Sequence Determinants of Quaternary Structure in Lumazine Synthase

María Silvina Fornasari,[1] Diego A. Laplagne,† Nicolás Frankel,‡ Ana A. Cauerhff,† Fernando A. Goldbaum,† and Julián Echave*[1]

*Universidad Nacional de Quilmes, Bernal, Argentina; †Fundación Instituto Leloir (IIBBA-CONICET, IIB-FCEN-UBA), Buenos Aires, Argentina; ‡Laboratorio de Fisiología y Biología Molecular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

Riboflavin, an essential cofactor for all organisms, is biosynthesized in plants, fungi and microorganisms. The penultimate step in the pathway is catalyzed by the enzyme lumazine synthase. One of the most distinctive characteristics of this enzyme is that it is found in different species in two different quaternary structures, pentameric and icosahedral, built from practically the same structural monomeric unit. In fact, the icosahedral structure is best described as a capsid of twelve pentamers. Despite this noticeable difference, the active sites are virtually identical in all structurally studied members. Furthermore, the main regions involved in the catalysis are located at the interface between adjacent subunits in the pentamer. Thus, the two quaternary forms of the enzyme must meet similar structural requirements to achieve their function, but, at the same time, they should differ in the sequence traits responsible for the different quaternary structures observed.

Here, we present a combined analysis that includes sequence-structure and evolutionary studies to find the sequence determinants of the different quaternary assemblies of this enzyme. A data set containing 86 sequences of the lumazine synthase family was recovered by sequence similarity searches. Seven of them had resolved three-dimensional structures. A subsequent phylogenetic reconstruction by maximum parsimony (MP) allowed division of the total set into two clusters in accord with their quaternary structure. The comparison between the patterns of three-dimensional contacts derived from the known three-dimensional structures and variation in sequence conservation revealed a significant shift in structural constraints of certain positions. Also, to explore the changes in functional constraints between the two groups, site-specific evolutionary rate shifts were analyzed.

We found that the positions involved in icosahedral contacts suffer a larger increase in constraints than the rest. We found eight sequence sites that would be the most important icosahedral sequence determinants. We discuss our results and compare them with previous work. These findings should contribute to refinement of the current structural data, to the design of assays that explore the role of these positions, to the structural characterization of new sequences, and to initiation of a study of the underlying evolutionary mechanisms.

Introduction

The enzymes lumazine synthase (LS) and riboflavin synthase (RS) also known as beta and alpha riboflavin synthases, consecutively catalyze the two last steps of the riboflavin (vitamin B_2) biosynthetic pathway. Animals are not able to synthesize riboflavin and, as has been proposed, some microorganisms cannot incorporate exogenous riboflavin (Kearney et al. 1979; Gerhardt et al. 2002). For these reasons, the two enzymes have been extensively studied, particularly as targets of chemotherapeutic agents (Bacher et al. 1996) and, in the case of Brucella spp. LS, as an antigen for serological diagnosis and for the design of vaccines against brucellosis (Goldbaum et al. 1999).

The three-dimensional structures of riboflavin synthase from *Escherichia coli* (Liao et al. 2001) and of lumazine synthases from different organisms have been determined. The structural analyses performed on *Bacillus subtilis* (Bacher et al. 1980; Ladenstein, Ludwig, and Bacher 1983; Bacher 1986; Ladenstein et al. 1988; Schott et al. 1990; Ladenstein et al. 1994; Ritsert et al. 1995) have revealed that two enzymes form a 1-Mda bifunctional complex, consisting of a homotrimer of alpha chains (α3) enclosed in a capsid made of 60 identical beta chains (β60). The β60 capsid is in fact arranged as dodecamer of pentamers giving rise to an icosahedral quaternary structure (fig. 1*a*). This particular enzyme association exhibits enhanced kinetic properties because of substrate channeling (Kis and Bacher 1995; Bacher et al. 2001; Huang, Holden, and Raushel 2001), and it has been proposed that the stabilization of de β60 capsid is mediated by ligand (Ladenstein et al. 1988). Besides *Bacillus subtilis*, lumazine synthases from *Escherichia coli* (Mörtl et al. 1996), spinach (Jordan et al. 1999; Persson et al. 1999), and *Aquifex aeolicus* (Zhang et al. 2001) also present the icosahedral capsid. However, they are presumably not complexed at all with the alpha trimeric chain, as it has been demonstrated in *E. coli* (Mörtl et al. 1996). In contrast, the LSs from the fungi *Magnaporthe grisea* (Persson et al. 1999), *Saccharomyces cerevisiae* (Meining et al. 2000), and *Schizosaccharomyces pombe* (Gerhardt et al. 2002), as well as those from the bacterium *Brucella abortus* (Baldi et al. 2000), display a different quaternary structure: a pentameric arrangement of five monomers (fig. 1*b*).

In spite of the different quaternary structures, LS forms rest on the same monomeric structural unit (fig. 1*c*). The fold of the lumazine synthase monomer consists of a central repetition of four β-α motifs in an arrangement of four parallel β-strands (β3β2β4β5) flanked by the helices α1, α4, and α5 on one side and the helices α2 and α3 on the other (Ladenstein et al. 1988). The fifth β1 strand, located in the N-terminal region, is not always present because certain secondary structure-disturbing residues are also present in this region. The root-mean-square deviation

[1] Address for correspondence and reprints: Universidad Nacional de Quilmes, R. Saénz Peña 180, B1876 BXD, Bernal, Argentina.

Key words: lumazine synthase, quaternary structure, structural constraints, evolutionary rates.

E-mail: silvina@unq.edu.ar.; je@unq.edu.ar.

Mol. Biol. Evol. 21(1):97–107. 2004
DOI: 10.1093/molbev/msg244
Molecular Biology and Evolution vol. 21 no. 1

