

Protein Modelling and Surface Folding by Limiting the Degrees of Freedom

Meir Israelowitz, Birgit Weyand, Syed W. H. Rizvi, Christoph Gille and Herbert P. von Schroeder

Abstract One aspect of tissue engineering represents modelling of the extracellular matrix of connective tissue as the fiber network arrangement of the matrix determines its tensile strength. In order to define the correct position of the e.g. collagen in a structure, an optimized tertiary structure must be characterized. Existing approaches of protein models consider random packing of rigid spheres. We propose an alternative strategy to model protein structure by focusing on the folding. Our model considers (a) segments of amino-acid peptides or beads, (b) hydrogen bond distances, and (c) the distance geometry as functional components rather than minimizing distances between the centers of atoms. We reduced the molecular volume by using concepts from low dimensional topology, such as *braids* and surfaces, via differential geometry. A *braid* group maintains the continuity of a sequence while the spatial minimization is performed, and guarantees the continuity during the process. We have applied this approach to different examples of known protein sequences using ab initio protocols of ProteoRubix SystemsTM. Sequence files of three different proteins types were tested and modeled by ProteoRubixTM and compared to models derived by other methods. ProteoRubixTM

M. Israelowitz · S. W. H. Rizvi · C. Gille · H. P. von Schroeder (✉)
Biomimetics Technologies Inc, Toronto, Canada
e-mail: herb.vonschroeder@uhn.ca

B. Weyand
Department of Plastic, Hand and Reconstructive Surgery, Hannover Medical School,
Hannover, Germany

C. Gille
Institute of Biochemistry, Charite Universitaetsmedizin, Berlin, Germany

H. P. von Schroeder
University Health Network, Toronto, Canada

H. P. von Schroeder
University of Toronto, Toronto, Canada

created near-identical models with minimal computational load. This model can be expanded to large, multi-molecular network structures.

1 Introduction

Tissue Engineering is an interdisciplinary field in the cross-section of biology, medicine, chemistry, physics, material sciences, engineering sciences and informatics. One aspect of tissue engineering represents modelling of extracellular matrix of connective tissue. The mechanical properties of the extracellular matrix are defined by its structure and composition. Collagen, for example, is the most abundant extracellular matrix protein in connective tissue, and is composed of three polypeptide strands each with 1,000 amino acids that are assembled to form a fibril [31]. The fiber network arrangement of the collagen matrix determines its tensile strength. In order to define the correct position of the collagen in a structure, an optimized tertiary structure must be characterized.

With the purpose of modelling a protein such as collagen, energy calculations must consider six degrees of freedom for each single atom. The calculations include hydrogen bonding and van der Waals forces, and moreover, because the interactions of the amino acids are both hydrophobic and hydrophilic forces in an aqueous environment, these calculations must also be considered in time [27]. The time for protein folding takes place in milliseconds to minutes [68]. In the case of collagen or other large proteins, there are assistance proteins, the so-called chaperones such as P52, which support protein folding and tertiary structure by holding initial amino acids for assembly of the total protein [61].

Protein modelling is considered to be a nondeterministic polynomial (NP) time problem that is regarded as inherently difficult if its solution requires significant resources, regardless of the algorithm that is used to solve the problem [45]. In order to solve a NP problem, we need to consider a program that makes 2^n operations before halting, since exponential-time algorithms might be unusable from the practical point of view. Collagen is an exceptionally large protein, but even for a small number of molecules (e.g. $n = 100$) and 10^{12} computer operations per second, a program would run for about 4×10^{10} years to solve a problem ([70] and Appendix 1) and therein lies the challenge for protein modelling.

Protein modelling is generally based on the fact that a structure is striving for the lowest state of energy based on the Lennard-Jones potential that defines energy as a function of the distances of the minimizations. Protein modelling uses free Gibbs energy to obtain the lowest state of energy. Currently, there are eight major methods used to denote protein conformation spaces or lowest state of energy:

1.1 *Molecular Dynamics* in which every coordinate position of the atoms contained within the amino acids in the sequence is taken into consideration. This

method calculates the gradient or steep-out slope to obtain an energy minimum [69, 75].

- 1.2 The *Monte Carlo* method uses computational algorithms that rely on repeated random sampling techniques to find a global minimum. The *Monte Carlo* method takes a sample of the averages on the conformation path; however, it is possible for the path to enter into local minima that are difficult to differentiate [13, 89].
- 1.3 *Smith's Microfibril model* calculates the conformation energy by finding the difference between the random state and the final conformation. The difference between the two is the objective function [17, 43, 67].
- 1.4 *Probabilistic techniques* based on Bayesian inference have been developed by Garnier et al. [30] as information theory. The method takes into account the probability of each amino acid having a particular secondary structure, and also considers the assumptions and probabilities of each structure contributing to that of its neighbours. The method is roughly 65 % accurate and is dramatically more successful in predicting alpha helices than beta sheets, which it frequently erroneously predicted as loops or disorganized regions [52].
- 1.5 *Neural Networks* use training sets of solved structures to identify common sequence motifs associated with particular arrangements of secondary structures. These methods are over 70 % accurate in their predictions, although beta strands are still often under-predicted due to the lack of three-dimensional structural information that would allow assessment of hydrogen bonding patterns that can promote formation of the extended conformation required for the presence of a complete beta sheet [52].
- 1.6 *Support Vector Machine (SVM)* is a method which analyzes data and recognizes patterns that are used for classification. SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input. SVM has proven particularly useful for predicting the locations of turns, which are difficult to identify with statistical methods data [60]. The SVM requirement for relatively small training sets has also been cited as an advantage to avoid over-fitting to existing structures. The major limitation for machine learning techniques is the attempt to predict more fine-grained local properties of proteins, such as multiple back bone dihedral angle in unassigned regions [87].
- 1.7 *Genetic Algorithm (GA)* is population of strings, which encode candidate solutions to optimize a problem to evolve a better solution. The fundamental *Genetic Algorithm* problems approach, due protein modelling interactions, are elaborate, hence *Genetic Algorithm* evaluations for complex problems are often the most prohibitive. *Genetic Algorithms* do not scale well with complexity, meaning that with increasing numbers of elements there is an exponential increase in the search space size. In general, the *Genetic Algorithm* has a tendency to optimize to local minima, and therefore becomes globally limited [79].
- 1.8 Ab initio models have been used exclusively to model binding sites of metals to proteins [80] or use statistical methods to find a secondary structures

[10, 35, 71]. *Ab initio* (from the beginning) refers to the beginning of an amino acid sequence.

There are several major limitations with all of these models. In general, all of the model simulations are restricted by computational time and cost. To minimize a structure, there is a need to consider every change of position in the configuration of a peptide or protein. In the case of establishing global minima, all possible energies need to be tested which can become an exponentially large problem [13, 36, 74, 81]. Further, existing models do not consider the dynamic behavior of proteins in an aqueous environment since they cannot consider water molecules [28, 38, 57, 76] primarily due to computer limitations [7, 12]. Secondary structure formation also depends on several other factors [90] including for example accessibility of residues to a solvent [47], the protein structural class [22], and even the organism from which the proteins are obtained [49]. These factors can further confound modelling paradigms.

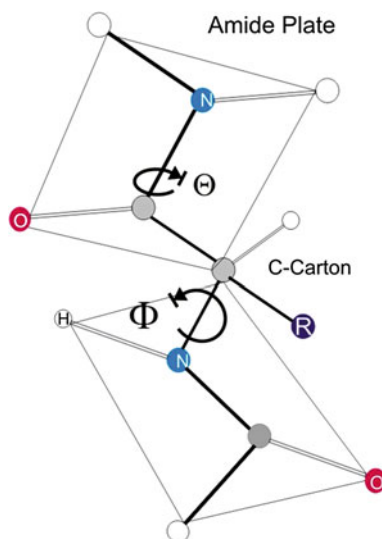
While X-ray crystallography is used to validate a protein structure, it only shows an aspect of a structure in that it does not map the real aqueous environment of the protein that affects protein interactions and structure. In addition, X-ray crystallography does not describe the actual interactions that occur during the folding process which is a further limitation [82]. In contrast, nuclear magnetic resonance (NMR) can be used to show structure in real time in an aqueous environment, but so far this method is limited to 70 amino acids [63].

Based on the limitations of existing protein modelling techniques, other approaches need to be considered; for example, if the *volume and area* of the protein can be optimized, this optimization can be used to calculate primary but also tertiary structures. The method presented uses the optimal volumes to bring a structure closest to the lowest energy state.

