Short communication

Structural Similarity Between Native Proteins and Chimera Constructs Obtained by Inverting the Amino Acid Sequence

Oliviero Carugo[1,2]

[1] Department of General Chemistry, University of Pavia, viale Taramelli 12, I-27100 Pavia, Italy

[2] Department of Structural and Computational Biology, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria

** Corresponding author: E-mail: oliviero.carugo@univie.ac.st*

Received: 08-04-2010

Abstract

The analysis of the symmetry of protein three-dimensional structures can be extremely useful in order to understand and classify the protein structural universe. The structures of proteins with back-traced amino acid sequence were modeled and compared to the structures of their native counterparts. Only in a very limited set of cases, the two objects showed a significant level of similarity. These extremely symmetric examples can be of any structural class and of any dimension. The lack of biunique "N to C" and "C to N" symmetry at the structural level mirrors that at the sequence level and we propose to design as a *dlof* symmetry the cases in which a protein structure is similar to its back-traced variant.

Keywords: Bioinformatics, computational structural biology, protein sequence, protein sequence alignment, protein structure, protein structure comparison and protein symmetry

1. Introduction

While various types and levels of symmetry are observed for amino acid sequences,[1,2] the symmetry of protein three-dimensional structures is known to be low,[3] with the exception of homo-oligomeric complexes.[4–6] A simple type of symmetry is the sequence inversion. Although proteins with sequence back-traced from the C-terminus to the N-terminus do not exist in Nature, they have received some attention. It was shown, for example, that their sequences can be used to improve the predictions of secondary structure;[7] the possibility to over-express them was also investigated;[8] and it was postulated that the fold remains unchanged in the case of a three-helix bundle after back-tracing of the sequence.[9] In the present communication, the attention is extended to the general problem of the similarity between the three-dimensional structures of the native protein and its variant obtained by reversing its sequence.

2. Results

The distribution of the 3451 mscore values, which measure the degree of similarity between two structures and were computed as it is described in the Methods, is depicted in Figure 1. It clearly appears that in the largest majority of the cases the mscores are very modest and that only very few of the proteins are similar to their inverted counterparts. The relationship between these mscores and the dimension of the protein, measured by the number of residues (nres), is shown in Figure 2. It can be seen that high mscores are observed more frequently for small proteins and that smaller mscores are usually observed for larger proteins. This can be more clearly seen in Figure 3 where the distributions over six dimension ranges are given. It appears that larger mscores are observed in the set of smaller proteins (for example if nres < 50 residues) and that smaller mscores are observed, on average, for larger proteins (for example when nres > 400 residues).

Figure 4 shows the relationships between the mscores and the percentage of sequence identity (pid), computed by comparing the structures and the sequences of a protein and of its inverted counterpart. While the mscores measure the degree of similarity between the real and the inverted structure the pid measure the degree of sequence similarity. It appears that while the mscores can be very large, though not frequently, and sometime approach their maximal value of 100, the pid values are nearly always very small and close to the values that, on average, indicate that two se-

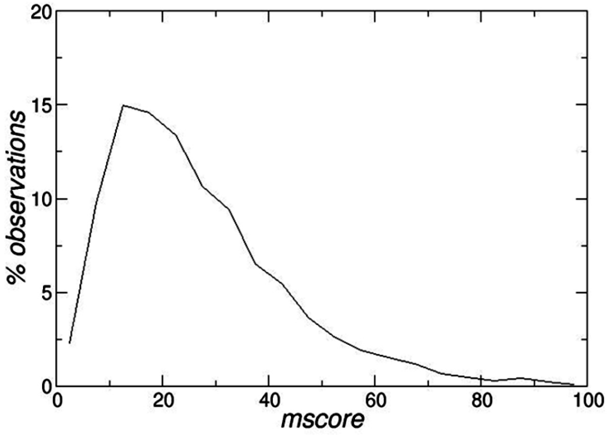

Figure 1. Distribution of the mscores that measure the degree of similarity between a protein and its inverted version.

