J Mol Model (2006) 12: 973–983
DOI 10.1007/s00894-005-0096-5

ORIGINAL PAPER

T. Pons · B. González · F. Ceciliani · A. Galizzi

FlgM anti-sigma factors: identification of novel members of the family, evolutionary analysis, homology modeling, and analysis of sequence-structure-function relationships

Received: 24 March 2005 / Accepted: 2 December 2005 / Published online: 4 May 2006

Abstract FlgM proteins, also known as Anti-sigma-28 factor (σ_{28}), are negative regulators of flagellin synthesis. Recently, a three-dimensional structure of the *Aquifexaeolicus* σ_{28}/FlgM complex (PDB code: 1rp3) was determined by X-ray crystallography at 2.3 Å resolution. Furthermore, experimental data on bacterial FlgM, including site-directed mutagenesis and structural characterization by NMR are also available. However, an interpretation of the sequence-structure-function relationships combining X-ray and NMR data with the evolutionary information extracted from the increasing number of FlgM-related sequences annotated in databases is not available. In the present study, we combined database sequence searches and sequence-analysis tools to update the multiple sequence alignment of a previously characterized cluster of orthologs (COG2747) and the PFAM classification of protein domains (PF04316) for the FlgM family. A phylogenetic analysis of 77 protein sequences revealed the presence of at least three major sequence clades within the FlgM family. Besides, we predicted functional residues using a SequenceSpace method. We also generated homology models for *Bacillus subtilis* and *Salmonella typhimurium* FlgM proteins, for which sequence-structure-function relationship data are available, and used the docking program ClusPro to hypothesize about the dimer association between FlgM proteins. In conclusion, the analysis presented in this work will be useful in designing new experiments to understand better protein–protein interactions between FglM, sigma factors, and putative molecules from the flagellar export apparatus. Electronic Supplementary Material is available in the online version of this article at http://link.springer.de/

Keywords SequenceSpace · Homology modeling · Docking · Protein–protein interaction

Abbreviations NMR: nuclear magnetic resonance · 3D: three-dimensional · PAGE: polyacrylamide gel electrophoresis · ASA: solvent accessibility surface area

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00894-005-0096-5

T. Pons (✉) · B. González
Centro de Ingeniería Genética y Biotecnología (CIGB), Havana, 10600, Cuba
e-mail: tirso.pons@cigb.edu.cu
Fax: +537-2714764
e-mail: tp_hernandez@yahoo.es

F. Ceciliani
Dipartimento di Patologia Animale, Igiene e Sanita, Pubblica Veterinaria, Università degli Studi, Milano, 20133, Italy

A. Galizzi
Dipartimento di Genetica e Microbiologia "A. Buzzati-Traverso", Universita degli Studi, Pavia, 27100, Italy

Present address:
T. Pons
Centro de Estudios de Proteinas (CEP)
Facultad de Biologia, Universidad de La Habana, Havana 10400, Cuba
e-mail: pons@fbio.uh.cu
Fax: +537-8321321

Introduction

Transcription initiation in bacteria is controlled by alternative sigma factors that bind to the catalytic core RNA polymerase (RNAP) to form the holoenzyme. FlgM is an anti-sigma factor of the flagellar-specific sigma-28 (σ^{28}) subunit of RNAP; it exerts its regulation by its direct interaction both to free σ^{28} to prevent it from forming a complex with core RNAP, and to σ^{28} holoenzyme (σ^{28}/RNAP) to destabilize the complex [1]. The σ^{28} homologs are present in a wide range of flagellated bacteria and many of these systems contain FlgM proteins. These FlgM proteins, as has been demonstrated in *Salmonella typhimurium*, *Escherichiacoli*, *Bacillus subtilis*, *Vibriocholerae*, *Helicobacterpylori* and *Pseudomonasaeruginosa*, participate in the regulation of the complex flagellar transcriptional circuit [2–7].

Concerning sequence-structure-function relationships studies, previous mutagenesis analysis suggested that the

N-terminal region, between amino acids Ser7 and Val25, of the *S. typhimurium* FlgM protein is essential for flagella-specific export [8], whereas, all mutations in FlgM that prevent σ^{28} inhibition are localized to a contiguous region in the C-terminal half of the protein [9]. Using deletion analysis in the *S. typhimurium* FlgM, a minimal binding domain was identified between Glu64 and Arg88 [8].

It is generally accepted that homologous proteins in a family share biological properties (e.g., function, ligand-binding specificity, post-translational modifications), depending on the degree of similarity in their amino-acid sequence. However, some differences appear when their amino-acid sequence identity decreases. Molecular characterization of FlgM proteins from both Gram-negative *S. typhimurium* and Gram-positive *B. subtilis* indicated different properties: N-terminal region of *B. subtilis* FlgM (residues 1–51) is structured, as was deduced from limited proteolysis studies [10]. However, *S. typhimurium* FlgM is largely unfolded in solution, as established by high resolution NMR experiments [9]. Upon interaction with σ^{28}, the C-terminal region of *S. typhimurium* FlgM becomes structured [9]. Also, it is postulated that the unfolded state of the FlgM proteins may facilitate their secretion through the channel of the basal body-hook structure [11].

In contrast to the FlgM of *S. typhimurium*, which actively dissociates core RNAP from the holoenzyme [12], the FlgM of *B. subtilis* can prevent the holoenzyme formation, but is not able to dissociate core RNAP from the holoenzyme [13]. On the other hand, *B. subtilis* FlgM is a dimer in solution, as determined by gel exclusion chromatography, native PAGE electrophoresis, and chemical cross-linking experiments [10]. However, there is no information concerning the possible oligomerization of FlgM of *S. typhimurium*. Another anti-sigma factor, AsiA anti-sigma-70, is a symmetric dimer in solution, and interacts with σ^{70} as a monomer via the same residues used for dimerization [14].

Here, we have identified using database searches and sequence analysis tools the amino-acid sequences that exhibit homology to known FlgMs, and proposed that bacterial FlgM proteins are distributed in their phylogenetic tree according to a combination of characteristics more than Gram-negative or Gram-positive classification. A computer analysis identified a group of specific residues that may be responsible for the biological differences between FlgM proteins. In addition, we generated homology models for the *B. subtilis* and *S. typhimurium* FlgM proteins, and used computational methods to provide additional and/or complementary information to the sequence-structure-function relationship data available for this family.

