Biologistics—Diffusion coefficients for complete proteome of
Escherichia coli

Tomasz Kalwarczyk, Marcin Tabaka and Robert Holyst*

Institute of Physical Chemistry, Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warsaw, Poland

ABSTRACT

Motivation: Biologistics provides data for quantitative analysis of transport (diffusion) processes and their spatio-temporal correlations in cells. Mobility of proteins is one of the few parameters necessary to describe reaction rates for gene regulation. Although understanding of diffusion-limited biochemical reactions in vivo requires mobility data for the largest possible number of proteins in their native forms, currently, there is no database that would contain the complete information about the diffusion coefficients (DCs) of proteins in a given cell type.

Results: We demonstrate a method for the determination of in vivo DCs for any molecule—regardless of its molecular weight, size and structure—in any type of cell. We exemplify the method with the database of in vivo DC for all proteins (4302 records) from the proteome of K12 strain of Escherichia coli, together with examples of DC of amino acids, sugars, RNA and DNA. The database follows from the scale-dependent viscosity reference curve (sdVRC). Construction of sdVRC for prokaryotic or eukaryotic cell requires ~20 in vivo measurements using techniques such as fluorescence correlation spectroscopy (FCS), fluorescence recovery after photobleaching (FRAP), nuclear magnetic resonance (NMR) or particle tracking. The shape of the sdVRC would be different for each organism, but the mathematical form of the curve remains the same. The presented method has a high predictive power, as the measurements of DCs of several inert, properly chosen probes in a single cell type allows to determine the DCs of thousands of proteins. Additionally, obtained mobility data allow quantitative study of biochemical interactions in vivo.

Contact: roholyst@ichf.edu.pl

Received on June 9, 2012; revised on August 2, 2012; accepted on August 27, 2012

1 INTRODUCTION

Biologistics and biochemistry in a crowded environment are two emerging interdisciplinary fields of science. They provide quantitative analysis of transport of proteins and their spatio-temporal correlations involved in gene expression and regulation. According to the current state-of-the-art theory of gene expression (activation or repression) in bacteria (Elf et al., 2007; Li et al., 2009), mobility of proteins is one of the few parameters necessary to describe reaction rates of gene regulation. The mobility is understood as a three-dimensional diffusion or one-dimensional sliding along DNA (for prokaryotes and eukaryotes), or by velocity of molecular motors (in eukaryotic cells). Understanding of diffusion-limited biochemical reactions requires accurate in vivo mobility data for the largest possible number of proteins in their native forms. The three-dimensional diffusion of different types of macromolecules in the cytoplasm of Escherichia coli has been experimentally studied in several cases (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Derman et al., 2008; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007), but experimental determination of the mobility of all proteins is technically an impossible task because of their large number in a given cell. For example, the proteome of the K12 strain of E. coli (Blattner et al., 1997) contains more than 4300 proteins. Moreover, most of the recent studies concern measurements mainly performed with the use of green fluorescent protein (GFP) (Elowitz et al., 1999; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007) or GFP fusion proteins (Jennifer et al., 2001).

Attempts to study the diffusion of many proteins simultaneously, under conditions resembling the interior of the cells, were performed in silico by McGuffee and Elcock (2010). Computational methods, however, have limitations arising from the speed and capacity of computing hardware and small number of interacting proteins in the system (~50 different types of proteins) (McGuffee and Elcock, 2010). An alternative approach is the quantitative analysis of available literature data. Mika and Poolman (2011) gathered literature data of diffusion coefficients (DCs) of ~20 different types of proteins in E. coli and proposed a power law dependence of the DC on the molecular weight of proteins. This power law, however (Mika and Poolman, 2011), can be applied only for the proteins in a narrow range of molecular weights, i.e. between 20 and 30 kDa.

