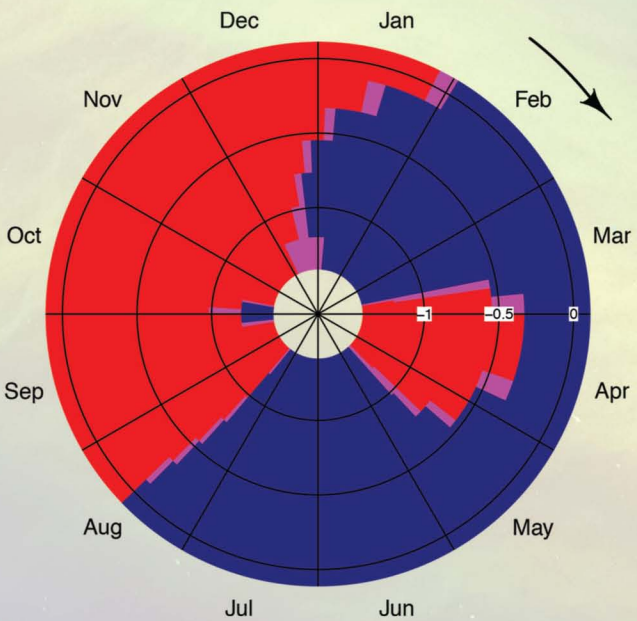
Chapman & Hall/CRC
Interdisciplinary Statistics Series

APPLIED DIRECTIONAL STATISTICS

MODERN METHODS AND CASE STUDIES



Edited by
Christophe Ley
Thomas Verdebout

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Applied Directional Statistics

Modern Methods and Case Studies

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

BAYESIAN ANALYSIS FOR POPULATION ECOLOGY
R. King, B. J.T. Morgan, O. Gimenez, and S. P. Brooks

DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION
D.L. Fairclough

MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS
J. P. Buonaccorsi

VISUALIZING DATA PATTERNS WITH MICROMAPS
D.B. Carr and L.W. Pickle

TIME SERIES MODELING OF NEUROSCIENCE DATA
Tohru Ozaki

FLEXIBLE IMPUTATION OF MISSING DATA
S. van Buuren

AGE-PERIOD-COHORT ANALYSIS
NEW MODELS, METHODS, AND EMPIRICAL APPLICATIONS
Y. Yang and K. C. Land

BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY, SECOND EDITION
A. B. Lawson

ANALYSIS OF CAPTURE-RECAPTURE DATA
R. S. McCrea and B. J.T. Morgan

MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION
S.Burgess and S.G.Thompson

POWER ANALYSIS OF TRIALS WITH MULTILEVEL DATA
M. Moerbeek and S.Teerenstra

STATISTICAL ANALYSIS OF QUESTIONNAIRES
A UNIFIED APPROACH BASED ON R AND STATA
F. Bartolucci, S. Bacci, and M. Gnaldi

MISSING DATA ANALYSIS IN PRACTICE
T. Raghunathan

SPATIAL POINT PATTERNS
METHODOLOGY AND APPLICATIONS WITH R
A. Baddeley, E Rubak, and R.Turner

CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION
S. Green, J. Benedetti, A. Smith, and J. Crowley

CORRESPONDENCE ANALYSIS IN PRACTICE, THIRD EDITION
M. Greenacre

STATISTICS OF MEDICAL IMAGING
T. Lei

CAPTURE-RECAPTURE METHODS FOR THE SOCIAL AND MEDICAL SCIENCES
D. Böhning, P. G. M. van der Heijden, and J. Bunge

THE DATA BOOK
COLLECTION AND MANAGEMENT OF RESEARCH DATA
Meredith Zozus

MODERN DIRECTIONAL STATISTICS
C. Ley and T. Verdebout

SURVIVAL ANALYSIS WITH INTERVAL-CENSORED DATA
A PRACTICAL APPROACH WITH EXAMPLES IN R, SAS, AND BUGS
K. Bogaerts, A. Komarek, E. Lesaffre

STATISTICAL METHODS IN PSYCHIATRY AND RELATED FIELD
LONGITUDINAL, CLUSTERED AND OTHER REPEAT MEASURES DATA
Ralitza Gueorguieva

APPLIED DIRECTIONAL STATISTICS
MODERN METHODS AND CASE STUDIES
Christophe Ley, Thomas Verdebout

For more information about this series, please visit: https://www.crcpress.com/go/ids

Applied Directional Statistics

Modern Methods and Case Studies

Edited by
Christophe Ley
Thomas Verdebout

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20180808

International Standard Book Number-13: 978-1-138-62643-0 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Ley, Christophe, author. | Verdebout, Thomas, author.
Title: Applied directional statistics / Christophe Ley and Thomas Verdebout.
Description: Boca Raton : CRC Press, 2018.
Identifiers: LCCN 2018010549 | ISBN 9781138626430 (hardback)
Subjects: LCSH: Mathematical statistics. | Circular data. | Spherical data.
Classification: LCC QA276 .L43 2018 | DDC 519.5--dc23
LC record available at https://lccn.loc.gov/2018010549

