Structural and evolutionary relationships among RuBisCOs inferred from their large and small subunits

Fu Xiang, Yuanping Fang, Jun Xiang

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is the key enzyme to assimilate CO_2 into the biosphere. The structural and evolutionary relationships among RuBisCOs were discussed at the domain level. The nonredundant sets for three superfamilies of RuBisCO, i.e. large subunit C-terminal domain (LSC), large subunit N-terminal domain (LSN) and small subunit domain (SS) were defined using QR factorization based on the structural alignment of the RuBisCO domains with QH as the similarity measure, respectively. The results suggest: (1) the core structures of LSC, LSN and SS are well conserved and homologies; (2) the LSC could have occurred naturally in both bacteria and Achaean kingdoms, and the carboxyl-terminal structure evolves increasingly complicated in both bacteria and Eukaryotae kingdoms; (3) the structural variations, such as coil structures at 67-82 positions of LSN and the βA-βB-loop of SS, could make attribution to the CO_2/O_2 specificity of RuBisCO from different species. Such findings provide insights on RuBisCO improvement.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

Structural and Evolutionary Relationships among RuBisCOs Inferred from Their Large and Small Subunits

Fu Xiang[1, 2*], Yuanping Fang[1, 2], Jun Xiang[1, 2]

[1] Hubei Key Laboratory of Economic Forest Germplasm Improvement and Resources Comprehensive Utilization, Huanggang Normal University, Huanggang 438000, P.R. China

[2] Hubei Collaborative Innovation Center for the Characteristic Resources Exploitation of Dabie Mountains, Huanggang 438000, P.R. China

[*] To whom correspondence should be addressed. E-mail: lc_xiangfu@163.com

Short title: Structural Relationships among RuBisCOs

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

Abstract: Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is the key enzyme to assimilate CO_2 into the biosphere. The structural and evolutionary relationships among RuBisCOs were discussed at the domain level. The nonredundant sets for three superfamilies of RuBisCO, i.e. large subunit *C*-terminal domain (LSC), large subunit *N*-terminal domain (LSN) and small subunit domain (SS) were defined using *QR* factorization based on the structural alignment of the RuBisCO domains with Q_H as the similarity measure, respectively. The results suggest: (1) the core structures of LSC, LSN and SS are well conserved and homologies; (2) the LSC could have occurred naturally in both bacteria and Achaean kingdoms, and the carboxyl-terminal structure evolves increasingly complicated in both bacteria and Eukaryotae kingdoms; (3) the structural variations, such as coil structures at 67-82 positions of LSN and the βA-βB-loop of SS, could make attribution to the CO_2/O_2 specificity of RuBisCO from different species. Such findings provide insights on RuBisCO improvement.

Keywords: RuBisCO; protein domain; nonredundant set; structural dendrogram; structural variation

Introduction

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO), the key enzyme to assimilate atmospheric CO_2 into the biosphere, is ubiquitous in phototrophic and chemoautotrophic organisms from the three kingdoms of life. As the name indicates, RuBisCO catalyses both carboxylation and oxygenation of the substrate ribulose 1,5-bisphosphate during photorespiration (Miziorko & Lorimer, 1983).

The comparison of primary sequences demonstrates that the four known forms or types of RuBisCO are placed in a separate category and have been designated the forms I, II, III, and IV (Tabita et al., 2008). Form I protein is the most abundant form of RuBisCO, and consists of eight large subunits and eight small subunits (Schneider, Lindqvist & Branden, 1992). Forms II, III, and IV comprise only large subunits. In fact, the fundamental catalytic unit of all forms of RuBisCO is the large subunit dimer in which the active site is formed from the interface between the *C*-terminal domain of one monomer and the *N*-terminal domain of another monomer. The small subunit can influence the conformation of the catalytic core of large subunit (Tabita et al., 2008).

Primary sequence is less conserved than protein structure (Chothia & Lesk, 1986; Gan et al., 2002). Many deeper evolutionary branches which would be difficult or impossible to determine with sequence data alone can be reconstructed by structural data (O'Donoghue & Luthey-Schulten, 2003, 2005; Sethi, O'Donoghue & Luthey-Schulten, 2005). On the other hand, protein domains evolve at different rates and can even be transposed among organisms such that evolutionary analysis could be done a domain each time. It is the basis for our motivation to discern the structural and evolutionary relationships among RuBisCOs based on

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

the domains in large and small subunits.

The kinetic limitations of RuBisCO with respect to its low carboxylation efficiency and poor CO_2/O_2 specificity aroused broad research to improve RuBisCO's selectivity for CO_2 over O_2 (Mueller-Cajar & Whitney, 2008). Discovering the functional/structural significance of evolution variation can be used to guide the improvement of RuBisCO. In fact, with the rapid increase of protein structures, it is possible to conduct a rather detailed evolutionary analysis to find the structural variations contributed to the CO_2/O_2 specificity of RuBisCO using structural approaches. In this study, the structural diversification and evolutionary relationships among RuBisCO domains will be presented in detail based on nonredundant structural data sets, structural alignment, structural homology measure, and structure-based phylogenetic analysis of RuBisCO domains.

Materials and Methods

Domain and Nonredundant Set

In order to discern the evolutionary course of RuBisCOs, we focus our attention on the protein domains of three superfamilies, i.e. large subunit C-terminal domain (LSC), large subunit N-terminal domain (LSN), and small subunit domain (SS) (see Table 1). Domain definitions were taken from the latest version of the Structural Classification of Proteins-extended, SCOPe 2.04 (Fox, Brenner & Chandonia, 2014). The ATOM and HETATM records corresponding to each SCOPe domain have been collated into PDB-style files which were used as a source of coordinates for domains of RuBisCOs.

Given that LSC, LSN and SS domains are represented by 239, 240, and 118 crystal forms

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

in the PDB respectively, it is necessary to determine a nonredundant set of domains by systematically remove some of the examples. According to the multidimensional *QR* factorization (O'Donoghue & Luthey-Schulten, 2003, 2005), the three-dimensional coordinates of the overlapped domain structures were used to compute the *QR* ordering for RuBisCO domains to define the nonredundant sets contained all organisms and specificities. As shown in Table 1, the nonredundant sets for three superfamilies are composed of 12, 12, and 8 domain structures, respectively, and the crystal structures with high sequence identity, i.e, ~100%, have been omitted.

Structural Alignment

Multiple structural alignments were computed using program STAMP (Russell & Barton, 1992), which aligns the most similar structures firstly, and moves along a dendrogram to add groups of aligned structures to the multiple alignments based on structural similarity. This kind of multiple structural alignments has been implemented in MultiSeq, which is a part of VMD (Roberts et al., 2006).