In this work, we present a method for predictions of the DCs of proteins for the proteome of any cell. We collected all available literature data (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Derman et al., 2008; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007) on diffusion of various probes, including small molecules (water, glucose), proteins and plasmids, in the cytoplasm of E. coli. We used those data and the scaling function of viscosity (Holyst et al., 2012).
2009; Kalwarczyk et al., 2011; Szymański et al., 2006a, b) to predict the mobility of macromolecules in the bacterial cytoplasm. We also predicted the DCs of amino acids, sugars, proteins and DNA. We created a unique database, including the DCs of all proteins of strain K12 of E. coli (4302 proteins), their oligomers and their potential complexes with translocation proteins; 6600 records in total.

2 METHODS

2.1 A brief description of the method

Our predictions of DCs of proteins in the bacterial cytoplasm are based on experimental data on diffusion in the cytoplasm of E. coli available in the literature (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Derman et al., 2008; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007). The method relies on the dependence $D_0/D_{cyto} = n/\eta_0$, where $D_0$ is the DC of macromolecule in water of viscosity $\eta_0$, and $D_{cyto}$ is the DC of macromolecule in the cytoplasm. $n$ is the effective viscosity experienced by the macromolecule during diffusion in the cytoplasm. The protocol of determination of DCs is graphically represented in Figure 1.

2.2 Calculation of hydrodynamic radii and DCs in water

Hydrodynamic radius of proteins was determined using the following formula (Dill et al., 2011):

$$r_p = 0.0515M_w^{0.392} \text{ [nm]},$$

while for RNA we used Equation (2) (Werner, 2011).

$$r_p = 0.0566M_w^{0.30} \text{ [nm]},$$

Dependence of the hydrodynamic radii of linear, circular or super coiled DNA on molecular weight [Equations (3)-(5), respectively] was obtained from DCs of DNA constructs (Robertson et al., 2006) using Equation (6).

$$r_p = 0.024M_w^{0.57} \text{ [nm]},$$

$$r_p = 0.0125M_w^{0.59} \text{ [nm]},$$

$$r_p = 0.0145M_w^{0.57} \text{ [nm]}.$$

Radii of amino acids and sugars have been calculated, assuming that the hydrodynamic radius $r_p$ corresponds to the van der Waals radius $r_w$ calculated according to the procedure described elsewhere (Zhao et al., 2003).

For each probe, we use the literature values of $D_{cyto}$, while the values of $D_0$ (if not available) were calculated using the Stokes–Sutherland–Einstein equation [Equation (6)].

$$D_0 = \frac{kT}{6\pi\eta_0 r_p}.$$

2.3 Calculation of DCs of various molecules in the cytoplasm of E. coli

Using the molecular weights from Uniprot protein database (Apweiler et al., 2011; Jain et al., 2009), we calculated the DCs for the complete proteome of E. coli (K12 strain). We identified the cellular localization of each protein as well as its quaternary structure (a single polypeptide chain or multiple chain aggregates or complexes). In the case of membrane or periplasmic proteins, we adopted the assumption that, after synthesis, the proteins diffuse via the cytoplasm to its target in the membrane, through one of two transport pathways [twin-arginine translocation (TAT) or the general secretion system (Sec)] (Driessen and Nouwen, 2008; Sargent, 2007). Consequently, these proteins were considered as single polypeptide chains (the TAT pathway) or protein complexes with SecB or Tig proteins (the Sec pathway). Hydrodynamic radius of proteins was determined using Equation (1). When the protein was composed of several subunits, the molecular weight of all polypeptide chains comprising the protein was added together. On the basis of cumulative molecular weight of the complex, hydrodynamic radius of the protein $r_p$ and further its DC $D_0$ was calculated [Equations (1) and (6)]. Then, using Equation (7), we calculated the relative DCs for all analysed proteins, and we calculated the DCs of proteins in the cytoplasm $D_{cyto}$. The calculated DCs of all proteins in the cytoplasm is summarized in Supplementary Table S1.

