Computational modeling of type I collagen fibers to determine the extracellular matrix structure of connective tissues

Meir Israelowitz1, Syed W.H. Rizvi1, James Kramer2 and Herbert P. von Schroeder1,3,4

1Biomimetics Technologies, Unit 101, 6 Fernwood Gardens, Toronto, Ontario M4K 2J9, Canada, 2Pittsburgh Safety and Health, Technology Center, Cochrans Mill Road, Pittsburgh, PA 15236, USA and 3University of Toronto, Toronto Western Hospital, 399 Bathurst Street, Toronto, Ontario M5T 2S8, Canada

A method is presented for generating computer models of biological tissues. The method uses properties of extracellular matrix proteins to predict the structure and physical chemistry of the elements that make up the tissue. The method begins with Protein Data Bank coordinate positions of amino acids as input into TissueLab software. From the amino acid sequence, a type I collagen-like triple helix backbone was computationally constructed and boundary spheres were added based on known chemical and physical properties of the amino acids. Boundary spheres determined the contact surface characteristics of the collagen molecules and intermolecular interactions were then determined by considering the relationships of the contact surfaces and by resolving the energy-minimum state using feasible sequential quadratic programming. From this, the software created fibrils that corresponded exactly to known collagen parameters and were further confirmed by finite element modeling. Computationally derived fibrils were then used to create collagen fibers and three-dimensional collagen matrices. By resolving the energy-minimum state, large complex components of the extracellular space as well as other structures can be determined to provide three-dimensional structure of molecules, molecular interactions and the tissues that they form.

Keywords: collagen/computation/extracellular matrix/modeling/tertiary structure

Introduction

Tissue engineering is an interdisciplinary field that applies the principles of engineering and the life sciences to the development of biological substitutes that restore, maintain or improve tissue function (Langer, 2000). A detailed understanding of the biochemical nature of the components of tissues rationalizes the development of such materials. Computer-generated models of biological tissues, in particular the extracellular matrix (ECM) of connective tissues, is a requirement for tissue engineering. The ECM of connective tissue is composed of networks of collagen and elastin fibers embedded in a ground matrix consisting of proteoglycans and water (Billiar and Sacks, 1997). The collagen fibers and their orientation provide the mechanical properties, including the tensile strength of the fibril (Kumar and Cochran, 1990).

Developing three-dimensional models of the ECM can define the mechanical properties of fibers and thereby characterize the functionality of the structure as a whole (Kumar and Cochran, 1987). Owing to the unique orientation of fibers in the tissue, different structures and scenarios can be created (Hartgerink et al., 2001) and tested. For example, in 1993, Sacks et al. described the geometric properties of the right ventricle of the heart (Sacks et al., 1993). In the process they defined the functionality of the ventricle in such disorders as chronic obstructive pulmonary disease, congenital defects and congestive heart failure. They created a geometric model showing how the unique geometry affected the functionality of the whole organ.

Previous attempts to model tissue have utilized experimental data (Sacks et al., 1997) and have applied differential equations to simulated tissues (Dallon et al., 1999). Modeling has also incorporated finite element analysis to account for the physical stresses during structure formation (Vorp et al., 2001). Towards this, computational chemistry is rapidly emerging as a subfield of theoretical chemistry, where the primary focus is on solving chemically related problems by calculations (Jensen, 1998) that can be applied to tissue modeling. One of the principal tools in computational chemistry that is used for the theoretical study of biological molecules such as proteins is the method of dynamic simulations at the molecular level. The molecular dynamics method was first introduced by Alder and Wainwright in the late 1950s (Alder and Wainwright, 1957, 1959) and the first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (McCammion et al., 1977). Molecular dynamics methods involve calculations of bonded and non-bonded energies between each atom in the system examined. Bonded energies include stretching, bending and torsional energies. Non-bonded energies include van der Waals energy, electrostatic energy and hydrogen bonding energy. Because of limited current computational ability and the enormous number of the atoms in a molecule, present molecular dynamics simulations allow for only a few molecules in their calculations and are therefore not applicable to large-scale systems, such as ECM structures and formation.