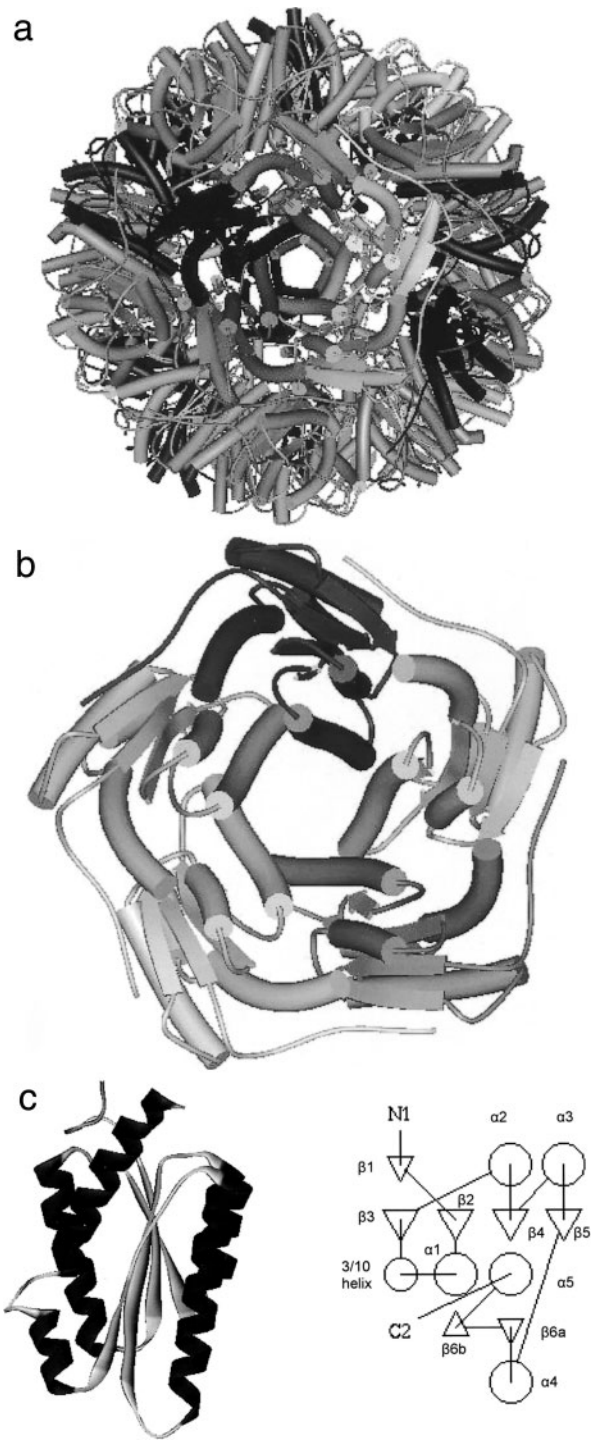

FIG. 1.—Lumazine synthase of *B. subtilis*. (*a*) Schematic representation of the icosahedral form. (*b*) Schematic representation of the pentameric form. α-helices are represented with cylinders and β-sheets with arrows. (*c*) Monomeric form and a topological representation of the secondary elements generated with TOPS (Westhead, Hatton, and Thornton 1998).

(r.m.s.d.) values derived from the structural comparison of the 142 α-carbon atoms in the monomers reveal their structural equivalence with values in the range of 0.67 to 1.01 Å (Gerhardt et al. 2002). Thus, all the lumazine synthases structurally characterized belong to the Lumazine Synthase fold superfamily in the SCOP database (Murzin et al. 1995).

The mechanism of the LS catalyzed reaction has been extensively studied, and the structural information contributed to the elucidation of the residues involved in the active site (Kis, Volk, and Bacher 1995; Ritsert et al. 1995; Bacher et al. 1996; Persson et al. 1999; Meining et al. 2000; Gerhardt et al. 2002). In all the lumazine synthases structurally characterized so far, the active site is located at the interface between two monomers in the pentamers and its topology is well preserved. This preservation reflects the constraints that function imposes onto the conservation of the fold and explains the equivalence observed in pentamer structures and in the active site topology. The sequence analysis should also support the structural patterns associated.

In this article we examine the sequence determinants responsible for the icosahedral quaternary structure of some lumazine synthases. For that purpose, a combined analysis including sequence, structural, and evolutionary information was performed. To recover all the putative homologous sequences included in current databases, sensitive sequence similarity searches were initially done. It is well known that sequence conservation observed in remote homologs with similar folds derives from structural and functional constraints. As very divergent sequences are evaluated, the study of the sequence patterns is facilitated, because sequence conservation from the common origin becomes negligible. Subsequently, different criteria were used to characterize this set of homologous sequences. First, the phylogenetic relationships between the sequences recovered were analyzed. Second, taking into account the location of known-structure representatives of the family, the phylogenetic tree was divided into two clusters, one containing the icosahedral forms and the other the non-icosahedral ones. Third, to find the sequence pattern behind the icosahedral quaternary structure, the sequence conservation in each group was explored in the context of structural parameters. Structural differences between the pentameric and icosahedral forms were characterized using the number of contacts per position as inferred from distance analysis between residues in each structure. To analyze variations in sequence conservation, a reduced entropy per position, a factor used extensively in sequence variability studies, was calculated and compared between the pentameric and icosahedral proteins (Shenkin, Erman, and Mastrandrea 1991; Atchley, Terhalle, and Dress 1999; Ptitsyn 1999; Larson et al. 2002). Finally, to determine whether the changes in quaternary structure are related to changes in the evolutionary constraints of particular residues, site-specific rate shifts between the two clusters mentioned above were estimated. It has been established that if the function or the structure of a protein changes, the evolutionary rates of the sites involved will be different in different parts of the phylogenetic tree (Casari, Sander, and Valencia 1995; Gu 1999; Landgraf, Fischer, and Eisenberg 1999). Also, it has been observed that in some cases only a few residues are involved in this type of divergence (Golding and Dean 1998). Thus, the sequence profile and the evolutionary changes associated with

a homologous alignment reflect the constraints that modulate sequence divergence. Such constraints can be related to structural or functional rate shifts and should be distinguishable from the merely neutral sequence variation (Gu 1999).

We found that sequence positions involved in icosahedral contacts are significantly more conserved in icosahedral proteins than in non-icosahedral ones, as compared to the rest of the sites. Even though such icosahedral sites are somewhat more conserved than non-icosahedral sites in icosahedral proteins, the main reason behind the increase in the degree of conservation is that these positions are significantly more variable than the rest in non-icosahedral proteins. Furthermore, some of these positions were found to have suffered rate shifts as a result of structural or functional constraint changes. The current structural analyses have postulated sequence positions responsible for icosahedron formation. We found that this oligomeric form is associated with a specific sequence signal. Although the functional role of the different quaternary structures is not yet well understood, the identification of specific sequence traits could contribute to initiate the analysis of the evolutionary origin of this unusual enzyme.

Materials and Methods

Sequence Similarity Search

To detect putative homologous proteins constituting the lumazine synthase family, searches in the non-redundant (nr) database using the National Center for Biotechnology Information (NCBI) server located at http://www.ncbi.nlm.nih.gov were performed using PSI-BLAST (Altschul et al. 1997) with default parameters (BLOSUM62 matrix, with an E-value threshold 0.005). The searches were initiated with a single query sequence and iterated until convergence. The sequences found were retrieved using Batch Entrez at the same location.

Sequence Alignments

Multiple-sequence alignments were built with the retrieved sequences using ClustalX (Thompson et al. 1997) under default parameter settings. From these sequences, those with known crystallographic structure were structurally aligned using the program MAPS (http://bioinfo1.mbfys.lu.se/TOP/maps.html).

Protein Structure Analysis

Protein structure similarity was explored using the information in the structure based database SCOP (Murzin et al. 1995). In those cases where the Protein Data Bank (PDB) file did not contain the native structure conformation, the quaternary protein structures were obtained from Protein Quaternary Structural Query (Henrick and Thornton 1998). The number of contacts for each residue in the structure was calculated using the XYZ coordinates of the protein structures. Two nonconsecutive amino acids were considered to be in contact if the distance between the geometric centers of their side chains was between 2 and 7 Å. The number of contacts per residue was calculated for the monomers, the pentamers, and the icosahedrons. The notation ''pentameric contact'' was adopted to describe a site involved in a contact that is present when a whole pentamer is considered but absent when only the structure of one monomer is analyzed. Likewise, a contact present in the icosahedron but absent in its pentamer was indicated as ''icosahedral contact.'' A detailed analysis of pentamer-pentamer interatomic contacts was performed using the program Contacts of Structural Units (CSU) (Sobolev et al. 1999) and confirmed with Swiss PDB Viewer (Guex, Diemand, and Peitsch 1999).

Phylogenetic Analysis

Phylogenetic trees were obtained using the aligned data set by the maximum parsimony (MP) and Neighbor-Joining (NJ) methods. MP analysis was done with the program PROTPARS of the PHYLIP package (Felsenstein 1993). To obtain the distances between protein sequences we used the PROTDIST module of PHYLIP with the Dayhoff 120 matrix option. The tree topology for these distances was built using the NEIGHBOR module. Bootstrapping (500 resamplings) to estimate the confidence limits of branching points was accomplished using the modules SEQBOOT and CONSENSE. The size of the data set made an exhaustive tree topology search by maximum likelihood prohibitive. Therefore, an exhaustive search with the eight members of three-dimensional known structure was performed with the MOLPHY software, version 2.3b3 (Adachi and Hasegawa 1996) using the JTT model with the frequencies estimated from the data (+F option).

