METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Biomolecular Simulations

Methods and Protocols

Edited by

Luca Monticelli

INSERM, DSIMB, UMR-S 665, Paris, France

Emppu Salonen

Department of Applied Physics, School of Science,
Aalto University, Espoo, Finland

Humana Press

Editors
Luca Monticelli, Ph.D.
INSERM, DSIMB, UMR-S 665
Paris, France

Université Paris Diderot
Sorbonne Paris Cité, UMR-S 665
Paris, France

Emppu Salonen
Department of Applied Physics
School of Science
Aalto University
Espoo, Finland

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-62703-016-8 ISBN 978-1-62703-017-5 (eBook)
DOI 10.1007/978-1-62703-017-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012946729

© Springer Science+Business Media New York 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The field of molecular simulations has evolved, during the past 40 years, from picosecond studies of isolated macromolecules in vacuum to studies of complex, chemically heterogeneous biomolecular systems consisting of millions of atoms, with the simulation time scales spanning up to milliseconds. Much of the impressive progress in biomolecular simulations has been simply due to more powerful computers. However, tackling the sheer complexity of macromolecular systems of biological interest has never relied on Moore's law alone. Among the most important factors contributing to such enormous progress are the development of faster simulation algorithms and entirely novel approaches to molecular modelling, such as multiscale methods employing different levels of theory for different regions of the system studied. As the number of researchers in the field has increased, limitations inherent in each methodology have also been described, creating awareness of the pros and cons of different techniques, and contributing significantly to the progress in the field. Quantum mechanical (QM) calculations, while in principle providing a rigorous description of the molecules of interest, necessitate the use of some approximations in the underlying theory. Practical methodological solutions within these frameworks (for example, computationally efficient basis sets) have been crucial in QM calculations becoming a powerful tool in biomolecular modelling. Classical simulations employing empirical force fields have steadily become more versatile and accurate, thanks to more sophisticated potential energy functions even including explicit functions for hydrogen bonding and electronic polarization, and to more accurate parameterizations. The general level of accuracy of the classical simulations, combined with their speed, has made them perhaps the most widely used method in molecular modelling of biological systems. Coarse-graining the atomistic description of the molecules of interest has simplified the computations immensely while still retaining some of the key physicochemical properties of the system studied.

One major challenge in modelling biological systems is the very large span of length and time scales involved. Depending on the problem at hand, researchers look at motions occurring on time scales from femtoseconds (10-15 s) to hours, and covering distances from sub-atomic scale to cell size. Another challenge is the great chemical complexity and heterogeneity of biological systems. Even the smallest biological sample contains a wide variety of molecules. Computational studies – just as simple in vitro experiments – require much simplification. It is assumed that by understanding the properties and behaviour of simple model systems one can learn about the properties and behaviour of real, much more complex systems. When such assumption and other simplifications are justified, simulations can make a significant contribution to understanding biological systems' structure and functioning.

Excellent books are available to students and researchers who venture into the field of molecular modelling, covering both the basic foundations as well as more specialized aspects. With the present volume, we aim to present the foundations of well-established simulation techniques together with some of the recent developments in methods and practices. The latter rarely find ample coverage in traditional textbooks, but are being used more and more by researchers in the biological field. We also aim at giving

some practical examples on how to carry out simulations of some particular systems of great biological interest, and particularly systems including biological macromolecules.

The book consists of three sections, with the division based on the predominant classes of methods used in modelling at various length and time scales. The covered methodologies include electronic structure calculations, classical molecular dynamics simulations and coarse-grained techniques. Each section comprises a methodological and an application part. The former provides an introduction to the basic physics and chemistry underlying the computational models, and focuses particularly on recent developments. The application part illustrates examples on the four main classes of biological macromolecules, namely proteins, nucleic acids, lipids, and carbohydrates. This subdivision is in line with the traditions of the *Methods in Molecular Biology* series, with an introductory overview, theoretical foundations, and good practices of the methodologies used, followed by chapters illustrating their practical application in studies of biological macromolecules.