3 RESULTS AND DISCUSSION

3.1 Construction of the scale-dependent viscosity reference curve

We collected the literature data (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007) for DCs of different solutes and macromolecules in the cytoplasm of E. coli (Fig. 2 and Table 1). We used the least squares method to fit those data with Equation (7) (Kalwarczyk et al., 2011).

$$\ln\left( \frac{D_0}{D_{cyto}} \right) = \ln\left( \frac{\eta}{\eta_0} \right) = \left( \frac{2}{R^2 + \frac{L^2}{r_p^2}} \right)^{1/2},$$

Fig. 1. Diagram of a method of predicting the DC of any molecule in the cell cytoplasm. To predict the DCs of molecules in the cytoplasm, it is essential to correctly select the probes that will be used to determine the reference curve. Next, one need to measure the DCs of selected probes in water (buffer) $D_0$ and the DC in the cytoplasm of studied cell $D_{cyto}$. Using $D_0$ and $D_{cyto}$, we create the sdVRC. To predict the DC of a given molecule, it is necessary to know its hydrodynamic radius $r_p$ or $D_0$. Although sdVRC depends on both $r_p$ and $D_0$, in practice, both parameters can be calculated knowing only one of them. Finally, by substituting the values of $r_p$ and $D_0$ to sdVRC, the DC in the cytoplasm $D_{cyto}$ can be determined.
Fig. 2. The sdVRC. The logarithm of viscosity \( \eta \) divided by the viscosity of water \( \eta_0 \) [ln(\( \eta/\eta_0 \) = ln(\( D_0/D_{298} \))] as a function of the hydrodynamic radius \( r_p \) of various probes (Table 1) of radii from 0.16 nm to 203 nm (closed square). The cytoplasmic DCs \( D_{298} \) of probes were taken from the literature (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007) (cf. Table 1). By fitting the data with Equation (7) (solid line), we determined two length scales: \( \xi = 0.51 \pm 0.09 \) nm and \( R_0 = 42 \pm 9 \) nm. We also determined the macroscopic viscosity of the cytoplasm \( \eta_m = 17.5 \) Pa s, i.e., 26,000 times higher than the viscosity of water \( \eta_0 \) at 310 K. Shading represents the maximum error of fitting

here \( r_p \) is the hydrodynamic radius of the probe, and \( R_0 \) and \( \xi \) are length scales characterizing the cytoplasm. \( \xi \) (an average distance between surfaces of proteins), \( R_0 \) (average hydrodynamic radius of the biggest crowders) and \( a \) (a constant of the order of one) are the fitting parameters whose values for the cytoplasm of \( E. coli \) are as follows: \( \xi = 0.51 \pm 0.09 \) nm, \( R_0 = 42 \pm 9 \) nm and \( a = 0.53 \pm 0.04 \). From the scale-dependent viscosity reference curve (sdVRC), we directly determined the macroscopic viscosity \( \eta_m \) of the cytoplasm. We found that \( \eta_m = 17.5 \) Pa s (26,000 times greater than the viscosity of water – \( \eta_0 \approx 0.7 \) mPa s at 310 K). \( R_0 \) is comparable to the radius of the loops (Kim et al., 2004) of DNA covered with proteins. The second length scale determined from sdVRC, \( \xi \), is comparable to the average distance between surfaces of proteins. \( R_0 \) determines the length scale above which the viscosity ceases to depend on the size of the probe and reaches the macroscopic value. For a probe smaller than \( \xi \), the experienced viscosity has a value comparable to the viscosity of water.

We used as-obtained sdVRC [Equation (7)] as a tool for prediction of DCs of all known proteins of K12 strain (Blattner et al., 1997) of \( E. coli \) as well as other molecules and macromolecules.

