Peptide and Protein Design for Biopharmaceutical Applications

Editor

Knud J. Jensen
*Faculty of Life Sciences, University of Copenhagen,
Denmark*

WILEY
A John Wiley and Sons, Ltd, Publication

Peptide and Protein Design for Biopharmaceutical Applications

Peptide and Protein Design for Biopharmaceutical Applications

Editor

Knud J. Jensen
Faculty of Life Sciences, University of Copenhagen,
Denmark

WILEY
A John Wiley and Sons, Ltd, Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of fitness for a particular purpose. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for every situation. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

The cover figure showing Cilengitide was kindly provided by Horst Kessler and Oliver Demmer.

Library of Congress Cataloging-in-Publication Data

Peptide and protein design for biopharmaceutical applications / editor, Knud J. Jensen.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-31961-1
1. Peptide drugs—Design. 2. Protein drugs—Design. 3. Peptides—Design.
4. Protein engineering. I. Jensen, Knud J.
[DNLM: 1. Peptides—chemistry. 2. Proteins—chemistry. 3. Drug Design.
QU 68 D4575 2009]
RS431.P38D47 2009
615′.19—dc22

2009019177

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-31961-1 (H/B)

Set in 10.5/13pt Sabon by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by TJ International Ltd., Padstow, Cornwall

Contents

List of Contributors

Jesper Brask, Novozymes A/S, 6Bs.98, Krogshøjvej 36, DK-2880 Bagsværd, Denmark

Oliver Demmer, Institute for Advanced Study at the Department of Chemistry, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany

Andreas O. Frank, Institute for Advanced Study at the Department of Chemistry, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany

Thomas Hoeg-Jensen, Novo Nordisk A/S, DK-2760 Maaloev, Denmark

Knud J. Jensen, Faculty of Life Sciences, University of Copenhagen, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Copenhagen, Denmark

Horst Kessler, Institute for Advanced Study at the Department of Chemistry, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany

Veronique Maes, Organic Chemistry Department, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium

Garland R. Marshall, Center for Computational Biology, Department of Biochemistry and Molecular Biophysics, Washington University, 700 South Euclid Ave., St. Louis, MO 63130, USA

Gregory V. Nikiforovich, MolLife Design LLC, 751 Aramis Drive, St. Louis, MO 63141, USA

Dirk Tourwé, Organic Chemistry Department, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium

Preface

Ever since the discovery of the therapeutic value of insulin, at the beginning of the twentieth century, peptides have been successfully applied as drugs. In the fifties and sixties, with the advent of new methodologies for isolation, identification and total synthesis, the range of peptide hormones was further expanded and additional peptide drugs were launched. Since then, a generally upward trend has been seen, despite occasional statements to the contrary. Improvements in techniques have allowed the identification of many naturally occurring peptides, which have provided the starting point for the design of peptide drug candidates. In addition, *de novo* design has emerged as a new approach for the invention of peptide drug candidates. Designed peptides and small proteins have become ubiquitous tools for biochemical and biophysical studies. The latter studies have had a significant impact on the design of peptides as potential drug candidates, and are thus to some extent covered in the present volume.

This book comprehensively presents central topics in the design of peptides and proteins, especially those with the goal of biopharmaceutical applications. It starts with an outline of computational methods, then moves on to cyclic peptides, which are often important in the development of peptide drug candidates; it provides an overview of peptidomimetics, carbohydrates in the design of peptides and proteins, *de novo* design of proteins, and finally, as a key example, the design of new insulin variants. This book is aimed at peptide scientists in academia and in industry, as well as at graduate students entering the field.

I wish to express my sincere gratitude to Gregory Nikiforovich, Garland Marshall, Horst Kessler, Oliver Demmer, Andreas O. Frank, Dirk Tourwé, Veronique Maes, Thomas Hoeg-Jensen and my former PhD student Jesper Brask for their excellent contributions. I am pleased to acknowledge Paul Deards and Richard Davies from Wiley for their support and guidance during the process. It is a privilege to be able to acknowledge Laura Quartara for her advice throughout this project.

1

Introduction

Knud J. Jensen

The aim of this book is to provide a comprehensive introduction to the concepts and methods behind the design of peptides and small proteins. The individual chapters are written by experts in each field. We have striven to coordinate the chapters to create coherence in the book. Inevitably, there is some constructive overlap between the topics in the chapters, and the chapters refer to one another.

