J Math Chem (2012) 50:1901–1919
DOI 10.1007/s10910-012-0010-1

ORIGINAL PAPER

Scaling laws in simple and complex proteins: size scaling effects associated with domain number and folding class

Parker Rogerson · Gustavo A. Arteca

Received: 13 February 2012 / Accepted: 30 March 2012 / Published online: 21 April 2012

Abstract The native states of the most compact globular proteins have been described as being in the so-called "collapsed-polymer regime," characterized by the scaling law $R_g \sim n^{\nu}$, where R_g is radius of gyration, n is the number of residues, and $\nu \approx 1/3$. However, the diversity of folds and the plasticity of native states suggest that this law may not be universal. In this work, we study the scaling regimes of: (*i*) one to four-domain protein chains, and (*ii*) their constituent domains, in terms of the four major folding classes. In the case of complete chains, we show that size scaling is influenced by the number of domains. For the set of domains belonging to the all-α, all-β, α/β, and $\alpha+\beta$ folding classes, we find that size-scaling exponents vary between $0.3 \leq \nu \leq 0.4$. Interestingly, even domains in the same folding class show scaling regimes that are sensitive to domain provenance, *i.e.*, the number of domains present in the original intact chain. We demonstrate that the level of compactness, as measured by monomer density, decreases when domains originate from increasingly complex proteins.

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

1 Introduction

The relation between the configurationally-averaged mean radius of gyration, $< R_g^2 >^{1/2}$, and the number of monomers, n, is well understood in simple homopolymer chains [1–3]. In this case, the value of the size-scaling exponent (ν) in the

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

asymptotic law $< R_g^2 >^{1/2} \sim n^{\nu}$ depends on the nature of the dominant interactions and the degree of polymer compactness.

In contrast, the relation between chain length and size presents a major challenge in proteins (*i.e.*, nonrandom heteropolymers). Here we find dominant native states that naturally select for optimal structural and biophysical functions. Therefore, instead of the statistically-averaged size $< R_g^2 >^{1/2}$ typical of diluted polymers, protein native states exhibit a unique, sharp value of the radius of gyration, R_g. Nevertheless, some native states (*e.g.*, small globular proteins) are known to follow the effective scaling behaviour of "collapsed polymers" [4–6]. It is not known whether this extends to specific families of proteins; the potential occurrence of such behaviour would hint at commonalities in compactness, native state selectivity, and the role of domains in the folding mechanism [7–10]. In this work, we combine experimental three-dimensional structures and data base fold classification to explore size scaling in protein domains with well-defined folding features.

Our goal is to study whether single protein domains (originating from single- and multi-domain chains), with folding topologies defined by the Structural Classification of Proteins (SCOP) data base (version 1.73) [11–14], have distinct scaling behaviours with respect to the global native state. For simplicity, a protein chain or domain is represented by its α-carbon trace, accordingly the radius of gyration is calculated as:

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

where $\mathbf{r}_i$ is the position vector of the ith α-carbon and $\mathbf{r}_0$ the centroid of the α-carbon chain, measured from the same origin; n is the total number of residues. The $\{\mathbf{r}_i\}$-coordinates are extracted from the Protein Data Bank (PDB) [15,16]. This simplified R_g-value quantifies global compactness in a protein fold, while omitting primary sequence and side chain conformation.

In free (unconfined) three-dimensional structures, the power law for R_g, averaged over all accessible configurations, follows a well-known dependence [1]:

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

where the size-scaling ν-exponent depends only on dominant interactions and polymer embedding (here, three-dimensional space). For simplicity, we indicate $< R_g^2 >^{1/2} \equiv R_g$.

Four exponents are well characterized in polymer theory and relevant to this work:

(*i*) Collapsed polymers (CP), *i.e.*, those dominated by attractive interactions and indicative of a poor solvent, scale with the compactness of a spherical volume, that is, $\nu_{CP} = 1/3$ [1].

(*ii*) Random-walk (RW) polymers, *i.e.*, those found at the Θ-temperature or Θ-solvent where repulsion and attraction are balanced, are characterized by $\nu_{RW} = 1/2$ [1].

(*iii*) Self-avoiding-walk (SAW) polymers, *i.e.*, those swollen in a good solvent or dominated by repulsions, exhibit the scaling exponent $\nu_{SAW} = 0.588 \pm 0.002$ [2,3].

(*iv*) In addition, we expect the scaling exponent $\nu_{rod} = 1$ in the case of rigid polymers; these structures are dominated by a single rod-like configuration and thus are nonrandom.

The pre-exponential length "L" in Eq. (1) depends on specific polymer properties, *e.g.*, type of monomer, composition, spatial correlations, and flexibility. From now on, we will characterize the "compactness regime" of a polymer (that is, collapsed, random-walk, or swollen) by the value of the size-scaling exponent. Polymers in the same compactness regime will appear as parallel lines in a logarithmic plot R_g vs n. This notion should not be confused with that of the "absolute compactness" of a chain, as specified by the actual R_g-value. Note that some polymers in the "swollen-compactness" (with $\nu \approx 0.6$) regime can be smaller in mean size than sufficiently-inflated spheroidal chains ($\nu \approx 0.333$).

Assessing true asymptotic scaling, as in Eq. (1), is not possible in proteins because single chains rarely surpass 1,000 residues. In practice, however, pseudo-scaling behaviour with an effective ν_{eff}-exponent can be recognized in some globular proteins over the range of chain lengths $50 < n < 1,000$ [4–6]. For example, the most compact short globular proteins, historically seen as those with $n < 300$, approach the $\nu_{CP} = 1/3$ law, while longer chains at the same level of compactness generate a larger ν-value [4,5]. Given that longer proteins often comprise several folding domains, it has been conjectured that this deviation may be associated with a distinct packing for multi-domain proteins [4]. On the other hand, we have recently shown that size-scaling exponents with values between ν_{CP} and ν_{CP} can be found within particular lineages of related single-domain proteins [17].

Previous studies of protein size pseudo-scaling have involved a small group of proteins with wildly different folds, functions, primary sequences, and number of domains. In this work, we expand these analyses by examining the scaling behaviour within well-defined folding classes, and evaluating any effects associated with the number of domains present.

Our goal is threefold: (*i*) establish whether distinct scaling differences can be recognized in full protein chains consisting of one to four domains; (*ii*) analyze the mean behaviour of the separated domains contained in those proteins; (*iii*) study how domain size scaling depends on the folding lineage, *i.e.*, whether they belong to the all-α, all-β, α/β, and $\alpha + \beta$ folding classes.

Using this information, we aim to address the following two fundamental questions:

(*a*) Do domains isolated from multi-domain proteins behave the same as a single-domain protein? In other words, can we distinguish between domains in terms of size-scaling depending on whether they have other neighbouring domains in the original protein?

(*b*) Can we distinguish domains with distinct folding features by their compactness regimes? That is, despite circumstantial differences in protein size, do different folding classes express a fundamentally different packing of amino acids?

