J Math Chem (2012) 50:169–186
DOI 10.1007/s10910-011-9904-6

ORIGINAL PAPER

Size scaling behaviour in protein domains belonging to the all-α, all-β, α / β, and α + β folding classes

Parker Rogerson · Gustavo A. Arteca

Received: 20 July 2011 / Accepted: 19 August 2011 / Published online: 9 September 2011

Abstract We studied the size scaling behaviour in an ensemble of 8,614 non-redundant protein domains belonging to the all-α, all-β, α / β, and α + β folding classes. We find that the most compact structural domains can be characterized by an effective exponent $\nu_{eff} = 0.39 \pm 0.01$, which is larger than the value for "collapsed-polymers," i.e., $\nu = 1/3$. We also show that the global ν_{eff}-exponent is an average of the scaling regimes for short and long compact chains, where the values change from $\nu_{eff} \approx 0.37$ to $\nu_{eff} \approx 0.45$ at chain length of *ca.* 269. A transition from short-chain to long-chain scaling behaviour is found in all major folding classes, over a window of chain lengths between 216 and 269 residues. In addition, variations in scaling exponent with respect to folding class indicates that the smallest domains in the (all-β) and (α / β) families appear to be more compact structures than the smallest (all-α)- and (α + β)-domains.

Keywords Polymer size · Protein folds · Folding families · Protein domains · SCOP database

1 Introduction

When averaged over all accessible configurations, a dilute random polymer solution at equilibrium shows a mean radius of gyration that scales with the number of monomers n as $< R_g^2 >^{1/2} \sim n^{\nu}$, where ν is the *size-scaling exponent* [1–3]. This exponent provides key information about the global interactions that dominate the formation of

P. Rogerson · G. A. Arteca (✉)
Département de Chimie et Biochimie and Biomolecular Sciences Program, Laurentian University, Ramsey Lake Road, Sudbury, ON P3E 2C6, Canada
e-mail: Gustavo@laurentian.ca

the three-dimensional structure of each isolated polymer chain. When the attractive interactions dominate (or in the presence of a poor solvent), chains resemble *collapsed polymers* (CP) characterized by the scaling exponent associated with spherical shape and compactness, $\nu_{CP} = 1/3$ [1]. In the Θ-condition (or in a Θ-solvent), where we find a balance of repulsion and attraction, chains resemble *random walks* whose size is characterized by the scaling exponent $\nu_{RW} = 1/2$ [1]. In contrast, when repulsions dominate over attractions (or in the presence of a good solvent), polymer chains adopt the more elongated form of *self-avoiding-walks*, whose mean size shows the scaling exponent $\nu_{SAW} = 0.588 \pm 0.002$ [2,3]. In the case of rigid rods or stretched polyelectrolytes, we find a dominant single configuration where the persistence length of the polymer is comparable to its contour length, leading to an exponent $\nu_{Rod} = 1$ [4].

In this work, we are interested in the size-scaling behaviour of a particular class of non-random heteropolymers, namely, single or individual proteins domains. Specifically, we deal with domains derived from the all-α, all-β, $\alpha\,/\,\beta$, and $\alpha + \beta$ folding classes. These structures possess a single native state, and their molecular sizes give rise to a narrow window of values of the radius of gyration (R_g) rather than the canonical average $< R_g^2 >^{1/2}$, used to describe and characterize polymers in solution.

For a given number of monomers n, protein native states exhibit a large diversity in size that cannot be captured by a single scaling behaviour [5–7]. A *subgroup of the smallest globular proteins* (i.e., those with the smallest radius of gyration within a window of chain lengths) do however resemble qualitatively the collapsed-polymer regime for relatively short chains (e.g., $n < 300$) [5–7]. Nevertheless, the significance of this result is somewhat ambiguous because the ensemble of proteins evaluated was derived from a small initial dataset comprising both single- and multi-domain proteins, as well as many different folding types. In this work, we work with a greatly expanded dataset, and focus on a better defined ensemble of structures. Our goal is to understand the scaling regimes for individual *structural domains*, i.e., separate folding units in proteins (extracted from both single and multidomain proteins) [8–13]. In particular, we deal with the domains belonging to the four major *folding classes* (*FC*s), that is, the all-α, all-β, $\alpha\,/\,\beta$, and $\alpha + \beta$ folds [14–17].

The existence of commonalities in mean chain size and residue packing in protein domains is important for understanding their structure, function, and role in folding mechanisms [18–23]. Our main tool of analysis is the occurrence of scaling behaviour between chain length and molecular size for the most compact single domains selected from both single- and multidomain chains.

This paper is organized as follows. In the next section, we discuss the protocol implemented to select a non-redundant ensemble of protein domains, and the shape descriptors used for their size scaling characterization. In Sect. 3, we discuss the general properties of the entire distribution of domain sizes within the present data set. In the case of "maximally-compact" domains (i.e., the domains with the smallest size within a fixed window of chain lengths), we discuss the role of chain length on size scaling behaviour. Section 4 presents the distinct scaling behaviours associated with protein domains belonging to each of the major folding classes: all-α, all-β, $\alpha + \beta$, and $\alpha\,/\,\beta$. Section 5 highlights the deviations from global scaling of selected "common folds" within these folding classes. We close with a summary of conclusions.

2 Shape characterization within an ensemble of non-redundant protein domains

2.1 Molecular size scaling

We deal with a selected ensemble of individual domains, comprising both single domain chains and individual domains extracted from multidomain chains. We follow the definitions of domain topology provided by the Structural Classification of Proteins (SCOP) data base (version 1.73) [14–17], and use the coordinate data of the selected domains extracted from the Protein Data Bank (PDB) [24,25]. Protein size and density is characterized by the radius of gyration (R_g) of the α-carbon trace. When considering a full chain containing n amino acid residues, the expression for the radius of gyration becomes:

$$R_g = \left\{ \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{r}_i - \mathbf{r}_0||^2 \right\}^{1/2}, \tag{1}$$

where $\mathbf{r}_i$ and $\mathbf{r}_0$ are, respectively, the position vector of the ith α-carbon and $\mathbf{r}_0$ the geometrical centre of the α-carbon chain. Protein chains whose mean size is denoted by R_g can include a number of domains. In contrast, the radius of gyration of an *individual domain* is denoted by r_g, where $r_g \leq R_g$; in the case of r_g, the set of $\{\mathbf{r}_i\}$-coordinates in Eq. (1) is restricted to the *contiguous* section of primary sequence that spans a particular structural domain. In this work, we are concerned with the domain radii r_g. Evidently, in the case of single-domain proteins we have $R_g = r_g$.

In systems with true asymptotic scaling behaviour (e.g., random homopolymers at high dilution), R_g is a configurationally-averaged value that follows the power law: [1]

$$R_g \sim L n^{\nu}, \tag{2}$$

where $n \gg 1$ and ν is the size-scaling exponent. As explained in Sect. 1, this exponent depends only on the dominant interactions between monomers, as well as on whether the polymer is two- or three-dimensional (i.e., flatted by adsorption or embedded in 3D-space). In contrast, the pre-exponential length L depends on the particular details of polymer shape, e.g., type of monomer, sequence, secondary-structural content and type of fold. Although protein chains seldom surpass $n = 1,000$ (prohibiting an assessment of true asymptotic scaling behaviour), it is still possible to estimate an effective scaling exponent that fits the law in Eq. (2) (within a range of n values) [5–7]. Whenever a subset of protein domains demonstrates a comparable scaling behaviour, we will write:

$$r_g \sim \ell n^{\bar{\nu}}, \tag{3}$$

where the ℓ-length and the $\bar{\nu}$-exponent play the same role as L and ν in Eq. (2), constrained to an individual domain.

