Definition of Systematic, Approximately Separable and Modular Internal Coordinates (SASMIC) for macromolecular simulation

Pablo Echenique[1,2] and J. L. Alonso[1,2]

[1] Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, Pedro Cerbuna 12, 50009, Zaragoza, Spain.

[2] Instituto de Biocomputación y Física de los Sistemas Complejos (BIFI), Edificio Cervantes, Corona de Aragón 42, 50009, Zaragoza, Spain.

January 26, 2006

Abstract

A set of rules is defined to systematically number the groups and the atoms of polypeptides in a modular manner. Supported by this numeration, a set of internal coordinates is defined. These coordinates (termed Systematic, Approximately Separable and Modular Internal Coordinates, SASMIC) are straightforwardly written in Z-matrix form and may be directly implemented in typical Quantum Chemistry packages. A number of Perl scripts that automatically generate the Z-matrix files are provided as supplementary material. The main difference with most Z-matrix-like coordinates normally used in the literature is that normal dihedral angles ("principal dihedrals" in this work) are only used to fix the orientation of whole groups and a different type of dihedrals, termed "phase dihedrals", are used to describe the covalent structure inside the groups. This *physical approach* allows to approximately separate soft and hard movements of the molecule using only topological information and to directly implement constraints. As an application, we use the coordinates defined and ab initio quantum mechanical calculations to assess the commonly assumed approximation of the free energy, obtained from "integrating out" the side chain degree of freedom χ, by the Potential Energy Surface (PES) in the protected dipeptide HCO-L-Ala-NH_2. We also present a sub-box of the Hessian matrix in two different sets of coordinates to illustrate the approximate separation of soft and hard movements when the coordinates defined in this work are used.

PACS: 87.14.Ee, 87.15.-v, 87.15.Aa, 87.15.Cc

1 Introduction

The choice of the coordinates used to describe a molecule is an important issue if computational considerations are to be taken into account and the efficiency of the simulations is pursued. This choice also affects the coding of applications. If cumbersomely defined coordinates are used, an unnecessary complexity may be added to the design of Monte Carlo movements, the construction and pruning of a database of structures [1,2] or the programming of molecular visualization and manipulation tools.

Suitable coordinates frequently used to describe arbitrary conformations of molecules are the so-called "internal" or "valence-type" coordinates [3]. Their adequacy stems from a number of characteristics: first, they are closely related to chemically meaningful structural parameters, such as bond lengths or bond angles; second, they are local, in the sense that each one of them involves only a small number of (normally close) atoms in its definition; and finally, there are only $3n-6$ of them (where n is the number of atoms in the molecule), in such a way that the overall rotation and translation have been naturally removed.

There also exists a family of coordinates [4–7], extensively used in the inner calculations of many Quantum Chemistry packages (such as Gaussian [8] or GAMESS [9]) and based on the "natural internal coordinates" originally proposed by Pulay and coworkers [10–12], which are defined through linear combinations of the original internals. These coordinates are specially designed to describe normal-mode vibrations in the immediate neighbourhood of energy minima and represent the best choice for accelerating convergence of geometry optimizations in a particular basin of attraction, via diagonal estimation of the Hessian matrix [7]. Accordingly, they maximally separate hard and soft movements in these conditions. However, if the conformation of the molecule is far from a minimum, this type of coordinates lose great part of their meaning and they introduce many computational difficulties without increasing the efficiency. Also, some of the definitions are *redundant* [6, 10–13], i.e., they use a number of coordinates larger than the number of degrees of freedom. In this work, we will only discuss coordinates, such as internals or Cartesian, that may be conveniently used to specify an *arbitrary* conformation of the system and that can be directly related to simple geometrical variables.

In macromolecules, such as proteins, the number of degrees of freedom is the main limiting factor when one tries to predict their behaviour via computer modeling. Therefore, it is also advisable that the set of coordinates chosen allows for a direct implementation of physically meaningful constraints that reduce the dimensionality of the conformational space considered. Most of the expressions used in Statistical Mechanics or in Molecular Dynamics are best written in Cartesian coordinates, however, the implementation of naturally appearing constraints is far from being straightforward in these coordinates. In internal coordinates, on the contrary, the approximate separation of hard and soft movements of the system allows to easily constrain the molecule [14–16] by setting the hard coordinates (those that require a considerable amount of energy to change noticeably) to constant values or to particular functions of the

soft coordinates. Moreover, in internal coordinates (and appealing to some reasonable approximations), the Statistical Mechanics formulae for the constrained system may be written in convenient closed form [17, 18].

Still, although the bond lengths and bond angles are customarily regarded as hard and their definition is unproblematic, the same is not true for dihedral angles. Some definitions of dihedrals may lead to difficulties or to worse separation of hard and soft modes. Let us exemplify this with a particular case:

Consider the definition of Z-matrix-like [19, 20] internal coordinates for the HCO-L-Ala-NH_2 molecule in fig. 6. Imagine that we "position" (i.e., we write the corresponding Z-matrix row) every atom up to the hydrogen denoted by H_9 and that we are now prepared to position the hydrogens in the side chain (H_{10}, H_{11} and H_{12}) via one bond length, one bond angle and one dihedral for each one of them. We will denote by (i, j) the bond length between atoms i and j; by (i, j, k), the bond angle between the vectors $\vec{r}_{jk}$ and $\vec{r}_{ji}$; and by (i, j, k, l) the dihedral angle between the plane defined by the atoms i, j and k and the one defined by j, k and l.

A choice to position the atoms that is frequently seen in the literature [1, 2, 21–23] is the one shown in table 1.

Atom name	Bond length	Bond angle	Dihedral angle
H_{10}	(10,8)	(10,8,5)	γ_1 :=(10,8,5,3)
H_{11}	(11,8)	(11,8,5)	γ_2 :=(11,8,5,3)
H_{12}	(12,8)	(12,8,5)	γ_3 :=(12,8,5,3)

Table 1: A part of the internal coordinates, in Z-matrix form, of the protected dipeptide HCO-L-Ala-NH_2, as frequently defined in the literature.

If we now perform the *gedanken experiment* that consists of taking a typical conformation of the molecule and slightly moving each internal coordinate at a time while keeping the rest constant, we find that any one of the three dihedrals in the previous definition is a hard coordinate, since moving one of them while keeping the other two constant distorts the internal structure of the methyl group. Hence, in these coordinates, the soft rotameric degree of freedom χ, which we know, for chemical arguments, that must exist[1], is ill-represented. In fact, it must be described as a *concerted* movement of the three dihedrals. In reference [24] this is clearly explained. In references [1, 2], the problem is recognized and the concept of "related dihedrals" is introduced, however, no action is taken to change the definition of the coordinates.

In this work, using the ideas of R. Abagyan and coworkers [14–16], we define a set of rules to uniquely and systematically number the groups, the atoms

[1]According to our calculations, at the RHF/6-31+G(d) level of the theory, the barrier for crossing from one of the three equivalent minima to any of the other two ranges from 3.1 to 6.8 kcal/mol, depending on the values of the Ramachandran angles ϕ and ψ. Compare with the barriers in ϕ or ψ which may be as large as 20 kcal/mol depending on the region of the Ramachandran map explored.

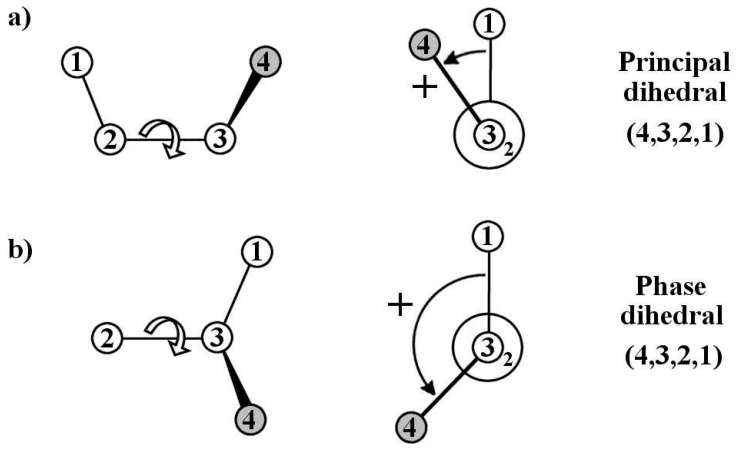

Figure 1: Two types of dihedral angles. **a)** *Principal dihedral*. Used to describe the rotation of whole groups around bonds. **b)** *Phase dihedral*. Used to describe the internal covalent structure of groups. The positive sense of rotation is indicated.

and define the internal coordinates of polypeptides[2] (a general set of rules to perform this numbering and the subsequent definition of internals in any organic molecule is included as supplementary material). The main difference with other Z-matrix-like coordinates normally used in the literature [1, 2, 21–23] is that, instead of positioning each atom with a bond length, a bond angle and a dihedral angle, we use normal dihedral angles (called, from now on, "principal dihedrals") only to fix the orientation of whole groups and a different type of dihedrals, termed "phase dihedrals" by R. Abagyan and coworkers [14–16] (see fig. 1), to describe the covalent structure inside a group[3]. This allows to *approximately* separate soft and hard movements of the molecule using only topological information (i.e., not knowing the exact form of the potential) and to easily implement constraints by forcing the coordinates that correspond to hard movements to take constant values or ones that depend on the soft coordinates. In reference [25] they correctly take this approach into account using out-of-plane angles instead of phase dihedrals, however, they do not describe any rules for a general definition and their numeration of the atoms is non-modular, as it

[2]IUPAC conventions only define a numeration system for the groups, for the branches and for some selected dihedral angles. They focus on functional considerations and not in computational problems. For related documents and references, see `http://www.chem.qmul.ac.uk/iupac/jcbn/`.

