PERIODICUM BIOLOGORUM
VOL. 118, No 4, 385–397, 2016
DOI: 10.18054/pb.v118i4.4847

UDC 57:61
CODEN PDBIAD
ISSN 0031-5362

original research article

Variations in amino acid composition in bacterial single stranded DNA–binding proteins correlate with GC content

TINA PARADŽIK*
ŽELIMIRA FILIĆ
DUŠICA VUJAKLIJA*

Ruđer Bošković Institute, Bijenička 54, 10 000 Zagreb

Correspondence:
tsimunov@irb.hr;
vujaklij@irb.hr

Received October 20, 2016
Revised December 20, 2016
Accepted December 30, 2016

Abstract

Background and purpose. *Single-stranded DNA binding proteins (SSBs) are essential for the maintenance of the genome in all domains of life. Most bacterial SSBs are active as homotetramers. Each monomer comprises N-terminal domain (OB-fold) which is responsible for ssDNA binding and a disordered C-terminal domain (Ct) with a conserved acidic tail responsible for protein interactions.*

The variations in these essential proteins prompted us to conduct in silico analyses of the amino acid (aa) composition and properties of two distinct SSB domains in relation to bacterial GC content.

Materials and methods. *SSB sequences were collected from genomes covering a wide range of GC content from 14 bacterial phyla. The maximum-likelihood (ML) trees were constructed for SSB sequences and corresponding 16S rRNA genes. The aa contents of OB folds and Ct domains were subsequently analysed.*

Results. *We showed that SSB followed predicted aa composition as a function of genomic GC content. However, two distinct domains of SSB exhibit significant differences to the expected aa composition. Variations in aa proportion were more prominent in Ct domains. Elevated accumulation of Gly (up to 60 %) and Pro (up to 24 %), significant drop in the proportion of basic Lys and reduction in hydrophobic Leu, Ile and Val were identified in Ct domains of SSBs from high GC genomes. Consequently, this influences the biochemical properties of Ct domains.*

Conclusions. *Based on this comparative study of SSBs we conclude that genomic GC content and two distinct domains with different functional roles participate in shaping aa composition of SSBs.*

INTRODUCTION

SSB proteins are indispensable for the survival of cells in all domains of life *(1)*. The members of SSB protein family are involved in various processes of DNA metabolism by binding to transiently formed ssDNA during DNA recombination, replication and repair. These proteins bind ssDNA with a high affinity and in a sequence independent manner. Thus, SSBs prevent degradation of ssDNA and the formation of unproductive secondary structures *(2)*. One of the most extensively studied bacterial SSB belongs to *Escherichia coli* (EcoSSB) *(3, 4)*. Apart from passive protection of cellular ssDNA, SSB also has a second less known

role. It interacts and modulates the activity of various proteins involved in all aspects of DNA metabolism. EcoSSB has become one of the standard models for studying ssDNA-SSB interactions, a comprehensive review of which can be found in the work of Shereda and co-authors *(5)*.

The crystal structure of EcoSSB proved that the functional protein exists as a homotetramer *(6)*. Since then, most eubacterial SSBs have been shown to function as homoteramers *(7)*. Each SSB monomer contains two distinct domains: an N-terminal domain and a C-terminal (Ct) domain.

The N-terminal domain consists of approximately 110 amino-acid residues, comprising a DNA-binding domain known as OB-fold (a structurally conserved folding motif) *(8)*. The OB-fold binds to ssDNA through a combination of electrostatic and base-stacking interactions with the phosphodiester backbone and nucleotide bases *(9)*. It comprises five antiparallel β-sheets which form a β-barrel capped by an α-helix.

The Ct domain of SSB is significantly less conserved among bacterial SSBs. It is often rich in glycine and proline residues and thus structurally dynamic. This unfolded region cannot be seen in the crystal structures of SSB proteins *(5, 10, 11)*. The Ct domain terminates with a conserved acidic hexapeptide motif (D-D-D-I/L-P-F) recognized as a critical binding site for SSB interactions *(12)*.

It was shown that deletion of the Ct has an effect on SSB binding mode *(13)*. Removal of the acidic Ct motif (Ct tail) increases the intrinsic affinity for ssDNA and decreases cooperative binding, indicating that the Ct has an inhibitory effect on ssDNA binding. It was also reported that the extension of EcoSSB by a C-terminal glycine residue results in slower cell growth, indicating impaired protein function in vivo *(14)*. A study on the phage T7 gene 2.5 SSB protein has shown that the Ct tail competes with ssDNA for binding to the OB-fold *(15)*. In the proposed model the Ct tail is bound to the OB fold in the absence of ssDNA, while in the presence of ssDNA the Ct tail is released, thus leaving it free for the interactions with other cellular proteins. Recently it was demonstrated that an intrinsically disordered C-terminal region of *E. coli* SSB protein participated in cooperative binding to ssDNA *(16)*. All these results suggest an important regulatory role of Ct domain. Although the crystal structures of SSB proteins from taxonomically distant bacteria reveal similar ssDNA binding domains (OB folds) and oligomeric states *(9, 17-20)* some interesting variations have been noticed. Due to the orientation between oligomeric subunits (AC and BD) it has been observed that a homotetramer in the case of EcoSSB is an approximate spheroid *(9)*, whereas the SSBs of *Mycobacterium spp.* and *Streptomyces coelicolor* are ellipsoid *(7, 21)*. *Mycobacterium sp. and Streptomyces* sp. belong to distantly related genera of the phylum Actinobacteria. Mycobacteria are a widespread slow growing bacteria with some pathogenic properties, while streptomycetes are soil-inhabiting filamentous bacteria best known for producing antibiotics. Both of these genera belong to the high GC ratio Gram positive bacteria. Additionally, the SSBs from both genera contain a short sequence of 7 highly conserved amino acids (aa) which form an additional β-strand at the C-terminal end of the OB-fold which is not found in Gram negative *E. coli*. These strands form two additional clamp-like structures in the homotetrameric SSB which contribute to the overall stability of the quaternary structure of SSB *(18, 21)*. It is thought that the mode of DNA binding of actinobacterial SSBs is different from that of EcoSSB partly on account of the difference in the shape of the tetramers *(22, 23)*.

During our previous studies, it has been observed that SSBs from high GC content Actinobacteria have an extremely high glycine ratio in their Ct domain. Sequence comparison of SSB proteins of the representative mycobacterial and streptomycetes species revealed their relatively high sequence similarity (67 %), and existence of additional motifs that contribute to the overall SSB structure stability. It has also been noted that their SSBs lack many of the highly conserved aas crucial for the EcoSSB structure-function relationship *(9)*.

We hypothesized that observed change in the aa composition of OB-fold and Ct- domain might be related to high GC content.

Therefore, in this study using a larger data set, we have examined composition and specific properties of aas present in the OB-fold and C-terminal regions in relationship to the GC content of selected species. The results were compared and discussed with respect to relative aa composition of 961 proteomes from different organisms.