The target audience of the book includes both graduate students and researchers interested in computational modelling of biomolecular systems in physics, chemistry, and biology. The structuring of the different sections has been made so that after reading the first methodological chapters in the section in question, a non-expert reader can understand and appreciate the following application-oriented chapters. An expert on a given methodology can, in turn, jump directly to the chapters on state-of-the-art applications of the methodology covered. We hope that the readers will find this structure of the book useful and easy to approach.

The general structure of the book is as follows. The QM section contains reviews covering the most central contemporary methodologies of biomolecular modelling. The section begins with an overview of different electronic structure calculation methods with an emphasis on methodological issues related to the investigation of biological systems. The so-called ab initio molecular dynamics methodology for dynamic electronic structure calculations is then introduced. The remaining three chapters in the section address a more practical side of QM calculations of biological systems, that is, the hybrid QM and molecular mechanics methodology (QM/MM). These chapters feature a particular emphasis on studies of proteins, reflecting the most common domain of application of QM/MM methods.

In the section on atomistic simulations we cover the basic ideas and the most common techniques (molecular dynamics and classical force fields), and we also give space to a few recent developments that gained more and more importance in recent years: enhanced sampling algorithms, allowing for crossing energetic barriers and speed up the sampling; free energy calculations, that were rarely found in the literature only a decade ago due to the high computational cost, but are now accessible to a wide audience; polarizable force fields, which aim to increase the accuracy of classical empirical simulations with a tolerable computational overhead, and have become increasingly popular during recent years.

Coarse-graining has been perhaps the fastest evolving area in biomolecular modelling over the past few years, with new techniques and new force fields published monthly. Coarse-graining rarely finds any space in traditional textbooks, but it is becoming attractive for a very wide public. The diversity in the techniques developed is such that a thorough review would require an entire volume. The book covers some of the most exciting recent developments in the area, with applications to lipid membranes and membrane proteins in particular.

The increasing speed of molecular simulations hardware and software, and the development of force fields and methodologies make it possible to describe increasingly

complex biological systems and processes. Progress in several directions can be foreseen in the near future, with the development of faster and more accurate methodologies for electronic structure calculations, more refined classical force fields, and improved coarse-graining techniques. Multiscale modelling, that currently is one of the most challenging problems in molecular simulations, will require substantial theoretical and methodological development. Yet, the great interest in combining descriptions at different length scales bears expectations of a significant growth in this area.

Paris, France — Luca Monticelli
Espoo, Finland — Emppu Salonen

Contents

Contributors

Hanne Antila • *Department of Chemistry, Aalto University, Espoo, Finland*

Volodymyr Babin • *Department of Physics, Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC, USA*

Svetlana Baoukina • *Department of Biological Sciences, University of Calgary, Calgary, AB, Canada*

Peter J. Bond • *Department of Chemistry, University of Cambridge, Cambridge, UK*

Andrea Bortolato • *Heptares Therapeutics Ltd, Hertfordshire, UK*

G. Andrés Cisneros • *Department of Chemistry, Wayne State University, Detroit, MI, USA*

Philippe Derreumaux • *Laboratory of Theoretical Biochemistry, Paris 7 University, Paris, France*

Marcus Elstner • *Institute of Physical Chemistry, Karlsruhe Institute of Technology, Karlsruhe, Germany*

Giray Enkavi • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Marco Fanton • *Dipartimento di Scienze Farmaceutiche, Universita di Padova, Padova, Italy*

Nicolas Foloppe • *Vernalis (R&D) Ltd, Granta Park, Abington, Cambridge, UK*

Gerrit Groenhof • *Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*

Marc Guéroult • *INSERM, Paris, FranceInstitut de Biologie Physico-Chimique, CNRS, Paris, France*

Brigitte Hartmann • *Laboratoire de biologie et pharmacologie appliquée (LBPA), CNRS, Paris, France*

