ACKNOWLEDGEMENTS
Chapter 10: ENERGY REFINEMENT OF TERTIARY STRUCTURE CHANGES CAUSED BY OXYGENATION OF HORSE HAEMOGLOBIN
Chapter 3: ENERGY MINIMIZATION: TEST ON A TRIPEPTIDE
Chapter 4: THE PROGRAMMING SYSTEM
Chapter 7: UNCONSTRAINED ENERGY REFINEMENT OF PROTEINS
Chapter 8: INTERPRETING PROBLEMATIC REGIONS OF ELECTRON DENSITY MAPS USING CONVERGENT ENERGY REFINEMENT
Chapter 9: ENERGY REFINEMENT OF ENZYME/SUBSTRATE COMPLEXES: LYSOZYME AND HEXA-N-ACETYLGUCOSAMINE
Chapter 9: ENERGY REFINEMENT OF ENZYME/SUBSTRATE COMPLEXES: LYSOZYME AND HEXA-N-ACETYLGLUCOSAMINE

Copyright Michael Levitt. Tue Feb 2 11:47:47 EST 1999
Preface
Preface

CONFORMATION ANALYSIS OF PROTEINS

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

Chapter 2: EMPIRICAL ENERGY FUNCTIONS

Chapter 3: ENERGY MINIMIZATION: TEST ON A TRIPEPTIDE

Chapter 4: THE PROGRAMMING SYSTEM
Chapter 5: PROTEIN COORDINATE REFINEMENT

Chapter 6: ENERGY PARAMETERS FROM PROTEINS

Chapter 7: UNCONSTRAINED ENERGY REFINEMENT OF PROTEINS

Chapter 8: INTERPRETING PROBLEMATIC REGIONS OF ELECTRON DENSITY MAPS USING CONVERGENT ENERGY REFINEMENT

Chapter 9: ENERGY REFINEMENT OF ENZYME/SUBSTRATE COMPLEXES: LYSOZYME AND HEXA-N-ACETYLGUCOSAMINE

Chapter 10: ENERGY REFINEMENT OF TERTIARY STRUCTURE CHANGES CAUSED BY OXYGENATION OF HORSE HAEMOGLOBIN

Chapter 11: CONCLUSIONS AND FUTURE WORK
Preface  
CONFORMATION ANALYSIS OF PROTEINS  
ACKNOWLEDGEMENTS  
TABLE OF CONTENTS  

Chapter 1: INTRODUCTION  

Chapter 2: EMPIRICAL ENERGY FUNCTIONS  

Chapter 3: ENERGY MINIMIZATION: TEST ON A TRIPEPTIDE  

Chapter 4: THE PROGRAMMING SYSTEM
Chapter 5: PROTEIN COORDINATE REFINEMENT

Chapter 6: ENERGY PARAMETERS FROM PROTEINS

Chapter 7: UNCONSTRAINED ENERGY REFINEMENT OF PROTEINS

Chapter 8: INTERPRETING PROBLEMATIC REGIONS OF ELECTRON DENSITY MAPS USING...

Preface

CONFORMATION ANALYSIS OF PROTEINS

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

Chapter 2: EMPIRICAL ENERGY FUNCTIONS

Chapter 3: ENERGY MINIMIZATION: TEST ON A
Chapter 4: THE PROGRAMMING SYSTEM

Chapter 5: PROTEIN COORDINATE REFINEMENT
Chapter 6: ENERGY PARAMETERS FROM PROTEINS

Chapter 7: UNCONSTRAINED ENERGY REFINEMENT OF PROTEINS
Chapter 8: INTERPRETING PROBLEMATIC REGIONS OF ELECTRON DENSITY MAPS USING CONVERGENT ENERGY REFINEMENT

Chapter 9: ENERGY REFINEMENT OF ENZYME/SUBSTRATE COMPLEXES: LYSOZYME AND HEXA-N-ACETYLGLUCOSAMINE

Preface
2.2.4 Hydrogen Bonds

Hydrogen bonds, which are stronger than long-range interactions and weaker than short-range interactions, are treated specially. Studies of small molecules indicate the following properties of hydrogen bonds between groups found in proteins (Ramakrishnan and Prasad, 1971): (a) the distribution of donor-acceptor distances is sharp for a given type of bond. For example, the O...N distance of a O...H—N hydrogen bond is between 2.7 Å and 3.0 Å. (b) The angle O...H—N is close to 180 degrees, giving linear hydrogen bonds. (c) Hydrogen bonding is most favourable in the direction of the long-pair electrons on the acceptor atom. (d) The energy of the bond is between 3 and 8 Kilocalories/mole. (e) At large separations or angle distortions, the hydrogen bond potential should transform smoothly into a normal non-bonded potential. Hydrogen bonds in a protein are always between atoms that are part of an extended group; many of the properties of hydrogen bonds described above may be due to other atoms of the group. Thus, for the simplest description, the donor-acceptor interaction will be spherically symmetric, and all angular dependence will be then caused by interactions between other atoms in the group. Many more complicated empirical functions have been used for hydrogen bonds, but none has all the right properties. (See Ramachandran and Sasisekharan, 1966, for a review.) This section introduces a general hydrogen bond function, which can have all the correct properties.

A weak bonding potential well is added to the normal non-bonded interaction of donor and acceptor. This potential well is deep for good hydrogen bonds, gets shallower as the hydrogen-bond geometry distorts, and is flat when no hydrogen bonds can form. If the normal non-bonded potential is $G(r)$, then the total hydrogen-bonding potential is
2.2.4 Hydrogen bonds

Hydrogen bonds, which are stronger than long-range interactions and weaker than short-range interactions, are treated specially. Studies of small molecules indicate the following properties of hydrogen bonds between groups found in proteins (Ramakrishnan and Presad, 1971): (a) the distribution of donor-acceptor distances is sharp for a given type of bond. For example, the O...N distance of a O...H—N hydrogen bond is between 2.7 Å and 3.0 Å. (b) The angle O...H—N is close to 180 degrees, giving linear hydrogen bonds. (c) Hydrogen bonding is most favourable in the direction of the long-pair electrons on the acceptor atom. (d) The energy of the bond is between 3 and 8 Kilocalories/mole. (e) At large separations or angle distortions, the hydrogen bond potential should transform smoothly into a normal non-bonded potential. Hydrogen bonds in a protein are always between atoms that are part of an extended group; many of the properties of hydrogen bonds described above may be due to other atoms of the group. Thus, for the simplest description, the donor-acceptor interaction will be spherically symmetric, and all angular dependence will be then caused by interactions between other atoms in the group. Many more complicated empirical functions have been used for hydrogen bonds, but none has all the right properties. (See Ramachandran and Sasisekharan, 1968, for a review.) This section introduces a general hydrogen bond function, which can have all the correct properties.

A weak bonding potential well is added to the normal non-bonded interaction of donor and acceptor. This potential well is deep for good hydrogen bonds, gets shallower as the hydrogen-bond geometry distorts, and is flat when no hydrogen bonds can form. If the normal non-bonded potential is $G(r)$, then the total hydrogen-bonding potential is
Figure 2.2. Graphs of the approximate Gaussian function given by

\[ F(x) = 1 - 3.5 x^2 + 4.5 x^4 - 2.5 x^6 + 0.5 x^8 \]

the solid curve is the function value, \( F \); the dashed curve is the first derivative, \( DF \), divided by 2; the long-dashed curve is the second derivative \( DDF \), divided by 6. All three curves are 0.0 at \( x = 1.0 \) or \( -1.0 \).

Figure 2.3. Plots of the hydrogen-bonding potential for three different values of the angle 180° - N-H-O, which is the deviation from a linear hydrogen bond. The energy parameters used are: \( r_0 = 2.95 \) Å; \( \Delta r^o = 0.5 \) Å; \( \Theta_1 = 180^\circ \); \( \Delta \Theta_1 = 35^\circ \); \( \Theta_2 = 130^\circ \); \( \Delta \Theta_2 = 60^\circ \); \( \epsilon^o = -4.5 \) kcal/mole and the normal non-bonded potential \( G_1(r) \) is a Lennard-Jones '12-6' potential with \( \epsilon = 0.1 \) kcal/mole and \( r_m = 3.3 \) Å. For each curve \( \Theta_2 = 130^\circ \). For the solid curve \( \Theta_1 = 0^\circ \), for the dashed curve \( \Theta_1 = 35^\circ \).
$17.5^\circ, \text{ and for the long-dashed curve } \theta, = 35^\circ$. 

Figure 2.2. Graphs of the approximate Gaussian function given by

\[ F(x) = 1 - 3.5 x^2 + 4.5 x^4 - 2.5 x^6 + 0.5 x^8 \]

developed by 2; the dash curve is the first derivative, \( DF \), divided by 2; the long-dashed curve is the second derivative
\( DDF \), divided by 8. All three curves are 0.0 at \( x = 1.0 \) or -1.0.

Figure 2.3. Plots of the hydrogen-bonding potential for three different values of the angle \( 180^\circ - 2 \cdot \theta \cdot 0 \), which is the deviation from a linear hydrogen bond. The energy parameters used were: \( \theta = 2 \cdot 95 \) \( \lambda \); \( \Delta \theta = 5 \lambda \); \( \theta_1 = 180^\circ \); \( \Delta \theta_1 = 35^\circ \); \( \theta_2 = 130^\circ \); \( \Delta \theta_2 = 60^\circ \). \( \theta = -4.5 \) kcal (mol), and the normal van der waals potential \( \phi^0 \) is 1.0 kcal (mol).
KCAL/mole and the nonlinear non-bonded potential $U(r)$ is a Lennard-Jones 12-6 potential with $\epsilon = 0.1$ kcal/mole and $r_0 = 3.3 \ \AA$. For each curve $\theta_2 = 130^\circ$. For the solid curve $\theta_1 = 0^\circ$, for the dashed curve $\theta_1 = 17.5^\circ$ and for the long-dashed curve $\theta_1 = 35^\circ$.
LYS-VAL-PHE-GLY-ARG4-CYS-GLU-LEU-ALA-ALA-ALA-MET-LYS
ARG-HIS-GLY-LEU-ASP-ASN-TYR-ARG-GLY-TYR-SER-LEU-GLY
ASN-TRY-VAL-CYS-ALA-ALA-LYS-PHE-GLU-SER-ASN-PHE-ASN
THR-GLN-ALA-THR-ASN-ARG-ASN-THR-ASN-GLY-SER-THR-ASP
TYR-GLY-ILU-LEU-GLN-ILU-ASN-SER-ARG-TRY-TRY-6-CYS-ASP
ASN-GLY-ARG-THR-PRO-GLY-SER-ARG-ASN-LEU7-CYS-ASN-ILU
PRO6-CYS-SER-ALA-LEU-LEU-SER-SER-ASP-ILU-THR-ALA-SER
VAL-ASN7-CYS-ALA-LYS-GLY-ILU-VAL-SER-ASP-GLY-ASP-GLY
MET-ASN-ALA-TRY-VAL-ALA-TRY-ARG-ASN-ARG5-CYS-LYS-GLY
THR-ASP-VAL-GLN-ALA-ILU-ARG-TRY-GLY4-CYS-ARG-LEU-END

Figure 2.7a. The abbreviated chemical formula of hen egg-white lysozyme.
All residues besides CYS are preceded by a dash. The same numeral, which cannot be '2' or '3', precedes the pair of cystine half-bridges that are chemically bonded through S-S bonds.

