Description of atomic burials in compact globular proteins by Fermi–Dirac probability distributions

Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions

Antonio L. C. Gomes[1], Júlia R. de Rezende[1]
Antônio F. Pereira de Araújo[1*] and Eugene I. Shakhnovich[2]
[1]Laboratório de Biologia Teórica, Departamento de Biologia Celular,
Universidade de Brasília, Brasília-DF 70910-900, Brazil and
[2]Department of Chemistry and Chemical Biology,
Harvard University, Cambridge-MA 02138, USA

running title: Description of atomic burials by Fermi functions
keywords: hydrophobicity, structural segregation, atomic burial, protein density, Fermi function, protein folding

Abstract

We perform a statistical analysis of atomic distributions as a function of the distance R from the molecular geometrical center in a non-redundant set of compact globular proteins. The number of atoms increases quadratically for small R, indicating a constant average density inside the core, reaches a maximum at a size-dependent distance R_{max} and falls rapidly for larger R. The empirical curves turn out to be consistent with the volume increase of spherical concentric solid shells and a Fermi-Dirac distribution in which the distance R plays the role of an effective atomic energy $\epsilon(R) = R$. The effective chemical potential μ governing the distribution increases with the number of residues, reflecting the size of the protein globule, while the temperature parameter β decreases. Interestingly, $\beta\mu$ is not as strongly dependent on protein size and appears to be tuned to maintain approximately half of the atoms in the high density interior and the other half in the exterior region of rapidly decreasing density. A normalized size-independent distribution was obtained for the atomic probability as a function of the reduced distance, $r = R/R_g$, where R_g is the radius of gyration. The global normalized Fermi distribution, $F(r)$, can be reasonably decomposed in Fermi-like sub-distributions for different atomic types τ, $F_\tau(r)$, with $\sum_\tau F_\tau(r) = F(r)$, which depend on two additional parameters μ_τ and h_τ. The chemical potential μ_τ affects a scaling pre-factor and depends on the overall frequency of the corresponding atomic type while the maximum position of the sub-distribution is determined by h_τ, which appears in a type-dependent atomic effective energy, $\epsilon_\tau(r) = h_\tau r$, and is strongly correlated to available hydrophobicity scales. Better adjustments are obtained when the effective energy is not assumed to be necessarily linear, or $\epsilon_\tau^*(r) = h_\tau^* r^{\alpha_\tau}$, in which case a correlation with hydrophobicity scales is found for the product $\alpha_\tau h_\tau^*$. These results indicate that compact globular proteins are consistent with a thermodynamic system governed by hydrophobic-like energy functions, with reduced distances from the geometrical center reflecting atomic burials, and provide a conceptual framework for the eventual prediction from sequence of a few parameters from which whole atomic probability distributions and potentials of mean force can be reconstructed.

1 Introduction

Recent theoretical studies and simulations of minimalist lattice models have suggested that a non-specific hydrophobic potential can reproduce many aspects of protein folding behavior for appropriate native structures [1–7]. According to this potential, the energetic contribution of different monomers in a given conformation is simply the negative product between the number of contacts they make and their "hydrophobicity". Since hydrophobic monomers (positive hydrophobicity) decrease the energy when forming a contact while the reverse is true for hydrophilic monomers (negative hydrophobicity), this function can be considered to mimic, in a very simplified way, the hydrophobic effect, which is believed to be the dominant factor in the stabilization of protein structures [8]. This simple functional form, in conjunction with the Z-score criterion for protein-like folding behavior, permits an interesting analysis where both sequences and structures are represented by vectors in N-dimensional space, where N is the number of monomers. Since unfolded conformations are expected to be, on average, close to the main diagonal direction in this space, appropriate native structures are predicted to have contact vectors pointing as far away from this direction as possible. The distance of a given vector from the diagonal is proportional to σ, the standard deviation of its individual components from their average, which implies that appropriate native conformations for the hydrophobic function should be structurally segregated, with monomers tending to be either completely buried or completely exposed to the solvent. In particular, maximally compact conformations in square and cubic lattices are predicted to be not appropriate, since they correspond to low σ values [1–3, 5].

From an informational perspective, it is interesting to note that the number of contacts made by different monomers, which is taken in these minimalist lattice models as an appropriate measure of exposure to the solvent or burial inside the globular structure, is at the same time directly encoded in the sequence of hydrophobicities and sufficiently informative to uniquely determine the native conformation. The first condition imposes the sequence information as an upper limit to the possible informational content of monomer burials while the second condition similarly imposes the native conformation as a lower limit to this quantity. In order to investigate to what extent the simple ideas described above in the context of minimalist models can be applied to real proteins it is mandatory, therefore, to identify an appropriate burial measure satisfying simultaneously these two conditions. Atomic contact numbers in continuous models fail to satisfy the second condition because large changes in contact numbers can result from subtle conformational variations and many different conformations are adjustable to the same set of contact numbers. Accessible surface areas would be more discriminative in this respect but their calculation during folding simulations would be computationally expensive. In a parallel study we are currently investigating if atomic distances to the molecular geometrical center, or central distances for short, in conjunction to physically motivated hydrogen bond restrictions, satisfy this second condition, at least for some small globular proteins. Although less general, since correlation with solvent exposure cannot be expected for proteins with arbitrary shapes, central distances are more discriminative than the number of contacts and easier to compute than accessible surface areas. It is also relevant that there is a single distance corresponding to each atom, implying that the informational content of atomic burials defined in this manner increases only linearly with protein size as does the information of amino acid sequences.

In the present study we analyze atomic radial distributions in a non-redundant set of 321 compact globular proteins, constructed as described in the Methods section, in order to identify eventual signatures of a hydrophobic-like energy function and structural segregation assuming atomic central distances as an appropriate measure of atomic burial. Attempts to decompose the global distribution in type-dependent sub-distributions are also performed as an initial step for the prediction of preferential atomic burials from primary sequences which are intended to be useful in protein folding simulations.

2 Results

2.1 The atomic density of compact globular proteins is described by a Fermi-Dirac probability distribution

Fig. 1(a) shows the radial atomic density distribution $n(R) = \delta N(R)/\delta R$, where $\delta N(R)$ is the number of atoms, labeled by i, with distance R_i from the geometrical center satisfying the relation $R - \delta R/2 \leq R_i < R + \delta R/2$, averaged inside each of five groups corresponding to different ranges for the number of monomers M: $M \leq 100$, $101 \leq M \leq 150$, $151 \leq M \leq 200$, $201 \leq M \leq 300$ and $M > 300$. Points represent the empirical data for $\delta R = 0.1$ Å. Error bars, shown for the fourth protein group, represent typical standard deviations from the mean for $\delta R = 0.5$ Å. All curves are coincident for small R, where they appear to increase quadratically, indicating a uniform atomic density independent of protein size. After reaching size-dependet maxima at size-dependent positions, which certainly reflect different globular dimensions, the number of atoms falls rapidly with a near-exponential behavior. Larger standard deviations at large R reflect the size dependence within each protein group. Log-log and semi-log plots of these curves shown in (c) and (d), respectively, confirm the quadratic increase for small R and the reasonably exponential decrease for large R. Fig 1(b) shows the volumetric density obtained from the same data divided by $4\pi R^2$. Wide fluctuations occur in this plot for small R because the associated experimental error diverges as R approaches 0 due to the vanishing term in the denominator.

The resulting curves are consistent with the quadratic volume increase of concentric solid shells around the geometric center and a constant atomic density for small radius R, ρ_0, which decreases exponentially from ρ_0 to 0 within a relatively small range around a size-dependent radius. A mathematical expression displaying these characteristics of the atomic density is the Fermi function, which describes the occupancy of different particle quantum states in a fermionic ideal gas as a function of their energies, ϵ:

$$\begin{aligned} F(\epsilon) &= \frac{e^{-\beta(\epsilon-\mu)}}{1+e^{-\beta(\epsilon-\mu)}} \\ &= \frac{1}{1+\exp(\beta(\epsilon-\mu))}, \end{aligned} \qquad (1)$$

where $\beta = 1/k_BT$ is the inverse absolute temperature, k_B is the Boltzmann constant, and μ is the chemical potential of the particles. As described in standard texts on statistical mechanics [9], when $\beta\mu >> 1$ and $\epsilon << \mu$ then $\exp(-\beta(\epsilon-\mu)) >> 0$ and $F(\epsilon) \approx 1$ while for $\epsilon >> \mu$, accordingly, $F(\epsilon) \approx \exp(\beta(\mu-\epsilon))$ falls exponentially. Moreover, $F(\mu) = 1/2$ and a sharp transition of $F(\epsilon)$ from almost unity to almost zero occurs in the interval $\mu - 1/\beta \leq \epsilon \leq \mu + 1/\beta$. As shown by the theoretical curves represented as lines in Fig. 1(a), the empirical data is well described by the following expression:

$$n(R) = AR^2F(R) = \frac{AR^2}{\exp(\beta(R-\mu))+1}, \qquad (2)$$

with $A = 0.65$ Å^{-3} and variable β and μ. The density inside the inner core around the geometrical center is therefore estimated as $\rho_0 = A/4\pi \approx 0.052$ atoms/Å^3, or 1 atom occupying the average volume of 19.3 Å^3.

The dependences of μ and β on protein size were obtained from independent fitting to all 321 proteins in the data bank. As shown in Fig. 2, μ appears to increase approximately as $N^{0.37}$. Importantly, this dependence is not identical to the dependence of the radius of gyration, R_g, for the present set of proteins, which is proportional to $N^{0.33}$. This conventional dependence of R_g on N was actually assumed *a priori* in the selection of globular proteins for the present study but, as described in the Methods section, it was previously observed for the set of structures formed by the most compact conformations of each size [10]. The increase in μ with N is to a great extent compensated by corresponding decrease in β, although individual variations in the temperature parameter are more pronounced. Interestingly, the product $\beta\mu$ appears to be much less dependent on protein size, as shown in the same figure.

2.2 Structural segregation and the absolute temperature of the atomic distribution

The description of the empirical data by eq. 2 permits an analysis using well known results for the Fermi function, defined by eq. 1. In particular, the atomic density should be essentially constant, $\rho(R) \approx \rho_0 = A/4\pi$, inside a central sphere of radius $R_0 = \mu - 1/\beta$ around the molecular geometrical center and should fall from ρ_0 to essentially 0 in the interval $\mu - 1/\beta < R < \mu + 1/\beta$, or $R_0 < R < R_0 + 2/\beta$.

Calling $N(R) = \int_0^R n(R')dR' = \int_0^R AR'^2F(R')dR'$ the number of atoms at a distance equal or smaller than R from the geometrical center, we can estimate the number of atoms inside the internal region of essentially constant density ρ_0 as

$$N(R_0) = \int_0^{R_0} AR^2 F(R)dR \approx \frac{A}{3}R_0^3, \tag{3}$$

since $F(R) \approx 1$ in this region, while the total number of atoms is given by

$$N_t = \int_0^{\infty} AR^2 F(R)dR \approx \frac{A}{3}\mu^3 + \frac{A\pi^2}{3\beta^2}\mu, \tag{4}$$

where we have used the useful general approximation [9]

$$\int_0^{\infty} \phi(x)F(x)dx \approx \int_0^{\mu} \phi(x)dx + \frac{\pi^2}{6\beta^2}\phi'(x)|_{x=\mu}. \tag{5}$$

Structural segregation would imply the total number of atoms to be approximately twice the number of internal atoms, $N_t = 2N(R_0)$, or

$$\mu^3 + \frac{\pi^2}{\beta^2}\mu = 2r_0^3 = 2(\mu - 1/\beta)^3. \tag{6}$$

After expanding the cubic term, rearranging and multiplying by β^3 we conclude that $\beta\mu$ should be the single real root of the following third degree polynomial equation:

$$p(\beta\mu) = (\beta\mu)^3 - 6(\beta\mu)^2 + (6 - \pi^2)\beta\mu - 2 = 0, \tag{7}$$

which happens to be approximately 6.63. Structural segregation, therefore, implies that $\beta\mu$ should be independent of protein size and, moreover, the crude above estimate predicts its value to be slightly below 7. As shown in Fig. 2, $\beta\mu$ appears to be independent or only weakly dependent on protein size, although displaying significant fluctuations in the entire size range for the present set of compact structures, with an overall average value of 10 and standard deviation of 3, which is higher but not completely inconsistent with the theoretical estimate. A more accurate estimate should take into consideration the density $\rho_c < \rho_0$ at which atoms become inaccessible to the solvent and, possibly, exact computation of the integrals above.