Zhijian Huang • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Susanna Hug • *Department of Applied Mathematics, University of Western Ontario, London, ON, Canada*

Anne Imberty • *Centre de Recherches sur les Macromolécules Végétales (CERMAV-CNRS), Grenoble cedex 9, France*

Mikael P. Johansso • *Department of Chemistry, University of Helsinki, Helsinki, Finland*

Ville R.I. Kaila • *Department of Chemistry, University of Helsinki, Helsinki, Finland*

Syma Khalid • *School of Chemistry, University of Southampton, Southampton, UK*

Kari Laasonen • *Department of Chemistry, Aalto University, Espoo, Finland*

Alessio Lodola • *Dipartimento Farmaceutico, Università degli Studi di Parma, Parma, Italy*

Jing Li • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Zhong-Yuan Lu • *State Key Laboratory of Theoretical and Computational Chemistry, Institute of Theoretical Chemistry, Jilin University, Changchun, China*

Paween Mahinthichaichan • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
Siewert-Jan Marrink • *Groningen Biomolecular Sciences and Biotechnology Institute & Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*
Hector Martinez-Seara • *Department of Physics, Tampere University of Technology, Tampere, Finland*
Jonathan S. Mason • *Heptares Therapeutics Ltd, Hertfordshire, UKLundbeck A/S, Copenhagen, Denmark*
Ayori Mitsutake • *Department of Physics, Keio University, Yokohama, Kanagawa, Japan*
David L. Mobley • *Department of Pharmaceutical Sciences, University of California, Irvine, CA, USA*
Luca Monticelli • *INSERM, UMR-S665, Paris, FranceUniv Paris Diderot, Sorbonne Paris Cité, UMR-S 665, Paris, FranceInstitut National de la Transfusion Sanguine (INTS), Paris, France*
Mahmoud Moradi • *Department of Physics, Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC, USA*
Stefano Moro • *Dipartimento di Scienze Farmaceutiche, Universita di Padova, Padova, Italy*
Yoshiharu Mori • *Department of Physics, Nagoya University, Nagoya, Aichi, Japan*
Adrian J. Mulholland • *Centre for Computational Chemistry, School of Chemistry, University of Bristol, Bristol, UK*
William G. Noid • *Department of Chemistry, The Pennsylvania State University, University Park, PA, USA*
Alessandra Nurisso • *Centre de Recherches sur les Macromolécules Végétales (CERMAV-CNRS), Grenoble cedex 9, France*
Yuko Okamoto • *Department of Physics, Nagoya University, Nagoya, Aichi, Japan*
Xavier Periole • *Groningen Biomolecular Sciences and Biotechnology Institute & Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*
Matej Praprotnik • *National Institute of Chemistry, Ljubljana, Slovenia*
Tomasz Róg • *Department of Physics, Tampere University of Technology, Tampere, Finland*
Christopher Roland • *Department of Physics, Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC, USA*
Celeste Sagui • *Department of Physics, Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC, USA*
Emppu Salonen • *Department of Applied Physics, Aalto University, Espoo, Finland*
Yves-Henri Sanejouand • *University of Nantes, Nantes, France*
Nicolas Sapay • *Centre de Recherches sur les Macromolécules Végétales (CERMAV-CNRS), Grenoble cedex 9, France*
Saher A. Shaikh • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
Julian C. Shillcock • *École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Michael R. Shirts • *Department of Chemical Engineering, University of Virginia, Charlottesville, VA, USA*
Luigi Delle Site • *Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany*
Thomas Steinbrecher • *Institute of Physical Chemistry, Karlsruhe Institute of Technology, Karlsruhe, Germany*
Dage Sundholm • *Department of Chemistry, University of Helsinki, Helsinki, Finland*
Emad Tajkhorshid • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
D. Peter Tieleman • *Department of Biological Sciences, University of Calgary, Calgary, AB, Canada*
Yong-Lei Wang • *State Key Laboratory of Theoretical and Computational Chemistry, Institute of Theoretical Chemistry, Jilin University, Changchun, China*
Po-Chao Wen • *Department of Biochemistry, Center for Biophysics and Computational Biology, College of Medicine, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

