Molecular Orbital Calculations for Amino Acids and Peptides

Anne-Marie Sapse

Molecular Orbital Calculations for Amino Acids and Peptides

With 32 Figures

Birkhäuser
Boston • Basel • Berlin

Anne-Marie Sapse
John Jay College and Graduate School
City University of New York
New York, NY 10019
and
Rockefeller University
New York, NY 10021
USA

Library of Congress Cataloging-in-Publication Data
Sapse, Anne-Marie.
Molecular orbital calculations for amino acids and peptides / Anne-Marie Sapse.
p. cm.
Includes bibliographical references and index.
ISBN 0-8176-3893-8 (hardcover: alk. paper)
1. Amino acids. 2. Peptides. 3. Molecular orbitals. I. Title.
QD431.S257 1999
547'.750448—dc21 99-26375
CIP

Printed on acid-free paper.

 Birkhäuser ®

ISBN 0-8176-3893-8
ISBN 3-7643-3893-8 SPIN 19901572

Typeset by Best-set Typesetter Ltd., Hong Kong.
Printed and bound by Sheridan Books, Inc., Ann Arbor, MI.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

Molecular Orbital Calculations for Amino Acids and Peptides

To my husband, Marcel Sapse, and to my daughter, Danielle Sapse, without whose support I could not have written this book.

Preface

This book is intended mainly for biochemists who would like to augment experimental research in the domain of amino acids and small peptides with theoretical calculations at the ab initio level.

The book does not require a profound knowledge of mathematics and quantum chemistry. It teaches one rather how to use computer software such as the Gaussian programs and gives examples of problems treated in this manner.

Chapter 1 describes the calculations and one of the programs used for ab initio work.

Chapter 2 describes calculations on small amino acids, such as glycine and alanine.

Chapter 3 discusses the biochemical properties of GABA (gamma amino butyric acid), which is one of the most important amino acids of the nervous system. Ab initio calculations performed in order to study the structure of GABA are presented.

Chapter 4 discusses an amino acid related to GABA, namely DABA (diaminobutyric acid), presenting information about its structure and transport properties.

A number of amino acids, essentials in the biochemistry of organisms, are discussed in Chapter 5. These acids have been subjected to ab initio investigation. Proline, a special amino acid as far as structure is concerned, is discussed in Chapter 6.

Chapter 7 discusses two sulfur-containing amino acids, taurine and hypotaurine, presenting some experimental studies on their mode of action and an ab initio study of their structure.

Starting with Chapter 8, small peptides of great importance are discussed. Glucagon, a small peptide that plays a role in diabetes, is the subject of Chapter 8.

Chapter 9 discusses the pheromone alpha factor, from an experimental and theoretical point of view.

Chapter 10 presents calculations on tight turns in proteins.

Chapter 11 discusses some small peptides that have been studied with ab initio methods.

Oligopeptides that feature anticancer activity, such as lexitropsins, are discussed in Chapter 12.

The book is addressed to graduate and postgraduate students as well as other researchers in the amino acid and peptide area.

New York, NY

Anne-Marie Sapse

Introduction

Knowledge about the origin of life requires the recapitulation of the steps of archaic molecular evolution. According to the protenoid model, proteinoids (copolyaminoacids) arose on earth from mixtures of self-sequencing amino acids. The structure of amino acids, of the peptides formed by their polymerization via the formation of peptidic bonds, as well as the structure of the proteins that are polypeptide chains in various numbers and conformations, have formed the subject of an enormous number of experimental and theoretical studies.

At present, both theoretical and experimental methods are taken seriously as useful sources of information. They compare results and confirm or dispute structural findings. While experimental results are usually not doubted, and computational results depend on such parameters as the quality of the basis sets used, there have been instances in which computational results have contradicted experimental ones regarding structural determination. However, in most instances the two types of methods complement each other. For instance, a laboratory search for intermediates in certain reactions can be avoided once large basis-set calculations show the intermediates not to be a stationary state, more exactly, a minimum on the energy hypersurface.

The application of computational methods to biological systems dates from the 1950s, when the pioneering work of Bernard and Alberte Pullman was first published. The biological systems studied with the quantum-chemical methods available at that time had to be small, and not all the conclusions derived were correct. However, this work opened the door to a whole new area of research.