As such, greater computational load or alternative strategies are required for the large-scale system. The basic principle of computational chemistry is still applicable and feasible by focusing on the energy-minimum state, not on the dynamic process of the energy state, and by considering only the dominant energy that determines the structure. By doing so, the computational load can be greatly reduced. It was our purpose to model type I collagen fibrils of the ECM by using energy-minimum calculations. The type I collagen-like molecule was selected based on the repetition and regularity of its molecular structure to reduce the computational load further. Given the unit structure of the collagen molecules, the alignment of the ECM can be theoretically calculated to achieve the energy-minimum configuration among molecules.
Type I collagen was also chosen for modeling since it is abundant in many organs in the human body and has a critical role in ECM formation and function. A complete understanding of the basis for collagen stability (and instability) could lead to new biomaterials and effective therapies for disorders (Jenkins and Raines, 2002). A collagen molecule is a left-handed helix consisting of three alpha chains (two alpha-1 and one alpha-2 chains); each alpha chain contains ~1040 amino acids. The basic amino acid sequence of collagen is the repetition of three amino acids: Gly–X–Y in which every third amino acid is a glycine (Gly). X and Y are typically proline; Y is typically modified to hydroxyproline in the polymer sequence after the collagen chain is built (Berg and Prockop, 1973) to increase collagen stability (Bella et al., 1994). Glycine has a small molecular size that fits perfectly inside the helix. Hydroxyproline is critical for collagen stability. Five collagen molecules are bound together to form a microfibril (Cowin, 2000) as shown in Figure 1.

Basic knowledge and principles of the structure of the collagen-based ECM have been acquired from X-ray crystallography and NMR spectroscopy. However, detailed structure and modeling cannot be completed based on existing information. The diameter of collagen microfibril is 3.5–4 nm (Piez and Trus, 1981), based on electron micrograph evidence, but the length of a microfibril is not well known. On electron micrographs, collagen fibrils show distinctive cross-striations, with an apparent periodicity, denoted by the letter D (Figure 1); D is used as a primary reference scale in describing structural levels. The true period is 5 D. The helical length of a collagen molecule is 4.34 D (~291 nm) and the discrete gaps are 0.66 D (~44 nm) (Cowin, 2000). The structure described above is known from preceding studies, but the underlying mechanism for the formation is not fully understood. Unlike the globular proteins, the fibril proteins including the collagens cannot be completely crystallized. Further, the underlying mechanism for the formation of three-dimensional structures is also not fully understood. For these reasons, the computational method presented here will provide insight for further discovery.

Methodology
Computational determination and rendering of collagen fibers were based on the consideration that proteins are rigid spheres that are packed together. The distance geometry between molecules and residues is a functional component that can be mathematically predicted. Reducing the molecular volumes to reduce the computational load is achieved by (a) considering the protein backbone as a braid that is converted to a mathematical model and (b) surface modeling that is based on differential geometry determined by calculation of energy sections that will determine intermolecular distances and hence define assembly and packing of molecules.

Computational modeling involved the following steps:

1. Protein Data Bank (PDB) file input.
2. Determination of backbone structure of polypeptide or molecule.
3. Representation of amino acid residues as spheres with radii based on known physical properties (polarity and charge) of residues (Gilbert, 1999). Excluded volumes (Connolly, 1985) are also represented. The assumption is that the residue volume can be treated as a boundary sphere around the actual structure of the residue as a result of the vibration of the atoms; the atoms in the structure spin about their bond-axes. The speed of the spin of the atoms about the bond-axes in the residue is much greater than the spinning of the molecule.
4. Representation of excluded volumes of amino acid side chains as boundary spheres based on physical properties.
5. Sequential quadratic programming is used to determine intermolecular interaction of closest pairs of residues. Calculations are based on the minimum total energy of interaction. Total energy is the sum of electrostatic, van der Waals and hydrogen bond energies. The assumptions are that the energy between the atoms, at the closest points on the surfaces of two residues examined, is much greater than the energy among other combinations of atoms. Also, the energy between the closest atoms is predominant in the amount of the energy between two residues. The predominant energy is sufficient for the determination of the overall energy-minimum alignment.
6. Energy minima and residue dimensions determine intermolecular distances, stacking geometry and three-dimensional shape of final structure.

