Diffusion Coefficients of Endogenous Cytosolic Proteins from Rabbit Skinned Muscle Fibers

Brian E. Carlson,[†] Jim O. Vigoreaux,[‡§] and David W. Maughan[‡]*

[†]Department of Molecular and Integrative Physiology, University of Michigan, Ann Arbor, Michigan; [‡]Department of Molecular Physiology and Biophysics, Health Science Research Facility, University of Vermont College of Medicine, Burlington, Vermont; and [§]Department of Biology, University of Vermont, Burlington, Vermont

SUPPORTING MATERIAL

A. Table of Symbols

Symbol	Description	Units
C	Protein concentration	μM
D	Diffusion coefficient of protein	cm^2 s^{-1}
r	Radial position in fiber	μm
t	Elapsed time of diffusion	s
a	Outer fiber radius	μm
C_0	Initial concentration of protein	μM
D_w	Diffusion coefficient of protein in aqueous solution	cm^2 s^{-1}
D_h	Diffusion coefficient of protein with steric hindrance	cm^2 s^{-1}
R_h	Hydrodynamic radius of protein	nm
R_o	Average myofilament radius	nm
L	Average half center to center spacing of myofilaments	nm
D_c	Diffusion coefficient of protein in a crowded cytosol	cm^2 s^{-1}
γ	Ease of vacancy of protein	unitless
V_h	Hydrodynamic volume of protein	dL g^{-1}
V_h^b	Average hydrodynamic volume of background proteins in cytosol	dL g^{-1}
ρ	Total protein density in the cytosol	g dL^{-1}
D_b	Diffusion coefficient of protein considering binding to cytomatrix	cm^2 s^{-1}
k_b	Average apparent binding constant of proteins to cytomatrix	unitless
k_s	Average apparent binding constant of supramolecular protein complex	unitless
$k_s^{\max}$	Maximal apparent binding constant of supramolecular protein complex	unitless
ρ_{50}	Total protein density at half maximal binding of the protein complex	g dL^{-1}
n_s	Supramolecular protein complex cooperativity parameter	unitless
M	Amount of protein diffused out of fiber	μmoles
M_∞	Total amount of diffusible protein in the fiber	μmoles
t/a^2	Diffusion time in fiber scaled by a cross-sectional area parameter	s/cm^2
$\overline{a}$	Average fiber radius over all experimental fibers	μm
φ	Ease of vacancy – hydrodynamic volume product	dL g^{-1}
k	Boltzmann's constant	J °K^{-1}
T	Experimental temperature	°K
η_w	Viscosity of water at 7°C	cP
ΔG	Free energy of protein binding to cytomatrix	kJ mol^{-1}

Table S1. Complete table of symbols, description and units used.

B. Test for Oil Dissipation Artifact

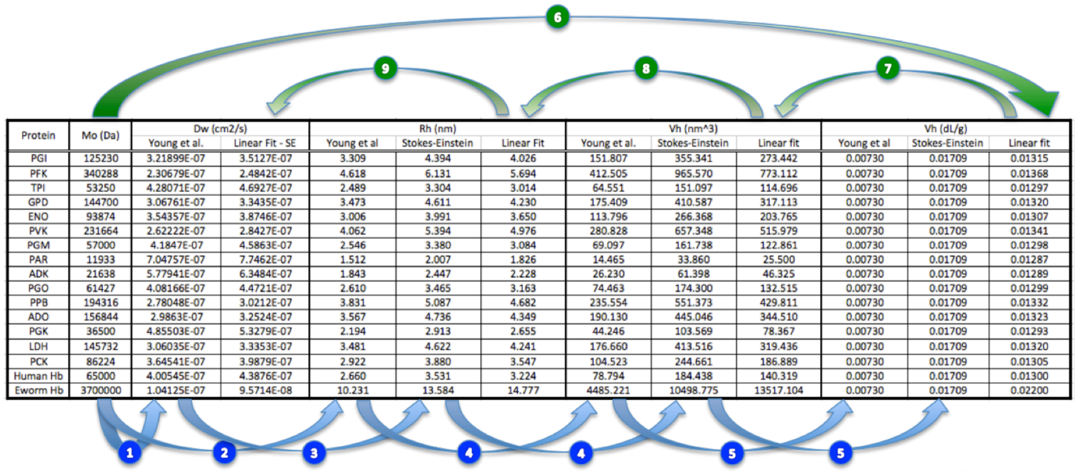
A freshly skinned fiber was transferred from oil to a drop of relaxing solution and, after two seconds, returned to oil. During this brief period in solution 3-9% (0.03-0.09: column 5) of the diffusible protein left the fiber, the exact amount depending on the protein species and fiber diameter. Following a 60 s equilibration period in oil, the fiber was then transferred to a second drop of relaxing solution for two seconds. That is, the initial conditions were re-established and the procedure repeated, during which the fiber was depleted of a slightly greater amount of diffusible protein (0.04-0.12: column 6). Finally, the fiber was transferred rapidly through oil to a third drop of relaxing solution for a final washout period of ten minutes. Values of $D_{t/a2=0.049}$ (i.e., D of the Constant D model evaluated at t/a^2=0.049 × 10^8) were calculated for a representative set of proteins by applying the Constant D model to drops 1-3 and to drops 2-3, respectively, as explained in the table caption. $D_{t/a2=0.049}$ derived from the second transfer (column 3) exceeded those from the first (column 2) by a factor of 2.8±0.6 (column 4). The fact that a significantly higher value of $D_{t/a2=0.049}$ was observed after a reduction in protein density is consistent with the breakup and dissipation of a protein complex and not with a diffusional delay due to dissipation of a layer of oil or polarized water (see main text)..