Structural segregation can be directly visualized, however, when the radial distributions shown in Fig. 1(a) are normalized, in such a way that central distances are expressed in terms of the radius of gyration, i.e. as reduced distances [11], and the number of atoms is expressed in terms of the corresponding fraction in each protein. Normalized radial distributions and corresponding cumulative distributions shown in Fig. 3(a) indicate that size dependence virtually disappears after normalization and that the cumulative atomic probability reaches 1/2 at the reduced distance $r \equiv R/R_g \sim 0.94$ which corresponds to density values between $0.8\rho_0$ and $0.9\rho_0$. The same figure also shows the adjustment of eq. 2 to the reduced distances averaged over all proteins, independently of size, with $\mu = 1.166 \pm 0.002$, $\beta = 9.37 \pm 0.07$ and $A = 1.729 \pm 0.008$. This normalized distribution will be considered to provide a satisfactory size-independent representation of atomic central distances in compact globular proteins.

2.3 Decomposition of the atomic distribution in Fermi-like contributions from different atomic types

When different atomic types are taken into consideration, we conjecture that the global Fermi distribution $F(R)$ might be decomposed into similar type-dependent functions of the form:

$$F_\tau(R) = \frac{e^{-\beta(\epsilon_\tau(R)-\mu_\tau)}}{1+\sum_\tau e^{-\beta(\epsilon_\tau(R)-\mu_\tau)}} \tag{8}$$

with the sum appearing in the denominator over all atomic types related to the global distribution parameters, β and μ, by:

$$\sum_\tau e^{-\beta(\epsilon_\tau(R)-\mu_\tau)} = e^{-\beta(R-\mu)}. \tag{9}$$

We initially assume the type-dependent atomic effective energy to be linear in R, or $\epsilon_\tau(R) = h_\tau R$, in which case

$$\begin{aligned} F_\tau(R) &= \frac{e^{-\beta(h_\tau R-\mu_\tau)}}{1+e^{-\beta(R-\mu)}} \\ &= \frac{e^{-\beta(h_\tau R-\mu_\tau)}e^{\beta(R-\mu)}}{1+e^{\beta(R-\mu)}} \\ &= e^{-\beta(\Delta h_\tau R-\Delta\mu_\tau)}F(R), \end{aligned} \tag{10}$$

and

$$\sum_\tau F_\tau(R) = F(R). \tag{11}$$

The radial type-dependent atomic density is therefore given by

$$n_\tau(R) = \frac{\delta N_\tau(R)}{\delta r} = AR^2F_\tau(R), \tag{12}$$

or, in the case of normalized curves,

$$p_\tau(r) = Ar^2F_\tau(r), \tag{13}$$

with central distances, R, being replaced by reduced distances, $r = R/R_g$, and the atomic radial densities, $n_\tau(R)$, being replaced by the probability densities $p_\tau(r)$.

The general shape and maximum position of this expression depend on $\Delta h_\tau = (h_\tau{-}1)$, which expresses how far is the given atomic type hydrophobicity from the unity effective hydrophobicity governing the global distribution, while the total area beneath the curve additionally depends on $\Delta\mu_\tau = (\mu_\tau - \mu)$, or how far is the atomic type chemical potential from the global chemical potential. Since this area must equal the total probability P_τ of atom type τ being at any r, or

$$\begin{aligned} P_\tau &= \int p_\tau(r)dr \\ &= \int Ar^2F_\tau(r)dr \\ &= Ae^{\beta\Delta\mu_\tau}\int r^2e^{-\beta\Delta h_\tau r}F(r)dr \\ &= Ae^{\beta\Delta\mu_\tau}I_\tau, \end{aligned} \tag{14}$$

where the integral $I_\tau = \int r^2 e^{-\beta \Delta h_\tau r} F(r) dr$ depends on Δh_τ but not on $\Delta\mu_\tau$, it is clear that for given h_τ and P_τ, which is estimated from the fraction of atomic type τ in the data bank, the chemical potential μ_τ is uniquely determined.

Fig. 3(b) shows the normalized distribution for atoms from hydrophobic and hydrophilic amino acids and corresponding adjusted curves by Eq. 13, averaged over all 321 structures in the data bank, in addition to the whole average distribution with the curve adjusted by the normalized version of Eq. 2. In this case A, β and μ were initially obtained from the curve adjusted to the whole distribution and the two parameters h_τ and μ_τ were adjusted independently of each other to both sub-distributions. Both sub-distributions are reasonably reproduced by the theoretical expression although hydrophilic atoms appear to be better adjusted than their hydrophobic counterparts, particularly for small r. The hydrophobicity for the hydrophobic distribution, $h_\tau = 1.138 \pm 0.008$, is indeed higher than for the hydrophilic distribution, $h_\tau = 0.879 \pm 0.005$, while their chemical potentials μ_τ scale the curves appropriately in such a way that their sum results in the whole distribution. We have also adjusted the same theoretical expression for 20 sub-distributions corresponding to all atoms from each of the amino acids. Fig. 3(c) shows the distributions and adjusted curves for phenilalanine and lysine, which are respectively the most hydrophobic and most hydrophilic residues according to the h_τ parameter. Distributions and adjusted curves for the remaining hydrophilic monomers ($h_\tau < 1$ or $\Delta h < 0$) are shown in Fig. 4 while for hydrophobic monomers ($h_\tau > 1$ or $\Delta h > 0$) are shown in Fig. 5.

Empirical distributions were reasonably reproduced by the theoretical expression given by eq. 13, including the correct maximum position, but their exact shapes are somewhat inaccurate for residues with large absolute values of Δh_τ, as clearly seen in Fig. 3(c), in which case adjusted curves tend to be broader than the empirical distributions. The quality of the adjustments, as seen Fig. 3(d), 6 and 7, is significantly improved with the addition of another adjustable parameter α_τ, which is introduced in the expression of $F_\tau(r)$ under the assumption that the effective central potential acting on each atom is not necessariy linear, $\epsilon_\tau(r) = h_\tau r$, but of a more general form, $\epsilon^*_\tau(r) = h^*_\tau r^{\alpha_\tau}$. The type-dependent probability distribution then becomes:

$$\begin{aligned} p^*_\tau(r) &= Ar^2 F^*_\tau(r) \\ &= Ar^2 \frac{e^{-\beta(h^*_\tau r^{\alpha_\tau} - \mu^*_\tau)}}{1 + e^{-\beta(r-\mu)}}. \end{aligned} \tag{15}$$

Note that all empirical distributions are very accurately reproduced by eq 15. The effective chemical potential μ^*_τ again affects only a scaling pre-factor and depends on the probability P_τ while the shape and maximum position now depend on two parameters: h^*_τ and α_τ.

As can be seen in Fig. 8(a, b), the hydrophobicity parameters h_τ obtained from the adjusted curves for the twenty amino acids under the assumption of linear effective energies, which are shown in Tab. 1, correlate with previous hydrophobicity scales obtained from effective partition coefficients between the interior and exterior of protein structures computed by Chotia and coworkers [12] and from experimental equilibrium constants of amino acid analogs between octanol and water obtained by Flauchére and Pliška [13], as quoted by Eisenberg and McLachlan [14], with Pearson's correlation coefficients $\mathcal{C} = 0.91$ and $\mathcal{C} = 0.90$, respectively. More recently, estimated average contributions of individual residues to protein stability, obtained by Zhou and Zhou both from experimental mutational data and computations with knowledge-based potentials, similarly suggested a linear dependence on solvent exposure [15] and resulting slopes, or "buriabilities", are also correlated to h_τ ($\mathcal{C} = 0.92$), as shown in Fig. 8(c). As can be seen in Fig. 8(d) and Tab. 1, h_τ values are also anticorrelated ($\mathcal{C} = -0.93$) to C_α average reduced central distances $< r >$ obtained more than two decades ago for a set of only 16 proteins by Scheraga and collaborators [11, 16], with perfect order agreement for the three most hydrophobic (phenilalanine, isoleucine and valine) and two most hydrophilic residues (lysine and glutamate), with some variation in between, particularly for serine, asparagine and cysteine. Anticorrelation in this last case reflects the trivial observation that large hydrophobicities tend to correspond to small central distances. The correlation coefficient between h_τ and $< r >$ when both quantities are computed for the current set of 321 structures is $\mathcal{C} = -0.98$ (not shown).

When effective energies are not necessarily linear the quantity that correlates with these hydrophobicity scales is not just h^*_τ but the product $h^*_\tau\alpha$, since $h^*_\tau\alpha$ is strongly correlated with h_τ, as seen in Fig. 9(a), with $\mathcal{C} = 0.99$. Consideration of the two independent parameters, however, permits a classification of the twenty amino acids that reflects extremely well some physical properties of different side chains that are not distinguished by their product $\alpha_\tau h^*_\tau$ or the single hydrophobicity parameter h_τ, as seen in the plot shown Fig. 9(b). In addition to the points corresponding to values of h^*_τ and α_τ for each amino acid taken from Tab. 1, curves of the form C/h_τ are also shown for $0.7 \leq C \leq 1.3$ at 0.1 intervals (thick line for $C = 1$, dotted lines for the others), representing lines of "isohydrophobicity", or constant $\alpha_\tau h^*_\tau$. The global reference distribution corresponds to point (1,1) in this plot but all other points at the dividing curve $\alpha_\tau = 1/h^*_\tau$ correspond to the same reference hydrophobicity. Points to the right of this curve correspond to hydrophobic atoms while points to the left correspond to hydrophilic atoms.

The differences between the theoretical expressions $p_\tau(r)$ and $p^*_\tau(r)$, given respectively by eqs. 13 and 15, can be visualized from the behavior of the effective atomic potentials $\epsilon(r)$ and $\epsilon^*(r)$. In the case of $p_\tau(r)$, shown in Fig. 10(a), the potential increases linearly with reduced distance $r = R/R_g$, $\epsilon(r) = h_\tau r$, and the only distinction between different atomic types is provided by the inclinations h_τ. For hydrophobic atoms the inclination is larger than unity while the reverse is true for hydrophilic atoms. The thick straight line with inclination 1 shown in the same plot corresponds to the effective hydrophobicity parameter of the whole distribution, which acts as a reference state. Since larger inclinations correspond to larger differences in the effective potential between the interior (small r) and exterior (large r) of the protein globule it is natural to consider $h_\tau = \frac{d\epsilon_\tau(r)}{dr}$ an appropriate measure of hydrophobicity in this case. For the second expression, $p^*_\tau(r)$, the effective central potential, $\epsilon^*_\tau(r) = h^*_\tau r^{\alpha_\tau}$, is not linear and its inclination changes with r depending on the exponent α_τ, as seen for charged residues (K, E, D and R) in Fig. 10(b), hydrophilic uncharged (N, Q, P) in Fig. 10(c), aromatic (W, F, Y) in Fig. 10(d), aliphatic (I, V, L, M, C) in Fig. 10(e), and indifferent (A, T, S, H and G) in Fig. 10(f). An appropriate measure of hydrophobicity, however, is provided by the inclination at $r = 1$, which is close to the imaginary border between interior (high constant atomic density) and exterior (rapidly decreasing density) regions in the protein globule, $\left.\frac{d\epsilon^*_\tau(r)}{dr}\right|_{r=1} = h^*_\tau \alpha_\tau$.