2 Model

The polypeptide chain of a protein can be represented by a low-dimensional topology structure called a *braid* group [1, 2, 8]. It is defined as the union of arc lengths forming a string, or *braid* that can readily model the fundamental properties of the peptide backbone [48, 53]. Rather than considering a peptide as a linear sequence of amino acid residues, the peptide bonds of the protein chain form the arc lengths of our *braid*. A peptide chain thus consists of a series of rigid arc lengths carrying various substitute groups. Each arc length runs from the oxygen to carbon of the amide bond to, but not including the next peptide carbonyl carbon [9]. Folding the polypeptide chain into different conformations simply results in changing the relative orientation of these arc lengths. Although this grouping does not follow the biosynthetic pattern, it limits orientation changes to movements about the rotating C–CO bond given by Φ and Θ as shown in Fig. 1. Next, the volume of the chain can be minimized [57]. The volume, given by a protein, can be

Fig. 1 Rotations around peptide groups. Two planar peptide groups are shown in this illustration. The only reasonably free movements are rotations around the C_α -ON bond (measured as Φ) and the C_α -OC bond (measured as Θ). By convention, Φ is 180° and Θ is 90° in the conformation shown and increase, as indicated, in the clockwise direction when viewed from C_α .



thought of as a backbone with additional groups attached to it. The C-N backbone is not straight because the bonds are not collinear. For example, carbon forms single bonds that are spaced equally apart from each other form a tetrahedron angle (109.5°) [83] rather than straight chains in the case of the residues.

With the distances between the C-N backbone being non-collinear, the angles between directional vectors are not parallel. From this, the groups are replaced with an outline of the atoms centered on the backbone so that we have strings of *beads* (though the bead shape is not round [21]). The lacing of the *beads* is the backbone of the protein is shown in Fig. 2a. Each of the amino acids has bonds that can rotate. In most cases there are two bonds that rotate. The R groups (amino acid side chains) can take one of several states. In the case of proline there is only one free rotating bond (to the H) [28], this is handled by an error function that adds a large penalty to the optimization function such that the bond will stay at the optimal angle (Fig. 2b). Generally, all bonds are considered to be of fixed length and only rotation is allowed. The angles are the parameter. The volume and surfaces are the results. The simplex method requires only functions given by the objective function [41, 55]. The bond lengths never change, the only change occurs in the two angles per residue in our configuration search [85].

The rotation of non-collinear bonds allows the molecule to twist, similar to a Rubik's cube puzzle toy where a set of angles are joined by rotating joints. This rotation allows the protein to take a shape [56]. The molecule can be twisted to nearly any shape, but the proper shape is achieved by optimization of an objective function. The objective function mirrors an energy function.

The method used for the geometrical model is *trilateration*, which is a type of measurement that determines a point by using the geometry of spheres, circles, or triangles. Unlike triangulation, which uses the measurement of angles to determine

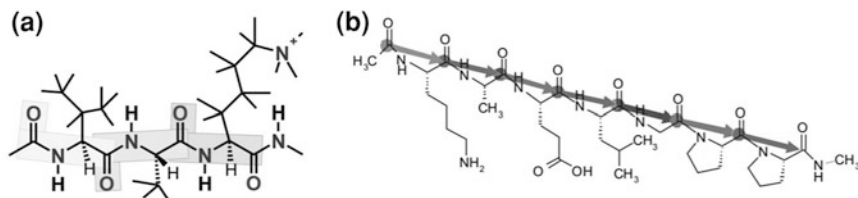


Fig. 2 **a** Beads are created by lacing the amino acid along the N-O backbone to construct an initial stage of the protein model. **b** Illustration of the optimized orientation given by the arrows of the minimization along the N-O backbone

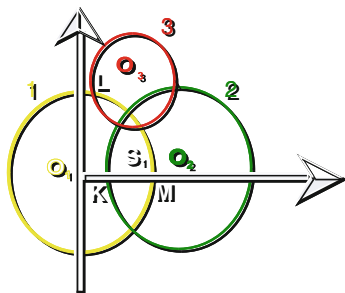


Fig. 3 Creating the surface for the bead building. The three overlapping/intersecting spheres with centers O_1 , O_2 and O_3 form a new object and the net area is given by $A = \left[\frac{1}{2} (A_{O_1} + A_{O_2} + A_{O_3}) \right] - \text{Internal Area}$

location, trilateration uses measures of distance. In two-dimensional geometry, the radii of circles can sometimes be used to find the location of a point. Hence, the surface is calculated from the intersection of the surfaces (atomic radii) of the atoms in the residue [64]. For one single sphere the volume is $V = \frac{4}{3}\pi r^3$ and surface is $S = 4\pi r^2$. For two spheres the volume is $V = V_1 - V_2 - V_L$ and the surface is $S = S_1 + S_2 - S_L$ (V_1, V_2 are volume one and volume two; S_1, S_2 are surface one and surface two; V_L, S_L are overlapping).

The model requires enough volume for the distances and surfaces. The surfaces are determined by allowing the 3 overlapping and intersecting spheres of centers O_1 , O_2 and O_3 to form a new object with defined volume. Its net area will be given by:

$$A = \left[\frac{1}{2} (A_{O_1} + A_{O_2} + A_{O_3}) \right] - \text{Internal Area}$$

where the internal area is given by (Fig. 3):

$$\text{Area}_{123} = \int_L^K \text{Area}(S_1) dS_1 + \int_K^M \text{Area}(S_1) dS_1 + \int_M^L \text{Area}(S_1) dS_1 \text{ and } 0 < K < M$$

We obtain the minimal free energy by minimizing the volume which we have accounted for by the surface areas as calculated above. Note that more than three overlapping spheres are improbable for atoms in molecules [20]. The objective function for the minimization is given by:

$$\begin{aligned} \text{Objective} = & \text{volume}(\text{volumeweighting}) \\ & - \sum \text{surfacearea}_{12}(\Phi, \Theta) \text{hydrophobicity}_1 \text{hydrophobicity}_2 \end{aligned}$$

The volume weights are proportional to the amount of energy needed to move the R from cyclohexane to water (0 is neutral, -1 is hydrophilic and 1 is hydrophobic), using the surface of the whole amino acid, rather than just the R group [65]. Each residue has a surface area with a known hydrophobicity [12]. The summation is over each set of residues that are touching (i.e. adjacent to each other). The surface area is the common surface area between the residues. This term will tend towards having hydrophobic residues together and hydrophilic ones together, but will avoid having hydrophilic next to hydrophobic residues.

There is also a volume term that minimizes the size of the molecule [11, 58]. This volume is given as the volume enclosed by the surface wetted by a solvent molecule with a $\sim 1.4 \text{ \AA}$ radius. The model is based on changing the distance and the angles and are adjusted to the distances. If we do not know the structure, we can calculate in parallel the α -coil and β -sheet simultaneously forming the braid used to obtain a globular or complex structure.

2.1 Constraints in the Standard Braid Theory

Prohibiting the braids from incidental intersection with themselves or with other braids is properly observed (rule of “no intersection”) in this application to keep the modeled peptide chains from overlapping each other (Appendix 2). The simple arc length model has been expanded to address the finite volume occupied by each amino acid residue [48]. While keeping the length and direction of the arc lengths constant, each segment is expanded into a *bead* enveloping the remainder of its amino acid residues. Each *bead* interacts with at most two other *beads*, and the intersection of any two sequential beads is a single point [39]. A *braid* now represents the peptide chain, which is a collection of *beads* (Appendix 3). Various peptide conformations can now be treated as changes in the relative orientation between pairs of *beads*. For large, single-chain proteins this is a significantly simplified approach to molecular modelling (Appendix 4; Appendix 5). The power of this approach is seen when considering protein structures composed of multiple

peptides, such as the collagen triple helix. The model presented in this section includes the collagen fibril as a three stranded *braid*.

2.2 Orientation of Protein Structures

Collagen is a 3_{10} helix which contains 3.6 residues per turn and has 10 atoms in the ring which is formed by making the hydrogen bond three residues up the chain [32]. The distance is determined by H bonds that lie parallel to the helix, and the carbonyl groups pointing along the axis in the opposite direction [66]. The opposite direction of nitrogen and the carbonyl will set the preference distance and define the α -helix. Since the direction is measured from the carbonyl, the distance between each turn is 3.6 residues [3, 26, 77, 86].

A protein β -sheet orientation is symmetric. The β -sheet is measured from the nitrogen terminal to the carbonyl terminal. The residue, carbonyl, and nitrogen are on the same side [5]. The inter-strand symmetric amide proton is the donor of the hydrogen bond to the carbonyl. For an anti-parallel orientation the exchange is perpendicular. This is not the case for a parallel orientation. The distance between residues is ~ 0.347 nm for anti-parallel and ~ 0.325 nm for parallel pleated sheet. Parallel β -sheets tend to be more regular than anti-parallel β -sheets [73]. The range of angles Φ and Θ angles for the peptide bonds in parallel sheets is much smaller than that for anti-parallel sheets [72]. Parallel sheets are typically large structures and sheets that are composed of less than five strands are rare [24]. Anti-parallel sheets however typically consist of only a few strands [50].