Sequence Conservation Study

Given a sequence alignment, the conservation profile was characterized by calculating reduced sequence entropies per position. Reduced entropy was preferred because, in terms of structure, physicochemically conservative replacements are often equivalent. Following Ptitsyn (1998), amino acids were grouped into the following physicochemical classes: aromatics (F, Y, and W), bulky aliphatics (L, I, V, and M), small non-polar (G and A), acidic or amides (E, D, Q, and N), basic (K, R, H), with hydroxyl (S and T), and others (P and C). Then, for each alignment position i, the reduced entropy is given by:

$$S_i = -\sum_{\sigma=1}^{m} p_\sigma(i)\ln(p_\sigma(i)) + \frac{m^* - 1}{2n},$$

where σ is a given class of amino acids, m is the number of classes considered, and $p_\sigma(i)$ is the frequency of residues belonging to class σ at position i. In the second term, m^* is the number of amino acid classes for which $p_\sigma(i) \neq 0$ and n is the number of sequences analyzed. This term corrects a systematic bias in the estimation of the entropy (Roulston 1999).

As explained elsewhere, the set of sequences was divided into two clusters: one containing the known pentameric proteins and another containing the icosahedral ones. For each cluster, reduced entropy profiles were calculated as explained above. Positions with more than

fifty per cent of gaps in each group were not included in the calculations. To study the change in constraints between the non-icosahedral and icosahedral proteins, for each alignment position i, we calculated the entropy difference

$$\Delta S_i = S_i^{ico} - S_i^{penta},$$

where the first term is the reduced entropy of position i in the icosahedral cluster and the second term that of the non-icosahedral cluster.

Evolutionary Rate Shifts

The DIVERGE program (Gu and Vander Velden 2002) was used to detect putative residues subjected to altered structural constraints along the tree. As input, a multiple alignment of amino acid sequences and their corresponding phylogeny divided into a given number of clusters are required. The underlying two-state probabilistic model calculates the probability of each site being in the state of different evolutionary rate between the defined clusters. In each state, the evolutionary rate among sites varies according to the gamma distribution. The coefficient of functional/structural divergence between the clusters, θ, is defined as the proportion of sites expected to be rate shifted. A likelihood ratio test is then performed to evaluate the significance of the estimated θ value. If θ is significantly greater than zero, then the posterior probability of each site to have different rates under the sequence pattern observed is used to detect the putative residues involved in evolutionary rate changes. In the present analysis, the sequence alignment and the MP and NJ topologies divided into the two clusters mentioned above were used as inputs.

Results and Discussion

Sequence Alignment

Initially, 102 homologous sequences were recovered from sequence similarity searches. After very short and redundant sequences were removed, 86 lumazine synthase homologous proteins were included in the final data set. Table 1 shows the species names, Swiss Prot or Ref_Seq database accession numbers, and quaternary structures of the sequences analyzed in this study.

The range of degrees of identity between the sequences recovered is very broad, including some pairs—like the one formed by lumazine synthase sequences from the archaea *Aeropyrum pernix* and the bacteria *Sinorhizobium meliloti*—with values as low as 9% and a conservative substitution value of 24%. Moreover, a high degree of divergence was also previously observed between the sequence from *Brucella abortus* and the other pentameric structurally characterized members of the family.

The sequence alignment, presented online as Supplementary Material, is in good agreement with the structural alignment performed on the seven members of known three-dimensional structures (not shown). Most of the gaps observed in the sequence alignment fall in bends or turns of the structures. All of the positions described in this paper are numbered following this alignment. In some cases, the numeration following the *B. subtilis* sequence is also included.

Contact Patterns

The average of the numbers of contacts per position was calculated for the known three-dimensional forms. In figure 2, the pentameric contacts in all the pentamers are shown as a function of the position in the alignment. Lumazine synthase from the hyperthermophile *Aquifex aeolicus* has a large number of additional ion-pair interactions as well as an increased charge to hydrophobic residues ratio in the accessible surface compared to the rest of lumazine synthases (Zhang et al. 2001). This differential non-covalent interaction profile has also been observed in other hyperthermophile proteins related to the enhanced thermal stability requirements (Karshikoff and Ladenstein 2001). Because of this distinctive characteristic, this enzyme was not included in the contact analysis. The enzyme from *E. coli* was not included either, because its structural characterization derives from hydrodynamic and electron microscopy studies.

It can be seen from figure 2 that the pentameric contacts in both quaternary forms, pentameric and icosahedral, fall into the same sequence regions, reflecting the equivalence of the topology of pentamers in pentameric proteins and that of pentamers in icosahedric proteins. Particularly, a noticeable number of shared contacts was found in the helices α2, α3, and α5, and strands β3, β4, and β5. It is important to note that, in the case of the pentameric quaternary forms, the first N-terminal positions were disordered in all except *S. cerevisiae* LS structures, so that the structural information for the first 5 to 11 residues is almost missing. Taking into account the information available for this N-terminal extreme, more pentameric contacts were found in the icosahedral forms than in the pentameric form.

Figure 2 shows that icosahedral contacts reside in different regions than pentameric ones. The main contact regions determined agree with previous structural studies (Persson et al. 1999; Meining et al. 2000; Gerhardt et al. 2002). Specifically, icosahedral contact sites belong to α1, α4, and α5 helices, the sheets β4 and β6a and b, as well as to the loops previous to β2 and connecting β2 to α1, α1 to β3 and, finally, β6a to β6b.

Phylogenetic Analysis

The multiple sequence alignment was used to reconstruct the phylogeny of the lumazine synthase family. The topology obtained by MP is shown in figure 3. The overall topology derived with the NJ method was highly congruent (data not shown). Some differences appeared in the arrangement of the terminal taxa, but the main branch points were maintained. The subsequent analyses were not affected by these differences (as discussed below). The total MP tree was divided into two clusters, one of 41 sequences including all the branches that contain at least one of the known icosahedral forms and another containing the non-icosahedral representatives. We shall call these the ''icosahedral cluster'' and the ''non-icosahedral

Table 1
Lumazine Synthase Family: Species Names, Accession Numbers, and Quaternary Structures of the Sequences Included in the Present Analysis

