Evolutionary drivers of protein shape

Gareth Shannon\textsuperscript{1}, Callum R. Marples\textsuperscript{2}, Rudesh D. Toofanny\textsuperscript{3} & Philip M. Williams\textsuperscript{2}

Diffusional motion within the crowded environment of the cell is known to be crucial to cellular function as it drives the interactions of proteins. However, the relationships between protein diffusion, shape and interaction, and the evolutionary selection mechanisms that arise as a consequence, have not been investigated. Here, we study the dynamics of triaxial ellipsoids of equivalent steric volume to proteins at different aspect ratios and volume fractions using a combination of Brownian molecular dynamics and geometric packing. In general, proteins are found to have a shape, approximately Golden in aspect ratio, that give rise to the highest critical volume fraction resisting gelation, corresponding to the fastest long-time self-diffusion in the cell. The ellipsoidal shape also directs random collisions between proteins away from sites that would promote aggregation and loss of function to more rapidly evolving nonsticky regions on the surface, and further provides a greater tolerance to mutation.

Interactions between diffusing proteins are central to the function of the cell. The rate at which reactions of proteins occur in the cell is proportional to the product of their concentration and their rate of diffusion\textsuperscript{1}. As protein concentration increases their translational diffusion $D_t$ slows due to interactions in the increasingly crowded environment\textsuperscript{2}. For a system of hard spheres, which undergo a glass transition at a volume fraction $\phi_c \approx 0.583$, the translational diffusion constant falls with volume fraction $\phi$ approximately as $D_t \sim D_0 \phi^{-1/4}$. The product of diffusion constant and concentration is maximum when $\phi = 19\%$\textsuperscript{1,2}, a value similar to that of proteins in the cell\textsuperscript{5–7}.

The diffusion constant of a protein depends on its size and shape and on the shape of other macromolecules in its environment\textsuperscript{8}. The translational diffusion constant of a spherical particle is proportional to its hydrodynamic radius, and the diffusion constant is smaller for an aspherical particle compared to the sphere of the same volume. Spherical proteins with their larger diffusion constants would, therefore, be expected to give rise to higher biochemical reaction rates than aspherical proteins of equivalent volume. Proteins are, however, generally not spherical\textsuperscript{9}.

The glass transition point $\phi_c$ also depends on particle shape. Mode coupled theory predicts that $\phi_c$ is higher for spheroids (ellipsoids of revolution) than it is for spheres, suggesting spheroids prevent crystallization by raising the glass transition point\textsuperscript{10}, and that $\phi_c$ of dumbbell-shaped particles increased with asymmetry up to a maximum when their length/diameter is around 1.4\textsuperscript{11–13}. Simulations of the liquid-to-solid phase transition also suggest that $\phi_c$ is higher for spheroids that it is for spheres\textsuperscript{14,15}. Raising the value of $\phi_c$ for protein would increase the diffusion constant and biochemical reaction rate. Studying the relationship between particle shape and $D_t$ and $\phi_c$ and hence the possible relationships between protein shape, their concentration and their reactions within a cell is the subject of this work.

Here, we first study the aspect ratios of globular proteins by taking both calliper measurements and determining triaxial ellipsoids of equivalent steric volume and diffusive properties. Using a combination of molecular dynamics and ellipsoidal packing and unpacking simulations we show how the critical volume fraction and translational diffusion of these triaxial ellipsoids depends on their shape and concentration. Assuming the diffusion of proteins follows that predicted for these ellipsoidal models these simulations suggest that proteins have evolved an ellipsoidal shape that limits aggregation and gelation in situations of high concentration, and to maximise diffusion-limited processes within the cell. We further analyse the collisions between diffusing ellipsoids and suggest that proteins have evolved binding sites on their surface at regions that limit the likelihood of aggregation during collisions.

\textsuperscript{1}NASA Ames Research Center, Moffett Field, CA, 94043, USA. \textsuperscript{2}Molecular Therapeutics and Formulation, School of Pharmacy, University of Nottingham, Nottingham, NG7 2RD, United Kingdom. \textsuperscript{3}Arzeda Corp., 3421 Thorndyke Ave W, Seattle, WA, 98119, USA. Correspondence and requests for materials should be addressed to P.M.W. (email: phil.williams@nottingham.ac.uk)
Materials and Methods

47,677 structures from the Protein Data Bank clustered at 95% sequence identity using CD-HIT were downloaded from RCSB.org on 01/23/2018. The proteins in this set were taken as the single-chain, and multimerisation ignored. The 3D Complexes (single chain proteins and protein dimers), PDB Select (asymmetric units), Dynomeomics, and species subsets were also studied.

Calliper measurements. To determine the extent of asphericity of proteins Feret diameters were measured by determining the largest and smallest differences between the maximum and minimum atomic coordinate values in the x, y, and z planes for all orientations of a protein around the x and y planes between 0 and π/2 in π/180 increments. The maximum error in this estimation is 1 − cos(π/180) = 0.02%.

Ellipsoid of equivalent steric volume. To calculate the principle radii of the ellipsoid of the equivalent steric volume to a protein we calculated the moment of inertia of a body representing the atoms.

