ColECM - Collagen ExtraCellular Matrix Simulation Documentation

Francis G.J. Longford

July 30, 2018

ColECM (Collagen ExtraCellular Matrix) is a computer program written in Python that models the dynamic behaviour of collagen fibres in the extracellular matrix (ECM) using a molecular dynamics (MD) style integrator. It has both serial and parallel implementations (written using the mpi4py library) which can be installed depending on system architecture. The main routines are as follows: 'simulation' creates, equilibrates and integrates system through time, 'analysis' fabricates and analyses second-harmonic generation (SHG) images based on simulation output files and 'editor' allows for editing of the simulation parameter and cell files.

1 Simulation Modelling

We propose that a bead-and-spring dynamical network approach to modelling the ECM could be a way forward to achieve this goal. In a similar methodology to previous studies, we suggest that a collagen fibril can be represented on a micro-metre scale by a series of atom-like beads connected by springs.

1.1 Potentials

Below describes the potentials used in our bead-and-spring fibril model.

1.1.1 Permanent Bonds

The potential energy of with respect to the length, r, of these springs is represented by a harmonic oscillator, with r_b as the equilibrium bond length and k_b as the bond force constant.

$$V_b(r) = k_b(r - r_b)^2 \tag{1}$$

Additionally, the chains can be allowed to bend linearly by using a similar harmonic potential to model the bond angle, θ, between three beads, with θ_0 as the equilibrium bond angle and k_a as the angle force constant. The "stiffening" of the fibrils can therefore be modified by increasing the force constants of both potentials. It could also be assumed that since all beads are homogeneous the equilibrium bond angle would be 180, so that $\theta_0 = \pi$. For computational ease we approximate the form of this harmonic potential by a sinusoidal function

$$V_a(\theta) = k_a \left(\cos(\theta) + 1\right) \tag{2}$$

1.1.2 Cross-links

Whereas these potentials have been used several times previously to model bonding within individual collagen fibrils and have clear physical analogies, the interstitial forces between fibrils are less well understood. Some studies have used cross-links between fibrils,(32),(35) although assigning these at the beginning of a dynamical simulation would predetermine the motion and ordering of the system if they were unable to break and form throughout. Other studies into the fibrillogenesis have suggested that lateral growth and fusion of two fibres side-by-side does not occur in the ECM, due to the twisted helical structure of fibrils[1, 2]. In which case, it would be inaccurate to explicitly assign bonds between fibres, rather than using a potential to describe these intra-fibriliar forces. Therefore, a Lennard-Jones potential could be appropriate to model the non-bonding interactions between fibrils, as it has been extensively used in molecular dynamics to model dispersion forces between non-bonded atoms.

$$V_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{3}$$

In order to model dispersion forces, this potential is applied to every pairwise distance between all beads in the simulation, although the energy parameter ϵ is significantly increase for interactions between beads at the ends of each linear chain. In this way we model possible linear bonding between fibrils. Additionally, in order to make sure that there is an overlap in potential between bonded beads, the equilibrium bond length r_0 is set equal to $2^{1/6}\sigma$, the value of r at which there is a minima in the potential, so that $V(r_0) = -\epsilon$.

The total potential acting upon each particle in the collagen chain is therefore given by the sum- mation the of bonded and non-bonded interactions:

$$V_{TOT} = \sum^{bonds} V_b(r_{ij}) + \sum^{angles} V_a(\theta_{ijk}) + \sum^{pairwise} V_{LJ}(r_{ij}) \tag{4}$$

1.2 Dynamic Bonds

NOTE - CURRENTLY IN DEVELOPMENT, NOT YET FULLY IMPLEMENTED

Simulation and experimental stress-strain studies have shown that collagen fibrils undergo hyperelastic deformation when under tensile stress[]. Therefore in order to model the transition between elastic and hyper-elastic behaviour, two sets of parameters for equation (1) can be used to describe different ranges of r. Previous simulation studies have used the same method to TC polymers within the fibrils, and report similar mechanical behaviour between the individual TC tensile strength and that of a fibril[]. Therefore, it seems appropriate to employ the same modelling technique for further coarse graining (figure ...). By default we set $r_1 = 1.5r_0$ and $k_1 = 4k_0$. We also include an r_{break} distance, at which point the bond between two beads in a fibril will break.

