T4, shown in Figure 2.8 (Leiman *et al.*, 2010). Comparing the structure in the relaxed and contracted state (not shown here) of the tail helps understanding the infectious mechanism.

However, these are just some examples on the applications of cryo-EM. The field has been developing rapidly over the past decade and the new experimental advances allow to address more and more challenging structures and molecular targets. If you are interested to know more about the possibilities, you should have a look at the review by Cheng (2018) on the rise and evolution of cryo-EM. The Panel “Type III Secretion System” shows one of the latest developments.

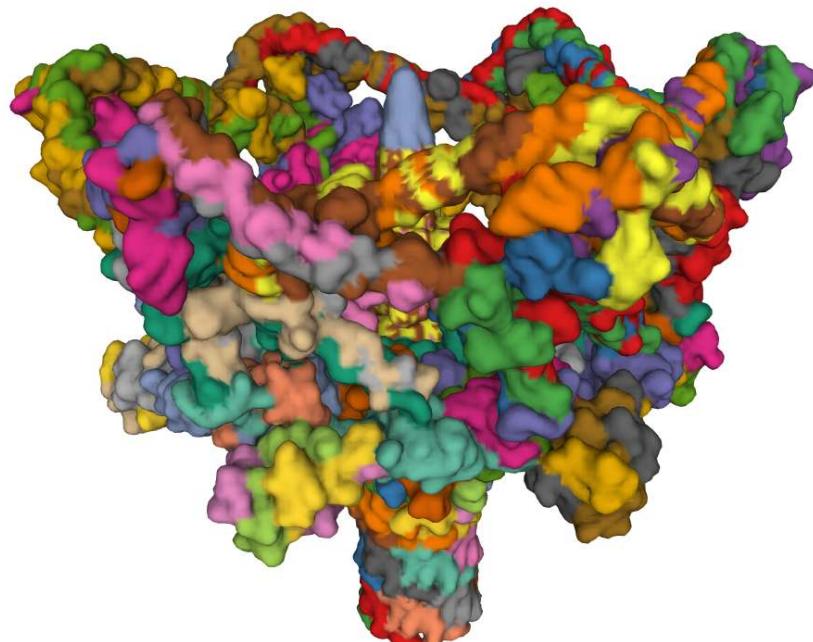
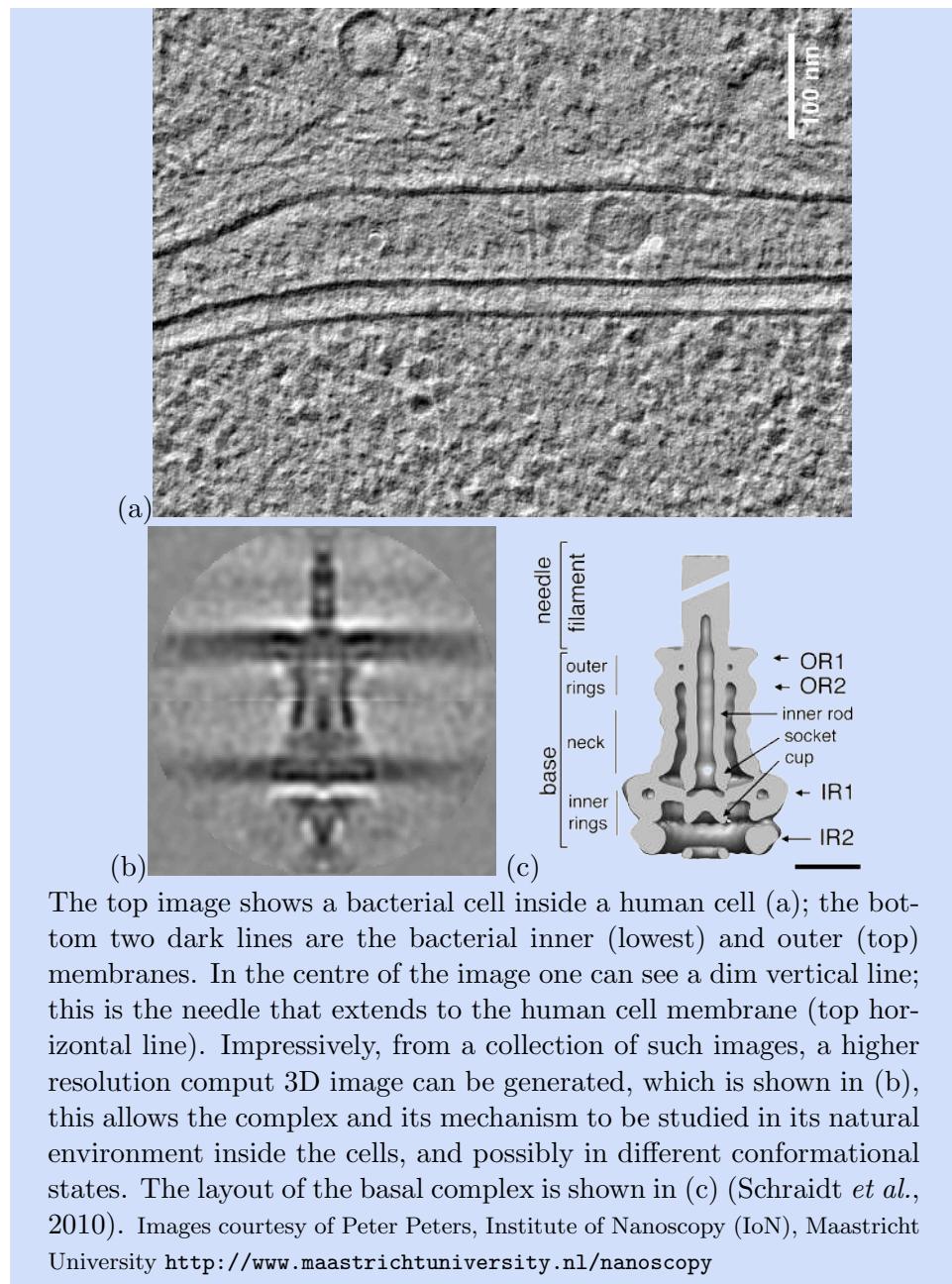


Figure 2.8: The complex is the tail of a bacteriophage (virus that infects bacteria); the tail contracts to insert the DNA into the host bacterium. It was imaged in EM at 17 Å resolution by Leiman et al. (2010); the image shown here is of the attachment baseplate and tube, also with cryo-EM at 4.1 Å (Taylor et al., 2016). The whole complex measures 1200 Å (120 nm) in length and has an atomic weight of 20 million Daltons (one amino acid on average is about 134 Dalton). There are about 20 different proteins present in the complex, most in (very many) multiple copies. Image generated from PDB:5IV5 using the PDB viewer (Berman et al., 2000).

### Type III Secretion System

The Type III Secretion System is one of the mechanisms by which pathogens infect human cells. It is anchored with a ‘basal plate’ in the bacterial inner and outer membrane, and extends a ‘needle’ filament towards the target (human) cell.



## 5 Other structure determination methods

Besides the “high-resolution” structure determination methods we have seen so far, there are many more techniques that allow us to obtain structural information of proteins or relevant biological systems. These are often used in combination with molecular simulations as they do not provide the same resolution as X-ray, NMR, and EM. Small-angle X-ray scattering (SAXS)

uses the scattering patterns of X-ray radiation to obtain information on the outside shape of molecules in solution. Several spectroscopic methods such as circular dichroism and infrared spectroscopy can be extremely useful to obtain qualitative structural insight. For example, one may follow conformational changes of proteins upon ligand binding or under different varying conditions, e.g. by increasing/decreasing the pH or by adding salts.

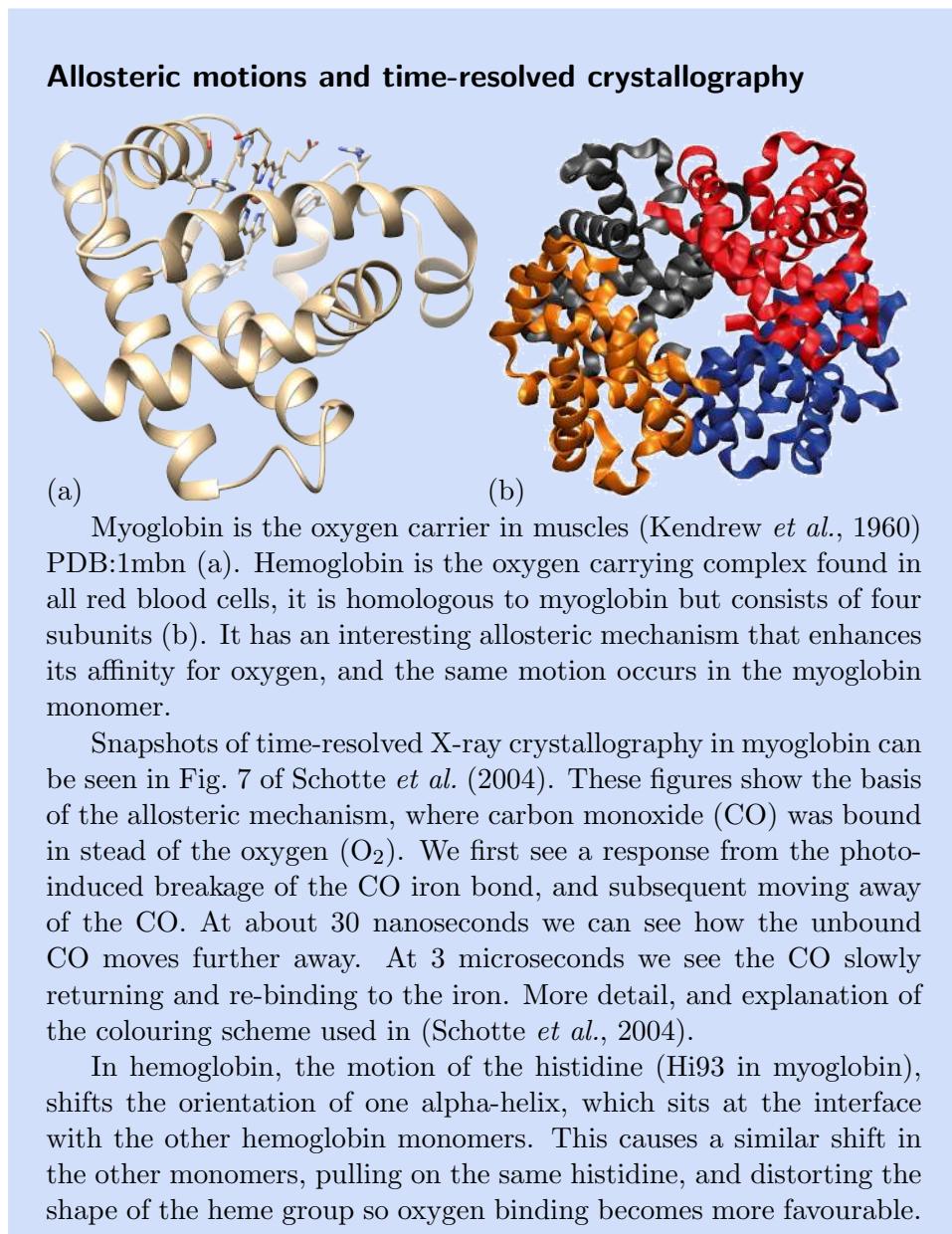
## 6 Dynamics and flexibility

In spite of the rigid-sounding name, protein crystal structures are actually quite dynamic. This also means proteins can retain some (or all) of their biological function, like ligand binding or enzyme turnover. For example, most enzymes are still active in the crystal form, although sometimes much slower than in solution. In a simulation of the native dynamics of a  $\beta$ -barrel fatty acid binding protein, we can see everything moves all the time but the overall structure of the protein remains the same folded  $\beta$ -barrel. Also the bound fatty acid molecule remains in its place, even though it wiggles back and forth a lot. Some water molecules that start inside the protein eventually ‘escape’. But also some that start outside the protein, in the ‘bulk’ water, can find their way into the interior during the course of the simulation.

Some enzymes, however, require large scale motions of the protein structure. Large-scale motions are not possible due to the tight packing of the protein molecules in the constraints of the crystal lattice. In some cases such a motion can be induced for example in enzymatic activity or activation of a receptor by adding the ligand. When a substrate is added to the crystal, and binds to the enzyme or receptor, this binding may then cause the crystal physically break. This also shows how strong some molecular motions can be.

Even though protein crystal structures give us initially a rigid picture of the protein, the data does contain some information that may also be interpreted in a dynamic sense. The B-factors, or temperature factors as they are also sometimes called, indicate how well the local electron density fits the atoms placed in it. A low B-factor means the fit is very good, a high B-factor means it is poor. There are two contributions to this goodness of fit. One is variations between the molecules in the crystal. The refinement process of X-ray crystallography assumes all molecules to have identical conformation and orientation throughout the crystal. For proteins this does not strictly hold, and any heterogeneity in the crystal arrangement will result in spreading of the electron density, and hence the diffraction signal. The second contribution is dynamics. Even if, in principle, all protein molecules are identically oriented, there will still be motion going on in them. This leads to a ‘smearing’ of the densities observed (similar to the blurring effect

you get when taking a picture of a fast moving object), and hence higher B-factors. This also is the reason that crystal structures are often recorded at low temperature; this slows down the molecular motions and diminishes the ‘blurring’ effect. From high-resolution low-temperature structures we now know that, typically, the effect of variation or heterogeneity between the protein molecules in the crystal is minor. So, in general it is relatively safe to interpret high B-factors as indicating high mobility in the structure.



For NMR there is a one-to-one relation between signals measured and particular atoms, bonds or angles. The width of these peaks can vary,

and this is usually entirely due to (local) structure and dynamics of the molecule, i.e. how the atoms are oriented and how much they move. In case of particularly dynamic molecules, one should realize that some atoms may alternatively be close to two different parts of the molecule. For example, a dynamic protein loop that has two conformations. Both distances will be short enough to yield a measurable NOE intensity; but no structure exists that can satisfy both short distances at the same time. For this, and other reasons, NMR experimental data is typically used to generate an ensemble of solution structures instead of a single one as is done for X-ray. One may think of this as reflecting the innate dynamics of the protein. However, be aware that sparsity of data may result in an under-defined structure, which will also yield larger variations in the ensemble generated, but this does not arise from dynamics.

X-ray crystallography and cryo-EM provide very detailed, but intrinsically static pictures of protein structures. The dynamics of protein structures are poorly represented by this static view. NMR and other spectroscopic techniques help remedy this, and these are often used in combination with molecular simulations, which we will cover in Chapter 12 and subsequent chapters on thermodynamics and simulations.

## 7 Key points

- 3D coordinates of the protein (PDB) are **not the primary experimental data**.
- For X-ray crystallography:
  - Electron density maps are also not the primary data
  - **Diffraction patterns** are the primary data
- For NMR:
  - Distances and angles are also not the primary data
  - **Spectra and intensities** are the primary data
- Everything else is (at least partly) **based on modelling**
- For the other techniques mentioned, this dependence on modelling is even stronger
- Proteins are not static, they are **dynamic**

## 8 Recommended further reading

- Branden and Tooze (1998) – “Introduction to protein structure” for a broader general introduction to protein structure and structure determination.
- Atkins and De Paula (2014) – “Physical Chemistry” for a more in-depth on structure determination of biological macromolecules, and other experimental approaches to elucidate functional, structural and

- chemical properties.
- Giacovazzo *et al.* (2011) – “Fundamentals of Crystallography” for an advanced account of modern crystallography, including the mathematical details of different approaches and techniques, highly recommended by practising crystallographers.
  - Shen *et al.* (2018) <https://cryoem101.org/chapter-1/> – an accessible introduction to the experimental and data processing basics of cryo-EM.
  - Teilmann *et al.* (2017) “(S)Pinning down protein interactions by NMR”
  - Kwan *et al.* (2011) “Macromolecular NMR for the non-spectroscopist”

## Author contributions

Wrote the text:	HM, BS, HI, KAF
Created figures:	HM, HI, JG, KW, KAF
Review of current literature:	HM, HI, KW, KAF
Critical proofreading:	BS, SA
Non-expert feedback:	JG, KW
Editorial responsibility:	HM, SA, KAF

The authors thank Arriën Symon Rauh  for creating Figure 2.3.

## References

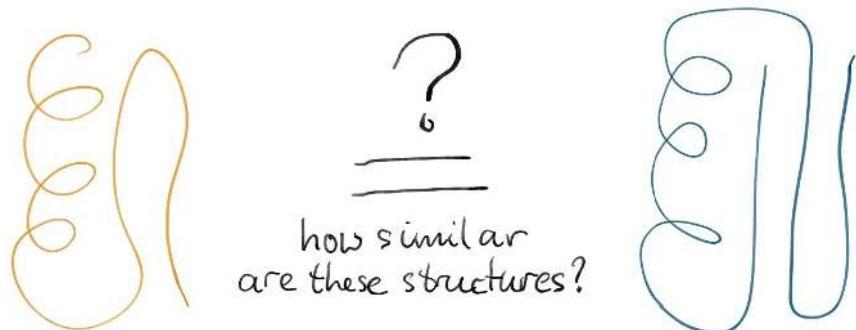
- Atkins, P.W. and De Paula, J. (2014). *Atkins' Physical chemistry*. Oxford University Press.
- Baker, T.S., Olson, N.H. and Fuller, S.D. (1999). Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and molecular biology reviews : MMBR*, **63**(4), 862–922.
- Barber, J., Nield, J., Orlova, E.V., Morris, E.P. et al (2000). 3D map of the plant photosystem II supercomplex obtained by cryoelectron microscopy and single particle analysis. *Nature Structural Biology*, **7**(1), 44–47.
- Baumeister, W. and Steven, A.C. (2000). Macromolecular electron microscopy in the era of structural genomics. *Trends in Biochemical Sciences*, **25**(12), 624–631.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. et al (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.
- Branden, C. and Tooze, J. (1998). *Introduction to protein structure*. garland publishing, New York.
- Cheng, Y. (2018). Single-particle cryo-EM-How did it get here and where will it go. *Science*, **361**(6405), 876–880.
- Corbeski, I., Dolinar, K., Wienk, H., Boelens, R. and van Ingen, H. (2018). DNA repair factor APLF acts as a H2A-H2B histone chaperone through binding its DNA interaction surface. *Nucleic Acids Research*, **46**(14), 7138–7152.
- Cowtan, K. (2003). Phase Problem in X-ray Crystallography, and Its Solution. In *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester.
- Earl, L.A., Falconieri, V., Milne, J.L. and Subramaniam, S. (2017). Cryo-EM: beyond the microscope. *Current Opinion in Structural Biology*, **46**, 71–78.
- Egelman, E.H. (2017). Cryo-EM of bacterial pili and archaeal flagellar filaments. *Current Opinion in Structural Biology*, **46**, 31–37.
- Ellis, R.J. (2001). Macromolecular crowding: Obvious but underappreciated.
- Feenstra, K., Peter, C., Scheek, R., Van Gunsteren, W. and Mark, A. (2002). A comparison of methods for calculating NMR cross-relaxation rates (NOESY and ROESY intensities) in small peptides. *Journal of Biomolecular NMR*, **23**(3).
- Frank, J. (2002). Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy. *Annual Review of Biophysics and Biomolecular Structure*, **31**(1), 303–319.

- Giacovazzo, C., Monaco, H.L., Artioli, G., Viterbo, D. et al (2011). *Fundamentals of Crystallography*. Oxford University Press, 3 edition.
- Herzik, M.A. (2020). Cryo-electron microscopy reaches atomic resolution.
- Jiang, W. and Tang, L. (2017). Atomic cryo-EM structures of viruses. *Current Opinion in Structural Biology*, **46**, 122–129.
- Kato, H., van Ingen, H., Zhou, B.R., Feng, H. et al (2011). Architecture of the high mobility group nucleosomal protein 2-nucleosome complex as revealed by methyl-based NMR. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(30), 12283–8.
- Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G. et al (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, **185**(4711), 422–427.
- Kwan, A.H., Mobli, M., Gooley, P.R., King, G.F. and Mackay, J.P. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal*, **278**(5), 687–703.
- Leiman, P.G., Arisaka, F., van Raaij, M.J., Kostyuchenko, V.A. et al (2010). Morphogenesis of the T4 tail and tail fibers. *Virology Journal*, **7**(1), 355.
- Li, X., Mooney, P., Zheng, S., Booth, C.R. et al (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods*, **10**(6), 584–590.
- Luby-Phelps, K. (1999). Cytoarchitecture and Physical Properties of Cytoplasm: Volume, Viscosity, Diffusion, Intracellular Surface Area. In *International Review of Cytology*, volume 192, pages 189–221. Academic Press.
- Malhotra, A., Penczek, P., Agrawal, R.K., Gabashvili, I.S. et al (1998). Escherichia coli 70 S ribosome at 15 Å resolution by cryo-electron microscopy: localization of fmet-tRNAAfMet and fitting of L1 protein. *Journal of Molecular Biology*, **280**(1), 103–116.
- Nogales, E., Wolf, S.G. and Downing, K.H. (1998). Structure of the  $\alpha\beta$  tubulin dimer by electron crystallography. *Nature*, **391**(6663), 199–203.
- Radermacher, M., Rao, V., Grassucci, R., Frank, J. et al (1994). Cryo-electron microscopy and three-dimensional reconstruction of the calcium release channel/ryanodine receptor from skeletal muscle. *The Journal of cell biology*, **127**(2), 411–23.
- Read, R.J. (1997). Model phases: Probabilities and bias.
- Schotte, F., Soman, J., Olson, J.S., Wulff, M. and Anfinrud, P.A. (2004). Picosecond time-resolved X-ray crystallography: Probing protein function in real time. *Journal of Structural Biology*, **147**(3), 235–246.
- Schraadt, O., Lefebre, M.D., Brunner, M.J., Schmied, W.H. et al (2010). Topology and Organization of the *Salmonella typhimurium* Type III Secretion Needle Complex Components. *PLoS Pathogens*, **6**(4), e1000824.
- Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M. et al (2001). A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Science*, **10**(4), 677–688.
- Sehnal, D., Deshpande, M., Vareková, R.S., Mir, S. et al (2017). LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nature Methods*, **14**(12), 1121–1122.
- Shen, P., Iwasa, J., Thuesen, A., Wambaugh, M. and Stewart, M. (2018). *CryoEM 101*. University of Utah, Utah.
- Taylor, N.M.I., Prokhorov, N.S., Guerrero-Ferreira, R.C., Shneider, M.M. et al (2016). Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature*, **533**(7603), 346–352.
- Teilum, K., Kunze, M.B.A., Erlendsson, S. and Kragelund, B.B. (2017). (S)Pinning down protein interactions by NMR. *Protein Science*, **26**(3), 436–451.
- Unwin, N. (1993). Nicotinic acetylcholine receptor at 9 Å resolution. *Journal of molecular biology*, **229**(4), 1101–24.
- Unwin, N. (2005). Refined Structure of the Nicotinic Acetylcholine Receptor at 4 Å Resolution. *Journal of Molecular Biology*, **346**(4), 967–989.
- Unwin, N. and Fujiyoshi, Y. (2012). Gating Movement of Acetylcholine Receptor Caught by Plunge-Freezing. *Journal of Molecular Biology*, **422**(5), 617–634.
- Van Heel, M., Brent, G., Matadeen, R., Orlova, E.V. et al (2000). Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, **33**(4), 307–369.
- Ward, A.B. and Wilson, I.A. (2017). The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunological Reviews*, **275**(1), 21–32.
- Yang, J., Yan, W., Yu, Y., Wang, Y. et al (2018). The compound millepachine and its derivatives inhibit tubulin polymerization by irreversibly binding to the colchicine-binding site in  $\beta$ -tubulin. *Journal of Biological Chemistry*, **293**(24), 9461–9472.
- Yu, X., Jin, L. and Zhou, Z.H. (2008). 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature*, **453**(7193), 415–419.
- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**(1), 40.
- Zhang, Y., Sun, B., Feng, D., Hu, H. et al (2017). Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature*, **546**(7657), 248–253.

# Chapter 3

## Structure Alignment

Olga Ivanova  Jose Gavaldá-García  Dea Gogishvili   
Isabel Houtkamp  Robbin Bouwmeester   
K. Anton Feenstra\*  Sanne Abeln\* 



\* editorial responsibility

## 1 Comparing protein structures

The Protein DataBank (PDB) contains a wealth of structural information (Berman *et al.*, 2000). In order to investigate the similarity between different proteins in this database, one can compare the primary sequence through pairwise alignment and calculate the sequence identity (or similarity) over the two sequences. This strategy will work particularly well if the proteins you want to compare are close homologs. Knowledge of sequence similarity is widely used in for example the Pfam database (Finn *et al.* 2014, see Chapter 4): Pfam cluster similar proteins into families based on the sequence of the protein. However, in this chapter we will explain that a structural comparison through structural alignment will give you much more valuable information, that allows you to investigate similarities between proteins that cannot be discovered by comparing the sequences alone.

Furthermore, we will discuss the challenges in understanding how similar two structures are based on structural information alone (i.e. the atomic coordinates, see also Figure 3.1); this means we do not use any sequence information on the two proteins. Using solely the structure for comparison poses a difficult computational problem due to the many possible ways to align the structure. However, structural comparison is generally considered to be more reliable and evolutionary accurate than a comparison based on sequence similarity.

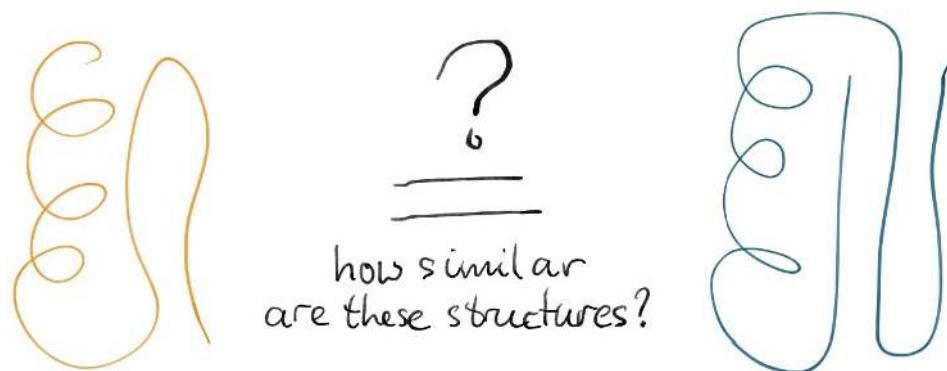


Figure 3.1: Structural alignment deals with the problem of determining how similar two structures are – based on the atomic coordinates alone (no sequence information).

### 1.1 Structure is more conserved than sequence

One of the reasons why structural comparison is valuable, is that structure is generally more conserved than sequence. In evolution, selection pressure acts on the function of a protein. If unfolded, the majority of proteins will

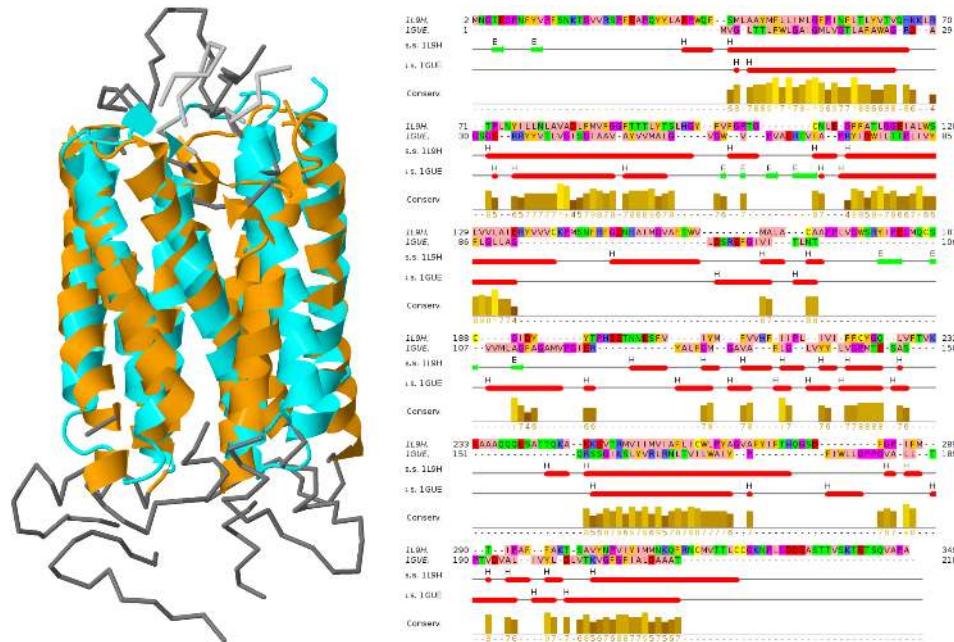


Figure 3.2: Structure is more conserved than sequence. Left: The output of a structural alignment program, Combinatorial Extension (CE). These two proteins (orange and cyan) have, as you see, a similar structure. They are both rhodopsins, and they have a similar function (light detection). However, their **sequence identity** (right) is **less than 5%**. This is below the similarity you would expect from two random sequences. Note that one would not be able to align these proteins using sequence identity alone. One can see that the positioning of the helices are very well conserved between the two structures, but that there is **much more variability** in the **loops** (both in structure and in length). The two proteins are bovine rhodopsin (PDB:1L9H, in orange) and sensory rhodopsin (PDB:1GUE, in cyan). Website at <http://www.rcsb.org/pdb/workbench/workbench.do>

lose their function. Moreover, it may be dangerous for the cell if a protein is partially folded or misfolded because hydrophobic residues would be exposed leading to the protein aggregation. This effectively puts an evolutionary pressure on the ability of a protein to fold into a functional state and on the folding process itself (Tartaglia and Vendruscolo, 2009).

Structure being more conserved than sequence has several important consequences for methods in structural bioinformatics. This importance is for example illustrated by the large number of structure prediction methods that use these principles as a foundation, and also by the design principles of structural classification databases. Importantly, distant evolutionary relations between two proteins are more easily observed by comparing structures rather than sequences.

An example of structure conservation without sequence conservation is demonstrated in Figure 3.2, which shows the typical output of a structural

alignment program. Two homologous proteins, i.e., with common ancestor, are aligned with the structural alignment program Combinatorial Extension (CE, by Shindyalov and Bourne, 1998, on the left) and superimposed on top of each other (on the right). Note that the sequence identity between the proteins is very low; so low in fact, that two *random* sequences aligned by sequence similarity would give a similar score. However, from the structural comparison in Figure 3.2, it is clear that the structures are much alike. In this case, only the structural alignment allows the correct identification of common ancestry for these two proteins.

## 2 Structural superposition

The simplest manner to compare two different configurations of a protein is by generating a *structural superposition* of the two structures. In this section we will focus on the superposition problem, i.e. how to "overlay" two protein structures in 3D space (see Figure 3.3). Keep in mind that if we want to superimpose two structures, we need a mapping between the residues of each structure ( i.e. an alignment - but not based on the sequence). In the next section, *Structural alignment*, we will see how we can compare structures if we do not have an alignment.

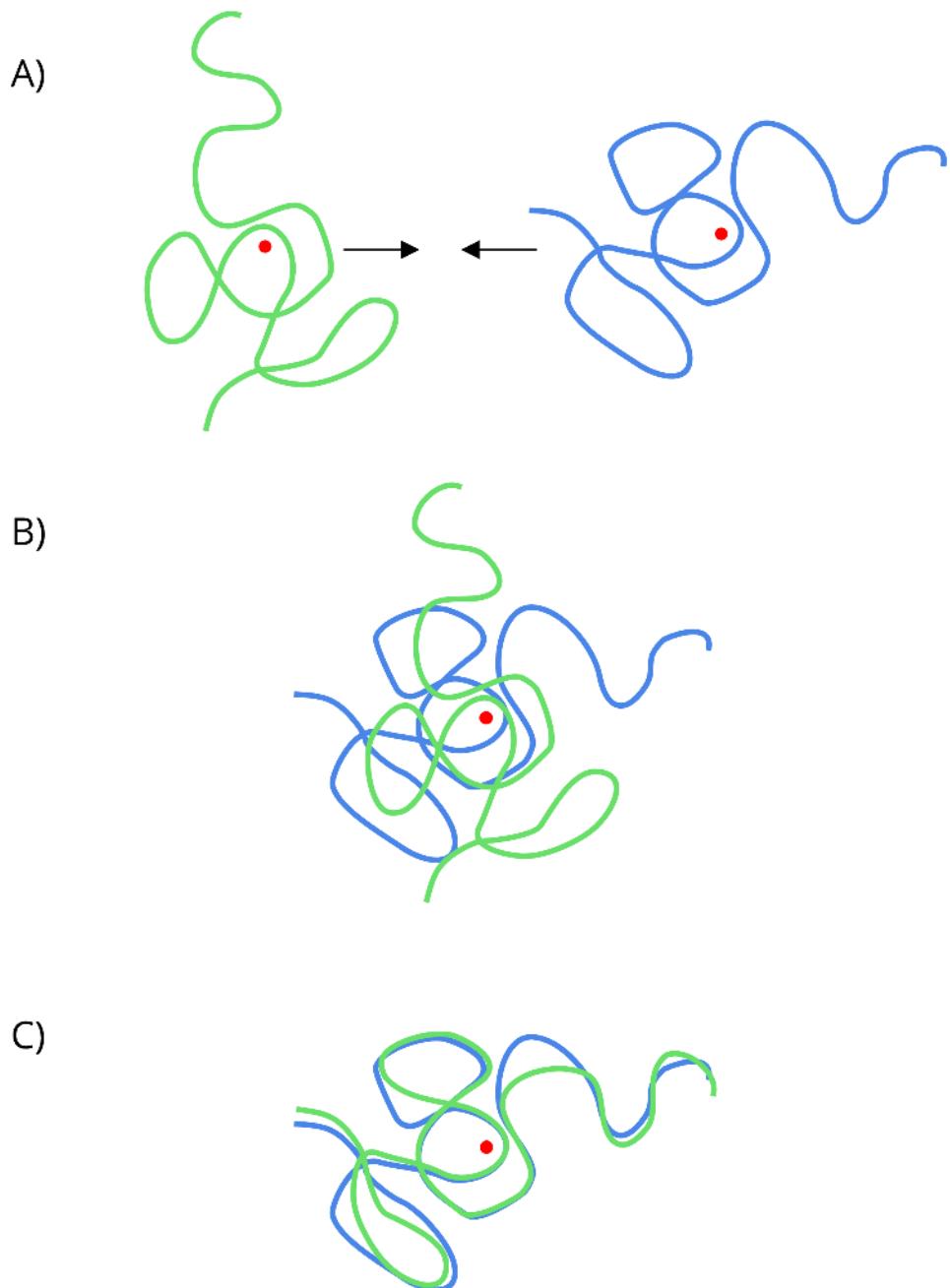
### 2.1 PDB coordinates as a structure representation

In Chapter 4 we will see how a protein structure may be represented and stored in a PDB record in more detail. For now it is enough to know that the structure is recorded in the PDB file by assigning an x, y and z coordinate for each atom (Equation 1). Hence each atom can be represented by a positional vector  $\mathbf{p} \in \mathbb{R}^3$  in three dimensional space, such that  $\mathbf{p} = (p_x, p_y, p_z)$ .

Note that a protein structure can be rotated as a rigid body, without changing the interatomic distances. It is up to the experimentalist who created the PDB file to record what the frame of reference is, as any arbitrary rotation and translation may be given. Comparing the rotation and translation of two different protein structures is a key part of the structural superpositioning. Before we go into any more detail, we will first need a score to evaluate the similarity of two protein structures.

### 2.2 A score for comparing protein structures – RMSD

The root mean square deviation (RMSD), is one of the simplest measures to score the similarity of two protein structures. RMSD calculates the squared difference between two sets of atoms. In practice, a single representative atom per residue is chosen, i.e., C-alpha or C-beta atoms. The RMSD of



**Figure 3.3: Superimposing two protein structures.** The lines represent proteins that need to be aligned and the red dots indicate their calculated center of mass. The superposition problem is explained by figures A), where we need to find the "best" overlay in which the the two structures can be compared. C) shows a solution to the problem. The process of superimposing two structures: A) The centers of mass for the proteins are calculated using Equation 2. B) The centers of mass of both proteins are put in the same spatial coordinate. C) The protein structures are superimposed.

two structures  $V$  and  $W$ , given the residues mapping  $M$ , is given by:

$$\begin{aligned} \text{RMSD}(V, W) &= \sqrt{\frac{1}{n} \sum_{i \in M} \|\mathbf{v}_i - \mathbf{w}_i\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i \in M} (v_{i,x} - w_{i,x})^2 + (v_{i,y} - w_{i,y})^2 + (v_{i,z} - w_{i,z})^2} \end{aligned} \quad (1)$$

Here we calculate the squared distance between atoms  $\mathbf{v}_i$  of structure  $V$  and atoms  $\mathbf{w}_i$  of structure  $W$ . Note the sum over  $i$  runs over all matched pairs  $M$  of atoms in the alignment of  $V$  and  $W$ ; the total number of matched (aligned) residues is  $n$ . This means we need to know how the residues correspond in the two different structures in order to calculate this score, i.e. we need to know which  $w_i$  we need to subtract from which  $v_i$ .

### 2.3 Structural superposition and RMSD

The method of **finding an optimal rotation and translation** is called **structural superposition**. An RMSD score between the two sets of atoms is minimised to determine the optimal rotation. Hence, a superposition algorithm provides the rotation that yields the best RMSD fit. Importantly, structural superposition cannot identify the alignment or residues between two protein structures; we will come back to this at the end of this section.

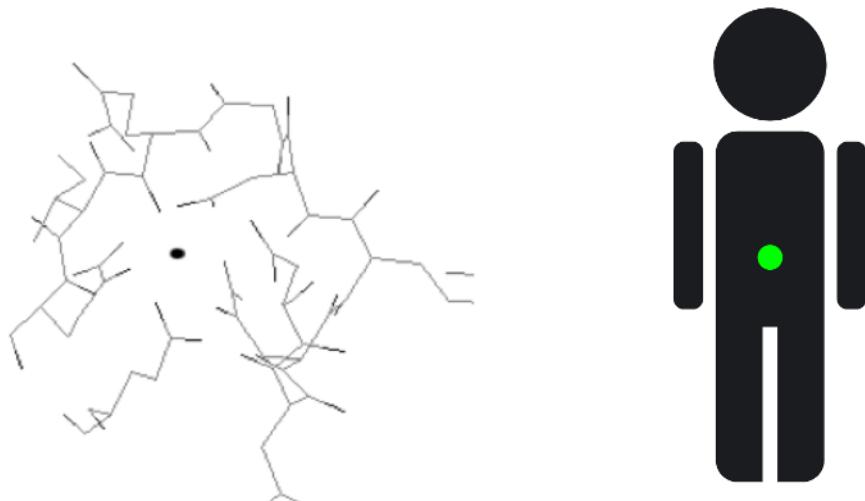
Note that the RMSD score only gives a measure of the dissimilarity between two protein structures, if they have first been superimposed; without superposition, one would merely measure how far apart the structures are in the given coordinate frame(s), as explained in Figure 3.3. More formally, a measure for structural dissimilarity should be described as the ‘RMSD **after least-squares-fitting**’, instead of simply using ‘RMSD’, but this is typically omitted in research papers (and even most text books).

Figure 3.3 also illustrates how one of the two structures needs to be translated and rotated in order to obtain a minimal RMSD between corresponding atom sets. The optimal translation can be found by simply ensuring that the **two centers of mass** will fall on top of each other. The coordinates for the center of mass of each structure,  $\mathbf{R} \in \mathbb{R}^3$ , may be found using Equation 2.

$$\mathbf{R} = \frac{\sum_{i=0}^{i=N} m_i \mathbf{r}_i}{\sum_{i=0}^{i=N} m_i} \quad (2)$$

This is the mass-weighted average of the atomic positions  $\mathbf{r}_i$ ; here  $N$  is the total number of atoms, and  $m_i$  is the mass of each atom  $i$ . Figure 3.4 exemplifies the meaning of the center of mass.

Since we are typically using a single type of atoms (i.e. C-alpha), the masses ( $m_i$ ) in Equation 2 cancel out. Now we can calculate the directional

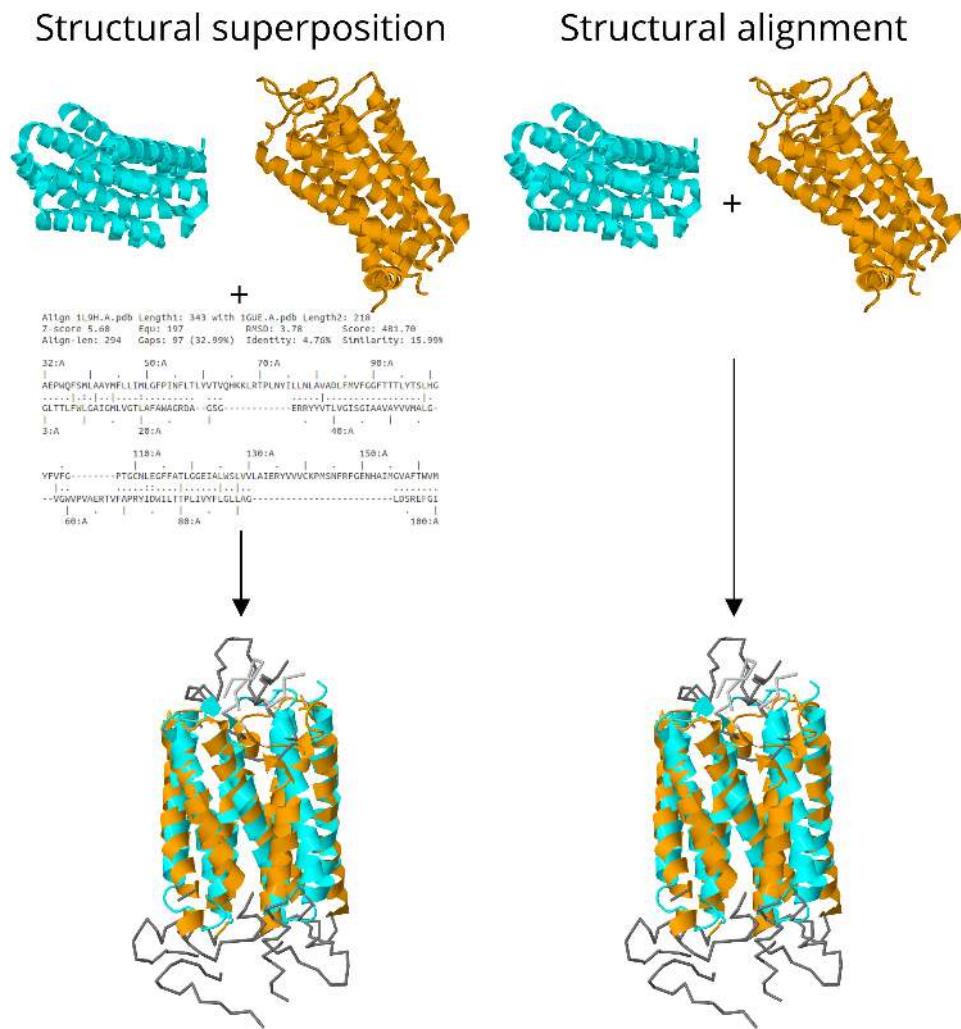


**Figure 3.4: Center of mass in molecules and humans.** The center of mass is the spatial position determined by the average of all the atoms in a system. The center of mass of a molecule (left, black dot) is calculated the same way that a center of mass for a human (right, green dot) would be calculated. Without going into further detail, the center of mass is placed in the same location as the center of gravity for most objects on Earth, including our bodies.

difference between the two centers of mass to calculate the required translation. Note that a translation in three-dimensional space means that one needs to add or subtract an identical vector, in this case corresponding to the center of mass, from the coordinates of all atoms.

Once the centers of masses have been superimposed, there exists a single rotation that will minimize the RMSD. This rotation can be found exactly; in practice, you need to transform three-dimensional coordinates, to 4D quaternions (Kearsley, 1989), and solve an eigenvalue problem using the Jacobi algorithm. We will not go through this method in detail, as it provides little extra understanding of the practical problem. Computationally, it is not very expensive to solve this problem: the exact solution can be found in polynomial time.

Structural super-positioning can be used to compare protein structures for which the mapping of residues is already known. For example, if we want to compare snapshots in a simulation to the native protein structure, we want to actually compare a protein with itself; hence the mapping of the residues is trivial. Similarly, if we want to compare a model from a structure prediction method, with the experimentally determined structure, we would also use structural super-positioning directly. However, if we want to compare (potentially) homologous proteins, the mapping of residues may not be known, and we need structural alignment to generate such a mapping.



**Figure 3.5: Structural superposition versus structural alignment.** Left: Structural superposition requires the structures of the proteins and an alignment of the residues as input. Note that if the two structures originate from one protein (and thus have the same sequence) the alignment of the residues is trivial. The superposition method works by minimizing the RMSD, for which we need a mapping (alignment) between the residues. The structural superposition will return two structures in the same frame of reference, such that the RMSD may be calculated. Right: Structural alignment takes the protein structures as its only input. The method will try to match similar substructures between the proteins. It will return an alignment, as well as a score for the (dis)similarity of two protein structures.

### 3 Structural Alignment

Structural alignment deals with the problem of generating a mapping between residues, or an alignment, based solely on the structure of two pro-

teins. Note that this mapping of residues is not a trivial task. One may be tempted here to use the sequence as a reference. However, as illustrated by Figure 3.2, this would be a poor choice, as the **sequence similarity may be very low for similar structures**. Hence we would like to use structure alone to map the corresponding residues (e.g. C-alpha or C-beta atoms) of two structures to be superimposed. The problem of **finding the optimal match or alignment between the residues of two protein structures** is called **structural alignment**.

A typical output of a structural alignment method will therefore provide an **alignment of the residues**, based on the structure of the backbone alone. Typically, also a **superposition of the two structures** will be given as an output, as well as a structural alignment score. Please also consider Figure 3.5 for an explanation of the difference between structural alignment and structural superposition.

The problem of structural alignment is computationally much more difficult to solve than the superposition problem. In fact, it has been suggested to belong to a class of computational problems that is called **NP-hard** (Hasegawa and Holm, 2009). In practice, this means that the time required for the computation grows **exponentially** with the size of the input (protein lengths in this case). A consequence is that for many real sized proteins, of about 200-300 residues, it would be very difficult to find the optimal solution, since it is computationally too expensive to search through all possible structural alignments. However, some methods can provide exact solutions for real size proteins (Wohlers *et al.*, 2012).

### Computational complexity of structural alignment versus sequence alignment

Note that the pairwise structural alignment problem is very different in terms of computational complexity from the pairwise sequence alignment problem. For the latter problem we can find an optimal solution, by filling in the dynamic programming matrix. Using dynamic programming, the pairwise sequence alignment problem can be solved exactly in  $O(n \cdot m)$  time for aligning one sequence of length  $n$  with one of length  $m$ , since  $n \times m$  operations need to be calculated to fill in the dynamic programming matrix.

The difference between these two alignment problems is that there are no natural local bounds in structural alignment. Imagine that we have an already aligned region of the sequence. For sequence alignment, if we change the alignment outside this region, no change of scoring will take place inside the already aligned region. However, with structural alignment, the introduction of a gap outside the aligned re-

gion may also affect the score of the already aligned region, due to residues outside the already aligned region that are close in 3D space (but not in sequence space).

### 3.1 The three key components of structural alignment

Structural alignment methods all contain three crucial parts: a suitable *structural representation* for the proteins, a method to *optimize* a similarity measure and a (statistical) *score for protein structure similarity* (Martin-Renom *et al.*, 2009). Plenty of different representations, methods and measures have been used to tackle the problem of structural alignment; here, we will give a brief overview of some of the most important strategies.

### 3.2 Structure representation and contact maps

In order to find a good structural alignment, methods typically encode the structure of both proteins into a representation that can be more easily compared between residues of the two proteins. Note that it is important that this representation is invariant to the frame of reference, otherwise it difficult to compare substructures; hence, the c-alpha coordinates are not suitable.

An example of a suitable representation is the *contact map*, or contact matrix. Note that such a contact map is defined for a single protein structure, not for a pair of structures.

A contact can be defined as two residues that are close in three-dimensional space. We can now create a contact map, by considering for each pair of residues in a protein if they make a contact, see for example Figure 3.6.

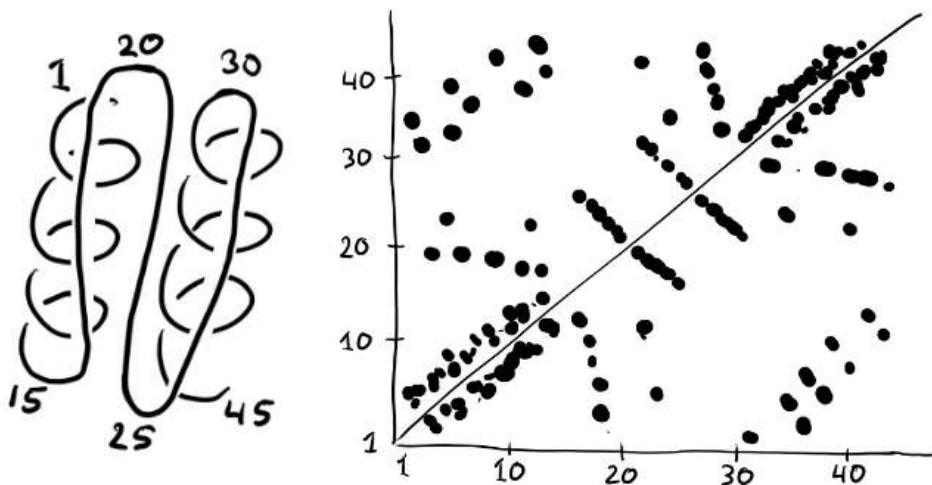
A contact matrix,  $C_{i,j}$ , can be mathematically formalized as follows.

$$C_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ make contact} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Here all possible residues pairs  $(i,j)$  are considered in a single protein structure. A cutoff of around 7 Å is typically used to define which residues are in contact.

From such a map, secondary structure elements may be recognized: for example, alpha-helices can be identified as a diagonal line adjacent to the diagonal. This is because in a helix there is a typical hydrogen bonding pattern of residue  $i$  and  $i + 4$  along the sequence, see also Figure 3.6.

We can compare the local three-dimensional surroundings of a residue by considering the contacts. If residues in two different protein structures have (locally) similar contact maps, they have similar substructures. Using contact maps, the structural alignment problem can be redefined as the problem to find an optimal alignment between two contact maps.



**Figure 3.6: Contact map of a single protein.** Here it can be observed which atoms are “in contact” (closer than a set distance) in a protein structure. From this figure it can be easily observed the the alpha helices close to the main diagonal, and the interaction between residues due to the tertiary structure.

Similarly, a *distance map* or *distance matrix* may be defined for a protein structure. In this case the matrix elements represent the *distance between residues*, instead of the boolean value indicating if there is a contact or not. The structural alignment method **Dali** (Holm and Laakso, 2016), which is still considered to be highly accurate to this day, uses distance matrices to represent structures. Other examples of structure representation are *c-beta* vectors as used in SSAP (Taylor and Orengo, 1989; Orengo and Taylor, 1996) and *URMSD vectors* as used by MAMMOTH (Lupyan *et al.*, 2005).

### 3.3 Heuristic optimization algorithms

Once we have a good structural representation in which two structures, or the substructures thereof, can be compared, we can formulate structural alignment as an *optimisation* problem and search for its solution. In the case of the contact matrix, we would look for an alignment of residues in which the *corresponding residues* have the *most similar contact patterns*. Since, searching for this optimal solution is thought to be an NP-hard problem (see previous section), performing an exhaustive search for the optimization problem is computationally not feasible. Instead, one needs to use a *heuristic* search strategy. Heuristic methods search for a good alignment, but cannot guarantee that the solution is optimal. Many of such heuristic search algorithms exist. Examples used for structural alignment are the *branch and bound algorithm* (Wohlers *et al.*, 2012) and *Monte Carlo optimization*.

In addition, double dynamic programming (Orengo and Taylor, 1996) and (single) dynamic programming (van Kempen *et al.*, 2023) are also used as optimisation strategies.

### 3.4 Statistical scoring of structural alignments

Next to the actual alignment, a structural alignment method will also provide a (dis)similarity measure for the two structures as an output. This is typically based on the alignment score, and is closely related to the used structure representation.

We described previously how RMSD can be used to score a structural superposition. Once a structural alignment is made, we could use RMSD as a measure to indicate how similar two proteins are. Most structural alignment programs will provide the RMSD as an output value for the superposition of the best structural alignment found.

However, there are some caveats with using RMSD as a measure for structural similarity. Note that, RMSD between two protein structures depends on the size of the proteins being compared. For two random structures, the average distance will become larger if the proteins being compared are large (e.g. Maiorov and Crippen, 1995). A similar effect may occur when one of the proteins that is being compared has many residues far away from the center of mass, for example if it has long loops. Hence, RMSD can be very sensitive to these ‘outlier’ atoms or residues in the structure.

Apart from the RMSD, structural alignment methods will typically also calculate a p-value or z-score, indicating how significant the structural alignment is. This is important, as the raw comparison scores (e.g. RMSD or overlap in contact maps) are intrinsically length dependent. If we want to see if two structures are significantly similar, this needs to be taken into account: two large structures only sharing two secondary structure elements are typically not homologous.

One can estimate distributions of raw alignment scores over a set of structural alignments between random (non-homologous) structures, for different sizes of structures. From these distributions, subsequent z-scores or p-values may be derived. Note that sequence alignment methods follow a similar strategy to derive statistical scores. For example, BLAST (Altschul *et al.*, 1997) uses statistical e-values in which database size, and sequence length are accounted for.

## 4 Applications of structure comparison

Applications of structural comparison and alignment are diverse but mostly concern the fields of evolutionary and structural biology. The majority of structural comparison applications are used to detect remote homology. Entire protein classification databases such as SCOP (Andreeva *et al.*, 2008),

CATH (Dawson *et al.*, 2017) and also the PDB (Berman *et al.*, 2000) use structural alignment methods to find and cluster structurally similar proteins. With increasing database sizes, and the availability of a larger number available predicted structures (Varadi *et al.*, 2022), the speed of these methods becomes increasingly important (van Kempen *et al.*, 2023). In Chapter 4 we will look at some of these resources in more detail.

The identification of distant homology with structure alignment also allows the prediction of protein's functions, (Roy *et al.*, 2012) and the classification of proteins with known structures. Structural comparison and alignment ease the prediction of protein structure from the sequence, (Ma *et al.*, 2012, 2013) and enable the identification of structure patterns and the observation of protein folding space (Yang and Honig, 2000; Kolodny *et al.*, 2006). A less known but significant application of structural comparison and alignment is the recognition of binding sites (Wass *et al.*, 2010).

## 5 Key points

- Structural superposition can be used to calculate the RMSD or another similarity metric between two structures, and requires a mapping between residues to do so.
- Structural alignment is used to determine a mapping between residues based on structure alone.
- Structure provides better insight into evolutionary processes because structure is more conserved than sequence.
- Contact maps can be used to represent a protein structure and its substructures in a rotation invariant fashion.
- Heuristic methods for score optimization are employed in structural alignment to make computational costs feasible.

## 6 Further reading

For a full overview please see Structural Alignment Chapter by Marti-Renom *et al.* (2009) in the book “Structural Bioinformatics” by Gu and Bourne (2009).

## Author contributions

Wrote the text:	OI, DG, SA, KAF
Created figures:	JG, SA, KAF
Review of current literature:	OI, RB, SA
Critical proofreading:	BS, SA, IH
Non-expert feedback:	HM, TL, RB
Editorial responsibility:	KAF, SA

The authors thank Halima Mouhib , Robbin Bouwmeester Robbin Bouwmeester, and Ting Liu  for critical proofreading.

## References

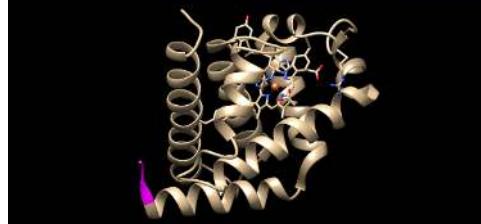
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J. et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–402.
- Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E. et al (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, **36**(SUPPL. 1), 419–425.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. et al (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.
- Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G. et al (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, **45**(D1), D289–D295.
- Finn, R.D., Bateman, A., Clements, J., Coggill, P. et al (2014). Pfam: the protein families database. *Nucleic acids research*, **42**(Database issue), 222–30.
- Gu, J. and Bourne, P.E. (2009). *Structural bioinformatics*. John Wiley & Sons, Hoboken, 2nd ed. nv edition.
- Hasegawa, H. and Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, **19**(3), 341–348.
- Holm, L. and Laakso, L.M. (2016). Dali server update. *Nucleic Acids Research*, **44**(W1), W351–W355.
- Kearsley, S.K. (1989). On the orthogonal transformation used for structural comparisons. *Acta Crystallographica Section A*, **45**(2), 208–210.
- Kolodny, R., Petrey, D. and Honig, B. (2006). Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, **16**(3), 393–398.
- Lupyan, D., Leo-Macias, A. and Ortiz, A.R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**(15), 3255–3263.
- Ma, J., Peng, J., Wang, S. and Xu, J. (2012). A conditional neural fields model for protein threading. *Bioinformatics*, **28**(12), 59–66.
- Ma, J., Wang, S., Zhao, F. and Xu, J. (2013). Protein threading using context-specific alignment potential. *Bioinformatics*, **29**(13), 257–265.
- Maiorov, V.N. and Crippen, G.M. (1995). Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Genetics*, **22**(3), 273–283.
- Marti-Renom, M.A., Capriotti, E., Shindyalov, I.N. and Bourne, P.E. (2009). Structure Comparison and Alignment. In J. Gu and P. E. Bourne, editors, *Structural Bioinformatics, 2nd Edition*, pages 397–418. John Wiley & Sons, Inc.
- Orengo, C.A. and Taylor, W.R. (1996). [36] SSAP: Sequential structure alignment program for protein structure comparison. In *Methods in Enzymology*, volume 266, pages 617–635. Academic Press.
- Roy, A., Yang, J. and Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, **40**(W1), 471–477.
- Shindyalov, I.N. and Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection*, **11**(9), 739–747.
- Tartaglia, G.G. and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Molecular BioSystems*, **5**(12), 1873–1876.
- Taylor, W.R. and Orengo, C.A. (1989). Protein structure alignment. *Journal of Molecular Biology*, **208**(1), 1–22.
- van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M. et al (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S. et al (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**(D1), D439–D444.
- Wass, M.N., Kelley, L.A. and Sternberg, M.J.E. (2010). 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Research*, **38**(SUPPL. 2), 469–473.
- Wohlers, I., Malod-Dognin, N., Andonov, R. and Klau, G.W. (2012). CSA: Comprehensive comparison of pairwise protein structure alignments. *Nucleic Acids Research*, **40**(W1), 303–309.
- Yang, A.S. and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology*, **301**(3), 665–678.

# Chapter 4

## Data Resources for Structural Bioinformatics

Jose Gavaldá-García  Bas Stringer  Olga Ivanova   
 Sanne Abeln\*  K. Anton Feenstra\*  and Halima Mouhib\* 

ATOM	1	N	ALA A	24	-2.158	-24.994	-20.089	1.00	26.29	N
ATOM	2	CA	ALA A	24	-1.385	-26.165	-20.158	1.00	26.29	C
ATOM	3	C	ALA A	24	-0.543	-26.935	-21.160	1.00	26.29	C
ATOM	4	O	ALA A	24	0.981	-27.992	-21.637	1.00	26.29	O
ATOM	5	CB	ALA A	24	-2.327	-27.069	-19.349	1.00	26.29	C
ATOM	6	CG	LYS A	25	0.672	-26.467	-17.987	1.00	22.24	N
ATOM	7	CD	LYS A	25	1.054	-26.542	-22.540	1.00	22.24	C
ATOM	8	C	LYS A	25	2.845	-27.512	-22.976	1.00	22.24	C
ATOM	9	O	LYS A	25	3.583	-26.837	-21.288	1.00	22.24	O
ATOM	10	CB	LYS A	25	1.610	-26.043	-23.704	1.00	22.24	C
ATOM	11	CG	LYS A	25	1.287	-26.551	-25.132	1.00	22.24	C
ATOM	12	CD	LYS A	25	1.998	-27.767	-25.656	1.00	22.24	C
ATOM	13	CE	LYS A	25	1.108	-29.069	-25.543	1.00	22.24	C
ATOM	14	NZ	LYS A	25	1.249	-29.768	-24.261	1.00	22.24	N
ATOM	15	CA	THR A	26	3.727	-28.750	-20.596	1.00	22.52	N
ATOM	16	CD	THR A	26	4.508	-28.257	-22.302	1.00	22.52	C
ATOM	17	C	THR A	26	5.529	-28.936	-23.547	1.00	22.52	C
ATOM	18	O	THR A	26	5.263	-29.275	-24.712	1.00	22.52	O
ATOM	19	CB	THR A	26	4.436	-30.770	-22.152	1.00	22.52	C
ATOM	20	OG1	THR A	26	3.796	-30.971	-26.880	1.00	22.52	O



\* editorial responsibility

Structural bioinformatics involves a variety of computational methods, all of which require input data. Typical inputs include protein structures and sequences, which are usually retrieved from a public or private database. This chapter introduces several key resources that make such data available, as well as a handful of tools that derive additional information from experimentally determined or computationally predicted protein structures and sequences.

## 1 Experimental protein structures

Experimentally determining the structure of a protein is no easy task, as was discussed in Chapter 2. Fortunately, researchers that do succeed in determining a protein’s atomic coordinates often submit their findings to a structure database. The Protein DataBank is the largest of such databases, providing a critically important resource to structural bioinformatics research.

### 1.1 The Protein DataBank

The Protein DataBank (Berman *et al.*, 2000) (or PDB) which was established in 1971, aims to provide a freely accessible, single global archive of experimentally determined structure data for biological macromolecules. The PDB provided one of first protein crystallography structures, sperm whale myoglobin (PDB-ID: 1MBN). Figure 4.1 shows how rapidly the number of determined structures has grown over the years. It is important to realize, that the total amount of available experimental protein structures is relatively small when compared to the amount of available sequences. For example, even though the PDB contained around 173,000 structures in 2020 (<https://www.rcsb.org/stats/summary>), UniProt already contained over 180 million sequences (<https://www.uniprot.org/statistics/TrEMBL>). This means that for only less than 0.1% of the known protein sequences an experimental structure of the protein is available. Note that the distribution of structures in the PDB is heavily biased towards structured proteins and proteins that are accessible for experimental structure determination. There is only little information on intrinsically disordered protein. Also, although 20-30% of all protein sequences contain transmembrane regions, less than 2% of protein structures in the PDB do. This is a direct consequence of experimental limitations. Since transmembrane regions usually contain large hydrophobic patches, their overexpression, purification and crystallisation is challenging and often not feasible during experimental workflows.

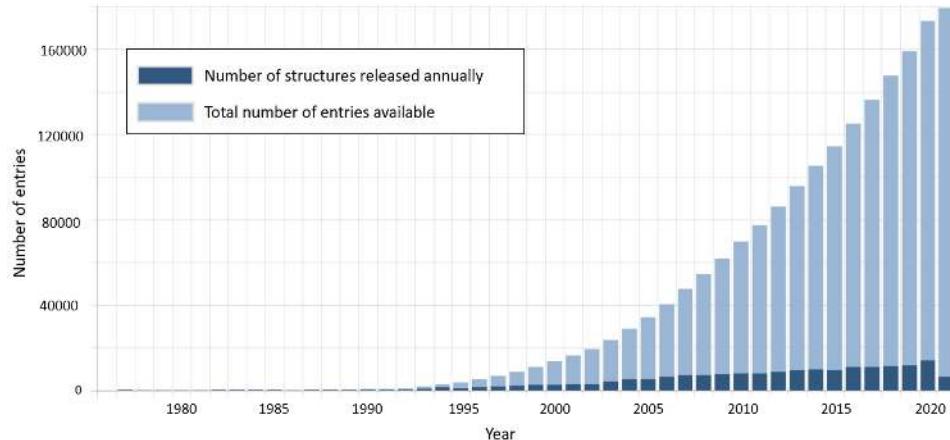
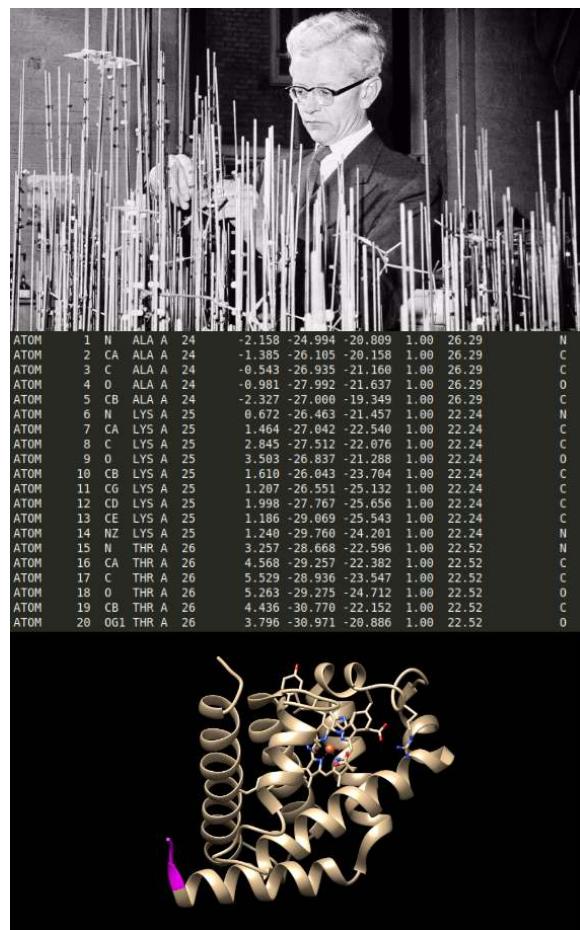


Figure 4.1: The availability of protein structures in the PDB over the years (June 2021).

### The PDB format

A typical PDB entry contains experimental data on the 3D structure of the heavy atoms in a protein, since hydrogen atoms are difficult to detect with X-ray radiation due to their low electron density. These sets of atomic coordinates are most commonly determined through one of three different techniques: X-ray crystallography tends to resolve structures with the highest resolution, followed by Nuclear Magnetic Resonance (NMR). Cryo-electron microscopy (cryo-EM) is more suitable for complexes. These techniques are fundamentally different in terms of what they can and cannot measure, and how accurately they do so. Therefore, it is important to be aware of the limitations of experimental structure determination techniques when we want to use them to build computational models.

One of the PDB's many important contributions to structural bioinformatics, is its standardised file format. Per PDB ID you can download the protein complex in "the PDB format", which, among other information, contains a list of atoms, their coordinates, and which molecule or chain they are a part of. This information is stored in the lines starting with "ATOM". Conceptually, this file contains all the information you need to place the atoms in a physical model, the same way John Kendrew did for sperm whale myoglobin, as shown in Figure 4.2 at the top. He elucidated its structure by means of X-ray crystallography, which granted him the Nobel Prize in Chemistry in 1962. The information allows to visualise the protein complex in more depth than the structure visualisation provided on the PDB web servers using visualization software such as UCSF ChimeraX(Goddard *et al.*, 2018). All visualization software is compatible with PDB format files.



*Figure 4.2: Different representation of the atomic coordinates of sperm whale myoglobin (PDB-ID: 1MBN). **Top:** John Kendrew working on his atomic model. **Middle:** small section of the corresponding PDB file (see text for further explanation). **Bottom:** Cartoon representation of the protein structure using UCSF-Chimera. The residues of the first 20 atoms in the PDB file are highlighted in magenta.*

The PDB file format also contains important additional data including the reference where the structure was published, updates that have been made since publication. Most importantly, it also contains all relevant experimental details on the method that was used for the structure determination, such as the structural resolution and the R value; see Chapter 2 for details on the different experimental methods and parameters.

### An ATOM line in the PDB file

An ATOM line in a PDB file contains extensive information about a resolved atom in the crystal structure. The meaning of each column is explained in detail below. (Note that Hydrogen atoms are not present in most structures, since X-ray crystallography can not resolve them due to their low electron density.)

	Record name	Atom serial number	Atom name	Residue	Chain	Residue sequence #	X Y Z coordinates (Å)	Occupancy	Temperature factor	Element symbol	Charge of the atom
ATOM	1314	C	GLY	A	150		0.015 31.316 -4.997	1.00	15.12	C	
ATOM	1315	O	GLY	A	150		0.306 31.382 -6.183	1.00	17.79	O	
ATOM	1316	N	TYR	A	151		0.741 30.645 -4.124	1.00	16.93	N	
ATOM	1317	CA	TYR	A	151		1.894 29.853 -4.417	1.00	20.10	C	
ATOM	1318	C	TYR	A	151		1.705 28.352 -4.172	1.00	26.94	C	
ATOM	1319	O	TYR	A	151		2.668 27.575 -4.301	1.00	41.24	O	
ATOM	1320	CB	TYR	A	151		3.020 30.468 -3.455	0.50	17.82	C	
ATOM	1321	CB	BTYR	A	151		3.073 30.254 -3.510	0.50	19.78	C	
ATOM	1322	CG	ATYR	A	151		4.370 29.863 -3.805	0.50	15.18	C	
ATOM	1323	CG	BTYR	A	151		3.490 31.695 -3.726	0.50	19.26	C	

Alternate location indicator      Code for insertion of residues

Charge of the atom

**Record name :** This field indicates the type of line in the PDB file.

In this figure, all the lines are ATOM. Other types of records like REMARK, ANISOU or HETATM exist and provide different information about the protein structure

**Atom serial number :** The number of the atom in the total structural complex. Note that this number is not renewed for the next molecule in the complex.

**Atom name :** The abbreviation of the atom name, e.g. CA is the alpha carbon.

**Alternate location indicator :** When an atom can be found in different locations, the PDB file describes each location as a different atom. A character (A, B, C, ...) indicates to which of the different locations each entry belongs. Since each line is interpreted as an atom, it is important to note that they will have different atom serial numbers. This data is directly related to the value of occupancy.

**Residue :** The name of the amino-acid residue to which the atom

belongs, in 3-letter notation.

**Chain** : Indicates to which molecule of the structural complex this atom belongs.

**Residue sequence number** : Indicates the position of the amino acid in the chain. Sometimes there is a “jump” in this series, which is due to a failed elucidation of the position of the residue. This value is renewed for each chain in the structure.

**Code for insertion of residues** : Rarely used. It is used to match the location of generally regarded “important” amino acids in a structure, for better comparison of this structure with its different versions. For example, this could be used to ensure that a catalytic pocket of an enzyme has the same residue sequence number in structures from different isoforms or species.

**X, Y and Z coordinates** : Spatial coordinates of the atom in the X, Y and Z axis. Given in Angstroms.

**Element symbol** : Chemical symbol of the respective atom in the protein structure (C, N, S, O, H for carbo).

**Charge of the atom** : If present, it indicates a non-neutral charge of the atom.

**Occupancy** : An occupancy of 1,0 indicates only this conformation was observed. An occupancy below 1,0 indicates that multiple conformations are possible. The occupancy of the different conformations for one atom should add up to 1,0.

**Temperature factor** or B-value: Measure of the smearing of the electron density. Higher values indicate higher excitation of the electrons and result in low precision in the determination of the atom’s position.

### Details, derived data and cross references

In addition to files like those described above, the PDB has a browser-based user interface that offers additional descriptions, derived data, and relevant cross references. For example, you can inspect a structure’s Ramachandran plot (see Chapter 1), structure validation reports and a detailed description of the experimental methods used to determine a structure. The feature viewer displays data derived by the PDB itself (e.g. secondary structure, disorder calculations, hydrophobicity) alongside information from other databases (UniProt, Pfam, Phosphosite) and homology models from the Structural Biology Knowledgebase (SBKB) and Protein Model Portal. Structural information on the PDB entries is also available on the web server PDBsum (De Beer *et al.*, 2014).

### The FAIR data principles

In recent years, there has been growing concern and awareness regarding the reusability of data. In order to improve the useful lifespan of data, more and more databases are adopting the FAIR principles (Wilkinson *et al.*, 2016). FAIR stands for Findable, Accessible, Interoperable and Reusable, and is meant to enable the sharing of data in such a way that consumers can more effectively locate, understand and reuse it. In the structural biology field, there was very early awareness that data should be shared in standardised formats and accompanied by ample provenance and metadata. In no small part because of this, the PDB has lived up to its vision of providing a consolidated source of experimental structure data. While this vision predates the FAIR principles by several decades, the PDB has since adopted said principles and continues to provide a wealth of structure data in a manner that suits the state of the art.

## 2 Structure analysis and annotation

As described in Chapter 2, experimental protein structure determination is a challenging endeavour, and computational models are often required to help and guide the process. Once protein structures are available from experiments (or from models, see Section 2.3) an important part of structural bioinformatics comprises a full characterization of these structures by further analysing and annotating them. Topics that should be discussed in this context include structure validation, secondary structure calling, structure classification and domain definition. Many of the resources previously mentioned in this chapter actually integrate precomputed analyses and annotations in their overviews. Secondary structure, domains and related families, and structure validation reports are frequently made available.

### 2.1 Structure validation

Validating the atomic model coming out of a structure determination experiment has three aspects: confirming validity of the actual measurements; confirming if the model is consistent with said measurements; and confirming if the model respects given physical and chemical constraints. In practice, users of experimentally determined models often don't have access to the raw measurements, so the third aspect is all they can consider. For modeled structures, there are no raw measurements, so only the third aspect is relevant.

These checks rely on tools similar to those used for homology modeling, which will be treated in depth in Chapter 7. Structural features, such as

bond lengths, angles, dihedrals, packing, H-bonds, should follow distributions similar to those observed across known (high resolution) structures. The assumption here is that these distributions will carry over to new structures, which may not hold true in case of novel features. Features that fall (far) outside the established distributions are potentially suspect, and the underlying data should be reviewed. Strong experimental evidence is required to accept an ‘outlier’ as a bona fide observation.

A visual representation of the quality of the backbone geometry can be obtained through a Ramachandran plot (see Chapter 1). Strict checks is done on whether the chemistry is in order, such as chirality of the backbone and side chains. Additional validation checks are performed using distributions which describe what is ‘chemically likely’, and whether hydrophobic residues are inside, and hydrophilic outside. Hydrogen bonds in the (hydrophobic) protein core have to be accounted for, as even a single unsatisfied hydrogen bond donor or acceptor in a hydrophobic environment may destabilize the entire protein. Sidechain packing in the protein interior is also checked using observed distributions from known, high resolution, structures.

There are several tools available for automated quality checking, either of novel structures or homology models. Frequently used tools include:

- WhatCheck (Hooft *et al.*, 1996)
- ProCheck (Laskowski *et al.*, 2012, 1993)
- MolProbity (Chen *et al.*, 2010; Hintze *et al.*, 2016)

Secondary structure has been introduced in Chapter 1. Assigning secondary structures to an experimentally determined structure is a fairly straightforward exercise, and can be done manually by expert crystallographers, or with programs such as DSSP (Kabsch and Sander, 1983) or Stride (Heinig and Frishman, 2004). Hydrogen bonds between beta sheets, within alpha helices, or within and between other parts of a protein and surrounding molecules can easily be assigned with programs like HBplus (McDonald and Thornton, 1994).

Note that *calling* a secondary structure from atomic coordinates is a very different exercise than *predicting* it from an amino acid sequence. State of the art methods, like NetSurfP-2.0 (Klausen *et al.*, 2019) and SPOT-1D (Hanson *et al.*, 2019), can make such predictions with reasonable accuracy. For more on secondary structure prediction, see Chapter 9.