Species Name	Accession #	Quaternary Structure
Actinobacillus pleuropneumoniae	P50856	
Aeropyrum pernix	Q9YC88	
Aeropyrum pernix	Q9YDC5	
Agrobacterium tumefaciens	Q8UG70	
Aquifex aeolicus	1HQK A	Icosahedral
Arabidopsis thaliana	O80575	
Archaeoglobus fulgidus	O28152	
Archaeoglobus fulgidus	O28856	
Bacillus amyloliquefaciens	Q44681	
Bacillus anthracis A2012	NP_658152	
Bacillus halodurans	Q9KCL4	
Bacillus subtilis	1RVV A	Icosahedral
Bartonella henselae	Q9REF4	
Brucella abortus	1DI0 A	Pentameric
Brucella melitensis	Q44668	
Brucella melitensis	Q8YGH2	
Buchnera aphidicola str. Sg (Schizaphis graminum)	Q8K9A6	
Buchnera sp. APS	Q9ZNM0	
Campylobacter jejuni	Q9PIB9	
Caulobacter crescentus CB15	Q9A8J4	
Caulobacter crescentus CB15	Q9A9S4	
Chlamydia muridarum	Q9PLJ4	
Chlamydia trachomatis	O84737	
Chlamydophila pneumoniae CWL029	Q9Z733	
Chlorobium tepidum TLS	Q8KAW4	
Clostridium acetobutylicum	Q97LG8	
Clostridium perfringens	Q8XMW9	
Corynebacterium ammoniagenes	O24753	
Corynebacterium glutamicum ATCC 13032	NP_600808	
Dehalospirillum multivorans	O68250	
Deinococcus radiodurans	Q9RXZ8	
Escherichia coli O157:H7 EDL933	P25540	Icosahedral[a]
Fusobacterium nucleatum ATCC 25586 subsp. nucleatum	Q8RIR4	
Haemophilus influenzae	P45149	
Halobacterium sp. NRC-1	Q9HRM5	
Helicobacter pylori 26695	O24854	
Helicobacter pylori J99	Q9ZN56	
Lactococcus lactis subsp. lactis	Q9CGU6	
Magnaporthe grisea	1C41 A	Pentameric
Mesorhizobium loti	NP_108510	
Mesorhizobium loti	NP_107635	
Methanococcus jannaschii	Q58584	
Methanococcus jannaschii	Q57751	
Methanopyrus kandleri AV19	Q8TYD9	
Methanopyrus kandleri AV19	Q8TYL5	
Methanosarcina acetivorans str. C2A	Q8TPT7	
Methanosarcina mazei Goe1	Q8Q093	
Methanothermobacter marburgensis str. Marburg	Q59587	
Methanothermobacter thermautotrophicus	O27443	
Methanothermobacter thermautotrophicus	O26237	
Mycobacterium leprae	Q9CCP3	
Mycobacterium tuberculosis CDC1551	P71685	
Neisseria meningitidis MC58	Q9JQV6	
Nicotiana tabacum	Q9XH13	
Nostoc sp. PCC 7120	Q8YQ43	
Pasteurella multocida	P57869	
Photobacterium leiognathi	Q93E92	
Photobacterium leiognathi	Q01994	
Photobacterium phosphoreum	P51963	
Pseudomonas aeruginosa PA01	Q9HWX5	
Pyrobaculum aerophilum	Q8ZTE3	
Pyrobaculum aerophilum	Q8ZW16	
Pyrococcus furiosus DSM 3638	Q8U4L8	
Ralstonia solanacearum	Q8Y1H8	
Rhodococcus erythropolis	Q53107	
Saccharomyces cerevisiae	1EJB A	Pentameric
Salmonella enterica subsp. enterica serovar Typhi	Q8XFI9	
Schizosaccharomyces pombe	1KYV A	Pentameric
Sinorhizobium meliloti	Q92NI1	
Sinorhizobium meliloti	Q92QU0	
Spinacia oleracea	1C2Y A	Icosahedral
Staphylococcus aureus subsp. aureus Mu50	Q931N8	
Staphylococcus aureus subsp. aureus N315	Q931N8	
Streptococcus pneumoniae TIGR4	Q97SY8	
Streptomyces coelicolor A3(2)	Q9EWJ9	
Sulfolobus solfataricus	Q980B4	
Sulfolobus solfataricus	Q989B5	
Sulfolobus tokodaii	Q975M6	
Sulfolobus tokodaii	Q975M5	
Synechocystis sp. PCC 6803	P73527	
Thermotoga maritima	Q9X2E5	
Vibrio cholerae	Q9KPU4	
Xanthomonas axonopodis pv. citri str. 306	Q8PPD6	
Xanthomonas campestris pv. Campestris str. ATCC 33913	Q8PCM7	
Xylella fastidiosa 9a5c	Q9PES4	
Yersinia pestis	Q8ZC41	

[a] Inferred from hydrodynamic and electron microscopy studies.

cluster,'' respectively. We preferred to use ''non-icosahedral'' and not ''pentameric'' because the known pentameric representatives are not uniformly distributed among the branches of the non-icosahedral cluster as is the case of icosahedral structures in the icosahedral cluster. In agreement with the MP topology, the exhaustive maximum likelihood topology (fig. 4) also allowed the division of the proteins of known structure into two well-defined groups according to their quaternary structures.

Entropy Differences

Entropy calculations were performed on the non-icosahedral and icosahedral clusters, and compared. After removal of all those positions for which entropies were not calculated (more than 50% of gaps in either of the clusters), there were 149 positions left. Figure 5*a* displays the profile of entropy differences between the icosahedral and non-icosahedral as a function of the alignment position, ΔS_i.

We were interested in studying the difference in constraint shifts of the sites involved in icosahedral contacts. The first step was to divide the 149 positions for which entropies were available into two sets, one containing the 27 positions involved in icosahedral contacts and the other including the remaining 122 positions. These two sets of alignment positions will be called ''icosahedral sites'' and ''non-icosahedral sites,'' respectively.

As shown in figure 5*a*, for most of the 149 positions the ΔS_i values are negative. This means that most sites are

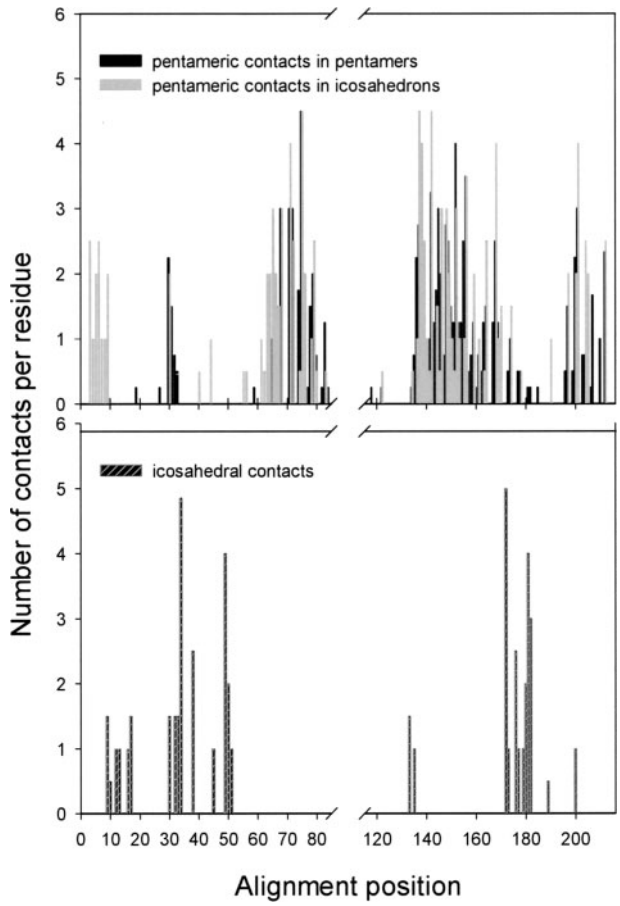

FIG. 2.—Number of contacts per residue as a function of the positions in the alignment. The notation ''pentameric contact'' represents a site involved in a contact that is present when a whole pentamer is considered but absent when only the structure of one monomer is analyzed. ''Icosahedral contact'' indicates a contact present in the icosahedron but absent in its pentamer.

more conserved in icosahedral proteins than in the non-icosahedral ones. This is a combined effect of changes in structural constraints and different degrees of evolutionary divergence of the two protein families. To focus only on changes of constraint, we compared the distributions of ΔS_i values of icosahedral sites with that of non-icosahedral sites. Phylogenetic effects should be similar for the two sets of sites, so that differences would mainly result from disparity in structural constraints. More precisely, we are interested in whether, when one compares icosahedral proteins with non-icosahedral ones, there is a significant increase of constraint in icosahedral sites, as compared with the rest of the sites.