Materials and methods

Sequence analysis

PSI-BLAST [15] was used to search the non-redundant version of current sequence databases (nr) at NCBI (http://www.ncbi.nlm.nih.gov/). All sequences were subsequently realigned using the profile menu of the CLUSTALW program [16] to include secondary structural information and improve the alignment quality. Manual adjustments were introduced based on BLAST pairwise comparison, secondary-structure prediction, and threading results.

The COG (COG2747) and Pfam (PF04316) classification of the FlgM proteins were obtained at URL sites:http://www.ncbi.nlm.nih.gov/COG/ and http://www.sanger.ac.uk/Pfam/, respectively.

Phylogenetic analysis

The evolutionary inference of the FlgM proteins and their related sequences was performed according to the Neighbor-Joining method of Saitou and Nei [17] implemented in the CLUSTALW program [16]. The sampling variance of the distance values was estimated from 1,000 bootstrap resampling of the alignment columns.

Functional residues prediction

The SequenceSpace method [18] was used to predict residues likely to be responsible for functional differences between protein subfamilies.

Structure prediction

In search for alternative alignments between the FlgM proteins and the structural template 1rp3 we used the MetaServer [19], available at http://bioinfo.pl/meta, with the default parameters. MetaServer uses fold recognition methods such as FFAS03 [20], ORFeus [21], 3D-PSSM [22], INBGU [23], mGenTHREADER [24], SAM-T02 [25], FUGUE2 [26], and the meta-predictors PCONS [27], 3D-SHOTGUN [28] and 3D-JURY [29]. Secondary structure was predicted independently with JPRED http://www.compbio.dundee.ac.uk/~www-jpred/submit.html), a consensus prediction server [30].

Modeling

Homology modeling was carried out using the SWISS-MODEL/PROMOD II server [31] and the GROMOS force field–;implemented in the SWISS-MODEL server–;for energy minimization [32]. The input alignment to generate three-dimensional (3D) models was based on a threading-derived pairwise target-template alignment obtained from the MetaServer [19]. The stereochemical and energetic parameters of the final 3D models were evaluated by the WHATCHECK [33] and PROSA II software embedded within PROMOD II [34].

Docking

To investigate the dimer association of the FlgM proteins, we used the ClusPro web-based method accessible at URL site:http://nrc.bu.edu/cluster/ [35]. ClusPro is a fully automated docking and discrimination server that filters docked conformations with good surface complementarity, and ranks them based on their clustering properties. The server was executed with default parameters, using the DOT [36] and ZDOCK [37] docking algorithms implemented in ClusPro, and requesting for a maximum of 30 solutions for each algorithm.

Results and discussion

Sequence similarity searches

FlgM proteins belong to the orthologous group COG2747, and define the PFAM domain classification PF04316, accessible at http://www.ncbi.nlm.nih.gov/COG/new and http://www.sanger.ac.uk/Pfam/, respectively. PFAM is a comprehensive collection of protein domains and families, freely available on the web, and with a range of well-established uses, including genome annotation. PFAM currently contains over 6,000 protein families and domains matching 75% of protein sequences in Swiss-Prot and TrEMBL databases.

The FlgM amino acid sequences of *B. subtilis* (FlgM_BACSU, P39809) characterized by our group [10], and *S. typhimurium* (FlgM_SALTY, P26477) an extensively studied molecule [8, 9] were used as query to retrieve related sequences, in a non-redundant version of current sequence databases (nr) at NCBI http://www.ncbi.nlm.nih.gov/) by the profile based PSI-BLAST program [15]. We also used the SRS system http://srs.ebi.ac.uk/), with the FlgM keyword as query, to retrieve additional and divergent amino-acid sequences annotated in the databases.

The combined database searches revealed 77 amino-acid sequences that included bacterial regulatory proteins of the flagellin synthesis, and several proteins annotated as hypothetical (Table 1), both with a relatively high similarity in their C-terminus. Some of the FlgM and hypothetical sequences were detected with *e*-value worse than threshold (BLAST *e*-value >0.005), which indicated the high variability of the amino-acid sequences into the FlgM family. The proteins listed in Table 1 were not arranged by confidence. In addition, most of the 77 proteins simultaneously identified by the three queries are from the Enterobacteriaceae family. The proposed new members of the family correspond to protein sequences without identifiers in COG and PFAM databases (Table 1). Also, the new members improved the evolutionary information contained in the PFAM multiple sequence alignment and would be very useful for scientists interested in the FlgM family.

The analysis presented here also includes several Helicobacter FlgM proteins that define the PFAM domain PF05998 [6]. It is noteworthy that differences in protein length are observed even between FlgM of the same species. The Helicobacter FlgM are 76- and 67-amino-acids for the *H. pylori* proteins, and 70-amino-acids for *H. hepaticus*, while for the rest of the anti-sigma factors the length fluctuates from 131- to 65-amino-acids (on average 99-amino-acids). These differences in protein length are localized mainly to the FlgM N-terminal region (discussed below). The results showed in Table 1 might serve as an update of the PFAM and COG databases.

Homology modeling of the *S. typhimurium* and *B. subtilis* FlgM proteins

First of all, we would like to introduce, briefly, the problem of protein modeling. To start the modeling process, we have to identify the template and define an alignment between the template and target sequences. This is the single most crucial step in a modeling process. Any errors at this stage are usually impossible to correct later and can lead to significant errors in the models. If the target and the template proteins are closely related, there is no problem both with the template identification and creation of the alignment. However, the problem becomes increasingly difficult as the sequence homology between the target and the possible template protein becomes more distant and sequence similarity weaker. Sequence alignments become unstable with sequence-similarity lower than 40% of identical residues. For the 3D-models of FlgM_BACSU and FlgM_SALTY proteins, presented below, the sequence identities with respect to the coordinates of *A. aeolicus* FlgM (PDB code: 1rp3) [38] are 32.05 and 28.6%, respectively.

The alignment provided by a database search (PSI-BLAST) is usually not optimal and often includes only regions of high similarity between the query sequence and the database hits, so that it is necessary to realign the selected template [39]. In addition, the PSI-BLAST algorithm was not optimized for alignment accuracy. The authors of this most popular and most effective sequence-similarity search method had to sacrifice to some extent the alignment accuracy for speed, which is the fundamental parameter when performing searches in large sequence databases [40].