By addressing these questions, this work seeks to contribute to a better understanding of compactness in protein folds, as well as shed light on the elusive notion of "domain." The very definition of what constitutes a domain continues to be a matter of debate,[13,14,18–27] be it in terms of their structural features, biochemical function, evolution, or simply the location of boundaries.

The work is organized as follows. In Sect. 2, we explain the fold classification of the SCOP data base, and the sampling criteria used to extract unique proteins and protein domains from this archive. Section 3 deals with the methodology applied to determine effective scaling exponents for intact proteins, and the procedure used to select protein domains that belong to different folding classes. Section 4 presents our results for multi-domain proteins, while Sect. 5 examines the details of the individual domains within each of those multi-domain proteins. We close with a summary of conclusions on the compactness regimes observed in protein domains, focusing on the effects due to folding class and number of neighbouring domains.

2 Classification of fold topology and sampling criteria

The SCOP data base classifies protein domains into hierarchical fold lineages based on structural relationships (*e.g.*, having similar "fold topologies") [11–14]; these units include known experimental structures derived from either single-domain proteins or individual components in multi-domain proteins (*i.e.*, "isolated" or "cut" domains). It should be noted that the SCOP classification scheme, though of widespread use, relies on subjective criteria [14]; every entry is visually inspected by the SCOP curators and classified according to a consensus of domain family definitions.

The SCOP data base contains eleven *root nodes*; the first four of these are considered in this work, namely the all-α, all-β, $\alpha+\beta$, and α/β classes. These principal folding classes are well represented in numbers and diversity, thus permitting a proper statistical characterization.

Domains in the aforesaid four folding classes are characterized by the content of their secondary and supersecondary structure:

(*i*) all-α domains consists mainly of helical supersecondary elements (*e.g.*, α-helical bundles);
(*ii*) all-β domains comprise β-sheet motifs;
(*iii*) $\alpha+\beta$ domains include independent α-helices and (most commonly) antiparallel β-sheets;
(*iv*) α/β domains consist of "$\beta-\alpha-\beta$" structural units, most often involving parallel β-sheets.

Our objective is to determine if there is a difference in the packing arrangement of domains derived from single-domain proteins and domains located within multi-domain proteins, and to evaluate these differences in terms of the four major folding classes. To this end, we rely on a subclass of independent, nonreduntant domains extracted from the PDB, and organized by the SCOP data base, in order to study their size-scaling regimes.

The SCOP and PDB archives contain a enormous number of entries, and, at the same time, they are highly redundant. This redundancy may take the form of closely

related sequences, or multiple structural entries associated with distinct resolutions, experimental methodology, temperature, or type of ligand binding. In order to avoid biases in the scaling behaviour, it is necessary to curtail this multiplicity. Duplication is eliminated by using an appropriate set of criteria to ensure only one entry per domain type. The following protocol was used to filter out redundancy and create our data set:

(*a*) Only single chains are considered (that is, quaternary structure is omitted in our analysis). If multiple data are available for the same chain, we chose the structure with the best resolution. In the case of X-ray data, structures with resolution above 3.2 Å were omitted. Proteins with missing residues or poorly-resolved areas were also excluded from our set.
(*b*) Chains under 35 residues are omitted, as they tend to resemble unstructured polypeptides.
(*c*) Proteins with the same chain length and over 90 % sequence identity are represented by a single entry.
(*d*) Domains with $> 90\,\%$ sequence identity, yet differing in more than 15 amino acids in chain length, are considered distinct entries.
(*e*) Due to the fact that termini regions are often labile, we allow chains to differ by up to a 12-residue segment at *one* end of the protein.

Using these criteria, we retain only one entry within such an ensemble for domains and chains belonging to the same SCOP classification.

We started our analysis with a total of 85,686 protein domains in the four major folding classes. After reorganizing the structures in the SCOP data base with the above criteria, we ended up with an ensemble of 8,614 non-redundant individual domains with the following breakdown in terms of folding classes (*FC*): all-α (1,741 out of 14,824), all-β (2,527 out of 23,547), $\alpha + \beta$ (2,099 out of 21,499), and α/β (2,247 out of 25,816). Finally, we have reclassified these units according to *provenance*, *i.e.*, the number of domains in the original intact protein chain. In order to study size scaling, each of these subgroups is reclassified in turn according to their R_g-value by the binning processes explained in the next section.

Note that while the SCOP data base includes a "multi-domain protein" section, it does not link these complex proteins to their constituent domains. This mapping is a necessary step in our analysis, permitting us to study the effect of domain provenance on size scaling. Multi-domain single-chain analysis was carried out after subjecting this data set to the same redundancy elimination protocol used for single domains.

3 Mean-size scaling in full chains and individual domains

3.1 Scaling laws

Consider a generic single chain within the ensemble of structures selected in the previous section. (As discussed before, quaternary structure is omitted in our analysis.) Each chain is characterized by having n residues and t-domains ($t \leq 4$). Moreover, using the SCOP data base, each of those individual t-domains is classified within one of four folding classes (*FC*), using a label $FC = \alpha, \beta, \alpha + \beta, \alpha/\beta$ [11–14].

The radius of gyration of the *full chain* will be denoted by $R_g^{(t)}$, highlighting the fact that it contains t-domains. Once a Δn bin interval is selected, *e.g.*, $\Delta n = 10$, each chain will be classified as belonging to a given j-bin, containing a total of N_j structures. To highlight this characteristic, the chain size is labeled as $[R_g^{(t)}]_j$.

In order to study scaling behaviour, we select the protein chain with the smallest $R_g^{(t)}$-value in each bin, denoted:

$$\min_{\{N_j\}} \left[R_g^{(t)}\right]_j = \left[R_g^{(t)}\right]_j^*, \tag{3}$$

and then seek the best correlation defined by an effective scaling exponent $\nu^{(t)}$:

$$\left[R_g^{(t)}\right]_j^* \sim L^{(t)} n_j^{\nu^{(t)}}, \tag{4}$$

where n_j is the number of residues of the selected protein belonging to the j-bin. If we make no distinction between the number of domains, we have a global scaling law as in Eq. (2) for the ensemble of the smallest proteins selected after binning.

Each native state with global size $[R_g^{(t)}]_j$ is a full chain comprising t-individual domains. The molecular size of individual units *extracted* from the original full chains will be represented by $r_g^{(t)}$, while retaining $R_g^{(t)}$ as a full chain designation.

For a single-domain protein, there is no difference between the two values, *i.e.*, $R_g^{(1)} = r_g^{(1)}$. On the other hand, each $R_g^{(t)}$ gives rise to a set of radii, $\{r_{g,i}^{(t)}, i \leq t\}$ for $t > 1$. Since each $r_{g,i}^{(t)}$-value describes the size and compactness of a *single* domain derived from a multidomain protein, it is possible that some of these domains are smaller than the smallest $r_g^{(1)}$-value extracted exclusively from single-domain proteins with the same chain length.