Families of domains that share an effective $\bar{\nu}$-exponent (cf. Eq. 3) will be said to be in the same "compactness regime." If we focus our analysis within a given folding class, the above scaling law can be rewritten with a more specific designation:

$$[r_g]_{FC} \sim \ell_{FC} n^{\bar{\nu}_{FC}}, \qquad (4)$$

where the subindex takes the values $FC = \alpha, \beta, \alpha + \beta, \alpha / \beta$ depending on the particular folding classes being considered. Domains in the "collapsed-polymer regime" satisfy $\bar{\nu}_{FC} \approx \bar{\nu}_{CP} = 1/3$. However, belonging to a scaling regime with an exponent $\bar{\nu}_{FC} = 1/3$ does not ensure that the domains will have the smallest possible absolute size. It is possible for some domains to belong to a less compact regime (i.e., $\bar{\nu}_{FC} > 1/3$) and have smaller r_g-values than some domains whose size is commensurate with the $\bar{\nu}_{FC} = 1/3$ regime.

2.2 Selection of non-redundant single domains

We used the SCOP data base as the basis to organize our selected ensemble of protein domains. This data base organizes domains into lineages of "common folds" within larger "folding classes" based on similarities in folding topology [14–17]. Entries in SCOP are manually curated; domains are inspected visually and classified according to consensual, albeit subjective criteria. In this work, we consider the four principal folding classes (or *root nodes*), corresponding to the (all-α), (all-β), (α + β), and (α / β)-folds. The all-α and all-β folds consist almost exclusively of helical and β-sheet structures, respectively, while the α + β and α / β classes contain varying degrees of both secondary-structural elements. In (α + β)-domains, helices and antiparallel-sheets are spatially segregated; in the (α / β)-folds, helices and β-sheets typically alternate, allowing the β-strands to organize in a parallel fashion, e.g., the TIM-barrels [26,27].

Given the high level of redundancy in the PDB and SCOP data bases, it is important to avoid biasing the size-scaling analysis by eliminating all duplicate entries from our data set. We devised the following selection protocol to ensure one entry per domain type:

(a) Only one structure was used per domain among entries with no missing residues and at least a 3.2Å-resolution.
(b) Domains with more than 90% sequence homology were represented by a single entry, unless they differed in chain length by more than fifteen residues, in which case they are considered distinct entries.
(c) Very short chains were deemed poorly-structured peptides and often omitted; typically, but not always, protein chains with less than 35 residues appear as "outliers" in our analysis.

The present study began with an ensemble of 85,686 single domains in the SCOP data base, with the following breakdown in terms of folding class: 14,824 for $FC = \alpha$ (i.e., (all-α)-domains), 23,547 for $FC = \beta$, 21,499 for $FC = \alpha + \beta$, and 25,816 for $FC = \alpha / \beta$. When subject to the above screening procedure, the set is reduced roughly to about 10% of the total entries. Specifically, we retain 8,614 non-redundant structures, with the following distribution according to folding class: 1,741 (all-α)-domains, 2,527 all-β, 2,099 $\alpha+\beta$, and 2,247 (α / β)-domains. This list includes individual domains associated with both single- and multi-domain proteins. The radius of gyration of each chain was computed from the α-carbon coordinates extracted from entries in the PDB archive. In the next section, we use these results to analyze the size-scaling behaviour in isolated protein domains.

3 Effect of chain length on the size scaling of individual protein domains

As noted in the literature [5–7], the wide span of molecular size distributions ensures that scaling behaviour can only be observed in restricted subensembles of proteins. Here, we focus on the scaling associated with the "most compact" protein domains, i.e., domains that have the smallest r_g-values for a given chain length. In practice, we have grouped the selected entries in bins of Δn-length, beginning from a minimum $n_0 = 26$, and then selected the smallest structure within those bins:

$$\min_{\{N_j\}} \left[r_g\right]_j = \left[r_g\right]_j^* \tag{5}$$

where $[r_g]_j^*$ is the domain with the smallest radius of gyration among the N_j-domains found within the jth-bin. If we consider the $\Delta n = 10$ case, a generic jth-bin will contain (at most) the ten structures with the smallest r_g-values for domains with chain lengths in the range $(n_0+(j-1)\Delta n) \leq n \leq (n_0+j \quad \Delta n)$ with $j \geq 1$; the $[r_g]_j^*$-value is the smallest value for the structures in that bin. By applying the criterion in Eq. (5), we obtain a set of $\{[r_g]_j^*, n_j\}$-values, where n_j denotes the number of residues for the selected domain in the jth-bin, and $[r_g]_j^*$ corresponds to the r_g-value of the selected smallest domain.

Figure 1 shows the radii of gyration for the entire set of 8,614 non-redundant domains (in small grey circles). In addition, the black squares in Fig. 1 denotes the domains satisfying criterion Eq. (5) for a $\Delta n = 10$ bin size.

The first point in Fig. 1 is the WW-domain of the YJQ8 yeast protein (PDB entry 1e0n, chain A). This (all-β)-structure is the smallest domain in the first bin (i.e., $26 \leq n \leq 35$), corresponding to $n_1 = 27$ and $[r_g]_1^* = 8.451\text{Å}$. We have sufficient number of entries to select compact structures up to the 66th bin ($685 \leq n \leq 695$), before we encounter the first "empty bin," i.e., no domains in the selected set. From that bin onwards, the $[r_g]_j^*$-values become progressively more sporadic up to $n \approx 800$ and rare after that; the last entry in Fig. 1 corresponds to nitrate reductase A in *E. coli* (PDB entry 1y5i, chain A), corresponding to $n = 1,074$ and $[r_g]^* = 30.841\text{Å}$.

Despite the large dispersion in r_g-values in Fig. 1, it is apparent that the smallest isolated domains exhibit not only an effective size-scaling behaviour, but also what appears to be more than one scaling regime in terms of $\bar{\nu}$-exponents. Upon close inspection, we find clear evidence of two *distinct* size-scaling exponents over the range of chain lengths considered, denoted here as $\bar{\nu}^{(short)}$ and $\bar{\nu}^{(long)}$.