MATERIALS AND METHODS

SSB sequences analysed in this work were retrieved from Uniprot database (http://www.uniprot.org/) *(24)*. 16S rRNA gene sequences were downloaded from NCBI *(25)*. The SSB dataset was constructed to cover bacterial genomes with wide range of GC content (13-75 %) belonging to 14 different phyla, including five classes of Proteobacteria. In total, 199 SSB sequences belonging to 199 sequenced genomes, were collected for this study (Table 1). This dataset was divided into three categories according to the GC content: low (<40 %), medium (40-60 %) and high (>60 %).

Evolutionary distances of the selected species were calculated using a standard molecular marker (16S rRNA gene) and a corresponding SSB sequences. Multiple sequence alignment (MSA) of 16S rRNA sequences was obtained using Clustal Omega *(26)*. Statistical selection of models of nucleotide substitution was performed under the AIC in JModeltest *(27)*. SSB sequences of all selected

species were aligned using 3D PROMALS *(28).* Only conserved motifs obtained from the multiple sequence alignment by Gblocks server under default conditions with included options for the less stringent selection, were used in further analysis *(29).* The length of pruned alignment for 16S rRNA was 1389 nucleotides and for SSB was 74 aa (Appendix 2). Substitution modelling was completed using AIC in ProtTest *(30)* and used for phylogenetic analysis. Phylogenetic trees for both 16S rRNA and SSBs were constructed using maximum likelihood method in PhyML *(31)* under the best-fit models selected by AIC (GTR+I+G for 16S rRNA and LG+I+G+F for SSB). aLRT values were used to infer branch support. The nodes with aLRT values over 0.9 were considered well supported. Programme Seaview *(32)* was used for statistical report and CorelDRAW® for graphic presentation of the results.

Aa composition was analysed using Protparam *(33)* separately for OB fold and Ct domain. The properties of the aas within Ct domain were calculated using web server peptide2.com/N_peptide_hydrophobicity_hydrophilicity.php. Statistical analysis was performed using GraphPad Prism version 5.00 (GraphPad Software, La Jolla California USA,) which included one-way ANOVA test for the analysis of aa composition in the three defined GC groups, and Pearson's correlation analysis for the peptide hydrophobicity/ hydrophilicity properties.

Table 1. *Number of SSBs retrieved from the representative members of 14 phyla with increasing ratios of GC content used in this study. Phylum Proteobacteria is divided into five classes.*

GC content		<40 %	40-60 %	>60 %	Total number of SSBs
Actinobacteria		1	7	10	18
Proteobacteria	Alphaproteobacteria	7	7	10	24
	Betaproteobacteria	4	7	11	22
	Epsilonproteobacteria	7	1	-	8
	Gammaproteobacteria	8	8	10	26
	Deltaproteobacteria	1	9	10	20
Fusobacteria		5	-	-	5
Tenericutes		8	1	-	9
Spirochaete		5	6	1	12
Bacteroidetes		5	8	2	15
Firmicutes		8	6	3	17
Aquificae		4	4	-	8
Veruccomicrobia		-	2	1	3
Chloroflexi		-	2	2	4
Nitrospira		1	2	-	3
Gemmatimonadetes		-	-	1	1
Planctomycetes		-	-	1	1
Acidobacteria		-	-	3	3
Total		64	70	65	199

RESULTS AND DISCUSSION

Two domains with three distinctive elements can be found in the SSBs: N-terminal domain which forms DNA-binding domain (OB-fold), and C-terminal domain which is a largely unstructured region often rich in glycine and proline residues with a conserved acidic C-terminal motif. While studying structure/function relationship of the paralogous SSBs in streptomycetes we noticed that bacteria with high GC content possess some structural elements previously reported for mycobacterial SSBs *(23)*. Moreover, SSBs from streptomycetes (GC content over 70 %) exhibit the high content of glycine (57 %) in their Ct domain (Table 2) and also show some additional specific structural variations which contribute to structural stability of respective SSBs *(18, 23)*. Based on this, we proposed that all reported variations may be of significance for SSB functioning in a high GC content bacteria *(18).*

In this study we aimed to examine whether SSBs underwent some additional evolutionary changes, not seen previously, as a result of GC adaptation, and would it be possible to observe variation in such adaptation between two SSB domains.

For our first analysis we compared the evolutionary relationships of 199 selected bacterial species using two molecular markers, standard 16S rRNA gene and corresponding sequences of SSB proteins. Two maximum likelihood (ML) trees were constructed and the 16S rRNA gene and SSB trees were designated according to taxonomic group and GC content (Figure 1). The coloured branches in Figure 1 represent an overview of species GC coverage within different phyla, as numerically presented in Table 1. As depicted in Figure 1 bacteria with various GC content are dispersed across bacterial phyla. Moreover, GC content varies even between closely related genera within single phylum.

16S rRNA gene tree is generally well supported giving distinct groups which belongs to different phyla (Figure 1A). SSB tree revealed two distinct group of SSB proteins: one belongs to SSB from Proteobacteria, Chloroflexi, Acidobacteria and Bacteroidetes, and the other to SSB from Firmicutes, Actinobacteria, Tenericutes, Aquificae, Fusobacteria and Verrucomicrobia (Figure 1B). However, SSB tree is not well supported as 16S rRNA gene tree. The possible reason for this poor resolving is too short alignment obtained after Gblocks server (only 18 % of starting alignment). This could result in random branching which does not follow phyla relationships obtained in 16S rRNA gene tree. Furthermore, mixed branching of some divisions of Proteobacteria, although well supported in some cases, could be ascribed to the preservation of their com-