The ensemble of all individual domains, extracted from all possible single- and multi-domain proteins, irrespective of folding class, can be binned and characterized by a scaling law. In analogy with Eq. (2), we write:

$$\left[r_g\right]_j^* \sim \ell\, n_j^{\bar{\nu}}, \tag{5}$$

where ℓ and $\bar{\nu}$ are single-domain parameters, while L and ν are reserved for full chains. Equation (5) forms a global reference (or "base line") to highlight any deviations from scaling behaviour associated with domain number and folding features.

We can then study the scaling behaviour with specific reference to number of domains in the original protein chain. To this end, we restrict our analysis to the subensembles of $\{r_g^{(t)}\}$-values (*i.e.*, single domains extracted exclusively from full chains with t-domains). In this case, the scaling law is independent of the type of fold, and denoted as follows:

$$\left[r_g^{(t)}\right]_j^* \sim \ell^{(t)} n_j^{\bar{\nu}^{(t)}}, \quad \text{where } t = 1, 2, 3, \text{ and } 4. \tag{6}$$

Any dependence on the folding features requires that we introduce the folding-class label.

Using the $\{r_g^{(t)}\}$-set of single chains, we study the scaling behaviour of domains belonging to each of the four folding classes. If we make no reference to domain index "t", and rather focus on folding class (FC), *i.e.*, omitting whether they were originally extracted from single- or multi-domain proteins, we write:

$$[r_g]^*_{j,FC} \sim \ell_{FC}\, n_j^{\bar{\nu}_{FC}}, \tag{7}$$

where $[r_g]^*_{j,FC}$ is the smallest r_g-value among all the proteins with the FC-fold in the j-bin.

When the scaling law omits the number of domains and reflects only the smallest r_g-values of a particular FC-fold, then we refer to Eq. (7) as the "base line" for all FC-domains. We compare this base line with the results corresponding to the domains extracted from proteins with a specific number of t-domains. This final scaling law illustrates the dependence of scaling on fold and number of original domains:

$$\left[r_g^{(t)}\right]^*_{j,FC} \sim \ell_{FC}^{(t)}\, n_j^{\bar{\nu}_{FC}^{(t)}}. \tag{8}$$

Using this approach, we have recently studied the size-scaling properties of isolated domains associated with short and long chains with variable folding characteristics [28]. In this case, we ignored the *provenance* of these domains, *i.e.*, domains generated from single- and multi-domain proteins were considered equivalent (cf. Eq. (5)) [28]. In the present work, we extend this approach in order to compare the scaling laws for intact proteins (Eq. (4)) and those for their isolated protein domains (Eqs. (6–8)).

3.2 Descriptors for deviation in scaling behaviour

As described previously, Eq. (7) serves as a reference (or "base line") to understand the effect of folding features on the scaling behaviour. On the other hand, Eq. (6) provides a reference to study effects associated with the number of domains in the original intact protein. To quantify the deviations from these reference lines, we monitor two types of descriptors.

3.2.1 Absolute displacement from the base lines

Two cases are considered inside this category:

(*i*) $s^{(t)}$, corresponding to the signed displacement with respect to the global behaviour of individual single domains (*i.e.*, the deviation between Eqs. (5) and (6)):

$$s^{(t)} = \frac{1}{N_{bin}} \sum_{j=1}^{N_{bin}} \left\{ \ln\left[r_g^{(t)}\right]^*_j - \ln\left[r_g\right]^*_j \right\}, \tag{9}$$

where N_{bin} is the number of bins where there are entries for *both* $[r_g^{(t)}]_j^*$ and $[r_g]_j^*$, *i.e.*, bins which lacked either one of these two values were excluded in Eq. (9) (and descriptors below). The descriptor $s^{(t)}$quantifies the deviation from the base line associated with the domain's provenance, *i.e.*, whether it came from a protein containing originally $t = 1, 2, 3,$ or 4 domains.

(*ii*) $s_{FC}^{(t)}$, measuring the signed displacement with respect to the global behaviour of single domains with *FC*-folds (*i.e.*, the deviation between Eqs. (7) and (8)):

$$s_{FC}^{(t)} = \frac{1}{N_{bin}} \sum_{j=1}^{N_{bin}} \left\{ \ln \left[r_g^{(t)} \right]_{j,FC}^* - \ln[r_g]_{j,FC}^* \right\}. \tag{10}$$

Note that the descriptors $s^{(t)}$ and $s_{FC}^{(t)}$ track the sign of the deviation. As we show in the next section, this is important because our results deviate systematically from the reference.

3.2.2 Relative displacement from the base lines

These descriptors complement the analysis in terms of $s^{(t)}$ and $s_{FC}^{(t)}$ above:

(*i*) $\rho^{(t)}$, corresponding to the relative displacement between Eqs. (5) and (6):

$$\rho^{(t)} = \frac{1}{N_{bin}} \sum_{j=1}^{N_{bin}} \frac{\ln \left[r_g^{(t)} \right]_j^*}{\ln[r_g]_j^*}. \tag{11}$$

(*ii*) $\rho_{FC}^{(t)}$, measuring the relative displacement between Eqs. (7) and (8):

$$\rho_{FC}^{(t)} = \frac{1}{N_{bin}} \sum_{j=1}^{N_{bin}} \frac{\ln \left[r_g^{(t)} \right]_{j,FC}^*}{\ln[r_g]_{j,FC}^*}. \tag{12}$$

3.3 Illustrative example of the notation for a multi-domain protein

Let us consider one example to illustrate the present notation. Figure 1 shows the α-carbon trace (chain A, $n = 135$) of the ε-subunit of the proton-translocating ATP synthase from *E. coli* (PDB code 1aqt, a dimeric protein). When using a binning interval $\Delta n = 10$, 1aqt is the seventh most-compact protein in the [131,140]-bin in our ensemble. Consequently, its corresponding radius of gyration is indicated as $[R_g^{(t)}]_{j=}[R_g^{(2)}]_7 = 15.87$ Å, denoting that it belongs to the seventh bin ($j = 7$) of two-domain proteins ($t = 2$). Since it is not the most compact in that bin, no asterisk is used.