Figure 2 highlights the results for the most compact $\{[r_g]_j^*, n_j\}$-pairs included in Fig. 1, emphasizing the two scaling regimes for short and long chains, as well as a

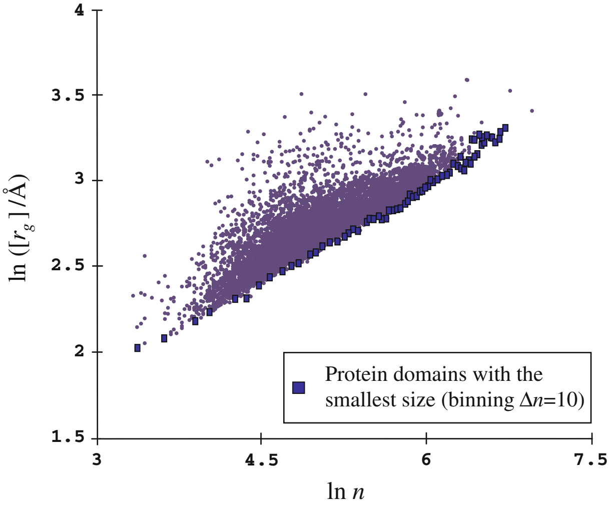

Fig. 1 Distribution of radius of gyration (r_g) for the ensemble of 8,614 non-redundant (isolated) protein domains. The *black squares* represent the most compact domains with the smallest r_g-values within the windows of chain length $(n, n + \Delta n)$, using a $\Delta n = 10$ bin window

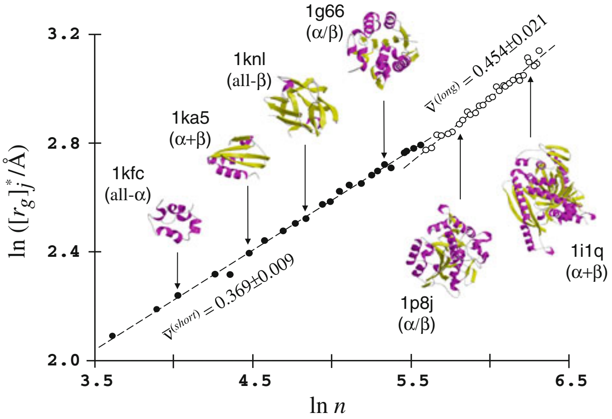

Fig. 2 Occurrence of two optimal scaling regimes in the most compact protein domains. The *black squares* in Fig. 1 are divided into two ensembles: "short chains" (*black circles*) and "long chains" (*white circles*). The *highlighted* structures illustrate that all folding classes are represented within this set. The figure gives the scaling $\bar{\nu}$-exponents arising from the linear regressions. See Fig. 3 and the text for the criterion used to determine the "transition" chain length $n^*_{1,2}$ that divides these two scaling regimes

number of representative structures. The insets show that all four major folding classes are represented among the most compact domains.

The black circles in Fig. 2 span the range of "short-chain" regime, while the white circles define the "long-chain" regime. The subsets provide optimal fittings to the scaling law in Eq. (3), when maximizing the correlation coefficients in a sequence of linear regressions with the criterion illustrated in Fig. 3.

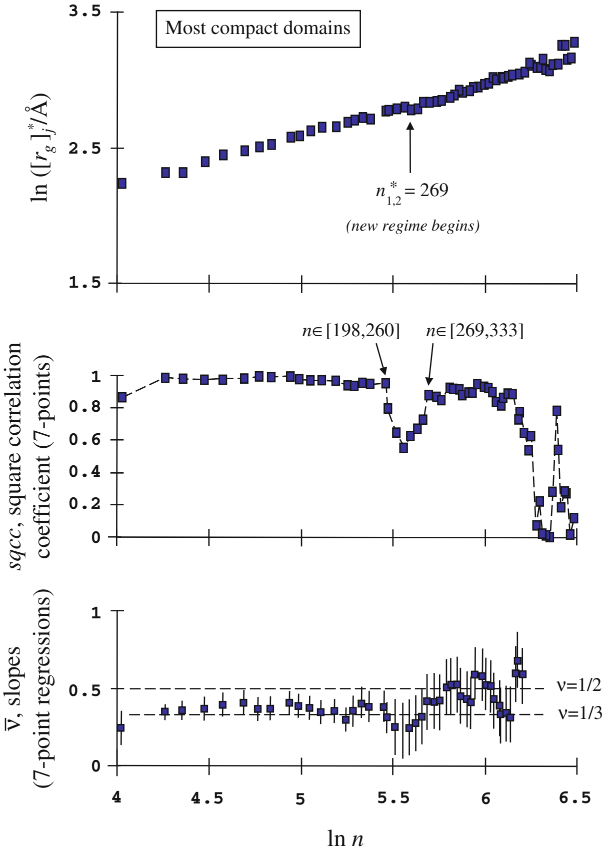

Fig. 3 Determination of the $n^*_{1,2}$ chain length for the transition between the short- and long-chain regimes in Fig. 2. The *middle* diagram gives the sequence of square correlation coefficients (*sqcc*) arising from consecutive 7-point fittings of ($[r_g]^*$, n)-values in the *top* diagram. (The figure indicates the range of n-values corresponding to the last fitting entry for the short-chain regime, $n \in [198, 260]$, and the first fitting entry for the long-chain regime, $n \in [269, 333]$.) The *bottom* diagram gives the corresponding sequence of slopes in the 7-point regressions (i.e., the $\bar{\nu}$-exponents), with their 95% confidence intervals. The exponents for collapsed polymers ($\nu = 1/3$) and random walks ($\nu = 1/2$) are given for reference. The *arrow* in the top diagram indicates the location of the transition, as marked by the sudden loss of correlation in the middle diagram

The top panel in Fig. 3 depicts the set of radii of gyration for $\ln n \geq 4$ (cf. Fig. 2). From this data set, we produce a series of linear fittings, each generated from *seven* consecutive data points. The square correlation coefficients (*sqcc*) arising from these regressions appear in the middle diagram, where each 7-point fitting is plotted at the n-value for the *middle* point in the set (i.e., the fourth point of the seven). For example, the first point in the diagram corresponds to the seven consecutive bins with entries $n \in [27, 88]$, while the second point corresponds to the entries $n \in [37, 97]$ (i.e., dropping the entry corresponding to $n = 27$ and incorporating the one for $n = 97$). The corresponding entries for the regression of these 7-point slopes (i.e., $\bar{\nu}$) are depicted in the bottom diagram; error bars represent 95% confidence intervals.

An analysis of the square correlation coefficients reveals two scaling regimes:

(a) Up to the 7-point fitting beginning with entry $n = 198$, i.e., the dataset $n \in [198, 260]$, we find a high-quality correlation with $0.932 \leq sqcc \leq 0.994$. The slopes associated with these fittings yield an average $\bar{\nu} = 0.371 \pm 0.014$ with 95% confidence, after omitting the first point in the middle diagram as an outlier.

(b) The quality of the 7-point correlations diminishes for $n > 198$; for example, including the next entry with $n = 207$ changes *sqcc*=0.943 to *sqcc*=0.796. This drop correlates with a change in $\bar{\nu}$-value and a larger statistical error (cf. Fig. 3, bottom diagram).

(c) The linear regressions improve again for longer chains. Within the range $269 \leq n \leq 441$, the square correlation coefficient satisfies $0.844 \leq sqcc \leq 0.944$; we also find a range of slopes with average $\bar{\nu} = 0.450 \pm 0.040$, consistently above the values for shorter chains in (a), despite the larger uncertainty. The beginning of this second scaling regime is indicated in Fig. 3 by the 7-point fitting with $n \in [269, 333]$ (i.e., corresponding to the bins with entries $n = 269, 279, 288, 297, 310, 318$, and 333). The first point in this set determines the transition from short- and to long-chain scaling behaviour (indicated by the arrow in Fig. 3 (top) at $n^*_{1,2} = 269$). Note that there are *no* "excluded" bins, since the last entry for the short-chain regime is $n = 260$ and the first entry for the long-chain regime is $n = 269$.