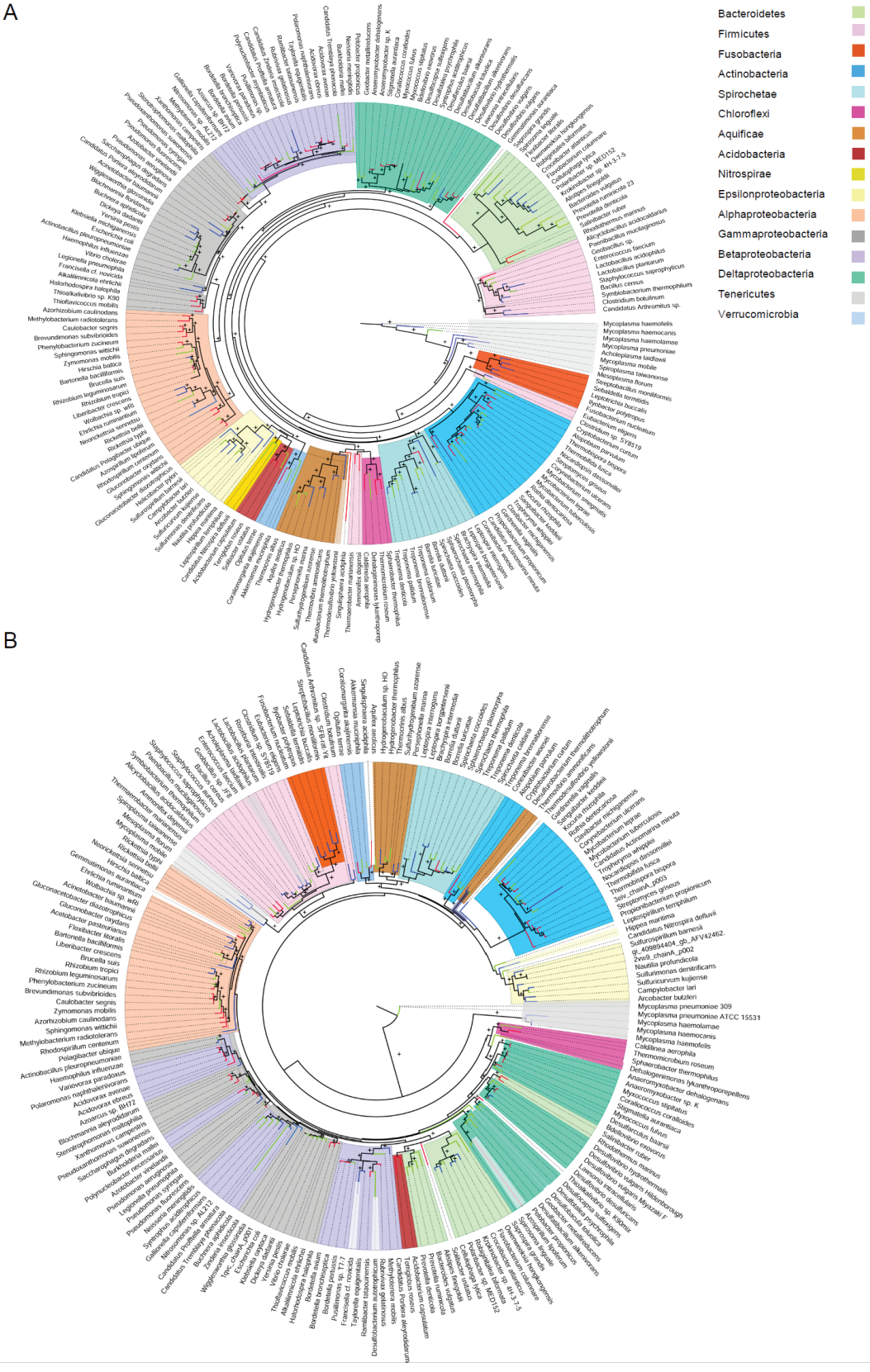


Figure 1. *ML trees constructed with 199 sequences of 16S rRNA gene (A), and corresponding SSBs (B). Branches are coloured depending on the GC content of species (blue – low GC, green – medium GC and red – high GC). Nodes with aLRT values equal or greater than 0.9 are indicated by asterisks.*

Table 2. *Comparison of SSB amino acid composition in three taxonomically distant bacterial species. Minimum and maximum values are with green background for OB folds and Ct domains.*

% GC	38	50	72	max/ min ratio	38	50	72	max/ min ratio
% AA		OB				Ct		
A	7,6	6,2	9,2	1,5	13,5	9,2	6,3	2,2
E	5,7	8,0	8,3	2,6	8,1	1,5	1,3	6,2
G	8,6	10,6	6,7	1,6	0,0	26,2	57,0	57,0
I	5,7	2,7	1,7	3,4	5,4	3,1	0,0	5,4
K	6,7	5,3	5,0	1,4	5,4	0,0	0,0	5,4
N	3,8	4,4	2,5	1,8	16,2	4,6	0,0	16,2
S	8,6	5,3	6,7	1,6	8,1	6,2	3,8	2,1
V	4,8	11,5	12,5	2,6	1,4	0,0	0,0	1.4
W	1,0	2,7	1,7	2,7	0,0	1,5	3,8	2,5

mon ancestral *ssb* gene. Some SSBs, e.g. two from extremophile species *Salinibacter ruber* and *Rhodothermus marinus* (phyla Bacteroidetes), branch within Deltaproteobacteria with support value over 0.9. This example could reflect some special SSB adaptations to extreme life conditions.

In addition, it has been reported that compositional bias may affect protein-based phylogenetic reconstructions *(34)*. Therefore, some SSB sequences could be branching outside their phyla due to GC-driven codon changes. Nevertheless, the majority of SSBs belonging to the same taxa branch together and follow the phylogenetic distribution as seen in the 16S rRNA gene tree (Figure 1).

Next, we examined the aa composition of the SSBs from three taxonomically distant species, *Helicobacter pylori*, (Epsilonproteobacteria), *Escherichia coli* (Gammaproteobacteria), and *Streptomyces coelicolor* (Actinobacteria). These species with solved SSB structures were selected since they possess 38 %, 50 % and 72 % GC ratio in their genomes, respectively *(6, 17, 18)*. Since OB fold is shown to be more conserved than Ct domain we have analysed aa composition for these two domains separately. Table 2 shows percentage of aa compositions (% aa) only for those aa which exhibit higher variation in percentages. Altogether, significantly greater variations in aa compositions are found between Ct domains. The most pronounced changes in the aa composition in Ct domain are observed for Gly and Asn, following Glu, Ile, Lys and Trp. With respect to the OB fold, the highest variations in the aa content were observed for Ile, Val and Trp. As stated above, the largest difference between species was observed for Gly and Asn residues; i.e. low GC bacteria *H. pylori* has high percentage of Asn in the Ct domain (16 %), while GC rich bacteria *S. coelicolor* has no Asn residues in its C-terminus. Contrary to this, *H. pylori* does not possess any glycine residues in its Ct, while *S. coelicolor* with high GC has 57 % of Gly residues within its Ct domain. On the other hand, *E. coli* with medium GC content (50 %) sits well between these values with an average of 4,6 % Asn and 26,2 % Gly.

In addition to observed interspecies differences, the composition of aa also differs between the OB fold and Ct domain of each species (Table 2).

To confirm that a change in aa content of the Ct domain can influence the chemistry of the region, we used *in silico* prediction tools to predict the nature of the Ct domain (Table 3). The data shows that in selected bacteria with high GC content there is a significant decrease in the percentage of the hydrophobic, acidic, and basic aa and increase in the neutral aa. The greatest differences are observed in the acidic (3 times) and basic (7 times) aa content of the Ct domains of *H. pylori* and *S. coelicolor* (Table 3).

In order to verify whether a similar trend could be observed in a much larger data set, we assembled 199 SSB proteins covering a broad range of GC content (13-75 %) and 14 phyla. It was not possible in this instance to find exactly equal number of SSBs from each GC category, however the numbers used were statistically comparable (Table 1).

Selected SSB sequences (199) divided into three GC categories were aligned to determine N- and Ct domains (Appendix 1). Based on these alignments, N- and C-terminal domains were separated for further analysis. As it has been reported previously *(5)*, in most cases N-terminal domain occupies approximately the first 110 aa. In addition to domain separation, the alignment did not reveal any obvious conserved motifs related to the GC content neither in the N- nor in the Ct domain. Since alignment of *S. coelicolor* SSB showed an extended Ct domain (approx. 20 aas) in comparison to *B. subtilis* and *E. coli* SSBs, we tested if this trend was conserved among all high GC content bacterial SSBs. Our data confirmed that selected GC rich bacteria have an extended Ct domain (P=0,002, n=199). However, when analysed taxa separately we have found that this trend is not conserved for all GC rich bacteria. For example, it was conserved for Alphaproteobacteria, but not for Actinobacteria. At present it is difficult to withdraw the final conclusion about the impor-

Table 3. *Biochemical properties of aa in Ct domain of SSBs from three bacterial species*

Biochemical properties of aa in Ct domain	*H. pylori*	*E. coli*	*S. coelicolor*
Hydrophobic	37,66 %	36,76 %	18,99 %
Acidic	14,29 %	7,35 %	5,06 %
Basic	7,79 %	2,94 %	1,27 %
Neutral	40,26 %	52,94 %	74,68 %

tance/correlation of the Ct domain length for the SSB proteins from bacteria with high GC content.