GLY NC(EO)  THR NC(CVC)EO
ALA NC(C)EO  VAL NC(C(C)C)EO
ASN NC(CEON)EO  PRO *NC(+(C)(+C)C)EO
ASP NC(CEOD)EO  MET NC(CCSC)EO
GLN NC(CEOD)EO  CYS NC(C=5)EO
GLU NC(CCEOD)EO  HIS NC(C*BNN*B)EO
ILU NC(C(C)CC)EO  PHE NC(C*AAAA*A)EO
LEU NC(CC(C)CC)EO  TYR NC(C*AAAAAV*A)EO
LYS NC(CCCCM)EO  TRY NC(C*BNN*GAAAA*G)EO
ARG NC(CCCCA(m)m)EO  END 0
SER NC(CV)EO

Figure 2.7b. The library used to translate the abbreviated chemical formula of lysozyme. The '=' that precedes the S of half cystines is replaced by the numeral that precedes the particular CYS in the abbreviated chemical formula above.
Figure 2.7a. The abbreviated chemical formula of hen egg-white lysozyme. All residues besides CYS are preceded by a dash. The same numeral, which cannot be '2' or '3', precedes the pair of cystine half-bridges that are chemically bonded through S-S bonds.

Figure 2.7b. The library used to translate the abbreviated chemical formula
of lysozyme. The '=' that precedes the S of half cystines is replaced by the numeral that precedes the particular CYS in the abbreviated chemical formula above.
2.4 Calculating derivatives

2.4.1 choosing a coordinate system

Minimization methods use a Taylor expansion of the energy about a point in conformation space to find the nearest minimum. These expansions require partial derivatives of the energy; for example, a second-order Taylor expansion requires first and second derivatives of the energy function with respect to the coordinates of the conformation space. The first derivatives of the energy, which are the forces in each coordinate direction, are zero for any equilibrium conformation. The matrix of second derivatives, which is positive definite at a minimum, gives the normal modes of vibration about the equilibrium (Lifson and Warshel, 1968).

Derivatives can be calculated in two ways: (a) Numerically from the mathematical definition of a derivative

$$\frac{\partial E}{\partial q} = \frac{E(q + \epsilon) - E(q)}{\epsilon} \quad \text{as} \quad \epsilon \to 0$$

This method, which involves little programming effort, requires n+1 evaluations of the energy function for the first derivatives, and $\frac{1}{2}n^2$ more evaluations for the second derivatives. Unless the computation is exceptionally accurate, the second derivatives will be too inaccurate to be useful. (b) Analytically, by programming the functions obtained by direct differentiation of the energy function. This is always more accurate and more efficient, and the complicated derivative formulae can be checked by comparing the analytical and numerical results.
2.4 Calculating derivatives

2.4.1 choosing a coordinate system

Minimization methods use a Taylor expansion of the energy about a point in conformation space to find the nearest minimum. These expansions require partial derivatives of the energy; for example, a second-order Taylor expansion requires first and second derivatives of the energy function with respect to the coordinates of the conformation space. The first derivatives of the energy, which are the forces in each coordinate direction, are zero for any equilibrium conformation. The matrix of second derivatives, which is positive definite at a minimum, gives the normal modes of vibration about the equilibrium (Lifson and Warshel, 1968).

Derivatives can be calculated in two ways: (a) Numerically from the mathematical definition of a derivative

\[
\frac{\partial E}{\partial \vec{q}} = \frac{E(\vec{q} + \epsilon) - E(\vec{q})}{\epsilon} \quad \text{as} \quad \epsilon \to 0
\]

This method, which involves little programming effort, requires n+1 evaluations of the energy function for the first derivatives, and \( \frac{1}{2}n^2 \) more evaluations for the second derivatives. Unless the computation is exceptionally accurate, the second derivatives will be too inaccurate to be useful. (b) Analytically, by programming the functions obtained by direct differentiation of the energy function. This is always more accurate and more efficient, and the complicated derivative formulae can be checked by comparing the analytical and
numerical results.
The empirical energy function, which depends explicitly on bond lengths, bond-angles, torsion-angles, and non-bonded interatomic distances, must be differentiated with respect to a set of independent variables. Two sets are available, (a) a set of independent internal coordinates chosen from all bonds, bond-angles, and torsion-angles, and (b) the set of atomic Cartesian coordinates. Although all non-bonded interatomic distances are really internal coordinates, they are derived easily from the smaller set of atomic Cartesian coordinates. Thus, the empirical energy can be written as the sum of a function, \( F(s) \), of internal coordinates, \( s \), (for each bond, bond-angle, and torsion-angle) and a function \( G(q) \) of Cartesian coordinates, \( q \), (for non-bonded distances)

\[
E = F(s) + G(q)
\]

The derivatives with respect to \( s \) and \( q \) are

\[
\frac{dE}{dq} = \sum_s \frac{dF(s)}{ds} \cdot \frac{ds}{dq} + \frac{dG(q)}{dq}
\]

\[
\frac{dE}{ds} = \frac{dF(s)}{ds} + \sum_q \frac{dG(q)}{dq} \cdot \frac{dq}{ds}
\]

Explicit differentiation of \( F(s) \) to give \( \frac{dF(s)}{ds} \), and of \( G(q) \) to give \( \frac{dG(q)}{dq} \) is trivial when all the internal coordinates, \( s \), are independent.
The empirical energy function, which depends explicitly on bond lengths, bond-angles, torsion-angles, and non-bonded interatomic distances, must be differentiated with respect to a set of independent variables. Two sets are available, (a) a set of independent internal coordinates chosen from all bonds, bond-angles, and torsion-angles, and (b) the set of atomic Cartesian coordinates. Although all non-bonded interatomic distances are really internal coordinates, they are derived easily from the smaller set of atomic Cartesian coordinates. Thus, the empirical energy can be written as the sum of a function, $F(s)$, of internal coordinates, $s$, (for each bond, bond-angle, and torsion-angle) and a function $G(q)$ of Cartesian coordinates, $q$, (for non-bonded distances)

$$
E = F(s) + G(q)
$$

The derivatives with respect to $s$ and $q$ are

$$
\frac{dE}{dq} = \sum_s \frac{dF(s)}{ds} \cdot \frac{ds}{dq} + \frac{dG(q)}{dq}
$$

$$
\frac{dE}{ds} = \frac{dF(s)}{ds} + \sum_q \frac{dG(q)}{dq} \cdot \frac{dq}{ds}
$$

Explicit differentiation of $F(s)$ to give $\frac{dF(s)}{ds}$, and of $G(q)$ to give $\frac{dG(q)}{dq}$ is trivial when all the internal coordinates, $s$, are independent.
Calculating derivatives in Cartesian coordinates needs the partial derivatives of each internal coordinate in Cartesian coordinates; whereas calculating derivatives in internal coordinates needs the partial derivatives of each Cartesian coordinate in internal coordinates. Thus, the choice of coordinates depends on the relative efficiency of these alternatives.

Each atomic Cartesian coordinate affects only those internal coordinates of which it is a part. A single internal coordinate depends on at most 12 Cartesian coordinates (the x, y and z of the four atoms forming the torsion-angle). Thus, the partial derivative matrix, \[ \frac{\partial s_i}{\partial q_j}, \] and tensor, \[ \frac{\partial^2 s_i}{\partial q_j \partial q_k}, \] which are needed to calculate energy derivatives in Cartesian coordinates, are very sparse: their number of non-zero elements is proportional to n, the number of atoms in the molecule. Atomic Cartesian coordinates are independent and equivalent, and the complicated derivative formulae are highly symmetric.

Each internal coordinate affects all the Cartesian coordinates of atoms after it in the chain (assuming that the beginning of the chain is fixed).

The matrix, \[ \frac{\partial q_i}{\partial s_j}, \] and tensor, \[ \frac{\partial^2 q_i}{\partial s_j \partial s_k}, \] needed to calculate energy derivative in internal coordinates, are not sparse: their number of non-zero elements is proportional to \( n^2 \) and \( n^3 \), respectively. Internal coordinates are not all independent: only four of the six bond-angles at a tetravalent atom are independent; six internal coordinates depend on the others in a closed ring. Finally, the full set of independent internal coordinates, which depends on the topology of the molecule, cannot be chosen easily for a general molecule.

When a molecule needs all degrees of freedom, Cartesian coordinates are better than internal coordinates for derivative calculation. Lifson and Warshel (1968), Warshel, Levitt, and Lifson (1969), and Warshel and Lifson (1970)
Calculating derivatives in Cartesian coordinates needs the partial derivatives of each internal coordinate in Cartesian coordinates; whereas calculating derivatives in internal coordinates needs the partial derivatives of each Cartesian coordinate in internal coordinates. Thus, the choice of coordinates depends on the relative efficiency of these alternatives.

Each atomic Cartesian coordinate affects only those internal coordinates of which it is a part. A single internal coordinate depends on at most 12 Cartesian coordinates (the x, y and z of the four atoms forming the torsion-angle). Thus, the partial derivative matrix, \( \frac{\partial S_i}{\partial q_j} \), and tensor, \( \frac{\partial^2 S_i}{\partial q_j \partial q_k} \), which are needed to calculate energy derivatives in Cartesian coordinates, are very sparse: their number of non-zero elements is proportional to \( n \), the number of atoms in the molecule. Atomic Cartesian coordinates are independent and equivalent, and the complicated derivative formulae are highly symmetric.

Each internal coordinate affects all the Cartesian coordinates of atoms after it in the chain (assuming that the beginning of the chain is fixed). The matrix, \( \frac{\partial q_i}{\partial S_j} \), and tensor, \( \frac{\partial^2 q_i}{\partial S_j \partial S_k} \), needed to calculate energy derivative in internal coordinates, are not sparse: their number of non-zero elements is proportional to \( n^2 \) and \( n^3 \), respectively. Internal coordinates are not all independent: only four of the six bond-angles at a tetravalent atom are independent; six internal coordinates depend on the others in a closed ring. Finally, the full set of independent internal coordinates, which depends on the topology of the molecule, cannot be chosen easily for a general molecule.

When a molecule needs all degrees of freedom, Cartesian coordinates are better than internal coordinates for derivative calculation. Lifson and
Warshel (1968), Warshel, Levitt, and Lifson (1969), and Warshel and Lifson (1970)
have used Cartesian derivatives in detailed calculations of the enthalpies, conformations, normal-mode vibration spectra, and crystal packing of small molecules. Levitt and Lifson (1969) refine protein conformations by moving all the atomic Cartesian coordinates simultaneously to minimize an empirical energy function.