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Contributors

Toshihiro Abe
Nanzan University

Jose Ameijeiras-Alonso
University of Santiago de Compostela

Richard Arnold
Victoria University of Wellington

Jolien Cremers
Utrecht University

Rosa M. Crujeiras
University of Santiago de Compostela

Marco Di Marzio
Chieti-Pescara University

Stefania Fensore
Chieti-Pescara University

Jesper Foldager
University of Copenhagen

Jes Frellsen
IT University of Copenhagen

Eduardo Garciá-Portugués
Carlos III University of Madrid

Alan Gelfand
Duke University

Michael Golden
University of Oxford

Thomas Hamelryck
University of Copenhagen

Jotun Hein
University of Oxford

Giovanna Jona-Lasinio
Sapienza University of Rome

Peter Jupp
University of St. Andrews

Irene Klugkist
Utrecht University and Twente University

Francesco Lagona
University of Roma Tre

Kanti V. Mardia
University of Leeds
University of Oxford

Kees Mulder
Utrecht University

Agnese Panzera
Florence University

Thanh Mai Pham Ngoc
Université Paris-Sud
Université Paris-Saclay

Gianluca Mastrantonio
Polytechnic of Turin

Karim Oualkacha
Université du Québec à Montréal

Arthur Pewsey
University of Extremadura C**á**ceres

Louis-Paul Rivest
Université Laval

Alberto Rodríguez Casal
University of Santiago de Compostela

Ichiro Ken Shimatani
The Institute of Statistical Mathematics

Michael Sørensen
University of Copenhagen

Suvrit Sra
Massachussets Institute of Technology

Charles C. Taylor
University of Leeds

Editors

Christophe Ley
Ghent University

Thomas Verdebout
Université libre de Bruxelles

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction

Aim of the book

Directional statistics are concerned with data that are directions. The typical supports for directional data are the unit circle and unit (hyper-)sphere, or more generally Riemannian manifolds. The nonlinear nature of these manifolds implies that the classical statistical techniques and tools cannot be used to analyze directional data, and this has given rise to the research flow called directional statistics which has been particularly active over the past two decades.

The present book is intended to be a companion book to our manuscript *Modern Directional Statistics* published by Chapman & Hall/CRC Press in 2017. While the latter book mainly covers theoretical aspects of recent developments in the field, the present book is dedicated to methodological advances and treatments of various modern real data applications.

Content of the Companion Book *Modern Directional Statistics*

In a nutshell, we now summarize the material described in *Modern Directional Statistics*. It begins with a very detailed description of the recently proposed probability distributions for data on the circle, sphere, torus, and cylinder. The book then focuses on more inferential aspects such as nonparametric density estimation, quantile-/depth-based inference, order-restricted inference, rank-based inference, tests of uniformity and symmetry, among others. The Le Cam methodology adapted to directional supports is described in detail and theoretical applications presented. Finally, the book deals with high-dimensional inference on hyperspheres.

Content of the present book

Various modern application areas will be described in this book, as well as the new methods that have been developed to analyze the corresponding directional data. These areas include protein bioinformatics (Chapter 1 by Mardia, Foldager and Frellsen, as well as Chapter 4 by García-Portugués, Golden, Sørensen, Mardia, Hamelryck and Hein), the study of sea regimes (Chapter 3 by Lagona and Chapter 7 by Gelfand, Jona Lasinio and Mastrantonio), biology (Chapter 8 by Abe and Shimatani), the study of wildfires (Chapter 9 by Ameijeiras-Alonso, Crujeiras and Rodríguez Casal), social and behavioural sciences (Chapter 10 by Klugkist, Cremers and Mulder), and machine learning (Chapter 12 by Sra). Specific topics with applications in diverse domains have also been addressed: ambiguous rotations (Chapter 2 by Arnold and Jupp), inference under noisy data (Chapter 5 by Pham Ngoc), the modeling of rotation matrices (Chapter 6 by Rivest and Oualkacha), and nonparametric classification (Chapter 11 by Di Marzio, Fensore, Panzera and Taylor). Finally, Chapter 13 by Pewsey provides an overview of existing `R` packages that are relevant for directional data analysis.

We wish to thank all contributors to the present book which, we hope, will please the reader and provide further motivation to delve into the passionating field of directional statistics.

1

Directional Statistics in Protein Bioinformatics

Kanti V. Mardia

University of Leeds

Jesper Illemann Foldager

University of Copenhagen

Jes Frellsen

IT University of Copenhagen

CONTENTS

1.1 Introduction

Directional statistics has been applied in several different branches of biology [2], including the modelling of periodic properties in biological tissues [19], movement of organisms [47], and in the study of circadian rhythms, such as wake-sleep cycles [32]. In this chapter we will outline several usages of directional statistics in protein bioinformatics, which is a field dealing with the

modelling and prediction of the three-dimensional structure of proteins. We will first give the biochemical background for studying these biological macromolecules and then we will outline different approaches to molecular modelling making use of methods from directional statistics. We will conclude the chapter with a discussion on related research and open challenges.

1.2 Protein Structure

In this chapter we will be considering proteins, which are a type of biological macromolecule. Proteins are essential to all living cells and they are often called the workhorses of cells, due to their central roles in cellular structures and activities. The functionality of proteins mainly arises through their structure, and in this section we take a closer look at both the terms used to describe the different levels of structure and some general aspects of what constrains the three-dimensional structure of molecules.

Molecules are made up of atoms connected by covalent bonds. The structure of a molecule consists of the three-dimensional positions of its atoms. At biological temperatures the atoms of a molecule do not remain at fixed positions, but evolve over time. The dynamics of most molecules are ergodic and as a consequence the probability of seeing a specific three-dimensional configuration is the same when sampling a slice from a time trajectory of a single molecule, and sampling a single molecule from a pool of molecules at a specific time. Therefore, simulation studies are often done using only a single molecule. Some molecules are more dynamic and can be found in a wide range of different configurations, while other molecules are more constrained and only vibrate around a single three-dimensional configuration. The distribution depends both on the temperature and the chemical properties of the molecule and its surroundings.

Proteins are biopolymers constructed from a linear sequence of monomeric subunits. In proteins these subunits are *amino acids*, which are joined by covalent bonding to form a single macromolecule. There are 20 different amino acids in naturally occurring proteins. They are identical in the part involved in the polymerization forming the *backbone*, but differ by what is called the *side chain*. The amino acids are compactly represented by a single letter code using the letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. For instance, A is the code for Alanine and C is the code for Cysteine. In the following sections we will use glutamic acid as an example, which is abbreviated E. The specific sequence of amino acids differs between different proteins and is called the *primary structure* of the molecule, which can be represented by a string of characters. An example of a primary structure is shown in Figure 1.1(a).