To address the issue raised in the previous paragraph, we used the Mann-Whitney nonparametric test (Spiegel 1998). First, the 149 positions were ranked from lowest (more negative) to highest ΔS_i. Then, we calculated the sum of the ranks of the 27 icosahedral positions, R_{ico}, and used it to calculate:

$$U = N_{ico}N_{no-ico} + \frac{N_{ico}(N_{ico}+1)}{2} - R_{ico}$$

where N_{ico} and N_{no-ico}, are, respectively, the number of icosahedral (27) and non-icosahedral (122) sites considered. U has a nearly normal distribution with mean $\mu = N_{ico}N_{no-ico}/2$ and variance $\sigma^2 = N_{ico}N_{no-ico}(N_{ico}+N_{no-ico}+1)/12$. Thus, $Z = U - \mu/\sigma$, which has a normal distribution with mean 0 and standard deviation 1 was calculated. In this way, we obtained $Z = 4.24$, which has a one-tailed significance P value of less than 10^{-4}. This means that when going from pentameric to icosahedral proteins, the increase in constraint of icosahedral positions is larger than that of non-icosahedral positions.

To study in more detail the origin of the previous difference in constraint shifts, as well as the entropy differences, we analyzed the entropy components. Table 2 shows the average values of S_i and the corresponding standard errors for the following four cases: icosahedral positions in icosahedral proteins, non-icosahedral positions in icosahedral proteins, icosahedral positions in non-icosahedral proteins, and non-icosahedral positions in non-icosahedral proteins. From the inspection of these values, a double effect is apparent: (1) icosahedral positions are significantly more variable than the rest in non-icosahedral proteins, and (2) icosahedral positions are somewhat less variable than the rest in icosahedral proteins. To verify these observations, the Mann-Whitney U test was again used to evaluate whether the S_i values of icosahedral positions are significantly different from those of the rest of the positions. The Mann-Whitney Z scores are also shown in table 2, together with their one-tailed significance levels. These values confirm the previous statement (1): icosahedral positions are freer to vary than the rest in non-icosahedral proteins, at a significance level better than 10^{-4}. In contrast, even though the icosahedral contact positions in the icosahedral cluster are less variable than the rest, difference is not significant at the 0.01 level used in the present work.

Substitution Rate Shifts

Another way to study changes in constraints is to analyze evolutionary rate shifts. DIVERGE (see *Materials and Methods*) was used to identify functional/structural divergence between the icosahedral and non-icosahedral branches of the phylogenetic tree. The maximum likelihood estimate for the gamma shape parameter for rate variation among sites, α, was 2.019 (Gu and Zhang 1997). The maximum likelihood estimate of functional divergence between the two clusters was $\theta = 0.415$, with an associated standard error of 0.054. The derived Z score value ($Z = \theta$ estimated/associated standard error) of 7.74 implies a significant altered degree of functional/structural constraint between icosahedral and non-icosahedral proteins ($P < 10^{-4}$) (Gu 1999). This fact was also reflected in the value of 59.92 obtained for the likelihood ratio test (LRT) against the null hypothesis of nonfunctional/structural divergence. This value is statistically significant for a difference of one in the degrees of freedom between the two compared models ($P < 10^{-4}$, χ distribution) (Gu 2001). In figure 5*b* the posterior probability values for predicting amino acids involved in functional/structural divergence are represented as a function of the alignment

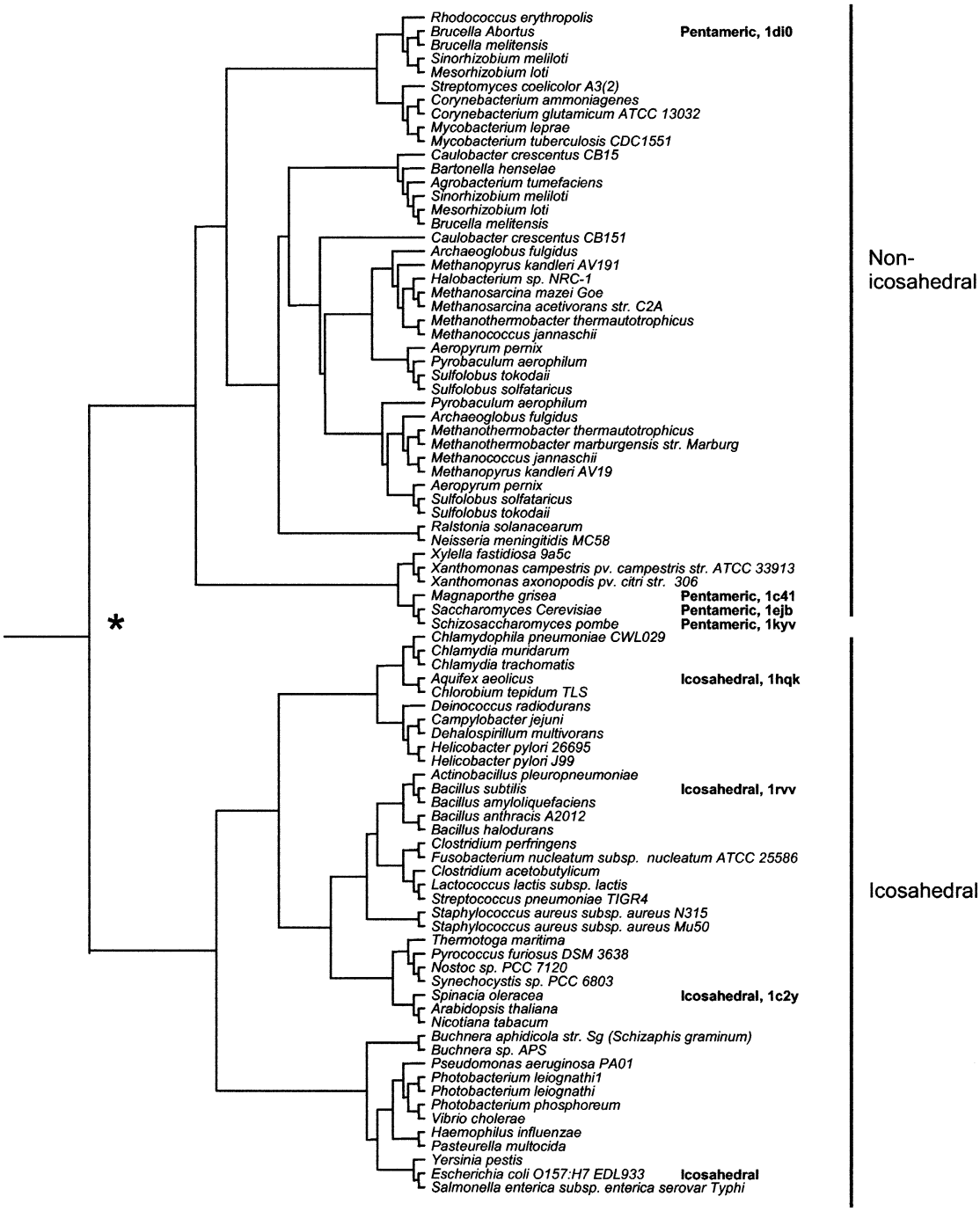

FIG. 3.—MP consensus topology based on the 86 sequence alignment. The node indicated with * had a percentage occurrence of 66 (500 bootstrap replicates of MP). This node splits the data into the two structural clusters. The tree was displayed using TreeView (Page 1996).

positions. The results obtained using the NJ topology as input were practically the same (data not shown).

Icosahedral Sequence Determinants

At this point, we are interested in identifying the specific sequence positions that are significantly different between the icosahedral and non-icosahedral proteins. Table 3 shows all positions that have a ΔS_i value smaller than that of non-icosahedral sites with significance level better than 0.01. The table also includes all positions whose rates have shifted between the icosahedral and non-icosahedral branches with a significance level better than 0.01, that is, DIVERGE posterior probabilities of being rate-shifted larger than 0.99.