Parallel sheets characteristically distribute hydrophobic side chains on both sides of the sheet, while anti-parallel sheets are usually arranged with all their hydrophobic residues on one side of the sheet [57]. This required an alteration of hydrophilic and hydrophobic residues in the primary structure of peptides involved in anti-parallel β -sheets because alternate side chains project to the same side of the sheet [37].

In general the N–H and the C=O (each with an individual dipole moment) need to be in the same plane to create a large net dipole for the structure whether it is an α , β or 3_{10} structure [46].

3 Computer Program

A computer program was developed to implement area minimization [14, 41]. The program accepts Protein Data Base (PDB) files or just amino acid sequences as input.

Input Parameters The data from the PDB or the amino acid sequence are randomized to simulate pre-coiling of the structure. The data are *relaxed* [23, 42, 54] and flattened by randomly given the Θ and Φ bonds a twist [4]. All the

atoms stay in relative position inside of a bead [44]. The energy minimization is determined by the *downhill simplex method*, which is a heuristic search method and which converges to a non-stationary point. The *objective function technique* is implemented in multiple-space. The method uses a simplex which is a generalization of a triangle or tetrahedron and can be used for line segments or triangles [55].

Output The output consists of a molecular coordinate system data file that is exported to RasMol (Molecular Visualization, Freeware) molecular graphics visualization tool for export to a GIF three-dimensional image. Five amino acid sequences for five proteins, including collagen, were tested using the minimizer and compared to their published three-dimensional structures for confirmation. The deviation between the ProteoRubixTM model and the known model was determined.

4 Results

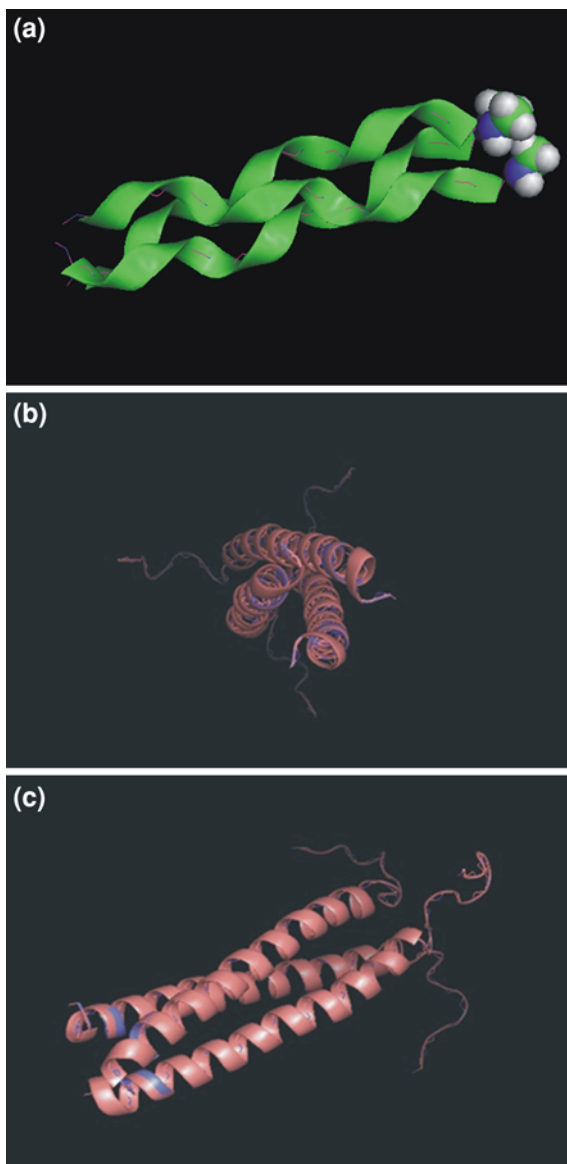
We considered the following examples of available Protein Data Base (PDB) files in three categories of proteins for the minimization process: 3_{10} coil (i.e. collagen consists of 3 coils and the coil turns every 10 amino acids); α -coil; and, globular structure protein. The minimization program was used to create three-dimensional structures of five proteins and was compared with data from the original protein crystallized structures. A correlation was achieved for each of the proteins as described below. For illustration purposes, *dumps* (i.e. the folding process export as data files for visualization) were added between the stages such that twisted regions *unfold*. The data from the original crystallized protein structures were compared to the data obtained after minimization. Different structural protein types were minimized by using the objective function and compared to known structures using a sequencer alignment.

4.1 3_{10} Coil Protein

For 1BBF glycoprotein, the source of the original reference structure was a theoretical model [56]. Figure 4a shows the matching superposition of the original file given by the green strand and the minimized construct shown by the purple backbone line along the green strand 1BBF-1. Figure 5 shows the sequence alignment given by Clustal-W. The sequences of the three coils are superimposed and differentiated by Coil (chain) A, Coil B and Coil C; note that the superposition shows the alignment between the minimized file and the original file. Figure 6 shows the local and global root mean square difference (RMSD) of the α -carbon and the back bone with distances in Angstroms in the structural alignment of the

Fig. 4 Correlation between known protein structures and the minimized files.

a Superposition of 1BBF collagen protein shown by the *green* is the known original and the *blue-green line* is minimized file showing excellent fit between the two models. **b** The *frontal view* of the superimposed files of original (*pink*) and the minimized (*purple*) models (with side chains) for 1AQ5 chicken cartilage matrix protein. Near-identical structural correlation was seen. **c** *Side view* of the superposition; of note are the tips of the coils that show the same alignment



superposition of the known and the modeled protein. The heavy atoms (one with a significantly higher atomic scattering factor than the others) are derivative structure factor amplitudes or changes along RMSD per molecule. The non-bonding energies used in order to minimize modelling of inter-molecular interactions can be integrated to the trilateration model [40].

Fig. 5 The Clustal-W sequence alignment denoted in *colour* show identical matched regions of alignment between the original known 1BBF collagen file and the “PDBB” minimized file

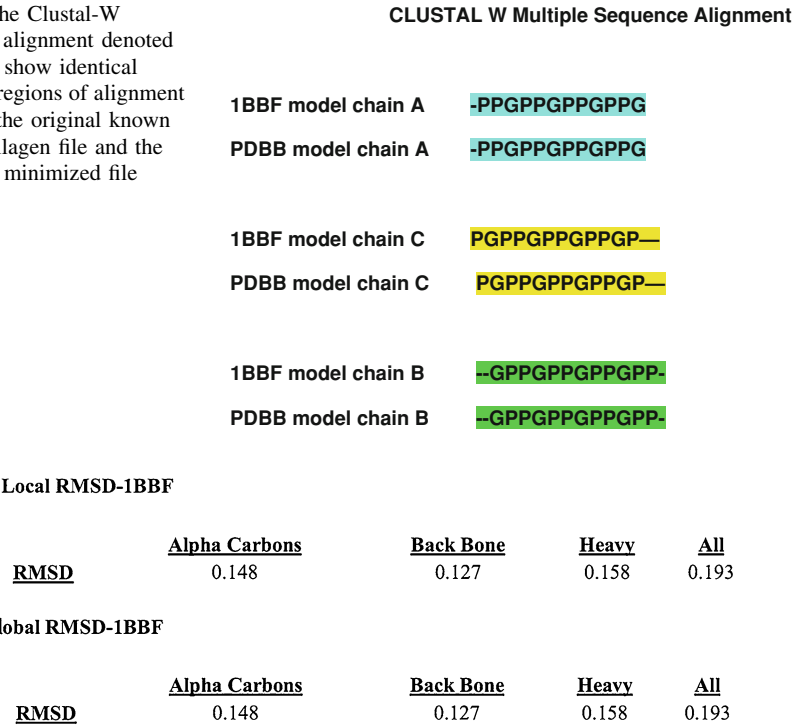


Fig. 6 Root Mean Square Difference (RMSD) resulting from the superposition of the 1BBF collagen and the minimized model. The RMSD considers local and global solutions for the C α , the back bone and the heavy portion of the atoms (i.e. every atom in the molecule, with exception of the hydrogen). The distances are in Angstroms

4.2 α -Coil Protein

1AQ5 is a coil–coil cartilage matrix protein of *Gallus gallus* (chicken). Based on NMR [84], Fig. 4b and c show frontal and side superposition of the known original file and the minimized file with excellent alignment. Figure 7 shows Clustal-W superposition of sequence of the two files with complete alignment. The adjoining coils are the sequence matched in the superposition and the others unmatched by Clustal-W. Figure 8 shows the local and global RMSD of the α -carbon and the back bone. The local minimization gives the RMSD per amino acid per carbon bone and global minimization gives the RMSD of the sequence. The RMSD compares atoms and the total number back bones of the PDBs. Figure 9 shows a graph of RMSD per amino acid.