[3]Another option may be to use, as a third internal coordinate for each atom, another bond angle. This is rather awkward, however, since two bond angles and a bond length do not specify the position of a point in space. Any values of these three coordinates (except for irrelevant degenerate cases) are compatible with two different symmetrical positions and a fourth number must be provided to break the ambiguity. Also out of plane angles may be used. In reference [24], different options are described.

proceeds first through the backbone (see sec. 2).

In addition, the coordinates herein defined, are straightforwardly cast into Z-matrix form and may be directly implemented in any Quantum Chemistry package, such as Gaussian [8] or GAMESS [9]. This is due to the fact that, although they involve atoms whose covalent structure is different, the mathematical construction of the two types of angles in fig. 1 is exactly the same, and the phase dihedrals are treated like principal ones without any problem by the applications.

A number of Perl scripts are provided that number the atoms and generate the coordinates herein defined for polypeptide chains. The applications read a sequence file in which the different ionization states of the titratable side chains, the tautomeric forms of Histidine and several terminal groups may be specified. Then, an output file is generated with the symbolic definition of the Z-matrix of the molecule which may be directly pasted into the input files of Gaussian [8] or GAMESS [9] (and, upon slight modifications, of any Quantum Chemistry package that is capable of reading Z-matrix format).

Now, if we redo the example in table 1 using phase dihedrals, we must write the rows of the Z-matrix for the hydrogens in the side chain as shown in table 2.

Atom name	Bond length	Bond angle	Dihedral angle
H_{10}	(10,8)	(10,8,5)	χ :=**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	α_1 :=(11,8,5,<u>10</u>)
H_{12}	(12,8)	(12,8,5)	α_2 :=(12,8,5,<u>10</u>)

Table 2: A part of the internal coordinates, in Z-matrix form, of the protected dipeptide HCO-L-Ala-NH_2, as defined by the rules given in sec. 2.

Where the angle **(10,8,5,3)** is now the principal dihedral χ describing the relative rotation of the methyl group around the bond (8,5) and the other two are phase dihedrals that describe the internal structure of the group and that are *pure* hard coordinates (as far as can be told only from topological information). However, one must point out that, although all bond lengths, bond angles and phase dihedrals may be regarded as hard coordinates, not all the principal dihedrals will be soft. Examples of hard principal dihedrals are the ones that describe the rotation around a double bond (or a triple one) or some of the principal dihedrals in cyclic parts of molecules.

The *physical approach* described in this section, which should be taken into account when designing internal coordinates, is embodied in a set of rules for polypeptide chains in sec. 2, a slightly different prescription for general organic molecules is provided as supplementary material. The systematic numeration introduced facilitates the computational treatment of this type of systems and the rules given for polypeptide chains ensure modularity [1,26], i.e., allows to add any residue with minimal modification of the already existing notation and to easily construct databases of structures or of Potential Energy Surfaces (PES).

The characteristics aforementioned have led us to term the coordinates

herein defined *Systematic, Approximately Separable and Modular Internal Coordinates* (SASMIC).

In this work, we will only deal with the numeration of one isolated molecule, however, the procedure described may be easily generalized (and will be in future works) to systems of many molecules (an important example being a macromolecular solute in a bath of solvent molecules). This could be done using *ghost atoms* in a similar manner to what is done in ref. [17], to position the center of mass of the system, and in refs. [14–16], to actually define the coordinates of a system of molecules.

Finally, in sec. 3, we use the new coordinates and ab initio quantum mechanical calculations in order to evaluate the approximation of the free energy, obtained from "integrating out" the rotameric degree of freedom χ, via the typical PES in the protected dipeptide HCO-L-Ala-NH_2. This will be relevant to design effective polypeptide potentials. We also present a small part of the Hessian matrix in two different sets of coordinates to illustrate the approximate separation of soft and hard movements when the SASMIC defined in this work are used. Sec. 4 is devoted to the conclusions.

2 Numeration rules for polypeptides

2.1 Definitions

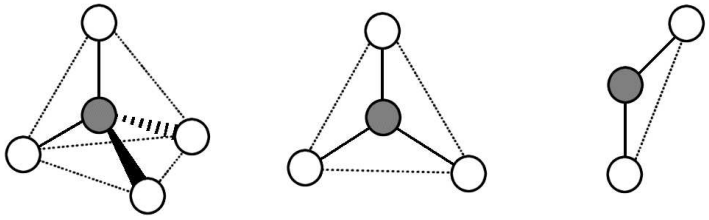

Figure 2: Schematic representation of the groups found in proteins (the angle in the linear group might as well be 180^o). From left to right: tetrahedral, triangular and linear.

First, we realize that any molecule may be formally divided in groups such as those in fig. 2. We will call "centers" the shaded atoms in the figure and "vertices" the white ones. In general, there may exist groups with more than four vertices, however, in proteins, only groups with four or less vertices occur. Examples of tetrahedral groups are the one whose center is the C_α in the backbone or the C_β in the side chain of alanine, triangular groups occur, for example, at the N or the C' in the backbone, finally, linear groups may be found at the O in the side chain of tyrosine or at the S in methionine (see fig. 10).

A particular atom may be vertex of different groups but may only be center of one group. There exist atoms that are only vertices but there do not exist

atoms that are only centers, except in the case of molecules with only one group. In the trivial case of diatomic molecules (in which the only internal coordinate is a bond length), neither of the previous definitions are possible, since we cannot identify a group.

Atoms that are covalently bonded to more than one atom will be called "internal atoms" and are indicated as shaded circles in fig. 3. Atoms that are covalently attached to only one internal atom will be called "external atoms" and are indicated as white-filled circles in fig. 3. In proteins, only H and O may be external atoms.

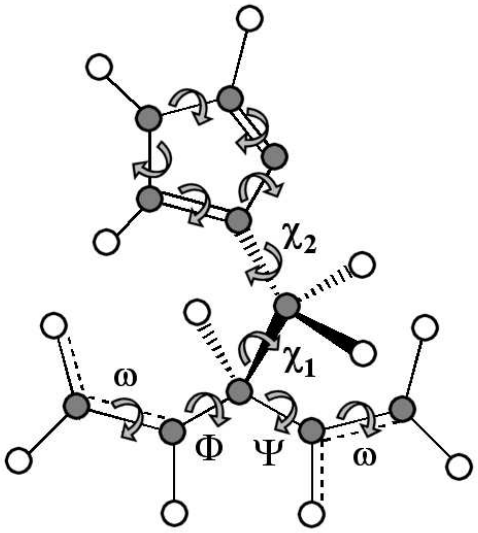

Figure 3: Schematic representation of the HCO-L-His-NH_2 model dipeptide (with the side chain in its uncharged ϵ tautomeric form). Internal atoms are shown as gray-filled circles, external ones as white-filled circles. Internal bonds are indicated with curved arrows. Typical biochemical definitions of some principal dihedrals are also shown.

In most macromolecular models (such as the Born-Oppenheimer approximation used in sec. 3), nuclei are considered point-like particles. Hence, rotation around bonds joining external and internal atoms (termed "external bonds" or "non-dihedral bonds") is neglected, i.e., there are no internal coordinates associated to this movement. On the other hand, rotation around bonds joining two internal atoms (called "internal bonds" or "dihedral bonds" and indicated with curved arrows in fig. 3) is relevant and there may exist internal coordinates describing it.

In order to conform with the *physical approach* stated in the introduction, only one *golden rule* must be followed when defining the internal coordinates:

One principal dihedral, at most[4]*, must be defined on each internal bond.*

The rest of the rules that will be given are mere tidy conventions and systematics.