The basic problem in the determination of the structure of biological systems is their size. In order to be able to handle such molecules as the nucleic acids or the proteins, new theoretical methods had to be developed, and the quantum-chemical methods, ab initio and semi-empirical, were augmented by the molecular mechanics method, which uses experimental parameters in order to determine the force fields of the systems.

Huge strides have been made in the development of computer programs that handle larger systems. Researchers are striving to find the optimum combination of accuracy and expediency, with the ultimate goal being the reduction of computational effort with no loss of accuracy.

All three of these types of theoretical methods are used in the description of amino acids and peptides. The size of proteins precludes the use of ab initio or semiempirical methods, so they are mainly described with computer modeling, with programs such as Sybil, Quanta, and Insight, augmented by energy calculations with the Charmm program and other molecular-mechanic calculations.

The primary structure of proteins, characterized by the amino acid composition and sequence, is determined experimentally by degradation via hydrolysis of the peptidic bonds. The classic method of determining the sequence involves Edman degradation, which is an end-labeling procedure. Physical methods used include mass spectrometry and nuclear magnetic resonance (NMR). Since the 1980s, sequencing of proteins has been performed by sequencing its mRNA or gene.

The three-dimensional structures of about 800 proteins have been determined by Max Perutz and John Kendrew using X-ray crystallography. Recently, NMR methods have also been used. The secondary structure of proteins, with 60% alpha helices or beta sheets and the rest random coils and turns, is determined by the propensity of the amino acids constituting the given protein to form either alpha helices or beta sheets. It is recognized now that the sequence of a protein determines its three-dimensional structure.

Given the size of proteins, quantum-chemical conformational and energy calculations are at present impossible. Some calculations on proteins are being performed at present in Dr. Lothar Schafer's laboratory. Undoubtedly, the increase in computer capacity and progress in computer algorithms will make it possible to perform many such calculations in the not too distant future. The theoretical methods used so far for proteins include molecular-mechanics methods that neglect electrons and describe the motion of nuclei under the influence of an empirical or quantum-mechanically calculated potential energy function, methods that do not use energy functions except in terms of stereochemical principles, computer graphics methods, and molecular-dynamic methods.

Smaller peptides have also been described by the above-mentioned methods, especially the empirical conformational energy program for peptides (ECEPP), written by Sheraga and his group, which has been applied to a large number of small peptides.

In recent years it has become possible to treat amino acids and small peptides with quantum-chemical calculations, as will be described in the next chapters.

Contents

1
Theoretical Background

Inadequate descriptions of atoms and molecules by the methods of classical physics led researchers to propose new ways to describe physical reality, giving birth to a totally new science, quantum mechanics. The methods of quantum mechanics are based on the introduction of a wave function, whose physical meaning is related to the probability of finding a certain particle, at a certain time in a volume element, positioned between x and $x + dx$ in the x = direction, between y and $y + dy$ in the y = direction, and between z and $z + dz$ in the z = direction at certain time t. This wave function Ψ satisfies the Schrödinger equation,

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v\right)\Psi = E\Psi, \quad \hbar = \frac{h}{2\pi},$$

or for short, $H\Psi = E\Psi$, where H, the Hamiltonian operator, is defined by the expression

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V;$$

h is Planck's constant; ∇^2 is the sum of the partial second derivatives with respect to x, y, and z; m is the mass of the particle; and V is the potential energy of the system. The Hamiltonian H represents the quantum equivalent of the sum of the kinetic energy and potential energy, with V being the potential energy operator and $\frac{-\hbar^2}{2m}\nabla^2$ the kinetic energy operator. Finally, E is the total energy of the system and is a number, not an operator.

The wave function, satisfying the Schrödinger equation, and the energy contain all the information about the system within the limits of the Heisenberg uncertainty principle, which states that the exact momentum and position of a particle cannot be known simultaneously. This is why the wave function represents a probability and not a certitude.

Applied to atoms, the Schrödinger equation describes the motion of the electrons in the electrostatic field created by the positive charge of the nucleus. In addition, each electron is subjected to the field created by the negative charge of the other electrons. When the Schrödinger equation is applied to molecules, the motion of the nuclei has also to be taken into consideration, but the fact that the nuclei are so much heavier than the electrons makes it possible to neglect their motion. This is embodied in the Born–Oppenheimer approximation. Accordingly, the electronic distribution in molecules does not depend on the motion of the nuclei, but only on their position. Indeed, the position of the nuclei determines the positive component of the electrostatic field to which electrons are subjected. The kinetic energy operator of the nuclei is considered to be zero.