(d) The linear fittings worsen for $n > 543$, except for a small window of high correlation in the interval $n \in [560, 723](sqcc > 0.894)$ with a larger effective scaling exponent, $\bar{\nu} = 0.83 \pm 0.13$. This result suggests that a third scaling regime may be present for single-domain compact proteins with very long chains; however, given the scarcity of entries beyond $n = 543$, the precise value of the corresponding scaling exponent is difficult to ascertain.

Our conclusions are summarized by the two linear fittings displayed in Fig. 2. The two regression lines, computed with entries before and after $n^*_{1,2} = 269$, provide a clear distinction between two effective size-scaling exponents for single domains:

$$\bar{\nu}^{(short)} = 0.369 \pm 0.009, \text{ (twenty-three points with } 37 \leq n \leq 260), \tag{6a}$$

$$\bar{\nu}^{(long)} = 0.454 \pm 0.021, \text{ (twenty-six points with } 269 \leq n \leq 543), \tag{6b}$$

with 95%-confidence intervals. If we ignore the $n^*_{1,2}$-transition, a fitting over all points in Fig. 2, for $37 \leq n \leq 543$, gives a global exponent $\bar{\nu} = 0.373 \pm 0.007$ (cf. Eq. 3). If extended to the entire ensemble of smallest domains in Fig. 1, we obtain $\bar{\nu} = 0.388 \pm 0.009$ (sixty-five points with $37 \leq n \leq 796$, with longer chains omitted as outliers due to low sampling). Both values match the effective $\bar{\nu}$-exponent estimated in full protein chains [5–7], independent of domain number. (Ref. [6] finds $\bar{\nu} = 0.38 \pm 0.02$ from a small set of proteins extracted with a $\Delta n = 50$ binning). However, Eq. (6) indicates that this exponent is an average of two distinct regimes: one of them slightly above that of collapsed chains (i.e., $\bar{\nu}^{(short)} \approx 0.37 > \nu_{CP} = 1/3$), while the other is closer to the size-scaling law for random chains ($\bar{\nu}^{(long)} \approx 0.45 > \nu_{RW} = 1/2$).

The consistency of these observations has been corroborated with two other analyses:

(i) We re-estimated the location of the $n^*_{1,2}$-transitions by analyzing the sequence of linear fittings comprising q-points, with $3 \leq q \leq 6$. Although these correlations have larger statistical errors, they indicate transitions between two distinct scaling regimes at roughly the same positions as the $q = 7$ fitting (i.e., within the same bin or its two nearest neighbours). However, the $q = 7$ case is "optimal" in the sense that it defines a two-regimes transition over *contiguous* bins, as the last entry used to fit the short-chain regime is $n = 260$ while the first entry in the fitting for the long-chain regime is $n = 269$.

(ii) We produced alternative sets of most compact domains by using other $\Delta n = 10$ binning schemes. For example, instead constructing our ensemble from a first bin $n \in [26, 35]$, we have repeated our study using all possible starting points for the first bins, i.e., $n \in [26, 35]$, $n \in [27, 36]$, and so forth until the last possible choice of initial $\Delta n = 10$ binning, i.e., $n \in [35, 44]$. These alternative binning schemes often produce a different $\{[r_g]^*_j\}$-selection, yet linear correlations that cannot be distinguished within statistical error.

The existence of distinct size-scaling regimes for short and long chains has been noted previously in the literature [5–7], and tentatively attributed to the more likely occurrence of an onset of multiple domains in long-chain proteins. The results in this section show that this difference is *an intrinsic property of single-domain* scaling. We can infer that short and long compact proteins (i.e., those with $n \geq 269$) are organized structurally in a different fashion. In the next section, we discuss how these scaling regimes relate to folding classes.

4 Effect of folding class on size-scaling behaviour

Using the SCOP classification, we have extended the previous analysis to determine the effect of folding class on the $\bar{\nu}_{FC}$-exponents introduced in Eq. (4). As in Sect. 3, we determine the domains with the smallest radius of gyration within a given bin of chain lengths. The set of molecular sizes for the most compact entries within a given folding class is denoted as $\{[r_g]^*_{j,FC}\}$, corresponding to the ensemble of jth-bins for FC-domains for a particular bin selection. Here, we use $\Delta n = 10$ for all folding classes.

Figure 4 sorts the original distribution of 8,614 non-redundant single domains into the four major folding classes (cf. Fig. 1). Despite the wide scattering, it is clear that the distributions are not equal. The biggest differences can be seen between the (all-α)- and (α / β)-domains: whereas the former contains the fewer number of entries (namely, 1,741) and the largest dispersion, the latter features a narrow, highly correlated distribution of 2,247 domains. Even a cursory inspection indicates the existence of scaling laws for $\{[r_g]^*_{j,FC}\}$-values within each of the *FC-families*. In the case of the (all-α)-domains, Fig. 4 highlights in black a selection of the most compact six-hairpin glycosidase (α / α)-toroids, a common fold that dominates the all-α scaling regime over the span of large n-values (*vide infra*).

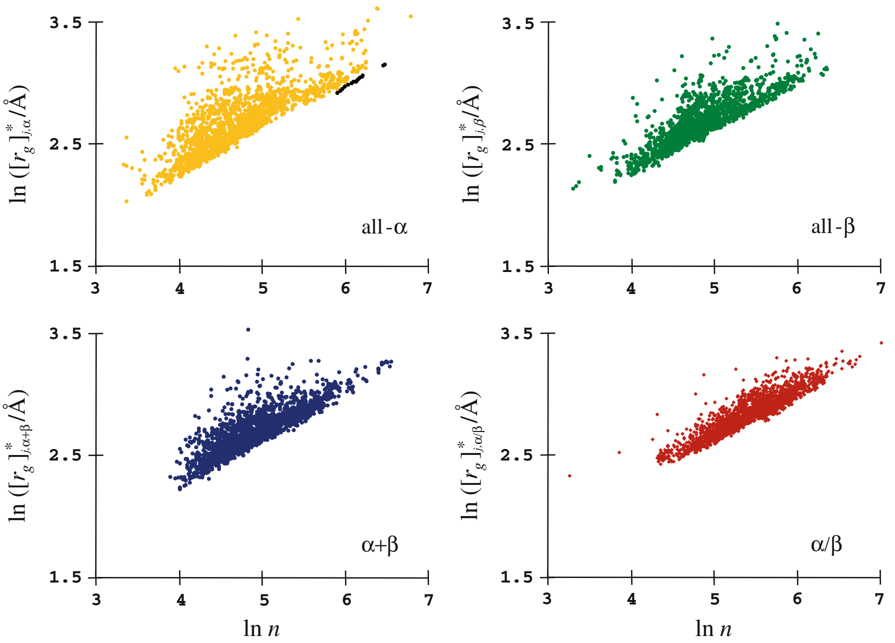

Fig. 4 Four distinct distributions of radius of gyration for the ensembles of non-redundant domains in the major folding classes. The distribution of (α / β)-folds can be distinguished from the others by its low dispersion. The points highlighted in *black* in the (all-α)-folds correspond to the α / α-toroidal (or α / α-barrel) common fold. (See text for the number of data points in each diagram)

Figure 5 shows the results for the $\{[r_g]^*_{j,FC}, n_j\}$-pairs extracted from the diagrams in Fig. 4. The two-colour coding indicates the distinct scaling behaviour for “short” and “long,” determined with the criterion of optimal squared correlation coefficients (*sqcc*) used in Sect. 3. The resulting $n^*_{1,2}$-transitions points are given by arrows in the diagrams of Fig. 5.