Sc is defined in STAMP to evaluate the overall alignment quality and structural similarity across a wide range of protein families, and the $Sc < 2.0$ generally indicates little structural similarity (Russell & Barton, 1992; Russell, 2006). All domains in three nonredundant sets were aligned with domains from *Alcaligenes eutrophus* (d1bxna1, d1bxna2, and d1bxni_, see Table 1) as the start, respectively.

Structural Similarity Measure

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

The similarity measures Q_H and Q_{res}, which had been implemented in MultiSeq (Roberts et al., 2006), were employed to measure structural similarity. Q_{res} computes similarity for each residue in a given set of aligned structures. The Q_{res} is a value between 0 and 1 with higher scores high similarity and lower scores low similarity. Q_H measures the similarity for paired protein structures. Q_H value ranges from 0 to 1 and Q_H =1 refers to identical structures.

Phylogenetic Analysis

In order to investigate the structural and evolutionary relationships between the RuBisCO domains, cluster analyses were carried out by UPGMA method in MEGA software (Tamura et al., 2007). UPGMA performs agglomerative clustering based on a pairwise similarity measure which can be represented as a dendrogram. With Q_H as the similarity measure, the distance matrix (see Table S1) required for the UPGMA method is a simply matrix of the pairwise structural dissimilarity values (1 - Q_H). For comparison purposes, the structural dendrograms were built also by neighbor joining (NJ) method based on the distance matrix. Other popular methods for reconstructing dendrograms, such as maximum parsimony, which depend on generating ancestral states of modern sequences have been developed for sequence-based comparisons and could not currently be applicable to structure-based dendrograms (O'Donoghue & Luthey-Schulten, 2003).

Results and Discussion

Structural Similarity of RuBisCOs

The distributions of the structural and sequence similarities of the nonredundant sets for

RuBisCO domain superfamilies were shown in Fig. 1. The domains d1bwvy_, d1rscp_, d1wdds_ from RuBisCO small subunit and all the domains from RuBisCO large subunit have a high structural similarity with *Sc* > 5.5 (see Fig. 1), which suggests a functional and/or evolutional relationship (Russell & Barton, 1992; Russell, 2006).

The distribution of Q_H is in more reasonable agreement with that of *Sc* measure than that of sequence identity, especially for the domains from RuBisCO large subunit in Fig. 1A and B. Meanwhile, both sequence and structural similarities of RuBisCO large subunit are greater than those of small subunit. Such results show that large subunit is more highly conserved than small subunit.

In addition, though the *Sc* and sequence identity values are distinguishable from some structures, the Q_H measure continues to give meaningful information about the similarity of two domain structures. For example, though most of sequence identity values are about 20% with *Sc* < 5.5, all Q_H values are greater than 0.6 in Fig. 1C. The results enhance the standpoint that sequence is less conserved than protein structure (Chothia & Lesk, 1986). Consequently, Q_H was used as a pairwise similarity measure to conduct structure-based evolution analysis of RuBisCO domains.

Structural Alignment

LSC is a large carboxy-terminal domain with stranded parallel αβ barrel structure, and LSN is a small amino-terminal domain containing stranded consecutive β sheet with helices on one side of the sheet. The functional structure of RuBisCO is located at the carboxy-terminal end of the β sheets (Andersson & Backlund, 2008). SS is a small subunit domain consisting of

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

four-stranded mixed β sheet with two helices on one side. SS not only assembles and concentrates the large catalytic subunits of RuBisCO, but also contributes substantially to the differences in kinetic properties among diverse RuBisCOs (Andersson & Backlund, 2008).

The sequence conservations resulting from the structural alignment of the nonredundant sets for RuBisCO domains were shown in Fig. 2. The overall folds of LSC are structural homologies with 115 conservative residues and 28 identical residues. The most significant difference is observed at loops inserted at the corners of the coils, such as those in positions 333-343, 355-369, and 432-445 in domain d1bxna1 (see Fig. S1). In positions 333-343 in d1bxna1, the corresponding residues in domain d1rbaa1 from bacteria are missed completely, while in positions 355-369, both residues in domains d1ykwb1and d2cwxe1 are lacked. In positions 432-445, domains d1geha1, d1rbaa1, d1ykwb1, and d2cwxe1 show similar mission. In other words, the structural variations in LSC are from archaea and bacteria, i.e. *T. kodakaraensis* (d1geha1), *R. rubrum* (d1rbaa1), *C. tepidum* (d1ykwb1), and *P. horikoshii* (d2cwxe1), which can be confirmed also by the sequence conservation in Fig. 2. The unshaded residues in LSC in Fig. 2 were mainly from domains d1geha1, d1rbaa1, d1ykwb1, and d2cwxe1. These structural variations can influence the catalytic properties of RuBisCO. For example, the residue V331 in green alga *C. reinhardtii* (d2v69h1), the counterpart of position 333 in domain d1bxna1, having been replaced by Ala brought about the 37% reduction in the CO_2/O_2 specificity of RuBisCO (Chen & Spreitzer, 1989). Similarly, the replacement of residue L335 in tobacco *N. tabacum* (d3rubl1), the counterpart of position 340 in domain d1bxna1, by Val caused considerable reduction in specificity and lessened sensitivity to inhibitors (Whitney et al., 1999; Pearce & Andrews, 2003).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

The secondary structure of LSN is also conserved among 12 domain structures. There are 43 conservative residues in LSN as shown in Fig. 2, of which 8 residues are identical and marked with *. In positions 67-82 in domain d1bxna2, many corresponding residues were missed or changed in LSN of *T. kodakaraensis* (d1geha2), *R. rubrum* (d1rusb2) and *C. tepidum* (d1ykwb2). Especially, as previously described disordered region of residues 47 to 58 in LSN of *C. tepidum* (d1ykwb2) (Tabita et al., 2007), the residues 63 to 68 and 53 to 65 were found to be disordered in *N. tabacum* (d1rlcl2) and *R. rubrum* (d1rusb2), respectively. Another specific region is positions 94 to 99 in domain d1bxna2. There is a long loop, the so called loop CD, consisting of 16 residues in LSN of *C. tepidum* (d1ykwb2), while a short loop connecting β-strands C and D was found in other *N*-terminal domains (see Fig. S2). The residues of loop CD are involved in multiple interactions close to the active site and hence may be critical for the function of RubisCO-like proteins in vivo (Tabita et al., 2007). In fact, the variant residues in LSN in Fig. 2 were mainly from domains d1geha2, d1rusb2 and d1ykwb2. Therefore, the residues' variations in LSN, like as LSC, are mostly from *T. kodakaraensis* (d1geha2), *R. rubrum* (d1rusb2) and *C. tepidum* (d1ykwb2).