## 2.2 Structural classification

With the set of available protein structures continuously expanding, it becomes insightful to compare and classify them. Done right, classification might for example allow us to find distant homologous relationships, since structure is generally more conserved than sequence. Well-designed structural classification schemes, such as those implemented by SCOP and CATH,

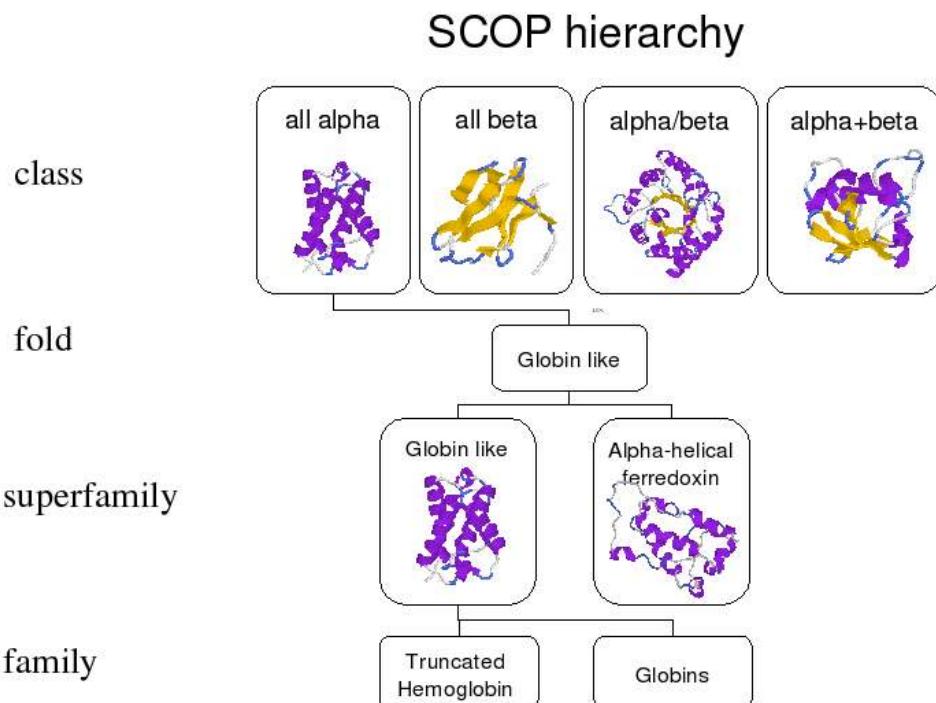


Figure 4.3: Different levels of classification in SCOP.

can provide a gold standard for curating homologous relations. This can in turn be used to validate sequence based homology search methods, such as (PSI-)BLAST, HMMer and HHBlitz, as we discuss in Chapter 7.

Many features can be considered to determine whether or not two proteins should be classified together. You could for example consider sequence similarity; shared functions or functional sites; conserved secondary structure elements and topology; or a high structural alignment score, as described in Chapter 3. Different classification schemes may use any combination of these and other factors to group and order protein structures into a hierarchy, ideally in such a way that different levels correspond to biologically relevant features (e.g. homology, or having equivalent functions).

For multi-domain proteins, structural classification works best when performed on each domain individually, rather than for the protein as a whole. Domains and domain calling are covered in the next section.

## SCOP

The Structural Classification of Proteins (SCOP) database (Andreeva *et al.*, 2008) implements four levels of hierarchy to classify protein structures. The aim is to group proteins in such a way that homologous proteins (i.e. those with shared evolutionary ancestry) will cluster on the superfamily level. In

order of decreasing specificity, the levels are family, superfamily, fold and class as shown in Figure 4.3. These are manually assigned to each structure by expert curators, abetted by computational methods.

- Proteins and domains are assigned to the same *family* if they *i)* have significant sequence similarity and *ii)* have a similar function and structure.
- Families are considered part of the same *superfamily* if they have low sequence identity, but their structure and functional features suggest common evolutionary origin is possible.
- Superfamilies share a *fold* if they have the same major secondary structure elements, in a similar arrangement and topology. Folds are thought to arise from certain chain topologies having specific packing arrangements that are energetically favourable.
- *Classes* are a coarse division based on the secondary structure elements: all- $\alpha$ , all- $\beta$ ,  $\alpha/\beta$  and  $\alpha+\beta$ , as was already introduced in Chapter 1 “Introduction to Protein Structure” (see Figure 1.9 for an overview).

## CATH

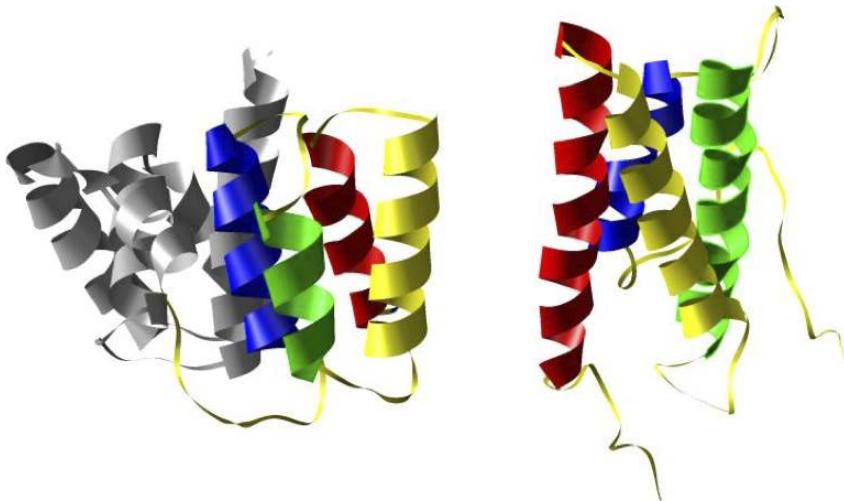
CATH (Dawson *et al.*, 2017) is, like SCOP, a hierarchical structural classification database, based on similarity of structure, function and sequence. There are, however, some significant differences.

The authors of CATH attempt to automate their classification process as much as possible, without losing biological relevance. Its hierarchy consists of the levels Class, Architecture, Topology and Homology (CATH). Here topology is similar to SCOP’s fold level, and homology is similar to SCOP’s superfamily level. The architecture level is specific to CATH and represents the shape defined by the assembly of secondary structures without considering their connectivity (see Figure 4.4).

The biggest challenges to fully automatic assignment are recognizing domain boundaries, distinguishing between the homology and topology level, and grouping families into homologous groups. Thus, manual curation is still necessary.

## Domain definitions

Domains are conserved protein regions between 25 and 500 amino acids in length, which can exist and evolve independently of the rest of the protein, as already introduced briefly in Chapter 1. Many (but not all) domains are self-contained, and will fold into a compact structure and function independently, not unlike a single-domain protein. In evolution, it is rather common for entire domains get duplicated, deleted or inserted next to other domains like building blocks: a phenomenon referred to as domain shuffling (see



*Figure 4.4: Left: ‘Influenza virus matrix protein’, PDB:1AA7. Right: ‘Solution structure of four helical up-and-down bundle domain of the hypothetical protein 2610208M17Rik similar to the protein FLJ12806’, PDB:1UG7. CATH classifies the N-terminal domain of 1AA7 (left, coloured region) and 1UG7 (right) into the same architecture: ‘up and down bundle’. Following the path of the secondary structure elements (coloured sequentially: red, yellow, green and blue) it is clear that the 4 helices are differently connected and have thus another topology. SCOP classifies both proteins under the same class: ‘all alpha’. CATH defines two separate domains for 1AA7 (grey, coloured), whereas SCOP defines the entire protein as a single domain.*

Figure 4.5, from Chothia (2003)).

Defining domains is no trivial task. Consider for example the rainbow coloured structure in Figure 4.6. At a glance, it is not immediately clear where the domain boundaries should fall. Several databases exist that provide domain definitions. An important distinction is domain definition that are derived from *structure* and those that are derived from *sequence*. We will return to domain prediction from sequence in Chapter 6, Section 1.5.

Structure-based domain definition can be found in CATH, SCOP and even in the PDB. Note that the definitions between these resources do not necessarily agree. In general SCOP is more conservative in splitting up a protein in several domains; it follows the general rule that an instance of a domain, or a homolog thereof, needs to have been observed to fold independently.

Both PFAM (Finn *et al.*, 2014) and the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2017) provide sequence based domain definitions. PFAM clusters protein sequences into sequence families using profile-based HMM alignments, where the seed profile has been manually curated. The CDD is a protein annotation resource that consists of domain defini-

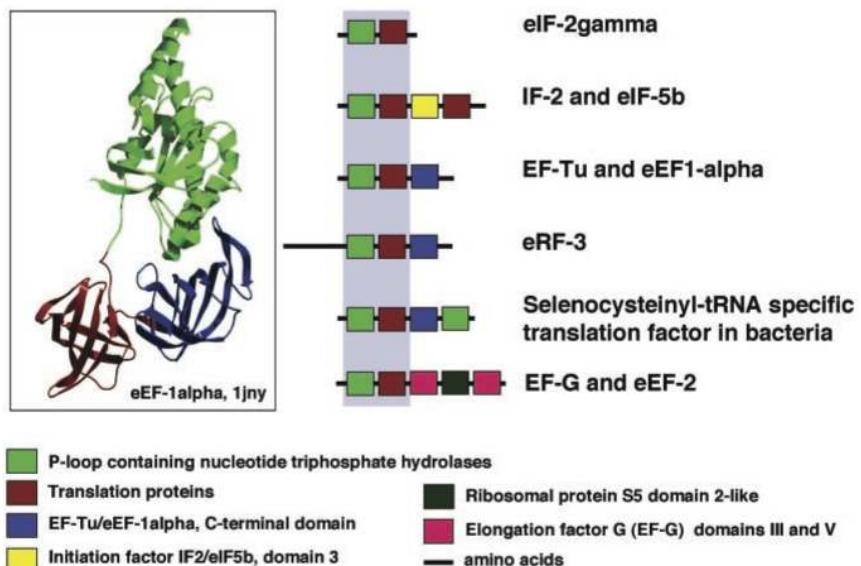


Figure 4.5: Domains being reused in different combinations is very common in evolution. This p-loop domain (green) occurs in at least 35 different domain combinations, six of which are shown above. From: Chothia (2003).

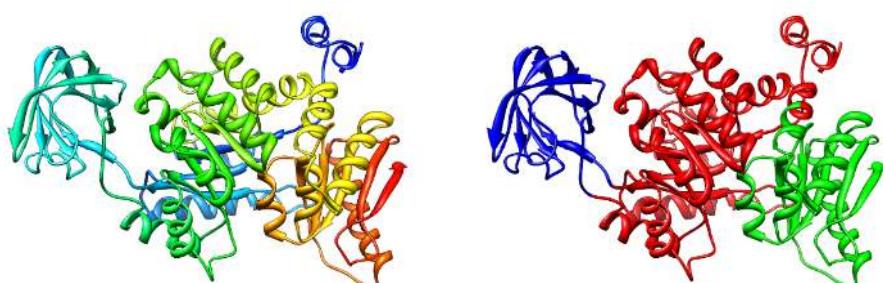


Figure 4.6: Structure of rabbit pyruvate kinase (1PKN). Left: rainbow from blue (N) to red (C). Right: coloured according to domain boundaries as assigned by SCOP. The left domain (blue) clearly shows a distinct compact structure, with its own hydrophobic core. From this angle, the C-terminal domain (green) is not immediately apparent as a separate domain in the rainbow colored structure. Note that the middle domain (red) is discontinuous, as the blue domain sits in the middle of it.

tions based on annotated multiple sequence alignment models. A subset of the CDD for which structures were used to define and validate domain boundaries is known as ‘NCBI-Curated Domains’. A position specific scoring matrix is associated with each conserved domain in the database, and can be used by PSI-BLAST for (remote) homology detection.

### 2.3 Protein sequences

Though not as immediately obvious as atomic coordinates, a protein’s sequence contains a lot of information about its structure, and by extension, its function. Despite predicted structures becoming much more accurate and more readily available, as discussed in Chapter 1, there is still much more sequence data and some form of sequence analysis is an integral part of many approaches commonly used in bioinformatics research.

Due to their ubiquity, sequence databases are not as consolidated as structure databases. Many databases offer some combination of protein, RNA and DNA sequences with relevant annotations and cross references, but within the context of structural bioinformatics, UniProt is arguably the most important one.

#### UniProt

The UniProt database (Bateman *et al.*, 2017) consolidates a vast amount of information about proteins from a various sources. Each entry contains, at least the protein’s amino acid sequence, name or description, taxonomic data, and citation information. This core data is then enriched with as many annotations as possible, among others spanning common biological ontologies, classifications and cross references, accompanied by an indication of annotation quality through experimental and computational evidence attribution. The database consist of two main parts: the UniProtKB/Swiss-Prot and the UniProtKB/TrEMBL database; resp. Swiss-Prot and TrEMBL for short here. Proteins stored in TrEMBL contain protein sequences for which annotations and function characterization are created by computational techniques. As of June 2021, almost 220,000,000 protein sequences were stored in TrEMBL, although it should be noted that the protein annotations therein are not manually curated. Once a sequence in TrEMBL is reviewed the protein will be stored in Swiss-Prot which provides human curated annotations for each of the protein sequences that have been manually annotated and filtered on redundancy. In June 2021, SwissProt contained almost 560,000 curated protein sequences.

The provided annotations in UniProt are essential to gain insight into the protein function, pathology, interaction partners, subcellular location, and post-translational modification or processing. However, not the same amount of information and annotations is provided for every sequence. The

UniProt database is an essential resource for structural bioinformatics as it provides a unified framework to find sequences and refer to them by unique identifiers. Furthermore, it conveniently cross references known 3D structures, secondary structure, homologues, and family and domain annotations and provides this information through easily accessible human- and machine-readable interfaces.

### Modeled structure resources

The PDB predominantly contains experimental structure data, but as mentioned before, structures are laborious and difficult to determine. Many methods and pipelines exist to predict, or model, a protein's structure in stead of direct measurements. This is discussed in greater detail in Chapter 7.

Databases containing modeled structures include the Swiss-Model Repository (SMR) (Bienert *et al.*, 2017) and ModBase (Pieper *et al.*, 2014) databases, which offer access to structures produced by the Swiss-Model homology-modelling and ModPipe pipelines, respectively. If you are looking for several modeled structures for a particular protein of interest, the Protein Modeling Portal (PMP) (Haas *et al.*, 2013) provides a single interface to simultaneously query SMR, ModBase, and models generated by several partners of the Protein Structure Initiative (PSI) (Montelione, 2012; Matthew Zimmerman *et al.*, 2017).

Since the emergence of AlphaFold v2.0 by DeepMind (Jumper *et al.*, 2021), the protein sequences in UniProt have been expanded with an unprecedented number of accompanying structural models, also known as AlphaFold DB (Varadi *et al.*, 2022). AlphaFold generates atomic coordinates directly from the amino acid sequence. Although these AI predicted structures are extremely helpful in many cases, it is important to be aware the not all predicted structures are equally good. The predictions will be lower quality for intrinsically disordered proteins and unstructured regions of proteins. As a measure to estimate the reliability of the models, the per-residue and pairwise model-confidence estimates, as well as the 'provided predicted aligned errors' should be used.

### Other sequence resources

Along with UniProt, the European Bioinformatics Institute (EMBL-EBI) also maintains UniRef (Suzek *et al.*, 2015) and UniParc (Leinonen *et al.*, 2004) – both very useful resources as far as protein sequences are concerned, though tightly integrated with UniProt itself. In parallel, the National Center for Biotechnology Information (NCBI) maintains a plethora of databases, along with the tools to search them, including RefSeq (O'Leary *et al.*, 2016), GenBank (Benson, 2004) and the Conserved Domain Database

(CDD) (Marchler-Bauer, 2003; Marchler-Bauer *et al.*, 2017). Also worth mentioning is the Protein Information Resource (PIR) (Wu *et al.*, 2003), hosted by the Georgetown University Medical Center.

Many tools and browser interfaces exist to query and visualize the contents of these databases, such as EMBL-EBI's *Ensembl* genome browser (Zerbino *et al.*, 2018), and NCBI's *Entrez* system (Maglott, 2004). Note that because of the high degree of cross referencing between all of the above, many search tools also yield results from, or at least refer to databases managed by other institutes.

### 3 Functional data resources

As will be further discussed in the next chapter, a protein's function is inseparably connected to its structure. Thus, an overview of important primary and tertiary structure resources would not be complete without considering correlated function annotation resources as well. Here, types of annotation to consider include, but are not limited to, protein function, interaction partners, the impacted biochemical pathways, and the location of protein expression at a cellular level. Two particularly useful examples are briefly introduced below.

#### Function annotation: Gene Ontology

The Gene Ontology (GO) (Ashburner *et al.*, 2000; Carbon *et al.*, 2017) is a structured vocabulary that was designed to annotate gene products with clearly defined terms. These terms span three categories: biological process, cellular component and molecular function. Using GO terms to describe a protein's function enables searching for specific or related functions without having to deal with natural language and free text, among other benefits. As such, GO terms are very popular a annotation tool, adopted by many databases including the PDB and UniProt.

#### Protein-protein interactions: STRING

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk *et al.*, 2015) is incredibly useful for compiling a list of proteins another protein of interest interacts with. STRING considers five sources of information to detect or predict your query's interaction partners: curated databases, experimental data, textmining, co-expression and homology. It additionally predicts gene neighborhoods, fusions and co-occurrence, and assigns confidence scores to each interaction based on the above. STRING also allows recursive queries, to extend the resulting network with interactions to additional protein binding partners.

In summary, there exists a plethora of gene and protein annotation resources, both structural and non-structural in nature. Other relevant databases, that were not mentioned before, include the Human Phenotype Ontology (HPO) (Köhler *et al.*, 2018) describes phenotypic abnormalities in human disease; WikiPathways (Slenter *et al.*, 2018), BioPax (Demir *et al.*, 2010) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) each contain a different type of pathway, network and model information; the Gene Expression Omnibus (GEO) (Barrett *et al.*, 2012), Expression Atlas (Petryszak *et al.*, 2016) and Human Protein Atlas (HPA) (Thul and Lindskog, 2018) track when and where certain genes are expressed.

## 4 Key points

- Understand the data you are working with, and its sources
- Experimental protein structures are available in the PDB database which provides pdb-files with atomic coordinates in 3D space.
- Structure validation is important to ensure that a structure is based on coherent experimental data and to ensure its consistency with physical chemical constraints.
- Databases with structural classification of proteins allow to find more distant homologues by comparison of domains and other secondary structure elements.
- Sequence databases are useful to find homologous proteins and their respective available information, such as a corresponding experimental or predicted structure.
- Other protein features can be retrieved from databases like the STRING (protein-protein interactions) or GO database (molecular function, biological process and cellular component).

## 5 Further reading

- Wilkinson *et al.* (2016)
- Further explanation on FAIR principles and list of resources to expand further in each of its principles: <https://www.dtls.nl/fair-data/fair-principles-explained/>

## Author contributions

Wrote the text:	JG, BS, OI, SA, KAF, HM
Created figures:	JG, BS
Review of current literature:	JG, BS, OI, SA
Editorial responsibility:	SA, KAF, HM

The authors thank Alumit Rodrigues Pereira  for critical proofreading.

## References

- Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E. et al (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, **36**(SUPPL. 1), 419–425.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. et al (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C. et al (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, **41**(D1), D991–D995.
- Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M. et al (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1), D158–D169.
- Benson, D.A. (2004). GenBank. *Nucleic Acids Research*, **33**(Database issue), D34–D38.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. et al (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.
- Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G. et al (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, **45**(D1), D313–D319.
- Carbon, S., Dietze, H., Lewis, S.E., Mungall, C.J. et al (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, **45**(D1), D331–D338.
- Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A. et al (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(1), 12–21.
- Chothia, C. (2003). Evolution of the Protein Repertoire. *Science*, **300**(5626), 1701–1703.
- Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G. et al (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, **45**(D1), D289–D295.
- De Beer, T.A.P., Berka, K., Thornton, J.M. and Laskowski, R.A. (2014). PDBsum additions. *Nucleic Acids Research*, **42**(D1), 292–296.
- Demir, E., Cary, M.P., Paley, S., Fukuda, K. et al (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, **28**(9), 935–942.
- Finn, R.D., Bateman, A., Clements, J., Coggill, P. et al (2014). Pfam: the protein families database. *Nucleic acids research*, **42**(Database issue), 222–30.
- Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F. et al (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, **27**(1), 14–25.
- Haas, J., Roth, S., Arnold, K., Kiefer, F. et al (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, **2013**.
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y. and Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, **35**(14), 2403–2410.
- Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, **32**(Web Server), W500–W502.
- Hintze, B.J., Lewis, S.M., Richardson, J.S. and Richardson, D.C. (2016). Molprobity's ultimate rotamer-library distributions for model validation. *Proteins: Structure, Function and Bioinformatics*, **84**(9), 1177–1189.
- Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996). Errors in protein structures [3].
- Jumper, J., Evans, R., Pritzel, A., Green, T. et al (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021* **596**:7873, **596**(7873), 583–589.
- Kabsch, W. and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577–2637.
- Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.
- Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K.K. et al (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, **87**(6), 520–527.
- Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B. et al (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**(2), 283–291.
- Laskowski, R.A., MacArthur, M.W. and Thornton, J.M. (2012). PROCHECK : validation of protein-structure coordinates. In E. Arnold, D. M. Himmel, and M. G. Rossmann, editors, *International Tables for Crystallography*, volume F,Ch.21.4, pages 684–687.

- Leinonen, R., Diez, F.G., Binns, D., Fleischmann, W. et al (2004). UniProt archive. *Bioinformatics*, **20**(17), 3236–3237.
- Maglott, D. (2004). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**(Database issue), D54–D58.
- Marchler-Bauer, A. (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Research*, **31**(1), 383–387.
- Marchler-Bauer, A., Bo, Y., Han, L., He, J. et al (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, **45**(D1), D200–D203.
- Matthew Zimmerman, Marek Grabowski, Wladek Minor, Helen M. Berman et al (2017). Protein Structure Initiative Publications, 2000–2016.
- McDonald, I.K. and Thornton, J.M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, **238**(5), 777–793.
- Montelione, G. (2012). The Protein Structure Initiative: achievements and visions for the future. *F1000 Biology Reports*, **4**, 7.
- O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S. et al (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**(D1), D733–D745.
- Petryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A. et al (2016). Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, **44**(D1), D746–D752.
- Pieper, U., Webb, B.M., Dong, G.Q., Schneidman-Duhovny, D. et al (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, **42**(D1), D336–D346.
- Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A. et al (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, **46**(D1), D661–D667.
- Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. and Wu, C.H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926–932.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K. et al (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**(D1), D447–D452.
- Thul, P.J. and Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science*, **27**(1), 233–244.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S. et al (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**(D1), D439–D444.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G. et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.
- Wu, C.H., Yeh, L.S.L., Huang, H., Arminski, L. et al (2003). The Protein Information Resource. *Nucleic acids research*, **31**(1), 345–7.
- Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R. et al (2018). Ensembl 2018. *Nucleic Acids Research*, **46**(D1), D754–D761.

# Chapter 5

## Protein Function & Interactions

Annika Jacobsen  Qingzhen Hou  Bas Stringer   
Sanne Abeln  K. Anton Feenstra 



\* editorial responsibility

**NOTE TO STUDENTS**

Many figures in this chapter are highlighted in light blue like this text; these are informative and/or background examples and explicitly not exam material!

In the preceding chapters of Part I we have considered the basics of protein structures. In Chapter 2 “Structure determination” we explained how they can be experimentally determined and subsequently validated, and in Chapter 3 “Structure Alignment” how they can be compared and classified. As introduced in Chapter 1 “Introduction to Protein Structure”, the structure of protein is key to its function and this topic will be further explored in this chapter. Moreover, in Chapter 4 “Data Resources for Structural Bioinformatics” we already saw that the functional site of a protein, and the conservation thereof, is key to detecting very remote homologous relationships between proteins at the SCOP superfamily level, or the CATH homology level.

In fact, most methods within structural bioinformatics will be used to answer questions about the function of a protein. For example, structural alignment and subsequent superposition is used to identify positions where the structure is most conserved. These conserved structures can point to functional sites of the protein (see Chapter 3 “Structure Alignment”). Similarly, homology modelling may be used to identify the functional part of a sequence, by overlaying it on a homologous structure (see Chapter 7 “Practical Guide to Model Generation” and Chapter 11 “Function Prediction”). A protein’s function is largely determined by its structure, and this should always be kept in mind when applying any structural bioinformatics method. Therefore an entire Chapter introduces this structure-function relationship, and explains in greater detail what may be considered a “functional site” of a protein (see also Table 2). This chapter discusses the major types of functional interactions of a protein, such as with another protein, with complexes, with small molecules (typically metabolites), and with DNA. We also introduce some experimental and bioinformatics approaches, which provide insight in the structure-function relationship. Finally, we consider different functions that proteins perform in the context of the whole cell.

## 1 Proteins are the machinery of the cell

Of the tens of thousands of possible proteins in, e.g., the human proteome, often several thousands are produced and present at the same time in a given cell. A snapshot of proteomics will determine the behaviour of the cell at the current moment, where function of proteins is defined by several processes: by production process, by environmental context, and by direct

interactions. In the complex process of protein production, the final product and its structure (and thus function) may be affected by regulation of gene expression, mature mRNA assembly, and post-translational modification. For instance, during mature RNA assembly, alternative splicing can result in slightly (or very) different form of protein. And post-translational modifications, as these modulate the physico-chemical properties of the protein surface. These alterations can cause a protein's to adopt a different structure or change its enzymatic activity. In addition, environmental factors like pH, salt concentration and temperature affect folding, dynamics, and thus ultimately, function. Despite their importance, thorough discussion of these factors falls outside the scope of this book.

After all, direct interactions are integral part of protein functions. Most, if not all, proteins make direct physical contact with other molecules: one protein may bind another to activate or deactivate it; or several proteins together may make an enzyme complex to catalyse a specific reaction; they can form a transcription complex that transcribes DNA into RNA; or proteins may build structures supporting the cells shape and extracellular environment. Biochemistry textbooks, such as ‘Stryer’ (Berg *et al.*, 2002), are filled with examples of interacting proteins involved in various processes, such as DNA replication, cell division, and cellular trafficking. A summary of data resources where such functional data and annotations may be retrieved is given in Chapter 2. Detailed descriptions of protein-protein, -DNA, -RNA, -small ligand and -membrane interactions are presented at the end of this chapter, with some concrete examples.

## 1.1 Protein function

Elucidating proteins functions requires various experimental and computational approaches because proteins functions are extremely diverse. Moreover, proteins might perform several (or many) tasks depending on the environmental context. Some of these functions are: reaction catalyst (enzymes) (e.g., hexokinase), messenger (hormones) (e.g., insulin), structure (e.g., keratin), contraction (e.g., actin), signals transducer (e.g., G-protein-coupled-receptor; GPCR), defence (antibodies) (e.g., immunoglobulin), transport (e.g., hemoglobin for oxygen) and storage (e.g., myoglobin also for oxygen); see Table 1 for an overview. All these functions revolve around binding sites at the protein surface. Recognising these functional sites in the structure is the first step in clarifying protein roles (even better if it could be done from the sequence, as we have much more sequence data than structures). Often, ‘irregularities’ in the structure indicate a functional role, so the problem shifts to finding these irregularities. We will come back to the issue of function prediction in Chapter 11 “Function Prediction”, although a comprehensive discussion of this topic goes beyond the scope of this book. In light of the diversity of protein function (the overview given here does not

*Table 1: Examples of particular functions that proteins may have. Some functions have a particular name for the class of proteins that perform them. Some examples will be elaborated further on in this chapter.*

Function	Class	Example
catalyzing reactions	enzymes	hexokinase
messenger	hormones	insulin
structure	typ. fibers	keratin
contraction	typ. motors	actin
transducing signals	membrane	G-protein-coupled receptor (GPCR)
defence	antibodies	immunoglobulin
transport		hemoglobin
storage		myoglobin
metabolism	e.g. metabolic network	glycolysis
signalling	e.g. signalling pathway	MAPK pathway

aim to be complete), it is always best to be specific when possible. For example, say ‘this region of the protein is involved in binding to actin’, or ‘we predict small molecule binding sites’, rather than to give a general statement that some protein or region is ‘functional’.

In addition to the type of functions mentioned above, we can also define the function of a protein at the level of their interactions with other proteins, or more loosely at the level of ‘pathways’, in which several functions are linked, often one occurring after the other. An example would be enzymes catalyzing biochemical (metabolic) reactions in a metabolic pathway. Another example is proteins involved in signalling (receptors, kinases, phosphatases, etc.) linked together into a signalling pathway.

In bioinformatics and systems biology, pathways are often represented as a network where nodes are proteins and edges are interactions. These networks are basically the static models of molecular interactions. The place of a protein in this network, and the local structure of this network, may tell us something about the particular function of that protein. For example local network structure seems to be linked to protein function in some cases (e.g. Alon, 2007). Clustering at the level of networks may also identify functional modules (e.g. Dittrich *et al.*, 2008; May *et al.*, 2016; El-Kebir *et al.*, 2015; Jacobsen *et al.*, 2018)

Table 2: Glossary of terms used in relation to protein Functional Sites.

Functional Site	Description
Binding Site	place where a ligand binds to a protein
Active Site	<i>binding site</i> in enzyme for ligand, where reaction takes place
Interface region	part of a protein that interacts with another protein
Allosteric site	<i>binding site</i> , which alters binding and/or reaction at the active site
Binding region	region (generic) of a protein where some molecule may bind
Binding pocket	Convex (hollow) <i>binding region</i> , typically for small molecule

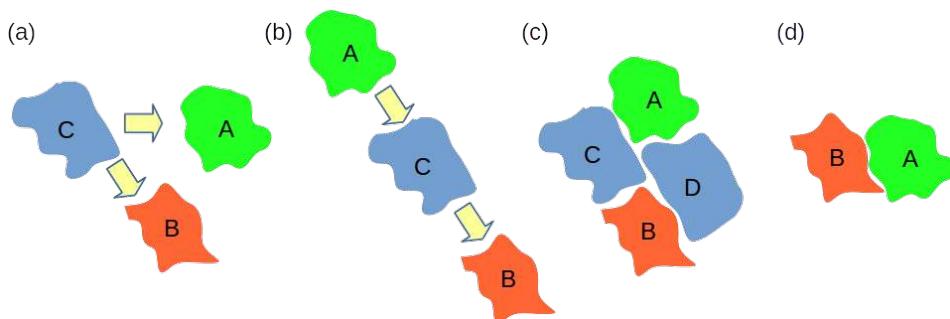
## 1.2 Functional site

A functional site of a protein is a region that enables a protein to perform a specific task. A single protein can even contain multiple functional sites that enables the protein to perform multiple tasks. In this chapter some of functional sites will be described. A general example of a functional site, is the interaction site where a protein binds DNA or small ligand, or another protein. In an enzyme, the functional site is called the ‘active site’ where it binds its substrate and performs a reaction on it. The binding region where a co-factors binds to an enzyme is also a functional site. Table 2 lists several types of functional sites.

## 2 Protein-protein interactions & complexes

Protein-protein interactions (PPIs), as the name implies, are interactions between two or more proteins. Thus, they can bind together to form a protein complex, also referred to as the quaternary protein structure. This is the natural extension of primary, secondary and tertiary structure, which were already introduced in Chapter 1 “Introduction to Protein Structure”. Often, protein function arises at the level of the quaternary structure, i.e., what function the complex performs that the protein is in.

A perhaps well-known example of such a quaternary protein structure is hemoglobin, which is a complex of four protein subunits. Here, the single hemoglobin subunits are not (fully) functional, hemoglobin usually is not observed in a monomeric (unbound) state. The Panel “Allosteric motions and time-resolved crystallography” (in Chapter 2) gives some more explanation on the role of protein interactions (and dynamics) in the function of hemoglobin.



*Figure 5.1: Protein-protein interactions inferred from experimental data are not always in direct physical contact. (a) The signal observed correlating A and B may have shared cause C. (b) The correlation may have an intermediate. (c) Proteins A and B may be member of the same complex, without being in contact directly. (d) Direct physical contact between proteins A and B.*

## 2.1 Inferring protein interaction from experimental data

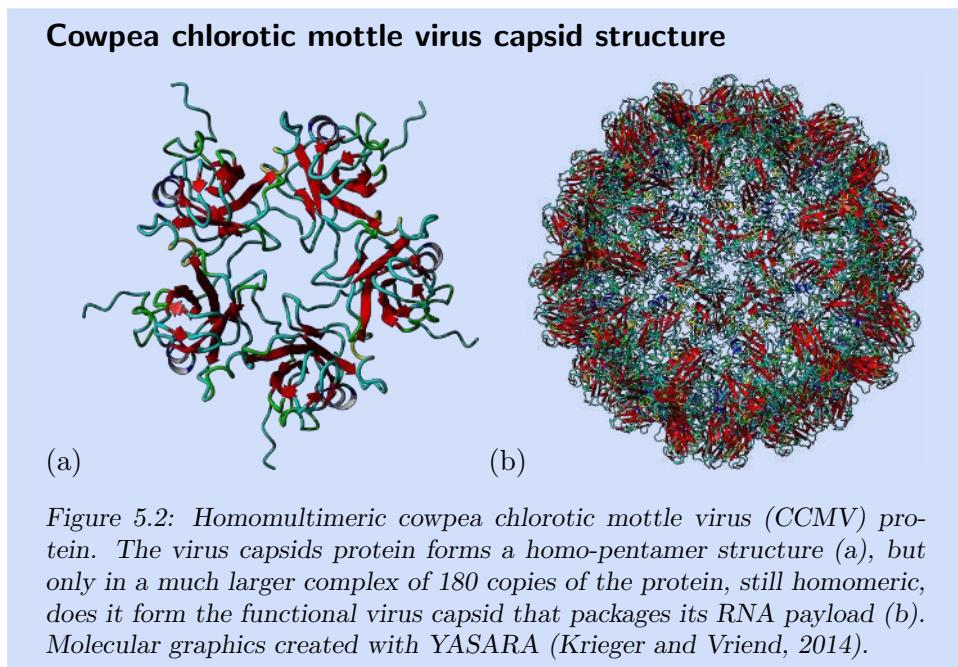
Various experimental methods exist to infer interactions between proteins. Some of these interrogate functional (or indirect) interactions, while others determine direct physical contacts. The functional interactions are usually inferred from omics data through calculating correlations. Besides direct physical contact, the cause of the correlation may be e.g. two proteins belonging to the same pathway. Figure 5.1a illustrates several scenarios by which a correlation occurs between proteins A and B, by: *i*) shared dependence on protein C (a); *ii*) indirect via intermediate protein C (b); *iii*) indirect via complex membership with protein C and D (c); and *iv*) direct contact physical interaction where A and B are bound together (d).

The interactions we will be discussing in this chapter are mainly in direct physical contact as shown in Figure 5.1d, which is most relevant for structural bioinformatics methods.

### Homomeric versus heteromeric complexes

Two interacting proteins can form a dimer complex, or simply a dimer. We distinguish between homo- and heterodimers. A homodimer is formed when two identical proteins interact, whereas a heterodimer is formed when two distinct proteins interact. Interactions typically also involve more than just two proteins, forming multimeric complexes (e.g., trimer, tetramer, etc.). Also for multimeric complexes, we distinguish between homomeric and heteromeric complexes.

A strong example of a complex consisting of multiple identical subunits is the virus capsid; Figure 5.2 shows a example for the cowpea chlorotic mottle plant virus (CCMV). Another example of a homomeric complex is F-actin which forms long filaments in the cell, as shown in Figure 5.3. Hemoglobin

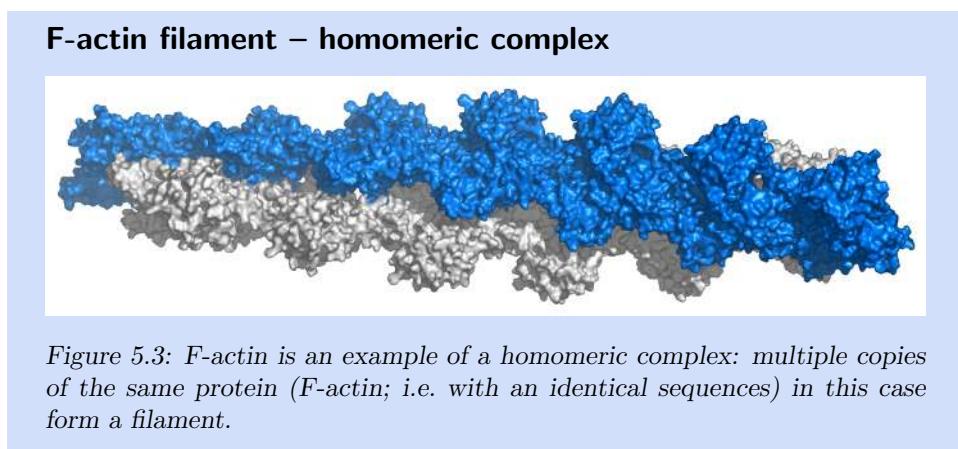


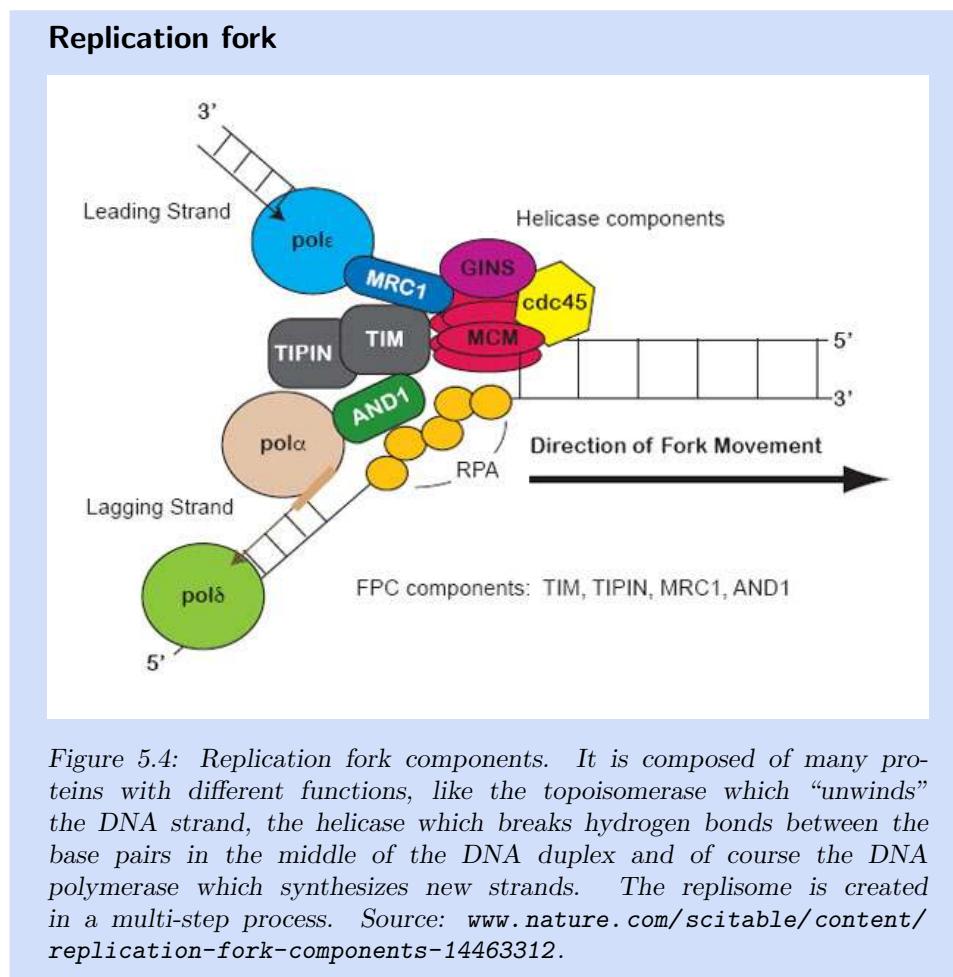
is an example of a heteromeric complex, as it usually contains two  $\alpha$  and two  $\beta$  subunits (Figure 1.1).

### Obligate versus non-obligate complexes

We can furthermore differentiate between two types of protein complexes: obligate and non-obligate. Obligate complexes form a functional structure only when all subunits are bound. A single subunit is not functional on its own.

Non-obligate complexes consist of proteins that also are functional when





not bound in the complex. Non-obligate complexes are formed based on specific temporal requirements e.g. nutritional status of the cell or the stage of the cell cycle. The interactions are described as transient and have a relatively short half-life. A classic example of that is the replication protein complex (replisome) Figure 5.4, which assembles on the DNA. It only starts performing its function of DNA replication when all components have been assembled, but individual proteins perform sub-tasks in smaller complexes or even in isolation.

On the other hand, obligate complexes are most often permanent; hemoglobin is a good example (Figure 1.1). Also an F-actin complex can be considered obligate, as single units do not have a clear function, even though the filaments may be actively regulated, so that they get built at some moment and later removed (Figure 5.3).

A rather impressive example of an obligate heteromeric complex is the ribosome which includes not only several tens of different proteins but also

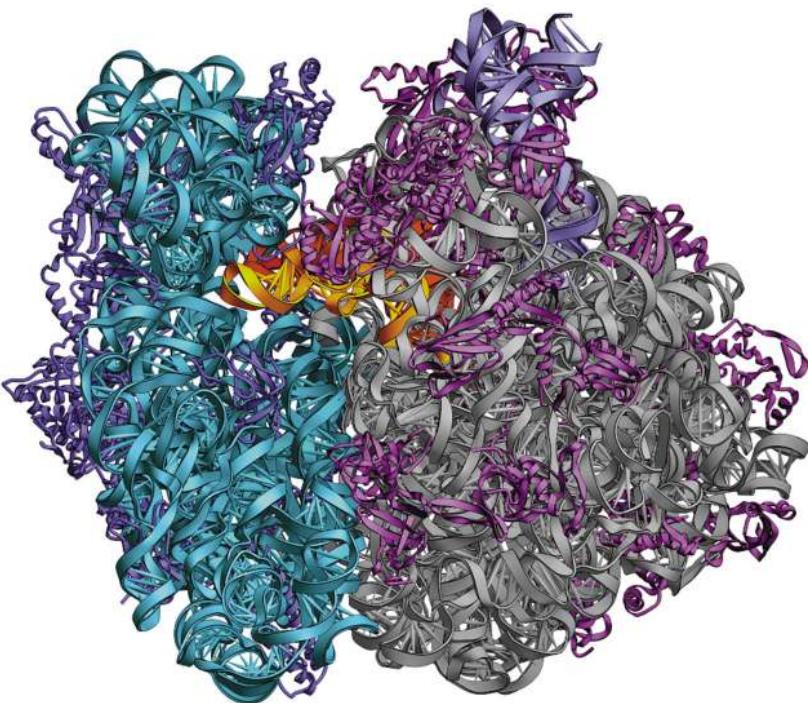
**70S Ribosome – heteromeric complex**

Figure 5.5: The 70S ribosome is an example of a heteromeric complex. It consists of several tens of different proteins and in addition one large and one small RNA molecule, which are the scaffolds of the large and small subunits of the ribosome. Source: [http://rna.ucsc.edu/rnacenter/ribosome\\_images.html](http://rna.ucsc.edu/rnacenter/ribosome_images.html).

two large RNA molecules. (Figure 5.5).

## 2.2 Contact and desolvation

The interface region where two proteins physically interact can be divided into three distinct areas, as illustrated in Figure 5.6: *i*) the contact area, *ii*) the desolvated area (which incidentally also includes the contact area), and *iii*) the solvent accessible area. The contact area is where atoms form both proteins touch. The desolvated area is the part that is not (or less) solvent accessible in the complex than in the free protein. The solvent accessible area the part of the protein that is still accessible when in complex with the other protein, that is, it is where the protein contacts the surrounding water (solvent).

Some general characteristics of PPI interfaces are given in the Panel “Protein-Protein interface characteristics”.

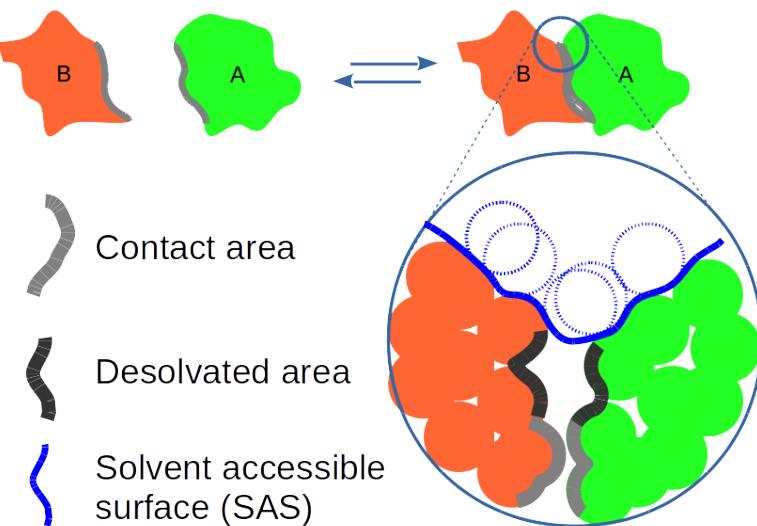


Figure 5.6: Contact and Desolvation. At the interface between two interacting proteins, A and B, we can discriminate the solvent accessible surface (either in the single protein, or as shown here in the complex), the desolvated surface and the contact surface. Solvent accessible surface is where the water molecules, typically represented by a  $1.2 \text{ \AA}$  radius sphere, can touch the protein atoms. Desolvated area is the part that is not (or less) solvent accessible in the complex than in the free protein. Contact area is where atoms from both protein touch.

### Protein-Protein interface characteristics

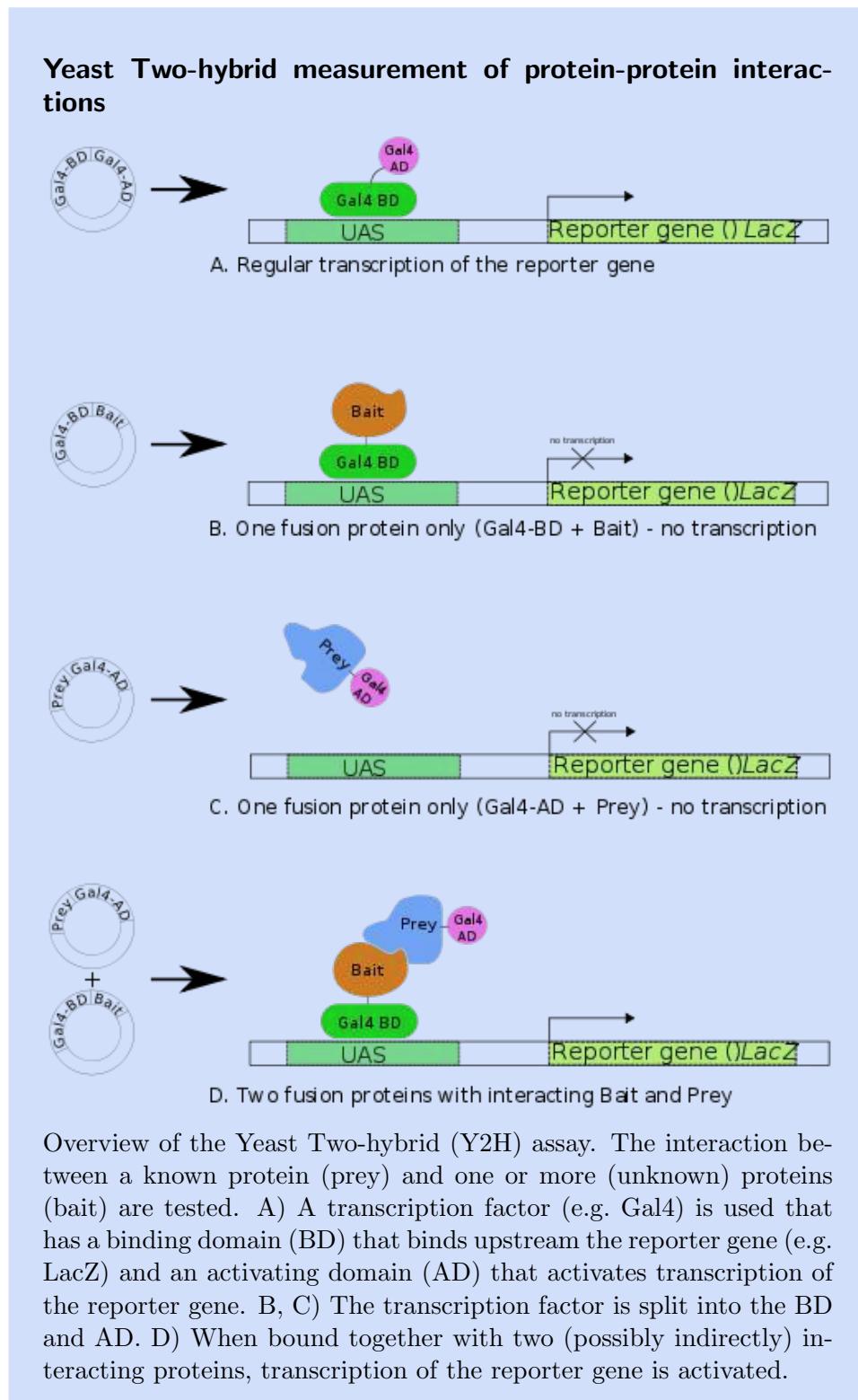
It is instructive to review some typical characteristics of protein interface regions. Here, we give a short summary, partly based on (Guharoy and Chakrabarti, 2005, 2007):

- They have relatively small size, compared to many proteins: only about 20 amino acids on average.
- The contact area is typically at least  $550 \text{ \AA}^2$
- The size of the desolvated area is typically around  $800 \text{ \AA}^2$ . Note that solvent accessible surface area lost (desolvated) is less than the contact area, as shown in Figure 5.6
- The fraction of protein surface involved is relatively low, typically around 10-20%. Not surprisingly this is lowest for dimers (12%), more for trimers (17%), but still only 20% on average for tetramers.
- Four out of five interfaces are quite flat. Note that protein-protein interactions are often described in terms of fitting structures together and several tools rely on surface shape complementarity; in reality that is a minor factor.

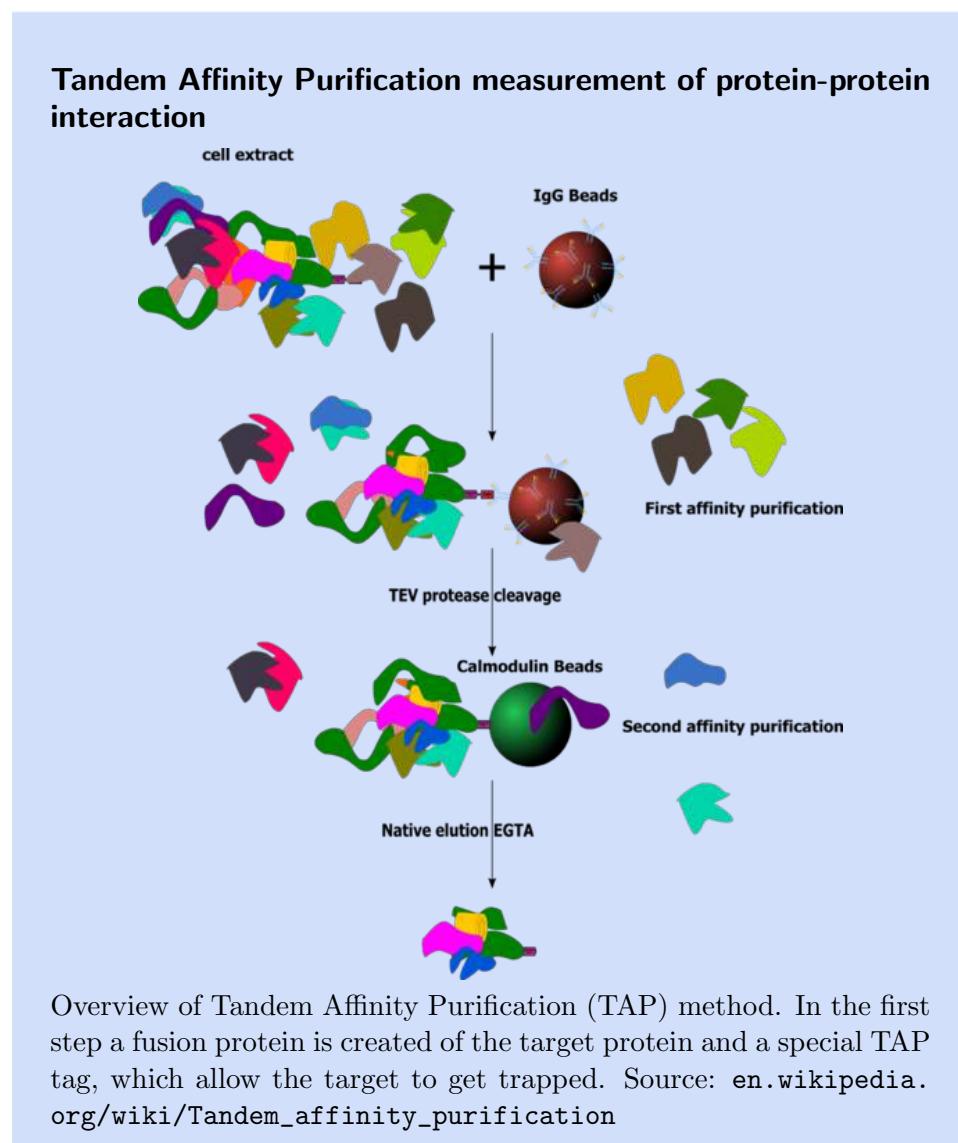
- $\alpha$ -helices seem to be overrepresented at the surface (50% vs. 30% on average), while coil/loop structures are underrepresented at the surface (20% vs. 50% on average). The reasons for this are not quite well understood, but please refer to Chapter 9 “Structural Property Prediction” for more on secondary structure propensities.

- While the interaction surface is more hydrophobic than the rest of the surface, it is still less hydrophobic than the protein core.

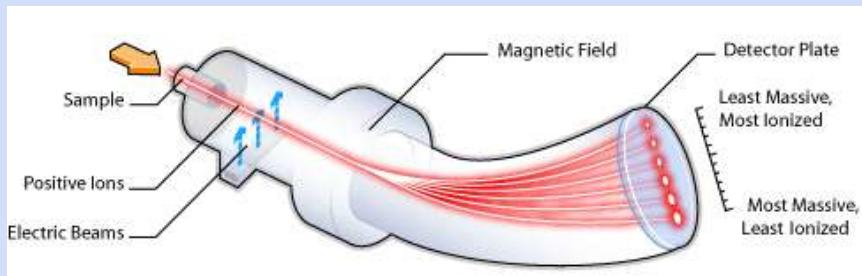
In addition, the number of interactions in different species varies. For example, datasets on yeast give rise to estimates of about 4-5 interactions on average per protein (Zotenko *et al.*, 2008), so the 6000 gene yeast genome would have 24,000-30,000 interactions, and we can estimate humans have around 100,000 (for 20,000 genes/proteins).



Source: [en.wikipedia.org/wiki/Two-hybrid\\_screening](https://en.wikipedia.org/wiki/Two-hybrid_screening)



## Mass Spectrometry



Schematic diagram of a Mass Spectrometer (MS). Currently, electrospray ionization MS and matrix-assisted laser desorption ionization (MALDI) MS are commonly used. The principle relies on an electrostatic field to accelerate charged protein molecules (or fragments), and then measuring the mass-dependent variation in their flight path (either direction or speed of flight) through a magnetic field. Proteins are first digested into peptides, which are then separated according to molecular mass and charges. A detector generates a so-called mass spectrum that displays ion intensity as a function of the mass-to-charge ratio (often denoted  $m/z$ ). As a step toward further identification, peptides can be identified in a subsequent MS phase of fragmentation and mass analysis. This technique is called tandem mass spectrometry (MS/MS). The molecular mass spectra, or fingerprints, of the peptides are more precisely determined in the second analyzer, which are used for identification by searching a database of theoretical spectra based on known digestion and fragmentation patterns of proteins. In practice, additional information such as the molecular mass of peptide fingerprints, peptide sequence, M/W, and pI of the intact protein, and even species names are used to obtain unique identifications of peptides from a particular protein. A basic requirement for peptide identification through database matching is the availability of the protein sequences. Thus, this method only works well with organisms that have been sequenced, and preferably well-annotated.

Source: <http://www.scq.ubc.ca/image-bank/>

## Experimental methods for determining PPIs

Many experimental methods exist that can measure PPIs, with a wide variety of detail and accuracy. The most basic type of information is whether two proteins interact. In addition, one can measure the binding affinity, or strength of the interaction. More detailed, one

can identify residues that participate in the interaction, i.e. define the binding region. And, finally, one may elucidate the full structure of the binding complex.

The last category, of elucidating the complex structure, is special, as it requires measuring the structure at some level of detail. For protein complexes, this can currently only be done using X-ray crystallography or electron microscopy, which are further discussed in Chapter 2 “Structure determination”. These two techniques immediately establish that these proteins can interact, and in addition define precisely what the binding region is. There is a caveat for crystallography, where it is often difficult to distinguish real interfaces from crystal packing contacts, as already mentioned in Chapter 2. Moreover, typically no information is gained on the strength of the interaction.

For the other categories of PPI information, typically high-throughput methods exist, that sacrifice accuracy for speed, as well as more accurate, but slower methods that are often used to validate the high-throughput data. Two of the best known high-throughput experimental techniques for detecting whether two proteins interact are yeast two-hybrid screening (Y2H), shown in Figure 2.2 and tandem affinity purification (TAP), see Figure 2.2. In Y2H, the binding of two proteins leads to activation of a reporter gene which is read out using a fluorescent signal. In TAP, a single target protein is filtered out of solution together with the target protein and any other proteins that may bind to it, and identified using mass spectrometry (MS), see Figure 2.2.

TAP determines protein partners quantitatively and *in vivo* without prior knowledge of complex composition, which is not easily done with Y2H. TAP may also be easier to perform, provides higher yields, and more resilient to contamination than Y2H. One of the main disadvantages is the required TAP tag, which might obscure binding to other proteins, or in other ways change relevant properties of the protein of interest.

In addition to its use in TAP for protein identification, MS is also used to analyze cross-linked complexes of bound proteins (Tang and Bruce, 2009). A recent but rather technical review of proteomic methods, mainly MS-based and including those for protein-protein interactions, is given in (Zhang *et al.*, 2013).

### 3 Functional classes

There are many ways in which protein function may be classified. Here, we will discuss protein function based on several broad features related to

protein function:

- small ligand binding
- protein-protein interactions
- protein-DNA interactions
- transmembrane proteins
- intrinsically disordered proteins
- post-translational modification
- functional motions

Protein-protein interactions were already covered in detail in the previous section, Section 2. For each of the others, we will describe examples in the sections below.

### 3.1 Protein Small Ligand Binding

Proteins can bind small ligands such as substrates, inhibitors, activators and neurotransmitters in specific sites. This binding can be important for the function of the proteins such as receptors, transporters, channels and enzymes. Receptors can be activated when they bind a small ligand. An example of such a receptor is the G-protein coupled receptor (GPCR) Figure 5.7. Transport proteins can bind small ligands and transport them from one location to another, like hemoglobin does for oxygen (Figure 1.1). Ion-channels translocate small ligands through cellular membranes, like the sodium-potassium pump Figure 5.8, discussed below.

Many enzymes are activated when they bind a cofactor. When a cofactor binds outside of the active site, and this leads to a conformational change of the protein which affects catalytic turnover of the enzyme, this is called allosteric regulation. The key notion of allostery is that binding of the cofactor happens at a different location than where the ligand binds. The example of hemoglobin was already mentioned above (Figure 1.1).

Related to allostery, but distinct from it, is that of induced fit. Here the binding site changes shape as the ligand binds to the same binding site, so the conformational change happens at the same place as where the binding occurs. An example of that is hexokinase (Figure 5.9).

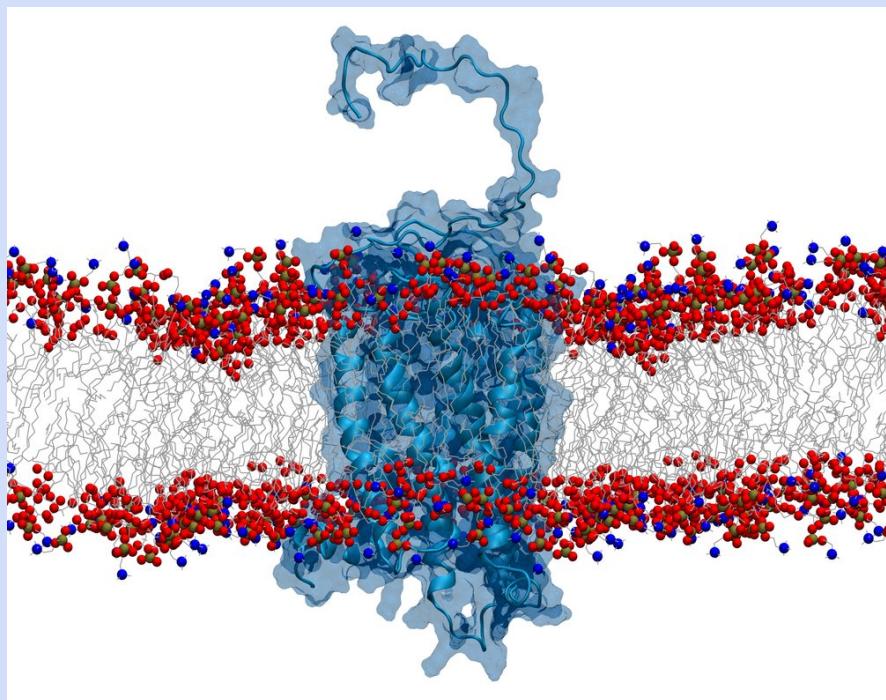
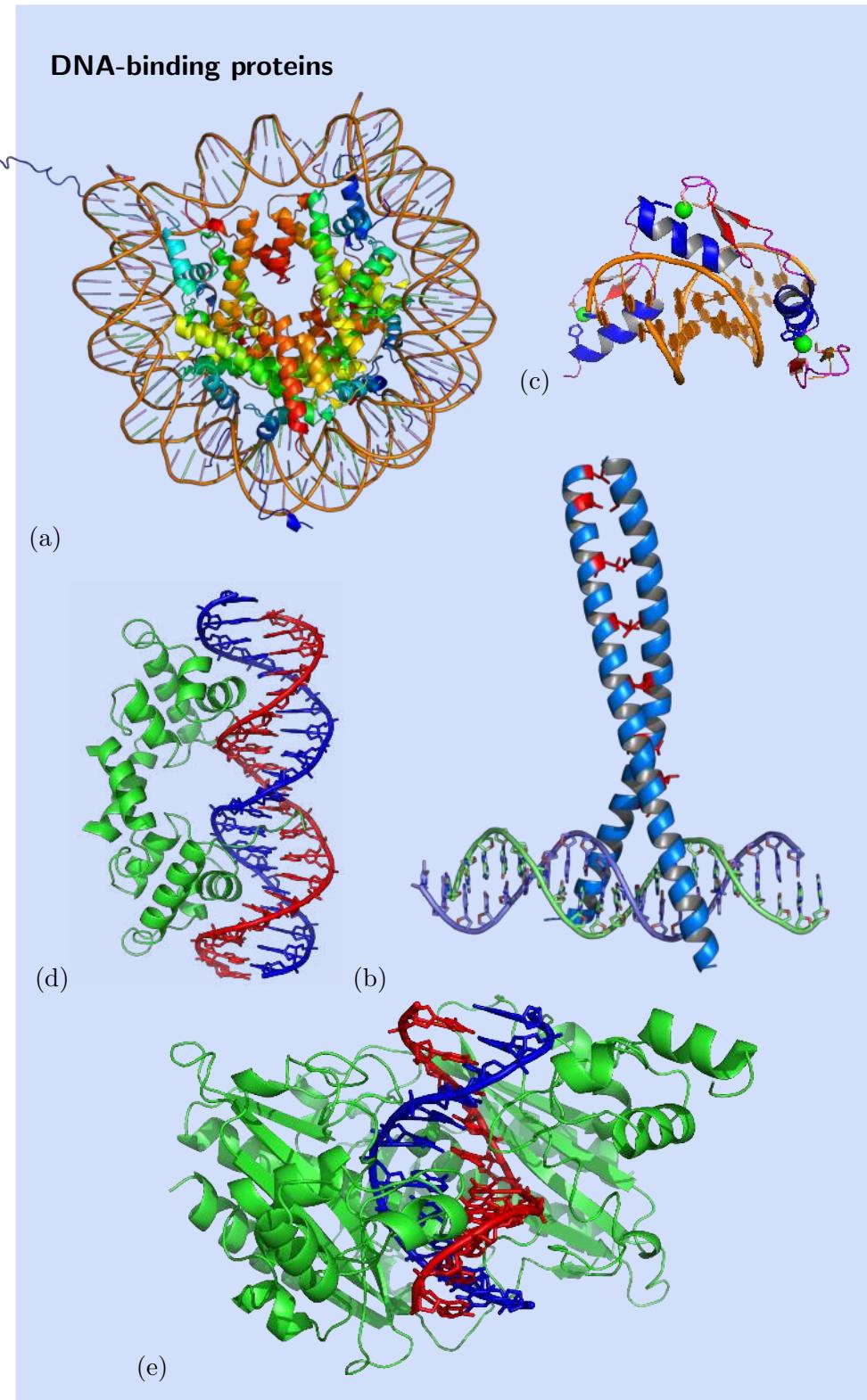
**G-protein coupled receptor membrane receptor family**

Figure 5.7: G-protein coupled receptor (GPCR) in a lipid bilayer membrane. This huge class of receptors typically receives a signal on the extra-cellular side (outside) of the membrane. The signal could be chemical (a molecule binding), physical (stress in the membrane, or even temperature), or other (e.g., light). Receiving the signal triggers a conformational change on the outside of the receptor, which is ‘mechanically’ relayed towards the inside, typically by the reorientation of a pair of helices. This causes a conformational change on the inside which is subsequently detected by other proteins (for example the ‘G-protein’ after which the GPCR family is named, but there are other downstream signalling routes as well). Source: <http://oldeurope.deviantart.com/art/GPCR-in-Lipid-Bilayer-focus-129477640>



DNA-binding proteins. (a) Histones are responsible for packing the DNA into chromatin and also play a role in gene regulation. Specific interactions are mediated by particular motifs, we give several examples here. (b) Leucine zippers consist of two halves, each one of them having an alpha-helix with a leucine at every 7th amino acid. When one leucine comes in direct contact with another leucine on the other strand every second turn adhesion forces hold them together, thus forming a “zipper”. The outside of this structure is hydrophilic, while the core formed by the leucins is hydrophobic and it is this core that is absolutely required for DNA binding. (c) Zinc-finger is a small motif characterized by being stabilized by one or more zinc ions, which coordinates the motif’s fold (the zinc atom is marked in green; note that this protein contains three zinc-finger domains). They are commonly found in transcription factors. The zinc ion is positively charged and contributes to interacting with the negatively charged phosphate residues. The specific structure, amount and placement of zinc-finger motifs in a binding domain characterize the specific DNA sequence recognized by that domain. (d) Helix-turn-helix in part of the  $\lambda$  repressor of bacteriophage lambda. It is quite simple and consists of two crossed alpha-helices joined by a short strand. In most cases, one of the helices is responsible for DNA recognition by binding to the major groove, while the other helice stabilizes the interaction between protein and DNA. (e) A more elaborate example of specific recognition of a target DNA sequence is the restriction endonuclease EcoRV (green) in complex with its substrate DNA (red and blue).

Sources (creative commons):

[http://en.wikipedia.org/wiki/Leucine\\_zipper](http://en.wikipedia.org/wiki/Leucine_zipper)

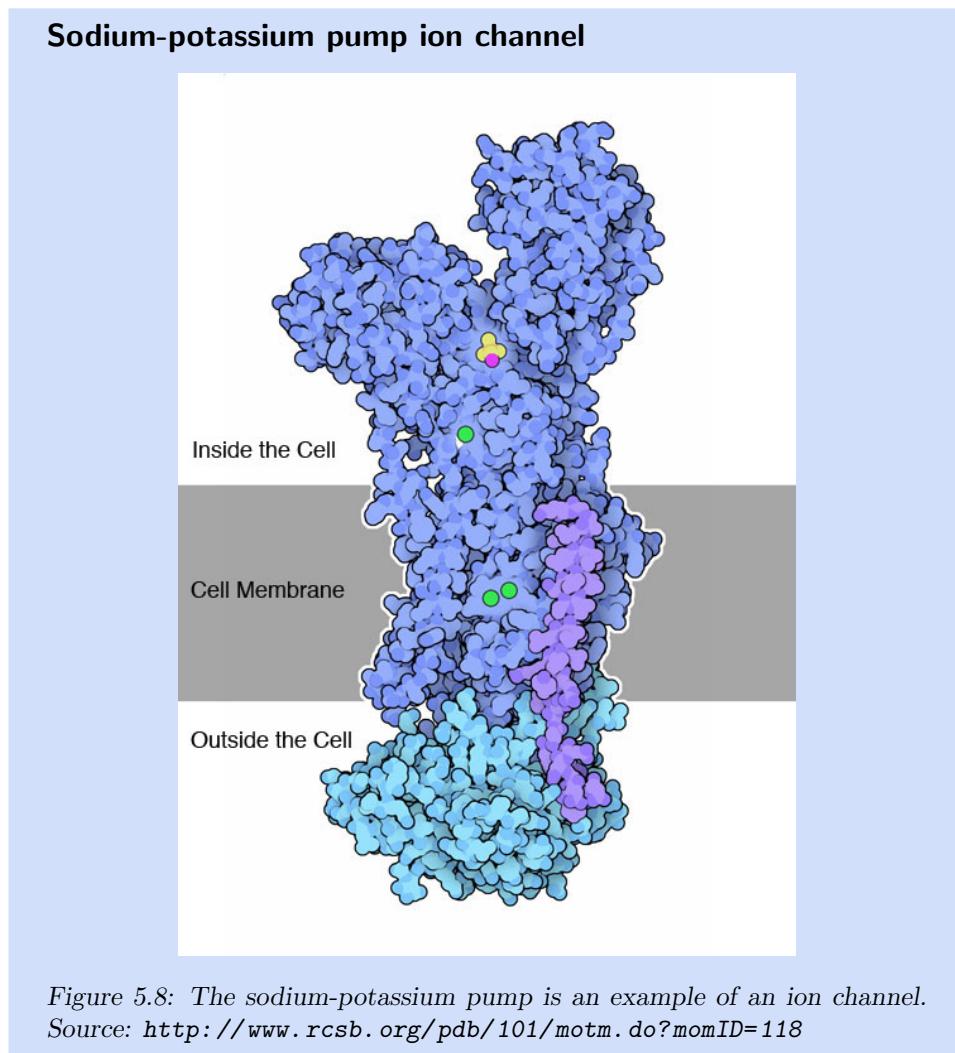
[http://en.wikipedia.org/wiki/Zinc\\_finger](http://en.wikipedia.org/wiki/Zinc_finger)

<http://en.wikipedia.org/wiki/Helix-turn-helix>

### 3.2 Protein-DNA interactions

Proteins that bind DNA contain specific domains that bind either to a single or double-stranded DNA. These proteins have different functions and can regulate gene transcription, DNA replication, DNA cleavage or DNA packing. The interaction between Protein and DNA can be either sequence specific or non-specific (Pabo and Sauer, 1984). Non-specific interactions are formed e.g. by basic residues of the protein interacting with acidic sugar-phosphate backbone of the DNA, such as histones Figure ??a. Specific interactions are mediated by particular motifs, we give several examples here. Leucine zippers are hydrophilic on the outside and have a hydrophobic core formed by the leucins which is required for DNA binding, see Figure ??b. Zinc-finger is a small motif containing a zinc ion that contributes to interacting with the DNA Figure ??c. The Helix-turn-helix consists of two crossed

alpha-helices joined by a short strand, with one of the helices interacting with the DNA Figure ??d. More elaborate modes of proteins binding to DNA exist, e.g. Figure ??d or Figure 5.4. Protein RNA interactions are a different class altogether, see e.g. Figure 5.5.

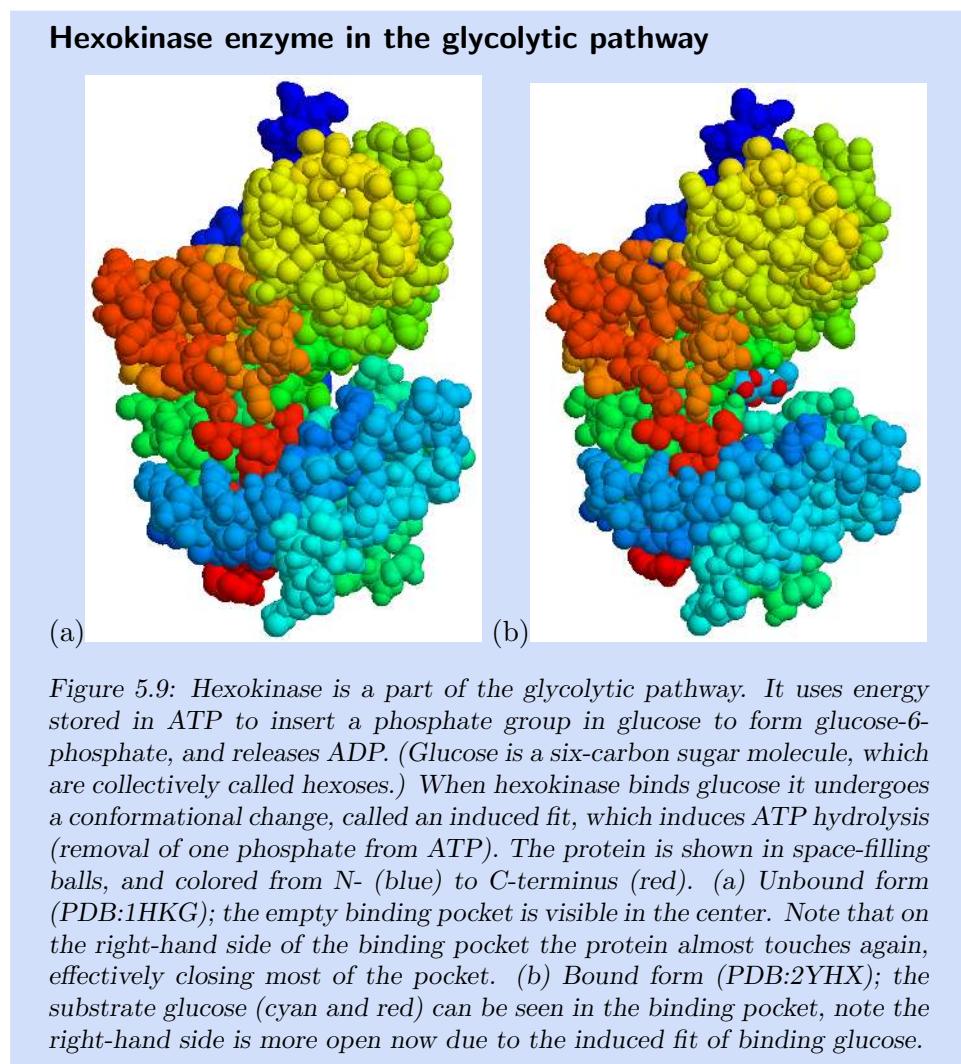


### 3.3 Transmembrane Proteins

Transmembrane (TM) proteins span through biological membranes and have a variety of different functions. Unlike other proteins, TM proteins have a hydrophobic surface, as the inside of the membrane consists of hydrophobic lipid tails. When transmembrane proteins form a pore through the membrane, the inside may even be partially charged or hydrophilic to facilitate the transport of ions. The structure is difficult to determine since TM pro-

teins are difficult to express, purify, crystallize and stabilize in solution, as we already saw in Chapter 2. Examples of TM proteins are the GPCR receptor family (Figure 5.7 in previous section), and the sodium-potassium pump shown in Figure 5.8.