(a) Primary structure

```
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR
```

(b) Secondary structure

```
CEEEEECCCCCEEEEECCCCCCHHHHHHHHHHHHCCCCCCEEEEECCEECCCCCCCCCCCCCCCCEEEEEEC
```

(c) Tertiary structure

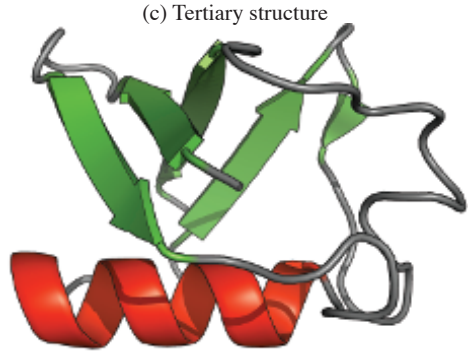


FIGURE 1.1
Protein structure is often divided into four levels, which here is illustrated with human ubiquitin [25]. The *primary structure* (a) is the linear sequence of amino acids, here shown by their one letter code. The *secondary structure* (b) is a classification of amino acids into local structural elements that they are part of. The secondary structure classification typically consists of three or eight classes; here we used the three classes helix (H) and strand (E) and coil (C). The *tertiary structure* (c) is the three-dimensional positions of all the atoms in the protein, which here is shown using a simplified cartoon representation. The cartoon is colored according to secondary structure, such that the helix is red, the strands are green, and coil regions are gray. The *quaternary structure* is the structure formed by multiple protein subunits (not shown here). The figure is reproduced from Frellsen [14].

Proteins typically fold into complex three-dimensional structures, as illustrated in Figure 1.1(c). The three-dimensional structure of a single protein is referred to as the *tertiary structure*. The tertiary structure mainly depends on the primary structure, i.e., the sequence of amino acids in the polymer, as the different amino acids have different biochemical properties. The function of a protein is normally dictated by its three-dimensional structure. However, experimentally determining the three-dimensional structure of a protein is typically a difficult and expensive process, and therefore substantial research efforts have been invested in developing computational methods for predicting the tertiary structure of proteins given their primary structure.

The individual subunits of proteins can be classified into reoccurring local structural patterns; this is called the *secondary structure*. Based on the hydrogen bond patterns in proteins each amino acid is classified into being a member of a helix, a sheet, or a coil as illustrated in Figure 1.1(b).

The covalent bonds in a molecule severely constrain the relative positions of the atoms. First of all, covalent bond lengths vary little, constraining the distance between bound atoms. Further, the geometry of the atoms covalently bound to a central atom is largely fixed, and depends on chemical properties that we will not describe here. In organic chemistry the most occurring geometry is that of a carbon atom with four elements bound, which forms a tetrahedron with the carbon atom at the center. There can be small deviations from a perfect tetrahedron, but this relatively fixed geometry translates to small variations in bond angles. The main degree of freedom in larger molecules, including proteins, comes from rotation around covalent bonds, which can be measured by dihedral angles [5]. The variations in the covalent bond angles and lengths are less significant, and sometimes fixed, ideal values are used when modelling proteins. In the following section (1.3) we will review the geometry and dihedral angles in proteins, and then we will return to the problem of structure prediction in Section 1.4.

1.3 Protein Geometry

The sequence of amino acids in a protein are linked together by *peptide bonds* between the carboxyl group of one amino acid and the amino group of the following, which form the *polypeptide backbone* of the protein. After amino acids form polypeptides the correct term for them is *amino acid residues* as the chemical groups they are named by have been modified. However, they are commonly still referred to as amino acids, which we will also do in this chapter when there is no ambiguity. As illustrated in Figure 1.2, the polypeptide backbone consists of a repeated sequence of three atoms: a nitrogen (N), a carbon (C_α), and another carbon (C). If we index the atoms by their position in the amino acid sequence, we can write the sequence of backbone atoms as

$$\mathrm{N}^{(1)} - \mathrm{C}_\alpha^{(1)} - C^{(1)} - \mathrm{N}^{(2)} - \mathrm{C}_\alpha^{(2)} - C^{(2)} - \cdots - \mathrm{N}^{(n)} - \mathrm{C}_\alpha^{(n)} - C^{(n)}.$$

In this sequence, the peptide bonds are $C^{(i)} - N^{(i+1)}$. The atoms of an amino acid that are not part of the backbone are part of the *side chain*, which is attached to C_α. The backbone is identical across all 20 amino acids, while the side chains are different, and it is this difference that gives the amino acids different biochemical properties. For each amino acid, i, there are three dihedral angles in the backbone:

- $\phi^{(i)}$ formed by $\mathrm{C}^{(i-1)} - \mathrm{N}^{(i)} - \mathrm{C}_\alpha^{(i)} - \mathrm{C}^{(i)}$,
- $\psi^{(i)}$ formed by $\mathrm{N}^{(i)} - \mathrm{C}_\alpha^{(i)} - \mathrm{C}^{(i)} - N^{(i+1)}$ and
- $\omega^{(i)}$ formed by $\mathrm{C}_\alpha^{(i)} - \mathrm{C}^{(i)} - N^{(i+1)} - \mathrm{C}_\alpha^{(i+1)}$.