3 Discussion

It is presently unclear to what extent the good agreement between experimental data and the mathematical expression given by eq. 2 is purely coincidental. If we imagine protein conformations to result from the attraction of small hard spheres to a point in space by a spherically symmetric effective potential, proportional to r, the situation could possibly be mapped into an ideal gas of fermionic particles, with small volume cells that can be either empty or occupied by a single atom and corresponding radial distances playing the role of energy levels and corresponding energies, respectively, and atomic volume exclusion in protein structures playing the role of Pauli's exclusion principle. Independently of any specific physical model, however, eqs. 2 and 12 provide a very reasonable description of the distributions of atomic central distances in compact globular proteins, even though they depend on just a small number of parameters. Global parameters β, μ and A characterize the whole distribution, independently of atomic type, while each type-dependent sub-distribution depends additionally on an individual chemical potential, μ_τ, and on the hydrophobicity parameter h_τ. It might be appropriate to stress at this point that terms used in the present study to describe adjustable parameters like chemical potential, temperature and energy, should be regarded within the context of a convenient analogy to the corresponding variables in the original Fermi function and not as the actual physical origin of these empirical quantities, even because no physical model for the observed distribution is being proposed.

The three global parameters do have simple physical interpretations, however. A is related to the atomic density inside the protein core, $\rho_0 = A/4\pi$, whose average over different proteins of the same size is independent of chain length and corresponds to an average volume of 19.3 Å^3 occupied by each atom. This average atomic volume is consistent with detailed volume computations involving Voronoi polyedra, like a recent study by Chotia and collaborators [17] which provides average volumes for buried atoms ranging from 14.4 Å^3, in the case of aspartate, to 22.3 Å^3, in the case of alanine. The chemical potential μ corresponds to the distance at which the atomic density is half of the interior density, or $\rho(\mu) = (1/2)\rho_0$, reflecting the size of the globule. It is only natural, therefore, that the chemical potential should increase with chain length. More interesting, however, is that the specific dependence, $\mu \sim N^{0.37}$, is different from the more conventional dependence observed for the radius of gyration, $R_g \sim N^{0.33}$. A similar situation was previously observed for the scaling behaviors of the radius of gyration, considered as a size descriptor of the "inner" region of the protein, and the maximal distance from the geometrical center, or backbone span, considered as a size descriptor of the "outer" region, which suggested a less compact exterior when compared to the interior of protein globules [18]. Differences in compactness, or packing, between buried and exposed amino acids have also been observed directly in protein structures, e.g. [19, 20].

The third global parameter appearing in eq. 2 is the temperature parameter β, which determines how abrupt is the transition from $\rho = \rho_0$ to $\rho = 0$ around $r = \mu$. β was found to display stronger oscilations than μ but it tends to decrease with chain length in such a manner that the product $\beta\mu$ is independent, or only weakly dependent, on protein size. A constant $\beta\mu$ would imply that the distributions of central distances in protein structures of different sizes are self-similar and, in particular, the ratio between buried and exposed atoms is a constant. The stronger dependence of μ on chain length when compared to the radius of gyration is therefore consistent with the previously suggested difference in compactness between interior and exterior regions of protein globules. The approximately constant value for $\beta\mu$, however, additionally suggests that the proportion of atoms in these two regions is reasonably invariant, as would be required if structural segregation does indeed act as a constrain of appropriate native conformations. The crude estimate for the $\beta\mu$ value required to result in maximal segregation was slightly below but not inconsistent with the observed average value.

It might also be relevant that 0.37, the exponent governing the dependence of μ on N equals 0.74/2, where 0.74 happens to be very close to the exponent governing the dependence of the accessible surface area, A_s, on molecular mass, M, previously observed both for monomeric ($A_s \sim M^{0.73}$) and oligomeric ($A_s \sim M^{0.76}$) globular proteins [12, 21], or $A_s \sim \mu^2$. This non-conventional dependence of accessible surface on protein mass has already been associated to the fractal dimension of protein surfaces [22]. Our results point out the possibility that structural segregation might be related to this fractal behavior. Moreover, a similar exponent $0.75 = 3/4$ has been claimed to occur in many allometric relations in biological systems, including the mass dependence of metabolic rates in animals, plants and even cells and organelles, while a possible origin for the ubiquitous exponent was suggested to arise from fractal space-filling vascular systems with invariant terminal units [23–26]. It is tempting to speculate that a similar mechanism could be involved in globular proteins and that the observed exponent would result from a space-filling network of water molecules that reach all protein atoms in the exterior region of the protein. The invariant terminal units would have the size of a single water molecule. The surface of this exterior region would scale with a 3/4 exponent when plotted as a function of its mass and also, since structural segregation implies that this exterior mass should be proportional (actually one half) to the total protein mass, when plotted as a function of total mass or number of atoms.

A more practical consequence of self-similarity of the distributions of central distances is the possibility of normalization. A single normalized curve describes the average size-independent behavior of all proteins in the data bank. Average quantities of interest, like frequencies of different atomic types, can then be computed in all proteins and combined in a single size-independent, statistically more significant, probability estimate. We have expressed central distances in terms of the radius of gyration, resulting in the reduced distance $r = R/R_g$, and the number of atoms has been replaced by probability densities estimated from corresponding frequencies in the data bank. We have also attempted other normalization procedures, involving μ instead R_g, but since no improvement was observed we decided to use the simpler procedure.

The reasonable description of type-dependent distributions by eq. 13 shows that the specific behavior of different atomic types, τ, depends essentially on the hydrophobicity parameter h_τ, which reflects the intrinsic tendency of the atom to be buried or exposed to the solvent, in addition to the chemical potential μ_τ which acts simply as a scaling pre-factor related to the overall probability P_τ of atomic type τ in compact globular proteins. For the same h_τ, μ_τ must therefore increase with P_τ. For the same P_τ, however, μ_τ is not independent of h_τ because curves with different shapes might require different scaling pre-factors in order to span the same area. These two tendencies can be observed in Tab. 1 for amino acid residues with similar hydrophobicities or probabilities. In this way, for example, $h_{\mathrm{W}} = 1.15 \sim 1.14 = h_{\mathrm{L}}$ but, since the probability of tryptophane atoms is much smaller than for leucine, $P_{\mathrm{W}} = 0.027 < 0.088 = P_{\mathrm{L}}$, it follows that the chemical potentials must also be different, $\mu_{\mathrm{W}} = 0.88 < 1.01 = \mu_{\mathrm{L}}$. On the other hand, phenilalanine and tyrosine have similar atomic probabilities, $P_{\mathrm{F}} = 0.057 \sim 0.056 = P_{\mathrm{Y}}$, but since phenilalanine is more hydrophobic, $h_{\mathrm{F}} = 1.19 > 1.09 = h_{\mathrm{Y}}$, its probability curve begins to decrease at smaller values of r and a larger chemical potential is required to provide an area similar to the one of tyrosine, $\mu_{\mathrm{F}} = 1.00 > 0.91 = \mu_{\mathrm{Y}}$. Note that atomic-type probabilities in the present case depend both on the amino acid frequencies and on their sizes.

The correlation between the hydrophobicity parameter obtained in the present study under the assumption of linear effective atomic energies, eq. 13, and previous hydrophobicity scales obtained by different methods, with examples shown in Fig. 8, corroborates the initial hypothesis that h_τ largely reflects a physical quantity usually associated with the term "hydrophobicity". For the twenty amino acid residues, Tab. 1 shows an almost continuous variation of h_τ from $h_{\mathrm{K}} = 0.79$, or $\Delta h_{\mathrm{K}} = -0.21$, to $h_{\mathrm{F}} = 1.19$, or $\Delta h_{\mathrm{F}} = 0.19$. Note that when h_τ is sufficiently different from 1, or Δh_τ from 0, the maximum of eq. 12 shifts to the right or the left, when compared to the global distribution given by eq. 2, and the atomic type can be considered to be "informative" in the sense that its knowledge decreases the initial uncertainty about its burial. Theoretical expressions for these informative atoms accurately reproduce the crucial shift in maximum position but the resulting curves, particularly for hydrophobic residues, tend to have slightly shorter and broader peaks when compared to the empirical distributions. Some residues whose distributions are very well adjusted by the theoretical expression, like serine and alanine, turn out however to have the hydrophobicity parameter very close to 1, or $\Delta\mu_\tau \sim 0$. In other words, the distributions for these well adjusted residues are very similar to the average distribution, except for a scaling pre-factor, and for this reason they are not very informative in the sense described above.

There is a significant improvement in the adjustments caused by the addition of the exponent α_τ and the resulting clustering of similar residues near each other in the plot shown in Fig. 9(b) strongly suggets that the curvature of the effective atomic potential is physically significant and not an artifact of the adjustment procedure. Aromatic residues (F, W and Y) and aliphatic residues (I, L, V, M and C) are all hydrophobic (large h_τ or $\alpha_\tau h^*_\tau$) and appear well to the right of the reference isohydrophobicity line $\alpha_\tau \mu^*_\tau = 1$. For the three aromatic residues, however, $\alpha_\tau > 1.35$, while for aliphatic residues $1 < \alpha_\tau < 1.35$. This distinction is likely to reflect experimentally observed qualitative differences in the thermodynamics of interaction between aromatic and aliphatic sidechains [27–30]. For hydrophilic residues, which appear to the left of the reference curve, charged amino acids (K, R, E and D) also have $\alpha_\tau > 1.35$ and can therefore be distinguished from uncharged residues (Q, P, N). Note, from the values shown in Tab. 1 or the isohysrophobicity lines in Fig. 9(b), that a single hydrophobicity parameter, h_τ or $\alpha_\tau h^*_\tau$, is unable to distinguish not only aromatic from aliphatic residues (e.g., $h_{\mathrm{F}} > h_{\mathrm{I}} > h_{\mathrm{W}} > h_{\mathrm{L}} > h_{\mathrm{Y}}$), but also hydrophilic charged from uncharged (e.g., $h_{\mathrm{K}} < h_{\mathrm{Q}} < h_{\mathrm{R}}$). Three residues (S, H and G) correspond $\alpha_\tau < 1$, or a negative curvature of the atomic effective energy. Since these three residues are close to the reference line we decided to group them as "indifferent" together with alanine and threonine, which although inside the hydrophobic aliphatic and hydrophilic uncharged regions, respectively, also have small absolute value of Δh_τ.

The use of a theoretical expression to describe the empirical distribution of atomic burials in compact globular proteins can be very important in the calculation of effective atomic potentials intended to be used in protein folding simulations. A clear advantage of this approach is the disentanglement of factors that contribute to the observed probabilities but are probably not intended to be implicitly included in the effective potential since they might be considered explicitly in folding simulations. This is the case for the volume of concentric shells, the atomic frequencies in the data bank and, possibly, the excluded volume effect. The atomic relevant intrinsic propensity to be buried or exposed to the solvent is provided by the effective potentials $\epsilon_\tau(b)$ or $\epsilon^*_\tau(b)$, which are determined by a small number of parameters that could be predicted from the amino acid sequence depending on a convenient atomic type classification, or typing scheme [31].