When a molecule must have degrees of freedom limited to one class of internal coordinate, then these internal coordinate variables should be used. For example, the torsion-angles about single bonds used in protein conformation analysis. Even when the variables are torsion-angles, the non-bonded energy term is still differentiated in Cartesian space, and then transformed to torsion-angle space. Equation 2d assumes that the summation, implicit in G(q), of terms that depend on each non-bonded distance is completed before the Cartesian derivatives are transformed to torsion-angle derivatives. This procedure is not used generally. Some workers transform the Cartesian derivatives of each non-bonded term individually, for then only terms that are affected by the particular torsion-angle need to be considered. The relative efficiency of the two approaches depends on the ratio of the number of affected non-bonded distances to the number of Cartesian coordinates. When there are fewer affected distances than Cartesian coordinates, as in small dipeptides, then the terms of the summation over non-bonded distances should be transformed individually. When the reverse is true, as in larger molecules, the Cartesian derivatives of non-bonded interactions should be summed and then transformed as in Equation 2d. The latter method has a second advantage when torsion-angles in rings of atoms are varied and constraints are needed to close the ring. The ring is broken at one bond; the energy contribution of the one bond, two bond-angles, and three torsion-angles affected by the break are included in the calculation of Cartesian derivatives; and the summed Cartesian derivative is finally transformed to torsion-angle space.
have used Cartesian derivatives in detailed calculations of the enthalpies, conformations, normal-mode vibration spectra, and crystal packing of small molecules. Levitt and Lifson (1969) refine protein conformations by moving all the atomic Cartesian coordinates simultaneously to minimize an empirical energy function.

When a molecule must have degrees of freedom limited to one class of internal coordinate, then these internal coordinate variables should be used. For example, the torsion-angles about single bonds used in protein conformation analysis. Even when the variables are torsion-angles, the non-bonded energy term is still differentiated in Cartesian space, and then transformed to torsion-angle space. Equation 2d assumes that the summation, implicit in $G(q)$, of terms that depend on each non-bonded distance is completed before the Cartesian derivatives are transformed to torsion-angle derivatives. This procedure is not used generally. Some workers transform the Cartesian derivatives of each non-bonded term individually, for then only terms that are affected by the particular torsion-angle need to be considered. The relative efficiency of the two approaches depends on the ratio of the number of affected non-bonded distances to the number of Cartesian coordinates. When there are fewer affected distances than Cartesian coordinates, as in small dipeptides, then the terms of the summation over non-bonded distances should be transformed individually. When the reverse is true, as in larger molecules, the Cartesian derivatives of non-bonded interactions should be summed and then transformed as in Equation 2d. The latter method has a second advantage when torsion-angles in rings of atoms are varied and constraints are needed to close the ring: The ring is broken at one bond; the energy contribution of the one bond, two bond-angles, and three torsion-angles affected by the break are included in the calculation of Cartesian derivatives; and the summed Cartesian derivative is finally transformed to torsion-angle space.
2.4.2 Cartesian derivatives

This section presents formulae for the partial derivatives of bond lengths, bond-angles, and torsion-angles with respect to Cartesian coordinates. Latin subscripts index the atoms; Greek superscripts indicate a component of the three-dimensional Cartesian coordinate vector. For example, \( q_i^\alpha \) is the \( x, y, \) and \( z \) coordinate of the \( i^{th} \) atom for \( \alpha = 1, 2, \) and \( 3, \) respectively.

The vector \( \overrightarrow{b}_{ij} \), which is directed from atom \( j \) to atom \( i \), has components \( q_i^\alpha - q_j^\alpha \); the bond length is the modulus of this vector. The bond-angle involving atoms \( i, j, \) and \( k \) is given by

\[
\gamma_{ijk} = \arccos \left\{ \frac{\overrightarrow{b}_{ij} \cdot \overrightarrow{b}_{kj}}{|\overrightarrow{b}_{ij}| \cdot |\overrightarrow{b}_{kj}|} \right\}
\]

The torsion-angle involving atoms \( i, j, k, \) and \( m \) is given by

\[
\Theta_{ijkm} = \arccos \left\{ \frac{(\overrightarrow{b}_{ij} \times \overrightarrow{b}_{kj}) \cdot (\overrightarrow{b}_{jk} \times \overrightarrow{b}_{km})}{|\overrightarrow{b}_{ij} \times \overrightarrow{b}_{kj}| \cdot |\overrightarrow{b}_{jk} \times \overrightarrow{b}_{km}|} \right\}
\]

Coordinate transformations simplify the expressions for the first and second derivatives of \( \overrightarrow{b}_{ij}, \ \gamma_{ijk}, \) and \( \Theta_{ijkm}. \) The first and second derivatives of function \( f(X), \) where \( X \) is itself a function of \( q_i, \) with respect to \( q \) are given by
2.4.2 Cartesian derivatives

This section presents formulae for the partial derivatives of bond lengths, bond-angles, and torsion-angles with respect to Cartesian coordinates. Latin subscripts index the atoms; Greek superscripts indicate a component of the three-dimensional Cartesian coordinate vector. For example, \( q_i^\alpha \) is the \( x, y, \) and \( z \) coordinate of the \( i^{\text{th}} \) atom for \( \alpha = 1, 2, \) and \( 3, \) respectively.

The vector \( \vec{b}_{ij} \), which is directed from atom \( j \) to atom \( i \), has components \( q_i^\alpha - q_j^\alpha \); the bond length is the modulus of this vector. The bond-angle involving atoms \( i, j, \) and \( k \) is given by

\[
\gamma_{ijk} = \arccos \left\{ \frac{\vec{b}_{ij} \cdot \vec{b}_{kj}}{||\vec{b}_{ij}|| \cdot ||\vec{b}_{kj}||} \right\}
\]

The torsion-angle involving atoms \( i, j, k, \) and \( m \) is given by

\[
\Theta_{ijkm} = \arccos \left\{ \frac{(\vec{b}_{ij} \times \vec{b}_{kj}) \cdot (\vec{b}_{jk} \times \vec{b}_{km})}{||\vec{b}_{ij} \times \vec{b}_{kj}|| \cdot ||\vec{b}_{jk} \times \vec{b}_{km}||} \right\}
\]

Coordinate transformations simplify the expressions for the first and second derivatives of \( \vec{b}_{ij}, \gamma_{ijk}, \) and \( \Theta_{ijkm}. \) The first and second derivatives of function \( f(X) \), where \( X \) is itself a function of \( q, \) with respect to \( q \) are given by
\[
\frac{df}{dq_i^\alpha} = \sum_\gamma \frac{df}{dX^\gamma} \frac{dX^\gamma}{dq_i^\alpha},
\]
\[
\frac{d^2f}{dq_i^\alpha dq_j^\beta} = \sum_{\gamma, \delta} \frac{d^2f}{dX^\gamma dX^\delta} \frac{dX^\gamma}{dq_i^\alpha} \frac{dX^\delta}{dq_j^\beta} + \sum_{\gamma} \frac{df}{dX^\gamma} \frac{d^2X^\gamma}{dq_i^\alpha dq_j^\beta}
\]

Thus, derivatives of each internal coordinate are calculated in two stages: (a) each internal coordinate is expressed as a function of \(X\), and (b) each dummy variable, \(X\), is expressed as a function of the Cartesian coordinates, \(q\). The dummy variables, \(X\), are introduced to replace the differentiation of one complex function by the differentiation of two simple functions.

The bond length is the modulus of \(\bar{X}\), where \(X^\alpha\) is defined as \(q_i^\alpha - q_j^\alpha\), for \(\alpha = 1, 2, 3\). The first and second derivatives of the bond length with respect to \(X\) are
\[
\frac{db}{dX^\gamma} = \frac{\bar{X}^\gamma}{b}
\]
\[
\frac{d^2b}{dX^\gamma dX^\delta} = \frac{1}{b} \left\{1 - \left(\frac{\bar{X}^\gamma}{b}\right)^2\right\} \quad \text{if} \quad \gamma = \delta
\]
\[
= -\frac{\bar{X}^\gamma \bar{X}^\delta}{b^3} \quad \text{if} \quad \gamma \neq \delta
\]
\[
\frac{df}{dq_i^\alpha} = \sum_{\gamma} \frac{df}{dX^{\gamma}} \frac{dX^{\gamma}}{dq_i^\alpha}
\]

\[
\frac{d^2f}{dq_i^\alpha dq_j^\beta} = \sum_{\gamma, \delta} \frac{d^2f}{dX^{\gamma}dX^{\delta}} \frac{dX^{\gamma}}{dq_i^\alpha} \frac{dX^{\delta}}{dq_j^\beta} + \sum_{\gamma} \frac{df}{dX^{\gamma}} \frac{d^2X^{\gamma}}{dq_i^\alpha dq_j^\beta}
\]

Thus, derivatives of each internal coordinate are calculated in two stages: (a) each internal coordinate is expressed as a function of \(X\), and (b) each dummy variable, \(X\), is expressed as a function of the Cartesian coordinates, \(q\). The dummy variables, \(X\), are introduced to replace the differentiation of one complex function by the differentiation of two simple functions.

The bond length is the modulus of \(\vec{X}\), where \(X^\alpha\) is defined as

\[q_1^\alpha = q_3^\alpha, \text{ for } \alpha = 1, 2, 3.\]

The first and second derivatives of the bond length with respect to \(X\) are

\[\frac{db}{dX^{\gamma}} = \frac{X^{\gamma}}{b}\]

\[\frac{d^2b}{dX^{\gamma}dX^{\delta}} = \frac{1}{b} \left\{ \begin{array}{ll} 1 - \left( \frac{X^{\gamma}}{b} \right)^2 & \text{if } \gamma = \delta \\
 X^{\delta} & \text{if } \gamma \neq \delta \end{array} \right.\]
The first derivatives of $X$ with respect to $q_i^\alpha$ and $q_j^\beta$ are given by

$$\frac{\partial X^\gamma}{\partial q_i^\alpha} = - \frac{\partial X^\gamma}{\partial q_j^\beta} = \delta^\gamma_\alpha,$$

where $\delta^\gamma_\alpha = 1$ if $\alpha = \gamma$,

$= 0$ if $\alpha \neq \gamma$.

All the second derivatives are zero. Coordinate transformation gives the derivative of $b_{ij}$ with respect to $q$. Formulæ for the derivatives of non-bonded distances are the same as for bond lengths.

Bond angles are computed by making the transformation

$$\cos\left(\tau_{ijk}\right) = \frac{\vec{X}_i \cdot \vec{X}_k}{|\vec{X}_i| |\vec{X}_k|}$$

where $X_i^\alpha = q_i^\alpha - q_i$, $X_k^\alpha = q_k^\alpha - q_k$, for $\alpha = 1, 2, 3$; $B_i = |\vec{X}_i|$, and $B_k = |\vec{X}_k|$. The derivative of the angle is calculated from the derivative of the cosine using

$$\frac{d\tau}{dX^\alpha} = -\frac{1}{\sin\tau} \frac{d(\cos\tau)}{dX^\alpha}.$$
The first derivatives of $X$ with respect to $q_i$ and $q_j$ are given by

$$\frac{dX^r}{dq_i^\alpha} = - \frac{dX^r}{dq_j^\alpha} = \delta^r_\alpha$$

where

$$\delta^r_\alpha = 1 \quad \text{if} \quad \alpha = r$$

$$= 0 \quad \text{if} \quad \alpha \neq r$$

All the second derivatives are zero. Coordinate transformation gives the derivative of $b_{ij}$ with respect to $q$. Formulae for the derivatives of non-bonded distances are the same as for bond lengths.