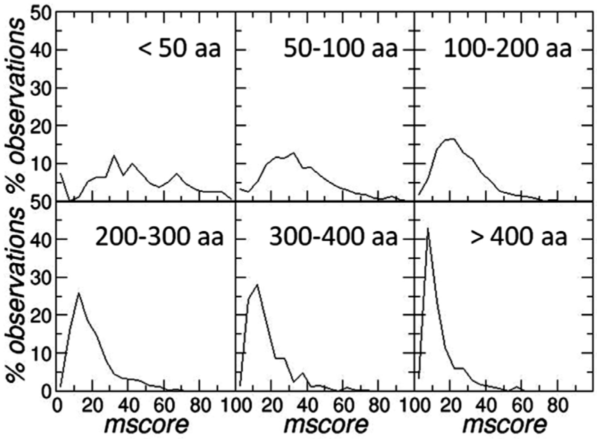

Figure 3. Distribution of the mscores that measure for protein that have different dimensions.

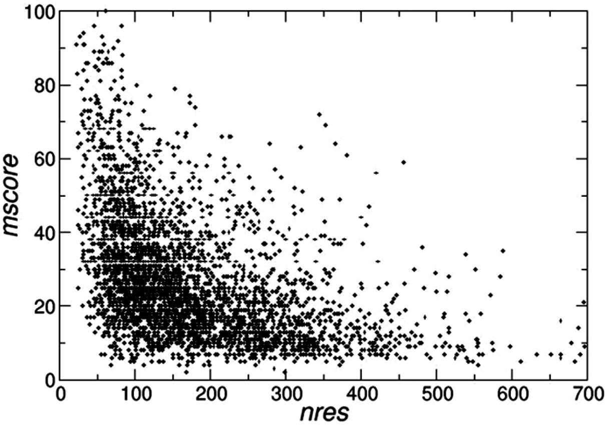

Figure 2. Relationship between the mscores between the real and the inverted protein structure and the number of residues (nres).

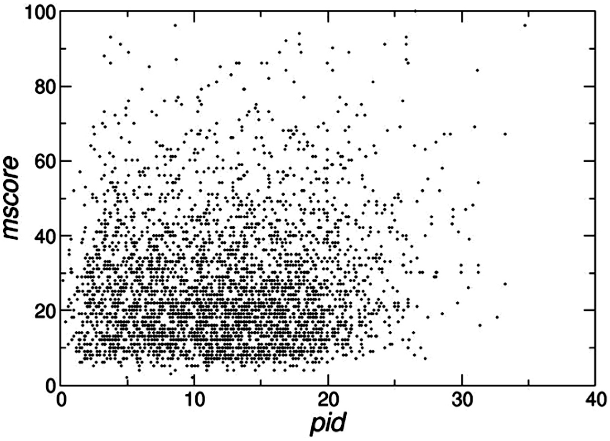

Figure 4. Relationship between the mscores the mscores and the percentage of sequence identity (pid).

quences are completely unrelated to each other.[10] Moreover, though it is possible to determine a positive correlation (Pearson´s correlation coefficient = 0.100) between mscore and pid, implying that the structure similarity increases if the sequence similarity increases, such a correlation is extremely weak and cannot be used for prediction purposes. Less obvious is the observation that inverse sequence similarity is unrelated to fold similarity.[2]

Given that the mscores are, on average, larger for smaller proteins (see Figure 2), the attention was focused on the cases with mscores larger than 90 for proteins smaller than 100 residues, mscores larger than 70 for proteins containing 100–200 residues, mscores larger than 60 for proteins containing 200–400 residues, and mscores larger than 50 for proteins with more than 400 residues. Although these thresholds are absolutely arbitrary, they can allow the identification of an ensemble of proteins that are much more similar to their inverted counterparts than what it is observed on average.