Part I

Quantum Mechanics Calculations

Chapter 1

Ab Initio, Density Functional Theory, and Semi-Empirical Calculations

Mikael P. Johansson, Ville R.I. Kaila, and Dage Sundholm

Abstract

This chapter introduces the theory and applications of commonly used methods of electronic structure calculation, with particular emphasis on methods applicable for modelling biomolecular systems. This chapter is sectioned as follows. We start by presenting *ab initio* methods, followed by a treatment of density functional theory (DFT) and some recent advances in semi-empirical methods. Treatment of excited states as well as basis sets are also presented.

Key words: Ab initio methods, Density functional theory, Wave-function methods, First-principles approach, Coupled-cluster methods, Second-order perturbation theory, Electronic excited states, Basis sets, Semi-empirical methods, Time-dependent density functional theory, Coupled-cluster linear response methods

1. Introduction

Application of quantum chemical electronic structure calculations has become feasible in biomolecular modelling during recent decades due to development of new methods and significant increase in computing power. Development of density functionals since the 1980s has had a considerable impact on modelling biomolecular systems, and it is now possible to obtain results close to chemical accuracy for systems comprising several hundred atoms. Moreover, as we shall discuss here, application of *ab initio* methods on ever-larger systems has also become feasible during recent years, and can now in many cases be considered as an alternative to DFT modelling. The aim of this chapter is to briefly discuss quantum mechanical (QM) approaches that can be employed in computational studies of biomolecular systems and reactions. Due to the large variety of quantum chemical methods, we offer merely an overview, a glimpse, of available methods and protocols.

Luca Monticelli and Emppu Salonen (eds.), *Biomolecular Simulations: Methods and Protocols*, Methods in Molecular Biology, vol. 924, DOI 10.1007/978-1-62703-017-5_1, © Springer Science+Business Media New York 2013

2. *Ab Initio* Methods

The starting point for most correlated *ab initio*, or wave-function (WF) calculations is the Hartree–Fock self-consistent field (HF-SCF) or mean-field approximation (1–3). The general assumption in the HF model is that the wave function can be expressed using one Slater determinant (SD) (4) or more generally, a linear sum of SDs with fixed coefficients called configuration state functions (CSFs). At the HF-SCF level, the one-particle functions (orbitals) are optimised by minimising the total HF electronic energy. Up front, we discourage large-scale applications at HF level, as the accuracy of the method is significantly lower compared to what can be obtained with density functional theory (DFT) which has the same, or even lower computational cost. The reason is of course that accounting for electron correlation is very important also when modelling biomolecules.

In WF theory, electron correlation is considered as an addition to the uncorrelated HF reference solution. The occupied and virtual HF orbitals (ϕ_i and ϕ_a) and the corresponding orbital energies (ε_i and ε_a) are used for estimating the correlation energy, defined as (5),

$$E_{\text{correlation}} = E_{\text{exact}} - E_{\text{HF}}. \tag{1}$$

E_{exact} is the energy obtained by solving the Schrödinger equation in the employed basis set, and E_{HF} is the total energy of the Hartree–Fock calculation obtained with the same basis set.

In *ab initio* correlation calculations, the wave function is constructed as a linear combination of the SDs or CSFs that can be constructed by permuting the occupation numbers of the HF orbitals. The *ab initio* correlated wave functions can be written as,

$$\Psi = \sum_I C_I \Psi_I, \tag{2}$$

where C_I is the expansion coefficients of the SDs or CSFs (Ψ_I). The difference between *ab initio* correlation approaches lies in the way the expansion coefficients are constructed and how large fraction of the complete set of SD's is considered in the calculation. The selected SD's to expand the wave function are obtained by replacing orbitals that are occupied in the HF wave function with unoccupied ones. In general, the replacement is done systematically by promoting one, two, three, etc., electrons from the occupied HF orbitals to the virtual space, corresponding to single (S), double (D), triple (T), etc., excitations. When all SD's are taken into account, the full configuration interaction (CI) model is obtained (6), the solutions of which correspond to the exact Schrödinger wave functions in the employed basis set.