Chapter 2, ‘Computational Approaches in Peptide and Protein Design: An Overview’, by Gregory V. Nikiforovich and Garland R. Marshall, provides a comprehensive overview of computational approaches to the modelling of peptides and proteins. This chapter surveys computational methods and principles as well as some of the available software. The authors illustrate their points with specific examples, such as the design of cyclopentapeptides as inhibitors of CXCR4. Here they describe the conformational study of both the cyclopentapeptides, as a 3D pharmacophore model for FC131, and of the G-protein-coupled receptor CXCR4, where they build on a 3D model of the transmembrane region of CXCR4. They then discuss docking of the peptide FC131 to CXCR4.

Chapter 3, ‘Aspects of Peptidomimetics’, by Veronique Maes and Dirk Tourwé, provides an overview of a hierarchical approach to peptidomimetic design. This includes the role of cyclic peptides in the development of peptidomomimetics, referring to Chapter 4. Maes and Tourwé then

Peptide and Protein Design for Biopharmaceutical Applications Edited by Knud J. Jensen

describe the concept of retroinverso structures and deliver an extensive overview of backbone modifications, before discussing side-chain constraints. They describe peptoids and secondary structure mimetics – a topic that is taken up again in Chapter 6 – as well as topomimetics. An important aspect of Chapter 3 is the examples given of modifications of peptide hormones, especially of somatostatin, as peptidomimetics. The chapter also covers a range of protease inhibitors.

Chapter 4, ‘Design of Cyclic Peptides’, by Oliver Demmer, Andreas O. Frank and Horst Kessler, provides a comprehensive overview of its topic. It starts with naturally-occurring cyclic peptides (cyclosporin A, for example) and moves on to different ways of cyclizing peptides. Then some backbone modifications are discussed, a theme covered in Chapter 3, as well as other modifications of cyclic peptides. A central part of the chapter is a description of the conformation and dynamics of cyclic peptides, especially the reduction in conformational space. The authors describe turn structures in cyclic peptides and concepts in the rational design of cyclic peptides, leading to the outline of a general strategy for finding active hits. The text exemplifies this with the development of the peptide drug candidate Cilengitide as an integrin inhibitor and CXCR4 antagonist.

Chapter 5, ‘Carbohydrates in Peptide and Protein Design’, by Jesper Brask and the editor, describes how carbohydrates are used to introduce new structural and conformational features to peptides and proteins. The topics in this chapter include sugar amino acids, cyclodextrins and carbohydrates as templates in the design of peptides and proteins.

Chapter 6, ‘*De Novo* Design of Proteins’, by the editor, gives an overview of concepts in the design of proteins from general principles, rather than through a redesign of natural structures. The focus is on structural aspects, especially rules for the design of secondary structural elements such as α-helical peptides, and the assembly of these into tertiary structures. Some *de novo* turn motifs, used to connect the secondary structural elements, are also included. The chapter features an introduction to foldamers, especially β- and γ-peptides. It ends with examples of biopharmaceutical applications of *de novo* design.

Chapter 7, ‘Design of Insulin Variants for Improved Treatment of Diabetes’, by Thomas Hoeg-Jensen, provides a comprehensive overview of the classical therapeutic peptide hormone insulin. The focus is on insulin as a modern biopharmaceutical drug and the development of new insulin variants with modulated therapeutic profiles, e.g. prolonged-acting vs. fast-acting insulins, either by modifications in the

51 AA structure or by appending moieties. Novel glucose-sensitive insulins and insulin mimetics are also covered.

As mentioned above, there is some constructive overlap between chapters. We have striven to make the index a powerful tool in accessing topics across chapters.

2

Computational Approaches in Peptide and Protein Design: An Overview

Gregory V. Nikiforovich and Garland R. Marshall

2.1 INTRODUCTION

Over the last decade, new examples of applications of computational methods to the design of peptides and proteins appeared in the literature literally every week, if not every day. Among the factors that contributed to this growth of computational design studies are the rapidly evolving capacity of computer systems, the availability of software packages for molecular modelling – both commercial and freeware – and ease of access to multiple databases, such as the Cambridge Structural Database (CSD), the Protein Data Bank (PDB) and Swiss-Prot/TrEMBL. Even more important, perhaps, is that nowadays there is almost universal agreement that computational approaches, along with experimental methods, are indispensable components of the general pipeline of drug design. Fifteen years ago, doubts about the potential practicality of computational methods were still widespread among the drug design community (see our earlier reviews on the subject [3,4]).