Though the sequence of SS is more diverse than those of LSC and LSN, the core structure of SS is well conserved. As shown in Fig. 2, there are 32 conservative residues in SS, and 13 of which are identical residues. The major variation is located at positions 39-45 in domain d1bxni_, the so-called βA-βB-loop (Andersson & Backlund, 2008). Interestingly, a loop was only observed in the corresponding positions in the domains from eukaryota, namely *C. reinhardtii* (d1gk8o_), *O. sativa* (d1wdds_), *N. tabacum* (d4rubv_), *S. oleracea* (d8rucl_), which contributes to the differences in kinetic properties between eukaryota and bacteria

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

RuBisCOs (see Fig. S3). In fact, the βA-βB-loop of SS has influence on catalytic performance and CO_2/O_2 specificity of RuBisCO (Spreitzer et al., 2001).

There is an interesting relationship between structure conservation and sequence identity from the alignments of the nonredundant sets for RuBisCO domain strucutres. Fig. 3 presents plots of Q_{res}, a measure for structural similarity, and sequence identity per residue averaged over the multiple alignments for LSC (d1bxna1), LSN (d1bxna2), and SS (d1bxni_), respectively. In Fig. 3, the areas of Q_{res} cover those of sequence similarity, and the Q_{res} peaks are wider and more frequent than those of high sequence similarity. In general, sequence conservation follows structure conservation (O'Donoghue & Luthey-Schulten, 2003). However, structure conservation does not always correspond to sequence similarity. For example, the sequence identity of the 5th residues in d1bxna1 is 9.1% with Q_{res} = 0.87; the sequence identity of 69th residue in d1bxna2 is 18.2% with Q_{res} = 0.89, and the sequence identity of 13th residue in d1bxni_ is low to zero with Q_{res} = 0.87 as shown in Fig. 3. Especially for the residues in 95th- 103rd positions in d1bxni_, the sequence similarity per residue is zero, and these residues are still conservative with Q_{res} ranging from 0.60 to 0.89 in Fig. 3C. These results show further that structure is significantly conserved more than sequence. Even though the sequence identity per residue is low to zero, the Q_{res} measure continues to give meaningful information about the structural similarity. It is worth noting that the sequence similarities of the residues in 105th- 127th positions in d1bxni_, as shown in Fig. 3C, are from 0 to 14.27% with Q_{res} ranging from 0.09 to 0.16, which means these residues are both structure and sequence variations. It is due to the two extra β-sheets at the end of domains d1bxni_ and d1bwvy_ (see Fig. S3).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

Structural Phylogeny of RuBisCOs

Protein domains within the same multidomain protein evolve at different rates, which are affected by the protein's contact density (Zhou, Drummond & Wilke, 2008). The structural dendrograms of the nonredundant sets for RuBisCO superfamilies were shown in Fig. 4-6. An unbiased profile of structure conservation at a different level of similarity can be obtained by *QR* algorithm (O'Donoghue & Luthey-Schulten, 2003). According to *QR* algorithm, the first protein domains in the *QR* order represent the major structures, following by the inclusion of domains with similar structures to each representative (O'Donoghue & Luthey-Schulten, 2005). Based on *QR* orders, the three dendrogram topologies are due to structural diversifications rather than species, and the key changes were mapped in contacts among domain structures in Fig. 4-6, respectively.

The structural dendrogram of LSC nonrendundant set can be divided into three clusters, namely d1rbaa1 (*R. rubrum*), C1 and C2 in Fig. 4 (the similar topology can be observed in Fig. S4 A). The cluster C1 is composed of LSC from bacteria *C. tepidum*, and two archaea, *T. kodakaraensis* and *P. horikoshii* (positions 1-3 in the second column of *QR* order). In other words, the common structure of LSC could be from the thermophilic green sulfur bacteria *C. tepidum* (position 1 in the first column of *QR* order), and then from photosynthetic bacteria *R. rubrum* (position 2 in the first column of *QR* order). The LSC carboxyl-terminus, from position 415 in domain d1bxna1 to the end as shown in Fig. 4, is important for the stability and activity of RuBisCO (Gutteridge, Rhoades & Herrmann, 1993; Andersson & Backlund, 2008). It is interesting that cluster C1 and *R. rubrum* display a 4-helix-2-helix-2-helix bundle

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

at the carboxyl-terminus of LSC, while cluster C2 (positions 1-8 in the third column of *QR* order) shows a 5-helix-3-helix-3-helix bundle. The structure variations observed in Fig. 4 throw some light on the evolution of LSC, especially on the structural evolution at the carboxyl terminus of LSC. The carboxyl-terminal structure of LSC could have occurred naturally in both bacteria and archaea kingdoms, and evolves increasingly complicated in both bacteria and eukaryota kingdoms, notably in *G. partita* (d1iwak1), tobacco (d3rubl1), spinach (d1auso1), rice (d1wdda1), and green alga (d2v69h1).

Like as LSC, 12 domains were selected to form the LSN nonrendundant set (see Table 1). The UPGMA dendrogram of LSN can be divided into d1rusb2 (*R. rubrum*), d1ykwb2 (*C. tepidum*), N1, N2 and N3 (see Fig. 5), and the NJ tree showed the similar topology (Fig. S4 B). Furthermore, such clusters are in keeping with positions 1-5 of *QR* order in Fig. 5, and can be contributed to the structural variations, as mentioned in section of structural alignment, at the corresponding region of 67-82 positions in domain d1bxna2. According to the *QR* order in Fig. 5, the common structure of LSN could be from *N. tabacum* (position 1 of *QR* order), and then from photosynthetic bacteria *R. rubrum* and thermophilic green sulfur bacterium *C. tepidum*. So, the evolution of LSN could be different from LSC. The structures at the corresponding region of 67-82 positions in domain d1bxna2 occurred not only in bacteria with a short coil, but also in eukaryota with a long coil. Interestingly, the archaea cluster N1 with a long coil attached 1-helix bundle, and the N3 branch with a long coil mixed 1-helix bundle, could share the ancestral structure with cluster N2 (as shown in Fig. 5). The evolution of the coil structure among different species could be attributed to the ability for assimilating CO_2. Therefore, the coil structures at the corresponding region of 67-82 positions in domain

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

d1bxna2 could show some light on the improvement of RuBisCO.

