J Math Chem (2011) 49:1493–1506
DOI 10.1007/s10910-011-9825-4

ORIGINAL PAPER

Molecular size scaling in families of protein native folds

Parker Rogerson · Gustavo A. Arteca

Received: 14 March 2011 / Accepted: 18 April 2011 / Published online: 9 June 2011

Abstract The mean size of the most compact native states of globular proteins, independent of folding type, follows the scaling law of collapsed polymers $R_g \sim n^{1/3}$, relating the radius of gyration R_g to the number of protein residues, n. Until now, this behaviour has only been observed within a small subset of unrelated single-domain proteins with $n < 300$. Here, we employ the SCOP database of protein folds to study systematically the scaling behaviour of well-defined families of domains that share structural and functional characteristics. In the particular case of helical proteins, we identify the folding types that can be associated with scaling laws corresponding to compact behaviour (e.g., the cytochrome-C monodomains) and noncompact behaviour (e.g., the immunoglobulin/albumin-binding and spectrin-repeat domains). Our results quantify the size variations within some folding families, as well as reveal that some distinct folds represent structures with equivalent compactness.

Keywords Polymer size · Protein folds · Alpha-proteins · Helical bundles · SCOP database

1 Introduction

The scaling between mean radius of gyration R_g and the number of monomers n is a simple indicator of the nature of the dominant interactions in a polymer and their degree of compactness. In free (unconfined) three-dimensional structures, the power law for R_g, averaged over all accessible configurations, follows a well-known dependence [1]:

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

$$\langle R_g^2\rangle^{1/2} \sim an^{\nu}, \tag{1}$$

and it corresponds to collapsed polymers (CP) for $\nu_{\mathrm{CP}} = 1/3$, random-walk (RW) polymers for $\nu_{\mathrm{RW}} = 1/2$, and the mean size of self-avoiding-walk (SAW, noncompact) polymers for $\nu_{\mathrm{SAW}} = 0.588 \pm 0.002$ [2,3], often approximated by the Flory exponent $\nu_{\mathrm{F}} = 3/5$ [1]. These exponents can also be related to solvent quality and polymer self-interactions, as they identify respectively: (i) chains collapsed in the attractive regime for poor solvents ($\nu_{\mathrm{CP}} = 1/3$), (ii) chains swollen by the excluded-volume repulsions in good solvents ($\nu \approx \nu_{\mathrm{F}}$), and ($iii$) and polymers at the *theta* (or "ideal") solvent, where attraction and repulsions are balanced ($\nu = 1/2$). In addition, we find the scaling exponent $\nu = 1$ in the case of rigid polymer chains with essentially a single rodlike configuration. For a given interaction type, the scaling exponent ν depends only on the spatial dimension [4], whereas the pre-exponential factor "a" in Eq. 1 depends on the particular polymer chain (e.g., chemical composition and detailed monomer-monomer interaction). In other words, although different in geometry and actual size, polymers in the same scaling regime will appear as parallel lines in a logarithmic plot R_g vs n.

In contrast to free unconfined homopolymers with nearly-degenerate low-energy states, protein native states span a very narrow range of configurations that dominate the value of the radius of gyration. These native states fold in specific classes which have been selected to satisfy structural and functional roles, and they can exhibit a wide range of compactness. If one focuses only on choosing the most compact available structures within a given window of residue numbers, the smallest globular proteins scale indeed as $\min_{\{n\}} \{R_g\} \sim n^{1/3}$, *i.e.*, as collapsed polymers [5–7]. Other independent shape descriptors are also consistent with this regime; for instance, the mean overcrossing number (or "average crossing number", $\bar{N}$) [5–7] of small globular proteins follows the law $\bar{N} \sim n^{4/3}$, in agreement with the scaling of the $\bar{N}$-value when averaged over the configurations accessible to compact and ideal knots [8,9], collapsed linear polymers, and dense polymer networks [10].

Nevertheless, the significance of these correlations remains unclear in biophysical terms, since it arises from grouping together a small group of unrelated proteins, spanning distinct folds, functions, and primary sequences, rather than accessible conformations of the same polymer. Moreover, some protein native states are *not* compact (e.g., viral coats, muscle proteins), and even those with the smallest R_g values appear to deviate from the $\nu_{\mathrm{CP}} = 1/3$ law at $n > 300$ values [5,6]. Since these proteins are unrelated both structural and functionally, it is not possible to assess whether the occurrence of scaling behaviour can be associated with a folding mechanism defined by native tertiary structure [11,12].

As far as we know, the relation between compactness and type of fold has never been probed in detail. Here, we study systematically the mean molecular size restricted to well-defined families of protein folds. Our goal is twofold: (i) establish whether scaling behaviour can be recognized in subgroups of native states, (ii) elucidate the nature of folding types that fall into established scaling laws commonly associated with polymer compactness.