Today, computational approaches cover the wide field ranging from suggesting plausible 3D structures of short oligopeptides in solution,

Peptide and Protein Design for Biopharmaceutical Applications Edited by Knud J. Jensen

through determining the peptide sequences most suitable for performing certain biological functions, to *de novo* predictions of large proteins interacting with one another. In fact, many of the most recent examples are described and reviewed in later chapters of this book. Therefore, we do not feel a real need to review the current state of computational approaches in detail. Instead, this chapter focuses only on the most general problems of computational design of proteins and especially peptides, as well as on the basic techniques required for the practical implementation of design methods. Our main goal is to introduce the basic elements of the computational approaches to those interested in peptide and protein design, as well as to share some of our thoughts and reflections on the current state of the field. In the first part of the chapter, we discuss computational tools and procedures connected with conformational flexibility of peptides and proteins; in our opinion, these problems can be satisfactorily resolved only by applying computational approaches. The second part contains a recent example of the application of these tools to the specific task of studying possible complexes between the CXCR4 G-protein-coupled receptor (GPCR) and its cyclopentapeptide inhibitors [6,7].

2.2 BASICS AND TOOLS

2.2.1 The Importance of Computational Approaches

From a very general point of view, most of the problems of modern peptide and protein drug design fall into two main areas: structure-based and target-based design. In the first case, one starts from a parent peptide ('ligand') with no details of the structural information on its specific target ('receptor'). In the second case, structural information on the receptor, at least on the receptor site that binds the ligand, is available at varying resolution. In both cases, the goal is to suggest compounds that would exhibit certain biological qualities (affinity, activity, etc.) as well as or better than the parent ligand. In structure-based design, the most essential requirement is to determine the 3D arrangement of the functional groups of ligand comprising the so-called '3D pharmacophore', which is responsible for ensuring correct interaction between the ligand and the receptor. In target-based design, one also aims to determine the correct binding mode of each ligand within the binding site of the common receptor.

However, in the frame of structure-based design, the available experimental methods of structural determination often fail to determine possible 3D pharmacophores characteristic for the interaction of the peptide ligand with its specific receptor. The apparent reason is the inherent conformational flexibility of peptides. Most peptides exist under physiological conditions as a mixture of more or less well-defined, interconverting conformers. The interconversion rate is such that, for instance, NMR spectroscopy with characteristic resolution times of 10^{-5}–10^{-3} seconds does not distinguish separate conformers of linear peptides in the kilodalton range in solution (with the exception of *cis/trans* peptide bond isomers). The same is true for CD, IR and ESR spectroscopy. In the absence of one highly predominant conformer, the 3D peptide structure deduced from physicochemical measurements (e.g. from NMR parameters such as NOEs, vicinal coupling constants, etc.) reflects the average over the ensemble of conformers present in solution and, in this sense, could not be related to any of the 'real' peptide conformers at all. On the other hand, the conformation of peptide ligands corresponding to the 3D pharmacophore, i.e. to the ligand conformation in the complex with receptor, may not necessarily be the one with the highest statistical weight in solution, since some other conformers may acquire the highest statistical weight in the peptide–receptor complex, being compensated by much more favourable interaction in the complex with the receptor. At the same time, X-ray crystallography produces only individual 'snapshots' of peptides, each representing a single 3D structure stabilized by the crystalline lattice from among the set of possible conformers existing in solution. For the highly flexible enkephalin molecule, for instance, X-ray crystallography obtained snapshots of the four drastically different 3D structures ranging from fully extended to various types of β-reversals (see review [8]).

On the other hand, computational methods, being applied to various analogues of the same peptide that differ by values of affinity (or activity) toward a specific receptor, may model all 3D structures feasible for the parent peptide and its analogues from the energetic and/or sterical point of view. Then one may compare sets of those structures to one another and select those among the biologically active analogues in which the important functional groups are arranged in space similarly. These structures may be regarded as reasonable candidates for 3D pharmacophores, which in turn may be stabilized by introducing constraints through chemical synthesis. The structure-based design employing this approach has been successful in developing novel cyclic analogues of linear peptides, many of which are biologically active (such as analogues of opioid peptides, angiotensin, α-melanotropin, etc.; see earlier review [3]).

Historically, target-based design came after structure-based design, since detailed information on the receptor molecules only became readily available in recent decades. This progress was made largely due to rapid development of technology for co-crystallization of ligand–receptor complexes (especially enzymes and their substrates/inhibitors), as well as advances in X-ray crystallography of proteins. Seemingly, experimental determination of ligand–receptor complexes by X-ray crystallography abolishes the need for computational approaches. Indeed, detailed information on both the recognition motif (3D pharmacophore) and the binding mode of the ligand within the receptor is readily available from the X-ray structure of the complex. In reality, however, there are several important limitations that still require emphasis on computational approaches in target-based design.