The 8 domains in SS nonrendundant set are from bacteria and eukaryota kingdoms (see Table 1). The dendrograms drawn by UPGMA and NJ are divided obviously into two clusters as shown in Fig. 6 and Fig. S4 C, and the UPGMA dendrogram gives more information about the structural specificities. According to the *QR* order in Fig. 6, the UPGMA dendrogram can be further divided into three clusters, namely S1, S2, and S3, with *QR* order 1, 2, and 3, respectively. Like as LSN, the three clusters of SS can be contributed to the structural specificities at the counterpart of the positions 39-45 in domain d1bxni_, i.e. the SS βA-βB-loop, which can influence the specificity and stability of RuBisCO (Spreitzer et al., 2001; Andersson & Backlund, 2008). Cluster S1 is a bacteria branch including *H. neapolitanus* and *Synechococcus sp.*, strain pcc 6301 with a short coil mixed helix bundle at the SS βA-βB-loop, while cluster S2 is composed of domains from *A. eutrophus* and *G. partita* with a short pure coil. Cluster S3 belongs to eukaryota kingdom, which includes domains from green alga (d1gk8o_), rice (d1wdds_), tobacco (d4rubv_) and spinach (d8rucl_). Interestingly, cluster S3 displays a long coil at the SS βA-βB-loop. Dissimilarly, a long coil mixed helix bundle occurred in the SS of green alga and a long pure coil in the SS of rice, tobacco and spinach. Such structural variations at the βA-βB-loop region of SS could make attribution to the CO_2/O_2 specificity of RuBisCO from different species, which further provide clues on the improvement of RuBisCO.

Conclusion

The core structures of LSC, LSN and SS are well conserved and homologies respectively.

The structural variations, such as loop residues inserted at the corners of the coils, loop CD and βA-βB-loop, can influence the catalytic properties and CO_2/O_2 specificity of RuBisCO.

The structural dendrogram can be rather easily understood in terms of structural diversification. The LSC could have occurred naturally in both bacteria and archaea kingdoms, and the carboxyl-terminal structure evolves increasingly complicated in both bacteria and eukaryota kingdoms. The structural variations, such as coil structures at 67-82 positions of LSN and the βA-βB-loop of SS, could make attribution to the CO_2/O_2 specificity of RuBisCO from different species, which could show some new light on the improvement of RuBisCO.

Acknowledgment

This work was supported by Hubei Provincial Natural Science Foundation of China (2014CFA129).

Table 1. The nonredundant sets for RuBisCO domain superfamilies

Superfamily	SCOPe sid	Specie source	Classification
RuBisCO, C-terminal domain (LSC)	d1bxna1	*Alcaligenes eutrophus*	Bacteria
	d1ykwb1	*Chlorobium tepidum*	Bacteria
	d1iwak1	*Galdieria partita*	Eukaryota
	d2v69h1	Green alga (*Chlamydomonas reinhardtii*)	Eukaryota
	d1svda1	*Halothiobacillus neapolitanus*	Bacteria
	d2cwxe1	*Pyrococcus horikoshii*	Archaea
	d1rbaa1	*Rhodospirillum rubrum*	Bacteria
	d1wdda1	Rice (*Oryza sativa*)	Eukaryota
	d1auso1	Spinach (*Spinacia oleracea*)	Eukaryota
	d1rbla1	*Synechococcus sp.*, strain pcc 6301	Bacteria
	d1geha1	*Thermococcus kodakaraensis*	Archaea
	d3rubl1	Tobacco (*Nicotiana tabacum*)	Eukaryota
RuBisCO, large subunit, N-terminal domain (LSN)	d1bxna2	*Alcaligenes eutrophus*	Bacteria
	d1ykwb2	*Chlorobium tepidum*	Bacteria
	d1bwvg2	*Galdieria partita*	Eukaryota
	d1uwar2	Green alga (*Chlamydomonas reinhardtii*)	Eukaryota
	d1svda2	*Halothiobacillus neapolitanus*	Bacteria
	d2cwxa2	*Pyrococcus horikoshii*	Archaea
	d1rusb2	*Rhodospirillum rubrum*	Bacteria
	d1wdde2	Rice (*Oryza sativa*)	Eukaryota
	d1aa1l2	Spinach (*Spinacia oleracea*)	Eukaryota
	d1rsch2	*Synechococcus sp.*, strain pcc 6301	Bacteria
	d1geha2	*Thermococcus kodakaraensis*	Archaea
	d1rlcl2	Tobacco (*Nicotiana tabacum*)	Eukaryota
RuBisCO, small subunit (SS)	d1bxni_	*Alcaligenes eutrophus*	Bacteria
	d1bwvy_	*Galdieria partita*	Eukaryota
	d1gk8o_	Green alga (*Chlamydomonas reinhardtii*)	Eukaryota
	d1svdm1	*Halothiobacillus neapolitanus*	Bacteria
	d1wdds_	Rice (*Oryza sativa*)	Eukaryota
	d8rucl_	Spinach (*Spinacia oleracea*)	Eukaryota
	d1rscp_	*Synechococcus sp.*, strain pcc 6301	Bacteria
	d4rubv_	Tobacco (*Nicotiana tabacum*)	Eukaryota

Fig. 1. Distributions of the structural and sequence similarities of the nonredundant sets for RuBisCO domain superfamilies. A is for the large subunit C-terminal domain superfamily, B is for the large subunit N-terminal domain superfamily, and C is for the small subunit superfamily. Both Q_H and sequence identity were amplified tenfold. The distribution of Q_H is reasonablely more consistent with that of *Sc* than that of sequence identity.

Fig. 2. The conservative residues resulting from the structural alignment of the nonredundant sets for RuBisCO domains. The identical residues are marked with *. Numbers above each sequence block indicate the position of each residue in domains d1bxna1, d1bxna2, and d1bxni_, respectively. The letters A, E, and B at the end of each sequence indicate the domain from Archaea, Eukaryota, and Bacteria, respectively. The SCOPe sid is used to identify each sequence. The conserved positions are shaded using GENDOC (Nicholas & Nicholas, 1997) with 90, 75, and 50% conserved, using PAM250 as scoring table.

Fig. 3. Conservation of structure and sequence averaged over the multiply aligned nonredundant set as a function of position in the domains for d1bxna1, d1bxna2, and d1bxni_, respectively. Residue index is related to PDB numbering. A is for the large subunit *C*-terminal domain superfamily, B is for the large subunit *N*-terminal domain superfamily, and C is for the small subunit superfamily.

Fig. 4. The structural dendrogram of LSC nonrendundant set (drawn with UPGMA method

in MEGA software). The ordering according to the *QR* transformation and the SCOPe sid are listed. Also listed are the structure variations at carboxyl-terminal end of LSC.

Fig. 5. The structural dendrogram of LSN nonrendundant set (drawn with UPGMA method in MEGA software). The ordering according to the *QR* transformation and the SCOPe sid are listed. Also listed are the coil variations at the corresponding region of 67-82 positions in domain d1bxna2.

Fig. 6. The structural dendrogram of SS nonrendundant set (drawn with UPGMA method in MEGA software). The ordering according to the *QR* transformation and the SCOPe sid are listed. Also listed are the structural variations at the βA-βB-loop region of SS.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

References

Andersson I, Backlund A. 2008. Structure and function of Rubisco. *Plant Physiology and Biochemistry* 46: 275-291. DOI: 10.1016/j.plaphy.2008.01.001.