\[
I = \frac{4\pi}{3} \sum_{i=1}^{N} r_i^4 \begin{vmatrix}
 y_i^2 + z_i^2 + \frac{2}{5} r_i^2 & -x_i y_i & -x_i z_i \\
 -x_i y_i & x_i^2 + z_i^2 + \frac{2}{5} r_i^2 & -y_i z_i \\
 -x_i z_i & -y_i z_i & x_i^2 + y_i^2 + \frac{2}{5} r_i^2
\end{vmatrix}
\]

where \( x, y, \) and \( z \) are the positions of the \( N \) atoms in the protein, and \( r \) is their vdW-radii (C 1.7 Å; N 1.55 Å; O 1.52 Å; S 1.8 Å; H 1.2 Å; other 1.7 Å). This tensor is diagonalized to give the three eigenvectors of the principle axes and their eigenvalues \( \lambda_1 \geq \lambda_2 \geq \lambda_3 \). The lengths \( a, b, \) and \( c \) of the semi-axes of the ellipsoid of equivalent steric volume are then found as

\[
a = \frac{5}{2M} (\lambda_1 + \lambda_2 - \lambda_3),
\]

\[
b = \frac{5}{2M} (\lambda_1 + \lambda_3 - \lambda_2),
\]

\[
c = \frac{5}{2M} (\lambda_2 + \lambda_3 - \lambda_1),
\]

where \( M = \sum_{i=1}^{N} \frac{4\pi}{3} r_i^4 \). Here, we express the ratios of these axes lengths by the parameters \( \alpha = a/c \), the aspect ratio of the ellipsoid, and \( \alpha' = b/c \), which describes the shape from prolate (\( \beta = 0 \)) to oblate (\( \beta = 1 \)) spheroid. Code for this was written in C using diagonalization routines of Kopp (http://arXiv:physics/0610206).

Brownian and Langevin dynamics. HOOMD-blue (v2.1.1-92) was used to simulate the diffusive motion of soft spheroids at volume fractions between 0.001% and 55%. An initial starting configuration was generated by packing 300 randomly oriented spheroids to a volume fraction of 30% using PackLSD (see below). The position and orientation of the packed spheroids were then used to generate 8100 Gay–Berne particles by replicating in a 3 × 3 × 3 array. The Gay–Berne anisotropic pair potential used in HOOMD-blue is

\[
V_{GB}(\vec{r}, \vec{r}', \vec{c}, \vec{c}') = \begin{cases} 
4\varepsilon[(\zeta^{-12} - \zeta^{-6})] & \zeta < \zeta_{cut} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\zeta = \left( \frac{r - \sigma + \sigma_{min}}{\sigma_{min}} \right)
\]

\[
\sigma^{-2} = \frac{1}{2} \vec{r} \cdot \vec{H}^{-1} \cdot \vec{r}
\]

\[
\vec{H} = 2\ell_{\perp}^2 \vec{1} + (\ell_{\parallel}^2 - \ell_{\perp}^2)(\vec{c}_i \otimes \vec{c}_i + \vec{c}_j \otimes \vec{c}_j)
\]

\[
\sigma_{min} = \min(\ell_{\parallel}, \ell_{\perp})
\]

with \( \ell_{\parallel} \) and \( \ell_{\perp} \) set such that \( V(2a) = V(2c) = 1 \).

Dynamics of unit-volume spheroids was performed for 10,000 steps with \( k_B T = 1 \), translational and rotational friction factors \( \gamma = 1 \), timestep \( \delta t = 0.0001 \) whilst the size of the periodic box was changed to reach the specified volume fraction, the system equilibrated for a further 10,000 steps and then particle diffusion monitored over 250,000 steps \( (t = 25) \). Translational diffusion was determined from the mean-squared displacement \( d \) of the particles

\[
D_x = \frac{4D}{6t}
\]

Eq. (14) was fitted to the simulated diffusion data to determine \( \phi_0 \) using OriginPro (OriginLab Corporation, Northampton, MA 01060).

In-house code was used to study the location of collisions between ellipsoids in Brownian motion described by Eq. (5) and Eq. (6). Here, a point \( x, y, \) and \( z \) on the surface of an ellipsoid in its body frame of reference (i.e. axis a aligns with x, b with y and c with z) is defined by two angles \( \theta \) and \( \varphi \), where \( x = t \cos(\theta) \), \( y = t \sin \theta \cos(\varphi) \), \( z = t \sin(\theta) \sin(\varphi) \) with \( t \) given by
\[
t = \frac{abc}{\sqrt{b^2 c^2 \cos^2(\theta) + a^2 c^2 \sin^2(\theta) \cos^2(\phi) + a^2 b^2 \sin^2(\theta) \sin^2(\phi)}}.
\]

**Packing and unpacking.** Maximally random jammed packings of ellipsoids were generated using the neighbour list collision-driven molecular dynamics algorithm PackLSD of Donev\(^{(26)}\). Unpackings were performed by running PackLSD on an ordered array of ellipsoids, as described in ref.\(^{(26)}\), to decompress to a final volume fraction of 35%.

**Estimation of the critical volume fraction \(\phi_c\) for ellipsoids.** To estimate the critical volume fraction \(\phi_c\), we studied the liquid-to-solid phase transition of triaxial ellipsoids. For a system of hard spheres the phase diagram of pressure as a function of volume fraction exhibits a freeze point \(\phi_F \approx 0.494\), below which the suspension is a liquid\(^{(27)}\). Between \(\phi_F\) and \(\phi_M \approx 0.545\), the melt point, for a system in equilibrium solid and liquid coexist and above which the system is a crystal. Forcing a system above \(\phi_c\) quickly enough to preclude equilibration sees the system enter a supercooled state of liquid until the glass transition, \(\phi_G \approx 0.58\) is reached, and between this and the maximally random jammed state \(\phi_{MRJ} \approx 0.64\) the system behaves as a glass\(^{(14)}\). There is some debate as which value is best for the glass transition: the value \(\phi_c \approx 0.58\) or the maximally jammed state value \(\phi_{MRJ} \approx 0.64\). Most experiments point to \(\phi_c \approx 0.58\), for example eye-lens spherical \(\alpha\)-crystallin multimers \((\phi_c = 0.579 \pm 0.004)\)^\(^{(38,29)}\), although some suggest the value of 0.64 is the true value\(^{(30,31)}\). Santamaria-Holek and Mendoza used \(\phi_c = \phi_{MRJ}\) when predicting the relative viscosity of ellipsoids\(^{(32)}\). Here, we assumed \(\phi_c = \phi_{MRJ}\) regardless of its value.