1	2	3	4	5	6
	D^1	D^2	D^2/D^1	$M^1/(M^1+M^2+M^3)$	$M^2/(M^2+M^3)$
Phosphoglucose isomerase	1.2 (1.0)	2.2 (1.9)	2.7 (2.0)	0.03 (0.02)	0.04 (0.03)
Triosephosphate isomerase	1.9 (1.3)	5.0 (4.7)	2.5 (0.7)	0.04 (0.02)	0.07 (0.05)
Pyruvate kinase	4.8 (5.8)	6.9 (3.9)	3.5 (2.3)	0.07 (0.05)	0.09 (0.04)
Enolase	2.1 (1.6)	8.5 (9.5)	3.8 (2.3)	0.04 (0.02)	0.09 (0.07)
Glyceraldehyde-3-phosphate dehydrogenase	4.1 (3.5)	10.0 (8.6)	3.5 (1.7)	0.06 (0.04)	0.11 (0.07)
Phosphoglycerate mutase	1.3 (1.4)	2.3 (2.5)	2.4 (2.2)	0.03 (0.03)	0.04 (0.03)
Phosphoglucose mutase	2.9 (2.6)	5.3 (5.7)	3.0 (2.2)	0.05 (0.04)	0.07 (0.06)
Adenylate kinase	1.8 (1.4)	3.5 (3.4)	2.0 (1.3)	0.04 (0.02)	0.06 (0.04)
Parvalbumin	5.3 (2.8)	12.0 (8.9)	2.2 (0.8)	0.09 (0.03)	0.12 (0.06)

Table S2. Comparison of diffusion coefficient $D_{t/a2}$ calculated before and after re-equilibration period in oil. Column 2 and 3: Radial diffusion coefficients ($D = D_{t/a2}$): × 10^{-8} cm^2 s^{-1} (means ± SD), 4 fibers, 7°C. Protocol: Skinned fiber transferred sequentially from oil to drop 1 (2 s), oil (60 s re-equilibration period), drop 2 (2 s), oil (<1 s), drop 3 (600 s), oil, fiber removed. Drop and fiber samples analyzed by SDS-PAGE. Diffusion coefficients calculated from equation [1] and [2] in text (Constant D model). Amount of each protein in each drop indicated by letter M; drop number indicated by superscript. D^1 calculated using the fraction $M^1/(M^1+M^2+M^3)$; D^2 calculated using the fraction $M^2/(M^2+M^3)$. Average radius of fiber subset, 25.5±3.7 µm. Phosphofructose kinase and glycogen phosphorylase not analyzed.

C. Estimation of Hydrodynamic Volume

Equation 4 and 7 in our simulation requires values for the hydrodynamic volume of each protein (as an estimate of the protein exclusion volume), as well as the average hydrodynamic volume of the background proteins in the cytosol. We estimated the hydrodynamic volume of each protein, V_h, for all 15 proteins tracked in this simulation, by means of a linear fit to the hydrodynamic volume of human and earthworm hemoglobin as a function of molecular weight from O'Leary (1). This method of estimating V_h was used because experimental values of the hydrodynamic volume are not available in the literature for these proteins. To incorporate the dependence of hydrodynamic volume on molecular weight a linear fit of hydrodynamic volume as a function of molecular weight is given by expression 6 in Figure 1S below. Of the 15 proteins in this study all but 6 of the proteins (TPI – 53.25 kDa, PGM – 57 kDa, PAR – 11.93 kDa, ADK – 21.64 kDa, PGO – 61.43 kDa and PGK – 36.5 kDa) fall in the molecular weight range spanned by human (65 kDa) and earthworm (3700 kDa) hemoglobin and of those 6 proteins, 3 are within 15 kDa of human hemoglobin (TPI,