Table 1 shows the estimates for the $\bar{\nu}^{(short)}_{FC}$- and $\bar{\nu}^{(long)}_{FC}$-exponents for the families of FC-domains; these two values characterize the size-scaling behaviour for $n < n^*_{1,2}$ (“short-chain” regime) and $n \geq n^*_{1,2}$ (“long-chain” regime), respectively. For completeness, Table 1 gives also the global $\{(\bar{\nu})^{(global)}_{FC}\}$-exponents for the linear regression over the entire range of change lengths, as well as the number of points and range of n-values used in the least-square fittings. Outliers for very short and very long chains have been eliminated in some cases (principally for the (α / β)-fold); structures are considered outliers when their inclusion in the sequences 7-point regressions breaks the pattern of large square correlation coefficients, as discussed in Sect. 3, or when they arise from bins with very low numbers of proteins.

Several important observations can be made from the results collected in Table 1:

(i) All folding classes exhibit two distinct regimes of $(\bar{\nu})_{FC}$-values (cf. Eq. 4), although the differences appear less well marked in the case of (all-α)-folds.

(ii) The short-chain regime for (α / β)-folds is statistically identical to that of collapsed polymers (i.e., maximally-compact structures with $(\bar{\nu})_{CP} = 1/3$). In

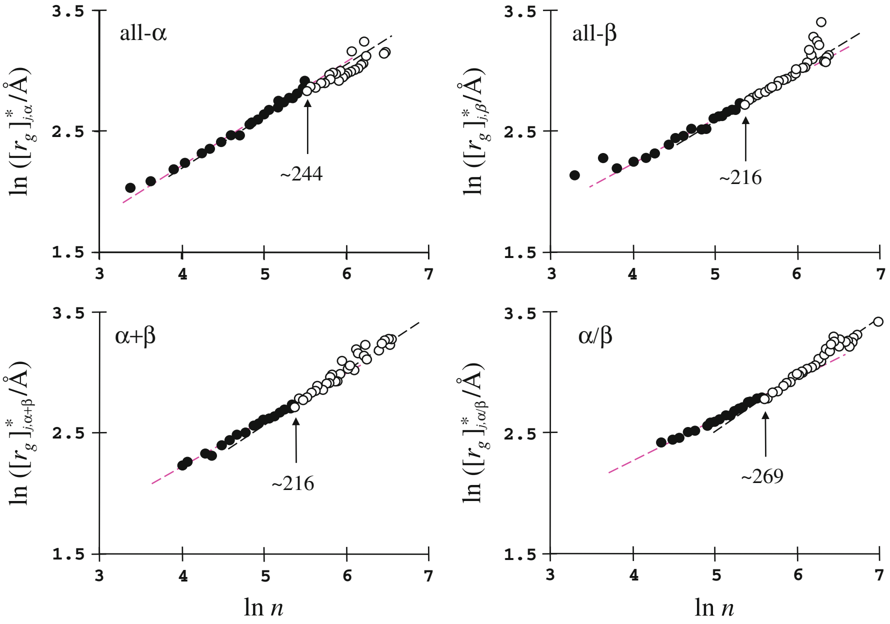

Fig. 5 Size scaling behaviour for the most compact domains extracted from the distributions in Fig. 4 (using $\Delta n = 10$ binning). The division into short- and long-chain scaling regimes was derived using the optimal-correlation criterion illustrated in Fig. 3. The *dashed lines* correspond to the least-square fittings within each regime

Table 1 Size-scaling exponents for the most compact domains in the four main folding classes (cf. Eq. 4)

FC	$\bar{\nu}_{FC}^{(short)}$	Range ($short$)	$\bar{\nu}_{FC}^{(long)}$	Range ($long$)[a]	$(\bar{\nu})_{FC}^{(global)}$	Points
all-α*	0.426 ± 0.018	[37, 235]	0.431 ± 0.046	[244, 510]	0.404 ± 0.011	45
all-β	0.362 ± 0.018	[45, 202]	0.416 ± 0.027	[216, 452]	0.372 ± 0.008	41
$\alpha + \beta$	0.382 ± 0.021	[55, 207]	0.449 ± 0.084	[216, 370]	0.392 ± 0.014	33
α/β	0.328 ± 0.020	[77, 260]	0.465 ± 0.029	[269, 534]	0.363 ± 0.013	45

The "short-chain" and "long-chain" regimes have been determined with the optimal-correlation approach illustrated in Fig. 3. The table gives the range of chain lengths used for the linear correlations; a $\Delta n = 10$ bin size was used to select representative domains with smallest radii of gyration $[r_g]^*$ in each bin interval. The exponent $(\bar{\nu})_{FC}^{(global)}$ characterizes the fitting over the complete ensemble of selected domains; the last column gives the total number of points used in estimating the $(\bar{\nu})_{FC}^{(global)}$-exponent. Error bars correspond to 95% confidence intervals

* The data for the all-α domains excludes the α / α toroids that appear in the long-chain regime (see text)

[a] The underlined entry represents the chain length value for the $n^*_{1,2}$-transition between the short- and long-chain regimes (see Fig. 5)

contrast, the short-chain regimes for (all-β)- and $(\alpha + \beta)$-folds exhibit slightly larger exponents, comparable to the global exponents for all domains, i.e., $\bar{\nu} = 0.388 \pm 0.009$ (cf. Sect. 3). The exponent for the (all-α)-domains proves to be slightly larger than this value (see (iv) below).

(iii) The long-chain domains appear in a less-compact regime for all folding classes, although the change in $\bar{\nu}$-value is small in (all-α)-domains. Within the 95%-confidence intervals, the $\bar{\nu}_{FC}^{(long)}$-exponents for $FC = \alpha + \beta$ and $\alpha\,/\,\beta$ are comparable to that of random polymers (i.e., $(\bar{\nu})_{RP} = 1/2$), while the corresponding value for long-chain all-β domains is slightly lower than these two. These values exclude data for $n > 534$, which show a systematic deviation from linear correlations according to the criteria discussed in Sect. 3.