$$V_b(r) = \begin{cases} k_0(r - r_0)^2 & if \quad r \leq r_1 \\ k_1(r - r_1)^2 & if \quad r_1 \leq r \leq r_{break} \\ 0 & if \quad r_{break} \leq r \end{cases} \tag{5}$$

A bond is subsequently able to form between the terminal beads of each fibril every time step at a probability determined by their radial distance $P_b(r)$ (equation (6)). The full range of r at which bonding forces may be applied lies between $2\sigma \leq r \leq r_{break}$, therefore we normalise $r^2 - 4\sigma^2$ by $r_{break}^2 - 4\sigma^2$ and use this ratio to determine the probability of bonding so that $P_b(2\sigma) = 1$.

$$P_b(r) = 1 - \frac{r^2 - \sigma^2}{r_{break}^2 - \sigma^2} \tag{6}$$

1.3 Forces

The force applied to each connected bead from this potential can be found by defining it as the negative of the derivative of potential with respect to bead position. Therefore, the bonded force vector $\mathbf{F}_b$ as a result of the displacement vector $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and radial distance $r_{ij} = |\mathbf{r}_{ij}|$ between beads i and j is given by:

$$\mathbf{F}_b(\mathbf{r}_i) = -\frac{\partial V_b(r_{ij})}{\partial \mathbf{r}_i} = -\frac{\partial V_b(r_{ij})}{\partial r_{ij}} \cdot \frac{\partial r_{ij}}{\partial \mathbf{r}_i} = \mathbf{r}_{ij}\frac{2k_b}{r_{ij}}(r_0 - r_{ij}) \tag{7a}$$

$$\mathbf{F}_b(\mathbf{r}_j) = -\frac{\partial V_b(r_{ij})}{\partial \mathbf{r}_j} = -\frac{\partial V_b(r_{ij})}{\partial r_{ij}} \cdot \frac{\partial r_{ij}}{\partial \mathbf{r}_j} = -\mathbf{r}_{ij}\frac{2k_b}{r_{ij}}(r_0 - r_{ij}) \tag{7b}$$

Performing a similar derivation as equation (7) results in an expression for the angular force $\mathbf{F}_a$ acting upon between beads i, j and k as a result of the angle θ_{ijk} that bisects vectors $\mathbf{r}_{ij}$ and $\mathbf{r}_{jk}$.

$$\mathbf{F}_a(\mathbf{r}_i) = -\frac{\partial V_a(\theta_{ijk})}{\partial \mathbf{r}_i} = -\frac{\partial V_a(\theta_{ijk})}{\partial \theta_{ijk}} \cdot \frac{\partial \theta_{ijk}}{\partial \mathbf{r}_i} = \frac{1}{r_{ij}} \left(\frac{\mathbf{r}_{jk}}{r_{jk}} + \frac{\mathbf{r}_{ij}}{r_{ij}} \sin(\theta_{ijk}) \right) \tag{8a}$$

$$\mathbf{F}_a(\mathbf{r}_j) = -\frac{\partial V_a(\theta_{ijk})}{\partial \mathbf{r}_j} = -\frac{\partial V_a(\theta_{ijk})}{\partial \theta_{ijk}} \cdot \frac{\partial \theta_{ijk}}{\partial \mathbf{r}_j} = -\frac{(\mathbf{r}_{ij} + \mathbf{r}_{jk})}{r_{ij} \cdot r_{jk}} - \sin(\theta_{ijk}) \left(\frac{\mathbf{r}_{ij}}{r_{ij}^2} - \frac{\mathbf{r}_{jk}}{r_{jk}^2} \right) \tag{8b}$$

$$\mathbf{F}_a(\mathbf{r}_k) = -\frac{\partial V_a(\theta_{ijk})}{\partial \mathbf{r}_k} = -\frac{\partial V_a(\theta_{ijk})}{\partial \theta_{ijk}} \cdot \frac{\partial \theta_{ijk}}{\partial \mathbf{r}_k} = \frac{1}{r_{jk}} \left(\frac{\mathbf{r}_{ij}}{r_{ij}} + \frac{\mathbf{r}_{jk}}{r_{jk}} \sin(\theta_{ijk}) \right) \tag{8c}$$