Bond angles are computed by making the transformation

$$\cos \left( \tau_{ijk} \right) = \frac{X_{ij}}{B_i} \cdot \frac{X_{jk}}{B_k}$$

where $X_i^\alpha = q_i^\alpha - q_j^\alpha$, $X_k^\alpha = q_k^\alpha - q_j^\alpha$, for $\alpha = 1, 2, 3$; $B_i = |X_i|$, and $B_k = |X_k|$. The derivative of the angle is calculated from the derivative of the cosine using

$$\frac{d\tau}{dX^\alpha} = -\frac{1}{\sin \tau} \frac{d(\cos \tau)}{dX^\alpha}$$
\[
\frac{\partial \tau}{\partial x^\alpha \partial x^\beta} = - \frac{\cos \tau}{\sin^3 \tau} \frac{\partial (\cos \tau)}{\partial x^\alpha} \frac{\partial (\cos \tau)}{\partial x^\beta} - \frac{1}{\sin \tau} \frac{\partial^2 (\cos \tau)}{\partial x^\alpha \partial x^\beta}
\]

The first derivative of cosine \((\tau_{ijk})\) with respect to \(x_i\) and \(x_k\) is given by

\[
\frac{\partial (\cos \tau)}{\partial x_i^\alpha} = \frac{1}{B_i} \left( \frac{x_k^\alpha}{B_k} - \cos \tau \frac{x_i^\alpha}{B_i} \right) \quad \text{for} \quad \alpha = 1, 2, 3 \quad \text{i & k interchange}
\]

The four distinct second derivatives are

\[
\frac{\partial^2 (\cos \tau)}{\partial x_i^\alpha \partial x_i^\alpha} = - \frac{1}{B_i^2} \left\{ 2x_i^\alpha \frac{\partial (\cos \tau)}{\partial x_i^\alpha} + \cos \tau \left( 1 - \frac{x_i^\alpha}{B_i} \right)^2 \right\} \quad \text{\(\alpha \& \beta, i \& k\ interchanged\)}
\]

\[
\frac{\partial^2 (\cos \tau)}{\partial x_i^\alpha \partial x_k^\alpha} = - \frac{x_i^\alpha}{B_i^2} \frac{\partial (\cos \tau)}{\partial x_k^\alpha} - \frac{x_k^\alpha}{B_k^2} \frac{\partial (\cos \tau)}{\partial x_i^\alpha} - \frac{1}{B_i B_k} \left( \frac{x_i^\alpha x_k^\alpha \cos \tau}{B_i B_k} - 1 \right) \quad \text{\(\alpha \& \beta\ interchanged\)}
\]

\[
\frac{\partial^2 (\cos \tau)}{\partial x_i^\alpha \partial x_j^\beta} = - \frac{1}{B_i^2 B_k} \left( x_k^\alpha x_i^\beta + x_i^\alpha x_k^\beta \right) + \frac{3}{B_i^4} \left( x_i^\alpha x_j^\beta \cos \tau \right)
\]
\[
\frac{d^2 \tau}{dX^\alpha dx^\beta} = -\frac{\cos \tau}{\sin^3 \tau} \frac{d \left( \cos \tau \right)}{dX^\alpha} \frac{d \left( \cos \tau \right)}{dx^\beta} - \frac{1}{\sin \tau} \frac{d^2 \left( \cos \tau \right)}{dX^\alpha dx^\beta}
\]

The first derivative of cosine \( \tau_{ijk} \) with respect to \( X_i \) and \( X_k \) is given by

\[
\frac{d \left( \cos \tau \right)}{dX_i^\alpha} = \frac{1}{B_i} \left( \frac{X_k^\alpha}{B_k} \cos \tau \frac{X_i^\alpha}{B_i} \right)
\]

for \( \alpha = 1, 2, 3 \)

\( i \) & \( k \) interchanged

The four distinct second derivatives are

\[
\frac{d^2 \left( \cos \tau \right)}{dX_i^\alpha dx_i^\alpha} = -\frac{1}{B_i^2} \left\{ 2X_i^\alpha \frac{d \left( \cos \tau \right)}{dX_i^\alpha} + \cos \tau \left( 1 - \frac{X_i^\alpha}{B_i} \right)^2 \right\}
\]

\( \alpha \& \beta, \ i \& k \) interchanged

\[
\frac{d^2 \left( \cos \tau \right)}{dX_i^\alpha dx_i^\alpha} = -\frac{X_i^\alpha}{B_i^2} \frac{d \left( \cos \tau \right)}{dx_i^\alpha} - \frac{X_k^\alpha}{B_k^2} \frac{d \left( \cos \tau \right)}{dx_k^\alpha} - \frac{1}{B_i B_k} \left( \frac{X_i^\alpha X_k^\alpha \cos \tau}{B_i B_k} - 1 \right)
\]

\( \alpha \& \beta \) interchanged

\[
\frac{d^2 \left( \cos \tau \right)}{dx^\alpha dx^\beta} = -\frac{1}{\sin \tau} \left( \frac{X_i^\alpha X_i^\beta + X_i^\alpha X_i^\beta}{2} \right) + \frac{3}{\sin \tau} \left( X_i^\alpha X_i^\beta \cos \tau \right)
\]
\[ \frac{d}{dx_i} \frac{d}{dx_j} B_i^2 B_k \]

i & k interchanged
\[
\frac{\partial (\cos \tau)}{\partial X_i^\alpha \partial X_k^\beta} = - \frac{X^\alpha_k X^\beta_k}{B_i B_k^3} - \frac{X^\alpha_i X^\beta_i}{B_i^3 B_k} + \frac{X^\alpha_i X^\beta_k \cos \tau}{B_i B_k^2}
\]

The derivatives of the transformed coordinates \(X\) with respect to the Cartesian coordinates are

\[
\frac{\partial X_i^\alpha}{\partial q^\beta_r} = \delta^\alpha_\beta \quad \text{if} \quad \tau = i
\]

\[
= -\delta^\alpha_\beta \quad \text{if} \quad \tau = j
\]

\[
= 0 \quad \text{if} \quad \tau = k
\]

\[
\frac{\partial^2 X_i^\alpha}{\partial q_j^\beta \partial q_k^\delta} = 0 \quad \text{always}
\]

Once again coordinate transformation gives the derivatives of \(\tau_{ijk}\) in Cartesian coordinates, \(q\).

Torsion-angles are computed by making the transformation

\[
\bar{X}_i = (\bar{q}_i - \bar{q}_j) \times (\bar{q}_k - \bar{q}_j); \quad \bar{X}_k = (\bar{q}_j - \bar{q}_k) \times (\bar{q}_m - \bar{q}_k)
\]

This makes the formula for a torsion-angle like that for a bond-angle, and the same derivative formulae apply for derivations with respect to the dummy variable, \(\bar{X}\). The first and second derivatives of \(\bar{X}\) with respect to \(q\) are of the form.
\[
\frac{\partial (\cos \zeta)}{\partial x_i^\alpha \partial x_k^\beta} = - \frac{X_i^\alpha X_k^\beta}{B_i B_k} - \frac{X_i^\alpha X_i^\beta}{B_i^3 B_k} + \frac{X_i^\alpha X_k^\beta \cos \zeta}{B_i B_k^2}
\]

The derivatives of the transformed coordinates \( X \) with respect to the Cartesian coordinates are

\[
\frac{\partial X_i^\alpha}{\partial q_j^\beta} = \delta^\alpha_j \quad \text{if} \quad \tau = i
\]

\[
= - \delta^\alpha_j \quad \text{if} \quad \tau = j
\]

\[
= 0 \quad \text{if} \quad \tau = k
\]

\[
\frac{\partial^2 X_i^\alpha}{\partial q_j^\beta \partial q_k^\gamma} = 0 \quad \text{always}
\]

Once again coordinate transformation gives the derivatives of \( \gamma_{ijk} \)

in Cartesian coordinates, \( q \).

Torsion-angles are computed by making the transformation

\[
\overline{X}_i = (q_i - q_j) \times (q_k - q_j) ; \quad \overline{X}_k = (q_j - q_k) \times (q_m - q_k)
\]

This makes the formula for a torsion-angle like that for a bond-angle, and

the same derivative formulae apply for derivations with respect to the dummy

variable, \( X \). The first and second derivatives of \( X \) with respect to \( q \) are

of the form
\[ E_{HB} = G(r) - d \cdot H(t) \]

where \( H(t) \) defines the potential well, and \( t \) is given by

\[
t^2 = \left( \frac{r - r^0}{\Delta r} \right)^2 + \left( \frac{\theta_1 - \theta_1^0}{\Delta \theta_1} \right)^2 + \left( \frac{\theta_2 - \theta_2^0}{\Delta \theta_2} \right)^2
\]

The donor-acceptor separation is \( r \), \( \theta_1 \) is the angle at the acceptor (for example, \( \text{N} \cdots \text{O} \cdots \text{C} \)), and \( \theta_2 \) is the deviation from linearity (for example \( \text{N} \cdots \text{H} \cdots \text{O} \) or \( \text{H} \cdots \text{N} \cdots \text{O} \)). Perfect hydrogen bonds occur when \( r = r^0 \), \( \theta_1 = \theta_1^0 \), and \( \theta_2 = \theta_2^0 \). Tolerated deviations from the perfect bonds are determined by \( \Delta r \), \( \Delta \theta_1 \), and \( \Delta \theta_2 \). The parameter \( d \) is chosen to make the energy of a perfect hydrogen bond equal to \( E^0 \). Since \( t = 0 \) and \( H(0) = 1 \) for a perfect hydrogen bond \( d = E^0 - G(r^0) \). The function \( H(t) \), which is an empirical approximation to a Gaussian function, is shown in Figure 2.2. When \( t^2 \) is greater than 1, the normal non-bonded potential is used. A smooth transfer to the hydrogen bonding potential as \( t^2 \) decreases is assured, for \( H(t) \) has first and second derivatives that are zero at \( t^2 = 1 \). Figure 2.3 shows plots of hydrogen-bonding potentials with a realistic set of parameters.