The many-electron molecule can be thus described by a Hamiltonian written as

$$H = K + V,$$

where K, the kinetic energy operator, is

$$-\frac{\hbar^2}{2m}\sum_i \frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2},$$

with the sum taken over the number of electrons, while V, the potential energy operator, is composed of two electronic terms. One is the attraction between the positive nuclei and the negative electrons, expressed as

$$-\sum_i \sum_I \frac{Z_I e^2}{\mathbf{R}_I - \mathbf{r}_i},$$

where i represents, as before, the summation over the electrons, and I is the summation over the nuclei. Here Z is the charge of the Ith nucleus, and $\mathbf{R}_I - \mathbf{r}_I$ is the distance between the Ith nucleus and the ith electron. The second term represents the repulsion between electrons:

$$\sum_i \sum_{j \neq i} \frac{e^2}{\mathbf{x}_i - \mathbf{x}_j},$$

where $\mathbf{r}_i$–$\mathbf{r}_j$ represents the distance between electron i and electron j, and e, as before, is the charge of the electron. In addition, one must consider the nuclear repulsion, which determines the nuclear potential energy. This can be expressed as

$$\sum_I \sum_{J \neq I} \frac{Z_I Z_J e^2}{\mathbf{R}_J - \mathbf{R}_I},$$

where $\mathbf{R}_J - \mathbf{R}_I$ is the distance between nucleus I and nucleus J.

The Schrödinger equation can be solved analytically only for one atom: the hydrogen atom. The solution, even for the lightest atom, is complicated, containing spherical harmonic functions and Hermite polynomials. When

the electron–electron interactions are involved, as they must be for any atom containing more than one electron, the Hamiltonian cannot be expressed any longer in terms of spherical coordinates, which allow the separation of the three-dimensional form into three one-dimensional solvable equations. Therefore, a number of approximations have to be introduced.

The main approximation used to solve the Schrödinger equation for systems larger than the hydrogen atom is the variation principle. Indeed, when the equation is applied to atoms, the wave function is composed of a set of functions called atomic orbitals, corresponding to given energy states, containing a number of electrons determined by Pauli's exclusion principle. If the exact form of Ψ is known, the energy of the system can be computed by using the expression

$$E = \frac{\int \Psi^{\alpha} H \psi d\tau}{\int \psi^{\alpha} \psi d\tau}.$$

If the exact form of Ψ is not known, an educated guess can be taken, and the approximate value of Ψ is used to compute an approximate E. The variation principle states that the expectation value of the energy thus obtained will always be higher than the exact energy of the system. This allows the energy to be minimized with some parameters characterizing the wave function, in order to obtain the closest possible energy to the exact energy of the system. This procedure establishes a number of equations whose solutions are the optimum values for the parameters of the system.

One of the methods to construct a good wave function is the Hartree–Fock method.

The Hartree–Fock method deals with the reason for the impossibility of solving the Schrödinger equation analytically: the term $e/r-r$, which is the term representing interelectronic repulsion. In the absence of this term, the equation for an atom with n electrons can be separated into n equations for the hydrogen atom. If the sum of these terms is replaced by the sum of terms describing the motion of each electron through a cloud of electric charge due to the other electrons, the equation becomes solvable through an iterative method. Indeed, the electronic cloud is characterized by its charge density, which depends on the atomic orbitals describing the electrons. Once the interaction between a given electron and the cloud of the other electrons is calculated making use of an initial approximated orbital, the equation can be solved, and a new, improved orbital is obtained. This new orbital replaces the initial guess in the equation, whose solution represents an even more improved orbital. This iteration procedure is repeated until a certain threshold is reached.

For molecules, molecular orbitals have to be used instead of atomic orbitals. These can be built out of atomic orbitals, and one of the most widely used methods is to construct the molecular orbitals out of a linear combination of atomic orbitals (LCAO). The total wave function of the

system has to be augmented by spin orbitals called alpha (for spin $\frac{1}{2}$) or beta (for spin $-\frac{1}{2}$).