2.2 Rules for numbering the groups

First of all, we will divide the peptide in groups and number them. To do this we proceed *by branches*, i.e., we choose the next group following a linear sequence of covalently attached groups until there is no possible next one, in which case, we either have finished the numeration process or we start another branch. Every group is numbered once and it cannot be renumbered as the process continues. This numeration is done for completeness and as a support for the numeration of atoms and coordinates. In fig. 4, we have implemented these rules in a protected histidine dipeptide.

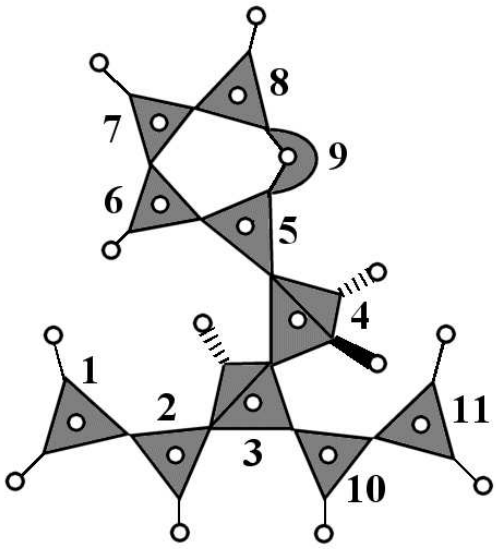

Figure 4: Group identification and numeration in the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form). The different types of groups are shown as gray-filled polyhedra.

The rules are as follows:

i) We select as the *first group* (and number it $j = 1$):

[4]It is not possible to define principal dihedrals for each internal bond for structures containing rings due to the well known limitation of Z-matrix internal coordinates.

- The *amino group* at the N-terminus (either charged or not) if the polypeptide is not N-protected.
- The *formyl group* at the N-terminus if the polypeptide is formyl-N-protected.
- The *methyl group* at the N-terminus if the polypeptide is acetyl-N-protected.

These three cases are the most frequent. If a different species is used to N-protect the polypeptide chain, a convention must be sought that also starts at the N-terminus. This choice takes into account that the primary structure of a polypeptide is normally presented from the N- to the C-terminus.

ii) If there is only one unnumbered group linked to group j, we number it as $j+1$, set $j = j+1$ and go to (ii).

iii) If there are two or more unnumbered groups linked to group j, we choose the next one as *the one with the greatest mass* (the mass of a group is defined as the sum of the atomic masses of its constituents). If two or more neighbouring unnumbered groups have the same mass, we add the mass of their first neighbours to break the tie. If this does not lead to a decision, we proceed to the second neighbours and so on. If we run out of neighbours and there is still a tie, we choose a group arbitrarily among the ones that have been selected via this process and we indicate the convention. We number the group chosen as $j+1$, set $j = j+1$ and go to (ii). EXCEPTION: When we must choose the next group to the one whose center is a C_α in the backbone, instead of applying the rule of greatest mass, which would yield the group at the C' as the next one, we choose *the first group in the side chain* (for residues that are different from glycine). Then, we number the group chosen as $j+1$, set $j = j+1$ and go to (ii). This is done in order to ensure modularity, since, otherwise, the backbone would be always numbered first and the whole numeration would have to be modified if we added a new residue to the chain.

iv) If there are no unnumbered groups linked to group j, we prepare to start another branch and have two choices: For modularity reasons, we want to completely number the side chain before proceeding into the backbone. Hence, if we are numbering side chain groups and there are still unnumbered groups in the same side chain, we set j to *the number of the lowest numbered group that has unnumbered neighbours and that belongs to the side chain of the residue whose groups we are numbering.*. If we are not numbering side chain groups or we are numbering side chain groups but there is no unnumbered groups in the same side chain left, we set j to *the number of the lowest numbered group that has unnumbered neighbours in the whole peptide*. Then, we go to (ii).

This process terminates when all the groups are numbered.

2.3 Rules for numbering the atoms

The atoms will be numbered in the order that they will be positioned via internal coordinates in the Z-matrix. In fig. 5, the rules given in this subsection are exemplified in a protected histidine dipeptide.

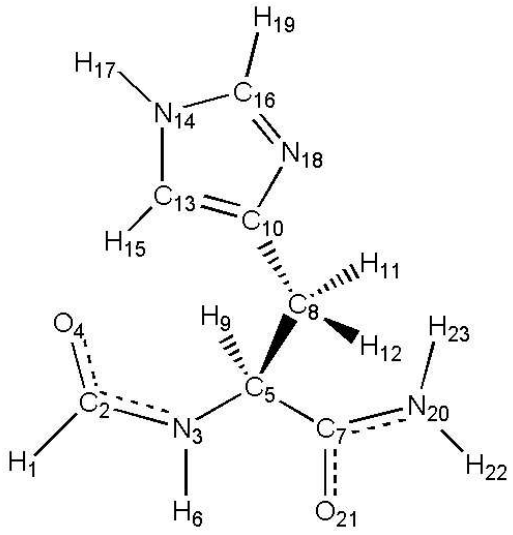

Figure 5: Atom numeration of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form).

The rules are as follows:

i) The first atom ($k = 1$), is chosen as *the heaviest of the external atoms in the first group*. If there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention. EXCEPTION: If the polypeptide is formyl-N-protected, instead of applying the rule, which would yield the oxygen at the formyl group, we choose *the hydrogen at the formyl group*

ii) The second atom ($k = 2$) is the center of the first group and we set $j = 1$ (the index of the group).

iii) If group $j + 1$ exists and is covalently attached to group j, we number the unnumbered vertices of group j *starting by the center of group* $j + 1$ *and, then, in order of decreasing mass.* If, otherwise, group $j + 1$ does not exist or it is not covalently attached to group j, *we simply number the unnumbered vertices of group* j *in order of decreasing mass.* If, at any point, there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention. EXCEPTION 1: If groups j and $j+$

1 belong to the same cyclic part of the molecule, the vertices of j that are centers of groups (other than $j+1$) belonging to the same cycle *must not be numbered at this step* (for an example of this rule, see the numeration of C_{13} and N_{18} in fig. 5). EXCEPTION 2: If the polypeptide is amide-C-protected, instead of applying the above rule and arbitrarily choosing one of the hydrogens in the terminal amide group before the other, we number *the trans hydrogen* before the other (see fig. 5). EXCEPTION 3: Due to the rules for the numeration of groups given in the previous subsection, the next group to the one at the C_α is the first one in the side chain. If we applied the general rule for numbering the vertices of the C_α-group, we would number first the center of the first group at the side chain and, then, the C' in the backbone. This would make the only principal dihedral defined on bond (C_α, N) different from the conventional Ramachandran angle ϕ. In order to avoid this, at this point, we number the C' first among the unnumbered vertices of the C_α-group and, then, resume the usual numeration process (see fig. 5).

iv) If group $j+1$ does not exist, we have finished. Otherwise, we set $j = j+1$ and go back to (iii).

The exception to rule (i) and the exceptions 2 and 3 to rule (iii) are introduced in order that the principal dihedrals that are to be defined after numbering the atoms conform to the biochemical IUPAC conventions for the dihedrals ϕ, ψ and ω in the backbone. At the termini, we have ensured that the atom where the C_α of the hypothetical residue 0 or $N+1$ would occur is used to define the principal dihedrals.

See fig. 10 for the numeration of the twenty naturally occurring amino acids with formyl-N- and amide-C-protection.

2.4 Rules for defining the internal coordinates

Using the numeration for the atoms given in the previous section, we give now a set of rules for defining the internal coordinates that conform with the *physical approach* discussed in the introduction of this work. The coordinates are written in Z-matrix form (see table 3) for convenience and the rules are applied to the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form) using the numeration given in fig. 5.

The rules are as follows:

i) The positioning of *the first three atoms* is special. The corresponding rows of the Z-matrix are *always* as the ones in table 3 (except, of course, for the chemical symbol in the first column, which may change).

ii) The positioning of the remaining vertices of group number 1 (if there is any) is also special, their rows in the Z-matrix are:

$$\mathrm{T}_i \quad (i,2) \quad (i,2,1) \quad (i,2,1,3)$$

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
C_{10}	(10,8)	(10,8,5)	**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
C_{13}	(13,10)	(13,10,8)	**(13,10,8,5)**
N_{14}	(14,13)	(14,13,10)	**(14,13,10,8)**
H_{15}	(15,13)	(15,13,10)	(15,13,10,14)
C_{16}	(16,14)	(16,14,13)	**(16,14,13,10)**
H_{17}	(17,14)	(17,14,13)	(17,14,13,16)
N_{18}	(18,16)	(18,16,14)	**(18,16,14,13)**
H_{19}	(19,16)	(19,16,14)	(19,16,14,18)
N_{20}	(20,7)	(20,7,5)	**(20,7,5,3)**
O_{21}	(21,7)	(21,7,5)	(21,7,5,20)
H_{22}	(22,20)	(22,20,7)	**(22,20,7,5)**
H_{23}	(23,20)	(23,20,7)	(23,20,7,22)

Table 3: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form), following the rules given in sec. 2.4. Principal dihedrals are indicated in bold face.