Theoretical expressions given by eqs. 13 and 15 appear to be sufficiently general to describe with reasonable accuracy distributions resulting from broad classifications, such as hydrophobic and hydrophilic, and also a more detailed classifications according to the twenty amino acid residues. In order to be useful in folding simulations, however, atomic types should, ideally, not only have their probability distributions accurately reproduced from sequence-dependent hydrophobicity parameters but should also be sufficiently informative or, in other words, should have the effective hydrophobicity, h_τ or $h^*_\tau \alpha_\tau$, sufficiently different from 1. More detailed classification schemes might reveal more informative types but this is not always the case. For large residues, like lysine or leucine, atoms from the end of the sidechain have a hydrophobicity significantly different from the atoms at or near the backbone but for small residues, like alanine and serine, different atoms tend to have similar values for this parameter (not shown).

The ideal classification scheme for folding simulations should maximize the number of informative atomic types. Hydrophobicities dependent on other factors in addition to the covalent structure of the protein molecule, like the presence of hydrogen bonds, salt bridges or any other environment condition, can also be easily investigated with the present theoretical framework and could be used to improve the classification scheme. A significant increase in information can be expected for backbone atoms, for example, if their burial tendency is considered to depend on the formation of hydrogen bonds. Finally, it should always be kept in mind that the local amino acid sequence certainly affects the burial tendency of individual residues although it is not clear at this point if these effects should be at least partly accounted for in the hydrophobicity values of each atom or if they should arise naturally during folding simulations as the sum of individual hydrophobicities of atoms that are near each other along the covalent structure.

4 Methods

We have used the publicly available list of protein chains PDB-SELECT which is intended to maximize coverage of conformational space minimizing redundancy [32,33]. We initially selected from the list structures determined by X-ray cristalography with resolution equal or better than 2.5 Åand coordinates for all side chains and obtained 766 structures. Since the present study is concerned with compact globular structures we have attempted to additionally filter out inadequate conformations based on two structural parameters: (a) the ratio between the radius of gyration and the cubic root of the number of monomers, $\mathcal{B} = R_g/M^{1/3}$, which decreases with overall chain compactness, where the radius of gyration for a chain with M monomers and N atoms is computed from the coordinates (x_a, y_a, z_a) of all atoms, labeled by a, relative to the geometrical center, (x_0, y_0, z_0)

$$R_g = \sqrt{\frac{1}{N}\sum_{a=1}^{N}(x_a - x_0)^2 + (y_a - y_0)^2 + (z_a - z_0)^2} \tag{16}$$

and (b) the "asphericity" parameter $\mathcal{A}$, which ranges from 0 to 1 and is small for spherically symmetric conformations [34]:

$$\mathcal{A} = \frac{(\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_2)^2}{2(\lambda_1 + \lambda_2 + \lambda_3)^2}, \tag{17}$$

where N is the number of atoms and λ_1, λ_2 and λ_3 are the three eigen values of the 3×3 shape (or gyration) tensor S, whose elements S_{ij} are given by

$$S_{ij} = \frac{1}{N} \sum_{a=1}^{N} (x_a^{(i)} - x_0^{(i)})(x_a^{(j)} - x_0^{(i)}) \tag{18}$$

with i and j varying from 1 to 3 and $x_a^{(1)}$, $x_a^{(2)}$ and $x_a^{(3)}$ representing, respectively, the x_a, y_a and z_a coordinates of atom a. Note that the trace of S is the square of the radius of gyration.

As previously observed [10] the dependence of the radius of gyration on chain length for the most compact protein structures at each length is well described by the expected scaling behavior for collapsed globular structures, $R_g \sim N^{1/3}$. $\mathcal{B}$ therefore is independent of N for these most compact conformations and, as shown in fig. 11(a), is close to 2.7. The observed strong deviations from this behavior tend to occur for non-globular conformations, such as structures consisting of single α-helices, for example, for which no correlation might be expected between solvent exposure, or burial, and reduced central distances. As shown in fig. 11(b), there is actually a general correlation between $\mathcal{B}$ and $\mathcal{A}$ although there are a few examples of very compact structures with relatively high asphericity and also of spherically symmetric structures not sufficiently compact. The presence of unusual structures in the original list might partially be related to the fact that PDB-SELECT is not a data bank of whole protein molecules but just of protein chains. It would actually be interesting to perform a more careful investigation of the dependence of our results on the composition of the structural data bank, distinguishing monomeric proteins from individual chains of oligomeric proteins, or even including whole quaternary structures. In the present study, however, we simply kept for our analysis conformations from PDB-SELECT satisfying simultaneously the conditions $\mathcal{B} < 2.9$ and $\mathcal{A} < 0.1$, resulting in the group of 321 protein chains. Our general results, however, are quite insensitive to the exact choice of these cutoff values. When we consider conformations satisfying $\mathcal{A} < 0.3$, for example, with no restriction on $\mathcal{B}$, resulting h_τ values are virtually identical ($\mathcal{C} = 0.998$) to the ones shown in Tab. 1. The distributions of $\mathcal{A}$ and $\mathcal{B}$ for the initial group of 766 structures are shown fig. 11 (c) and (d), respectively.

All adjustments were performed with the publicly available plotting program gnuplot which uses an implementation of the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm. Estimated errors for adjusted parameters are typically in the third decimal position and they are for this reason shown conservatively in Tab. 1 with two decimal digits.

5 Acknowledgements

This research was supported by a FIRCA award under NIH Grant Number R03 TW006859 and by CNPq, including a Millennium Institute for Structural Biology in Biomedicine and Biotechnology (MISB3). ALCG is supported by a graduate fellowship from CAPES.

References

1. Pereira de Araújo, A. F. Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. Proc. Natl. Acad. Sci. USA 96(22):12482–12487, Oct, 1999.

2. Pereira de Araújo, A. F. Sequence rotation in N-dimensional space and the folding of hydrophobic protein models: Surpassing the diagonal unfolded state approximation. J. Chem. Phys. 114(1):570–578, 2001.

3. Garcia, L. G., Treptow, W. L., and Pereira de Araújo, A. F. Folding simulations of a three-dimensional protein model with a non-specific hydrophobic energy function. Phys. Rev. E 64:011912, 2001.

4. Treptow, W. L., Barbosa, M. A. A., Garcia, L. G., and Pereira de Araújo, A. F. Non-native interactions, effective contact order and protein folding: A mutational investigation with the hydrophobic model. Proteins: Struct., Funct. and Genet. 49:167–180, 2002.

5. Barbosa, M. A. A. and Pereira de Araújo, A. F. Relevance of structural segregation and chain compaction for the thermodynamics of folding of a hydrophobic exact protein model. Phys. Rev. E 67:051919, 2003.

6. Garcia, L. G. and Pereira de Araújo, A. F. Folding pathway dependence on energetic frustration and interaction heterogeneity for a three dimensional hydrophobic protein model. Proteins: Struct., Funct. and Bioinf. 62:46–63, 2006.

7. Barbosa, M. A. A., Garcia, L. G., and Pereira de Araújo, A. F. Entropy reduction effect imposed by hydrogen bond formation on protein folding cooperativity: Evidence from a hydrophobic minimalist model. Phys. Rev. E 72:051903, 2005.

8. Dill, K. A. Dominant Forces in Protein Folding. Biochemistry 29(31):7133–7155, aug, 1990.

9. Reif, F. Fundamentals of Statistical and thermal Physics, chapter 3. McGraw-Hill, 1965.

10. Arteca, G. A. Scaling behavior of some molecular shape descriptors of polymer chains and protein backbones. Phys. Rev. E 49:2417–2428, 1994.

11. Meirovitch, H., Rackovsky, S., and Scheraga, H. A. Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. Macromolecules 13:1398–1405, 1980.

12. Miller, S., Janin, J., Lesk, A. M., and Chotia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.

13. Flauchère, J. L. and Pliška, V. Hydrophobic paprameters π of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur. J. Med. Chem. - Chim. Ther. 18:369–375, 1983.

14. Eisenberg, D. and McLachlan, A. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.

15. Zhou, H. and Zhou, Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins: Struct., funct. and bioinf. 54:315–322, 2004.

16. Meirovitch, H. and Scheraga, H. A. Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral and ambivalent amino acids in the interior and exterior layers of native proteins. Macromolecules 13:1406–1414, 1980.

17. Tsai, J., Taylor, R., Chotia, C., and Gerstein, M. The packing density in proteins: Standard radii and volumes. J. Mol. Biol. 290:253–266, 1999.

18. Arteca, G. A. Different molecular size scaling regimes for inner and outer regions of proteins. Phys. Rev. E 54:3044–3047, 1996.

19. Flemming, P. J. and Richards, F. M. Protein packing: Dependence on protein size, secondary structure and amino acid composition. J. Mol. Biol. 299:487–498, 2000.

20. Liang, J. and Dill, K. A. Are proteins well packed ? Biophys. J. 81:751–766, 2001.

21. Janin, J., Miller, S., Lesk, A. M., and Chotia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204:155–164, 1988.

22. Timchenko, A. A., Galzitskaya, O. V., and Serdyuk, I. N. Roughness of the globular protein surface: Analysis of high resolution X-ray data. Proteins 28:194–201, 1997.

23. West, G. B., Brown, J. H., and Enquist, B. J. A general model for the origin of allmetric scaling laws in biology. Science 276:122–126, 1997.

24. West, G. B., Brown, J. H., and Enquist, B. J. A general model for the structure and allometry of plant vascular systems. Nature 400:664–667, 1999.

25. West, G. B., Brown, J. H., and Enquist, B. J. The fourth dimensionof life: Fractal geometry and allometric scaling of organisms. Science 284:1677–1679, 1999.

26. West, G. B. and Brown, J. H. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. The Journal of Experimental Biology 208:1575–1592, 2005.

27. Pochapsky, T. C. and Gopen, Q. A chromatographic approach to the determination of relative free energies of interaction between hydrophobic and amphiphilic amino acid side chains. Protein Sci. 1:786–795, 1992.

28. Makhatadze, G. I. and Privalov, P. L. Energetics of interactions of aromatic hydrocarbons in water. Biophys. Chem. 50:285–291, 1994.

29. Makhatadze, G. I. and Privalov, P. L. Hydration effects in protein folding. Biophys. Chem. 51:291–309, 1994.

30. Pereira de Araújo, A. F., Pochapsky, T. C., and Joughin, B. Thermodynamics of interaction between aromatic-aromatic, aromatic-aliphatic and aliphatic-aliphatic amino acid side chains in water. Biophys. J. 76:2319–2328, 1999.

31. Chen, W. W. and Shakhnovich, E. I. Lessons from the design of a novel atomic potential for protein folding. Protein Sci. 14:1741–1752, 2005.