3.2 Interpretation of sdVRC

For more than a decade, diffusion of various proteins in the cytoplasm of \( E. coli \) has been studied (Table 1) (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007). Those experimental data show that the DCs exponentially depend on the size of the diffusing molecule. For example, GFP with a molecular weight \( M_w = 27 \) kDa and hydrodynamic radius \( r_p = 2.8 \) nm is characterized by cytoplasmic DC (Elowitz et al., 1999) \( D_{298} = 7.7 \pm 2.5 \mu m^2/s \). On the other hand, the DC of large oligomeric protein consisting of four subunits of GFF-tagged \( \beta\)-galactosidase (\( \beta\)-gal-GFP), of radius almost three times greater than GFP (\( M_w \approx 580 \) kDa, \( r_p = 7.3 \) nm), is equal to 0.7 ± 0.22\( \mu m^2/s \) (Mika et al., 2010). The above differences are explained in terms of scale-dependent viscosity (Kalwarczyk et al., 2011) experienced by the diffusing molecule [cf. sdVRC, Equation (7)]. Equation (7) is an empirical equation primarily found for synthetic systems such as polymer or micellar solutions (Holyst et al., 2009; Kalwarczyk et al., 2011; Szymański et al., 2006a, b). Interpretation of four parameters in Equation (7) \(( R_0, \xi, \eta_m \) and \( \eta_0 \)) is taken from those studies (Holyst et al., 2009; Kalwarczyk et al., 2011; Szymański et al., 2006a, b). In synthetic systems, \( \xi \) is the average distance between macromolecular components of the complex liquid and \( R_0 \) is equal to the hydrodynamic radius of a polymer random coil or of a micelle. In sdVRC, both \( \xi \) and \( R_0 \) determine the viscosity experienced by a probe diffusing in the investigated liquid. For \( r_p \gg R_0 \), the probe experiences the macroscopic viscosity \( \eta_m \). A probe of radius \( r_p \) smaller than \( \xi \) moving in the liquid experiences the viscosity of the solvent \( \eta_0 \). On the other hand, a probe of \( r_p > \xi \) will experience a viscosity higher than the viscosity of the solvent. Finally, the effective viscosity \( \eta \) experienced by a probe of radius between \( \xi \) and \( R_0 \) \(( \xi < r_p < R_0 \)) depends exponentially on \( r_p \). In case of the cytoplasm of mammalian cells, \( R_0 \) corresponds to the hydrodynamic radius of the filaments forming the cellular cytoskeleton in the volume of the cytoplasm (Kalwarczyk et al., 2011). The bacterial cytoskeleton (Shih and Rothfield, 2006), however, is located directly next to the inner membrane (Pogliano, 2008). We can therefore assume that it should not have a large contribution to the viscosity experienced by the proteins diffusing across the cytoplasm. This assumption is also supported by the value of \( R_0 = 42 \pm 9 \) nm determined from fitting, which is similar to the radius of the objects identified as fragments of the bacterial nucleoid (around 40 nm) (Kim et al., 2004), i.e., loops of DNA covered with structural proteins. This value can be compared with the value of the hydrodynamic radius of the filaments forming the bacterial cytoskeleton (Hou et al., 2012; Pogliano, 2008) (fragments of length \( L = 100 \) nm and a radius \( r = 2.5 \) nm), which is ~17 nm (Vandesande and Persoons, 1985), well below \( R_0 \), obtained from the fit. Therefore, the length scale, \( R_0 \), is neither correlated with the hydrodynamic radius of the filaments nor with the proteins whose highest hydrodynamic radius is about 10 nm. \( \xi \) in the cytoplasm of \( E. coli \) equals 0.51 ± 0.09 nm and is comparable with the average distance between proteins. Parameters of the sdVRC \(( \xi \) and \( R_0 \)) depend on the internal structure of the cytoplasm (proteins density, size of the nucleoid, etc.). Thus, each cell type will be characterized by a different shape of the reference curve (due to differences in parameters \( \xi \) and \( R_0 \)), while the mathematical form of the sdVRC will not change, and such curve can be constructed for other cell types.
3.3 Other models of diffusion in the cytoplasm