For computational modeling of type I collagen we utilized the model simplification method illustrated in Figure 2. The structure of the type I collagen-like molecule [(Gly–Pro–Pro–)10]3 (Berisio et al., 2002) was provided computationally by ProteoRubix software that used a volume minimization approach to determine optimal positions of each amino acid in the protein (Campbell et al., 2003). The initial structure consisting of 1000 amino acids was input into our modeling program as a PDB file (Figure 2a) and any PDB file can be entered. The backbones of each of three alpha chains were extracted (Figure 2b). The backbone was based on the regular structure resulting from linking the amino acids through amide bonds (Creighton, 1993). It was created by (a) using appropriate geometry for shorter amide bonds (C=O–N 1.36 Å) in which resonance structure was a significant contributor compared with regular amide bonds (C=O–N 1.47 Å) and (b) creating a barrier to C–N bond rotation of ~17 kcal/mol (72 kJ/mol) (Nemethy et al., 1992). The three backbones were represented with a single backbone by averaging the nodes on three backbones (Figure 2c).

The residues present at the surface of the triple helix define the characteristics of the molecular surface and determine...
intemolecular interactions. The amino acid residues were represented as boundary spheres and were added to the single backbone (Figure 2d). Boundary spheres were defined by the known chemical and physical properties of the residues, such as hydrophobicity and electrostatic charge. It was assumed that the backbones of the amino acid chains do not contribute significantly to the force interaction with other molecules because the backbones sit inside the molecule.

For intermolecular interactions and fibril formation, this model considered the contact between surfaces of different molecules. The surface interactions were determined by the common boundary sphere interactions between adjacent molecules to determine the energy minimum configuration between molecules. The non-bonding energies between molecules consisted of electrostatic forces \( E_{\text{coul}} \), van der Waals forces \( E_{\text{vdw}} \) and hydrogen bonds \( E_{\text{H-bond}} \). The minimization of the cost function of the total of non-bonding energy is the sum of the three and was used to determine energy-minimizing configurations between molecules in the model:

\[
f(x) = E_{\text{tot}}(x) = E_{\text{coul}}(x) + E_{\text{vdw}}(x) + E_{\text{H-bond}}(x)
\]

where \( x \) is the transformation of the molecule with six degrees of freedom; \( f(x) \) is the objective cost function that is the total amount of energy between collagen molecules, specifically between the boundary spheres representing the residues on the collagen triple helix. Optimization software was developed for the computational experiment. The software read the PDB file, converted the model to a simplified one and ran the optimization routine to locate the energy minimum configuration.

The program, written in C++, used feasible sequential quadratic programming (FSQP) (Panier and Tits, 1993) for the optimization routine to locate the energy minimum configuration of the cost function. FSQP effectively found the solution, i.e. determined energy minima, by solving quadratic subproblems via two-phase operations. Phase I generated iterates to satisfy all linear constraints and nonlinear inequality constraints. Phase II minimized the maximum of the objectives while maintaining satisfaction of linear constraints and nonlinear inequality constraints, with nonlinear equality constraints being satisfied asymptotically (Zhou and Tits, 1993). To handle nonlinear equality constraints, SQP incorporated a modification scheme to generate feasible iterates (Panier and Tits, 1993). The result was a differentiable exact penalty function that was used to determine energy minimum configurations.

Assuming that the minimization of volume and energy is the only principle for calculation and the molecule is a rigid object that has six degrees of freedom, the cost function was solved by determining and measuring distances for every combination of residues on the interacting collagen backbones. The distances increased the value of the cost function if the combination of the residues had an attracting force, such as the interaction between a hydrophobic and a hydrophilic residue. It decreased the value of the cost function if the combination of the residues caused a repelling force, such as the combination of a positively charged residue with another positively charged residue.