6
9
8
7
1
2
3
4
4
5
5

Protein	Mo (Da)	Dw (cm2/s)		Rh (nm)			Vh (nm^3)			Vh (dL/g)		
		Young et al.	Linear Fit - SE	Young et al	Stokes-Einstein	Linear Fit	Young et al.	Stokes-Einstein	Linear fit	Young et al	Stokes-Einstein	Linear fit
PGI	125230	3.21899E-07	3.5127E-07	3.309	4.394	4.026	151.807	355.341	273.442	0.00730	0.01709	0.01315
PFK	340288	2.30679E-07	2.4842E-07	4.618	6.131	5.694	412.505	965.570	773.112	0.00730	0.01709	0.01368
TPI	53250	4.28071E-07	4.6927E-07	2.489	3.304	3.014	64.551	151.097	114.696	0.00730	0.01709	0.01297
GPD	144700	3.06761E-07	3.3435E-07	3.473	4.611	4.230	175.409	410.587	317.113	0.00730	0.01709	0.01320
ENO	93874	3.54357E-07	3.8746E-07	3.006	3.991	3.650	113.796	266.368	203.765	0.00730	0.01709	0.01307
PVK	231664	2.62222E-07	2.8427E-07	4.062	5.394	4.976	280.828	657.348	515.979	0.00730	0.01709	0.01341
PGM	57000	4.1847E-07	4.5863E-07	2.546	3.380	3.084	69.097	161.738	122.861	0.00730	0.01709	0.01298
PAR	11933	7.04757E-07	7.7462E-07	1.512	2.007	1.826	14.465	33.860	25.500	0.00730	0.01709	0.01287
ADK	21638	5.77941E-07	6.3484E-07	1.843	2.447	2.228	26.230	61.398	46.325	0.00730	0.01709	0.01289
PGO	61427	4.08166E-07	4.4721E-07	2.610	3.465	3.163	74.463	174.300	132.515	0.00730	0.01709	0.01299
PPB	194316	2.78048E-07	3.0212E-07	3.831	5.087	4.682	235.554	551.373	429.811	0.00730	0.01709	0.01332
ADO	156844	2.9863E-07	3.2524E-07	3.567	4.736	4.349	190.130	445.046	344.510	0.00730	0.01709	0.01323
PGK	36500	4.85503E-07	5.3279E-07	2.194	2.913	2.655	44.246	103.569	78.367	0.00730	0.01709	0.01293
LDH	145732	3.06035E-07	3.3353E-07	3.481	4.622	4.241	176.660	413.516	319.436	0.00730	0.01709	0.01320
PCK	86224	3.64541E-07	3.9879E-07	2.922	3.880	3.547	104.523	244.661	186.889	0.00730	0.01709	0.01305
Human Hb	65000	4.00545E-07	4.3876E-07	2.660	3.531	3.224	78.794	184.438	140.319	0.00730	0.01709	0.01300
Eworm Hb	3700000	1.04125E-07	9.5714E-08	10.231	13.584	14.777	4485.221	10498.775	13517.104	0.00730	0.01709	0.02200

1 $D_w = C_{Young}\left(\frac{T}{\eta M_o^{1/3}}\right)$

2 $R_h = \left[\left(\frac{3}{4\pi}\right)\left(\frac{v}{N_A}\right)M_o\right]^{1/3}\left[\frac{1\text{x}10^7\text{nm}}{\text{cm}}\right]$

3 $R_h = \frac{k_bT}{6\pi\eta D_w}\left[\frac{1\text{x}10^{16}\text{nm}}{\text{J}/\text{cP}(\text{cm}^2/\text{s})}\right]$

4 $V_h^{\text{nm}^3} = \frac{4\pi}{3}R_h$

5 $V_h^{\text{dL/g}} = V_h^{\text{nm}^3}\left(\frac{N_A}{M_o}\right)\left(\frac{1\text{x}10^{-23}\text{dL}}{\text{nm}^3}\right)$