Where T is the chemical symbol of the i-th atom, and $(i, 2, 1, 3)$ is a phase dihedral.

iii) We set i to *the number that follows that of the last vertex of the first group.*

iv) We choose j as *the lowest numbered atom that is covalently linked to* i.

v) We choose k as *the lowest numbered atom that is covalently linked to* j.

vi) If no principal dihedral has been defined on the bond (j, k) (we say that a principal dihedral (i, j, k, l) *is* "on the bond (j, k)"), we choose l as *the lowest numbered atom that is covalently linked to* k. Otherwise, we choose l as *the second lowest numbered atom that is covalently linked to* j (i.e., the lowest numbered atom that is covalently linked to j and that is different from k, or, equivalently, the atom that was used to define the only principal dihedral on the bond (j, k)).

vii) The row of the Z-matrix that corresponds to atom i is:

$$\mathrm{T}_i \quad (i,j) \quad (i,j,k) \quad (i,j,k,l)$$

Where T is the chemical symbol of atom i, (i,j) is a bond length, (i,j,k) is a bond angle and (i,j,k,l) is a principal dihedral if the first case in point (vi) has occurred or a phase dihedral otherwise.

viii) If $i+1$ does not exist, we have finished. Otherwise, we set $i = i+1$ and go to (iv).

3 Application

3.1 Theory

When a number of degrees of freedom are removed from the description of the conformations of a physical system via their integrating out in the partition function, the energy function that remains, which describes the behaviour of the system only in terms of the rest of the degrees of freedom, is a *free energy*. It depends on the temperature and contains the entropy of the information that has been averaged out as well as the enthalpy. However, it is frequent, when studying the conformational preferences of model dipeptides in order to use the information for designing effective potentials of polypeptides [27–33], that the energy of these molecules be approximated by the Potential Energy Surface (PES) in the bidimensional space spanned by the Ramachandran angles ϕ and ψ [25, 33–35]. If we recognize that the potential energy of the system in the Born-Oppenheimer approximation (denoted by V_{3n-6}) depends on the $3n-6$ internal coordinates, this surface (denoted by V_2) may be defined as:

$$V_2(\phi,\psi) := \min_{Q^\alpha} V_{3n-6}(\phi,\psi,Q^\alpha) \ . \tag{1}$$

Where Q^α denotes the rest of the internal coordinates.

The use of this surface, instead of a free energy function with the Q^α degrees of freedom integrated out, is justified in the approximation that these internal coordinates are *hard* and that they are comparably much more difficult to excite at room temperature than ϕ and ψ. If we assume that this is correct, these hard degrees of freedom may be easily eliminated [17] and the partition function of the system may be written as follows:

$$Z = C \int \mathrm{d}\phi\, \mathrm{d}\psi\, \mathrm{d}Q^\alpha\, \mathrm{e}^{-\beta V_{3n-6}(\phi,\psi,Q^\alpha)} \simeq C' \int \mathrm{d}\phi\, \mathrm{d}\psi\, \mathrm{e}^{-\beta V_2(\phi,\psi)} \ . \tag{2}$$

Where $\beta := 1/RT$.

Note however that, in the "flexible" picture for the constraints, this expression is correct only if we assume that the Jacobian determinant of the change of coordinates from Cartesians to $\{\phi,\psi,Q^\alpha\}$ and the determinant of the potential

second derivatives matrix with respect to the hard coordinates, both evaluated at the equilibrium values, do not depend on ϕ and ψ (see ref. [17]). If, alternatively, we accept the "rigid" picture for the constraints, we must ask that the determinant of the induced metric tensor in the constrained hypersurface do not depend on ϕ and ψ [36]. If these approximations (which will be reexamined in future works) do not hold but the hardness of the Q^α degrees of freedom is still assumed, the expressions in eq. 2 must be modified by adding some correction terms to $V_2(\phi, \psi)$.

In eq. 2 for the partition function, one also may see that, apart from the different multiplicative constants C and C', which do not affect the expected values of observables, the use of the PES $V_2(\phi, \psi)$ as the fundamental energy function of the system is justified because it plays the same role as the whole potential energy of the system in the first integral.

However, although the hardness of the bond lengths, the bond angles and even the dihedral ω in the peptide bond may be assumed, this is not a good approximation for the rotameric degrees of freedom in the side chains of residues. In the frequently studied [25, 34] example of HCO-L-Ala-NH_2 (see fig. 6), as it has already been said in footnote 1, the side chain degree of freedom χ must be regarded as *soft*. Still, although it is more complex, a soft degree of freedom may also be averaged out if it is considered convenient.

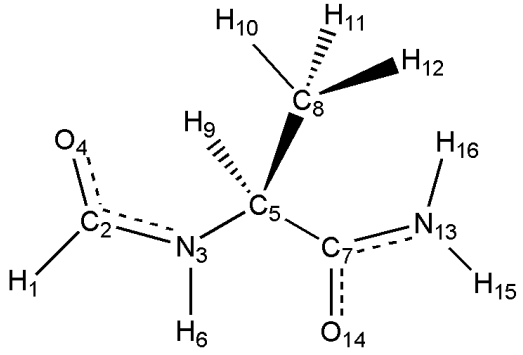

Figure 6: Atom numeration of the protected dipeptide HCO-L-Ala-NH_2.

In this section, we will assume that the energy of the formyl-alanine-amide dipeptide may be correctly approximated by a Potential Energy Hypersurface (PEH) (denoted by V_3) that depends on the Ramachandran angles ϕ and ψ but also on the principal dihedral χ that describes the rotation of the methyl group in the side chain. Analogously to eq. 1, its definition in terms of the whole energy of the system is:

$$V_3(\phi, \psi, \chi) := \min_{Q'^\alpha} V_{3n-6}(\phi, \psi, \chi, Q'^\alpha) \ . \tag{3}$$

Where Q'^α represents the internal coordinates that are not ϕ, ψ or χ.

Note, in addition, that the two definitions are related by the following expression:

$$V_2(\phi,\psi) = \min_{\chi} V_3(\phi,\psi,\chi) \ . \tag{4}$$

We will also assume for $V_3(\phi,\psi,\chi)$ the aforementioned approximations that lead to eq. 2, in such a way that we can write (deliberately omitting the irrelevant multiplicative constants):

$$\begin{aligned} Z &\simeq \int \mathrm{d}\phi\,\mathrm{d}\psi\,\mathrm{d}\chi\,\mathrm{e}^{-\beta V_3(\phi,\psi,\chi)} = \int \mathrm{d}\phi\,\mathrm{d}\psi\,Z(\phi,\psi) := \\ &:= \int \mathrm{d}\phi\,\mathrm{d}\psi\,\mathrm{e}^{-\beta F(\phi,\psi)} \ . \end{aligned} \tag{5}$$

Where we have defined:

$$Z(\phi,\psi) := \mathrm{e}^{-\beta F(\phi,\psi)} := \int \mathrm{d}\chi\,\mathrm{e}^{-\beta V_3(\phi,\psi,\chi)} \ . \tag{6}$$

This is what must be done in general when a soft degree of freedom is needed to be integrated out in Statistical Mechanics [37] and the approximations in ref. [17] cannot be made. The function $F(\phi,\psi)$ is a free energy because, in general, it depends on the temperature and it contains the entropy of the degree of freedom χ whose influence has been averaged.

We must remark at this point that, to integrate out the side chain angle χ could be reasonable if one's aim is to use the ab initio obtained information from a single dipeptide to include it in an effective potential for simulating polypeptides. There is no point in integrating out the Ramachandran angles ϕ and ψ, since the conformation of the larger system will depend crucially on their particular values, because they lie in the backbone of the molecule and there are as many pairs (ϕ,ψ) as residues in the chain. The side chain angle χ, on the contrary, will only influence its immediate surroundings and its importance could be of different magnitude depending on the treatment that the side chains are given in the model for the polypeptide.