To improve the multiple sequence alignment (Fig. 1) and to provide a structural framework for the interpretation of sequence-structure-function relationships, we attempted to predict the 3D-structure of *S. typhimurium* and *B. subtilis* FlgM, using sequence-to-structure threading and homology modeling. The rationale behind this approach is that most of the alignment errors that are undetectable at the level of primary and secondary structure would manifest themselves in the model. They could be identified and corrected by computer software for the evaluation of 3D-structures, followed by the analysis of graphic representations with a trained eye.

The amino-acid sequences of FlgM_SALTY and FlgM_BACSU that represented individual subfamilies were submitted to the MetaServer, which combines several

Table 1 FlgM proteins and its homologous sequences used in this study

Identifier	Length	COGs identifier	UniProt/NCBI acc. no.	PFAM identifier	Source
FlgM_AERHY	106 aa	–	Q8GLQ2	PF04316	*Aeromonas hydrophila*
FlgM_AQUAE	88 aa	COG2747	O66683	PF04316	*Aquifex aeolicusVF5*
Hyp_AZOVI	95 aa	–	ZP_00089092	–	*Azotobactervinelandii*
FlgM_BACHA	86 aa	COG2747	Q9K6V2	PF04316	*BacillushaloduransC-125*
FlgM_BACLI	88 aa	–	Q65EB1	PF04316	*Bacillus licheniformis* DSM 13
FlgM_BACSU	88 aa	COG2747	P39809	PF04316	*Bacillussubtilis168*
FlgM_Bdebac	109 aa	–	Q6MQD5	PF04316	*Bdellovibrio bacteriovorus* HD100
FlgM_Borpa	96 aa	–	Q7WA99	PF04316	*Bordetella parapertussis* 12822
FlgM_Borpe	96 aa	–	Q7VYH1	PF04316	*Bordetella pertussis* Tohama I
FlgM_Burce2	110 aa	COG2747	ZP_00224113	PF04316	*Burkholderia cepacia* R1808
FlgM_Burce1	109 aa	COG2747	ZP_00213011	–	*Burkholderia cepacia* R18194
Hyp_BURFU	117 aa	–	ZP_00032168	–	*Burkholderiafungorum*
FlgM_Burps	114 aa	–	Q62ET6	PF04316	*Burkholderia pseudomallei* K96243
Hyp_Camje	65 aa	–	Q9PMJ6	–	*Campylobacter jejuni* NCTC 11168
FlgM_Chrvi	99 aa	–	Q7NZC2	PF04316	*Chromobacterium violaceum* DSM 30191
FlgM_CLOAC	93 aa	COG2747	Q97H01	PF04316	*Clostridiumacetobutylicum DSM 792*
Hyp_CLOTH	100 aa	–	ZP_00060838	–	*Clostridium* thermocellum *ATCC 27405*
FlgM_Decar	98 aa	COG2747	ZP_00348428	–	*Dechloromonas aromatica* RCB
Hyp_DESHA	91 aa	–	ZP_00098295	–	*Desulfitobacteriumhafniense*
FlgM_Desps	93 aa	–	Q6AJR4	–	*Desulfotalea psychrophila* LSv54
FlgM_Desvu2	104 aa	–	Q72EP6	PF04316	*Desulfovibrio vulgaris* NCIMB 8303
FlgM_Erwca	98 aa	–	Q6D6I1	–	*Erwinia carotovora* SCRI1043
FlgM_ECOLI	97 aa	COG2747	P43532	PF04316	*EscherichiacoliK-12*
FliA_ECOLB	97 aa	–	Q8X8M4	PF04316	*Escherichia coli* O157:H7
FlgM_Exisp	72 aa	COG2747	ZP_00183361	–	*Exiguobacterium sp*. 255–15
Hyp_Geome1	97 aa	–	ZP_00080517	–	*Geobacter metallireducens*
Hyp_Geome2	104 aa	–	ZP_00079933	–	*Geobacter metallireducens*
FlgM_Geosu1	95 aa	–	Q748F7	–	*Geobacter sulfurreducens* PCA
FlgM_Geosu2	102 aa	–	Q74B56	PF04316	*Geobacter sulfurreducens* PCA
FlgM_HELHE	70 aa	–	AAP77471	–	*Helicobacter hepaticus* 3B1
FlgM_Helpi1	67 aa	–	Q8VN27	–	*Helicobacter pyroli* CC28C
FlgM_Helpi2	76 aa	–	Q8VN34	–	*HelicobacterpyroliBO242*
FlgM_Helpi3	76 aa	–	Q8VN30	PF05998	*Helicobacter pyroli* N6
FlgM_Helpi4	76 aa	–	Q8VN31	PF05998	*Helicobacter pyroli* SS1
FlgM_Helpi5	76 aa	–	Q8VN32	PF05998	*HelicobacterpyroliCC7C*
FlgM_Helpi6	76 aa	–	Q8VN33	PF05998	*Helicobacter pyroli NCTC11637*
FlgM_Helpi7	76 aa	–	Q8VN35	PF05998	*Helicobacter pyroli* BO255
FlgM_Helpi8	76 aa	–	Q8VN29	PF05998	*Helicobacter pyroli* RE8029
FlgM_Helpi9	67 aa	–	Q8VN26	PF05998	*Helicobacter pyroli* CC48A
FlgM_Helpi10	67 aa	–	Q8VN28	PF05998	*Helicobacter pyroli* CC29C
FlgM_Legpn	106 aa	–	YP_094941	–	*Legionella pneumophila*
Hyp_MAGSP	110 aa	–	ZP_00044273	–	*Magnetococcussp.MC-1*
FlgM_Metfl	100 aa	COG2747	ZP_00173433	–	*Methylobacillus flagellatus* KT
Hyp_MICDE	105 aa	–	ZP_00065049	–	*Microbulbiferdegradans2–40*
FlgM_Mooth	96 aa	COG2747	ZP_00329913	–	*Moorella thermoacetica* ATCC 39073
Hyp_NITEU	107 aa	–	Q82S76	PF04316	*Nitrosomonas europaeaIFO 14298*
FlgM_OCEIH	85 aa	–	Q8ENH4	PF04316	*Oceanobacillus iheyensis* HTE831
FlgM_Phopr2	104 aa	–	Q6LTR7	PF04316	*Photobacterium profundum* SS9
FlgM_Pholu	100 aa	–	Q7N5N5	PF04316	*Photorhabdus luminescens* TT01
FlgM_PROMI	99 aa	COG2747	P96974	PF04316	*Proteus mirabilisU6540*
FlgM_Pseae1	107 aa	–	Q79SU4	PF04316	*Pseudomonas aeruginosa* PAK
FlgM_PSEAE	107 aa	COG2747	Q9HYP5	PF04316	*Pseudomonas aeruginosaPAO1*
Hyp_PSEFL	131 aa	–	ZP_00084200	–	*Pseudomonas fluorescens* PfO-1
FlgM_PSEPU	104 aa	–	Q88EQ8	PF04316	*Pseudomonas putida* KT2440