The present work focuses on single-chain single-domain proteins with α-helical native folds, specifically those classified in the Structural Classification of Proteins

(SCOP) database (version 1.73) as belonging to the "three-helix common fold." It should be noted, however, that this class includes proteins not restricted exclusively to having three α-helices [13–15].

Since single-chain proteins are limited in practice to $n < 1{,}000$, it is not possible to assess a truly asymptotic scaling law. We show below that restricted pseudo-scaling is nevertheless possible in some cases, thereby allowing one to estimate effective scaling exponents ν_{eff}. In this work, we focus our analysis on a set of α-helical proteins derived from the SCOP data base with $n < 225$ amino acids.

The work is organized as follows. The next section lays out the classification of protein folds by SCOP and the sampling criteria used in this study to select unique α-proteins. We focus on groups where sampling allows sufficient information to estimate the occurrence effective scaling over a range of n-values. Next, we present the results for the cytochrome-C family of helical bundles, the fold which exhibits the best defined power law for near-compact structures. The occurrence of noncompact scaling is observed in the immunoglobin/albumin-binding domains and spectrin-repeat motifs. The two limit cases (namely, the cytochrome-C and spectrin-repeat families) provide the framework to interpret the results observed in other families; the behaviour of the latter is illustrated using the histone fold. Finally, we examine the distribution of all unique single-chain single-domain α-helical proteins that are part of the "three-helix common fold" of SCOP lineages. Within that ensemble, we uncover the scaling behaviour for the subgroup of most compact structures. We close with a discussion of our observations in the context of the classification of protein folds based on structural and functional criteria.

2 Classification of fold topology and sampling criteria

We use the "Structural Classification of Proteins" (SCOP) database to select the protein folding families whose molecular sizes will be monitored. The choice of SCOP is motivated by the fact that it represents a standard of protein domain classification against which all other methods and classification criteria are compared.

The SCOP system organizes proteins into lineages based on structural relationships (e.g., having similar "fold topologies") and evolutionary relationships (e.g., sequence homologies and shared biological function) [13–15]. The SCOP scheme is commonly described as "manually curated" [16], as it results from the visual inspection of every entry and a consensus of subjective assessments of domain family definitions.

Our goal is to study whether protein domains with similar fold topology, as defined by the SCOP database, follow distinct scaling with respect to the radius of gyration of the native state, R_g. For simplicity, the protein is represented by an α-carbon chain:

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

where $\mathbf{r}_i$ represents the position vector of the ith α-carbon, and $\mathbf{r}_0$ the centroid of the α-carbon chain, and n is the total number of residues. The $\{\mathbf{r}_i\}$-coordinates are extracted from the Protein Databank (PDB) [17,18]. This simplified R_g omits all information irrelevant to the fold's global compactness, e.g., primary sequence, side chain conformation, and the peptide bond.

Previous work indicates that *no* global size scaling can be associated with protein native states because the dispersion of R_g-values is very large over non-redundant structures deposited in the PDB [5]. Only when restricting oneself to the subset of proteins with minimal R_g can a scaling law be defined, corresponding to $R_g \sim n^{1/3}$ in $n < 250$, and $R_g \sim n^{1/2}$ thereafter. As mentioned before, this result sheds no light as to the relation between scaling and folding type since it emerges from a selected group of totally unrelated proteins.

The version of the PDB considered in our work holds *ca.* 12,900 unique (non-redundant) single-domain entries; this value may vary depending on the criteria used to establish and eliminate redundancy [19]. In our study, using the criteria explained below, we have characterized 308 "unique" single-domain proteins representing twenty-seven three-helical core lineages within the *all-* α classes of the SCOP database. These entries are then used to test: (*a*) whether proteins from similar fold families (or "lineages") share a single scaling law for R_g, and (*b*) whether any native folds in that group can be assigned univocally the size-scaling exponent associated with the compact-polymer regime.

2.1 The SCOP database

We have computed the R_g-values for native states belonging to selected lineages, extracted from the SCOP database. The SCOP classification system uses the *protein domain* as its fundamental unit. Structures are grouped according to the following hierarchy:

(i) *domain*: objects at this base level exhibit either complete sequence homology or virtually identical folds;
(ii) *family:* proteins in this category exhibit > 30% sequence homology or a very similar tertiary structure [20];
(iii) *superfamily*: these entries are characterized by low sequence homology but either their structure or function suggests a common evolutionary origin;
(iv) *common fold*: these entries have common major secondary structures with the same topological connections;
(v) *class*: this broadest designation is determined by the content of secondary structure. For instance, all the proteins from our study belong to the *all-* α *class*, meaning that all the proteins and protein domains found therein consist entirely of α-helical structural elements.

The SCOP database sorts proteins within the above categories in a nonautomated fashion, as it relies on the manual inspection by human experts [13–15]. This subjective analysis is a weak point when it comes to assessing the degree of similarity in folds. In our study, we analyze whether size scaling is a valid complementary criterion that can be applied to characterize the SCOP hierarchy.