First, many biologically active peptides (and over 30% of drugs in clinical use [9]) act through interactions with GPCRs, which are integral membrane proteins that include seven transmembrane helical stretches (TM helices) connected by loops that form the intracellular (IC) and extracellular (EC) domains, together with the fragments containing the N- and C-termini. Being membrane proteins, GPCRs are extremely difficult to express and to extract from the membrane in quantity, and have resisted chemical synthesis. Accordingly, X-ray structures are known presently only for four GPCRs, namely the photoreceptor rhodopsin, the β2- and β1-adrenergic receptors and the A2A adenosine receptor [10,11, 114, 115]. Therefore, the only way to address ligand–receptor interactions involving other GPCRs (the largest human gene family) for target-based design is to apply computational approaches to model (either by homology or by *de novo* approaches) the 3D structure of GPCR.

Second, while many ligand-receiving sites in receptors feature more or less well-defined 3D ‘pockets’ inside the protein globule (as, for example, in enzymes), other recognition sites are formed by flexible loops protruding away from the bulk of the protein (as, for example, in antibodies or GPCRs). In the latter case, one faces the same problem as in determining the 3D structure of a flexible peptide: namely, even if the X-ray structure of the loops in the receptor–ligand complex is resolved, the X-ray snapshot may capture only a single specific conformation out of several that may be more characteristic for the given complex and more representative of a functional complex. For instance, the conformations of the IC loops connecting TM helices in the five X-ray structures of rhodopsin published so far drastically differ from one another. Again, determining the set of plausible conformations of the loops in question requires use of computational modelling.

Third, depending on the shape and rigidity of the receiving site of the receptor, the binding modes of the ligand also may be nonunique. That may be especially important for small ligands with medium affinity toward the receptor (e.g. in micromolar range). In this case, again, the X-ray snapshot of the receptor cocrystallized with the ligand may not represent the optimal binding mode, nor the ensemble of binding conformations, for the ligand. In these cases, the binding modes have to be refined by computational sampling of various spatial positions of the ligand within the receiving site of the receptor to sample orientations in which the ligand binds the receptor more tightly. The roles of entropy and enthalpy of complex formation are not independent and simple comparisons of affinity (ΔG) without dissection into its components, ΔH and ΔS, can be misleading [12].

These considerations illustrate the point that in peptide and protein design, whether structure-based or target-based, there are certain problems for which we require computational modelling. At the same time, one should not overestimate the precision of current computational results. For typical applications, computational approaches generate plausible suggestions regarding structural aspects of the functional recognition of peptides and proteins. In all cases, these suggestions have to be independently validated either by direct experimental structural measurements or by confirmation of predictions through biological experiments.

2.2.2 Tools and Procedures: Force Fields and Sampling

General protocols of any computational approach to peptide and protein design regularly include two key aspects: the generation of possible molecular conformations and relative orientations of interacting molecules (sampling) and the evaluation of the plausibility of the generated conformations or orientations in terms of their relative energies (scoring). The more thorough the sampling protocol and the more accurate the scoring function, the more reliable the predictions. Ultimately, the best results may be obtained when all possible states of a system in question (conformations and relative orientations) are sampled and the energy for each of the states is calculated employing high-level quantum calculations. However, this best-case scenario is seldom applicable in peptide and protein design, simply because of the system size, which overwhelms available computer resources. Even with current rapid expansion of computer capacity, it is unrealistic to expect adequate quantum chemical calculations for, say, a linear octapeptide in water – the system featuring thousands of possible conformations of the peptide and millions of configurations of the solvent – within the next decade.

2.2.2.1 Force fields

Atom–atom force fields currently in use: validation and applicability In the so-called Born–Oppenheimer approximation, interactions within molecular systems are limited to atom–atom interactions that allow the estimation of energies for various states of peptides and proteins, with an accuracy sufficient for many practical applications. In this approximation (molecular mechanics, MM), peptide molecules are considered systems of points (atoms) in space, which interact with each other by different types of forces. The forces can be divided into two main classes, namely those between atoms that are bonded and those that are nonbonded in the valence structure. Usually, the forces between nonbonded atoms include at least two terms: van der Waals and electrostatic forces. A special term describing hydrogen bonding between corresponding atoms is often included, as well as an additional dihedral angle ('torsional') term. The forces between bonded atoms include bond stretching, valence angle bending and improper dihedral angle forces. Summarily, molecular energy calculated with a typical atom–atom force field in MM may be expressed as follows:

$$V(r) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[cos(n\phi + \delta) + 1] + \sum_{\substack{nonbond\\pairs}} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \right] \tag{2.1}$$