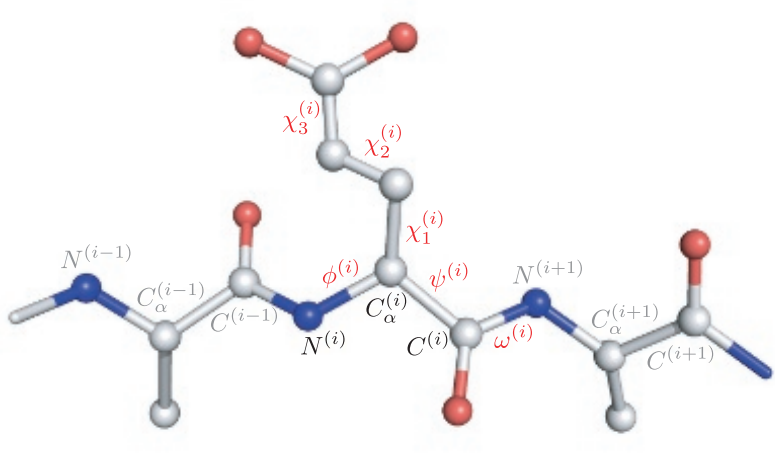


FIGURE 1.2
A protein fragment with selected atom and dihedral angle names. Atoms are colored by element: nitrogen, blue; carbon, gray; oxygen, red; and hydrogens are left out. Backbone atom names are in gray for the noncentral amino acids and black for the central amino acid. The central amino acid is a glutamic acid with dihedral angle names in red. The glutamic acid has the three dihedral angles in the side chain, χ_1, χ_2, and χ_3. The dihedral angles are shown on the covalent bond they specify rotation around. The figure is adapted from Frellsen [14].

The three dihedral angles of the backbone are illustrated in Figure 1.2.

The peptide bonds have a partial double-bond character, which means that the bond is planar and the ω angle is concentrated around two modes. The two modes are denoted *cis* isomer with $\omega \approx 0°$ and *trans* isomers with $\omega \approx 180°$. The main flexibility in the backbone is due to the rotational degrees of freedom of the ϕ and ψ angles. However, due to stereochemical properties of the polypeptide chain, only a limited number of conformations of these angles are energetically allowed. This is typically illustrated in a Ramachandran plot [50], a scatter plot where the ψ angle is plotted against the ϕ angle for all amino acids in a set of proteins. The Ramachandran plot from the original work by Ramachandran et al. [50] is shown in Figure 1.4. The Ramachandran plot is normally divided into energetically favorable and energetically disallowed regions, making it a useful tool in structure quality assessment. If most of the amino acids of a protein are not in the favorable regions, the model is likely in poor agreement with the real protein structure. Figure 1.3 shows the Ramachandran plot in the plan and on the two-dimensional torus.

There are between zero and four free dihedral angles in the side chain depending on the amino acid type. These angles are denoted χ_1, χ_2, χ_3, and χ_4,

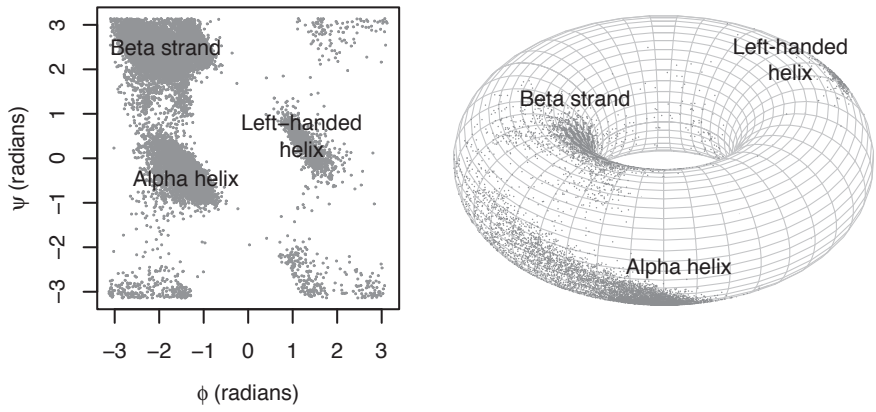


FIGURE 1.3
Two Ramachandran plots constructed from the Top 100 database [58]. The left is a scatter plot in the plane and the right is a scatter plot on the two-dimensional torus. In the torus plot, ϕ is the toroidal angle and ψ is the poloidal angle. Regions for right-handed α-helices, left-handed α-helices, and β-strands are shown on both plots. Figure reproduced from Mardia and Frellsen [38].

and they are also constrained to a limited number of allowed conformations. The side chain angles are shown in Figure 1.2.

1.4 Structure Determination and Prediction

There exist a number of experimental biophysical methods for determining the structure of proteins, where the two most predominantly used methods are X-ray crystallography and NMR spectroscopy. Both methods obtain measurements of physical quantities that can be used to infer the tertiary structure, often at atomic level resolution. Predicting the structure of proteins is a major research challenge in molecular biology, computational chemistry and bioinformatics. The problem can be formulated as: given the primary structure of a molecule (the sequence of amino acids), predict the secondary and the tertiary structure.

Due to very active research in the past decades, the secondary structure of proteins can be predicted with great accuracy. Traditionally, artificial neural networks have been used for secondary structure prediction [24], and today

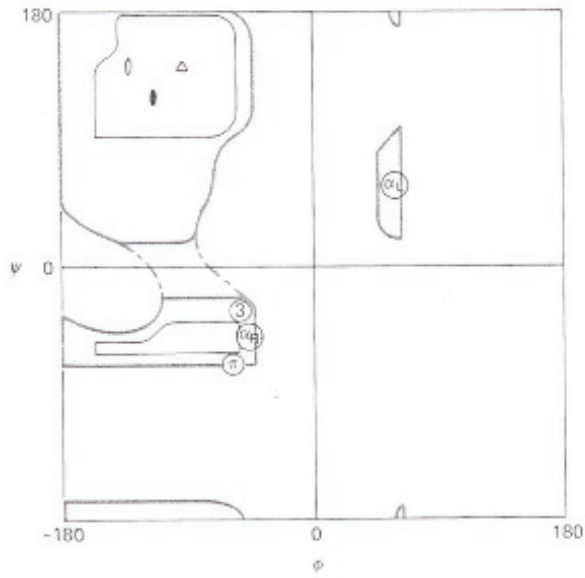


FIGURE 1.4
The original Ramachandran plot reproduced from Ramachandran et al. [50]. The regions are labelled according to secondary structure, including α_R for right-handed helix, α_L for left-handed helix, and the upper left region for strands.

the three classes of secondary structure can be predicted with more than 80% accuracy [57].