Table1 (continued)

Identifier	Length	COGs identifier	UniProt/NCBI acc. no.	PFAM identifier	Source
Hyp_PSESY	104 aa	–	ZP_00127307	–	*Pseudomonas syringae* B728a
FlgM_PSESY	104 aa	–	Q885B0	PF04316	*Pseudomonas syringae* DC3000
FlgM_Raleu	102 aa	COG2747	ZP_00167918	–	*Ralstonia eutropha* JMP134
Hyp_RALME	97 aa	–	O51794	–	*Ralstonia(Wautersia)metalliduransCH34*
Identifier	Length	COGs identifier	UniProt/NCBI acc. no.	PFAM identifier	Source
FlgM_RALSO	106 aa	–	Q8XSX7	PF04316	*Ralstonia solanacearumGMI1000*
Hyp_RHOSP	104 aa	–	ZP_00006078	–	*Rhodobacter sphaeroides*
FlgM_Rubge	102 aa	COG2747	ZP_00242030	–	*Rubrivivax gelatinosus* PM1
FlgM_SALEN	97 aa	–	Q8Z7K6	PF04316	*Salmonella typhiCT18*
FlgM_SALTY	97 aa	–	P26477	PF04316	*Salmonella typhimuriumLT2*
FlgM_SHEON	106 aa	–	Q8EC90	PF04316	*Shewanella oneidensis MR-1*
FlgM_Symth	99 aa	–	Q67K33	–	*Symbiobacterium thermophilum* IAM 14863
Hyp_TERMA	93 aa	COG2747	Q9WXU0	PF04316	*Thermotoga maritima strain MSB8*
FlgM_THETE	91 aa	–	Q8RCE0	–	*Thermoanaerobacter tengcongensisMB4*
FlgM_Thide	98 aa	COG2747	ZP_00335231	–	*Thiobacillus denitrificans* ATCC 25259
FlgM_VIBCH	107 aa	COG2747	Q9KQ03	PF04316	*Vibrio cholerae* O1
FlgM_VIBFI	103 aa	–	Q8GM70	PF04316	*Vibrio fischeriES114*
LfgM_Vibpa1	105 aa	COG2747	Q9X9K5	PF04316	*Vibrio parahaemolyticusBB22*
LfgM_Vibpa2	93 aa	COG2747	Q56717	PF04316	*Vibrio parahaemolyticus* O3:K6
FlgM_VIBVU	108 aa	–	Q8DFI1	PF04316	*Vibrio vulnificus CMCP6*
Flg_XANAX1	103 aa	–	Q8PL17	–	*Xanthomonas axonopodis306*
Flg_XANAX2	103 aa	–	Q8P9B0	PF04316	*Xanthomonas campestris NCPPB 528*
FlgM_YEREN	99 aa	COG2747	Q57401	PF04316	*Yersinia enterocolitica O:8*
FlgM_YERPE	100 aa	–	Q8ZFC0	PF04316	*Yersinia pestisCO-92*

programs for predicting secondary structure, solvent accessibility and fold recognition (i.e. detection of the known structure) that are most compatible with the query sequence (see Materials and methods). The goal is to find alternative alignments between the FlgM_SALTY and FlgM_BACSU amino-acid sequences, and the template structure 1rp3.

In the case of FlgM_BACSU, the MetaServer found alternative alignments with significant 3D-JURY scores for FFAS03 (score=25.14), 3D-PSSM (score=24.86), FUGUE2 (score=24.86), mGenTHREADER (score=23.14), BLAST (score=30.14), and 3D-SHOTGUN (score=34.86). Consequently, the amino-acid sequence of FlgM_BACSU and the alignments between FlgM_BACSU and 1rp3B produced by the MetaServer were submitted to the Swiss-Model server for homology modeling. For the FlgM_SALTY sequence, the MetaServer found alternative alignments with significant 3D-JURY scores for FUGUE2 (score=27.14), 3D-PSSM (score=26.14), mGenTHREADER (score=25.43), FFAS03 (score=23.29), PCONS (score=38.00), and 3D-SHOTGUN (score=36.57). We therefore submitted the resulting alignments between FlgM_SALTY and 1rp3B to the SWISS-MODEL server.

Refinement of the models was done in parallel with refinement of the multiple sequence alignment. This procedure was utilized until the final model could not be improved further. The final multiple sequence alignment is shown in Fig. 1, and 3D models are available as supplementary material ESM 1. We also provide the alternative alignments obtained by MetaServer as supplementary material ESM 2.

Sequence conservation and evolutionary relationships in the FlgM family

We have compared the 77 amino-acid sequences of FlgM anti-sigma factors and their related sequences using the profile-alignment option of the CLUSTALW program. The pattern of the secondary structures (helix 2 to helix 4) observed in the 3D-structure of the *A. aeolicus* FlgM (1rp3B) [38] and that predicted for individual subfamilies using JPRED agreed very well with the alignment reported in this work. Only helix 1 is missing from the consensus prediction of JPRED (see the supplementary material ESM 3).