To estimate \(\phi_c\) for triaxial ellipsoids we followed the method of Donev where a dense crystal arrangement of packed ellipsoids is unpacked and their order parameter and the pressure of the system monitored\(^{(26)}\). As the system was unpacked, the point at which order is lost precipitously was taken as the freeze point, \(\phi_F\), which also corresponds to an increase in pressure\(^{(24,25)}\). We took the melt point, \(\phi_M\), as the volume fraction of the unpacked ellipsoids that exerted the same pressure as this raised value of the freeze point. The value of \(\phi_{MRJ}\) was determined by packing the ellipsoids from the random (liquid) state obtained at the end of the unpacking. We then studied how ellipsoid diffusion varied when taking the values found for \(\phi_p\), \(\phi_M\), and \(\phi_{MRJ}\) as \(\phi_c\).

**Mapping.** Maps of the location of contacts between ellipsoids and of residue type were generated by representing each contact \(\theta\) point as a Gaussian spot. To calculate the Gaussian, geodesic distances on the surface of the ellipsoid between two integer \(\theta\) points were estimated by considering the map as an image of 180 \(\theta\) slices presented as a 2D image. For \(\phi = \phi_{MRJ}\) the system behaves as a glass\(^{(14)}\). For \(\phi = \phi_M\) the system behaves as a liquid\(^{(27)}\). Between \(\phi = \phi_F\) and \(\phi = \phi_M\) the maximally random jammed state \(\phi_{MRJ}\) is reached, and between this and \(\phi = \phi_{MRJ}\) the system behaves as a glass\(^{(14)}\). There is some debate as which volume fraction is best for the glass transition: the value \(\phi_c \approx 0.58\) or the maximally jammed state value \(\phi_{MRJ} \approx 0.64\). Most experiments point to \(\phi_c \approx 0.58\), for example eye-lens spherical \(\alpha\)-crystallin multimers \((\phi_c = 0.579 \pm 0.004)\)^\(^{(38,29)}\), although some suggest the value of 0.64 is the true value\(^{(30,31)}\). Santamaria-Holek and Mendoza used \(\phi_c = \phi_{MRJ}\) when predicting the relative viscosity of ellipsoids\(^{(32)}\). Here, we assumed \(\phi_c = \phi_{MRJ}\) regardless of its value.

**Positional evolutionary rates.** The rates of residue evolution were calculated for the 382 orthologous sequences taken from the genomes of *Saccharomyces cerevisiae* and nine closely related species: *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii*, *Saccharomyces kluveri*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Ashbya gossypii* using Rate4Site\(^{(33)}\) as described in ref.\(^{(34)}\). In that work, evolutionary rates are normalised to the average of all positions in all proteins in the set, and presented as log_{10}[normalised rate].

**Equations of diffusion.** Rotational and diffusional constants of triaxial ellipsoids have been derived to predict the translational and rotational diffusion of proteins\(^{(35-39)}\). Here, using the semi-axis lengths of the ellipsoid of equivalent steric volume we calculated the translational \(D^0_{ia,b,c}\) and rotational \(D^0_{ia,b,c}\) diffusion coefficients along and around each semi-axis as (in the example of a)\(^{(40)}\)

\[
D^0_{ia} = \frac{k_B T}{6\pi \eta} \left[ 3R_F(a^2, b^2, c^2) + R_G(b^2, c^2, d^2) a^2 \right],
\]

\[
D^0_{ia} = \frac{k_B T}{8\pi \eta} \left[ R_F(c^2, a^2, b^2) b^2 + R_G(a^2, b^2, c^2) c^2 \right],
\]

respectively, for a protein where \(a', b'\) and \(c'\) are the semi-axis lengths of the equivalent ellipsoid increased by an amount \(\delta H\) to reflect the width of a stationary hydration layer surrounding the protein, and \(R_F\) and \(R_G\) are the Carlson symmetric elliptic integrals of the first \(R_F(x, y, z)\) and second \(R_G(x, y, z)\) kind

\[
R_F(x, y, z) = \frac{1}{2} \int_0^\infty \frac{d\lambda}{\sqrt{(x + \lambda)(y + \lambda)(z + \lambda)}},
\]

\[
R_G(x, y, z) = \frac{3}{2} \int_0^\infty \frac{d\lambda}{(z + \lambda)\sqrt{(x + \lambda)(y + \lambda)(z + \lambda)}}.
\]
The algorithms to solve these were taken from Numerical Recipes\textsuperscript{41}. The long-time translational and rotational diffusion constants are the arithmetic mean of the values for each axis

\[
D^0_\alpha = \frac{(D^0_{\alpha a} + D^0_{\alpha b} + D^0_{\alpha c})}{3},
\]

\[
D^0_\beta = \frac{(D^0_{\beta a} + D^0_{\beta b} + D^0_{\beta c})}{3}.
\]

In terms of \(\alpha' = (a + \delta H)/(c + \delta H)\), \(\alpha'\beta = (b + \delta H)/(c + \delta H)\) and \(r' = (a'b'c')^{\frac{1}{3}}\)

\[
D^0_\alpha = \frac{k_B T}{6\pi \eta r'} R_p \left( (\alpha'^2 - \beta^2)^{\frac{2}{3}}, (\alpha'^2\beta^2 - 1)^{\frac{2}{3}}, (\alpha'^3 + 1)^{\frac{2}{3}} \right).
\]