6 $V_h^{\text{dL/g}} = mM_o + b$

7 $V_h^{\text{nm}^3} = V_h^{\text{dL/g}}\left(\frac{M_o}{N_A}\right)\left(\frac{1\text{x}10^{23}\text{nm}^3}{\text{dL}}\right)$

8 $R_h = \frac{3}{4\pi}V_h^{\text{nm}^{1/3}}$

9 $D_w = \frac{k_bT}{6\pi\eta R_h}\left[\frac{1\text{x}10^{16}\,\text{cm}^2/\text{s}}{\text{J}/(\text{cP nm})}\right]$

$C_{\text{Young}} = 8.34\text{x}10^{-8}\,\frac{\text{cm}^2\text{cP g}^{1/3}}{°\text{K s mole}^{1/3}}$

$v = 0.73\,\frac{\text{cm}^3}{\text{g}}$

From Young et al. 1980

$N_A = 6.022\text{x}10^{23}\,\frac{\text{molecules}}{\text{mole}}$

$k_b = 1.3806\text{x}10^{-23}\,\frac{\text{J}}{°\text{K}}$

$T = 280\,°\text{K}$

$\eta = 1.45\text{ cP}$

$m = 2.476\text{x}10^{-9}\,\frac{\text{dL/g}}{\text{g/mole}}$

$b = 0.01284\,\frac{\text{dL}}{\text{g}}$

Linear fit to values in O'Leary 1987

Figure S1. Comparison between several methods of calculating, diffusion of various proteins in an aqueous solution D_w, hydrodynamic radius R_h, and hydrodynamic volume V_h. Values of diffusion in water and hydrodynamic radius of each protein using linear estimation of V_h as a function of molecular weight (expression 6) is contrasted against that obtained from empirical expressions derived from Young et al. (expressions 1 and 2) using the Stokes-Einstein relation (expressions 3 and 9) to perform intermediate calculations.

PGM and PGO). A comparison of the estimated diffusion in water of each protein from Young et al. (2), shown in Expression 1, Figure 1S, with that obtained from the linear fit (Expressions 6-9, Figure 1S) shows between 7.5 to 10% higher rate of diffusion is predicted from the linear fit. In addition, the estimated hydrodynamic radius using the linear fit lies between that estimated directly from Young et al. (Expression 2, Figure 1S) and that obtained estimating D_w from Young et al. and then calculating R_h using the Stokes-Einstein expression (Expressions 1 and 3, Figure 1S). These observations suggest that the estimates of hydraulic volume used in our simulation fall within the range of the values that would be expected experimentally.