The phylogenetic relationship of FlgM and their related sequences is shown in Fig. 2. The unrooted tree revealed that bacterial FlgM originates from at least three phylogenetically distinct groups. Group I includes FlgM mainly from Gram-negative enteric bacteria, and many of them from human and plant pathogens. Group II includes FlgM from gram-positive and gram-negative non-pathogenic bacteria with the exception of pathogens *H. pylori*, *H. hepaticus*, *Campylobacter jejuni*, and *Vibrioparahaemolyticus*. Group 3 comprises FlgM proteins from the archaea *Thermotogamaritima* and the hyperthermophilic bacterium *A. aeolicus*, two organisms, which

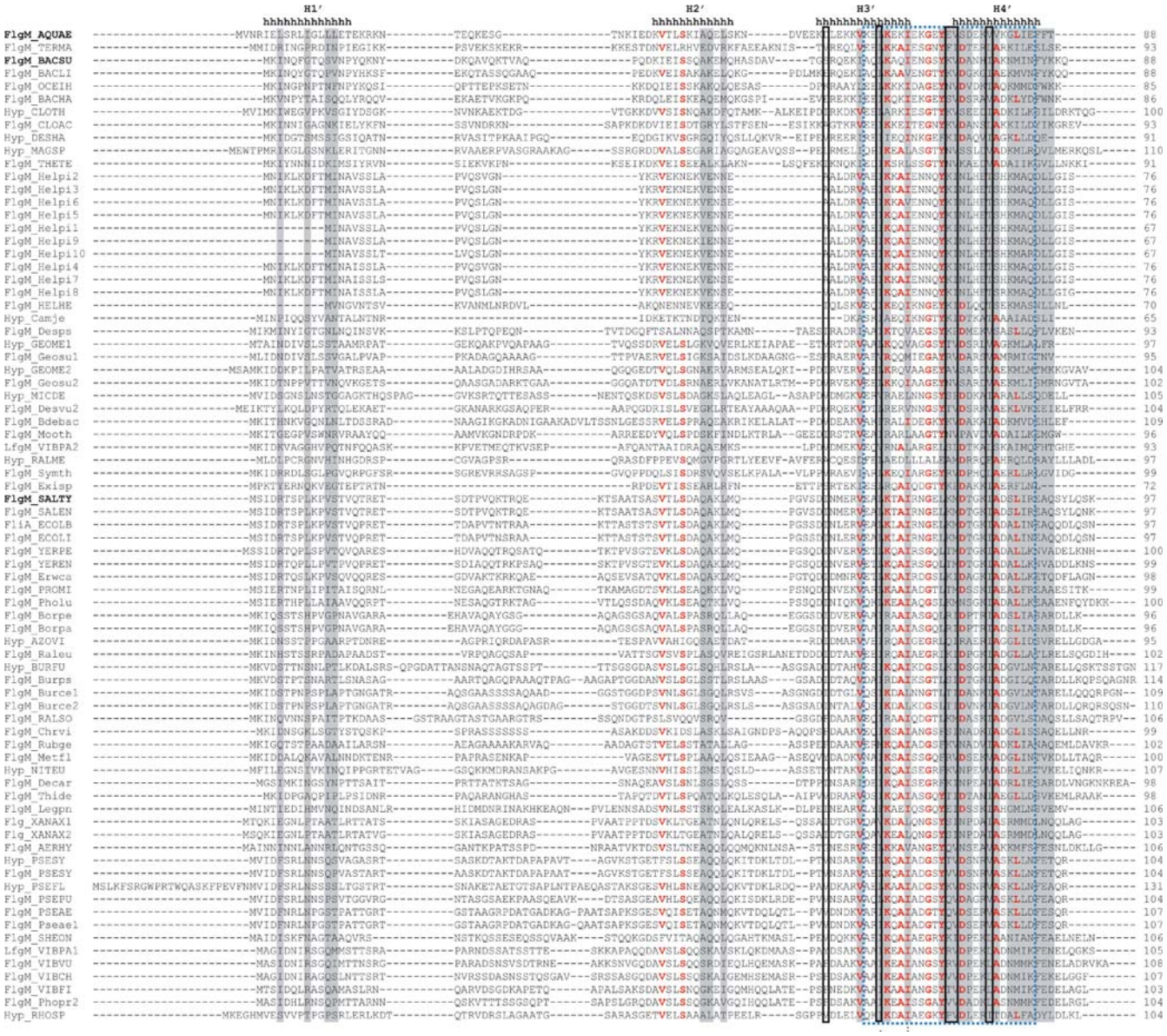
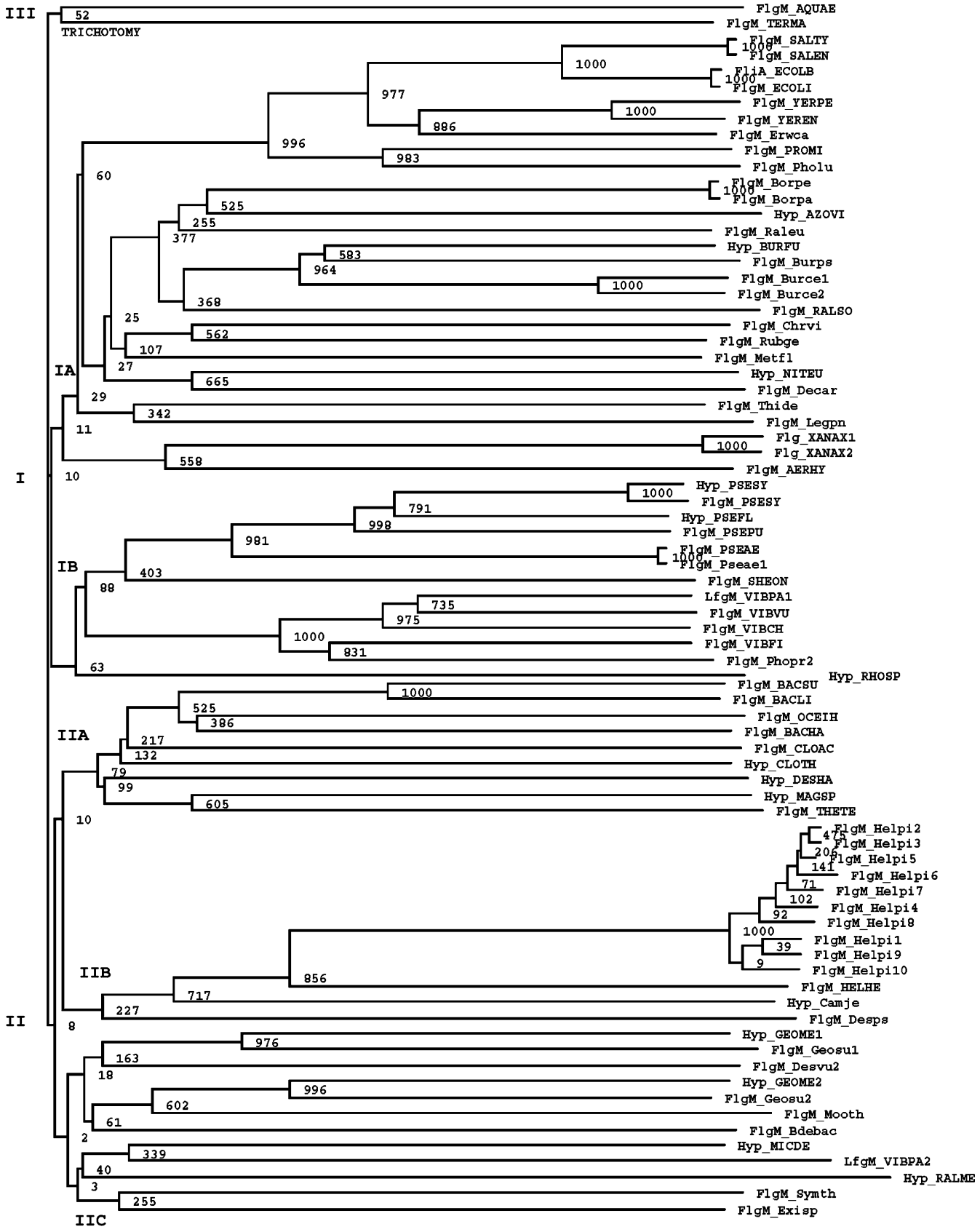