**Diffusion at finite concentration.** Interactions between diffusing bodies leads to diffusive motion that is dependent on timescale and concentration\textsuperscript{42}. There are several descriptions of the correlation between intrinsic viscosity or long-time diffusion of spheres and volume fraction, all sharing a critical volume fraction at the divergent point\textsuperscript{43–45}. Tokuyama \textit{et al.} described the short-time self-diffusion of spheres by

\[
D^0_\phi(\phi) = \frac{D^0_\phi}{1 + L(\phi)},
\]

where \(\phi\) is the volume fraction of the particles in the suspension and \(L(\phi)\) defined as

\[
L(\phi) = \frac{2B^2}{1 - B} - \frac{C}{1 + 2C} - \frac{BC(2 + C)}{(1 + C)(1 - B + C)},
\]

with \(B = (9\phi/8)^{\frac{1}{2}}\) and \(C = 11\phi/16\). At longer times, a cage-effect of proteins surrounding others decreases diffusion further with proteins needing to transit from cage to cage. The long-time self-diffusion coefficient of a soft particle is described well by the expression

\[
D_\phi^L(\phi, \phi_c) = \frac{D_\phi^L(\phi)}{1 + \kappa \frac{D_\phi^L(\phi)}{D_\phi^L(\phi_c)}} \left( \frac{\phi}{\phi_c} \right) \left( \frac{\phi_c}{\phi} \right)^{-\kappa},
\]

where \(\phi_c\) is a singular point of the cessation of long-time self-diffusion; the critical volume fraction. Eqs (12) and (14) with a value \(\kappa = 2\) have been shown to describe the volume fraction dependence of translational diffusion for a number of proteins, where the function has been fitted to the experimental data to determine the value of \(\phi_c\)\textsuperscript{44–48}.

**Results and Discussion**

**Proteins are naturally aspherical and have an aspect ratio around 1.6.** The aspect ratio of proteins, calculated as the ratio of the largest-to-smallest calliper diameters, is broadly distributed around 1.6 and ranges from 1.2 to 18.6 (Fig. 1). The distribution of aspect ratio of the single chains of 47,677 PDB structures clustered at 95% sequence similarity is represented well by a log-lognormal with a modal value of 1.639 (±0.001). We divided the aspect ratio distributions into those of proteins that are generally prolate and of those generally oblate by determining the moment of inertia of the protein represented by its atoms as unit density van-der Waals radius spheres and determining the semi-axis lengths \(a \geq b \geq c\) of the ellipsoid with the same inertia; the so-called ellipsoid of equivalent steric volume\textsuperscript{20}. We found the aspect ratio of the calliper measurements as longest-to-shortest of prolate and shortest-to-longest of oblate are distributed around 1.6 and 0.7 (1/1.6), respectively (Fig. 1A). The modal aspect ratios are similar to that of the Golden ratio \(\Phi = (1 + \sqrt{5})/2 = 1.618...\), and its reciprocal \(\Phi^{-1}\). The distributions of aspect ratios of the 13,052 monomers (3D Monomers), 19,148 dimer components (3D Dimers A & B) and their 9,574 dimeric complexes (3D Dimers AB) of the structures in 3D Complex database\textsuperscript{16}, 3272 single chain asymmetric units in the PDBselect database\textsuperscript{17}, 701 proteins in the Dynameomics database\textsuperscript{18}, and 1,243 proteins from different species (667 \textit{H. sapiens}; 396 \textit{E. coli}; 180 \textit{S. cerevisiae})\textsuperscript{19} are similarly shaped (Fig. 1). The modes of these distributions are given in Table 1.

This separation of the distribution shows that approximately 25% of proteins are oblate, a value similar to that found by Dima and Thirumalai who studied the proteins in the PDBselect\textsuperscript{17} subset of the Protein Data Bank\textsuperscript{8}. As an unfolded chain is generally prolate we may question why some proteins fold to oblate structures as such would need to undergo significant changes in volume and/or surface area on folding\textsuperscript{49–51}. This misconception arises from the approximation to purely oblate and prolate spheroids of revolution that suggests the change between the two requires a transition through a spherical state with corresponding differences in surface area or volume. The distribution of protein shape is better represented in two dimensions of \(\alpha\beta\), where \(\alpha = a/c\) and \(\beta = b/c\). Figure 2 shows this 2D distribution for the 47,677 protein chains. The dashed line \(\beta = (\ln(\alpha + 1) - 2)/(\ln(\alpha))\) for \(\alpha > 1\) shows the boundary between prolate (\(\beta = 0\); lower) and oblate (\(\beta = 1\); upper) proteins. We also plot (dotted) contours of the isoperimetric quotient \(36\pi V^2/A^3\) that represent ellipsoids of equivalent surface area and volume\textsuperscript{52}. The \(\alpha\beta\) aspect ratios of proteins are broadly distributed around a modal value of \(\alpha = 1.63\) and \(\beta = 0.34\). These values of aspect ratio correspond to the asphericity parameter \(\Delta \approx 0.1\) reported in ref.\textsuperscript{7}.

The surface area \(A\) of an ellipsoid defined by \(a\), \(b\) and \(c\) with volume \(V = \frac{4}{3}\pi abc\) is approximately...
with \( p = 1.6075 \), which for all values of \( \beta \) increases monotonically with \( \alpha \) from the smallest (sphere) value at \( \alpha = 1 \). For an ellipsoid of fixed volume and surface area there is a unique value of \( \alpha \) for each value of \( \beta \) between 0 and 1. Notwithstanding constraints due to necessary rearrangements of the polypeptide chain a molten globule could morph from prolate to oblate maintaining constant volume and surface area (i.e. along an isoperimetric contour). This suggests, therefore, that the shape of the folded protein may not necessarily reflect the general shape of its denatured state. Interestingly, \( A \) has a minimum in \( \beta \) between 0 and 1 for values of \( \alpha > 1 \), so the transition between prolate and oblate spheroids at constant volume and surface area requires a seemingly paradoxical increase in aspect ratio \( \alpha \). A value of \( \alpha > 1 \), where the surface area of the protein is greater than the sphere of equivalent volume, affords a greater tolerance to mutation than a spherical protein would, as any mutations which cause a change in volume can be accommodated without a change in surface area, and vice versa\(^{33} \).