At the CI level, the total electronic energy is minimised with respect to the state parameters C_I. In the related multi-configuration self-consistent-field (MCSCF) calculations, the C_I coefficients and the occupied orbitals are optimised until the energy minimum is reached. The non-integer orbital occupation numbers are calculated using the C_I coefficients (7). The MCSCF calculations are time-consuming, which restricts the size of the orbital space that can be afforded. The remaining correlation effects omitted in the MCSCF calculation can be estimated by employing second-order perturbation theory (PT2) or by using low-order CI corrections, called multi-reference CI (MRCI) (8). The most popular combination of MCSCF and perturbation theory is complete-active-space SCF calculations (CASSCF) (9) and second-order perturbation theory (CASPT2) originally developed by Roos et al. and implemented by others in a variety of versions (10). These state-of-the-art approaches are far from black-box methods, due to the difficulty in choosing the active space. This is currently limited to about 15 electrons in 15 orbitals due to the size of the CI calculations. As there are already ten d-electrons in transition metals, treating bioinorganic complexes with this method is very challenging. Despite this, the CASPT2 approach has been applied to treat excited states of many biomolecular systems (11–13). In addition, the electronic structure of porphyrin model systems (14, 15) and the oxoheme (16) has recently been treated with the CASPT2 methodology.

Before performing large-scale CASPT2 calculations, a thorough training in these advanced computational approaches is recommended. For this purpose, we highlight the European Summer School in Quantum Chemistry. The theory behind MCSCF perturbation theory calculations will not be further discussed here; comprehensive reviews have recently been published elsewhere (17, 18, 19).

At coupled-cluster (CC) levels of theory (20), an exponential ansatz is used for the wave function (Ψ),

$$\Psi = e^{\hat{T}_1+\hat{T}_2+\hat{T}_3+\cdots}\Psi \tag{3}$$

where Ψ is the HF reference SD and $\hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots$ are the cluster operators containing S, D, T, etc. excitation operators and the corresponding cluster amplitudes, which are obtained by solving the CC equations (7). The cluster operators containing, for example, the single and double excitation amplitudes are given by,

$$\hat{T}_1 = \sum_{ia} t_i^a a_a^\dagger a_i \qquad \hat{T}_2 = \frac{1}{4}\sum_{ijab} t_{ij}^{ab} a_a^\dagger a_b^\dagger a_i a_j \tag{4}$$

where t_i^a and t_{ij}^{ab} are the single and double excitation amplitudes, respectively. The SDs considered in the CC calculations are selected by the creation and annihilation operators. $a_a^\dagger$ creates an electron in the virtual orbital a, and a_i annihilates the electron in the i:th occupied orbital. The computational requirements of CI and CC

calculations are comparable. High-order CI and CC calculations, thus those considering SD's obtained by T, Q, etc. replacements from the HF reference are computationally very demanding. They are so far not applicable on actual biomolecular systems as they are limited to systems with <20 atoms (21). Truncated CI calculations involving S, D, T, etc. replacements are not size extensive (22, 23), a flaw which has drastically diminished their popularity.

The simplest *ab initio* approaches considering electron correlation effects are the second-order Møller–Plesset perturbation theory (MP2) (24) and the CC approximate singles and doubles (CC2) methods (25). The MP2 and CC2 expressions can be derived using several approaches, but it is probably most convenient to relate them as special cases to the coupled-cluster singles and doubles (CCSD) model. The MP2 method has fixed double excitation amplitudes and vanishing single excitation amplitudes. The MP2 double excitation amplitudes can be expressed using the orbital energies of the HF calculations and two-electron interaction energies as,

$$t_{ij}^{ab} = \frac{\langle ab||ij\rangle}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j} \tag{5}$$

where $\langle ab||ij\rangle$ denotes the antisymmetrised Coulomb integrals $(ai|bj) - (aj|bi)$, given in the Mulliken notation. The MP2 energy expression then becomes,

$$E_{MP2} = -\frac{1}{4}\sum\nolimits_{ijab} t_{ij}^{ab}\langle ab||ij\rangle \tag{6}$$