Chen ZX, Spreitzer RJ. 1989. Chloroplast intragenic suppression enhances the low CO2/O2 specificity of mutant ribulose-bisphosphate carboxylase/oxygenase. *Journal of Biological Chemistry* 264: 3051-3053.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5: 823-826.

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucl Acids Res* 42: 304-309. DOI: 10.1093/nar/gkt1240.

Gan HH, Perlow RA, Roy S, Ko J, Wu M, EA. 2002. Analysis of protein sequence/structure similarity relationships. *Biophys. J.* 83: 2781-2791. DOI: 10.1016/S0006-3495(02)75287-9.

Gutteridge S, Rhoades DF, Herrmann C. 1993. Site-specific mutations in a loop region of the C-terminal domain of the large subunit of ribulose bisphosphate carboxylase/oxygenase that influence substrate partitioning. *Journal of Biological Chemistry* 268: 7818-7824.

Miziorko HM, Lorimer GH. 1983. Ribulose-1,5-bisphosphate carboxylase/oxygenase. *Ann Rev Biochem* 52: 507-535.

Mueller-Cajar O, Whitney SM. 2008. Directing the evolution of Rubisco and Rubisco activase: first impressions of a new tool for photosynthesis research. *Photosynthesis Research* 98: 667-675. DOI: 10.1007/s11120-008-9324-z.

Nicholas KB, Nicholas HBJ. (1997). GeneDoc: A tool for editing and annotating multiple sequence alignments. Distributed by the authors.

O'Donoghue P, Luthey-Schulten Z. 2003. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiology and Molecular Biology Reviews* 67: 550-573. DOI: 10.1128/MMBR.67.4.550-573.2003.

O'Donoghue P, Luthey-Schulten Z. 2005. Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. *Journal of Molecular Biology* 346: 875-894. DOI: 10.1016/j.jmb.2004.11.053.

Pearce FG, Andrews TJ. 2003. The relationship between side reactions and slow inhibition of ribulose-bisphosphate carboxylase revealed by a loop 6 mutant of the tobacco enzyme. *Journal of Biological Chemistry* 278: 32526-32536. DOI: 10.1074/jbc.M305493200.

Roberts E, Eargle J, Wright D, Luthey-Schulten Z. 2006. MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics* 7: 382. DOI: 10.1186/1471-2105-7-382.

Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* 14: 309-323.

Russell RB. (2006). STAMP Structural Alignment of Multiple Proteins Version 4.3 User Guide.

Schneider G, Lindqvist Y, Branden CI. 1992. RUBISCO: structure and mechanism. *Annual Review of Biophysical and Biomolecular Structure* 21: 119-143.

Sethi A, O'Donoghue P, Luthey-Schulten Z. 2005. Evolutionary profiles from the QR factorization of multiple sequence alignments. *Proc Natl Acad Sci USA* 102: 4045-4050. DOI: 10.1073/pnas.0409715102.

Spreitzer RJ, Esquivel MG, Du Y, McLaughlin PD. 2001. Alanine-scanning mutagenesis of the small-subunit βA-βB loop of chloroplast ribulose-1, 5-bisphosphate carboxylase/oxygenase: Substitution at Arg-71 affects thermal stability and CO2/O2 specificity. *Biochemistry* 40: 5615-5621. DOI: 10.1021/bi002943e.

Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. 2007. Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiology and Molecular Biology Reviews* 71: 576-599. DOI: 10.1128/MMBR.00015-07.

Tabita FR, Satagopan S, Hanson TE, Kreel NE, Scott SS. 2008. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *Journal of Experimental Botany* 59: 1515-1524. DOI: 10.1093/jxb/erm361.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24: 1596-1599. DOI: 10.1093/molbev/msm092.

Whitney SM, von Caemmerer S, Hudson GS, Andrews TJ. 1999. Directed mutation of the Rubisco large subunit of tobacco influences photorespiration and growth. *Plant Physiology* 121: 579-588. DOI: 10.1104/pp.121.2.579.

Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *Journal of Molecular Evolution* 66: 395-404. DOI: 10.1007/s00239-008-9094-4.