The volume of the ellipsoidal approximation of protein was found to scale with the number of residues \( N \) as \( V \approx 203N \, \text{Å}^3 \) (Fig. 3C), equal to an equivalent radius of 3.1 Å-per-residue at a packing density of 0.64, and
the surface area of this ellipsoid of equivalent steric volume approximates to $A \approx 47N_s \text{Å}^2$ (equivalent radius of 3.7 Å-per-residue at the maximum disc packing density of 0.9), where $N_s$ is the number of residues at the surface. These findings are in agreement with those of others. As expected, the extent to which amino acids are buried was found to be correlated with their hydrophobicity and, in general, anticorrelated with their ‘stickiness’ as defined in ref.19 (Fig. 3A, B).

The average fractional radial extent of the alpha carbon of each residue toward the surface of the equivalent ellipsoid afforded an estimation of the likelihood that an amino acid of a type would be at the surface. By summing the product of these values by the number of each residue type in the protein we could estimate $N_s$, and hence surface area (Fig. 3D). This estimation could be improved by using the actual frequencies of each residue type being at the surface. It is possible, therefore, to estimate both volume and surface area of a protein’s equivalent ellipsoid, and therefore, from Eq. (15) its aspect ratio from knowledge of the amino acid content alone. Since the volume of a protein is related to the number of amino acids and its surface area related to the amino acid composition, Eq. (15) reveals that a protein aspect ratio $\alpha$ is dependent on the fractional content of surface-exposed residues. This indicates that for a chain of a given length there are far more different amino acid compositions that give rise to ellipsoids than give rise to spheres, and therefore suggest that proteins are naturally ellipsoidal irrespective of secondary structure. To support this we calculated the order parameter $\theta = <\cos \theta>$ where $\theta$ is the angle between each secondary structural unit (helix or sheet) and the $a$ axis of the protein (all, and

### Table 1. Calliper aspect ratios and shape parameters measured for various subsets of proteins in the PDB.

| Dataset                    | Number | Calliper oblate | Calliper prolate | Equivalent ellipsoid $\alpha$ | $\beta$ |
|----------------------------|--------|-----------------|-----------------|------------------------------|--------|
| All PDB (single chains)    | 47 677 | 0.649           | 1.676           | 1.65                         | 0.34   |
| D Complex (dimers A & B)   | 19 148 | 0.635           | 1.676           | 1.54, 1.78                   | 0.44, 0.44 |
| D Complex (dimers AB)      | 9 574  | 0.627           | 1.749           | 1.93                         | 0.17   |
| D Complex (monomers)       | 13 052 | 0.659           | 1.612           | 1.35, 1.65                   | 0.45, 0.30 |
| PDBselect                  | 3272   | 0.664           | 1.657           |                              |        |
| Dynameomics                | 701    | 0.620           | 1.665           |                              |        |
| H. sapiens                 | 667    | 0.658           | 1.602           |                              |        |
| E. coli                    | 396    | 0.630           | 1.588           |                              |        |
| S. cerevisiae              | 180    | 0.678           | 1.583           |                              |        |

**Figure 2.** The distribution of the aspect ratios for the ellipsoids of equivalent steric volume of the 47 677 structures taken from the PDB database (single chains clustered at 95% sequence similarity). For each ellipsoid with semi-axis lengths $a \geq b \geq c$, $\alpha = a/c$ and $\alpha^2 = b/c$. The Gaussian-smoothed distribution has a maximum at $\alpha = 1.65, \beta = 0.34$. The dashed line left-to-right for $\alpha > 1$ shows the boundary between generally prolate ($\beta = 0$) and generally oblate ($\beta = 1$) structures. 75% of the structures are generally prolate. The dotted isoperimetric contour lines between values of $\beta$ represent ellipsoids of equal volume and surface area.
The average fractional radial distance of the $C_{\alpha}$ of each type of residue from the core to the surface of a protein (A orange) is independent of protein shape and is ordered similarly to the frequency at which they are found at the surface (A magenta; B closed symbols), and opposite to their stickiness (A violet; B open symbols). (C) The volume of the ellipsoid of equivalent steric volume can be accurately predicted from the number of residues. (D) The number of residues at the surface of the protein can be predicted from the sum of each residue content and their surface propensities (partial least squares with one component, closed symbols; Bayesian Regularized Neutral Network with 20 neurons, open symbols).

The ellipsoids of equivalent steric volume allow the accurate prediction of protein diffusion. Calculated values of the translational and rotational diffusion constants of the equivalent steric triaxial ellipsoid of proteins, using Eqs (5) and (6), respectively, matched the experimentally determined values after increasing the semi-axis lengths of the ellipsoid by 2.32 Å, for PDB structures without hydrogens, or 2.30 Å for those with (translation rRMSE = 4.7%, rotation rRMSE = 9.1%) (Fig. 6A). Thus, the ellipsoid of equivalent steric volume with a hydration layer of 2.3 Å, equivalent to a hydration level of 0.38 g/g (volume of equivalent protein divided by mass of equivalent protein) between 0.35 and 0.45 has a correlation coefficient of 0.95 with experimentally determined values.

By considering the volume of the protein is related to its chain length and its surface area related to its composition we suggest that the general ellipsoidal shape of a protein (and its molten globule) is inherent. The fact that proteins can fold successfully without the assistance of chaperones and can successfully refold following denaturation and hence not co-translationally proximal to the ribosomal vestibule, suggests that their shape is not defined by the physical environment in which they fold. We estimated the extent of this ellipticity of proteins by considering the random sequence polypeptide chain as a binary sequence of either surface (polar) or buried (hydrophobic) residues, similar to a HP polymer model, which are also been to collapse into ellipsoidal shapes. By the binomial theorem shows that the largest number of combinations of H and P is when they are in equal amounts. The dependence of aspect ratio $\alpha$ on chain length of this simple model when $\beta = 0.5$ is shown in Fig. 5A. For a chain of 400 residues, equal to the average length in the human proteome, the most frequent composition has 200 surface exposed residues giving rise to ellipsoid aspect ratios for varying $\beta$ of $\alpha_{\beta=0} = 1.57$ to $\alpha_{\beta=0.4} = 1.66$ to $\alpha_{\beta=0.6} = 1.52$ (coincident with the contour line through the maximum of the distribution in Fig. 2). The ability of this model protein to accommodate mutations causing changes in volume and/or surface area is shown in Fig. 5B.