The non-bonded Lennard-Jones type forces between each bead are subsequently derived in the same fashion.

$$\mathbf{F}_{LJ}(\mathbf{r}_i) = -\frac{\partial V_b(r_{ij})}{\partial \mathbf{r}_i} = -\frac{\partial V_b(r_{ij})}{\partial r_{ij}} \cdot \frac{\partial r_{ij}}{\partial \mathbf{r}_i} = -\mathbf{r}_{ij} \frac{24\epsilon}{r_{ij}^2} \left[2 \left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{9a}$$

$$\mathbf{F}_{LJ}(\mathbf{r}_j) = -\frac{\partial V_b(r_{ij})}{\partial \mathbf{r}_j} = -\frac{\partial V_b(r_{ij})}{\partial r_{ij}} \cdot \frac{\partial r_{ij}}{\partial \mathbf{r}_j} = \mathbf{r}_{ij} \frac{24\epsilon}{r_{ij}^2} \left[2 \left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{9b}$$

This results on the net force acting upon a single bead i defined by:

$$\mathbf{F}_i = -\frac{\partial V_b(r_{ij})}{\partial \mathbf{r}_i} - \frac{\partial V_a(\theta_{ijk})}{\partial \mathbf{r}_i} - \frac{\partial V_{LJ}(r_{ij})}{\partial \mathbf{r}_i} \tag{10}$$

In such a way, a molecular dynamical simulation, with particle trajectories being determined by pairwise force interactions and integrated using the verlocity-Verlet algorthim could be used to simulate a simplified system of collagen fibrils in the ECM.

1.4 Calculation of Angles

In order to efficiently calculate the force and energy components of each multi-body term, we neglect to calculate each angle θ_{ijk} between connected beads i, j and k explicitly, and instead use sinusoidal functions (equations (2), (8)). We then calculate the terms $\sin(\theta_{ijk})$ and $\cos(\theta_{ijk})$ from the cross and dot products respectively of the displacement vectors $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{r}_{jk} = \mathbf{r}_j - \mathbf{r}_k$.

$$\sin(\theta_{ijk}) = \frac{\mathbf{r}_{ij} \otimes \mathbf{r}_{jk}}{|\mathbf{r}_{ij}||\mathbf{r}_{jk}|} \qquad \cos(\theta_{ijk}) = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}_{jk}}{|\mathbf{r}_{ij}||\mathbf{r}_{jk}|} \tag{11}$$

The cross-product for two dimensional vectors is undefined, and so when running a simulation in 2D, we used the following trigonometric definitions.

$$\sin(\theta_{ijk}) = \frac{\det(\mathbf{r}_{ij}, \mathbf{r}_{jk})}{|\mathbf{r}_{ij}||\mathbf{r}_{jk}|} \qquad \cos(\theta_{ijk}) = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}_{jk}}{|\mathbf{r}_{ij}||\mathbf{r}_{jk}|} \tag{12}$$

Where the determinant of two 2D vectors is defined by.

$$\det(\mathbf{a}, \mathbf{b}) = \mathbf{a}_y \mathbf{b}_x - \mathbf{a}_x \mathbf{b}_y \tag{13}$$

We use numerical python (NumPy) to calculate each sinusoidal term, due to its speed when handling vector arrays, which is close to that of compiled C.

1.5 Fibre Dynamics

As mentioned previously, MD simulations are propagated through time by the calculation of net intra- and inter-molecular forces acting upon each particle. These forces are defined as proportional to the negative of the derivative of the potential energy function for each interaction. Therefore the net force per particle is a vector summation of all the individual covalent and non-covalent force contributions from its surrounding neighbours. The net force in each atom is then used to propagate the system through time using the atomic mass m from Newton's equations of motion.

$$\mathbf{F}_i(t) = m_i\mathbf{a}_i(t) \tag{14}$$

This leads to a set of differential equations for the position $\mathbf{r}_i(t)$ and velocity $\mathbf{r}_i(t)$ of each bead at any given moment in time t.