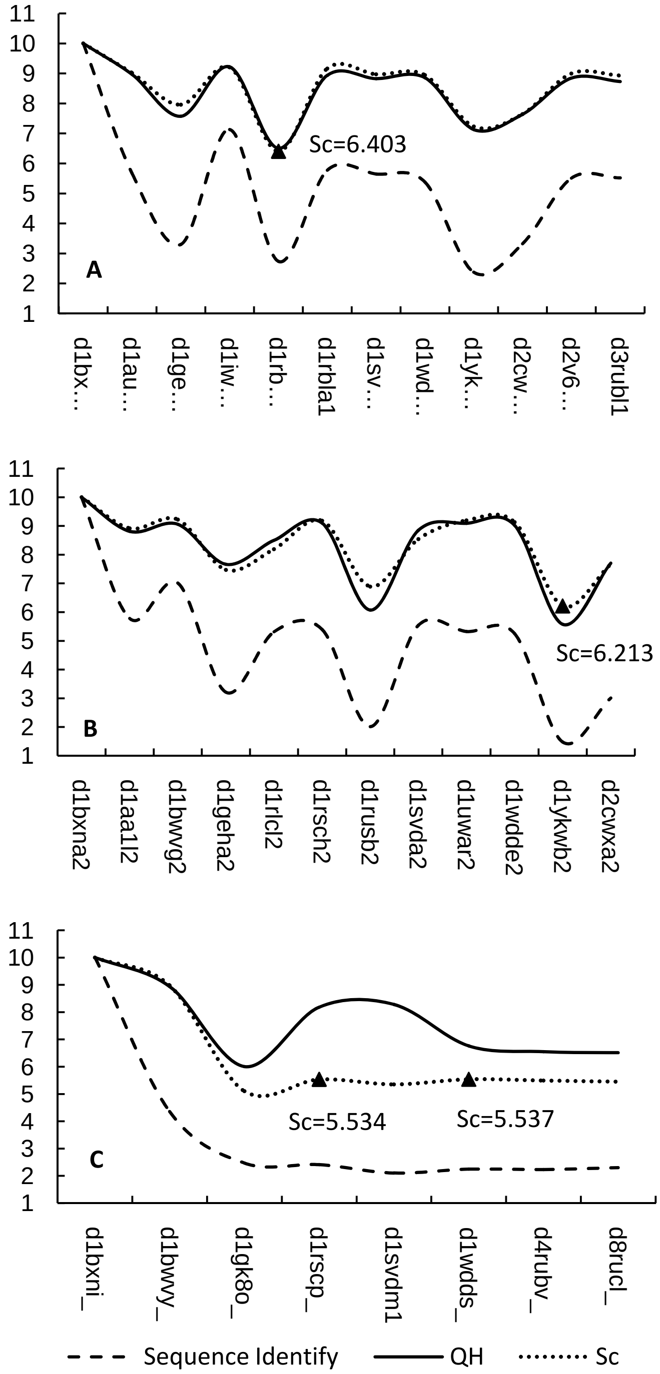

Fig. 1. Distributions of the structural and sequence similarities of the nonredundant sets for RuBisCO domain superfamilies. A is for the large subunit C-terminal domain superfamily, B is for the large subunit N-terminal domain superfamily, and C is for the small subunit superfamily. Both Q_H and sequence identity were amplified tenfold. The distribution of Q_H is reasonablely more consistent with that of *Sc* than that of sequence identity.

```
           153 158  169 174  180 183  193 198  201 205  208 213  218 230  235 239  247 256  264 275
            154 162  170 178  181 184  194 199  202 206  211 214  220 233  236 244  252 257  269
             157 165  171 179  182 188  196 200  204 207  212 216  227 234  237 246  253 260  271
LSC
                    **          * *    *            ******         *
d1bxna1 : GPGIRLGRPGKPKLGLSYYELGGLDFKDDENSQPFHRRAKAATGEKNTAEMRAAGMDG : B
d1auso1 : GPGIRLGRPGKPKLGLSYYELGGLDFKDDENSQPFRRRAKAETGEKNTADMRAAGMDG : E
d1geha1 : GPGIRLDRPGKPKVGYSFYDLNGADYKDDENSPWYRERIKNETGEKNTAEMRLLGMDG : A
d1iwak1 : GPGVRLGRPGKPKLGLSYYELGGLDFKDDENSQPFRRRAKAATGEKNTAEMRAAGMDG : E
d1rbaa1 : GPNIWLDGLGKPKLGLRFHAWGG-DFKNNEPNQPFPRTARDETGEKNTAEIRGVELDG : B
d1rbla1 : GPGIRLGRPGKPKLGLSYYELGGLDFKDDENSQPFRRRAKAETGEKNTAEMRAAGMDG : B
d1svda1 : GPGIRMGRPGKPKLGLSYYELGGLDFKDDENSQPFRRRATAQTGEKNTAEMRAAGMDG : B
d1wdda1 : -PGIRLGRPGKPKLGLSYYELGGLDFXDDENSQPFRRRAKAETGEKNTAEMRAAGMDG : E
d1ykwb1 : GPGIRLGRPFKP-N-LSFYQWGGLDIKDDEMDVTWSERAKAETGEKNTDSLKHAGLNG : B
d2cwxe1 : GPGVRMDRPAKPKMGWSYYEWGGIDLKDDENSFPFRERVRAETGEKNTGIMRAVGMDG : A
d2v69h1 : GXGIRLGRGGKPKLGLSYYELGGLDFXDDENSQPFRRRAKAETGEKNTAEMRAAGMDG : E
d3rubl1 : GPGIRLGRPGKPKLGLSYYELGGLDFKDDENSQPFRRRAKAETGEKNTAEMRAAGMDG : E

           278 294  298 305  315 324  331 370  378 381  385 397  402 406  410 414  421 427  456
            288 296  300 309  318 326  332 372  379 382  389 399  403 407  411 418  423 428  457
             293 297  301 310  321 329  368 376  380 383  392 401  405 408  412 419  425 436  460
              *      * ***            *         *   ** *** ** *  *
d1bxna1 : CQLHHRAHGRHGVKRGDHGTQWSVPVASGGHMLGDVLQGGGTGHPGGAARAEAGPLA : B
d1auso1 : ADLHHRAHARHGVKRGDHGTQWSVPVASGGHMLGDVLQGGGTGHPGGAARAEAGELA : E
d1geha1 : ADIHHRAHARHGVKRGDHGTQFSAPTSSGGHIVGDVLQGGGTGHPGGAARADAGELA : A
d1iwak1 : ADLHHRANSRHGVKRGDHGTMWSVPVASGGHMLGDVLQGGGTGHPGGAARAEANALA : E
d1rbaa1 : A-LHHRAHGSRGVKRGSHGTQWGCPIISGGNMFNNILTGGGAGHIGGASRAQAGELA : B
d1rbla1 : ADLHHRAHARHGVKRGDHGTQWSVPVASGGHMLGDVLQGGGTGHPGGAARAEAGELA : B
d1svda1 : ADLHHRAHARHGVKRGDHGTQWSVAVASGGHMLGDVLQGGGTGHPGGAARAEAGELA : B
d1wdda1 : ADLHHRAHARHGVKRGDHGTQWSVPVASGGHMLGDVLQGGGTGHPGGAARAEAGELA : E
d1ykwb1 : A-LIHFPIARYGVKRGDIPG-MRCPVPGGSSLVNDGFVGRGVGHPGGASRAEAGELM : B
d2cwxe1 : ADIHHRAHARHGAKRGDHGT-WHVPVASGGHMLGDVIQGGGVGHPGGAARADAGELS : A
d2v69h1 : ADLHHRAHARHGVKRGDHGTQWSVPVASGGHMLGDCLQGGGTGHPGGAARAEAGELA : E
d3rubl1 : ADLHHRAHARHGVKRGDHGTQWSVPVASGGHMLGDVLQGGMTGHPGGAARAEAGELA : E

           27  40  47  55  59  64  88  107  112  115  125  128  136  139  147
            32  41  50  56  61  65  105  110  113  118  126  129  137  140
             38  43  51  58  63  66  106  111  114  122  127  131  138  144
LSN
                 *      * *  *   *      *         **
d1bxna2 : YYDLAFPGVEAAAAESSTYAYDLFEEGSNSGNVFSKARLEDPY : B
d1aa1l2 : YYDLAFPGVEAAAAESSTYAYPLFEEGSNSGNVFGKLRLEDPY : E
d1bwvg2 : YYDLAFPGVEAAAAESSTYAYELFEEGSNSGNVFGKLRLEDPY : E
d1geha2 : -YDIAFPGYQAGAAESSTYAYPAFEEANGSGNIFGKLRLEDPL : A
d1rlcl2 : YYDLAFPGVEAAAAESSTYAYPLFEEGSNSGNVFGKLRLEDPY : E
d1rsch2 : YYDLAFPGVEAAAAESSTYAYPLFEEGSNSGNVFGKLRLEDPL : B
d1rusb2 : YLHLCYPGYATAHAESSTYAYPLFDRKMSLGNNQGGAKMHDPY : B
d1svda2 : YYDLAFPGVEAAAAESSTYAYPLFEEGSNSGNVFGKLRLEDPY : B
d1uwar2 : YYDLAFPGVECAAAESSTYAYXLFEEGSNSGNVFGKLRLEDPY : E
d1wdde2 : YYDLAFPGVEAAAAESSTYAYPLFEEGSNSGNVFGKLRLEDPY : E
d1ykwb2 : FRYVLYSG-TAAHCEQSTIAHPNF-GPKNAGEGYFPVKLMDPY : B
d2cwxa2 : FYEIVYPGVEAGRAESSIFAYPLFEEGSQAGNVFGKLRLLDPY : A

           6  10  14  22  33  48  52  56  65  81  96
            7  11  17  25  36  50  53  60  67  85  101
             9  12  18  31  47  51  54  64  76  86
SS
           ** **    *  ***   ***  **
d1bxni_ : GTSFLPLEQQYWVEYWMFGLPFDILEPRFDMN : B
d1bwvy_ : GTSFLPLEQQYLIEYWIWGLPFDVLESKFSFP : E
d1gk8o_ : ETSYLPLEQQYWPEYWMWKLPFDVLEPRFDFP : E
d1rscp_ : ETSYLPLRQQYFPEYWMWKLPFSVLEGRFDFP : B
d1svdm1 : ETSYLPMERQYWPEYWMWKLPFNVLEPKYDFG : B
d1wdds_ : ETSYLPLEDQYWPEYWMWKLPFDVLEPRFDFP : E
d4rubv_ : ETSYLPLEQEYWPEYWMWKLPFDVLEPRFDFP : E
d8rucl_ : ETSYLPLDQQYWPEYWMWKLPFDVLEPRFDFP : E
```