We compared our results with three models of diffusion in the cytoplasm of *E. coli*, available in the literature (Figures 3 and 4). McGuffee and Elcock (2010) proposed two models of diffusion in the cytoplasm: the ‘steric’ model, which takes into account only steric interactions between diffusing proteins, and the ‘full’ model, which includes steric, electrostatic and hydrodynamic interactions between diffusing entities. Comparison of the results (Figure 3) shows that the model we propose takes into account possible interactions between the diffusing probes and the surrounding environment. Moreover, we show that the full information needed to build the sdVRC can be obtained only after taking into account the probes whose $r_p$ greatly exceeds $R_h$. For example, simulations conducted by McGuffee and Elcock (2010) include proteins that are most abundant in the cytoplasm, but the absence of large objects such as the nucleoid leads to underestimated values of $D_0/D_{cyto}$. The effect starts to be meaningful for probes whose $r_p > 10$ nm. In that case, the values of $D_0/D_{cyto}$ are lower by an order of magnitude with respect to experimental results.

We also compared our results with the model proposed by Mika and Poolman (2011), where $D_{cyto} \propto M_w^{0.7}$. As can be seen, the power law dependence of $D_{cyto}$ on $M_w$ may also lead to underestimated values of $D_0/D_{cyto}$. For example, for the ribosome 70S $D_0/D_{cyto}$ measured experimentally is five times higher than predicted using power law dependence. Therefore, the power law dependence proposed by Mika and Poolman (2011) holds for the proteins in a small range of molecular weights 20–30 kDa and, moreover, is not applicable to macromolecules other than proteins. This is because each type of macromolecules (DNA, RNA, proteins, polymers, etc.), has different shape and thus different dependence of $r_p$ on $M_w$ [Equations (1)–(5)]. The shape of the macromolecule and in consequence its radius translates into the DC. The dependence of DC $D_{cyto}$ of different types of macromolecules (proteins, RNA and DNA) on their molecular weight is shown in Figure 4.

3.4 Accuracy of the model

Accuracy in determination of the sdVRC strongly depends on the amount of available data. One would expect that increasing the amount of data for probes of $r_p \gg R_h$ and $r_p < \xi$, would significantly decrease the maximum error of the sdVRC (compare Fig. 2).
generated by McGuffee and Elcock (2010) were fitted by Equation (7), (open diamond, dashed line). The model proposed by Mika and et al., 2009; van den Bogaart Equation (7) with parameters: $a = 0.51 \pm 0.09$, $R_c = 42 \pm 9$ nm and $a = 0.53 \pm 0.04$. We compared our results with data generated by McGuffee and Elcock (2010) and Mika and Poolman (2011). The data generated by McGuffee and Elcock (2010) were fitted by Equation (7), yielding the following parameters: for the ‘full’ model $\xi = 0.2 \pm 0.2$ nm, $R_c = 20 \pm 48$ nm and $a = 0.32 \pm 0.12$ (dotted circle, dotted line), for the ‘steric’ model $\xi = 3.57 \pm 0.1$ nm, $R_c = 17 \pm 6$ nm and $a = 0.59 \pm 0.05$ (open diamond, dashed line). The model proposed by Mika and Poolman (2011) where $D_{\text{cyto}} \propto M_w^{-0.7}$ is plotted as dashed–dotted line.