30% of all genes contain a TM region so finding a way to determine the structure of these proteins is very important. However, TM regions, and in particular TM spanning helices, are relatively easy to predict, due to their strong hydrophobic signal, and specific length for crossing the membrane. We will return briefly to TM proteins in Chapter 9 “Structural Property Prediction”.



### 3.4 Functional motions

We already saw that conformational changes in proteins are associated with their function; we can refer to this broadly as ‘functional motions’. In terms of dynamics or flexibility, proteins can exhibit very different behaviours. Examples are the activation of a GPCR receptor upon binding of the ligand (Figure 5.7), the allosteric binding of oxygen to hemoglobin (Figure 1.1), or the induced fit upon binding of glucose to hexokinase (Figure 5.9).

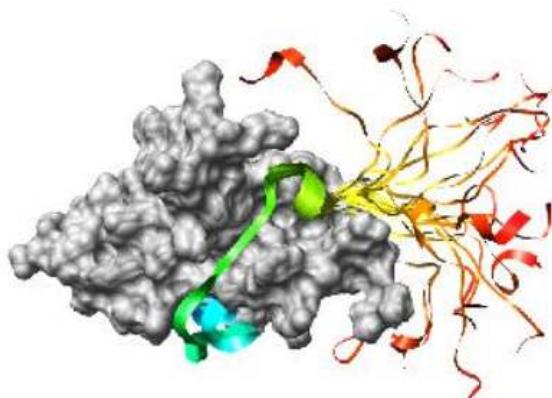
Not all proteins are equally dynamic. For example, the axin protein that forms axin-filaments is considerably less flexible than lysozyme, which has an active site that closes upon ligand binding. Figure 1.9C shows the two domains of lysozyme that move upon ligand binding). Crystallin, the main protein component of the eye lens in vertebrates, is as stiff as a protein can get as it should survive throughout the lifespan of the organism – that is, up to 100 years or longer!

We will return to protein motions, flexibility and dynamics in Part III.

### 3.5 Intrinsically disordered Proteins

So far we have thought in terms of a well-defined protein structure leading to the function of that protein, for example by forming a suitable binding site upon folding. However, there is an important class of proteins that will occur unfolded in the cell. These proteins are often referred to as disordered or natively unfolded proteins. There is an even larger group of protein that contains regions, or sometimes very long stretches that have a disordered structure. Some of these disordered proteins or protein regions may take a specific structure upon binding to another protein or other substrate (Fuxreiter *et al.*, 2007). In fact, this transition from an unfolded to a folded state, may be important for the function of the protein.

Historically, the importance of intrinsically disordered proteins was only recognised a couple of decades after the determination of the first protein structures. One of the reasons for this late recognition is that we tend to understand the functional mechanism of a protein by considering its structure as determined by X-ray or NMR experiments. If the structure is not observed, it may simply be due to a failed experiment. Moreover, such regions may actually make such structure resolving experiments much more difficult, for example, because the protein will not crystallize, if they contain a large disordered region (see Figure 5.10a). For this reason, computational programs were developed that could discover such disordered regions, see Chapter 9. By using such disorder prediction methods on fully sequenced genomes we now know that many proteins contain large disordered segments: in 33% of eukaryotic, in 2% archaea, in 4.2 % bacteria (Ward *et al.*, 2004). Generally speaking, the more complex the organism, the larger the fraction of proteins with disordered regions.



*Figure 5.10: Many disordered regions contain small binding motifs (green region) that can bind to structured binding partners, and that obtain a fixed structure upon binding. Note that the flanking regions often remain unstructured, even upon binding. Here multiple possible configurations of this flanking regions are shown in orange/red.*

Moreover, proteins that contain disordered regions have been associated with regulatory and signalling functions, and are often hubs in protein-protein interaction networks. This can be understood, if we recognise that such disordered regions can contain many small binding regions (see Figure 5.10b), that can bind to other proteins, typically with a native structure.

### Post-translational modification

Post-translational modification (PTM) is all the (chemical) changes that may happen to a protein structure after it comes off the ribosome. Common examples are the attachment of a chemical group, like phosphorylation, ubiquitination and glycosylation. These groups are attached to particular sidechains in the protein. Specialized enzyme systems exist to perform each of these PTM reactions. A bit more detail on the three examples mentioned above:

**Phosphorylation** is the coupling of phosphate group to a hydroxyl (OH) group in serine, threonine or tyrosine, or to an amide (NH) group in histidine, arginine or lysine. Depending on the protein, and the specific location of the phosphorylation, this can modulate the function of the protein in various ways (e.g. activating or inactivating).

**Ubiquitin** is a protein by itself, and can be linked via its C-terminus to a lysine sidechain (using an isopeptide bond, which is similar to the way aminoacids are linked together), or to the N-terminus of the protein via a peptide bond, except for the fact

that normal peptide bonds are created at the ribosome. Polyubiquitination is possible by stringing additional ubiquitin proteins together. This typically targets the protein for degradation through the lysosomal system, either the protein due to conformational changes in the protein structure.

**Glycans** are a broad class of sugars, which can be attached in many different ways to a protein. They can be linked to the nitrogen in asparagine or arginine, or to the hydroxyl in serine, threonine or tyrosine. Single sugar units can be attached, but these may also be expanded into linear or branched structures of varying sizes. They can be as large as the protein itself, and can have a wide variety of effects on the protein's function.

Another class of PTM are modifications of sidechains, rather than additions to them. Examples are hydroxylysine and hydroxyproline. Also the formation of so-called cysteine-bridges, also known as sulfur bridges, between two cysteine residues that are adjacent in the structure. Here, the hydrogen atoms at the sulfur are released, and a chemical bond is formed between the sulfur atoms of both cysteines (hence, sulfur bridge). This motif often occurs in proteins that need to enhance their stability for various reasons, for example for acidic or high-temperature environments, or to protect against degradation by proteases.

## 4 Protein function in the context of the living cell

The insides of a cell are more like a thick broth than the watery solutions one may imagine, or that one may have seen in the lab. Figure 5.12 shows a stunning and beautifully detailed three-dimensional model of how proteins are jumbled together inside a neuronal synapse. Proteins were put inside the volume at their approximate known locations (precise locations cannot be determined), and the number of each protein added are also based on real data (Wilhelm *et al.*, 2014). This picture would not look much different in any type of cell. So, it is fair to say that the typical protein in a typical cell has ample opportunity for interaction.

To be able to perform all the functions described in the previous sections, proteins need to be dynamic and interact. Ultimately, all transfer of information within living cells (signalling, regulation), and also in between cells, depends on proteins interacting, and (dynamically) responding to those interactions. In Chapter 11 “Function Prediction” we will dive deeper into prediction of functional regions and protein interactions, and in Part III we will devote several chapters to protein dynamics.

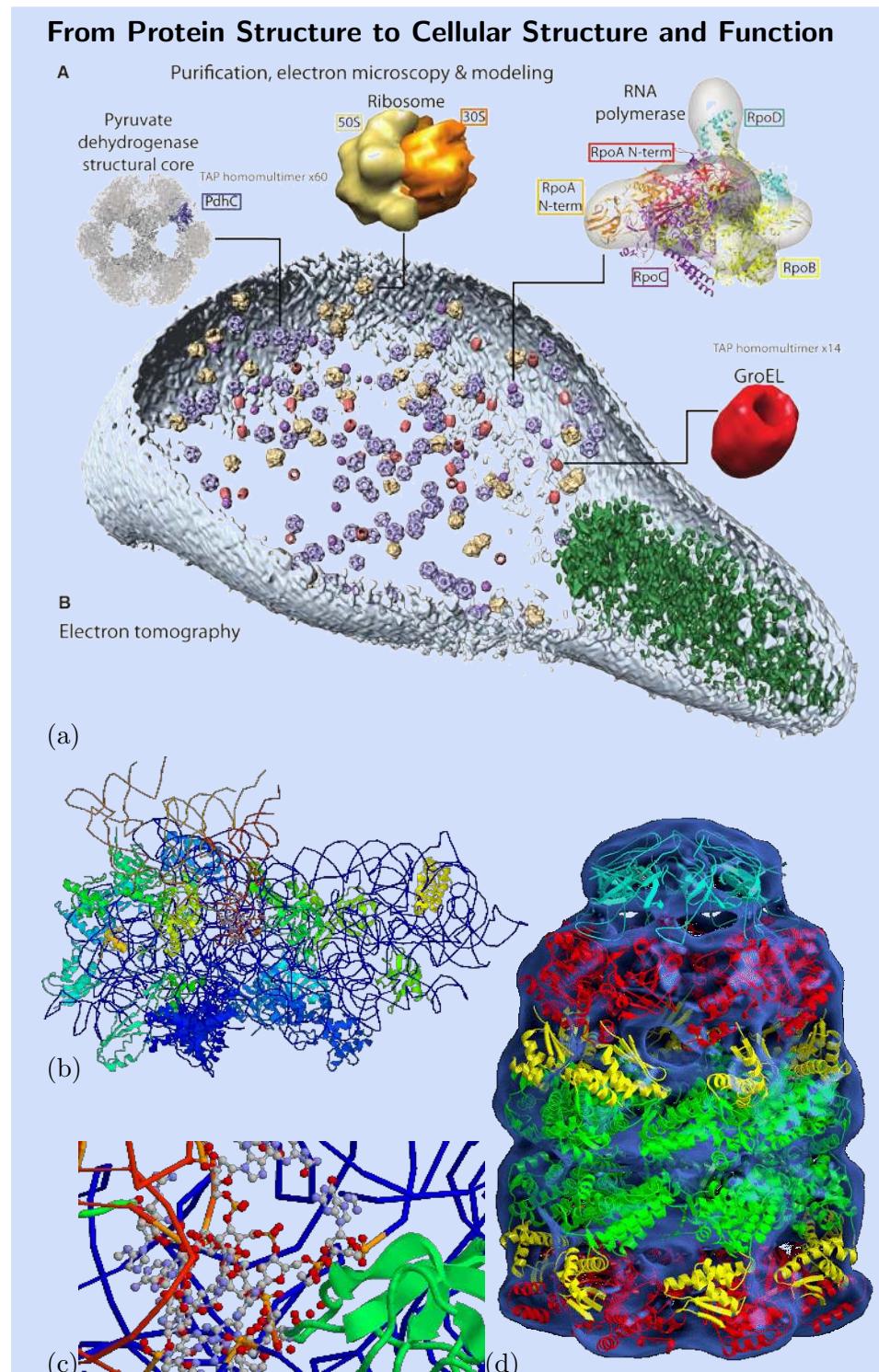


Figure 5.11: protein structure with cellular function (caption on next page)

Figure 5.11 (on previous page) This image connects – quite directly – protein structure with cellular function. (a) In this electron tomography image of mycobacterium pneumonia, it is possible to count individual protein complexes (only the larger ones which have an identifiable shape), so we know for these complexes exactly how many there are in one cell. Here, they identified 100 copies of the pyruvate dehydrogenase enzyme complex (core element of energy metabolism); 140 ribosomes (which translate messenger RNA (mRNA) into protein molecules); 300 copies of the RNA polymerase complex, which transcribes active genes into mRNA; and 100 GroEL complexes, which is a folding chaperone that recognizes misfolded proteins, induces unfolding and subsequent re-folding. Taken from (Kühner et al., 2009). (b) The whole ribosome, and (c) a close-up of the codon-binding area of the ribosome; here we have made the final step from cellular function down to atomic detail. (d) The GroEL folding chaperone complex. Taken from (Ranson et al., 2001).

## 5 Key points

- Protein function is defined on several levels: by direct interactions with other molecules (e.g. being part of protein complex) and by indirect interactions (e.g. being part of signalling pathway).
- A ligand is the molecule a protein binds to perform its function.
- The ‘active site’ or ‘functional site’ is the region with which a protein binds to its ligand.
- Protein function is strongly linked to protein-protein interaction, when the ligand is a protein, but also when being regulated by another protein.
- Protein-protein interactions can be represented by networks. Network methods then applied to build on protein functions.
- For many proteins, function is coupled to conformational changes (motion, dynamics or flexibility).
- Proteins function within the context of a living cell, which is crowded with many molecules: proteins and other types.

## 6 Further Reading

- ‘Biochemistry’, Berg *et al.* (2002)
- ‘Molecular Biology of The Cell’, Alberts (2015)

## References

- Alberts, B., editor (2015). *Molecular Biology of The Cell*. Garland Science, Taylor & Francis Group, LLC.,
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**(6), 450–461.
- Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002). *Biochemistry*. W H Freeman, New York.



Figure 5.12: (A) A section through the synaptic bouton, showing 60 proteins in the estimated copy numbers, and in positions determined according to imaging data and literature. (B) High-zoom view of the active zone area. (C) High-zoom view of one vesicle within the vesicle cluster. (D) High-zoom view of a section of the plasma membrane in the vicinity of the active zone. Clusters of syntaxin (yellow) and SNAP 25 (red) are visible, as well as a recently fused synaptic vesicle (top). The graphical legend indicates the different proteins (right). Displayed synaptic vesicles have a diameter of 42 nm. Taken without permission from Wilhelm et al. (2014).

- Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T. and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**(13), i223–i231.
- El-Kebir, M., Soueidan, H., Hume, T., Beisser, D. et al (2015). xHeinz: an algorithm for mining cross-species network modules under a flexible conservation model. *Bioinformatics*, **31**(19), 3147–3155.
- Fuxreiter, M., Tompa, P. and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**(8), 950–956.
- Guharoy, M. and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(43), 15447–15452.
- Guharoy, M. and Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics*, **23**(15), 1909–1918.
- Jacobsen, A., Bosch, L.J.W., Martens-de Kemp, S.R., Carvalho, B. et al (2018). Aurora kinase A (AURKA) interaction with Wnt and Ras-MAPK signalling pathways in colorectal cancer. *Scientific Reports*, **8**(1), 7522.
- Krieger, E. and Vriend, G. (2014). YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, **30**(20), 2981–2982.
- Kühner, S., van Noort, V., Betts, M., Leo-Macias, A. et al (2009). Proteome Organization in a Genome-Reduced Bacterium. *Science*, **326**, 1235–1240.
- May, A., Brandt, B.W., El-Kebir, M., Klau, G.W. et al (2016). metaModules identifies key functional subnetworks in microbiome-related disease. *Bioinformatics*, **32**(11), 1678–1685.
- Pabo, C. and Sauer, R. (1984). Protein-DNA recognition. *Annu Rev Biochem*, **53**, 293–321.
- Ranson, N.A., Farr, G.W., Roseman, A.M., Gowen, B. et al (2001). ATP-Bound States of GroEL Captured by Cryo-Electron Microscopy. *Cell*, **107**(7), 869–879.
- Tang, X. and Bruce, J.E. (2009). Chemical Cross-Linking for Protein-Protein Interaction Studies. *Methods In Molecular Biology*, **492**, 283–293.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, **337**, 635645.
- Wilhelm, B.G., Mandad, S., Truckenbrodt, S., Krohnert, K. et al (2014). Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. *Science*, **344**(6187), 1023–1028.
- Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. and Yates, J.R. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.*, **113**(4), 2343–2394.
- Zotenko, E., Mestre, J., O’Leary, D.P. and Przytycka, T.M. (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Comput Biol*, **4**, e1000140.



**Part II**

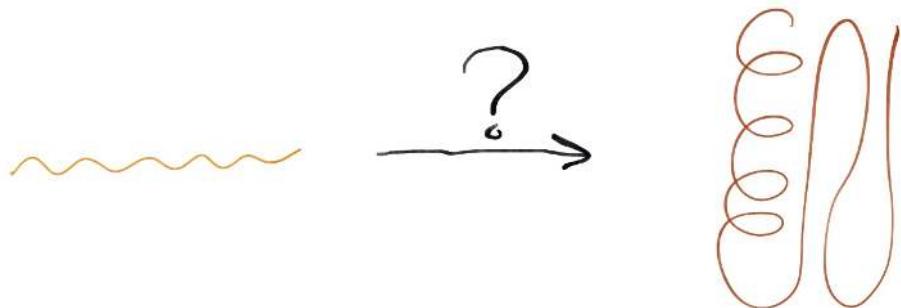
**Structure Prediction**



# Chapter 6

## Introduction to structure prediction

Sanne Abeln  Jaap Heringa  K. Anton Feenstra 



[arxiv.org/abs/1712.00407](https://arxiv.org/abs/1712.00407)

\* editorial responsibility



# 1 What is the protein structure prediction problem?

## 1.1 Predicting the structure for a protein sequence

This chapter revolves around a simple question: “given an amino acid sequence, what is the folded structure of the protein?” (Figure 6.1) Even though this seems like a simple question, the answer is far from straightforward. In fact, whether we can give an answer at all depends heavily on the **sequence in question** and the availability of protein structures that can be used as **modelling templates**. The question is of growing importance, however, as the gap between protein **structures and sequences** continues to widen; while the number of structures deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000) continues to rise rapidly<sup>1</sup>, the number of sequenced genes rises much faster. Fortunately, recently developed methods can use these large resources of sequence data to increase the quality of some predictions. Here, we will give an overview of **current** structure prediction methods, and describe some tools that provide insight into **how reliable** the structure predicted will be.

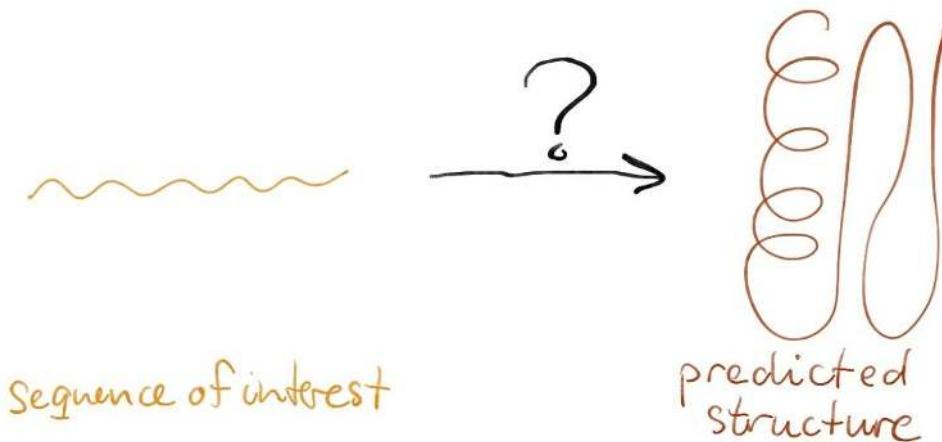


Figure 6.1: Structure prediction methods try to answer the question: given an amino acid sequence, what is the folded protein structure?

The typical problem is that we want to generate a structural model for a protein with a sequence but without an experimentally determined structure. In this chapter, we will build up a workflow for tackling this problem, starting from the easy options that, if applicable, are likely to generate a good structural model, and gradually working up to the more hypothetical options whose predictions are much more uncertain.

Another very important remark is in place here: **the modelling strategy** should depend heavily on **what we want to do with the structure**. Do we

<sup>1</sup><https://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100>

want to predict where the functional site of the protein is, whether a specific substrate binds, or if a certain residue may be exposed to the surface? These different questions imply a different degree of accuracy we need for an answer, and may lead to choices regarding technology and methods to carry out these predictions.

Throughout this discussion, it is important to keep in mind that one of the most important aspects of any scientific model is whether a research question may be answered with the model produced or not. Even if we do have an experimental structure available, some of these questions may not be straightforward to answer; we will come back to this issue later in the chapter.

## 1.2 Structure is more conserved than sequence

Almost all structure prediction relies on the fact that, for two homologous proteins, structure is more conserved than sequence (see Figure 6.2). The real power of this observation manifests itself when we turn this statement around: if two protein sequences are similar, these two proteins are likely to have a very similar structure. The latter statement has very important consequences. It means that if our sequence of interest is similar to a protein sequence with a known structure, we have a good starting point for a structural model. We thus exploit that sequence similarity points towards a homologous relationship to predict structure.

The vast majority of accurate structure prediction methods use structure conservation as an underlying principle; methods that have been developed to deal with more difficult modelling questions to exploit the sequence-structure-conservation relation in an advanced manner, as discussed towards the end of this chapter.

## 1.3 Terminology in structure prediction

## 1.4 Different classes of structure prediction methods

We can classify structure prediction strategies into two categories of difficulty: template-based modelling, and template-free modelling (see Figure 6.4). In the first case, it is possible to find a suitable template for the target sequence in the PDB as a basis for the model, whereas for template-free modelling no such experimental structure is available. Note that it may not be trivial at all to find out in which of these two categories a structure prediction problem falls. Only if we can find a close homolog – based on sequence similarity – in the PDB we can be sure that a template-based modelling strategy - also known as homology modelling - will suffice. With

---

<sup>2</sup>Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

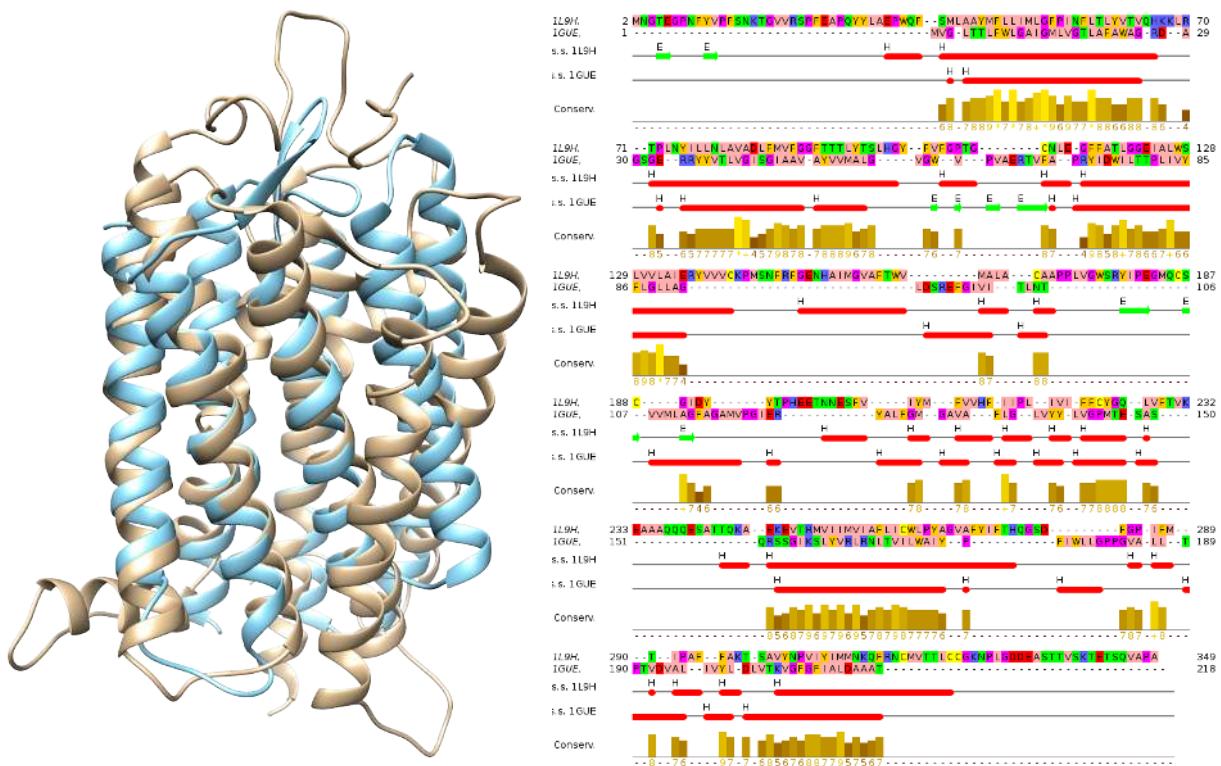


Figure 6.2: Protein structure more conserved than sequence. Here the output of a structural alignment is shown on the left, created using Chimera<sup>2</sup> (Pettersen et al., 2004). The structural alignment shows both proteins are highly similar; the RMSD is 2.3 Å over 144 aligned residues (root mean square deviation, introduced in Chapter 3). Furthermore, the function of the two proteins, one from cattle (PDB:1L9H, light brown) and one from an archaeon (PDB:1GUE, light blue), is similar: both are light-sensitive rhodopsins, used for vision and phototaxis, respectively. However, as can be seen in the sequence alignment on the right, the sequence identity is only 7%. This is lower than would be expected for any two random sequences. The alignment shown is based on the structural alignment on the left, and visualised using JalView (Waterhouse et al., 2009).

a template, the **constraints** from the alignment between the model and the template sequence, in addition to the **template structure**, will give sufficient constraints to build a structural model for the target sequence. Even in this case, however, small missing substructures in the alignment such as loops may require a **template-free modelling** strategy.

If no close homologs are available in the PDB, we may need to use more advanced template finding strategies, such as **remote homology detection** or **fold recognition** methods. If such a search is neither fruitful, we will need to resort to a **template-free** modelling strategy. In the “ab initio” approach knowledge-based energy terms are used to generate structural models based

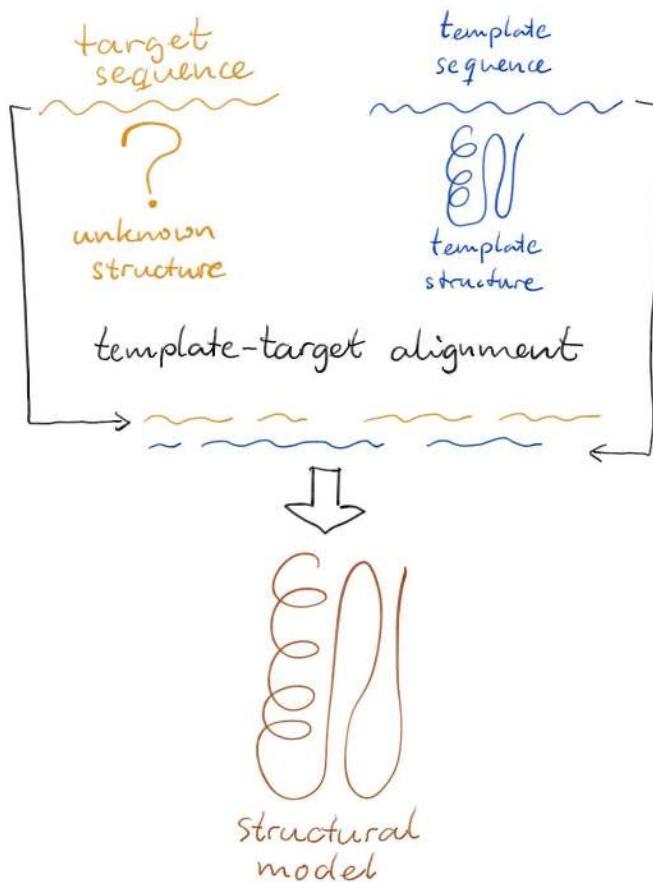


Figure 6.3: Terminology used in protein structure prediction. We start from our protein of interest (with no known structure): the target sequence. First step is to find a matching protein: a template sequence with known structure; the template structure. We then create a template-target sequence alignment, and from this alignment create the structural model which is the solution structure for our target protein.

on the sequence of the template alone. Small, suitable fragments, from various PDB structures, are then assembled to generate a large set of possible structural models. How to pick a model from this set and refine models further will be discussed in the next chapter.

Sometimes experimental data such as NMR, cryoEM or contact prediction methods can provide additional constraints for modelling protein structure. In such cases, we have a much better chance of finding a suitable model (Moult *et al.*, 2016). In fact, such constraints can even be considered an alternative to the constraints provided by homology modelling.

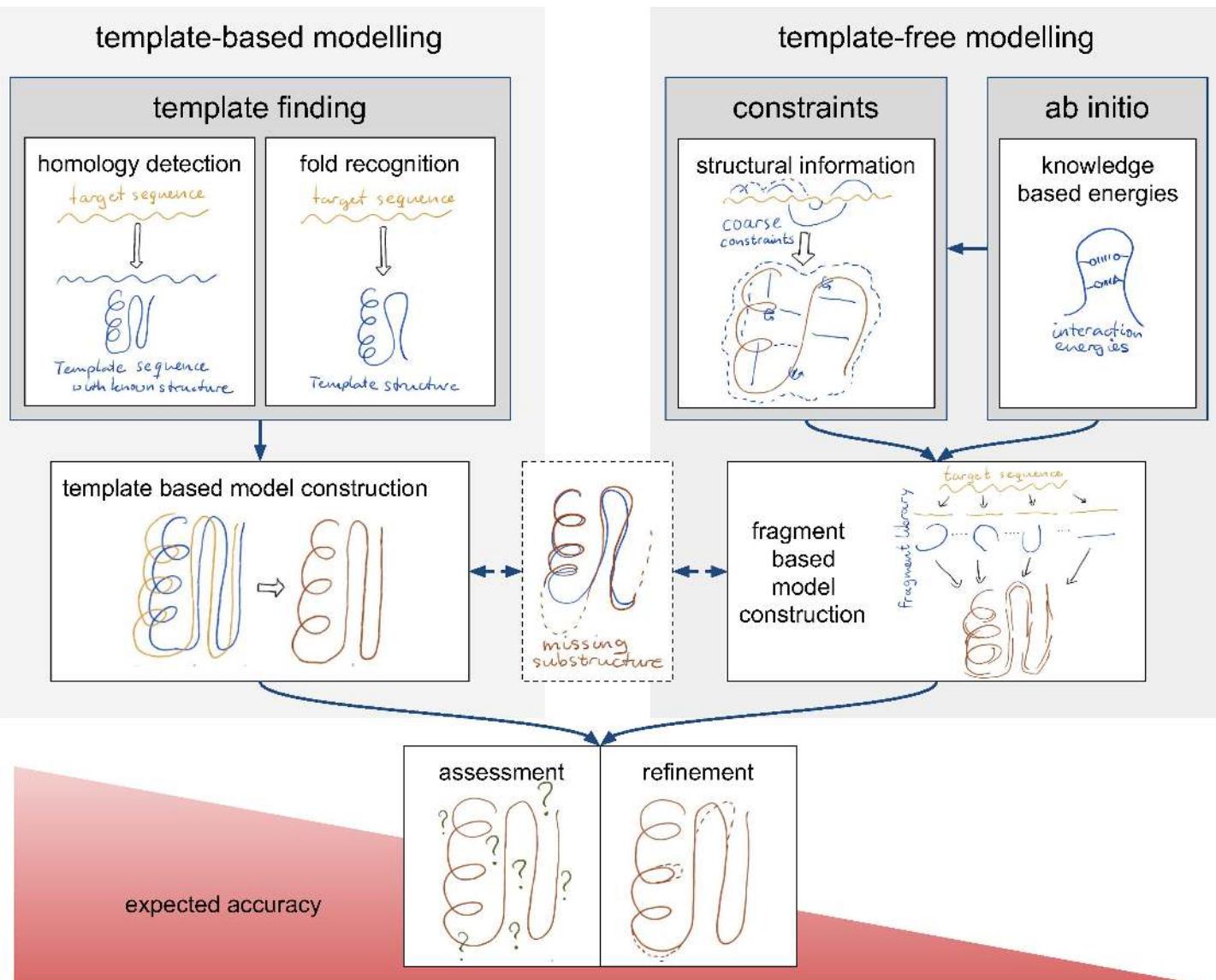


Figure 6.4: Overview of Structure Prediction. Template-based modelling: a template is found on the basis of homology between the template and the target. Fold recognition: no obvious homologous structure can be found in the PDB, we need **fold recognition** methods to find a suitable template. Template-free modelling: no suitable template for protein domains can be found. Without a template, we need to use a combination of coarse constraints from experiment or co-evolution analysis, and *ab initio* prediction. *Ab initio* methods typically work with taking fragment templates from various proteins, and assemble these into a model or decoy structure. Expected model accuracy declines from left to right: good accuracy is expected if based on homology; in contrast, *ab initio* modelling should only be considered if no other options remain.

## 1.5 Domains

So far we have implied that we may follow the above strategy for an entire protein. This is in general, however, not very sensible if a protein consists of **multiple domains** as most structure prediction methods only work well at the domain level.

This means that a sequence first needs to be split in multiple domains before we can start to make models. Such domain splitting is often ambiguous, both at the sequence and structural level (see the Panel “Domain prediction”). It also means that in practice **multiple templates might be needed** to model the structure for a **single** target sequence.

Adequate models of the individual domains being within reach also do not imply that we succeed in obtaining an adequate model for the entire protein structure; combining models built for various domains into a single structure is far from trivial. In fact, **resolving the orientation of modelled domains** with respect to each other is an unsolved problem, unless there is a suitable, homologous, template available where the domains have the same orientation.

Sometimes coarse constraints on the domain orientations, such as data from **small-angle scattering** experiments, **distance restraints** from NMR, **chemical cross-links** or **co-evolutionary relationships** can help to put different homology models in the correct orientation. Such constraints have also helped the CAPRI community progress, who focuses on the closely related problem of modelling protein assemblies.

## 2 Assessing the quality of structure prediction methods

Generally, as with any prediction problem, we can assess the quality of a prediction if we have a true answer to the question. Here, the truth will be represented by an experimentally determined protein structure (of high quality). Fortunately, nowadays there are very many structures in the PDB (see Section 1.1). Simply assessing how well a method performs over this set is, however, problematic as methods have been **trained on (a subset of)** this dataset. This means in particular that they generally will be able to yield good models for sequences that are within this dataset and homologs of those sequences. To assess how well a method performs, a completely independent data set - also **not containing homologs** of the training set - is thus required.

A second problem is that even if we assess performance on an independent test set, we cannot be sure performance generalizes well to proteins with structures that remain unknown; the test set consists of proteins whose structures could be experimentally determined, and may therefore not be

### Domain prediction

One of the challenges for domain prediction is that, even if the protein structure is known, it is not entirely trivial to identify where domains begin and end. What is worse, domains are defined at several levels (e.g., Kirillova *et al.*, 2009): *i*) domains are structurally compact regions, *ii*) domains move in a rigid-body like fashion, *iii*) domains are independently folding units, *iv*) domains are evolutionary units.

Methods that predict domain regions or domain boundaries are usually exploiting one or more of these features:

**Scooby-domain** (George *et al.*, 2005) uses a multi-level smoothing window over residue hydrophobicity to find regions of sequence that are more hydrophobic than their surrounding regions. This delineates sequence regions that will most likely form a globular structure, independent of the rest of the protein sequence.

**Domaination** (George and Heringa, 2002a) uses the evolutionary conservation level, as detected by finding domain-level sequence matches using PSI-Blast, to find (conserved) domains.

**SnapDRAGON** (George and Heringa, 2002b) generates coarse distance restraints from predicted secondary structure, and then uses those to predict an ensemble of possible folds. From the consistency of observed (structural) domains in the predicted folds, finally, the domain boundaries for the input sequence are predicted.

representative of nature's entire proteomic toolbox.

## 2.1 Critical Assessment of protein Structure Prediction

Every other year CASP, a Critical Assessment of protein Structure Prediction, provides such an independent validation benchmark. CASP is a blind test or competition: experimentalists provide sequences for which they know the structures will be solved imminently; modelling groups and servers try to predict the structure (Moult *et al.*, 1995). Once the structure is solved, the models can be evaluated using the solution structure of the target (see also Figure 6.5).

CASP was started because the protein structure prediction problem was claimed to have been solved several times. Problematic was, however, that the algorithms were trained on databases that contained the structures that were evaluated in benchmarking tests. CASP overcomes this problem.

Note that the very first step in any practical structure prediction approach should be to inspect the results from the latest CASP round (Moult

*et al.*, 2016) via the CASP website<sup>3</sup> to see what the state of the art methods are, and what their expected performance is.

## 2.2 Comparing structures – RMSD and GDT\_TS

To assess the quality of a modelling method, we need to compare the structure of the predicted model and the experimentally resolved protein structure. A straightforward way to do this is to compare the atomic coordinates of the model and the solution structure, and quantify their (dis)similarity.

In Chapter 3, we noted that prior to the comparison of two structures based on their coordinates we need a (structural) alignment between the structures and structural superpositioning. As the solution structure and model share the same amino acid sequence, obtaining an alignment is trivial; we know which residues and atoms correspond between the model and solution structure. If we then superimpose the model and structure, we can proceed by quantifying the similarity in atomic coordinates.

One way to quantify such similarity, discussed in-depth in Chapter 3, is the Root Mean Square Deviation (RMSD). The RMSD is, however, not a very good measure to assess the quality of a model; averaging the distance squared over all residues means that the RMSD is notoriously sensitive to outliers. For instance, if a model gets a loop very wrong, it tends to stick out and can be positioned quite distant from the true structure. The RMSD penalizes heavily for this, even though the overall structure may be reasonably accurate.

A structural similarity measure more robust to such outliers is the global distance test total score (GDT\_TS). The key idea is to count the number of residues that can maximally be fitted within a certain distance cutoff, see also Figure 6.5. The GDT score will therefore produce a percentage. In the formula below, the final score is the average over four different distance cutoffs (1, 2, 4, 8 Å).

$$GDT\_TS = \frac{1}{4} \sum_{v=1,2,4,8\text{\AA}} \frac{G(v)}{t} \quad (1)$$

Here,  $G(v)$  is the number of aligned residues within given RMSD cutoff  $v$  (in Ångstrom –  $10^{-10}m$ ) and  $t$  is the total number of aligned residues. A related score called GDT\_HA was introduced in CASP some time ago (Read and Chavali, 2007) using stricter distance cutoffs (0.5, 1, 2, 4 Å) to cater to targets in the template-based modelling category where very high accuracies may be realized.

A word of caution should be given to not think of the GDT\_HA superseding the GDT\_TS; for a typical “difficult” CASP target no model comes

---

<sup>3</sup>CASP website: <http://predictioncenter.org/>

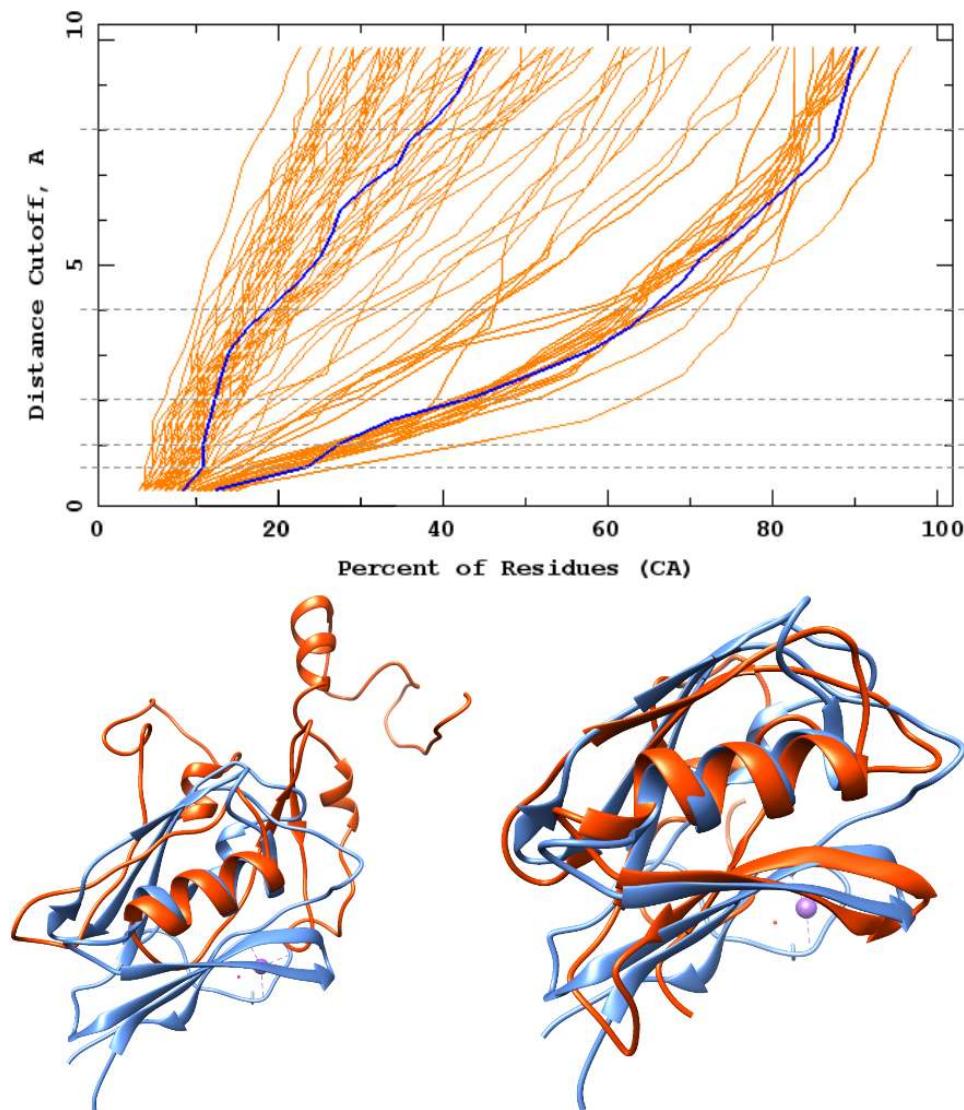


Figure 6.5: Example of structural comparison for the target T0886-D2 and two models submitted to CASP12. The top panel shows individual traces for all models generated for this target; the distance cutoff (vertical axis, in Å) is plotted against the fraction of residues (horizontal axis, in %) that can be aligned within this cutoff. The traces were obtained from [predictioncenter.org/casp12](http://predictioncenter.org/casp12). The dotted lines indicate the thresholds used in the GDT\_TS (1, 2, 4, 8 Å) and GDT\_HA (0.5, 1, 2, 4 Å) scores. Two models are highlighted in blue: a bad model (TS236, GDT\_TS=18.90) on the left, and a good model (TS173; GDT\_TS=51.97) on the right. Both model structures are also shown in the panels below in red, superposed onto the solution crystal structure in blue (PDB:5FHY). Structural superposition created using LGA at [proteinmodel.org/AS2TS/LGA/](http://proteinmodel.org/AS2TS/LGA/) (Zemla, 2003), 3D visualisation using Chimera 1.11.2 (Pettersen et al., 2004).

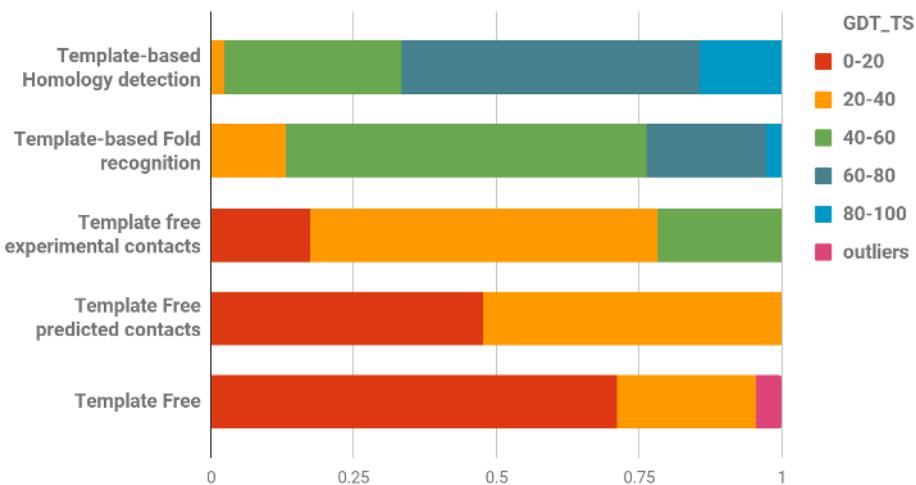


Figure 6.6: Distribution of GDT\_TS scores for the different model categories in CASP11 for template-based (Modi et al., 2016), template-free with contact information (Kinch et al., 2016a) and template-free (Kinch et al., 2016b). The legend coloring corresponds to the GDT\_TS scores, the bars indicate the fraction of models in each GDT\_TS range for the six categories (GDT\_TS scores for (Modi et al., 2016) were estimated from the reported GDT\_HA scores using their Figure 4A). “Outliers” targets have unusually high GDT\_TS due to being very short (~ 50 residue) with extended structures. Targets selected for server prediction (top bar) were considered easier than those for human prediction (second from top), average sequence identity was 26% vs. 20%, respectively. It is clear that overall prediction accuracy sharply declines going down this list of categories. For template-free modelling, the quality of contact information used is crucial. Experimental information (from chemical cross-linking or simulated NMR) can give reasonable models. Predicted contacts do not guarantee that an acceptable model can be obtained, but without even predicted contacts, more than two-thirds of models are at most 20% correct.

REMARK: copyright OK – from scratch by Anton

even close to the experimentally solved structure. Results for such targets are similar to the left-most model in Figure 6.5, and performances of GDT\_TS < 20% are no exceptions. Thus, especially for **targets without good template structures, the protein structure prediction problem has NOT yet been solved.**

### 2.3 How difficult is it to predict?

Overall, if one can find an appropriate template, the quality of the predicted model will be **relatively good**. CASP results show that for homology modelling based on close homologues, it is possible to obtain models similar to the experimentally determined structure (Moult *et al.*, 2016). The modelled structure will typically have a good accuracy for the regions that can be well

aligned between the target and template (using the sequences). The top two bars in Figure 6.6 show that one may expect the majority of such models to be accurate for > 50% of their residues. Gaps in an alignment will typically lie in loop regions of a structure and are more difficult to model. So, if we are interested in a large loop region that is not present in our template, we still may not be able to answer our scientific question with the resulting model structure (Moult *et al.*, 2016).

If no acceptable template can be found, the chances of successfully answering our scientific question will become very low. As a last resort, ab initio modelling can provide us with structural models. Typically, ab initio methods use very small templates from various proteins (see Figure 6.4). The state of the art is that on average one may expect to find one structure that looks somewhat like the solution structure for the target among the top five or ten models (Moult *et al.*, 2016). However, be aware that the best model is typically not recognised as being the best through the scores of the prediction program. In Figure 6.6 one sees that very clearly in the bottom few bars: without a template, even with predicted contacts, one may have less than 20% of the structure correct in the majority of models; even in the best cases at most 40% of the residues are modelled accurately.

## 2.4 For which gene sequences can we predict a three-dimensional structure?

If and only if there is a structure of a homologous protein present in the PDB, it is possible to generate a structural model of reasonable accuracy. Based on this notion, we can estimate for which (fraction of) gene sequences it is possible to predict a structure. This way it has been estimated that for about 44% of residues in Eukaryotic gene sequences, we cannot yet make a homology model, and 15% of these residues lie within a gene for which we can not make a homology model for a single domain (Perdigão *et al.*, 2015). Especially membrane proteins are underrepresented in the PDB, due to the experimental difficulty of determining these structures. Note that these residues may also lie in natively disordered regions (see also Section 3).

Similarly, it is possible to predict the range of protein structures present in an organism, based on the gene sequences in their completed genome. This reveals that there is a subset of protein structures, that is present in nearly all organisms, for example, TIM-barrels or Rossmann-folds (Abeln and Deane, 2005; Edwards *et al.*, 2013). Nevertheless, there is also a group of structures that is extremely lineage-specific. It is to be expected that for this type of protein structure, many new structures remain to be discovered. This also implies that it will remain difficult to find suitable templates for homology modelling for these lineage specific protein families.

## 2.5 How accurate do we need to be?

We already mentioned that we may approach the modelling of a protein structure of interest differently, depending on the **biological question** we want to ask, e.g. which residues are likely to be crucial for the functioning of the protein. Sometimes an answer to the research question is attainable without full-scale prediction of the protein structure, e.g. by direct prediction of the **impact of certain mutations** or of **protein-protein interaction sites**. Examples of fully-automated web servers that do just that, are HOPE – (Venselaar *et al.*, 2010) and SeRenDIP (Hou *et al.*, 2017). In some cases, a rough homology model inspires the understanding of experimental results, spurring forward the project and eventually ending with crystal structures highlighting the protein function (in this case, protein-protein interactions) of interest (e.g., De Vries-van Leeuwen *et al.*, 2013). Also, specifically for enzymes, such as for example cytochromes P450, modelling of the protein structure should be done in combination with that of the **ligand** (Graaf *et al.*, 2005).

Protein structural models are regularly used for the prediction of **protein-protein interactions** or **protein-ligand interactions by docking**. We will return on this topic in Chapter 11.

## 3 Is there such a concept as a single native fold?

Before we conclude, we should consider a more physical description of protein structure. In fact, protein folding from a physical point of view is a very interesting process: given a sequence, a protein tends to fold always, and exactly into the same functional structure. In material design, it is extremely difficult to mimic such high specificity. The apparent observation of folding specificity also leads to the question, is there such a concept as a single native fold? Or, more pragmatically, is **sequence-to-structure truly a one-to-one relation?**

In fact, if one wants to start making quantitative predictions, such as the stability of a protein fold, or the binding strength between two proteins in terms of free energy, it is much more helpful to think in **ensembles of structural configurations for a protein sequence** (e.g. May *et al.*, 2014; Pucci and Roonan, 2017). The probability to find a protein in a specific ensemble of structural configurations will depend on conditions such as the presence or absence of **binding partners, the pressure, the pH or the temperature** (e.g. van Dijk *et al.*, 2015, 2016). There are a few specific cases, common cases, for which even the functional or biologically relevant structural ensembles do not resemble a single globular folded structure.

### 3.1 Intrinsically disordered proteins

Not all proteins fold into single configurations, some proteins stay natively unfolded, i.e. they can take up a large variety of more extended, and very different configurations (Uversky *et al.*, 2000; Mészáros *et al.*, 2007). Some disordered regions contain elements that do form stable structures upon binding. The regions that remain disordered are thought to be important to prevent aggregation within the cell (Abeln and Frenkel, 2008). Missing residues in X-ray structures are typically removed for crystallization; for this reason disorder prediction methods have been developed. Disordered regions are relatively easy to predict in protein sequences just like secondary structures; broadly speaking, prediction can be based on a large amount of charged/polar (hydrophilic) amino acids in combination with the presence of amino acids that disrupt the secondary structure (proline and glycine) in these regions (Oldfield *et al.*, 2005; Wang *et al.*, 2016). We know sequences of many proteins contain large disordered segments (33% of eukaryotic, 2% archaeal, and 4% bacterial proteins).

### 3.2 Allostery and functional structural ensembles

It is important to realize that one protein, typically, does not correspond to one defined three-dimensional structure. Disordered regions or proteins are one particularly salient case, but also proteins that fold into specific three-dimensional configurations, may exist in multiple functional states each with a specific structure. The biological question of interest dictates which state is relevant. Most proteins have only been crystallized in one particular state, and often it is not known to which biological condition this crystal structure may correspond. One may have cases where a homology model of the relevant state may be preferred over a crystal structure of a different or unknown state (e.g., Graaf *et al.*, 2005).

### 3.3 Amyloid fibrils

Lastly, we should consider a competing state of folded proteins: the aggregated state, where multiple peptide chains clog together in fibrillar structures or amorphous aggregates. Amyloid fibres are formed by  $\beta$ -strands formed between different protein or peptide (small protein) chains. Fibril formation is associated with various neurodegenerative diseases, such as Alzheimer's, Creutzfeldt-Jakob and Parkinson's (Chiti and Dobson, 2006). In fact, the fibrillar state is more favorable than the state of separately folded structures for several protein types. The general cellular toxicity of such aggregates puts evolutionary pressure on avoiding structural characteristics on the surface of proteins; hence it is extremely rare to observe solvent-accessible  $\beta$ -strand edges or large hydrophobic surface patches (Richardson and Richardson, 2002; Abeln and Frenkel, 2011). The propensity proteins

have to form Amyloid fibrils is relatively easy to predict (Graña-Montes *et al.*, 2017). However, reference databases are still small so it is difficult to verify such methods.

## 4 Key Points

- Several claims have been made that the problem of predicting protein structure from amino acid sequence was solved.
- The Critical Assessment of protein Structure Prediction (CASP) competition was set up in response to such claims.
- The Root Mean Square Deviation (RMSD) is sensitive to outlier residues such as those in the loop regions. CASP uses GDT\_TS and GDT\_HA to address this problem.
- Results show homology modeling can yield adequate structural models, but depends on the availability of good templates.
- Homology modeling benefits from constructing models per domain.
- It is estimated that 44% of the eukaryotic proteome templates are not available, precluding homology modeling for these proteins.
- In absence of a (remote) template, we can resort to “ab initio” modelling, but resulting predictions are often poor.
- Whether unattainability of structural models for entire proteins is a bad thing depends on the scientific question we intend to answer with our model.

## 5 Further Reading

### Author contributions

Wrote the text: SA, JH, KAF  
Created figures: SA, KAF  
Review of current literature: SA, KAF  
Editorial responsibility: SA, KAF  
The authors thank Nicola Bonzanni , Kamil K. Belau, Ashley Gallagher, Jochem Bijlard , Hans de Ferrante  for insightful discussions and critical proofreading.

## References

- Abeln, S. and Deane, C.M. (2005). Fold usage on genomes and protein fold evolution. *Proteins*, **60**(4), 690–700.  
Abeln, S. and Frenkel, D. (2008). Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, **4**(12), e1000241.  
Abeln, S. and Frenkel, D. (2011). Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins. *Biophys. J.*, **100**(3), 693–700.

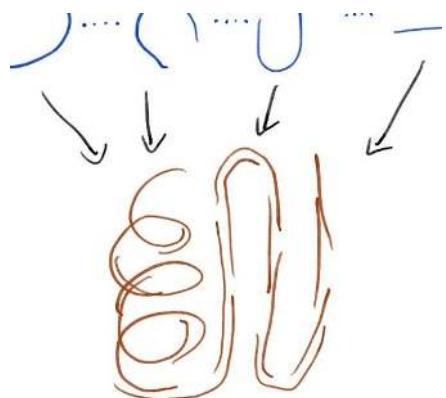
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. et al (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235–242.
- Chiti, F. and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.
- De Vries-van Leeuwen, I.J., da Costa Pereira, D., Flach, K.D., Piersma, S.R. et al (2013). Interaction of 14-3-3 proteins with the estrogen receptor alpha F domain provides a drug target interface. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(22), 8894–9.
- Edwards, H., Abeln, S. and Deane, C.M. (2013). Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. *PLoS computational biology*, **9**(11), e1003325.
- George, R.A. and Heringa, J. (2002a). Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins: Structure, Function, and Genetics*, **48**(4), 672–681.
- George, R.A. and Heringa, J. (2002b). SnapDRAGON: a method to delineate protein structural domains from sequence data. *Journal of Molecular Biology*, **316**(3), 839–851.
- George, R.A., Lin, K. and Heringa, J. (2005). Scooby-domain: prediction of globular domains in protein sequence. *Nucleic Acids Research*, **33**(Web Server), W160–W163.
- Graaf, C.d., Vermeulen, N.P.E. and Feenstra, K.A. (2005). Cytochrome P450 in Silico: An Integrative Modeling Approach. *Journal of Medicinal Chemistry*, **48**(8), 2725–2755.
- Grana-Montes, R., Pujols-Pujol, J., Gómez-Picanyol, C. and Ventura, S. (2017). Prediction of Protein Aggregation and Amyloid Formation. In *From Protein Structure to Function with Bioinformatics*, pages 205–263. Springer Netherlands, Dordrecht.
- Hou, Q., De Geest, P., Vranken, W., Heringa, J. and Feenstra, K. (2017). Seeing the trees through the forest: Sequencebased homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics*, **33**(10).
- Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A. and Grishin, N.V. (2016a). Assessment of CASP11 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 164–180.
- Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A. and Grishin, N.V. (2016b). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 51–66.
- Kirillova, S., Kumar, S. and Carugo, O. (2009). Protein domain boundary predictions: a structural biology perspective. *The open biochemistry journal*, **3**, 1–8.
- May, A., Pool, R., van Dijk, E., Bijlard, J. et al (2014). Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics (Oxford, England)*, **30**(3), 326–334.
- Mészáros, B., Tompa, P., Simon, I. and Dosztányi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol*, **372**(2), 549–561.
- Modi, V., Xu, Q., Adhikari, S. and Dunbrack, R.L. (2016). Assessment of template-based modeling of protein structure in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 200–220.
- Moult, J., Pedersen, J.T., Judson, R. and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics*, **23**(3), ii–iv.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, **84**(S1), 4–14.
- Oldfield, C.J., Cheng, Y., Cortese, M.S., Brown, C.J. et al (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**(6), 1989–2000.
- Perdigão, N., Heinrich, J., Stoltze, C., Sabir, K.S. et al (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, **112**(52), 15898–15903.
- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S. et al (2004). UCSF Chimera – A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13), 1605–1612.
- Pucci, F. and Rooman, M. (2017). Physical and molecular bases of protein thermal stability and cold adaptation. *Current Opinion in Structural Biology*, **42**, 117–128.
- Read, R.J. and Chivali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function, and Bioinformatics*, **69**(S8), 27–37.
- Richardson, J.S. and Richardson, D.C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A*, **99**(5), 2754–2759.
- Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**(3), 415–427.
- van Dijk, E., Hoogeveen, A. and Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, **11**(5), e1004277.
- van Dijk, E., Varilly, P., Knowles, T.P.J., Frenkel, D. and Abeln, S. (2016). Consistent Treatment of

- Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, **116**(7), 078101.
- Venselaar, H., te Beek, T.A., Kuipers, R.K., Hekkelman, M.L. and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, **11**(1), 548.
- Wang, S., Ma, J. and Xu, J. (2016). AUCpred: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics (Oxford, England)*, **32**(17), i672–i679.
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.
- Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research*, **31**(13), 3370–4.

## Chapter 7

# Practical Guide to Model Generation

Sanne Abeln  K. Anton Feenstra 



[arxiv.org/abs/1712.00425](https://arxiv.org/abs/1712.00425)

\* editorial responsibility



Here we consider strategies for a typical protein structure prediction problem: we want to generate a structural model for a protein with a sequence, but without an experimentally determined structure. In the previous chapter on “Introduction to protein structure prediction” we introduced the problem of how to obtain the folded structure of the protein, given only an amino acid sequence. In this chapter, we will build up a workflow for tackling this problem, starting from the easy options that, if applicable, are likely to generate a good structural model, and gradually working up to the more hypothetical options whose results are much more uncertain.

An overview of protein structure modelling, including both template-based and template-free modelling is given in Figure 7.1; see also Chapter 7 for the terminology used here. In short, homology-based models (far left in Figure 7.1) are most accurate, while ab-initio modelling (far right) is notoriously unreliable. Template-based (by homology or fold recognition) models require an alignment with the target sequence, from which the initial model will be built. Course constraints may sometimes be incorporated in this stage. The raw model may need to be completed by separate modelling of the missing substructures. Template-free modelling can benefit greatly if such constraints are available, if not the only sources left are fragment libraries and knowledge-based energy functions (there will be more on energy functions in Chapter 12 and Chapter 14). The model, whether template-based or -free, is usually refined until the desired level of (estimated) quality is reached to produce the final predicted structure for our target protein sequence of interest.

Below, we will discuss in detail first template-based and then template-free modelling.

## 1 Template based protein structure modelling

### 1.1 Homology based Template Finding

Homology modelling is a type of template-based modelling with a template that is homologous to the target protein. As mentioned before, homology modelling works well, because structure is more conserved than sequence. Typically, we will use a sequence-based homology detection method, such as BLAST, to search for homologous protein sequences in the full PDB dataset. If we find a sequence that has significant sequence similarity to the full length of our target sequence we have found a template. Of course, it is possible that a template **only covers part of the target sequence**, see also the section on domains in Chapter 6 “Introduction to structure prediction”.

If a simple BLAST search against the PDB gives no good results we may need to start using alignment methods that can detect more distant homologues based on sequence comparison, such as PSI-BLAST or HMMs. PSI-BLAST uses hits from a previous iteration of BLAST to create a profile

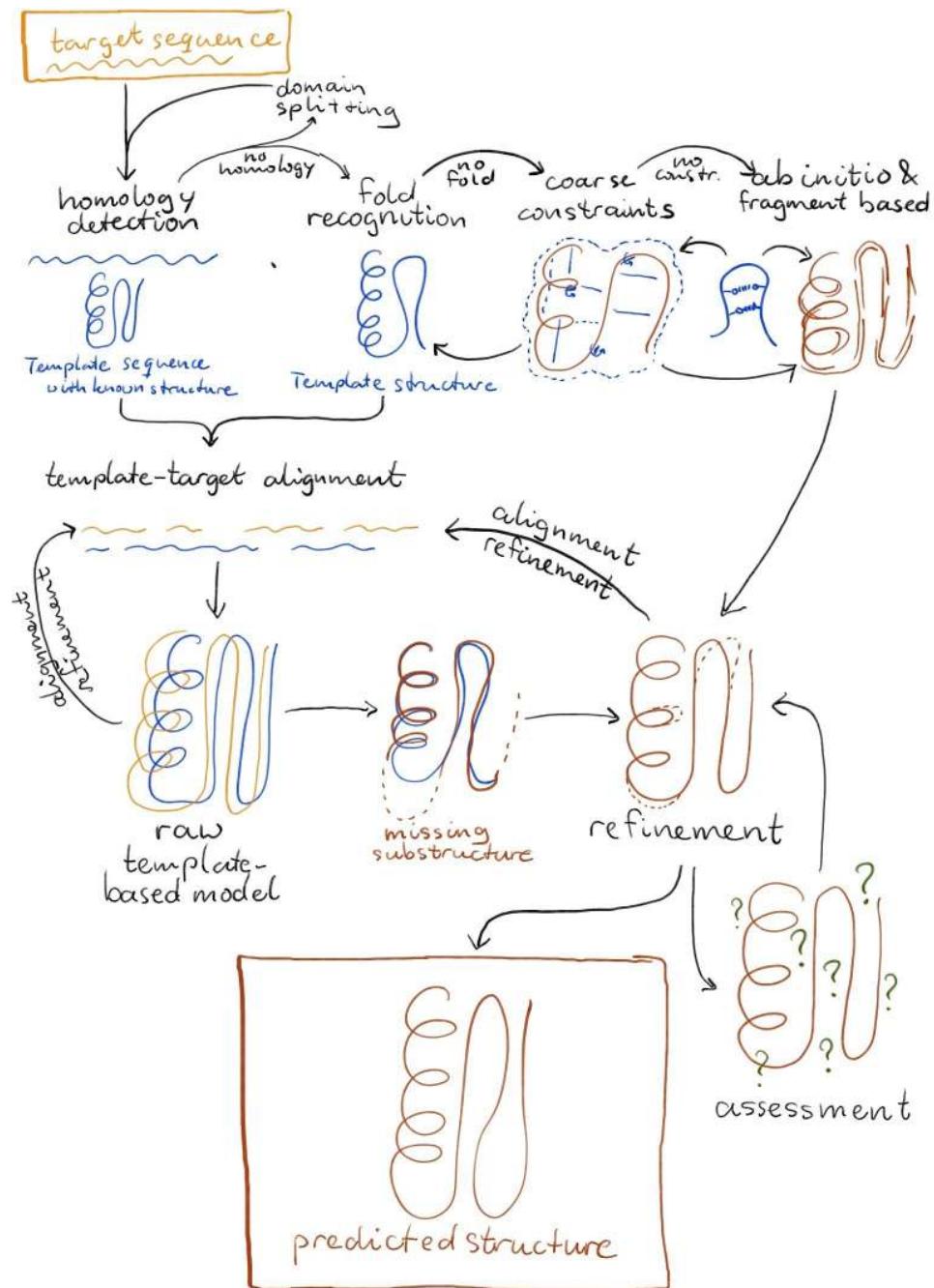


Figure 7.1: Flowchart of protein three-dimensional structure prediction. It starts at the top left with the target protein sequence of interest, and ends with a predicted 3D structure at the bottom. Depending on the availability of a homologous template, a suitable fold, or coarse/experimental constraints, different options are available, with sharply decreasing expected model accuracy for each step. See text for more details

for our query sequence (Altschul *et al.*, 1997). This allows successive iterations to give more weight to very conserved residues, allowing to find more distant homologues. Even more sensitive homology detection tools typically consider sequence profiles of both the query and the template sequence, and try to align or score these profiles against each other (Pietrokovski, 1996), with more recent implementations including Compass (Sadreyev and Grishin, 2003), HMMer (Finn *et al.*, 2011) and HHpred (Söding *et al.*, 2005). Profile-profile matching is known to increase detection of evolutionary sequence signals, leading to optimised alignment or sensitive sequence searching (Wang and Dunbrack, 2004). Note that to make full use of such methods it is important that the searching profiles are generated using extended sequence databases (so not just on the PDB). The evolutionary sequence profiles used in these methods can ensure that the most conserved, and therefore structurally most important, residues are more likely to be aligned accurately.

Once we have a good template, and an alignment of the target sequence with the template structure, we can build a structural model; this is called homology modelling in cases of clear homology between the target and the template.

## 1.2 Fold recognition

If no obvious homologs with a known structure can be found in the PDB, it becomes substantially more difficult to predict the structure for a sequence. We may then use another trick to **find a template**: remember that the template does not only have a sequence, but also a structure. Therefore, we may be able to use structural information in this search. **Fold recognition** methods look explicitly for a plausible match between our query sequence and a **structure** from a database. Typically, such methods use structural information of the putative template to determine a match between the target sequence and the putative template sequence and structure.

Note that, in contrast to homology-based template search, it is not strictly necessary for a target sequence and a template sequence to be homologous – they may have obtained similar structures through convergent evolution. Therefore some fold recognition methods are trained and optimised for the detection of structural similarity, rather than homology. In fact, for very remote homologs we may not be able to differentiate between convergent and divergent evolution. For the purpose of structure prediction this may not be an important distinction, nevertheless divergent evolution, i.e., homology through a shared common ancestor, invokes the principle of structure being more conserved, giving more confidence in the final model.

**Threading** is an example of a fold recognition method (Jones *et al.*, 1992). The query sequence is threaded through the proposed template structure. This means structural information of the template can be taken into account

when **score** if the (threaded) alignment between the target and the potential template is a good fit . Typically threading works by **scoring the pairs in the structure that make a contact**, given the sequence composition. A sequence should also be aligned (threaded) onto the structure giving the best score. Scores are typically based on **knowledge based potentials**. For example, if two hydrophobic residues are in contact this would give a better score than contact between a hydrophobic and polar amino acid. Note that threading does not necessarily search for homology. Threading remains a popular fold recognition methods, with several implementations available (Jones *et al.*, 1992; Zhang, 2008; Song *et al.*, 2013).

Another conceptually different fold recognition approach is to consider that **amino-acid conservation rates** may strongly differ between different structural environments. For example, one would expect residues on the **surface to be less conserved**, compared to those buried in the core. Similarly, residues in a  $\alpha$ -helix or  $\beta$ -strand are typically **more conserved** than those in loop regions. In fact, the chance to form an insertion or deletion in loop regions is seven times more likely than within (other) secondary structure elements . Since we know the structural environment of the residues for the potential template, we can use this to score an alignment between the target sequence and potential template sequence. **FUGUE** is a method that scores alignments using structural environment-specific substitution matrices and structure-dependent gap penalties (Shi *et al.*, 2001).

### 1.3 Generating the target-template alignment

Once a suitable template has been found, one can start building a structural model. Typical model building methods will need the following inputs: (1) the **target sequence**, (2) the **template structure**, (3) **sequence alignment** between target and template and (4) any **additional known constraints**. The output will be a structural model, based on the constraints defined by the template structure and the sequence alignment.

Here, it is important to note that the methods, that recognize **good potential templates**, are not necessarily the methods that will produce the **most accurate alignments** between the target and template sequence. As the final model will heavily depend upon the alignment used, it is important to consider different methods, including for example **structure and profile-based** methods, **multiple sequence alignment** programs or methods that can include **structural information** of the template in constructing the alignment. Examples of good sequence-based alignment methods, which can also exploit evolutionary signals and profiles, are **Praline** (Simossis *et al.*, 2005) and **T-coffee** (Notredame *et al.*, 2000). The T-coffee suite also includes the structure-aware alignment method **3d-coffee** (O'Sullivan *et al.*, 2004). Some aligners are context-aware, taking into account for example secondary structure (Simossis and Heringa, 2005) or **trans-membrane** (TM) regions ex-

plicitly (Pirovano *et al.*, 2008; Pirovano and Heringa, 2010; Floden *et al.*, 2016), which are useful extension as TM proteins are severely underrepresented in the PDB. Of particular interest to the general user are automated template detection tools such as **HHpred** (Söding *et al.*, 2005). Bawono *et al.* (2017) give a good and up-to-date overview of multiple sequence alignment methodologies, including profile-based and hidden-Markov-based methods.

Moreover, once a first model is created, it may be wise to interactively adapt the alignment, based on the resulting model, which might lead to an updated model. This procedure may also be carried out in an interactive fashion.

## 1.4 Generating a model

Here we consider the *MODELLER* software to generate template based alignments (Sali and Blundell, 1993), which is one of several alternative approaches to construct homology models (Schwede *et al.*, 2003; Zhang, 2008; Song *et al.*, 2013). Firstly, the **known template 3D structures** should be aligned with the **target sequence**. Secondly, **spatial features**, such as C $\alpha$ -C $\alpha$  distances, hydrogen bonds, and mainchain and sidechain dihedral angles, are transferred from the template(s) to the target. *MODELLER* uses “knowledge-based” constraints. The constraints are based on the **template distances**, the **alignment**, but also on knowledge- based **energy functions** (probability distribution). The constraints are optimised using molecular dynamics with simulated annealing. Finally, a 3D model can be generated by satisfying all the constraints as well as possible.

## 1.5 Loop or missing substructure modelling

We now have a model for **all the residues that were aligned well** between the template and the target. The remaining substructure(s), that are not covered by the template, will show as gaps in the alignment between target sequence and template. “**Loop modelling**” is used to determine the structure for these missing parts. Loop models are typically based on fragment libraries, knowledge-based potentials and constraints from the aligned structure. This problem is in fact closely related to the template-free modelling procedure, as we need to generate a structure without a readily available template (template-free approaches are discussed in more detail in the next section). In CASP11, consistent refinement overall as well as for loop regions was achieved; the limiting factor for effective refinement was concluded to be the **energy functions** used, in particular, missing physicochemical effects and balance of energy terms (Lee *et al.*, 2016).

## 2 Template-free protein structure modelling

### 2.1 What if no suitable template exists?

If on the other hand, no suitable template is available for our target protein of interest, we will need to follow a ‘template-free’ modelling strategy. Without a direct suitable template, we need an “*ab initio*” strategy that can suggest possible structural models based on the sequence of the template alone. In this case, we need to resort to fragment-based approaches. Here small, suitable fragments, from various PDB structure, are assembled to generate possible structural models. As the fragments are typically matched using sequence similarity, one may even consider this as template-based modelling at a smaller scale. However, since the sequence match is based on a limited number of residues, this would not generally imply a homologous relation between the fragment template and the target sequence. It is also important to note that this type of *ab initio* modelling is still “knowledge based”: the structural models are generated from small substructures present in the PDB, assessed by energy scoring functions generated by mining the PDB. In other words, these models are not based on physical principles and physicochemical properties alone. This also means that such models are likely to share any of the biases that are present in the PDB, such as lack of trans-membrane proteins and absence of disordered regions.

### 2.2 Generating models from structural fragments

Here we follow the key ideas in the Rosetta based *ab initio* modelling suite (Simons *et al.*, 1999). Quark, another fragment-based approach provides a similar performing alternative (Xu and Zhang, 2012). The overall approach is to split the target sequence into 3 and 9-residue overlapping sequence fragments, i.e., sliding windows, and find matching structural fragments from PDB.

A fragment library is generated by taking 3 and 9-residue fragments from the PDB and clustering these together into groups of similar structure. For each fragment in the database sequence profiles are created. These profiles are subsequently used to search for suitable fragments for our query/target sequence, as we only have sequence information (see Figure 7.2 on the left-hand side).

The target sequence also will be split into 3 and 9-residue overlapping sequence fragments. These target sequence fragments are then matched with the structural PDB based fragment library using profile-profile matching. Note that this procedure will generate multiple fragments for each fragment window in the sequence. The fragment windows on the sequence typically also overlap (see Figure 7.2 middle panel).

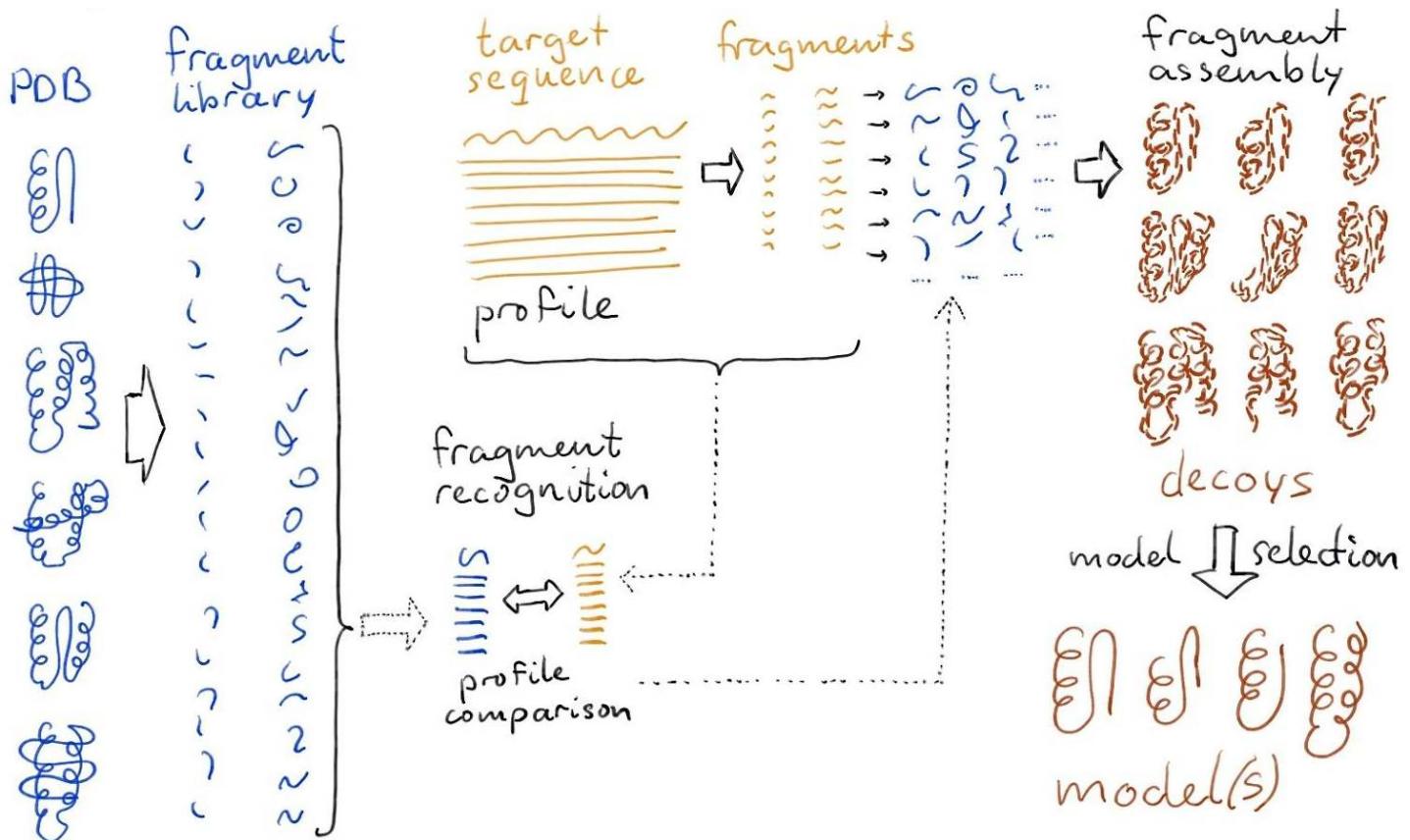


Figure 7.2: Overview of the fragment-based modelling strategy. A library of structure fragments was created once from the PDB; all small 3-residues and larger 9-residue fragments are collected and clustered. A target sequence of interest is also separated into 3- and 9-residue sequence fragments. For each of these, a profile-profile search is performed to find matching fragments from the fragment library; typically for each target fragment, multiple hits with different structure are retrieved. This collection of fragments of alternate structure are then assembled through a Monte Carlo algorithm into a large set of possible structures, called ‘decoys’. Using knowledge-based potentials and overall statistics, from the decoy set, a final selection of model structures is made.