2.2 Criteria for eliminating database redundancy

We began with a total of 2,824 SCOP entries with three helices as a core. In order to eliminate redudancy, we represent each protein entry by *a unique single chain with one domain only.* Quaternary structure is ignored. In the case of homomultimeric proteins, only the first chain entry in each PDB file is considered; this condition excludes 778 entries. Partial single domain entries from multi-domain chains are also excluded (832 entries). Proteins that contain "gaps" (e.g., missing amino acids) were excluded; there are 483 entries in this category, with 160 of those not included in the previous trimming categories. This leaves a total of 1,072 entries to sort through for sequence-related redundancy.

The SCOP database contains many closely-related sequences that need to be screened out in order to avoid a skewed sampling for size-scaling relationships.

In order to generate a unique and representative entry, we use the following criteria:

(1) First, all proteins with $> 90\%$ sequence similarity are replaced by a single representative.
(2) We allow for variations in sequence length in related proteins up to 15 amino acid residues. In other words, proteins with the same *domain* classification, yet differing in less than 15 residues are not considered distinct entries, and are represented by a single entry.
(3) Given that termini regions can be relatively unstructured, we also allow that two chains can differ by a segment at either end of the protein with up to 12 amino acids. Any two such entries belonging to the same *domain* classification are not considered independent.

Both NMR and X-ray diffraction data were used in our study. Whenever multiple experimental entries exist, we chose the highest resolution; structures with resolution above 3.2 Å were ignored. In cases where α-carbon coordinates are missing, another protein from the same cluster was selected. Only *all*-α protein chains with $n > 35$ were considered, thereby excluding peptides that may lack a well-defined native state.

The application of these criteria results in 308 proteins in the *all-* α *class* of the SCOP database, with three helices as the core, that are deemed to be distinct. These species were grouped according to their folding topologies as *domains* in the SCOP classification, and then the R_g value was computed for each. Each "domain cluster" spans a range of n values. Depending on the particular structures and fold types, some clusters may be characterized by an effective size-scaling exponent ν_{eff}. The presence of a common scaling behaviour could be strong indication of a similar native fold, and possibly a related folding mechanism [11,12]. On the other hand, the lack of a well-defined size scaling over a large range of n values may indicate a merely superficial similarity in their folding topologies. In other words, our analysis can serve to complement the SCOP classification by shedding light on folding homologies.

It should be noted that not only the classification and evolution of domains remains a matter of debate [16,20–23], but also the distinction between structural or functional domains, as well as the location of their putative boundaries within a protein chain [24–29]. Understanding the occurrence of size scaling may be helpful at clarifying this ellusive notion.

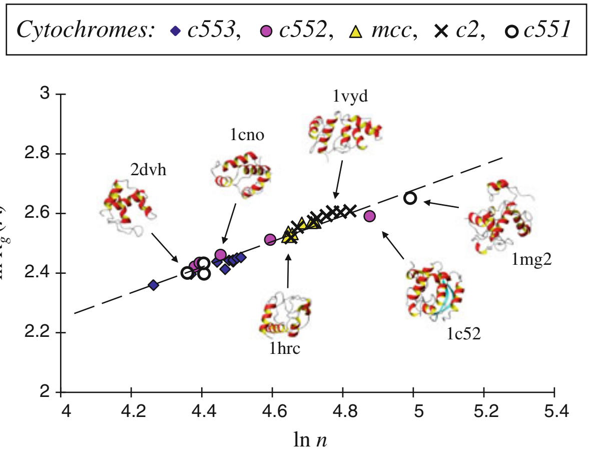

Fig. 1 Size scaling behaviour in the *cytochrome-C monodomain* (CCM) folding family. The plot relates the radius of gyration R_g of the α-carbon trace of an single protein chain to the number of amino acid residues, n. Some proteins within the CCM lineages are highlighted, denoted by their PDB codes. All chains correspond to "chain A" in the PDB, except for 3cyt and 1 mg2 where we show "chain O" and "chain D," respectively

3 Compact scaling behaviour: the cytochrome-*C* helical bundles

The *cytochrome-C monodomain* (CCM) is a *folding family* characterized by the occurrence of a special packing of α-helices. Including redundancy, there are 369 structures in this category. After applying the criteria for removing redundancy discussed in the previous section, the CCM lineage yields 44 "unique" entries with the same class, common fold, superfamily, and family. It the most well-represented *all*- α class lineage considered here.

The SCOP database lists several domain shapes within the CCM family; five of these are present in sufficient numbers to evaluate size scaling at the domain level: cytochrome *c*6, denoted as *c*553, with 10 entries, cytochrome *c*552 (5 entries), cytochrome *c*551 (5 entries), cytochrome *c*2 (8 entries), and the mitochondrial cytochrome-*C*, denoted as *mcc*, with 6 entries. All these species appear in Fig. 1. Within the all-α class, the CCM produces the most clearly defined example of size scaling behaviour for a folding family of small proteins. The set spans chain lengths $70 < n < 150$, and includes structures with three up to five α-helices.