As shown in Fig. 1, 1aqt comprises two contiguous domains. The first domain (right-hand side, an all-β fold) extends from amino acid 2 to 86; with a length $n = 85$,

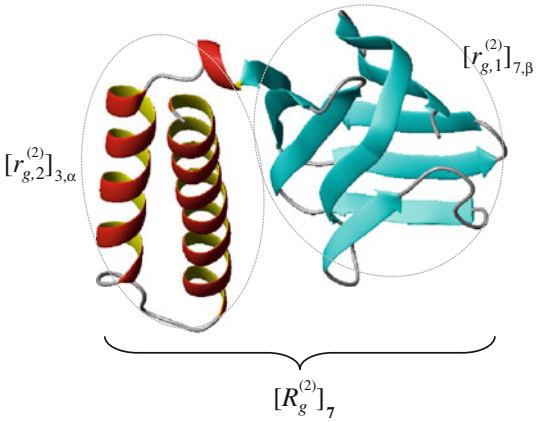

Fig. 1 ATP synthase from *E. coli*, a two-domain protein with $n = 135$ amino acids (PDB code 1aqt, chain A). The scheme illustrates the radii of gyration for the individual domains, and the notation used in this work. The radius of gyration for the entire chain is $[R_g^{(2)}]_7 = 15.87$ Å, whether the subindex "7" indicates that the protein belongs to the seventh sample bin (for $\Delta n = 10$). Domain 1 belongs to the seventh bin of all-β units ($n = 85$), with a local radius $[r_{g,1}^{(2)}]_{7,\beta} = 11.32$ Å. Domain 2 belongs to the all-α folding class ($n = 50$, third bin), with a local radius $[r_{g,2}^{(2)}]_{3,a} = 12.11$ Å

it is found in the [81,90]-bin in our ensemble of single domains. For this reason, its local radius of gyration is indicated as $[r_{g,i}^{(t)}]_{j,FC} \equiv [r_{g,1}^{(2)}]_{7,\beta} = 11.32$ Å, to denote that it is the first domain ($i = 1$) derived from a two-domain protein ($t = 2$) in the all-β folding class ($FC = \beta$), located in the seventh-bin ($j = 7$) of collected single domains. Since this structure is not the most compact among the sixty structures in the $j = 7$ bin, the r_g-symbol omits the asterisk. Note that the seventh bin of *isolated* single domains is different from the seventh bin of *complete* two-domain proteins. At $\Delta n = 10$, the former begins with proteins as short as $n = 20$. The latter case, restricted exclusively to two-domains proteins, involves fewer proteins and begins with chains with minimum length $n = 70$.

The second domain (left-hand side, an all-α fold) extends from amino acid 88 to 137; with a length $n = 50$, it is in the [41,50]-bin within the ensemble of single domains. The corresponding radius of gyration in Fig. 1 is denoted thus $[r_{g,i}^{(t)}]_{j,FC} \equiv [r_{g,2}^{(2)}]_{3,a} = 12.11$ Å, to highlight that it is the second domain ($i = 2$) extracted from a two-domain protein ($t = 2$) in the all-α folding class ($FC = \alpha$), located in the third-bin ($j = 3$) of single domains. Again, the omitted asterisk indicates that this domain is not the most compact among the fourteen structures in that bin.

In the next section, we apply this notation for discussing the mean size of complete multi-domain chains, as well as the scaling behaviour of their constituent domains broken down in terms of folding classes.

4 Results for the scaling behaviour of complete multi-domain chains

We first studied the molecular size characteristics of entire protein chains. When selecting the most compact structure within each chain, the corresponding scaling law is

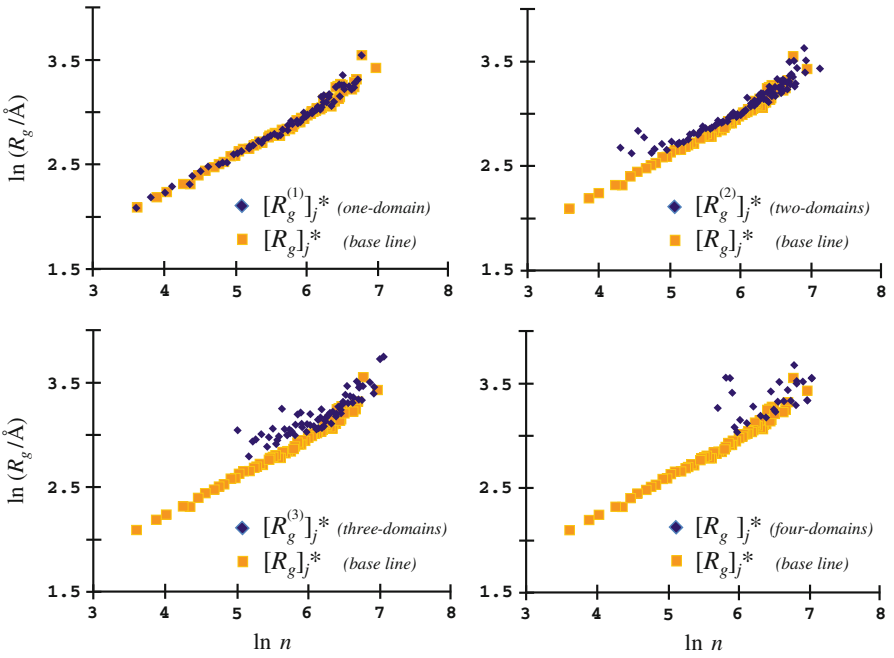

Fig. 2 Scaling behaviour for the most compact single chains, comprising t-domains ($t \leq 4$). Results are compared with a "base line" (*squares*) generated from the most compact units derived from either single- or multi-domain proteins. Single-domain proteins match this base line, but more complex proteins deviate systematically while preserving the same qualitative trend in scaling. (See text for more details)

represented by Eq. (4), with an effective size-scaling exponent $\nu^{(t)}$ associated with an intact protein chain with t-domains.

Figure 2 summarizes our findings with respect to the scaling behaviour of complete chains comprising a number of domains from $t = 1$ to $t = 4$. The light squares correspond to the smallest values of radius of gyration selected within each $\Delta n = 10$ bin, independent of the number of domains in the protein; these are the values that define the global reference "base line" (present in each of the four panels in Fig. 2). The dark rhomboids correspond to the minimal R_g-values associated with protein chains containing t-domains (*i.e.*, $[R_g^{(t)}]^*$). Perhaps not surprisingly, single-domain proteins provide the majority of the base-line points. The sample is sufficiently large to draw statistically significant conclusions for $t \leq 3$; in contrast, the number of four-domain proteins is rather limited, and thus produces a much larger dispersion.

The following observations can be made from the results in Fig. 2:

(*i*) There is unequivocal evidence of size-scaling behaviour for proteins containing t-domains, $t < 4$. Despite the large scattering, the $t = 4$ case is not inconsistent with this trend.