CLUSTAL W multiple sequence alignment for 1AQ5

```
1AQ5_model_1_chain_A
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
1AQ5_model_1_chain_C
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
PDBB-AQ5_model_default_chain_A
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
PDBB-AQ5_model_default_chain_B
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
PDBB-AQ5_model_default_chain_C
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
1AQ5_model_1_chain_B
GSHMEEDPCECKSIVKFQTKVEELINTLQQKLEAVAKRIEALENKII
```

Fig. 7 The Clustal-W for the alignment between the original sequence of 1AQ5 and the minimized structure labeled as PDBB-1AQ5; *colors* identify each chain and matching of the files

Local RMSD-1AQ5				
	<u>Alpha Carbons</u>	<u>Back Bone</u>	<u>Heavy</u>	<u>All</u>
<u>RMSD</u>	0.02	0.03	0.03	0.12
<u>Atoms</u>	47	188	303	414
Global RMSD-1AQ5				
	<u>Alpha Carbons</u>	<u>Back Bone</u>	<u>Heavy</u>	<u>All</u>
<u>RMSD</u>	0.02	0.03	0.03	0.12
<u>Atoms</u>	47	188	303	414

Fig. 8 The RMSD calculations resulting from the superposition of the 1AQ5 and the minimized model. The RMSD considers local and global solutions for the C α , the back bone and the heavy portion of the atoms (i.e. every atom in the molecule, with exception of the hydrogen). The distances are in Angstroms

4.3 Globular Proteins

1CQD, a globular protein, is a hydrolase from *Zingiber officinale* (ginger). Choi et al. [19] studied the structure via X-ray diffraction. Figure 10a shows the superposition of the original known file (green) with the minimized file (purple). Figure 11 shows the Clustal-W superposition of the sequence of the two files with 1,294 total number of amino acids. The colors identify the individual chains and the similar strands that are shown together. The superimposed structures of the PDB and the minimize file match. Figure 12 shows the local and global RMSD of the α -carbon and back bone. The local gives the RMSD per amino acid per carbon bone and global gives the RMSD of the sequence. The RMSD compare atom and the total number back bones of the PDBs. Figure 13 shows the graph of RMSD per residue and number of residues.

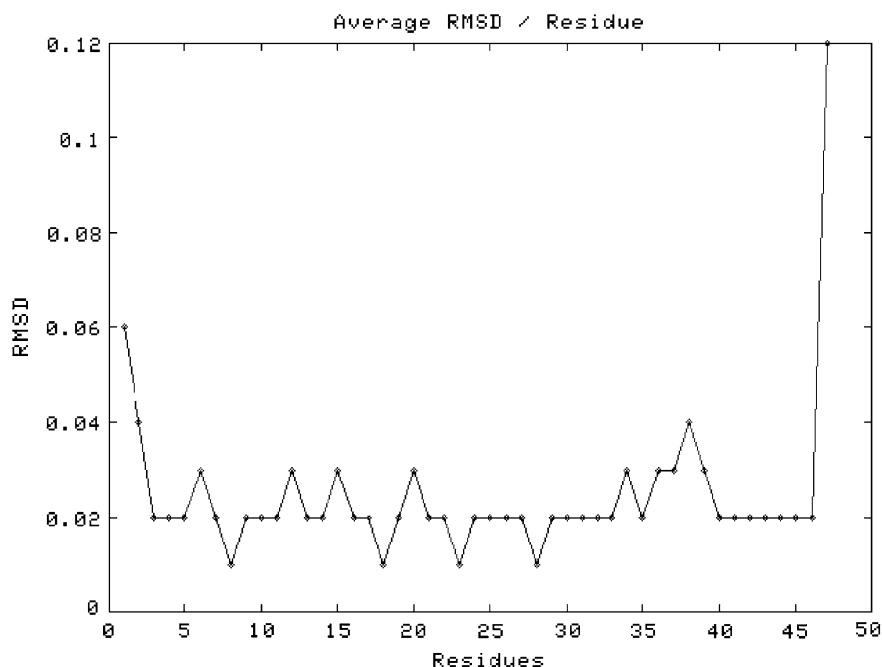
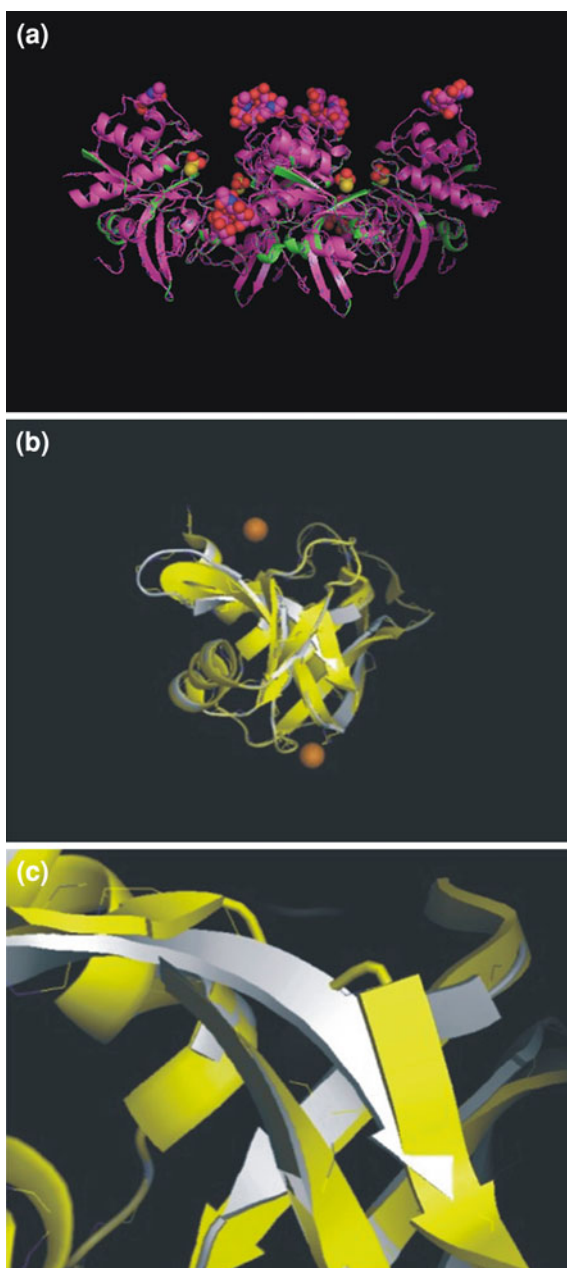


Fig. 9 The RMSD per residue, with the spikes showing the difference between the original file and the minimized file for 1AQ5

1AQP, a blood clotting protein from *Bos Taurus* (cattle), has been modeled in 1997 via X-ray diffraction [6]. Figure 10b shows the superposition of the known model and the minimized model. The grey color shows the superimposed model matching the original file. Figure 10c shows ribbons to depict the protein and the minor differences between the original file in yellow and the minimized file in gray; the distances of the differences are at ~ 9 Å. Figure 14 shows the local and global RMSD of the α -carbon and Back Bone. The local gives the RMSD per amino acid per carbon bone and the global gives the RMSD of the sequence of 124 residues.

To further validate our methods, we compared the predicted structures with the respective PDB entries using the alignment program STRAP [29, 33, 34] freely obtainable from <http://3d-alignment.eu/>. To compare two given 3D-structures of proteins, STRAP uses the method of 3D-superposition. STRAP implements several different back-ends for computation. Here TM-align [88] was used to perform a rigid superposition. TM-align moves one of the two models in space until both structures coincide as best as possible. There is a trade off between a low RMSD and a high percentage of assigned C-alpha positions. The result is a translation vector and rotation matrix. STRAP processes the result. It determines the RMSD which is a measure for the dissimilarity of two structures, derives a (multiple) sequence alignment and visualizes the result in Pymol [25].