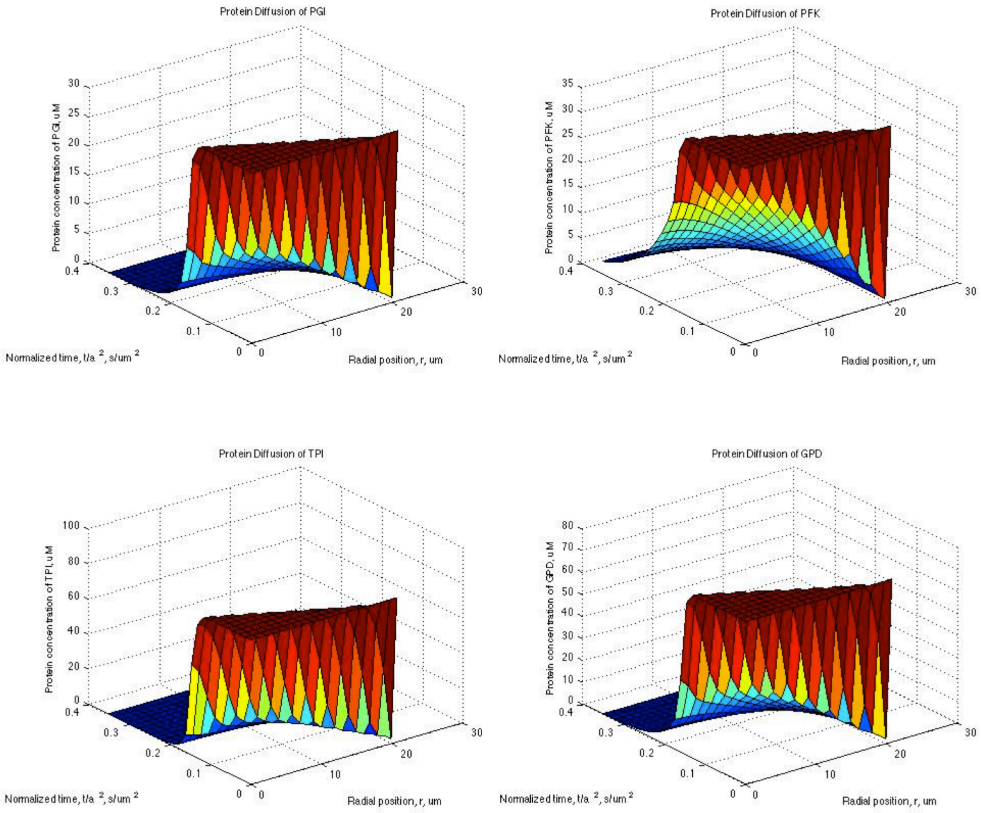


Figure S2. Surface plots of protein concentration across the fiber radius as a function of time for phosphoglucose isomerase (PGI), phosphofructose kinase (PFK), triosephosphate isomerase (TPI), and glyceraldehyde-3-phosphate dehydrogenase (GPD).

D. Surface plots of protein concentration across myofibril radius as a function time

In Figures 2S, 3S, 4S and 5S the simulation profiles of protein concentration as a function of radial position and time are shown for the optimized parameters of the Variable D model given in Table 1 in the text. It can be observed from these figures that parvalbumin and adenylate kinase (Figure 4S), which were not included in the

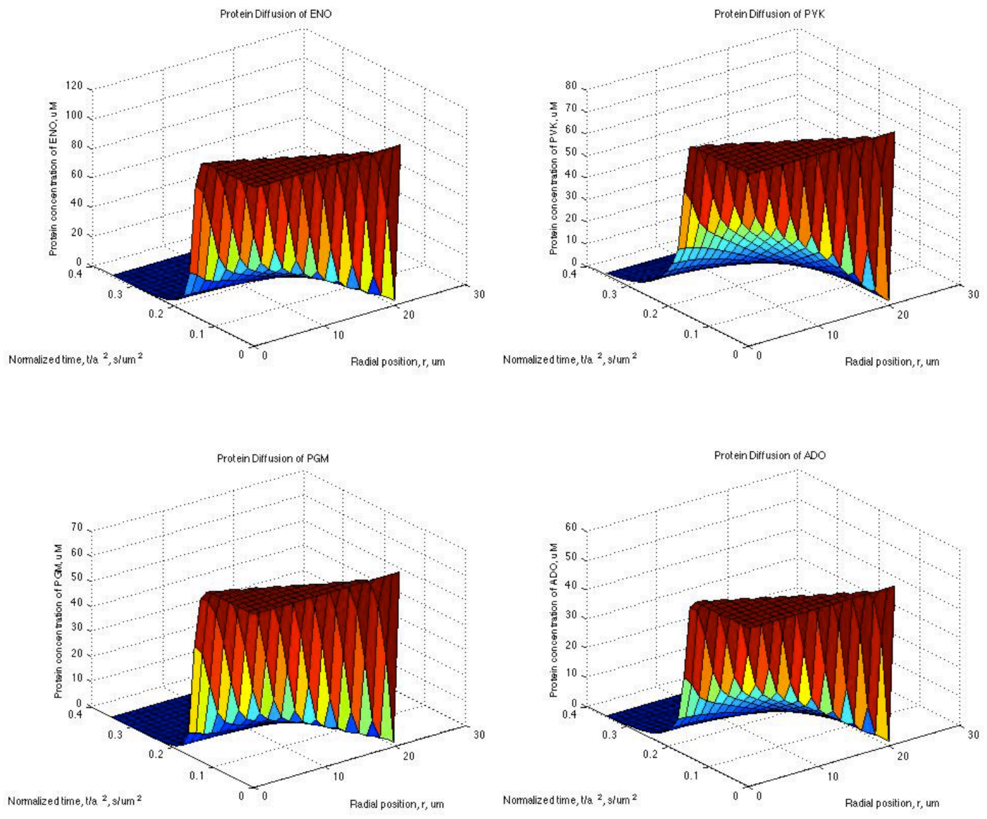


Figure S3. Surface plots of protein concentration across the fiber radius as a function of time for enolase (ENO), pyruvate kinase (PVK), phosphoglycerate mutase (PGM), and aldolase (ADO).

supramolecular protein complex, diffuse out of the myofibril rapidly and have nearly completely diffused out of the myofibril at t/a^2 of 0.1 s/μm^2. The remaining proteins take much longer to diffuse out, with concentrations approaching zero across the myofibril at $t/a^2 > 0.2$ s/μm^2. Phosphofructose kinase takes the longest to diffuse out of the myofibril, with near complete diffusion out of the myofibril at $t/a^2 \sim 0.4$ s/μm^2 (Figure 2S). In all proteins participating in the supramolecular complex the steep drop in the concentration that progresses along the radial direction from the outer fiber radius with time is a direct result of the complex dissociating as local total protein concentration decreases.