Recently, it was shown that shortening of C-terminal region of *E. coli* SSB protein had an impact on cooperative binding to ssDNA *(16)*. It was also reported that extension of Ct domain slowed cell growth rate, indicating impaired protein function in vivo *(14)*. Thus SSB C-terminal elongation could be only partially explained by adaptation of SSB proteins to varying GC content. Possibly within certain groups of bacteria, where this elongation is not significant, this could be reflection to some other adaptations. For example, slow growing bacteria need different SSB adaptation compared to fast growing bacteria.

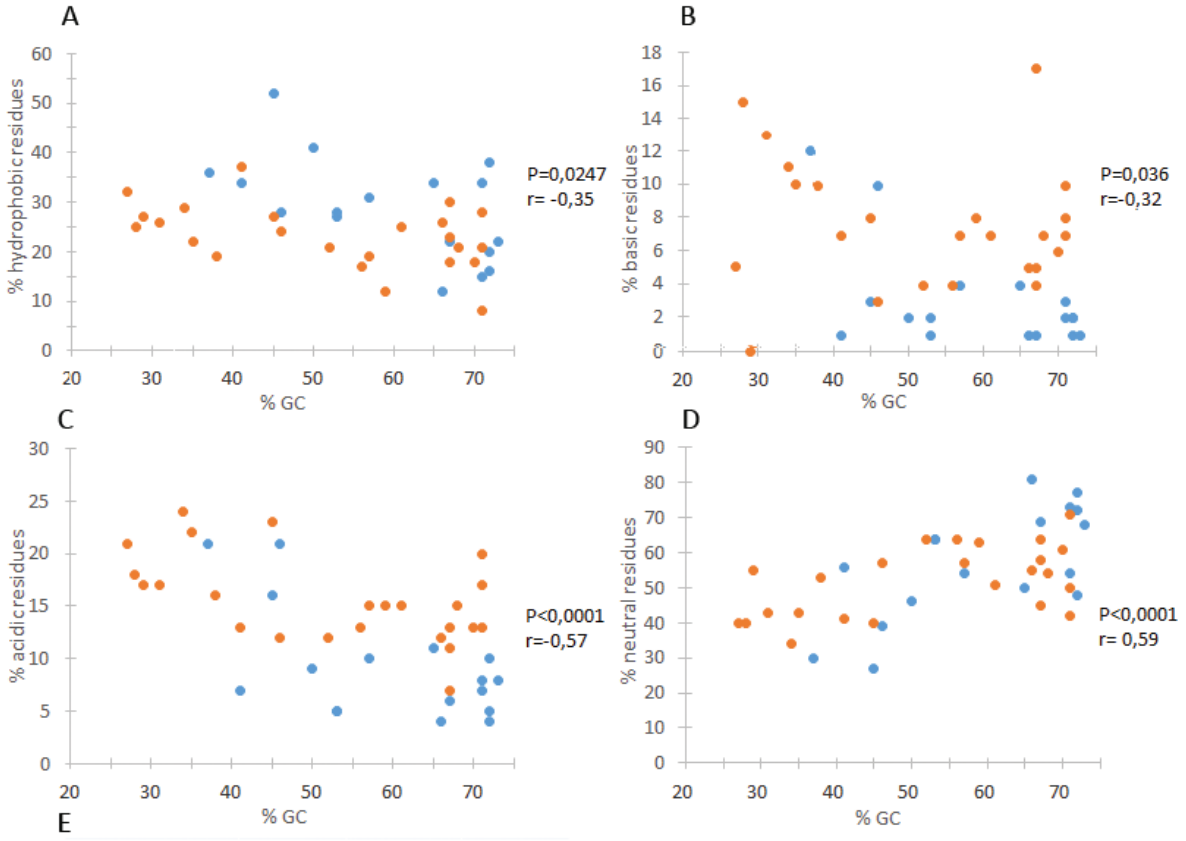
Next, as in the previous analysis (Table 3), the percentage of each aa in OB fold and Ct domain of SSB proteins collected from 199 proteomes was calculated and expressed as average within each of three groups (low, medium and high GC) (Table 4). Composition of amino acids in SSB proteins in our dataset were compared to composition of aas in the overall proteomes of 961 species *(35)*, as shown in Table 4. Observed changes in aa content reported for 961 proteomes were explained with enrichment of GC rich/poor codons *(35)*. Significant correlation between genomic GC composition and proteome aa content was well documented *(36, 37)*. It was reported that AT-rich genome would encode proteins rich in the Phe, Tyr, Met, Ile, Asn, and Lys (FYMINK), whereas GC-rich genomes would encode proteins rich in the Gly, Ala, Arg, and Pro (GARP) *(36)*. Indeed, this trend is present in SSB proteins (Table 4) *(34, 35)*. As shown previously for a trial sample (Table 2) the changes were much more pronounced for Ct domain.

Next, the trend of Gly accumulation in C-terminal domains, observed previously in SSBs in bacteria with high GC content, was also confirmed on this large data set (P<0.0001, n=199) (Table 4). However, it is much higher (up to 60 %, on average 30 %) than it has been expected from the reported proteome analyses; i.e. up to 10 % of Gly residues were found in GC rich proteomes *(35)*. This suggests that elevated Gly accumulation in the Ct domain evolved with some specific functional request of SSBs to high GC content genomes. The increase in the Gly content can contribute to the flexibility of this region while the extended Ct domain is possibly important to accommodate OB fold – Ct domain interaction in the ellipsoidal structure of SSBs from high GC content bacteria.

Next, although it has been expected that high GC content bacteria will accumulate Pro due to the GC codon enrichment *(35)*, OB folds from high GC content bacteria (Table 4) do not show this trend. This could be ascribed to the fact that proline, due to its unique chemical and structural properties, belongs to the group of the aas known to have „disorder-promoting" residues, and as such Pro can have a negative influence on the classical secondary elements which form OB-fold *(38)*. Thus, accumulation of Pro in Ct domains which are structurally disordered is not surprising. However, the proportion of Pro in Ct domain is higher (up to 24 %) than expected (proteomes possess approx. 6 % in GC rich organisms) *(35)*. Pro rich Ct domains have already been reported by other researchers *(4)*. Another aa which does not follow the expected trend is Tyr. This aa is encoded by GC poor codons and according to the Moura and collaborators *(35)* its content in high GC content bacterial genomes should decrease. However, the percentage value of Tyr seems to be independent on the GC content or SSB domain. Tyr residue is found to be phosphorylated in SSBs of phylogenetically distant bacteria *(39)*. It is reported that this