Fig. 2. The conservative residues resulting from the structural alignment of the nonredundant sets for

RuBisCO domains. The identical residues are marked with *. Numbers above each sequence block indicate the position of each residue in domains d1bxna1, d1bxna2, and d1bxni_, respectively. The letters A, E, and B at the end of each sequence indicate the domain from Archaea, Eukaryota, and Bacteria, respectively. The SCOPe sid is used to identify each sequence. The conserved positions are shaded using GENDOC (Nicholas and Nicholas, 1997) with 90, 75, and 50% conserved, using PAM250 as scoring table.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1525v1 | CC-BY 4.0 Open Access | rec: 24 Nov 2015, publ: 24 Nov 2015

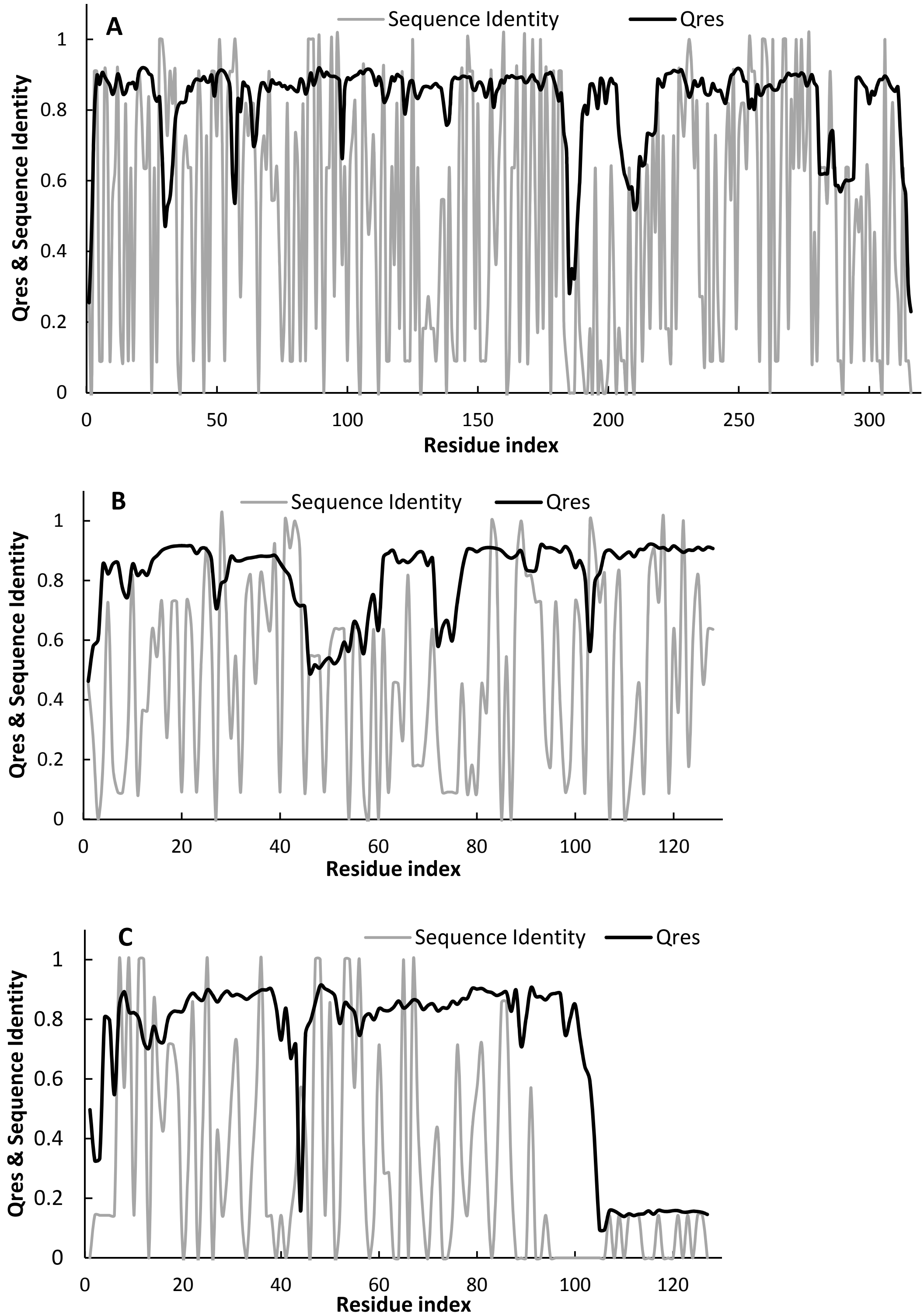

Fig. 3. Conservation of structure and sequence averaged over the multiply aligned nonredundant set as a function of position in the domains for d1bxna1, d1bxna2, and d1bxni_, respectively. Residue index is related to PDB numbering. A is for the large subunit *C*-terminal domain superfamily, B is for the large subunit *N*-terminal domain superfamily, and C is for the small subunit superfamily.