Table 3 includes eight of the 27 icosahedral contact sites: alignment positions 10, 32, 45, 176, 177, 180, 182, and 200. In contrast, only three of the non-icosahedral sites are represented: 42, 197, and 204, showing the power of the present analysis to detect the icosahedral sequence

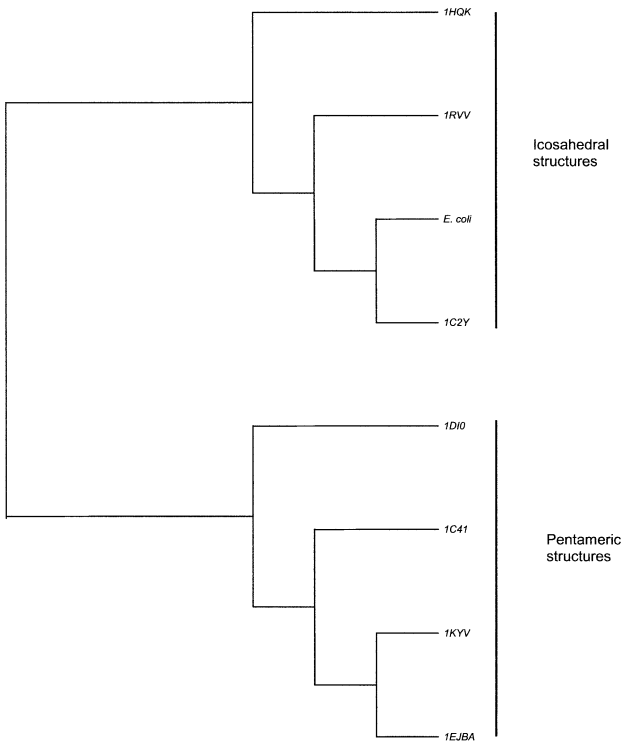

FIG. 4.—Topology obtained from a ML exhaustive search of the eight lumazine synthases with known three-dimensional structures. The tree was displayed using TreeView (Page 1996).

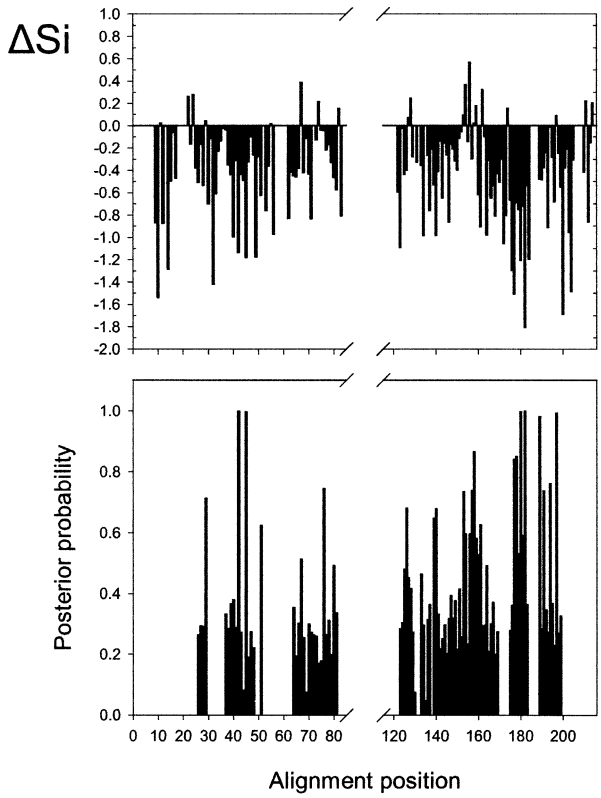

FIG. 5.—(*a*) ΔS_i values as a function of the positions in the alignment. (*b*) Posterior probability of a site being rate shifted because of functional/structural divergence.

determinants. Note that because we are working at an 0.01 significance level, of the 149 sites we would expect only one or two to be singled out if all sites belonged to the same distribution. Moreover, these one or two sites would probably be non-icosahedral, as there are 122 non-icosahedral sites and only 27 icosahedral sites.

The eight icosahedral positions are implicated in a variable number of contacts with neighboring pentamers in the enzyme from *B. subtilis*. Thus, following *B. subtilis* numbering, Gly 6 establishes 2 van der Waals contacts and 1 hydrogen bond; Asn 23, 13 van der Waals contacts and 1 hydrogen bond; Asp 36, 4 van der Waals contacts and 1 hydrogen bond; Ile 125, on average, 7 van der Waals contacts; Gly 129, 9 van der Waals contacts and 1 hydrogen bond; Lys 131, 20 van der Waals contacts; Glu 145, 3 van der Waals contacts and 1 hydrogen bond. In contrast, Glu 126 makes only 1 van der Waals contact.

Previous structural analyses postulated icosahedral determinants on two main sequence regions, the extreme N-terminal region and the last C- terminal α-helix. The change in the orientation of the β1 strand induced by proline in *M. grisea* has been proposed as the main cause for the failure to form icosahedrons (Persson et al. 1999). It was postulated that icosahedron formation would be prevented by the presence of one or more proline residues among the first 10 positions in the N-terminal extreme (Gerhardt et al. 2002). The present study does not support this proposition. First, even though all known pentameric representatives exhibit one or more prolines among the first residues, if one considers the proteins whose structures have not yet been determined, one finds some sequences in the non-icosahedral group that do not have prolines in the N-terminal region. Second, some of the putative icosahedral proteins studied do have prolines in this region. Furthermore the icosahedral LS of *Escherichia coli* has a proline residue position 11, which one would think should be as capable of perturbing the icosahedral structure as positions 1–10. Finally, it should be noted that only one of the positions of this region, position 10, has been detected as a sequence determinant of the icosahedral structure in the present analysis. Thus, the present study casts some doubt on the proposition that the presence of prolines in the N-terminal sequence is one of the signals of non-icosahedral structures. More experimental work will be needed to clarify this issue.

The comparison of crystal structures of pentameric and icosahedral LSs reveals important differences in the last α-helix (Ritsert et al. 1995; Persson et al. 1999; Braden et al. 2000; Meining et al. 2000; Zhang et al. 2001; Gerhardt et al. 2002). In the icosahedral LS of *B. subtilis* the C-terminal α-helix that starts at position 121 (172 in the alignment) is interrupted by the presence of a five residue kink. Because of this interruption this helix is also described as split into α4 and α5 helices. The sequence pattern associated with this kink was reported as G(T/G)K(A/H)G (positions 180 to 184 in the alignment, 129 to 133 following *B. subtilis* numbering). The present study seems to support the importance of this kink in the capacity to form icosahedrons. Of the five residues involved in this kink, two (180 and 182) were detected

Table 2
Differential Conservation of Icosahedral Positions. Comparison of Entropy (S_i) Distributions for Icosahedral and Non-Icosahedral Sites in the Two Protein Clusters Considered

Cluster	Sites	$\langle S_i \rangle$, σ[a]	Z score, P value[b]
Icosahedral	Icosahedral	0.5195, 0.4982	1.41, 0.08
	Non-icosahedral	0.6381, 0.4859	
Non-icosahedral	Icosahedral	1.3366, 0.4215	-3.94, 4×10^{-5}
	Non-icosahedral	0.9759, 0.4669	

[a] $\langle S_i \rangle$ represents the average of the S_i values calculated for each set of positions, icosahedral and non-icosahedral, in each protein set, icosahedral and non-icosahedral. σ represents the corresponding standard deviation.

[b] Z score is the Z score calculated from applying the Mann-Whitney U test to assess whether the S_i distribution for icosahedral positions is different from the one for non-icosahedral sites. P value is the one-tailed probability value corresponding to the calculated Z score in each set of samples.