32. Hobohm, U., Scharf, M., Schneider, R., and Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.

33. Hobohm, U. and Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.

34. Baumgärtner, A. Shapes of flexible vesicles at constant volume. J. Chem. Phys. 98:7496–7501, 1993.

τ	P_τ	h_τ	μ_τ	h^*_τ	μ^*_τ	α_τ	H_1	H_2	H_3	H_4
K	0.07	0.79	0.63	0.40	0.24	1.82	-2.00	-1.35	6.1	1.05
E	0.07	0.80	0.65	0.48	0.34	1.57	-1.09	-0.87	7.3	1.02
D	0.06	0.83	0.66	0.58	0.42	1.38	-0.72	-1.05	8.2	1.01
Q	0.04	0.87	0.66	0.68	0.47	1.28	-0.74	-0.30	8.5	1.02
N	0.05	0.87	0.68	0.77	0.58	1.12	-0.69	-0.82	7.6	0.98
P	0.04	0.88	0.67	0.73	0.53	1.19	-0.44	0.98	9.9	1.00
R	0.06	0.90	0.74	0.64	0.48	1.41	-1.34	-1.37	8.5	0.98
S	0.05	0.95	0.74	0.98	0.77	0.97	-0.34	-0.05	8.2	1.02
G	0.04	0.96	0.74	1.18	0.95	0.81	0.06	0.00	7.0	1.00
T	0.05	0.98	0.79	0.90	0.71	1.10	-0.26	0.35	10.3	0.99
A	0.05	1.03	0.84	0.96	0.78	1.08	0.20	0.42	13.4	0.93
H	0.03	1.05	0.79	1.10	0.85	0.95	0.04	0.18	11.3	0.89
Y	0.06	1.09	0.91	0.82	0.65	1.41	-0.21	1.31	19.5	0.93
M	0.02	1.12	0.81	1.00	0.70	1.15	0.71	1.68	15.7	0.84
L	0.09	1.14	1.01	0.95	0.82	1.27	0.65	2.32	20.8	0.85
W	0.03	1.15	0.88	0.86	0.60	1.47	0.45	3.07	24.5	0.83
C	0.01	1.16	0.78	1.04	0.67	1.15	0.67	1.34	22.6	0.88
V	0.06	1.17	0.99	1.00	0.82	1.23	0.61	1.66	19.5	0.81
I	0.06	1.18	0.99	0.96	0.77	1.33	0.74	2.46	20.3	0.79
F	0.06	1.19	1.00	0.94	0.75	1.39	0.67	2.44	23.9	0.78

Table 1: Amino acid types, τ, corresponding to all atoms of each of the twenty amino acids labeled by their one-letter code, their frequencies in the data bank, P_τ, type-dependent hydrophobicities and chemical potentials, h_τ and μ_τ, obtained from curve adjustments shown in in figs. 4 and 5, parameters h^*_τ, μ^*_τ and α_τ, obtained from curve adjustments shown in in Figs. 6 and 7, and four hydrophobicity scales available in the literature to which h_τ (or $h^*_\tau\alpha$) can be compared. H_1 is the free energy of transfer $\Delta G_{in\rightarrow out}$, in Kcal/mol, from the interior to the exterior of globular proteins estimated by Chotia and coworkers from the frequencies observed in a set of protein structures [12]. H_2, also in Kcal/mol, is the experimental free energy of transfer from octanol to water, $\Delta G_{oct\rightarrow wat}$, obtained from the partitioning of amino acid analogs by Flauchére and Pliška [13], as quoted by Eisenberg and McLachlan [14]. H_3 is the "buriability" scale, in cal mol^{-1} Å^{-2} proposed by Zhou and Zhou [15] and H_4 is the average reduced distance $< r >$ obtained by Scheraga and collaborators for a small set of globular proteins [11].

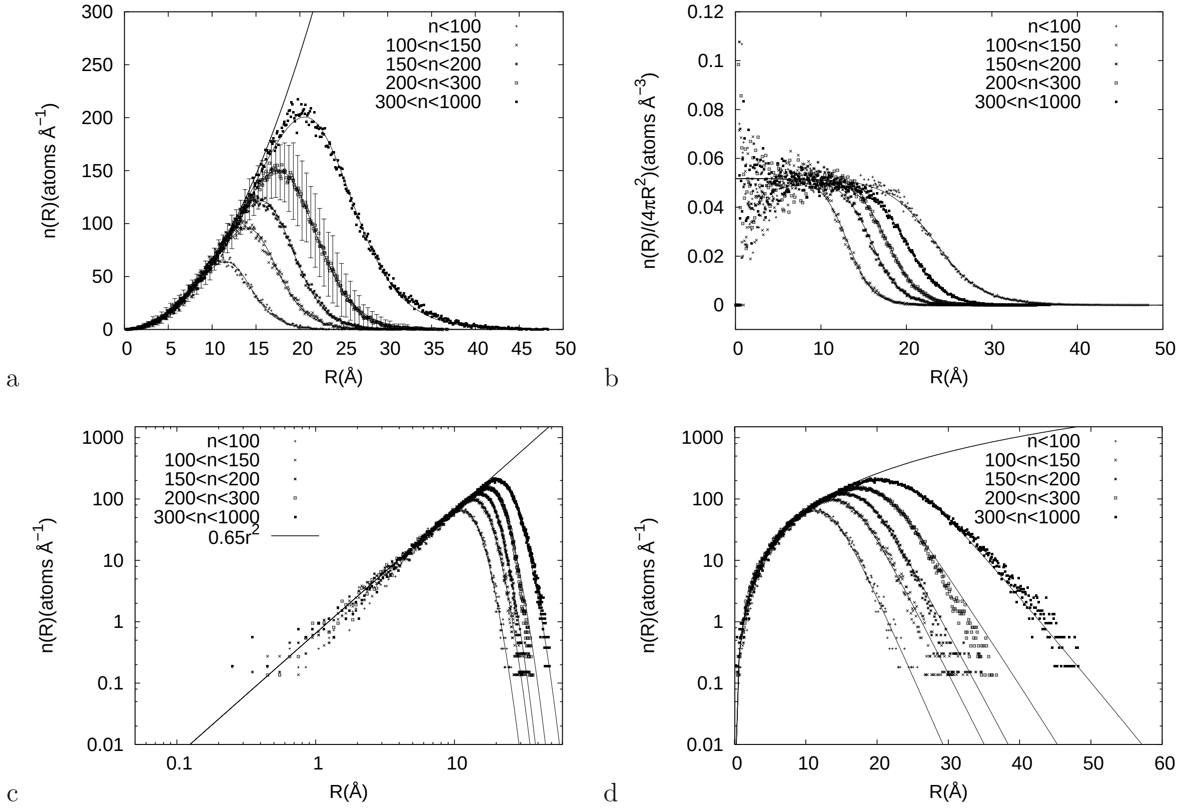

Figure 1: (a) Radial density distributions, $n(R) = \delta N(R)/\delta R$, where $\delta N(R)$ is the number of atoms with central distance within an interval of width $\delta R = 0.1$Å around R. The 321 structures in the data bank were divided into five groups according to chain length M and points represent the average of $n(R)$ within each of these groups. Typical standard deviations from the average are shown for one of the groups. Continuous lines were adjusted by the theoretical expression given by eq. 2. The parabola AR^2 with $A = 0.65\text{Å}^{-3}$ is also shown. Volumetric densities, $\frac{1}{4\pi R^2}n(R)$, obtained by division of data points and adjusted lines by $4\pi R^2$, are shown in (b). The size-independent quadratic increase of $n(R)$ observed for small R in (a) is better visualized when the same data and adjusted curves are plotted in double-log scale (c) while the near exponential decrease for large R is better visualized in mono-log scale (d).

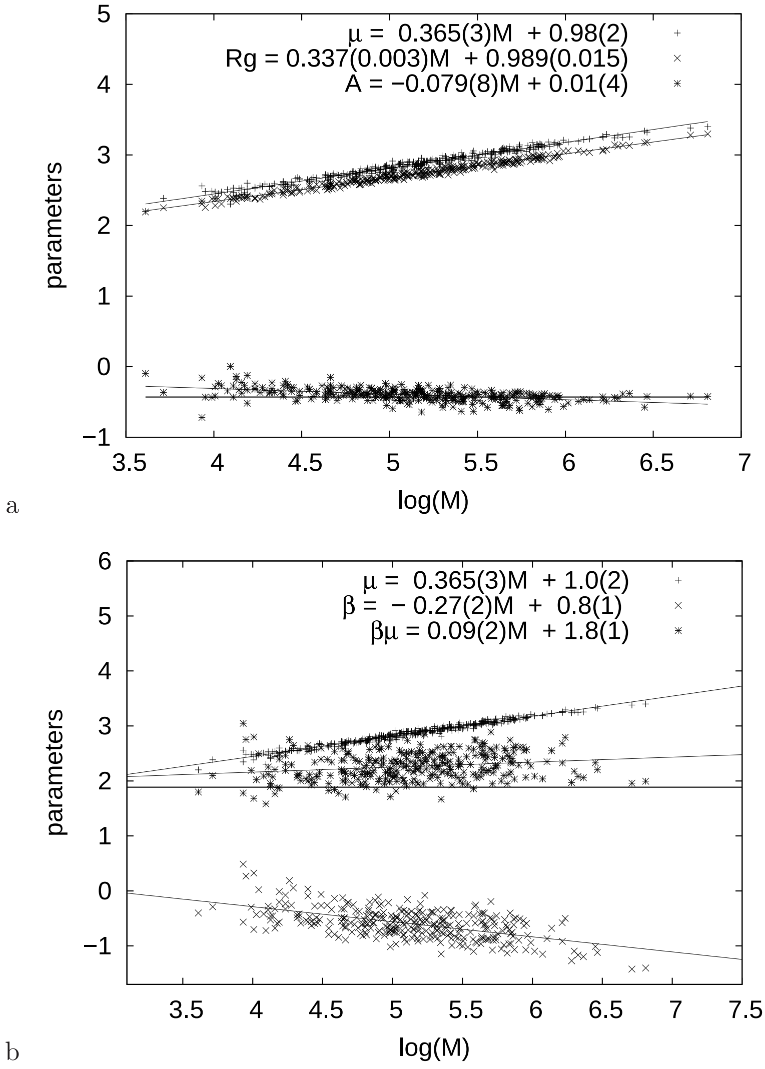

Figure 2: Dependence of the logarithm of radius of gyration, R_g, and adjusted parameters (μ, β, $\beta\mu$ and A), on the logarithm of chain length, M, for all 321 structures in the data bank. Each point represents the adjusted value by eq. 2 for a single structure while straight lines are linear fits to the data. Radius of gyration, R_g, chemical potential, μ, and the atomic density parameter, A, are shown in (a). Different scaling dependences for chemical potential, $\mu \sim M^{0.37}$, and radius of gyration, $R_g \sim M^{0.33}$, are observed while A is essentially constant and close to 0.65. The increase of μ with chain length is shown again in (b) together with the associated decrease in the temperature parameter β and resulting product $\beta\mu$ only weakly dependent on protein size. The constant line represents the theoretical prediction for perfect structural segregation.