Figure 1 indicates a unique scaling law for all CCM entries, despite their differences in sequence, composition, domain folds, as well as biological origin. A least-squares fitting over the entire ensemble gives an effective scaling exponent $\nu_{\text{eff}}^{\text{CCM}} = 0.432 \pm 0.013$ (95% confidence levels are used throughout). For the entries with the largest span of n-values (namely *c*552, *c*551, *c*2 in Fig. 1), their individual scaling behaviour agrees well with the mean value for the entire ensemble. The statistically more diverse sets (*i.e.*, *c*552 and *c*551) are consistent with smaller exponents: $\nu_{\text{eff}}^{\text{C552}} = 0.34 \pm 0.02$ and $\nu_{\text{eff}}^{C551} = 0.39 \pm 0.03$, respectively.

The correlation is poorer in the cases with smaller span of n-values ($c553$, mcc). Yet, although the $c553$ and mcc data are too limited in span for a scaling analysis by themselves, they contribute at the mean (global) scaling behaviour of the CCM family.

The $\nu_{\mathrm{eff}}^{\mathrm{CCM}}$-exponent suggests compact behaviour ($\nu_{\mathrm{eff}} < 1/2$), yet not at the level of the known globular proteins with the smallest R_g-values and similar chain length. For native states with $n < 306$ (with diverse folding features), the effective exponent is $\nu_1 = 0.34 \pm 0.05$ [6]. Instead, the $\nu_{\mathrm{eff}}^{\mathrm{CCM}}$-exponent for the CCM ensemble is comparable to that of *longer* globular proteins, where the native states with the smallest R_g-values and $n > 306$ give $\nu_2 = 0.41 \pm 0.05$ [6,7]. No other family (to date) of all-α proteins is as well represented, and yields such a well-defined scaling law, as the CCM ensemble. In particular, the $c552$ proteins are consistent with the collapsed-polymer regime.

In contrast, the native states for α-helical bundles will scale differently. We expect the latter to become elongated structures as the chain lengthens, thus adopting asymptotically a cylinder-like (or rod-like) shape with a size-scaling exponent $\nu_{\mathrm{rod}} \to 1$. In other words, a level of compactness consistent with small-ν exponents cannot be met by the packing a few long parallel helices; high compactness is achieved instead by packing a larger number of shorter (nonparallel) helices. Since there are no $n > 150$ structures in the CCM folding family, it would appear that this high level of compactness cannot be maintained for long by packing only α-helices. We can conjecture that longer protein chains with the same compactness will either fold differently or have a different content of secondary structure.

4 Noncompact scaling behaviour: the immunoglobulin and spectrin-repeat helical bundles

The SCOP database includes two well-represented lineages of all-α proteins whose native states resemble noncompact polymers: (i) The *immunoglobin/albumin-binding domain* (IABD), depicted in Fig. 2, and (ii) the *spectrin-repeat domain* (SRD), depicted in Fig. 3. Both domains appear as helical bundles with similar rod-like behaviour for size scaling.

4.1 The immunoglobulin/albumin-binding domain

The IABD folds span chains with length $49 \leq n \leq 187$ amino acids. They include two lineages at the level of superfamily; one of them comprises three distinct families of folds. Their common motif is a three-helix left-handed twist up-down bundle. The ranges of chain length between the families are noncontiguous.

The first seven native folds in Fig. 2 are very short chains ($49 < n < 57$) characterized by an effective size-scaling exponent $\nu_{\mathrm{eff}}^{\mathrm{IABD}} = 0.915 \pm 0.025$. The three highlighted structures in that group (1t60, 1edj (chain A), and 1ℓp1 (chain B)) are regular helical bundles. In contrast, 1gjt (chain A) breaks away from the local scaling law because a terminal coil attached to the three-helix bundle yields a less compact structure. From our point of view, this latter structure should be considered *a new fold* as it deviates from those of the shorter immunoglobulins by the occurrence of a

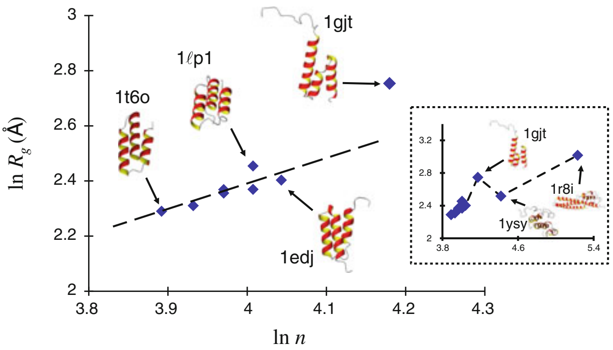

Fig. 2 Size scaling behaviour in the *immunoglobin/albumin-binding domain* (IABD) folding family. Proteins are represented by their corresponding "chains A" in the PDB entries, except for 1ℓp1 whose "chain B" belongs to this folding family. The protein 1gjt is an outlier due to the existence of an extended disorganized coil at a terminus that removes it from the scaling law of the other three-helical bundles. The inset shows the proteins 1gjt, 1ysy and 1r8i; the SCOP database includes these as part of the IABD family, yet are clearly outliers in our analysis

disorganized section. In contrast, the previous seven structures share a unique motif: three antiparallel helices which lengthen from 1t6o to 1edj, hence the rod-like scaling behaviour ($\nu_{\text{eff}}^{\text{IABD}} \approx \nu_{\text{rod}} = 1$).