### 2.3 Fragment Assembly into decoys

The above scenario leads to many possible structural fragments to cover a sequence position. To find the best structural model is a combinatorial problem in terms of fragment combinations (see Figure 7.2, top right). A Monte Carlo algorithm is used to search through different fragment combinations. Good combinations are those that give low energy. Each MC run will produce a different model, since it is a stochastic algorithm. Note that to be able to optimise a model, we need a scoring function. A knowledge-based en-

ergy function is used, including the number of neighbours, given amino acid type, residue pair interactions, backbone hydrogen bonding, strand arrangement, helix packing, radius of gyration, Van der Waals repulsion. These are all terms that are relatively cheap to compute when a new combination of fragments is tried. The structural model is optimised by slowly lowering the MC temperature – this is also called simulated annealing. The generated models are called ‘decoys’.

Thereafter, decoys are refined using additional Monte Carlo cycles, and a more fine-grained energy function including: backbone torsion angles, Lennard-Jones interactions, main chain and side-chain hydrogen bonding, solvation energy, rotamers and a comparison to unfolded state.

Finally, the most difficult task is to select, from all the refined decoys, a structure that is a suitable model for the target sequence (see Figure 7.2 bottom right). Again, using a more detailed, knowledge-based energy function, decoys can be scored to assess how ‘protein-like’ they are. Such a selection procedure may get rid of very wrong models. However, selecting the best model, without any additional information (from for example experiments or co-evolution-based contact-prediction), is likely to lead to poor results (see Chapter 6).

## 2.4 Constraints from co-evolution based contact prediction or experiments

As already mentioned, valuable additions to the modelling process are coarse constraints from experimental data or contact prediction. Experimental data from NMR and chemical cross-linking can yield distance restraints that are particularly useful in the template-free modelling to narrow down the conformational space to be searched; still average accuracy of models produced remains extremely limited, as shown in Chapter 6 Figure 6.6 (Kinch *et al.*, 2016). Other sources of information are contours or surfaces that can be obtained from cryo-EM, or small angle scattering experiments, either with electrons, neutrons, or x-ray radiation. However, since these techniques are employed mostly for elucidating larger macromolecular complexes, they are considered out of scope for the current chapter.

Of more general applicability may be methods for predicting intra-protein residue contacts; the main approach currently is based on some form of co-evolution information obtained by direct-coupling methods from ‘deep’ alignments (Marks *et al.*, 2011; Jones *et al.*, 2012; Morcos *et al.*, 2011). The depth here signifies the amount of sequence variation present in the alignment in relation to the length of the protein (the longer the protein, the more variation is needed). Ovchinnikov *et al.* (2017) expresses this as the *effective protein length*:  $Nf = N80\%ID/\sqrt{l}$  where  $l$  is the protein length, and  $N80\%ID$  the number of cluster at 80% sequence identity. They showed that  $Nf$  can be greatly enhanced by the use of metagenomic sequencing data,

and that this leads to a marked improvement in model quality, and estimate that this would triple the number of protein families for which the correct fold might be predicted (Ovchinnikov *et al.*, 2017). Wuyun *et al.* (2016) investigated ‘consensus’-based methods, which combine both direct-coupling and machine-learning approaches, and find that the machine-learning methods are less sensitive to alignment depth and target difficulty, which are crucial factors for success for the direct-coupling methods.

### 3 Selecting and refining models from structure prediction

Once we have created (several) models, we need to assess which model is the best one. Typically this can be done by scoring models on several properties using model quality assessment programs and visual inspection with respect to “protein like” features. Moreover, if any additional knowledge about the structure or function of our target protein is available, this may also help to assess the quality of the model(s). In addition, one may in some cases want to improve a model, or parts of it; this is called model refinement.

#### 3.1 Model refinement

For many years in CASP, model refinement was a no-go area; the rule of thumb was: build our homology model and do not touch it! An impressive example of the failure of refinement methods was shown by the David Shaw group, who concluded that “simulations initiated from homology models drift away from the native structure” (Raval *et al.*, 2012). Since CASP10 in 2014 (Nugent *et al.*, 2014) and continuing in the latest CASP11, there is a reason for moderate optimism. General refinement strategies report small but significant improvements of 3-5% over 70% of models (Modi and Dunbrack, 2016). Interestingly, and in stark contrast to the earlier results by (Raval *et al.*, 2012), the average improvement of GDT\_HA using simulation-based refinement now also is about 3.8, with an improvement (more than 0.5) for 26 models (Feig and Mirjalili, 2016). For five models, the scores became worse (by more than 0.5), and another five showed no significant change. Particularly, for very good initial models ( $\text{GDT\_HA} > 65$ ), models were made worse. Moreover, they also convincingly showed that both more and longer simulations consistently improved these results; note however that protocol details such as using  $\text{C}\alpha$  restraints, are thought to be the limiting factor (Feig and Mirjalili, 2016), as already used previously (e.g., Keizers *et al.*, 2005; Feenstra *et al.*, 2006), and replicated by others (e.g., Cheng *et al.*, 2017). Most successful refinement appears to come from correctly placing  $\beta$ -sheet or coil regions at the termini (Modi and Dunbrack, 2016).

### 3.2 Model quality assessment strategies

It may be generally helpful to compare models generated by different prediction methods; if models from different methods look alike (more precisely if the pair has a low RMSD and/or high GDT\_TS) they are more likely to be correct. Similarly, templates found by different template finding strategies, found for example both by homology sequence searches and fold recognition methods are more likely to yield good modelling results (Moult *et al.*, 2016; Kryshtafovych *et al.*, 2016). Such a consensus template is generally more reliable than the predictions from individual methods – especially if the individual scores are barely significant. Lastly, one can consider the biological context to select good models.

Whether a model is built using homology modelling, fold recognition and modelling or *ab initio* prediction, all models can be given to a Model Quality Assessment Program (MQAP) for model validation. A validation program provides a score predicting how reliable the model is. These scores typically take into account to what extent a model resembles a “true” protein structure. The best performing validation programs take a large set of predicted models, and indicate which out of these is expected to be the most reliable.

Validation scoring may be based on similar ideas as validation for experimental structures or may be specific to structure prediction. For example, it can be checked if the amount of secondary structure, e.g., helix and strand vs. loop, has a similar ratio as in known protein structures; if a model for a sequence of 200 amino acids does not contain a single helix or  $\beta$ -strand, the model does not resemble true protein structures, and is therefore very unlikely to be the true structural solution for the sequence. A similar type of check may be done for the amount of buried hydrophobic groups and globularity of the protein.

Different models may also be compared to each other. One trick that is commonly used, is that if multiple prediction methods create structurally similar models, these models are more likely to be correct. Hence, a good prediction strategy is to use several prediction methods, and pick out the most consistent solution. A pitfall here is that if all models are based on the same, or very similar, templates, they will look similar but this may not reflect that they are correct.

### 3.3 Secondary Structure Prediction

Secondary structure prediction is relatively accurate (see e.g. the review by Pirovano and Heringa, 2010). This problem is in fact much easier to solve than three-dimensional structure prediction, as is shown in Chapter 9 “Structural Property Prediction”. The accuracy of assigning strand, helix or loops to a certain residue can go up to 80% with the most reliable methods. Typically such methods use (hydrophobic) periodicity in the sequence

combined with phi and psi angle preferences of certain amino acid types to come to accurate predictions. The real challenge lies in assembling the secondary structure element in a correct topology. Nevertheless, secondary structure prediction may be used to assess the quality of a model built with a (tertiary) structure prediction method. Many (automated) methods also incorporate secondary structure information during alignment (Simossis and Heringa, 2005), homology detection (Söding *et al.*, 2005; Shi *et al.*, 2001) and contact prediction (Terashi and Kihara, 2017; Wang *et al.*, 2017).

## 4 Key points

- Fold recognition
  - Fugue
  - Environment specific substitution tables
- Ab initio prediction (free modelling)
  - Rosetta
  - MC
  - Fragments
- CASP
  - GDT\_TS score
  - How well are we doing?

## 5 Further Reading

Since writing this chapter, a lot of progress has happened in structure prediction, particularly around AlphaFold and trRosetta. In anticipation of a full update of this chapter, some recent papers on this are listed below:

- Hanson *et al.* (2020): Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning. *Journal of Computational Biology*
- Senior *et al.* (2020): Improved protein structure prediction using potentials from deep learning. *Nature*
- Humphreys *et al.* (2021): Computed structures of core eukaryotic protein complexes. *Science*
- Tunyasuvunakool *et al.* (2021): Highly accurate protein structure prediction for the human proteome. *Nature*
- Jumper *et al.* (2021): Highly accurate protein structure prediction with AlphaFold. *Nature*
- Jumper and Hassabis (2022): Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods*
- Jones and Thornton (2022) The impact of AlphaFold2 one year on. *Nature Methods*

- Thornton *et al.* (2021): AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine*

## Author contributions

Wrote the text: SA, KAF

Created figures: SA, KAF

Review of current literature: SA, KAF

Editorial responsibility: SA, KAF

The authors thank Olga Ivanova  and Hans de Ferrante  for insightful discussions and critical proofreading.

## References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J. et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–402.
- Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A. et al (2017). Multiple Sequence Alignment. In *Methods in Molecular Biology – Bioinformatics – Volume I: Data, Sequence Analysis, and Evolution*, pages 167–189. Humana Press, New York, NY.
- Cheng, Q., Joung, I. and Lee, J. (2017). A Simple and Efficient Protein Structure Refinement Method. *Journal of Chemical Theory and Computation*, **13**(10), 5146–5162.
- Feenstra, K.A., Hofstetter, K., Bosch, R., Schmid, A. et al (2006). Enantioselective substrate binding in a monooxygenase protein model by molecular dynamics and docking. *Biophysical journal*, **91**(9), 3206–16.
- Feig, M. and Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 282–292.
- Finn, R.D., Clements, J. and Eddy, S.R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, **39**(SUPPL. 2), W29–W37.
- Floden, E.W., Tommaso, P.D., Chatzou, M., Magis, C. et al (2016). PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic acids research*, **44**(W1), 339–43.
- Hanson, J., Paliwal, K.K., Litfin, T., Yang, Y. and Zhou, Y. (2020). Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning. *Journal of Computational Biology*, **27**(5), 796–814.
- Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A. et al (2021). Computed structures of core eukaryotic protein complexes. *Science*.
- Jones, D.T. and Thornton, J.M. (2022). The impact of AlphaFold2 one year on. *Nature Methods*, **19**(1), 15–20.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature*, **358**(6381), 86–89.
- Jones, D.T., Buchan, D.W.A., Cozzetto, D. and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–190.
- Jumper, J. and Hassabis, D. (2022). Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods 2022 19:1*, **19**(1), 11–12.
- Jumper, J., Evans, R., Pritzel, A., Green, T. et al (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, **596**(7873), 583–589.
- Keizers, P.H.J., Graaf, C.d., Kanter, F.J.J.d., Oostenbrink, C. et al (2005). Metabolic Regio- and Stereoselectivity of Cytochrome P450 2D6 towards 3,4-Methylenedioxy-N-alkylamphetamines: in Silico Predictions and Experimental Validation. *Journal of Medicinal Chemistry*, **48**(19), 6117–6127.
- Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A. and Grishin, N.V. (2016). Assessment of CASP11 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 164–180.
- Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K. et al (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 349–369.

- Lee, G.R., Heo, L. and Seok, C. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 293–301.
- Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.a. et al (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.
- Modi, V. and Dunbrack, R.L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 260–281.
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A. et al (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(49), 1293–301.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, **84**(S1), 4–14.
- Notredame, C., Higgins, D.G. and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.
- Nugent, T., Cozzetto, D. and Jones, D.T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function, and Bioinformatics*, **82**(S2), 98–111.
- O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G. and Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *Journal of Molecular Biology*, **340**(2), 385–395.
- Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S. et al (2017). Protein structure determination using metagenome sequence data. *Science*, **355**(6322), 294–298.
- Pietrovovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, **24**(19), 3836–3845.
- Pirovano, W. and Heringa, J. (2010). Protein secondary structure prediction.
- Pirovano, W., Feenstra, K.A. and Heringa, J. (2008). PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, **24**(4), 492–497.
- Raval, A., Piana, S., Eastwood, M.P., Dror, R.O. and Shaw, D.E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, **80**(8), 2071–2079.
- Sadreyev, R. and Grishin, N. (2003). COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of Molecular Biology*, **326**(1), 317–336.
- Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**(3), 779–815.
- Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, **31**(13), 3381–3385.
- Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J. et al (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, **577**(7792), 706–710.
- Shi, J., Blundell, T.L. and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**(1), 243–257.
- Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **Suppl 3**, 171–176.
- Simossis, V.A. and Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, **33**(Web Server), W289–W294.
- Simossis, V.A., Kleinjung, J. and Heringa, J. (2005). Homology-extended sequence alignment. *Nucleic Acids Research*, **33**(3), 816–824.
- Söding, J., Biegert, A. and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, **33**(SUPPL. 2), 244–8.
- Song, Y., Dimaio, F., Wang, R.Y.R., Kim, D. et al (2013). High-resolution comparative modeling with RosettaCM. *Structure*, **21**(10), 1735–1742.
- Terashi, G. and Kihara, D. (2017). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function and Bioinformatics*.
- Thornton, J.M., Laskowski, R.A. and Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine 2021* **27**:**10**, 1666–1669.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T. et al (2021). Highly accurate protein structure prediction for the human proteome. *Nature 2021* **596**:**7873**, **596**(7873), 590–596.
- Wang, G. and Dunbrack, R.L. (2004). Scoring profile-to-profile sequence alignments. *Protein Science*, **13**(6), 1612–1626.
- Wang, S., Sun, S. and Xu, J. (2017). Analysis of deep learning methods for blind protein contact prediction in CASP12.
- Wuyun, Q., Zheng, W., Peng, Z. and Yang, J. (2016). A large-scale comparative assessment of methods for residue-residue contact prediction. *Briefings in Bioinformatics*, page bbw106.
- Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure

- fragments and optimized knowledge-based force field. *Proteins: Structure, Function and Bioinformatics*, **80**(7), 1715–1735.
- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**(1), 40.

# Chapter 9

## Structural Property Prediction

Maurits Dijkstra  Katharina Waury  Dea Gogishvili   
Punto Bawono  Juami H. M. van Gils  Jose Gavaldá-García   
Mascha Okounev  Robbin Bouwmeester  Bas Stringer   
Jaap Heringa  Sanne Abeln  K. Anton Feenstra 

\* editorial responsibility



## 1 Introduction

The previous two chapters (Chapter 6 and Chapter 7) have shown us that predicting the three dimensional structure of a protein molecule from its amino acid sequence has been largely solved in recent years, although some challenges remain (Liu *et al.*, 2021). Some structural properties, however, may be much easier to predict from sequence. Like tertiary structure, structural properties such as secondary structure, surface accessibility, flexibility and disorder, may be more strongly conserved than the primary sequence. Serving as building blocks for the native protein fold, these structural properties also contain important structural and functional information not apparent from the amino acid sequence directly.

Note that, with all the predictors mentioned in this chapter, one can only predict the propensities of amino acids to be part of a certain structural component such as:  $\alpha$ -helix,  $\beta$ -sheet or coil, protein-protein interaction (PPI) interface site, epitope or a hydrophobic patch. There are a few major reasons why structural property prediction is still a complicated task: *i*) the large fraction of the structural data used to train various machine learning models is coming from static X-ray crystallography studies. However, globular proteins are dynamic and static information does not capture its characteristics completely. *ii*) Proteins are part of a complex living system and do not exist in isolation: they may undergo various post-translational modifications, interactions or environmental alterations, leading to conformational changes not taken into account when considering it sequence or structure without context. Importantly, when predicting structural properties by training on PDB structures, one assumes a single structural conformation for each residue in a protein, not taking into account its dynamics. As such, knowledge of structural properties of a protein can contribute to tasks like fold recognition, but also be useful for multiple sequence alignment to find distant homologs, analysis of protein stability, and more generally for function prediction. In the next chapter (Chapter 11) we will return to function prediction.

Here, we will first give an introduction into the application of machine learning for structural property prediction, and explain the concepts of cross-validation and benchmarking. Subsequently, we will discuss two major concepts that play a key role in the characterization and prediction of structural properties: *i*) the patterns in hydrogen bonding observed in  $\alpha$ -helices and  $\beta$ -sheets, and *ii*) the intrinsic preference of different amino acids to be in certain types of structural environments. Next, we will review various methods that incorporate knowledge of these concepts to predict those structural properties, such as secondary structure, surface accessibility, disorder and flexibility, and aggregation.

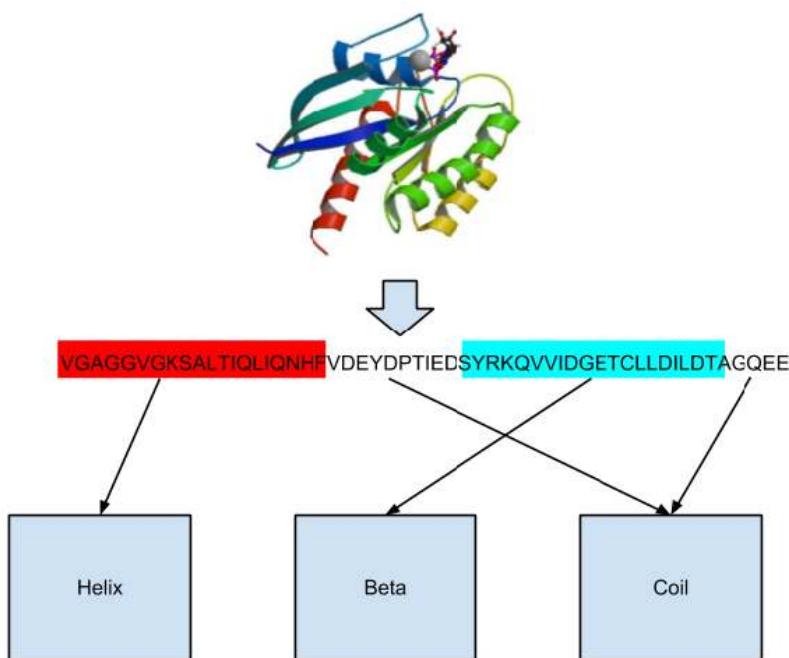


Figure 9.1: Secondary structure prediction as a classification problem. Each of the residues in a protein sequence will be classified as being either of the prediction classes – helix, strand or coil. Structure information (top part) is used as a reference for (supervised) learning, and as a gold standard for testing the accuracy of the predictions.

## 2 Structural property prediction as a machine learning problem

Structural property prediction can be approached as a supervised machine learning problem; the aim here is to find patterns in the data that can explain the associated outcome. For this aim labeled data to learn on is required that already contains the true outputs, the *labels*. Supervised learning can be divided further: Prediction of a specific *class*, for instance the secondary structural component, as shown in Figure 9.1 is a *classification* task. The prediction of a continuous *value* such as disorder or solvent accessibility is a *regression* task. The output of a supervised machine learning method is a predictor or model which allows us to predict the classes or values of an outcome variable. A more in-depth explanation of key concepts of machine learning are provided in Panel “Key concepts and typical tasks in machine learning”.

## Key concepts and typical tasks in machine learning

We often use machine learning algorithms to try to increase our understanding of complex biological problems. In this box, we introduce some basic terminology in machine learning for those that have little experience in this field.

When we pose a biological question, we are often interested in what characteristics are specific to a certain group of samples, or in other words, what separates one group from another. In this case, we usually have a certain number of samples for each of the groups that we want to compare. Usually, the more samples we have per subgroup the better, as it may enable us to better separate biological signal from technological and biological noise.

Machine learning algorithms can broadly be differentiated based on whether or not an algorithm requires a ground truth to find patterns in the data:

**Supervised learning:** When we train a model between groups for which the ground truth values (class label or continuous value) are known in advance, we are doing supervised learning. When we apply a supervised learning method, the data is split up into a training and a test set. The training set is used to train the model, whereas the test set is used to assess the performance of the model on data that it has not encountered before. This is necessary to avoid overfitting of the model. To prevent biased predictions, for most machine learning methods, the split in train and test should be balanced (similar percentage of labels or distribution of continuous values in train and test sets).

**Unsupervised learning:** When our aim is to identify interesting patterns in the data without prior knowledge about subgroups or correlations, we use unsupervised learning. Examples are principle component analysis (PCA), which calculates a weighted combination of variables that explains the largest variation in the dataset, and hierarchical clustering, which describes the similarity between samples using features of the dataset.

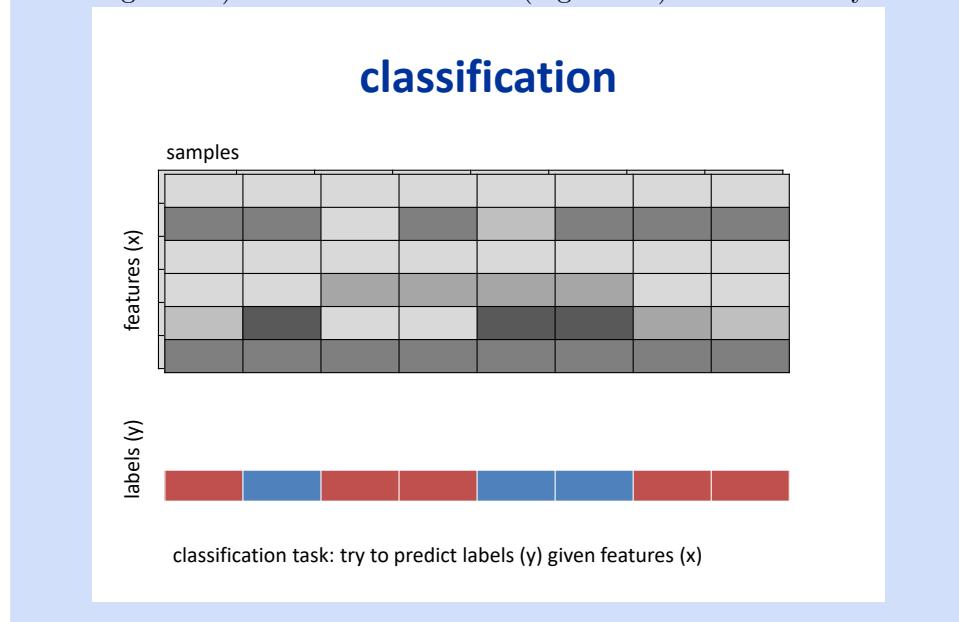
Supervised machine learning tasks commonly belong to one of the two following approaches depending on the type of outcome to predict:

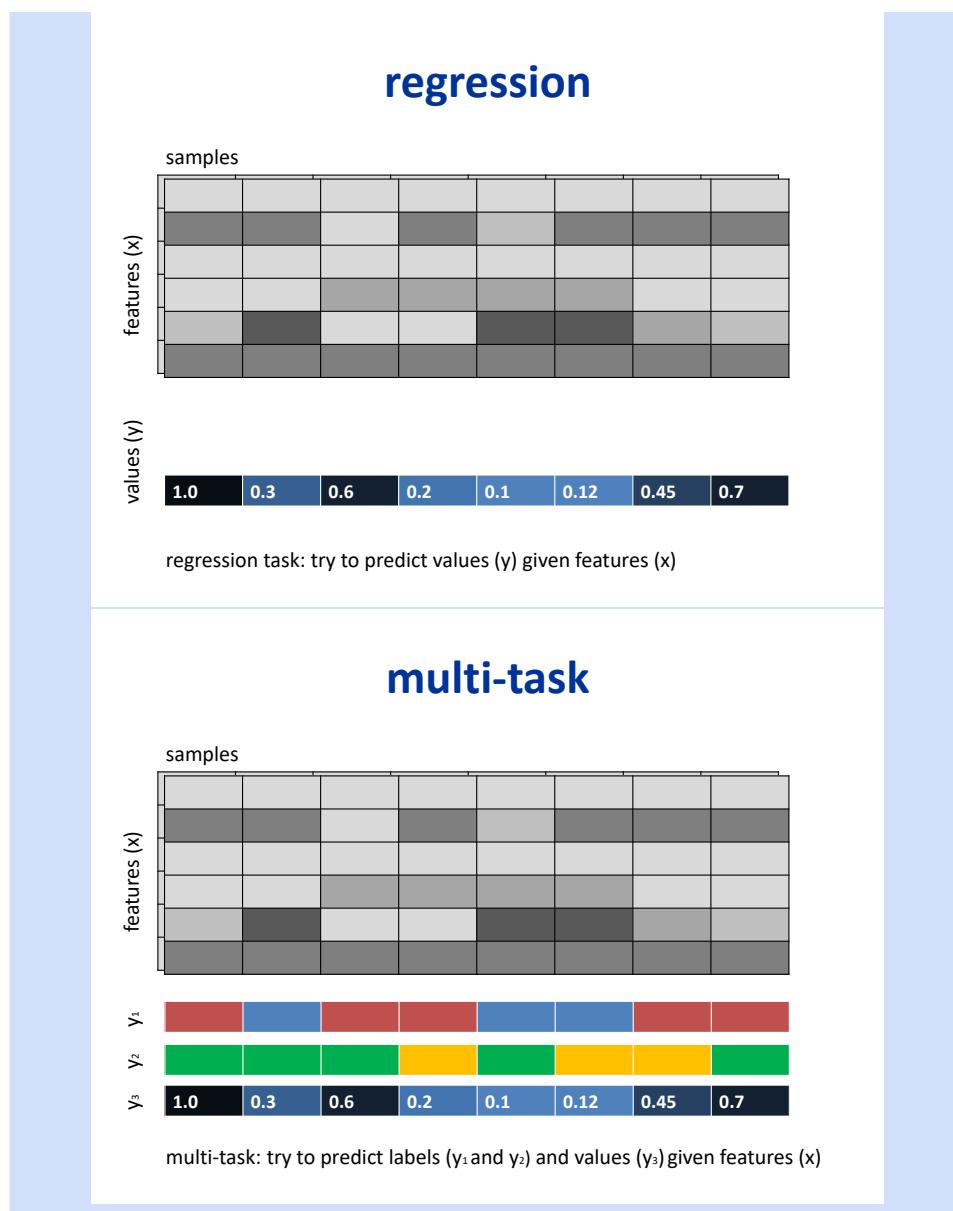
**classification:** when we try to classify samples into known classes, we want to predict the class labels from known variables. The variables in the dataset are called *features*, and the process of identifying the most relevant features to be used for a particular prediction task is called *feature selection*. Features can be either *continuous* (e.g. percentage of aromatic amino acids) or

*categorical* (e.g. known binding to RNA). Although the input features can be continuous or categorical, the output is always categorical.

**regression:** Alternatively, we may be interested in the relationship between some continuous features. For example, we might ask whether protein length is related to protein aggregation likelihood. To answer such questions, we use *regression* models. The most simplest is *linear regression*, which approximates a relationship between two variables with a linear model. If the slope of this fit is significantly different from zero, we can say that there is an association between the two variables. The prediction model then uses combinations of such linear fits to predict the value of the output variable from the input features.

**multi-task:** Moreover, we could be interested in classification and/or regression of multiple related output variables given one set of input features. This is called multi-task learning. With such methods we could for example predict secondary structural elements (classification), surface accessibility (classification or regression) residues and disorder (regression) simultaneously.





Most algorithms make structural property predictions per amino acid residue (e.g. NetSurfP-2.0). However, from the previous section you may already realise that we cannot predict the structural property of a residue in isolation: we need information on the surrounding residues, and potentially evolutionary conservation profiles around these positions in order to obtain accurate predictions. We refer to these kind of inputs to be used in a supervised learning method as *features*.

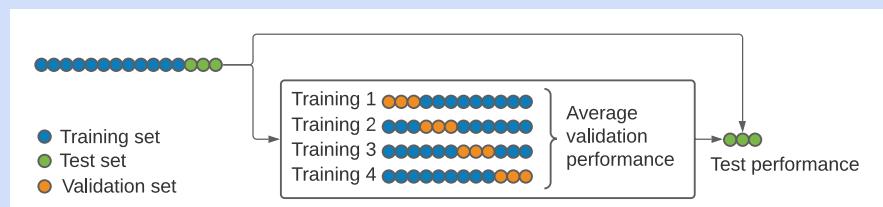
## 2.1 Training and benchmarking structural property predictions

Machine learning methods excel at finding patterns and relationships in data. However, for this they need to be trained on a big enough dataset of relevant samples, the *training dataset*. This training process is done on labeled data, this means the predictor can “see” the correct class or value of the samples as it is already known. By this the machine learning predictor can recognize which features are associated in what way with the output to be predicted and derive rules that can be used on yet unlabeled data to predict the output.

An important aspect of training a supervised machine learning model is to estimate its performance while adapting during the training process. A dataset independent from the training data is needed for this, as reported performance in the training set may be inflated due to overfitting. A frequently used method in machine learning to estimate the performance is cross-validation, in which part of the training dataset is intentionally left out to be used as a *validation dataset* (see Panel “N-fold cross-validation”). Furthermore, to measure the performance of the final machine learning model, a part of the available data is usually completely kept out from the training process to be used afterwards as the *test dataset*.

It is important to realise that close homologs typically have both high sequence, functional and structural similarity. As a result, there is a danger that models become biased towards a certain sequence composition and are not representative of the full spectrum of sequence variation that may be encountered when applying the method to new data. Therefore, it is important to use a training and test set that do not contain (very) close homologs. In other words, the PDB structures should be filtered for sequence similarity before they can be used in model training (Rost and Sander, 1993). Additionally, a sequence conservation profile may be generated for each of the PDB structures to include evolutionary information in the model, e.g. using a PSSM or an HMM model.

### N-fold cross-validation

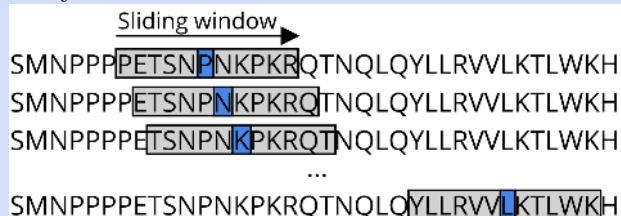


N-fold cross-validation splits the dataset in  $N$  equally-sized subsets. For every training instance, one of the subsets is taken out of the dataset as a test set. The remaining samples constitute the training

set. This set is used to train the model. The performance of the model trained on the training set is measured by comparing predictions of the samples in the test set to their actual values. This training is performed  $N$  times, until every subset has been used as a test set. Finally, the performances of all training instances are averaged to obtain a global measure of the model's performance.

### Sliding window, convolution and recurrent units

Non-locality is an interesting feature for predicting structural properties from a sequence, especially due to long-range interactions. Some ways of dealing with this are the use of a sliding window, convolutions or recurrent layers.



A sliding window is used to average the values of a property (e.g. hydrophobicity) in a range of amino acids of a predetermined length.

Convolutions are the defining elements of Convolutional Neural Networks (CNNs). They consider a position's surroundings by processing its information through the application of filters or kernels. The values of a region of the input are multiplied by those in a kernel, summed and stored. Then, the filters are applied to an adjacent region of the input. These steps are continued until all the positions of the input have been visited. It is the weights of the kernels that are learned in the training stage of such methods. The kernels can be thought of as units that can learn specific motifs. The creation of kernels is usually automated with most implementations, such as `pytorch`<sup>a</sup>, so we do not actually define which operations we are being performed, in contrast with a pure sliding window approach. Convolutions are often combined with other convolutions and pooling steps prior to passing their outputs into a neural network, which will predict a label. These methods will not be expanded here, but resources for more information about them can be found at the end of this chapter.

Another way to predict long-range interactions is the recurrent units, the defining element of Recurrent Neural Networks (RNNs) (Yu *et al.*, 2019). These units retain information from the context of a sequence in a trainable manner. In contrast with sliding window,

information is not obtained up to a predefined sequence distance, but in a more flexible manner. Additionally, there are no pre-defined operations to be executed in each area, as is the case for CNNs. If a piece of information may be useful to improve a prediction, it may be obtained even from further away in the sequence. Often, these gates include some kind of forget gate in most implementations, like the `pytorch` implementation of Long-Short Term Memory (LSTM)<sup>b</sup>.

None of the methods described above is necessarily superior to another. There are, however, methods more appropriate for a given problem. The table below shows a summary of the main characteristics of each of the 3 methods described in this box, to provide good starting points to identify which strategy is suitable for a learning task.

	Sliding window	CNN	RNN
Pre-defined learning size	Yes	Yes	No
Pre-defined operation applied	Yes	No*	No
Need for sequential order	No	No	Yes

\* Not with usual implementation

<sup>a</sup>[pytorch.org/docs/stable/generated/torch.nn.Conv1d.html](https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html)

<sup>b</sup>[pytorch.org/docs/stable/generated/torch.nn.LSTM.html](https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html)

## 2.2 Sequence signatures

Remember that the backbone parts of all naturally occurring amino acids are chemically identical (with the exception of proline, see Chapter 1 Panel “Amino acids, residues, and the peptide bond”). As secondary structure is stabilised by hydrogen bonding patterns between backbone atoms, the ability to form secondary structure is in essence a generic property of the peptide backbone. However, side chains have preferences for particular structural environments. Therefore, it is important to consider which patterns in the protein sequence are associated with specific structural elements. Those are the patterns that can serve as information sources from which structural property predictions can be made. Below we list these sources of information, and briefly put them in context of protein structures.

## 2.3 Evolutionary information

Sequence conservation patterns form a very strong indicator to recognise specific structural property types. All current state-of-the-art methods for structural property prediction take an evolutionary sequence profile as an

input instead of a single sequence. Such profiles may be provided as the output of a multiple sequence alignment (MSA), a position-specific scoring matrix (PSSM) from BLAST or provided as a Hidden Markov Model (HMM) profile (Woo *et al.*, 2004; Jones *et al.*, 1992). All of these provide information on the probability of observing a residue type at a certain position. For example, in loop regions it is seven times more likely to have gaps (in a multiple alignment) than in an  $\alpha$ -helix or  $\beta$ -strand.

### 3 Secondary structural element (SSE) prediction

Secondary structure prediction is distinct from the task of structure assignment. Structural property assignment is the task of assigning a structure label to each of the amino acids in a protein given the protein's three-dimensional (tertiary) structure, i.e. the full atomic coordinates as found in the PDB structure. Structural property assignment can be viewed as a way to define a structure, and is thus used to create a benchmark or gold standard against which to evaluate the performance of prediction methods; such assignment methods are also used to analyse protein structures. See Chapter 1 for further details on programs which perform secondary structure assignment, such as DSSP (Kabsch and Sander, 1983), DEFINE (Richards and Kundrot, 1988) and Stride (Heinig and Frishman, 2004).

#### 3.1 Hydrophobicity patterns

Sequence patterns of hydrophobicity can be a very strong indicator for secondary structure types. For example, helices that are partially exposed to the solvent, will have hydrophobic residues on the buried side of the helix and polar or charged residues on the exposed side. This will lead to a periodic pattern of alternating hydrophobic and hydrophilic residues, with a period of (on average) 3.6 residues, see Figure 9.2. (For more background, please refer to Chapter 1.) Similarly, a  $\beta$ -sheet with one side exposed to the solvent will show a sequence of alternating hydrophobic and hydrophilic residues (with a period of 2), see Figure 9.2. Loops are generally exposed to the solvent and therefore contain many more charged and polar residues than  $\alpha$ -helices and  $\beta$ -strands.

#### 3.2 Intrinsic preference of amino acids for certain secondary structure types

Each secondary structure type has a different preference for amino acid types. For example, amino acids with side chains that are **bulky** close to the backbone – more precisely, that have a branched structure at the C $\beta$  atom – tend to favour  $\beta$ -strands, **smaller amino acids** tend to favour  $\alpha$ -helices and loops. We will return to this in more detail in Section 3.4, and

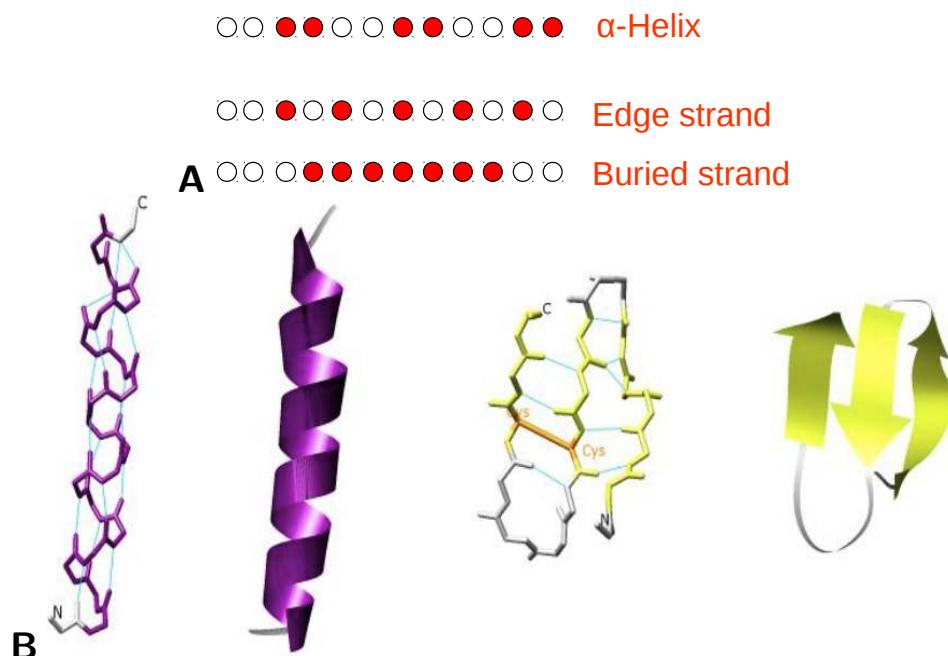


Figure 9.2: (A) Schematically and simplified, the hydrophobicity patterns in the sequence one may expect for different types of secondary structure elements; here, hydrophobic residues are indicated in red. (B) Examples of  $\alpha$ -Helical (left two) and  $\beta$ -strand (right two) structures. An  $\alpha$ -helix is often found at the protein surface, so that one side will be exposed to the solvent; this yields a sequence pattern of two hydrophobic, two hydrophilic residues, alternating. A  $\beta$  strand will often be buried, with only the first and last residues hydrophylic;  $\beta$  strands at the edge of the sheet, will have sidechains alternatingly sticking ‘back’ towards the protein (hydrophobic) and ‘out’ into the solvent (hydrophilic).

resulting propensities are shown there, in Figure 9.4. Furthermore, residues with non-standard backbone configurations such as glycine and proline are often named ‘helix breakers’ since they disrupt the helical pattern (Aurora and Rosee, 1998), and may often be found at the ends (caps) of helices. Both residues, glycine and proline, are also enriched in loop regions, as they generally disrupt regular secondary structure patterns (Branden and Tooze, 1998; Imai and Mitaku, 2005).

### 3.3 Locality of secondary structure

The interactions in an  $\alpha$ -helix are more local than those within a  $\beta$ -sheet, which have long-range interactions between residues on different strands as illustrated in Figure 9.3. Moreover,  $\beta$ -strands tend to be smaller continuous regions within the protein sequence compared to  $\alpha$ -helices, due to the extended conformation of the  $\beta$ -strand. Overall, this makes  $\alpha$ -helices relatively easier to predict than  $\beta$ -sheets. In order to properly identify non local

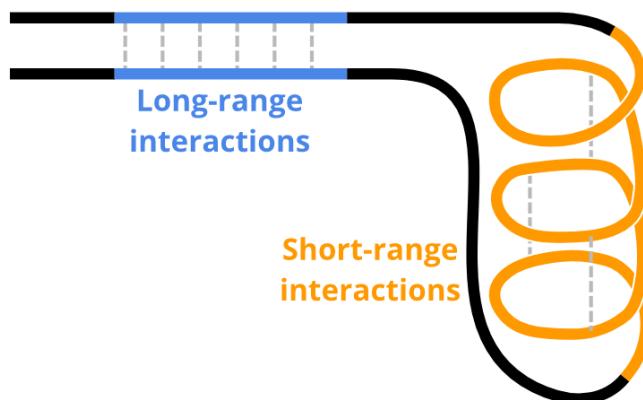


Figure 9.3: Interactions (dotted lines) in helical structure (orange, on the right) are always local; those between the strands in a sheet structure may be highly non-local (blue, on the left).

interactions one has to take the protein-wide context into account.

Most prediction methods suffered from a relatively poor performance when predicting  $\beta$ -strands because they only took the local context (in terms of primary structure) into account. Incorporating non-local interactions in a prediction method, however, is **far from trivial** if we use window based approaches (see also Panel “History of secondary structure prediction” below) (Baldi *et al.*, 1999; Magnan and Baldi, 2014). Recent progress in contact prediction (see Chapter 7 Section 2.4) enables use of additional input for  $\beta$ -strand allocation, which can have a high impact on (secondary) structure prediction.

### 3.4 Deriving Amino Acid Propensities

The first secondary structure prediction methods were developed during the 1970s, a time when only a handful of solved protein structures were available (Berman *et al.*, 2000). The general idea behind these early methods is that **each amino acid type has a different preference for different types of secondary structure**. We can quantify these preferences using amino acid **propensities**, derived from a large set of PDB structures with assigned secondary structures. Also, we can **calculate such propensities** for other types of structural features, such as **buried or interface regions**.

Generally, a propensity aims to reflect how much more likely a given amino acid is to be found in a certain environment than randomly. Let's first introduce the fraction (or probability) of amino acids  $p(\text{total})_{SS}$  in a particular structure type  $SS$ , e.g. the fraction of residues (in a protein) that are in an  $\alpha$ -helix:

$$p(\text{total})_{SS} = \frac{N(\text{total})_{SS}}{N(\text{total})}, \quad (1)$$

where  $N(\text{total})_{SS}$  is the total number of residues in structure type  $SS$ , and  $N(\text{total})$  is the total number of residues in the dataset. Let us furthermore consider the fraction of a specific amino acid type  $aa$  in a particular structure type  $SS$ :

$$p(aa)_{SS} = \frac{N(aa)_{SS}}{N(aa)}, \quad (2)$$

where  $N(aa)_{SS}$  is the number of amino acid type  $aa$  in secondary structure type  $SS$ , and  $N(aa)$  is the total number of amino acid type  $aa$  in all residue positions.

We can now calculate a **propensity**  $P$  for amino acid type  $aa$  for a **specific type of structure  $SS$** , by dividing the fraction of  $aa$  found in  $SS$  ( $p(aa)_{SS}$ ) by the overall fraction of  $SS$  ( $p(\text{total})_{SS}$ ), as follows:

$$P(aa)_{SS} = \frac{p(aa)_{SS}}{p(\text{total})_{SS}} \quad (3)$$

Note that the *propensity*  $P(aa)_{SS}$  is not the same as the *probability*  $p(aa)_{SS}$ ; propensity is a relative probability. When we calculate the propensity, we divide the fraction of residues of a specific amino acid in a secondary structure type by the total fraction of positions in that secondary structure type. Thus, a propensity below one indicates that an amino acid **avoids** that type of secondary structure, a propensity of around one indicates **no preference**, and a propensity larger than one indicates **a (strong) preference** of that amino acid for that secondary structure type. See also Panel “Example residue propensity” for an example of how to calculate a propensity in practice.

### Example residue propensity

If 30% of glutamic acid residues occur in an  $\alpha$ -helix, thus  $p(Glu)_\alpha = 0.3$ , and 20% of all residues are in an  $\alpha$ -helix, thus  $p_\alpha = 0.2$ , then the propensity of glutamate for  $\alpha$ -helix becomes:

$$P(Glu)_\alpha = \frac{0.3}{0.2} = 1.5$$

So, (in this example) glutamate has a preference for the  $\alpha$ -helix.

Given that we have a database of sequences for which the secondary structure has been assigned from 3D structure, we can **calculate the propensity** of observing an amino acid in a given type of secondary structure, see Figure 9.4.

A very simple method for secondary structure prediction then becomes a matter of checking **which propensity is biggest** for a given subset of sequence positions. We could do this is by looking at every residue in isolation, but

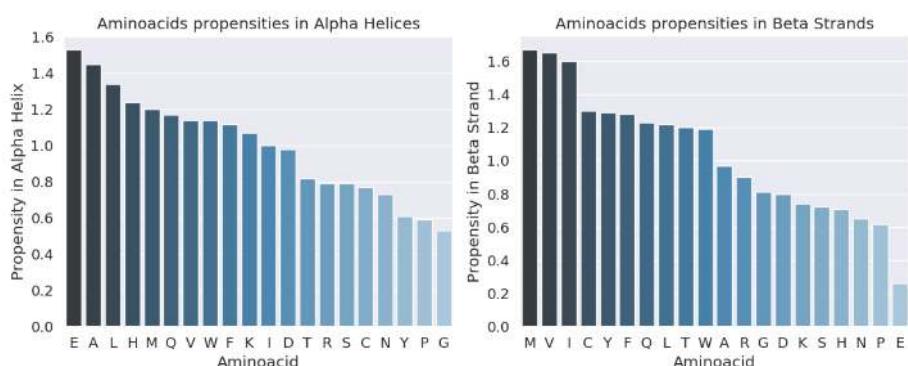


Figure 9.4: Propensities of every amino acid type in  $\alpha$ -helix and  $\beta$ -strand. Based on data from <http://www.bmrb.wisc.edu/referenc/choufas.shtml>.

this ignores the fact that the secondary structure of a residue is largely determined by its neighbours. A slightly more advanced approach is to average the propensity over a sequence window, as illustrated in the Panel “History of secondary structure prediction”. Note that every residue is still only considered in isolation, so implausible configurations such as a single helical residue in isolation could still be predicted. Early methods based on these principles are the Chou-Fasman algorithm (Chou and Fasman, 1978) and the GOR family of algorithms (Garnier *et al.*, 1996), see for more details the review by Pirovano and Heringa (2010).

### 3.5 Secondary structure prediction methods

To train and validate structural property prediction methods, large datasets are needed with labelled data. For this purpose, we can use the PDB in combination with (secondary) structure assignment software; some of the most used methods for secondary structure, interactions and buriedness are DSSP (Kabsch and Sander, 1983) or Stride (Heinig and Frishman, 2004).

Increasing numbers of available protein structures in the PDB allowed for the use of more complicated methods. These deep learning models namely require a large amount of training data. Neural net architectures that have been shown to be successful are e.g. (artificial) neural networks (ANN) (Rost and Sander, 1993), recursive neural nets (RNN) (Pollastri *et al.*, 2002; Baldi, 2003) and convolutional neural networks (CNN) (Wang *et al.*, 2016). An ANN consist of multiple connected nodes in which information flows from the input, through intermediate nodes to the final output node. An RNN is a specific kind of neural network which do not only have forward connections but also feedback connections (Goodfellow *et al.*, 2016). RNNs are able to capture similar trends as HMMs. A CNN is a neural network in which convolution instead of general matrix multiplication is used in at least one layer (Goodfellow *et al.*, 2016). Convolution is a mathematical operation

in which one function shifts over the input function to form together the output function (Burrus and Parks, 1985). CNNs can therefore capture similar trends as sliding windows. Which model can be used best depends on the dataset, the prediction task and the aim of the study. Besides, different methods may be combined into one architecture.

The most commonly observed secondary structures were introduced in Chapter 1. Methods which perform secondary structure prediction typically approach it as a three-state classification task: the problem of affixing one of three labels,  $\alpha$ -helix,  $\beta$ -sheet or coil, to each residue in a protein sequence.

Traditionally, secondary structure prediction has been approached as a classification task. More recent methods such as NetSurfP2.0 (Klausen *et al.*, 2019) have started to include prediction of other attributes, such as surface accessibility and backbone dihedral angles, which are regression tasks. Other methods such as PSIPRED (Jones, 1999; Buchan and Jones, 2019) treat secondary structure as a regression task, and predict propensities of an amino acid to be in sheet, coil or helix conformation.

### History of secondary structure prediction

Many different machine learning algorithms have been developed to tackle the problem of structural property prediction. Many of these methods incorporate evolutionary relationships by creating sequence profiles using (PSI-)BLAST, which are then fed into the model as training data (Rost and Sander, 1993; Woo *et al.*, 2004). Besides, evolutionary information can be stored in Hidden Markov Model (HMM) profiles, which also includes position dependent penalties for amino acid deletions and insertions (Eddy, 1996; Bystroff and Krogh, 2008).

In the beginning simple prediction models were used to make structural property prediction, e.g. multiple sequence alignment (Frishman and Argos, 1996), k-nearest neighbor (Salamov and Solovyev, 1995), decision trees (Rost *et al.*, 1994), and support vector machines (Hua and Sun, 2001). For many structural prediction tasks, like secondary structure prediction, it is beneficial to include non-linearity. Different approaches have been developed, e.g. by identifying correlations that indicate  $\beta$ -strand connectivity. These correlations can be used to strengthen the signal to predict the existence of a  $\beta$ -sheet in the sequence. For example, Predator uses weak sequence signals that indicate correlations between (contacting) residues in adjacent  $\beta$ -strands (Frishman and Argos, 1996). Propensity values for hydrogen bonding in a sliding window are used to predict  $\beta$ -strands. Other examples include PHD, which uses homologous sequences identified by BLAST to incorporate evolutionary information and SSPro, which uses three

sliding windows to identify possible  $\beta$ -strand interactions (Rost *et al.*, 1994; Pollastri *et al.*, 2002).

One of the earliest sophisticated methods - YASPIN (Lin *et al.*, 2005) utilised evolutionary profiles, HMMs and a neural network architecture with a slightly different strategy. The advantage of the YASPIN method was high speed mostly due to its simplicity: Instead of using an alignment algorithm directly, YASPIN method applied a 15-residue PSSM window generated from PSI-BLAST. The 7-state (instead of the common 3-state) secondary structural output was generated by the neural network in order to obtain more information that was afterwards filtered by an HMM to ultimately output a 3-state secondary structures. The major strength of the method was its ability to predict  $\beta$ -strands with high accuracy.

More recently, multi-task prediction architectures have been widely applied, during which multiple learning tasks (classification & regression) are solved simultaneously by using similarities and differences across these tasks. Such architectures have been created that are able to use sequence information to predict secondary structure in combination with other structural properties such as solvent accessibility, exposed vs. buried, disorder, backbone angle and residue contacts (Pollastri *et al.*, 2002; Heffernan *et al.*, 2015). The two leading examples are NetSurfP2.0 (Petersen *et al.*, 2000, 2009) and OPUS-TASS (Klausen *et al.*, 2019; Xu *et al.*, 2020). Multi-task learning can be used to improve a specific structural prediction tasks by learning related prediction tasks at the same time (Caruana, 1997). In previous methods, the structural annotations for a specific task were often used as input feature for the prediction of another task. Learning both tasks at the same time can transfer the same information and requires less pre-processing steps. Another recent developed model that tries to transfer information between multiple tasks and has received a lot of interest is the Transformer model (Devlin *et al.*, 2019; Rao *et al.*, 2019; Vig *et al.*, 2020). Importantly, these deep learning models require a substantial amount of training data that is becoming more easily approachable by the increasing amount of protein structural data.

### 3.6 Special cases

Apart from typical secondary structural components, there are multiple methods predicting 'meta-secondary' structures - motifs. Coiled-coil is a pair of helices that together form a twisted rod and are typically DNA binding (Chapter 1 Figure 1.7e). Due to the twist in the coiled-coil, there is a seven-residue repetitive element where residues of both helices are in direct contact. Typically, a leucine is found at each seventh residue, and in between (at each third and fourth residue), a valine or isoleucine. Due to

the repeating leucine, these structures are also known as ‘leucine zippers’ or ‘leucine-rich repeats’. This pattern makes them fairly easy to detect and predict (Lupas, 1997). COILS (Lupas, 1997), a profile based methods was the first coiled coil prediction algorithm developed. Later, HMM based methods like Marcoil (Delorenzi and Speed, 2002) were developed which improved predictions for short coiled coil regions. Recent methods like DeepCoil (Ludwiczak *et al.*, 2019) use deep learning and have higher sensitivity and accuracy than profile based and HMM based methods.

## 4 Other structural properties

### 4.1 Surface accessibility prediction

Among various structural properties, surface accessibility predictions of amino acids is of major importance. Residues that are exposed to the environment can have many different functions. For example, they can be part of the catalytic site of an enzyme or participate in PPIs. Furthermore, knowing which residues are on the surface of a protein can be important in drug design, e.g. in molecular docking (Ferreira *et al.*, 2015; Naderi-Manesh *et al.*, 2001). For proteins for which the 3D structure has been solved, we can easily determine which residues are on the surface of the protein. However, as we have seen in Chapter 4 “Data Resources for Structural Bioinformatics”, the structure of the large fraction of sequences has not been solved yet. Therefore, a lot of research is ongoing to improve our ability to predict the surface accessibility of residues using only the primary sequence.

One way to approach the problem of surface accessibility prediction is as a classification problem. Such methods predict whether a residue is buried, exposed or partially exposed based on a threshold and do not regard the absolute value of surface exposure (e.g. Naderi-Manesh *et al.*, 2001; Ahmad *et al.*, 2003). Other methods approach the prediction of surface accessibility as a regression problem and aim to predict which fraction of a residue is exposed (e.g. Petersen *et al.*, 2009; Wagner *et al.*, 2005).

Because secondary structure prediction and solvent accessibility prediction are methodologically similar problems, several methods have been developed that aim to tackle both (Heffernan *et al.*, 2015; Klausen *et al.*, 2019). Finally, solvent accessibility prediction can also be used to improve the accuracy of other structural property predictions and vice versa (Faraggi *et al.*, 2012; Klausen *et al.*, 2019).

### 4.2 Disorder and flexibility prediction

Disordered proteins or protein regions (Figure 1.7c,d), are those that lack a folded structure. Disorder prediction is relatively easy compared to 3D structure prediction. A simple but effective approach is to count amino acids

with high propensities (Oates *et al.*, 2013) – these are charged and polar (hydrophilic) amino acids – inside a sliding window over the sequence. There are more advanced predictors which use hidden Markov models (HMMs) (Cheng *et al.*, 2005) and Recurrent Neural Networks (RNN), for example, DisoMine (Orlando *et al.*, 2018). Recently, the Critical Assessment of protein Intrinsic Disorder prediction (CAID) experiment was designed to assess the prediction methods for intrinsic disorderd proteins (IDPs) (Necci *et al.*, 2021). Deep learning methods like and RawMSA outperformed the physicochemical based methods in the first CAID experiment (Mirabello and Wallner, 2019).

Flexibility is related to disorder, but not necessarily the same (e.g. Pancsa *et al.*, 2016). Protein flexibility influences a protein's biological activity like catalysis and stability. DynaMine is a dedicated method that aims to predict backbone and sidechain flexibility from sequence (Cilia *et al.*, 2014, 2013; Raimondi *et al.*, 2017).

### 4.3 Transmembrane prediction

Transmembrane (TM) proteins exist inside a membrane environment which is largely non-polar. Therefore, the membrane-spanning region of the TM protein will tend to have amino acids with hydrophobic side chains *on the outside* to match the apolar lipid environment of the membrane. In case of pore or channel proteins, the inside (enclosed by a ‘ring’ of helices or a  $\beta$ -barrel) will tend to be hydrophilic (Krogh *et al.*, 2001). The earlier methods in transmembrane topology prediction were based on hydrophobicity analysis (Yuan *et al.*, 2004). The recent methods utilize machine learning approaches including HMM (for example, TMHMM (Krogh *et al.*, 2001), HMMTOP (Tusnády and Simon, 1998)) and SVM (SVMtm) (Yuan *et al.*, 2004) for the prediction of alpha helices. Neural network approaches have been developed for the prediction of beta-barrels (for example, Jacoboni’s (Jacoboni *et al.*, 2001) and Gromiha’s (Jacoboni *et al.*, 2001) prediction methods).

### 4.4 Aggregation propensity prediction

Amyloid fibrils consist of insoluble protein structures that can form fibril structures through aggregation, see Chapter 1 Figure 1.7a,b. Early results show that several proteins associated with disease also have a high propensity for amyloid fibril formation (Chiti and Dobson, 2006). There are multiple amyloid fibril prediction tools (Zibaee *et al.*, 2007), however, reference databases are still small, making it difficult to validate such methods (Missonai *et al.*, 2015). Since protein aggregation is mostly linked to amyloid fibrils with cross- $\beta$  structure, various algorithms have been developed to predict aggregation-prone parts from the primary sequences. PASTA algo-

rithm, for example, is a protein aggregation predictor that was trained on a dataset of globular proteins of known native structure and predicts propensities of two residues to be a part of a cross-beta structure of neighbouring stands (Walsh *et al.*, 2014).

## 5 Practical advice

In the previous sections multiple methods for secondary structure prediction are discussed. In this section, we will provide some tips for end users of structural property-prediction algorithms:

- Firstly, check the recent literature for the latest best-performing methods, preferably from a review where all methods have been vetted in the same way on the same dataset.
- If possible, benchmark some of the best-scoring tools on relevant cases to get an idea of their accuracy for your purposes.
- Finally, check the similarity between the different methods, but also the similarity in prediction of specific regions. Methods or regions that get the same prediction with very different methods are generally the most reliable ones.

### Caveats

Using the most accurate secondary structure prediction methods, we predict more than 8 out of 10 residues correctly. Is it possible to do better still or is this close to the maximum attainable performance? There are a number of fundamental reasons why it is difficult, if not impossible, to achieve perfect predictive performance:

**Biases in the reference set.** A large fraction of the structures was experimentally solved using X-ray crystallography. This process, which requires a stable protein conformation to succeed, may lead to biases towards a more stable secondary structure. For example, a region that is disordered under normal conditions may be removed completely or be stabilized as e.g. a  $\beta$ -sheet in crystal form. Theoretically, NMR structure determination should suffer less from these problems as the proteins are measured in solution, but the heuristic algorithms used to find the most plausible conformations may still lead to biases.

**A static picture of the native state provides too little information.** This problem is related to the first in that it illustrates the limits of using purely static X-ray structure information as a reference. A globular protein in solution is not fixed in its

native conformation, but shows significant internal motion. Furthermore, the secondary structure may not even be stable in the native state. For example, a region may show constant transitions between a disordered state and a metastable helical state (Linding *et al.*, 2003; Kagami *et al.*, 2021b,a)

**A protein molecule does not exist in isolation in vivo.** Many proteins are post-translationally modified in a way that may induce a change in conformation (Xin and Radivojac, 2012). Other examples of environmental interactions include binding to other molecules after which a particular conformation is stabilized or even a conformational transition based on the acidity or temperature of the environment. When we try to predict the secondary structure from sequence we take none of such factors into account.

**The secondary structure may be determined in part by the tertiary structure.** We assume that secondary structure formation is always driven, and only driven by the primary structure alone. Some examples in nature show us that this may not be justified for all proteins (Holley and Karplus, 1989). In such cases, it may be that existing in a particular tertiary conformation stabilizes the secondary structure in a certain state. This leaves us with a chicken-and-egg problem where we would like to use the secondary structure as a stepping stone to solving the tertiary structure, but doing so would require solving the tertiary structure first.

In short, current methods reach an accuracy of 86-88%, which is (by far) accurate enough for most applications. The main remaining limitation seems to be that methods assume a single (secondary) structure exists for each residue in a protein, while this may in many cases be dynamic and depend on the environment or other conditions.

## 6 Key Points

- With structural property predictors, one can predict amino acid propensities to be a part of a certain structure, site or a patch, that can be either a label or a probability score.
- Structural property prediction exploits patterns of hydrophobicity and other amino acid propensities, as well as evolutionary patterns of conservation.
- Current secondary structure prediction methods routinely reach accuracies 86-88% for many structural properties, showing that this is

- much easier than tertiary (3D) structure prediction.
- Secondary structure and other structural property prediction methods are all based on machine learning approaches; this necessitates rigorous testing on independent test data.

## 7 Further Reading

- For further reading on secondary structure patterns – Branden and Tooze (1998), in particular Chapter 2 “Motifs of Protein Structure”
- “Biological Sequence Analysis” – Durbin *et al.* (1998)
- Review on secondary structure prediction methods – Pirovano and Heringa (2010)
- Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review – Rawat and Wang (2017)

## Author contributions

Wrote the text:	MD, PB, KW, DG, JG, JvG, SA, JH, KAF
Created figures:	JG, JvG, JH
Review of current literature:	JH, MD, JG, MO
Critical proofreading:	BS, SA, KAF
Non-expert feedback:	MO, RB
Editorial responsibility:	SA, KAF

The authors thank Ting Liu  for critical proofreading.

## References

- Ahmad, S., Gromiha, M.M. and Sarai, A. (2003). RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**(14), 1849–1851.
- Aurora, R. and Rosee, G.D. (1998). Helix capping. *Protein Science*, **7**(1), 21–38.
- Baldi, P. (2003). The Principled Design of Large-Scale Recursive Neural Network Architectures – DAG-RNNs and the Protein Structure Prediction Problem. *Journal of Machine Learning Research*, **4**, 575–602.
- Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**(11), 937–946.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. et al (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.
- Branden, C. and Tooze, J. (1998). *Introduction to protein structure*. garland publishing, New York.
- Buchan, D.W.A. and Jones, D.T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research*, **47**(W1), W402–W407.
- Burrus, C.S. and Parks, T.W. (1985). *Convolution Algorithms*. Citeseer.
- Bystroff, C. and Krogh, A. (2008). Hidden Markov Models for Prediction of Protein Features. In *Protein Structure Prediction*, volume 413, pages 173–198. Human Press.
- Caruana, R. (1997). Multitask Learning. *Machine Learning*, **28**(1), 41–75.
- Cheng, J., Sweredoski, M.J. and Baldi, P. (2005). Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery*, **11**(3), 213–222.
- Chiti, F. and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.
- Chou, P.Y. and Fasman, G.D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*.

- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. and Vranken, W.F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nature Communications*, **4**(1), 2741.
- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. and Vranken, W.F. (2014). The DynaMine webserver: Predicting protein dynamics from sequence. *Nucleic Acids Research*.
- Delorenzi, M. and Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**(4), 617–625.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL).
- Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- Eddy, S.R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3), 361–365.
- Faraggi, E., Zhang, T., Yang, Y., Kurgan, L. and Zhou, Y. (2012). SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry*, **33**(3), 259–267.
- Ferreira, L.G., dos Santos, R.N., Oliva, G. and Andricopulo, A.D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, **20**(7), 13384–13421.
- Frishman, D. and Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. "Protein Engineering, Design and Selection", **9**(2), 133–142.
- Garnier, J., Gibrat, J.F. and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology*, **266**, 540–553.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*. MIT press Cambridge.
- Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A. et al (2015). Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, **5**(1), 11476.
- Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, **32**(Web Server), W500–W502.
- Holley, L.H. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, **86**(1), 152–156.
- Hua, S. and Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, **308**(2), 397–407.
- Imai, K. and Mitaku, S. (2005). Mechanisms of secondary structure breakers in soluble proteins. *BIOPHYSICS*, **1**, 55–65.
- Jacoboni, I., Martelli, P.L., Fariselli, P., Pinto, V.D. and Casadio, R. (2001). Prediction of the transmembrane regions of  $\beta$ -barrel membrane proteins with a neural network-based predictor. *Protein Science*, **10**(4), 779–787.
- Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices 1 Edited by G. Von Heijne. *Journal of Molecular Biology*, **292**(2), 195–202.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature*, **358**(6381), 86–89.
- Kabsch, W. and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577–2637.
- Kagami, L., Roca-Martínez, J., Gavaldá-García, J., Ramasamy, P. et al (2021a). Online biophysical predictions for SARS-CoV-2 proteins. *22*(1), 1–7.
- Kagami, L.P., Orlando, G., Raimondi, D., Ancien, F. et al (2021b). b2bTools: online predictions for protein biophysical features and their conservation. *Nucleic Acids Research*, **49**(W1), W52–W59.
- Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K.K. et al (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, **87**(6), 520–527.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, **305**(3), 567–580.
- Lin, K., Simossis, V.A., Taylor, W.R. and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics (Oxford, England)*, **21**(2), 152–159.
- Linding, R., Jensen, L.J., Diella, F., Bork, P. et al (2003). Protein disorder prediction: Implications for structural proteomics. *Structure*, **11**(11), 1453–1459.
- Liu, J., Wu, T., Guo, Z., Hou, J. and Cheng \$, J. (2021). Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *bioRxiv*, page 2021.01.28.428706.
- Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. and Dunin-Horkawicz, S. (2019). DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, **35**(16),

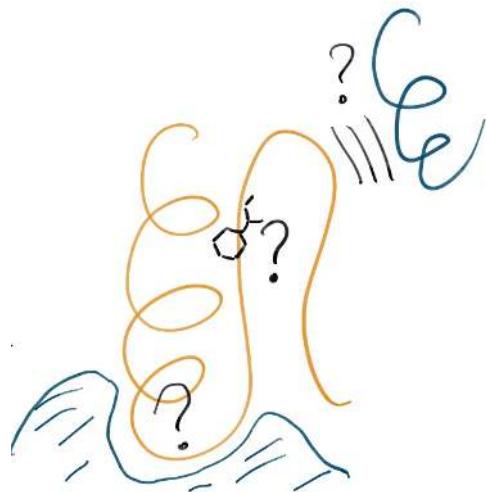
- 2790–2795.
- Lupas, A. (1997). Predicting coiled-coil regions in proteins. *Current Opinion in Structural Biology*, **7**(3), 388–393.
- Magnan, C.N. and Baldi, P. (2014). SSPro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**(18), 2592–2597.
- Micsonai, A., Wien, F., Kernya, L., Lee, Y.H. et al (2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, **112**(24), E3095–E3103.
- Mirabello, C. and Wallner, B. (2019). rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE*, **14**(8), e0220182.
- Naderi-Manesh, H., Sadeghi, M., Arab, S. and Moosavi Movahedi, A.A. (2001). Prediction of protein surface accessibility with information theory. *Proteins: Structure, Function, and Bioinformatics*, **42**(4), 452–459.
- Necci, M., Piovesan, D. and Tosatto, S.C.E. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods* 2021 18:5, **18**(5), 472–481.
- Oates, M.E., Romero, P., Ishida, T., Ghalwash, M. et al (2013). D2P2: Database of disordered protein predictions. *Nucleic Acids Research*, **41**(D1), D508–D516.
- Orlando, G., Raimondi, D., Codice, F., Tabaro, F. and Vranken, W. (2018). Prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics. *bioRxiv*, page 2020.05.25.115253.
- Pancsa, R., Raimondi, D., Cilia, E. and Vranken, W.F. (2016). Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity. *Biophysical Journal*, **110**(3).
- Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology*, **9**(1), 1.
- Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H. et al (2000). Prediction of protein secondary structure at 80% accuracy. *Proteins: Structure, Function and Genetics*, **41**(1), 17–20.
- Pirovano, W. and Heringa, J. (2010). Protein secondary structure prediction.
- Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function and Genetics*, **47**(2), 228–235.
- Raimondi, D., Orlando, G., Pancsa, R., Khan, T. and Vranken, W.F. (2017). Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. *Scientific Reports*, **7**(1), 8826.
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y. et al (2019). Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, page 676825. Neural information processing systems foundation.
- Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, **29**(9), 2352–2449.
- Richards, F.M. and Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Genetics*, **3**(2), 71–84.
- Rost, B. and Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, **90**(16), 7558–7562.
- Rost, B., Sander, C. and Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Bioinformatics*, **10**(1), 53–60.
- Salamov, A.A. and Solovyev, V.V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments.
- Tusnády, G.E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of Molecular Biology*, **283**(2), 489–506.
- Vig, J., Madani, A., Varshney, L.R., Xiong, C. et al (2020). BERTology Meets Biology: Interpreting Attention in Protein Language Models.
- Wagner, M., Adamczak, R.R., Porollo, A. and Meller, J.J. (2005). Linear Regression Models for Solvent Accessibility Prediction in Proteins. *Journal of Computational Biology*, **12**(3), 355–369.
- Walsh, I., Seno, F., Tosatto, S.C.E. and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, **42**(W1), W301–W307.
- Wang, S., Peng, J., Ma, J. and Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*, **6**(1), 18962.
- Woo, H.J., Dinner, A.R. and Roux, B. (2004). Grand canonical Monte Carlo simulations of water in protein environments. *The Journal of Chemical Physics*, **121**(13), 6392–6400.
- Xin, F. and Radivojac, P. (2012). Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics*, **28**(22), 2905–2913.
- Xu, G., Wang, Q. and Ma, J. (2020). OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, **36**(20), 5021–5026.

- Yu, Y., Si, X., Hu, C. and Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, **31**(7), 1235–1270.
- Yuan, Z., Mattick, J.S. and Teasdale, R.D. (2004). SVMtm: Support vector machines to predict transmembrane segments. *Journal of Computational Chemistry*, **25**(5), 632–636.
- Zibaee, S., Makin, O.S., Goedert, M. and Serpell, L.C. (2007). A simple algorithm locates  $\beta$ -strands in the amyloid fibril core of  $\alpha$ -synuclein, A $\beta$ , and tau using the amino acid sequence alone. *Protein Science*, **16**(5), 906–918.

# Chapter 11

## Function Prediction

Bas Stringer  Annika Jacobsen  Qingzhen Hou   
Hans de Ferrante  Olga Ivanova  Katharina Waury   
Jose Gavaldá-García  Sanne Abeln\*  K. Anton Feenstra\* 



\* editorial responsibility

## 1 Introduction

As mentioned in Chapter 1, the main motivation underlying our interest in studying protein structures is that structure relates more closely to protein function than protein sequence does. However, there are still huge gaps in our knowledge and in the mechanistic understanding of molecular function of proteins. This raises the question on how well we can predict protein function, when little to no knowledge from direct experiments is available.

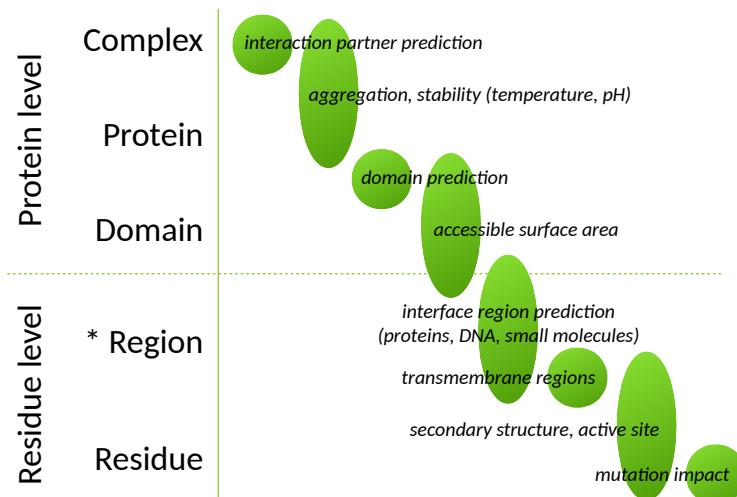
Function is a broad concept which spans different scales: from quantum scale effects for catalyzing enzymatic reactions, to phenotypes that can only be measured at the organism level, e.g. comparing healthy versus diseased state. In fact, there are different research areas which focus on the elucidation of function at different scales: biochemistry at the protein level, biology at the organism level, and (bio)medicine at the level of health and disease. In this chapter, we will consider prediction of a smaller range of functions, roughly spanning the protein residue-level up to the pathway level. We will give a conceptual overview of which functional aspects of proteins we can predict, which methods are currently available, and how well they work in practice.

## 2 Different types of function prediction tasks

Just as ‘function’ is a really broad concept, so is the field of protein function prediction. Examples of function prediction tasks include:

- cellular location for a protein;
- which molecular pathway a protein acts in;
- if the protein has alternative splice forms;
- proteins regions, e.g. transmembrane, and functional sites;
- if a pair of proteins is likely to interact or bind;
- if the protein is likely to form amyloid fibrils;
- protein stability at different temperatures, which will affect functionality;
- of a single nucleotide polymorphism (SNP).

We see that the purpose or output of the functional predictions acts at different levels of detail, roughly spanning residues, sequence regions, domains, proteins, complexes and pathways. Some tasks are very specific, such as prediction of the propensity to form amyloid fibrils, others are generic, such as predicting the likely functional impact upon mutation. In the next sections we will focus on three types of methods *i*) those that make functional predictions at residue-level and *ii*) those that make functional predictions at protein-level, and *iii*) those that make predictions on complexes (protein-protein interactions). Figure 11.1 gives an overview of general function prediction tasks organised according to the scale of output.



\* a region can be contiguous in structure, without being so in sequence.

Figure 11.1: Protein function prediction can be performed at different levels. Level of detail goes from the top quaternary complexes, e.g. proteins interacting to form a complex of multiple proteins, down to residue-level, e.g. which specific amino acid residues are important for a particular function. The different types of functional features that may be predicted range from overall prediction of aggregation or stability, down to the impact of a single residue mutation.

According to Bork *et al.* (1998), protein function may be best understood in terms of protein interactions. Protein interaction may mean quite different things in different contexts, i.e. at different levels. In Figure 11.2 we give an overview of which types of functional relations as well as physical interactions may be captured under the notion of PPI.

## 2.1 Different function prediction methods

Function prediction methods have some fundamental differences, both in terms of *input* and *methodology*. Some methods may be able to predict function from the sequence alone, where others need homology profiles or protein structures. The underlying methodology of function prediction can span many techniques, including molecular dynamics simulations, optimisation methods, various types of machine learning, structure prediction, sequence alignment, homology searches and network analysis.

Just as with structure prediction (see Chapter 7 “Practical Guide to Model Generation”), the highest accuracy for function predictions may be expected from methods that are based on homology; either by direct transfer of functional annotations from homologs, or by structure-based prediction of function based on predicted models of the structure of the proteins. However, for many proteins no additional structural or functional information

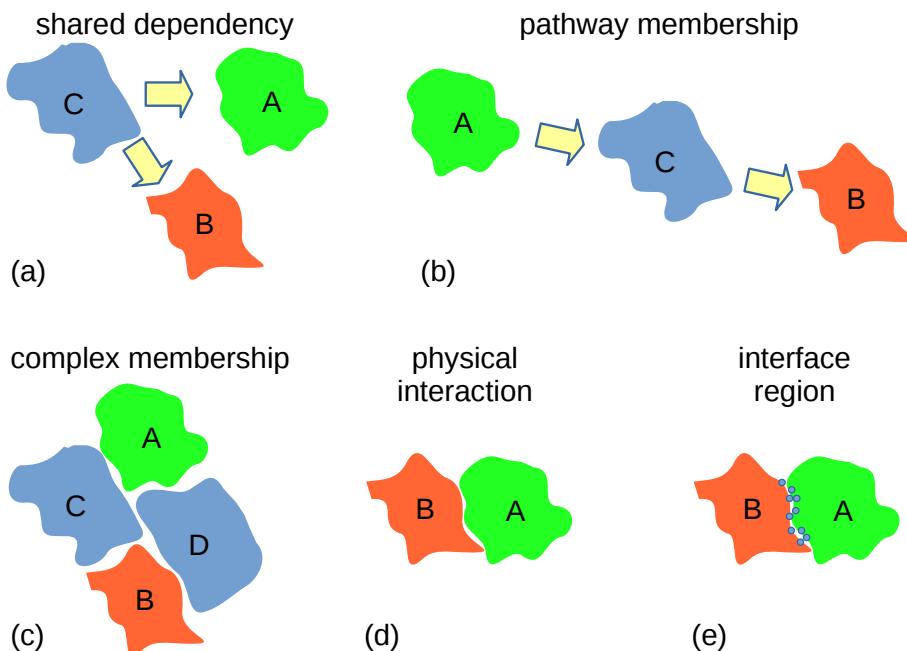


Figure 11.2: Overview of protein-protein interaction at different levels, and with different functional implications. (a) Mutual dependence: a correlation is observed between proteins A and B, caused by mutual dependence on protein C. (b) Indirect/cascade: the observed correlation between proteins A and B is mediated by protein C. (a) and (b) may arise through being in the same pathway. (c) Complex membership: proteins A and B are physically connected, but via intermediates C and D. (d) Direct interaction: proteins A and B are in direct physical contact. (e) The location of the interacting interface region.

is available, making it necessary to predict functional annotations based on sequence alone.