Fig. 3. The comparison of sdVRC with other existing models. The plot shows the literature values for the logarithm of $D_p/D_{\text{cyto}}$ (open squares) (Bakshi et al., 2012; Campbell and Mullins, 2007; Cluzel et al., 2000; Elowitz et al., 1999; English et al., 2011; Golding and Cox, 2004; Jasnin et al., 2008; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; van den Bogaart et al., 2007). Black solid line represents Equation (7) with parameters: $\xi = 0.51 \pm 0.09$ nm, $R_c = 42 \pm 9$ nm and $a = 0.53 \pm 0.04$. We compared our results with data generated by McGuffee and Elcock (2010) and Mika and Poolman (2011). The data generated by McGuffee and Elcock (2010) were fitted by Equation (7), yielding the following parameters: for the ‘full’ model $\xi = 0.2 \pm 0.2$ nm, $R_c = 20 \pm 48$ nm and $a = 0.32 \pm 0.12$ (dotted circle, dotted line), for the ‘steric’ model $\xi = 3.57 \pm 0.1$ nm, $R_c = 17 \pm 6$ nm and $a = 0.59 \pm 0.05$ (open diamond, dashed line). The model proposed by Mika and Poolman (2011) where $D_{\text{cyto}} \propto M_w^{-0.7}$ is plotted as dashed–dotted line.

To test the accuracy of the presented method, we perform an analysis of the error of calculation of DC $\delta D_{\text{cyto}}$ for GFP as a function of the number of experimental data points. Using Equation (7), we generated 10 datasets, where the number of data points ranges from 6 to 100. The generated data were uniformly distributed on a logarithmic scale and were randomly drawn on the assumption that measurement error is described by a normal distribution with standard deviation $\sigma = 0.1$. We assumed that the error of $R_p$ equals to 5%. We found that 20 data points are sufficient to obtain $\delta D_{\text{cyto}}$ at the level of 20% for the GFP (averaged over 10 generated datasets). In comparison, $\delta D_{\text{cyto}}$ obtained from the analysis of the literature data was at the level of 40% (cf. Fig. 2). This is mainly because of the small number of available experimental data. Furthermore, most of the experimental data are available for a narrow range of hydrodynamic radii (around 3 nm, cf. Fig. 2) which is not preferred in this type of analysis. To date, however, there is no experimental data which would improve the accuracy of the sdVRC. Therefore, to improve the accuracy, additional experiments are needed to cover a wider range of $R_p$ of the probes and also uncertainties of $D_p$, $D_{\text{cyto}}$, and $r_p$ should be minimized.

3.5 DCs of proteins

Preparing a database of DCs of the entire proteome, one should keep in mind that about 45% of the proteome are proteins forming a larger macromolecular complex (homo-, hetero-oligomers and complexes of membrane proteins with translocation proteins). Thus, the calculation of DCs of proteins should be carried out also for protein complexes. The Uniprot protein database (Apweiler et al., 2011; Jain et al., 2009) contains information on the molecular weight of their proteins, their quaternary structure and their location in cell. Using these data and sdVRC (cf. Fig. 2) we calculated the DCs $D_{\text{cyto}}$ of all proteins in E. coli (Supplementary Table S1) present in the cytoplasm as monomers (single polypeptide chains) or as multimers (homo- or hetero-oligomers) or complexes composed of many chains, see Fig. 5). Figure 5A shows the histogram of molecular weights of cytoplasmic proteins, including homo- and hetero-multimers. Distribution of molecular weights of proteins is given by log-normal distribution with probability density function $q(M_w) = (\sqrt{2\pi}\sigma M_w)^{-1}\exp\left[-(\ln(M_w/\mu))^2/(2\sigma^2)\right]$, where standard deviation $\sigma = 0.825 \pm 0.007$ and mean molecular weight $\mu = 31.9 \pm 0.3$ kDa. The relationship between the DC and the molecular weight of protein is expressed by the Equations (1) and (7). A histogram of DCs of cytoplasmic proteins is shown in Figure 5B (same proteins as in Fig. 5A). The distribution follows the curve given by the probability density function: $p(D_{\text{cyto}}(M_w)) = q(M_w)\frac{dM(D_{\text{cyto}})}{dD_{\text{cyto}}}$.  