In this context, if we wanted to use an energy function that does not depend on χ (in some circumstances, a computational must), we would have to perform the integral in the last term of eq. 6 and use $F(\phi,\psi)$ instead of $V_2(\phi,\psi)$, since, as it has already been remarked, χ is not a hard coordinate and the approximations needed to write eq. 2 do not hold. Therefore, if we compare the last term in eq. 5 with the last term in eq. 2, we see that, apart from additive constants that do not depend on ϕ and ψ and that come from the multiplicative constants omitted, the PES $V_2(\phi,\psi)$ must be understood as a candidate for *approximating* the more realistic $F(\phi,\psi)$ and saving much computational effort.

There are several ways of computing the potential energy, see, for example, the hindered rotor approximation in reference [38]. In the following subsections, the validity of this approximation will be assessed in the particular case of formyl-alanine-amide with ab initio Quantum Mechanics calculations.

3.2 Methods

The ab initio quantum mechanical calculations have been done with the package GAMESS [9] under Linux. The coordinates used for the HCO-L-Ala-NH_2 dipeptide in the GAMESS input files and the ones used to "move" the molecule in the the automatic Perl scripts that generated the input files are the SASMIC defined in sec. 2. They are presented in table 4 indicating the name of the conventional dihedral angles (see also fig. 6 for reference). In the energy optimizations, on the contrary, they have been converted to Delocalized Coordinates [4] to accelerate convergence.

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	ω_0 :=**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	ϕ :=**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
H_{10}	(10,8)	(10,8,5)	χ :=**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
N_{13}	(13,7)	(13,7,5)	ψ :=**(13,7,5,3)**
O_{14}	(14,7)	(14,7,5)	(14,7,5,13)
H_{15}	(15,13)	(15,13,7)	ω_1 :=**(15,13,7,5)**
H_{16}	(16,13)	(16,13,7)	(16,13,7,15)

Table 4: SASMIC internal coordinates in Z-matrix form of the protected dipeptide HCO-L-Ala-NH_2. Principal dihedrals are indicated in bold face and their typical biochemical name is given.

First, we have calculated the typical PES $V_2(\phi, \psi)$ defined in eq. 1 in a regular 12x12 grid, with both ϕ and ψ ranging from -165^o to 165^o in steps of 30^o. This has been done by running energy optimizations at the RHF/6-31+G(d) level of the theory, freezing the two Ramachandran angles at each value on the grid, starting from geometries previously optimized at a lower level of the theory and setting the gradient convergence criterium to `OPTTOL`=0.0001 and the self-consistent Hartree-Fock convergence criterium to `CONV`=0.00001.

Then, at each grid point, we have defined another one-dimensional grid in the coordinate χ that ranges from $\chi_0(\phi, \psi) - 50^o$ to $\chi_0(\phi, \psi) + 60^o$ in steps of 10^o, where $\chi_0(\phi, \psi)$ is one of the three equivalent equilibrium values (selected arbitrarily) of this degree of freedom at each point of the original PES. This partition in 12 points spans one third of the χ-space, but it is enough

for computing the integrals because the surface $V_3(\phi,\psi,\chi)$ has exact three-fold symmetry in χ (note, for example, that the value of V_3 at $\chi_0(\phi,\psi)-60^o$ would be equal to the one at $\chi_0(\phi,\psi)+60^o$). Next, we have run energy optimizations, with the same parameters described above and at the same level of theory, at each point of the χ-grid for every grid-value of the PES (i.e., freezing the three angles). The starting geometries have been automatically generated via Perl scripts taking the final geometries in the (ϕ,ψ)-grid and systematically changing χ. Note that this amounts to only changing the principal dihedral (10,8,5,3) in the Z-matrix in table 4; with poorly designed coordinates that did not separate the hard modes from the soft ones, this process would have been more difficult and rather unnatural.

After all the optimizations ($\sim$ 54 days of CPU time in 3.20 GHz PIV machines), we have 12x12x12=1728 points with grid coordinates (ϕ_i,ψ_j,χ_k) $i,j,k=1\ldots12$ of the function $V_3(\phi,\psi,\chi)$ and we may approximate the integral defining $F(\phi,\psi)$ in eq. 6 by a finite sum:

$$F(\phi_i,\psi_j) := -RT\,\ln\left(\sum_k \mathrm{e}^{-\beta V_3(\phi_i,\psi_j,\chi_k)}\right) =$$
$$= -RT\,\ln\left(\sum_k \mathrm{e}^{-\beta[V_3(\phi_i,\psi_j,\chi_k)-\langle V_3\rangle(\phi_i,\psi_j)]}\right) + \langle V_3\rangle(\phi_i,\psi_j)\ . \qquad (7)$$

Where the additive constants produced by the three-fold symmetry in the coordinate χ have been omitted.

The quantity $\langle V_3\rangle(\phi,\psi)$, defined as:

$$\langle V_3\rangle(\phi_i,\psi_j) := \frac{1}{12}\sum_k V_3(\phi_i,\psi_j,\chi_k)\ , \qquad (8)$$

has been introduced in order for the values of the exponential function to be in the precision range of the computer.

Analogously, the average energy may be computed via:

$$U(\phi_i,\psi_j) := \frac{\sum_k V_3(\phi_i,\psi_j,\chi_k)\mathrm{e}^{-\beta V_3(\phi_i,\psi_j,\chi_k)}}{\sum_k \mathrm{e}^{-\beta V_3(\phi_i,\psi_j,\chi_k)}} =$$
$$= \frac{\sum_k V_3(\phi_i,\psi_j,\chi_k)\mathrm{e}^{-\beta V_3(\phi_i,\psi_j,\chi_k)-\langle V_3\rangle(\phi_i,\psi_j)]}}{\sum_k \mathrm{e}^{-\beta[V_3(\phi_i,\psi_j,\chi_k)-\langle V_3\rangle(\phi_i,\psi_j)]}} = \ . \qquad (9)$$

And, finally, we extract the entropy from:

$$F(\phi_i,\psi_j) = U(\phi_i,\psi_j) - TS(\phi_i,\psi_j)\ . \qquad (10)$$

Additionally, apart from the calculations needed to integrate out χ, we have also performed an unconstrained geometry optimization in the basin of attraction of the local minima of the PES normally known as $\gamma_{\rm L}$ or $C7_{\rm eq}$ depending on the author [35]. This calculation was done at the MP2/6-31++G(d,p) level of the theory and with the same values of the variables `OPTTOL` and `CONV` than the ones used in the PES case. The starting geometry was the final structure corresponding to the point $(-75^o, 75^o)$ of the PES calculations at the lower level of the theory described in the preceding paragraphs.

In the local minimum found, we have computed the Hessian matrix (also at MP2/6-31++G(d,p)) in two different sets of coordinates: the properly defined SASMIC shown in table 4 and an ill-defined set in which the lines corresponding to the hydrogens H_{10}, H_{11} and H_{12} in the side chain have been substituted by those in table 1. This is done to numerically illustrate the better separation of the hard and soft modes achieved by the internal coordinates defined in this work with respect to other Z-matrix-like coordinates.

3.3 Results

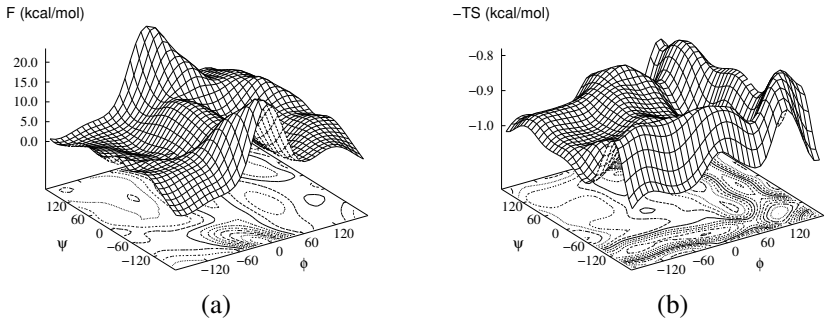

Figure 7: Ramachandran plots of **(a)** the free energy $F(\phi, \psi)$ and **(b)** $-TS(\phi, \psi)$ in the model dipeptide HCO-L-Ala-NH_2.

In order to assess if $V_2(\phi, \psi)$ could be considered a good approximation of $F(\phi, \psi)$, we have used a statistical quantity, defined in [39], which measures the typical error that one makes in the energy differences between arbitrary pairs of conformations of the system if one effective potential is used instead of the other. If we measure this *distance* between $F(\phi, \psi)$ and $V_2(\phi, \psi)$, using the 144 points in the (ϕ, ψ)-grid, we obtain:

$$d(F, V_2) = 0.098\ RT\ . \tag{11}$$

We present the result in units of RT (at 300^o K, where $RT \simeq 0.6$ kcal/mol) because it has been argued in [39] that, if the distance between two different approximations of the energy of the same system is less than RT, one may safely

substitute one by the other without altering the relevant physical properties. *In this case, this criterium is widely satisfied.* Moreover, if one assumes that the effective energy studied will be used to construct a polypeptide potential and that the latter will be designed as simply the sum of mono-residue ones (making each term suitably depend on different pairs of Ramachandran angles), then, the number N_{res} of residues up to which one may go keeping the distance between the two approximations of the N-residue potential below RT is (see ref. [39]):

$$N_{res}(F, V_2) = \left(\frac{RT}{d(F, V_2)} \right)^2 \simeq 104 \ . \tag{12}$$

The goodness of the approximation in this case is much due to the simplicity and small size of the side chain of the alanine residue and also to the fact that the dipeptide is isolated. For bulkier residues included in polypeptides, we expect the difference between $F(\phi, \psi)$ and $V_2(\phi, \psi)$ to be more important.