Taking $N_s/N$ to be 0.55, the value most found in the proteins and equal to the eleven out of twenty residues that have a surface preference value exceeding 0.5 (Fig. 3A), we find the length of protein where the predicted surface area is at least equal that of the sphere of equivalent volume is $36\pi(203)^2/0.55^347^3 = 270$ (the value of $N$ when in Eq. 1 $A = 47(0.55N)$, $V = 203N$, and $\alpha = 1$) residues. Proteins maintaining a 0.55/0.45 surface/core ratio of residues are predicted, therefore, to be at least approximately 270 residues in length. Larger proteins can be formed maintaining this surface/core ratio by becoming elliptical, but smaller proteins only formed through an increased prevalence of surface-preferred residues. Eukarya have median protein lengths longer at 361 residues, bacteria at 267 residues, and archaea are shorter at 247 residues.

The dependence of aspect ratio $\alpha$ on chain length of this simple model when $\beta = 0.5$ is shown in Fig. 5A. For a chain of 400 residues, equal to the average length in the human proteome, the most frequent composition has 200 surface exposed residues giving rise to ellipsoid aspect ratios for varying $\beta$ of $\alpha_{\beta=0} = 1.57$ to $\alpha_{\beta=0.4} = 1.66$ to $\alpha_{\beta=0.6} = 1.52$ (coincident with the contour line through the maximum of the distribution in Fig. 2). The ability of this model protein to accommodate mutations causing changes in volume and/or surface area is shown in Fig. 5B.

Figure 3. (A,B) The average fractional radial distance of the $C_{\alpha}$ of each type of residue from the core to the surface of a protein (A orange) is independent of protein shape and is ordered similarly to the frequency at which they are found at the surface (A magenta; B closed symbols), and opposite to their stickiness (A violet; B open symbols). (C) The volume of the ellipsoid of equivalent steric volume can be accurately predicted from the number of residues. (D) The number of residues at the surface of the protein can be predicted from the sum of each residue content and their surface propensities (partial least squares with one component, closed symbols; Bayesian Regularized Neutral Network with 20 neurons, open symbols).
ellipsoid with boundary layer $\approx 2.39 M_p$, density of bound water 1.104 g/mL), facilitates the prediction of the translational and rotational diffusive properties of globular proteins, indicating the effective hydrodynamic radii of a protein for each of its semi-axes is equivalent to this ellipsoid of equivalent steric volume. The hydration value of 0.38 g/g is commensurate with a monolayer coverage of water. Taking the example of lysozyme with equivalent ellipsoid semi-axis lengths of $a = 23.4$ Å, $b = 15.3$ Å and $c = 13.9$ Å ($\alpha = 1.69$, $\beta = 0.19$) we calculate a hydration shell volume of at most $9874$ Å$^3$ containing up to $365$ water molecules. Microparticle dehydration studies estimate that lysozyme is hydrated by a stationary layer of $\approx 380$ water molecules$^{60}$. By noting that the volume of the equivalent hydrated ellipsoid $\frac{4}{3} \pi r^3 \approx 2.39 M_p$, and that $\alpha = 1.65$ and $\beta = 0.34$, we can approximate the translational and rotational diffusion constants from protein length $N$ or weight $M_p$ as (20°C, $\rho = 1.0016$ mPa s)

$$D^0_{\pi} \approx \frac{52}{\sqrt{N}} \approx \frac{252}{\sqrt{M_p}} \times 10^{-7} \text{ cm}^2 \text{ s}^{-1},$$

(16)
The numerator in this approximation of the translational diffusion coefficient lies between the value of 244 given by Young-Carroad-Bell and 285 given by Polson. It was noted by Hem and Neimeyer that the equation derived by Tyn and Gusek for the approximation of diffusion based on a protein's radius of gyration implies a spheroidal geometry with an aspect ratio of 1.4 if prolate and 0.66 (1/1.5) if oblate.

Diffusion within the crowded cell is greater for ellipsoidal proteins. The Brownian and Langevin dynamics simulations of the diffusion of soft spheroidal particles, represented by the Gay-Berne potential, at various volume fractions confirmed the aspect ratio dependence of the diffusion of spheroids. Figure 6B is a plot of the critical volume fraction $\phi_c$ from both simulation methods show an identical dependence on shape (Fig. 6C), starting at a value of 0.565 for $\alpha = 1.0$ (spheres) increasing to a maximum near $\alpha = 1.6$.

The crystal unpacking simulations gave values for $\phi_F$, $\phi_M$ and $\phi_{MRJ}$ for spheres as 0.49, 0.56 and 0.68, respectively (Fig. 7). The value of $\phi_F$ is maximal at $\alpha = 1.71$, $\beta = 0.4$. The value of $\phi_M$ has three maxima in this $\alpha\beta$ profile, with the highest of $\phi_M = 0.68$ at $\alpha = 1.72$, $\beta = 0.36$. As an extrapolation, however, we expect a degree of variability in the determination of $\phi_M$. Ellipsoids with shape $\alpha \approx 1.7$, $\beta \approx 0.4$ are expected, therefore, to have the highest value of $\phi_M$, which suggests that the translational diffusion proteins of a similar...
shape is least retarded by crowding. Proteins of aspect ratio of $\alpha \approx 1.7$ and $\beta \approx 0.4$ have both optimal diffusion and limited propensity to form a glass when at high concentration. Such heterogeneous crowding in the cell can turn diffusion from normal to anomalous, where mean-squared displacement is no-longer linear in time $^{62}$. A more complete model of the dependence of diffusion-limited reactions on cellular concentration that incorporates the

Figure 7. Estimation of the critical freeze $\phi_{F}$, melt $\phi_{M}$, and maximum jammed $\phi_{MRJ}$ volume fractions. The ordered crystal state for each ellipsoid defined by $\alpha \beta$ (top) is unpacked during which both the pressure of the system (blue) and order (black) are recorded (left). At the freeze point there is a slight increase in pressure as order is lost precipitously. The melt point is taken as the more packed state that exerted the same pressure as the raised value at the freeze point. After unpacking, the system is then packed to determine the maximum random jammed state (left, red).
subdiffusion due to crowding is required\textsuperscript{63,64}, although we expect the dependence on aspect ratio of the diffusing particles to be similar to that found here.