Fig. 4. Comparison of measured and predicted $D_{\text{cyto}}$ as a function of molecular weight of the investigated probes. Predicted dependencies shown in the graph are expressed by Equation (7). The hydrodynamic radius $R_p$ of each type of macromolecules is given by the relationship: $R_p = CM_w^n$, where $M_w$ is the molecular weight of the macromolecule. For proteins $C = 0.0514$ and $a = 0.392$—Equation (1); RNA $C = 0.0566$ and $a = 0.38$—Equation (2), linear DNA $C = 0.024$ and $a = 0.57$—Equation (3); circular DNA $C = 0.0125$ and $a = 0.59$—Equation (4); super coiled $C = 0.0145$ and $a = 0.57$—Equation (5). For comparison, we present experimental data on DCs of proteins (Cluzel et al., 2000; Elowitz et al., 1999; English et al., 2011; Konopka et al., 2006; Kumar et al., 2010; Mika et al., 2010; Mullineaux et al., 2006; Nenninger et al., 2010; Slade et al., 2009; Slade et al., 2009, 2010; RNA (Golding and Cox, 2004), plasmid (Campbell and Mullins, 2007) and ribosomes 30S and 70S (Bakshi et al., 2012). The dashed-dotted straight line indicates the relationship $D \propto M_w^{-0.7}$ proposed by Mika and Poolman (2011). The dependence of $D_{\text{cyto}}$ on $M_w$ proposed by Mika and Poolman (2011), when applied to large plasmids ($M_w \sim 2 \times 10^6$ kDa), yields several orders of magnitude overestimation of DC.
We also calculated $D_{cyto}$ of membrane proteins that are $\sim$30% of the proteome of *E. coli*. Membrane proteins, after synthesis by the ribosome, are transported to the membrane, according to one of the two pathways: the TAT (Sargent, 2007) in which proteins are transported as single polypeptides in a folded state and the Sec (Driessen and Nouwen, 2008) in which unfolded proteins are complexed mainly by one of the two proteins: SecB or Tig. We created a database (Supplementary Table S1) listing the $D_{cyto}$ of all proteins, including their monomeric forms, the possible homo- and hetero-multimers, and in the case of membrane proteins also the complexes with translocation proteins (SecB and Tig). Apart from DCs of proteins, we calculated $D_{cyto}$ of small molecules such as amino acids or sugars and for macromolecules such as RNA or DNA (linear, circular and super coiled). Calculated values of DCs are listed in Table 2.

The predicted DCs refer only to three-dimensional diffusion. In cells, particularly eukaryotes, there are also other types of transport such as molecular motors (Vale, 2003). Nevertheless, mobility, irrespective of the type of motion, is inversely proportional to the viscosity of the surrounding environment. Since the viscosity is dependent on the scale (Høllyset et al., 2009; Kalwarczyk et al., 2011; Szyman’ski et al., 2006a, b), each type of motion will depend exponentially [Equation (7)] on the size of a moving object.

3.6 Application of DC database in studies of biochemical processes occurring in cells

Using the database of DCs, one can determine quantitatively whether the protein diffuses freely or interacts and forms complexes with much larger macromolecules, e.g. plasmids. Capoulade et al. (2011) performed diffusion measurements and showed that, in the nucleus of eukaryotic cell, euchromatin creates domains of high and low affinity for heterochromatin protein (HP1α).

Another kind of analysis was performed by Elf et al. (2007). Authors compared in vivo DCs of both: the lactose repressor in its native form and the lactose repressor devoid of the DNA-binding domain. Order of magnitude difference in the coefficient of diffusion of both proteins led to the conclusion that the native lactose repressor spends 87% of the time attached to the DNA. This shows that the presence of attractive interactions between diffusing particles will result in a slowdown of diffusion of molecules.