Table 4. *Comparison of aa composition of SSB sequences between different GC content groups. Average values (%) for each aa in OB-fold and Ct domain were shown for each GC category. The last column shows the overall trend in the aa composition in 961 proteomes collected from the bacteria with wide range of GC content (approx. 25-75 %). Statistically significant changes (P<0.05) between three groups for OB fold or Ct domains are with green background. Trends in overall proteomes are shown in the right column (↑ increase in aa ratio with an increase in GC, ↓ decrease in aa ratio with an increase in GC and – no overall change in aa ratio with increase in GC).*

% GC	<40	40-60	>60	<40	40-60	>60	
average % AA	OB fold			C-tail			overall proteomes (35)
A	5,4	6,4	6,5	4,7	8,2	9,8	↑
C	0,9	0,6	0,8	0,5	0,0	0,0	-
D	4,6	4,5	5,0	11,6	9,8	11,0	-
E	7,1	8,0	8,2	8,4	6,7	3,4	-
F	4,0	2,8	2,4	5,3	4,5	3,9	↓
G	7,7	8,3	8,9	7,1	17,5	29,9	↑
H	0,8	0,9	0,8	0,6	0,4	0,4	-
I	6,2	5,3	4,4	4,1	2,2	1,3	↓
K	7,2	6,2	6,1	5,2	1,8	0,2	↓
L	7,0	7,3	7,1	2,5	2,0	1,1	-
M	2,3	2,5	2,3	1,5	1,3	1,3	-
N	6,5	5,3	4,8	9,6	5,7	2,0	↓
P	1,5	1,9	1,9	7,1	10,8	10,7	↑
Q	4,4	4,9	5,1	7,5	7,2	6,2	-
R	7,3	8,4	9,0	2,3	4,0	6,5	↑
S	6,2	5,1	4,9	12,1	11,5	7,0	-
T	7,4	7,3	7,5	4,1	2,2	1,0	-
V	8,3	8,6	9,3	2,3	1,7	0,9	↑
W	1,8	2,2	2,4	0,3	0,4	0,5	↑
Y	3,3	3,4	3,0	2,7	2,1	2,4	↓



E

aminoacid type	low GC	medium GC	high GC
hydrophobic	30,86	28,43	22,73
acidic	20,14	12,57	10,30
basic	10,57	4,57	4,74
neutral	37,00	52,57	60,44

Figure 2. *Chemical properties of amino acids of Ct domains in two distantly related phyla, Actinobacteria and Alphaproteobacteria. Graphs show plotted percentage of each aa type (A hydrophobic, B basic, C acidic or D neutral) for SSB from single species against its GC content. All the correlations are statistically significant (P<0.05). Table on the bottom left (E) represents average percentage of certain aa type within single GC group.*

modification has an impact on ssDNA binding and thus the lack of Tyr can have a great impact on SSB binding properties. Contrary to trend predicted for the whole proteome, Val is an example of the aa with decreased percentage in high GC content bacteria, but only in SSBs Ct (Table 4). This is not surprising since Val is hydrophobic and usually found in the interior of proteins. Furthermore, Trp is recognized as an aa important for ssDNA binding *(40)*. As shown in Table 4, its content slightly increases in OB fold in bacteria with higher GC content, although this change is not highly statistically significant (P=0.001). Interestingly, the average content of Trp is higher for low GC SSBs than for the overall proteomes of high GC organisms. This possibly has some implication for the interactions of OB fold with ssDNA. For example in EcoSSB, Trp 40, 54 and 88 binds ssDNA *(9)* and these aa positions are not preserved in mycobacterial/streptomycetes SSBs. The ratio of Trp is low in Ct domain and is not influenced by the change in GC content (Table 4). The binding site of SSB-ssDNA in a low GC content bacterium *H. pylori* was determined by crystal structure, and instead of Trp (40 and 54), Phe (37, 50, and 55) predominantly participates in ssDNA binding (17). As depicted in the Table 4, our result also confirms higher content of Phe in low GC bacteria. Additional amino acids whose proportion was expected to be less affected by the GC content are Asp, Cys, Glu, Gln, His, Leu, Met and Ser *(35)*. Our data partially correlate with this observation indicating that variations in aa composition of SSB proteins are not only dependent on genomic GC content. As shown, the proportion of Asp in SSBs is fairly constant through different GC content but elevated in Ct domain for all GC categories. This is in agreement with the fact that acidic tip of Ct domain is essential for protein inter-

actions *(5)*. Interestingly, Glu which also contributes to acidity of Ct domain is significantly decreased within Ct domain (Table 4). Next, the proportion of hydrophobic Leu is not affected by the change in GC content of OB fold, but it is significantly decreased in Ct domain (Table 4). This is not surprising since other hydrophobic aa such as Ile are decreased within Ct of GC rich SSBs as well thus allowing higher flexibility of this domain. The disordered regions of proteins (such as Ct of SSB) are known to possess less hydrophobic aas (38).

Finally, discrepancy between proteome analyses *(35)* and our data was observed for Ser; proportion of this aa significantly decreased in OB fold and Ct domain in bacteria with high GC content (Table 4). In addition, the proportion of Thr is also affected by the change of GC content, but only in Ct of SSBs. In comparison to this result the overall proteome content of Thr is not changing with respect to GC content (Table 4).

In this analysis we show how two distinct domains of an essential protein exhibit significant differences in the aa composition with respect to the expected distribution of the aas for a defined GC content *(35)*. We demonstrated that some aas in the OB fold are affected by GC content, but not to the same extent as Ct domain of SSBs. This is expected since SSB has to preserve its core function in all living cells.

We additionally examined the biochemical properties of the aas that compose Ct domains of the SSBs from two distantly related bacterial clusters: Actinobacteria and Proteobacteria (α-division). Representative members with high, medium and low GC content were selected for this analysis. As shown in Figure 2, representative members of both groups showed statistically highly significant (P<0,0001) reduction in percentage of acidic aa content in dependence to GC content (Figure 2). In addition to this, percentage of neutral aa is increasing with an increase in GC, which is also highly statistically significant (P<0,0001). In contrast, the change of basic and hydrophobic aa content in dependence of GC content, although statistically significant (P=0,036 and P=0,0247 respectively), is not so pronounced. These properties are the result of changed aa composition and probably have an impact on the regulatory function of Ct-domain.

It has been reported that genes that evolve slowly are less affected by aa composition changes due to the changed GC content than the more rapidly diverging genes *(36)*.

As reported, the strongly conserved housekeeping genes, *gap* and *tuf* show amino acid composition changes in the predicted directions, although to a more moderate degree than non-essential genes *(36)*.

Since SSB proteins also belong to the housekeeping genes we expected the same trend at least in the OB fold domain. Indeed, in this study, it was shown that the OB fold tended to be more conserved, although some specific changes had been observed. In contrast, the Ct domain displayed a greater variability in aa composition with respect to GC content, with the exception of the acidic tail motif *(15)*. Properties of Ct domain are presumably changed predominantly due to elevated content of Gly residues and reduced composition of other aas such as Glu, Asn and Ser. Accumulation of the Gly not only changes the aa ratio but also promotes the extension of the Ct domain. It has been reported that long disordered regions increase the complexity of protein interacting networks *(41)*. Such regions within proteins are often found to be evolving faster than ordered regions *(42)*. This is in agreement with our data and with the biological role of Ct domain, which is essential for SSB network.