Generally, all forces depend on the distance between interacting atoms, and on the parameters selected for each term. There is no one single atom–atom force field universally adapted as standard for calculating energies in peptides and proteins. Several force fields are currently used for this purpose, differing mostly in the sets of parameters. The parameters are usually selected to fit the experimental data on crystal packing of amides and amino acids, such as in AMBER (Assisted Model Building with Energy Refinement) [13,14] and ECEPP (Empirical Conformational Energy Program for Peptides) [15–17], or on properties of organic liquids, such as in OPLS (Optimized Potentials for Liquid Simulations) [18,19] and GROMOS (GROningen MOlecular Simulation) [20], as well as to fit the results of quantum chemistry calculations, such as in AMBER, OPLS and CHARMM (Chemistry at HARvard Molecular Mechanics) [21]. Historical development of atom–atom force fields was strongly impacted by the availability of computer resources. The oldest force fields, such as ECEPP, employ rigid valence geometry and therefore do not include the first two terms of the potential expression, which results in significant reduction of the computer time needed for energy calculations. All of the other force fields include flexibility of valence geometry. Options are to consider all hydrogens separately (slower calculations) or consider aliphatic hydrogens united with their bonded carbon atoms (united-atom assumption; speedier calculations). The force fields with flexible valence geometry are utilized also for other classes of biochemical compounds, such as nucleic acids, carbohydrates, etc. Specifically, force fields for peptides and proteins have been reviewed many times, emphasizing different aspects of their applications; the reader is referred to the recent excellent reviews by Ponder and Case [22] and Mackerell [23].

It is difficult to determine which force field is preferable for conformational calculations involving peptides and proteins. On the one hand, obviously, the force fields with flexible valence geometry and those calibrated to fit the results of quantum chemistry calculations are more likely to yield an accurate value of energy for a given state of a peptide system including solvent. But on the other hand, computational approaches are especially valuable for problems involving conformational flexibility of peptide chains and orientations of ligands within the binding site of the receptor. In both cases, estimations of energy for a large number of possible states of the system are required in order to select the most plausible states as fast as possible with as few as possible false positives and false negatives, providing some justification for the more computationally efficient approximations.

In this regard, a convenient test utilizes reconstruction of the Ramachandran map for the simplest element of the peptide chain, the acetyl-N-Methyl-L-alanine (Ac-Ala-OMe), by various force fields. The Ramachandran map is a function of the two torsional angles, ϕ(rotation around the bond NH-C^{α}) and ψ(C^{α}-CO), adjacent to the α-carbon that maps the potential surface of peptide backbone conformations. The Ramachandran map can be roughly divided into four quadrants: the upper-left, corresponding to ($\phi < 0°$, $\psi > 0°$); the lower-left ($\phi < 0°$, $\psi < 0°$); the upper-right ($\phi > 0°$, $\psi > 0°$); and the lower-right ($\phi > 0°$, $\psi < 0°$). The upper-left quadrant contains the (ϕ, ψ) points corresponding to an extended structure, such as a β-strand ($\phi \sim -140°$, $\psi \sim 140°$), and the lower-left and upper-right quadrants contain points corresponding to the right- ($\phi \sim -60°$, $\psi \sim -60°$) and left-handed ($\phi \sim 60°$, $\psi \sim 60°$) α-helices, respectively. The upper-left quadrant also contains the (ϕ, ψ) points ($-75°$, $140°$) and ($-80°$, $80°$), corresponding to conformations P_{II} (polyproline II) and $C^{7}{}_{eq}$ (the inverted γ-turn). The lower-right quadrant contains the (80°, −80°) point that corresponds to the C^7_{ax} conformation (the γ-turn). At this rough 'quadrant' approximation, all quadrants except the lower-right are considered sterically allowed for the L-amino acid residues. All four quadrants are allowed for Gly residues but only two of them, namely those at the left side of the plot, are allowed for L-Pro residues. The Ramachandran maps for L- and D-amino acid residues are symmetrical with respect to rotation by 180° around an orthogonal axis at the centre of the map.