(*ii*) Short chains exhibit a similar effective scaling law, with exponents not far from the collapsed-polymer regime $\nu_{CP} = 1/3$. However, two differences are clear. First, single-domain proteins have a slightly larger exponent. Secondly,

an increase in the number of domains causes an *upward* displacement in the scaling line, that is, these complex proteins have *larger* absolute sizes for a given chain length. These two characteristic are captured in best defined linear regressions, corresponding to the shorter chains in each group:

$$\ln\left[R_g^{(1)}\right]^* \approx (0.357 \pm 0.006)\ln n + (0.81 \pm 0.03)\,,\ n \leq 260, \tag{13a}$$

$$\ln\left[R_g^{(2)}\right]^* \approx (0.294 \pm 0.013)\ln n + (1.22 \pm 0.07)\,,\ n \leq 348, \tag{13b}$$

$$\ln\left[R_g^{(3)}\right]^* \approx (0.305 \pm 0.009)\ln n + (1.22 \pm 0.05)\,,\ n \leq 433\ (\text{bin } \Delta n = 25), \tag{13c}$$

$$\ln\left[R_g^{(4)}\right]^* \approx (0.293 \pm 0.019)\ln n + (1.30 \pm 0.13)\,,\ n \leq 1073\ (\text{bin } \Delta n = 50), \tag{13d}$$

where the error bars correspond to standard errors. In other words, the $\nu^{(t)}$-values cannot be distinguished for $t \geq 2$; they are consistently lower than $\nu^{(1)}$, and the $\nu_{CP} = 1/3$ exponent is statistically bracketed between the single- and multi-domain exponents. In contrast, the deviation in linear intercept is a clear indication that the smallest multi-domain proteins provide a packing that is less compact, in absolute terms, than that provided by the smallest single-domain proteins with the same chain length. However, the fact that the $\nu^{(t)}$-exponents ($t \geq 2$) are closer to 1/3 than $\nu^{(1)}$ indicates that every domain added maintains an overall spheroidal packing in multi-unit proteins with the smallest possible size. (Note that our discussion deals with *effective* numerical ν-exponents; exact ν-exponents in infinite random polymers cannot be below 1/3 as they would imply a fractal dimension larger than three [29]).

(*iii*) Longer chains appear to deviate from the effective scaling in Eq. (13). Due to the smaller sample size, uncertainties are large but the trend is still well defined. The following estimates (with standard errors), made from the smallest R_g-values, describe the behaviour for longer chains: $\nu^{(1)} \approx 0.47 \pm 0.02$ ($269 \leq n \leq 551$), $\nu^{(2)} \approx 0.51 \pm 0.07$ ($390 \leq n \leq 535$), and $\nu^{(3)} \approx 0.50 \pm 0.06$ ($455 \leq n \leq 862$). Data is insufficient to decide whether a second scaling regime exists for $t = 4$.

(*iv*) These results can be compared with the scaling behaviour for the entire set of individual domains. If we disregard differences in domain provenance, units with the smallest r_g-values yield an effective *global* exponent $\nu = 0.391 \pm 0.006$ (Eq. (2)) for $29 \leq n \leq 876$ (with outliers removed). A breakdown into short- and long-chain domain scaling indicates a similar shift away from the collapse-polymer regime: $\nu = 0.368 \pm 0.004$ in short chains ($37 \leq n \leq 260$), $\nu = 0.45 \pm 0.01$ for longer chains ($269 \leq n \leq 543$), and finally $\nu = 0.8 \pm 0.2$ for the longest ones ($555 \leq n \leq 696$). It is not possible to say whether this represents several distinct scaling regimes, or a crossover between spheroidal and elongated chains. The length at which the first switch in

exponent takes place (*i.e.*, $n \approx 269$) improves on the previous rough estimate of $n \sim 300$ [4,5].

The results from (*i*)–(*iv*) indicate that the number of domains has an effect on protein fold compactness, leading to chains that are *larger in absolute size* as the number of domains increases. In addition, we find that all maximally compact proteins with up to three domains have at least two scaling regimes, corresponding to short and long chains (*cf.* (*iii*)). The observation of these two scaling regimes is consistent with trends in the literature, where the change in ν-exponent was conjectured to be caused by the fact that most long proteins include several domains [4–6]. Our results show, however, that this is not the case; the occurrence of a ν_{eff}-exponent distinct from the small-n scaling regime seems to be an intrinsic characteristic of longer chains, regardless of their number of domains. This may indicate there exists a maximum value in chain size beyond which reorganization in protein structure must take place, an effect that holds true for all folding classes and domain numbers.

5 Results for the scaling behaviour of individual single domains

Up until now, we have studied the molecular size of single- and multi-domain proteins without considering the context within which the domains occur. In other words, the emphasis has been on the *complete* chains. By changing our focus from full chains to the individual isolated domains we can draw our attention to their folding features, length, *and* provenance (*i.e.*, whether they have been extracted from single- or multi-domain proteins). Our results are derived from subensemble of domains with smallest local radius of gyration extracted after a $\Delta n = 10$ binning, as in the previous section.

Figure 3 shows the main results for size-scaling behaviour in terms of folding class and provenance (Eq. (8)) as derived from the $[r_g^{(t)}]^*_{j,FC}$-values (*i.e.*, the domains with the smallest radius of gyration). For a common reference, the panels in Fig. 3 include a single dashed line with the same intercept and slope, the latter taken as the exact scaling exponent for collapsed polymers ($\nu_{CP} = 1/3$).

Three qualitative trends emerge from Fig. 3: (*i*) evaluated as a global family, isolated domains follow a similar pattern, regardless of having originated from single- or multi-domain proteins; (*ii*) the salient characteristic is a qualitative scaling closer to the collapsed-polymer regime in short domains, and closer to random-polymer regime in longer-chain domains; (*iii*) domains appear to move away from the reference line as the t-index increases, *i.e.*, multi-domain proteins have a *larger absolute size* for the same chain length.

A careful analysis of the data allows us to quantify these observations, as well as reveal some subtle variations in scaling behaviour. For instance, if one studies the data in Fig. 3 without reference to provenance (*i.e.*, exclusively in terms of folding classes, as in Eq. (7)), we obtain:

$$\ln [r_g]^*_\alpha = (0.401 \pm 0.008) \ln n + (0.64 \pm 0.04)\,, \tag{14a}$$

$$\ln [r_g]^*_\beta = (0.385 \pm 0.008) \ln n + (0.68 \pm 0.05)\,, \tag{14b}$$

$$\ln [r_g]^*_{\alpha+\beta} = (0.420 \pm 0.007) \ln n + (0.51 \pm 0.04)\,, \tag{14c}$$