The cost function thus determined the energy minimum state as the triple helix moved with six degrees of freedom, which computationally represents a stable interaction and position between the collagen molecules (Lawrence et al., 1997). Any given molecule can change its energy in any of its degrees of freedom. There is one degree of freedom per independent mode of motion, i.e. translational (rigid ‘whole body’ movement in the \( x, y \) or \( z \) directions), rotational (whole body rotation around one or two independent axes) and vibrational (oscillation of bond length or angle). Orientation, however, becomes fixed as a result of the energy minimizing function. The backbone was fixed and therefore did not contribute to the degrees of freedom.

Total energy calculations \( [f(x) = E_{\text{coul}}] \) were based on the following calculations for residue interactions based on inherent physical and chemical properties of the residues. As seen from the equations, energy \( (E) \) minima were a function of distances \( (r) \).

### Electrostatic energy

Electrostatic energy describes the repulsion or attraction between atoms based on atomic charge (Gilbert, 1999). \( E_{\text{coul}} \) is calculated as

\[
E_{\text{coul}} = \frac{q_1 q_2}{4 \pi \varepsilon_0 r}
\]

where \( q_1 \) and \( q_2 \) are the charge on each residue, assuming that the charge on a residue is uniform, \( r \) is the distance between the residue surfaces and \( \varepsilon_0 \) is the dielectric constant.

### van der Waals interaction

van der Waals energy is the non-electrostatic energy that describes the repulsion or attraction between atoms that are not covalently bonded. To calculate van der Waals energy, the Lennard–Jones potential was used:

\[
E_{\text{vdw}} = \varepsilon \left[ \left( \frac{r_0}{r} \right)^{12} - 2 \left( \frac{r_0}{r} \right)^6 \right]
\]

where \( r_0 \) is the sum of the van der Waals radii of two contacting atoms, \( r \) is the distance between the molecular surfaces and \( \varepsilon \) is
the depth of the minimum parameter (Alder and Wainwright, 1957; Jensen, 1998).

Hydrogen bond

A hydrogen bond is formed between hydrogen atoms bonded to electronegative atoms and negatively charged heteroatoms. A hydrogen bond (2–5 kcal/mol) is significantly stronger than van der Waals interactions (0.1–0.2 kcal/mol). The amino acids lysine, arginine and histidine all have one or more hydrogen-bond donor sites. Glutamate, aspartate and methionine have one or more possible hydrogen-bond acceptor sites. Glutamine, asparagine, threonine, serine, cysteine and tyrosine have both donor and acceptor sites. In cases in which residues have multiple hydrogen donors and acceptors, the number of donors and acceptors actively involved in hydrogen bonding to other residues depends on the relative position of the residues and the environment. To simplify the problem, we calculated only the energy of one hydrogen bond, occurring at the closest pair of the point on the surfaces of the donor and acceptor residue. We also assumed that the directional term in hydrogen bonding is negligible in such a large-scale simulation since the residues are fixed to the backbone. To calculate the energy, we adopted one commonly used function modified from the Lennard–Jones potential:

\[
E_{\text{H-bond}} = \varepsilon \left[ 5 \left( \frac{r}{r_0} \right)^{12} - 6 \left( \frac{r}{r_0} \right)^{10} \right]
\]

where \(r_0\) is the equilibrium distance of the donor atom and that of the acceptor atom, respectively (Jensen, 1998).

Finite element model construction

To confirm the structures created by our software, residue boundary spheres and backbones were converted to a finite element mesh to construct the surface of the collagen molecule (Oden and Carey, 1983). This model also allowed for the consideration of the contact between surfaces of different molecules. Specifically, model surface interactions were determined by common node interactions between adjacent molecules. The nodes corresponded to the attraction/repelling interactions between various molecular, van der Waals and hydrophobic forces at corresponding residue positions. The finite element method is basically concerned with the determination of field variables. The most important ones are the stress and strain fields determined in CalculiX (Dhondt, 2004). This software computationally utilizes Lagrangian strain tensor for elastic media, Eulerian strain tensor for deformation plasticity and deviatoric elastic left Cauchy–Green tensor for incremental plasticity to determine a Cartesian coordinate systems. All CalculiX input (e.g. distributed loading) and output are in terms of true stress.