Earlier calculations of the Ramachandran map for Ac-Ala-OMe using CHARMM and AMBER, but not those using ECEPP, showed that conformation C^7_{ax} possessed relative energies close to those of C^7_{eq}, the conformation with the lowest energy [24]. Moreover, the CHARMM and AMBER maps showed fairly large regions of energetically-allowed conformations in the lower-right quadrant, which was contradictive to data on the X-ray structures of amino acid residues in proteins available at the time (1989). It was argued that the calculations were performed without proper account for solvent and, in fact, modelled the Ac-Ala-OMe in the gas phase, for which no experimental data exists. On the other hand, high-level quantum calculations for Ac-Ala-OMe also found that conformation C^7_{ax} possesses energy close to C^7_{eq} and lower than conformations corresponding to the right- or left-handed α-helical structures (e.g. [25]).

Recently, the Ramachandran map for Ac-Ala-OMe was extensively sampled by molecular dynamics simulations employing several force fields with flexible valence geometry [1]. Simulations included interactions with water molecules described explicitly. Figure 2.1(a)–(d) depicts the Ramachandran maps obtained with the AMBER, CHARM22, GROMOS and OPLS-AA force fields, respectively. Figure 2.1(e) depicts the distribution of the (φ, ψ) positions for each of 97 368 residues derived from 500 high-resolution X-ray structures of proteins [2]; to avoid bias, only the data for residues not involved in regular secondary structures such as β-strands or α-helices were included in the map in Figure 2.1(e). Assuming that the distribution of the X-ray data on residues in proteins is a good approximation of the general distribution of plausible (φ, ψ) values for peptides and proteins, one can utilize this distribution to evaluate the relative validity of various force fields for computational studies of peptides and proteins.

Experimental data presented in Figure 2.1(e) suggest that most plausible conformations of the L-amino acid residues in proteins are concentrated into three regions. The largest and most populated region is located in the upper-left quadrant of the Ramachandran map, encompassing conformations corresponding to the β-strand and P_{II} (the 'beta' region).