The MP2 and CC2 calculations formally scale as N^5, where N is the size of the basis set. CC2 is an iterative method, where the single and double excitation amplitudes are obtained by solving the CC2 equations. The CC2 expression for the double excitation amplitudes is structurally identical to the one for the MP2 amplitudes. However, in the CC2 expression, the Coulomb integrals are transformed using the single excitation amplitudes. The CC2 double excitation amplitudes can be recalculated using given single excitation amplitudes, two-electron integrals, and orbital energies when needed. The CC2 double excitation amplitudes are therefore not stored rendering calculations on large molecules feasible (25, 26). The obtained CC2 energies are generally of the same quality as the MP2 ones, whereas the CC2 computational costs are larger. For molecules where correlation effects are small, the accuracy of ground-state structures and vibrational frequencies obtained in the MP2 and CC2 calculations are on average the same (27). However, for systems with large correlation effects the CC2 bond lengths are somewhat longer, the bond strengths weaker, and the vibrational frequencies lower than those obtained in MP2 calculations, and the MP2 model is to be preferred. The main use of the single excitation amplitudes in the CC2 model is to make it possible

to calculate properties of excited states and to introduce orbital polarisation effects. In general, it makes little sense to replace MP2 by CC2 in ordinary ground-state calculations. The CC2 method should mainly be used in calculations of excited-state properties.

The MP2 and CC2 models are in practice the only correlated *ab initio* computational methods that can be applied to large biomolecules today. This could well change in the near future as computers become faster and more efficient divide-and-conquer algorithms are being developed (28, 29, 30). Another reason for the success of the MP2 and CC2 methods is the efficient resolution-of-the-identity (RI) MP2 and CC2 algorithms that have been developed and implemented (26, 31, 32). The RI algorithms, also known as density-fitting methods, are regularly used in applications on biomolecular systems (33). A linear-scaling MP2 implementation already exists, so biomolecules consisting of more than a thousand atoms can indeed be investigated with present-day computer resources (34). Other *ab initio* correlation approaches that can be applied on large biomolecules are e.g., the molecular orbital fragmentation approaches (35–37).

2.1. Ground-State Calculations

The main advantage with *ab initio* correlation approaches is that the accuracy and reliability can be systematically improved by taking higher-order correlation effects into account. Thus, the reasons for deviations between calculated values and experimental results can be understood by employing methods that consider higher-order correlation effects. However, the huge computational costs of these methods makes biomolecular calculations unfeasible. The basis set size significantly affects the accuracy. The use of small basis sets in *ab initio* correlation calculations renders the obtained results unreliable despite a mathematically accurate treatment of electron correlation.

The simplest correlated WF method, MP2, *does* already account for one of the most important correlation effects for biomolecules, that is, dispersion interactions. Benchmark calculations on small molecules have shown that MP2 generally overestimates non-bonding interactions, however. This flaw can be corrected for by introducing semi-empirical scaling factors to the same-spin and opposite-spin contributions of the MP2 (38, 39) and CC2 correlation energies (40). In many applications, the spin-component scaled (SCS) and scaled opposite-spin (SOS) MP2 methods have been found to provide an accuracy comparable to the CCSD model perturbationally corrected for triple excitation, CCSD(T) (41). One should naturally be aware of the caveats of relying on an accuracy rooted in cancellation of errors.

The SOS-MP2 and SOS-CC2 can be made faster than the uncorrected versions as they can be formulated to formally scale as N^4 (42). For example, the SOS-MP2 methodology has recently been applied to study peptide isomerisation for model systems comprising ~400 atoms (43).