Although the essential result is the one stated in the previous paragraphs, we wanted to look in more detail at the origin of the differences between $F(\phi, \psi)$ and $V_2(\phi, \psi)$. For this, we have first subtracted from $F(\phi, \psi)$, $U(\phi, \psi)$ and $V_2(\phi, \psi)$ the same constant reference (min $F(\phi, \psi)$)[5] in order to render the numerical values more manageable and to minimize the statistical error of the y-intercept in the linear fits [40, 41] that will be made in the following.

Then, fitting $U(\phi, \psi)$ against $V_2(\phi, \psi)$, we have found that they are more correlated than $F(\phi, \psi)$ and $V_2(\phi, \psi)$ (compare the Pearson's correlation coefficient, $r(U, V_2) = 0.999999$ vs. $r(F, V_2) = 0.999954$, and the aforementioned distance, $d(U, V_2) = 0.015\ RT$ vs. $d(F, V_2) = 0.098\ RT$), and that they are separated by an almost constant offset: $V_2(\phi, \psi)$ is ~ 0.3 kcal/mol lower that $U(\phi, \psi)$ (on the other hand, $V_2(\phi, \psi)$ is ~ 0.6 kcal/mol higher than $F(\phi, \psi)$). Hence, the three Ramachandran surfaces $F(\phi, \psi)$, $U(\phi, \psi)$ and $V_2(\phi, \psi)$ are very similar, except for an offset. In fig. 7a, $F(\phi, \psi)$ is depicted graphically and, in fig 8, the relative offsets among the three energies are schematically shown.

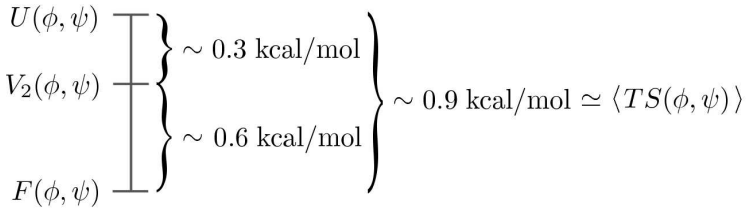

Figure 8: Relative offsets among the thermodynamical surfaces involved in the study.

Contrarily, the entropy (we use $TS(\phi, \psi)$ in order to deal with quantities that have units of energy), which may be found in fig. 7b, and whose average

[5] At the level of the theory used in the calculations, the minimum of $F(\phi, \psi)$ in the grid is -414.7985507934 hartree.

magnitude is ~ 0.9 kcal/mol, is almost uncorrelated with $F(\phi,\psi)$, $U(\phi,\psi)$ and $V_2(\phi,\psi)$, being the correlation coefficients $r(TS,F) = 0.382$, $r(TS,U) = 0.379$ and $r(TS,V_2) = 0.381$, respectively. Hence, given that $d(U,V_2)$ is almost an order of magnitude lower than $d(F,V_2)$, it is reasonable to conclude that the greatest part of the (little) noise between $F(\phi,\psi)$ and $V_2(\phi,\psi)$ comes from the entropic term $-TS(\phi,\psi)$. This is supported by the fact that the difference $F(\phi,\psi) - V_2(\phi,\psi)$ is highly correlated with $TS(\phi,\psi)$, being the correlation coefficient $r(F - V_2, TS) = 0.998$.

Finally, and in order to illustrate the better separation of the hard and soft modes achieved by the internal coordinates defined in this work, we have calculated the Hessian matrix in the minimum γ_{L} (also $\mathrm{C7_{eq}}$) in two different sets of coordinates. They are described at the end of sec. 3.2 and they correspond to the SASMIC set, defined according to the rules given in sec. 2, and a set in which the coordinates that position the hydrogens in the side chain have been ill-defined.

In fig. 9, we present the sub-boxes of the two Hessian matrices corresponding to the coordinates defined in tables 2 and 1.

Properly defined coordinates

	χ	α_1	α_2
χ	15.74	1.40	8.71
α_1	1.40	110.98	−54.23
α_2	8.71	−54.23	115.37

Ill-defined coordinates

	γ_1	γ_2	γ_3
γ_1	113.49	−55.55	−52.60
γ_2	−55.55	110.98	−54.23
γ_3	−52.60	−54.23	115.37

Figure 9: Sub-boxes of the Hessian matrix in the minimum γ_{L} (also $\mathrm{C7_{eq}}$) corresponding to the coordinates defined in tables 2 and 1. The quantities are expressed in kcal/mol $\cdot$ rad^{-2}. See the text for more details.

From the values shown, one can conclude that, in the "properly defined coordinates", some convenient characteristics are present: on one side, the relatively low values of the elements $H_{\chi\alpha_1}$ and $H_{\chi\alpha_2}$ (and their symmetric ones) indicate that the soft degree of freedom χ and the hard ones α_1 and α_2, which describe the internal structure of the methyl group, are uncoupled to a reasonable extent; on the other side, the relatively low value of $H_{\chi\chi}$ compared to $H_{\alpha_1\alpha_1}$ and $H_{\alpha_2\alpha_2}$ (a difference of almost an order of magnitude) proves that χ may be regarded as soft when compared to the hard degrees of freedom α_1 and α_2.

On the contrary, in the "ill-defined coordinates", the three dihedrals are hard, considerably coupled and equivalent.

4 Conclusions

Extending the approach of refs. [14–16] and the ideas stated in [1, 2, 26], we have defined a systematic numeration of the groups, the atoms and the internal

coordinates (termed SASMIC) of polypeptide chains. The advantages of the rules herein presented are many-fold:

- The internal coordinates may be easily cast into conventional Z-matrix form and they can be directly implemented into quantum chemical packages.
- The algorithm for numbering allows for automatizing and facilitates the coding of computer applications.
- The modularity of the numeration system in the case of polypeptides permits the addition of new residues without essentially changing the already numbered items. This is convenient if databases of peptide structures need to be designed.
- The set of internal coordinates defined reasonably separate the hard and soft movements of polypeptides for arbitrary conformations using only topological information.

A number of Perl scripts that automatically generate these coordinates for polypeptide chains are provided as supplementary material. Also, a supplementary document is provided that states the rules for numbering the groups, atoms and defining the internal coordinates in general organic molecules.

In addition, we have used the coordinates herein defined and ab initio Quantum Mechanics to assess the approximation of the free energy obtained from averaging out the rotameric degree of freedom χ via the conventional PES in the protected dipeptide HCO-L-Ala-NH_2. Applying the criterium in ref. [39], we have found that approximating $F(\phi, \psi)$ by $V_2(\phi, \psi)$ is justified up to polypeptides of medium length ($\sim$ 100 residues) and much computational effort may be saved using the PES instead of the more realistic free energy. However, the small size of the side chain of the alanine residue and the fact that the dipeptide is isolated do not allow to extrapolate this result. For bulkier residues included in polypeptides, we expect the difference between $F(\phi, \psi)$ and $V_2(\phi, \psi)$ to be more important.

We would like to thank I. Calvo and G.A. Chass, for illuminating discussions. We would also like to thank the reviewers of the manuscript for much useful suggestions regarding the structure of the article. The numerical calculations have been performed at the BIFI computing facilities. We thank I. Campos, for the invaluable CPU time and the efficiency at solving the problems encountered.

This work has been supported by the Aragón Government ("Biocomputación y Física de Sistemas Complejos" group) and by the research grants MEC (Spain) FIS2004-05073 and MCYT (Spain) BFM2003-08532. P. Echenique is supported by a MEC (Spain) FPU grant.