Given that the mscores are, on average, larger for smaller proteins (see Figure 2), the attention was focused on the cases with mscores larger than 90 for proteins smaller than 100 residues, mscores larger than 70 for proteins containing 100–200 residues, mscores larger than 60 for proteins containing 200–400 residues, and mscores larger than 50 for proteins with more than 400 residues. Although these thresholds are absolutely arbitrary, they can allow the identification of an ensemble of proteins that are much more similar to their inverted counterparts than what it is observed on average.

3. Discussion

By using these selection criteria, only 31 protein structures (less than the 1% of the structures examined in the present work) are detected to be quite similar to their inverted counterparts (Figure 5 and Table 1). They can be divided into three groups.

The first includes the protein domains that contain only alpha helices. They have a variable degree of complexity. The simplest are those containing only one helix and cannot be considered to be genuine folds. They are in general moieties of larger proteins, like for example 1dp5b_, which is considered to be a non-globular subunit of globular proteins in the SCOP classification, or 1l2pa_, 1be3k_, 1vf5e_, 1vf5f_ and 1vf5h_, which are single trans-membrane helices. A higher level of complexity is reached by the alpha hairpins, where two anti-parallel helices of comparable length are separated by a short turn. This is observed for example in the SCOP domains 1a36a1, 1b6qa_, 1yzma1, and 2caza1. Slightly more distorted, though not completely different, is the domain 2g38b1, that adopts the ferritin-like fold. Higher levels of complexity are observed when other helices are added to the C-terminus, in such a way that a series of anti-parallel helices separated by short turns is a constant theme. For example, three anti-parallel helices, in some cases quite distorted, are observable in the SCOP domains 1br0a_, 1e2aa_, 1fewa_, 1hx1b_, 1urua_, and 1wdza1. Five helices are present in the domain 1aepa_ and six in 1sumb_. The highest level of complexity is reached in the alpha-alpha super-helices, like the SCOP domains 1hz4a_, 2aw6a2, and 1xm9a1, in which a series of alpha-alpha hairpins follows each other in a large structure that assumes the shape of a large helical ribbon made by two layers of helices.

The second group of domains that are quite similar to their inverted counterparts includes the protein domains that contain only beta strands. All of them are single-stranded beta-helices, in two cases they are left-handed (SCOP domains 1fwya1 and 2f9ca1) and in two cases they are right-handed (SCOP domains 1ezga_ and 2bm4a1). In these domains, the helical ribbon is made by a long series of parallel beta strands separated, in general, by short turns.

The third and last group of domains that are quite similar to their back-traced counterparts includes the protein domains that have both helices and strands. For example, the SCOP domain 1uynx_ contains a N-terminal helix that is followed by a beta-barrel. On the contrary, the domains 1dfji_, 1g9ua_, 1k5dc_, and 1o6sa adopt the

Table 1. Protein domains of the SCOP database which are structurally similar to their inverted versions. The following data are given: the domain identification code within the SCOP database (domain), the fold type according to the SCOP database (fold), the number of residues contained in the domain (nres), the number of residues of thee domain that can be structurally aligned with equivalent residues of the inverted domain by using the program SHEBA (nali), the root-mean-square-distance between the Calpha atoms of the equivalent residues after optimal superposition (rmsd), the similarity score between the real and the inverted structure computed with the program SHEBA (mscore), and the name of the protein (protein).