2.2.5 truncating long-range interactions

All the non-bonded interactions probably can be neglected at atomic
\[ E_{H^6} = G(r) - d \cdot H(t) \]

where \( H(t) \) defines the potential well, and \( t \) is given by

\[ t^2 = \left( \frac{r - r^0}{\Delta r} \right)^2 + \left( \frac{\Theta_1 - \Theta_1^0}{\Delta \Theta_1} \right)^2 + \left( \frac{\Theta_2 - \Theta_2^0}{\Delta \Theta_2} \right)^2 \]

The donor-acceptor separation is \( r \), \( \Theta_1 \) is the angle at the acceptor (for example, \( N \cdots H \cdots O \)), and \( \Theta_2 \) is the deviation from linearity (for example \( N \cdots H \cdots O \) or \( H \cdots N \cdots O \)). Perfect hydrogen bonds occur when \( r = r^0 \), \( \Theta_1 = \Theta_1^0 \), and \( \Theta_2 = \Theta_2^0 \). Tolerated deviations from the perfect bonds are determined by \( \Delta r \), \( \Delta \Theta_1 \), and \( \Delta \Theta_2 \). The parameter \( d \) is chosen to make the energy of a perfect hydrogen bond equal to \( E^0 \). Since \( t = 0 \) and \( H(0) = 1 \) for a perfect hydrogen bond \( d = E^0 - G(r^0) \). The function \( H(t) \), which is an empirical approximation to a Gaussian function, is shown in Figure 2.2. When \( t^2 \) is greater than 1, the normal non-bonded potential is used. A smooth transfer to the hydrogen bonding potential as \( t^2 \) decreases is assured, for \( H(t) \) has first and second derivatives that are zero at \( t^2 = 1 \). Figure 2.3 shows plots of hydrogen-bonding potentials with a realistic set of parameters.

### 2.2.5 truncating long-range interactions

All the non-bonded interactions probably can be neglected at atomic
\[
\frac{\partial X_i}{\partial q_j} = q_k - q_m \quad \text{where} \quad r, s, k, m \text{ depend on } \alpha, \beta, i, j.
\]

\[
\frac{\partial^2 X_i}{\partial q_j^2} = 1, 0, -1 \quad \text{depending on } i, k, j, \alpha, \beta, r.
\]

Again the coordinate transformation is used to get the derivatives of the torsion-angle \( \theta_{ijkm} \) in Cartesian coordinates.

2.4.3 Torsion-angle derivatives

When exploring conformation space or convergently minimizing the energy of large molecules, the number of variables must be small. A sub-set of the internal coordinates, the torsion-angles about single bonds, becomes important, as a protein with 3n Cartesian coordinates has only about \( \frac{1}{4}n \) torsion-angles about single bonds. In addition, the topological arrangement of single bonds in a protein with no S-S bridges is simple, and all the torsion-angles about the single bonds are independent. When torsion-angles in closed rings are fixed, they cannot change bond lengths and bond-angles. Thus, the energy function in torsion-angle space only depends explicitly on torsion-angles, \( \Theta \) (for bond twisting), and Cartesian coordinates, \( q \) (for non-bonded interactions).
\[
\frac{\partial X_i}{\partial q_j} = Q_k - Q_m \quad \text{where } \alpha, \beta, \kappa, \mu \text{ depend on } \alpha, \beta, i, j
\]

\[
\frac{\partial^2 X_i}{\partial q_j \partial q_k} = 1, 0, -1 \quad \text{depending on } i, k, j, \alpha, \beta, \tau
\]

Again the coordinate transformation is used to get the derivatives of the torsion-angle \( \Theta_{i,j,k,m} \) in Cartesian coordinates.

2.4.3 torsion-angle derivatives

When exploring conformation space or convergently minimizing the energy of large molecules, the number of variables must be small. A sub-set of the internal coordinates, the torsion-angles about single bonds, becomes important, as a protein with \( 3n \) Cartesian coordinates has only about \( \frac{3}{2}n \) torsion-angles about single bonds. In addition, the topological arrangement of single bonds in a protein with no S-S bridges is simple, and all the torsion-angles about the single bonds are independent. When torsion-angles in closed rings are fixed, they cannot change bond lengths and bond-angles. Thus, the energy function in torsion-angle space only depends explicitly on torsion-angles, \( \Theta \) (for bond twisting), and Cartesian coordinates, \( q \) (for non-bonded interactions).
\[ E = F(\Theta) + G(q) \]

\[
\frac{\partial E}{\partial \Theta_i} = \frac{\partial F(\Theta)}{\partial \Theta_i} + \sum_k \frac{\partial G(q)}{\partial q_k} \frac{\partial q_k}{\partial \Theta_i}
\]

\[
\frac{\partial^2 E}{\partial \Theta_i \partial \Theta_j} = \frac{\partial^2 F}{\partial \Theta_i \partial \Theta_j} + \sum_k \frac{\partial G(q)}{\partial q_k} \frac{\partial^2 q_k}{\partial \Theta_i \partial \Theta_j} + \sum_{k, l} \frac{\partial^2 G(q)}{\partial q_k \partial q_l} \frac{\partial q_k}{\partial \Theta_i} \frac{\partial q_l}{\partial \Theta_j}
\]

Explicit differentiation is trivial, and formulae are needed for the first and second derivatives of a Cartesian coordinate with respect to torsion-angles.

Changing a torsion-angle about the unit vector \( \mathbf{b} \) by an angle \( \Theta \) affects the position vector, \( \mathbf{r}_k \), of an atom \( k \) further along the chain. The position vector \( \mathbf{r}_k \) has Cartesian components, and the first and second derivatives of \( \mathbf{r}_k \) in torsion-angle space are needed to transform the derivatives of the non-bonded part of the energy function from Cartesian space to torsion-angle space. Figure 2.8 shows how the movement of atom \( k \) can be expressed in terms of three orthogonal vectors \( V_1, V_2, \) and \( V_3 \).
\[ E = F(\Theta) + G(q) \]

\[ \frac{\partial E}{\partial \Theta_i} = \frac{\partial F(\Theta)}{\partial \Theta_i} + \sum_k \frac{\partial G(q_i)}{\partial q_k} \frac{\partial q_k}{\partial \Theta_i} \]

\[ \frac{\partial^2 E}{\partial \Theta_i \partial \Theta_j} = \frac{\partial^2 F}{\partial \Theta_i \partial \Theta_j} + \sum_{k} \frac{\partial G(q_i)}{\partial q_k} \frac{\partial^2 q_k}{\partial \Theta_i \partial \Theta_j} + \sum_{k,l} \frac{\partial^2 G(q_i)}{\partial q_k \partial q_l} \frac{\partial q_k}{\partial \Theta_i} \frac{\partial q_l}{\partial \Theta_j} \]

Explicit differentiation is trivial, and formulae are needed for the first and second derivatives of a Cartesian coordinate with respect to torsion-angles.

Changing a torsion-angle about the unit vector \( b \) by an angle \( \Theta \) affects the position vector, \( r_k \), of an atom \( k \) further along the chain. The position vector \( r_k \) has Cartesian components, and the first and second derivatives of \( r_k \) in torsion-angle space are needed to transform the derivatives of the non-bonded part of the energy function from Cartesian space to torsion-angle space. Figure 2.8 shows how the movement of atom \( k \) can be expressed in terms of three orthogonal vectors \( V_1, V_2, \) and \( V_3 \).
Figure 2.8. The change in position is expressed by the three components $\Delta x$, $\Delta y$, and $\Delta z$. How rotation occurs about the unit vector $\mathbf{b}$ moves around $\mathbf{b}$.
and \( Z_1 \) and \( Z_2 \) are orthogonal to both \( \Delta \) and \( \Delta \), and \( \Delta \) is antiparallel to \( \Delta \).

The change in position \( \mathbf{r}' \) is expressed by the change in position \( \mathbf{r} \). Figure 2.6. How rotation \( \Theta \) affects the unit vector \( \mathbf{b} \) moves atom \( \mathbf{k} \) to
\[ \mathbf{V}_1 = \mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_i) \]

\[ \mathbf{V}_2 = -\mathbf{b}_i \times \mathbf{V}_1 = -\mathbf{b}_i \times [\mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_i)] \]

\[ \mathbf{V}_3 = \mathbf{b}_i \left[ \mathbf{b}_i \cdot (\mathbf{r}_k - \mathbf{r}_i) \right] \]

The unrotated position vector of atom \( k \) is

\[ \mathbf{r}_k = \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{r}_i \]

and the rotated position vector is

\[ \mathbf{r}_k' = \mathbf{V}_1 \sin \theta_i + \mathbf{V}_2 \cos \theta_i + \mathbf{V}_3 + \mathbf{r}_i \]

This transformation can be written in matrix notation as

\[ \mathbf{r}_k' = T_{\theta_i} (\mathbf{r}_k - \mathbf{r}_i) + \mathbf{r}_i \]
\[ \vec{V}_1 = \vec{b}_i \times (\vec{r}_k - \vec{r}_i) \]

\[ \vec{V}_2 = -\vec{b}_i \times \vec{V}_1 = -\vec{b}_i \times [ \vec{b}_i \times (\vec{r}_k - \vec{r}_i) ] \]

\[ \vec{V}_3 = \vec{b}_i [ \vec{b}_i \cdot (\vec{r}_k - \vec{r}_i) ] \]

The unrotated position vector of atom k is

\[ \vec{r}_k = \vec{V}_2 + \vec{V}_3 + \vec{r}_i \]

and the rotated position vector is

\[ \vec{r}_k' = \vec{V}_1 \sin \Theta_i + \vec{V}_2 \cos \Theta_i + \vec{V}_3 + \vec{r}_i \]

This transformation can be written in matrix notation as

\[ \vec{r}_k' = \mathbf{T}_{\Theta_i} (\vec{r}_k - \vec{r}_i) + \vec{r}_i \]
where $\mathbf{T}_{\theta_i}$ is the appropriate rotation matrix, which is used to change the conformation as the torsion-angle changes.

As $\theta_i$ is the change in torsion-angle, the first and second derivates of $\mathbf{r}'_k$ at $\theta_i = 0$, are the coefficients of $\theta_i$ and $\theta_i^2$ in the expansion of $\mathbf{r}'_k$ about $\theta_i = 0$. At $\theta_i = 0$, the derivatives of $\mathbf{r}'_k$ become those needed to transform Cartesian derivatives to torsion-angle space, for $\mathbf{r}'_k$ then equals $\mathbf{r}_k$.

$$\mathbf{r}'_k = \mathbf{V}_1 \sin \theta_i + \mathbf{V}_2 \cos \theta_i + \mathbf{V}_3 + \mathbf{r}_i$$

$$= \mathbf{V}_1 \left( \theta_i - \frac{\theta_i^3}{3!} + \ldots \right) + \mathbf{V}_2 \left( 1 - \frac{\theta_i^2}{2!} + \ldots \right) + \mathbf{V}_3 + \mathbf{r}_i$$

Therefore

$$\left( \frac{\partial \mathbf{r}'_k}{\partial \theta_i} \right)_{\theta_i = 0} = \mathbf{V}_1 = \mathbf{b}_i \times (\mathbf{r}'_k - \mathbf{r}_i)$$

and

$$\left( \frac{\partial^2 \mathbf{r}'_k}{\partial \theta_i^2} \right)_{\theta_i = 0} = -\mathbf{V}_2 = -\mathbf{b}_i \times \left[ \mathbf{b}_i \times (\mathbf{r}'_k - \mathbf{r}_i) \right]$$

Partial derivatives of $\mathbf{r}_k$ with respect to two different torsion-angles, $\theta_i$ and $\theta_j$, are obtained as follows. For small changes of torsion-angle $\theta_i$ only, the new position vector $\mathbf{r}'_k$ is given by
where \( T_{\theta_i} \) is the appropriate rotation matrix, which is used to change the conformation as the torsion-angle changes.