(iv) As shown in Fig. 5, the long-chain (all-α)-domains exhibit two distinct populations, one of them being the compact (α / α)-toroids highlighted in Fig. 4 (black circles). The latter structures comprise eleven domains with a scaling exponent $\bar{\nu} = 0.432 \pm 0.042$ for $363 \leq n \leq 488$. If we exclude the (α / α)-toroids to avoid any structural bias, we find twenty-three distinct structures which produce $\bar{\nu}_{\alpha}^{(long)} = 0.431 \pm 0.046$ in the range $244 \leq n \leq 510$ (cf. Table 1). The latter two exponents are statistically indistinguishable, and comparable with $\bar{\nu}_{\alpha}^{(short)} = 0.426 \pm 0.018$. However, it should be noted that, if all structures were included, the size-scaling exponent would be $\bar{\nu}_{\alpha}^{(long)} = 0.326 \pm 0.078$. This uncharacteristically low value arises from the fact that the (α / α)-toroid population begins at $n = 326$, while the non-toroids start at $n = 248$; this mismatch skews the results in the long-chain region when both protein sets are considered together. Evaluated in isolation, each separate population produces $\bar{\nu} = 0.43 \pm 0.04$, although they can still be distinguished by their distinct ℓ_{FC}-values in Eq. (4).

(v) Although the $(\bar{\nu})_{FC}^{(global)}$-exponents are less sensitive to folding class, it is clear that $(\bar{\nu})_{\beta}^{(global)}$ and $(\bar{\nu})_{\alpha/\beta}^{(global)}$ are the smallest. The $(\bar{\nu})_{FC}^{(global)}$-exponents for $FC = \beta, \alpha{+}\beta$, and α/β are averages of the short-chain and long-chain regimes. In contrast, the scaling behaviour for all-α domains is the least sensitive to chain length (that is, $\bar{\nu}_{\alpha}^{(long)} \approx \bar{\nu}_{\alpha}^{(short)}$), in addition to being the *least* compact of all the FC-families (at least, with respect to the compactness level of collapsed polymers).

In summary, we observe that all folding classes show a distinct two-regime scaling behaviour that depends on chain length. The actual values for the scaling exponents vary however with the folding features, where the short-chain (α / β)-domains are the most compact and the (all-α)-domains are the least. It is likely that these differences in spatial organization relate to intrinsic properties of secondary-structural elements: [26–28]

(a) α-helices are approximately ten amino-acid long, relatively rigid objects; they adopt a restricted number of relative orientations, favouring the formation of elongated bundle-like objects with $\bar{\nu} > 1/3$.

(b) On average, β-strands are five amino-acid long objects; their smaller size permits a larger variation in relative orientations, leading to the formation of curved β-sheets that ensure spheroidal compact structure with $\bar{\nu} \approx 1/3$. Considering that the (α / β)-domains alternate the α- and β-contents, instead of segregating them as in the (α + β)-domains, it is then reasonable that the former adopt more compact structures than the latter, thus $\bar{\nu}_{\alpha\,/\,\beta}^{(short)} < \bar{\nu}_{\alpha} + \beta^{(short)}$.

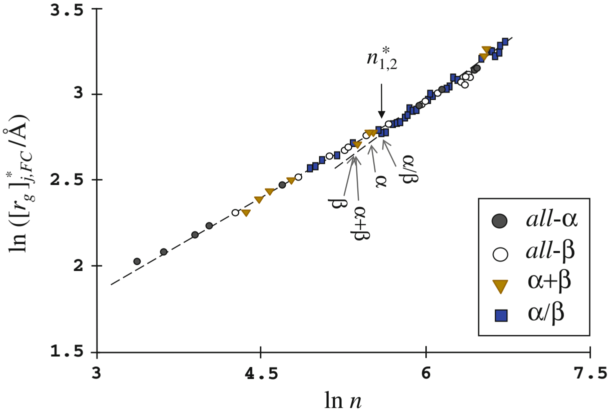

Fig. 6 Breakdown of the most compact domains in Fig. 1 according to folding class. The *thin arrows* indicate the corresponding $n^*_{1,2}$-transition lengths for folding classes (cf. Fig. 5). The *thick arrow* on top indicates the global $n^*_{1,2}$-transition length (i.e., without reference to folding class). The *dashed lines* correspond to the linear regressions in Fig. 2. Note that, although the (α / β)-folds dominate the longer chains, all foldings classes are represented throughout the ensemble

(c) The fact that all $\bar{\nu}^{(long)}_{FC}$-exponents are larger than their $\bar{\nu}^{(short)}_{FC}$ counterparts suggests that longer-chain domains have a lower monomer density, regardless of their folding features. This distinct behaviour matches experimental evidence that shows that long-chain and short-chain proteins fold by different mechanisms: while most small proteins fold cooperatively in a two-state process resembling qualitatively a *collapsed transition*, long-chain proteins fold via a three-state process involving partly-folded, *less compact intermediates* [18,22,26,27].

Figure 6 complements our analysis by showing the breakdown of the most-compact domains in Fig. 2 (black squares) in terms of the four major folding classes. For clarity, Fig. 6 also indicates the location of the $n^*_{1,2}$-transition points between the short- and long-chain regimes for the *FC*-families (cf. Table 1). This diagram indicates clearly that, while the line of most-compact domains includes representatives of all folds, the (α / β)-folds represent a majority (57%) and their distribution is not equal in terms of chain length.

A detailed breakdown shows:

(i) In the range of short chains, namely $37 \leq n \leq 260$, all folding classes contribute a similar amount, with 21% all-α, 25% all-β, 29% (α + β), and 25% (α / β).

(ii) In the region of long chains (i.e., $269 \leq n \leq 543$, excluding the low-sampling areas $n > 543$), we find a strong bias for (α / β)-folds (72%), while 10% are all-α, 14% all-β, and only 3% (α + β)-domains. The fact that the vast majority of most-compact long-chain domains belong to the (α / β)-family is consistent with our previous observation that the latter fold has the smallest $\bar{\nu}^{(long)}_{FC}$-exponent (cf. Table 1).

5 Deviation from the global scaling behaviour in particular protein lineages

Figures 2 and 6 illustrate the scaling behaviour for structural domains with the smallest r_g-values within the four major folding classes. Folding class, however, represents only a top level of domain organization. In this section we want to address briefly the following question: How much do individual protein families contribute to the most-compact scaling regime? Do the scaling behaviours described in Sect. 4 extend to other lower levels of domain organization, e.g., the protein lineages within each of the major *FC*-families?

Within each structural class, we find the "common folds," i.e., distinct motifs that share a similar secondary-structural content and overall global spatial arrangement [14–16,29,30]. In turn, common folds are associated with different protein lineages. In this section, we explore whether the scaling behaviour in Sect. 3 and 4 extends to lineages of common folds. In other words, suppose that the most compact domain in a given lineage belongs to a family of structures satisfying a scaling law as in Eq. (2). Then, do all other domains in the same protein lineage share the same $\bar{\nu}$-exponent? In order to address this issue, we have considered examples of lineages that contain at least one of the most compact domains in Figs. 2 and 6.