Fig. 1 Multiple alignment of the FlgM family. Protein sequences for which amino-acid identity was 100% were not included. Highly conserved residues are highlighted in *red*, according to the column score parameters option in the quality menu of CLUSTALX [45]. *Gray* shading of residues denotes those contacting σ^{28} [38]. *Boxes* represent the amino-acid positions where mutants impaired in σ^{28} binding [9]. *Numbers* show the size of the proteins. The binding region (Glu64–Arg88, *dotted line* in *blue*) identified in *Salmonella-typhimurium* [8] is shown

grow at very high temperatures, 80 and 95°C, respectively. Interestingly, the *Thermoanaerobacter tengcongensis* MB4 grows at 80°C, similar to organisms included into group 3, but its genes are similar to those of *Bacillus halodurans* (group IIA; see http://www.genomenewsnetwork.org/). Therefore, FlgM from *T. tengcongensis* MB4 is included into group IIA that contains proteins from gram-positive bacteria.

On the other hand, the distribution of FlgM proteins into the unrooted tree in Fig. 2 is not explained only by simple Gram-positive or Gram-negative classification, but also by a combination of characteristics such as: pathogenic, non-pathogenic, thermophilic or hyperthermophilic bacteria, and the regulation of genes by sigma-factors σ^{28} and σ^{54}. It should be noted that sequences obtained from preliminary data and cDNA clones may contain errors that can influence the outcome of the comparative analysis. How-

Fig. 2 Unrooted tree showing phylogenetic branches of the FlgM ▶ family. The numbers at the nodes indicate the statistical support of the branching order by the bootstrap criterion. The nodes with bootstrap support <50% are shown as unresolved. The *bar* at the *bottom* of the phylogram indicates the evolutionary distance, to which the branch lengths are scaled based on the estimated divergence


III
52
TRICHOTOMY
FlgM_AQUAE
FlgM_TERMA
FlgM_SALTY
1000
FlgM_SALEN
1000
FliA_ECOLB
1000
FlgM_ECOLI
977
FlgM_YERPE
1000
FlgM_YEREN
886
FlgM_Erwca
996
FlgM_PROMI
983
FlgM_Pholu
60
FlgM_Borpe
1000
FlgM_Borpa
525
Hyp_AZOVI
255
FlgM_Raleu
377
Hyp_BURFU
583
FlgM_Burps
964
FlgM_Burce1
1000
FlgM_Burce2
368
FlgM_RALSO
25
FlgM_Chrvi
562
FlgM_Rubge
107
FlgM_Metfl
27
IA
Hyp_NITEU
665
FlgM_Decar
29
FlgM_Thide
342
FlgM_Legpn
11
Flg_XANAX1
1000
Flg_XANAX2
558
FlgM_AERHY
I
10
Hyp_PSESY
1000
FlgM_PSESY
791
Hyp_PSEFL
998
FlgM_PSEPU
981
FlgM_PSEAE
1000
FlgM_Pseae1
IB
403
FlgM_SHEON
88
LfgM_VIBPA1
735
FlgM_VIBVU
975
FlgM_VIBCH
1000
FlgM_VIBFI
831
FlgM_Phopr2
63
Hyp_RHOSP
FlgM_BACSU
1000
FlgM_BACLI
525
FlgM_OCEIH
386
FlgM_BACHA
IIA
217
FlgM_CLOAC
132
Hyp_CLOTH
79
Hyp_DESHA
99
Hyp_MAGSP
605
FlgM_THETE
10
FlgM_Helpi2
475
FlgM_Helpi3
206
FlgM_Helpi5
141
FlgM_Helpi6
71
FlgM_Helpi7
102
FlgM_Helpi4
92
FlgM_Helpi8
1000
FlgM_Helpi1
39
FlgM_Helpi9
9
FlgM_Helpi10
IIB
856
FlgM_HELHE
717
Hyp_Camje
227
FlgM_Desps
II
8
Hyp_GEOME1
976
FlgM_Geosu1
163
FlgM_Desvu2
18
Hyp_GEOME2
996
FlgM_Geosu2
602
FlgM_Mooth
61
FlgM_Bdebac
2
Hyp_MICDE
339
LfgM_VIBPA2
40
Hyp_RALME
3
FlgM_Symth
255
FlgM_Exisp
IIC

.10

ever, the topology of the branching pattern is supported by bootstrap analysis, and the phylogenetic groups correlate well with the presence of sequence signatures derived from a common ancestor.

Transfer of function annotation from one member of the family to others, based only on amino-acid sequence identity values, is not a trivial task mainly when sequence identity values decrease [41]. The phylogenetic analysis presented here assists to this purpose. For example, the flagellar biogenesis in *V. parahaemolyticus*, *V. cholerae*, and *Pseudomonasaeruginosa* (group IB), and *H. pyroli* (group IIB) is regulated by both σ^{28} and σ^{54}, in contrast to the regulation described in the *S. typhimurium* system (group IA), where none of the flagellar genes are regulated by σ^{54} [7, 42]. The other members into the phylogenetic groups IA, IB and IIB would share the σ^{28} and/or σ^{54} regulation property more reliably than FlgM outside the groups.