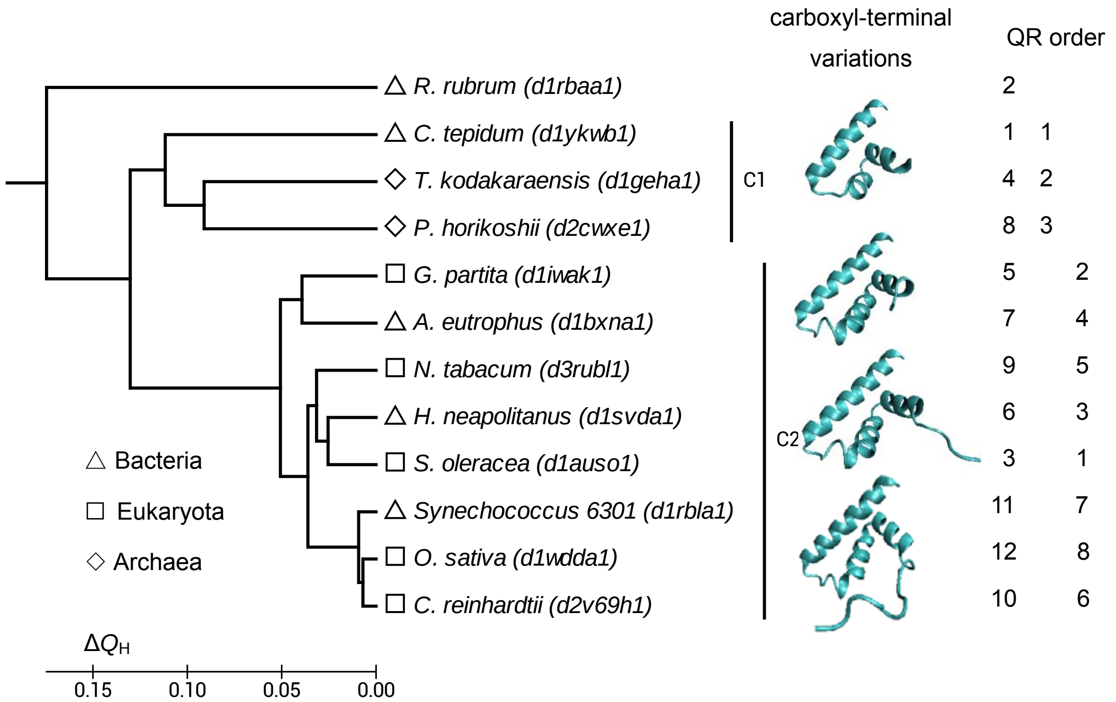

Fig. 4. The structural dendrogram of LSC nonrendundant set (drawn with UPGMA method in MEGA software). The ordering according to the QR transformation and the SCOPe sid are listed. Also listed are the structure variations at carboxyl-terminal end of LSC.

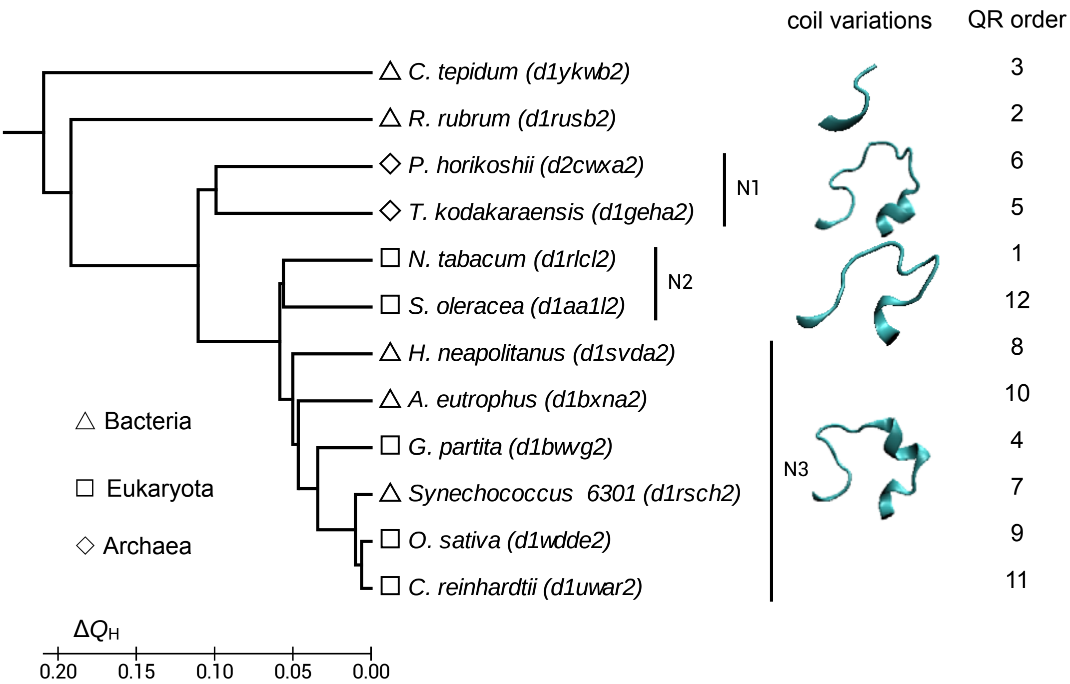

Fig. 5. The structural dendrogram of LSN nonrendundant set (drawn with UPGMA method in MEGA software). The ordering according to the QR transformation and the SCOPe sid are listed. Also listed are the coil variations at the corresponding region of 67-82 positions in domain d1bxna2.

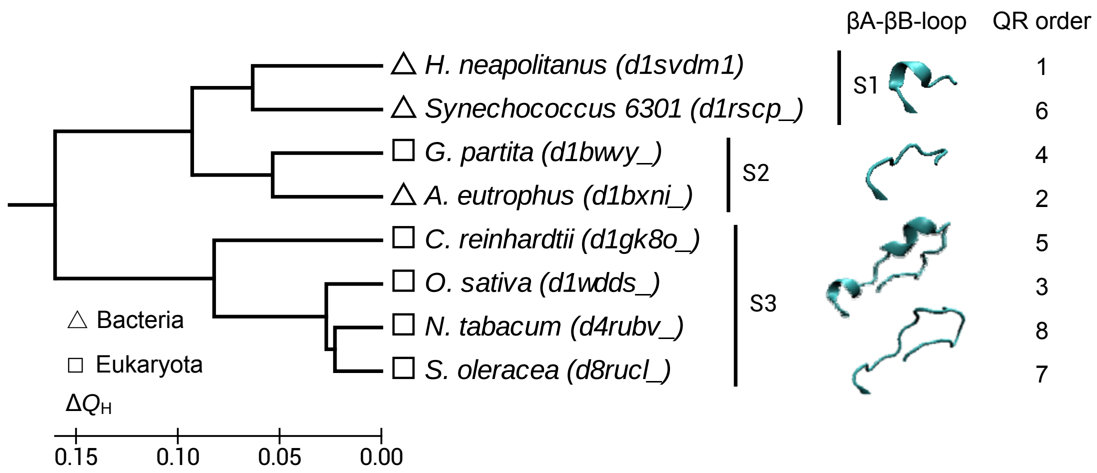

Fig. 6. The structural dendrogram of SS nonrendundant set (drawn with UPGMA method in MEGA software). The ordering according to the QR transformation and the SCOPe sid are listed. Also listed are the structural variations at the βA-βB-loop region of SS.