Tertiary structure prediction is a more challenging problem. If the molecule has a close sequence homolog (i.e., a molecule with shared ancestry) for which the structure is known, the tertiary structure of the molecule can often be well predicted by using the structure of the homolog as a scaffold. This approach is denoted *homology modelling* or *template-based modelling*, and if the sequence similarity between the two molecules is high the prediction can be quite accurate [59]. If no close sequence homolog exists, tertiary structure prediction is hard and this is the problem addressed by so-called *de novo* structure prediction.

De novo structure prediction methods often make use of a parametrized physical force field, which facilitates calculating the potential energy $U(\mathbf{x})$ of the molecule from the 3D Cartesian coordinates $\mathbf{x} \in \mathbb{R}^{3m}$ of all the m atoms in the molecule. Popular force fields include AMBER [49], CHARMM [7], and OPLS [30]. If we assume that the volume and temperature T of the system are constant, then according to statistical physics [13, 20, 51] the probability of observing a particular configuration $\mathbf{x}$ of the molecule can be expressed by

the Boltzmann distribution

$$p(\mathbf{x} \mid \beta) = \frac{\exp(-\beta U(\mathbf{x}))}{Z_\beta}, \tag{1.1}$$

where $Z_\beta = \int_{\mathbb{R}^{3m}} \exp(-\beta U(\mathbf{x}))\,\mathrm{d}\mathbf{x}$ is the normalization constant (partition function in physics), $\beta = (k_b T)^{-1}$ is the thermodynamic beta, k_b the Boltzmann constant, and T the temperature. In Equation (1.1) we left out the kinetic contribution, since it is trivial in Cartesian space [13].

The Boltzmann distribution can be used to make inference about the system. For instance, the tertiary structure(s) of the molecule can then be inferred from $p(\mathbf{x} \mid \beta)$ by finding the mode(s) of the distribution. The *Helmholtz free energy* at a given temperature can be calculated for the normalization constant as $F(\beta) = -\beta^{-1} \ln(Z_\beta)$ [13, 51]. Furthermore, given a function $g : \mathbb{R}^{3m} \to \mathbb{R}$, we may be interested in calculating the expectation

$$\mathbb{E}_\beta[g] = \int_{\mathbb{R}^{3m}} g(\mathbf{x}) p(\mathbf{x} \mid \beta)\,\mathrm{d}\mathbf{x}. \tag{1.2}$$

For instance we may want to find the mean potential energy $\mathbb{E}_\beta[U]$ at given β-value. This can be used to calculate the *thermodynamic entropy* of the system, $S(\beta) = \beta k_b(\mathbb{E}_\beta[U] - F(\beta))$ [13, 51].

However, for all nontrivial energy functions, inference in $p(\mathbf{x} \mid \beta)$ is analytically intractable. Molecular dynamics (MD) is a simulation-based method for probing $p(\mathbf{x} \mid \beta)$, which assumes that the system follows Newton's laws of motion and performs inference by numerically solving Newton's equations. These simulations assume that $\nabla_{\mathbf{x}} U(\mathbf{x})$ is readily available, which is usually the case. While MD methods have been very successful [53], their main disadvantage is that they are very computationally demanding. In practice, the time step size in an MD simulation is of the magnitude of femtoseconds (10^{-12}s), while proteins fold in the order of magnitudes of microseconds (10^{-6}s) to seconds (s). This means that simulating a single folding event with MD simulations requires millions to trillions of simulation steps.

1.4.1 Markov Chain Monte Carlo Simulations of Proteins

Monte Carlo (MC) based methods do not have the time scale limitation of MD. They work by drawing L samples $\{\mathbf{x}_\ell\}_{\ell=1}^{L}$ from $p(\mathbf{x} \mid \beta)$, and then approximating the integral $\mathbb{E}_\beta[g]$ in equation (1.2) by

$$\hat{I}_L(g) = \frac{1}{L} \sum_{\ell=1}^{L} g(\mathbf{x}_\ell). \tag{1.3}$$

It can be shown that $\hat{I}_L(g)$ is an unbiased estimator for $\mathbb{E}_\beta[g]$ and that it almost surely converges to $\mathbb{E}_\beta[g]$ as $L \to \infty$ [1]. This is known as the *Monte*

Carlo principle. The principle can also be used to obtain estimates of ratios of normalization constants: An unbiased estimate of

$$\frac{Z_{\beta'}}{Z_\beta} = \int_{\mathbb{R}^{3m}} \frac{\exp(-\beta' U(\mathbf{x}))}{Z_\beta}\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^{3m}} \frac{\exp(-\beta' U(\mathbf{x}))}{\exp(-\beta U(\mathbf{x}))} p(\mathbf{x} \mid \beta)\,\mathrm{d}\mathbf{x} \tag{1.4}$$

can be obtained by $\hat{I}_L\left(\frac{\exp(-\beta' U(\mathbf{x}))}{\exp(-\beta U(\mathbf{x}))}\right)$.

As mentioned before, it is usually not straightforward to efficiently generate samples from $p(\mathbf{x} \mid \beta)$. In Markov chain Monte Carlo (MCMC) this is resolved by constructing a Markov chain that has $P(\mathbf{x} \mid \beta)$ as stationary distribution. In principle this can be done using the Metropolis–Hastings (MH) algorithm [23]. In this algorithm a sequence of states $\{\mathbf{x}_\ell\}_{\ell=1}^L$ is generated one at a time. At the $(\ell+1)^{\text{th}}$ time step, a new state $\mathbf{x}'$ is sampled from a proposal distribution $q(\mathbf{x}' \mid \mathbf{x}_\ell)$ and then either accepted or rejected as the $(\ell+1)^{\text{th}}$ realization of the chain. The probability of accepting the proposed state $\mathbf{x}'$ is

$$\alpha(\mathbf{x}' \mid \mathbf{x}_\ell) = \min\left(1, \frac{p(\mathbf{x}' \mid \beta) \cdot q(\mathbf{x}_\ell \mid \mathbf{x}')}{p(\mathbf{x}_\ell \mid \beta) \cdot q(\mathbf{x}' \mid \mathbf{x}_\ell)}\right). \tag{1.5}$$

If the proposed state is accepted, then $\mathbf{x}_{\ell+1} = \mathbf{x}'$, otherwise the chain stays in the previous state, $\mathbf{x}_{\ell+1} = \mathbf{x}_\ell$.