Fig. 10 Correlation between known protein structures and the minimized files. **a** The superposition of known (green) and minimized (purple) models for 1CQD globular protein from ginger showing high match pattern between the two structures. **b** The superposition of the known original model (yellow) and the minimized model (grey) for 1AQP blood clotting protein from cattle, illustrating the matched regions and locations. **c** Zoomed image of the superposition at 9 Angstroms



The analysis was done using SuperPose software [51]. Minimizer output was in the format of PDB. STRAP software was used for the PDB alignment and Clustal-W software (Clustal-W: Multiple Alignment for DNA or Proteins, Freeware) [78] was

CLUSTAL W (1.83) multiple sequence alignment for 1CQD

```

PDBB_model_default_chain_A LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
PDBB_model_default_chain_B LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
PDBB_model_default_chain_C LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
1CQD_model_default_chain_D LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
PDBB_model_default_chain_D LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
1CQD_model_default_chain_C LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
1CQD_model_default_chain_A LPDSIDWRENGAVVPVKNQGGCGSCWAFSTVAAVEGINQIVTGDLSLSE
*****
PDBB_model_default_chain_A QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
PDBB_model_default_chain_B QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
PDBB_model_default_chain_C QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
1CQD_model_default_chain_D QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
PDBB_model_default_chain_D QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
1CQD_model_default_chain_C QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
1CQD_model_default_chain_A QQLVDCCTANHGCRGGWMNPAFQFIVNNGGINSEETYPYRGQDGICNSTV
*****
PDBB_model_default_chain_A NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
PDBB_model_default_chain_B NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
PDBB_model_default_chain_C NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
1CQD_model_default_chain_D NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
PDBB_model_default_chain_D NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
1CQD_model_default_chain_C NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
1CQD_model_default_chain_A NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
*****
1CQD_model_default_chain_B NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
1CQD_model_default_chain_A NAPVVSIDSYENVP SHNEQSLQKAVANQPVSVTMDAAGRDFQLYRSGIFT
*****
PDBB_model_default_chain_A GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
PDBB_model_default_chain_B GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
PDBB_model_default_chain_C GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
1CQD_model_default_chain_D GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
PDBB_model_default_chain_D GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
1CQD_model_default_chain_C GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
1CQD_model_default_chain_A GSCNISANHALTVVGYGTENDKDFWIVKNSWGKNWGSGGYIRAERNIENP
*****
PDBB_model_default_chain_A DGKCGITRFASYPVKK
PDBB_model_default_chain_B DGKCGITRFASYPVKK
PDBB_model_default_chain_C DGKCGITRFASYPVKK
1CQD_model_default_chain_D DGKCGITRFASYPVKK
PDBB_model_default_chain_D DGKCGITRFASYPVKK
1CQD_model_default_chain_C DGKCGITRFASYPVKK
1CQD_model_default_chain_B DGKCGITRFASYPVKK
1CQD_model_default_chain_A DGKCGITRFASYPVKK

```

Fig. 11 The Clustal-W file chains are identified by *colors* and show alignment matches between 1CQD (control file) and PDBB (minimized output)

used for the sequence between the known structure and the minimized structure. The STRAP software followed the structural trace through to the C-termini and takes the C-termini of both files to compare. No regions were ignored or skipped (i.e., even loops were carefully considered and aligned). The alignment was constructed with the primary aim of maximizing the aligned positions between structures, provided that there was a rational basis for the alignment. Excellent alignment was achieved.

5 Discussion

The implementation of the *ab initio* model as a geometry solution was considered since other approaches for protein modelling were fraught with difficulty due mainly to insufficient computation resources. The complexity of the protein

Local RMSD -1CQD

<u>RMSD</u>	<u>Alpha Carbons</u>	<u>Back Bone</u>	<u>Heavy</u>	<u>All</u>
	0.01	0.01	0.21	0.21
<u>Atoms</u>	216	864	1294	1294

Global RMSD -1CQD

<u>RMSD</u>	<u>Alpha Carbons</u>	<u>Back Bone</u>	<u>Heavy</u>	<u>All</u>
	0.01	0.01	0.21	0.21
<u>Atoms</u>	216	864	1294	1294

Fig. 12 The RMSD calculations for superposition of the 1CQD with the minimized model. The RMSD considers local and global solutions for the C α , the back bone and the heavy portion of the atoms. The number of residues that were considered was 1,294

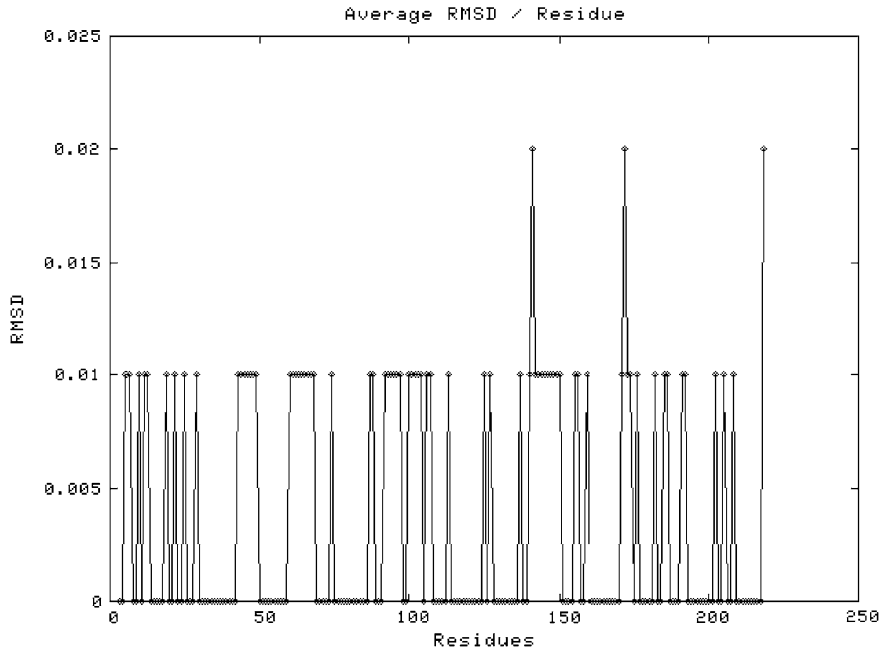


Fig. 13 RMSD per residue. The *spikes* show the difference between the original file and the minimize file for 1CQD

folding, and also the numbers of interactions needed to give a solution, leads to the NP problem. Furthermore, the actual environmental conditions such as interaction of protein in an aqueous environment increase the complexity of this challenge.

Local RMSD-1AQP

<u>RMSD</u>	<u>Alpha Carbons</u> 0.01	<u>Back Bone</u> 0.01	<u>Heavy</u> 0.21	<u>All</u> 0.21
<u>Atoms</u>	124	496	785	929
<u>Structure</u> PDBA PDBB		<u>Residues</u> 1-124 1-124		

Global RMSD-1AQP

<u>RMSD</u>	<u>Alpha Carbons</u> 0.01	<u>Back Bone</u> 0.01	<u>Heavy</u> 0.02	<u>All</u> 0.01
<u>Atoms</u>	124	496	785	929
<u>Structure</u> PDBA PDBB		<u>Residues</u> 1-124 1-124		

Fig. 14 The RMSD calculations from the superposition of the 1AQP with the minimized model for 929 residues

Based on our alternative approach, we were able to achieve accurate models for known proteins based on PDB files or amino acid sequences. The alignments were shown between the modeled and known structures via Clustal-W. The alignment differences given by the standard deviation (RMSD) showed minimal differences.

To summarize, the peptide *braid* provided the mathematical foundation that described the model and the three dimensional structure. The union arc-lengths for the model were given by the dipole moment (plane) of the N–H and the C=O. The differential geometry was used to create the algorithm for the projections in the rotations of the angles and the surface of the structure.

The *minimizer* calculated complex structures based the interactions between each of the 20 types of residues. This only gives 210 possible interactions and includes the interactions between the solute with the sodium or other charged ions. The residue interaction was obtained by assigning properties to the surface charges and hydrophobicity.

For the examples considered, we used Clustal-W and STRAPS for the alignment and RMSD for the verification. The simulations assumed residue interactions but ignored charges and sulfur bridges that are not systematic (all high by an eV or two). To make the model more physiological, we added the solute component, and assumed that the structure was formed in plain water with no ionic charge, and a size of $\sim 1.8 \text{ \AA}$. As an example, if we add a charged item to the solute, this did influence charge attraction by shielding the charged residues. If the charged species interacted at the site of hydrogen or other ions, the ionic size would alter the

protein size and additional terms would be required to account for this based on species size. These changes can readily be accomplished within the software and hence numerous further applications are possible.

One limitation that we encountered with our approach was an objective function that required readjustment when distances approached gapping in the protein structure that were over 10 %. A solution for this would be to add the functionality of AMBER software (Assistance Model Building with Energy Refinement) where the energy for the distance is predicted and can therefore control or refine the distance [16, 18, 59, 62]. The added approach does not consider the atoms, but simply the distance, and can deal with large or small sequences.

6 Conclusion

We present a new model to solve three-dimensional modelling of proteins that can create supra structures within tissues. By considering amino acid peptides as beads, hydrogen bond distances, the surface and distance geometry as functional components, and by focusing on folding, the Ab initio protocols can successfully model complex proteins in three dimensions with minimal computational load.

Acknowledgments The authors thank: Isabella Verdinelli and Lauren Ernst for the discussions related to the model and Troy Wymore from the Pittsburgh Supercomputing Center for his suggestions in verifying the models and a special thanks to Dr. Alex Cohen from ProteoRubix Systems who developed the minimizer and for discussion and development of ProteoRubix™ software for the Geometry Modelling. Chris Holm for editing the manuscript.