Proteins are ellipsoidal to maximise their translational diffusion in the crowded cell. A biomolecular reaction limited by diffusion has a rate proportional to the product of the relative diffusion constant (i.e. the sum of the diffusion constants of the reactants) and the reactant concentrations. By combining Eqs (11), (12) and (14) we found the product $\phi_D(\alpha, \beta)$ is maximum at $\alpha = 1.58$ and $\phi = 0.19$ for prolate spheroids, and is maximum when $\alpha = 1.60$, $\beta = 0.5$, $\phi = 0.20$ for ellipsoids of equivalent steric volume, we recover the cell volume fraction of 19% and additionally the optimum protein aspect ratio of $\alpha/c \approx 1.7$ and $b/c \approx 1.3$, similar to that measured of proteins that form dimers (Table 1).

The predicted translational diffusion of the protein at a volume fraction of 19% is 35% of its value at infinite dilution, giving a simple expression for the approximate translational diffusion $D_{\text{cell}}(x, y, z)$ of a protein in the cell at 37 °C, $\rho = 0.6913$ mPa s of

$$D_{\text{cell}} \approx \frac{28}{\sqrt{N}} \approx \frac{135}{\sqrt{M_p}}.$$

Ellipsoidal shape helps proteins avoid non-functional interactions. Figure 9B shows the normalized frequency distribution in $\theta, \varphi$ space of the location of collisions between ellipsoids ($\alpha = 1.78$, $\beta = 0.44$) undergoing Brownian dynamics. Each collision at $\theta, \varphi$ on the surface of the ellipsoid is represented by a Gaussian spot with variance 10° arc. The locations of the collisions are distributed unevenly across the surface with the fewest collisions made at the poles of the $c$ axis.

The distribution is similar in form to that of the effective translational diffusion constants over short times (i.e. dominated by the rotation) $D_{\varphi}(x, y, z)$ of points across the surface of the ellipsoid (Fig. 9A), calculated as
Figure 9. Proteins have surfaces that are nonsticky at the sites where they are most likely to make collisions during diffusion. (A) The effective translational diffusion over short time of points on the surface of a tumbling ellipsoid, defined by two angles $\theta$ from the $a$ axis and $\varphi$ from $b$, is lowest (light) at the poles of the $c$ axis, the flattest face, and highest (dark) at the poles of the $a$ axis. (B) Plot of the normalized distribution of collision frequency (low light to high dark) for each $\theta, \varphi$ point on the surface of a diffusing ellipsoid. (C) Plot of the frequency distribution of the location of points at the centre of the interface between monomers in a protein dimer (low light to high dark). (D) Plot of the average radial extent values (Fig. 3A) for the residues found on the surface of the monomers in protein dimers. (E) The most conserved residues on the surface of homologs are coincident with the centre of the pole at $c$ where binding and active sites are common. (F) The $a/c$ cross-section through the evolutionary rate data reveals that mutations of the faster-evolving residues on the surface away from the active site facilitate changes in the core.

$$
D_{\psi}(x, y, z) \approx D_t + \frac{1}{2} \left[ D_{tb}(y^2 + z^2) + D_{ta}(x^2 + z^2) + D_{ta}(x^2 + y^2) \right].
$$

(19)
Points that are translating the quickest, i.e. the poles of the $a$ axis, make more contacts than those which are translating the slowest, i.e. the poles of the $c$ axis. Molecular dynamics simulations support this importance of rotational diffusion in directing collisions between proteins.

We compared the collision probabilities to the points of the centres of protein binding patches measured from the 3D Complex database. We calculated the ellipsoid of equivalent steric volume of each partner in a binary complex and determined their centre of contact by shrinking the ellipsoids until their surfaces just touched. Figure 9C shows the distribution of the $\theta\varphi$ contact points of 9958 protein pairs, where each point is mapped as a Gaussian-spot with variance $10^\circ$ arc across the surface. Figure 9D shows the map of the average extent to which the amino acid at the contact point is normally found at the surface (Fig. 3A), which reflects the patch stickiness as calculated by Levy et al. Protein–protein interaction sites were found predominately on the face of the poles of the $c$ axis, at residues that are generally normally found more buried in a protein. Whilst the face at $c$ presents the greatest surface area it is the area involved in the fewest collisions during diffusion. An ellipsoidal shape to protein directs random collisions between them to nonsticky regions of their surface and thereby limits the formation of non-functional interactions and aggregation.