Figure 7 displays typical examples for families of common folds in each of the four major folding classes. The following observations can be made:

(i) The protein 1c9b (chain Q) provides one of the domains included in the scaling law for the most compact structures (dashed line). This unit belongs to a family of sixteen all-α cyclins highlighted in the top left diagram. Despite the dispersion, it is evident that the cyclins follow a different size-scaling law with a larger effective exponent $\bar{\nu} = 0.52 \pm 0.14$, consistent with that for polymers swollen in a good solvent. The other highlighted compact domains, indicated with black circles, belong to the family of (α / α)-toroids. This lineage comprises forty-two structures, and it includes some of the most compact long-chain (all-α)-domains, among them the six-hairpin glycosidase (protein 1ks8, chain A, in Fig. 7). After eliminating the first protein in this set as an outlier, the (α / α)-toroids are characterized by an effective scaling exponent $\bar{\nu} = 0.29 \pm 0.06$, corresponding to thirty-seven domains in the range $271 \leq n \leq 642$. As discussed in Sect. 4, the *most compact* (α / α)-toroids (dominated by six- and seven-hairpin glycosidases) are characterized by $\bar{\nu} \approx 0.43$; the entire lineage, however, follows a collapsed-polymer regime ($\bar{\nu} \approx 1/3$), in agreement with previous findings in other common folds [31].

(ii) Two lineages of all-β proteins appear in the top right diagram, i.e., the 7-bladed β-propellers and the β-trefoils. The two highlighted structures, 1knℓ for the β-propellers and protein 1a12 for the β-trefoils are among the globally most compact domains. It is clear that the rest of the proteins in these families move away from the most-compact behaviour (dashed line). We estimate similar effective scaling exponents: $\bar{\nu} = 0.54 \pm 0.07$ for the β-trefoils and $\bar{\nu} = 0.45 \pm 0.15$ for the 7-bladed β-propellers, which qualitatively border Θ-state behaviour.

(iii) The zincin-like proteins are a family of (α + β)-folds which include the highlighted protein 1s4b among those that determine the scaling law for the

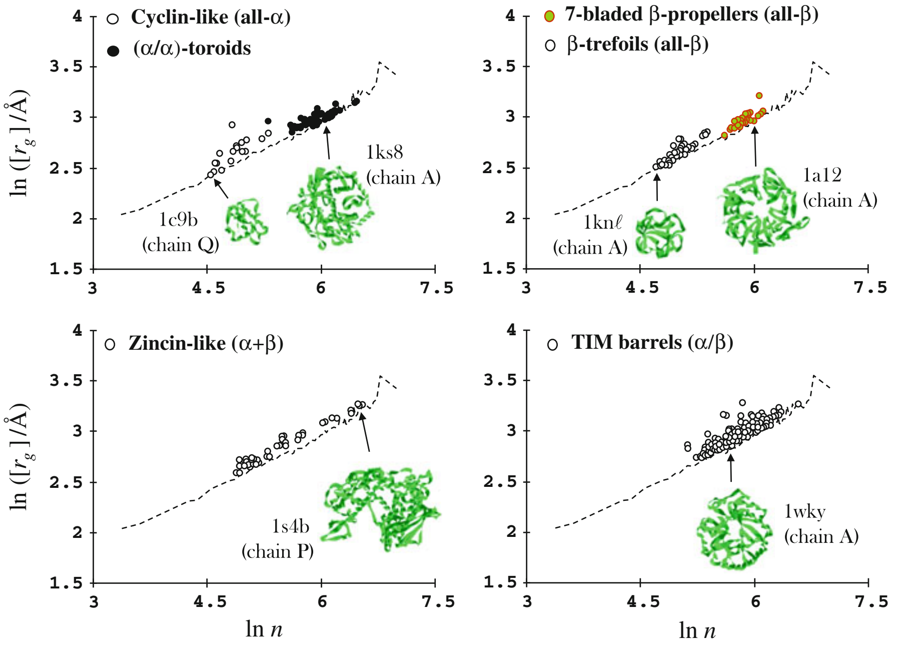

Fig. 7 Size scaling behaviour for selected protein lineages in the four major folding classes. The *arrows* indicate structures that belong to the regime for the most-compact domains (*dashed line*). The cyclins, β-propellers, and β-trefoils deviate from the compact-domain regime with approximate exponents $\bar{\nu} \approx 0.54$ and $\bar{\nu} \approx 0.45$, respectively, while the highlighted (α + β)- and (α / β)-folds also deviate but maintain a similar exponent (i.e., $\bar{\nu} = 0.4$). The (α / α)-toroid lineage (*black circles*) includes the most compact structures *highlighted* in *black* in the *top-left panel* in Fig. 4; this common fold approaches the collapsed-polymer regime with $\bar{\nu} \approx 0.3$ (see text)

most-compact domains (dashed line). In contrast to (i) and (ii), the fifty-two structures in this lineage match the dashed-baseline behaviour despite their larger radii of gyration, leading to an estimated $\bar{\nu} \approx 0.40 \pm 0.02$.

(iv) The TIM-barrels in the lower right diagram includes many of the most-compact (α / β)-domains (e.g., the highlighted protein 1wky). Despite its systematically larger r_g-values, we find that the TIM-barrel lineage follows the scaling behaviour of the most compact domains: the 357 structures in this set produce $\bar{\nu} \approx 0.38 \pm 0.02$. This situation is comparable to that of zincin-like proteins in (iii).

A similar trend can be observed in Fig. 8, which displays the molecular size regime for the α–α superhelices, a common fold within the all-α folding class that spans the entire range of n values considered here [14–16]. Five representative cases are highlighted on the right-hand side, corresponding to the black circles in the main diagram. Two of these structures are among the most compact domains in Fig. 6, namely, the globular and spheroidal proteins 1c9ℓ and 1ukℓ (denoted as A and E). In contrast, helices are arranged in an elongated fashion in structures B, C, and D, and thus they are distinctly less compact. A linear fitting over the ensemble of seventy-eight α–α

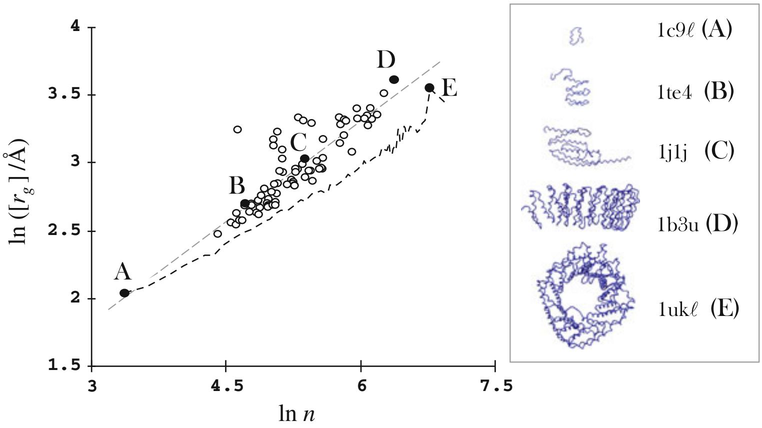

Fig. 8 The α–α superhelices, a common fold within the all-α folding class, demonstrates a systematic deviation from the most-compact scaling behaviour. The structures in black circles (A–E) are highlighted on the *right-hand* side. Proteins A and E belong to the most-compact domains (*dashed line*); the other structures are less spheroidal and are far removed from this line

superhelices in Fig. 8 gives:

$$\ln r_g = (0.498 \pm 0.061) \ln n + (0.37 \pm 0.28), \tag{7}$$

with 95% confidence intervals over a span of $29 \leq n \leq 876$. In summary, although protein 1c9ℓ belongs to the compact domains with effective scaling exponent $\bar{\nu} = 0.40 \pm 0.01$ (cf. Sect. 3), the scaling law restricted to the α–α superhelices gives an exponent $\bar{\nu} = 0.50 \pm 0.06$, which is in the range of polymers in the Θ-condition. The contrast between these two exponents is well illustrated by proteins 1b3u (point D) and 1ukℓ (point E) in Fig. 8: while the former's non-spheroidal form places it clearly in line with $\bar{\nu} \approx 0.5$ slope, the latter's globular shape makes it an outlier closer to the line with $\bar{\nu} \approx 0.4$ slope.