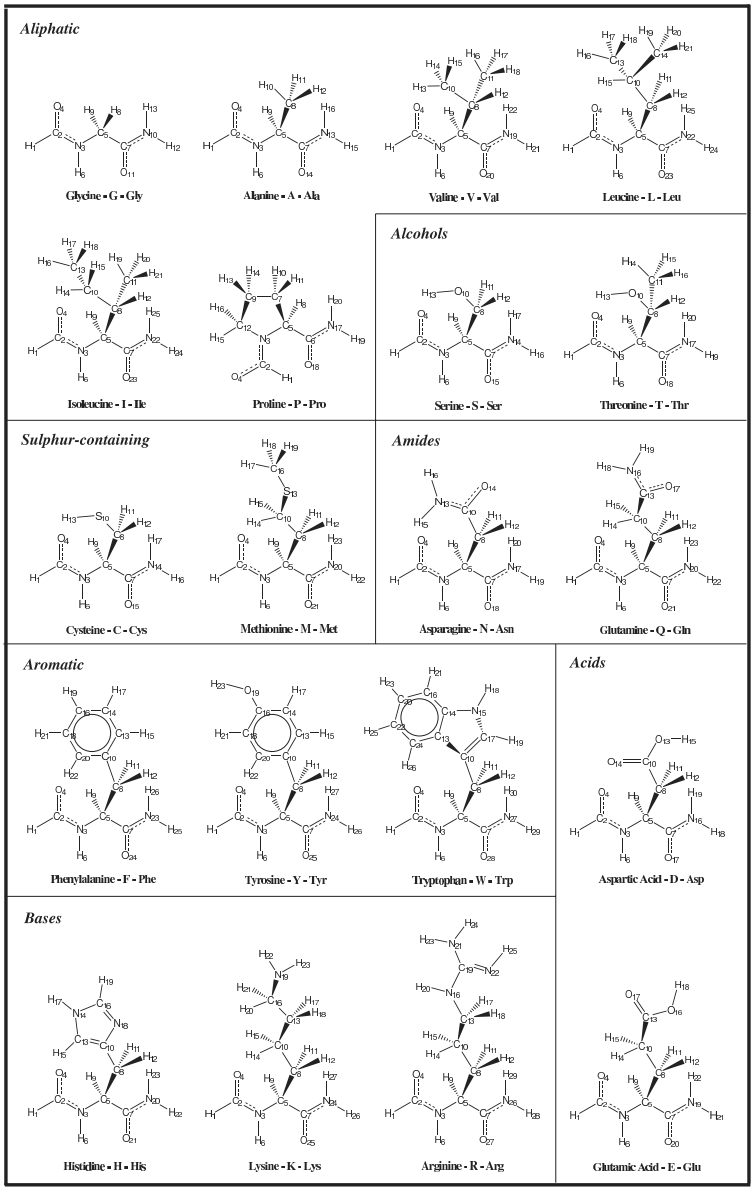

Figure 10: Numeration of the left-handed dipeptides HCO-L-X-NH_2, where X runs on the twenty naturally occurring amino acids (except for Glycine, which is the achiral species HCO-Gly-NH_2). Uncharged side chains are displayed and Histidine is shown in its ϵ tautomeric form.

Appendix

SASMIC rules for general organic molecules

The definitions found in sec. 2.1 are kept. The only changes affect the rules for numbering the groups and the atoms, since, when this is achieved, the rules for defining the internal coordinates are the same as the ones in sec. 2.4.

Rules for numbering the groups

As we have done for peptides, first of all, we will divide the molecule in groups and number them. To do this we proceed *by branches*, i.e., we choose the next group following a linear sequence of covalently attached groups until there is no possible next one, in which case, we either have finished the numeration process or we start another branch. Every group is numbered one time and it cannot be renumbered as the process continues.

In fig. 11, we have implemented these rules for general organic molecules in a protected histidine dipeptide.

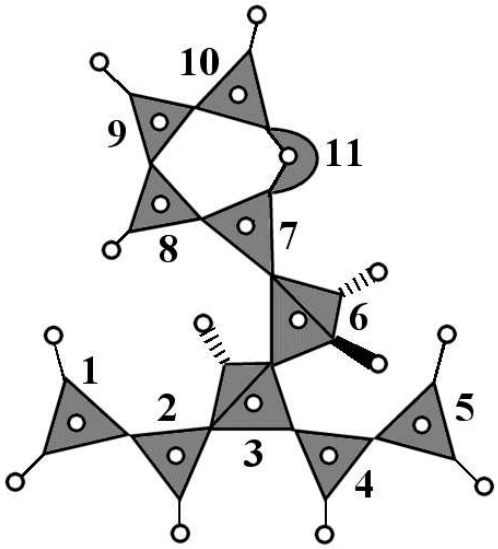

Figure 11: Group identification and numeration in the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form), following the rules for general organic molecules. The different types of groups are shown as gray-filled polyhedra.

The rules are as follows:

i) The first group ($j = 1$), is chosen, among those that are linked to the molecule via only one internal bond (termed "terminal groups"), as *the*

one that has the greater mass (the mass of a group is defined as the sum of the atomic masses of its constituents). If two or more terminal groups have the same mass, we add the mass of their first neighbours to break the tie. If this does not lead to a decision, we proceed to the second neighbours and so on. If we run out of neighbours and there is still a tie, we choose a group arbitrarily among the ones that have been selected via this process and we indicate the convention. If there are no terminal groups, we perform this selection process among those groups that have *at least one external atom*[6].

ii) If there is only one unnumbered group linked to group j, we number it as $j+1$, set $j = j+1$ and go to (ii).

iii) If there are two or more unnumbered groups linked to group j, we choose *the one with the greater mass* as in point (i), we number it as $j+1$, set $j = j+1$ and go to (ii).

iv) If there are no unnumbered groups linked to group j but there are still unnumbered groups in the molecule, we set j to *the number of the lowest numbered group that has unnumbered neighbours* (we prepare to start another branch) and we go to (ii).

This process terminates when all the groups are numbered.

Rules for numbering the atoms

The atoms will be numbered in the order that they will be positioned via internal coordinates in the Z-matrix. As in the previous section, in fig. 12, these rules for a general organic molecule are exemplified in a protected histidine dipeptide.

The rules are as follows:

i) The first atom ($k = 1$), is chosen as *the heaviest of the external atoms in the first group*. If there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention.

ii) The second atom ($k = 2$) is the center of the first group and we set $j = 1$ (the index of the group).

iii) If group $j+1$ exists and is covalently attached to group j, we number the unnumbered vertices of group j *starting by the center of group* $j+1$ *and, then, in order of decreasing mass*. If, otherwise, group $j+1$ does not exist or it is not covalently attached to group j, *we simply number the unnumbered vertices of group* j *in order of decreasing mass*. If, at any point, there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention. EXCEPTION: If groups j and $j+1$

[6]The rare case in which there are neither terminal groups nor external atoms (such as C_{60} fullerene) will not be treated here, although it would require only a small number of adjustments to the rules.

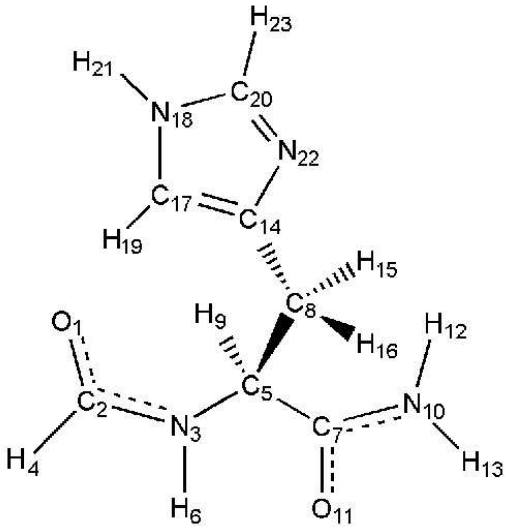

Figure 12: Atom numeration of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form), following the rules for general organic molecules.

belong to the same cyclic part of the molecule, the vertices of j that are centers of groups (other than $j+1$) belonging to the same cycle *must not be numbered at this step* (for an example of this rule, see the numeration of C_{17} and N_{22} in fig. 12).

iv) If group $j+1$ does not exist, we have finished. Otherwise, we set $j = j+1$ and go back to (iii).

Rules for defining the internal coordinates

Using the numeration for the atoms given in the previous section, the rules for defining the SASMIC internal coordinates that conform with the *physical approach* discussed in the introduction are the same as the ones given in the sec. 2.4.

The coordinates, written in Z-matrix form, of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form) using the numeration given in fig. 12 are given in table 5.