Acknowledgements: *The authors would like to thank dr.sc. Gerrard A. Quinn for a critical reading of this manuscript, to dr.sc. Bruna Pleše for help in phylogeny tree construction, and to dr.sc. Nevenka M. Radan for helpful discussion.*

REFERENCES

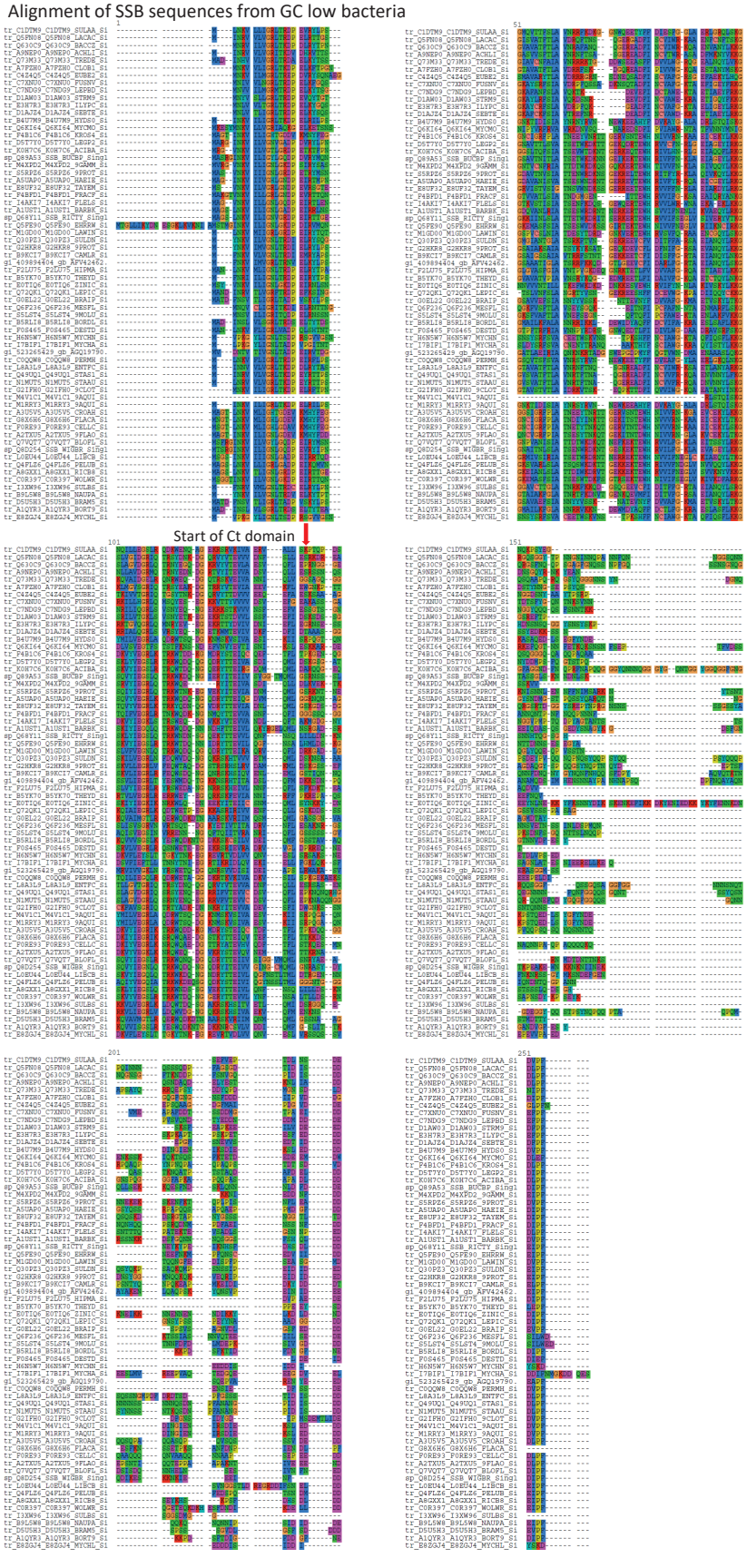
1. Szczepankowska AK, Prestel E, Mariadassou M, Bardowski JK, Bidnenko E. Phylogenetic and complementation analysis of a single-stranded DNA binding protein family from lactococcal phages indicates a non-bacterial origin. PloS one. 2011;6(11):e26942. https://doi.org/10.1371/journal.pone.0026942
2. Glassberg J, Meyer RR, Kornberg A. Mutant single-strand binding protein of Escherichia coli: genetic and physiological characterization. J Bacteriol. 1979;140(1):14-9.
3. Sigal N, Delius H, Kornberg T, Gefter ML, Alberts B. A DNA-unwinding protein isolated from Escherichia coli: its interaction with DNA and with DNA polymerases. Proc Natl Acad Sci U S A. 1972;69(12):3537-41. https://doi.org/10.1073/pnas.69.12.3537
4. Meyer RR, Laine PS. The single-stranded DNA-binding protein of Escherichia coli. Microbiol Rev. 1990;54(4):342-80.
5. Shereda RD, Kozlov AG, Lohman TM, Cox MM, Keck JL. SSB as an organizer/mobilizer of genome maintenance complexes. Crit Rev Biochem Mol Biol. 2008;43(5):289-318. https://doi.org/10.1080/10409230802341296
6. Raghunathan S, Ricard CS, Lohman TM, Waksman G. Crystal structure of the homo-tetrameric DNA binding domain of Escherichia coli single-stranded DNA-binding protein determined by multiwavelength x-ray diffraction on the selenomethionyl protein at 2.9-A resolution. Proc Natl Acad Sci U S A. 1997;94(13):6652-7. https://doi.org/10.1073/pnas.94.13.6652
7. Stefanic Z, Vujaklija D, Andrisic L, Mikleusevic G, Andrejasic M, Turk D, et al. Preliminary crystallographic study of Streptomyces coelicolor single-stranded DNA-binding protein. Croat Chem Acta. 2007;80(1):35-9.
8. Murzin AG. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. 1993;12(3):861-7.
9. Raghunathan S, Kozlov AG, Lohman TM, Waksman G. Structure of the DNA binding domain of E. coli SSB bound to ssDNA. Nat Structur Biol. 2000;7(8):648-52. https://doi.org/10.1038/77943
10. Savvides SN, Raghunathan S, Futterer K, Kozlov AG, Lohman TM, Waksman G. The C-terminal domain of full-length E. coli SSB is disordered even when bound to DNA. Protein Sci. 2004;13(7):1942-7. https://doi.org/10.1110/ps.04661904
11. Lu D, Keck JL. Structural basis of Escherichia coli single-stranded DNA-binding protein stimulation of exonuclease I. Proc Natl Acad