1.6 Velocity-Verlet Integration

A Verlet integration routine is used to propagate the acceleration $\mathbf{a}_i(t)$ and velocity $\mathbf{v}_i(t)$ vectors through a time step Δt in order to solve the positions $\mathbf{r}_i(t+\Delta t)$ and velocities $\mathbf{v}_i(t+\Delta t)$. It is assumed that the forces remain constant during a finite difference in time Δt, which is reversible and therefore the change in dynamical variables can be approximated as a Taylor series expansion.

$$\begin{aligned}
\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2 + ...\\
\mathbf{r}_i(t-\Delta t) &= \mathbf{r}_i(t) - \mathbf{v}_i(t)\Delta t - \frac{1}{2}\mathbf{a}_i(t)\Delta t^2 + ...\\
\therefore \quad \mathbf{r}_i(t+\Delta t) &= 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \mathbf{a}_i(t)\Delta t^2 \qquad (15)
\end{aligned}$$

If we wish to explicitly solve $\mathbf{v}_i(t)$, this normally requires both $\mathbf{r}_i(t-\Delta t)$ and $\mathbf{r}_i(t+\Delta t)$ to be already known. However, the Velocity Verlet algorithm[3] uses a half-step velocity calculation, allowing for $\mathbf{v}_i(t+\Delta t)$ to be estimated at the same time as $\mathbf{r}_i(t+\Delta t)$.

$$\begin{aligned}
\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2 && \text{(16a)}\\
\mathbf{a}_i(t+\Delta t) &= \frac{\mathbf{F}_i(t+\Delta t)}{m_i} && \text{(16b)}\\
\mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \frac{1}{2}\left[\mathbf{a}_i(t) + \mathbf{a}_i(t+\Delta t)\right]\Delta t && \text{(16c)}
\end{aligned}$$

Note that the forces are recalculated halfway through the algorithm based on the updated nuclear positions. An appropriate length of time step Δt is critical to maintain the conservation of energy and ensure the integration method is time-reversible. A larger time step will increase the speed of simulation and computational efficiency, though at the cost of numerical accuracy. In order to avoid bead overlaps, leaping to physically unrealistic interactions and excessive potential energies Δt is required to be shorter than the fastest motions in the system.

1.7 Temperature/Pressure

In a dynamical system particles obey the fluctuation-dissipation theorem, whereby deviations of properties away from the ensemble average are counter-acted statistically by an equivalent reverse process. Thermal fluctuations alter particle velocities, which impact upon the particle collision rate and so vary the amount of kinetic energy dissipated. Therefore the average temperature of a system can be represented in terms of kinetic energy and consequently average particle velocities. At equilibrium, classical systems adhere to the Maxwell-Boltzmann distribution of energy states. This means that the total kinetic and potential energies will fluctuate around their ensemble averages with a probability given by a normal distribution. The total kinetic energy K_{TOT} is given in terms of individual particle velocities by

$$K_{TOT} = \frac{1}{2}\sum_{i}^{N} m_i |\mathbf{v}_i|^2 \tag{17}$$

The instantaneous temperature T of a system containing N particles is related to K_{TOT} via the Boltzmann constant k_B and number of degrees of freedom of each particle N_f.

$$k_B T = \frac{1}{2}\frac{K_{TOT}}{N_f} = \frac{1}{2N_f}\sum_{i}^{N} m_i |\mathbf{v}_i|^2 \tag{18}$$

Therefore we can constrain the velocities of each bead in the system by setting a fixed parameter $k_B T$, to which this relation must hold. Volume and pressure fluctuations are controlled in a similar way by barostats, using a statistical mechanical description of the pressure or stress tensor across the global system.