### 3 Residue level function predictions

The lowest level of protein function prediction we consider here is at the residue level. Prediction tasks that fall within this category are mutation impact analysis, active site prediction, and structural annotation predictions.

#### 3.1 Mutation impact analysis

Single base-changes (mutations) in the coding regions of a protein that result in an amino acid change in the protein are known as nonsynonymous SNPs (pronounced ‘snips’). A single SNP can have detrimental effects for the function of a protein, but most SNPs are functionally neutral (e.g., Mah

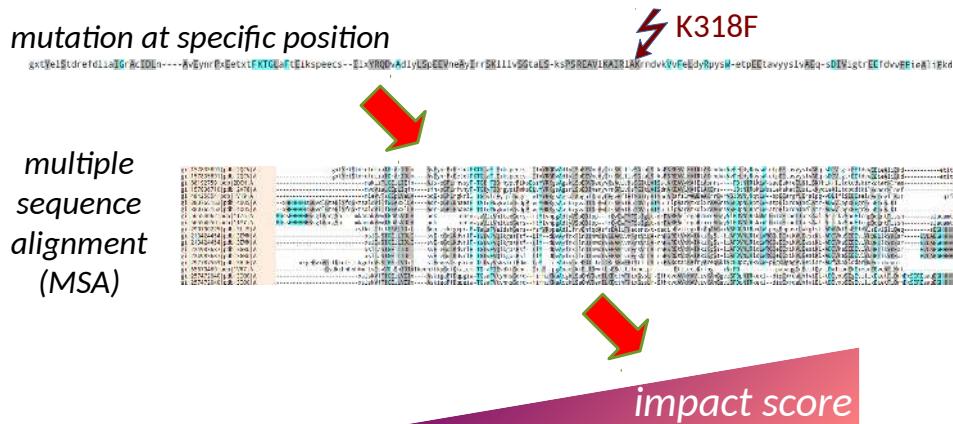


Figure 11.3: The concept of mutation impact prediction. From a given single amino acid change, using evolutionary information from multiple sequence alignment, one aims to assess the possible influence (impact) the mutation may have on the function of the protein.

*et al.*, 2011). This, and the fact that their abundance prohibits experimental analysis of all SNPs, has motivated development of a large body of bioinformatics tools that aim to predict the impact of a SNP.

Most of these methods exploit that functionally relevant residues are more strongly conserved, and SNPs of these residues with physicochemically distant residues are more likely to be deleterious (see Figure 11.3). Many tools also include (predicted) structural information to disentangle SNPs that affect structure from those that affect molecular function, and from those that are unlikely to affect either structure or function. Some well-known tools are from the Baker lab (Cheng *et al.*, 2005), Condel (González-Pérez and López-Bigas, 2011), PolyPhen-2 (Adzhubei *et al.*, 2013), and IMHOTEP (Knecht *et al.*, 2017); for an overview, including other tools such as SIFT, MAPP, PANTHER and MutPred, please refer to Brown and Tastan Bishop (2017). Some tools also allow assessment of impact of multiple SNPs, such as PROVEAN (Choi and Chan, 2015), others provide a comprehensive summary with all results explained for use by e.g. clinicians (Venselaar *et al.*, 2010).

### 3.2 Active site prediction

A similarly important prediction task at the residue level is to annotate the residues in a protein that are important for the function of the protein. The goal in this task is to identify residues in the active site(s) of the protein. Here we will focus on the binding of small ligands, for example in a receptor or an enzyme. Protein-protein interactions, or PPI, and their interaction interfaces will be covered later in this chapter.

The annotation of active site binding residues in a protein can be made on the basis of three types of information (Gherardini and Helmer-Citterich, 2008; Mills *et al.*, 2015): *i*) sequence, *ii*) local structure, and *iii*) global structure. Methods can use only one of these sources of information, or a combination of them. Roughly speaking sequence based methods find small sequence motifs (Lelieveld *et al.*, 2016), the local structure based methods inspect local curvature of a protein structure (Zhang *et al.*, 2011), and docking methods typically included an (ad-hoc) simulation of the full three-dimensional structure of the protein and ligand (Lensink *et al.*, 2016).

### 3.3 Structural annotation predictions

Many forms of structure annotation can be used, directly or indirectly, to infer protein function. For example, the presence of a transmembrane region would give a strong suggestion on the cellular location of a protein. Similarly, disorder prediction and surface accessibility prediction may provide clues on the molecular function of a protein. Such structural feature prediction methods are covered in Chapter 9.

One important thing to remember is that many of these structural annotations can be predicted accurately (70-85% accuracy), making them a reliable source of information.

#### Epitope prediction

An epitope is a region of a protein that is recognized by the immune system; i.e. it is a region to which an antibody binds. Epitope prediction is important for vaccine development, assay development for protein biomarker detection and antibody design for other purposes (Sanchez-Trincado *et al.*, 2017). Epitope prediction may both be sequence (Jespersen *et al.*, 2017) and structure (Kringelum *et al.*, 2012; Lin *et al.*, 2013) based. A good review of epitope prediction is Backert and Kohlbacher (2015). A recent method from our group is SeRenDIP-CE, which predicts conformational epitopes (Hou *et al.*, 2021).

## 4 Protein level function predictions

At the intermediate level, we can aim to predict function for a protein. Typically, such protein functions are predicted by making inferences through homology.

## 4.1 Inferring function through homology

With its three ontologies, Cellular Components, Molecular Function and Biological Process, the Gene Ontology consortium (Ashburner *et al.*, 2000; Carbon *et al.*, 2017) aims to enable exhaustive mapping of gene function for any protein (see also Chapter 4). Generation of such annotations however, requires costly and time-consuming experiments, and continues to be outpaced by the number of genes sequenced. To address this growing gap, researchers have aimed to automate functional annotation of proteins for already more than two decades (Bork *et al.*, 1998).

A first idea for this task would be to transfer functional annotations of proteins' closest homologs identified with such annotations (Radivojac *et al.*, 2013; Bernardes and Pedreira, 2013). We may infer these annotations since homologous proteins are more likely to take part in the same biological process, to be located in similar cellular compartments, and to have the same or similar molecular functions. Although this approach provides a good start for identifying the protein function, it does not take into account two things: *i*) sequence is less conserved than structure/function, *ii*) homology correlates imperfectly with functional annotations. Consequently, inference of functional annotations by homology transfer alone is error prone, even with levels of sequence similarity as high as >60% (Rost *et al.*, 2003)).

Often, protein-level functions are associated with a particular protein domain. When using a function prediction method it is important to realise if the methods have been developed to make predictions on a domain or full protein level. Note that there are also methods that predict the location of domains or domain boundaries; some further information is listed in the Panel “Domain prediction” in Chapter 6.

## 4.2 Critical Assessment of Function Annotation

The *Critical Assessment of Functional Annotation (CAFA)* experiment is CASP's equivalent for the protein function annotation task. In this large-scale experiment, different computational methods that automate protein function annotation with Gene Ontology terms are compared (CASP is introduced in Chapter 6). The experiment found that the field has progressed significantly from annotation by homology transfer alone, with top-performing methods combining statistical learning with data beyond sequence similarity such as protein-protein interactions, gene expression data, and protein structure (Tian and Skolnick, 2003; Radivojac *et al.*, 2013).

Limitations of the CAFA experiment include that quality and completeness of the GO annotations vary widely, making interpretation and usefulness of different tools dependent on the application and on which other supporting information is available (Jiang *et al.*, 2016).

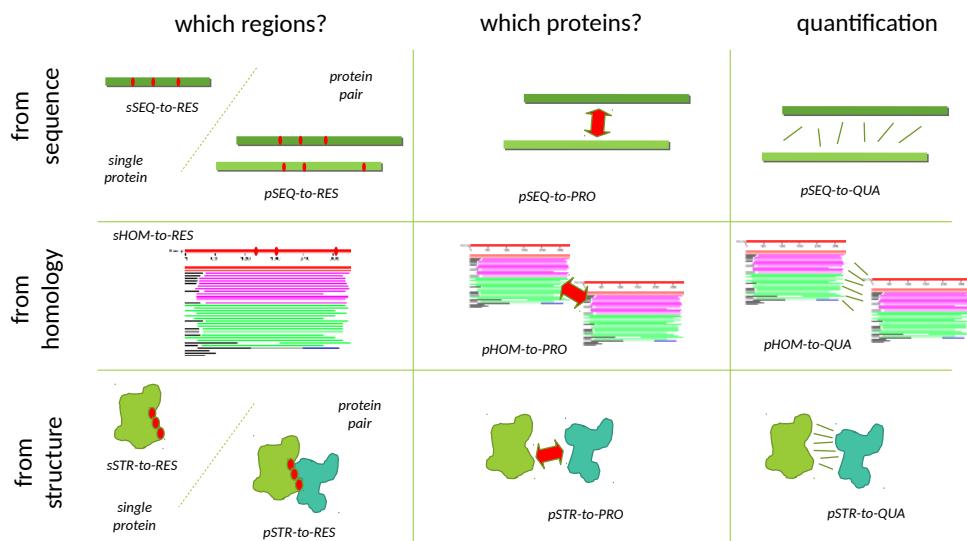


Figure 11.4: Levels of protein interaction prediction and types of input information. At the region level, one can predict which residues in a protein are most likely to participate in the interaction; this may be done for an individual protein without considering possible interaction partners, or for a putative interacting pair of proteins. At the protein level, one can predict which (pair of) proteins may interact, and one may furthermore quantify the interaction for example by interaction strength. Such predictions may be made from sequence data as input, from homologous sequences, or from structure data (or a combination).

## 5 Protein-protein interaction predictions

Knowledge of protein-protein interactions (PPIs) can help to narrow down the biological context of proteins, e.g. by suggesting in which pathways a protein is involved. PPIs may also be predicted, and such predictions can give information about function. Moreover, prediction of interaction interface, i.e. the residues of either protein that make contact with the other one inside a PPI, can help to assess the impact upon mutation (Ashworth and Baker, 2009). An overview of this is given in Figure 11.4.

### 5.1 Prediction of PPI from structure – docking method

PPI prediction, i.e. the computational prediction of the complex structure of interacting proteins where the protein structures are known, is called docking. Protein-protein docking is a hard and largely unsolved problem, even though we already have the structures for both proteins. In the docking method, one strong assumption is made out of necessity: the proteins do not change their conformation; typically only side chain rearrangements in the interface region are allowed. Without this assumption, the computational

cost of the predictions quickly becomes prohibitive.

Large conformational changes are hard to predict. In some way this is the same problem as for protein folding. Here, structures of homologs of the protein are used as a proxy for the possible conformations this protein could adopt, thus homology modeling is used to predict these conformations for our protein of interest. Then, regular ‘rigid’ protein-protein docking is used. (Lensink and Mendez, 2008; Lensink *et al.*, 2016)

## 5.2 Prediction of PPI from sequence

Prediction of PPI from sequence is an extensively studied field. Evolutionary and functional relation between proteins can be used for this purpose, since genes with closely related functions encode potentially interacting proteins. In prokaryotes functionally related proteins are often located in the same operon and thereby transcribed as a single unit. These genes can be predicted since the intergenic distance within an operon often is shorter than between operons (Shoemaker and Panchenko, 2007; De Juan *et al.*, 2013).

The phylogenetic profiling method exploits the fact that functionally related proteins, during evolution, sometimes get fused as domains into a single protein. Reversing that logic, when we observe two domains together in one protein, and we also observe homologs of the two domains as separate proteins, then we may assume these proteins are functionally related, and probably also interacting. (De Juan *et al.*, 2013).

## 5.3 Protein interface prediction

The goal of protein-interface prediction is to predict which residues constitute the interaction interface of proteins. The first task is to arrive at a definition of which residues are part of the interaction interface. Common definitions are that residues should fall below an intermolecular distance threshold, or that upon forming of the complex the accessible surface area of residues is reduced more than some threshold (Esmaielbeiki *et al.*, 2016).

After annotation, input of most protein-interface prediction tools is a single protein sequence (although some methods also take a pair of sequences as input) (Zhang *et al.*, 2018). Roughly four approaches are then taken conceptually to predict protein-protein interfaces for a given protein:

- i)* a sequence-based where only sequence information is used to predict interface residues (e.g., Murakami and Mizuguchi, 2010; Hou *et al.*, 2019),
- ii)* structure-based where structural information is included,
- iii)* a combination of sequence and structural information, and
- iv)* template-based where known interfaces of homologous proteins are used for interface prediction (e.g., Xue *et al.*, 2011).

The last option is by far the most reliable, if a suitable homolog complex is available. The combined option is a good second choice, if reliable structures of one or both of the interacting proteins are available (Melo *et al.*, 2016; Esmaielbeiki *et al.*, 2016).

Methods that predict PPI interface from sequence may utilize various classic Machine Learning (ML) (Cheng *et al.*, 2008; Hou *et al.*, 2017, 2021) and Deep Learning (DL) architectures (Shi *et al.*, 2021; Hanson *et al.*, 2018; Stringer *et al.*, 2021). Most of these methods use related structural features which are first predicted separately from sequence, such as secondary structure and solvent accessibility, as input features (Ofran and Rost, 2007; Li *et al.*, 2012; Hou *et al.*, 2017); Chapter 9 “Structural Property Prediction” gives an overview of methods that may be used to predict these structural features. The PPI interface prediction methods use conservation (Hou *et al.*, 2017; Zhang and Kurgan, 2019), secondary structure (Ofran and Rost, 2007; Zhang and Kurgan, 2019), surface accessibility (Chen and Zhou, 2005; Hoskins *et al.*, 2006; Zhang and Kurgan, 2019), backbone flexibility (Cilia *et al.*, 2013, 2014) or a combination of these (Hou *et al.*, 2017, 2019) as input features. We have recently investigated the different performances obtained from several different neural network architectures, and found that dilated convolutional networks (DCN) work well for protein interface prediction, but an ensemble network trained over the output of six other architectures (including DCN) always work best (Stringer *et al.*, 2021). Further improvements are expected by using multi-task approaches (Capel *et al.*, 2022).

## 5.4 CAPRI

CASP was already introduced in Chapter 6. In the CASP11 round, three functional aspects were explicitly scored: multimeric state, (small) ligand binding, and mutation impact. The multimeric state of proteins is which type of quaternary complex they participate in, or in other words, which and how many (other) proteins interact. These aspects were selected on being able to qualitatively evaluate them. Targets were selected that in solved crystal structure were dimeric, had a ligand bound, were from the crystallographers or in literature interest was expressed for evaluating mutants (Huwe *et al.*, 2016). For prediction of dimer structures, only in two cases out of ten a dimer model with reasonable accuracy could be generated for the majority of monomer model structures (Huwe *et al.*, 2016).

A related community for the critical assessment of prediction of protein interaction (CAPRI) explicitly deals with the prediction of PPI. In the 2015 round for ‘easy’ dimer PPI targets between 30-80% of models generated were of ‘acceptable’ or ‘medium’ quality out of a top 10 models per participating predictor method. However, for harder targets (difficult dimers, multimers and heteromers), this fraction dropped to below 10% (Lensink *et al.*, 2016). Encouragingly, it was seen that also protein 3D structure models of lower

quality could sometimes lead to acceptable or even medium quality models of the bound proteins (Lensink *et al.*, 2016).

For ligand binding, it was found in CASP11 that the accuracies of even the best models ( $\sim 2\text{\AA}$ ) are not good enough for accurate ligand docking (Huwe *et al.*, 2016). It was also the case for mutation impact prediction; for most targets, model accuracy did not correlate with accuracy of impact prediction (Huwe *et al.*, 2016). Apparently, either homology models are not yet accurate enough for these purposes, or methods are tuned to particular characteristics of crystal structures.

### Structure-Based Drug Design

If we have knowledge about the three-dimensional structure of the target protein (preferably obtained through experimental methods, like crystallography) it is possible to design ligands that have a high probability of binding to it. If those ligands perform a certain task, e.g. form an active complex or just the opposite - inactivate the target protein, then that ligand can be used as a drug. This is known as structure-based drug design (SBDD) (Blundell *et al.*, 1987; Blundell, 1996).

While it is possible to design a new drug based only on other (known) ligand structures (ligand-based drug design), there will always be a significant level of uncertainty whether the performed comparative analysis is correct. For a broad overview of related methods, both (protein) structure-based and ligand-based, please refer to this review on Cytochrome P450 modelling by Graaf *et al.* (2005) and the more recent ones by Sliwoski *et al.* (2014) and by (Ferreira *et al.*, 2015).

There are two basic approaches of designing a new drug based on a known structure. The first is a specific database search, where many potential ligands are screened, docked and scored based on how well they fit the binding site. Then, if needed, the found molecules may be modified in a desired manner and then scored again to see if they still fit. The second approach is to build a new molecule based solely on the binding site structure (its chemical and physical constraints), step by step, using a library of known fragments and applying a strategy (like growing the ligand from a “seed” fragment or linking best-fitting fragments). This approach has a significantly higher level of difficulty and computational complexity but allows to develop completely new molecules, not present in any database.

## 6 Key points

- Function prediction is an extremely diverse field
- Due to large gaps in our knowledge of (molecular) functions, function prediction algorithms and techniques are in high demand.
- Predictions may be made at residue, protein or pathway level
- Methods can be sequence or structure based, or combined
- Structural Bioinformatics methods such as molecular dynamics, homology recognition, sequence and structure alignment, and structural feature prediction are all used in function prediction.

## 7 Further reading

- Mah *et al.* (2011): SNP impact prediction
- Backert and Kohlbacher (2015): Epitope prediction
- Mills *et al.* (2015): Molecular function from structure
- Ferreira *et al.* (2015): Structure-based drug design
- Jiang *et al.* (2016): Critical assessment of function prediction

## 8 Author contributions

Wrote the text:	BS, JG, AJ, QH, OI, KW
Created figures:	JG, BS, KAF
Review of current literature:	BS, JG, KW, HdF, KAF
Critical proofreading:	SA, AJ, HdF
Non-expert feedback:	OI
Editorial responsibility:	SA, KAF

## References

- Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, **76**(1), 1–7.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. et al (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25.
- Ashworth, J. and Baker, D. (2009). Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic acids research*, **37**(10), e73.
- Backert, L. and Kohlbacher, O. (2015). Immunoinformatics and Epitope Prediction in the Age of Genomic Medicine. *Genome Medicine*, **7**, 119.
- Bernardes, J.S. and Pedreira, C.E. (2013). A Review of Protein Function Prediction under Machine Learning Perspective. *Recent Patents on Biotechnology*, **7**(2), 122–141.
- Blundell, T.L. (1996). Structure-based drug design. *Nature*, **382**, 23–26.
- Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**(6111), 347–352.
- Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F. et al (1998). Predicting Function: From Genes to Genomes and Back. *Journal of Molecular Biology*, **283**(4), 707–725.
- Brown, D.K. and Tastan Bishop, Ö. (2017). Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis: Analyzing Variation at the Protein Level. *Global Heart*, **12**(2), 151–161.
- Capel, H., Weiler, R., Dijkstra, M., Vleugels, R. et al (2022). ProteinGLUE multi-task benchmark suite for self-supervised protein modeling. *Scientific Reports*, **12**(1), 16047.

- Carbon, S., Dietze, H., Lewis, S.E., Mungall, C.J. et al (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, **45**(D1), D331–D338.
- Chen, H. and Zhou, H.X. (2005). Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins: Structure, Function, and Bioinformatics*, **61**(1), 21–35.
- Cheng, C.W., Su, E.C.Y., Hwang, J.K., Sung, T.Y. and Hsu, W.L. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. In *BMC Bioinformatics*, volume 9.
- Cheng, G., Qian, B., Samudrala, R. and Baker, D. (2005). Improvement in Protein Functional Site Prediction by Distinguishing Structural and Functional Constraints on Protein Family Evolution Using Computational Design. *Nucleic Acids Research*, **33**(18), 5861–5867.
- Choi, Y. and Chan, A.P. (2015). PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics*, **31**(16), 2745–2747.
- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. and Vranken, W.F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nature Communications*, **4**(1), 2741.
- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. and Vranken, W.F. (2014). The DynaMine webserver: Predicting protein dynamics from sequence. *Nucleic Acids Research*.
- De Juan, D., Pazos, F. and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**(4), 249–261.
- Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.C. and Deane, C.M. (2016). Progress and challenges in predicting protein interfaces. *Briefings in bioinformatics*, **17**(1), 117–131.
- Ferreira, L.G., dos Santos, R.N., Oliva, G. and Andricopulo, A.D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, **20**(7), 13384–13421.
- Gherardini, P.F. and Helmer-Citterich, M. (2008). Structure-Based Function Prediction: Approaches and Applications. *Briefings in Functional Genomics*, **7**(4), 291–302.
- González-Pérez, A. and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*, **88**(4), 440–449.
- Graaf, C.d., Vermeulen, N.P.E. and Feenstra, K.A. (2005). Cytochrome P450 in Silico: An Integrative Modeling Approach. *Journal of Medicinal Chemistry*, **48**(8), 2725–2755.
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y. and Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**(23), 4039–4045.
- Hoskins, J., Lovell, S. and Blundell, T.L. (2006). An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Science*, **15**(5), 1017–1029.
- Hou, Q., De Geest, P., Vranken, W., Heringa, J. and Feenstra, K. (2017). Seeing the trees through the forest: Sequencebased homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics*, **33**(10).
- Hou, Q., De Geest, P.F.G., Griffioen, C.J., Abeln, S. et al (2019). SeRenDIP: SEquential REmasteriNg to DerIve profiles for fast and accurate predictions of PPI interface positions. *Bioinformatics*.
- Hou, Q., Stringer, B., Waury, K., Capel, H. et al (2021). SeRenDIP-CE: sequence-based interface prediction for conformational epitopes. *Bioinformatics*, **37**(20), 3421–3427.
- Huwe, P.J., Xu, Q., Shapovalov, M.V., Modi, V. et al (2016). Biological function derived from predicted structures in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 370–391.
- Jespersen, M.C., Peters, B., Nielsen, M. and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, **45**(W1), W24–W29.
- Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R. et al (2016). An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy. *Genome Biology*, **17**(1).
- Knecht, C., Mort, M., Junge, O., Cooper, D.N. et al (2017). IMHOTEP—a Composite Score Integrating Popular Tools for Predicting the Functional Consequences of Non-Synonymous Sequence Variants. *Nucleic Acids Research*, **45**(3), e13–e13.
- Kringelum, J.V., Lundsgaard, C., Lund, O. and Nielsen, M. (2012). Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Computational Biology*, **8**(12), e1002829.
- Lelieveld, S.H., Schütte, J., Dijkstra, M.J., Bawono, P. et al (2016). ConBind: Motif-aware cross-species alignment for the identification of functional transcription factor binding sites. *Nucleic Acids Research*, **44**(8).
- Lensink, M. and Mendez, R. (2008). Recognition-induced Conformational Changes in Protein-Protein Docking. *Current Pharmaceutical Biotechnology*, **9**(2), 77–86.
- Lensink, M.F., Velankar, S., Kryshtafovych, A., Huang, S.Y. et al (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 323–348.

- Li, B.Q., Feng, K.Y., Chen, L., Huang, T. and Cai, Y.D. (2012). Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PloS one*, **7**(8), e43927.
- Lin, S., Cheng, C.W. and Su, E. (2013). Prediction of B-cell epitopes using evolutionary information and propensity scales. *BMC Bioinformatics*, **14**(Suppl 2), S10.
- Mah, J.T.L., Low, E.S.H. and Lee, E. (2011). In Silico SNP Analysis and Bioinformatics Tools: A Review of the State of the Art to Aid Drug Discovery. *Drug Discovery Today*, **16**(17-18), 800–809.
- Melo, R., Fieldhouse, R., Melo, A., Correia, J.D. et al (2016). A machine learning approach for hot-spot detection at protein-protein interfaces. *International Journal of Molecular Sciences*, **17**(8).
- Mills, C.L., Beuning, P.J. and Ondrechen, M.J. (2015). Biochemical Functional Predictions for Protein Structures of Unknown or Uncertain Function. *Computational and Structural Biotechnology Journal*, **13**, 182–191.
- Murakami, Y. and Mizuguchi, K. (2010). Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, **26**(15), 1841–1848.
- Ofran, Y. and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput. Biol.*, **3**(7), e119.
- Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M. et al (2013). A Large-Scale Evaluation of Computational Protein Function Prediction. *Nature Methods*, **10**(3), 221–227.
- Rost, B., Nair, R., Liu, J., Wrzeszczynski, K.O. and Ofran, Y. (2003). Automatic Prediction of Protein Function. *Cellular and Molecular Life Sciences (CMLS)*, **60**(12), 2637–2650.
- Sanchez-Trincado, J.L., Gomez-Perez, M. and Reche, P.A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of Immunology Research*, **2017**, 1–14.
- Shi, Q., Chen, W., Huang, S., Wang, Y. and Xue, Z. (2021). Deep learning for mining protein data. *Briefings in Bioinformatics*, **22**(1), 194–218.
- Shoemaker, B.A. and Panchenko, A.R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, **3**(4), e43.
- Sliwoski, G., Kothiwale, S., Meiler, J. and Lowe, E.W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews*, **66**(1), 334–395.
- Stringer, B., Ferrante, H.d., Abeln, S., Heringa, J. et al (2021). PIPENN: Protein Interface Prediction with an Ensemble of Neural Nets. *bioRxiv*, page 2021.09.03.458832.
- Tian, W. and Skolnick, J. (2003). How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *Journal of Molecular Biology*, **333**(4), 863–882.
- Venselaar, H., te Beek, T.A., Kuipers, R.K., Hekkelman, M.L. and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, **11**(1), 548.
- Xue, L.C., Dobbs, D. and Honavar, V. (2011). HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, **12**(1), 1.
- Zhang, J. and Kurgan, L. (2019). SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. In *Bioinformatics*, volume 35, pages i343–i353. Oxford University Press.
- Zhang, J., Ma, Z. and Kurgan, L. (2018). Comprehensive Review and Empirical Analysis of Hallmarks of DNA-, RNA- and Protein-Binding Residues in Protein Chains. *Briefings in Bioinformatics*, **20**(4), 1250–1268.
- Zhang, Z., Li, Y., Lin, B., Schroeder, M. and Huang, B. (2011). Identification of Cavities on Protein Surface Using Multiple Computational Approaches for Drug Binding Site Prediction. *Bioinformatics*, **27**(15), 2083–2088.



## **Part III**

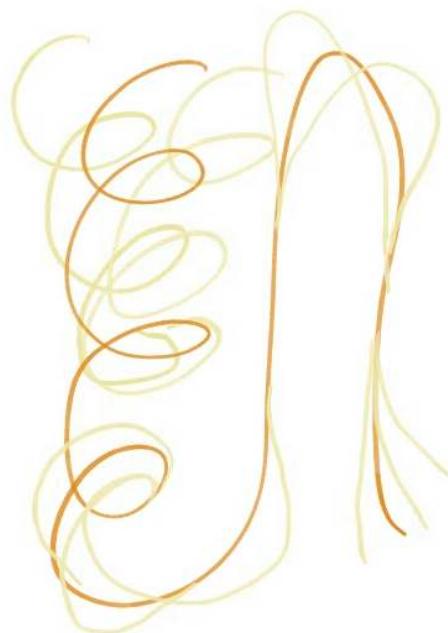
# **Dynamics and Simulation**



# Chapter 12

## Introduction to Protein Folding

Juami H. M. van Gils\*  Erik van Dijk  Ali May   
Halima Mouhib  Jochem Bijlard  Annika Jacobsen   
Isabel Houtkamp  K. Anton Feenstra\*  Sanne Abeln\* 



\* editorial responsibility



In this chapter we explore basic physical and chemical concepts required to understand protein folding. We introduce major (de)stabilising factors of folded protein structures such as the hydrophobic effect and backbone entropy. In addition, we consider different states along the folding pathway, as well as natively disordered proteins and aggregated protein states. In this chapter, an intuitive understanding is provided about the protein folding process, to prepare for the next chapter on the thermodynamics of protein folding. In particular, it is emphasized that protein folding is a stochastic process and that proteins unfold and refold in a dynamic equilibrium. The effect of temperature on the stability of the folded and unfolded states is also explained.

## 1 Protein folding and restructuring

### 1.1 Flexibility of protein chains & structural ensembles

In structural biology, it is generally believed that the protein sequence determines the 3D structure, which determines the function of the protein. Thus, a protein acquires its function once it is folded into its three-dimensional structure, the native state. This provides us with a very rigid view of protein structures. As we have seen in the previous chapters of this book, many methods in structural bioinformatics rely on this rigid view; examples are structure prediction, structure comparison and structure validation. However, proteins should in fact be viewed as flexible molecules that can take up a whole ensemble of different structural conformations (see Figure 12.1)

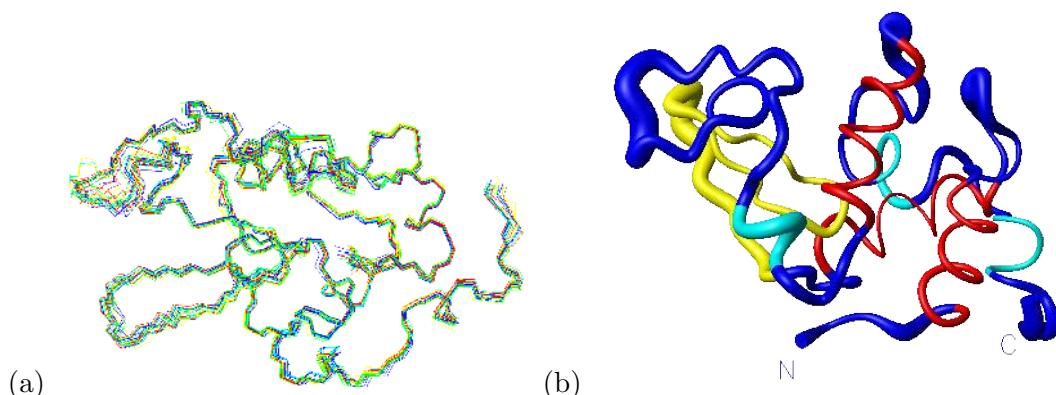


Figure 12.1: Proteins do not necessarily take one single structural conformation, but may instead be flexible. The native, functional state may contain many different structural conformations. (a) An ensemble of conformations based on NMR experimental data, shown as a set of overlaid backbone traces. (b) Another ensemble of conformations where the variation is shown as the thickness of the backbone - this is also known as the sausage representation.

for a simple illustration of a conformational ensemble for a protein structure). This ensemble may change depending on physical conditions such as temperature, pH, salt concentration, or the concentration of the protein in question. The conformational ensemble also may change depending on the presence of binding partners (e.g. small ligands, DNA or other proteins), a membrane, or crowding of the cell cytoplasm. In fact, exactly this dynamic interaction between the structural ensemble of a protein with its environment, is what allows for the functionality of proteins. Several examples of this are discussed in some detail in Chapter 5 “Protein Function & Interactions”.

It is not difficult to see why it is important to take the flexibility of proteins into account; for example protein folding, allostery (conformational change upon ligand-binding) and complex formation would not be possible without a protein changing its shape. However, most experimental and bioinformatics methods described in the previous chapters cannot (explicitly) deal with this flexibility. In this part of the book, we will study *simulation methods*, and explain the underlying *thermodynamic principles* behind these simulations. Moreover, we will consider how protein flexibility can be modeled explicitly: simulations allow us to consider the ensemble of different structural conformations of a protein. Simulations, in combination with experimental observations, can give us insight into the process of folding and, perhaps more importantly, investigate how the flexibility of its structure allows a protein to perform its function. We should therefore not think of protein structure and folding as deterministic or static phenomena, but as stochastic and dynamical processes.

## 1.2 Defining the folded and unfolded states

Before we go any further, it may be helpful to define the folded, or *native state* of a protein. Intuitively, you may have a good idea what such a folded state looks like, as most of the experimentally determined structures resemble a uniquely folded conformation (if the experiment was performed at very low temperatures); the folded state actually covers a small ensemble of conformations at physiological temperatures. In fact, it is the unfolded state that may be less intuitive: this state covers a large ensemble of (possibly extended) conformations of the peptide backbone. The exact nature of this ensemble may depend on the specific conditions in the system. For example, at low temperatures, the unfolded state may be more compact than at high temperatures (van Dijk *et al.*, 2016).

There are different experimental methods that can observe if a solution contains folded proteins, e.g. NMR (Nuclear Magnetic Resonance) or CD (Circular Dichroism) (Wüthrich, 1989; Kelly *et al.*, 2005). For example, NMR can be performed in solution, allowing full structural information to be resolved for small proteins; but only if the conformations are similar (hence

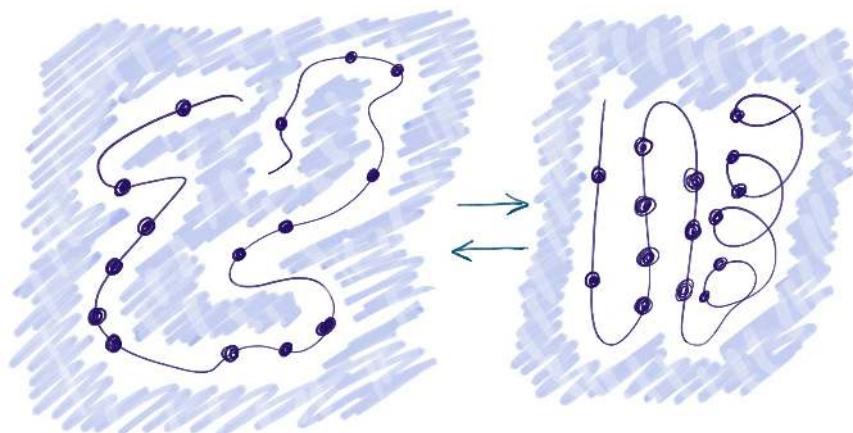


Figure 12.2: Denatured, unfolded protein chain on the left and the native, folded state on the right. The protein is shown in dark blue, water is shown in light blue. Dots on the protein indicate hydrophobic residues. One may observe that in the unfolded state interactions with the water (solvent) are far more extensive; more precisely there is a large interface between the solvent and the residues in the protein. In the folded state, only the outside of the protein interacts with the water, while hardly any solvent is present in the core of the protein. This is a result of the hydrophobic effect (see Section 3.1).

when a protein is folded). Alternatively, more indirect measurements can determine if the proteins in a solution are fully folded: for example, Green Fluorescent protein (GFP) will only show fluorescence when fully folded. This protein is therefore often used for folding experiments. Similarly, enzymes typically only show catalytic activity in their fully folded native form; for enzymes, enzymatic activity can report if the protein is folded.

In simulations on the other hand, we typically compare a simulation snapshot, or conformation, to the experimentally determined structure in order to see if it is folded. Two commonly used measures to indicate if a protein is folded are: 1) RMSD to the native structure - determined by superpositioning the conformation onto the native structure (see Chapter 3) or 2) the topological similarity of internal contacts - calculated by taking the intersection between the contact map of the conformation and the contact map of the native structure.

By analysing these measures over a simulation run, we can see that the number of possible conformations that is similar to a particular native structure is much lower than the number of conformations that are dissimilar to the native structure, with the latter ensemble of conformations defining the unfolded structure.

## 2 Folding and refolding

To further illustrate the dynamic and flexible nature of protein molecules, we will first try to sketch a picture of what we mean by protein folding. Consider we have a simple, *in vitro*, system of proteins dissolved in water (see Figure 12.2). These proteins, all with the same sequence, are two-state folders, meaning they are only stable in the folded or unfolded state. The folding and unfolding rate of the proteins is rapid, such that the proteins will reversibly unfold and refold over time. Moreover, the system is in a dynamic equilibrium, meaning that the fraction of unfolded proteins (and therefore also the fraction of folded proteins) remains stable over time (see the Chapter 13 for a detailed explanation of equilibrium). On a path from unfolding to folding, a single molecule will visit a large number of different conformations. Nevertheless, it will spend the majority of its time either in the folded or unfolded state.

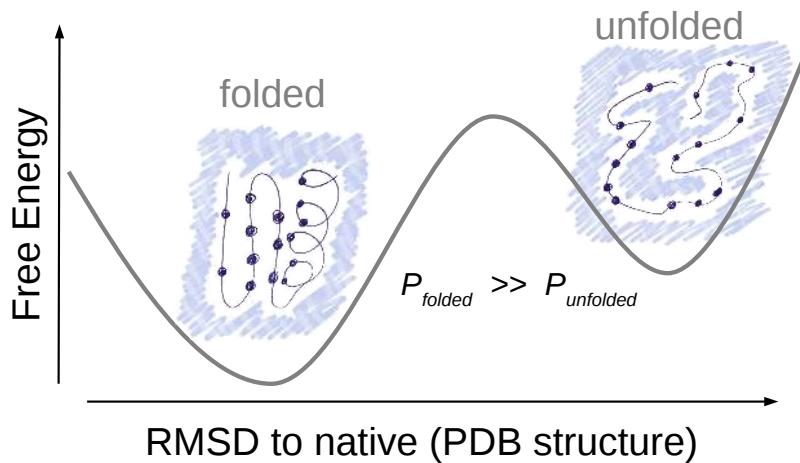
### 2.1 Stability and probability

Now, one of the most important observations to make is the relation between the *probability* of being in a certain state and the *stability* of that state in our simple system. Typically, under physiological conditions, around 30 °C, proteins would be most stable in their native, folded state. This means that the chance of finding a single molecule in the folded state would be much higher than finding it in the unfolded state; see Figure 12.3. Moreover, the fraction of folded molecules will, under these conditions, be much higher than the fraction of unfolded proteins. Another way of phrasing this is that the *free energy* of the folded state is lower than the free energy of the unfolded state. In Chapter 13 we will see that the probability of finding a molecule in a given state, is quantitatively directly related to the free energy of that state.

### 2.2 Changing conditions

In the system described above we can change physical conditions, for example the temperature. If we were to raise the temperature to about 70 °C, the unfolded state would become more stable than the folded state. In the next chapter we will see that this is an *entropic* effect of the peptide backbone: at higher temperatures states that are comprised of many possible conformations (in this case the unfolded state) are favoured. Note that other changes in conditions, such as altering the pH, salt concentration or adding a denaturant, e.g. urea, may also have a strong effect on the relative stabilities of the folded and unfolded states (McNay *et al.*, 2001; Sahin *et al.*, 2010).

Finally, please note that in an equilibrium situation the fractions of



*Figure 12.3: Sketch of a free energy landscape for a protein under physiological conditions. The protein is said to spend most of its time in the native or folded state (left well, low RMSD to native), as this state has the lowest free energy. Note that under these conditions, the native state is not exactly the same as the PDB structure but nevertheless very similar. The other local minimum (right well, high RMSD to native) represents the unfolded state.  $P_{folded}$  is the probability to find the protein in the folded state, which here is higher than  $P_{unfolded}$ : the probability to find the protein in the unfolded state.*

folded and unfolded conformations will be completely determined by the probability of these states, under the given conditions. However, when equilibrium has not been reached we get an unstable situation: for example, if a protein solution has just been heated up very quickly (known as a temperature jump experiment)(French and Hammes, 1969), the fraction of unfolded protein molecules is still small, even though the unfolded state may be more favourable under these new conditions. In this case, the system will relax over time, until an equilibrium is reached again.

### 3 Factors that (de)stabilize the native fold

So, why and how do proteins fold into their unique native structures? What are the important physical factors contributing to the transition of an unstructured polypeptide chain to a specific three-dimensional shape?

Since we know that the folded state should be the most stable state under native conditions, we can rephrase this question: why is the native state more stable than the unfolded state? There are many factors that

contribute to the stability of the native state.

### 3.1 Hydrophobic effect

It is thought that for the majority of globular proteins, the burial of hydrophobic side chains is the most important stabilizing effect on the protein structure (Tsong *et al.*, 1972; Baldwin, 2007): water molecules are strongly attracted to each other, due to the possibility to form intermolecular hydrogen bonds between the electropositive H and electronegative O atoms. Hydrophobic particles, cannot form such hydrogen bonds. Therefore a water molecule generally prefers to be situated among other water molecules, instead of forming an interface with a hydrophobic substance. In the unfolded state, many hydrophobic residues will form an interface with the water (see the left panel of Figure 12.2). In the folded state, on the other hand, the hydrophobic side chains do not form an interface with the water (see the right panel of Figure 12.2). Hence, with respect to the hydrophobic effect, the folded state is most favourable.

### 3.2 Hydrogen bonds, salt-bridges and packing

Other important factors that may stabilize the native state are van der Waals forces between the side chain atoms, hydrogen bonds between side chain atoms and/or backbone atoms, and salt bridges between charged side chains (Baldwin, 2007). Lastly, also some quantum effects can stabilize protein structures, a good example are the  $\pi$ - $\pi$  interactions formed by aromatic residues. Chapter 1 “Introduction to Protein Structure” explained many of these interactions in detail, and in Chapter 14 “Molecular Dynamics” we will describe some more detail on how these interactions can be modeled and calculated.

### 3.3 Backbone entropy

Lastly, there is also a factor that favours the unfolded state: the entropy of the backbone. In the next Chapter we consider this effect in more detail. For now, it is enough to understand that there are way more possibilities to generate an unfolded conformation than to exactly match the native conformation of a protein. So if none of the factors (e.g. hydrophobic effect and backbone hydrogen bonding) mentioned in the previous sections would favour the folded state, all proteins would be unfolded.

#### Anfinsen’s Theorem

Environmental factors such as solvent properties, temperature and pH are known to contribute to the specific three-dimensional structures of

the native protein. However, the most important determinant of the folded structure is the amino acid sequence. Anfinsen showed, in his Nobel Prize winning denaturation-renaturation experiments, that proteins can be denatured and then will spontaneously refold to their native forms when conditions are changed back (Anfinsen, 1973). These findings resulted in the general acceptance of what is now called the “thermodynamic hypothesis”, which states that the folded structure of a protein is fully encoded by its sequence, and the protein finds this structure due to thermodynamic laws.

We can rephrase the idea behind this theorem by saying that the folded state is the most likely, lowest free energy, state in the native conditions. For this to hold there are three important conditions:

- i) uniqueness of the free energy minimum, given the sequence,
- ii) stability of the free energy minimum,
- iii) kinetic accessibility of the free energy minimum.

Point i) suggests that a (naturally evolved) protein sequence, folds specifically into a specific structure; in other words the sequence is the recipe for the exact structure the protein will take. Point (iii) suggests that the folding and unfolding rates are sufficiently high or - in other words - that the barrier between the unfolded and the native state should not be too high. High barriers may in practice prevent a protein from reaching the folded state. Note that some proteins may require special conditions to fold; we will return to this in Section 5. In the last section of this Chapter we will see that, although the thermodynamic hypothesis seems to hold true for most naturally evolved proteins, there are also many proteins where the functional state, which is observed in nature, is not the one with the lowest free energy measured in *in vitro* in experiments.

## 4 Folding pathways

Previously, we considered a the case of a two-state folder: a protein for which the folding pathway only contains two stable states: the **folded and unfolded** state; moreover we assumed fast transitions between the folded and unfolded state. For many proteins, the folding pathways may be more complex, with **intermediate stable states**. For example, a multi-domain protein may fold one domain at a time, in a specific order. The state, in which only the first domain is folded would typically be a (meta-)stable state; this state is extremely likely to be visited on the path from the unfolded to the folded state and vice versa. More recently, it has been shown that for several proteins there exist smaller **intermediate folding structures**, or **foldons**, that appear as meta-stable states on the folding path (Englander and Mayne, 2017).

It is important to note that for different proteins, very different pathways have been observed experimentally (Hartl and Hayer-Hartl, 2009; Dobson, 2003). Moreover, generally folding and refolding is a **stochastic** process (e.g. Baclayon *et al.*, 2016). We will look at this in more detail in the last section of this Chapter.

## 4.1 Free Energy Landscapes

We have already shown a simple free energy landscape Figure 12.3. Many processes, like protein folding or protein-protein interactions, can be described as two-state processes. That means, there are **two free energy minima**, which are separated by a barrier. The top of the barrier is referred to as the **transition state**, i.e. the state through which the system must progress to go from one state to another. The height of the barrier determines the rate of the transition from one state to another. This makes knowing the **barrier height** important. However, from simulations it typically is difficult to sample a barrier, because the transition state is often **very unstable and therefore rarely visited**. The system will spend **most** of its time in the lowest of the two free energy **minima** (this is true for simulations and experiments). We will go in detail into the relation between free energy and probabilities in the next chapter, “Thermodynamics of Protein Folding”. Even though a folded protein will also visit the other (unfolded) minimum, states that cross the barrier are short-lived. There are several techniques to improve the sampling of the transition state in simulations, which we will return to in Chapter 15 “Monte Carlo for Protein Structures”.

### Levinthal’s paradox

How does this spontaneous folding occur? Levinthal argued that if a small protein would have to sample every possible three-dimensional conformation before obtaining its native structure, it would take more time than the age of the universe for it to find its native structure (Levinthal, 1969). To understand this, let us consider a protein of 100 amino acids, where each peptide bond in between two amino acids has two possible torsion angles, and each of these angles can assume three different values. The protein then has  $3^{99 \times 2} \approx 2.9 \times 10^{94}$  possible conformations. If each conformation can be visited in one picosecond ( $10^{-12}s$ ) it would take about  $10^{75}$  years for the protein to visit all possible conformations (our universe is  $13.8 \times 10^9$  years old).

However, it is known that small proteins like this can fold into their native structures in a matter of seconds. This phenomenon, known as “Levinthal’s paradox”, suggests that the folding protein only **samples a very small fraction of all possible conformations** before it finds its most

stable state. A typical protein would have a path towards the folded state that is relatively smooth, without very large barriers. Such a folding path may allow early formation of stable interactions, allowing the molecule to obtain its lowest energy state within reasonable time. Note that inside the cell, other factors, such as chaperones, may in fact make a folding path more smooth, see section Section 5.

## 5 Folding in the cell

Inside the cell (*in vivo*), the folding process occurs during and after the synthesis of the polypeptide chains in the ribosomes. The correct folding is necessary for the protein to perform its biological function. Up until this point, we have assumed protein folding to be a reversible process. Note that in practice some proteins are not able to refold *in vivo* after unfolding (or denaturation); some proteins only fold directly while being synthesized at the ribosome, and some proteins require chaperones to fold from an unfolded state. Other proteins, or protein regions, may only fold, upon binding a specific binding partner.

### 5.1 Chaperones

To prevent misfolding, folding *in vivo* is often aided by chaperones. GroEL is one of them, but there are several others (Horwich *et al.*, 2006). Most chaperones, including GroEL, are so-called heat-shock proteins, which were given this name because bacteria upregulate them as temperature increases. This makes sense as the chances of protein misfolding and aggregation increases with temperature. For many proteins this aid from chaperones is a necessity for reaching the native state or for refolding after denaturation. Chaperones have several functions in the cell. In addition to folding, there are also chaperones that prevent aggregation of misfolded proteins.

### 5.2 Folded proteins are only marginally stable

Most naturally occurring proteins are only marginally stable under physiological conditions (Privalov and Khechinashvili, 1974; Pucci and Rooman, 2017). This means that there is only a small free energy difference between the folded and unfolded state. This is most likely the result of evolution: proteins only need to be stable enough to perform their function, and have thus found a balance between stability and flexibility in order to be able to move and function. In fact, making them more stable may result in an additional cost when proteins need to be ‘cleaned up’ by degradation in the ubiquitin-mediated proteasome (Wilson *et al.*, 2020).

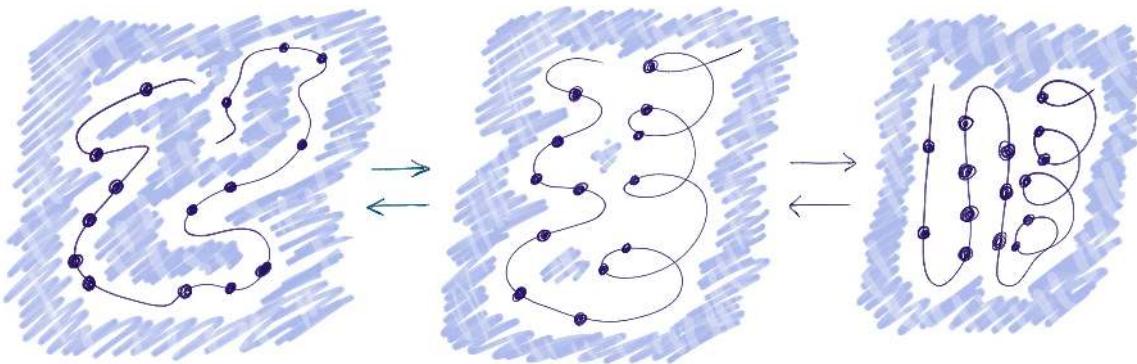


Figure 12.4: The transition from denatured (on the left) to folded (on the right) goes via some transition state which involves a hydrophobic collapse: all (or most) of the hydrophobic residues (here drawn as circles) are on the inside, but not all of the secondary structure has been formed yet. This intermediate state is often referred to as ‘molten globule’.

## 6 Alternative stable states of proteins

Besides the folded and unfolded state, there may be alternative (meta)stable states for a protein. As previously suggested, some of these may lie on the pathway from the folded to the unfolded state. There are also some states that, depending on the conditions, may actually compete with the native functional state. Some important alternative states are listed below.

### 6.1 Molten globules

So far we have considered a single unfolded state. In fact, different types of unfolded states can be defined ranging from states with mostly extended conformation, to states that are considerably compact. If such states are separated by a free energy barrier - we can truly observe different properties both by experiments and simulations.

The molten globule state is a compact state in which hydrophobic amino acid residues are clustered as drawn schematically in Figure 12.4. Local hydrophobic groups are formed to avoid unfavorable interactions with water. However, the core of the protein may still be somewhat more permeable than the fully folded state, because this state is typically less compact than the fully folded state and not all secondary structure elements are formed (Baldwin and Rose, 2013). Molten globule states may be observed under specific conditions, in which the molten globule state is more stable than the folded state. Alternatively molten globules may be meta-stable states along a folding pathway (Hartl and Hayer-Hartl, 2009; Dijkstra *et al.*, 2018). This latter concept is also referred to as ‘the hydrophobic collapse’.

## 6.2 Natively disordered proteins

Proteins for which the functional conformation is (largely) unfolded are called natively, or intrinsically, disordered proteins. Such proteins were already discussed in Chapter 5 “Protein Function & Interactions” Section 3.5.

## 6.3 Misfolding

*In vivo*, cellular processes may occasionally fail, including protein folding. Some proteins may end up in a misfolded state. Such a state would not be functional, but may be rather stable. From these states the barrier to a correctly folded state may be so high, that a transition is very unlikely, in this scenario the protein is kinetically trapped in a misfolded state.

Proteins may become misfolded by chance, through a ‘casualty’ on the folding pathway, through an abrupt change in conditions (e.g. heat shock), or through interaction with other proteins (e.g. prions).

Misfolded proteins could be very dangerous in the cell, as many hydrophobic residues may be exposed to the surface. In a crowded environment, such a misfolded protein would be very sticky, and could disrupt other parts of the cell (e.g. by making a membrane leaky or sticking to other proteins) (Relini *et al.*, 2014; Bondarev *et al.*, 2018).

## 6.4 Aggregation and amyloid formation

An even more dangerous form of misfolding, is where several protein molecules start aggregating together. Specific forms of such aggregates are amyloid fibrils, where multiple chains of proteins form large beta sheets (Chiti and Dobson, 2006; Dobson, 2003). This amyloid state, is - under several conditions - actually more stable than the folded state (Buell *et al.*, 2014). Meaning that if you wait long enough (think years), many proteins would end-up in amyloid fibrils.

In Chapter 1, we gave an example of misfolding of prion proteins, which leads to formation of  $\beta$  fibrils that disrupt cellular function and even kills cells. In general, misfolding of proteins is a problem that cells need to avoid. Therefore, in order to protect other elements in cell from misfolded proteins, a cell typically has an extensive machinery to aid folding, or to target misfolded and/or aggregated proteins for degradation. Chaperones (discussed in the previous section), are a part of this machinery.

### Protein folding in experiment and simulation

It is not straightforward to study the folding of proteins, neither in experiment nor in simulation. Specifically, it is extremely difficult to observe intermediate stages of the folding pathway. These intermedi-

ates will typically not be very stable and therefore are only present for short times, and at low concentrations. Experimental procedures, such as X-ray crystallography, NMR and various spectroscopic methods typically give information on structure, but limited detail on dynamical processes like folding (see also Chapter 2 “Structure determination” for an overview of these methods). Nevertheless, it is possible to obtain insight into which residues are important for intermediate states on the folding pathway. One trick is to see how the rate of folding changes, when mutations are made in a protein sequence. As the speed of folding is directly related to the height of the energy barrier between the folded and the unfolded states, one can infer if a mutation stabilizes or destabilizes the transition state (state on top of the folding barrier). This analysis, called ‘Phi ( $\phi$ ) analysis’ or ‘Alanine scanning’, is therefore an important trick to get an insight into the transition state and folding nucleus (Fersht, 1999). You see an example of it in Shaw *et al.* (2010), where it is applied to the small FiP35  $\beta$ -sheet peptide folding.

On the other hand computational folding simulations offer a possibility to study the underlying mechanisms of protein folding. For example, molecular dynamics (MD) simulations may be used to study protein folding for some specific, small proteins (Shaw *et al.*, 2010). However, generally speaking it is not possible to study the full folding of a protein using direct simulation techniques and fully detailed atomistic models. One of the reasons is that the length of the computational time needed to fold a protein is too large. Another problem is that the models used may not be accurate enough. To give an indication what is currently possible: Molecular dynamics (MD) simulations were used to obtain milliseconds-scale folding events of a small peptide (Daura *et al.*, 1998) and small proteins (Shaw *et al.*, 2010), however the folding of large proteins remains out of reach. More about molecular dynamics simulations and protein folding in Chapter 14.

Nevertheless, the combination of experimental observations and computational simulation can give us very good insight into the nature of folding proteins (Vendruscolo and Dobson, 2005; Knowles *et al.*, 2014; Fersht, 1999; Tompa and Fersht, 2009; Shaw *et al.*, 2010).

## 7 Key concepts

- Proteins can take many different conformations
- Under physiological conditions the native (functional) state is typically most stable
- Proteins fold, unfold and re-fold continuously → dynamic equilibrium
- Protein folding is a stochastic rather than a deterministic process

- In the crowded environment of the cell, special precautions must be taken to allow proteins to fold properly, and to avoid problems due to accumulation of misfolded proteins.
- Increasing the temperature increases the stability of the entropically favourable state. In protein folding, this is typically the unfolded state.
- Decreasing the temperature increases the stability of the enthalpically favourable state.

## 8 Further reading

- “Converging concepts of protein folding in vitro and in vivo” – Fersht (1999)
- “Energetics of Protein Folding.” – Baldwin (2007)
- “Physical and molecular bases of protein thermal stability and cold adaptation.” – Pucci and Roonan (2017)
- “Protein folding and misfolding.” – Dobson (2003)
- “Protein misfolding, functional amyloid, and human disease.” – Chiti and Dobson (2006)

## Author contributions

Wrote the text: JvG, EvD, HM, KAF, AM, SA

Created figures: JvG, JB, KAF, SA,

Review of current literature: JvG, IH, KAF, SA

Critical proofreading: AF, HM, JvG, SA

Non-expert feedback: JB, AJ

Editorial responsibility: JvG, KAF, SA

The authors thank Reza Haydarlou  for non-expert feedback.

## References

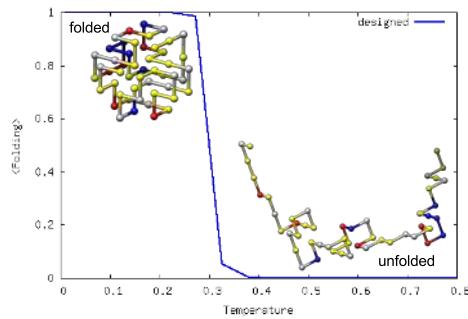
- Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223–230.
- Baclayon, M., van Ulzen, P., Mouhib, H., Shabestari, M.H. et al (2016). Mechanical Unfolding of an Autotransporter Passenger Protein Reveals the Secretion Starting Point and Processive Transport Intermediates. *ACS Nano*, page acsnano.5b07072.
- Baldwin, R.L. (2007). Energetics of Protein Folding. *Journal of Molecular Biology*, **371**(2), 283–301.
- Baldwin, R.L. and Rose, G.D. (2013). Molten globules, entropy-driven conformational change and protein folding. *Current Opinion in Structural Biology*, **23**(1), 4–10.
- Bondarev, S.A., Antonets, K.S., Kajava, A.V., Nizhnikov, A.A. and Zhouravleva, G.A. (2018). Protein co-aggregation related to amyloids: Methods of investigation, diversity, and classification.
- Buell, A.K., Dobson, C.M. and Knowles, T.P.J. (2014). The physical chemistry of the amyloid phenomenon: thermodynamics and kinetics of filamentous protein aggregation. *Essays in biochemistry*, **56**, 11–39.
- Chiti, F. and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.
- Daura, X., Jaun, B., Seebach, D., van Gunsteren, W.F. and Mark, A.E. (1998). Reversible Peptide Folding in Solution by Molecular Dynamics Simulation. *J. Mol. Biol.*, **280**(5), 925–932.

- Dijkstra, M., Fokkink, W., Heringa, J., van Dijk, E. and Abeln, S. (2018). The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. *Molecular Physics*, **116**(21-22), 3173–3180.
- Dobson, C.M. (2003). Protein folding and misfolding. *Nature*, **426**(6968), 884–890.
- Englander, S.W. and Mayne, L. (2017). The case for defined protein folding pathways. *Proceedings of the National Academy of Sciences*, **114**(31), 8253–8258.
- Fersht, A. (1999). *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. W.H. Freeman, New York.
- French, T.C. and Hammes, G.G. (1969). [1] The temperature-jump method. *Methods in Enzymology*, **16**(C), 3–30.
- Hartl, F.U. and Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.*, **16**(6), 574–581.
- Horwich, A.L., Farr, G.W. and Fenton, W.A. (2006). GroEL-GroES-mediated protein folding.
- Kelly, S.M., Jess, T.J. and Price, N.C. (2005). How to study proteins by circular dichroism.
- Knowles, T.P., Vendruscolo, M. and Dobson, C.M. (2014). The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, **15**(6), 384–396.
- Levinthal, C. (1969). How to fold graciously. In *Mössbau Spectroscopy in Biological Systems Proceedings*, volume 67 of *Univ. of Illinois Bulletin*, pages 22–24, Urbana, IL 61801.
- McNay, J.L., O'Connell, J.P. and Fernandez, E.J. (2001). Protein unfolding during reversed-phase chromatography: II. Role of salt type and ionic strength. *Biotechnology and Bioengineering*, **76**(3).
- Privalov, P.L. and Khechinashvili, N.N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *Journal of Molecular Biology*, **86**(3), 665–684.
- Pucci, F. and Roonan, M. (2017). Physical and molecular bases of protein thermal stability and cold adaptation. *Current Opinion in Structural Biology*, **42**, 117–128.
- Relini, A., Marano, N. and Gliozzi, A. (2014). Probing the interplay between amyloidogenic proteins and membranes using lipid monolayers and bilayers.
- Sahin, E., Grillo, A.O., Perkins, M.D. and Roberts, C.J. (2010). Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. *Journal of Pharmaceutical Sciences*, **99**(12).
- Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S. et al (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, **15**, 341–346.
- Tompa, P. and Fersht, A. (2009). *Structure and Function of Intrinsically Disordered Proteins*. Chapman and Hall/CRC.
- Tsong, T.Y., Baldwin, R.L., McPhie, P. and Elson, E.L. (1972). A sequential model of nucleation-dependent protein folding: Kinetic studies of ribonuclease A. *Journal of Molecular Biology*, **63**(3), 453–469.
- van Dijk, E., Varilly, P., Knowles, T.P.J., Frenkel, D. and Abeln, S. (2016). Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, **116**(7), 078101.
- Vendruscolo, M. and Dobson, C.M. (2005). Towards complete descriptions of the free-energy landscapes of proteins. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **363**(1827), 433–452.
- Wilson, A.E., Kosater, W.M. and Liberles, D.A. (2020). Evolutionary Processes and Biophysical Mechanisms: Revisiting Why Evolved Proteins Are Marginally Stable.
- Wüthrich, K. (1989). [6] Determination of three-dimensional protein structures in solution by nuclear magnetic resonance: An overview. *Methods in Enzymology*, **177**(C), 125–131.

# Chapter 13

## Thermodynamics of Protein Folding

Juami H. M. van Gils\*  Halima Mouhib  Erik van Dijk   
Maurits Dijkstra  Isabel Houtkamp  Arthur Goetze   
Sanne Abeln\*  K. Anton Feenstra\* 



\* editorial responsibility



In the previous chapter, “Introduction to Protein Folding”, we introduced the concept of *free energy* in the context of protein folding. The goal of that chapter was to start building an intuitive feeling of the meaning of the fundamental concepts that are important for this topic. Given a **free energy landscape** for the **conformational space** of a protein, we explained that conformations with low free energy are the most **stable** states of a protein (see Figure 12.3). In other words, these represent the states that are most **probable**, and proteins will typically spend most of the time in these low free energy states. The most stable state of a protein under normal conditions (e.g., room temperature and neutral pH) is often referred to as the **native state** of that protein. This is typically also the functional state of the protein. Although the protein is mostly in its native state under normal conditions, some fraction of the molecules may in fact be unfolded (typically as little as **1%** or even less).

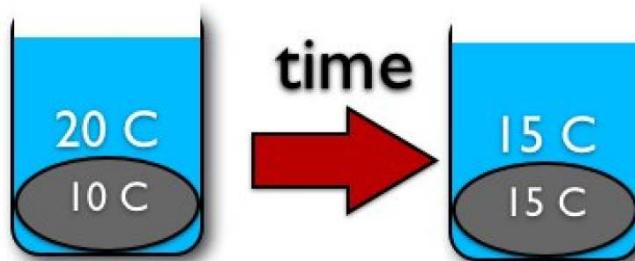
In this chapter, we will take a step back and try to get a deeper, more formal understanding of the concept of free energy in terms of the **entropy** and **enthalpy**; to this end, we will first need to better define the meaning of **equilibrium, entropy and enthalpy**. When we understand these concepts, we will come back for a more quantitative explanation of protein folding and dynamics.

## 1 Equilibrium and Dynamics

In molecular simulations of proteins, we consider proteins in a *dynamic equilibrium*. We do this under the assumption that over time, all systems move towards equilibrium. This means that the system will eventually reach a state where there is no net flow of energy or molecules between different parts of the system.

Take for example thermal equilibrium. When a cold metal object is placed into a warm water bath Figure 13.1, the two subsystems will exchange energy in the form of heat. However, initially the warm water will have higher thermal energy than the metal object, and there will be larger energy flow from the water into the metal. Therefore, the water will (gradually) cool down and the metal object will become warmer. At a certain point, the temperature of the water and the metal object have become equal, and there will be no net energy flow. The system has now reached thermal equilibrium. Note that this is a macroscopic property of the system; although there is no net flow of energy in the water bath, at the **microscopic level**, individual atoms in the system can (and will) still exchange energy when they bump into one another. We will elaborate on macro- vs microscopic properties later in this chapter.

Equilibrium also applies to protein folding and unfolding. When the system is in equilibrium, individual molecules can still (stochastically) switch



*Figure 13.1: Thermal Equilibrium.* If a cold metal object ( $T_1 = 10^\circ\text{C}$ ) is placed into a warm-water bath ( $T_2 = 20^\circ\text{C}$ ), the two components will eventually exchange energy until they have reached the same temperature ( $T_{\text{equilibrium}} = 15^\circ\text{C}$ ).

between the folded and the unfolded state, but the total number of molecules in each state remains **constant**. In other words, although protein molecules are still switching between states, the probability to be in a certain state is constant. As highlighted before in Chapter 12, the probability of being in a given state in the system is directly related to the free energy of that state: states with a low free energy have a high probability to be encountered, and are therefore **considered to have high stability**.

Finally, before we dive into the detailed material, we will provide a brief explanation of the difference between **classical and statistical thermodynamics**. We will need both, and understanding this difference will help you while reading the rest of the chapter. Classical thermodynamics is used to describe **macroscopic properties of the system**, such as the distribution of energy in the warm water bath described above, but does not consider **microscopic fluctuations**, such as the folding or unfolding of individual protein molecules. Statistical thermodynamics, also referred to as **(equilibrium) statistical mechanics**, describes how fluctuations at the microscopic (molecular) level of the system **lead** to the macroscopic behaviors that can be observed in experiments. Therefore, if we want to gain more mechanistic insight into properties of a process like protein folding, stability or function, we need to include the principles of statistical thermodynamics.

In the next sections, we begin by explaining the fundamental laws of thermodynamics, followed by the entropy and enthalpy. Finally, we will bring these concepts together to understand protein folding and dynamics from a thermodynamic perspective.

## 2 Thermodynamic laws

The laws of thermodynamics were first formulated in the 19th and early 20th century. They have several fundamental implications for protein dynamics, and many checks, tricks and assumptions used in simulations are based on

these laws. (Note that in literature the numbering of these rules is not always consistent.)

**Zeroth law of thermodynamics.** The zeroth law of thermodynamics states that if we have three states (A, B and C) in a system, and both A & B and B & C are in equilibrium, then A & C must also be in equilibrium.

**First law of thermodynamics.** The first law is equivalent to the law of conservation of energy; energy cannot be created or destroyed, just transformed from one form to another. This can be rephrased by saying that the amount of total energy in the system does not change:

$$\frac{\partial E_{tot}}{\partial t} = 0 \quad (1)$$

Examples of consequences of the first law:

- 1) Motor proteins in a cell use ATP to move along a microtubule (Hirokawa *et al.*, 1998; Mondal *et al.*, 2017). This process converts internal chemical energy (in the form of ATP) into kinetic energy or work (moving forward with a certain velocity), and heat.
- 2) The F-ATPase found in bacteria and mitochondria can use ATP to move protons ( $H^+$ ) across the cell membrane, or, reversibly, use a concentration gradient of protons to synthesize ATP (Mitchell, 1961; Frasch *et al.*, 2022). Here, chemical energy from the ATP molecule is converted into a different kind of chemical energy in the form of an electrochemical (charge & pH) gradient across the cell membrane.
- 3) Friction converts kinetic energy (movement) into thermal energy (heat). For example, if you rub your hands together because you are feeling cold), they become warmer.

**Second law of thermodynamics.** The second law of thermodynamics states that, for an isolated system, the entropy will never decrease (we will define entropy in the next section; for now, think of it as a measure for the amount of chaos/disorder). In other words, the entropy will keep increasing until it reaches the maximum possible value within the physical constraints of the system and then remain constant:

$$\frac{\partial S}{\partial t} \geq 0 \quad (2)$$

This also means that an isolated system always evolves to thermodynamic equilibrium (a state with no net flow of energy or molecules).

Examples of consequences of the second law:

- 1) The gas atoms in a room will not spontaneously go to one corner in that room.

- 2) A purely hydrophilic peptide sequence will not spontaneously fold in a hydrophilic solution, as the folded state is entropically unfavorable.
- 3) Mixed red and blue marbles in a box will not separate spontaneously when shaking the box (we will use this example in Section 3 “Entropy”).

**Third law of thermodynamics.** The entropy of a system at absolute zero (zero Kelvin), is constant and zero (actually close to zero due to quantum mechanical effects, and kinetically trapped states in so-called glassy systems) (Kittel and Kroemer, 1980). The third law has an interesting consequence: later in this chapter, we will see that **entropy becomes unimportant near zero Kelvin** (as the product of  $T$  and  $S$  approaches zero in Equation 6).

Examples of consequences of the third law:

- 1) There is no vibration of atoms in molecules at zero Kelvin (aside from quantum-mechanical effects).
- 2) At low temperature, a protein will optimize its internal interactions, i.e., minimise its total internal energy. (This property is the foundation of a simulation technique called ‘simulated annealing’, as we will see later in Chapter 14 and Chapter 15).
- 3) At high temperature, a protein will unfold, i.e., maximise its entropy.

Note that these laws apply to all physical systems, not just to the atoms, molecules and other details that we find important when investigating biological protein structures. Think of for example steam engines, the weather or galactic systems.

### 3 Entropy

In this section we will explain the concept of entropy in more detail. Entropy is a quantification of the amount of **conformational freedom** of the system. Loosely speaking, one may think of entropy as quantifying the amount of chaos or disorder in the system. More strictly, it is the **number of possible microstates** (structural conformations) that are **accessible** in a given macrostate (e.g., folded/unfolded state of a protein), as determined by the physical conditions.

To illustrate this concept on a simple system, take for example a box with spherical red and blue marbles that do not interact (i.e., there are no attractive or repulsive forces between the marbles). Initially, the marbles are sorted and all red marbles are at one side of the box and all blue marbles are at the other side (left panel Figure 13.2). If you shake the box, the marbles

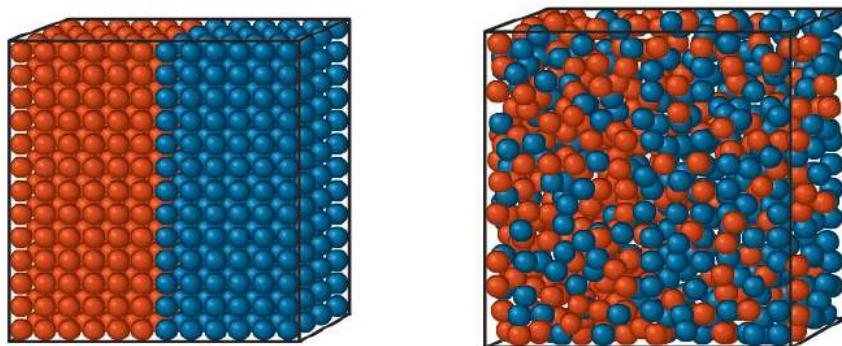


Figure 13.2: Box with marbles. Initially, the marbles are sorted with the red marbles on one side and the blue marbles on the other side of the box (left panel). If the box is shaken, the marbles will move around randomly. In the equilibrium state, the marbles are distributed homogeneously over the box (right panel).

will start moving around. After the box has been shaken for a sufficient amount of time, the red and blue marbles will be randomly distributed over the box. The number of red and blue marbles at each side of the box is now approximately equal. If you keep shaking the box, marbles will keep moving from one side to the other and back, but the total number of red and blue marbles at each side will **no longer change**. The system has reached **equilibrium** (right panel Figure 13.2).

### How to calculate the multiplicity of a state

The multiplicity of a state  $A$  can be calculated by

$$\Omega_A = \frac{N!}{k_A!(N - k_A)!} \quad (3)$$

where  $N$  equals the **total number of conformations** and  $k_A$  equals the **number of conformations in one particular state**.

In our example of the marbles, imagine we have ten red and ten blue marbles in the box and that only ten marbles can fit into one side of the box (see Figure 13.2). In this case, the number of conformations equals the total number of marbles at one side of the box, hence  $N = 10$ .  $k_A$  represents the number of red marbles at that side of the box. There are only two ways to have all red marbles at the same side, namely  $k_A = 0$  and  $k_A = 10$ . However, there are 252 ways to distribute the marbles equally over the two sides of the box (if  $k_A = 5$ ,  $\Omega_A = \frac{10!}{5!5!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{30240}{120} = 252$ ). Note that we do not care exactly which marbles are in which position, but only the total distribution of the differently coloured marbles over the sides of the box.

Is it impossible for the red and blue marbles to spontaneously return to the initial (sorted) state? Technically, no. However, the probability of such an event is extremely low. The probability of being in a certain state is dependent on the number of ways to reach that state. This is called the multiplicity of a state, and is denoted by  $\Omega$  (omega). For example, if four out of ten blue marbles are on the left side of the box, it does not matter which four these are, you will observe the same state, namely four blue marbles on the left side and six on the right side of the box.

Another example is coin tossing. If you flip a coin three times and throw two heads and a tail, it does not matter if you throw two heads first and then a tail or first a tail and then two heads. Either way you reach the same state, namely two out of three heads. In the context of protein folding, the multiplicity would indicate the number of conformations in each state (e.g., folded/unfolded). The probability of each state can be calculated by dividing the multiplicity of that state by the sum of the multiplicities of all states in the system (i.e., the total number of ways to arrange the marbles in the box – see Panel “How to calculate the multiplicity of a state”):

$$p_A = \frac{\Omega_A}{\sum_X \Omega_X} \quad (4)$$

where  $p_A$  equals the probability of state  $A$ ,  $\Omega_A$  is the multiplicity of state  $A$  and  $\sum_X \Omega_X$  is the sum of the multiplicities of all states in the system.

From this equation one can see that in a system lacking attracting or opposing forces the largest multiplicity also has the largest probability. A large multiplicity implies a lot of conformational freedom, i.e., high entropy, which is directly related to the multiplicity  $\Omega$  by:

$$S_A = k_B \ln \Omega_A \quad (5)$$

where  $S_A$  is the entropy of state  $A$ ,  $k_B$  is the Boltzmann constant (which relates temperature to energy per molecule (Fischer, 2019)) and  $\Omega_A$  is the multiplicity of state  $A$ . Thus, a state with a high multiplicity also has a high entropy, which makes that state more favorable. Note that in complex continuous systems like proteins in solution, the number of conformations is infinite and so we cannot count the absolute number of conformations. However, we can use the relation between multiplicity and probability to get an estimate of entropic differences between states from simulations (more on that later, in Section 5 “Free energy”).

## 4 Enthalpy

The enthalpy is the internal energy of the system plus the product of pressure and volume. The internal energy consists of kinetic energy (movement),

thermal energy (heat), potential and interaction energies, amongst others. The more favourable interactions a molecule has, the lower the enthalpy. An example of interaction energy is the Van der Waals interaction, which can be described by the Lennard-Jones interaction potential (more detail on interaction potentials, including the Lennard-Jones in Chapter 14). Other forms of interactions are polar interactions (e.g., hydrogen bonds), electrostatic interactions and hydrophobic interactions (e.g.,  $\pi$ -stacking).

Note, that a favourable interaction reduces the enthalpy. The interaction energy at infinite distance is typically defined as zero; this means that a favorable interaction will have a negative energy.

## 5 Free energy

So far, we have explained two thermodynamic concepts that affect the systems of interest: entropy and enthalpy. Now we will combine these two concepts to understand the stability of states in the system. As we have seen in Chapter 12, the more stable a state is, the lower the free energy associated with that state is. If the entropy and enthalpy of the state are known, we can directly calculate the free energy of that state using

$$F = H - TS \quad (6)$$

where  $H$  is the enthalpy,  $T$  is the temperature in Kelvin and  $S$  is the entropy. One can see from this formula that a macroscopic state with a high number of favorable interactions (low enthalpy) and/or low number of unfavorable interactions (which would increase the enthalpy), has a low free energy. Conversely, a state with fewer favorable interactions or more unfavourable interactions, will have a higher free energy.