The inset in Fig. 2 shows the occurrence of "outliers" within the proteins that the SCOP database classifies as belonging to the IABD folding family. It is clear that these long-chain "immunoglobins" scale differently. On the one hand, 1ysy (chain A) has loop and coil regions that offset considerably the regular three-helix geometry. In contrast, 1r8i is a *five*-helix bundle, with two small helices attached to the three-helix left-handed twist up-down bundle. The inclusion of 1gjt, 1ysy, and 1r8i into the IABD protein fold produces a large dispersion ($\sigma(\nu_{\text{eff}}^{\text{IABD}}) = 0.35$, 95% confidence error bar), which prevents any meaningful assessment of a scaling behaviour. Clearly, these structures belong to another scaling regime.

It is an open question whether the similarities in these folds imply a common folding mechanism to that of the seven shortest immunoglobins. From our perspective, the absence of a unique scaling law for the entire group implies that the proteins in the "IABD fold" do not share a structural motif with common compactness. In other words, the lineages defined according to the SCOP database may be *structurally too broad*; they include proteins with radically different size scaling, hence possibly characterized by distinct folding mechanisms.

4.2 The spectrin-repeat domain

The folding family of the spectrin-repeat domain (SRD) comprises a three-helix left handed twist up-down bundle; representative proteins are highlighted in Fig. 3. Among the SCOP lineages in this work, the SRD fold has the most branches: eleven superfamilies, eleven families, and twelve different domains interspersed among twenty entries for our analysis.

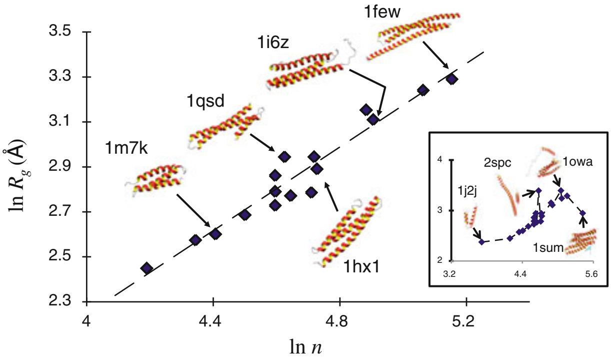

Fig. 3 Size scaling behaviour in the *spectrin-repeat domain* (SRD) folding family. The highlighted proteins in the main body of the figure illustrate how the scaling law arises from a lengthening bundle of three (anti)parallel helices. The inset demonstrates however that this behaviour is not representative of the entire ensemble of proteins classified as a "spectrin-repeat domain" in the SCOP database. Some proteins (e.g., 1j2j, 2spc, 1owa, and 1sum) are organized in a sufficiently distinct fashion that they appear as outliers in our analysis

As indicated in Fig. 3 (see inset), this lineage includes structures that do not follow a common size-scaling law. The main diagram in Fig. 3 collects the results for the SRD-proteins where size scaling can be recognized: a subgroup of sixteen entries that exhibit a tight three-helix antiparallel bundle (see highlighted structures). Between 1rrz (first point in Fig. 3) and 1few (last point), chain length increases by *ca.* 2.5 and we find an effective scaling exponent $\nu_{\text{eff}}^{\text{SRD}} = 0.929 \pm 0.064$, clearly in the regime of rod-like polymers ($\nu_{\text{rod}} = 1$).

The inset in Fig. 3 indicates however that the above behaviour is *not* representative of the entire ensemble. As in Fig. 2, the SCOP classification of SRD entries includes structures with distinct levels of compactness. Some of these "outliers" include:

(a) Protein 1j2j (the smallest in the inset) is classified as belonging to the SRD-fold, even though it contains only two helices, which are of different length and not antiparallel. This structure lies *above* the least-square line with $\nu_{\text{eff}}^{\text{SRD}} \approx 0.929$ (*i.e.*, it is less compact). Protein 2spc behaves similarly, albeit the difference between its two helices is even more dramatic.

(b) Protein 1owa includes the antiparallel three–helix motif, but it is less compactness due to the occurrence of a fourth ("unbundled") helix and a terminal coil. Were this part of the chain disregarded, the resulting segment would fit with the rod-like scaling described previously.