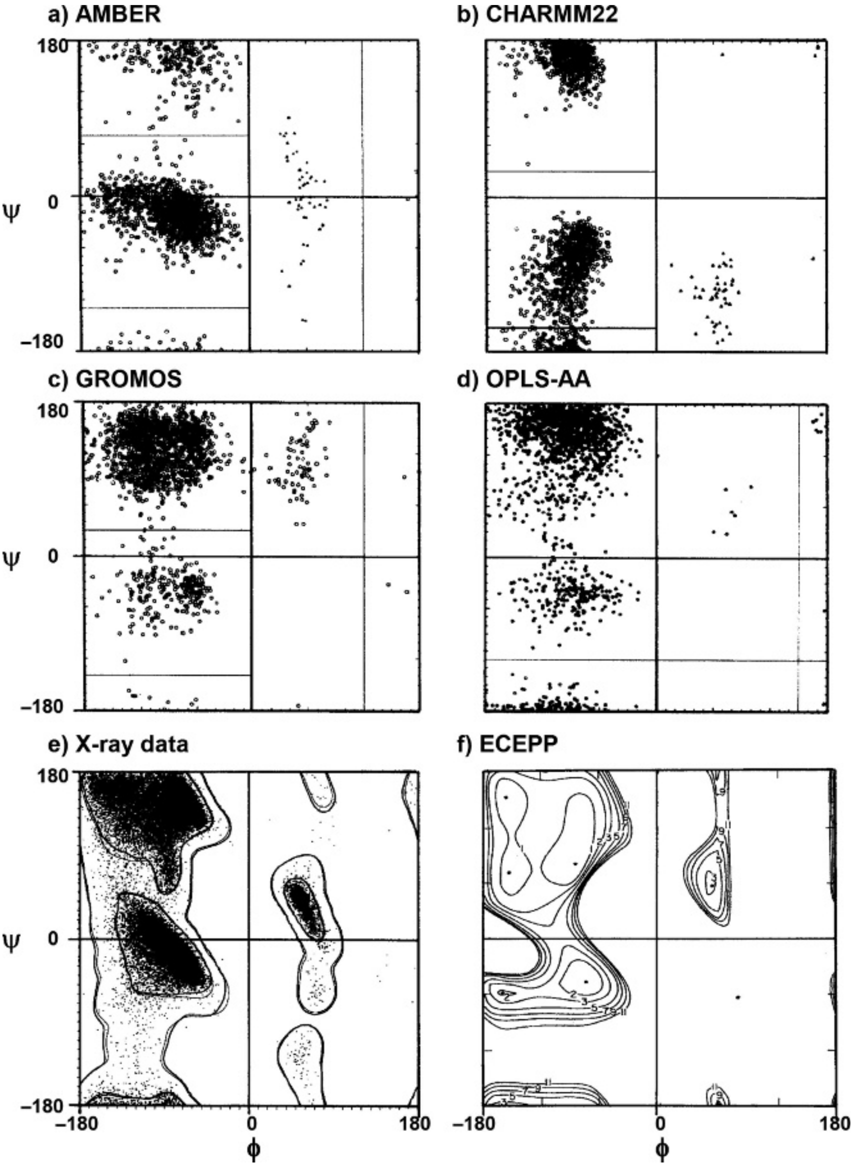

Figure 2.1 (a)–(d) Sampled conformational distributions of Ac-Ala-Me obtained with the AMBER, CHARMM22, GROMOS and OPLS-AA force fields, respectively (adapted from [1], Figures 1, 3, 4 and 5, with permission from John Wiley & Sons, Inc). (e) The (ϕ, ψ) data points for the amino acid residues derived from the X-ray structures of 500 proteins (adapted from [2], Figure 7, with permission from John Wiley & Sons, Inc). (f) The Ramachandran map calculated with the ECEPP force field showing equipotential levels of relative energy (adapted from [5], Figure 9, with permission from Wiley-VCH)

It also includes conformation C^7_{eq}, located into the 'pass' zone (that with ψ values around 60 ± 30°) extended toward the second main region of plausible conformations centred around (−90°, 0°). This region spreads through both left quadrants and contains conformations close to the right-handed α-helix (the 'alpha R' region). The third region is the smallest one, located in the upper-right quadrant, and contains conformations close to the left-handed α-helix (the 'alpha L' region). Only a few conformations were experimentally found in the lower-right quadrant, which includes conformation C^7_{ax}. Interestingly, experimental distributions in the regions corresponding to the right- and left-handed α-helices show a somewhat diagonal shape.

Sampling of the Ramachandran map with the AMBER and CHARMM22 force fields (Figure 2.1(a,b)) resulted in two main regions located in the left half of the map. For AMBER, population of the alpha R region was significantly higher than that of the beta region, and only a few conformations were found in the alpha L region (after performing additional molecular dynamics simulations; see [1] for details). These results are not consistent with the experimental data in Figure 2.1(e); one may expect that energy estimations involving the AMBER force field will lead to large numbers of false positives in the alpha R region when sampling conformational possibilities of peptide systems. The results obtained with CHARMM22 (Figure 2.1(b)) showed a high population of the beta region; however, the second equally populated region is shifted down to the lower-left quadrant and possesses only a small overlap with the experimentally determined alpha R region.

Much more consistent with experimental data were the Ramachandran maps obtained with the GROMOS and OPLS-AA ('AA' stands for 'all atoms', meaning 'including all hydrogens') force fields (Figure 2.1(c,d)). In both cases, the sampled beta and alpha R regions were located close to the experimentally determined ones, and populations of the beta regions clearly exceeded populations of the alpha R regions. Sampling also found conformations belonging to the pass zone between two regions, as well as conformations in the upper-right quadrant. These latter conformations were more frequent in the GROMOS map, but more consistent with the experimentally observed alpha L region in the OPLS-AA map. One may expect that energy estimations based on the GROMOS or OPLS-AA force fields will lead to more adequate sampling of peptide systems with less false positives or negatives. In fact, both force fields are currently more frequently used in conformational calculations of peptides and proteins than either AMBER or CHARMM.

At the same time, the Ramachandran map in Figure 2.1(f), calculated with the much older ECEPP force field using rigid valence geometry without accounting for solvent, whether implicit or explicit, shows at least as good consistency with experimental data in Figure 2.1(e) as the GROMOS and OPLS-AA maps. Indeed, the allowed areas of this map cover all three regions, determined experimentally with approximately the same relative populations. The zone region was also populated; in fact, conformation $C^7{}_{eq}$ had the lowest relative energy. Even the diagonal shape of the experimental distribution was preserved in the alpha R region (and, though just slightly, in the alpha L region) of Figure 2.1(f). The main difference from the experimental data was in the extension of the alpha R region in the ECEPP map toward (ϕ, ψ) values around ($-60° \pm 30°$, $-150° \pm 30°$); these are virtually unpopulated in Figure 2.1(e). One reason for the good consistency of the ECEPP map with the experimental distribution was that the parameters of the ECEPP force field were calibrated specifically to reproduce the X-ray data on crystal packing of amino acids. This made the ECEPP force field limited in applications to molecules other than peptides and proteins, which precluded its acceptance by commercially available modelling packages, such as SYBYL, INSIGHT or MacroModel. Nevertheless, comparison of maps in Figure 2.1(e) and (f) clearly suggests that the ECEPP force field, though less sophisticated than GROMOS or OPLS, may be successfully used in the sampling of possible conformational states of peptide and protein systems. At the same time, practical applications of the ECEPP force field require significantly less computer resources, with the additional advantage that it is able to perform sampling in dihedral angle space, which is much less complex than Cartesian coordinate space (see Section 2.2.2.3).