Appendix 1

NP

Suppose the bead involves d dihedral angles. Let $\phi^* = (\phi_1^*, \dots, \phi_n^*) \in [0, 360]^d$ be an optimal solution to the constrained optimization problem

$$(1.1) \quad \min_{\phi \in [0, 360]^d} \{v(\phi : \phi \text{ is a rotation about the bond } i)\}$$

Then there is a maximal number $n > 0$.

(1.2) The p^n hard problem of an exhaustive search over the angles $0 \leq \phi_i^1 < \dots < \phi_i^p \leq 360$ to find an approximate optimizer $\bar{\phi}$ to ϕ^* may be possible for a modern computer.

(1.3) There is exactly one solution to (4) in $|\bar{\phi}_i - p, \bar{\phi}_i + p|$ which would be ϕ^* and can be approximated using a given constrained optimization algorithm (**B 1**).

The convergence ball for the constrained optimization algorithm provides a candidate for p in the proposition. Using this proposition, we can obtain an acceptable initial condition for a constrained optimization algorithm.



Fig. 15 The sequence $[(a, b)]$ is length of a single coil) gives an approximation to the polygon arc P ; the length between two points (a, b) , where D is segments of arc-length

Appendix 2

Boundary Determination to Prevent Overlap

A path is a one-dimensional sub-manifold M of \mathbb{R}^3 , so that, for any point $x \in M$ there is a local parameterization near x . $C^k (k \geq 2)$ denotes the curvature of the path and D denotes the coordinates identifying the path. The output of each iteration is a set of coordinates in three dimensions, $(D = x_1, x_2, \dots, x_n)$ identifying a path. We denote by *length bond* is the polygonal arc around the path (Fig. 15). The curvature C^k and the arc-length are non-regular. Let $x = x(t)$, with $a \leq t \leq b$ and consider a partition [15]:

$a = t_0 < t_1 < \dots < t_n = b$, of an interval (a, b) .

The sequence (a, b) are the boundaries of a single coil) gives an approximation to the polygon arc C . As illustrated the length between two points (a, b) , where D are segments of arc-length given by:

$$\lambda(D) = \sum_{j=1}^n D_j = \sum_{i=1}^n \|x_i - x_{i-1}\| = \sum_{i=1}^n \|x(t_i) - x(t_{i-1})\| \quad (2.1)$$

The arc-length can be bounded from above and from below. The upper bound is given by:

$$\rho_+(K, D) = \frac{1}{\lambda(D)} \sum_{(K \circ D_j) \cap D \neq \emptyset} \lambda(K \circ D_j) \quad (2.2)$$

And the lower bound is:

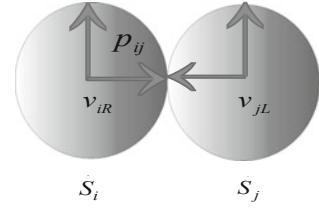
$$\rho_-(K, D) = \frac{1}{\lambda(D)} \sum_{(K \circ D_j) \subset D} \lambda(K \circ D_j) \quad (2.3)$$

where $\rho_+(K, D)$ is the ratio of the total measure of the set in the system K (is the volume minimization) so that the transformation \circ (*projection*) of the segments and the curve C give the lower and the upper bound (a, b) .

$$b = \rho_+ = \lim_{\lambda(D) \rightarrow \infty} \sup \rho_+(K, D) = \lim_{\lambda \rightarrow \infty} \sup_{\lambda(D) \geq \lambda} \rho_+(K, D) \quad (2.4)$$

$$a = \rho_- = \lim_{\lambda(D) \rightarrow \infty} \inf \rho_-(K, D) = \lim_{\lambda \rightarrow \infty} \inf_{\lambda(D) \geq \lambda} \rho_-(K, D) \quad (2.5)$$

Fig. 16 The radii are chosen so that the intersection of the closure of any two beads \bar{S}_i and \bar{S}_j is a single point p_{ij} . The point p_{ij} is the origin of a right and a left vector v_{iR}, v_{jL}



Hence the boundaries of C are given in (2.4) and in (2.5).

Appendix 3

The geometry structure of the protein is defined by a braid (see **B 2** for a description of a chain as a collection of *beads* forming a *braid*). The j th molecule of the chain is fitted to a conveniently shaped open *bead* S_j (see **B 3**) with its center located at the center of the *bead* and the radius r_i has size such that the i th *bead* does not overlap with the j th *bead* when $i \neq j$.

The radii in Fig. 16 r_i are chosen so that the intersection of the closure of any two beads \bar{S}_i and \bar{S}_j is a single point p_{ij} , (see **B 3**). The point p_{ij} is the origin of a right and a left vector v_{iR}, v_{jL} . In this process it is important to translate (projection) and rotate these vectors. The mathematics of this construction justifies geometry of the bead construction.

Appendix 4

With our model of collagen in mind, we next introduced the concept of the *braid* group. The *braid* was defined as the union of the backbones creating a string representing the amino acids. The collagen has three strands (as a group) or *coils* and each strand has a back bone, represented as the union of all points $x(t_i - 1, t_i)$ that are generated:

$$\text{Bonds} = \bigcup_{n=1}^N \{x(t_i - 1, t_i)\} \quad (4.1)$$

A braid is a collection of *beads* for which two operators $(\circ, =)$ can be defined. The *bead* in the collection can be *projected* using least of the squares. Let B denote this collection of *beads*, so $B = (\text{braids})$, and (B, \circ) is a group. We are checking the segments of the radius of bead of a single *braid*. The *enclosed volume* shrinks driven by minimization and through the homoeopathy is guaranty [2] (see **B 5**). We are modelling three coils, and their geometrical configuration has an equivalence class denoted by σ_i and σ_i^{-1} . A *braid* is equivalent and it is called *isotope* if the three coils cannot pass each other or themselves without intersecting [8] Fig. 17.

Fig. 17 A *braid* is equivalent and it is called *isotope* if the three coils cannot pass each other or themselves without intersecting

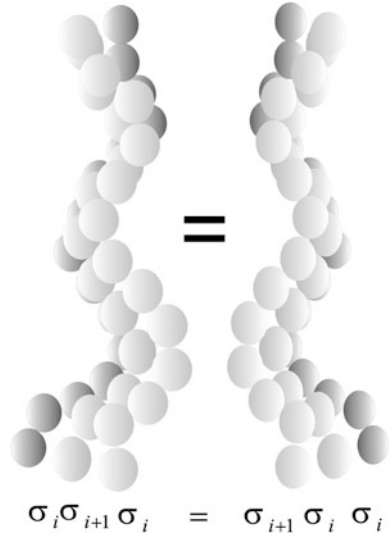
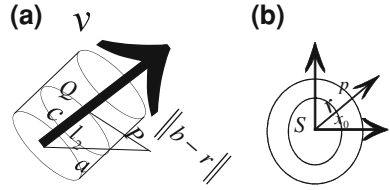


Fig. 18 The projection of the rotation of the rotation angles. The *enclosed volume* shrinks by the minimization and via homoeopathy which is guaranteed



$$\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1} \text{ if } 1 \leq i \leq n-2 \text{ [1]}$$

Appendix 5

The distances of the projection to P is given by $\|b - r\|$, where $v(x - p) = 0$ and by Pythagorean gives us that $b^2 = c^2 - a^2$, where, $a^2 = \left\| \frac{b}{\|b\|_2} \times (Q - P) \right\|_2^2$, $c^2 = \|Q - P\|_2^2$ or $\frac{b(Q-P)^2}{\|v\|_2}$ shown in Fig. 18a.

B 1

Let $\bar{\phi}$ the solution to $\phi^* = (\phi_1^*, \dots, \phi_n^*) \in [0, 360]^d$ and d dihedral angle, $\phi_n^* = \sum^* \rightarrow N$, $1 \leq n \leq k$; Let q, r be polynomial such $\phi_n^*(I) \leq q(|I|)$, where I is

the instance of the angle in our problem. Then test instance construction system for all the angles of our problem (*TICA*), then $P = NP$.

Conversion We know that $\phi^* = (\phi_1^*, \dots, \phi_n^*) \in [0, 360]^d$ is the optimal solution, where d dihedral angles then $\delta = \frac{1}{2} \min_{\phi \in [0, 360]^d} \{v(\phi : \phi \text{ is a rotation about the bond } i)\}$ n is the maximum number of angles, $n > 0$ and $\delta > 0$. Let $\epsilon > 0$ be given. Where ϕ^* is continuous, there is a point $p \in \phi^*$, $\phi^* \leq \frac{1}{2} \phi(p)$ where implies $|\overline{\phi_i} - p, \overline{\phi_i} + p| < \epsilon$ and $v(p) \leq \frac{1}{2} \phi(p)$, we have $|\overline{\phi_i} - p, \overline{\phi_i} + p| \leq \phi^* + |v(p)| < \delta + \frac{1}{2} \phi(p) < \epsilon$

Uniqueness using the existence and uniqueness theorem, we know that ϕ^* is continuous, in the interval $|\overline{\phi_i} - p, \overline{\phi_i} + p|$ then converges.