A special case of MH is the Metropolis algorithm [41]. In this algorithm a symmetric proposal distribution is used, such that $\frac{q(\mathbf{x}_\ell|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}_\ell)} = 1$, which simplifies Equation (1.5). A proposal is then accepted if $\mathbf{x}'$ is at least as probable as $\mathbf{x}_\ell$, and otherwise $\mathbf{x}'$ is accepted with the probability $\frac{p(\mathbf{x}'|\beta)}{p(\mathbf{x}_\ell|\beta)}$. This choice of acceptance criterion still ensures the correct stationary distribution. When the proposal distribution is not symmetric, the ratio of proposals in (1.5) can be seen as a correction factor that allows the chain to have the correct stationary distribution no matter the choice of proposal distribution.

For most molecular systems the Markov chain constructed by the MH algorithm will not mix well, i.e., the sample will be highly correlated and the chain will only explore the relevant part of the sample space slowly. So in practice more advanced MCMC methods are used [6, 18], see for example the reviews by Iba [26], Murray [43], or Ferkinghoff-Borg [13].

One of the challenges in MCMC based simulation is designing a good proposal distribution $q(\mathbf{x}' \mid \mathbf{x}_\ell)$. Many methods [6, 27] assume that bond angles and bond lengths of the molecule are constant and represent the molecule by internal dihedral angles $\mathbf{\Phi} \in \mathbb{T}^p$, taking value on the p-dimensional hypertorus $\mathbb{T}^p = [-\pi, \pi)^p$. Changes to the molecule are then proposed as changes in dihedral angles and the proposal distribution takes the form $q(\mathbf{\Phi}' \mid \mathbf{\Phi}_\ell)$.

The most straightforward proposal distributions to use are concentrated Gaussian perturbations [27]. However, a proposal distribution is better when it is closer to the stationary distribution, and to take advantage of this, most proposal distributions incorporate protein structure information. A simple way to achieve this is by proposing angles, or stretches of angles, observed in

real proteins. Such methods are very successful, but come with the statistical problem that $q(\mathbf{\Phi}' \mid \mathbf{\Phi}_\ell)$ is not meaningful for continuous $\mathbf{\Phi}$. The solution is often to disregard the term $\frac{q(\mathbf{\Phi}_\ell|\mathbf{\Phi}')}{q(\mathbf{\Phi}'|\mathbf{\Phi}_\ell)}$ in equation (1.5), and abide that this changes the stationary distribution of the chain and therefore results in biased Monte Carlo estimators, c.f. Equation (1.3). The standard choices for such proposal distributions are to use fragment libraries for backbone angles [28, 29, 54], and rotamer libraries for side chain angles. Rotamer libraries exist in both backbone-independent version [34] and backbone-dependent version [11, 31, 52], where the frequency of each rotamer depends on the backbone angles (ϕ, ψ).

In the following sections we are going to review a number of tractable statistical models for describing the distributions over the dihedral angles in proteins. Ideally we would be interested in a fully tractable model for the conditional distribution of the dihedral angles in a protein $p(\mathbf{\Phi} \mid \mathbf{a})$ given the amino acid sequence $\mathbf{a}$. However, tractable models have only been developed for marginals or conditionals of this distribution.

1.5 Generative Models for the Polypeptide Backbone

Generative models for protein structure draw samples from the joint probability distribution of internal angles. As discussed in Section 1.3, the main degrees of freedom in the protein backbone are the angles ϕ, ψ, and ω. Due to the partial double bond character of the peptide bond, ω can be closely approximated by a discrete two-state variable, leaving most of the variation in the Ramachandran angles (ϕ, ψ). Any generative model either implicitly or explicitly defines the joint probability distribution over all the backbone dihedral angles, $p(\mathbf{\Psi})$, where $\mathbf{\Psi} \in \mathbb{T}^r$ and r is the total number of backbone angles. From a modelling perspective, the full joint distribution is difficult to work with directly without simplifying assumptions, both in terms of functional form and dependency structure. Reducing the problem to independent distributions of Ramachandran angle pairs on $\mathbb{T}^2$ is a good starting point for modelling. A generative model for protein structure must as a minimum recuperate the Ramachandran empirical distribution. However, such a generative model turns out to be very poor for predicting local protein structure, as there is a strong dependency between (ϕ, ψ)-angle pairs at different positions in the protein sequence. A tractable way of introducing dependency structure was introduced by Boomsma et al. [4] using a hidden Markov model, denoted TorusDBN, which encodes dependency along the sequence while any (ϕ, ψ)-pair remain conditionally independent from other angle pairs given the sequence of hidden (latent) variables.

1.5.1 Bivariate Angular Distributions

There are multiple options for the functional form of angle distributions involved in protein structures. Here we focus on bivariate forms, due to the strong dependency between ϕ and ψ observed in Ramachandran plots. Bivariate distributions can be used as a basic building block for more complicated models such as the generative model for protein backbones.