Table 3
Differentially Constrained Sites. Alignment Positions with Either Significant Entropy Increase or Evolutionary Rate Shift, to 0.01 Level

Residue Position[a]	Location[b]	ΔS_i, P[c]	Rate-Shift Probability[d]
10 (6)	Loop between β1 and β2	−1.5382, 0.000	—
32 (23)	Bend between β2 and α1	−1.4184, 0.008	—
42 (33)	α1	−1.1361, 0.025	0.9993
45 (36)	α1	−1.1778, 0.016	0.9976
176 (125)	α4	−1.2978, 0.008	0.3622
177 (126)	α4	−1.5075, 0.000	0.8417
180 (129)	β6a	−1.2062, 0.008	0.9982
182 (131)	Loop between β6a and β6b	−1.8072, 0.000	0.9999
197 (142)	α5	0.0912, 0.885	0.9934
200 (145)	α5	−1.6990, 0.000	—
204 (149)	α5	−1.4846, 0.008	—

[a] The numbers in parentheses indicate the residue positions following *B. subtilis* sequence numbering.

[b] Location is the secondary structure element associated with the reported position.

[c] ΔS_i is the entropy difference between the icosahedral and non-icosahedral sets of sequences in the reported position. P is the associated significant level obtained as the fraction of non-icosahedral sites that have ΔS_i smaller than the site reported (that is, the probability of false positives).

[d] Posterior probability that a site will be involved in functional/structural divergence under an observed amino acid pattern.

in our analysis meeting the three requirements considered, that is, (1) to be an icosahedral contact position, (2) to exhibit a significant difference in sequence entropy, and (3) to be rate-shifted when the two clusters are compared. Furthermore, a structural superposition of this five-residue kink between the three solved icosahedral structures (PDB codes: 1RVV, 1HQK and 1C2Y) shows a high degree of conformational similarity (r.m.s.d. of 0.34 Å). The most remarkable feature is that, in the three cases, the loop folds in a way that displays the highly conserved lysine 182 pointing outward toward a neighboring pentamer. The K182 residue is 100% conserved among icosahedral LSs (41/41) and is practically absent in the non-icosahedral enzymes (43/45). In agreement, position 182 shows the highest value of entropy difference ($\Delta S_i = -1.8072$; see table 3 and fig. 5*a*).

The situation is different for the C-terminal α-helix of the pentameric lumazine synthases of *B. abortus*, *S. cerevisiae*, *M. grisea*, and *S. pombe*. All these sequences present insertions of variable length in this region, from one in *S. pombe* to four in *S. cerevisiae*. In *B. abortus*, the three-residue insertion does not affect the last α helix structure that is almost continuous without interruptions. These insertions have also been suggested as responsible for the lack of icosahedral assembly (Mörtl et al. 1996; Persson et al. 1999; Meining et al. 2000). However, inspection of the present sequence alignment reveals that not all non-icosahedral sequences display similar insertions. Furthermore, in the cases where there are insertions, a variable number of residues are involved. It is true, however, that none of the sequences belonging to the icosahedral group exhibit insertions in the same region as non-icosahedral proteins.

Before we finish this discussion, we will consider briefly those positions of table 3 that are not involved in icosahedral contacts: positions 42, 197, and 204. Position 197, was detected by its high rate-shift posterior probability value. According to its entropy change value, this position is more conserved in the non-icosahedral cluster than in the icosahedral one. This position takes part of the active site in *B. subtilis* (Ritsert et al. 1995). Furthermore, it is involved in pentameric contacts in the seven structures analyzed, and in all cases it contacts position 68—that is, it is part of the active site in all the structurally studied lumazine synthases. Thus, either directly or indirectly, position 197 is involved with the active site, so that the observed significant rate-shift probability could be related to changes in functional constraints rather than changes in structural constraints related to the ability to form icosahedral structures. In the case of position 204, even though it is not directly involved in icosahedral contacts, it exhibits a larger number of pentameric contacts in icosahedral proteins than in non-icosahedral ones. This would result in increased structural constraints, explaining the significant entropy decrease. Finally there is no apparent reason for the increased conservation of position 42. However, with the significance level used, one of the 122 non-icosahedral positions can be expected to be included in table 3: a false positive.

Conclusions

To detect the sequence determinants of the capacity to form icosahedral quaternary assemblies, we performed a comparative analysis of icosahedral and non-icosahedral lumazine synthases. In view of the entropy differences, we found that icosahedral sites face a larger constraint increment than non-icosahedral sites. Furthermore, this difference is mainly due to these sites being significantly more variable than the rest in non-icosahedral proteins, rather than being more conserved than the rest in icosahedral ones.

Regarding the sequence determinants of icosahedral structure, we found eight out of 27 icosahedral positions that display a significant degree of increased conservation in icosahedral proteins, as compared with non-icosahedral proteins. Considering the high degree of conservation found in these positions, the loss of this signal rather than the gain of additional sequence fragments appears to be the

most likely origin of the inability to form the icosahedral capsid. Thus, the gradual loss of this sequence pattern would explain the appearance of the pentameric forms and perhaps other intermediate oligomeric forms. Furthermore, the irregularities in the pattern of insertions in the last α-helix region and in the N terminus, both characteristics of the sequences belonging to the non-icosahedral group, support our view. In this sense, as mentioned above, in LS of *B. abortus* the three-residue insertion does not affect the structure of the last helix. This continuity supports the formation of a dimer of pentamers with a distinctive pattern of contacts confined to this region. This characteristic, apparently confined to the evolutionarily closer *B. abortus* relatives, reinforces the decisive role of the five-residue kink in the icosahedron formation. (F. A. Goldbaum, unpublished results).

The phylogenetic division of the sequences situated most of archae LSs in the cluster that includes the non-icosahedral forms. At first glance, judging from the sequence signals found, these sequences seem unable to form stable icosahedrons. The structural characterization of some of them will be important to confirm or redefine the organization of this group. The findings presented should contribute to refining the current structural data, to design assays to explore the role of these positions, and to the structural characterization of new sequences.

From an evolutionary perspective, some incipient explanations have been proposed in terms of the advantages of the icosahedral assembly (Bacher et al. 1996). However, a more exhaustive analysis is required to understand the relationship between life style of the species involved and the quaternary structure distribution. The knowledge of the sequence determinants of the different oligomeric forms of the lumazine synthase should contribute to addressing the ancestral state and the mechanisms involved in the evolution of this family.

Supplementary Material

An additional figure provided online as Supplementary Material includes the sequence alignment of the 86 sequences analyzed in the present work. Secondary structure elements of the seven known three-dimensional structure representatives and positions detected as sequence determinants of quaternary structure are also included.

Acknowledgments

The authors thank the two anonymous reviewers for their helpful suggestions. This work was supported by the Universidad Nacional de Quilmes, the Fundación Antorchas, the Agencia Nacional de Promoción Científica, Tecnológica y de Innovación. F.A.G. is the recipient of Carrillo-Oñativia and Howard Hughes Medical Institute International Research fellowships.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Atchley, W. R., W. Terhalle, and A. Dress. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48**:501–516.

Bacher, A., R. Baur, U. Eggers, H. D. Harders, M. K. Otto, and H. Schnepple. 1980. Riboflavin synthases of Bacillus subtilis. Purification and properties. J. Biol. Chem. **255**: 632–637.

Bacher, A. 1986. Heavy riboflavin synthase from Bacillus subtilis. Methods Enzymol. **122**:192–199.

Bacher, A., M. Fischer, K. Kis, K. Kugelbrey et al. (13 co-authors). 1996. Biosynthesis of riboflavin: structure and mechanism of lumazine synthase. Biochem. Soc. Trans. **24**: 89–94.