E. Simulation Model Code and Optimization Notes

All code and datasets used to simulate the best fits of the two different model versions to the experimental data are available upon request from the authors. Constant D and Variable D simulation codes are very similar but files in each folder differ even though they may be named the same. Therefore each simulation should be run out of its respective folder. All code is written in MATLAB and requires the Partial Differential Equation toolbox to run. Actual optimization code used to obtain these best fits is also

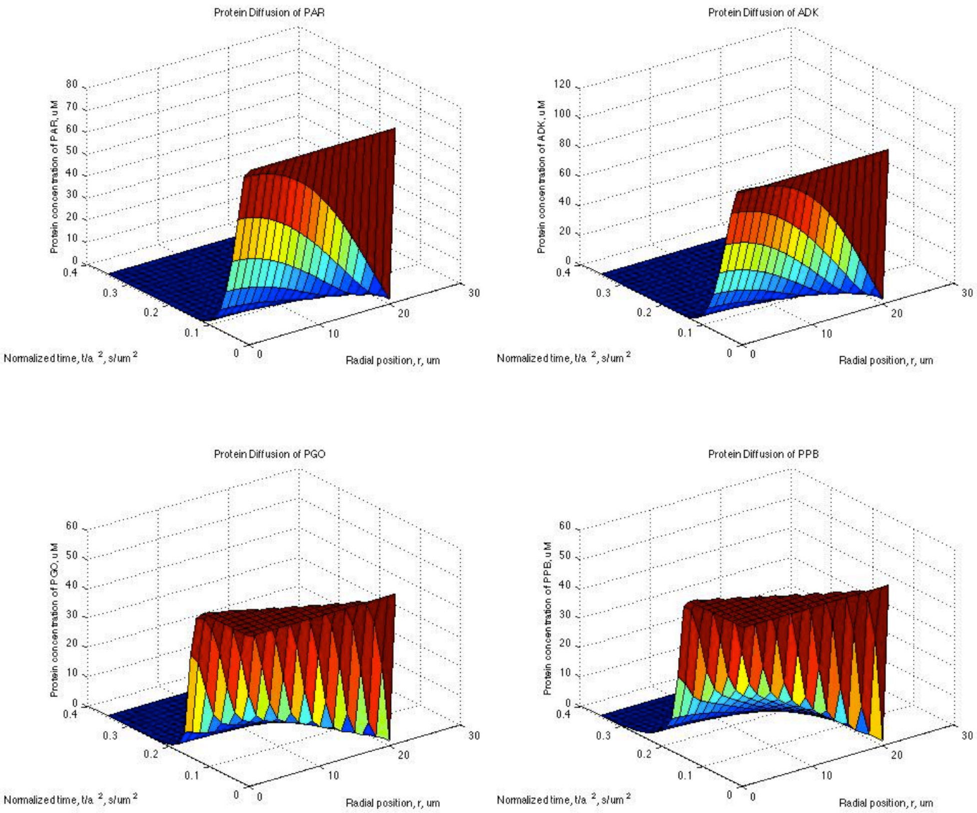


Figure S4. Surface plots of protein concentration across the fiber radius as a function of time for parvalbumin (PAR), adenylate kinase (ADK), phosphoglucose mutase (PGO), and glycogen phosphorylase (PPB).

available upon request and requires the additional Optimization toolbox to run locally on a single core and the Parallel Computing Toolbox to run the optimization in a parallel manner either locally on a multicore processor or on a computational cluster. These file are extensively documented and could be rewritten in other programming languages. In MATLAB help can be obtained for all scripts and functions by typing "help" followed by the script or function name. A brief one-line description of all the scripts and functions in a folder can be obtained by typing "help" followed by the folder name. This documentation and attached model code aims at facilitating model transparency and model replication.

There are three major differences from typical optimization methods, which are worth noting but are not covered in the original paper.