B 2

D is said to be covering itself if $\bigcup_j D_j \supset D$ and each elements of at least one of D belongs to d_j . The system D_j is packing if $D_i \cap D_j = \emptyset (i \neq j)$, $\bigcup_j D_j \supset D$

If two sets D_1, D_2, \dots have the same elements in common then each element D_1, D_2, \dots belong to D .

B 3

Each segment can be treated as open *beads*, as such the coordinates belong to a set X and for any point $p \in D_j$ and $\delta = D_j$ where the measure is positive.

So, the definition of the *bead* is:

$$D = \{x: d(p, x) < \delta\}$$

B 4

Let A and B be a disjoint convex set in a convex space, then

$A = \{x: (x - D_i)^2 < r_i\}$ and $B = \{x: (x - D_j)^2 < r_j\}$, the distance is given by:

$\text{dis}(D_i, D_j) = r_i + r_j$. The closure of B is given by $\overline{B} = \{x: (x - D_j)^2 \leq r_j\}$ then

$A \cap \overline{B} = \emptyset$.

A is an open set by construction. A & B are the convex hull, also by construction, and then:

$\exists l(x) = a$ if

$$x \in A \quad l(x) \leq a$$

$$x \in B \quad l(x) \geq a$$

where a is

$$v_j = (a - D_j)^2$$

$$v_i = (a - D_i)^2$$

where

$$D_i = \text{dist}(a - v_{Ri}) \text{ and } D_j = \text{dist}(a - v_{Lj})$$

B 5

Let $x \in S(r, x_0)$, $S \in \mathbb{R}^n$ and $x_0 \neq 0$ i.e. $p(x) = x_0 + r \frac{x}{\|x\|}$ (Fig. 18b) then; $r = \|p - x_0\| = \left\| x_0 - r \frac{x}{\|x\|} - x_0 \right\| = \frac{r}{\|x\|} \|x\| = r$ hence $p \in S(r, x_0)$ [2].

References

1. Adams, C., Hildebrand, M., Weeks, J.: Hyperbolic invariants of knots and links. *Trans. Amer. Math. Soc.* **1**, 1–56 (1991)
2. Adams, J.F.: Vector fields on spheres. *Ann. Math.* **75**, 603–632 (1962)
3. Andrew, C.D., Penel, S., Jones, G.R., Doig, A.J.: Stabilizing nonpolar/polar side-chain interactions in the α -helix. *Proteins Struct. Funct. Genetics* **45**, 449–455 (2001)
4. Arge, E., Bruaset, A.M., Langtangen, H.P.: *Modern Software tools for Scientific Computing*. Birkhauser Press, Boston (1997)
5. Baker, E.N., Hubbard, R.E.: Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **179**, 97–177 (1984)
6. Balakrishnan, R., Ramasubu, N., Varughese, K., Parthasathy, R.: Crystal structure of the cooper and nickel complexes of RNAase A: metal-induced interprotein interactions and identification of a novel cooper binding motif. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 9620–9625 (1997)
7. Biegler, T.F., Mumenthaler, C., Wener, B.: Folding proteins by energy minimization and Montecarlo simulations with hydrophobic surface area potentials. *J. Mol. Model.* **1**, 1–10 (1995)
8. Birman, J.S.: Recent developments in Braid and Link theory. *Math. Intell.* **13**, 52–60 (1991)
9. Blaney, J.M., Dixon, J.S.: *Distance Geometry in Molecular Modelling*. CRC Press, Boca Raton (1993)
10. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E., Baker, D.: Rosetta in CASPA 4: progress in ab initio protein structure prediction. *Proteins* **5**, 119–126 (2001)
11. Bondi, A.: van der Waals volumes and radii. *J. Phys. Chem.* **68**, 441–451 (1964)
12. Bryngelson, J.D., Billing, E.M.: Interatomic interactions to protein structure. *Rev. Comput. Chem.* **5**, 84 (1995)
13. Bussemaker, H.J., Thirumalia, D., Bhattacharjee, J.K.: Thermodynamic stability of folding protein against mutation. *Phys. Rev. Lett.* **79**, 3530–3533 (1997)
14. Campbell, P.G., Cohen, A.P., Ernst, L.A., Ernsthansen, J., Farkas, D.L., Galbraith, W., Israelowitz, M.: US Patent Application. USPTO Patent Application Number 20030216867 (2003)

15. Carmo, M.P.: Differential Geometry of Curves and surfaces. Prentice-Hall, Englewood Cliffs (1976)
16. Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005)
17. Chen, J.M., Kung, C.E., Fearheller, S.H., Brown, E.M.: An energetic evaluation of a “Smith” collagen microfil model. *J. Protein Chem.* **10**(5), 535–551 (1991)
18. Cheatham III, T.E., Young, M.A.: Molecular dynamics simulation of nucleic acids: successes, limitations and promise. *Biopolymers* **56**, 232–256 (2001)
19. Choi, H.K., Laursen, R.A., Allen, K.N.: The 2.1 Å structure of a cysteine protease with proline specificity from ginger rhizome, *Zingiber officinale*. *Biochemistry* **38**, 11624–11633 (1999)
20. Connolly, M.L.: Computation of molecular volume. *J. Am. Chem. Soc.* **107**, 1118–1124 (1985)
21. Connolly, M.L.: Adjoin volumes. *J. Math. Chem.* **15**, 339–352 (1994)
22. Costantini, S., Colonna, G., Facchiano, A.M.: Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.* **342**, 441–451 (2006)
23. Cramer, C.J.: Essentials of computational chemistry: theories and models. Wiley, West Sussex (2004)
24. Das, A.K., Cohen, P.W., Barford, D.: The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein–protein interactions. *EMBO J.* **17**, 1192–1199 (1998)
25. DeLano, W.L.: The PyMOL molecular graphics system on World Wide Web (2002). <http://www.pymol.org>
26. Eaton, W.A., Munoz, V., Thompson, P.A., Henry, E.R., Hofrichter, J.: Kinetics and dynamics loops, α -helices, β -hairpins, and fast-folding proteins. *Acc. Chem. Res.* **31**, 745–753 (1998)
27. Ege, S.: Organic Chemistry, pp. 18–71. D.C. Heath and Company, Lexington (1984)
28. Fernandez, A., Sinanoglu, O.: Denaturation of proteins in methanol/water mixtures. *Biophys. Chem.* **21**, 163–164 (1985)
29. Frömmel, C., Gille, C., Goede, A., Gröpl, C., Hougardy, S., Nierhoff, T., Preissner, R., Thimm, M.: Accelerating screening of 3D protein data with a graph theoretical approach. *Bioinformatics* **19**, 2442–2447 (2003)
30. Garnier, J., Osguthorpe, D.J., Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–112 (1978)
31. Garrett, R., Grisham, C.M.: Biochemistry, p. 150. Brooks/Cole, Belmont (2005)
32. Gibson, K.D., Scheraga, H.A.: An algorithm for packing multistrand polypeptide structures by energy minimization. *J. Comput. Chem.* **15**, 1414–1428 (1994)
33. Gille, C., Lorenzen, S., Michalsky, E., Frömmel, C.: KISS for STRAP: user extensions for a protein alignment editor. *Bioinformatics* **19**, 2489–2491 (2003)
34. Gille, C.: Structural interpretation of mutations and SPNs using STRAP-NT. *Protein Sci.* **15**, 208–210 (2006)
35. Gong, H., Porter, L.L., Rose, G.: Counting peptide-water hydrogen bonds in unfolded proteins. *Protein Sci.* **574**, 417–427 (2011)
36. Havel, T.F.: An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* **56**, 43–78 (1991)
37. Hill, B.R., Raleigh, D.P., Lombardi, A., Degrado, W.F.: De novo design of helical bundles as models for understanding protein folding and function. *Acc. Chem. Res.* **33**, 745–754 (2000)
38. Hummer, G., Garde, S., García, A.E., Paulaitis, M.E., Pratt, L.R.: The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1522–1555 (1998)
39. Hunt, A.J., Gittes, F., Howard, J.: The force exerted by a kinesin molecule against a viscous load. *Biophys. J.* **67**, 766–781 (1994)