1.8 Langevin Dynamics

The Langevin equation is a stochastic differential equation, describing a slower-moving system coupled to one containing fast degrees of freedom that are implicitly taken into account by a random force. This seems an appropriate description of motion for large collagen proteins containing 1000s of atoms immersed in the ECM blood-solution, mainly comprised of water. In order to replicate these dynamics we employ a Langevin thermostat, which includes an explicit term for the drag, or friction force as a function of bead velocity, as well as a stochastic force $\mathbf{R}(t)$, representing thermal "noise". The drag force felt by particle i is defined as proportional to its velocity and a friction coefficient ξ.

$$F_i(t) = -\xi \mathbf{v}_i(t) \tag{19}$$

The friction coefficient is related to the collision frequency γ_i of a particle via $\gamma_i = \xi/m_i$, so that γ_i^{-1} is a measure of the velocity relaxation time. Relaxation times are usually expressed, since they can be parametrised using fluid viscosities via Stoke's law. Therefore (14) can be amended with these terms included.

$$F_i(t) = m_i \mathbf{a}_i(t) - \gamma_i m_i v_i(t) + \mathbf{R}_i(t) \tag{20}$$

We include this constraint in the velocity-Verlet algorithm via the second-order integration method proposed by Goga *et al*[4], which is accurate to Δt^2. The stochastic force at each time step is modelled by $\xi_i(t)$, sampled from a Gaussian distribution with $\langle \xi \rangle = 0$ and $\langle \xi_n \xi_m \rangle = \delta(n, m)$, where δ is the Dirac delta. This noise, along with the friction term determined by $0 \leq \gamma_i \leq 1$ is applied via an impulsive leap-frog step.

$$\begin{aligned}
\mathbf{v}_i(t) &= \mathbf{v}_i(t - \frac{1}{2}\Delta t) + \mathbf{a}_i(t)\Delta t && \text{(21a)}\\
\Delta \mathbf{v}_i &= -\gamma_i \mathbf{v}_i(t) + \sqrt{\gamma_i(2-\gamma_i)k_B T/m_i}\; \xi_i(t) && \text{(21b)}\\
\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \left(\mathbf{v}_i(t) + \frac{1}{2}\Delta \mathbf{v}_i\right)\Delta t && \text{(21c)}\\
\mathbf{a}_i(t+\Delta t) &= \frac{\mathbf{F}_i(\mathbf{r}_i(t+\Delta t))}{m_i} && \text{(21d)}\\
\mathbf{v}_i(t+\frac{1}{2}\Delta t) &= \mathbf{v}_i(t) + \Delta \mathbf{v}_i && \text{(21e)}
\end{aligned}$$

1.9 Berendsen Barostat

We employ the Berendsen barostat, which couples the pressure of the system with an external bath, amending particle trajectories to control the internal pressure of the simulation cell[5]. Strictly speaking, the Berendsen barostat is not able to reproduce a correct NPT ensemble of particle trajectories, as it is not stochastic. However, for our investigation we only employ pressure constraints to equilibrate the density of our liquid phase during preparation final system and therefore do not sample any trajectories or report measurements of properties taken whilst applying the barostat.

The global pressure of a molecular simulation can be defined at each time step in terms of the net force exerted by N particles upon the walls of the system. For an n dimensional many body system, this can be represented as a $n \times n$ Gauchy stress tensor $\mathbf{P}$, where each component, $P_{\alpha\beta}$, represents the force in β direction acting on the surface plane that has a normal vector in the α direction. The pressure components include the kinetic energy of the system and the virial sum of the distance $\mathbf{r}_{ij}$ and force $\mathbf{F}_{ij}$ vectors of each pairwise interaction between particles i and j (equation 22, where m_i is the mass of particle i, v_{i_α} the α component of the particle's Cartesian velocity $\mathbf{v}_i$, and r_{ij_α} and F_{ij_α} are the α components of the Cartesian distance and force vectors respectively between particles i

and j).

$$P_{\alpha\beta} = \frac{1}{2V}\left[\sum_{i}^{N} m_i v_{i_\alpha} v_{i_\beta} + \sum_{j>i} r_{ij_\alpha} F_{ij_\beta}\right] \tag{22}$$