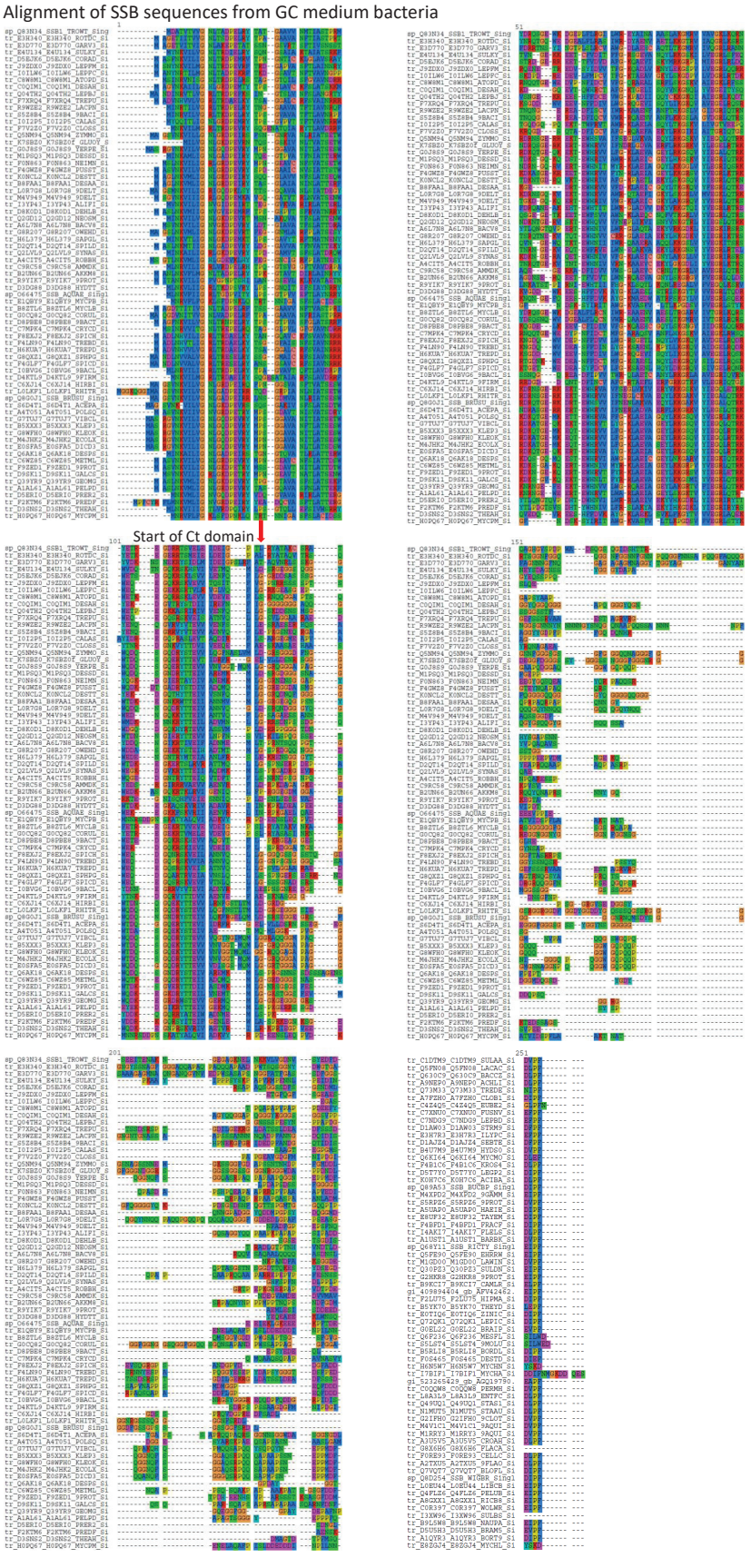
Sci U S A. 2008;105(27):9169-74. https://doi.org/10.1073/pnas.0800741105

12. Curth U, Genschel J, Urbanke C, Greipel J. In vitro and in vivo function of the C-terminus of Escherichia coli single-stranded DNA binding protein. Nucleic Acids Res. 1996;24(14):2706-11. https://doi.org/10.1093/nar/24.14.2706
13. Kozlov AG, Cox MM, Lohman TM. Regulation of single-stranded DNA binding by the C termini of Escherichia coli single-stranded DNA-binding (SSB) protein. J Biol Chem. 2010;285(22):17246-52. https://doi.org/10.1074/jbc.M110.118273
14. Naue N, Fedorov R, Pich A, Manstein DJ, Curth U. Site-directed mutagenesis of the chi subunit of DNA polymerase III and single-stranded DNA-binding protein of E. coli reveals key residues for their interaction. Nucleic Acids Res. 2011;39(4):1398-407. https://doi.org/10.1093/nar/gkq988
15. Marintcheva B, Marintchev A, Wagner G, Richardson CC. Acidic C-terminal tail of the ssDNA-binding protein of bacteriophage T7 and ssDNA compete for the same binding surface. Proc Natl Acad Sci U S A. 2008;105(6):1855-60. https://doi.org/10.1073/pnas.0711919105
16. Kozlov AG, Weiland E, Mittal A, Waldman V, Antony E, Fazio N, et al. Intrinsically disordered C-terminal tails of E. coli single-stranded DNA binding protein regulate cooperative binding to single-stranded DNA. J Mol Biol. 2015;427(4):763-74. https://doi.org/10.1016/j.jmb.2014.12.020
17. Chan KW, Lee YJ, Wang CH, Huang H, Sun YJ. Single-stranded DNA-binding protein complex from Helicobacter pylori suggests an ssDNA-binding surface. J Mol Biol. 2009;388(3):508-19. https://doi.org/10.1016/j.jmb.2009.03.022
18. Stefanic Z, Vujaklija D, Luic M. Structure of the single-stranded DNA-binding protein from Streptomyces coelicolor. Acta Crystallogr D Biol Crystallogr. 2009;65(Pt 9):974-9. https://doi.org/10.1107/S0907444909023634
19. George NP, Ngo KV, Chitteni-Pattu S, Norais CA, Battista JR, Cox MM, et al. Structure and cellular dynamics of Deinococcus radiodurans single-stranded DNA (ssDNA)-binding protein (SSB)-DNA complexes. J Biol Chem. 2012;287(26):22123-32. https://doi.org/10.1074/jbc.M112.367573
20. Jedrzejczak R, Dauter M, Dauter Z, Olszewski M, Dlugolecka A, Kur J. Structure of the single-stranded DNA-binding protein SSB from Thermus aquaticus. Acta Crystallogr D Biol Crystallogr. 2006;62(Pt 11):1407-12. https://doi.org/10.1107/S0907444906036031
21. Saikrishnan K, Manjunath GP, Singh P, Jeyakanthan J, Dauter Z, Sekar K, et al. Structure of Mycobacterium smegmatis single-stranded DNA-binding protein and a comparative study involving homologus SSBs: biological implications of structural plasticity and variability in quaternary association. Acta Crystallogr D Biol Crystallogr. 2005;61(Pt 8):1140-8. https://doi.org/10.1107/S0907444905016896
22. Reddy MS, Guhan N, Muniyappa K. Characterization of single-stranded DNA-binding proteins from Mycobacteria. The carboxyl-terminal of domain of SSB is essential for stable association with its cognate RecA protein. J Biol Chem. 2001;276(49):45959-68. https://doi.org/10.1074/jbc.M103523200
23. Paradzik T, Ivic N, Filic Z, Manjasetty BA, Herron P, Luic M, et al. Structure-function relationships of two paralogous single-stranded DNA-binding proteins from Streptomyces coelicolor: implication of SsbB in chromosome segregation during sporulation. Nucleic Acids Res. 2013. https://doi.org/10.1093/nar/gkt050
24. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database-Oxford. 2011. https://doi.org/10.1093/database/bar009
25. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7-D19. https://doi.org/10.1093/nar/gkv1290
26. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. https://doi.org/10.1038/msb.2011.75
27. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25(7):1253-6. https://doi.org/10.1093/molbev/msn083
28. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 2007;23(7):802-8. https://doi.org/10.1093/bioinformatics/btm017
29. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540-52. https://doi.org/10.1093/oxfordjournals.molbev.a026334
30. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21(9):2104-5. https://doi.org/10.1093/bioinformatics/bti263
31. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21. https://doi.org/10.1093/sysbio/syq010
32. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27(2):221-4. https://doi.org/10.1093/molbev/msp259
33. Elisabeth Gasteiger CH, Alexandre Gattiker, S'everine Duvaud, Marc R. Wilkins, Ron D. Appel, Amos Bairoch. The Proteomics Protocols Handbook. Humana Press; 2005.
34. Foster PG, Hickey DA. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol. 1999;48(3):284-90. https://doi.org/10.1007/PL00006471
35. Moura A, Savageau MA, Alves R. Relative Amino Acid Composition Signatures of Organisms and Environments. PloS one. 2013;8(10). https://doi.org/10.1371/journal.pone.0077319
36. Singer GA, Hickey DA. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol. 2000;17(11):1581-8. https://doi.org/10.1093/oxfordjournals.molbev.a026257
37. Lightfield J, Fram NR, Ely B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PloS one. 2011;6(3):e17677. https://doi.org/10.1371/journal.pone.0017677
38. Theillet F, Kalmar L, Tompa P, Han K, Selenko P, Dunker A, et al. The alphabet of intrinsic disorder. Intrinsically Disord Proteins. 2013;1(1). https://doi.org/10.4161/idp.24360
39. Mijakovic I, Petranovic D, Macek B, Cepo T, Mann M, Davies J, et al. Bacterial single-stranded DNA-binding proteins are phosphorylated on tyrosine. Nucleic Acids Res. 2006;34(5):1588-96. https://doi.org/10.1093/nar/gkj514
40. Casas-Finet JR, Khamis MI, Maki AH, Ruvolo PP, Chase JW. Optically detected magnetic resonance of tryptophan residues in Escherichia coli ssb gene product and E. coli plasmid-encoded single-stranded DNA-binding proteins and their complexes with poly(deoxythymidylic) acid. J Biol Chem. 1987;262(18):8574-83.
41. Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. 2013;22(6):693-724. https://doi.org/10.1002/pro.2261
42. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 2002;55(1):104-10. https://doi.org/10.1007/s00239-001-2309-6