Additionally, Equation 6 shows that increasing the entropy of the system also reduces the free energy. Usually there is a trade-off between reducing the enthalpy and increasing the entropy of a system. Having many favorable interactions (low enthalpy) reduces the conformational freedom of the protein, which will lead to a lower entropy, which is entropically unfavourable. On the other hand, the state with the maximum entropy usually consists of unfolded conformations, where there are more interactions with the solvent. But for a protein this also means more interactions between hydrophobic residues and the solvent, which is enthalpically unfavourable. An example of this is shown in Figure 13.3. The figure shows a model of a six-residue peptide on a square lattice, where the orange circles indicate hydrophobic residues, and blue circles hydrophilic residues. There is one conformation with the lowest enthalpy (the highest number of favourable internal interactions), this is the folded state at the bottom in Figure 13.3. The highest entropy state is the unfolded state with fourteen conformations, which all have no internal interactions, shown at the top of Figure 13.3. Depending

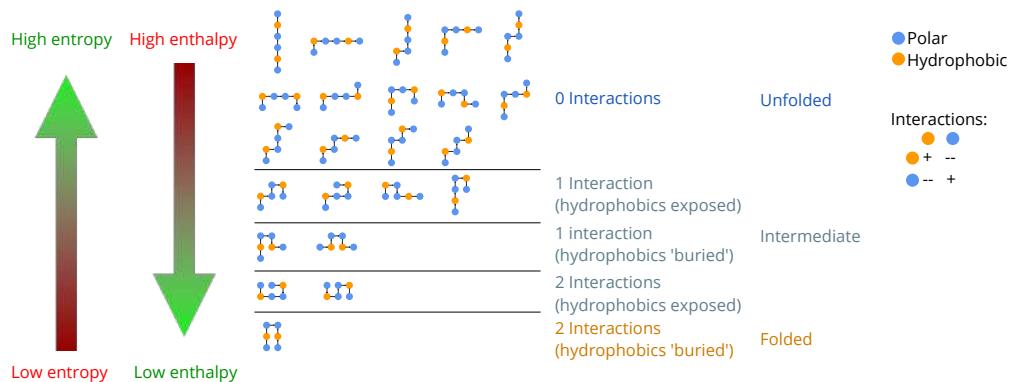


Figure 13.3: Conformational enthalpy and entropy in an hydrophobic-polar (HP) lattice model of protein folding. Here, atoms can only be at the intersections of a two-dimensional square lattice. The lowest energy is shown at the bottom, which corresponds to the ‘folded’ state. Energy here is counted as the number of interactions, where H-H and P-P are favorable and H-P is not. The top row has no interactions, the second row one P-P contact, the third row one H-H contact, and the bottom (native state) has one H-H plus one P-P contact. The number of ‘possible’ conformations per energy level (state) decreases as well, going from top (unfolded) to bottom (native/folded). Image adapted from Martin Gruebele, University of Illinois, USA (Ballew et al., 1996).

on the **conditions** of the system (e.g., **temperature, pH, other solvents**), the folded, unfolded or one of the intermediate states will be the **most stable**.

Now that we know the relation between free energy, entropy and enthalpy (Equation 6), the next question is then, how do we determine these properties from a system. As explained in Section 3, the entropy  $S_A$  of a state A is directly related to its **multiplicity** via Equation 5. The **total energy** of the state  $E_A$  can be calculated from the **weighted sum of the energies of the conformations in state A**:

$$E_A = \sum_{i \in A} E_i p_i \quad (7)$$

However, the calculation of the absolute values of the free energy contributions is far too complex for **continuous** systems such as proteins in solution, because the number of conformations to be considered is infinite. Nevertheless, the **probability of a macroscopic state**, such as the folded or unfolded state of a protein, can be **approximated** from simulations by determining the **fraction of time spent** in that state (see Panel “Derivation of free energy using statistical thermodynamics” to see how we get here using statistical thermodynamics):

$$F_A \propto -k_B T \ln(p_A) \quad (8)$$

where  $F_A$  is the **free energy of state A**,  $k_B$  is the **Boltzmann constant**,  $T$  is

the **temperature** in Kelvin and  $p_A$  is the **probability** of the system to be in state  $A$ . Note that  $F_A$  is **proportional**, not equal, to  $p_A$ . Thus, we cannot calculate the absolute value of the free energy of state  $A$  using this approach. In practice, this is not a problem, since we are usually only interested in the **free energy difference** between two states (e.g folded vs unfolded) to see which one is more favourable, rather than the absolute free energy of an individual state. This difference can be obtained using:

$$\begin{aligned}\Delta F_{A \rightarrow B} &= F_B - F_A \\ &= -k_B T \ln(p_B) - (-k_B T \ln(p_A))\end{aligned}\quad (9)$$

Thus, the **free energy difference** between two states equals:

$$\Delta F_{A \rightarrow B} = -k_B T \ln\left(\frac{p_B}{p_A}\right)\quad (10)$$

Equation 10 is very important, since it tells us that the only thing we need to know to calculate the free energy difference between two states, is the **relative amount of time** proteins spend in each state;  $p_A$  and  $p_B$ . As we will see in Chapter 14 and Chapter 15, these probabilities can easily be determined from **simulations** (provided all states are **sampled sufficiently**).

### Derivation of free energy using statistical thermodynamics

Combining Equation 5, Equation 6, and Equation 7, the formula for the free energy of state  $A$  ( $F_A$ ) becomes

$$F_A = E_A - TS_A = \sum_i E_i p_i - k_B T \ln(\Omega_A)\quad (11)$$

where  $E_i$  is the energy of a conformation in the ensemble of conformations belonging to state  $A$ ,  $p_i$  is the probability of that conformation,  $k_B$  is the Boltzmann constant,  $T$  is the temperature in Kelvin and  $\Omega$  is the multiplicity of the state. In a complex system such as a protein in solution, the multiplicity can usually not be calculated explicitly. However, in Panel “How to calculate the multiplicity of a state”, we explained that  $\Omega_A = N!/(k_A!(N - k_A))$ . Applying this to Equation 5 ( $S_A = k_B \ln \Omega_A$ ), using Stirling’s approximation, which says that the **logarithm of the factorial** of a very large number can be **approximated** as  $\ln(N!) \approx N \ln(N) - N$ , (Glazer *et al.*, 2002) and plugging the results into Equation 11, eventually yields:

$$F_A = \sum_i E_i p_i + k_B T \sum_i p_i \ln(p_i)\quad (12)$$

Equation 4 shows how to calculate the probability of a state if there are no interactions between the particles in the system (i.e., no enthalpy, so the system is **entropy-driven**), such as the example with the marbles in Section 3. In a system where the particles do interact, such as a protein in solution, the probability of a conformation also depends on the **energy** of that conformation. This is captured in the **Boltzmann distribution**, which describes the **probability  $p_i$  of a state  $i$**  as function of its energy  $E_i$  in an equilibrium situation:

$$p_i = \frac{e^{-E_i/k_B T}}{\sum_i e^{-E_i/k_B T}} \quad (13)$$

where  $p_i$  is the probability of a conformation,  $E_i$  is the energy of that conformation,  $k_B$  is the Boltzmann constant and  $T$  is the temperature (in Kelvin). The denominator is called the **Partition function**, denoted as  $Z$ :

$$Z = \sum_i e^{-\frac{E_i}{k_B T}} \quad (14)$$

which simplifies Equation 13 to:

$$p_i = \frac{1}{Z} e^{-\frac{E_i}{k_B T}} \quad (15)$$

Substituting Equation 15 back into Equation 12 and simplifying the result gives the **total free energy**  $F$  of the system:

$$F_A = -k_B T \ln(Z) \quad (16)$$

We can now directly relate the **free energy** to the **probability of being in a certain state**. Equation 14 defines the partition  $Z$  as the **sum of all Boltzmann factors of the conformations in the system**. For a system like a solution of protein molecules, we can think of this as describing the how likely it is to find one molecule in a particular microstate at any given moment in time, or in other words, how the system is divided, **partitioned**, across the **different microstates**. Then, we can reverse Equation 15 to obtain the probability  $p_i$  of being in that state  $i$ , then

$$F_i \propto -k_B T \ln(p_i) \quad (17)$$

This probability can, at least in principle, be obtained from **simulations or experiments**. We will show examples of this, from simulation, in the next section, “Temperature Dependence of Free Energy Landscapes”, and more in Chapter 14 “Molecular Dynamics” and Chapter 15 “Monte Carlo for Protein Structures”.

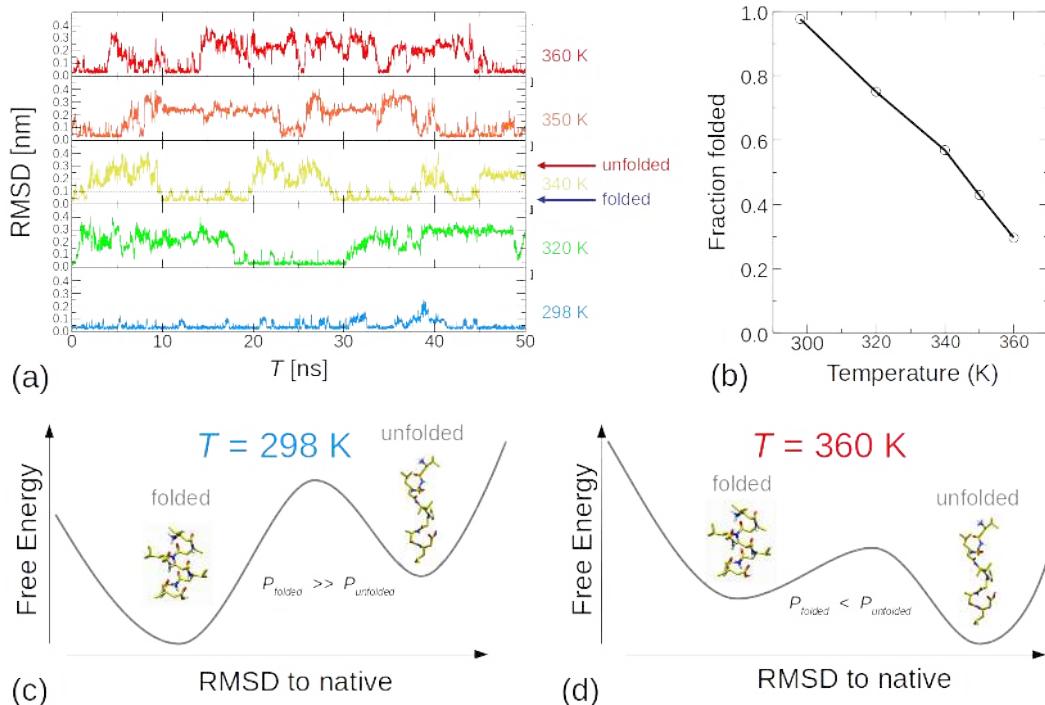


Figure 13.4: Temperature dependence of protein stability. (a) RMSD vs. time for 5 different temperatures: 298, 320, 340, 350 and 360 K. (b) Fraction folded as function of temperature, derived from the simulations shown in (a). At room temperature the protein is almost always in its folded state. As the temperature increases the protein is more in its unfolded state. (c) Schematic Free Energy diagrams corresponding to the lowest temperature (298K). The reaction coordinate used is the same as for (a): the RMSD to native. The folded state (left) has a lower RMSD, while for the unfolded state (right) it is high. The free energy of the folded state is lower, which indicates it is more stable than the unfolded state at this temperature. The barrier between folded and unfolded states limits the rate at which folding and unfolding events may happen. (d) Same, but for the highest temperature (360K). Now, the free energy of the unfolded state is lower, which indicates it is more stable than the folded state at 360K. The barrier between folded and unfolded states is somewhat lower, reflecting the higher rate at which folding and unfolding can be observed in panel (a). Panel (a), data for panel (b) and structures in (c) and (d) with permission from Daura & Oostenbrink (Daura et al., 1998).

## 5.1 Temperature Dependence of Free Energy Landscapes

One of the reasons we are interested in the thermodynamics of protein folding is so we can understand the **temperature dependence** of the stability. Proteins can fold reversibly, as we know from Anfinsen (1973). Figure 13.4a shows simulations of a small (7 amino acids) peptide to illustrate this phenomenon. Each of the colors in the graph represents a time trace of the **RMSD** to the native structure over time for simulations at a different tem-

perature. One can see in the figure that throughout the simulations, the RMSD increases and decreases, indicating that the protein **unfolds and refolds** several times. It is clear that as the **temperature increases** the protein spends more time in the **unfolded state** (high RMSD in Figure 13.4a) which can be quantified by the **probability** (time spent) of the folded state as a function of temperature (Figure 13.4b).

Let us see if we can understand this from a thermodynamic perspective. Recall, from Equation 6, that  $\Delta F = \Delta H - T\Delta S$ . In this example, we are talking about the free energy of folding, so the  $\Delta$  here indicates the difference between the folded state and the unfolded state. A negative value of  $\Delta F$  indicates the unfolded state is more favourable, whereas a positive value for  $\Delta F$  indicates the folded state is more favourable. For the first term of the equation, the unfolded state has fewer favorable internal contacts than the folded state, so the unfolded state has a higher enthalpy:  $\Delta H = H_{unfolded} - H_{folded} > 0$ . For the second term of the equation, the unfolded state has more conformational freedom and thus a higher entropy than the folded state:  $\Delta S = S_{unfolded} - S_{folded} > 0$ . Since temperature in Kelvin ( $T$ ) is always positive, in the formula  $\Delta F = \Delta H - T\Delta S$ , both  $\Delta H$  and  $T\Delta S$  are **positive**.

At **low  $T$** , the absolute value of  $T\Delta S$  will be **small**. Therefore, the favorable enthalpy of the folded protein structure to  $\Delta F$  will outweigh the unfavourable entropy of the folded state at low temperatures, and the protein will spend most time in the **folded** conformation, as shown in Figure 13.4c.

As  $T$  increases, the absolute value of  $T\Delta S$  will increase as well. At a sufficiently high value of  $T$ , the favorable enthalpy of the folded protein structure to  $\Delta F$  will no longer outweigh the unfavourable entropy of the folded state, and at higher temperatures, the protein will spend most time in the **unfolded** conformation, as shown in Figure 13.4d. This decrease in stability of the folded state at higher temperatures is exactly what is observed in Figure 13.4a and b: as temperature increases, the protein spends relatively more time in the unfolded state because the unfolded state becomes more stable than the folded state. Thus, the importance of entropy **increases** with increasing temperature.

Figure 13.5 shows another example of temperature dependence of protein folding, but here a 3D lattice model is used. Despite its simplicity, we can observe the transition from a stable folded state at low temperature, through a transition temperature at which both states are stable to some extent, to high temperatures where the unfolded state is the most stable (van Dijk *et al.*, 2015). Although the model is very simple, it is realistic in the sense that such transition temperature effects are also observed for real proteins (Thiriot *et al.*, 2005).

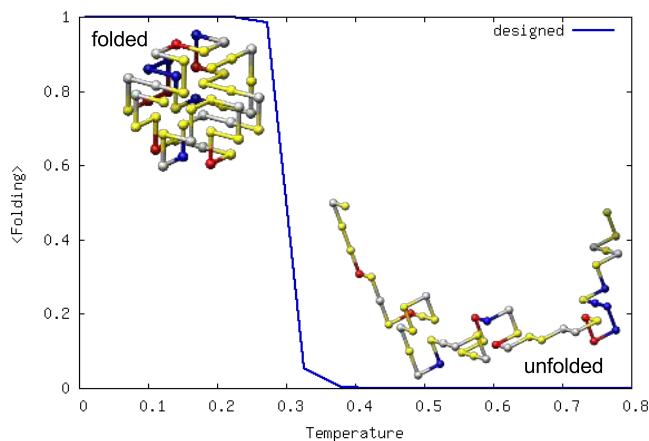


Figure 13.5: Temperature dependence of protein folding. The horizontal axis shows temperature (the results are from a simplified lattice model of protein folding with reduced units for temperature). The vertical axis shows the extent of folding, 1 meaning fully folded and 0 unfolded (measured by the fraction of native contacts formed). At low temperature ( $T < 0.2$ ), the native state is stable and therefore the protein is folded. This folding is driven by the energetically favourable conformation where hydrophobic residues (yellow) are ‘shielded’ in the interior of the protein structure. At high temperature ( $T > 0.4$ ), entropic effects win out over the energetic effects. This makes the unfolded state more stable. The unfolded state, naturally, has a higher entropy than the folded state, but has to pay the energetic cost of exposing hydrophobic residues to the water (van Dijk et al., 2015, 2016).

## 6 From Microstates to Macrostates

So far, we have discussed the formal definitions of the free energy, enthalpy and energy of a system and how they are related. An important distinction that has been mentioned, but not yet emphasized, is the distinction between ***microstates*** and ***macrostates***. Microstates are individual conformations (in the previous sections of this chapter indicated with the subscript  $i$ ), whereas macrostates describe the behavior of **ensembles of conformations** with very similar properties (indicated with the subscripts  $A$  and  $B$ ). For example, the folded and unfolded states with local energy minima in Figure 13.4c,d are macrostates. The probability of a conformation calculated using the Boltzmann equation is an example of the probability of a microstate. Another example can be found in Figure 12.1a, where each individual conformation shown of the structure is a microstate, and the collection of conformations is a macrostate. In the next two subsections we will explain two concepts that are important to understand how to relate microstates to macrostates, namely ***order parameters*** and ***ensemble averages***.

## 6.1 Order Parameters

In a free energy profile (such as shown in Figure 13.4c,d) the free energy is determined for different values of an **order parameter**. An order parameter is a quantitative measure that can distinguish the relevant states of a system (also see panel “Evaluating your MD simulations – Order Parameters”). For protein folding, frequently used order parameters are for example the **number of native contacts** (e.g., the number of internal hydrogen bonds), the **RMSD to the native structure** or the **radius of gyration** (a measure of the diameter of a flexible molecule, which indicates how compact or extended a protein conformation is). Order parameters are used to define the **macrostates** in the system. For example, we can define an **RMSD threshold** to determine whether a protein is folded or unfolded.

Choosing a suitable order parameter is very important when setting up a research project involving simulations. Not every order parameter is able to help answer every research question. For example, say that we are interested in the folded and unfolded state of a protein with a large disordered loop region (e.g., a protein with two domains separated by a linker or some transmembrane proteins). During simulation, this region would be relatively flexible and move around a lot. As a result, the RSMD of this structure would be relatively high in both the folded and the unfolded state. Thus, the RMSD might not be able to distinguish the states of interest that well, and would not be a **suitable** order parameter. The number of native contacts on the other hand, would be less sensitive to the dynamics of the loop region(s), and might be more suitable in this case.

For the simple two-state examples of protein folding considered so far, a single order parameter suffices, but in many cases **two or several more order parameters** are used to establish a multi-dimensional free energy landscape. We will discuss an example of this in Chapter 15 “Monte Carlo for Protein Structures”, where we use native vs. non-native contacts to describe an two-dimensional free energy landscape (see Section 6.1 “A simple protein lattice model”).

## 6.2 Ensemble Average

Apart from free energies, we may also be interested in other properties that are **characteristic of the system** (e.g., enthalpy, radius of gyration, secondary structure content); Especially properties that can be measured experimentally, because those show that the simulations can reproduce behavior observed *in vitro* or *in vivo*, while providing mechanistic insights into the observed processes.

Because the values of the property of interest may not **equal for all conformations within a state**, and the **probabilities of observing each conformation** within a state may not be equal, we calculate a weighted **average**

to describe that property for a state. This weighted average is called the **ensemble average**. The ensemble average for a **property**  $X$  in **state**  $A$  is calculated by:

$$\langle X \rangle_A = \frac{\sum_{i \in A} X_i p_i}{\sum_{i \in A} p_i} \quad (18)$$

where  $X_i$  is the value of the property of interest  $X$  in conformation  $i$  and  $p_i$  is the probability of conformation  $i$ .  $i \in A$  is a notation indicating all conformations  $i$  in state  $A$ . Note that  $i$  here denotes microstates, and  $A$  the (observable) macrostate, as introduced at the start of this section. From our simulation, we can directly calculate an ensemble average, by averaging a certain property  $X$  over all the sampled conformations.

One example of the application of ensemble averages is the calculation of the **enthalpy of a state**. The enthalpy of a state is the ensemble average of the internal energies of all the conformations in the ensemble of that state:

$$H_A = \langle E \rangle_A = \frac{\sum_{i \in A} E_i p_i}{\sum_{i \in A} p_i} \quad (19)$$

where  $H$  is the enthalpy and  $E$  is the internal energy (more on the relation between the two in the next section). To calculate the **difference in enthalpy** we can subtract the ensemble averages of the energies of the two states:

$$\Delta H_{A,B} = \langle E \rangle_B - \langle E \rangle_A \quad (20)$$

Thus, ensemble averages can help us quantify structural properties related to macrostates.

## 7 Ensembles

In molecular simulations, we generally consider the molecules of interest in an isolated environment (e.g., protein in a box of water with constant temperature and volume). We will not go into detail here, but generally this

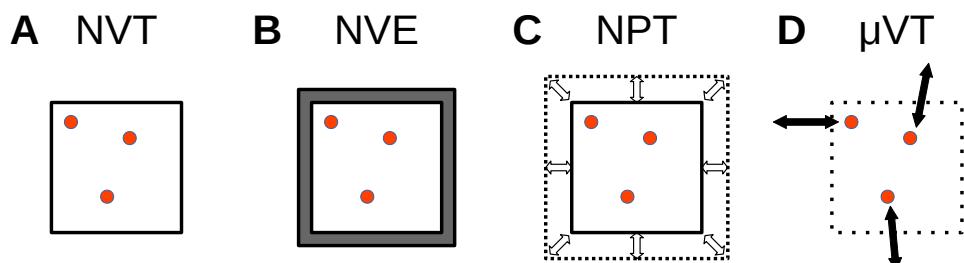


Figure 13.6: A schematic representation of different ensembles. A: NVT ensemble, B: NPT ensemble, C:  $\mu$ VT ensemble, D: NVE ensemble.

helps significantly simplify the calculations needed during the simulations. It is important to realise that thermodynamic relations are slightly different depending on the physical conditions chosen for the system. This gives rise to what is known in thermodynamics as different ensembles (not to be confused with ensemble averages), each with their own specific definition of free energy. In this section, we will (briefly) discuss several of these ensembles, as shown schematically in Figure 13.6. Each of them are defined by the three parameters that are constant under the specific conditions:

**NVT**: Constant number of particles ( $N$ ), volume ( $V$ ) and temperature ( $T$ ), typically encountered in Monte Carlo simulations (Chapter 15). Known as the “canonical” ensemble.

**NVE**: The “microcanonical ensemble” is the natural situation in molecular dynamics simulations (Chapter 14), where we typically define a fixed system ( $N$ ), in a fixed size environment ( $V$ ) and no exchange of energy allowed ( $E$ ).

**NPT**: Instead of volume, here the pressure ( $P$ ) is constant, as well as the number of particles ( $N$ ) and temperature ( $T$ ). This is closer to laboratory conditions, and still convenient for simulations. Therefore, in MD simulations, usually a thermostat and barostat are applied to the system such that the simulations are performed in NPT rather than NVE (Chapter 14).

**$\mu$ VT**: Instead of number of particles, here the chemical potential ( $\mu$ ) is constant, which presumes exchange of particles with the surroundings. Additionally, the volume ( $V$ ) and temperature ( $T$ ) are constant. This so-called “grand canonical” ensemble is close to typical laboratory conditions, but hard to achieve in simulations.

Note, that we do not go into why there are three parameters, and why they occur in these combinations (although there are more possible). For this, please refer to Schroeder (1999).

We will consider two ensembles, the NVT and NPT, in some more detail, as these are common in simulations. In the NVT ensemble, the number of particles  $N$ , the volume  $V$  and the temperature  $T$  are kept constant. This corresponds to the so called Helmholtz ensemble or canonical ensemble, and the Helmholtz free energy; which is what we defined in Equation 6, and is often written as:

$$\Delta F = \Delta E - T\Delta S \quad (21)$$

Here,  $F$  is the Helmholtz free energy.  $E$  is the internal energy in the system. We typically cannot calculate or measure absolute values for energies, but only differences between states. This is not a problem since these relative free energies between states determines their stability, and therefore, the difference is the variable that needs to be determined. In protein folding for instance, these can be between the folded and unfolded states (e.g., Figure 12.3), or for the case of protein interactions, between the

bound and unbound states.

In the NPT ensemble, the number of molecules  $N$ , pressure  $P$  and temperature  $T$  are constant, and the Gibbs free energy  $G$  is minimized. The Gibbs free energy is defined as

$$\Delta G = \Delta E - T\Delta S + PV = \Delta H - T\Delta S \quad (22)$$

The Gibbs free energy is the Helmholtz free energy plus the product of the pressure and volume. The term  $E + PV$  is called the enthalpy and is typically denoted as  $H$ . Due to the low compressibility of water, the  $PV$  term can often be neglected in the systems we are interested in. Therefore, we sometimes consider the free energy without specifying whether we use the Helmholtz or the Gibbs ensemble.

## 8 Conclusion

In this chapter we have elaborated on the concepts introduced in Chapter 12 in a more formal way using principles from thermodynamics and statistical mechanics. We have mainly focused on how the free energy is related to the entropy and enthalpy, and how this affects the probability that a system is in a certain state. We have also illustrated how changes in the temperature affect the free energy landscape. Finally, we highlighted some key differences between microstates (individual conformations) and macrostates (ensembles of conformations with similar properties). In the next chapters, we will go into more depth on how simulation methods can be applied to study free energy landscapes.

## Key concepts

- Free energy  $F$  is a function of the internal energy  $E$ , entropy  $S$  and temperature  $T$ :  $F = E - TS$
- Boltzmann's relation gives us the probability  $p_i$  of conformation  $i$  as function of its energy  $E_i$ :  $p_i = \frac{1}{Z} e^{-E_i/k_B T}$
- The free energy of a state is proportional to the probability of encountering that state:  $F_A \propto -k_B T \ln(p_A)$
- The free energy difference between two states can be calculated by:  $\Delta F_{A \rightarrow B} = -k_B T \ln(p_B/p_A)$
- Ensemble averages can be used to describe the general behaviour of the microstates belonging to a macrostate
- Different thermodynamic ensembles are used to describe the physical properties of the system of interest
- An order parameter is needed to distinguish different states of the system; only then does it become possible to calculate free energies

## Further Reading

- “Physical Biology of the Cell” – Phillips *et al.* (2012)
- “Statistical Mechanics - A survival guide” – Glazer *et al.* (2002)
- “The Real Reason Why Oil and Water Don’t Mix” – Silverstein (1998)
- “An Introduction to Thermal Physics” – Schroeder (1999)

## Author contributions

Wrote the text:	JvG, EvD, HM, KAF, SA
Created figures:	JvG, AG, KAF, SA
Review of current literature:	JvG, JV, IH, KAF, SA
Critical proofreading:	JvG, HM, KAF, JV, SA
Non-expert feedback:	JB, JG
Editorial responsibility:	JvG, KAF, SA

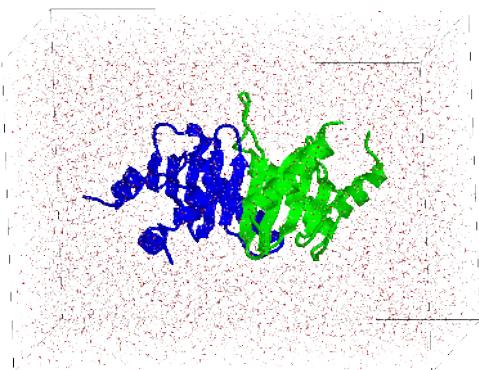
## References

- Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223–230.
- Ballew, R.M., Sabelko, J. and Gruebele, M. (1996). Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(12), 5759–64.
- Daura, X., Jaun, B., Seebach, D., van Gunsteren, W.F. and Mark, A.E. (1998). Reversible Peptide Folding in Solution by Molecular Dynamics Simulation. *J. Mol. Biol.*, **280**(5), 925–932.
- Fischer, J. (2019). The Boltzmann Constant for the Definition and Realization of the Kelvin. *Annalen der Physik*, **531**(5), 1800304.
- Frasch, W.D., Bukhari, Z.A. and Yanagisawa, S. (2022). F1FO ATP synthase molecular motor mechanisms. *Frontiers in Microbiology*, **13**.
- Glazer, M., Wark, J. and Schmittmann, B. (2002). Statistical Mechanics: A Survival Guide. *American Journal of Physics*, **70**(12), 1274–1275.
- Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**(4), 378–379.
- Kittel, C. and Kroemer, H. (1980). Thermal physics. *WIT Freeman: San Francisco*.
- Mitchell, P. (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature*, **191**(4784), 144–148.
- Mondal, P., Khamo, J.S., Krishnamurthy, V.V., Cai, Q. et al (2017). Drive the Car(go)s-New Modalities to Control Cargo Trafficking in Live Cells.
- Phillips, R., Kondev, J., Theriot, J., Garcia, H.G. and Orme, N. (2012). *Physical biology of the cell*.
- Schroeder, D.V. (1999). *An Introduction to Thermal Physics*. Addison-Wesley Publishing Company, San Francisco, CA.
- Silverstein, T.P. (1998). The Real Reason Why Oil and Water Don’t Mix. *Journal of Chemical Education*, **75**(1), 116.
- Thiriot, D.S., Nezvorov, A.A. and Opella, S.J. (2005). Structural basis of the temperature transition of Pfl bacteriophage. *Protein Science*, **14**(4), 1064–1070.
- van Dijk, E., Hoogeveen, A. and Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, **11**(5), e1004277.
- van Dijk, E., Varilly, P., Knowles, T.P.J., Frenkel, D. and Abeln, S. (2016). Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, **116**(7), 078101.

# Chapter 14

## Molecular Dynamics

Halima Mouhib\*  Juami H. M. van Gils   
Jose Gavaldá-García  Qingzhen Hou  Ali May   
Arriën Symon Rauh  Jocelyne Vreede   
Sanne Abeln\*  K. Anton Feenstra\* 



\* editorial responsibility



## 1 Introduction to molecular dynamics

We know that many proteins have functional motions, and we already introduced this in Chapter 2, Section 6. One famous example presented there in Panel “Allosteric motions and time-resolved crystallography”, is the cooperative binding of oxygen to hemoglobin, where the first oxygen binding induces a conformational change throughout the tetrameric structure, which make subsequent oxygen binding much more favourable. However, experimentally, such motions are hard to observe. Exceptions are when a protein can be crystallized in multiple distinct conformations, when fluorescence experiments are able to capture fast events, or in the very rare cases where time-resolved crystal structures can be recorded. A dramatic manifestation of functional motion occurs for some enzymes that are still active in the crystal form (which is relatively common). Large scale motions of the protein structure may occur during catalysis of the reaction, and sometimes these motions physically break the crystal when the substrate is added to the crystal! (This shows how strong molecular motions can be.) As an alternative to experimentally studying protein motions, molecular dynamics (MD) simulations have been used since the 1960s. These allow us to simulate all atomic motions in detail, but of course within the restrictions of the accuracy of the molecular models used.

In this chapter we will introduce MD simulations, which represent an important method to investigate the dynamic behaviour of proteins and polymers. Basically, in such simulations, the forces and interactions between particles are used to numerically derive the resulting three-dimensional movement of these particles over a certain time-scale. The main emphasis will lie on explaining the basic method of MD simulations, when and how to use them, and provide some examples of how to evaluate the results. More specifically, we will describe the physical interactions between atoms and the algorithms used to perform these simulations. Secondly, we will see how simulation results may be interpreted. For this purpose, some basic knowledge on thermodynamics and statistical mechanics, as introduced in Chapter 13 “Thermodynamics of Protein Folding” is essential. It is particularly important to understand the relation between free energy and probability, in order to analyse simulation results.

Although a wide variation of problems (not necessarily of biological relevance) can be addressed using MD simulations, in the course of this book we will focus mainly on proteins.

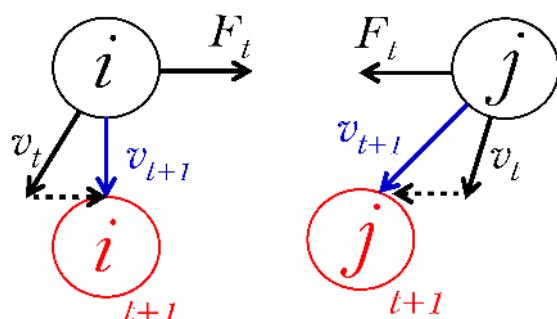
### 1.1 Simulating a protein by classical physics

Molecular dynamics (MD) simulations, using classical or Newtonian physics, is currently the most common tool to investigate the dynamics and dynamic ensembles of proteins and other large biomolecules at a molecular level (e.g.,

Adcock and McCammon, 2006). There are several branches of the technique that can be used to elucidate complicated or ambiguous results from biophysical experiments such as **Atomic Force Microscopy** (AFM) or **Nuclear Magnetic Resonance (NMR)** (Kumar and Li, 2010). Most importantly, MD can provide insights into the molecular mechanisms through which proteins **perform their functions**; this can range from **binding to a specific substrate**, to **proteins that can bind to each other**, to the complete **cycle of movement** of molecular motors (e.g., Karplus and Kuriyan, 2005).

Even though MD is used ubiquitously, it is important to remember that simulation methods and descriptions of interactions used are necessarily an **abstraction** of the real world (see in particular also the Panel “Limitations of Newtonian physics and force fields”). Therefore, any hypothesis on mechanistic workings obtained from MD simulations should be **verified** with experimental evidence. This way, the biological, physical and scientific relevance of the results can be guaranteed. Fortunately, Newtonian physics suffices for the vast majority of biologically relevant phenomena, because we are typically interested in macroscopic properties which depend on **averages obtained from the ensembles** observed in the simulation. Please refer to Chapter 13 for more detail on the statistical dynamics of ensembles.

In Newtonian physics, interactions are described as forces, and a force applied to an object triggers an effect. Figure 14.1 shows how the interaction (by way of force  $F$ ) between two particles leads to changes in positions of both. This way, the structural and dynamical properties of large biological systems can be investigated at a molecular level. However, to simulate the large collection of atoms that make up biological systems, we need to resort to **numerical approximations**, as analytical solutions are not available (this is known as the ‘**multi-body**’ problem in physics). Moreover, to obtain an



*Figure 14.1: Two particles  $i$  and  $j$  at time  $t$  with initial position (indicated by black spheres) and velocities  $v_t$  (black arrows), and exerting a force  $F_t$  on each other (also black arrows; note that between two particles  $F_{t;i,j} = -F_{t;j,i}$ ). These forces cause the velocities to change at the next time step  $t + 1$ , as indicated by  $v_{t+1}$  in blue arrows, and the new velocities cause the positions to change as well (red spheres).*

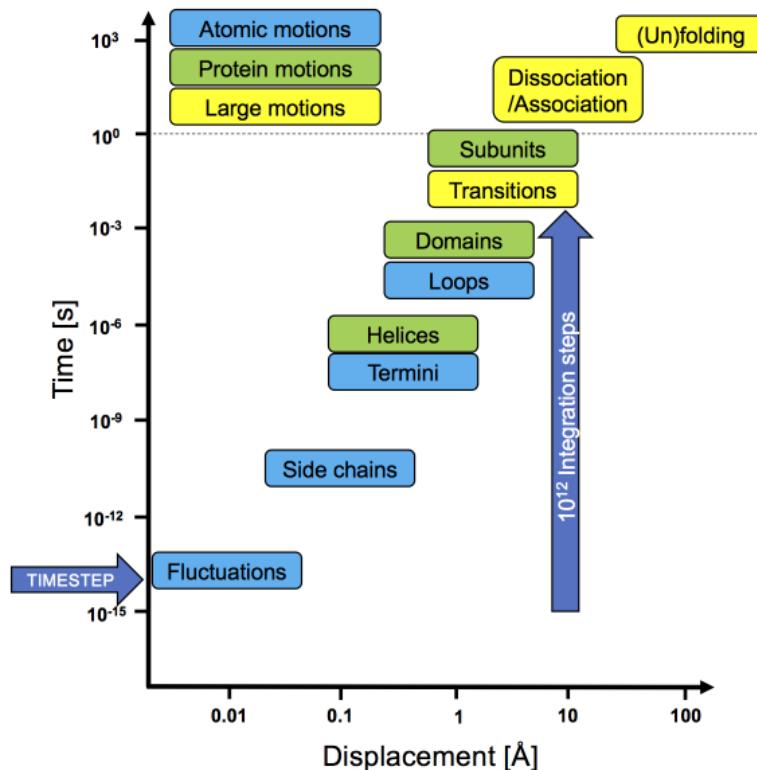


Figure 14.2: A simplified overview of the order-of-magnitude time and length scales of molecular motions that are of biological interest; the boxes indicate atomic motions (blue), protein domain motions (green), and large biological motions (yellow). The x axis corresponds to displacement of atoms or molecules, and the y axis indicates the approximate time scales. Note that both axes are on a logarithmic scale. The size of the MD timestep (2 fs) is indicated by the arrow on the left; the arrow on the right indicates that a thousand billion ( $10^{12}$ ) timesteps are needed to reach the shortest biologically relevant timescales.

accurate approximation, we need very small integration (time) steps, which is the main limit on computational efficiency. This is what much of the current chapter will deal with.

## 2 Relevant time and length scales

To give some perspective to the phenomena we will be describing in molecular simulations, it is good to realize which time and length scales are involved. Figure 14.2 shows an overview of the relevant time and length scales of several classes of biological processes. Please note that these are order-of-magnitude indications that should not be taken as absolute values.

Summarizing:

- At the lower end of the scale are small-scale fluctuations of individual atoms over fractions of an Ångstrom ( $\text{\AA}$ ,  $10^{-10} \text{ m}$ ) distance and at timescales of tens to hundreds of femtoseconds ( $10^{-15} \text{ s}$ ).
- Amino acid side chains that involve several to tens of atoms move tenths of an Ångstrom per picoseconds ( $10^{-12} \text{ s}$ ).
- Rotamer changes, conformational changes arising through rotation around rotatable bonds and sidechains, may produce larger motions, i.e., in the order of Ångstroms, and will be slower, tens to hundreds of picoseconds.
- The termini and loop regions, which involve several residues and thus many more atoms than a single sidechain are still slower, moving up to several Angstroms anywhere in the range of nano- ( $10^{-9}$ ) and microseconds ( $10^{-6}$ ).
- As helices are relatively stable structural (and dynamical) units, they move around the same scales as termini regions and loops.
- Larger parts, such as domains or subunits are again larger and therefore move slower, around tens of Ångstroms or even more and may take up to seconds in the slowest of cases.
- Conformational transitions take place in the same time-range as domain motives. These transitions often involve the motion of domain-size parts of a protein, or for example the association or dissociation of protein binding (e.g. in protein-protein interactions).
- The typical folding time for a protein is on the order of seconds, nevertheless ‘fast folders’ may take only milliseconds.

For accurate simulation, the timestep used should be smaller than the fastest motions. Fastest vibrations involve hydrogen atoms (which are light), and/or particularly stiff angles, with vibrational periods of around 10 femtoseconds. Therefore, typically timesteps of 1 or 2 fs ( $10^{-15} \text{ s}$ ) are used, as indicated with an arrow at the bottom left in Figure 14.2. It should be noted that an order of  $10^{12}$  integration steps are then required to start getting into biologically relevant (millisecond) timescales as, indicated with a vertical arrow at the right in Figure 14.2. This computational power is slowly coming within range of being feasible. But, in practice this means we are not routinely able to fold or unfold proteins in MD simulations, simply because we can not simulate for long enough. Note, also, that just simulating for longer may also not simply solve the problem, as the results would still depend on the accuracy of the force fields we use. We will deal with force fields in section 3.1. Moreover some effects, such as quantum effects, are hard to model accurately. In a large-scale survey of the ability of current force fields to improve upon homology models, (Lindorff-Larsen *et al.*, 2012; Raval *et al.*, 2012) observed the accumulation of errors in long simulations, highlighting that this is an as yet unsolved problem.

Nevertheless, numerous interesting problems of biological relevance such

as Amyloid aggregation in Alzheimer's (Lemkul and Bevan, 2013) and T-cell receptor-MHC interactions (Cuendet *et al.*, 2011) have been investigated using simulations that might strictly be considered too short. These simulations have fortunately shown the interesting and relevant dynamical molecular behaviour that helps answering several important biological questions.

### Historical background

A historical overview on the evolution and applications of MD simulations is given in (van Gunsteren *et al.*, 2006). The rough estimates of future simulation times are extrapolated from past achievements based on the size of those systems and assuming the continued growth of computer power by Moore's law by doubling every 18 months (Moore, 1965). Assuming that will hold, it may take till the end of this century before we can simulate protein folding at the speed it occurs in nature – but even then only a single protein, and only if the accuracy of our force fields do not remain limiting.

year	system	time-scale	reference
1936	Gelatine balls		Morrell and Hildebrand
1953	MC Simulations		Metropolis <i>et al.</i>
1957	MC of Lennard-Jones spheres		Wood and Parker
1964	MD of liquid Argon	10 ps	Rahman
1970's	Non-equilibrium methods		
1970's	Stochastic dynamics methods		
1974	MD of liquid water		Stillinger and Rahman
1977	MD of protein in vacuum	20 ps	McCammon <i>et al.</i>
1980's	Quantum-mechanical effects		
1983	MD of protein in water	20 ps	van Gunsteren <i>et al.</i>
1998	MD of reversible peptide folding	100 ns	Daura <i>et al.</i>
1998	MD of protein folding	1 $\mu$ s	Duan and Kollman
2010	MD of reversible small protein folding	1 ms	Shaw <i>et al.</i>
Today	Large proteins or complexes in water or membrane	up to milliseconds ( $\sim 10^{12}$ – $10^{14}$ slower than nature)	
2029	Protein folding	1 ms	???
2034	E-coli, $\sim 10^{11}$ atoms	1 ns	
2056	Eukaryotic cell, $\sim 10^{15}$ atoms	1 ns	
2080	Protein folding	as fast as in nature	

The field of molecular simulations was initiated around 1930, when Morrell and Hildebrand (Morrell and Hildebrand, 1936) investigated the distribution of gelatine spheres, compared to X-ray experiments on atoms. The subsequent rapidly growing number of applications was enabled through the continuous development and improvement of computational facilities over the last 50 years. This allowed for more and more detailed theoretical investigations of larger systems over longer timescales. Along with this development, simulations have moved from hard-sphere systems (Alder and Wainwright, 1957) and simple mono-atomic systems (e.g. argon) in the 60's (Rahman, 1964), to proteins and water separately in the 70's (Stillinger and Rahman, 1974; McCammon *et al.*, 1976), to proteins in water in the 80's (van Gunsteren *et al.*, 1983), and finally, to something close to the equilibrium behaviour of peptides and (small) proteins in water in the 90's and early 2000's (Daura *et al.*, 1998; Duan and Kollman, 1998; Shaw *et al.*, 2010). Generally it can be said that MD simulation became important for biophysics in the late 70's and early 80's with the first simulations of a protein in water. More biological relevance perhaps can be said to start with the simulation of reversible, equilibrium peptide folding (late 90's) and protein folding (early 2000's).

### 3 Forces & interactions

If we want to understand how particles move, we need to understand how they interact with each other, or more precisely, with what forces they repel or attract each other. The basis of all molecular simulations is the so-called *force field*. Force fields consist of a collection of all forces which are considered to occur in a system of interest, for example all atoms within a protein and the surrounding solvent. Forces, as noted before, correspond to interactions, in this case between individual atoms in the system. Note that these interactions and therefore the energies associated with them are dependent on the types of atoms involved, as well as their positions with respect to each other. Formally, the force  $\mathbf{F}$  on particle  $i$  is the derivative of the energy  $U$ , depending on the position  $r_i$  of particle  $i$ , and is given as:

$$\mathbf{F}_i = -\frac{\partial U}{\partial r_i}, \quad (1)$$

#### 3.1 Force fields

Force fields can be written as interaction energies; in the next section we will see how interaction forces and interaction energies are related. In molecular mechanics, interaction energies between individual atoms in the

system contain the following terms for interaction energies:

$$U_{total} = U_{bonded} + U_{non-bonded} + U_{crossterm} \quad (2)$$

Thus, in molecular systems, a distinction is made between bonded interactions, non-bonded interactions and ‘other’ interactions called ‘crossterm’ in the formula. Bonded interactions act in between atoms close by in the same molecules; most notably two atoms in a bond, three in an angle and four in a dihedral:

$$U_{bonded} = U_{bond} + U_{angle} + U_{dihedral} \quad (3)$$

Non-bonded interactions act in between molecules and also in between atoms in the same molecule: these are Coulomb’s (electrostatic) and Van der Waals’ (atomic contact) interactions.

$$U_{non-bonded} = U_{Coulomb} + U_{VanderWaals} \quad (4)$$

Combining Equations 2-4, the total energy may thus be specified as:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{Coulomb} + U_{VanderWaals} + U_{crossterm}. \quad (5)$$

Each of these energy terms will be specified in more detail in the next section. The main assumption here is that the total interaction (e.g. between molecules) can be described as a sum of pairwise atomic interactions. For some specific interactions, e.g., weak non-covalent interactions between two different bonds, a pairwise description is not a good approximation; these are represented as the ‘crossterm’ interactions in the formula. For protein simulations these terms are, usually, negligible and therefore not included in the most commonly used force fields to save computational resources.

### 3.2 Interactions

The most commonly used energy functions for the bonded interactions are bond, angle, and dihedral; for the non-bonded these are Coulomb and Van der Waals. Bond and angle interactions are shown in Figure 14.3a and b. Proper dihedrals correspond to rotatable bonds, for the proteins backbone these are the N-C<sub>α</sub> and C<sub>α</sub>-C=O bonds, as well as the non-aromatic ones in the side chains, Figure 14.3c. Improper dihedrals are used to constrain a fixed geometry, as for example the peptide bond or the tetrahedral arrangements of the hydrogens in an -NH<sub>3</sub> group, see Figure 14.3d. For bonds, angles and improper dihedrals, typically quadratic (harmonic) functions are used:

$$\text{Bonds: } U(r) = \frac{1}{2}k_b(r - r_0)^2 \quad (6)$$

$$\text{Angles: } U(\theta) = \frac{1}{2}k_\theta(\theta - \theta_0)^2 \quad (7)$$

$$\text{Improper dihedrals: } U(\chi) = \frac{1}{2}k_\chi(\chi - \chi_0)^2 \quad (8)$$

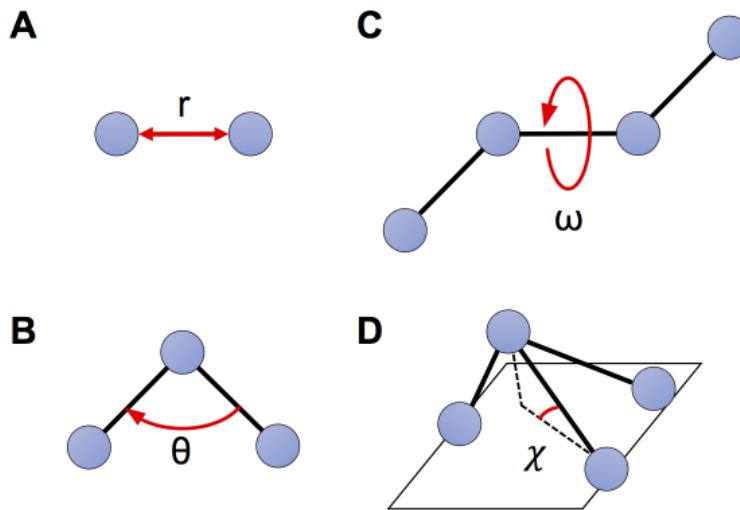


Figure 14.3: Schematic of common bonded interactions: **A** bond between two atoms, **B** angle between three atoms, **C** dihedral angle between four atoms (rotatable bond between the middle two), **D** improper dihedral, used to fix particular orientation, e.g. in-plane, or as in the drawing one atom out of the plane of three other atoms.

For these interactions, **r** denotes distance, and  **$\theta$**  (theta), and  **$\chi$**  (chi) denote angles. The subscript 0 refers to the equilibrium value in each case. The  **$k$ 's** are force constants. These parameters depend on the atoms involved.

(Proper) dihedrals are expressed as (periodic) cosines:

$$\text{Dihedrals: } U(\omega) = \frac{1}{2} \sum k_j [1 + (-1)^{j+1} \cos(j\omega + \phi)] \quad (9)$$

Here,  $\omega$  (omega) denotes the angle. Note that angle (Equation 7) and dihedral (Equations 8 and 9) bonded interactions are not strictly speaking pairwise as they involve three and four atoms, respectively. They are the main non-pairwise contribution that is usually included for protein simulations.

Non-bonded interactions are modelled according to Coulomb's law for electrostatic forces, and the Lennard-Jones potential describing Van der Waals forces, respectively as:

$$\text{Coulomb: } U(r_{ij}) = \frac{\epsilon_0(q_i \cdot q_j)}{r_{ij}} \quad (10)$$

$$\text{Lennard-Jones: } U(r_{ij}) = 4\epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right] \quad (11)$$

For the non-bonded interactions,  $r_{i,j}$  denotes the distance between atoms  $i$  and  $j$ , and  $q_i$  and  $q_j$  the charges on atoms  $i$  and  $j$ . In Equation 10  $\epsilon_0$  (epsilon)

Harmonic vs.  
periodic

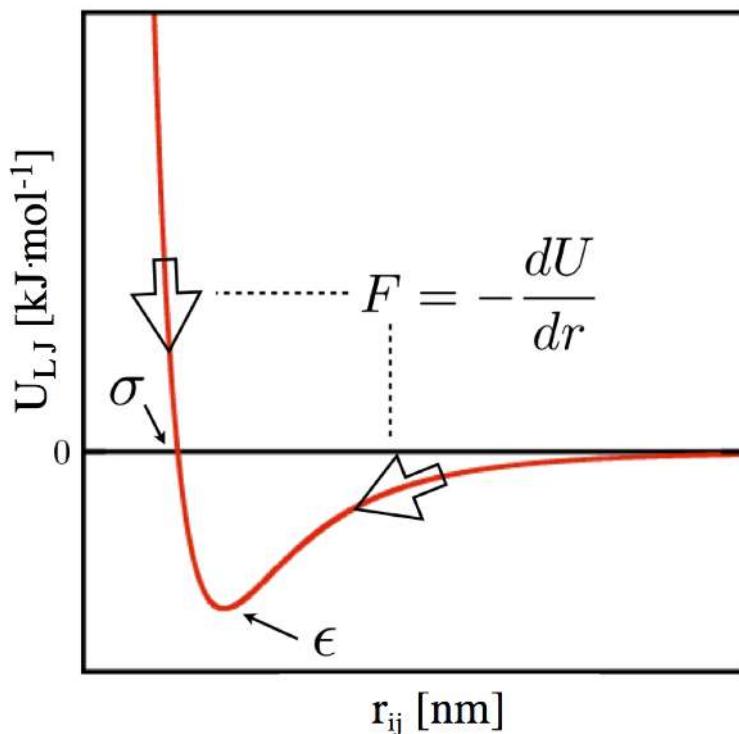


Figure 14.4: The Lennard-Jones potential  $U_{LJ}$  as function of  $r_{ij}$ . The force  $F$  is the derivative of  $U$  with respect to  $r$ , or in other words the slope of the function  $U$  in this plot (indicated by the two arrows).

indicates Coulomb's constant and in Equation 11  $\sigma$  (sigma) and  $\epsilon$  indicate the Van der Waals radius and the strength of the interaction, respectively. These parameters depend on the atoms involved. The Lennard-Jones potential is a combination of repulsive interactions originating from the Pauli exclusion principle, and attractive interactions resulting from London dispersion forces. Figure 14.4 shows an example of a Lennard-Jones interaction potential. Generally speaking, an energy is associated with the forces of the system: the slope of the energy determines the direction in which the atom is pushed as shown in Figure 14.4.

### 3.3 Parameters

The parameters included in force fields are usually derived from experimental techniques and (expensive) quantum chemical calculations (Leach, 2001). For example, bond lengths and angles, i.e.  $r_0$ ,  $\theta_0$ ,  $\chi_0$ , as introduced in the previous section “Interactions”, are based on distances in small molecule crystal structures, while force constants are based on a combination of infrared spectroscopy (which measures molecular vibrations) and quantum

chemical calculations. Existing force fields are continuously validated and improved to properly describe the nature of a given system. Many **modern force fields** like **GROMOS** (van Gunsteren *et al.*, 1996; Oostenbrink *et al.*, 2004), **AMBER** (Case *et al.*, 2014), **CHARMM** (Brooks *et al.*, 1983) and **OPLS** (Kaminski *et al.*, 2001) are extensively validated on thermodynamic parameters like **partition coefficients** (affinity for amino acids side chains for either water or oil phase), and peptide and protein folding **equilibria** (Daura *et al.*, 1999; van Gunsteren *et al.*, 2001; Oostenbrink *et al.*, 2004; Case *et al.*, 2014; Brooks *et al.*, 1983; Shaw *et al.*, 2010; Lindorff-Larsen *et al.*, 2012). This makes these force fields ready to simulate biological phenomena, since we are usually interested in the **equilibrium states and transition events** which are thermodynamic quantities. See Chapter 13 for more on thermodynamics.

### Limitations of Newtonian physics and force fields

MD simulations take a classical mechanical approach by using **Newton's second law of motion** to determine the collective dynamics of a set of particles, such as the atoms in a protein and the surrounding solvent (typically water); this is known as the **Molecular Mechanics** (MM) approach. Most simulations are performed without considering **quantum** effects explicitly, such as pi-pi interactions between aromatic rings (e.g. Phenylalanine), electron transfer, proton transfer, bond breakage, and even H-bonds in certain conditions.

For some chemical processes such as **enzyme** catalysis, **proton** transfer, and **pH** effects (typically involving hydrogen atoms) simulations using more advanced models that incorporate quantum physics may be required. Unfortunately, when compared to a classical mechanical approach, quantum mechanical (QM) methods require much **longer computation times** to investigate the **conformational space and energy landscape** for large systems such as proteins and polymers, which typically consist of many tens of thousands of atoms. QM methods are currently limited to hundreds of atoms at the very best. Hybrid QM/MM approaches may also be employed, but these also remain strongly limited by the expense on the QM part.

Fortunately, many biological effects that we may be interested in are macroscopic properties, such as protein folding or ligand binding, for which classical or Newtonian physics suffices. Moreover, for such macroscopic properties, our models can be, and typically also have been, validated with experiments.(van Gunsteren *et al.*, 2006)

However, you have to keep in mind that force fields cannot automatically adapt to new conditions, which means that they will work

only for atoms and structures for which parameters have been appropriately determined. If your system includes anything not described in the force field, e.g., a ligand or metal ions, this will typically make the software crash. Therefore, you need to parametrize your ligands and metal coordination centers before launching such a simulation. This is not trivial although several approaches and web tools are available to generate reasonably good topologies for small organic molecules. One of the major bottlenecks is that classical force fields describe electrostatic interactions in physical systems as static atom-centered charges that do not change during your simulation. In other words, the polarization that a real physical system undergoes in a dielectric medium such as water is entirely neglected. To address this problem, polarizable force fields that aim to describe the variations in charge distribution within the dielectric environment a very promising for future applications (Halgren and Damm, 2001). However, these force fields are not easy to set up and therefore currently not available for large systems.

If you need to include metal ions (for example in metalloproteins), things may even get more complicated. Here, transition metals such as copper, zinc and iron are particularly challenging as they exhibit several oxidation states. Often, a subtle equilibrium between two co-existing redox states will be present in a biological systems. Until now, even when using hybrid methods such as QM/MM, it is not possible to get sufficiently realistic parameters for metalloprotein simulations. However, there exist several approaches that approximate these systems as best as possible. For the interested reader, the review by Li and Merz (2017) provides a detailed overview of the advances, state-of-the-art and bottlenecks of modelling metal ions in classical dynamics.

Now, because of the assumption of additive pairwise interactions, you can combine everything (Equations 6-11) by simply adding all energy terms, as was already suggested by Equation 5, which leads to the following result:

$$\begin{aligned}
 U(r) = & \frac{1}{2} \sum_b k_b (r - r_0)^2 + \frac{1}{2} \sum_\theta k_\theta (\theta - \theta_0)^2 + \\
 & \frac{1}{2} \sum_\omega \sum_j k_j [1 + (-1)^{j+1} \cos(j\omega + \phi)] + \\
 & \frac{1}{2} \sum_\chi k_\chi (\chi - \chi_0)^2 + \epsilon \sum_i \sum_j (q_i \cdot q_j) / r_{ij} + \\
 & 4\epsilon \sum_i \sum_j [(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^6]
 \end{aligned} \tag{12}$$

Although not all details of this formula are equally important, you should notice the character of the different summations. The first four are single

Table 1: Overview of important molecular dynamics force fields.

Force fields	Description	Ref.
AMBER	Full-atomistic, used for proteins and DNA	Ponder and Case 2003; Tian <i>et al.</i> 2020
CHARMM	Full-atomistic, used for small molecules and macromolecules	Brooks <i>et al.</i> 1983, 2009
GROMOS	Full-atomistic, bio-molecular systems: solutions of proteins, nucleotides, and sugars	Oostenbrink <i>et al.</i> 2004; Reif <i>et al.</i> 2012
OPLS	Full-atomistic, Optimized Potential for Liquid Simulations	Kaminski <i>et al.</i> 2001; Shivakumar <i>et al.</i> 2012
MARTINI	Coarse grained, molecular dynamics simulations of large bio-molecular systems ( $\sim 4$ heavy atoms represented by a single bead)	Marrink <i>et al.</i> 2007; Monticelli <i>et al.</i> 2008; Souza <i>et al.</i> 2021

Note that force fields are constantly improved, so that there are several versions available for each. Therefore, it is recommended to use the latest version of a force field (unless you have a valid reason not to). The choice of the force field itself is not strictly predetermined and may often depend on the system of interest, the program package or the history and experience of the research team. A relatively recent review of force fields can be found in Lindorff-Larsen *et al.* 2012.

sums over all bonds ( $b$ ), angles ( $\theta$ ) and proper ( $\omega$ ) and improper ( $\chi$ ) dihedrals. Here, as above, we have simply left out cross terms between bonds, angles or dihedrals; this is the ‘other’ class already mentioned which is typically negligible for protein simulations. The last two sums go over all unique pairs of atoms  $ij$  for both the Van der Waals (first) and Coulomb (second) interactions. The number of bonds, angles and dihedrals in a system increases approximately linearly with the number of atoms. However, the pairs of atoms is quadratic in the number of atoms. Therefore, the non-bonded interactions are typically the vast majority ( $> 95\%$ ) of the computational effort of any biomolecular simulation (Leach, 2001).

## 4 Dynamics

Now that we have the forces (and energies) of our system, we can look at the mechanics behind the simulations. We will first look into dynamics in the form of MD simulations. In the next chapter (“Monte Carlo for Protein Structures”) we will cover Monte Carlo (MC) simulations, which can be seen as a statistical approach to the same problem that MD simulations address.

To describe a dynamic system, the coordinates (positions,  $r$ ) and momenta (velocities times the mass,  $p = m \cdot v$ ) of all particles are needed. These

Table 2: Overview of important molecular dynamics software packages.

Software packages	Description	Ref.
<b>AMBER</b> – Assisted Model Building with Energy Refinement	Molecular dynamics simulations of bio-molecules	Case <i>et al.</i> 2014
<b>GROMACS</b> – GRONingen MACHine for Chemical Simulations	Molecular dynamics simulations of proteins, lipids and nucleic acids	Pronk <i>et al.</i> 2013; Abraham <i>et al.</i> 2015
<b>NAMD</b> – NAnoscale Molecular Dynamics program	Molecular dynamics simulations of large bio-molecular systems	Phillips <i>et al.</i> 2005, 2020
<b>YASARA</b> – Yet Another Scientific Artificial Reality Application	Molecular dynamics program, including molecular visualization and modeling	Krieger and Vriend 2014, 2015
<b>CP2K</b> – Open Source molecular Dynamics	Atomistic and molecular simulations of solid state, liquid, molecular, and biological systems including QM/MM	Warshel and Levitt 1976

two sets of variables describe the ‘phase space’. Previously, in Chapter 13 we have already introduced the *conformational space* which only includes the *coordinates* (and not the velocities or momenta). In the course of this chapter, we will not go into the subtle difference between these, so for simplicity we will consider the conformational space an adequate description of our molecular system.

In addition to the coordinates and momenta, the *energies* of the particles are needed. These are split into two parts: the *potential* and the *kinetic* energies,  $U$  and  $K$  respectively. The potential energy derives from the *force field* description we introduced in the previous section, while the kinetic energy simply derives from the *temperature* of the system. For example at 0 K (in a frozen system) all velocities are zero, whereas at higher temperature the (average) velocity of all atoms increases.

The final ingredient required to understand the course of MD simulations is the relation between the *forces* (derivative of the potential energy) and the *positions of the atoms*, which goes according to Newton’s famous law of motion:

$$F = m \cdot a. \quad (13)$$

Here,  $F$  is the force,  $a$  the acceleration and  $m$  the mass. As we can primarily calculate the change in velocity, this equation can be rearranged to

$$a = F/m. \quad (14)$$

Actually, Newton wrote his law as a differential equation

$$\frac{\partial a(r)}{\partial r} = \frac{1}{m} \frac{\partial F(r)}{\partial r}, \quad (15)$$

with  $r$  the position,  $F$  and  $a$  as before, and  $\partial$  indicating the (partial) first derivative. Remember that the differential of the energies is the force, so for dynamics we require a differentiable function for the energy.

Now, solving Newton's equation of motion yields a description of the motion of the particles (atoms) involved, which is what we want when we perform an MD simulation. However, this differential equation can only be solved analytically for a system of two particles. Therefore, in a typical MD simulation dealing with large number of particles, a numerical approximation is required to obtain the changes in velocity and position over time. The analytical solution to Newton's equation has nice properties: it is energy-conserving, reversible and deterministic. The numerical approximation, however, may not be so well behaved. Moreover, for a system of more than two particles, we know that this behaves intrinsically non-deterministic or 'chaotic'; meaning that even the tiniest differences in the starting situation will eventually diverge into vastly different behaviour of the simulation. In practice however, for our biomolecular simulations, this is not a problem as average and/or aggregate behaviour should be independent of the details of the integration scheme (as long as the errors are not systematic).

## 4.1 Integrating equations of motion

### The Verlet integration scheme

The most used numerical scheme to integrate the equations of motion in MD simulations is the 'Verlet integration' scheme, which is based on a Taylor expansion of Newton's differential equation (Verlet, 1967). Using a Taylor expansion is one way of writing a differential equation, which uses infinitesimally small increments, as a numerical approximation using finite size time step increments (See Panel "Derivation of the Verlet integration scheme" for more details). Note that there exist a variety of additional molecular dynamics integrators that can be used. However, a detailed discussion of the differences between the methods, their benefits and limitations goes beyond the scope of this chapter and further details may easily be found elsewhere (see for instance Leach (2001) or Frenkel and Smit (2002)).

### Choosing timesteps

The time-step  $\Delta t$  needs to be chosen carefully. If a too small timestep is used, precious computational effort is wasted. A too large timestep on the other hand will yield integration errors and prevent the proper dynamics of the system, as shown in Figure 14.6. Strictly, 10 to 20 integration steps are

```

1 def molecular_dyanamics(r ,num_steps ,N,V,T):
2     # generate initial velocities ,
3     # depending on set Temperature
4     v = initial_velocities(T)
5     for i in range(num_steps):
6         # compute forces , from current coordinates r ;
7         # depends on (globally defined) force field
8         F = calc_forces(r)
9         # calculate thermostat scaling factor for set
10        # temperature (depends on velocities )
11        λ = temp_scaling(v)
12        # calculate barostat scaling factor for set
13        # pressure (depends on coordinates and forces )
14        μ = pressure_scaling(r , F)
15        # using calc'd forces F, update velocities v
16        v' = λ*(v + F/m*Δt)
17        # using calculated velocities v ,
18        # update coordinates r
19        r' = r + v'·Δt
20        # apply constraints – certain atomic distances
21        # and angles that are not allowed to change ,
22        # like for hydrogen atoms
23        r'' = apply_constraints(r' , r)
24        # correct v for constraints
25        v = (r'' - r)/Δt
26        # apply barostat and obtain final coordinates
27        r = μ*r"
28        # every set number of steps , perform output
29        # of system variables , r , v , or F
30        if (i%nst_log) do_print_log(r , v)
31        if (i%nst_r) do_output_r(r)
32        if (i%nst_v) do_output_v(v)
33        if (i%nst_F) do_output_F(F)
34    # end for loop
35 # end Molecular Dynamics

```

Figure 14.5: Molecular Dynamics algorithm for molecular simulations in pseudo code Python style.

### Derivation of the Verlet integration scheme

The trick to deriving the Verlet integration scheme is to take the Taylor expansion of Newton's equation of motion for the forward ( $t + \Delta t$ ) and backward ( $t - \Delta t$ ) step in time. Adding these cancels the third order term (eliminating any approximations there), and yields a formula of positions at the next time point ( $t + \Delta t$ ) as a function of the current ( $t$ ) and previous ( $t - \Delta t$ ) time points, which contains only zero, second and fourth order terms. If our final equations include only the zeroth and second order terms, we thus make only a fourth order error.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{r}'(t)\Delta t + \frac{1}{2!}\mathbf{r}''(t)\Delta t^2 + \frac{1}{3!}\mathbf{r}'''(t)\Delta t^3 + \dots \quad (16)$$

where the first derivative (the velocity) is  $\mathbf{r}'(t) = \mathbf{v}(t)$  and the second derivative (the acceleration) is  $\mathbf{r}''(t) = \mathbf{a}(t)$ .

$$\begin{aligned} \mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{r}'(t)\Delta t + \frac{1}{2!}\mathbf{r}''(t)\Delta t^2 + \frac{1}{3!}\mathbf{r}'''(t)\Delta t^3 \\ \mathbf{r}(t - \Delta t) &= \mathbf{r}(t) - \mathbf{r}'(t)\Delta t + \frac{1}{2!}\mathbf{r}''(t)\Delta t^2 - \frac{1}{3!}\mathbf{r}'''(t)\Delta t^3 \\ \hline \mathbf{r}(t + \Delta t) + \mathbf{r}(t - \Delta t) &= 2\mathbf{r}(t) + 2\frac{1}{2!}\mathbf{r}''(t)\Delta t^2 \end{aligned} \quad (17)$$

Which, using  $2\frac{1}{2!} = 1$  and  $\mathbf{r}''(t) = \mathbf{a}(t)$  can then be rearranged to:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \mathbf{a}(t)\Delta t^2 \quad (18)$$

One can further rewrite this using the velocity  $\mathbf{v}(t - \frac{1}{2}\Delta t) = \mathbf{r}(t) - \mathbf{r}(t - \Delta t)$  into:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t - \frac{1}{2}\Delta t) + \mathbf{a}(t)\Delta t^2 \quad (19)$$

needed per period of a harmonic vibration to integrate accurately. However, due to Verlet's third-order accuracy, 5 integration steps suffice in practice. Typically, the fastest vibrations in biomolecular simulations, are the hydrogen atoms that vibrate at 10 fs (Feenstra *et al.*, 1999). Therefore, 1 or 2 fs timesteps are typically used for MD simulations.

To put these time scales in perspective, the 2 fs timestep is at the bottom of the scale in Figure 14.2. Biologically relevant behaviour starts at the right-hand block; around milliseconds. That means, we need an order of  $10^{12}$  integration steps to start getting into biological timescales. This is typically not feasible, however, fortunately there is still a lot of interesting and biologically relevant behaviour that we may observe at achievable timescales of nanoseconds to microseconds.

### Temperature and pressure

The integration schemes used in MD simulations conserve energy, and the simulations are performed in a box of fixed dimensions. Thermodynamically speaking, these are thus in the *NVE* ensemble, where the number of particles ( $N$ , atoms), volume ( $V$ ) and energy ( $E$ ) are constant. As noted in Chapter 13, most biological experiments, however, are done in the so-called ‘grand canonical ensemble’ where chemical potential ( $\mu$ ), Pressure and Temperature are constant. This is very hard to achieve in general in a simulation (Pool *et al.*, 2012), so as the next best option, a typical MD simulation is set up in the *NPT* ensemble where, as before, the number of particles is constant, and pressure ( $P$ ) and temperature ( $T$ ) are kept constant on average over some relatively short time period. There are simple solutions for maintaining constant temperature and pressure, called ‘weak coupling’, that have been used for many decades (Berendsen *et al.*, 1984), but currently more recent methods are generally used as they avoid some serious problems in long time-scale simulations (Cheng and Merz, 1996; Mor *et al.*, 2008; Lingenheil *et al.*, 2008).

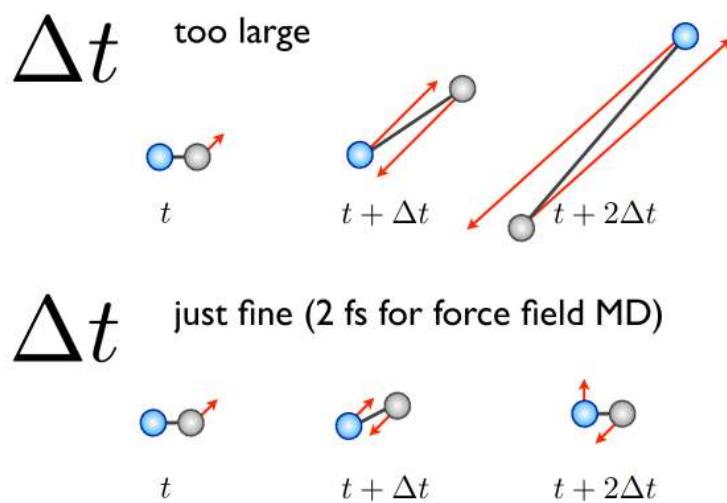


Figure 14.6: Effect of different time steps on the evolution of two particles (atoms) in a molecular dynamics simulation. The forces acting on the atoms are depicted with the red arrows. Note that if the time step is too large ( $> 2$  fs for full-atomistic MD simulations) the forces increase gradually and the positions of the particles are swapped at each step, preventing a physically accurate description of the dynamics.

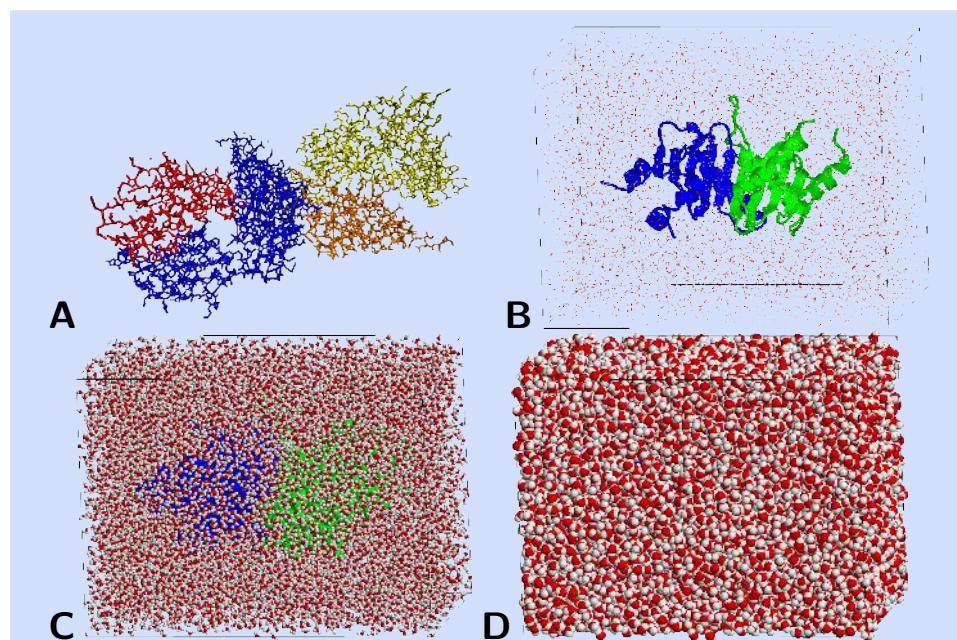


Figure 14.7: Illustration of an exemplary set-ups for an MD simulation. **A** a protein in **vacuum**. **B-D** solvated in a box with **water molecules** illustrated as **two (O-H) bonds** (**B**), as **small spheres** (**C**), or using the **full Van der Waals radii** of the atoms (**D**).

### Biological systems: water needed

When we simulate a protein, the largest part of the investigated system consists of **just water**. This water is absolutely crucial to the **proper behaviour** of the protein, as was already introduced in Chapter 1 Section 3.1, but not something one is actually interested in by itself. There are two different ways to treat water in a simulation: using either an **implicit** or an **explicit** water model. In an explicit water model, the water is modelled as **discrete H<sub>2</sub>O molecules**, with a triangular arrangement of the oxygen and two hydrogen atoms. In an **implicit** model, the water surrounding the protein is described using an **average ‘field’ description**, which means that water molecules are not treated as actual molecules consisting of particles, but are instead approximated by a **dielectric constant** which effectively **dampens the electrostatic interactions**. In the case of explicit water, for two protein molecules of about 3.5 nm diameter, we need a box of 9.3 nm by 6.9 nm by 5.6 nm filled with 10865 water molecules. Figure 14.7A shows the two proteins in vacuum, B shows the proteins with the wa-

ter molecules drawn as lines. In C the water atoms are shown as small spheres where we can still see the protein somewhere in the middle. And in D the full Van der Waals radii of the atoms are drawn which completely hides the protein, thus emphasizing the amount of water needed relative to the size of the protein molecule. This example illustrates why explicit water simulations are much slower and thus computationally more expensive than simulations using an implicit water model. However, depending on the research question and on the experimental counterpart used to validate the simulations, it may be justified to use an implicit model.

### Common approaches for performance enhancement

Figure 14.2 highlights the enormous gap between the integration timestep needed, and biologically relevant timescales that we would like to approach in our simulations. So, many tricks have been invented over the years to improve efficiency of the calculations, as well as a wide variety of approximations that enhance performance typically by reducing complexity of the simulated system.

- Trivially: The forces derived from pairwise potentials are related as follows  $F_{ij} = -F_{ji}$ , which means the interaction between an atom pair needs to be computed only once.
- The non-bonded potentials (Coulomb:  $\sim 1/r$ ; Lennard-Jones:  $\sim 1/r^6$ ) tend to zero at large distances, so we can use a distance cut-off beyond which interactions are not calculated.
  - Atoms only move a little in one step, so we use a pair-list to keep track of atoms within the cut-off distance which we only re-evaluate every so many steps (calculating distances is relatively expensive because there are quadratically many pairs for the number of atoms).
  - Evaluating  $r$  is relatively expensive due to the square-root; so compare the square of the cut-off radius with the square of the distance.
- Large distances change less, which forms the basis for twin-range and multiple time-step methods.
- Many Processor/Compiler/Language specific optimizations:
  - use of Fortran vs. C (and even assembly code) in performance critical parts of the code
  - optimize cache performance for very large arrays (positions, velocities, forces, parameters)
  - compiler optimizations
  - efficient use of multi-core systems

- use of GPU nodes for efficient execution of so-called ‘vector’ and grid operations
- Maximum Time step is limited by vibrational frequencies (see also Figure 14.2):
  - The carbon–hydrogen bond angle vibration is  $10^{-14}\text{s} = 10\text{ fs}$ . When using 10-20 integration steps per vibrational period, we would need a time step of 0.5 fs, but in practice 1 or 2 fs time steps are used (1.000.000 or 500.000 steps for 1 ns)
  - Bond distances also have high frequencies of vibration, but for protein simulations these are already always constrained.
  - If we also remove other fast vibrations (hydrogen atom bond and angle motion and some out of plane motions of aromatic groups) using constraints we can use a timestep of 2 or 4 fs, and even stretch it to 6 or 7 (Feenstra *et al.*, 1999; Hopkins *et al.*, 2015).
- When using simplified force fields, fewer particles are required, thus reducing the time needed for the force computations.
  - United atoms: aliphatic groups (a carbon bound to only other carbons and hydrogen atoms) are uncharged and nearly spherical, because the hydrogens are so much smaller than the carbon atoms. These can be treated as single particles with a radius slightly larger than just the carbon atom and a larger mass (van Gunsteren *et al.*, 1996; Yang *et al.*, 2006).
  - Coarse-graining: multiple atoms are combined into one particle with so-called effective interaction potentials, resulting in fewer atoms, and also in (much) larger timesteps (May *et al.*, 2014; Singh and Li, 2019).
  - Implicit water: water may also be described not as a collection of molecules, but instead the water molecules are described as an average ‘field’ (Roux and Simonson, 1999; Onufriev and Case, 2019).

In this list we mention only a few strategies to reduce the simulation time. More comprehensive and technical treatises may be found in Hess *et al.* (2008) and Pronk *et al.* (2013), both focused on the GROMACS MD simulation software, and Bowers *et al.* (2006), Dror *et al.* (2012) and Grossman *et al.* (2013) from the D. E. Shaw research group who developed their own simulation software but also their own supercomputer based on specialized MD hardware called ANTON.