The collagen model from Tissue Lab provided the values for the energy nodes. Repulsive and attractive stresses were located at corresponding residue positions on the surface. The stress zones modeled the hydrophilic and hydrophobic areas of the collagen. The finite element acts to minimize and balance stress throughout the structure and modeled the final outcome of the redistribution. In the initial study the cylindrical mesh was made up from 20 node brick elements. There were 16 elements per cylinder. For the initial test of the software, the cylinder was fixed and acted as a cantilever only on its \(\pm\zeta\) edge; all other surfaces were free to move. A force was applied to the top of the cylinder at the \(\pm\zeta\) edge. The force was directed in the \(y\)-direction and the stress contours for this situation were determined (Eringen, 1988; Simo and Hughes, 1997). This was described as a multi-dimensional nonlinear optimization problem (Kumar and Cochran, 1990; Nawrocki et al., 1996; Beltyutschko et al., 2001) to determine optimal interaction for final modeling of the collagen.

Results

For collagen fibril formation, the software examined the energy function to identify the region holding the global minimum from other regions that may have held local minima. Thus the optimizer narrowed the search region stochastically. Figure 3 shows the plots of energy function with two types of translations and two types of rotations. The molecule used here was a common collagen-like structure \([(\text{Gly–Pro–Pro–})_{10}]\) (Berisio et al., 2002). It was clear from the energy equations of large surface areas that the energy decreased as: collagen molecules came together until they overlapped (Figure 3a); lined up in parallel as opposed to perpendicular (Figure 3b); lined up at a similar but slightly shifted level (Figure 3c) which agreed with the known periodicity of collagen (Figure 1); and attained a rotational orientation (Figure 3d). The collagen molecule is known to be a helical structure. Addition of subsequent collagen molecules resulted in fiber formation (see below). It was possible to restrict the intermolecular energy minimization of the transformation according to the nature of the energy function so that the optimizer only searched the region with a high probability of containing the global minimum.

Through the optimization process, Figure 4a shows five collagen molecules forming a five-molecule-based fibril in agreement with the known structure and periodicity of collagen (Figure 1). The molecules are shown to be tightly aligned together. As the molecules formed fibrils, these were computationally linked, again by determining energy minima to build three-dimension fibers shown in Figure 4b–d.

For confirmation of a stable three-dimensional structure our energy calculations, which were based on known data of thermal stability, were compared with known experimental data for the denaturation process involving over 30 residues (Voet and Voet, 1990). From our calculations, the micro process (micro unfolding) has a Gibbs energy of the order of 7–11 kJ/mol; the macro process (macro unfolding) energy is 200–400 kJ/mol; the total value for enthalpy \(\Delta H\) is 4000–6500 kJ/mol; and entropy \(\Delta S\) is 14–21 kJ/mol. For the enthalpy \(\Delta H\), we used two main criteria for estimating the strength of the H-bond in the A-HO¨O¨O¨O¨O¨O¨B (backbone-residue) system: the A-H stretching frequency or its relative shift \((\nu_0–\nu_0)^o\) where \(\nu_0\) is stretching frequency of the free A-H group) and the distances A–H and A000B. According to these criteria H-bonds are regarded as weak, intermediate and strong. For the OHOOB.O bonds this approximate classification is given in Table I. The length of H-bonds in collagen is \(\sim 1.5\) Å (Richards, 1979). In the case of AHO00B the theoretical potential energy curve (Voet and Voet, 1990) corresponds directly to our simulated energy curves (Figure 3) and therefore validates our results for fiber formation.

For confirmation of our computer-generated model, we converted collagen molecules into finite element constructs (Figure 5a) with nodes based on those generated by our software. Interactions of the finite element ‘molecules’ resulted in
Fig. 3. Energy plots during fibril formation: global energy minima were calculated and plotted as molecules were brought into proximity during fibril formation. (a) Energy decreased to reach increased stability as molecules came together. (b) Energy decreased as molecules were oriented into a parallel position. (c) Slight shifting further reduced energy and resulted in periodicity of the structure. (d) Correct rotational alignment also contributed to the energy-minimum state.