As \( \theta_i \) is the change in torsion-angle, the first and second derivatives of \( r'_{k,i} \) at \( \theta_i = 0 \), are the coefficients of \( \theta_i \) and \( \theta_i^2 \) in the expansion of \( r'_{k,i} \) about \( \theta_i = 0 \). At \( \theta_i = 0 \), the derivatives of \( r'_{k,i} \) become those needed to transform Cartesian derivatives to torsion-angle space, for \( r'_{k,i} \) then equals \( r_{k,i} \).

\[
\frac{\overline{r'}_{k,i}}{\overline{r}_{k,i}} = \overline{v}_1 \sin \theta_i + \overline{v}_2 \cos \theta_i + \overline{v}_3 + \overline{r}_i
\]

\[
= \overline{v}_1 (\theta_i - \frac{\theta_i^3}{3!} + \cdots) + \overline{v}_2 (1 - \frac{\theta_i^2}{2!} + \cdots) + \overline{v}_3 + \overline{r}_i
\]

Therefore

\[
\left( \frac{d\overline{r'}_{k,i}}{d\theta_i} \right)_{\theta_i=0} = \frac{d\overline{r}_{k,i}}{d\theta_i} = \overline{v}_1 = \overline{b}_i \times (\overline{r}_{k,i} - \overline{r}_i)
\]

and

\[
\left( \frac{d^2\overline{r'}_{k,i}}{d\theta_i^2} \right)_{\theta_i=0} = \frac{d^2\overline{r}_{k,i}}{d\theta_i^2} = -\overline{v}_2 = \overline{b}_i \times [\overline{b}_i \times (\overline{r}_{k,i} - \overline{r}_i)]
\]

Partial derivatives of \( r_{k,i} \) with respect to two different torsion-angles, \( \theta_i \) and \( \theta_j \), are obtained as follows. For small changes of torsion-angle \( \theta_i \) only, the new position vector, \( \overline{r}_{k,i} \), is given by
only, the new position vector \( \mathbf{r}_k \) is given by
Figure 2.9. Showing how a pair of small rotations $\Theta_i$ and $\Theta_j$ about unit vectors $\vec{b}_i$ and $\vec{b}_j$, respectively, affect the position of atom $k$. The vectors $\vec{b}_i$ and $\vec{b}_j$ point along the chain towards the free end.
Figure 2.9. Showing how a pair of small rotations $\Theta_i$ and $\Theta_j$ about unit vectors $\vec{b}_i$ and $\vec{b}_j$, respectively, affect the position of atom $k$. The vectors $\vec{b}_i$ and $\vec{b}_j$ point along the chain towards the free end.
\[ \tilde{r}_k' = \tilde{V}_1 \Theta_i + \tilde{V}_a + \tilde{V}_3 + \tilde{r}_i \] (as \( \Theta_i \) is small)

\[ = \Theta_i \tilde{V}_1 - \overline{b}_i \times \left[ \overline{b}_i \times (\tilde{r}_k - \tilde{r}_i) \right] - \overline{b}_i \cdot \left[ \overline{b}_i \cdot (\tilde{r}_k - \tilde{r}_i) \right] + \tilde{r}_i \]

For any three vectors \( \vec{a}, \vec{b}, \vec{c} \)

\[ \vec{a} \times (\vec{b} \times \vec{c}) = \vec{b} (\vec{a} \cdot \vec{c}) - \vec{c} (\vec{a} \cdot \vec{b}) \]

and

\[ \tilde{r}_k' = \Theta_i \tilde{V}_1 + (\tilde{r}_k - \tilde{r}_i) (\overline{b}_i \cdot \overline{b}_i) + \tilde{r}_i \]

\[ = \Theta_i \tilde{V}_1 + \tilde{r}_k \quad \text{as} \quad \overline{b}_i \cdot \overline{b}_i = 1 \]

\[ = \Theta_i \overline{b}_i \times (\tilde{r}_k - \tilde{r}_i) + \tilde{r}_k \]

Figure 2.9 shows how the position of atom \( k \) depends on two small rotations \( \Theta_i \) about \( \overline{b}_i \), and \( \Theta_j \) about \( \overline{b}_j \). Because \( \Theta_i \) affects \( \tilde{r}_j \) and \( \tilde{r}_k \) while \( \Theta_j \) only affects \( \tilde{r}_k \), it is not clear that small rotations about \( \Theta_i \) and \( \Theta_j \) are independent of the order of the rotations.

Firstly, apply a rotation \( \Theta_i \) followed by \( \Theta_j \)

\[ \tilde{r}_j' = \Theta_i \overline{b}_j \times (\tilde{r}_j - \tilde{r}_i) + \tilde{r}_j \]
\[ \mathbf{\bar{r}}_k' = \mathbf{V}_1 \Theta_i + \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{r}_i \quad (\text{as } \Theta_i \text{ is small}) \]

\[ = \Theta_i \mathbf{V}_1 - b_i \times \left[ b_i \times (\mathbf{r}_k - \mathbf{r}_i) \right] - b_i \cdot \left[ b_i \cdot (\mathbf{r}_k - \mathbf{r}_i) \right] + \mathbf{r}_i \]

For any three vectors \( \mathbf{a}, \mathbf{b}, \mathbf{c} \)

\[ \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b} (\mathbf{a} \cdot \mathbf{c}) - \mathbf{c} (\mathbf{a} \cdot \mathbf{b}) \]

and

\[ \mathbf{\bar{r}}_k' = \Theta_i \mathbf{V}_1 + \left( \mathbf{\bar{r}}_k - \mathbf{r}_i \right) \left( b_i \cdot b_i \right) + \mathbf{r}_i \]

\[ = \Theta_i \mathbf{V}_1 + \mathbf{\bar{r}}_k \quad \text{as } b_i \cdot b_i = 1 \]

\[ = \Theta_i b_i \times \left( \mathbf{\bar{r}}_k - \mathbf{r}_i \right) + \mathbf{r}_k \]

Figure 2.9 shows how the position of atom \( k \) depends on two small rotations \( \Theta_i \) about \( \mathbf{\bar{b}}_i \) and \( \Theta_j \) about \( \mathbf{\bar{b}}_j \). Because \( \Theta_i \) affects \( \mathbf{\bar{r}}_j \) and \( \mathbf{\bar{r}}_k \) while \( \Theta_j \) only affects \( \mathbf{\bar{r}}_k \), it is not clear that small rotations about \( \Theta_i \) and \( \Theta_j \) are independent of the order of the rotations.

Firstly, apply a rotation \( \Theta_i \) followed by \( \Theta_j \)

\[ \mathbf{\bar{r}}_j' = \Theta_i \mathbf{b}_j \times \left( \mathbf{\bar{r}}_j - \mathbf{r}_i \right) + \mathbf{r}_j \]
separations greater than six Angstroms. The energy of the repulsive term tends to zero as $\frac{1}{r^6}$, whereas that of the attractive term tends to zero as $\frac{1}{r^\delta}$. The electrostatic energy between dipoles varies as $\frac{1}{r^5}$, and is further reduced by an increasing dielectric constant. Solvent molecules in shells outside the first have little effect on the solvation energy of a substance (Gibson and Scheraga, 1967). In any case, if long-range forces are not neglected when calculating the energy of a large molecule, much more computing time is needed. Because minimization of the energy requires continuous first and second derivatives, all non-bonded potentials must approach zero smoothly at a cut-off separation. Gibson and Scheraga (1967) truncate non-bonded interactions by multiplying all potential functions by the function

$$U(r) = \frac{(1 - \frac{r^3}{r_0^2})^4}{0.0001 \left( \frac{r}{r_0} \right)^8 + \left(1 - \frac{r^2}{r_0^2} \right)^4}$$

for $0 \leq r < r_0$

$$= 0 \quad \text{for} \quad r \geq r_0 \quad (2c)$$

This procedure increases the non-bonded calculation time as $U(r)$ must be evaluated for all separations, $r$, up to the cut-off, $r_0$. I use a polynomial function, $S(t)$, which has the following properties: $S(t) = 1$, $t \leq -1$; $S(t) = 0$, $t \geq 1$; and the first three derivatives are also zero at both these points. The variable $t$ is defined as $\frac{r - r_0 + \delta}{\delta}$. Only those non-bonded interactions at separations between $r_0 - 2\delta$ and $r_0$ are multiplied by the truncating function; at larger separations no interaction is calculated, and at smaller separations a normal interaction is calculated. $\delta$ can be
separations greater than six Angstroms. The energy of the repulsive term tends to zero as \( \frac{1}{r^2} \), whereas that of the attractive term tends to zero as \( \frac{1}{r^6} \). The electrostatic energy between dipoles varies as \( \frac{1}{r^3} \), and is further reduced by an increasing dielectric constant. Solvent molecules in shells outside the first have little effect on the solvation energy of a substance (Gibson and Scheraga, 1967). In any case, if long-range interactions are not neglected when calculating the energy of a large molecule, much more computing time is needed. Because minimization of the energy requires continuous first and second derivatives, all non-bonded potentials must approach zero smoothly at a cut-off separation. Gibson and Scheraga (1967) truncate non-bonded interactions by multiplying all potential functions by the function

\[
U(r) = \frac{(1 - \frac{r^2}{r_0^2})^4}{0.0001 \left( \frac{r}{r_0} \right)^8 + \left( 1 - \frac{r^2}{r_0^2} \right)^4} \quad \text{for} \ 0 \leq r < r_0
\]

\[
= 0 \quad \text{for} \ r \geq r_0 \quad (2c)
\]

This procedure increases the non-bonded calculation time as \( U(r) \) must be evaluated for all separations, \( r \), up to the cut-off, \( r_0 \). I use a polynomial function, \( S(t) \), which has the following properties: \( S(t) = 1, t < -1; S(t) = 0, t > 1; and \) the first three derivatives are also zero at both these points. The variable \( t \) is defined as \( \frac{r - r_0 + \delta}{\delta} \). Only those non-bonded interactions at separations between \( r_0 - 2\delta \) and \( r_0 \) are multiplied by the truncating function; at larger separations no interaction is calculated.
and at smaller separations a normal interaction is calculated. It can be
Figure 2.4. Graphs of the approximate step-function given by

\[ F(x) = 0.5 + 0.0026 x - 0.8203 x^3 + 0.3828 x^5 - 0.0651 x^7 \]

The solid curve is the function value, \( F \); the dashed curve is the first derivative, \( DF \); and the long-dashed curve is the second derivative, \( DDF \), divided by 2. The three curves are 0.0 at \( x = 1.0 \), the two derivative curves are 0.0 at \( x = -1.0 \).
Figure 2.4. Graphs of the approximate step-function given by

\[ F(x) = 0.5 + 0.026 x - 0.8203 x^3 + 0.3828 x^5 - 0.0651 x^7 \]

The solid curve is the function value, \( F \); the dashed curve is the first derivative, \( DF \); and the long-dashed curve is the second derivative, \( DDF \), divided by 2. The three curves are 0.0 at \( x = 1.0 \), the two derivative curves are 0.0 at \( x = -1.0 \).
made small (say $r_0/20$) so that only a few interactions need be truncated.