## 4.2 Convergence of state properties

Viewing the movie resulting from the MD simulation is not an efficient way to analyse simulations, and many details are not registered by the human eye. Moreover, the particular order of events, or even a particular conformation, observed in a simulation, is not of interest in itself. When analysing multiple simulations of the same system, one will observe that the **order is not conserved**, and many **small variations** on conformational states exist, due to the chaotic nature of the system, as described in the section on dynamics.

Thus, in this part, we will introduce how to **quantify simulation results**. An important overall feature of molecular simulations is some measure of **convergence**. Strictly speaking, convergence indicates **how well the sampled conformations represent the system at equilibrium**. It can also be seen as a measure of the **exploration of the energy landscape (or conformational space)** of the system of interest. Moreover, not all parts of the energy landscape are equally relevant to the behavior of the simulated system; (very) high energy parts are (very) **unlikely** to be visited, in reality and as well as in the simulation, as we saw in the previous Chapter 13. The parts that do not have (very) high energies are often called the '**reachable**' conformational space.

So, **sampling** can be considered as the extent to which this reachable part of conformational space has been **visited** during the **simulation**. Finally, **complete sampling** then corresponds to **(complete) convergence**. However, you should always keep in mind that the dynamics of proteins are stochastic. When starting a simulation from a given structure you might get stuck in a **local minimum**, in which case, you will not be able to explore all the relevant regions of your energy landscape. Now, we will show some examples of how to (not) **recognize convergence** from the results of a simulation.

### Enhanced Sampling techniques

When launching a molecular dynamics simulation from a **given starting point** (like a protein crystal structure), there is the danger of getting stuck in a local minimum during the simulations. In that case, you won't be able to **sample sufficient parts of the conformational space** of your system to study what is going on. To avoid this multiple minima problem, several advanced sampling techniques can be used (Yang *et al.*, 2019). Here, we want to briefly introduce the concepts of **umbrella sampling** and **replica exchange molecular dynamics** (see Panel "Replica exchange molecular dynamics (REMD)"); we will also come back to these in Chapter 15.

Umbrella sampling requires you to have some **prior information** of what is going on in your system (Kästner, 2011). This is different from REMD which allows you to freely explore the conformational space. In umbrella

sampling, you need to pre-define a specific order parameter (a collective variable) that will properly describe relevant changes in your system. What is a good order parameter depends on your research questions; it may for instance be the distance between a ligand and its protein binding site, or between opposite sides of an opening and closing channel protein, or the end-to-end distance in an unfolding event. Once this is clearly defined, your system will be constrained using harmonic potentials (umbrellas) all along this order parameter to sample the parts of space you are interested in. Similar to REMD, this technique is computationally expensive. However, it does allow you to get a quantitative estimation of the free energy for a given event.

### Replica exchange molecular dynamics (REMD)

Replica exchange molecular dynamics (REMD) allows you to sample large portions of your configurational space by running several identical copies (replicas) of your system (e.g., the protein and water) simultaneously at different temperatures (e.g., between 300 K and 700 K). While the higher temperature replicas are provided with sufficient energy to jump over high energy barriers, the low temperature replicas are able to properly sample the local minima. Over time, neighboring replica are allowed to swap their positions which lets you sample different portions of the surface. The major bottleneck of this techniques is the computational costs due to the fact that all the replica need to be simulated in parallel.

### Observing sampling – RMSD

Root mean square deviation (RMSD, introduced in Chapter 3) is one way to chart the progress in sampling the conformational space, by effectively measuring the distance from the starting point. Figure 14.8 shows the RMSD of the conformation of a protein with respect to the starting (crystal) structure during the simulation. It is clear that the RMSD increases over time, as expected (because the atoms move, the structure changes). Importantly, the three panels show this simulation at three different time scales: 0-100 ps (top left), 0-1 ns (top right) and 0-5.5 ns (bottom right). At each of these timescales, the RMSD appears to reach a plateau roughly half-way through the plot. But the plots are all based on the same simulation, and on the next plot with a longer timescale one sees the increase in RMSD continuing. In other words, each time the plateau we see is ‘temporary’, and the impression it gives of reaching a converged value (in this case for the RMSD) is false. We can be sure this holds for the plateau observed in the bottom right figure as well. If we would run this simulation for much longer, say 100 ns, we

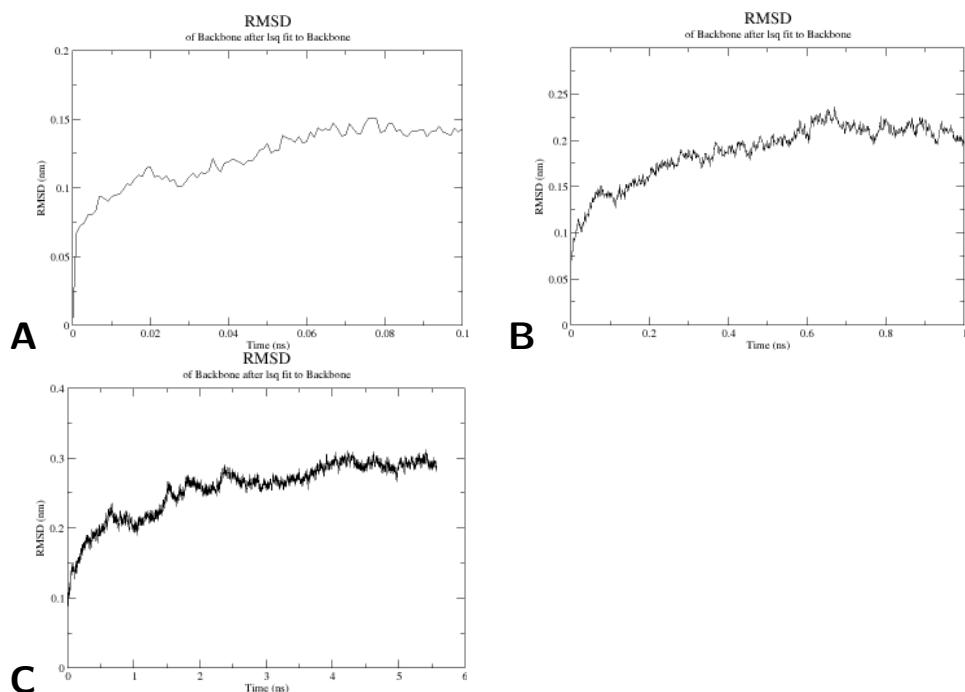
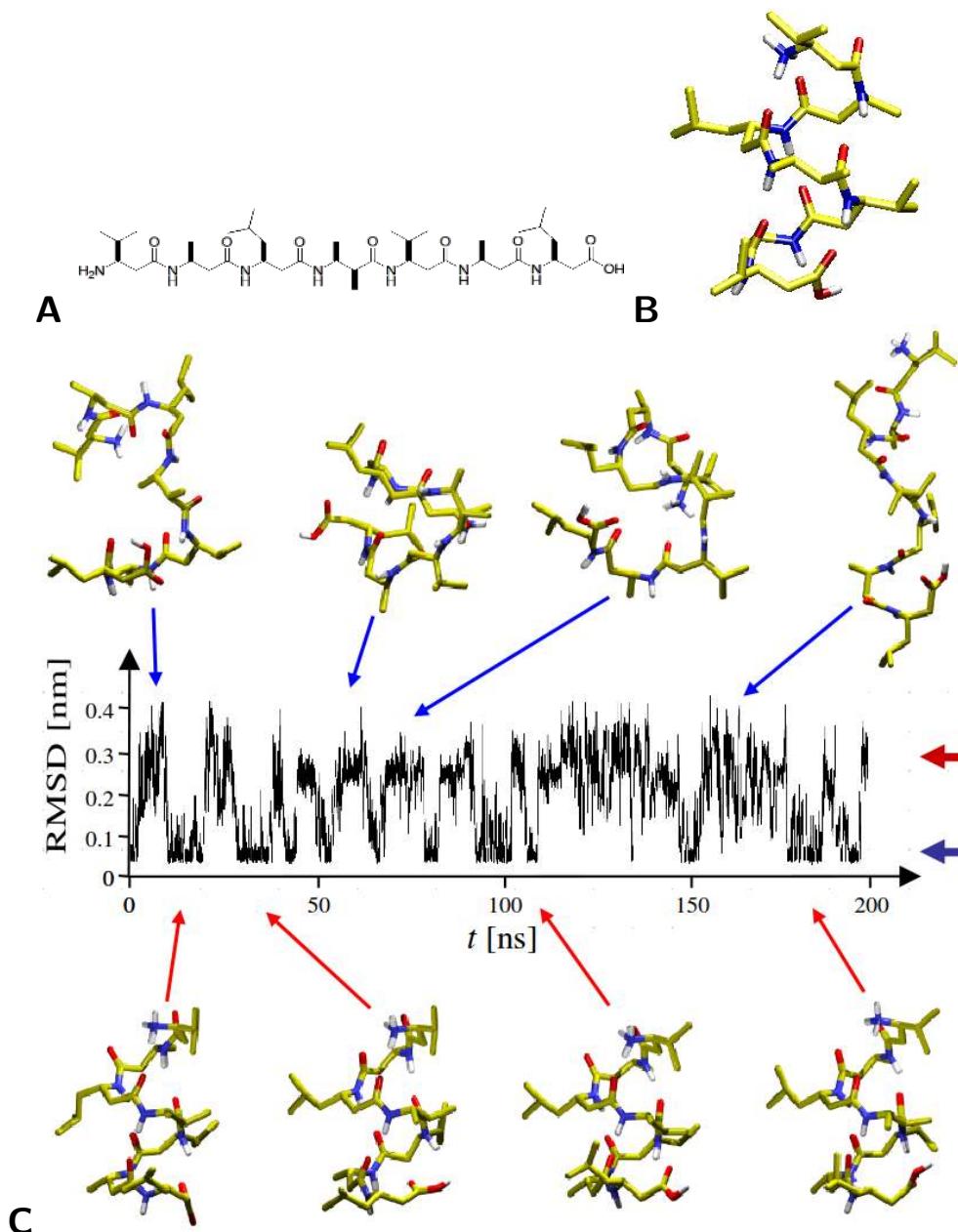


Figure 14.8: Apparent convergence in root-mean-square-deviation (RMSD) may be observed in protein simulations at different timescales. The three plots each derive from the same simulation. At the shortest timescale, up to 0.1 ns or 100 ps **A**, the RMSD appears to reach a plateau, suggesting convergence. However, extending the simulation to 1 ns **B**, and to 5.5 ns **C** shows this convergence is **transient**.

would probably still see similar behaviour. These are all symptoms of **diffusive motion in a high-dimensional space**; as the distance from the starting point increases, the sampling spreads out **more and more at a given RMSD** distance from the starting point. This is similar to **diffusion** of molecules in a solution where the distance reached from the starting point increases with the square root of the time elapsed – but the effect is much stronger in high-dimensional conformational space, as we would take a higher power root of the elapsed time here. Therefore, the **change in distance gets slower**, but the actual changes in conformation still occur as much as in the beginning.

When sampling of the system reaches timescales at which these **conformational transitions start to equilibrate**, we should see a change in this behavior. We expect the **distances to become smaller again** after some time. In fact, Figure 14.9 shows that for a small peptide this is achieved at timescales of a **few hundred ns**. This seven-residue peptide was studied at the ETH in Zürich in the late 1990's. The **NMR** predicted an  $\alpha$ -helical conformation for the peptide, shown in Figure 14.9b, but the side chains could not be resolved due to lack of data, and some violations of the distances obtained from the experiment.



**Figure 14.9:** **A** Small beta-peptide of seven residues (**beta-peptides** have an additional carbon atom in the backbone compared to normal (**alpha**) peptides). **B** The peptide forms a **helical structure** according to NMR. **C** Simulations show a very dynamical behaviour, where the RMSD to the helical structure **increases**, but also **decreases again** repeatedly. This indicates the peptide **unfolding** (increasing RMSD) and **(re-)folding**; importantly this is the very first reversible folding simulation ever (Daura et al., 1999; van Gunsteren et al., 2001). Reproduced with permission from Daura & Oostenbrink (pers. comm.).

Very extensive simulations, for those times, of this peptide were performed in the group of Wilfred van Gunsteren (also ETH Zürich) (Daura *et al.*, 1999); note that in those days this was still a major investment of computational power, requiring almost half a year on a large compute cluster. The simulation shows the peptide going through various different conformations, some of them helical (or helix-like), others decidedly different, as can be seen in Figure 14.9c. The RMSD with respect to the helical (NMR) structure shows that this simulation, for the first time in history, reached equilibrium behaviour. The RMSD not only went up, but also went down again. In other words, during this simulation the system visits different states, but also returns to states visited before. It is important to realize that, in this plot, a low RMSD always means the same state (the native, folded helical state, as can be seen in the snapshots below the plot), but a high RMSD can mean many different things. The snapshots above the plot show that these states are all clearly not helical, but they are also not similar to each other. Clustering of these unfolded structure showed that there are (only) about 1000 distinct conformational states in this unfolded ensemble (for details, please refer to van Gunsteren *et al.* 2001).

Ten years later in 2010, Shaw *et al.* showed that now we can obtain reversible folding also for a small (70 residue) protein. The timescale needed is much larger than that for the peptide; here we see the simulation goes up to 200  $\mu$ s (note, that this is a similarly major investment of computational power as the peptide was a decade earlier, of the order of many months in parallel on a large number of highly specialized CPU's).

### Evaluating your MD simulations – Order Parameters

To be able to track the progress of a molecular dynamics (MD) simulation, several powerful analysis tools have been developed over the years. Although it may be fun and also seemingly intuitive to visualize your results as a movie by loading the MD trajectories into programs such as VMD or Chimera, it is not sufficient to understand what is going on and you need a quantitative way of analysing your results. Luckily, MD packages provide a large number of ready-implemented parameters that can be used to track changes between states of interest in your simulations. First, you should check the general properties of your system, such as the average pressure and the temperature, to make sure everything went well. In a next step, tracking the changes of the following order parameters may help you pinpoint potential events during your simulations:

- **Root Mean Square Deviation:** Measure of dissimilarity (distance) between two molecular conformations used to compare

the conformation of each frame with respect to, e.g., the **starting point** of the simulation

- **Root Mean Square Fluctuation:** Measure of structure variation, it calculates the **standard deviation** of the deviation of an atomic position over time.
- **Radius of gyration:** Measure for the **compactness** of a structure, powerful to identify **unfolding** event in your protein structures.
- **Solvent Accessible Surface Area:** Measure to determine the **surface** of the biomolecules that are accessible to the surrounding solvent (usually water). This is very helpful in protein **folding** and stability studies.
- **Number of hydrogen bonds:** Compute the numbers of **hydrogen bond contacts** in your system. Particularly interesting when looking at complexes (ligand-protein, protein-protein) and (un)folding events.
- **Essential Dynamics** (ED) analysis (Amadei *et al.*, 1993; van Aalten *et al.*, 1997): This technique is a principal components analysis (**PCA**) of the **atomic coordinates**. It takes the covariance matrix of the coordinates of the atoms. This covariance matrix is diagonalized, yielding eigenvalues and eigenvectors; the eigenvectors describe **collective motions** of the atoms analyzed. The eigenvalues (or loadings) represent the **magnitude** of each of these motions. We can now simply focus analysis on the eigenvector(s) with the largest eigenvalues, or several of the largest – these are also referred to as '**Essential Modes**'. Finally, if we choose only the two largest, we can project motions onto a two-dimensional plot for visualization.

For all these measures, the parameter of interest does not have to be calculated for the whole system – often for example only the **C<sub>α</sub>-atoms** or the **active site or binding region** are analyzed.

Depending on the problem at hand, looking at additional order parameters may be interesting. For example to analyse the simulation of a ligand-protein complex, the **contact frequency** of the ligand with specific amino acid inside the binding site or the **distance between the centers of mass**. Finally, despite the large number of available tools that come with most simulation packages, in some cases, your simulations may require you to do some **self-coding** to be able to follow important changes during your simulations. A Python library to analyse your simulations that may come handy can be found here: <https://www.mdanalysis.org/> (Michaud-Agrawal *et al.*, 2011; Gowers *et al.*, 2016).

### 4.3 Temperature dependence

In the previous Chapter 13, Figure 13.4a,b, we already saw how the folding-unfolding **equilibrium** that we are sampling in these simulations for this peptide, depends on the **temperature** of the simulation. The middle plot (at 340 K) is the one also shown in Figure 14.9c, where on average the peptide is folded 50% of the time (and unfolded the other 50%). At lower temperatures in Figure 14.10a, we see the peptide tends to spend a larger fraction of time in the folded state. At **higher** temperatures we see the opposite; a larger fraction of time spent in the **unfolded** state(s). Note that the temperature changes are fairly small, only 10-20K up, and 20-40K down, and it is reassuring that simulations are able to capture these relatively subtle effects. Figure 14.10b shows that also **high pressure** has the effect of unfolding. Both effects of high temperature and high pressure unfolding of proteins are also experimentally well known facts, for by far the majority of proteins.

#### Simulated annealing

One way to overcome free energy barriers and escape from free energy minima in a molecular system is to **increase the temperature**. At room temperature a protein system can be trapped in a free energy minimum from which it can escape only after a very long time, as we already discussed previously. For MD simulations, a microsecond is already considered a long time. When performing a simulation at higher temperature, 400 K for example, the system can explore many more different configurations than at room temperature. However, such configurations do not necessarily represent configurations of the system at room temperature, as then they might be very **unlikely**. To overcome this, one can take a number of **snapshots** from the high temperature simulation, and use these as a **starting point** for MD simulations at room temperature. The snapshots will then equilibrate to configurations that are likely to occur at room temperature. This approach is known as **simulated annealing** and is used in for example resolving NMR structures by letting high temperature conformations **anneal to the constraints** obtained from various NMR spectra.

One problem with the simulated annealing approach is that the resulting conformations **do not represent the equilibrium distribution**. The selection of the snapshots chosen to equilibrate to room temperature does not follow the Boltzmann distribution. In principle, the MD simulation at room temperature should sample the equilibrium distribution, but may require a long time to achieve that. To overcome this issue, but still make use of the enhanced sampling provided by performing MD at high temperature, the procedure to select snapshots for equilibration at lower temperature should follow the **Boltzmann distribution**. The solution to this is parallel temper-

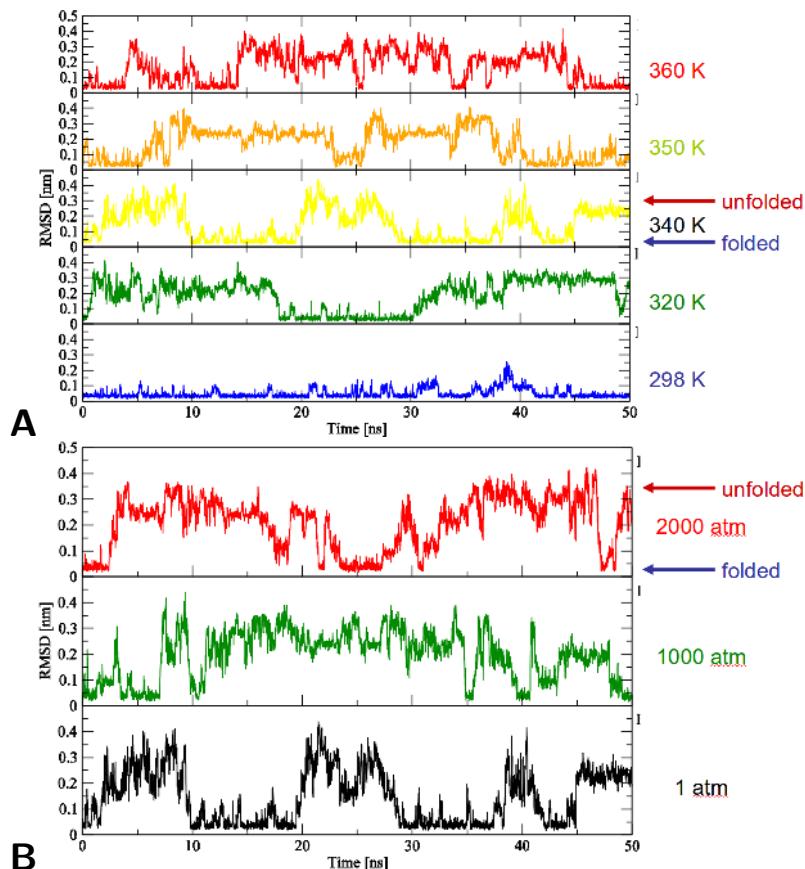


Figure 14.10: **A** The folding equilibrium of the beta-peptide depends on temperature: at lower temperatures a larger fraction of time is spent in the folded (low RMSD) state, at higher temperatures a smaller fraction is folded. **B** Also at higher pressure, the fraction folded decreases (Daura et al., 1999; van Gunsteren et al., 2001). Reproduced with permission from Daura & Oostenbrink (pers. comm.).

ing or replica exchange, which we already touched on in Panel “Replica exchange molecular dynamics (REMD)”.

#### 4.4 Homology model optimization

Now we will illustrate simulation, sampling and convergence in practice. Here we have used MD simulations to optimize details of a homology model of an enzyme (Feenstra *et al.*, 2006). Figure 14.11a is an overview picture of the enzyme Styrene mono-oxygenase (SMO) homology model protein structure. If you look closely, you will observe two additional molecules in the center: the styrene (STY) ligand, and the flavin-adenin-dinucleotide (FAD) co-factor.

The homology model for this structure was built based on two templates

of about 23% sequence identity, which is far enough to be a difficult target for homology modelling. Docking of the ligand into the model yielded binding conformations that did not correspond to known products of this enzyme. We therefore used MD simulations to optimize the model, with a main aim to improve the shape of the active site pocket, but to do that we would also need to relax strain due to bad contacts in the initial model. The procedure was therefore aimed to relax this strain, without leading to distortion of the binding site region. We first allowed the water to relax around the protein during a 1-ps MD where the positions of the atoms in the protein were restrained. Subsequently, we released the constraints on parts of the protein, so these were also allowed to relax. First, only the side chains of the residues outside the binding pocket were released, and simulated for 1 ps. Next, also the backbone of these residues were released and simulated for 10 ps. Then, the side chains of the binding residues were released, and simulated for another 10 ps. During this whole procedure the positions of the FAD co-factor and the styrene substrate were also restrained, this allowed the protein structure to relax around the bound co-factor and substrate. Finally, only the backbone atoms of the binding residues were restrained for 100 ps, while the rest of the protein as well as FAD and styrene were free.

To be able to track progress during the optimization simulations for this homology model, we used Essential Dynamics (ED) analysis (Amadei *et al.*, 1993; van Aalten *et al.*, 1997), see also Panel “Evaluating your MD simulations – Order Parameters”. We can now simply focus analysis on the two largest eigenvalues (the essential modes) and project these motions onto a two-dimensional plot for visualization.

To assess stability of the protein, and to see if the optimization procedure was successful in improving stability, we performed two sets of simulations, one started from the non-optimized ‘raw’ homology model, while the second set of simulations was started from the optimized model. We visualized the behaviour in these two sets of simulations using ED analysis on both. Figure 14.11b shows the ED plot representing the motion during these simulations, projected onto the two largest eigenvectors; i.e. the two largest collective motions going on in these simulations. It is clear, that both sets of simulations (‘raw’ vs. ‘optimized’) have different behaviour. The ‘raw’ simulations tend to start out by moving rapidly away from their starting point, and in different directions; this indicates strain in the initial conformation. In contrast the ‘optimized’ simulations move more gradually and more similarly. Finally, it is clear that both sets of simulations have a different starting point: this of course corresponds to the changes in the structure due to the optimization procedure.

The motion of the protein structure corresponding to each of these two eigenvectors can be visualized in a movie. In such a movie, you will be able to see that all atoms contribute to the overall motion, although not all atoms equally. You will also see that the two (eigenvector) motions are

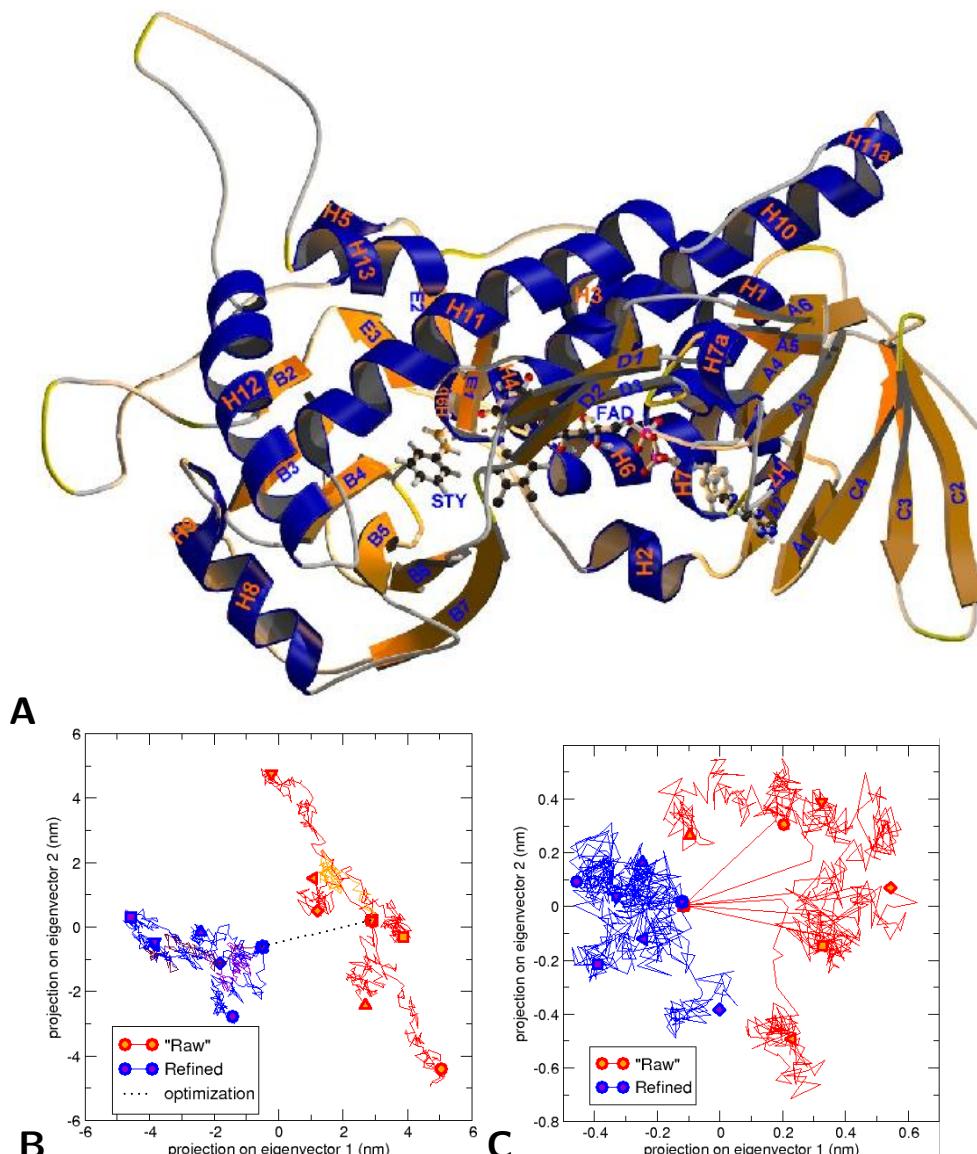


Figure 14.11: **A** Structure of the Homology Model Styrene Mono-Oxygenase (SMO) Enzyme. **B** Essential Dynamics (ED) analysis of the  $C\alpha$  atoms, showing backbone rearrangements during simulations starting from the ‘raw’ structure (blue), which are distinct from those started from the ‘refined’ homology model (red). The dotted line indicates the optimization path. **C** The same, but now the ED analysis was performed on the active site region only. Here, structural effects (difference between starting points) are small - the optimization path can not even be seen here. Nevertheless, overall behaviour of the ‘refined’ simulations is still distinctly different from that of the ‘raw’ simulations. The long straight lines ‘shooting’ out from the ‘raw’ starting point indicate high levels of strain in the ‘raw’ structure (red), which is relaxed in the refined structure (blue). Figure modified from Feenstra et al. (2006).

distinct; different atoms contribute more to each of them, and the motions are (often) in different directions. Finally, you will also see that different parts, sometimes quite far away in the protein structure, move in and out together in a sort of concerted way.

## 5 Outlook and summary

Molecular dynamics simulations have shown their usefulness in giving a detailed view on how proteins work; in the words of Shaw c.s., a “Computational Microscope for Molecular Biology” (Dror *et al.*, 2012). With ever increasing computational power, microsecond simulations are within reach and millisecond simulations have already been produced for a few selected systems. Current challenges lie in making available methods accessible, and in handling and analysing the huge amounts of data produced from the simulations. Before starting to run an MD simulation, there are two important things that you need to think about, namely which simulation technique is adequate to solve your scientific question and do you have experimental data to validate your models. As numerous techniques have been developed over the years and are now available to the scientific community, you should make sure to understand the basic concepts, usefulness and limitation of the technique you want to use. Later, it is extremely important to support your results and validate your models to guarantee their physical relevance. Often, you will be able to work in close collaboration with experimental groups or even perform experiments yourself. Alternatively, experimental data can be used from published results in the literature. In any case, make sure to be aware of what is going on in the experiment, the information it is able to provide, its benefits, bottlenecks, errors sources, and limitations.

## 6 Key concepts

Concepts that should be known after reading this chapter:

- The importance of molecular motions:
  - Proteins are dynamic and understanding their motions is crucial to be able to understand their biological function
  - Always remember and be aware of the relevant length and time scales to observe a biological event
- MD Simulations are based on:
  - Newton’s classical mechanics
  - force fields that contain the description (parametrization) of the bonded and non-bonded interactions
  - integrating the equations of motion
  - describing thermodynamics ensembles of the system
  - Trajectory on Energy Surface

- In protein Dynamics:
  - make sure that you have an adequate experimental counterpart to guide and validate your simulations
  - the level of detail of your forcefield, e.g. atomistic or coarse-grained, should match the problem you are interested in
  - attainable timescales and level of convergence should match the problem you are interested in

## 7 Further reading

- Understanding Molecular Simulation – Frenkel and Smit (2002)  
→ General, in-depth introduction to simulation
- Simulating the Physical World – Berendsen (2007)  
→ Technical treatise on physical properties, simulation algorithms and statistical thermodynamics
- Molecular Modelling: Principles and Applications – Leach (2001)  
→ Overview of Molecular Modelling and Computational simulations.
- The Art of Molecular Dynamics Simulation – Rapaport (2004)

## Author contributions

Wrote the text:	HM, JvG, QH, AM, JV, KAF
Created figures:	HM, ASR, JV, KAF
Review of current literature:	HM, ASR, JV, KAF
Critical proofreading:	HM, JvG, ASR, JV, SA
Non-expert feedback:	QH, AM
Editorial responsibility:	HM, SA, KAF

## References

- Abraham, M.J., Murtola, T., Schulz, R., Páll, S. et al (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1-2**.
- Adcock, S.A. and McCammon, J.A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.*, **106**(5), 1589–1615.
- Alder, B.J. and Wainwright, T.E. (1957). Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, **27**(5), 1208.
- Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C. (1993). Essential Dynamics of Proteins. *PROTEINS: Struct. Funct. Gen.*, **17**, 412–425.
- Berendsen, H.J.C. (2007). *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge Univ. Press, Leiden.
- Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**(8), 3684.
- Bowers, K.J., Chow, E., Xu, H., Dror, R.O. et al (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*, New York, NY. IEEE.
- Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J. et al (1983). CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculation. *J. Comput. Chem.*, **4**, 187–217.
- Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L. et al (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, **30**(10).

- Case, D., Babin, V., Berryman, J., Betz, R. et al (2014). AMBER 14. University of California, San Francisco.
- Cheng, A. and Merz, K.M. (1996). Application of the Nosé-Hoover Chain Algorithm to the Study of Protein Dynamics. *The Journal of Physical Chemistry*, **100**(5), 1927–1937.
- Cuendet, M.A., Zoete, V. and Michielin, O. (2011). How T cell receptors interact with peptide-MHCs: a multiple steered molecular dynamics study. *Proteins*, **79**(11), 3007–3024.
- Daura, X., Jaun, B., Seebach, D., van Gunsteren, W.F. and Mark, A.E. (1998). Reversible Peptide Folding in Solution by Molecular Dynamics Simulation. *J. Mol. Biol.*, **280**(5), 925–932.
- Daura, X., Gademann, K., Jaun, B., Seebach, D. et al (1999). Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Intl. Ed.*, **38**(1/2), 236–240.
- Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H. and Shaw, D.E. (2012). Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics*, **41**(1), 429–452.
- Duan, Y. and Kollman, P.A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**, 740–744.
- Feeenstra, K., Hess, B. and Berendsen, H. (1999). Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of Computational Chemistry*, **20**(8).
- Feeenstra, K.A., Hofstetter, K., Bosch, R., Schmid, A. et al (2006). Enantioselective substrate binding in a monooxygenase protein model by molecular dynamics and docking. *Biophysical journal*, **91**(9), 3206–16.
- Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Pr, San Diego, second edition.
- Gowers, R., Linke, M., Barnoud, J., Reddy, T. et al (2016). MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, number Scipy, pages 98–105.
- Grossman, J.P., Kuskin, J.S., Bank, J.A., Theobald, M. et al (2013). Hardware Support for Fine-grained Event-driven Computation in Anton 2. In *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '13, pages 549–560, New York, NY, USA. ACM.
- Halgren, T.A. and Damm, W. (2001). Polarizable force fields. *Current Opinion in Structural Biology*, **11**(2).
- Hess, B., Kutzner, C., van der Spoel, D. and Lindahl, E. (2008). Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**(3), 435–447.
- Hopkins, C.W., Le Grand, S., Walker, R.C. and Roitberg, A.E. (2015). Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, **11**(4), 1864–1874.
- Kaminski, G.A., Friesner, R.A., Tirado-Rives, J. and Jorgensen, W.L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B*, **105**(28), 6474–6487.
- Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, **102**(19), 6679–6685.
- Kästner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **1**(6), 932–942.
- Krieger, E. and Vriend, G. (2014). YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, **30**(20), 2981–2982.
- Krieger, E. and Vriend, G. (2015). New ways to boost molecular dynamics simulations. *Journal of Computational Chemistry*, **36**(13).
- Kumar, S. and Li, M.S. (2010). Biomolecules under mechanical force. *Phys. Rep.*, **486**, 1–74.
- Leach, A. (2001). *Molecular Modelling: Principles and Applications*. Pearson.
- Lemkul, J.A. and Bevan, D.R. (2013). Aggregation of Alzheimer's amyloid  $\beta$ -peptide in biological membranes: a molecular dynamics study. *Biochemistry*, **52**(29), 4971–4980.
- Li, P. and Merz, K.M. (2017). Metal Ion Modeling Using Classical Mechanics. *Chemical Reviews*, **117**(3), 1564–1686.
- Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P. et al (2012). Systematic validation of protein force fields against experimental data. *PLoS ONE*, **7**(2), e32131.
- Lingenheil, M., Denschlag, R., Reichold, R. and Tavan, P. (2008). The “Hot-Solvent/Cold-Solute” Problem Revisited. *Journal of Chemical Theory and Computation*, **4**(8), 1293–1306.
- Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P. and de Vries, A.H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, **111**(27), 7812–7824.
- May, A., Pool, R., van Dijk, E., Bijlard, J. et al (2014). Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics (Oxford, England)*, **30**(3), 326–334.
- McCammon, J.A., Gelin, B.R., Karplus, M. and Wolynes, P.G. (1976). Hinge bending mode in

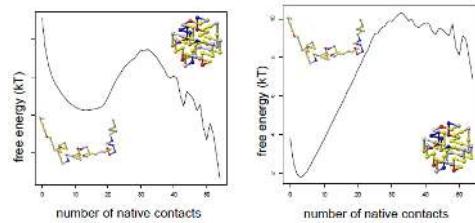
- lysozyme. *Nature*, **262**, 325–326.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.
- Michaud-Agrawal, N., Denning, E.J., Woolf, T.B. and Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, **32**(10), 2319–2327.
- Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G. et al (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput*, **4**(5), 819–834.
- Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, **38**(8), 4.
- Mor, A., Ziv, G. and Levy, Y. (2008). Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *Journal of Computational Chemistry*, **29**(12), 1992–1998.
- Morrell, W.E. and Hildebrand, J.H. (1936). The Distribution of Molecules in a Model Liquid. *The Journal of Chemical Physics*, **4**(3), 224–227.
- Onufriev, A.V. and Case, D.A. (2019). Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics*, **48**(1), 275–296.
- Oostenbrink, C., Villa, A., Mark, A.E. and van Gunsteren, W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem*, **25**(13), 1656–1676.
- Phillips, J.C., Braun, R., Wang, W., Gumbart, J. et al (2005). Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**(16), 1781–1802.
- Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E. et al (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, **153**(4).
- Ponder, J.W. and Case, D.A. (2003). Force fields for protein simulations. *Adv. Protein Chem.*, **66**, 27–85.
- Pool, R., Heringa, J., Hoefling, M., Schulz, R. et al (2012). Enabling grand-canonical Monte Carlo: Extending the flexibility of GROMACS through the grompy python interface module. *Journal of Computational Chemistry*, **33**(12).
- Pronk, S., Pall, S., Schulz, R., Larsson, P. et al (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**(7), 845–854.
- Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.*, **136**(2A), A405–A411.
- Rapaport, D.C. (2004). The Art of Molecular Dynamics Simulation. *The Art of Molecular Dynamics Simulation*.
- Raval, A., Piana, S., Eastwood, M.P., Dror, R.O. and Shaw, D.E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, **80**(8), 2071–2079.
- Reif, M.M., Hünenberger, P.H. and Oostenbrink, C. (2012). New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *Journal of Chemical Theory and Computation*, **8**(10).
- Roux, B. and Simonson, T. (1999). Implicit solvent models. *Biophysical Chemistry*, **78**(1-2).
- Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S. et al (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, **15**, 341–346.
- Shivakumar, D., Harder, E., Damm, W., Friesner, R.A. and Sherman, W. (2012). Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *Journal of Chemical Theory and Computation*, **8**(8).
- Singh, N. and Li, W. (2019). Recent advances in coarse-grained models for biomolecules and their applications. *International Journal of Molecular Sciences*, **20**(15).
- Souza, P.C.T., Alessandri, R., Barnoud, J., Thallmair, S. et al (2021). Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods*, **18**(4).
- Stillinger, F.H. and Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, **60**(4), 1545–57.
- Tian, C., Kasavajhala, K., Belfon, K.A.A., Ragquette, L. et al (2020). ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, **16**(1).
- van Aalten, D.M.F., de Groot, B.L., Berendsen, H.J.C., Findlay, J.B.C. and Amadei, A. (1997). A Comparison of Techniques for Calculating Protein Essential Dynamics. *J. Comput. Chem.*, **18**(2), 169–181.
- van Gunsteren, W.F., Berendsen, H.J., Hermans, J., Hol, W.G. and Postma, J.P. (1983). Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proceedings of the National Academy of Sciences*, **80**(14), 4315–4319.
- van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H. et al (1996). *Biomolecular Simulation: The GROMOS96 Manual and User Guide*. Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland.

- van Gunsteren, W.F., Burgi, R., Peter, C. and Daura, X. (2001). The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State. *Angew. Chem. Int. Ed. Engl.*, **40**(2), 351–355.
- van Gunsteren, W.F., Bakowies, D., Baron, R., Chandrasekhar, I. et al (2006). Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed. Engl.*, **45**(25), 4064–4092.
- Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, **159**, 98–103.
- Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, **103**(2), 227–249.
- Wood, W.W. and Parker, F.R. (1957). Monte Carlo Equation of State of Molecules Interacting with the Lennard-Jones Potential. I. A Supercritical Isotherm at about Twice the Critical Temperature. *The Journal of Chemical Physics*, **27**(3), 720–733.
- Yang, L., Tan, C.h., Hsieh, M.J., Wang, J. et al (2006). New-Generation Amber United-Atom Force Field. *The Journal of Physical Chemistry B*, **110**(26).
- Yang, Y.I., Shao, Q., Zhang, J., Yang, L. and Gao, Y.Q. (2019). Enhanced sampling in molecular dynamics. **151**(7).

# Chapter 15

## Monte Carlo for Protein Structures

Juami H. M. van Gils\*  Maurits Dijkstra  Halima Mouhib   
Arriën Symon Rauh  Jocelyne Vreede   
K. Anton Feenstra\*  Sanne Abeln\* 



\* editorial responsibility

## 1 Introduction

In the previous chapter, Chapter 14, we have considered protein simulations from a **dynamical point of view**, using **Newton's laws**. In this Chapter, we first take a step back and return to the bare minimum needed to simulate proteins, and show that proteins may be simulated in a **more simple fashion**, using the **partition function** directly, as given in Chapter 13. We will assume basic knowledge on thermodynamics and statistical mechanics, as introduced there as well. It is particularly important to understand the relation between **free energy and probability**, in order to understand this chapter. This means we do not have to calculate explicit forces, velocities, moments and do not even consider time explicitly. Instead, we heavily rely on the fact that for most systems we will want to simulate, the system is in a **dynamic equilibrium**; and that we want to find the **most stable states** in such systems by determining the **relative stabilities** between those states.

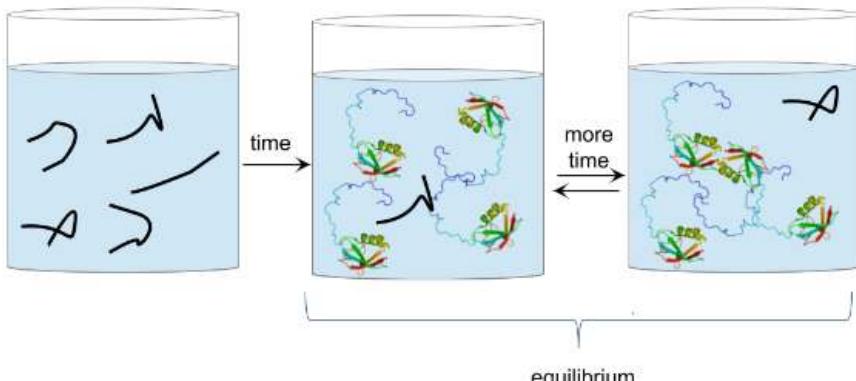
## 2 Proteins in equilibrium

Firstly, we will briefly revise our conceptual understanding of a dynamic equilibrium. In **equilibrium**, for each state in the system the number of particles moving into that state is equal to the number of particles moving from that state to a different state.

Proteins in solution are dynamic systems, see Figure 15.1. Proteins constantly unfold and refold. Once in equilibrium, the number of proteins moving from a folded to an unfolded state equals the number of proteins moving from an unfolded to a folded state, such that the fraction of folded and unfolded proteins will remain constant over time. We will see later in this chapter, that this also needs to **hold for simulations in equilibrium**; this concept is called '**detailed balance**' (see Panel "Detailed balance" later in this chapter for more detail).

In this Chapter we will consider two systems: *i)* **particles freely moving in a box**; see Figure 13.2 in Chapter 13, and *ii)* **a simplified protein chain freely moving**; see Figure 12.2 in Chapter 12.

In the first system, we consider the two **macrostates**: the colour separated and mixed states; here the positions of the particles define the specific **configurations or microstates**. In the second system we consider the folded and unfolded macrostate; here the positions of the particles (residues) in the **chain** define the specific configurations or microstates; for definitions of micro- and macrostates see Chapter 13



*Figure 15.1: Proteins in equilibrium.* Proteins are non-static entities. Over time, proteins constantly unfold and refold. When the proper folding of proteins is experimentally determined by for example by measuring the activity of the protein, the average behaviour over the ensemble of protein configurations in solution is determined rather than the behaviour of individual molecules. An equilibrium simulation of a single particle over time is equivalent to measurements on an ensemble of multiple proteins in equilibrium - provided that they do not interact.

### 3 The Purpose of Simulations

Before we go into the technical details of simulations, we first reconsider what we typically want to learn from them. In Monte Carlo and Molecular Dynamics simulations, the main goal is to understand what the most stable state of the system is under certain conditions. For example, one can determine the stability of a certain fold, calculate the interaction strength of protein-protein or protein-ligand interactions, or the phase of the particles in the system under different conditions. If these interaction strengths are known, one can for example calculate the concentration needed for two proteins to start binding at a given temperature. In addition, determining the transition states between the most stable states in a system can recover mechanisms of function, when for example considering a binding or a folding process. The most stable state of a system is defined as the state with the highest probability and the lowest free energy. As discussed in Chapter 13, the free energy  $F_A$  and probability  $p_A$  of a macrostate  $A$  are related as:

$$F_A = -k_B T \ln(p_A) \quad (1)$$

where  $k_B$  is the Boltzmann constant,  $T$  is the temperature in Kelvin and  $p_A$  is the probability of state  $A$ .

Moreover, as we previously discussed that the difference in free energy between two states calculated over a statistical ensemble approximates the difference in Gibbs free energy (i.e.,  $\Delta F_{A,B} \approx \Delta G_{A,B}$ ), we also have:

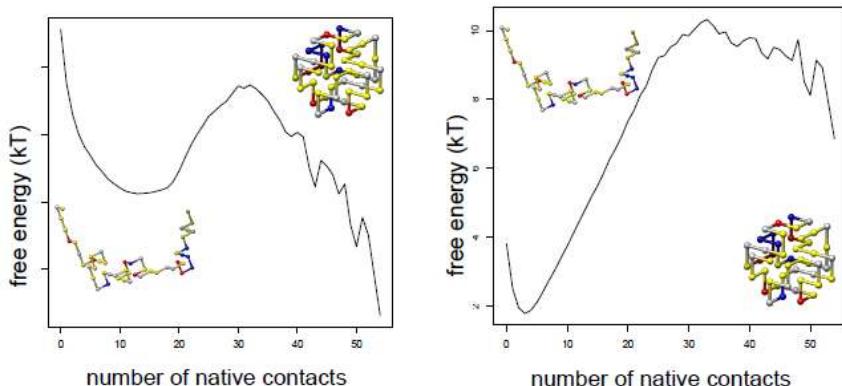


Figure 15.2: Free energy of a protein in a 3D cubic lattice model of a protein at high and low temperature. Left: at low temperature, the system with the **largest number of native contacts** is the most stable. The low enthalpy has the largest influence on the free energy of the system, and therefore the configuration with the largest number of favourable interactions is the most stable. Right: at high temperature, the state with the largest entropy has the lowest free energy and is therefore more stable than the native state.

$$\Delta G_{A,B} \approx -k_B T \ln \frac{p_A}{p_B} \quad (2)$$

This means that once we have **sampling the statistical ensemble of configurations appropriately**, we can make a good estimate of the relative free energy between states.

Figure 15.2 illustrates the difference in free energy between the folded and unfolded state at two different conditions. It is this relative free energy that determines the stability of the respective states.

Note that, with any simulation technique, it is only possible to calculate **relative free energies**. If we wanted to get absolute free energies - we would need to calculate the **full partition function**, which is (computationally) **intractable**. Nevertheless, absolute free energies may be estimated from reference points for which the full partition function can be calculated. Such calculations go beyond the scope of this book, but are described in Frenkel and Smit (2002).

## 4 Comparison to experiments

Similar to simulations, relative free energies between well defined states can be obtained from **experiments**. Differences in enthalpy ( $\Delta H$ ) can also be **measured directly** between states (e.g., Kardos *et al.*, 2004).

With some experiments, we can obtain information about the configurational ensemble of proteins in solution. For example, Hydrogen-Deuterium

exchange experiments can reveal the fraction of surface exposed residues of an ensemble in solution (Englander and Mayne, 2017). In simulations, we can estimate such observables on the macrostate through ensemble averages, by averaging over the microstates:

$$\langle a \rangle = \frac{\sum_i a_i p_i}{\sum_i p_i} \quad (3)$$

From a simulation, we can simply calculate an ensemble average  $\langle a \rangle$ , by averaging a certain property  $a$  over all the sampled microstates (or configurations)  $i$ . See Chapter 13 Section 6.2 for a more detailed explanation of ensemble averages.

## 5 Monte Carlo Algorithm

The Monte Carlo algorithm can be used in simulations with a constant number of particles, volume and temperature, also referred to as NVT ensemble; see Chapter 13 Section 7. In the Metropolis Monte Carlo algorithm one can sample the partition function directly, which means we do not need to consider forces, velocities or time. What we do need in order to sample the partition function, is a way to obtain the potential energy of specific configurations.

### 5.1 Potential energies

We can calculate the potential energy  $E_i$  for a micro state  $i$ , if we consider all pairwise interactions between the particles:

$$E_i = \frac{1}{2} \sum_{k=0}^{k=N} \sum_{l=0}^{l=N} \epsilon_{(k,l)} C_{(k,l)} \quad (4)$$

Here  $\epsilon_{(k,l)}$  are the pairwise interaction energies between particles  $k$  and  $l$ , and  $C_{(k,l)}$  indicates if the two particles interact with each other, which would depend on the distance of the two particles.

We can also use continuous interaction potentials, such as the Lennard-Jones potential. In that case, the pairwise particle interaction energies ( $\epsilon_{(k,l)}$ ) also depend on the distances between particles as shown in Figure 14.4 in Chapter 13.

### 5.2 Sampling the partition function

As explained in Chapter 13, the partition function  $Z$  can be used to calculate the free energy and describe the state of the system (i.e the macrostate). From the Boltzmann distribution we have:

$$p_i = \frac{e^{-\frac{E_i}{k_B T}}}{Z} \quad (5)$$

where  $Z = \sum_i e^{-\frac{E_i}{k_B T}}$ .

If we know all the possible configurations (microstates) of the system, it is possible to calculate the absolute free energy landscape of the system from Equation 12 in Chapter 13. Note that for a continuous three-dimensional system ( $\mathbb{R}^3$ ) with a constant finite number of particles the partition function becomes an integral over the full three-dimensional space, rather than a sum over all possible configurations.

However, in a simulation, computation of the full partition function is intractable. Instead, we aim to sample those configurations (microstates) with the largest contribution to the total free energy; from Equation 5 we can see that the microstates with the **highest probabilities** are the microstates with **low energies**. However the contribution low energy microstate become smaller at **high temperatures**.

### 5.3 The Metropolis Monte Carlo algorithm

The Monte Carlo algorithm is a stochastic algorithm that only depends on the potential energy of the system. The temperature, volume and number of particles in the system are kept constant. Additionally, the algorithm assumes the system is in equilibrium.

The key idea in the Monte Carlo algorithm is to make sure the probabilities of the sampled (micro)states **follow the Boltzmann distribution**. This can be achieved in a simple manner: by generating **a random move**, and **consistent rule** - the **Boltzmann acceptance criterion**.

In the algorithm random moves are proposed to change the configuration of the system: randomly chosen particles are moved by a random, but typically small, displacement, as shown in Figure 15.3. Now we have two configurations, the ‘old’ configuration and the ‘new’ configuration. For both configurations we can calculate an **explicit potential energy**, using Equation 4. These energies of the microstates can be used to calculate the **Boltzmann factor**  $B$ :

$$B = e^{-\frac{E_{\text{new}} - E_{\text{old}}}{k_B T}} \quad (6)$$

where  $E_{\text{new}}$  is the energy of the new state,  $E_{\text{old}}$  is the energy of the previous state,  $k_B$  is the Boltzmann constant and  $T$  is the temperature.

When the new configuration has a **lower** energy than the old configuration, i.e.,  $E_{\text{new}} \leq E_{\text{old}}$  we always **accept** the move, note that in that case  $B > 1$ . If, on the other hand  $E_{\text{new}} > E_{\text{old}}$  we use the Boltzmann factor and a **random variable**  $r \in [0, 1]$  to **determine if the move will be accepted**: the move will only be accepted if  $r < B$ .

Similar to simulated annealing

**e^0 = 1**

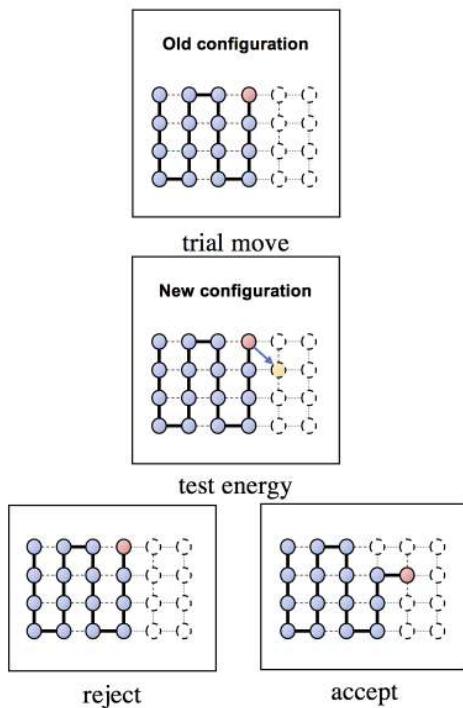


Figure 15.3: Trial move in a Monte Carlo simulation. Based on whether the change in energy of a random configurational change is favourable or not, it will be either accepted or rejected as the new state of the system. Unfavourable moves are accepted with a probability equal to the Boltzmann factor. Here a coarse-grained model of a protein on a 2D square lattice is shown to exemplify the algorithm.

Note that in the latter case, the system will actually get a **more unfavourable energy after the move**. At **high** temperatures, the Boltzmann factor will be **close to one** even if the energy difference between the old and new state is large; hence, at high temperatures the majority of moves will be accepted. This will lead to the enthalpic contribution becoming less dominant. This can be directly compared to the classical thermodynamics relation  $\Delta G = \Delta E - T\Delta S$ , which states that the entropy becomes more dominant at higher temperatures. The full MC algorithm is listed in Figure 15.4.

To obtain a correct sampling of the partition function, **sampling needs to be performed after every move**, regardless of whether it is accepted or rejected; this means that for a rejected move, we count (sample) the old configuration again (!). Note that this may make more intuitive sense if you consider a state that is already close to the free energy minimum (e.g., a folded state, and try to move away from this state (e.g., partially unfold the protein), which may be rejected in most trial moves. In this case, the **low free energy state** (e.g., folded state) will be **sampled very often** - but only if

```
1 # num_cycles: how many cycles of random sampling
2 # N: number of particles (or residues)
3 # V: volume
4 # T: the temperature
5 # C: initial configuration of the particles (protein)
6 def monte_carlo(num_cycles,N,V,T,C):
7     config_old = C
8     for x in range(num_cycles):
9         # pick a particle (residue) to displace
10        # randint() is random integer generator
11        x = randint(0, len(N)-1)
12        # move the chain by generating
13        # a new configuration for particle x
14        # note that the new configuration is generated
15        # within a constant volume (V)
16        config_new = generate_config(config_old,x,V)
17        # calculate the old and new interaction energies
18        # for particle x
19        E_new = Energy(config_new)
20        E_old = Energy(config_old)
21        # calculate Boltzmann factor , given kT
22        boltz = exp(-(E_new - E_old)/k*T)
23        # acceptance criterion:
24        acc = min(1.0, boltz))
25        # rand() gives random number between 0 and 1
26        # accept move if rand() is smaller than the
27        # acceptance criterion
28        if(rand() < acc):
29            # move is accepted
30            config_old = config_new
31            system_Energy += (E_new - E_old)
32        #end if
33        #sample at every step , to calculate p_i
34        sample(config_old)
35    # end for loop
36 # end Monte Carlo
```

Figure 15.4: Monte Carlo algorithm for molecular simulations in pseudo code Python style.

we also sample the old configuration after a rejected move.

From the simulation, the probability for a particular macrostate can be determined by calculating the fraction of configurations within the state, and those sampled outside of this state. Subsequently, the relative free energy of that state can be calculated using Equation 1.

As the simulation should be in equilibrium, in theory the starting state of the system should not matter. In practice, it is wise to check if there is indeed no flux during the simulation: if the simulation starts from a high free energy (unlikely) state, it may get stuck in a local minimum for a while, effectively not sampling the partition function evenly.

### Detailed balance

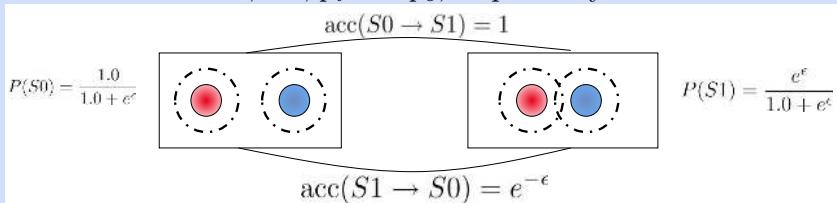
Detailed balance is a way of making sure equilibrium is kept in a Monte Carlo simulation. In other words, it ensures there is no net flux between states over time. Hence the number of accepted moves from a state  $S_1$  to state  $S_2$  needs to equal the number of accepted moves from the state  $S_2$  into that state  $S_1$ , for any two states  $S_1$  and  $S_2$  in the system. This can be expressed as follows:

$$N_{S_1} * P_{acc}(S_1, S_2) = N_{S_2} * P_{acc}(S_2, S_1) \quad (7)$$

Here  $N_{S_1}$  and  $N_{S_2}$  represent the number of times states A and B are visited, respectively, and  $P_{acc}(i, j)$  is the probability that the move from state  $i$  to state  $j$  is accepted. One can show that the Boltzmann acceptance criterion used in the Monte Carlo algorithm,

$$P_{acc}(S_1, S_2) = \min\left(e^{-\frac{E_j - E_i}{k_B T}}, 1\right)$$

Note that  $N_{S_1}$  and  $N_{S_2}$  can be replaced by the probabilities that the states are visited, i.e.,  $p_i$  and  $p_o$ , respectively.



Here, we will simply demonstrate that the Monte Carlo acceptance criterion satisfies detailed balance with the simple example shown above. We consider a system with two particles in solution and only two possible states: either they are separated and do not interact (left) or they are bound and have a favourable interaction, with an interaction energy  $-\epsilon$  (right). In this case we have two states:  $S_0$

(separated) and  $S1$  (bound), hence  $E_0 = 0$  and  $E_1 = -\epsilon$ . For simplicity, we can set  $k_B T = 1$ . Using the probabilities from Equation 5 , we get  $p_0 = e^0/(e^0 + e^\epsilon)$ ,  $p_1 = e^\epsilon/(e^0 + e^\epsilon)$ ,  $P_{\text{acc}}(S0 \rightarrow S1) = 1$  and  $P_{\text{acc}}(S1 \rightarrow S0) = e^{-\epsilon}$ . Substituting this into Equation 7 gives:

$$\frac{e^0}{e^0 + e^\epsilon} * 1 = \frac{e^\epsilon}{e^0 + e^\epsilon} * e^{-\epsilon} \quad (8)$$

Since  $e^0 = e^\epsilon * e^{-\epsilon} = 1$ , the left and right hand side of the equation are equal. Therefore, the system is in equilibrium and detailed balance is satisfied.

For Monte Carlo simulations it is essential that detailed balance is kept, else the results of the simulation will be non-physical as the partition function will not be sampled correctly. Note that there are many ways to break detailed balance, for example by not sampling after rejected moves.

## 6 Applications of Monte Carlo for proteins

### 6.1 A simple protein lattice model

Full-atomistic simulations are computationally very demanding; in fact so demanding that it is still computationally too expensive to simulate the folding of proteins or realistic size ( $\sim 100$  residues) that form fully hydrophobic cores, as explained at length in Chapter 14. Therefore, it is very useful to simplify such a system into a lattice model (Sali *et al.*, 1994; Coluzza *et al.*, 2003; Coluzza and Frenkel, 2004; Abeln and Frenkel, 2008, 2011; Abeln *et al.*, 2014; van Dijk *et al.*, 2016). The residues are placed onto a regular cubic-lattice, which means we have a discrete rather than a continuous three dimensional space. This greatly reduces the number of possible configurations for the protein chain. Nevertheless, for real size proteins the number of possible configurations is still computationally intractable, even on a discrete lattice.

Figure 15.5 shows an example of a 3D lattice model. Two residues are considered in contact when they are on neighbouring positions on the lattice but are not linked with a peptide bond. Using this criterion, all pairwise interactions can be determined using

$$C_{k,l} = \begin{cases} 1 & \text{if } k \text{ and } l \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The strength of the interactions are defined in the matrix in Figure 15.5. Now we can calculate the full potential energy over a specific configuration using Equation 4. The model can be simulated with a Monte Carlo algorithm as shown in Figure 15.4. To generate a new configuration, we should only

consider moves, that are feasible on the cubic lattice. A set of possible moves, that do not break the chain, on a cubic lattice, are shown in Figure 15.6.

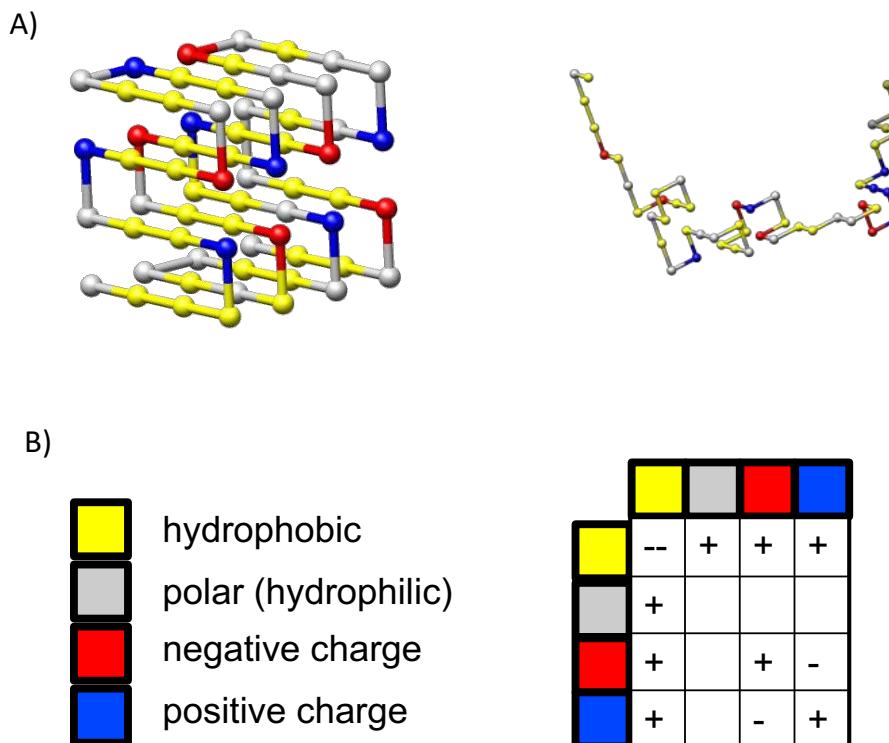


Figure 15.5: Simple 3D lattice model of a protein. A) a folded and unfolded configuration on the cubic lattice. The residues in the protein are placed on a 3D grid. Note that on the cubic lattice a residue has a maximum of four contacts with other residues - this is relatively similar for the average contact number of residues in real proteins. B) Schematic interaction energies. For simplicity, the amino acid pair potential is schematically shown in terms of interaction energies ( $\epsilon_{(k,l)}$ ) for Hydrophobic residues indicated in yellow, polar residues in grey, positively charged residues in red and negatively charged residues in blue.

Monte Carlo and lattice models can be used to determine the most stable states of a protein under different physiological conditions. Dijkstra *et al.* (2018), applied a Monte Carlo algorithm to a 3D protein lattice model to study the stability of a protein at different temperatures. Due to the simplified model, it becomes possible to obtain very extensive sampling of the conformational landscape, and allows details of the free energy landscape to be mapped out, as shown in Figure 15.7. The model describes three main states: the native folded state, molten globule state, and unfolded state. As shown in Figure 15.7, the native, molten globule and folded states are all present at lower temperatures, whereas at high temperatures only the

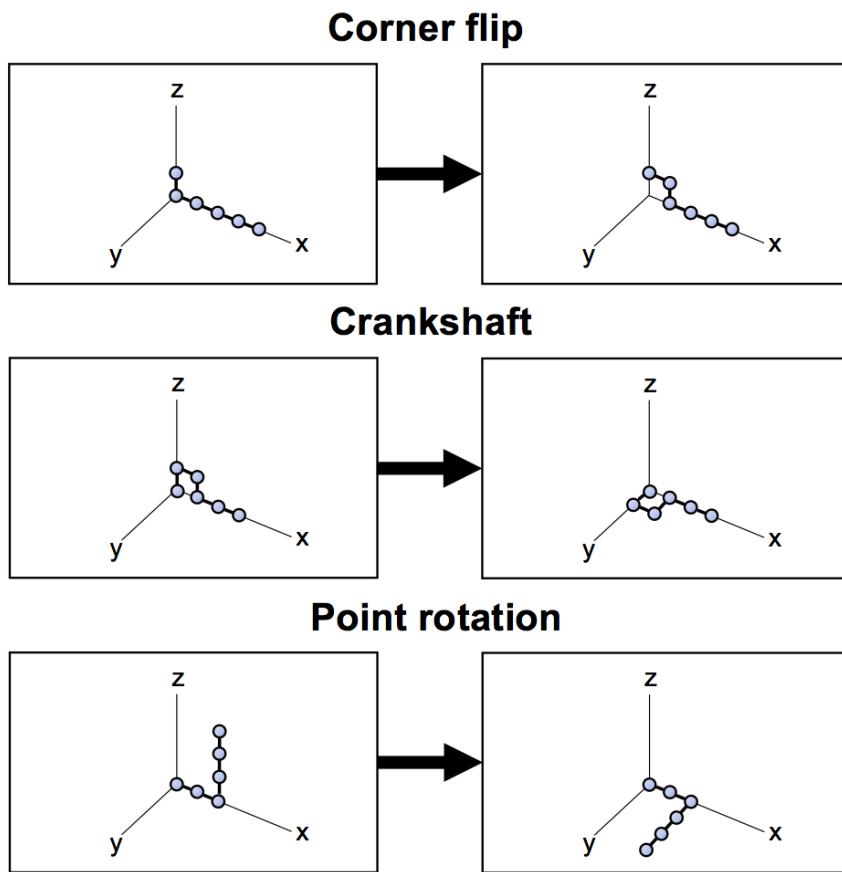


Figure 15.6: Moves on a cubic lattice. **Three different moves** on a cubic lattice are shown: the corner flip, crankshaft and point rotation. Each of the moves ensure the chain is not broken after the move. In order to keep detailed balance the reverse move needs to be equally probable as the forward move.

**unfolded** state has a low free energy.

Using such simulations, we can observe behaviour that is very similar to proteins in experimental settings: at high temperatures proteins unfold, due to the chain entropy. In this particular work, it was shown that proteins with the same fold, but with a different sequence, could have very different folding pathways and different intermediate molten-globule like states.

## 6.2 Other applications in bioinformatics

**Fragment based structure prediction** methods typically use Monte Carlo simulations to assemble **decoy structures** from the **structural fragments** (Song *et al.*, 2013), see also Chapter 6. Here, Monte Carlo sampling is used as a **search and optimisation** technique. The simulation starts at a medium to high temperature, which is decreased step-wise throughout the procedure

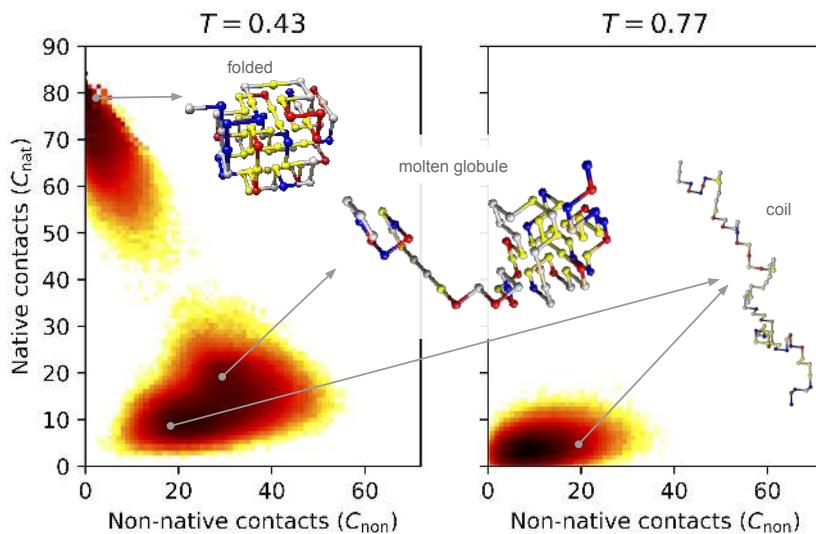


Figure 15.7: Free energy landscape as a function of the number of native and non-native contacts in a lattice model, with the free energy values shown as heatmap colors (dark red is very low free energy; white is high free energy). At a high number of native contacts, the protein is in its native folded state (top left in the plots). At intermediate values of native and non-native contacts, the protein is in a molten globule state. At very low numbers of native and non-native contacts, the protein is an unfolded, coil-like state. The figure shows that at a low temperature (left), the free energy is low for the folded state, the molten-globule state and the unfolded state. At even lower temperatures (not shown here), both the molten globule state and the unfolded state become unstable. At high temperature (right) the free energy is lowest when there are very few native and non-native contacts in the protein, indicating that the unfolded state is the most stable.

until  $T = 0$  and an energy minimum is reached. This process is called ‘simulated annealing’.

Simulated annealing is also used in homology model building by MOD-ELLER (Sali and Blundell, 1993), and in proposing moves via Molecular Dynamics. Here, the goal is to optimise a configuration that adheres to structural constraints from a template structure, see also Chapter 7.

It is important to note that such optimisation procedures are fundamentally different from molecular simulation approaches that try to sample the partition function. In simulated annealing only the (potential) energy is minimised, and not the free energy. In other words, entropy is not considered in the simulated annealing derived predictions. Moreover, typically non-physical energies are included in the energy function, such as distant constraints on specific residues. It is important to realise that we cannot use such optimisation techniques to consider folding or binding mechanisms.

### Hybrid MC & MD simulations

Proteins are very long molecules (polymers, or polypeptides). This means that any moves along the chain are generally correlated: neighbouring atoms in the chain cannot move independently from each other. This means that Monte Carlo moves on **single atoms** or residues – in case of coarse grained models – can be **very inefficient**. One way of overcoming this is to generate **collective moves**; in structure prediction the fragment based approach of Rosetta (Song *et al.*, 2013) is very efficient. For molecular simulations, often a hybrid approach gives extremely efficient sampling (Woo *et al.*, 2004; Pool *et al.*, 2012; Yang *et al.*, 2016): here, the smaller moves are implemented as **a series of MD steps**. These **trajectories** may then be rejected or accepted according to the rules based on the **Boltzmann factor**, making the higher level moves stochastic. The advantage of such an approach is that a multitude of **enhanced sampling** techniques can easily be applied within a **high level MC simulation**, using **low level MD moves** and a **force field parametrised for MD**. In such hybrid simulations, time development and (hydro)dynamics are **not conserved**.

## 7 Enhanced sampling techniques

As explained in previous sections, the **relative free energy** of a state can be calculated from the fraction of **time spent** in that state during a simulation. **Low free energy states** correspond to a **high probability of sampling**. This means that during a simulation, mainly the most stable states are sampled. On the other hand, sampling of high energy states is much more difficult: in severe cases, there may be no sampling of such states all together. This is particularly troublesome if these higher energy states lie in between two stable states, since such states form a ‘**barrier**’ between two stable states. An example of this was already shown in Figure 12.3. In order to calculate the relative free energy of the two stable states, it is **essential to also sample the path connecting them**. There are different tricks that can be applied to improve sampling in these regions and obtain a free energy landscape over the entire region of an order parameter.

Here, we will discuss two methods for enhanced sampling: Umbrella Sampling and Replica Exchange/Parallel Tempering. Both methods can be applied within **MD simulations as well as MC**, but are more easily implemented in **MC**. Moreover, the exchange steps in Replica Exchange are essentially **Monte Carlo moves**.

## 7.1 Umbrella Sampling in MC

Umbrella Sampling is one of the more simple enhanced sampling techniques. In Umbrella sampling, a value of the **order parameter** (e.g., the distance between two interacting proteins) is **chosen around which one wants to sample**.

In a Monte Carlo simulation this is extremely easy to implement. The only thing we need is a good order parameter. If for an order parameter  $x$  we want to sample a barrier region between  $a$  and  $b$ , we need to ensure that the **path** sampled by the MC algorithm rejects any steps going to a microstate where  $x < a$  or  $x > b$ . Remember that the (sampling) probability of a state has a direct relation with the **free energy** of that state: from Equation 10 in Chapter 13 we can derive  $p_A = e^{F_A/k_B T}$ , where  **$F_A$  is the free energy relative to the other sampled states**. Now, we can easily understand that the sampling probability of a state will go up, if the system is not allowed to visit the low free energy states of the system. In other words, if we choose the interval between  $a$  and  $b$  to be small enough, such that sampling is focused on the high free energy states only, the **probability of sampling the barrier goes up**. Now we can split the entire free energy landscape in **multiple intervals**. For each interval, we can approximate a free energy curve, which can be stitched together in a final step. Generally, the **steeper** the slope of the free energy curve with respect to the order parameter, the **more intervals we need**. Once we have all the free energy curves for the intervals, we need to stitch them together, this can be done by **curve fitting**; this will work much better, if there is overlap between the intervals. For more details, please see Frenkel and Smit (2002)

### Umbrella sampling using quadratic potentials

In MD simulations, we cannot simply reject moves or add a "hard wall". Instead, an **artificial energy penalty** is added around a **selected point**, such that it becomes **very unfavourable** for a protein to **deviate far from this point**. This penalty is called the '**Umbrella potential**' ( $E_{\text{Umbrella}}$ ) and takes the form of a **quadratic equation**:

$$E_{\text{Umbrella}} = k_{\text{umbrella}} (d - d_0)^2 \quad (10)$$

where a **higher** value of  $k_{\text{umbrella}}$  indicates a steeper penalty for **deviating distance**  $d - d_0$  from the selected point  $d_0$ .

Now we can draw such umbrellas over the entire range of the order parameter of interest, as shown in Figure 15.8. The name 'Umbrella sampling' originates from the shape of the penalty curve.