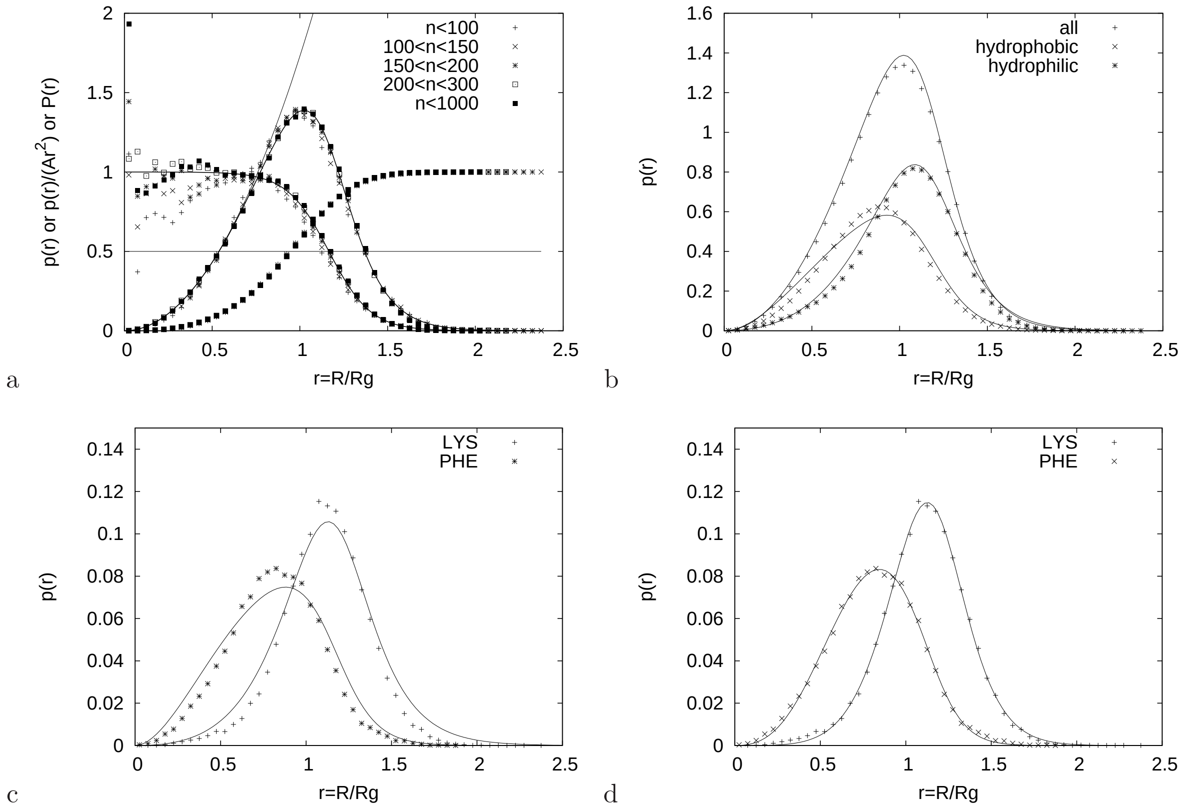

Figure 3: Curves that have a maximum near $r = 1$ in the first plot (a) represent normalized probability atomic densities $p(r)$, where $p(r)\delta r$ is the probability of an atom being found at a reduced distance between r and $r + \delta r$, while cumulative probabilities $P(r)$ of an atom being found at a reduced distance smaller than r increase from 0 to 1 and normalized atomic densities, $\rho/\rho_0 = p(r)/(Ar^2)$, decrease from 1 to 0 in the same panel. Different types of points in each set of curves correspond to averages within the same size groups shown in Fig. 1. The continuous line represents the adjustment eq. 2 to all proteins in the data bank. The same size-independent normalized probability density $p(r)$ is shown in (b) with two sub-distributions, hydrophobic and hydrophilic, adjusted by eq. 13. Sub-distributions for phenilalanine and lysine atoms are shown in (c) adjusted by eq. 13 and in (d) by eq. 15.

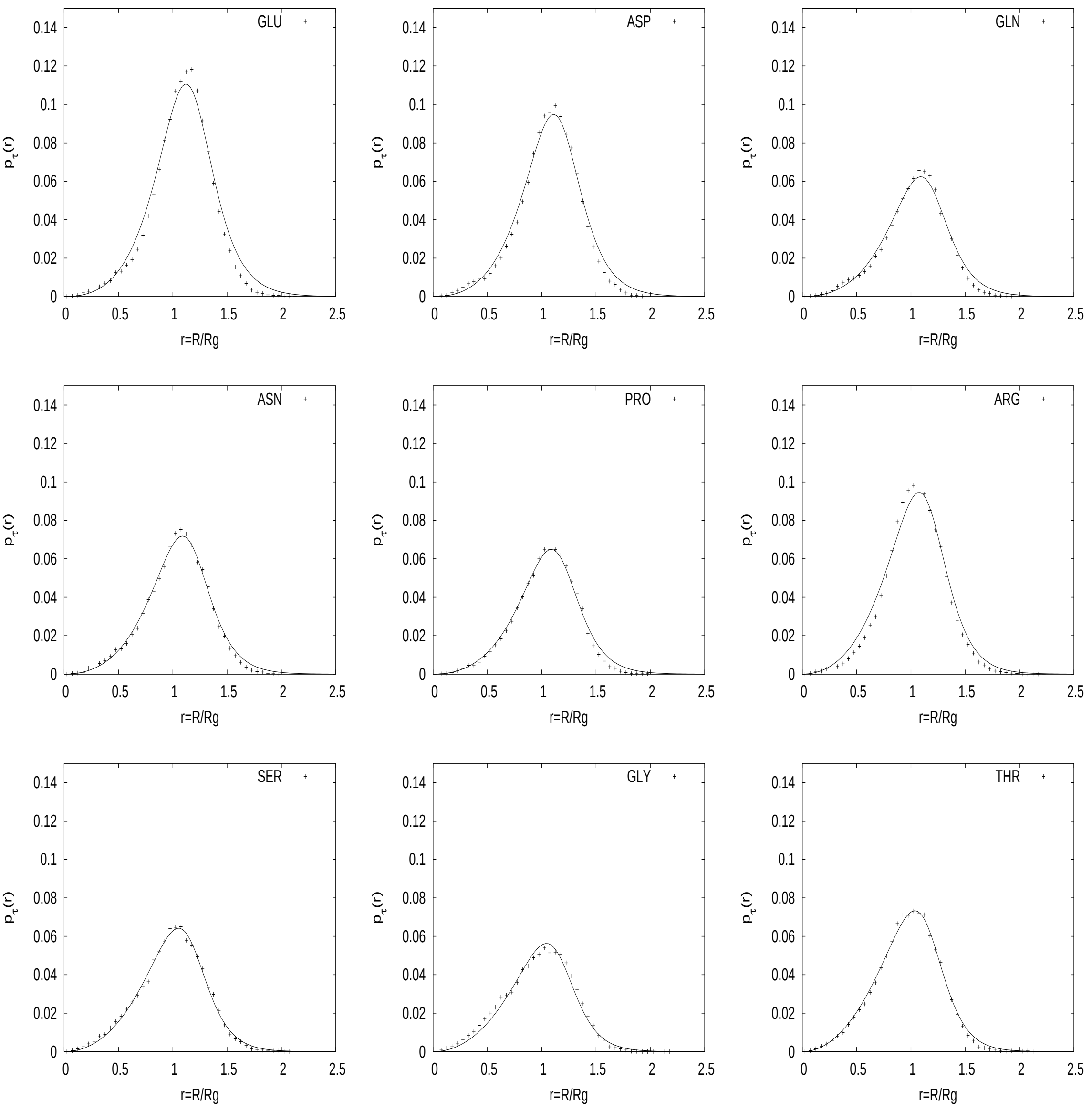

Figure 4: Radial atomic probability densities, $p_\tau(r)$, as a function of reduced distance, $r = R/R_g$, for hydrophilic residues. Points are obtained from empirical frequencies in the data bank and curves are adjusted according to eq. 13. The area beneath each curve is the total probability P_τ of the corresponding atomic type in the data bank. Global parameters A, β and μ are the same in all curves and were obtained from the adjustment of the global distribution. Adjusted parameters μ_τ affect only the height of the curves and are therefore dependent on P_τ. The shape and maximum position of the curves depend uniquely on the hydrophobicity parameter h_τ.

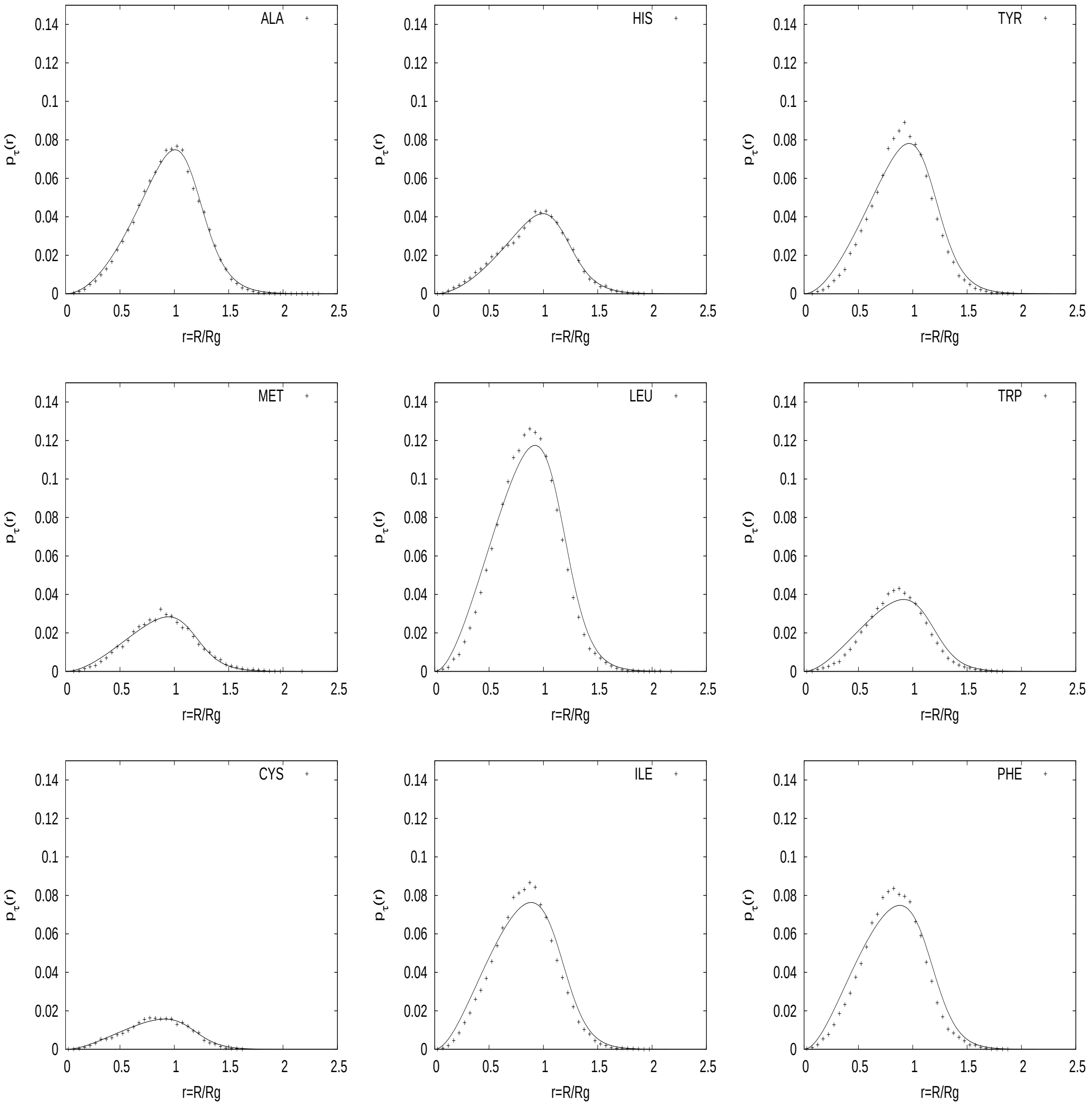

Figure 5: The same as in fig. 4 but for hydrophobic residues.