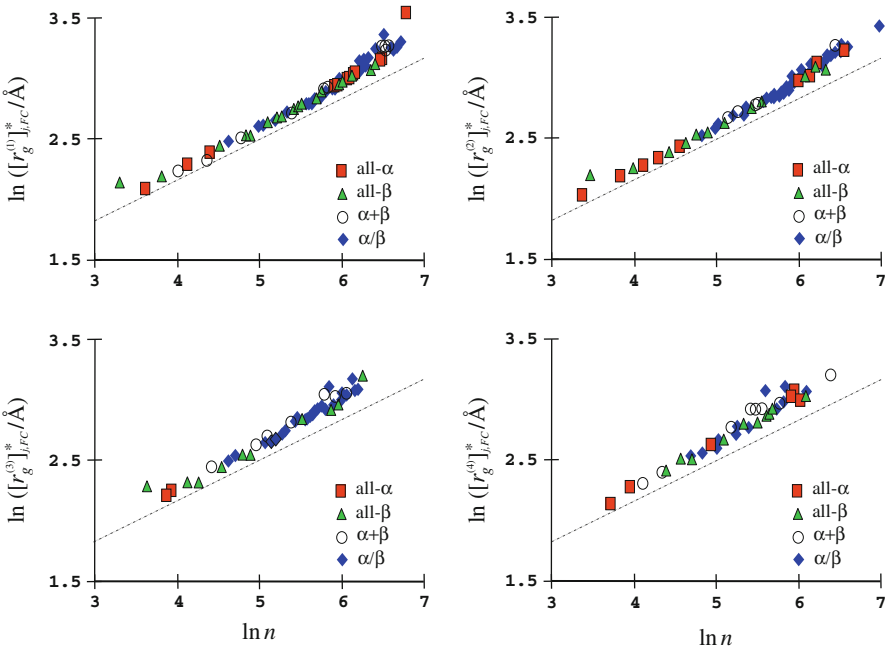

Fig. 3 Scaling behaviour for the most compact domain units extracted from proteins containing t-domains. Results in each panel include domains belonging to the four major folding classes (FC), using the classifications supplied by the SCOP data base. The notation $[r_g^{(t)}]^*_{j,FC}$ indicates thus the domain with the smallest radius of gyration among those in the jth bin for the FC family (with $FC = \alpha, \beta, \alpha+\beta, \alpha/\beta$), and originating from a protein chain initially containing t-domains. The dashed line corresponds to the ideal slope for collapsed polymers (*i.e.*, $\nu \equiv \nu_{CP} = 1/3$); the same line is used in all four panels for easier comparison. The results indicate the same trend in scaling for all domains, although units originating from complex proteins have larger sizes for the same chain length. (See text for more details)

$$\ln [r_g]^*_{\alpha/\beta} = (0.396 \pm 0.008) \ln n + (0.62 \pm 0.05) , \tag{14d}$$

where the error bars correspond to standard errors. Within the statistical confidence, it is clear that $(\alpha + \beta)$-domains are characterized by the largest-$\bar{\nu}_{FC}$ value, while the (all-β) and (α/β) folding classes are closer to full-protein-chain behaviour discussed in Sect. 4. It should be noted that the correlation for (α/β)-folds includes several long-chain bins with low sampling numbers ($n > 551$). If these structures were excluded as outliers, due to the scarcity of long-chain proteins, the scaling exponent would decrease to $\bar{\nu}_{\alpha/\beta} = 0.363 \pm 0.007$. Similar exclusion of outliers in the (all-β)-folds produces $\bar{\nu}_{\beta} = 0.372 \pm 0.004$. These revised global estimates confirm Eqs. (14) in making the (all-β)- and (α/β)-domains the more spheroidal of all folds.

Other effects can be inferred from the effective $\bar{\nu}^{(t)}_{FC}$-exponents for individual classes of domains. Table 1 collects these scaling exponents, together with the $\bar{\nu}^{(t)}$-values (last column in Table 1, *cf.* Eq. (6)). The last column translates the quantitative trends displayed in Fig. 3 (*i.e.*, ignoring the different folding classes). From Table 1, we can state:

Table 1 Dependence of the mean-size scaling exponent length ($\bar{\nu}_{FC}^{(t)}$), Eq. (8) in the scaling behaviour for the most compact domains within the *FC*-family (folding class) and extracted from proteins containing t-domains

t	*FC*				*All FC-folds*
	α	β	$\alpha+\beta$	α/β	
1	0.387 ± 0.008	0.380 ± 0.010	0.420 ± 0.011	0.397 ± 0.009	0.402 ± 0.008
2	0.386 ± 0.011	0.376 ± 0.019	0.402 ± 0.008	0.398 ± 0.011	0.398 ± 0.009
3	0.425 ± 0.019	0.380 ± 0.017	0.408 ± 0.025[a]	0.339 ± 0.020	0.394 ± 0.009
4	0.382 ± 0.025	0.361 ± 0.022	0.399 ± 0.024[a]	0.341 ± 0.023	0.393 ± 0.014

The results correspond to the $[r_g^{(t)}]_{j,FC}^{*}$-values extracted with a $\Delta n = 10$ binning (unless otherwise noted). The last column corresponds to the $\bar{\nu}^{(t)}$-exponent, *i.e.*, without any reference to folding class (Eq. (6)). Error bars correspond to standard errors in the regression lines

[a] If extracted with a $\Delta n = 50$ binning (to account for a more even statistics), we obtain: $\bar{\nu}_{\alpha+\beta}^{(3)} = 0.381 \pm 0.011 (62 \leq n \leq 430)$ and $\bar{\nu}_{\alpha+\beta}^{(4)} = 0.381 \pm 0.022 (59 \leq n \leq 597)$

(*i*) The scaling exponent for (all-β)-domains is systematically smaller than that for ($\alpha + \beta$)-domains, but both values appear to be insensitive with respect to the provenance of the chain. This may also be the case for the all-α domains, but the dispersion makes it hard to assess whether there is a slight drift towards larger scaling exponents in the all-α fold as the t-index increases. On the other hand, the larger global exponent $\bar{\nu}_\alpha = 0.401 \pm 0.008$ (Eq. (14a)) suggests that may be the case.

(*ii*) Table 1 indicates that (α/β)-domains originating from one- and two-domain proteins are statistically identical to the global (α/β) value while possessing a significantly larger scaling exponent than those extracted from three- and four-domain proteins ($\bar{\nu}_{\alpha/\beta} \approx 0.40$ vs 0.34, respectively).

(*iii*) The last column in Table 1 shows that the differences between $\bar{\nu}^{(t)}$-values disappear once we consider the smallest domains, regardless of their folding features. A simple mean over the last column gives $\bar{\nu}^{(t)} = 0.40 \pm 0.01$ (*i.e.*, an average for $t \leq 4$); this single value that translates the global similarities observed in Fig. 3 and in Eqs. (14).

Table 2 completes the analysis of scaling laws by providing the different behaviour of the pre-exponential lengths $\ell_{FC}^{(t)}$ (see Eq. (8)). The table indicates that $\ell_{FC}^{(t)}$ increases with t for $FC = \beta, \alpha + \beta, \alpha/\beta$, while it remains relatively constant in the all-α domains. The net result (last column in Table 2) shows a small increase in $\ell^{(t)}$-values for domains originating from multi-domain proteins. These results, and those in Table 1, indicate that scaling in isolated domains depends on folding types. On the other hand, if one neglects the folding classes, only the pre-exponential constant $\ell^{(t)}$ (and not $\bar{\nu}^{(t)}$) is sensitive to the domain's provenance.