Internal contributions to the virial are cancelled out on average by the internal kinetic energy[5], therefore we choose to omit bonded forces between beads, so that $F_{ij_\beta} = F_{LJ}(r_{ij_\beta})$. The global pressure of an orthorhombic system is defined as the statistical average of the diagonal components $P(t) = \langle P_{\alpha\alpha}\rangle$, since it is assumed that the off diagonal elements remain negligible. The Berendsen barostat uses a pressure scaling factor $\mu(t)$ at each time step to amend the positions and cell dimensions, which is proportional to the deviation of $P(t)$ from the reference pressure P_0.

$$\mu(t) = \left[1 + \left(\frac{\beta_p \Delta t}{\tau_p}\left[P(t) - P_0\right]\right)\right]^{1/3} \tag{23}$$

Where β_p is related to the isothermal compressibility of the system and τ_p is a non-critical time constant for the change in pressure. The form of $\mu(t)$ is similar to that of the scaling factor that appears in the Andersen thermostat[6]. Since neither are measurable properties, inaccuracies in either can be overcome by choosing a suitable factor $\lambda_p = \frac{\beta_p \Delta t}{\tau_p}$. We set $\tau_p = 10\Delta t$ and $\beta_p = 10^{-4}$, yielding $\lambda_p = 10^{-5}$. Incorporating the Berendsen barostat into our Langevin dynamics routine is achieved via the following procedure, where $V(t)$ is the cell volume at time t:

$$\mathbf{v}_i(t) = \mathbf{v}_i(t - \frac{1}{2}\Delta t) + \mathbf{a}_i(t)\Delta t \tag{24a}$$

$$\mu(t) = \left(1 + \lambda_p \left[P(\mathbf{r}_i(t), \mathbf{v}_i(t)) - P_0\right]\right)^{1/3} \tag{24b}$$

$$\Delta \mathbf{v}_i = -\gamma_i \mathbf{v}_i(t) + \sqrt{\gamma_i(2-\gamma_i)k_BT/m_i}\, \xi_i(t) \tag{24c}$$

$$\mathbf{r}_i(t + \Delta t) = \mu(t) \left[\mathbf{r}_i(t) + \left(\mathbf{v}_i(t) + \frac{1}{2}\Delta \mathbf{v}_i\right) \Delta t\right] \tag{24d}$$

$$V(t + \Delta t) = \mu(t)V(t) \tag{24e}$$

$$\mathbf{a}_i(t + \Delta t) = \frac{\mathbf{F}_i(\mathbf{r}_i(t + \Delta t))}{m_i} \tag{24f}$$

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t) + \Delta \mathbf{v}_i \tag{24g}$$

$$\tag{24h}$$

1.10 Parametrisation

Following SHG studies we use a van der Waals radius of $\sigma = 0.5$ μm

1.11 Simulation Procedure

Our simulation procedure is then as follows:

1. Create a fibril template of length l beads
2. Fill a cuboid with n_x, n_y and n_z (3D only) repeating units of this template.
3. Heat the system under NVT conditions until temperature converges to k_BT.
4. Equilibrate the system under NPT conditions at k_BT until density converges to ρ_0, by amending the reference pressure P_0.
5. Equilibrate the system under NVT conditions until internal energy converges.

6. Record simulation of the system under NVT conditions at k_BT lasting for n_{step} Δt, where system configurations are recorded every 1000 Δt.

2 Simulation Analysis

2.1 Fibre Orientation

Our simulation allows us to calculate fibril orientations directly without needing to use image analysis techniques. This is a useful tool to demonstrate the reliability of assessing fibre network structures. Considering that each of our fibres are linear, we simply define the average vector $\mathbf{R}_i$ for each fibre as a sum of the individual vectors $\mathbf{r}_k$ of n_i bonds between beads.

$$\mathbf{R}_i = \frac{1}{n_i} \sum_{k=0}^{n-1} \mathbf{r}_k \tag{25}$$

Considering that the directional components are arbitrary, the magnitude R of the average $\mathbf{R}_i$ vector $\langle \mathbf{R} \rangle$ will then inform us of the strength of alignment for each fibril.

$$\langle \mathbf{R} \rangle = \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{R}_i \tag{26}$$

$$R = \sqrt{\langle \mathbf{R} \rangle_x^2 + \langle \mathbf{R} \rangle_y^2 + \langle \mathbf{R} \rangle_z^2} \tag{27}$$

Consequently, the greater the magnitude of R, the higher the degree of alignment.