40. Israelowitz, M., Rizvi, S.W.H., Kramer, J., von Schroeder, H.P.: Computational modelling of type I Collagen fibers to determine the extracellular matrix structure of connective tissues. *Protein Eng. Des. Sel.* **18**, 329–335 (2005)
41. Jacoby, S.L.S., Kowalik, J.S., Pizzo, J.T.: Interactive methods for nonlinear optimization problems. Prentice-Hall, Englewood Cliffs (1972)
42. Tang, C.J.K., Alexandrov, V.: Relaxed Monte Carlo linear solver. In: Alexandrov, V.N., Dongarra, J.J., Juliano, B.A., Tan, C.J.K. (eds.) *Lecture Notes in Computer Science*, vol. 2073, p. 1289. Springer, Heidelberg (2001)
43. King, G., Brown, E.M., Chen, J.M.: Computer model of a bovine type I collagen microfibril. *Protein Eng.* **1**, 43–49 (1996)
44. Kuntz, I.D., Thomason, J.F., Oshiro, C.M.: Distance geometry. In: Openheimer N.J., James T.L. (eds.) *Methods in Enzymology*, vol. 177, pp. 159–204. Academic press, New York (1989)
45. Lyngsø, R.B., Pedersen, C.N.: RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* **7**, 409–427 (2000)
46. Liu, W., Chou, K.: Prediction of protein secondary structure content. *Protein Eng.* **12**, 1041–1050 (1999)
47. Macdonald, J.R., Johnson Jr, W.C.: Enviromental features are important in determining protein secondary structure. *Protein Sci.* **10**, 1172–1177 (2001)
48. Maritan, A., Micheletti, C., Triovato, A., Banava, J.B.: Optimal shapes of compact strings. *Nature* **406**, 287–290 (2000)
49. Marashi, S.A., Behrouzi, R., Pezeshk, H.: Adaptation of proteins to different environments: a comparison of proteome structural properties in *Bacillus subtilis* and *Escherichia coli*. *J. Theor. Biol.* **244**(1), 127–132 (2007)
50. MacCallum, P.H., Poet, R., Milner-White, J.E.: Coulombic interaction between partially charged main-chain atoms hydrogen-bonded to each other influence the confirmations of α -helices and antiparallel β -sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the coulombic interactions. *J. Mol. Biol.* **248**, 361–373 (1995)
51. Maiti, R., von Domselear, G.H., Zang H., Wisshart DS.: Super pose: a simple server sophisticated structural superposition. *Nucleic Acid Res.* **32**, W590–W594 (2004)
52. Mount, D.M.: Bioinformatics: Sequence and Genome Analysis, vol. 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2004)
53. More, J.J., Wu, Z.: Issues in large scale global minimization. In: Biegler, L.T., Coleman, T. F., Conn, A.R., Santosa, F.N. (eds.) *Large-scale optimization with applications*, Part III, p. 99. Springer-Verlag, New York (1997)
54. Morgan, D., Ceder, G., Curtarolo, S.: High-throughput and data mining with ab initio methods. *Meas. Sci. Technol.* **16**, 296–301 (2005)
55. Nelder, J.A., Mead, R.: A simplex method for function minimization. *Comput. J.* **7**, 308–313 (1965)
56. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A.: Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* **96**, 6472–6484 (1992)
57. Nicholls, A., Sharp, K.A., Honig, B.: Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991)
58. Pavanì, R., Ranghino, G.: A method to compute the volume of a molecule. *Comput. Chem.* **6**, 133–135 (1982)
59. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.R., Cheatham III, T.E., DeBolt, S., Ferguson, D., Seib, G., Kollman, P.: AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comput. Phys. Commun.* **91**, 1–41 (1995)
60. Pham, T.H., Satou, K., Ho, T.B.: Support vector machines for prediction and analysis of beta and gamma-turns in proteins. *J. Bioinform. Comput. Biol.* **3**, 343–358 (2005)

61. Polanowska-Grabowska, R., Simon Jr, C.G., Shabanowitz, J., Hunt, D.F., Gear, A.R.L.: Platelet adhesion to collagen under flow causes dissociation of a phosphoprotein complex of heat-shock proteins and protein phosphatase 1. *Blood* **90**, 1516–1526 (1997)
62. Ponder, J.W., Case, D.A.: Force fields for protein simulations. *Adv. Protein Chem.* **66**, 27–85 (2003)
63. Ochsebein, F., Gilquin, B.: NMR for protein analysis. *CLEFS CEA* **56**, 52–55 (2008)
64. Ott, R., Bijma, J., Hemleben, C.: A computer method for estimating volume and surface areas of complex structure consisting of overlapping spheres **16**, 83–98 (1992)
65. Radzicka, A., Wolfeden, R.: Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**, 1664–1670 (1988)
66. Ramachandra, G.N., Ramakrishnan, C., Sasisekharan, V.: Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963)
67. Raspanti, M.: Different architectures of collagen fibrils enforce different fibrillogenesis mechanisms. *J. Biomed. Sci. Eng.* **3**, 1169–1174 (2010)
68. Rojnuckarin, A., Santae, K., Shankar, S.: Brownian dynamics simulations of protein folding: Access to milliseconds time scale and beyond. *PNAS* **68**, 4288–4292 (1998)
69. Roux, B.: Perspective in molecular dynamics and computational method. *J. Cell Biol.* **135**, 547–548 (2010)
70. Sanjeev, A., Barak, B.: *Computational Complexity: A Modern Approach*, pp. 50–59. Cambridge University Press, Cambridge (2009)
71. Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P.I.W., Wittenstein, E., Sali, A.: MODBase, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **28**, 250–253 (2000)
72. Sharman, G.J., Searle, M.S.: Cooperative interaction between the three strands of a designed antiparallel β -sheet. *J. Am. Chem. Soc.* **120**, 5291–5300 (1998)
73. Schenck, H.L., Gelman, S.H.: Use of a designed tripled-stranded antiparallel β -sheet to probe β -sheet cooperativity in aqueous solution. *J. Am. Chem. Soc.* **120**, 4869–4870 (1998)
74. Shakhnovich, E.I., Farztdinov, G., Gutin, A.M., Karplus, M.: Protein folding bottlenecks: a lattice Monte Carlo simulation. *Phys. Rev. Lett.* **67**, 1665–1668 (1991)
75. Shreraga, H., Gibson, K.D.: An algorithm for packing regular multistrand polypeptide structures by energy minimization. *J. Comput. Chem.* **15**, 1414–1428 (1994)
76. Sinaglou, O.: Microscopic surface tension down to molecular dimensions and microthermodynamic surface areas of molecules or clusters. *J. Chem. Phys.* **1**, 463–468 (1981)
77. Srinivasan, R.: Helix length distribution from protein crystallographic data. *Indian J. Biochem. Biophys.* **13**, 192–193 (1976)
78. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994)
79. Ting, C.-K.: On the mean convergence time of multi-parent genetic algorithms without selection. In: Freitas, A.A., Bentley, P.J., Johnson, C.G., Timmis, J. (eds.) *Advances in artificial life*, p. 403 Springer-Verlag, Berlin (2005)
80. Tiraboschi, G., Gresh, N., Giessner-Pretre, C., Pedersen, L.G., Deerfield II, D.W.: A joint ab initio and molecular mechanics investigation of polycoordinated Zn(II) complexes with model hard and soft ligands. Variations of the binding energy and of its components with the number and the charges of the ligands. *J. Comput. Chem.* **21**, 1011–1039 (2000)
81. Torda, A.E., Van Gunsteren, W.F.: Molecular modelling using nuclear magnetic resonance data. In: Lipkowitz, K.B., Boyd, D.B. (eds.) *Reviews in Computational Chemistry*, vol 3, pp. 143–172. VCH Publishers, New York (1992)
82. Voelz, V.A., Bowman, G.R., Beauchamp, K., Pande, V.S.: Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1–39). *J. Am. Chem.* **132**, 1526–1528 (2010)
83. Wade, L.G.: *Structure and Stereochemistry of Alkanes: Organic Chemistry*, 6th edn, pp. 103–122. Pearson Prentice Hall, Upper Saddle River (2006)

84. Wiltschek, R., Kammerer, R.A., Dames, S.A., Schulthess, T., Blommers, M.J., Engel, J., Alexandrescu, A.T.: NMR assignments and secondary structure of the coiled coil trimerization domain from cartilage matrix protein in oxidized and reduced forms. *Protein Sci.* **6**, 1734–1745 (1997)
85. William, P., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes in Fortran*, 2nd edn., pp. 312–326. Cambridge University Press, Cambridge (1992)
86. Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., Richardson, D.C.: Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogens. *J. Mol. Biol.* **285**, 1711–1733 (1999)
87. Zimmermann, O., Hansmann, U.H.: Support vector machines for prediction of dihedral angle regions. *Bioinformatics* **22**, 3009–3015 (2006)
88. Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**(7), 2302–2309 (2005)
89. Zhang, X.Q., Jansem, A.P.: Kinetic Monte Carlo method for simulating reactions in Solution. *Phy. Rev. E Stat. Nonlin. Softw. Matter Phys.* **82**, 046704 (2010)
90. Zhong, L., Johnson Jr, W.C.: Environment affects amino acid preference for secondary structure. *PNAS* **89**, 4462–4465 (1992)



<http://www.springer.com/978-3-642-32562-5>

Computational Modeling in Tissue Engineering

Geris, L. (Ed.)

2013, XII, 444 p., Hardcover

ISBN: 978-3-642-32562-5