Bacher, A., S. Eberhardt, W. Eisenreich, M. Fischer, S. Herz, B. Illarionov, K. Kis, and G. Richter. 2001. Biosynthesis of riboflavin. Vitam. Horm. **61**:1–49.

Baldi, P. C., C. A. Velikovsky, B. C. Braden, G. H. Giambartolomei, C. A. Fossati, and F. A. Goldbaum. 2000. Structural, functional and immunological studies on a polymeric bacterial protein. Braz. J. Med. Biol. Res. **33**:741–747.

Braden, B. C., C. A. Velikovsky, A. A. Cauerhff, I. Polikarpov, and F. A. Goldbaum. 2000. Divergence in macromolecular assembly: x-ray crystallographic structure analysis of lumazine synthase from Brucella abortus. J. Mol. Biol. **297**:1031–1036.

Casari, G., C. Sander, and A. Valencia. 1995. A method to predict functional residues in proteins. Nat. Struct. Biol. **2**: 171–178.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Gerhardt, S., I. Haase, S. Steinbacher, J. T. Kaiser, M. Cushman, A. Bacher, R. Huber, and M. Fischer. 2002. The structural basis of riboflavin binding to Schizosaccharomyces pombe 6,7-dimethyl-8-ribityllumazine synthase. J. Mol. Biol. **318**: 1317–1329.

Goldbaum, F. A., C. A. Velikovsky, P. C. Baldi, S. Mörtl, A. Bacher, and C. A. Fossati. 1999. The 18-kDa cytoplasmic protein of Brucella species—an antigen useful for diagnosis—is a lumazine synthase. J. Med. Microbiol. **48**:833–839.

Golding, G. B., and A. M. Dean. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. **15**:355–369.

Gu, X., and J. Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664–1674.

Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18**:453–464.

Gu, X., and K. Vander Velden. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics **18**:500–501.

Guex, N., A. Diemand, and M. C. Peitsch. 1999. Protein modelling for all. Trends Biochem. Sci. **24**:364–367.

Henrick, K., and J. M. Thornton. 1998. PQS: a protein quaternary structure file server. Trends Biochem. Sci. **23**:358–361.

Huang, X., H. M. Holden, and F. M. Raushel. 2001. Channeling of substrates and intermediates in enzyme-catalyzed reactions. Annu. Rev. Biochem. **70**:149–180.

Jordan, D. B., K. O. Bacot, T. J. Carlson, M. Kessel, and P. V. Viitanen. 1999. Plant riboflavin biosynthesis. Cloning, chloroplast localization, expression, purification, and partial

characterization of spinach lumazine synthase. J. Biol. Chem. **274**:22114–22121.
Karshikoff, A., and R. Ladenstein. 2001. Ion pairs and the thermotolerance of proteins from hyperthermophiles: a ''traffic rule'' for hot roads. Trends Biochem. Sci. **26**:550–556.
Kearney, E. B., J. Goldenberg, J. Lipsick, and M. Perl. 1979. Flavokinase and FAD synthetase from Bacillus subtilis specific for reduced flavins. J. Biol. Chem. **254**:9551–9557.
Kis, K., R. Volk, and A. Bacher. 1995. Biosynthesis of riboflavin. Studies on the reaction mechanism of 6,7-dimethyl-8-ribityllumazine synthase. Biochemistry **34**:2883–2892.
Kis, K., and A. Bacher. 1995. Substrate channeling in the lumazine synthase/riboflavin synthase complex of Bacillus subtilis. J. Biol. Chem. **270**:16788–16795.
Ladenstein, R., H. C. Ludwig, and A. Bacher. 1983. Crystallization and preliminary X-ray diffraction study of heavy riboflavin synthase from Bacillus subtilis. J. Biol. Chem. **258**:11981–11983.
Ladenstein, R., M. Schneider, R. Huber, H. D. Bartunik, K. Wilson, K. Schott, and A. Bacher. 1988. Heavy riboflavin synthase from Bacillus subtilis. Crystal structure analysis of the icosahedral beta 60 capsid at 3.3 A resolution. J. Mol. Biol. **203**:1045–1070.
Ladenstein, R., K. Ritsert, R. Huber, G. Richter, and A. Bacher. 1994. The lumazine synthase/riboflavin synthase complex of Bacillus subtilis. X-ray structure analysis of hollow reconstituted beta-subunit capsids. Eur. J. Biochem. **223**:1007–1017.
Landgraf, R., D. Fischer, and D. Eisenberg. 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng. **12**:943–951.
Larson, S. M., I. Ruczinski, A. R. Davidson, D. Baker, and K. W. Plaxco. 2002. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. J. Mol. Biol. **316**:225–233.
Liao, D. I., Z. Wawrzak, J. C. Calabrese, P. V. Viitanen, and D. B. Jordan. 2001. Crystal structure of riboflavin synthase. Structure (Camb.) **9**:399–408.
Meining, W., S. Mörtl, M. Fischer, M. Cushman, A. Bacher, and R. Ladenstein. 2000. The atomic structure of pentameric lumazine synthase from Saccharomyces cerevisiae at 1.85 A resolution reveals the binding mode of a phosphonate intermediate analogue. J. Mol. Biol. **299**:181–197.
Mörtl, S., M. Fischer, G. Richter, J. Tack, S. Weinkauf, and A. Bacher. 1996. Biosynthesis of riboflavin. Lumazine synthase of Escherichia coli. J. Biol. Chem. **271**:33201–33207.
Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**: 536–540.
Page, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**:357–358.
Persson, K., G. Schneider, D. B. Jordan, P. V. Viitanen, and T. Sandalova. 1999. Crystal structure analysis of a pentameric fungal and an icosahedral plant lumazine synthase reveals the structural basis for differences in assembly. Protein Sci. **8**:2355–2365.
Ptitsyn, O. B. 1998. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? J. Mol. Biol. **278**:655–666.
———. 1999. Protein evolution and protein folding: non-functional conserved residues and their probable role. Pacific Symposium on Biocomputing, Hawaii, January 4–9;494–504.
Ritsert, K., R. Huber, D. Turk, R. Ladenstein, K. Schmidt-Bäse, and A. Bacher. 1995. Studies on the lumazine synthase/riboflavin synthase complex of Bacillus subtilis: crystal structure analysis of reconstituted, icosahedral beta-subunit capsids with bound substrate analogue inhibitor at 2.4 A resolution. J. Mol. Biol. **253**:151–167.
Roulston, M. S. 1999. Estimating the errors on measured entropy and mutual information. Physica D **125**:285–294.
Schott, K., R. Ladenstein, A. Konig, and A. Bacher. 1990. The lumazine synthase–riboflavin synthase complex of Bacillus subtilis. Crystallization of reconstituted icosahedral beta-subunit capsids. J. Biol. Chem. **265**:12686–12689.
Shenkin, P. S., B. Erman, and L. D. Mastrandrea. 1991. Information-theoretical entropy as a measure of sequence variability. Proteins **11**:297–313.
Sobolev, V., A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics **15**:327–332.
Spiegel, M. R. 1998. Statistics. McGraw-Hill-Education, Europe.
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.
Westhead, D. R., D. C. Hatton, and J. M. Thornton. 1998. An atlas of protein topology cartoons available on the World-Wide Web. Trends Biochem. Sci. **23**:35–36.
Zhang, X., W. Meining, M. Fischer, A. Bacher, and R. Ladenstein. 2001. X-ray structure analysis and crystallographic refinement of lumazine synthase from the hyperthermophile Aquifex aeolicus at 1.6 A resolution: determinants of thermostability revealed from structural comparisons. J. Mol. Biol. **306**:1099–1114.

William Taylor, Associate Editor

Accepted August 9, 2003