(c) Protein 1sum is a larger six-helix bundle, significantly more compact than the "outliers" above. In fact, this protein can be regarded as comprising *two* tighly packed three-helix bundles. The resulting structure shows a level of compactness ($n \approx 220$, $R_g \approx 19.1$Å) that falls somewhat closer to the cytochrome-C folding family than to the SRD-fold.

From our point of view, there is sufficient commonality among many of the SRD-proteins to define a characteristic scaling behaviour; however, there are other entries

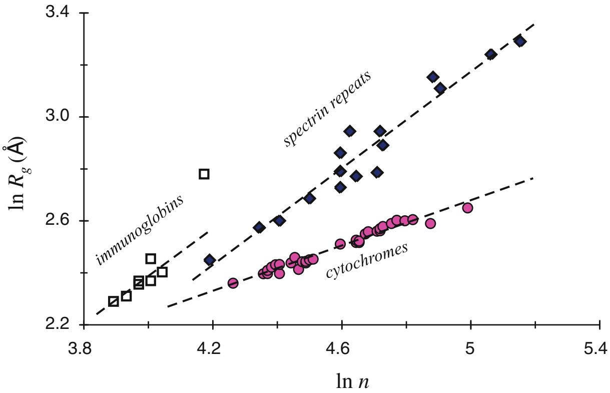

Fig. 4 Comparison of well-defined size scaling behaviour in different lineages of single domain proteins with three α-helices. The diagram includes the three scaling laws depicted in Figs. 1–3, excluding outliers. Two laws emerge: **a** the cytochrome-*C* proteins, with a level of compactness approaching, yet inferior to, that of collapsed polymers ($\nu_{\text{eff}}^{\text{CCM}} \approx 0.43 > \nu_{\text{CP}} = 1/3$); **b** the spectrin-repeat and the immunoglobins, with a low level of compactness approaching that of rod-like polymers ($\nu_{\text{eff}}^{\text{IABD}} \approx \nu_{\text{eff}}^{\text{SRD}} \approx 1$)

within this classification whose compactness level would place them in a category of objects with possibly distinct folding histories.

Within its 95% confidence interval, the effective $\nu_{\text{eff}}^{\text{SRD}}$-exponent extracted from Fig. 3 is indistinguishable from $\nu_{\text{eff}}^{\text{IABD}}$-value in Fig. 2, (i.e., the scaling regimes form parallel lines). In other words, in the cases where scaling behaviour can be recognized, the immunoglobulin and spectrin-repeat folds are found in the same compactness regime. On the other hand, the level of compactness in SRD and IABD is quite distinct from that observed in the cytochrome-*C* folding motif, even though these folds comprise three α-helices. These trends are summarized in Fig. 4, which collects the data for all three folding families (excluding the "outliers" described before). Given that these exponents define a lower bound for flexible polymers ($\nu_{\text{CP}} = 1/3$, collapsed) and an upper bound for rigid polymers ($\nu_{\text{rod}} = 1$, rod-like), we conjecture that folding families of all-α proteins with *well-defined size scaling* should likely fall between these two limit cases.

Note that the two linear trends in Fig. 4 intersect at similar n-values: the least-square lines for CCM and IABD meet at $ln\ n \approx 3.75$, whereas CCM and SRD do at $ln\, n \approx 4.0$. These results define a minimum hypothetical chain length which could belong to *either* scaling regime: $n_{\text{min}} = 48 \pm 5$. This value is consistent with the smallest proteins known to adopt a native state via a two-state folding mechanism [11,12,30].

5 Folding families without a unique scaling behaviour: the histone domain fold

We have contrasted the behaviour in the previous two sections with the histone domain (HD) fold, a folding family that includes 24 unique entries over a relatively large range

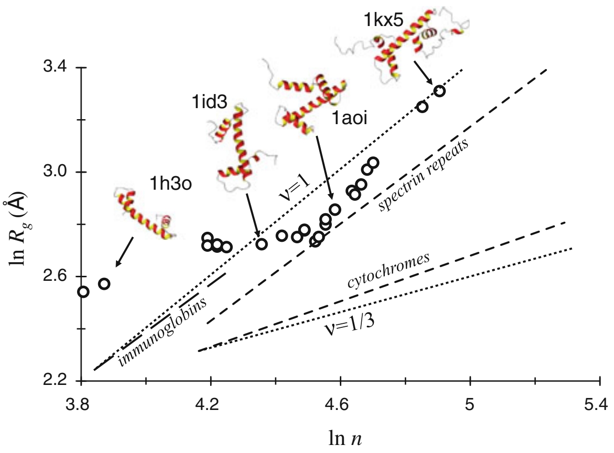

Fig. 5 Size variation for proteins of the histone domain fold (HD, open circles), compared with the scaling behaviours for the three folding families depicted in Fig. 4. The *dashed lines* correspond to the least-square fittings for the latter. The *dotted lines* correspond to the *ideal* behaviour of a rod-like scaling (top, $\nu_{\text{rod}} = 1$) and collapse-polymer scaling (bottom, $\nu_{\text{CP}} = 1/3$), fitted to the smallest R_g-value in immunoglobins and cytochromes-C, respectively

of chain length values. The histone fold is characterized by the occurrence of (at least) three α-helices with different lengths and nonparallel packing. Our results indicate that this fold is structurally quite diverse, and it is *not* characterized by a unique size scaling behaviour.