Figure 4 illustrates a set of results collected in Tables 1 and 2. In this case, we show the linear regressions for the ($\alpha + \beta$) folding class for $t = 1, 2, 3,$ and 4, contrasted with the values for the "base line" (dark circles). Whereas the nearly-parallel lines show that this folding class has a unique $\bar{\nu}_{FC}^{(t)}$-exponent regardless of provenance

Table 2 Dependence of the pre-exponential length ($\ell_{FC}^{(t)}$, Eq. (8)) for the most compact domains within the *FC*-family (folding class), and extracted from proteins containing t-domains

t	*FC*				*All FC-folds*
	α	β	$\alpha+\beta$	α/β	
1	0.71 ± 0.05	0.71 ± 0.05	0.59 ± 0.06	0.62 ± 0.05	0.66 ± 0.05
2	0.76 ± 0.05	0.78 ± 0.10	0.64 ± 0.04	0.64 ± 0.07	0.71 ± 0.04
3	0.58 ± 0.09	0.77 ± 0.09	0.65 ± 0.13	0.98 ± 0.11	0.93 ± 0.07
4	0.79 ± 0.12	1.19 ± 0.20	0.72 ± 0.12	0.97 ± 0.12	0.89 ± 0.09

The uncertainties correspond to standard errors. Results correspond to binning $\Delta n = 10$. (See Table 1 for more details on the notation)

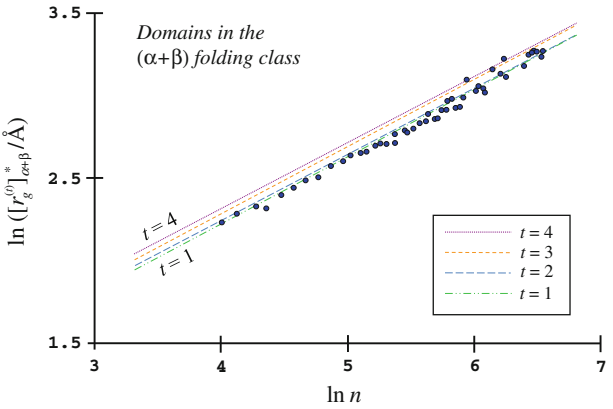

Fig. 4 Effect of domain provenance on the size scaling for domains in the $(\alpha+\beta)$ folding class. *The circles* represent the set of domains with the smallest radius of gyration, irrespective of provenance (*i.e.*, the "base line"). The four linear regressions give the scaling behaviour for domains originating from proteins with t-domains ($t \leq 4$). *The parallel lines* indicate a unique $\bar{\nu}_{\alpha+\beta}^{(t)}$-exponent scaling; their systematic shift with t indicates that the domains are increasingly larger when they originate from multi-domain proteins

(*cf.* Table 1), their shift as a function of t indicates that domains originating from complex proteins increase in size as the number of neighbouring domains also increases (*cf.* Table 2). This deviation can be characterized quantitatively for all folding classes using the $s_{FC}^{(t)}$ and $\rho_{FC}^{(t)}$ descriptors introduced in Sect. 3.

Table 3 shows the deviation from the established base-line for the most compact domains (Eqs. (9)–(12)). The top value in each entry measures the absolute displacement descriptor $s_{FC}^{(t)}$ (*cf.* Eq. (10)), while the entry in parenthesis gives the relative displacement $\rho_{FC}^{(t)}$ (*cf.* Eq. (12)). The last column presents the folding-independent deviations $s^{(t)}$ and $\rho^{(t)}$ (*cf.* Eqs. (9) and (11), respectively). Both descriptors agree that there is a very clear trend: as t increases, the radius of gyration moves away from the reference behaviour defined by our "base line." The deviation, though observable in all cases, is largest for (all-α)-domains. In other words, there is no doubt that, although the $\bar{\nu}$-exponents are relatively insensitive to provenance (*cf.* Table 1), domains are *larger* in absolute terms (for the same chain length) when there are other domains present in a given protein. In other words, the monomer density is *lower* in compact domains

Table 3 Deviation in scaling behaviour for the most compact single domains within the *FC*-family (folding class) and extracted from proteins with t-domains

t	*FC*				*All FC-folds*
	α	β	$\alpha+\beta$	α/β	
1	0.0156	0.0082	0.0067	0.0090	0.0088
	(1.0063)	(1.0030)	(1.0022)	(1.0032)	(1.0033)
2	0.0344	0.0413	0.0235	0.0255	0.0210
	(1.0128)	(1.0150)	(1.0082)	(1.0092)	(1.0077)
3	0.0932	0.0650	0.0510	0.0854	0.0628
	(1.0344)	(1.0250)	(1.0187)	(1.0316)	(1.0238)
4	0.1316	0.0885	0.0890	0.1091	0.0823
	(1.0493)	(1.0353)	(1.0328)	(1.0415)	(1.0310)

The top value in each entry is the absolute displacement $s_{FC}^{(t)}$ from the base line (Eq. (10)); the entry in parenthesis corresponds to the relative displacement $\rho_{FC}^{(t)}$ (Eq. (12)). The results show a systematic deviation in compactness elicited by the increasing number of domains. The systematic shift in $\rho_{FC}^{(t)}$ values indicate a loss of compactness caused by increasing values of the radius of gyration, while conserving the same scaling law. The last column gives the results for the deviations in $[r_g^{(t)}]^*$, *i.e.*, without reference to folding class (Eq. (6)); these results highlight size effects associated exclusively with the number of neighbouring domains in the original protein chain

originating from complex proteins. Table 2 suggests that this behaviour arises from the fact that $\ell^{(t)}$-values increase with t, while keeping $\bar{\nu}^{(t)}$ constant.

6 Further comments and conclusions

The existence of a power law relating mean size and chain length within a family of polymers indicates a unifying principle that directs their spatial organization. In this context, the present results provide insight into a number of issues relating to protein structure. We summarize here the principal observations and their significance.

In intact full chains, we find that single- and multi-domain proteins follow the same trends in molecular size. The main variations in scaling laws appear to be caused by the number of residues, where *short* proteins exhibit *smaller* size-scaling exponents (closer to $\nu_{CP} = 1/3$), *regardless* of the number of domains present. This result is in line with our findings for *isolated* domains, where we find a transition in size-scaling behaviour at a critical chain length estimated at $260 \leq n \leq 269$ amino acids [28]. In addition, these results shed light on the significance of other observations in the literature [30,31], which indicate that various protein properties (*e.g.*, the exposed surface area, the distribution of voids inside the protein) also appear to scale differently with chain size in small and large protein chains, with a transition estimated between $n = 255$ and $n = 270$.

With respect to the general questions posed in Sect. 1, we find that units "cut away" from multi-domain proteins resemble qualitatively single-domain chains in terms of size scaling. Variations in $\bar{\nu}$-exponent depend principally on the domain's folding

class, rather than on the number of neighbouring units in the original protein. We find, however, that pre-exponential lengths (*cf.* Eq. (5)) do depend on the number of units; domains extracted from complex proteins are *larger* in absolute size (for the same chain length) than those in single-domain proteins. In other words, domains in complex proteins are packed more loosely (*i.e.*, have a lower monomer density). This could be for one of two basic reasons: (a) all "common folds" within the same *FC*-category [11–14] are more swollen in multi-domain proteins; (b) proteins with several domains feature mostly a subset of common folds which are less compact. A preliminary inspection (data not shown) indicates that both possibilities do occur, though the latter situation is more frequent, *i.e.*, we find distinct common folds depending on domain provenance.