A protein's interaction or active site is expected to be the most conserved during evolution, since mutation of a residue at the site is more likely to lead to loss of function than a mutation elsewhere on the surface. Therefore, we expected the residues at the poles of the $c$ axis to be most conserved in homologs across species of known phylogeny. Following the method of Tóth-Petróczy and Tawfik we calculated the evolution rates per-position of 382 protein domains of known structure in orthologs in 10 yeast species. In Fig. 9E we plot the average rates of the surface residues mapped onto $\theta\varphi$ (the protein is rotated around $a$ by $0^\circ$ or $180^\circ$ so the face at $c$ with the slower rates is at $0 < \varphi < 180$). The residues around the centre of the $c$ face at $\theta = 90^\circ$, $\varphi = 90^\circ$, coincident with the location of the active site most often, were found to have an evolution rate around half that of the other surface residues. A cross section taken through the data taken at $\varphi = 90^\circ$ highlights the association of evolutionary rates between the residues at the surface and those at the core. The evolutionary rates of residues decrease with their distance from the surface with core residues exchanging on average fourfold slower than those on the surface (away from the centre of $c$ at $\theta = 90^\circ$, $\varphi = 90^\circ$). The core region of the protein that is most conserved is offset from the centre of the protein towards the interface. This too is as expected following the surface–core association of evolutionary rates revealed by Tóth-Petróczy and Tawfik. Mutations of the faster evolving residues on the surface away from the active site facilitate changes in the core. The rate of evolution decreases with distance from the surface, converging on a point that is offset the $c$ axis (coincident with $\Phi$). If core mutations were facilitating surface changes we would expect the most conserved residues to be found at the centre of the protein, which is not the case.

Summary

We have shown that proteins are generally aspherical with an aspect ratio distributed around 1.6; a value close to the Golden ratio. We have shown that proteins of random sequence greater than 270 residues in length are likely to be ellipsoidal irrespective of secondary structure and that a simple model of a protein of 400 residues has a shape similar to that most common in the PDB.

We have shown that the translational and rotational diffusion of proteins can be accurately modelled by considering the protein as a triaxial ellipsoid of equivalent steric volume. We have shown that the translational diffusion of such ellipsoids decreases with increasing concentration and that the critical volume fraction corresponding to the glass transition, where translational diffusion becomes negligible, is dependent on the shape of the ellipsoid. We found that the shape of ellipsoid that gives the highest predicted value for the critical volume fraction is coincident with the modal value found for proteins. We suggested, therefore, that proteins have a shape that maximizes their translational diffusion within the cell and limits the likelihood of gelation at high concentration.

We studied the location of contacts between diffusing ellipsoids and found a correlation between the location where collisions are least likely to occur and where the protein/protein interface is found in protein dimers. We found the same association between the type of residues found at these interaction sites as others, and showed in a series of proteins that the residues that are found at these sites have evolved less quickly than those on other regions of the protein where random collisional contacts are more likely to occur. We suggested that proteins have evolved a shape which directs non-functional random collisions away from their sites of interaction to non-sticky residues that are least conserved.

References

1. Schmit, J. D., Kamber, E. & Kondov, J. Lattice model of diffusion-limited biomolecular reactions in confined environments. *Phys. Rev. Lett.* **102**, 218302 (2009).
2. Dill, K. A., Ghosh, K. & Schmit, J. D. Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. USA* **108**, 17876–17882 (2011).
3. Weeks, E. R., Crocker, J. C., Levitt, A. C., Schofield, A. & Weitz, D. A. Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. *Science* **287**, 627–631 (2000).
4. Tokuyama, M., Moritik, T. & Kimura, Y. Self-diffusion of biomolecules in solution. *Phys. Rev. E* **83**, 051402 (2011).
5. Fulton, A. B. How crowded in the cytoplasm? *Trends Biochem. Sci.* **26**, 597–604 (2001).
6. Balbo, J., Merighetti, P., Herten, D.-P. & Wade, R. C. The shape of protein crowders is a major determinant of protein diffusion. *Biophys. J.* **104**, 1576–1584 (2013).
7. Dima, R. I. & Thirumalai, D. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B* **108**, 6564–6570 (2004).
8. Jain, A. K. Macromolecular crowding: Obvious but underappreciated. *Trends Biochem. Sci.* **26**, 597–604 (2001).
9. Balbo, J., Merighetti, P., Herten, D.-P. & Wade, R. C. The shape of protein crowders is a major determinant of protein diffusion. *Biophys. J.* **104**, 1576–1584 (2013).
10. Dima, R. I. & Thirumalai, D. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B* **108**, 6564–6570 (2004).
11. Chong, S. H. & Gotze, W. Idealized glass transitions for a system of dumbbell molecules. *Phys. Rev. E* **65**, 041503 (2002).
61. Polson, A. & Vanderreyden, D. Relationship between diffusion constants and molecular weight. *Biochim. Biophys. Acta.* 5, 358–366 (1950).
62. Ghosh, S. K., Cherstvy, A. G., Grebenkov, D. S. & Metzler, R. Anomalous, non-gaussian tracer diffusion in crowded two-dimensional environments. *New J. Phys.* 18, 013027 (2016).
63. Weiss, M., Elsner, M., Kartberg, F. & Nilsson, T. Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells. *Biophys. J.* 87, 3518–3524 (2004).
64. Matzler, R. & Kläffer, J. The restaurant at the end of the random walk: Recent developments in the description of anomalous transport by fractional dynamics. *J. Phys. A: Math. Gen.* 37, R161–R208 (2004).
65. Nawrocki, G., Karaboga, A., Sugita, Y. & Feig, M. Effect of protein-protein interactions and solvent viscosity on the rotational diffusion of proteins in crowded environments. *Phys. Chem. Chem. Phys.* 21, 876–883 (2019).

**Acknowledgements**

This work was supported by the Biotechnology and Biological Sciences Research Council and the University of Nottingham via a PhD sponsorship for Gareth Shannon; and the Engineering and Physical Sciences Research Council and the University of Nottingham via PhD sponsorships for Callum Marples and Rudesh Toofanny. High-performance computing infrastructure and support were provided by the University of Nottingham.

**Author Contributions**

R.D.T. posed the question whether the Golden ratio is found in proteins; G.S. and P.M.W. designed the research; C.R.M., G.S. and P.M.W. performed the research; C.R.M., G.S. and P.M.W. analysed the data; and G.S. and P.M.W. wrote the manuscript. All authors reviewed the manuscript.

**Additional Information**

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019