Analysis of the multiple sequence alignment (Fig. 1) revealed amino-acid residues that are conserved among all or most of the individual family members as well as some differences between the subfamilies (Fig. 2). A Sequence-Space analysis identified a group of specific residues, which might be responsible for the biological differences between FlgM proteins (see the supplementary material ESM 4).

Another peculiarity emerging from the alignment is that despite the differences in protein length, which are localized mainly to the FlgM N-terminal region, all proteins listed in Table 1 would have the capacity to interact with sigma factors because of their high sequence similarity into the minimal binding domain [8] that is represented as a discontinuous blue line in Fig. 1. For example, experimental evidence indicates that *H. pylori* FlgM is able to interact with the *S. typhimurium* σ^{28} (FliA) and inhibits the expression of FliA-dependent genes in Salmonella, although it lacks about 20 amino-acid residues at the N-terminal region of enterobacterial FlgM proteins [6].

Sequence-structure-function relationships of FlgM proteins

The recently determined crystal structure of the *A. aeolicus* σ_{28}/FlgM complex, provided a detailed explanation of the inhibition mechanism of RNAP by FlgM proteins [38]. However, combining the crystallographic data, the new FlgM-related sequences (Table 1), and their multiple alignment, we provide additional and/or complementary information to the analysis of the RNAP inhibitory mechanism, and other aspects concerning the FlgM recognition by the flagellar export apparatus and the dimer association (see below).

In order to explain some of the sequence-structure-function relationship data available for the *B. subtilis* and *S. typhimurium* FlgM proteins, and since the efficiency of the experimental approach for identifying functional residues is enhanced considerably by the insights that the 3D-structure of the protein can provide, we used the crystal structure of the *A. aeolicus* FlgM to generate homology models. Fig. 3a shows the structural superposition between the homology models and the crystallographic structure, and reveals that the main differences are localized to the N-terminal region.

From the alignment in Fig. 1, we observed a few amino-acid positions (Val37, Ser40, Val60, Lys64, Ala66, Ile67, Gly70, Tyr72, Ala80, and Leu83) that are highly conserved in the FlgM family (Table 2, Fig. 3b,c). Only four of them (Val37, Ser40, Ala66, and Gly70; highlighted in red without gray shading in Fig. 1) are not included in the list of residues interacting with σ_{28} (Fig. 3 in [38]) (highlighted in gray in Fig. 1). According to the 3D-structure of *A. aeolicus* σ_{28}/FlgM complex, Val60, Ile67, and Tyr72 display van der Waals interactions with conserved hydrophobic residues in σ_{28} [38]. The amino acids Val37 and Ser40 localize to the FlgM-H2' region, while Ala66 (FlgM-H3') and Gly70 (the loop connecting H2'-H3') are in the minimal binding domain [8].

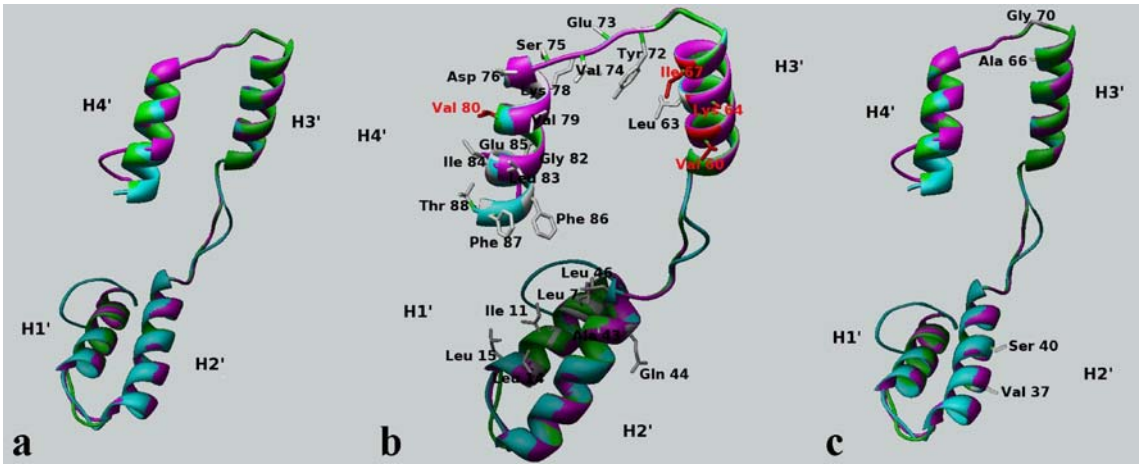


Fig. 3 Structural superposition between *A. aeolicus*, *S. typhimurium*, and *B. subtilis* FlgMs, and the spatial localization of the interacting residues between FlgM and σ^{28}.**a** Ribbon diagram showing the structural superposition between *A. aeolicus* (*green*), *S. typhimurium* (*magenta*), and *B. subtilis* (*cyan*) FlgMs. Highly conserved residues are in *red*, according to the column score parameters option in the quality menu of CLUSTALX [45]. Helices are labeled H1' to H4' as in [38]. **b** Residues listed in [38], which interact with σ^{28}. **c** Conserved residues not included in [38]. The figure was generated using the CHIMERA program [47]

Table 2 Interaction between FlgM and σ^{28}

Residue	Location in 1rp3	Contacting in σ^{28}	Equivalent in St	Equivalent in Bs	Variability	ASA (%)*
Leu7	H1'	(σ_2)	Pro8	Thr8	9	6.4
Ile11	H1'	(σ_2)	Val12	Val11	12	1.9
Leu14	H1'	(σ_2)	Val15	Tyr14	13	5.9
Leu15	H1'	(σ_2)	Gln16	Gln15	15	21.1
Ala43	H2'	(σ_2)	Asp45	Ala40	30	0
Gln44	H2'	(σ_2)	Ala46	Lys41	29	41.1
Leu46	H2'	(σ_2)	Ala48	Met43	18	12.1
Val60	H3'	(σ_4)	Val63	Ile58	75	7.7
Leu63	H3'	(σ_4)	Leu66	Leu61	62	13.2
Lys64	H3'	(σ_4)	Lys67	Lys62	68	15.9
Ile67	H3'	(σ_4)	Ile70	Ile65	77	23.7
Tyr72	H4'	(σ_4)	Leu75	Tyr70	26	20.5
Glu73	H4'	(σ_4)	Lys76	Lys71	36	59.6
Val74	H4'	(σ_4)	Met77	Val72	51	37
Ser75	H4'	(σ_4)	Asp78	Asp73	50	34.8
Asp76	H4'	(σ_4)	Thr79	Ala74	21	24.2
Lys78	H4'	(σ_4)	Lys81	His76	33	30.6
Val79	H4'	(σ_4)	Ile82	Ile77	35	0.7
Val80	H4'	(σ_4)	Ala83	Ala78	76	5.5
Gly82	H4'	(σ_4)	Ser85	Asn80	29	3.6
Leu83	H4'	(σ_4)	Leu86	Met81	57	0.6
Ile84	H4'	(σ_4)	Ile87	Ile82	31	14.8
Glu85	H4'	(σ_4)	Arg88	Asn83	33	65.4
Phe86	H4'	(σ_2)	Glu89	Phe84	8	5.6
Phe87	H4'	(σ_3–σ_4) linker	Ala90	Tyr85	11	3.4
Thr88	H4'	(σ_4)	Gln91	Lys86	14	8.9