domain	fold	nres	nali	rmsd	mscore	protein
d1a36a1	a.2.8.1	72	65	1.60	90	Topoisomerase I
d2caza1	a.2.17.1	58	53	1.54	91	Vacuolar protein sortin-associated protein)
d1yzma1	a.2.19.1	46	44	2.14	96	Rabenosyn RAB4 binding domain
d1b6qa_	a.30.1.1	56	52	1.63	93	ROP protein
d1dp5b_	a.137.7.1	31	29	1.53	94	Phospholipase A2
d1ezga_	b.80.2.1	82	79	1.94	96	Antifreeze proteins
d1fwya1	b.81.1.4	77	71	1.98	92	UPD-N-acetylglucosamine pyro.
d1l2pa_	f.23.21.1	61	61	1.06	100	ATP synthase B chain
d1be3k_	f.23.15.1	22	20	0.72	91	Ubiquinol cytochrome C oxid. complex
d1vf5e_	f.23.24.1	32	29	1.43	91	Protein PET L, cytochrome B6F complex
d1vf5f_	f.23.25.1	33	30	1.13	91	Protein PET M, cytochrome B6F complex
d1vf5h_	f.23.27.1	27	25	1.61	93	Protein PET G, cytochrome B6F complex
d1e2aa_	a.7.2.1	102	82	3.26	80	Enzyme IIA
d1fewa_	a.7.4.1	173	130	4.06	75	Activator of caspase
d1hx1b_	a.7.7.1	112	82	1.65	73	BAG-family molecular chaperone regulator
d2g38b1	a.25.4.2	173	134	2.88	77	PPE family protein
d1br0a_	a.47.2.1	120	92	2.85	77	N-terminal domain of syntaxin 1A
d1aepa_	a.63.1.1	153	121	3.14	79	Apolipophorin III
d2bm4a1	b.80.8.1	180	133	2.45	74	Pentapeptide repeat family protein
d1sumb_	a.7.12.1	225	149	1.92	66	PHOU homolog 2
d1urua_	a.238.1.1	215	141	3.05	66	Amphiphysin
d1wdza1	a.238.1.3	227	150	4.54	66	Insulin receptor substrate P53
d1hz4a_	a.118.8.2	366	234	2.87	64	Transcription factor MALT domain III
d2aw6a2	a.118.8.4	218	131	3.44	60	PRGX
d2f9ca1	b.81.1.7	320	202	2.04	63	Hypothetical protein YDCK
d1k5dc_	c.10.1.2	344	246	2.53	72	RAN GTPase activating protein 1
d1o6sa2	c.10.2.1	381	232	2.46	61	Internalin A
d1g9ua_	c.10.2.6	353	242	2.41	69	Outer protein YOPM
d1uynx_	f.4.5.1	279	178	1.92	64	Translocator domain of NALP
d1xm9a1	a.118.1.24	420	235	2.79	56	Plakophilin 1
d1dfji_	c.10.1.1	456	267	2.56	59	Ribonuclease inhibitor

leucine-rich repeat fold, also known as a right-handed beta-alpha super-helix. In this case a series of beta-alpha-beta motifs (two adjacent and parallel beta strands separated by an anti-parallel helix) form a sort of long helical ribbon, consisting of two layers. One of them, internal and close to the helical axis, is made by a parallel beta sheet. The other, more external, is made by a series of parallel helices.

It can also be observed that many of the domains similar to their variants obtained by inverting their sequences belong to the so-called solenoid protein family, where a short structure motif is repeated several times.[11] Interestingly, in Table 1 and Figure 5 there are no β-propellors or α/β barrels, typical examples of solenoid structures. However, this might also depend on the procedure used to build the non-redundant set of protein structures (one random picking/fold; see methods). Moreover, it is also interesting to observe that the inverted sequences had no homologous sequences in UniProt, the most exhaustive protein sequence database.[12] On the one hand this is not surprising, given that Nature used an infinitesimal number of the possible proteins that can be generated with the alphabet of 20 amino acids. On the other side, one must also consider that it is in general difficult to use tools like BLAST or PSI-BLAST with solenoid proteins, where low complexity segments are abundant and residue conservation appears to depend on basic features (like, for example, hydrophobicity) rather than on any specific amino acid type.[11]

It can thus be concluded that the structural similarity between a protein domain and its counterpart obtained by inverting its sequence is extremely modest. This mirrors the lack of similarity between the sequences of the native proteins and of the back-traced variant and might suggest that proteins explored a relatively small fraction of their possible sequences and structures during evolution. However, in very few cases, the structural similarity can be very high, like for example in helix bundles, alpha-alpha super-helices, single-stranded beta helices or in the leucine-rich repeat fold. It is worth nothing that the elegance of this very uncommon symmetry might deserve a name: *dlof* (reverse of fold).