1) The optimization is performed in two stages. First a simulated annealing method explores the parameter space as defined by the upper and lower bounds on the parameters giving us a more global perspective on the minimization. This method randomly selects parameter values so is independent of the initial guess. The exact

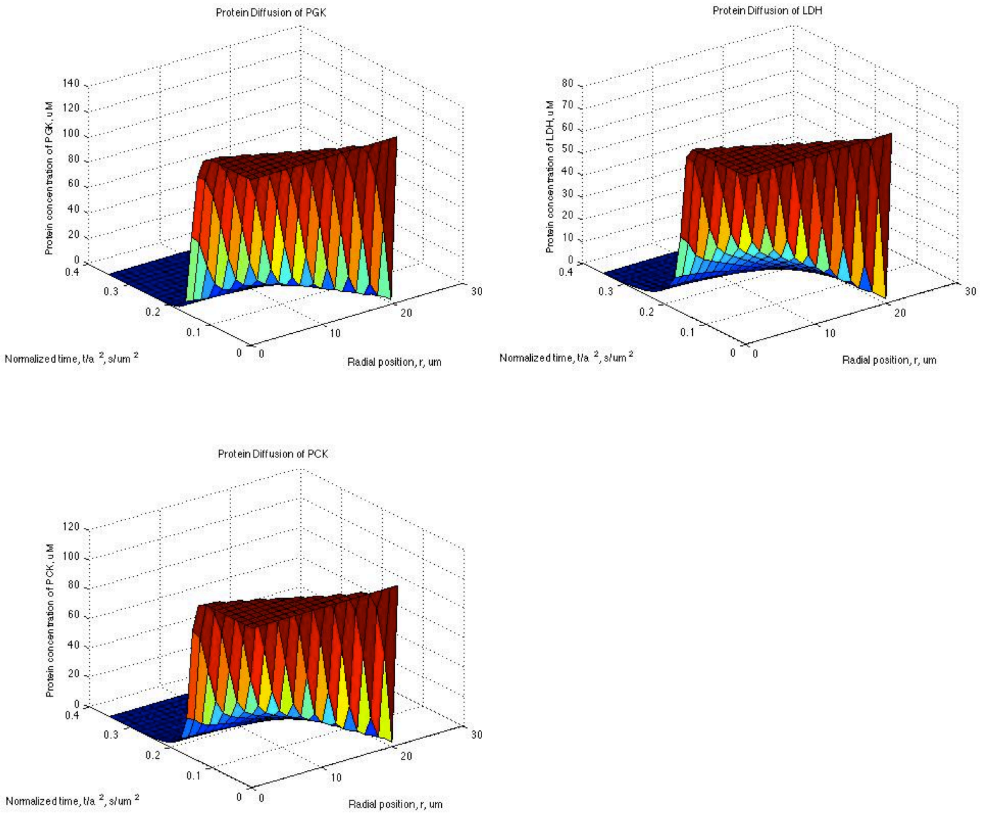


Figure S5. Surface plots of protein concentration across the fiber radius as a function of time for phosphoglycerate kinase (PGK), lactate dehydrogenase (LDH), and glycogen phosphocreatine kinase (PCK).

method is very similar to that developed by Boyan (3) which utilizes a Modified Lam annealing schedule and does not impose a neighborhood function that is dependent on the annealing temperature for each random selection of parameter values. Therefore this simulated annealing code is relatively model-independent and does not require a good initial guess; however, it does require the user to define the minimal bounded parameter space to search. Another modification made is that the user can specify the number of parameters between two iterations that are allowed to change. The exact parameters that do change between iterations is randomly selected and these parameters are allowed to change anywhere within their bounds. After the simulated annealing optimization returns a value we use *fmincon*, a gradient-based minimization function included in the MATLAB optimization toolbox, to find the local minima in the region identified by the global simulated annealing method. For more details on *fmincon* see MATLAB documentation (4).

2) The form of the least-squares minimization used to determine the residual error between simulation and data is not the traditional least-squares error in the y variable given by:

$$RE = \frac{\sqrt{\sum_{i=1}^{N}\left(y_{sim}\left(x_i\right)-y_i\right)^2}}{Ny_{\max}}$$

where *RE* is the residual error between the simulation evaluated at x_i, $y_{sim}(x_i)$, and y_i where x_i and y_i are the values of x and y for data point i, $y_{\max}$ is the maximal value of the y variable and N is the total number of data points. If this residual error was to be used the sigmoidal distribution of the data points would lead to an optimization that would tend to fit the large number of points at long t/a^2 ($t/a^2 > 0.10 \times 10^8$ s/cm^2) and would not try to minimize the error generated by the dramatic increase in diffusion around $t/a^2 \sim 0.05 \times 10^8$ s/cm^2. This type of data requires a residual error function that approximates the perpendicular distance of the data points to the simulation curve. This is given by:

$$RE = \frac{\sqrt{\sum_{i=1}^{N}\frac{\left(x_{sim}\left(y_i\right)-x_i\right)^2\left(y_{sim}\left(x_i\right)-y_i\right)^2}{\left(\frac{x_{sim}\left(y_i\right)-x_i}{x_{\max}}\right)^2+\left(\frac{y_{sim}\left(x_i\right)-y_i}{y_{\max}}\right)^2}}}{Nx_{\max}y_{\max}}$$

where $x_{sim}(y_i)$ is the simulation x value at y_i and $x_{\max}$ is the maximum value of the x variable. This approximate perpendicular distance residual error function resulted in much better fits to the sigmoidal shape of the data.

3) The optimization was run on either a multi-node remote cluster or a single/multi core local workstation. In the cluster configuration the computational jobs are distributed across the cores in an “embarrassingly parallel” manner, meaning that each job executes to completion without communicating with other jobs run at the same time. This is the simplest of cluster configurations and requires no message passing interface (MPI). However if no cluster is available to the user the optimization code (available on request) can be run on a multi core workstation using the Parallel Computing Toolbox.

F. Akaike and Bayesian Information Criteria and Residual Error

Akaike and Bayesian information criteria (AIC and BIC) were performed on the simulations that most closely replicated the data for both the Constant D and Variable D models. Both the AIC and BIC employ a penalty term that is a function of the number of parameters in each model in order to assess whether a better fit is justified in a model with more adjustable parameters. The expression for AIC is:

$$AIC = n\log\left(\sigma^2\right)+2K$$

where n is the number of datapoints, σ^2 is the sum of the residual error squared divided by the number of datapoints and K is the number of adjustable parameters in the model.

The expression for BIC is similar but has a larger penalty on additional free parameters and is given by:

$$BIC = n\log\left(\sigma^{2}\right) + K\log n$$

In the Variable D model there are 4 adjustable parameters, which determine the varying diffusion coefficients for all 11 proteins while in the Constant D model each proteins diffusion coefficient is determined separately yielding a model with 11 adjustable parameters. In addition the Variable D model yields a smaller residual error so it is evident that the AIC and BIC evaluations will favor the Variable D model. Despite this we have evaluated these criteria and they are given in Table 3S below and support the use of the Variable D over the Constant D model. When comparing the AIC and BIC values calculated here the more negative values indicate better fits. A difference of over 10 between the two models suggest that the model with the lower AIC or BIC value is highly preferred.

In addition we have evaluated the residual error in two regions of the data, at t/a^2 less than and greater than 0.125 × 10^8 s cm^{-2}, to compare how closely each model represents the data in these regions. It can be seen from Table 3S below that the Variable D model fits the data at the shorter time scales much more successfully than the Constant D model and almost equally at the longer time scales.

	Variable D Model	**Constant D Model**
Akaike Information Criteria	-1493	-1398
Bayesian Information Criteria	-1477	-1355
Delta AIC	0	95
Delta BIC	0	122
Residual Error	0.0204	0.0229
Residual Error ***t/a*** ^{2} **< 0.125**	0.0186	0.0232
Rsidual Error ***t/a*** ^{2} **> 0.125**	0.0784	0.0679

Table S3. Akaike and Bayesian Information Criteria and residual error comparisons between the Variable D and Constant D models. For both the AIC and BIC values a larger negative number is preferred. These values are negative because the residual errors were normalized by the maximal values on each axis to represent percentage error. In addition residual error was calculated over the data points where *t*/*a*^{2} is less than and greater than 0.125 × 10^8 s cm^{-2}. Note that the residual error at short time and long times is calculated in a slightly different manner than the overall residual error leading to larger predicted error in the Constant D model for both short and long times than is found overall.

SUPPORTING REFERENCES

1) O'Leary TJ. Concentration dependence of protein diffusion. *Biophysical Journal* **52**:137-139, 1987.

2) Young ME, Carroad PA and Bell RL. Estimation of diffusion coefficients of proteins. *Biotechnology and Bioengineering* 22: 947-955, 1980.

3) Boyan JA. Learning evaluation functions for global optimization. Ph.D, Thesis, Carnegie Mellon University, August 1998

4) MATLAB documentation. http://www.mathworks.com/access/helpdesk/help/helpdesk.html