The polynomial step-function, $S(t)$, is shown in Figure 2.4.

Truncation can cause large spurious forces at the cut-off separation if the non-bonded energy at this point is very different from zero. For example, the slope of the total energy in Fig. 2.1c is large near the cut-off because the electrostatic energy is appreciable. The absolute level of the potential energy is usually not important as the equilibrium conformation is defined by a balance of forces not energies. These spurious forces are greatly reduced if the level of the non-bonded energy is altered to make the energy close to zero near the cut-off. For example, when the step function, $S(t)$, is used for truncation, the non-bonded potential is altered as follows:

$$G'(r) = G(r) - G(r_0 - \delta)$$

With the Gibson and Scheraga truncation function, the value of the non-bonded function at $r'$ must be subtracted from the normal non-bonded function before truncation, where $r'$ is the separation at which the slope of $U(r)$ is maximum. This correction, which has not been used by Gibson and Scheraga, is needed only if electrostatic terms contribute to the energy; other energy contributions fall off too rapidly to generate appreciable forces on truncation.

2.3 Specifying the empirical energy function of any molecule

2.3.1 Introduction

The energy function of any molecule depends on the topological relationship of its atoms. As the short-range contributions to the energy
made small (say \( r_0/20 \)) so that only a few interactions need be truncated. The polynomial step-function, \( S(t) \), is shown in Figure 2.4.

Truncation can cause large spurious forces at the cut-off separation if the non-bonded energy at this point is very different from zero. For example, the slope of the total energy in Fig. 2.1c is large near the cut-off because the electrostatic energy is appreciable. The absolute level of the potential energy is usually not important as the equilibrium conformation is defined by a balance of forces not energies. These spurious forces are greatly reduced if the level of the non-bonded energy is altered to make the energy close to zero near the cut-off. For example, when the step function, \( S(t) \), is used for truncation, the non-bonded potential is altered as follows:

\[
G'(r) = G(r) - G(r_0 - \delta)
\]

With the Gibson and Scheraga truncation function, the value of the non-bonded function at \( r' \) must be subtracted from the normal non-bonded function before truncation, where \( r' \) is the separation at which the slope of \( U(r) \) is maximum. This correction, which has not been used by Gibson and Scheraga, is needed only if electrostatic terms contribute to the energy; other energy contributions fall off too rapidly to generate appreciable forces on truncation.

2.3 Specifying the empirical energy function of any molecule

2.3.1 introduction
The energy function of any molecule depends on the topological relationship of its atoms. As the short-range contributions to the energy
Figure 2.5. Deriving the symbolic chemical formula of cubane. Five bonds are broken at atoms and the end of each break marked by \(+\), \(\circ\), \(=\), \(*\), \(+\); some atoms are marked by more than one symbol. The molecule is now simply connected, and the atoms, preceded by the broken-bond symbols, are written from left to right starting at the lower right-hand corner of the figure.

**TABLE 2.1**

Definition of atom symbols when hydrogens are excluded

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Atom and bonding environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>Carboxyl oxygen</td>
</tr>
<tr>
<td>C</td>
<td>Carboxyl oxygen</td>
</tr>
<tr>
<td>V</td>
<td>Hydroxyl group, water</td>
</tr>
<tr>
<td>N</td>
<td>Amide nitrogen</td>
</tr>
<tr>
<td>M</td>
<td>All other nitrogens</td>
</tr>
<tr>
<td>G</td>
<td>Tetravalent carbon with associated hydrogens</td>
</tr>
<tr>
<td>A</td>
<td>Trivalent carbon in 6 membered ring</td>
</tr>
<tr>
<td>B</td>
<td>Trivalent carbon in 5 membered ring</td>
</tr>
<tr>
<td>G</td>
<td>Trivalent carbon in 6 and 5 membered ring</td>
</tr>
<tr>
<td>E</td>
<td>Carboxyl or carboxyl trivalent carbon</td>
</tr>
<tr>
<td>S</td>
<td>Sulphur</td>
</tr>
</tbody>
</table>

\[ \text{CH} = \text{CH}\$ + \text{CH} + \text{CH}\$ = \text{CH} * + \text{CH} \]
Figure 2.5. Deriving the symbolic chemical formula of cubane. Five bonds are broken at atoms and the end of each break marked by $+$, $-$, $=$, $\ast$, $\ast+$; some atoms are marked by more than one symbol. The molecule is now simply connected, and the atoms, preceded by the broken-bond symbols, are written from left to right starting at the lower right-hand corner of the figure.

**TABLE 2.1**

Definition of atom symbols when hydrogens are excluded

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Atom and bonding environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>Carboxyl oxygen</td>
</tr>
<tr>
<td>G</td>
<td>Carboxyl oxygen</td>
</tr>
<tr>
<td>V</td>
<td>Hydroxyl group, water</td>
</tr>
<tr>
<td>N</td>
<td>Amide nitrogen</td>
</tr>
<tr>
<td>M</td>
<td>All other nitrogens</td>
</tr>
<tr>
<td>C</td>
<td>Tetravalent carbon with associated hydrogens</td>
</tr>
<tr>
<td>A</td>
<td>Trivalent carbon in 6 membered ring</td>
</tr>
<tr>
<td>B</td>
<td>Trivalent carbon in 5 membered ring</td>
</tr>
<tr>
<td>G</td>
<td>Trivalent carbon in 6 and 5 membered ring</td>
</tr>
<tr>
<td>E</td>
<td>Carbonyl or carboxyl trivalent carbon</td>
</tr>
</tbody>
</table>
are summations over all bonds, all bond-angles, and all torsion-angles, it is computationally convenient to supply the indices (positions in a list) of those atoms that form each bond, bond-angle, and torsion-angle. The lists of bond indices and atom types, which uniquely specify the topology of any molecule, can be used to generate the indices of bond-angles and torsion-angles. Bond-angle indices come from pairs of bond indices that have an atom in common; torsion-angle indices come from pairs of bond indices that have an atom in common with each of the atoms of a third bond. If the chain branches, the first and fourth atoms of a torsion-angle can be chosen in several ways. The present work follows the convention proposed by the IUPAC-IUB Sub-commission on Biological Nomenclature (Kendrew et al., 1970). The list of bond indices and atom types can be prepared by hand. This would have to be done once for each residue of a repeating polymer, as the full information could be generated by linking the residue information in the appropriate order. Nevertheless, as the aim of this work is a completely general formulation of the empirical energy function, an easier method was developed.

2.3.2 the notation

Common chemical notation, with additional symbols, is used to describe the atomic connectivity uniquely. Each atom is represented by a single letter; different letters are used to distinguish similar types of atoms in different bonding environments (Table 2.1). Some of the letters used in the notation are defined as side-atoms (atoms bonded to one atom only), and others as chain-atoms (atoms bonded to more than one atom). Certain atoms, like hydrogen, are always side-atoms; others, like oxygen, are sometimes side-atoms (in a peptide group), and sometimes chain-atoms (in a -OH group). A period converts the atom symbol that it precedes into a chain-atom.
are summations over all bonds, all bond-angles, and all torsion-angles, it is computationally convenient to supply the indices (positions in a list) of those atoms that form each bond, bond-angle, and torsion-angle. The lists of bond indices and atom types, which uniquely specify the topology of any molecule, can be used to generate the indices of bond-angles and torsion-angles. Bond-angle indices come from pairs of bond indices that have an atom in common; torsion-angle indices come from pairs of bond indices that have an atom in common with each of the atoms of a third bond. If the chain branches, the first and fourth atoms of a torsion-angle can be chosen in several ways. The present work follows the convention proposed by the IUPAC-IUB Sub-commission on Biological Nomenclature (Kenirew et al., 1970). The list of bond indices and atom types can be prepared by hand. This would have to be done once for each residue of a repeating polymer, as the full information could be generated by linking the residue information in the appropriate order. Nevertheless, as the aim of this work is a completely general formulation of the empirical energy function, an easier method was developed.

2.3.2 the notation

Common chemical notation, with additional symbols, is used to describe the atomic connectivity uniquely. Each atom is represented by a single letter; different letters are used to distinguish similar types of atoms in different bonding environments (Table 2.1). Some of the letters used in the notation are defined as side-atoms (atoms bonded to one atom only), and others as chain-atoms (atoms bonded to more than one atom). Certain atoms, like hydrogen, are always side-atoms; others, like oxygen, are sometimes side-atoms (in a peptide group), and sometimes chain-atoms (in a -CH group). A period
converts the atom symbol that it precedes into a chain-atom.
* A + BB(AB)B(C)B(+N456F)A + BB(AB)B(C)B(+4N)A + BB(C)B

(CCAOV)B(+5N)A + BB(CCAOV)B(C) + B(6+N)

Figure 2.6. Deriving the symbolic chemical formula of the haem group without hydrogen atoms. The atom symbols come from Table 2.1. After marking bonds that are broken, the formula is written down, starting at the ringed atom. Many different combinations of bonds can be broken as long as the resulting network is connected and has no loops.
Figure 2.6. Deriving the symbolic chemical formula of the haem group without hydrogen atoms. The atom symbols come from Table 2.1. After marking bonds that are broken, the formula is written down starting at the ringed atom. Many different combinations of bonds can be broken as long as the resulting network is connected and has no loops.
Any molecular system is symbolized by first breaking all the closed rings of atoms to give one or more simply connected networks. The broken bonds are marked by preceding each atom of the separated pair by the same undefined symbol; any symbol different from those indicating atoms or those with a special meaning (left and right parentheses and period) can be used. One chain, preferably the longest, of each of the simply connected networks now representing the molecular system, is defined as the main-chain. The atom symbols of this chain are written consecutively from left to right; each chain-atom is followed by the side-atoms connected to it. Several side-atoms of the same type that are bonded to one chain-atom are symbolized by a single letter followed by the number of side-atoms of that type; for example, a methyl group is written CH₃. All the symbols of side-chains are enclosed by parentheses, and placed immediately after any side-atoms attached to the particular branch-point atom. Once both atoms separated by breaking a closed ring have been written down, the closure symbol can be used again for a different break. The symbolic formula of each complete molecule is enclosed in parentheses; formulae of different molecules of the system can be in any order. A molecular system can be defined uniquely by many symbolic formulae, which differ only in the order of the atoms. This order is critical when operating on a set of atomic coordinates, as the order of atoms and coordinates must agree.