In order to perform quantum-chemical calculations without using any approximations, such as neglecting integrals of interaction between atomic orbitals located at different centers or using experimental parameters, one has to use the ab initio method, which uses a theoretically constructed wave function from the beginning. Since the Hartree–Fock method involves the calculation of integrals over atomic functions, the computational time is proportional to N^4, where N is the number of atoms of the system. For amino acids, and especially for peptides, this is an enormous task, since the atomic orbitals are exponential functions of the form $e^{-\zeta r}$, where r is the distance of each electron from the nuclei. This form, called the Slater orbital, requires a large amount of computer time for the computation of the integrals. To shorten the time, these functions have been replaced by expansions in a certain number of Gaussian functions of the form e^{-ar^2}. The integrals over Gaussian functions are much easier to compute. To reproduce better the form of a Slater orbital, which is the real dependence of the functions on r, as large a number of Gaussians as possible has to be used for the expansion.

So the function will take the form

$$\psi = ce^{-ar^2},$$

where α is a constant determining the radial extent and c is another constant.

Among the computer programs devised for performing ab initio calculations are the Gaussian programs, written at Carnegie Mellon University, in Pittsburgh, Pennsylvania. These programs make use of the expansion of Slater-type orbitals into a series of Gaussians, thereby establishing different basis functions for describing the system.

The smallest basis set used by the programs is the STO-3G basis set. The name comes from "Slater-type orbital," expanded into a series of three Gaussians. For hydrogen atoms, the orbital is the *s* orbital, while for heavier atoms *s* and *p* orbitals are used, as appropriate for a given electron in the atom. For large systems, the STO-3G basis set is the only possible one. Slightly larger minimal basis sets include the STO-4G, STO-5G, and STO-6G, where only one Slater orbital is used, expanded into 4, 5, and 6 Gaussians, respectively. It has been found that the energy decreases with the number of Gaussians, but such important information as optimum geometry, energy differences, and atomic charges are fairly insensitive to this number. In most cases bond distances calculated by STO-3G are very close to the experimental ones.

A larger series of basis sets are the split-valence basis sets. Among these, the double-zeta basis sets consist of two Slater-type orbitals for the valence electrons, one expanded in a number of Gaussians, the other approximated

by one Gaussian. The core electrons are described by one Slater-type orbital, expanded in a number of Gaussians. For instance, one of the most widely used basis sets, the 6-31G basis set, has the core electrons described by a Slater-type orbital expanded in a series of six Gaussians, while the valence electrons are described by two Slater-type orbitals, one expanded in a series of three Gaussians and the other one approximated by one Gaussian function. The functions used are *s* for hydrogen and *s* and *p* for nonhydrogen atoms. Triple-zeta basis sets feature three Slater-type orbitals for the description of valence electrons. An example is the 6-311G basis set, which uses three Slater-type orbitals for the description of the valence electrons, one expanded in a series of three Gaussians and the other two approximated by one Gaussian each.

The larger the basis set, the lower is the predicted energy of the system, and thus the closer to the real energy. However, the optimized geometries predicted by minimal basis sets are sometimes no worse than those predicted by double-zeta basis sets. For instance, the double-zeta basis sets predict too-large bond angles for water, ammonia, and the HOC angle in alcohols. The minimal basis sets predict that these angles will have values that are too small but closer to the experimental values than the ones predicted by the double-zeta basis sets. Energy differences and reaction energies are predicted better by double-zeta basis sets than by the minimal basis sets.

In order to improve even further the results obtained through the use of Gaussian basis sets, polarization functions are introduced. These are *d* functions on nonhydrogen atoms and *p* functions on hydrogens. Polarization functions possess angular momentum beyond that required for the ground state of the atom, while split-valence basis sets allow the orbitals to change size but not shape. The use of polarization functions increases greatly the accuracy of the results, especially where the bond angles are concerned. An even greater improvement due to polarization functions is observed in the prediction of the puckering of rings. This problem will be discussed in more detail in the next chapters. Basis sets containing polarization functions predict values too short for certain bond lengths. This problem is remedied by using them in conjunction with correlation energy calculations, as will be shown later.

For species rich in electrons, such as anions, it is advisable to add diffuse functions to the basis set in order to provide a better description of the system. Such basis sets, for instance 6-31+G*, add diffuse *s*- and *p*-type functions to nonhydrogen atoms, while the 6-31++G* set also adds *p* functions to the hydrogen. Negatively charged amino acids, such as aspartic and glutamic, are particularly prone to requiring the use of diffuse functions.