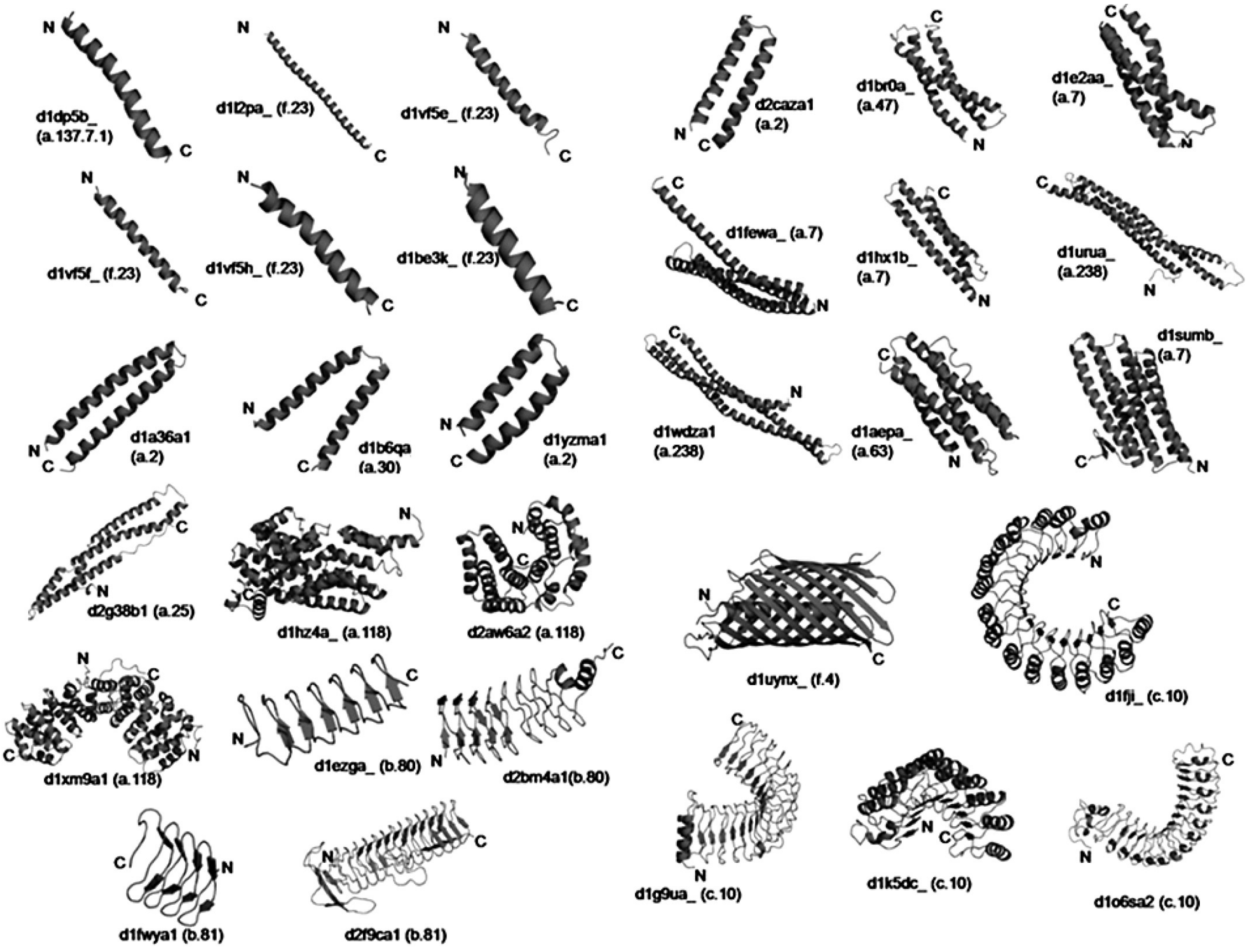

Figure 5. Ribbon view of 31 protein shuchural domains taken from the database SCOP that have a shucture remarkably similar to that of their inverted counterpart.