Figures 2.5 and 2.6 show the derivation of the symbolic chemical formula for two molecules. Molecules having repeated chemical units in a linear chain may be defined more compactly. Thus, a protein is represented by a sequence of the common three letter abbreviations for the different amino acids. A library containing the symbolic formula of each residue is used to prepare the full chemical formula (Fig. 2.7).
Any molecular system is symbolized by first breaking all the closed rings of atoms to give one or more simply connected networks. The broken bonds are marked by preceding each atom of the separated pair by the same undefined symbol; any symbol different from those indicating atoms or those with a special meaning (left and right parentheses and period) can be used. One chain, preferably the longest, of each of the simply connected networks now representing the molecular system, is defined as the main-chain. The atom symbols of this chain are written consecutively from left to right; each chain-atom is followed by the side-atoms connected to it. Several side-atoms of the same type that are bonded to one chain-atom are symbolized by a single letter followed by the number of side-atoms of that type; for example, a methyl group is written CH3. All the symbols of side-chains are enclosed by parentheses, and placed immediately after any side-atoms attached to the particular branch-point atom. Once both atoms separated by breaking a closed ring have been written down, the closure symbol can be used again for a different break. The symbolic formula of each complete molecule is enclosed in parentheses; formulae of different molecules of the system can be in any order. A molecular system can be defined uniquely by many symbolic formulae, which differ only in the order of the atoms. This order is critical when operating on a set of atomic coordinates, as the order of atoms and coordinates must agree.

Figures 2.5 and 2.6 show the derivation of the symbolic chemical formula for two molecules. Molecules having repeated chemical units in a linear chain may be defined more compactly. Thus, a protein is represented by a sequence of the common three letter abbreviations for the different amino acids. A library containing the symbolic formula of each residue is used to prepare the full chemical formula (Fig. 2.7).
\[ \mathbf{r}_k' = \Theta_i \mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_i) + \mathbf{r}_k \]

\[ \mathbf{b}_j' = \Theta_i \mathbf{b}_i \times \mathbf{b}_j + \mathbf{b}_j \]

as the unit vector \( \mathbf{b}_j \) is also moved by rotation \( \Theta_i \).

\[ \mathbf{r}_k'' = \Theta_j \mathbf{b}_j' \times (\mathbf{r}_k' - \mathbf{r}_j) + \mathbf{r}_k' \]

\[ = \Theta_j (\Theta_i \mathbf{b}_i \times \mathbf{b}_j + \mathbf{b}_j) \times (\Theta_i \mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_j) + (\mathbf{r}_k - \mathbf{r}_j)) \]

\[ + \Theta_i \mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_i) + \mathbf{r}_k \]

hence

\[ \frac{\partial^2 \mathbf{r}_k}{\partial \Theta_j \partial \Theta_i} = \text{coeff. of } \Theta_i \Theta_j = \mathbf{b}_j \times [\mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_j) + (\mathbf{b}_i \times \mathbf{b}_j) \times (\mathbf{r}_k - \mathbf{r}_j)] \]

Secondly, apply a small rotation \( \Theta_j \) followed by \( \Theta_i \).

\[ \mathbf{r}_k' = \Theta_j \mathbf{b}_j \times (\mathbf{r}_k - \mathbf{r}_j) + \mathbf{r}_k \]

\[ \mathbf{r}_k'' = \Theta_i \mathbf{b}_i \times (\mathbf{r}_k' - \mathbf{r}_i) + \mathbf{r}_k' \]

\[ = \Theta_i \mathbf{b}_i \times [\Theta_j \mathbf{b}_j \times (\mathbf{r}_k - \mathbf{r}_j) + \mathbf{r}_k - \mathbf{r}_i] + \Theta_i \mathbf{b}_i \times (\mathbf{r}_k - \mathbf{r}_j) + \mathbf{r}_i \]
\[ \bar{r}_k' = \Theta_i \bar{b}_i \times (\bar{r}_k - \bar{r}_i) + \bar{r}_k \]

\[ \bar{b}_j' = \Theta_i \bar{b}_i \times \bar{b}_j + \bar{b}_j \]

as the unit vector \( \bar{b}_j \) is also moved by rotation \( \Theta_i \)

\[ \bar{r}_k'' = \Theta_j \bar{b}_j' \times (\bar{r}_k' - \bar{r}_j) + \bar{r}_k' \]

\[ = \Theta_j (\Theta_i \bar{b}_i \times \bar{b}_j + \bar{b}_j) \times [\Theta_i \bar{b}_i \times (\bar{r}_k - \bar{r}_j) + (\bar{r}_k - \bar{r}_j)] \]

\[ + \Theta_i \bar{b}_i \times (\bar{r}_k - \bar{r}_j) + \bar{r}_k \]

hence

\[ \frac{\partial^2 \bar{r}_k}{\partial \Theta_j \partial \Theta_i} = \text{coeff. of } \Theta_i \Theta_j = \bar{b}_j \times [\bar{b}_i \times (\bar{r}_k - \bar{r}_j) + (\bar{b}_i \times \bar{b}_j) \times (\bar{r}_k - \bar{r}_j)] \]

Secondly, apply a small rotation \( \Theta_j \) followed by \( \Theta_i \)

\[ \bar{r}_k' = \Theta_j \bar{b}_j \times (\bar{r}_k - \bar{r}_j) + \bar{r}_k \]

\[ \bar{r}_k'' = \Theta_i \bar{b}_i \times (\bar{r}_k' - \bar{r}_i) + \bar{r}_k' \]
\[
\mathcal{U}_i \ b_i \times \left[ \mathcal{U}_j \ b_j \times (\mathbf{r}_k - \mathbf{r}_j) + \mathbf{r}_k - \mathbf{r}_i \right] + \mathcal{U}_j \ b_j \times (\mathbf{r}_k - \mathbf{r}_j) + \mathbf{r}_j
\]
Therefore \[
\frac{\partial^2 \overline{r}_k}{\partial \Theta_i \partial \Theta_j} = b_i \times \left[ b_j \times (\overline{r}_K - \overline{r}_j) \right]
\]
and \[
\frac{\partial^2 \overline{r}_K}{\partial \Theta_i \partial \Theta_j} = \frac{\partial^2 \overline{r}_K}{\partial \Theta_j \partial \Theta_i}
\]
under the vector identity
\[
\overline{a} \times (\overline{b} \times \overline{c}) + \overline{b} \times (\overline{c} \times \overline{a}) + \overline{c} \times (\overline{a} \times \overline{b}) = 0
\]

2.5 Generating initial coordinates

2.5.1 Introduction

Cartesian coordinates can be computed from the internal coordinates of a molecule. As each atomic position depends on all the internal coordinates in the chain leading to that atom, small errors in internal coordinates result in unacceptable errors in Cartesian coordinates. Unless the growing chain is fitted to occasional Cartesian guide-points, large structures cannot be built from internal coordinates. Diamond (1966) has developed a procedure that builds a protein chain from images of amino acids as determined in small structures, and then fits the chain to guide points by changing torsion-angles of single bonds. I have developed a general model-building procedure that
Therefore \[ \frac{\partial^2 \mathbf{r}_k}{\partial \theta_i \partial \theta_j} = b_i \times \left[ b_j \times (\mathbf{r}_k - \mathbf{r}_j) \right] \]

and \[ \frac{\partial^2 \mathbf{r}_k}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 \mathbf{r}_k}{\partial \theta_j \partial \theta_i} \]

under the vector identity

\[ \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b}) = 0 \]

2.5 Generating initial coordinates

2.5.1 Introduction

Cartesian coordinates can be computed from the internal coordinates of a molecule. As each atomic position depends on all the internal coordinates in the chain leading to that atom, small errors in internal coordinates result in unacceptable errors in Cartesian coordinates. Unless the growing chain is fitted to occasional Cartesian guide-points, large structures cannot be built from internal coordinates. Diamond (1966) has developed a procedure that builds a protein chain from images of amino acids as determined in small structures, and then fits the chain to guide points by changing torsion-angles of single bonds. I have developed a general model-building procedure that
References for Chapter 2


References for Chapter 2


Chapter 3

ENERGY MINIMIZATION: TESTS ON A TRIPPEPTIDE

3.1 Introduction

Protein conformations are stable for there is no force on any atom, and the energy is at a minimum value. When protein atoms are disturbed in any way, they will move to eliminate any forces and minimize the energy. The dynamic behaviour of a protein molecule is copied by minimizing an energy function. Thus, efficient well characterized minimization procedures are needed for the analysis of conformations of proteins and indeed all other molecules.

All energy minimization methods are based on rather simple concepts. The empirical energy function is a complex function that is defined in conformation space and depends on the variables that describe the conformation, (i.e. the atomic Cartesian coordinates or the single-bond torsion-angles of the molecule). It is infinitely time-consuming to calculate the energy for all possible spatial arrangements of the atoms to find what point in conformation space has the lowest energy. Because the energy function is complex analytically, there is no simple way to calculate algebraically where the minima are. Thus, limited information must be collected at a small number of points in conformation space (i.e. at different conformations), and this information must be used to guess where in conformation space the
Chapter 3

ENERGY MINIMIZATION: TESTS ON A TRIPEPTIDE

3.1 Introduction

Protein conformations are stable for there is no force on any atom, and the energy is at a minimum value. When protein atoms are disturbed in any way, they will move to eliminate any forces and minimize the energy. The dynamic behaviour of a protein molecule is copied by minimizing an energy function. Thus, efficient well characterized minimization procedures are needed for the analysis of conformations of proteins and indeed all other molecules.

All energy minimization methods are based on rather simple concepts. The empirical energy function is a complex function that is defined in conformation space and depends on the variables that describe the conformation, (i.e. the atomic Cartesian coordinates or the single-bond torsion-angles of the molecule). It is infinitely time-consuming to calculate the energy for all possible spatial arrangements of the atoms to find what point in conformation space has the lowest energy. Because the energy function is complex analytically, there is no simple way to calculate algebraically where the minima are. Thus, limited information must be collected at a small number of points in conformation space (i.e. at different conformations),
and this information must be used to guess where in conformation space the
minimum should be. Although such a guess is unlikely to get to the minimum energy conformation, it may well get to a conformation with lower energy. The guessing is repeated at the new conformation until new guesses do not make any significant change to the conformation. If the guesses are sensible then the final conformation will be a minimum energy conformation, and the energy gradient will be zero.

Guessing where the minimum may be is rationalized as follows:

(a) Assume that a simple analytical function describes how the energy varies as the atoms move and the conformation changes. If this approximation is simple it can only be valid for small changes of conformation. (b) Find the parameters of the simple analytical function. (c) Calculate algebraically the conformation where the analytical approximation is a minimum, and take this conformation as the guessed minimum of the energy function. The analytical function that approximates the energy when the atoms move slightly from their current position is

\[ E'(x) = a + bx + cx^2 + dx^3 + \ldots + \ldots \]

When \( E'(x) \) is the approximate energy at a conformation displaced by \( x \) from the current conformation. In this example \( x \) is one-dimensional; for a general problem \( x \) will be \( n \)-dimensional when the conformation is defined by \( n \) variable coordinates. If more terms are included in the expansion, then the approximate energy function will be reliable for larger changes in conformation, \( x \).

The values of the parameters \( a, b, c, \) and \( d \) etc. that define the energy approximation can be calculated in two different ways: The first method assumes that the approximate energy function is only reliable very close to the current conformation. All the information that is needed to calculate all the parameters \( a, b, c \) etc. is obtained at the current conformation. For example,