Further developments of molecular force fields The MM force fields outlined above are routinely used in computational design of peptides and proteins. However, as was concluded in a recent review, ‘Currently, force fields are not perfect (even the all-atom ones). It is possible to obtain different results with different force fields. Therefore, improving force fields (both the all-atom and reduced ones, and the water potential) is a priority.’ [26]. Further development of force fields has occurred recently, offering several possible improvements.

One approach is to combine quantum mechanics with molecular mechanics (QM/MM), whereby selected parts of the molecular system under study, for instance the active sites of enzymes, are treated by QM approximations, and the larger surrounding molecular areas are represented with MM [27,28]. This approach was applied to Ac-Ala-OMe and showed excellent consistency with the experimental data in Figure 2.1(e) [1]. Specifically, the molecule of Ac-Ala-OMe was treated with the fast approximate QM method SCCDFTB [29], whereas interactions between Ac-Ala-OMe and the explicit water molecules were calculated either with SCCDFTB (electrostatic interactions) or with force fields with flexible valence geometry (such as AMBER). Interactions between the water molecules were calculated with the same force fields. Figure 2.2(a) presents the results of sampling overlapped with contours corresponding (from purple to pink) to 99.8%, 99.5%, 98%, 95% and 90% of the levels of the experimental distribution of the (ϕ, ψ) points for the alanine residues from the experimental distribution in Figure 2.1(e). Distribution of sampled conformations in Figure 2.2(a) is remarkably close to that in Figure 2.1(e). All main regions of the experimental distributions have comparable relative populations, including the pass zone. Also, the diagonal shapes of distributions in the alpha R and alpha L regions are reproduced, whereas only a few conformations appear in the region of the (ϕ, ψ) values around (−60° ± 30°, −150° ± 30°). One slight discrepancy is the relatively low population of the narrow zone around ψ ~100°.

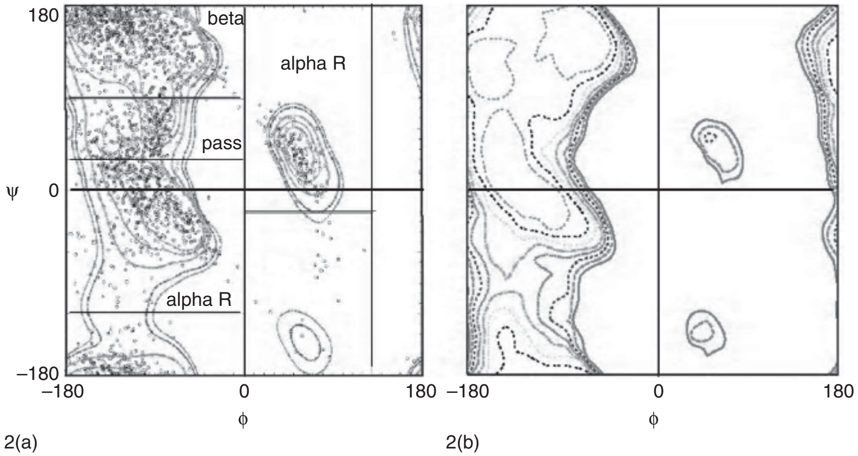

Figure 2.2 (a) Sampled conformational distribution of Ac-Ala-OMe obtained with the QM/MM approach. Lines show 99.8%, 99.5%, 98%, 95% and 90% (from purple to pink) levels of the experimental distribution of the (ϕ, ψ) points for alanine residues (adapted from [1], Figure 6). (b) Ramachandran map of free energy for Ac-Ala-OMe in water, obtained with the AMOEBA polarizable force field. Lines show energy levels of 3.2, 2.8, 2.4, 2.0, 1.6, 1.2, 0.8 and 0.4 kcal/mol, from red to dashed orange (map courtesy of Prof. Jay Ponder). Adapted with permission from John Wiley & Sons, Inc (see colour Plate 1)