Atom name	Bond length	Bond angle	Dihedral angle
O_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
H_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
N_{10}	(10,7)	(10,7,5)	**(10,7,5,3)**
O_{11}	(11,7)	(11,7,5)	(11,7,5,10)
H_{12}	(12,10)	(12,10,7)	**(12,10,7,5)**
H_{13}	(13,10)	(13,10,7)	(13,10,7,12)
C_{14}	(14,8)	(14,8,5)	**(14,8,5,3)**
H_{15}	(15,8)	(15,8,5)	(15,8,5,14)
H_{16}	(16,8)	(16,8,5)	(16,8,5,14)
C_{17}	(17,14)	(17,14,8)	**(17,14,8,5)**
N_{18}	(18,17)	(18,17,14)	**(18,17,14,8)**
H_{19}	(19,17)	(19,17,14)	(19,17,14,18)
C_{20}	(20,18)	(20,18,17)	**(20,18,17,14)**
H_{21}	(21,18)	(21,18,17)	(21,18,17,20)
N_{22}	(22,20)	(22,20,18)	**(22,20,18,17)**
H_{23}	(23,20)	(23,20,18)	(23,20,18,22)

Table 5: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ϵ tautomeric form), following the rules for general molecules. Principal dihedrals are indicated in bold face.

References

[1] G. A. Chass, M. A. Sahai, J. M. S. Law, S. Lovas, Ö. Farkas, A. Perczel, J.-L. Rivail, and I. G. Csizmadia, Toward a computed peptide structure database: The role of a universal atomic numbering system of amino acids in peptides and internal hierarchy of database, *Intl. J. Quant. Chem.* **90**, 933 (2002).

[2] M. A. Sahai, S. Lovas, G. A. Chass, P. Botond, and I. G. Csizmadia, A modular numbering system of selected oligopeptides for molecular computations: using pre-computed amino acid building blocks, *J. Mol. Struct.* **666-667**, 169 (2003).

[3] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons, Chichester, 2nd edition, 2002.

[4] J. Baker, A. Kessi, and B. Delley, The generation and use of delocalized internal coordinates in geometry optimization, *J. Chem. Phys.* **105**, 192 (1996).

[5] M. von Arnim and R. Ahlrichs, Geometry optimization in generalized natural internal coordinates, *J. Chem. Phys.* **111**, 9183 (1999).

[6] B. Paizs, J. Baker, S. Suhai, and P. Pulay, Geometry optimization of large biomolecules in redundant internal coordinates, *J. Chem. Phys.* **113**, 6566 (2000).

[7] K. Németh and M. Challacombe, The quasi-independent curvilinear coordinate approximation for geometry optimization, *J. Chem. Phys.* **121**, 2877 (2004).

[8] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.

[9] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, H. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, General Atomic and Molecular Electronic Structure System, *J. Comp. Chem.* **14**, 1347 (1993).

[10] P. Pulay, G. Fogarasi, F. Pang, and J. E. Boggs, Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives, *J. Am. Chem. Soc.* **101**, 2550 (1979).

[11] P. Pulay and G. Fogarasi, Geometry optimization in redundant internal coordinates, *J. Chem. Phys.* **96**, 2856 (1992).

[12] G. Fogarasi, X. Zhou, P. W. Taylor, and P. Pulay, The calculation of ab initio molecular geometries: Efficient natural internal coordinates and empirical correction by offset forces, *J. Am. Chem. Soc.* **114**, 8191 (1992).

[13] C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, Using redundant internal coordinates to optimize equilibrium geometries and transition states, *J. Comp. Chem.* **17**, 49 (1996).

[14] A. K. Mazur and R. A. Abagyan, New methodology for computer-aided modelling of biomolecular structure and dynamics. 1. Non-cyclic structures, *J. Biomol. Struct. Dyn.* **6**, 815 (1989).

[15] R. A. Abagyan and A. K. Mazur, New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles, *J. Biomol. Struct. Dyn.* **6**, 833 (1989).

[16] R. A. Abagyan, M. M. Totrov, and D. A. Kuznetsov, ICM: A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation, *J. Comp. Chem.* **15**, 488 (1994).

[17] N. Gō and H. A. Scheraga, On the use of classical statistical mechanics in the treatment of polymer chain conformation, *Macromolecules* **9**, 535 (1976).

[18] M. Karplus and J. N. Kushick, Method for estimating the configurational entropy of macromolecules, *Macromolecules* **14**, 325 (1981).

[19] W. J. Hehre, W. A. Lathan, R. Ditchfield, M. D. Newton, and J. A. Pople, Gaussian 70, Quantum Chemistry Program Exchange, 1970, Program No. 237.

[20] I. N. Levine, *Quantum Chemistry*, Prentice Hall, Upper Saddle River, 5th edition, 1999.

[21] R. F. Frey, J. Coffin, S. Q. Newton, M. Ramek, V. K. W. Cheng, F. A. Momany, and L. Schäfer, Importance of correlation-gradient geometry optimization for molecular conformational analyses, *J. Am. Chem. Soc.* **114**, 5369 (1992).

[22] K. J. Jalkanen and S. Suhai, N-acetyl-L-alanine N'-methylamide: A density functional analysis of the vibrational absorption and vibrational circular dichroism spectra, *Chem. Phys.* **208**, 81 (1996).

[23] K. J. Jalkanen, R. M. Nieminen, M. Knapp-Mohammady, and S. Suhai, Vibrational analysis of various isotopomers of L-alanyl-L-alanine in aqueous solution: Vibrational Circular Dichroism, Raman, and Raman Optical Activity spectra, *Intl. J. Quant. Chem.* **92**, 239 (2002).

[24] H. B. Schlegel, Some practical suggestions for optimizing geometries and locating transition states, in *New Theoretical Concepts for Understanding Organic Reactions*, edited by J. Bertrán and I. G. Csizmadia, pp. 33–53, Kluwer Academic, The Netherlands, 1989.

[25] C.-H. Yu, M. A. Norman, L. Schäfer, M. Ramek, A. Peeters, and C. van Alsenoy, Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation, *J. Mol. Struct.* **567–568**, 361 (2001).

[26] J. L. Alonso, G. A. Chass, I. G. Csizmadia, P. Echenique, and A. Tarancón, Do theoretical physicists care about the protein folding problem?, in *Meeting on Fundamental Physics 'Alberto Galindo'*, edited by R. F. Álvarez-Estrada et al., Aula Documental, Madrid, 2004, (arXiv:q-bio.BM/0407024).

[27] T. A. Halgren, Merck molecular force field. I. Basis, form, scope, parametrization, and performance of MMFF94, *J. Comp. Chem.* **17**, 490 (1996).

[28] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc.* **118**, 11225 (1996).

[29] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. Merz, K. M., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* **117**, 5179 (1995).

[30] J. W. Ponder and D. A. Case, Force fields for protein simulations, *Adv. Prot. Chem.* **66**, 27 (2003).

[31] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comp. Chem.* **4**, 187 (1983).

[32] A. D. MacKerell Jr., B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, CHARMM: The energy function and its parameterization with an overview of the program, in *The Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer et al., pp. 217–277, John Wiley & Sons, Chichester, 1998.

[33] A. R. MacKerell Jr., M. Feig, and C. L. Brooks III, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, *J. Comp. Chem.* **25**, 1400 (2004).

[34] A. M. Rodríguez, H. A. Baldoni, F. Suvire, R. Nieto Vázquez, G. Zamarbide, R. D. Enriz, Ö. Farkas, A. Perczel, M. A. McAllister, L. L. Torday, J. G. Papp, and I. G. Csizmadia, Characteristics of ramachandran maps of L-alanine diamides as computed by various molecular mechanics, semiempirical and ab initio MO methods. A search for primary standard of peptide conformational stability, *J. Mol. Struct.* **455**, 275 (1998).

[35] A. Perczel, O. Farkas, I. Jakli, I. A. Topol, and I. G. Csizmadia, Peptide models. XXXIII. Extrapolation of low-level Hartree-Fock data of peptide conformation to large basis set SCF, MP2, DFT and CCSD(T) results. The Ramachandran surface of alanine dipeptide computed at various levels of theory, *J. Comp. Chem.* **24**, 1026 (2003).

[36] W. K. Den Otter and W. J. Briels, Free energy from molecular dynamics with multiple constraints, *Mol. Phys.* **98**, 773 (2000).

[37] T. Lazaridis and M. Karplus, Thermodynamics of protein folding: a microscopic view, *Biophys. Chem.* **100**, 367 (2003).

[38] P. Y. Ayala and H. B. Schlegel, A nonorthogonal ci treatment of symmetry breaking in sigma formyloxyl radical, *J. Chem. Phys.* **108**, 7560 (1998).

[39] J. L. Alonso and P. Echenique, A physically meaningful method for the comparison of potential energy functions, *J. Comp. Chem.* **27**, 238 (2006).

[40] P. R. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, Mc. Graw–Hill, New York, 3rd edition, 2003.

[41] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C. The art of scientific computing*, Cambridge University Press, New York, 2nd edition, 2002.