The structures with the histone fold appear as open circles in Fig. 5. For reference, Fig. 5 includes also the least-square fittings from Fig. 4 (denoted by the dashed lines for cytochromes, immuglobulins, and spectrins). For the sake of comparison, we have added two lines representing the hypothetical upper and lower bounds to R_g for three-helix proteins (dotted lines). These two reference lines were obtained by fitting the ideal limiting slopes to the smallest R_g-values in the family of IABD ($\nu_{\text{rod}} = 1$) and CCM proteins ($\nu_{\text{CP}} = 1/3$). The results for the HD-fold allow us to recognize at least three different size regimes, only one of which defines something approaching a scaling law:

(a) If the histone proteins with $4.5 < ln\, n < 5.0$ (i.e., between *ca.* 90 and 150 amino acids) were assigned an effective exponent $\nu_{\text{eff}}^{\text{HD90}-150}$, it would be absurdly larger than ν_{rod}, namely, $\nu_{\text{eff}}^{\text{HD90}-150} \sim 1.48$. This lack of a recognizable scaling law indicates that these proteins do not share a well-defined common fold. Instead, the entries in this group exhibit structural inconsistencies, e.g., large variations in secondary structural motifs. For instance, 1aoi and 1kx5 (highlighted in Fig. 5) present four main α-helices in nonuniform angular arrangements.

(b) The HD proteins with $n < 60$ have *two* α-helices instead of three. From our perpective, they belong to different fold, and should not have been included in the same category.

(c) The R_g-values for the HD native states with $60 < n < 90$ follow no recognizable pattern in terms of n. These HD-entries resemble spatial accommodations

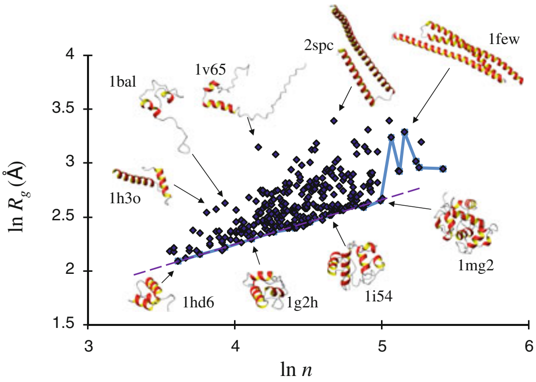

Fig. 6 Radius of gyration (R_g) for the total distribution of all 308 "unique" single-domain single-chain proteins with a three α-helix core, as a function of the number of amino acids, n. The most compact proteins in the range of $40 \leq n \leq 150$ ($\ln n < 5$) yield a unique scaling regime for the ensemble, depicted with the *dashed line*. Note that some of the large-R_g outliers highlighted (e.g., 1h3o, 1bal, 2spc) are noncompact structures classified within the "three-helical folds" by the SCOP database, even though they contain two α-helices

of three α-helical that occupy the *same effective volume* despite having different numbers of residues.

This diversity in compactness is incompatible with a single size-scaling law. It hints to the presence of significant differences in tertiary contacts and, possibly, folding mechanisms.

6 Scaling behaviour in the entire set of three-helix proteins (as classified by SCOP)

In order to put in context the distinct scaling behaviour found within the CCM, IABD, SRD and HD families, Fig. 6 shows the results for the entire distribution of proteins classified by SCOP as belonging to three α-helix protein lineages.

Figure 6 gives the radius of gyration as a function of n for the entire working set of 308 unique single-chain single-domain proteins. This ensemble includes twenty-seven distinct lineages, classified as "three-helical" by the SCOP database. As it is the case in the broader set of all native states [5–7], the natural dispersion in R_g indicates the lack of a single scaling law.

It is clear, however, that a well-defined size-scaling exponent can be derived for the proteins whose α-backbone have the smallest R_g-values within windows of amino acid residues, $(n, n + \Delta)$. We have computed this exponent by using $\Delta = 10$, i.e., finding the smallest R_g value in each of the intervals in the sequence (40,50), (50,60), and so forth. Two distinct scaling behaviour can be observed:

(i) Proteins within the range $40 \leq n < 150$ yield $v_{\text{eff}}^{(40-150)} = 0.414 \pm 0.010$.
(ii) Longer proteins in the ensemble ($150 \leq n \leq 225$) clearly move away from that scaling law, although their number is insufficient to derive a reliable scaling exponent.