4. Methods

To reduce cpu usage, the attention was limited to only one representative entry, randomly chosen from each SCOP family (3451 domains).[13] The inverted protein was built as it follows. The residues were renumbered back from the latest to the first and only the Calpha atoms were kept. Each residue was mutated to a glycine and the other backbone atoms were positioned with the program CTRIP of the JACKAL package.[14]

Structural similarities were estimated with the program SHEBA.[15] The structural similarity was measured with the mscore, defined as 100 times the ratio between the number of aligned residues and the total number of residues (which is the same in the real protein and the inverted protein). Sequence alignments were done with the NEEDLE program of the EMBOSS package.[16]

5. Acknowledgements

This work was supported financially by the BIN-III project of the Austrian GEN-AU initiative. The late Dr Tzvicech is gratefully acknowledged for helpful and stimulating discussions.

6. References

1. Sheari, A.; Kargar, M.; Katanforoush, A.; Arab, S.; Sadeghi, M.; Pazashk, H.; Eslahchi, C.; Marashi, S.-A., *BMC Bioinformatics* **2008**, *9*, 274.
2. Preiszner, R.; Goede, A.; Michalski, E.; Frommel, C., *FEBS Lett.* **1997**, *414*, 425–429.
3. Chothia, C., *CIBA Found. Symp.* **1991**, *162*, 36–57.
4. Pinotsis, N.; Willmanns, M., *Cell Mol. Life Sci.* **2008**, *65*, 2953–2956.
5. André, I.; Strauss, C. E. M.; Kaplan, D. B.; Bradley, P.; Baker, D., *Proc. Natl. Acad. Sci. U S A* **2008**, *105*, 16148–16152.
6. Goodsell, D. S.; Olson, A. J., *Annu. Rev. Biophys. Niomol. Struct.* **2000**, 29, 105–153.
7. Park, J.; Dietman, S.; Heger, A.; Holm, L., *Bioinformatics* **2000**, *16*, 978–987.
8. Lacroix, E.; Viguera, A. R.; Serrano, L., *Fold. Des.* **1998**, *3*, 79–85.
9. Olszewski, K. A.; Kolinski, A.; Skolnick, J., *Prot. Eng.* **1996**, *9*, 5–14.
10. Rost, B., *Prot. Eng.* 1999, *12*, 85–94.
11. Marsella, L.; Sirocco, F.; Trovato, A.; Seno, F.; Tosatto, S. C. E., *Bioinformatics* **2009**, *25*, i289–i295.
12. Consortium UniProt, *Nucleic Acids Res.* **2007**, *35*, D193–197.
13. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C., *J. Mol. Biol.* **1995**, *247*, 536–540.
14. Petrey, D.; Xiang, X.; Tang, C. L.; Xie, L.; Gimpelev, M.; Mitros, T.; Soto, C. S.; Goldsmith-Fischman, S.; Kernytsky, A.; Schlessinger, A.; Koh, I. Y.; Alexov E.; Honig, B., *Proteins* **2003**, *53*, 430–435.
15. Jung, J.; Lee, B., *Protein Eng.* **2000**, *13*, 535–543.
16. Rice, P.; Longden, I.; Bleasby, A., *Trends Genet.* **2000**, *16*, 276–7.

Povzetek

Analiza simetrije trodimenzionalnih struktur proteinov je lahko zelo pomembna pri razumevanju in klasificiranju strukturnega prostora. Strukture proteinov z obrnjenim aminokislinskim zaporedjem smo modelirali in jih primerjali z njihovimi nativnimi analogi. Le v redkih primerih sta analoga imela signifikantni nivo podobnosti. Ti zelo simetrični primeri so lahko iz kateregakoli strukturnega razreda in katerekoli dimenzije. Pomankanje edinstvene N do C in C do N simetrije na strukturnem nivoju se kaže tudi na nivoju zaporedja, zato predlagamo načtova »dlof« simetrije v primerih ko sta struktura in struktura z obrnjenim zaporedjem podobni.