1.5.1.1 Bivariate von Mises

The von Mises distribution can be used for univariate angular data, and has the attractive feature that it is a close approximation to the wrapped normal distribution. By analogy to the normal distribution we would like a bivariate von Mises distribution with five parameters: two mean parameters and three parameters for variance and covariance. However, the "full" bivariate von Mises distribution introduced by Mardia [36] has eight parameters,

$$p(\phi, \psi \mid \mu, \nu, \kappa_1, \kappa_2, \boldsymbol{A}) \propto \exp(\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \left[\cos(\phi - \mu), \sin(\phi - \mu)\right] \mathbf{A} \left[\cos(\psi - \nu), \sin(\psi - \nu))\right]^{\mathsf{T}}), \quad (1.6)$$

where μ and ν are mean parameters, κ_1 and κ_2 are concentration parameters, and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is a two-by-two matrix. Different submodels of Equation (1.6) have been proposed, of which most attention has been given to the sine model by Singh et al. [55]

$$p_{\mathrm{s}}(\phi, \psi \mid \mu, \nu, \kappa_1, \kappa_2, \lambda) \propto \exp(\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \lambda \sin(\phi - \mu) \sin(\psi - \nu)), \quad (1.7)$$

and the cosine model explored by Mardia et al. [40]

$$p_{\mathrm{c}}(\phi, \psi \mid \mu, \nu, \kappa_1, \kappa_2, \kappa_3) \propto \exp(\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \kappa_3 \cos(\phi - \mu - \psi + \nu)). \quad (1.8)$$

Both submodels have five parameters and both approximate a normal distribution for higher concentrations. For details on their properties see Mardia et al. [40], Mardia and Jupp [39], Mardia and Frellsen [38], or Ley and Verdebout [33].

1.5.1.2 Histograms and Fourier Series

Initially models for (ϕ, ψ) were histogram based, which involves discretising $\mathbb{T}^2$ into regions and assuming equal density within a region. Histogram methods are very flexible, as the bin size can be arbitrarily small. However, the number of parameters is equal to the number of bins, and grows large for finer meshes. Besides the large number of parameters needed, a histogram using an adequately fine mesh would also suffer from a lack of precision due to the

limited number of available protein structures according to Pertsemlidis et al. [48]. Continuous models can have a more compact formulation, and have the inherent advantage of being smooth.

A simple unimodal distribution cannot be used to approximate the quite complicated multimodal Ramachandran empirical distribution, shown in Figure 1.3. Pertsemlidis et al. [48] proposed a parametrized continuous model for the distribution using two-dimensional Fourier series for the log likelihood, which can approximate any density function arbitrarily well by increasing the order of the series. The density across the identity lines was ensured to be continuous by using a period of 2π for the basis functions. A good approximation to the Ramachandran density required 80 parameters [48]. However, the parameters are not easily interpretable in a biological context.

1.5.1.3 Mixture of von Mises

Mardia et al. [40] suggested using a mixture of bivariate von Mises distributions for the Ramachandran empirical distribution, where either Equation (1.8) or (1.7) was the building block for the individual components. The graphical representation of the model is shown in Figure 1.6 and the mixture is defined as

$$p(\phi, \psi \mid \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \boldsymbol{\lambda}, \boldsymbol{w}) = \sum_i w^{(i)} p_{\mathrm{s}}(\phi, \psi \mid \mu^{(i)}, \nu^{(i)}, \kappa_1^{(i)}, \kappa_2^{(i)}, \lambda^{(i)}), \quad (1.9)$$

where i denotes the different components, $w^{(i)}$ is a positive weight such that $\sum_i w^{(i)} = 1$, and $p_{\mathrm{s}}(\cdot)$ is the sine based bivariate von Mises distribution (1.7), as it is used by Mardia [37]. A mixture can also be constructed using the cosine model (1.8) as seen in Mardia et al. [40] and in TorusDBN described later [4]. The mixture model needs multiple components to approximate the Ramachandran empirical distribution, but the components are consistent with partitions traditionally assigned to Ramachandran plots [40]. A maximum likelihood fit of the model can be obtained using the expectation-maximization (EM) algorithm [10], and a fitted model using the sine bivariate von Mises is seen in Figure 1.5.

A mixture model is also compatible with how the Ramachandran empirical distribution is thought to arise. In a protein the local context dictates and restricts the possible dihedral angles for a residue. If the possible contexts with reasonable accuracy can be discretised, then each discrete context corresponds to a mixture component, and the distribution over Ramachandran angles is obtained by marginalization. As we will see in the next section, this idea was used in a hidden Markov model, where a discrete hidden state encapsulates the contexts [4].

1.5.2 A Dynamical Bayesian Network Model: TorusDBN

Boomsma et al. [4] proposed a generative model for backbone angles that is similar to a hidden Markov model, but it has more emission nodes. This type

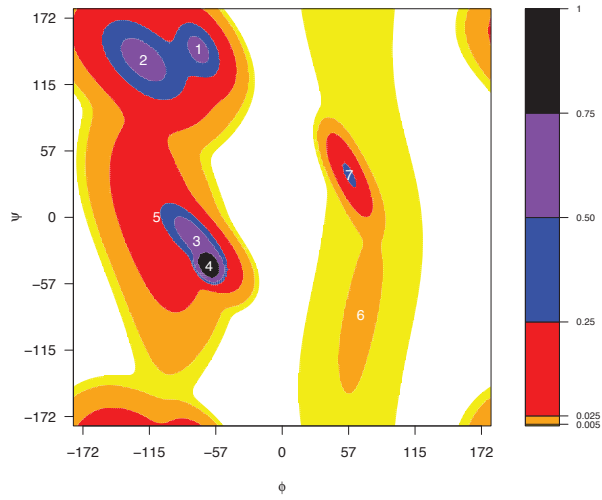


FIGURE 1.5
The Ramachandran plot for a mixture model, Equation (1.9), fitted to a subset of the Top 500 database [35]. The model is a mixture of seven bivariate von Mises sine components. Figure reproduced from Mardia [37].