As conjectured before, we find that the IABD and CCM lineages form an upper a lower bounds for the entire distribution. In fact, the least-squares fitting of the most compact lineages include several points already seen in the discussion of the CCM protein group.

If we broaden the definition of domain by including single domains "extracted" from multi-domain proteins, then the three-helix lineages may have sufficient data points to distinguish the ν_{eff}-values in (i) and (ii) above. However, this new ensemble is fundamentally different from the present one, and it will be analyzed elsewhere.

7 Further comments and conclusions

The existence of a power law between the total number of monomers and the mean size of a polymer is an indication of a simple principle directing its shape and organization (namely, a dominant interaction between monomers). When applied to protein native states, the occurrence of well-defined size scaling within a group of *related* proteins may indicate that they share common structural domains. Given that small proteins comprise typically a single domain, we propose that scaling behaviour can serve to recognize the presence (or lack) of an identical domain within a family of small-protein folds.

Our approach serves as a complementary criterion in the debate of what constitutes a domain, and what its boundaries are within a protein [23]. Automated recognition algorithms for domains are often inconsistent with each other [24–29], and the non-automated classification schemes rely on subjective, consensual criteria based on visual inspection [13–16,20]. Our results show that some of these classifications are not necessarily compatible with a single regime in compactness as defined by a size scaling law with a physically-meaningful ν-exponent. A classification that considered size-scaling behaviour would split some folding families into subgroups, and reveal deeper similarities among the constituent proteins.

Domains, however defined, are assumed to be spatial blocks within a protein which fold (at least often) independently of each other [31]. The occurrence of a tight scaling law within a "folding family" of small proteins is consistent with having a common folding mechanism; deviations from the scaling law may then hint at a protein that either folds differently or possesses more than one domain. In addition, the present approach may be useful in the context of techniques for *de novo* protein design [32] and folding predictions [33] which rely on the accommodation of individual domains within a multi-domain protein. Currently we are working on extending the present analysis to folding families of β- and α / β proteins, as well as to comparing the scaling behaviour of single- and multi-domain proteins.

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. **50**, 1365 (1989)
4. G.A. Arteca, S. Zhang, Phys. Rev. E **59**, 4209 (1999)
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. G. Buck, Nature **392**, 238 (1998)
9. J. Cantarella, R.B. Kusher, J.M. Sullivan, Nature **392**, 237 (1998)
10. G.A. Arteca, J. Chem. Inf. Comput. Sci. **42**, 326 (2002)
11. V. Grantcharova, E.J. Alm, D. Baker, A.L. Horwich, Curr. Op. Struct. Biol. **11**, 70 (2001)
12. D.N. Ivankov, S.O. Garbuzynskiy, E. Alm, K.W. Plaxco, D. Baker, A.V. Finkelstein, Protein Sci. **12**, 2057 (2003)
13. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
14. A. Andreeva, D. Howorth, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **32**, D226 (2004)
15. A. Andreeva, D. Howorth, J.M. Chandonia, S.E. Brenner, T. Hubbard, C. Chothia, A.G. Murzin, Nucl. Acid Res. **36**, D419 (2008)
16. A. Heger, L. Holm, J. Mol. Biol. **328**, 749 (2003)
17. 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)
18. H.M. Berman, K. Henrick, H. Nakamura, Nature Struct. Biol. **10**, 980 (2003)
19. C. Gong, P. Rogerson, G. Arteca, Trimmed version of the PDB based on different criteria for selecting structurally-redundant entries. Unpublished results, (2007)
20. L. Holm, C. Sander, Proteins **33**, 88 (1998)
21. A.M. Lesk, *Introduction to Protein Architecture* (Oxford UP, Oxford, 2001)
22. G.A. Petsko, D. Ringe, *Protein Structure and Function* (New Science Press, London, 2004)
23. C.P. Ponting, R.R. Russell, Annu. Rev. Biophys. Biomol. Struct. **31**, 45 (2002)
24. S.J. Wodak, J. Janin, Biochemistry **20**, 6544 (1981)
25. M.B. Swindells, Protein Sci. **4**, 103 (1995)
26. M.H. Zehfus, Protein Sci. **6**, 1210 (1997)
27. C.J. Tsai, R. Nussinov, Protein Sci. **6**, 24 (1997)
28. M. Dumontier, R. Yao, H.J. Feldman, C.W. Hogue, J. Mol. Biol. **350**, 1061 (2005)
29. L.S. Wyrwicz, G. Koczyk, L. Rychlewski, D. Plewczynski, J. Phys. Condens. Matter. **19**, 285222(8p) (2007)
30. P.F.N. Faisca, R.C. Ball, J. Chem. Phys. **117**, 8587 (2002)
31. M.Y. Shen, F.P. Davis, A. Sali, Chem. Phys. Lett. **405**, 224 (2005)
32. B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Science **302**, 1364 (2003)
33. A.M. Wollacott, A. Zanghellini, P. Murphy, D. Baker, Protein Sci. **16**, 165 (2007)