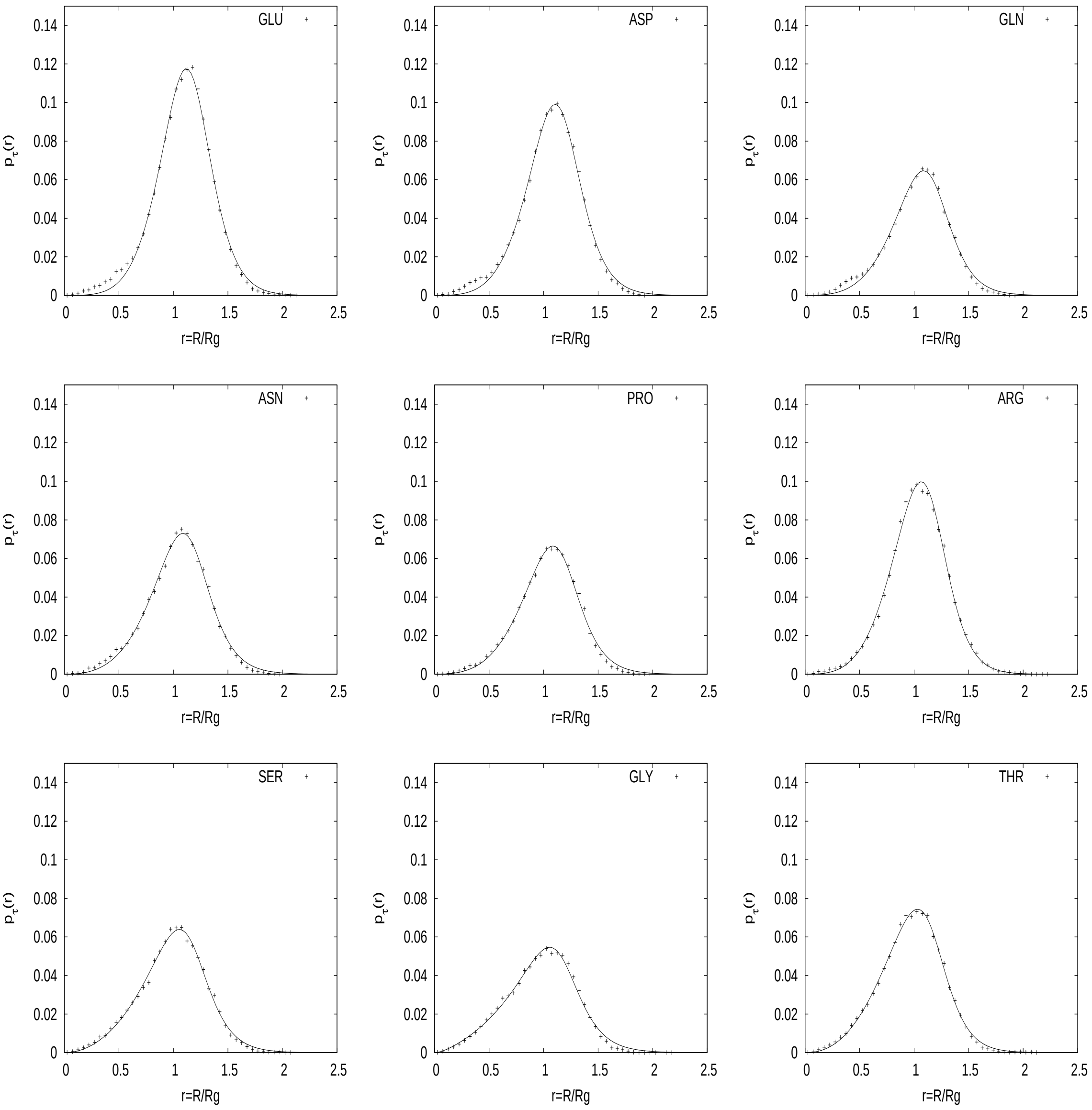

Figure 6: Radial atomic probability densities, $p^*_\tau(r)$, as a function of reduced distance, $r = R/R_g$, for hydrophilic residues. Points are obtained from empirical frequencies in the data bank and curves are adjusted according to eq. 15. The area beneath each curve is the total probability P_τ of the corresponding atomic type in the data bank. Global parameters A, β and μ are the same in all curves and were obtained from the adjustment of the global distribution. Adjusted parameters μ^*_τ affect only the height of the curves and are therefore dependent on P_τ. The shape and maximum position of the curves depend on the h^*_τ and α_τ.

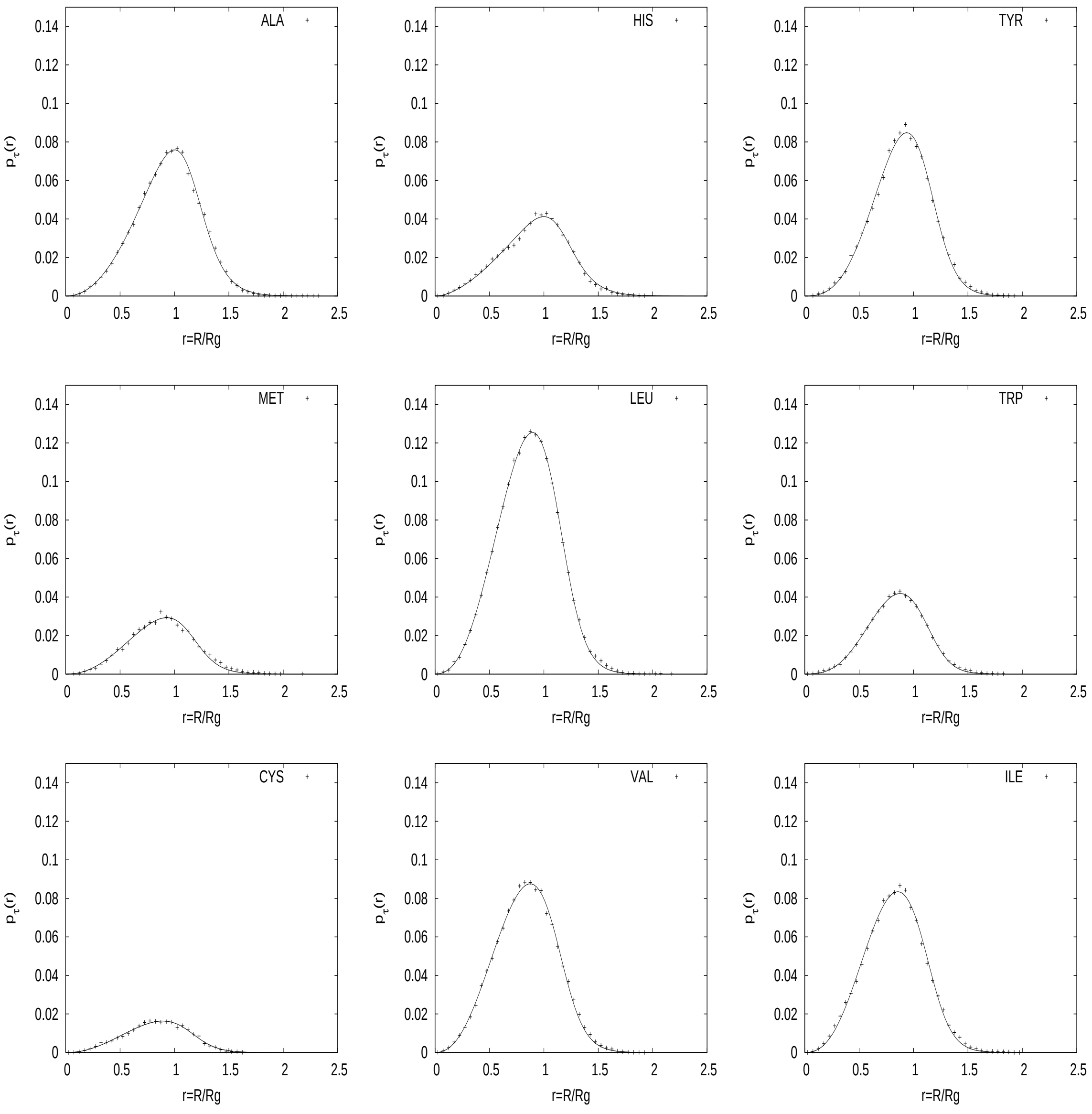

Figure 7: The same as in fig. 6 but for hydrophobic residues.

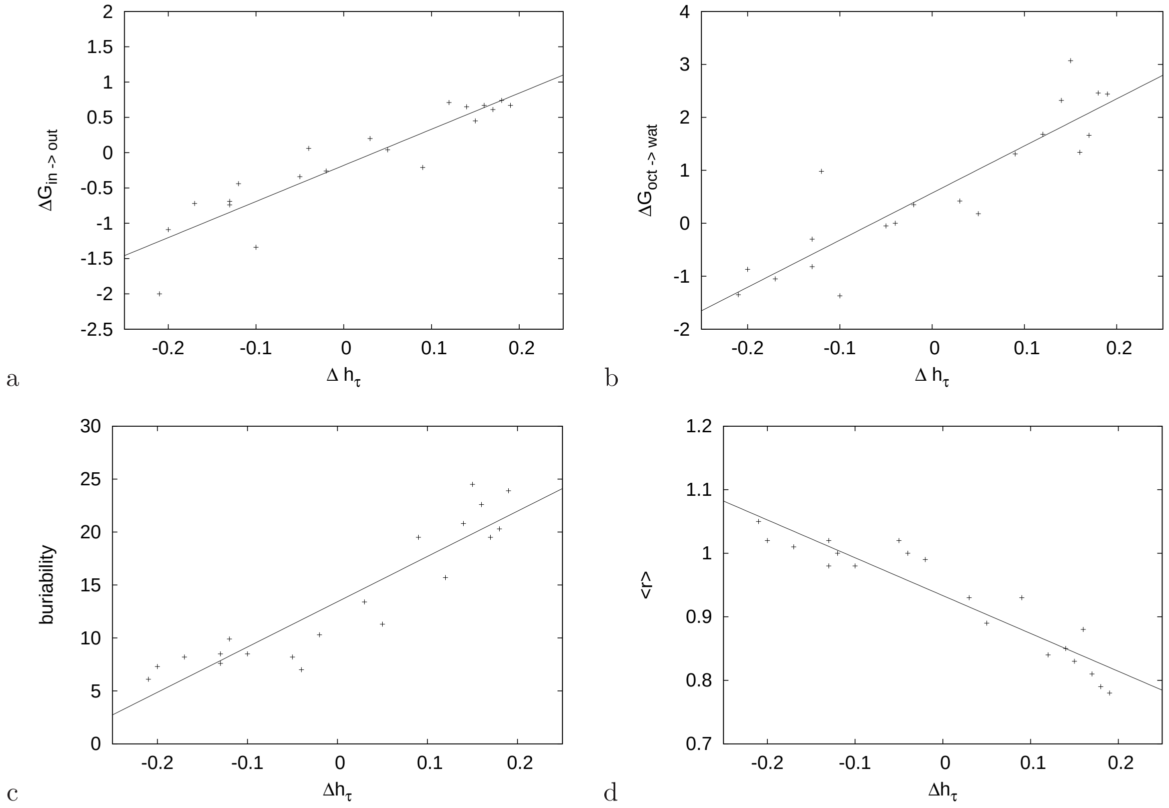

Figure 8: Correlation between the distance of the hydrophobicity parameters from the effective unity value governing the global distribution, $\Delta h_\tau = h_\tau - 1$ and hydrophobicity scales $\Delta G_{in \rightarrow out}$ [12] (a), $\Delta G_{oct \rightarrow wat}$ [13] (b), buriability [15] (c) and average reduced distance in a different set of proteins [11] . Values plotted in (a), (b), (c) and (d) are also shown in Tab. 1, where they are labeled H_1, H_2, H_3 and H_4. Pearson's correlation coefficients are $\mathcal{C} = 0.91$ in (a) and $\mathcal{C} = 0.90$ in (b), $\mathcal{C} = 0.92$ in (c) and $\mathcal{C} = -0.93$ in (d).