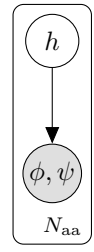


FIGURE 1.6
Graph representation of a mixture model. h is a discrete variable, each state emitting a distinct distribution for (ϕ, ψ). N_{aa} is the total number of amino acids.

of model belongs to the broader group of models called dynamical Bayesian networks [9, 42], however for most purposes it is simpler to consider it a hidden Markov model. The structure of the model can be seen in Figure 1.7. The model consists of a sequence of hidden nodes, h_i, each connecting to four emission nodes. The sequence represents the protein chain and N denotes the sequence length. Each hidden node is a discrete variable and can only take on

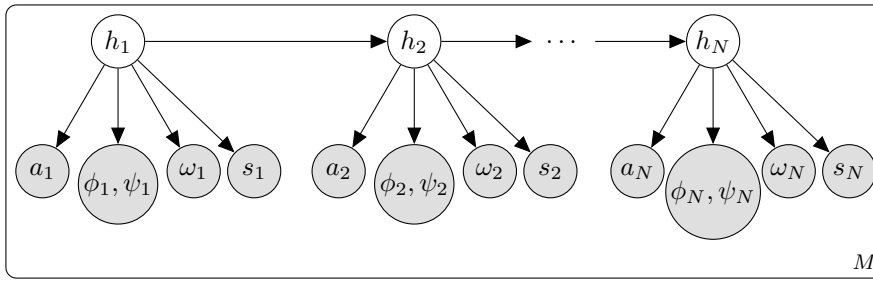


FIGURE 1.7
The independence structure for the protein backbone model TorusDBN by Boomsma et al. [4]. The lack of an arrow between two nodes indicates they are conditionally independent. Hidden nodes are connected to emission nodes for amino acid (a), dihedral angles (ϕ, ψ), cis/trans configuration (ω), and secondary structure (s). M is the total number of proteins and N is the length of individual proteins.

a limited number of states. Each hidden state corresponds to a distinct emission distribution for amino acid (a), Ramachandran angles (ϕ, ψ), cis/trans conformation (ω), and secondary structure (s). The model can be used in multiple ways; it can both be used for evaluating probabilities and generate samples for whole proteins or parts of a protein. When used as a generative model the hidden state sequence is sampled first, followed by sampling of the emission nodes. This method is also applicable for partial resampling, where nodes are resampled conditioned on the remaining fixed part of the sequence.

If some of the emission nodes are known a priori, they can be used to inform the sequence of hidden states. In protein structure prediction the amino acid sequence is known, and using Bayes theorem the hidden state sequence can be sampled conditioned on the observed amino acid sequence. This can be done for any of the emission nodes making the model ideal for generating proposals in MCMC sampling.

The factorization of the joint probability distribution can be read directly from the directed acyclic graph (DAG) in Figure 1.7. Each node contributes with the probability of the node itself conditioned on any input nodes. Starting with the top node h_1 the full factorization reads

$$\begin{aligned} p(\boldsymbol{a}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{h}) = p(h_1)p(a_1 \mid h_1)p(\phi_1, \psi_1 \mid h_1)p(\omega_1 \mid h_1)p(s_1 \mid h_1) \\ p(h_2 \mid h_1)p(a_2 \mid h_2) \dots p(s_2 \mid h_2) \\ \vdots \\ p(h_N \mid h_{N-1})p(a_N \mid h_N) \dots p(s_N \mid h_N). \end{aligned} \tag{1.10}$$

In the TorusDBN model, the transition probabilities $p(h_i \mid h_{i-1})$ are assumed to be a categorical distribution, which means that the parameters can be described by a $\mathbb{R}^{K \times K}$ transition matrix, where K is the number of hidden states.

Similarly, the emission probabilities for the amino acid $p(a_i \mid h_i)$ and the secondary structure $p(s_i \mid h_i)$ are categorical distributions. By assuming that the ω-angle is highly concentrated in the two modes, the cis/trans conformation probability $p(\omega_i \mid h_{i-1})$ can be assumed to be binomial. Finally, the probability of the Ramachandran angles $p(\phi_i, \psi_i \mid h_i)$ is assumed to be a cosine model bivariate von Mises distribution.

As implied by the name, the hidden node sequence is not directly observable, and when evaluating the probability of a protein structure the marginalized distribution is used,

$$\begin{aligned} p(\boldsymbol{a}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{s}) &= \sum_{h} p(\boldsymbol{a} \mid \boldsymbol{h}) p(\boldsymbol{\phi}, \boldsymbol{\psi} \mid \boldsymbol{h}) p(\boldsymbol{\omega} \mid \boldsymbol{h}) p(\boldsymbol{s} \mid \boldsymbol{h}) p(\boldsymbol{h}) \\ &= \sum_{\boldsymbol{h}} \prod_{i} p(a_i \mid h_i) p(\phi_i, \psi_i \mid h_i) p(\omega_i \mid h_i) p(s_i \mid h_i) p(h_i \mid h_{i-1}), \end{aligned} \tag{1.11}$$

where we conveniently define $p(h_1 \mid h_0) = p(h_1)$ since h_1 has no incoming edges. The calculation scales poorly as $\boldsymbol{h}$ has K^N possible states where K is the number of possible states for each hidden node, and N is the length of the sequence. Fortunately the complexity can be reduced by using the forward-backward algorithm that takes advantage of dynamic programming, see, e.g., Durbin et al. [12]. Similarly, the forward-backtrack algorithm can be used for efficient sampling [8].

The Markov property makes for a simple model but comes at the cost of a short memory. A larger state space could make up for this, but would not scale well. The model can generate protein structures that locally are protein-like which combined with the ability to evaluate exact probabilities makes it a perfect proposal distribution for MCMC sampling.

The parameters of the TorusDBN model can be estimated from data. A maximum likelihood estimate of the parameters can be obtained using the EM algorithm [10] or the stochastic EM algorithm [45].

1.6 Generative Models for Amino Acids Side Chains

In the previous section, we described generative models for the protein backbone. In this section we will review generative models for the amino acids side chains. As mentioned earlier, the main degrees of freedom in the amino acids side chains are the dihedral angles denoted χ_1, χ_2, χ_3, and χ_4. The number of angles varies between amino acids, for example, glutamic acid has three χ-angles, as illustrated in Figure 1.2.

Here we will consider a continuous model for the amino acid side chain angles called BASILISK [22]. The independence assumptions in this model follow an input output hidden Markov model (IOHMM) structure [3, 15, 42], which has previously also been used successfully for modelling the dihedral angles in