2.2 SHG Image Recreation

In order to produce imitation images to compare with SHG results, we convolute a 2D collagen density distribution $\rho(x, y)$ with a Gaussian function. The resulting image intensity map $I(x, y)$ is designed to mimic experimental images of collagen intensity produced by SHG analysis of prostate biopsies[7].

$$I(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \iint \rho(x'-x, y'-y) \exp\left(-\frac{x'^2+y'^2}{2\sigma^2}\right) dx'dy' \tag{28}$$

For simulations in 3D, we mimic a biopsy “slice” by including a density distribution along the z axis $\rho(x,y,z)$, weighted by the radial distance viewed when from the plane $z=0$. In both cases we are left with an intensity function containing two variables $I(x,y)$ only.

$$I(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \iiint \rho(x'-x, y'-y, z') \exp\left(-\frac{x'^2+y'^2+z'^2}{2\sigma^2}\right) dx'dy'dz' \tag{29}$$

In practice we create a histogram of H_{ij} from a set of bead positions $\mathbf{r}(x,y)$ for n beads, by binning the frequency of each position across a discrete pixel grid $i=0,1..N$, $j=0,1..M$. The number of grid points in each dimension is given by selecting an appropriate level of resolution. Considering diameter of each collagen fibre is typically 1 μm, we set the radius of one bead as $\sigma = 0.5$ μm. A typical SHG slide possesses a resolution of ...,[8] leading to the conversion factor...

The value of this discrete density distribution at each pixel H_{ij} is then convoluted by integrating a Gaussian function with a variance of σ^2 to yield the intensity map I_{ij}.

$$I_{ij} = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{NM}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1} H_{uv} \exp\left(-\frac{(u-i)^2+(v-j)^2}{2}\right) \tag{30}$$

For simulations in 3D viewed from the plane $z=0$, this becomes.

$$I_{ij} = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{NMO}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\sum_{w=0}^{O-1} H_{uvw} \exp\left(-\frac{(u-i)^2+(v-j)^2+w^2}{2}\right) \tag{31}$$

2.3 Nematic Tensor Analysis

Considering the definition of our Gaussian convoluted intensity map $I(x,y)$ is another Gaussian convoluted intensity map $I(x,y)$, we choose to implement the FibrilTool methodology ourselves, rather than exporting images for analysis with ImageJ. Consequently, we define the derivatives in equation (**??**) by the following:

$$\frac{\partial I(x,y)}{\partial x'} = -\frac{1}{\sqrt{2\pi}\sigma^3}\iint x' H(x'-x, y'-y)\exp\left(-\frac{x'^2+y'^2}{2\sigma^2}\right)dx'dy' \tag{32a}$$

$$\frac{\partial I(x,y)}{\partial y'} = -\frac{1}{\sqrt{2\pi}\sigma^3}\iint y' H(x'-x, y'-y)\exp\left(-\frac{x'^2+y'^2}{2\sigma^2}\right)dx'dy' \tag{32b}$$

Which can then be easily transformed into $\mathbf{t}(x,y)$ and $\mathbf{n}(x,y)$ via equations (**??**) and (**??**) respectively. For our discrete images I_{ij}, we form the discrete derivative distributions dIx_{ij} and dIy_{ij}.

$$dIx_{ij} = -\frac{1}{\sqrt{2\pi}\sigma^2}\frac{1}{NM}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}(u-i)H_{uv}\exp\left(-\frac{(u-i)^2+(v-j)^2}{2}\right) \tag{33a}$$

$$dIy_{ij} = -\frac{1}{\sqrt{2\pi}\sigma^2}\frac{1}{NM}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}(v-j)H_{uv}\exp\left(-\frac{(u-i)^2+(v-j)^2}{2}\right) \tag{33b}$$

With each discrete nematic tensor $\mathbf{n}_{ij}$ defined in terms of the tangential unit vector $\mathbf{t}_{ij}$:

$$\mathbf{t}_{ij} = \frac{(-dIy_{ij}\; dIx_{ij})}{\sqrt{(dIx_{ij})^2 + (dIy_{ij})^2}} \tag{34}$$