The results in Figs. 7 and 8 confirm that lineages of common folds can also exhibit size scaling behaviour [31]. However, having one of their members among the most compact domains does not imply that the entire family will follow the same behaviour. Whether or not a group of related proteins can be characterized by the scaling exponent $\bar{\nu} \approx 0.4$ depends entirely on its average spatial organization and the globularity of the folding motif.

6 Conclusions

In this work, we have shown the existence of a scaling behaviour relation between the radius of gyration and the chain length in subgroups of isolated domains. Although protein domains exhibit a wide distribution of molecular size, the domains with the smallest radii for a given length exhibit well-defined scaling laws.

We find two distinct scaling laws for the most compact domains, characterized by two exponents: (i) short-chain domains belong to a scaling regime for structures slightly *less* compact than collapsed polymers, $\bar{\nu}^{(short)} = 0.37 \pm 0.01 > \bar{\nu}_{CP} = 1/3$; (ii) the long-chain regime shows a level of compactness slightly *above* that for random polymers, $\bar{\nu}^{(long)} = 0.45 \pm 0.02 < \bar{\nu}_{RW} = 1/2$. Recent work indicates that these differences do not depend on whether the domains are extracted from simple or complex (i.e., multi-domain) proteins [32]. On the other hand, we have shown here that these different scaling behaviours depend on folding class.

When we consider the set of individual domains with the minimal r_g-values in each Δn-bin (i.e., without reference to folding class), the transition between these two scaling regimes can be estimated at $n \approx 269$ (within the certainty of a $\Delta n = 10$ binning). When the type of fold is taken into account, our results show again the occurrence of two scaling laws within each *FC*-family, with the (all-β)- and (α / β)-folds providing the most compact behaviours. In the case of short chains, the (α / β)-domains approach the $\bar{\nu}_{CP} = 1/3$ value for most compact polymers. Even in the cases where we find similar $\bar{\nu}$-values, the role of folding class can be recognized by the distinct ℓ_{FC}-parameters (cf. Eq. 4) and a shift in $n^*_{1,2}$-transition chain lengths.

In regard to the latter $n^*_{1,2}$-transitions, we have shown that the global value $n^*_{1,2} \approx 269$ is an intrinsic property of individual domains, and not associated with the transition from single- to multi-domain proteins, as previously speculated [5–7]. From the results in Sect. 4, it is clear that the global $n^*_{1,2}$ -value is determined by a single structural class, i.e., the most compact (α / β)-domains.

Finally, we have also shown that scaling behaviour can also be found within families of proteins that share a common fold or biological function. In these cases, however, the power laws observed are often characterized by larger $\bar{\nu}$-exponents, corresponding to less globular and compact domains. Examples of linear scaling ($\nu \approx 1$) have also been noted in the literature for selected protein families [31].

Our work provides insight into the spatial organization of the most compact domains. The fact that size behaviour can be captured by a single $\bar{\nu}$-exponent suggests a common principle underlying domain organization. Given that the $\bar{\nu}$-exponent is determined by the nature of the monomer-monomer interaction [1], we can conjecture that the structure of compact proteins arises from the same dominant "forces," despite differences in fold, function, or lineage among proteins. In other words, proteins must share the same balance of monomer attraction and repulsion, with the addition of spatial restrictions imposed by the presence of secondary structure.

On the other hand, the variation of $\bar{\nu}$-exponents with chain length is consistent with the experimental observation that small and large globular proteins fold differently (according to two- and three-state mechanisms, respectively) [26,27]. Our results would indicate that three-state mechanisms (i.e., those associated with long-chain protein domains) lead to the formation of less compact native structures.

Acknowledgments This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell UP, Ithaca, 1985)
2. J.-C. LeGuillou, J. Zinn-Justin, Phys. Rev. B **21**, 3976 (1980)

3. J.-C. LeGuillou, J. Zinn-Justin, J. Phys. (France) **50**, 1365 (1989)
4. Y. Kantor, M. Kardar, Europhys. Lett. **14**, 421 (1991)
5. G.A. Arteca, Phys. Rev. E **49**, 2417 (1994)
6. G.A. Arteca, Phys. Rev. E **51**, 2600 (1995)
7. G.A. Arteca, Phys. Rev. E **54**, 3044 (1996)
8. C.P. Ponting, R.R. Russell, Annu. Rev. Biophys. Biomol. Struct **31**, 45 (2002)
9. S.J. Wodak, J. Janin, Biochemistry **20**, 6544 (1981)
10. M.B. Swindells, Protein Sci. **4**, 103 (1995)
11. M.H. Zehfus, Protein Sci. **6**, 1210 (1997)
12. C.J. Tsai, R. Nussinov, Protein Sci. **6**, 24 (1997)
13. J. Liu, B. Rost, Proteins **55**, 678 (2004)
14. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
15. A. Andreeva, D. Howorth, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **32**, D226 (2004)
16. A. Andreeva, D. Howorth, J.M. Chandonia, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **36**, D419 (2008)
17. A. Heger, L. Holm, J. Mol. Biol. **328**, 749 (2003)
18. M.O. Lindberg, M. Oliveberg, Curr. Opin. Struct. Biol. **17**, 21 (2007)
19. S.W. Englander, L. Mayne, M.M. Krishna, Q. Rev. Biophys. **40**, 287 (2007)
20. C.F. Wright, S.A. Teichmann, J. Clarke, C.M. Dobson, Nature **438**, 878 (2005)
21. M.Y. Shen, F.P. Davis, A. Sali, Chem. Phys. Lett. **405**, 224 (2005)
22. D. Baker, Nature **405**, 39 (2000)
23. W.J. Netzer, F.U. Hartl, Nature **388**, 343 (1997)
24. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, J. Mol. Biol. **112**, 535 (1977)
25. H.M. Berman, K. Henrick, H. Nakamura, Nat. Struct. Biol. **10**, 980 (2003)
26. A.M. Lesk, *Introduction to Protein Architecture* (Oxford UP, Oxford, 2001)
27. G.A. Petsko, D. Ringe, *Protein Structure and Function* (New Science Press, London, 2004)
28. C. Chothia, M. Levitt, D. Richardson, Proc. Natl. Acad. Sci. USA **74**, 4130 (1977)
29. C.A. Orengo, F. Pearl, J.M. Thornton, Meth. Biochem. Anal. **44**, 249 (2002)
30. F. Pearl, C. Bennett, C.A. Orengo, in *Dictionary of Bioinformatics and Computational Biology*, ed. by J.M. Hancock, M.J. Zvelebil (Wiley, Chichester, 2004)
31. P. Rogerson, G.A. Arteca, J. Math. Chem. **49**, 1463 (2011)
32. P. Rogerson, G.A. Arteca, (submitted)