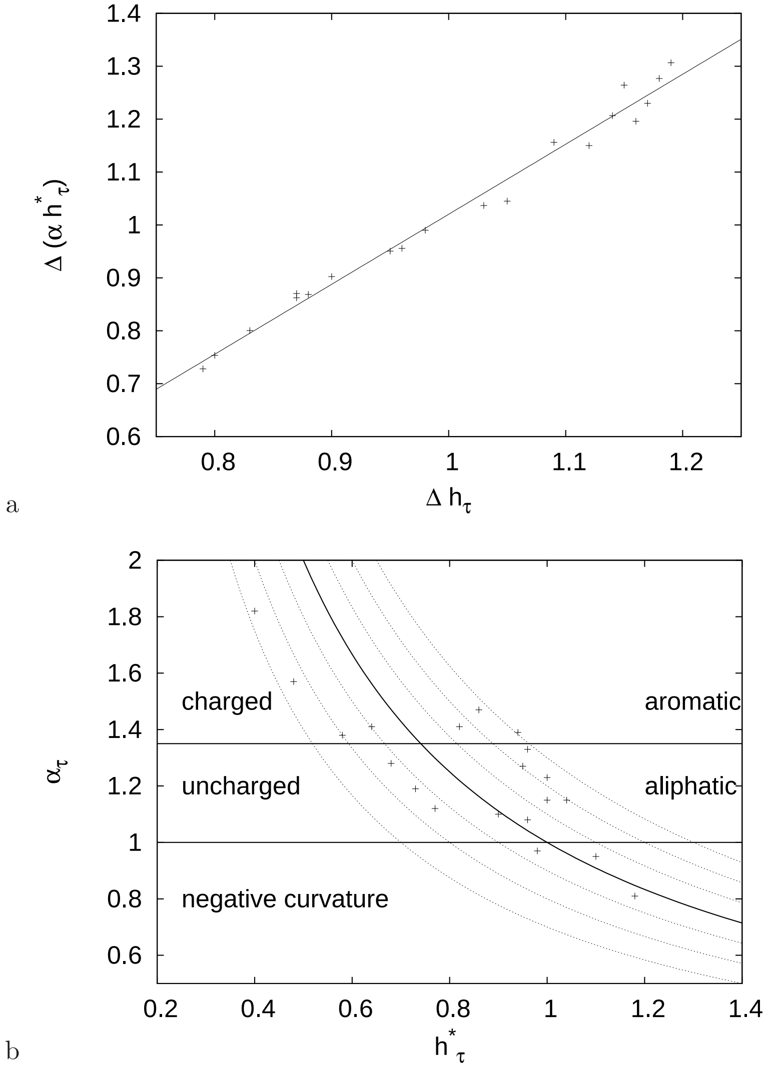

Figure 9: Correlation between the single hydrophobicity parameter h_τ and the product $\alpha_\tau h^*_\tau$, with Pearson's correlation coefficient $\mathcal{C} = 0.99$ (a) and a physically meaningful classification of amino acid residues according to h^*_τ and α_τ parameters (b). All plotted values are also shown in Tab. 1. Hyperboles of the form $\alpha_\tau h_\tau = C$ with C ranging from 0.7 to 1.3 are shown in (b) to indicate curves of isohydrophobicity. The reference hydrophobicity $\alpha_\tau h^*_\tau = 1$ is shown by a thick line.

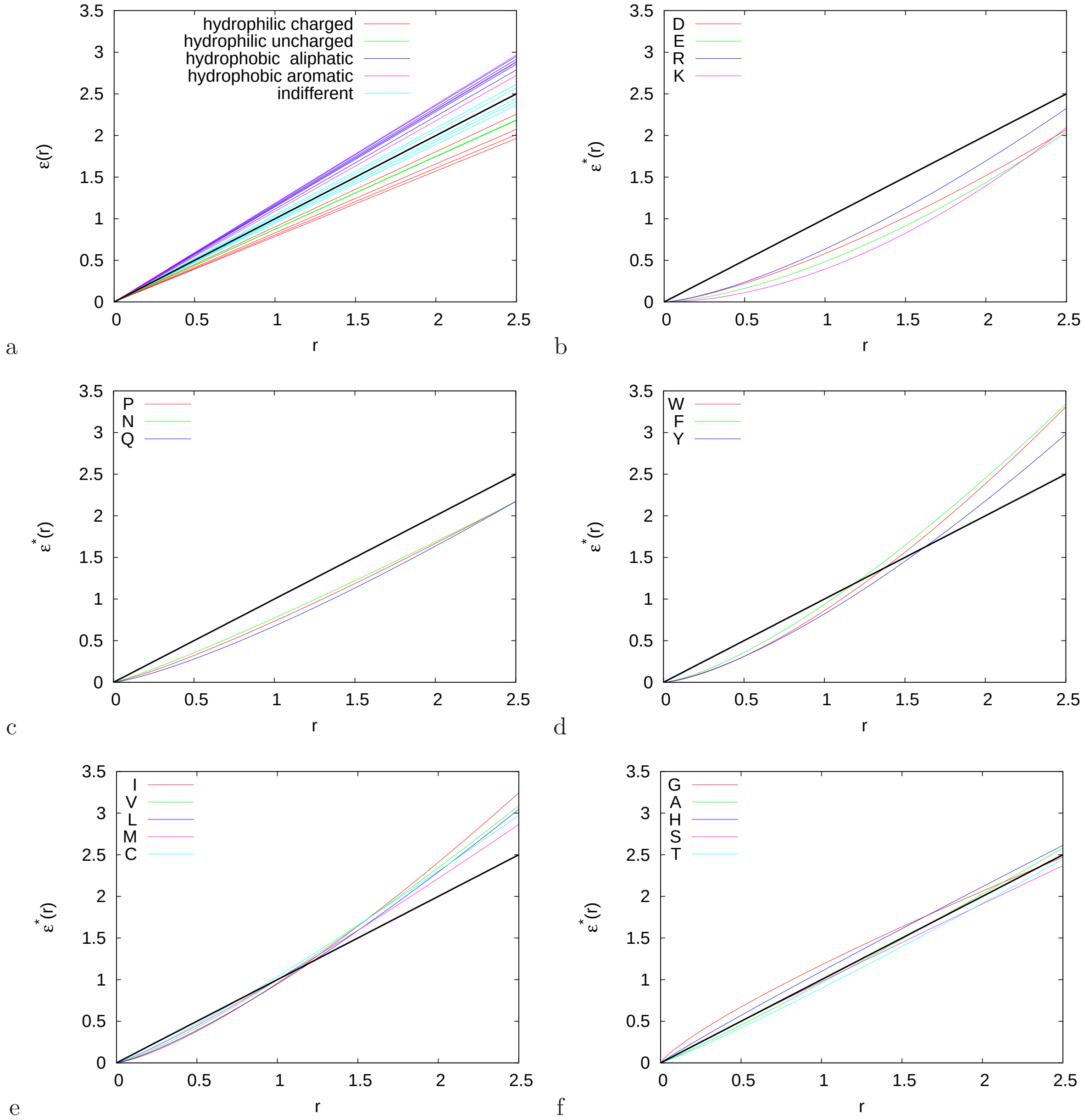

Figure 10: Type-dependent effective atomic energies for the twenty amino acid residues under the assumption of a linear dependence on reduced central distance $\epsilon(r) = h_\tau r$ (a), and under the assumption of a more general dependence $\epsilon^*_\tau(r) = h^*_\tau r^{\alpha_\tau}$ for hydrophilic charged (b), hydrophilic uncharged (c), aromatic (d), aliphatic (e) and indifferent amino acid residues (f).

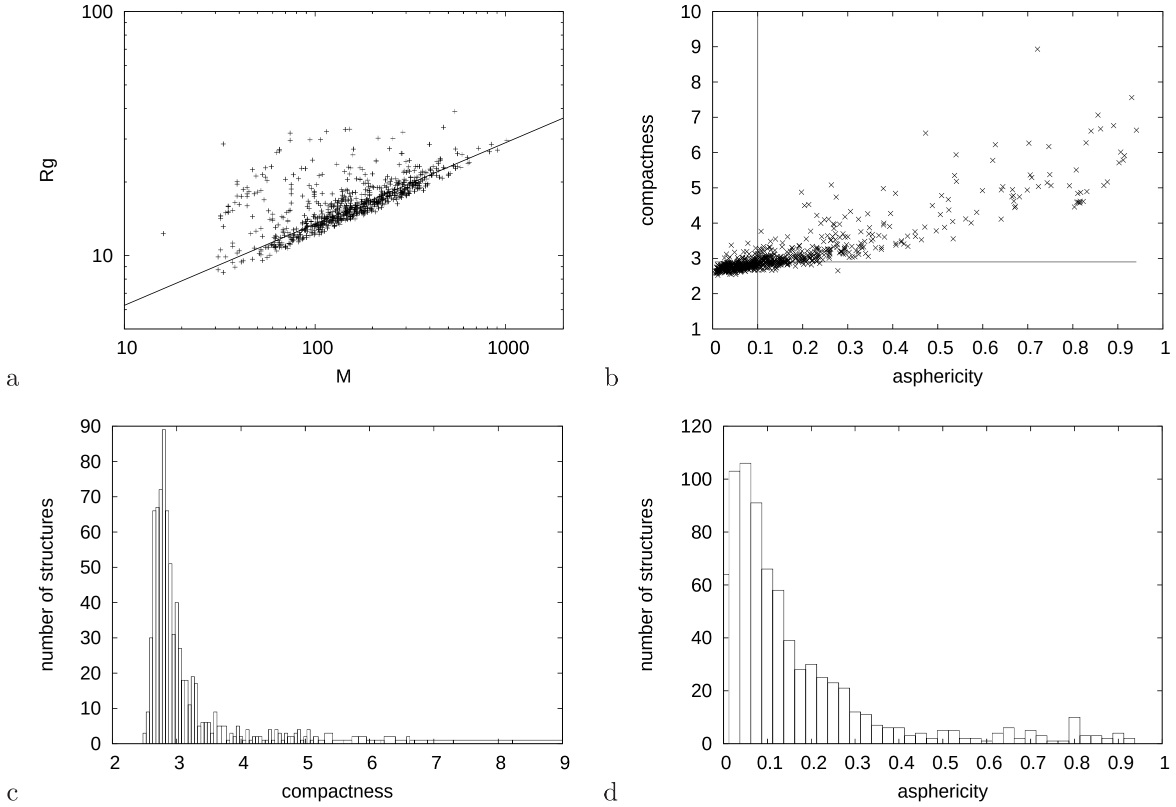

Figure 11: The dependence of the radius of gyration, R_g, on the number of monomers, M, for the initial 766 obtained from the PDB-SELECT list is shown in (a) in double-log scale. Points below the straight line correspond to sufficiently compact structures, with $\mathcal{B} = R_g/N^{1/3} < 2.9$. The dependence of the compactness, as measured by $\mathcal{B}$, on the asphericity parameter $\mathcal{A}$ is shown in (b). The 321 structures used in the present study correspond to the points with $\mathcal{A} < 0.1$ and $\mathcal{B} < 2.9$. The number of structures from the initial set as a function of compactness and asphericity are shown in (c) and (d), respectively.