These results support two important concepts that have been debated in the literature:

(*i*) First, a similarity in scaling exponents hints at a common underlying global folding mechanism; this is in line with the conjecture that spatial domains fold independently [7–9,32]. Moreover, the small variation in $\bar{\nu}_{FC}$-exponents suggests that distinct folding classes may not fold in exactly the same fashion; this observation agrees well with the notion that folding rates and mechanisms are related, at least in part, to the distinct topologies of native states [33]. We also find that all-β and (α/β)-domains are characterized by slightly smaller scaling exponents (once outliers are removed), and consequently are the most spheroidal folds. This observation agrees well with the reported fact that all-β folds have a higher density of tertiary contacts [34,35] and a larger number of buried residues [35]. In contrast, (all-α)-folds, or folds with segregated α-helical content (*e.g.*, $(\alpha+\beta)$-folds), have slightly larger $\bar{\nu}_{FC}$-values and are thus less globular; this is consistent with the notion that these types of proteins expose a larger fraction of residues to the solvent [35]. To some degree, these distinctions are linked to fundamental differences between helices and strands [18,19]. Helices are bulkier, on average ten amino-acid rigid cylinders with a limited number of relative orientations, favouring the formation of nearly parallel bundles in (all-α)-folds, *i.e.*, rod-like structures as opposed to spheroidal [36]. Strands, on the other hand, are shorter and more flexible, with a wider range of accessible relative orientations, thus allowing the formation of curved arrangements (*e.g.*, β-barrels). As a result, the (all-β)-folds, and to some extent (α/β)-folds, are intrinsically better equipped to form more spheroidal globular structures.

(*ii*) Our results also indicate that some domain properties (*e.g.*, the $\ell_{FC}^{(t)}$-lengths) have "traces" of the original multi-domain protein architecture. In other words, there must exists at least some degree of interaction between the domains that is responsible for the consistency in these variations. This observation may provide support to the conjecture that the increased propensity of large proteins to misfold (relative to small proteins) is caused by interactions between domains [9,37]. According to this notion, domains in complex proteins, although separated in the native state, interact during folding, thereby perturbing the mechanism and increasing the possibility of misfolding or aggregation. This work shows that the effects of these interactions on folding can be recognized in the

$\ell_{FC}^{(t)}$-lengths, while the $\bar{\nu}_{FC}^{(t)}$-exponents are nearly independent of the number of domains in the original protein.

From the point of view of molecular modeling, our approach complements other approaches used in the literature to establish the boundaries of a domain within a protein [21,38]. Often, algorithms that rely on domain automated recognition provide different answers [22–25]; non-automated classification relies on visual inspection and is thus subjective [11–14]. Our results provide an additional criterion to select a domain region, *i.e.*, requesting that it fits within an established scaling law for a given folding class.

It should be noted that the present analyses focus on proteins with the smallest R_g-values for a given $(n, n + \Delta n)$-bin. It is possible that different constraints may reveal other laws, as may be the case for proteins within specific functional families or more restricted folding classes. This analysis is currently underway, and it will be communicated elsewhere.

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 University Press, 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. G.A. Arteca, Phys. Rev. E **49**, 2417 (1994)
5. G.A. Arteca, Phys. Rev. E **51**, 2600 (1995)
6. G.A. Arteca, Phys. Rev. E **54**, 3044 (1996)
7. M.O. Lindberg, M. Oliveberg, Curr. Opin. Struct. Biol. **17**, 21 (2007)
8. S.W. Englander, L. Mayne, M.M. Krishna, Q. Rev. Biophys. **40**, 287 (2007)
9. C.F. Wright, S.A. Teichmann, J. Clarke, C.M. Dobson, Nature **438**, 878 (2005)
10. F.U. Hartl, M. Hayer-Hartl, Nat. Struct. Mol. Biol. **16**, 574 (2009)
11. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
12. A. Andreeva, D. Howorth, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **32**, D226 (2004)
13. A. Andreeva, D. Howorth, J.M. Chandonia, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **36**, D419 (2008)
14. A. Heger, L. Holm, J. Mol. Biol. **328**, 749 (2003)
15. 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)
16. H.M. Berman, K. Henrick, H. Nakamura, Nat. Struct. Biol. **10**, 980 (2003)
17. P. Rogerson, G.A. Arteca, J. Math. Chem. **49**, 1493 (2011)
18. L. Holm, C. Sander, Proteins **33**, 88 (1998)
19. A.M. Lesk, *Introduction to Protein Architecture* (Oxford University Press, Oxford, 2001)
20. G.A. Petsko, D. Ringe, *Protein Structure and Function* (New Science Press, London, 2004)
21. C.P. Ponting, R.R. Russell, Annu. Rev. Biophys. Biomol. Struct. **31**, 45 (2002)
22. S.J. Wodak, J. Janin, Biochemistry **20**, 6544 (1981)
23. M.B. Swindells, Protein Sci. **4**, 103 (1995)
24. M.H. Zehfus, Protein Sci. **6**, 1210 (1997)
25. C.J. Tsai, R. Nussinov, Protein Sci. **6**, 24 (1997)
26. M. Dumontier, R. Yao, H.J. Feldman, C.W. Hogue, J. Mol. Biol. **350**, 1061 (2005)
27. L.S. Wyrwicz, G. Koczyk, L. Rychlewski, D. Plewczynski, J. Phys. Condens. Matter **19**, 285222 (2007)
28. P. Rogerson, G.A. Arteca, J. Math. Chem. **50**, 169 (2012)
29. M. Leeuw, S. Reuveni, J. Klafter, R. Granek, PLoS ONE **4**, e7296 (2009)

30. J.P. Zbilut, G.H. Chua, A. Krishnan, C. Bossa, K. Rother, C.L. Webber, A. Giuliani, Proteins **66**, 621 (2007)
31. A. Szilágyi, Proteins **71**, 2086 (2008)
32. M.Y. Shen, F.P. Davis, A. Sali, Chem. Phys. Lett. **405**, 224 (2005)
33. D. Baker, Nature **405**, 39 (2000)
34. J.D. Bloom, D.A. Drummond, F.H. Arnold, C. Wilke, Mol. Biol. E **23**, 1751 (2006)
35. T. Begum, T.C. Ghosh, J. Mol. Evol. **71**, 60 (2010)
36. C. Chothia, M. Levitt, D. Richardson, Proc. Natl. Acad. Sci. USA **74**, 4130 (1977)
37. W.J. Netzer, F.U. Hartl, Nature **388**, 343 (1997)
38. J. Liu, B. Rost, Proteins **55**, 678 (2004)