Fig. 4. (a) Computer graphic showing alignment of five type I collagen-like microfibrils, also based on achieving energy-minimum status as determined by molecular interactions. (b)–(d) Microfibril alignment to produce a three-dimensional computer-generated extracellular matrix.
The results show the successful implementation of the physical and chemical principles to build three-dimensional collagen fibrils and an ECM scaffolding. Although the structure of the type I collagen is well known theoretically, the actual measurements are difficult to obtain because the technical challenges of obtaining diffraction of fibrous proteins. Available diffraction measurements of collagen-like polypeptide [(Pro- Pro-Gly)]\textsubscript{10} obtained from crystals grown in a microgravity and having the highest resolution description of a collagen triple helix reported to date (Bersio et al., 2002) are, however, in agreement with our findings. The computed fibril was verified by finite element analysis; further, the ECM structure can be, in part, improved if necessary using direct measurement data to modify computational calculations. When the robustness of the simulation is assured, it can be applied as a design tool for experiments and simulations of new biomaterials and other applications.

Although finite element analysis provided confirmation of fibril assembly, the applications of this technique to tissue modeling should generally be considered as rudimentary. Finite element analysis does not allow for robust calculations of interactions including the significant bending or torsional molecular changes that occur.

It must be kept in mind that our modeling was based on a collagen-like molecule (Gly–Pro–Pro)\textsubscript{n} that is less complex and more stable than that which occurs in the ECM. This structure was chosen because it could be compared with known crystallographic data on the molecule. Although \textit{in situ} collagen has a slightly differing amino acid sequence and interacts with several other molecules ranging from proteoglycans to other collagen types and hydroxyapatite, our model did mimic collagen with high accuracy including the formation of a microfibril consisting of five collagen molecules. Any PDB file can be used as input for modeling in our software. Further, our model can readily be modified to account for alterations in amino acids and intermolecular interactions, as well as interactions with other ECM molecules (Brodsky and Ramshaw, 1994), to improve accuracy further. The use of type I collagen-like molecule as an example shows that our computational methods are an intriguing first approach towards analyzing complex tertiary and quaternary structures of proteins.

In addition to energy minima on intermolecular interactions and interactions with other ECM molecules, several other properties also predispose the orientation of collagen fibers, including stress vectors, which are related to the effect of mechanical loads on a forming tissue. These loads can give result in directional orientation of the fibers that form the tissue (Hughes and DeVore, 1980) and which thereby provide specific structure and hence functionality to the tissue. Additional programming can be readily added to determine preferred fiber orientation(s) during fibril formation based on physical stress.

<table>
<thead>
<tr>
<th>H-bond</th>
<th>(\Delta\psi_{\text{eq}}(%))</th>
<th>ROOO (Å)</th>
<th>(\Delta H)</th>
<th>(kcal/mol)</th>
<th>(kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>12</td>
<td>2.7</td>
<td>5</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>12–22</td>
<td>2.7–2.6</td>
<td>6–8</td>
<td>25–33</td>
<td></td>
</tr>
<tr>
<td>Strong</td>
<td>25–83</td>
<td>2.6–2.4</td>
<td>8</td>
<td>33</td>
<td></td>
</tr>
</tbody>
</table>
As a starting point, the molecule was initially assumed to be a rigid object; however, it was expected that one molecule will deform to fit tightly to the other. Local deformations were initially calculated after locating the energy-minimum state by optimization, without allowing the molecule to deform. This was done by modeling the functions between the energy in the local domain and the bending and torsional angles of the segment of the backbone of the molecule. The software effectively determined local deformation for tighter packing. Although the same problem simplification method may not be applicable for all molecules, the use of feasible sequential quadratic programming can be successfully used to determine tissue structure based on molecular configuration. As such, this modeling can be applied to computer-assisted tissue engineering, proteomics, protein folding-related diseases, effects of mutations and the interactions between proteins with ligands, substrates, DNA, pharmaceuticals and other proteins or peptides. Molecules with less symmetry do require additional programming and software steps for the increased complexity. Modeling has direct applications to investigate the fundamental question of how amino acid sequences of peptides change to functional tertiary or quaternary structures and the four-dimensional process that they undergo.

References


Received December 23, 2004; revised May 14, 2005; accepted May 17, 2005

Edited by Albert Berghuis