Finally, to obtain the true free energy landscape from the different simulations, the **obtained energies** need to be **corrected for the umbrella potential**. This can be done using

$$\langle A \rangle = \frac{\langle \frac{A}{w} \rangle_w}{\langle \frac{1}{w} \rangle_w} \quad (11)$$

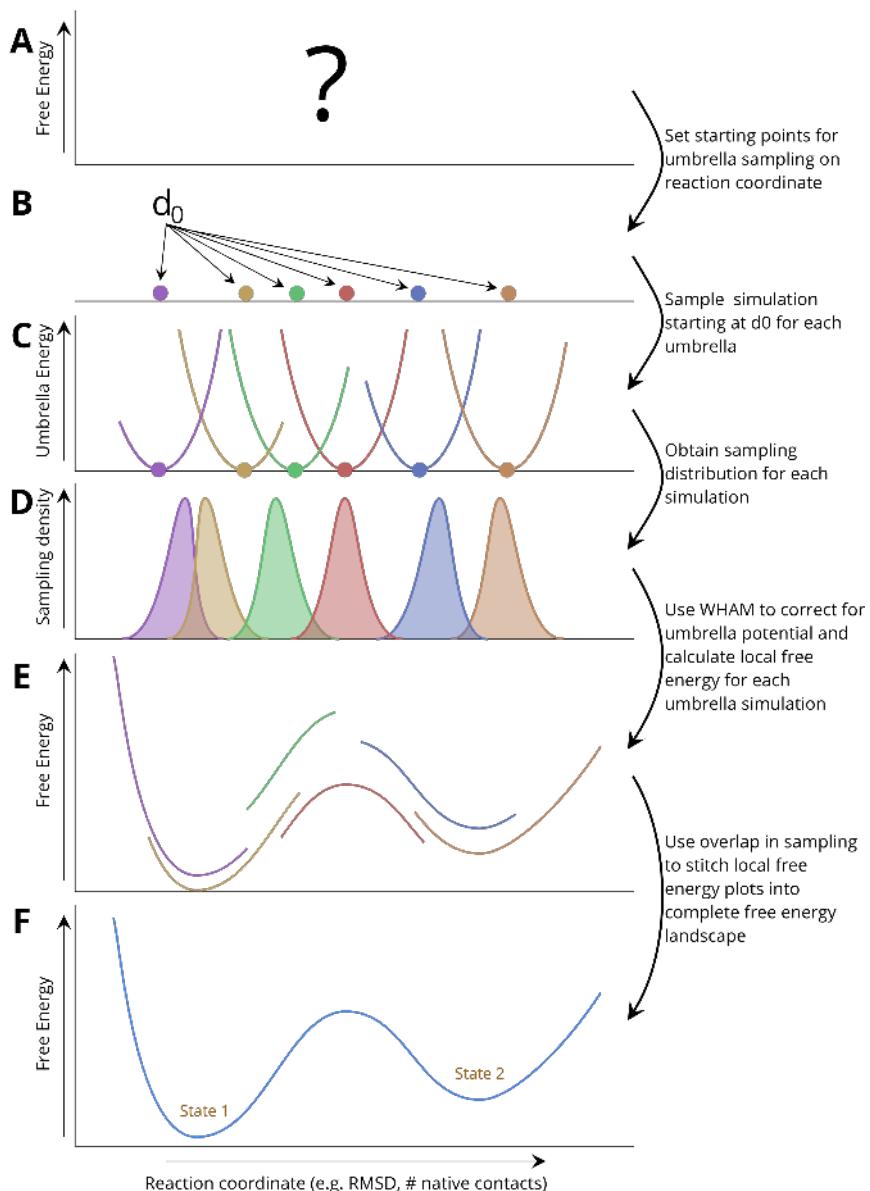


Figure 15.8: Schematic overview of an umbrella sampling for an MD simulation (see main text for further details). (A) Choice of the reaction coordinate (RC). (B) Apply umbrella potentials on selected values of the RC.  $d_0$  is the minimum of the umbrella in terms of the RC (C) Individual sampling around selected coordinates. (D) Density of sampling along the RC for each simulation. (E) Using weighted histogram analysis method (WHAM). (F) Joining the local free energy landscapes into a complete free energy landscape around the RC. Note that the sampling overlap is essential to create the final free energy landscape.

where  $A$  is the property of interest and  $w$  the weights of the sampling. The correctness of the final fit, stitching the intervals together to obtain a free energy curve can be improved using the WHAM method (Grossfield, 2003).

### Umbrella sampling procedure

The umbrella sampling procedure can be summarised as follows (Figure 15.8):

- A First, a reaction coordinate needs to be defined and an estimate of a range of values of this reaction coordinates that captures the relevant protein dynamics needs to be made.
- B Subsequently, multiple points within this range of the reaction coordinate are chosen to initiate the simulations. On each of these starting points, an umbrella potential is applied that adds an energy penalty to the simulation whenever the value of the reaction coordinate deviates from the starting point. The penalty is zero at the starting point, and increases quadratically as the distance from this value of the reaction coordinate increases ( $E_{umbrella} = k_{umbrella} (d - d_0)^2$ ), though other functions for the energy penalty may be chosen as well.
- C While running the simulations, at each point the value of the reaction coordinate and corresponding umbrella energy is sampled.
- D After the simulations are completed, the density of sampling along the reaction coordinate is calculated for each simulation.
- E Using weighted histogram analysis method (WHAM), the free energy profile can be corrected for the added umbrella potential and a local free energy landscape can be created.
- F Using the overlap in the regions of the reaction coordinate that were sampled between simulations, the local free energy landscapes can be stitched together into a complete free energy diagram of the sampled region of the reaction coordinate. Note that the sampling overlap between simulations is necessary to be able to create the final free energy landscape. If overlap is insufficient or lacking in any area, additional simulations need to be run initiated in this area to obtain a higher sampling density.

### Replica Exchange or Parallel tempering

Parallel tempering, also known as temperature replica exchange, is another enhanced sampling technique. The key idea is that some transitions may be more easily sampled at different, typically higher,

temperatures than the temperature of interest.

This approach consists of letting a number simulation boxes run simultaneously, while each box visits different temperatures during the parallel tempering procedure. These simulations are referred to as **replicas**, that can run in parallel. At fixed **time intervals (MD)** or **number of steps (MC)**, attempts are made to **exchange temperatures** between the different simulation boxes. Attempting to exchange temperatures between replicate simulations follows a **Monte Carlo** procedure, which is best described as performing a **Monte Carlo move in temperature space**.

With this Monte Carlo move, we need to ensure that detailed balance is observed, such that we have equal probabilities for the forward and backward swaps. It can be shown that the following rule for accepting moves, indeed **keeps detailed balance**.

$$P_{acc}(S1 \rightarrow S2) = \min(1, e^{(\beta_1 - \beta_2)(E_1 - E_2)}) \quad (12)$$

Here the variable  $\beta_i = \frac{1}{kT_i}$ . Note that  $\beta$  is often used instead of  $T$ , to make manipulation of equations in thermodynamics easier.  $E_i$  are the potential energies of the **states to be swapped**. A formal proof to show this acceptance rule adheres to detailed balance, which can be found in more details in (Frenkel and Smit, 2002).

A little care needs to be taken, how the temperatures of the different simulation replicas are chosen. It is important that the temperatures are **swapped sufficiently**. As a rule of thumb, **one accepted exchange out of three trials** is considered reasonable. A replica exchange procedure can be considered to be **finished** if **all replica boxes have visited all temperatures several times**. Then, the system has **heated up and cooled down** several times. If swaps between specific temperatures do not occur during the procedure, this suggests that these temperature may **lie close to a transition point**, and typically the interval between temperatures need to be made smaller, to allow for sufficient sampling.

## 8 Monte Carlo vs. Molecular Dynamics

Now we have considered two simulation protocols, Molecular Dynamics (MD) and Monte Carlo (MC), both can be used to study the same properties of a system, namely the **stability of states** and the **transitions between them**. Using either technique, the free energy landscape can be calculated along a **chosen order parameter** (or multiple order parameters). However, in practice it is **not possible** to sample a complete folding pathway of a real-size protein in a full-atomistic model with either of the two techniques. Thus, we cannot exhaustively cover the whole free energy landscape, and we typically

refer to the simulation process as **sampling states** in the free energy landscape. Both techniques should **maintain detailed balance**, and sample the **Boltzmann distribution**. A short summary of main differences is provided in Table 1.

MC is an intrinsically stochastic method that depends on **random moves** to determine a simulation path. To calculate the next state of a system, only the **energy difference** between the old and the new state needs to be known. Any forces, velocities, momenta, and time are ignored in MC. This large simplification of the system makes MC simulations **much faster** to execute and much easier to code than MD.

MC simulations natively sample an **NVT** ensemble, while MD on the other hand natively samples an **NVE** ensemble, see also Frenkel and Smit (2002) for more details.

MD is theoretically a **deterministic** simulation, however, in practice, due to limits in computational precision, and the use of a **thermostat and/or barostat**, MD is it is **not deterministic**.

Most biological systems are naturally exposed to an environment with constant temperature, i.e., they exist within larger systems with constant exchange of heat between the system and its surroundings, leading to a constant temperature of the considered system. Therefore, **NVT** is often a more natural choice. This means that for most practical cases we will need a **thermostat** in MD simulations; this (re)tunes the **velocities of particles** in such a way that the temperature is kept constant throughout the simulation.

Since MD captures dynamics explicitly, it is possible to include effects such as **hydrodynamics** (e.g., movements of **water** in direct vicinity to a moving part of the protein). In MC, because the forces, speeds, and momenta of all the particles are not known, **collective moves**, incorporating multiple particles, often need to be **added explicitly** to speed up the simulation.

Lastly, due to the simplicity of the MC algorithm, it is much more straightforward to implement enhanced sampling techniques (see section below) in an MC simulation. If we want to consider large systems, such as proteins that (re)fold, **enhanced sampling** techniques are essential to allow even sampling within a range of the order parameter during the simulation.

## 9 Key points

- When a system is in **equilibrium** we do not have to simulate **velocities and time explicitly** in order to obtain relative free energies
- Monte Carlo samples the **partition function** of systems in equilibrium
- Monte Carlo is a **stochastic** sampling method
- It is **straightforward** to use enhanced sampling techniques in the Monte Carlo framework
- Molecular simulations need to keep **detailed balance** in order to adhere

	MC	MD
algorithm	stochastic	deterministic
native ensemble	NVT	NVE
advantages	easier to code easier to implement enhanced sampling	explicit dynamics time development
disadvantages	need collective moves for efficient sampling fewer simulation packages available	need integrable forces thermostat required for NVT

Table 1: Monte Carlo (MC) versus Molecular Dynamics (MD) simulations.

to statistical mechanics

- In structural Bioinformatics many ideas of molecular simulation are used, sometimes with shortcuts that mean the sampled ensembles may be non-physical.

## 10 Further reading

- Vlugt *et al.* (2008)
- Frenkel and Smit (2002)

## Author contributions

Wrote the text:	JvG, MD, HM, AF, JV, SA
Created figures:	JvG, MD, AR, AF, SA,
Review of current literature:	JvG, JV, AF, SA
Critical proofreading:	AF, JV, HM, SA
Non-expert feedback:	AR
Editorial responsibility:	JvG, SA

## References

- Abeln, S. and Frenkel, D. (2008). Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, **4**(12), e1000241.
- Abeln, S. and Frenkel, D. (2011). Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins. *Biophys. J.*, **100**(3), 693–700.
- Abeln, S., Vendruscolo, M., Dobson, C.M. and Frenkel, D. (2014). A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. *PLoS ONE*, **9**(1), e85185.
- Coluzza, I. and Frenkel, D. (2004). Designing specificity of protein-substrate interactions. *Phys Rev E Stat Nonlin Soft Matter Phys*, **70**(5 Pt 1), 51917.

- Coluzza, I., Muller, H.G. and Frenkel, D. (2003). Designing refoldable model molecules. *Phys Rev E Stat Nonlin Soft Matter Phys*, **68**(4 Pt 2), 46703.
- Dijkstra, M., Fokkink, W., Heringa, J., van Dijk, E. and Abeln, S. (2018). The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. *Molecular Physics*, **116**(21-22), 3173–3180.
- Englander, S.W. and Mayne, L. (2017). The case for defined protein folding pathways. *Proceedings of the National Academy of Sciences*, **114**(31), 8253–8258.
- Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Pr, San Diego, second edition.
- Grossfield, A. (2003). WHAM: the weighted histogram analysis method.
- Kardos, J., Yamamoto, K., Hasegawa, K., Naiki, H. and Goto, Y. (2004). Direct measurement of the thermodynamic parameters of amyloid formation by isothermal titration calorimetry. *Journal of Biological Chemistry*, **279**(53), 55308–55314.
- Pool, R., Heringa, J., Hoefling, M., Schulz, R. et al (2012). Enabling grand-canonical Monte Carlo: Extending the flexibility of GROMACS through the grompy python interface module. *Journal of Computational Chemistry*, **33**(12).
- Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**(3), 779–815.
- Sali, A., Shakhnovich, E. and Karplus, M. (1994). Kinetics of Protein Folding : A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.*, **235**(5), 1614–1638.
- Song, Y., Dimai, F., Wang, R.Y.R., Kim, D. et al (2013). High-resolution comparative modeling with RosettaCM. *Structure*, **21**(10), 1735–1742.
- van Dijk, E., Varilly, P., Knowles, T.P.J., Frenkel, D. and Abeln, S. (2016). Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, **116**(7), 078101.
- Vlugt, T.J., Eerden, J.P.v.d., Dijkstra, M., Smit, B. and Daan Frenkel (2008). *Introduction to Molecular Simulation and Statistical Thermodynamics*. Delft.
- Woo, H.J., Dinner, A.R. and Roux, B. (2004). Grand canonical Monte Carlo simulations of water in protein environments. *The Journal of Chemical Physics*, **121**(13), 6392–6400.
- Yang, K., Różycki, B., Cui, F., Shi, C. et al (2016). Sampling enrichment toward target structures using hybrid molecular dynamics-Monte Carlo simulations. *PLoS ONE*, **11**(5).





# References

- Abeln, S. and Deane, C. M. (2005). Fold usage on genomes and protein fold evolution. *Proteins*, **60**(4), 690–700.
- Abeln, S. and Frenkel, D. (2008). Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, **4**(12), e1000241.
- Abeln, S. and Frenkel, D. (2011). Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins. *Biophys. J.*, **100**(3), 693–700.
- Abeln, S., Molenaar, D., Feenstra, K. A., Hoefsloot, H. C. J., Teusink, B., and Heringa, J. (2013). Bioinformatics and Systems Biology: bridging the gap between heterogeneous student backgrounds. *Briefings in Bioinformatics*, **14**(5), 589–598.
- Abeln, S., Vendruscolo, M., Dobson, C. M., and Frenkel, D. (2014). A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. *PLoS ONE*, **9**(1), e85185.
- Abeln, S., Heringa, J., and Feenstra, K. A. (2017a). Introduction to Protein Structure Prediction. *arXiv*, **1712.00407**.
- Abeln, S., Heringa, J., and Feenstra, K. A. (2017b). Strategies for protein structure model generation. *arXiv*, **1712.00425**.
- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1-2**.
- Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.*, **106**(5), 1589–1615.
- Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, **76**(1), 1–7.
- Ahmad, S., Gromiha, M. M., and Sarai, A. (2003). RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**(14), 1849–1851.
- Alberts, B., editor (2015). *Molecular Biology of The Cell*. Garland Science, Taylor & Francis Group, LLC,.
- Alder, B. J. and Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, **27**(5), 1208.
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**(6), 450–461.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–402.
- Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993). Essential Dynamics of Proteins. *PROTEINS: Struct. Funct. Gen.*, **17**, 412–425.
- Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, **36**(SUPPL. 1), 419–425.
- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223–230.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25.
- Ashworth, J. and Baker, D. (2009). Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic acids research*, **37**(10), e73.
- Atkins, P. W. and De Paula, J. (2014). *Atkins' Physical chemistry*. Oxford University Press.
- Aurora, R. and Rosee, G. D. (1998). Helix capping. *Protein Science*, **7**(1), 21–38.
- Backert, L. and Kohlbacher, O. (2015). Immunoinformatics and Epitope Prediction in the Age of Genomic Medicine. *Genome Medicine*, **7**, 119.

- Baclayon, M., van Ulsen, P., Mouhib, H., Shabestari, M. H., Verzijden, T., Abeln, S., Roos, W. H., and Wuite, G. J. (2016). Mechanical Unfolding of an Autotransporter Passenger Protein Reveals the Secretion Starting Point and Processive Transport Intermediates. *ACS Nano*, page acsnano.5b07072.
- Baker, T. S., Olson, N. H., and Fuller, S. D. (1999). Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and molecular biology reviews : MMBR*, **63**(4), 862–922.
- Baldi, P. (2003). The Principled Design of Large-Scale Recursive Neural Network Architectures – DAG-RNNs and the Protein Structure Prediction Problem. *Journal of Machine Learning Research*, **4**, 575–602.
- Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**(11), 937–946.
- Baldwin, R. L. (2007). Energetics of Protein Folding. *Journal of Molecular Biology*, **371**(2), 283–301.
- Baldwin, R. L. and Rose, G. D. (2013). Molten globules, entropy-driven conformational change and protein folding. *Current Opinion in Structural Biology*, **23**(1), 4–10.
- Ballew, R. M., Sabelko, J., and Gruebele, M. (1996). Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(12), 5759–64.
- Banner, D., Bloomer, A., Petsko, G., Phillips, D., and Wilson, I. (1976). Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochemical and Biophysical Research Communications*, **72**(1), 146–155.
- Barber, J., Nield, J., Orlova, E. V., Morris, E. P., Gowen, B., and Heel, M. v. (2000). 3D map of the plant photosystem II supercomplex obtained by cryoelectron microscopy and single particle analysis. *Nature Structural Biology*, **7**(1), 44–47.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippe, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, **41**(D1), D991–D995.
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggiali, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shyptsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbott, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognoli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1), D158–D169.
- Baumeister, W. and Steven, A. C. (2000). Macromolecular electron microscopy in the era of structural genomics. *Trends in Biochemical Sciences*, **25**(12), 624–631.
- Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S., and Heringa, J. (2017). Multiple Sequence Alignment. In *Methods in Molecular Biology – Bioinformatics – Volume I: Data, Sequence Analysis, and Evolution*, pages 167–189. Humana Press, New York, NY.
- Benson, D. A. (2004). GenBank. *Nucleic Acids Research*, **33**(Database issue), D34–D38.
- Berendsen, H. J. C. (2007). *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge Univ. Press, Leiden.
- Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**(8), 3684.
- Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*. W H Freeman, New York.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.
- Bernardes, J. S. and Pedreira, C. E. (2013). A Review of Protein Function Prediction under Machine Learning Perspective. *Recent Patents on Biotechnology*, **7**(2), 122–141.
- Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, **45**(D1), D313–D319.
- Blundell, T. L. (1996). Structure-based drug design. *Nature*, **382**, 23–26.

- Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**(6111), 347–352.
- Bondarev, S. A., Antonets, K. S., Kajava, A. V., Nizhnikov, A. A., and Zhouravleva, G. A. (2018). Protein co-aggregation related to amyloids: Methods of investigation, diversity, and classification.
- Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting Function: From Genes to Genomes and Back. *Journal of Molecular Biology*, **283**(4), 707–725.
- Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., Klepeis, J. L., Kolossvary, I., Moraes, M. A., Sacerdoti, F. D., Salmon, J. K., Shan, Y., and Shaw, D. E. (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*, New York, NY. IEEE.
- Branden, C. and Tooze, J. (1998). *Introduction to protein structure*. garland publishing, New York.
- Brooks, B. R., Brucoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculation. *J. Comput. Chem.*, **4**, 187–217.
- Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, **30**(10).
- Brown, D. K. and Tastan Bishop, Ö. (2017). Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis: Analyzing Variation at the Protein Level. *Global Heart*, **12**(2), 151–161.
- Buchan, D. W. A. and Jones, D. T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research*, **47**(W1), W402–W407.
- Buell, A. K., Dobson, C. M., and Knowles, T. P. J. (2014). The physical chemistry of the amyloid phenomenon: thermodynamics and kinetics of filamentous protein aggregation. *Essays in biochemistry*, **56**, 11–39.
- Burrus, C. S. and Parks, T. W. (1985). *Convolution Algorithms*. Citeseer.
- Bystroff, C. and Krogh, A. (2008). Hidden Markov Models for Prediction of Protein Features. In *Protein Structure Prediction*, volume 413, pages 173–198. Humana Press.
- Capel, H., Weiler, R., Dijkstra, M., Vleugels, R., Bloem, P., and Feenstra, K. A. (2022). ProteinGLUE multi-task benchmark suite for self-supervised protein modeling. *Scientific Reports*, **12**(1), 16047.
- Carbon, S., Dietz, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Basu, S., Chisholm, R. L., Dodson, R. J., Fey, P., Thomas, P. D., Mi, H., Muruganujan, A., Huang, X., Poudel, S., Hu, J. C., Aleksander, S. A., McIntosh, B. K., Renfro, D. P., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Marygold, S. J., McQuilton, P., Ponting, L., Millburn, G. H., Rey, A. J., Stefancsik, R., Tweedie, S., Falls, K., Schroeder, A. J., Courtot, M., Osumi-Sutherland, D., Parkinson, H., Roncaglia, P., Lovering, R. C., Foulger, R. E., Huntley, R. P., Denny, P., Campbell, N. H., Kramarz, B., Patel, S., Buxton, J. L., Umrao, Z., Deng, A. T., Alrohaif, H., Mitchell, K., Ratnaraj, F., Omer, W., Rodríguez-López, M., C. Chibucus, M., Giglio, M., Nadendra, S., Duesbury, M. J., Koch, M., Meldal, B. H., Melidoni, A., Porras, P., Orchard, S., Shrivastava, A., Chang, H. Y., Finn, R. D., Fraser, M., Mitchell, A. L., Nuka, G., Potter, S., Rawlings, N. D., Richardson, L., Sangrador-Vegas, A., Young, S. Y., Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Harris, M. A., Hayles, J., Oliver, S. G., Rutherford, K., Wood, V., Bahler, J., Lock, A., De Pons, J., Dwinell, M., Shimoyama, M., Laulederkind, S., Hayman, G. T., Tutaj, M., Wang, S. J., D'Eustachio, P., Matthews, L., Balhoff, J. P., Balakrishnan, R., Binkley, G., Cherry, J. M., Costanzo, M. C., Engel, S. R., Miyasato, S. R., Nash, R. S., Simison, M., Skrzypek, M. S., Weng, S., Wong, E. D., Feuermann, M., Gaudet, P., Berardini, T. Z., Li, D., Muller, B., Reiser, L., Huala, E., Argasinska, J., Arighi, C., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Bely, B., Blatter, M. C., Bonilla, C., Bougueret, L., Boutet, E., Breuza, L., Bridge, A., Britto, R., Hye-A-Bye, H., Casals, C., Cibrán-Uhalte, E., Couder, E., Cusin, I., Duek-Roggli, P., Estreicher, A., Famiglietti, L., Gane, P., Garmiri, P., Georghiou, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Holmes, A., Hulo, C., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Mac-Dougall, A., Magrane, M., Martin, M. J., Masson, P., Natale, D. A., O'Donovan, C., Pedruzzi, I., Pichler, K., Poggioli, D., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Speretta, E., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Wu, C., Xenarios, I., Yeh, L. S., Chan, J., Gao, S., Howe, K., Kishore, R., Lee, R., Li, Y., Lomax, J., Muller, H. M., Raciti, D., Van Auken, K., Berriman, M., Stein, Paul Kersey, L., W. Sternberg, P., Howe, D., and Westerfield, M. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, **45**(D1), D331–D338.
- Caruana, R. (1997). Multitask Learning. *Machine Learning*, **28**(1), 41–75.
- Case, D., Babin, V., Berryman, J., Betz, R., Cai, Q., Cerutti, D., Cheatham III, T., Darden, T., Duke, R., Gohlke, H., Goetz, A., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, K., Paesani, F., Roe,

- D., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C., Smith, W., Swails, J., Walker, R., Wang, J., Wolf, R., Wu, X., and Kollman, P. (2014). AMBER 14. University of California, San Francisco.
- Chen, H. and Zhou, H.-X. (2005). Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins: Structure, Function, and Bioinformatics*, **61**(1), 21–35.
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., Richardson, D. C., and IUCr (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(1), 12–21.
- Cheng, A. and Merz, K. M. (1996). Application of the Nosé-Hoover Chain Algorithm to the Study of Protein Dynamics. *The Journal of Physical Chemistry*, **100**(5), 1927–1937.
- Cheng, C. W., Su, E. C. Y., Hwang, J. K., Sung, T. Y., and Hsu, W. L. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. In *BMC Bioinformatics*, volume 9.
- Cheng, G., Qian, B., Samudrala, R., and Baker, D. (2005a). Improvement in Protein Functional Site Prediction by Distinguishing Structural and Functional Constraints on Protein Family Evolution Using Computational Design. *Nucleic Acids Research*, **33**(18), 5861–5867.
- Cheng, J., Sweredoski, M. J., and Baldi, P. (2005b). Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery*, **11**(3), 213–222.
- Cheng, Q., Joung, I., and Lee, J. (2017). A Simple and Efficient Protein Structure Refinement Method. *Journal of Chemical Theory and Computation*, **13**(10), 5146–5162.
- Cheng, Y. (2018). Single-particle cryo-EM—How did it get here and where will it go. *Science*, **361**(6405), 876–880.
- Chiti, F. and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.
- Choi, Y. and Chan, A. P. (2015). PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics*, **31**(16), 2745–2747.
- Chothia, C. (2003). Evolution of the Protein Repertoire. *Science*, **300**(5626), 1701–1703.
- Chou, P. Y. and Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*.
- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nature Communications*, **4**(1), 2741.
- Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: Predicting protein dynamics from sequence. *Nucleic Acids Research*.
- Coluzza, I. and Frenkel, D. (2004). Designing specificity of protein-substrate interactions. *Phys Rev E Stat Nonlin Soft Matter Phys*, **70**(5 Pt 1), 51917.
- Coluzza, I., Muller, H. G., and Frenkel, D. (2003). Designing refoldable model molecules. *Phys Rev E Stat Nonlin Soft Matter Phys*, **68**(4 Pt 2), 46703.
- Corbeski, I., Dolinar, K., Wienk, H., Boelens, R., and van Ingen, H. (2018). DNA repair factor APLF acts as a H2A-H2B histone chaperone through binding its DNA interaction surface. *Nucleic Acids Research*, **46**(14), 7138–7152.
- Cowtan, K. (2003). Phase Problem in X-ray Crystallography, and Its Solution. In *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester.
- Cuendet, M. A., Zoete, V., and Michielin, O. (2011). How T cell receptors interact with peptide-MHCs: a multiple steered molecular dynamics study. *Proteins*, **79**(11), 3007–3024.
- Daura, X., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1998). Reversible Peptide Folding in Solution by Molecular Dynamics Simulation. *J. Mol. Biol.*, **280**(5), 925–932.
- Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int'l Ed.*, **38**(1/2), 236–240.
- Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, **45**(D1), D289–D295.
- De Beer, T. A. P., Berka, K., Thornton, J. M., and Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Research*, **42**(D1), 292–296.
- De Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**(4), 249–261.
- De Vries-van Leeuwen, I. J., da Costa Pereira, D., Flach, K. D., Piersma, S. R., Haase, C., Bier, D., Yalcin, Z., Michalides, R., Feenstra, K. A., Jiménez, C. R., de Greef, T. F. A., Brunsved, L., Ottmann, C., Zwart, W., and de Boer, A. H. (2013). Interaction of 14-3-3 proteins with the estrogen receptor alpha F domain provides a drug target interface. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(22), 8894–9.
- Delorenzi, M. and Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**(4), 617–625.
- Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemire, C., Vastrik, I., Wu, G., D'Eustachio, P.,

- Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, Ö., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovsky, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le Novère, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, **28**(9), 935–942.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL).
- Diamond, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *Journal of Molecular Biology*, **82**(3), 371–391.
- Dijkstra, M., Fokkink, W., Heringa, J., van Dijk, E., and Abeln, S. (2018). The characteristics of molten globule states and folding pathways strongly depend on the sequence of a protein. *Molecular Physics*, **116**(21-22), 3173–3180.
- Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**(13), i223–i231.
- Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**(6968), 884–890.
- Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012). Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics*, **41**(1), 429–452.
- Duan, Y. and Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**, 740–744.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- Earl, L. A., Falconieri, V., Milne, J. L., and Subramaniam, S. (2017). Cryo-EM: beyond the microscope. *Current Opinion in Structural Biology*, **46**, 71–78.
- Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3), 361–365.
- Edwards, H., Abeln, S., and Deane, C. M. (2013). Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. *PLoS computational biology*, **9**(11), e1003325.
- Egelman, E. H. (2017). Cryo-EM of bacterial pili and archaeal flagellar filaments. *Current Opinion in Structural Biology*, **46**, 31–37.
- El-Kebir, M., Soueidan, H., Hume, T., Beisser, D., Dittrich, M., Müller, T., Blin, G., Heringa, J., Nikolski, M., Wessels, L. F. A., and Klau, G. W. (2015). xHeinz: an algorithm for mining cross-species network modules under a flexible conservation model. *Bioinformatics*, **31**(19), 3147–3155.
- Ellis, R. J. (2001). Macromolecular crowding: Obvious but underappreciated.
- Englander, S. W. and Mayne, L. (2017). The case for defined protein folding pathways. *Proceedings of the National Academy of Sciences*, **114**(31), 8253–8258.
- Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.-C., and Deane, C. M. (2016). Progress and challenges in predicting protein interfaces. *Briefings in bioinformatics*, **17**(1), 117–131.
- Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., and Zhou, Y. (2012). SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry*, **33**(3), 259–267.
- Feeenstra, K., Hess, B., and Berendsen, H. (1999). Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of Computational Chemistry*, **20**(8).
- Feeenstra, K., Peter, C., Scheek, R., Van Gunsteren, W., and Mark, A. (2002). A comparison of methods for calculating NMR cross-relaxation rates (NOESY and ROESY intensities) in small peptides. *Journal of Biomolecular NMR*, **23**(3).
- Feeenstra, K. A., Hofstetter, K., Bosch, R., Schmid, A., Commandeur, J. N. M., and Vermeulen, N. P. E. (2006). Enantioselective substrate binding in a monooxygenase protein model by molecular dynamics and docking. *Biophysical journal*, **91**(9), 3206–16.
- Feeenstra, K. A., Abeln, S., Westerhuis, J. A., Brancos dos Santos, F., Molenaar, D., Teusink, B., Hoefsloot, H. C. J., Heringa, J., Brancos, F., Molenaar, D., Teusink, B., Hoefsloot, H. C. J., and Heringa, J. (2018). Training for translation between disciplines : a philosophy for life and data sciences curricula. *Bioinformatics*, **34**(13), 1–9.
- Feig, M. and Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations:

- What works and what does not? *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 282–292.
- Ferreira, L. G., dos Santos, R. N., Oliva, G., and Andricopulo, A. D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, **20**(7), 13384–13421.
- Fersht, A. (1999). *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. W.H. Freeman, New York.
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, **39**(SUPPL. 2), W29–W37.
- Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). Pfam: the protein families database. *Nucleic acids research*, **42**(Database issue), 222–30.
- Fischer, J. (2019). The Boltzmann Constant for the Definition and Realization of the Kelvin. *Annalen der Physik*, **531**(5), 1800304.
- Floden, E. W., Tommaso, P. D., Chatzou, M., Magis, C., Notredame, C., and Chang, J.-M. (2016). PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic acids research*, **44**(W1), 339–43.
- Frank, J. (2002). Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy. *Annual Review of Biophysics and Biomolecular Structure*, **31**(1), 303–319.
- Frasch, W. D., Bukhari, Z. A., and Yanagisawa, S. (2022). F1FO ATP synthase molecular motor mechanisms. *Frontiers in Microbiology*, **13**.
- French, T. C. and Hammes, G. G. (1969). [1] The temperature-jump method. *Methods in Enzymology*, **16**(C), 3–30.
- Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Pr, San Diego, second edition.
- Frishman, D. and Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. "Protein Engineering, Design and Selection", **9**(2), 133–142.
- Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**(8), 950–956.
- Garnier, J., Gibrat, J. F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology*, **266**, 540–553.
- George, R. A. and Heringa, J. (2002a). Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins: Structure, Function, and Genetics*, **48**(4), 672–681.
- George, R. A. and Heringa, J. (2002b). SnapDRAGON: a method to delineate protein structural domains from sequence data. *Journal of Molecular Biology*, **316**(3), 839–851.
- George, R. A., Lin, K., and Heringa, J. (2005). Scooby-domain: prediction of globular domains in protein sequence. *Nucleic Acids Research*, **33**(Web Server), W160–W163.
- Gherardini, P. F. and Helmer-Citterich, M. (2008). Structure-Based Function Prediction: Approaches and Applications. *Briefings in Functional Genomics*, **7**(4), 291–302.
- Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Milanesio, M., Gilli, G., Gilli, P., Zanotti, G., Ferraris, G., and Catti, M. (2011). *Fundamentals of Crystallography*. Oxford University Press, 3 edition.
- Glazer, M., Wark, J., and Schmittmann, B. (2002). Statistical Mechanics: A Survival Guide. *American Journal of Physics*, **70**(12), 1274–1275.
- Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, **27**(1), 14–25.
- González-Pérez, A. and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*, **88**(4), 440–449.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press Cambridge.
- Gowers, R., Linke, M., Barnoud, J., Reddy, T., Melo, M., Seyler, S., Domański, J., Dotson, D., Buchoux, S., Kenney, I., and Beckstein, O. (2016). MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, number Scipy, pages 98–105.
- Graaf, C. d., Vermeulen, N. P. E., and Feenstra, K. A. (2005). Cytochrome P450 in Silico: An Integrative Modeling Approach. *Journal of Medicinal Chemistry*, **48**(8), 2725–2755.
- Graña-Montes, R., Pujols-Pujol, J., Gómez-Picanyol, C., and Ventura, S. (2017). Prediction of Protein Aggregation and Amyloid Formation. In *From Protein Structure to Function with Bioinformatics*, pages 205–263. Springer Netherlands, Dordrecht.
- Grossfield, A. (2003). WHAM: the weighted histogram analysis method.
- Grossman, J. P., Kuskin, J. S., Bank, J. A., Theobald, M., Dror, R. O., Ierardi, D. J., Larson, R. H., Schafer, U. B., Towles, B., Young, C., and Shaw, D. E. (2013). Hardware Support for Fine-grained Event-driven Computation in Anton 2. In *Proceedings of the Eighteenth International Conference*

- on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '13, pages 549–560, New York, NY, USA. ACM.
- Gu, J. and Bourne, P. E. (2009). *Structural bioinformatics*. John Wiley & Sons., Hoboken, 2nd ed. nv edition.
- Guharoy, M. and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(43), 15447–15452.
- Guharoy, M. and Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics*, **23**(15), 1909–1918.
- Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, **2013**.
- Halgren, T. A. and Damm, W. (2001). Polarizable force fields. *Current Opinion in Structural Biology*, **11**(2).
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**(23), 4039–4045.
- Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, **35**(14), 2403–2410.
- Hanson, J., Paliwal, K. K., Litfin, T., Yang, Y., and Zhou, Y. (2020). Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning. *Journal of Computational Biology*, **27**(5), 796–814.
- Harris, L. J., Larson, S. B., Hasel, K. W., and McPherson, A. (1997). Refined Structure of an Intact IgG2a Monoclonal Antibody.
- Hartl, F. U. and Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.*, **16**(6), 574–581.
- Hasegawa, H. and Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, **19**(3), 341–348.
- Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, **5**(1), 11476.
- Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, **32**(Web Server), W500–W502.
- Herzik, M. A. (2020). Cryo-electron microscopy reaches atomic resolution.
- Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**(3), 435–447.
- Hintze, B. J., Lewis, S. M., Richardson, J. S., and Richardson, D. C. (2016). Molprobity's ultimate rotamer-library distributions for model validation. *Proteins: Structure, Function and Bioinformatics*, **84**(9), 1177–1189.
- Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**(4), 378–379.
- Holley, L. H. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, **86**(1), 152–156.
- Holm, L. and Laakso, L. M. (2016). Dali server update. *Nucleic Acids Research*, **44**(W1), W351–W355.
- Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996). Errors in protein structures [3].
- Hopkins, C. W., Le Grand, S., Walker, R. C., and Roitberg, A. E. (2015). Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, **11**(4), 1864–1874.
- Horwich, A. L., Farr, G. W., and Fenton, W. A. (2006). GroEL-GroES-mediated protein folding.
- Hoskins, J., Lovell, S., and Blundell, T. L. (2006). An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Science*, **15**(5), 1017–1029.
- Hou, Q., De Geest, P., Vranken, W., Heringa, J., and Feenstra, K. (2017). Seeing the trees through the forest: Sequencebased homo- and heteromeric protein–protein interaction sites prediction using random forest. *Bioinformatics*, **33**(10).
- Hou, Q., De Geest, P. F. G., Griffioen, C. J., Abeln, S., Heringa, J., and Feenstra, K. A. (2019). SeRenDIP: SEquential REmasteriNg to DerIve profiles for fast and accurate predictions of PPI interface positions. *Bioinformatics*.
- Hou, Q., Stringer, B., Waury, K., Capel, H., Haydarlou, R., Xue, F., Abeln, S., Heringa, J., and Feenstra, K. A. (2021). SeRenDIP-CE: sequence-based interface prediction for conformational epitopes. *Bioinformatics*, **37**(20), 3421–3427.

- Hua, S. and Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, **308**(2), 397–407.
- Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X.-H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernández, I. S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E. C., Biggins, S., Keeney, S., Miller, E. A., Fromme, J. C., Hendrickson, T. L., Cong, Q., and Baker, D. (2021). Computed structures of core eukaryotic protein complexes. *Science*.
- Huwe, P. J., Xu, Q., Shapovalov, M. V., Modi, V., Andrade, M. D., and Dunbrack, R. L. (2016). Biological function derived from predicted structures in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 370–391.
- Imai, K. and Mitaku, S. (2005). Mechanisms of secondary structure breakers in soluble proteins. *BIOPHYSICS*, **1**, 55–65.
- Jacoboni, I., Martelli, P. L., Fariselli, P., Pinto, V. D., and Casadio, R. (2001). Prediction of the transmembrane regions of  $\beta$ -barrel membrane proteins with a neural network-based predictor. *Protein Science*, **10**(4), 779–787.
- Jacobsen, A., Bosch, L. J. W., Martens-de Kemp, S. R., Carvalho, B., Sillars-Hardebol, A. H., Dobson, R. J., de Rinaldis, E., Meijer, G. A., Abeln, S., Heringa, J., Fijneman, R. J. A., and Feenstra, K. A. (2018). Aurora kinase A (AURKA) interaction with Wnt and Ras-MAPK signalling pathways in colorectal cancer. *Scientific Reports*, **8**(1), 7522.
- Jespersen, M. C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, **45**(W1), W24–W29.
- Jiang, W. and Tang, L. (2017). Atomic cryo-EM structures of viruses. *Current Opinion in Structural Biology*, **46**, 122–129.
- Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., Koo, D. C. E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, S. M. E., Martelli, P. L., Profiti, G., Casadio, R., Cao, R., Zhong, Z., Cheng, J., Altenhoff, A., Skunca, N., Dessimoz, C., Dogan, T., Hakala, K., Kaewphan, S., Mehryary, F., Salakoski, T., Ginter, F., Fang, H., Smithers, B., Oates, M., Gough, J., Törönen, P., Koskinen, P., Holm, L., Chen, C.-T., Hsu, W.-L., Bryson, K., Cozzetto, D., Minneci, F., Jones, D. T., Chapman, S., Bkc, D., Khan, I. K., Kihara, D., Ofer, D., Rappoport, N., Stern, A., Cibrian-Uhalte, E., Denny, P., Foulger, R. E., Hieta, R., Legge, D., Lovering, R. C., Magrane, M., Melidoni, A. N., Mutowo-Meullenet, P., Pichler, K., Shypitsyna, A., Li, B., Zakeri, P., ElShal, S., Tranchevent, L.-C., Das, S., Dawson, N. L., Lee, D., Lees, J. G., Sillitoe, I., Bhat, P., Nepusz, T., Romero, A. E., Sasidharan, R., Yang, H., Paccanaro, A., Gillis, J., Sedeño-Cortés, A. E., Pavlidis, P., Feng, S., Cejuela, J. M., Goldberg, T., Hamp, T., Richter, L., Salamov, A., Gabaldon, T., Marcket-Houben, M., Supek, F., Gong, Q., Ning, W., Zhou, Y., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Toppo, S., Ferrari, C., Giollo, M., Piovesan, D., Tosatto, S. C. E., del Pozo, A., Fernández, J. M., Maietta, P., Valencia, A., Tress, M. L., Benso, A., Di Carlo, S., Politano, G., Savino, A., Rehman, H. U., Re, M., Mesiti, M., Valentini, G., Bargsten, J. W., van Dijk, A. D. J., Gemovic, B., Glisic, S., Perovic, V., Veljkovic, V., Veljkovic, N., Almeida-e Silva, D. C., Vencio, R. Z. N., Sharan, M., Vogel, J., Kansakar, L., Zhang, S., Vucetic, S., Wang, Z., Sternberg, M. J. E., Wass, M. N., Huntley, R. P., Martin, M. J., O'Donovan, C., Robinson, P. N., Moreau, Y., Tramontano, A., Babbitt, P. C., Brenner, S. E., Linial, M., Orengo, C. A., Rost, B., Greene, C. S., Mooney, S. D., Friedberg, I., and Radivojac, P. (2016). An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy. *Genome Biology*, **17**(1).
- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices 1 Edited by G. Von Heijne. *Journal of Molecular Biology*, **292**(2), 195–202.
- Jones, D. T. and Thornton, J. M. (2022). The impact of AlphaFold2 one year on. *Nature Methods*, **19**(1), 15–20.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**(6381), 86–89.
- Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–190.
- Jumper, J. and Hassabis, D. (2022). Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods* 2022 19:1, **19**(1), 11–12.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 2021 596:7873, **596**(7873), 583–589.
- Kabsch, W. and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of

- Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577–2637.
- Kagami, L., Roca-Martínez, J., Gavaldá-García, J., Ramasamy, P., Feenstra, K. A., and Vranken, W. F. (2021a). Online biophysical predictions for SARS-CoV-2 proteins. **22**(1), 1–7.
- Kagami, L. P., Orlando, G., Raimondi, D., Ancien, F., Dixit, B., Gavaldá-García, J., Ramasamy, P., Roca-Martínez, J., Tzavella, K., and Vranken, W. (2021b). b2bTools: online predictions for protein biophysical features and their conservation. *Nucleic Acids Research*, **49**(W1), W52–W59.
- Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B*, **105**(28), 6474–6487.
- Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.
- Kardos, J., Yamamoto, K., Hasegawa, K., Naiki, H., and Goto, Y. (2004). Direct measurement of the thermodynamic parameters of amyloid formation by isothermal titration calorimetry. *Journal of Biological Chemistry*, **279**(53), 55308–55314.
- Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, **102**(19), 6679–6685.
- Kästner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **1**(6), 932–942.
- Kato, H., van Ingen, H., Zhou, B.-R., Feng, H., Bustin, M., Kay, L. E., and Bai, Y. (2011). Architecture of the high mobility group nucleosomal protein 2-nucleosome complex as revealed by methyl-based NMR. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(30), 12283–8.
- Kearsley, S. K. (1989). On the orthogonal transformation used for structural comparisons. *Acta Crystallographica Section A*, **45**(2), 208–210.
- Keizers, P. H. J., Graaf, C. d., Kanter, F. J. J. d., Oostenbrink, C., Feenstra, K. A., Commandeur, J. N. M., and Vermeulen, N. P. E. (2005). Metabolic Regio- and Stereoselectivity of Cytochrome P450 2D6 towards 3,4-Methylenedioxy-N-alkylamphetamines: in Silico Predictions and Experimental Validation. *Journal of Medicinal Chemistry*, **48**(19), 6117–6127.
- Kelly, S. M., Jess, T. J., and Price, N. C. (2005). How to study proteins by circular dichroism.
- Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, **185**(4711), 422–427.
- Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016a). Assessment of CASP11 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 164–180.
- Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016b). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 51–66.
- Kirillova, S., Kumar, S., and Carugo, O. (2009). Protein domain boundary predictions: a structural biology perspective. *The open biochemistry journal*, **3**, 1–8.
- Kittel, C. and Kroemer, H. (1980). Thermal physics. *WIT Freeman: San Francisco*.
- Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., Marcattili, P., Soenderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., and Petersen, B. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, **87**(6), 520–527.
- Knecht, C., Mort, M., Junge, O., Cooper, D. N., Krawczak, M., and Caliebe, A. (2017). IMHOTEP-a Composite Score Integrating Popular Tools for Predicting the Functional Consequences of Non-Synonymous Sequence Variants. *Nucleic Acids Research*, **45**(3), e13–e13.
- Knowles, T. P., Vendruscolo, M., and Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, **15**(6), 384–396.
- Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Laulederkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Louighi, H., Della Rocca, M. G., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., and Robinson, P. N. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*.

- Kolodny, R., Petrey, D., and Honig, B. (2006). Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, **16**(3), 393–398.
- Krieger, E. and Vriend, G. (2014). YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, **30**(20), 2981–2982.
- Krieger, E. and Vriend, G. (2015). New ways to boost molecular dynamics simulations. *Journal of Computational Chemistry*, **36**(13).
- Kringelum, J. V., Lundsgaard, C., Lund, O., and Nielsen, M. (2012). Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Computational Biology*, **8**(12), e1002829.
- Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, **305**(3), 567–580.
- Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 349–369.
- Kühner, S., van Noort, V., Betts, M., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castaño-Diez, D., Chen, W.-H., Devos, D., Güell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Böttcher, B., Frangakis, A., Russell, R., Serrano, L., Bork, P., and Gavin, A.-C. (2009). Proteome Organization in a Genome-Reduced Bacterium. *Science*, **326**, 1235–1240.
- Kumar, S. and Li, M. S. (2010). Biomolecules under mechanical force. *Phys. Rep.*, **486**, 1–74.
- Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F., and Mackay, J. P. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal*, **278**(5), 687–703.
- Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**(2), 283–291.
- Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (2012). PROCHECK : validation of protein-structure coordinates. In E. Arnold, D. M. Himmel, and M. G. Rossmann, editors, *International Tables for Crystallography*, volume F, Ch.21.4, pages 684–687.
- Leach, A. (2001). *Molecular Modelling: Principles and Applications*. Pearson.
- Lee, G. R., Heo, L., and Seok, C. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 293–301.
- Leiman, P. G., Arisaka, F., van Raaij, M. J., Kostyuchenko, V. A., Aksyuk, A. A., Kanamaru, S., and Rossmann, M. G. (2010). Morphogenesis of the T4 tail and tail fibers. *Virology Journal*, **7**(1), 355.
- Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. (2004). UniProt archive. *Bioinformatics*, **20**(17), 3236–3237.
- Lelieveld, S. H., Schütte, J., Dijkstra, M. J., Bawono, P., Kinston, S. J., Göttgens, B., Heringa, J., and Bonzanni, N. (2016). ConBind: Motif-aware cross-species alignment for the identification of functional transcription factor binding sites. *Nucleic Acids Research*, **44**(8).
- Lemkul, J. A. and Bevan, D. R. (2013). Aggregation of Alzheimer's amyloid  $\beta$ -peptide in biological membranes: a molecular dynamics study. *Biochemistry*, **52**(29), 4971–4980.
- Lensink, M. and Mendez, R. (2008). Recognition-induced Conformational Changes in Protein-Protein Docking. *Current Pharmaceutical Biotechnology*, **9**(2), 77–86.
- Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., Grudinin, S., Popov, P., Neveu, E., Lee, H., Baek, M., Park, S., Heo, L., Rie Lee, G., Seok, C., Qin, S., Zhou, H.-X., Ritchie, D. W., Maigret, B., Devignes, M.-D., Ghoorah, A., Torchala, M., Chaleil, R. A., Bates, P. A., Ben-Zeev, E., Eisenstein, M., Negi, S. S., Weng, Z., Vreven, T., Pierce, B. G., Borrman, T. M., Yu, J., Ochsenebein, F., Guerois, R., Vangone, A., Rodrigues, J. P., van Zundert, G., Nellen, M., Xue, L., Karaca, E., Melquiond, A. S., Visscher, K., Kastritis, P. L., Bonvin, A. M., Xu, X., Qiu, L., Yan, C., Li, J., Ma, Z., Cheng, J., Zou, X., Shen, Y., Peterson, L. X., Kim, H.-R., Roy, A., Han, X., Esquivel-Rodriguez, J., Kihara, D., Yu, X., Bruce, N. J., Fuller, J. C., Wade, R. C., Anishchenko, I., Kundrotas, P. J., Vakser, I. A., Imai, K., Yamada, K., Oda, T., Nakamura, T., Tomii, K., Pallara, C., Romero-Durana, M., Jiménez-García, B., Moal, I. H., Fernández-Recio, J., Joung, J. Y., Kim, J. Y., Joo, K., Lee, J., Kozakov, D., Vajda, S., Mottarella, S., Hall, D. R., Beglov, D., Mamonov, A., Xia, B., Bohnuud, T., Del Carpio, C. A., Ichishi, E., Marze, N., Kuroda, D., Roy Burman, S. S., Gray, J. J., Chermak, E., Cavallo, L., Oliva, R., Tsvchigrechko, A., and Wodak, S. J. (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 323–348.
- Levinthal, C. (1969). How to fold graciously. In *Mössbau Spectroscopy in Biological Systems Proceedings*, volume 67 of *Univ. of Illinois Bulletin*, pages 22–24, Urbana, IL 61801.

- Li, B.-Q., Feng, K.-Y., Chen, L., Huang, T., and Cai, Y.-D. (2012). Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PloS one*, **7**(8), e43927.
- Li, P. and Merz, K. M. (2017). Metal Ion Modeling Using Classical Mechanics. *Chemical Reviews*, **117**(3), 1564–1686.
- Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., Agard, D. A., and Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods*, **10**(6), 584–590.
- Lin, K., Simossis, V. A., Taylor, W. R., and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics (Oxford, England)*, **21**(2), 152–159.
- Lin, S., Cheng, C.-W., and Su, E. (2013). Prediction of B-cell epitopes using evolutionary information and propensity scales. *BMC Bioinformatics*, **14**(Suppl 2), S10.
- Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: Implications for structural proteomics. *Structure*, **11**(11), 1453–1459.
- Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS ONE*, **7**(2), e32131.
- Lingenheil, M., Denschlag, R., Reichold, R., and Tavan, P. (2008). The “Hot-Solvent/Cold-Solute” Problem Revisited. *Journal of Chemical Theory and Computation*, **4**(8), 1293–1306.
- Liu, J., Wu, T., Guo, Z., Hou, J., and Cheng \$, J. (2021). Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *bioRxiv*, page 2021.01.28.428706.
- Luby-Phelps, K. (1999). Cytoarchitecture and Physical Properties of Cytoplasm: Volume, Viscosity, Diffusion, Intracellular Surface Area. In *International Review of Cytology*, volume 192, pages 189–221. Academic Press.
- Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V., and Dunin-Horkawicz, S. (2019). DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, **35**(16), 2790–2795.
- Lupas, A. (1997). Predicting coiled-coil regions in proteins. *Current Opinion in Structural Biology*, **7**(3), 388–393.
- Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**(15), 3255–3263.
- Lutz, M. (2013). *Learning Python*. O'Reilly.
- Ma, J., Peng, J., Wang, S., and Xu, J. (2012). A conditional neural fields model for protein threading. *Bioinformatics*, **28**(12), 59–66.
- Ma, J., Wang, S., Zhao, F., and Xu, J. (2013). Protein threading using context-specific alignment potential. *Bioinformatics*, **29**(13), 257–265.
- Maglott, D. (2004). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**(Database issue), D54–D58.
- Magnan, C. N. and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**(18), 2592–2597.
- Mah, J. T. L., Low, E. S. H., and Lee, E. (2011). In Silico SNP Analysis and Bioinformatics Tools: A Review of the State of the Art to Aid Drug Discovery. *Drug Discovery Today*, **16**(17-18), 800–809.
- Maiorov, V. N. and Crippen, G. M. (1995). Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Genetics*, **22**(3), 273–283.
- Malhotra, A., Penczek, P., Agrawal, R. K., Gabashvili, I. S., Grassucci, R. A., Jünemann, R., Burkhardt, N., Nierhaus, K. H., and Frank, J. (1998). Escherichia coli 70 S ribosome at 15 Å resolution by cryo-electron microscopy: localization of fmet-tRNAfMet and fitting of L1 protein. *Journal of Molecular Biology*, **280**(1), 103–116.
- Marchler-Bauer, A. (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Research*, **31**(1), 383–387.
- Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y., and Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, **45**(D1), D200–D203.
- Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. a., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.
- Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, **111**(27), 7812–7824.
- Marti-Renom, M. A., Capriotti, E., Shindyalov, I. N., and Bourne, P. E. (2009). Structure Comparison and Alignment. In J. Gu and P. E. Bourne, editors, *Structural Bioinformatics, 2nd Edition*, pages 397–418. John Wiley & Sons, Inc.
- Matthew Zimmerman, Marek Grabowski, Wladek Minor, Helen M. Berman, Margaret J. Gabanyi,

- Robert Lowe, Raship Shah, Wendy Yi-Ping Tao, John D. Westbrook, and Protein Structure Initiative network of Investigators (2017). Protein Structure Initiative Publications, 2000-2016.
- May, A., Pool, R., van Dijk, E., Bijlard, J., Abeln, S., Heringa, J., and Feenstra, K. A. (2014). Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics (Oxford, England)*, **30**(3), 326–334.
- May, A., Brandt, B. W., El-Kebir, M., Klau, G. W., Zaura, E., Crielaard, W., Heringa, J., and Abeln, S. (2016). metaModules identifies key functional subnetworks in microbiome-related disease. *Bioinformatics*, **32**(11), 1678–1685.
- McCammon, J. A., Gelin, B. R., Karplus, M., and Wolynes, P. G. (1976). Hinge bending mode in lysozyme. *Nature*, **262**, 325–326.
- McDonald, I. K. and Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, **238**(5), 777–793.
- McNay, J. L., O'Connell, J. P., and Fernandez, E. J. (2001). Protein unfolding during reversed-phase chromatography: II. Role of salt type and ionic strength. *Biotechnology and Bioengineering*, **76**(3).
- Melo, R., Fieldhouse, R., Melo, A., Correia, J. D., Cordeiro, M. N. D., Gümus, Z. H., Costa, J., Bonvin, A. M., and Moreira, I. S. (2016). A machine learning approach for hot-spot detection at protein-protein interfaces. *International Journal of Molecular Sciences*, **17**(8).
- Mészáros, B., Tompa, P., Simon, I., and Dosztányi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol*, **372**(2), 549–561.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.
- Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, **32**(10), 2319–2327.
- Micsonai, A., Wien, F., Kernya, L., Lee, Y.-H., Goto, Y., Réfrégiers, M., and Kardos, J. (2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, **112**(24), E3095–E3103.
- Mills, C. L., Beuning, P. J., and Ondrechen, M. J. (2015). Biochemical Functional Predictions for Protein Structures of Unknown or Uncertain Function. *Computational and Structural Biotechnology Journal*, **13**, 182–191.
- Mirabello, C. and Wallner, B. (2019). rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE*, **14**(8), e0220182.
- Mitchell, P. (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature*, **191**(4784), 144–148.
- Modi, V. and Dunbrack, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 260–281.
- Modi, V., Xu, Q., Adhikari, S., and Dunbrack, R. L. (2016). Assessment of template-based modeling of protein structure in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 200–220.
- Mondal, P., Khamo, J. S., Krishnamurthy, V. V., Cai, Q., Zhang, K., Diao, J., Wang, Y., and Qiu, R. (2017). Drive the Car(go)s-New Modalities to Control Cargo Trafficking in Live Cells.
- Montelione, G. (2012). The Protein Structure Initiative: achievements and visions for the future. *F1000 Biology Reports*, **4**, 7.
- Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tielemans, D. P., and Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput*, **4**(5), 819–834.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, **38**(8), 4.
- Mor, A., Ziv, G., and Levy, Y. (2008). Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *Journal of Computational Chemistry*, **29**(12), 1992–1998.
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(49), 1293–301.
- Morrell, W. E. and Hildebrand, J. H. (1936). The Distribution of Molecules in a Model Liquid. *The Journal of Chemical Physics*, **4**(3), 224–227.
- Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics*, **23**(3), ii–iv.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, **84**(S1), 4–14.
- Murakami, Y. and Mizuguchi, K. (2010). Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, **26**(15), 1841–1848.

- Naderi-Manesh, H., Sadeghi, M., Arab, S., and Moosavi Movahedi, A. A. (2001). Prediction of protein surface accessibility with information theory. *Proteins: Structure, Function, and Bioinformatics*, **42**(4), 452–459.
- Necci, M., Piovesan, D., and Tosatto, S. C. E. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods* 2021 18:5, **18**(5), 472–481.
- Nogales, E., Wolf, S. G., and Downing, K. H. (1998). Structure of the  $\alpha\beta$  tubulin dimer by electron crystallography. *Nature*, **391**(6663), 199–203.
- Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.
- Nugent, T., Cozzetto, D., and Jones, D. T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function, and Bioinformatics*, **82**(S2), 98–111.
- Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztányi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., and Gough, J. (2013). D2P2: Database of disordered protein predictions. *Nucleic Acids Research*, **41**(D1), D508–D516.
- Ofran, Y. and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput. Biol.*, **3**(7), e119.
- Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**(6), 1989–2000.
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**(D1), D733–D745.
- Onufriev, A. V. and Case, D. A. (2019). Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics*, **48**(1), 275–296.
- Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem*, **25**(13), 1656–1676.
- Orrego, C. A. and Taylor, W. R. (1996). [36] SSAP: Sequential structure alignment program for protein structure comparison. In *Methods in Enzymology*, volume 266, pages 617–635. Academic Press.
- Orlando, G., Raimondi, D., Codice, F., Tabaro, F., and Vranken, W. (2018). Prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics. *bioRxiv*, page 2020.05.25.115253.
- O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *Journal of Molecular Biology*, **340**(2), 385–395.
- Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyripides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, **355**(6322), 294–298.
- Pabo, C. and Sauer, R. (1984). Protein-DNA recognition. *Annu Rev Biochem*, **53**, 293–321.
- Pancsa, R., Raimondi, D., Cilia, E., and Vranken, W. F. (2016). Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity. *Biophysical Journal*, **110**(3).
- Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., Schafferhans, A., and O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, **112**(52), 15898–15903.
- Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology*, **9**(1), 1.
- Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G. P., and Lund, O. (2000). Prediction of protein secondary structure at 80% accuracy. *Proteins: Structure, Function and Genetics*, **41**(1), 17–20.
- Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A. M.-P., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H. E., and Brazma, A. (2016). Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, **44**(D1), D746–D752.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and

- Ferrin, T. E. (2004). UCSF Chimera – A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13), 1605–1612.
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**(16), 1781–1802.
- Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C. R., Radak, B. K., Skeel, R. D., Singhary, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L. V., Schulten, K., Chipot, C., and Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, **153**(4).
- Phillips, R., Kondev, J., Theriot, J., Garcia, H. G., and Orme, N. (2012). *Physical biology of the cell*.
- Pieber, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., Tainer, J. A., Nilges, M., and Sali, A. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, **42**(D1), D336–D346.
- Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, **24**(19), 3836–3845.
- Pirovano, W. and Heringa, J. (2010). Protein secondary structure prediction.
- Pirovano, W., Feenstra, K. A., and Heringa, J. (2008). PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, **24**(4), 492–497.
- Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function and Genetics*, **47**(2), 228–235.
- Ponder, J. W. and Case, D. A. (2003). Force fields for protein simulations. *Adv. Protein Chem.*, **66**, 27–85.
- Pool, R., Heringa, J., Hoefling, M., Schulz, R., Smith, J., and Feenstra, K. (2012). Enabling grand-canonical Monte Carlo: Extending the flexibility of GROMACS through the grompy python interface module. *Journal of Computational Chemistry*, **33**(12).
- Privalov, P. L. and Khechinashvili, N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *Journal of Molecular Biology*, **86**(3), 665–684.
- Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**(7), 845–854.
- Pucci, F. and Rooman, M. (2017). Physical and molecular bases of protein thermal stability and cold adaptation. *Current Opinion in Structural Biology*, **42**, 117–128.
- Radermacher, M., Rao, V., Grassucci, R., Frank, J., Timerman, A. P., Fleischer, S., and Wagenknecht, T. (1994). Cryo-electron microscopy and three-dimensional reconstruction of the calcium release channel/ryanodine receptor from skeletal muscle. *The Journal of cell biology*, **127**(2), 411–23.
- Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W. A., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kaßner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Höngschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Björne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J. E., Škunca, N., Supek, F., Bošnjak, M., Panov, P., Džeroski, S., Šmuc, T., Kourmpetis, Y. A. I., van Dijk, A. D. J., ter Braak, C. J. F., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzi, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D., and Friedberg, I. (2013). A Large-Scale Evaluation of Computational Protein Function Prediction. *Nature Methods*, **10**(3), 221–227.
- Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.*, **136**(2A), A405–A411.
- Raimondi, D., Orlando, G., Pancsa, R., Khan, T., and Vranken, W. F. (2017). Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. *Scientific Reports*, **7**(1), 8826.
- Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L., and Saibil, H. R. (2001). ATP-Bound States of GroEL Captured by Cryo-Electron Microscopy. *Cell*, **107**(7), 869–879.
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, page 676825. Neural information processing systems foundation.

- Rapaport, D. C. (2004). The Art of Molecular Dynamics Simulation. *The Art of Molecular Dynamics Simulation*.
- Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, **80**(8), 2071–2079.
- Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, **29**(9), 2352–2449.
- Read, R. J. (1997). Model phases: Probabilities and bias.
- Read, R. J. and Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function, and Bioinformatics*, **69**(S8), 27–37.
- Reif, M. M., Hünenberger, P. H., and Oostenbrink, C. (2012). New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *Journal of Chemical Theory and Computation*, **8**(10).
- Relini, A., Marano, N., and Gliozzi, A. (2014). Probing the interplay between amyloidogenic proteins and membranes using lipid monolayers and bilayers.
- Richards, F. M. and Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Genetics*, **3**(2), 71–84.
- Richardson, J. S. and Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A*, **99**(5), 2754–2759.
- Rost, B. and Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, **90**(16), 7558–7562.
- Rost, B., Sander, C., and Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Bioinformatics*, **10**(1), 53–60.
- Rost, B., Nair, R., Liu, J., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic Prediction of Protein Function. *Cellular and Molecular Life Sciences (CMLS)*, **60**(12), 2637–2650.
- Roux, B. and Simonson, T. (1999). Implicit solvent models. *Biophysical Chemistry*, **78**(1-2).
- Roy, A., Yang, J., and Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, **40**(W1), 471–477.
- Sadreyev, R. and Grishin, N. (2003). COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of Molecular Biology*, **326**(1), 317–336.
- Sahin, E., Grillo, A. O., Perkins, M. D., and Roberts, C. J. (2010). Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. *Journal of Pharmaceutical Sciences*, **99**(12).
- Salamov, A. A. and Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments.
- Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**(3), 779–815.
- Sali, A., Shakhnovich, E., and Karplus, M. (1994). Kinetics of Protein Folding : A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.*, **235**(5), 1614–1638.
- Sanchez-Trincado, J. L., Gomez-Perez, M., and Reche, P. A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of Immunology Research*, **2017**, 1–14.
- Schotte, F., Soman, J., Olson, J. S., Wulff, M., and Anfinrud, P. A. (2004). Picosecond time-resolved X-ray crystallography: Probing protein function in real time. *Journal of Structural Biology*, **147**(3), 235–246.
- Schraudt, O., Lefebvre, M. D., Brunner, M. J., Schmied, W. H., Schmidt, A., Radics, J., Mechtler, K., Galán, J. E., and Marlovits, T. C. (2010). Topology and Organization of the *Salmonella* typhimurium Type III Secretion Needle Complex Components. *PLoS Pathogens*, **6**(4), e1000824.
- Schroeder, D. V. (1999). *An Introduction to Thermal Physics*. Addison-Wesley Publishing Company, San Francisco, CA.
- Schwalbe, H., Grimshaw, S. B., Spencer, A., Buck, M., Boyd, J., Dobson, C. M., Redfield, C., and Smith, L. J. (2001). A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Science*, **10**(4), 677–688.
- Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, **31**(13), 3381–3385.
- Sehnal, D., Deshpande, M., Vareková, R. S., Mir, S., Berka, K., Midlik, A., Pravda, L., Velankar, S., and Koča, J. (2017). LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nature Methods*, **14**(12), 1121–1122.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, **577**(7792), 706–710.
- Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A.,

- Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, **15**, 341–346.
- Shen, P., Iwasa, J., Thuesen, A., Wambaugh, M., and Stewart, M. (2018). *CryoEM 101*. University of Utah, Utah.
- Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**(1), 243–257.
- Shi, Q., Chen, W., Huang, S., Wang, Y., and Xue, Z. (2021). Deep learning for mining protein data. *Briefings in Bioinformatics*, **22**(1), 194–218.
- Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection*, **11**(9), 739–747.
- Shivakumar, D., Harder, E., Damm, W., Friesner, R. A., and Sherman, W. (2012). Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *Journal of Chemical Theory and Computation*, **8**(8).
- Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, **3**(4), e43.
- Silverstein, T. P. (1998). The Real Reason Why Oil and Water Don't Mix. *Journal of Chemical Education*, **75**(1), 116.
- Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **Suppl 3**, 171–176.
- Simossis, V. A. and Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, **33**(Web Server), W289–W294.
- Simossis, V. A., Kleijnjung, J., and Heringa, J. (2005). Homology-extended sequence alignment. *Nucleic Acids Research*, **33**(3), 816–824.
- Singh, N. and Li, W. (2019). Recent advances in coarse-grained models for biomolecules and their applications. *International Journal of Molecular Sciences*, **20**(15).
- Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L. M. T., Evelo, C. T., Pico, A. R., and Willighagen, E. L. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, **46**(D1), D661–D667.
- Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews*, **66**(1), 334–395.
- Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, **33**(SUPPL. 2), 244–8.
- Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*, **21**(10), 1735–1742.
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C., Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., Best, R. B., Vattulainen, I., Monticelli, L., Periole, X., Tielemans, D. P., de Vries, A. H., and Marrink, S. J. (2021). Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods*, **18**(4).
- Stillinger, F. H. and Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, **60**(4), 1545–57.
- Stringer, B., Ferrante, H. d., Abeln, S., Heringa, J., Feenstra, K. A., and Haydarlou, R. (2021). PIPENN: Protein Interface Prediction with an Ensemble of Neural Nets. *bioRxiv*, page 2021.09.03.458832.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926–932.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**(D1), D447–D452.
- Tang, X. and Bruce, J. E. (2009). Chemical Cross-Linking for Protein-Protein Interaction Studies. *Methods In Molecular Biology*, **492**, 283–293.
- Tartaglia, G. G. and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Molecular BioSystems*, **5**(12), 1873–1876.
- Taylor, N. M. I., Prokhorov, N. S., Guerrero-Ferreira, R. C., Shneider, M. M., Browning, C., Goldie, K. N., Stahlberg, H., and Leiman, P. G. (2016). Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature*, **533**(7603), 346–352.
- Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology*,

- 208**(1), 1–22.
- Teilum, K., Kunze, M. B. A., Erlendsson, S., and Kragelund, B. B. (2017). (S)Pinning down protein interactions by NMR. *Protein Science*, **26**(3), 436–451.
- Terashi, G. and Kihara, D. (2017). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function and Bioinformatics*.
- Thiriot, D. S., Nevzorov, A. A., and Opella, S. J. (2005). Structural basis of the temperature transition of Pfl bacteriophage. *Protein Science*, **14**(4), 1064–1070.
- Thornton, J. M., Laskowski, R. A., and Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine* **27**:10, **27**(10), 1666–1669.
- Thul, P. J. and Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science*, **27**(1), 233–244.
- Tian, C., Kasavajhala, K., Belfon, K. A. A., Ragquette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., and Simmerling, C. (2020). f19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, **16**(1).
- Tian, W. and Skolnick, J. (2003). How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *Journal of Molecular Biology*, **333**(4), 863–882.
- Tompa, P. and Fersht, A. (2009). *Structure and Function of Intrinsically Disordered Proteins*. Chapman and Hall/CRC.
- Tsong, T. Y., Baldwin, R. L., McPhie, P., and Elson, E. L. (1972). A sequential model of nucleation-dependent protein folding: Kinetic studies of ribonuclease A. *Journal of Molecular Biology*, **63**(3), 453–469.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* **596**:7873, **596**(7873), 590–596.
- Tusnády, G. E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of Molecular Biology*, **283**(2), 489–506.
- Unwin, N. (1993). Nicotinic acetylcholine receptor at 9 Å resolution. *Journal of molecular biology*, **229**(4), 1101–24.
- Unwin, N. (2005). Refined Structure of the Nicotinic Acetylcholine Receptor at 4 Å Resolution. *Journal of Molecular Biology*, **346**(4), 967–989.
- Unwin, N. and Fujiyoshi, Y. (2012). Gating Movement of Acetylcholine Receptor Caught by Plunge-Freezing. *Journal of Molecular Biology*, **422**(5), 617–634.
- Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**(3), 415–427.
- van Aalten, D. M. F., de Groot, B. L., Berendsen, H. J. C., Findlay, J. B. C., and Amadei, A. (1997). A Comparison of Techniques for Calculating Protein Essential Dynamics. *J. Comput. Chem.*, **18**(2), 169–181.
- van Dijk, E., Hoogeveen, A., and Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, **11**(5), e1004277.
- van Dijk, E., Varilly, P., Knowles, T. P. J., Frenkel, D., and Abeln, S. (2016). Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, **116**(7), 078101.
- van Gunsteren, W. F., Berendsen, H. J., Hermans, J., Hol, W. G., and Postma, J. P. (1983). Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proceedings of the National Academy of Sciences*, **80**(14), 4315–4319.
- van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P. A., Mark, E., Scott, W. R. P., and Tironi, I. G. (1996). *Biomolecular Simulation: The GROMOS96 Manual and User Guide*. Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland.
- van Gunsteren, W. F., Burgi, R., Peter, C., and Daura, X. (2001). The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State. *Angew. Chem. Int. Ed. Engl.*, **40**(2), 351–355.
- van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glattli, A., Hunenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F., and Yu, H. B. (2006). Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed. Engl.*, **45**(25), 4064–4092.
- Van Heel, M., Brent, G., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., and Patwardhan, A. (2000). Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, **33**(4), 307–369.

- van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., and Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**(D1), D439–D444.
- Vendruscolo, M. and Dobson, C. M. (2005). Towards complete descriptions of the free-energy landscapes of proteins. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **363**(1827), 433–452.
- Venselaar, H., te Beek, T. A., Kuipers, R. K., Hekkelman, M. L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, **11**(1), 548.
- Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, **159**, 98–103.
- Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. (2020). BERTology Meets Biology: Interpreting Attention in Protein Language Models.
- Vlugt, T. J., Eerden, J. P. v. d., Dijkstra, M., Smit, B., and Daan Frenkel (2008). *Introduction to Molecular Simulation and Statistical Thermodynamics*. Delft.
- Wagner, M., Adamczak, R. R., Porollo, A., and Meller, J. J. (2005). Linear Regression Models for Solvent Accessibility Prediction in Proteins. *Journal of Computational Biology*, **12**(3), 355–369.
- Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic acids research*, **42**(W1), W301–W307.
- Wang, G. and Dunbrack, R. L. (2004). Scoring profile-to-profile sequence alignments. *Protein Science*, **13**(6), 1612–1626.
- Wang, S., Ma, J., and Xu, J. (2016a). AUCpred: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics (Oxford, England)*, **32**(17), i672–i679.
- Wang, S., Peng, J., Ma, J., and Xu, J. (2016b). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*, **6**(1), 18962.
- Wang, S., Sun, S., and Xu, J. (2017). Analysis of deep learning methods for blind protein contact prediction in CASP12.
- Ward, A. B. and Wilson, I. A. (2017). The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunological Reviews*, **275**(1), 21–32.
- Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, **337**, 635645.
- Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, **103**(2), 227–249.
- Wass, M. N., Kelley, L. A., and Sternberg, M. J. E. (2010). 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Research*, **38**(SUPPL. 2), 469–473.
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.
- Wilhelm, B. G., Mandad, S., Truckenbrodt, S., Krohnert, K., Schafer, C., Rammner, B., Koo, S. J., Classen, G. A., Krauss, M., Haucke, V., Urlaub, H., and Rizzoli, S. O. (2014). Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. *Science*, **344**(6187), 1023–1028.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.
- Wilson, A. E., Kosater, W. M., and Liberles, D. A. (2020). Evolutionary Processes and Biophysical

- Mechanisms: Revisiting Why Evolved Proteins Are Marginally Stable.
- Wohlers, I., Malod-Dognin, N., Andonov, R., and Klau, G. W. (2012). CSA: Comprehensive comparison of pairwise protein structure alignments. *Nucleic Acids Research*, **40**(W1), 303–309.
- Woo, H.-J., Dinner, A. R., and Roux, B. (2004). Grand canonical Monte Carlo simulations of water in protein environments. *The Journal of Chemical Physics*, **121**(13), 6392–6400.
- Wood, W. W. and Parker, F. R. (1957). Monte Carlo Equation of State of Molecules Interacting with the Lennard-Jones Potential. I. A Supercritical Isotherm at about Twice the Critical Temperature. *The Journal of Chemical Physics*, **27**(3), 720–733.
- Wu, C. H., Yeh, L.-S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B. E., Vinayaka, C. R., Zhang, J., and Barker, W. C. (2003). The Protein Information Resource. *Nucleic acids research*, **31**(1), 345–7.
- Wüthrich, K. (1989). [6] Determination of three-dimensional protein structures in solution by nuclear magnetic resonance: An overview. *Methods in Enzymology*, **177**(C), 125–131.
- Wuyun, Q., Zheng, W., Peng, Z., and Yang, J. (2016). A large-scale comparative assessment of methods for residue-residue contact prediction. *Briefings in Bioinformatics*, page bbw106.
- Xin, F. and Radivojac, P. (2012). Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics*, **28**(22), 2905–2913.
- Xiong, J. (2006). *Essential Bioinformatics*. Cambridge University Press.
- Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function and Bioinformatics*, **80**(7), 1715–1735.
- Xu, G., Wang, Q., and Ma, J. (2020). OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, **36**(20), 5021–5026.
- Xue, L. C., Dobbs, D., and Honavar, V. (2011). HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, **12**(1), 1.
- Yang, A. S. and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology*, **301**(3), 665–678.
- Yang, J., Yan, W., Yu, Y., Wang, Y., Yang, T., Xue, L., Yuan, X., Long, C., Liu, Z., Chen, X., Hu, M., Zheng, L., Qiu, Q., Pei, H., Li, D., Wang, F., Bai, P., Wen, J., Ye, H., and Chen, L. (2018). The compound millepachine and its derivatives inhibit tubulin polymerization by irreversibly binding to the colchicine-binding site in  $\beta$ -tubulin. *Journal of Biological Chemistry*, **293**(24), 9461–9472.
- Yang, K., Rózycki, B., Cui, F., Shi, C., Chen, W., and Li, Y. (2016). Sampling enrichment toward target structures using hybrid molecular dynamics-Monte Carlo simulations. *PLoS ONE*, **11**(5).
- Yang, L., Tan, C.-h., Hsieh, M.-J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P. A., and Luo, R. (2006). New-Generation Amber United-Atom Force Field. *The Journal of Physical Chemistry B*, **110**(26).
- Yang, Y. I., Shao, Q., Zhang, J., Yang, L., and Gao, Y. Q. (2019). Enhanced sampling in molecular dynamics. *151*(7).
- Yu, X., Jin, L., and Zhou, Z. H. (2008). 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature*, **453**(7193), 415–419.
- Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, **31**(7), 1235–1270.
- Yuan, Z., Mattick, J. S., and Teasdale, R. D. (2004). SVMtm: Support vector machines to predict transmembrane segments. *Journal of Computational Chemistry*, **25**(5), 632–636.
- Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research*, **31**(13), 3370–4.
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, **46**(D1), D754–D761.
- Zhang, J. and Kurgan, L. (2019). SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. In *Bioinformatics*, volume 35, pages i343–i353. Oxford University Press.
- Zhang, J., Ma, Z., and Kurgan, L. (2018). Comprehensive Review and Empirical Analysis of Hallmarks of DNA-, RNA- and Protein-Binding Residues in Protein Chains. *Briefings in Bioinformatics*, **20**(4), 1250–1268.
- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**(1), 40.
- Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C., and Yates, J. R. (2013). Protein analysis by

- shotgun/bottom-up proteomics. *Chem. Rev.*, **113**(4), 2343–2394.
- Zhang, Y., Sun, B., Feng, D., Hu, H., Chu, M., Qu, Q., Tarrasch, J. T., Li, S., Sun Kobilka, T., Kobilka, B. K., and Skiniotis, G. (2017). Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature*, **546**(7657), 248–253.
- Zhang, Z., Li, Y., Lin, B., Schroeder, M., and Huang, B. (2011). Identification of Cavities on Protein Surface Using Multiple Computational Approaches for Drug Binding Site Prediction. *Bioinformatics*, **27**(15), 2083–2088.
- Zibaee, S., Makin, O. S., Goedert, M., and Serpell, L. C. (2007). A simple algorithm locates  $\beta$ -strands in the amyloid fibril core of  $\alpha$ -synuclein, A $\beta$ , and tau using the amino acid sequence alone. *Protein Science*, **16**(5), 906–918.
- Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Comput Biol*, **4**, e1000140.
- Zvelebil, M. and Baum, J. (2008). *Understanding Bioinformatics*. Garland Science, Taylor & Francis Group, New York – London.