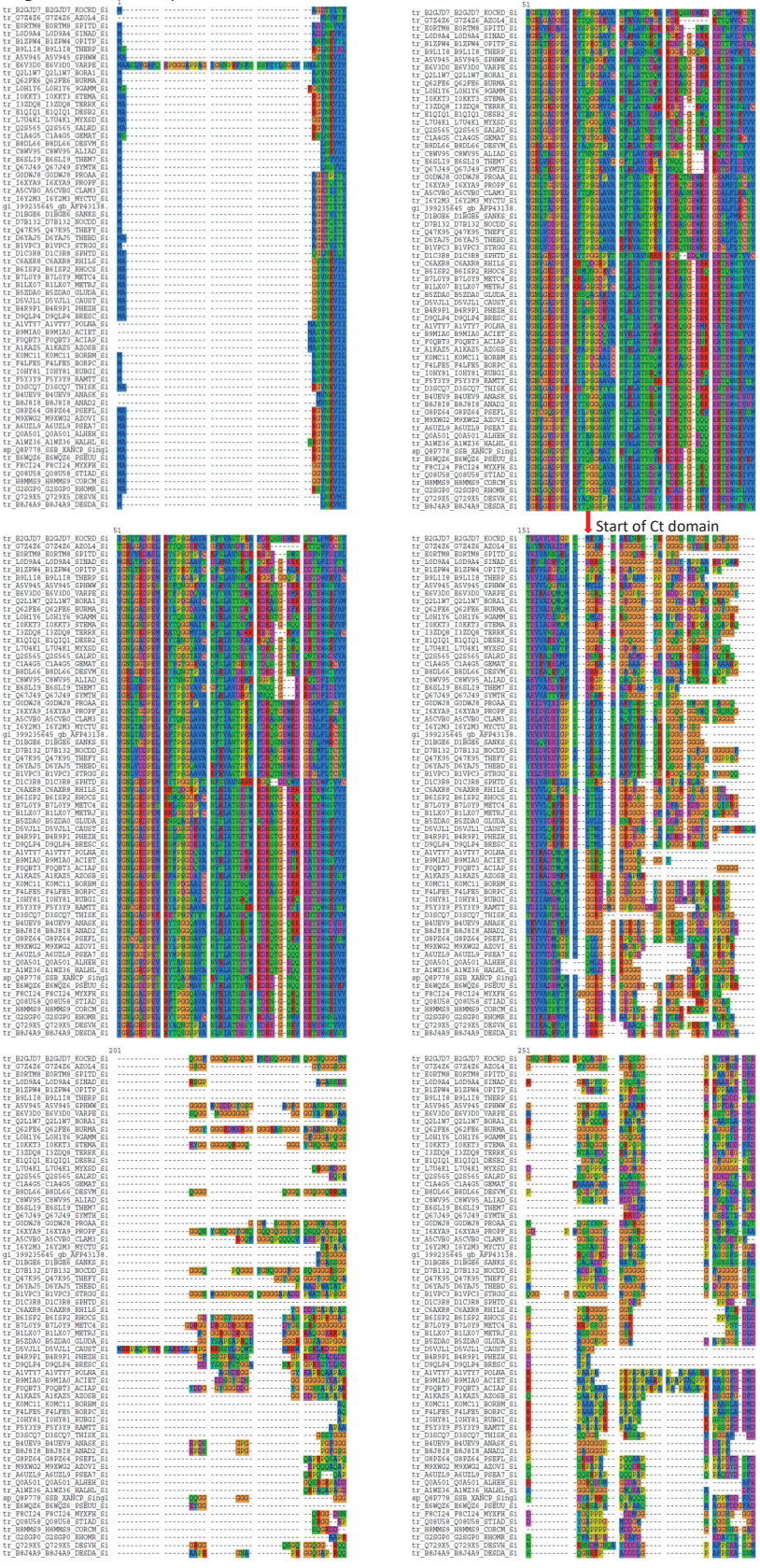
APPENDIX 1. ALIGNMENTS OF SSB SEQUENCES

Alignment of SSB sequences from GC low bacteria




Alignment of SSB sequences from GC medium bacteria
Start of Ct domain

Alignment of SSB sequences from GC rich bacteria



APPENDIX 2. FINAL ALINGMENT OF SSB PROTEINS FROM 199 BACTERIAL SPECIES