Larger basis sets make use of more than one *d* function and of *f* functions, such as 6-311G* (*2df,2pd*), which uses two *d* functions, or basis sets with 3 *df*, which use three *d* functions besides the *f*.

Atomic Electrical Charges

Parameters of great importance for the description of a molecule are the electrical charges on each atom. These are of particular interest when the system to be studied is an amino acid or peptide molecules that can be neutral or charged or that exhibit the structure of a zwitterion. Two of the methods to evaluate the net atomic charges are Mulliken population analysis and the Merz–Kollman–Singh method. The Gaussian programs use the former as default and the latter if the command Pop = MK is given.

Mulliken population analysis calculates the total atomic charge on an atom X as the atomic number of X minus the gross atomic population expressed as the sum of the net population of the functions associated only with atom X and half of the overlap population of the functions associated with both atom X and any atom bound to it. This method uses the concept of electron density functions.

The Merz–Kollman–Singh method fits the electrostatic potential to points selected on a set of concentric spheres around each atom.

Two other methods besides the Merz–Kollman–Singh that are used to select the points where point charges are assigned to fit the computed electrostatic potential are CHelp and CHelpG.

Another method to obtain atomic charges, natural population analysis, is carried out in terms of localized electron pairs that act as bonding units.

An example of the difference between the net atomic charges predicted by Mulliken population analysis and by the Merz–Kollman–Singh method can be observed in the charges obtained for one of the conformations of glycine. This conformation does not feature hydrogen bonds and sets the N–C–C–O atoms in the same plane, as shown in Figure 1.1. The optimization was performed at the Hartree–Fock level, using the 6-31G* basis set. The following results were obtained for the net atomic charges (the units are eu):

Atom	Mulliken	Merz–Kollman–Singh
C1	–0.215	0.349
C2	0.748	0.765
N	–0.838	–1.090
O1	–0.702	–0.735
O2	–0.550	–0.583
H1	0.208	0.044
H2	0.178	–0.043
H3	0.468	0.487
H4	0 345	0.409
H5	0.358	0.397

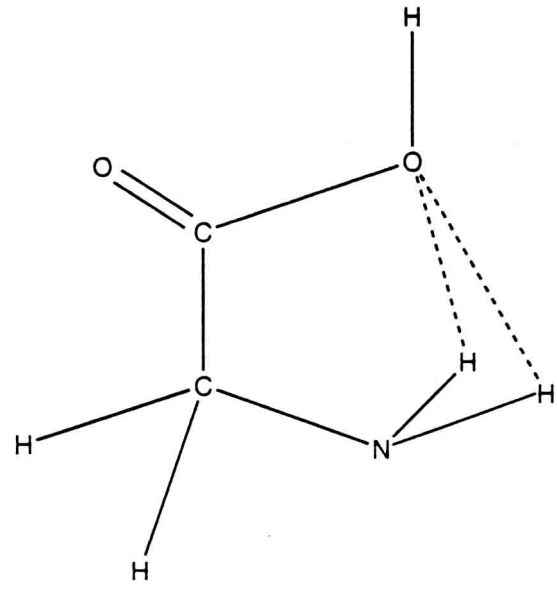

FIGURE 1.1 A conformer of glycine.

The basic difference between the two sets of charges is, as can be seen, the charge on C1 and the hydrogens attached to it. While the Mulliken population analysis method predicts a high polarity to the C–H bond, with the negative charge set on the carbon and a positive charge on the hydrogen, for the same bond the Merz–Kollman–Singh method predicts the hydrogen to be almost neutral and sets the positive charge on the carbon. The negative charge is significantly set on the nitrogen attached to the same carbon.

However, a set of calculations on acetylene, at HF/6-311G** optimized geometry, features the predicted by Mulliken population analysis charges as –0.129 on the carbons and 0.129 on the hydrogens. The Merz–Kollman–Singh method increases the charge separation to –0.302 on the carbons and 0.302 on the hydrogens. If the positive ion of acetylene is investigated by the same method in the Mulliken population analysis case, the charge of 1 is spread almost equally among the carbons and the hydrogens (0.237 and 0.263, respectively), while the Merz–Kollman–Singh method sets more charge on the hydrogens (0.167 on the carbons and 0.333 on the hydrogens). Therefore, it is hard to derive general conclusions about the trend of the differences between the two methods.