To clarify the method, consider a hypothetical protein of hydrodynamic radii $r_p = 3$ nm. The DCs of this protein $D_{cyto}$ (calculated from sdVRC) is approximately equal to $D_{cyto} = 8.7 \mu m^2/s$. The time required by the protein to visit every place in the cell volume [for *E. coli* $V \sim 0.6 \mu m^3$ (Kubitschek, 1990)] is approximately equal to $t = V/4\pi D_{cyto}r_p \approx 1.8$ s. Now suppose that the protein binds to a
plasmid whose molecular weight equals to 10,000 kDa, the DC of the plasmid is of the order of $D_{\text{plasm}} = 10^{-4} \mu \text{m}^2/\text{s}$. Suppose further that the protein spends one-tenth of the time diffusing freely $\tau_f$ and the remaining 90% of time $\tau_c$ as a complex with the plasmid ($\tau_c = 10\tau_f$). The effective DCs of the complexes $D_{\text{eff}}$, defined as $D_{\text{eff}} = (D_{\text{cyto}} + D_{\text{plasm}})/\tau_c$, is approximately $6 \times 10^{-7} \mu \text{m}^2/\text{s}$. Analysis of the sDvRC allows to determine the characteristic length scales $R_0$ and $\xi$, and the DC of any (macro)molecule in the cytoplasm. For the cytoplasm of E. coli, we found that $R_0$ is surprisingly well correlated with the average radius of the DNA loops forming the nucleoid. This suggests that the nucleoid is the main crowding agent (responsible for the macroscopic viscosity) in the cytoplasm of E. coli.

Finally, it should be noted that there are no additional requirements (except experimental data) to construct analogous database of DCs in other systems such as the nucleus or mitochondria of eukaryotic cells. We also believe that sDvRC can be easily adopted to calculate other types of mobility, including one-dimensional sliding, velocity of molecular motors, etc., as they all are inversely proportional to the viscosity.

**ACKNOWLEDGEMENTS**

The authors would like to thank Prof. Marcin Fialkowski for inspiring discussions. R.H. conceived the study; R.H. directed the project with input from T.K and M.T.; T.K. made data analysis and processing with inputs from R.H and M.T.; T.K. and R.H. wrote the manuscript.

**Funding** T.K. thanks the National Science Center for funding the project from the funds granted on the basis of the decision number: DEC-2011/01/N/ST3/00865, and Foundation for Polish Science for START scholarship. M.T. thanks the Ministry of Science of Poland for support within the Iuventus-Plus program IP2010 052570 (2011). R.H. thanks the National Science Center for funding the project from the funds granted on the basis of the decision number: 2011/02/A/ST3/00143 (Maestro grant).

**Conflict of Interest** none declared.

**REFERENCES**

Apweller, R. et al. (2011) Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.*, 39, D214–D219.

Bakshi, S. et al. (2012) Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Mol. Microbiol.*, 85, 21–38.

Blattner, F. et al. (1997) The complete genome sequence of *Escherichia coli* k-12. *Science*, 277, 1453–1462.

Campbell, C.S. and Mullins, R.D. (2007) In vivo visualization of type II plasmid segregation: bacterial actin filaments pushing plasmids. *J. Cell Biol.*, 179, 1059–1066.

Capoulade, J. et al. (2011) Quantitative fluorescence imaging of protein diffusion and interaction in living cells. *Nat. Biotechnol.*, 29, 835–842.

Cluzel, P. et al. (2000) An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science*, 287, 1652–1655.

Derman, A. et al. (2008) Intracellular mobility of plasmid DNA is limited by the para family of partitioning systems. *Mol. Microbiol.*, 67, 935–946.

Dill, K.A. et al. (2011) Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. USA*, 108, 17876–17882.

Driessen, A.J.M. and Nouwen, N. (2008) Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.*, 77, 643–667.
Elowitz,M. et al. (2000) Diffusion of green fluorescent protein in three cell environments in Escherichia coli. J. Bacteriol., 188, 3442–3448.

Hinrichs,R. and Karp,R. (2004) Predicting translational diffusion of evolutionary conserved RNA structures by the nucleotide number. Nucleic Acids Res., 39, e17.

Zhao,Y.H. et al. (2003) Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. J. Org. Chem., 68, 7368–7373.