2.4 Non-Negative Matrix Factorisation (NMF)

We consider a matrix solely comprised of non-negative elements $\mathbf{V}$ to be able to be factorised into two smaller matrices $\mathbf{W}$ and $\mathbf{H}$.

$$\mathbf{V} = \mathbf{WH} \tag{35}$$

The features matrix $\mathbf{W}$, represents a set of key (hidden) features, and the coefficients matrix $\mathbf{H}$, provides the weighting of these features in the original $\mathbf{V}$. The factorisation relies on the property that all matrices do not possess any negative elements, making it particularly applicable to image analysis. Solutions to the matrices $\mathbf{H}$ and $\mathbf{W}$ are estimated by the following criteria, where $||A||_F$ signifies the Frobenius norm (equation (37)).

$$\min_{W,H} ||V - WH| |_F^2 \qquad \text{where} \qquad W \geq 0, H \geq 0 \tag{36}$$

$$||A||_F^2 = \sum_{i,j} A_{ij}^2 \tag{37}$$

In practice, there are several algorithms available to find a solution to equation (36). One of the most popular and straight forward methods is the multiplicative update rule, developed by Lee and Seung[12]. Using an initial (non-negative) guess of **W** and **H**, each index i, j is iteratively updated at step n using the following scheme, until both matrices are stable.

$$H_{ij}^{n+1} \leftarrow \mathbf{H}_{ij}^{n} \frac{((W^n)^T V)_{ij}}{((W^n)^T W^n H^n)_{ij}} \tag{38}$$

$$W_{ij}^{n+1} \leftarrow \mathbf{W}_{ij}^{n} \frac{(V(H^{n+1})^T)_{ij}}{((W^{n+1}H^{n+1}(H^{n+1})_{ij}^T} \tag{39}$$

References

[1] M. Raspanti, “Different architectures of collagen fibrils enforce different fibrillogenesis mechanisms.,” *Journal of Biomedical Science and Engineering*, vol. 3, pp. 1169–1174, 2010.

[2] M. Raspanti, M. Reguzzoni, M. Protasoni, and D. Martini, “Evidence of a discrete axial structure in unimodal collagen fibrils,” *Biomacromolecules*, vol. 12, no. 12, pp. 4344–4347, 2011.

[3] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, “A computer simula-

tion method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics*, vol. 76, no. 1, pp. 637–649, 1982.

[4] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen, "Efficient algorithms for langevin and dpd dynamics," *Journal of Chemical Theory and Computation*, vol. 8, no. 10, pp. 3637–3649, 2012.

[5] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[6] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, 1980.

[7] A. M. Garcia, F. L. Magalhes, J. S. Soares, E. Paulino-Jr, M. F. de Lima, M. Mamede, and A. M. de Paula, "Second harmonic generation imaging of the collagen architecture in prostate cancer tissue," *Biomed. Phys. Eng. Express*, vol. 4, p. 025026, 2018.

[8] P. Campagnola, "Second harmonic generation imaging microscopy: Applications to diseases diagnostics," *Analytical Chemistry*, vol. 83, no. 9, pp. 3224–3231, 2011.

[9] A. Boudaoud, A. Burian, D. Borowska-Wykret, M. Uyttewaal, R. Wrzalik, D. Kwiatkowska, and O. Hamant, "Fibriltool, an imagej plug-in to quantify fibrillar structures in raw microscopy images," *Nat. Protocols*, vol. 9, p. 457–463, 2014.

[10] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nihimage to imagej: 25 years of image analysis," *Nat. Methods*, vol. 9, p. 671–675, 2012.

[11] A. Ghazaryan, H. F. Tsai, G. Hayrapetyan, W.-L. Chen, Y.-F. Chen, M. Y. Jeong, C.-S. Kim, C. S-J., and D. C-Y., "Analysis of collagen fiber domain organization by fourier second harmonic generation microscopy," *J. Biomed. Opt.*, vol. 18, p. 031105, 2012.

[12] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *In NIPS*, pp. 556–562, MIT Press, 2000.