The last two columns show the amino acid variability values calculated by the CLUSTALX program [45], and the solvent accessibility surface area (ASA) calculated by WHAT IF [46], respectively. *According to the *Aquifexaeolicus* σ_{28}/FlgM complex (PDB code: 1rp3). *Salmonellatyphimurium* (St), *Bacillussubtilis* (Bs). High conserved residues are highlighted in red, according to the column score parameters option in the quality menu of CLUSTALX

Among the *A. aeolicus* FlgM residues interacting with the σ_{28} subunit, only Val60, Lys64, and Ile67 located in helix 3, and Val80 located in helix 4 are the highly conserved ones. These residues are placed in the same face of the helix-loop-helix motif (Fig. 3b) and interact with the σ_4 domain of the σ_{28} subunit. Therefore, our results suggest an essential role for the FlgM residues Val60, Lys64, Ile67, and Ala80.

FlgM recognition by the flagellar export apparatus

FlgM proteins not only regulate the transcription of flagellar genes, but also sense the developmental state of the flagellum, being a substrate for secretion through the flagellum-specific type III secretion pathway [3].

Since a portion of the N-terminal 40 amino acids of *S. typhimurium* FlgM are essential for export (Ser7-Val25) [8] and the NMR resonances for these residues show no significant chemical-shift or line-shape changes in the presence of σ^{28}, Daughdrill and colleagues proposed that the export apparatus might recognize an export signal in the N-terminal portion of FlgM both when FlgM is free in solution or bound to σ^{28} [9]. Based on the multiple alignment presented here (Fig. 1), we observed that *A. aeolicus* FlgM region Arg9-Glu27 is equivalent to the *S. typhimurium* FlgM region Ser7-Val25, and in the 3D structure of *A. aeolicus* σ28/FlgM complex, Arg9-Glu27 is localized in the FlgM H1'-H2' region that occludes the β' coiled-coil binding determinant for the σ_2^{28} domain. In addition, although the region Glu18–Glu27 in the *A. aeolicus* FlgM H1'–H2' was not modeled, the interaction of σ_2^{28} with the RNAP β' coiled-coil or FlgM H1'–H2' appears to be mutually exclusive [38], and according to the 3D-structure, only FlgM Glu16 (ASA=70.8%), in the FlgM H1'–H2' region, has a side chain completely exposed (ASA≥50%). Therefore, based on the sequence and structural analysis presented in this study, we suggest that proteins from the export apparatus may recognize an export signal in the N-terminal portion of FlgM only when FlgM is free in solution, in contrast to the suggestion by Daughdrill and colleagues [9].

Dimer association between FlgM proteins

Previously, our group characterized the *B. subtilis* FlgM and found this molecule associated as a dimer in solution [10]. To the best of our knowledge, similar results for other FlgM molecules have not been published. However, oligomerization was reported for a different subfamily, the anti-σ^{70} (AsiaA). AsiA is a symmetric dimer in solution, and interacts with σ^{70} as a monomer via the same residues used for dimerization [14]. Both, *B. subtilis* FlgM and AsiA, have an α-helical fold; FlgM contains four α-helices, while Asia contains six. In which way the dimer association is a general mechanism for the anti-sigma factors has not been investigated yet.

In the present work, we used the ClusPro web-based method to study the possible orientation between two FlgM protomers (dimer association) for the *A. aeolicus*, *S. typhimurium*, and *B. subtilis*. In Table 3 we summarize the results of the ClusPro web server. The total number of putative conformations analyzed was 180, which are distributed as follows: 60 solutions for the DOT and ZDOCK programs (a maximum of 30 for each algorithm) using the crystal structure of the *A. aeolicus* FlgM, and 120 solutions using the homology models of the *S. typhimurium* and *B. subtilis* FlgM (see the supplementary material ESM 5).

The five different conformations showed in Table 3 are those compatible with the helix–helix packing regularities [43], and they are frequently observed in nature. The rest of the putative conformations are less 'realistic' and they are not shown in Table 3 (see the supplementary material).

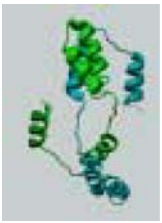
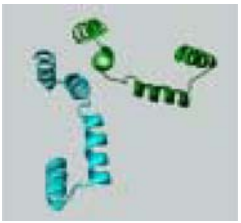
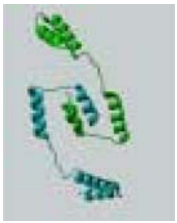
This analysis revealed that the most frequent solutions are the three- and four-helix bundle (H'_{2A}–H'_{3B}–H'_{1A}, and H'_{1A}–H'_{2A}–H'_{1B}–H'_{2B}, respectively) involving the FlgM N-terminal region. Interestingly, these conformations agree with our previous results about the limited proteolysis of the *B. subtilis* FlgM [10]. Furthermore, in agreement with the results of Urbauer and colleagues [14], where AsiA interacts with σ^{70} via the same residues used for dimerization; the FlgM N-terminal region (H1'–H2') involved in dimer association, occludes the RNAP β' coiled-coil binding determinant for the σ_2^{28} domain [38].

On the other hand, this result also concurs with the fact that the C-terminal half of *S. typhimurium* FlgM gains structure inside *E. coli* cells under physiologically relevant conditions in vitro [44]. The results provided by Dedmon and colleagues [44] support the hypothesis that there are two classes of intrinsically disordered proteins, with FlgM providing an example of each class. One class, exemplified by the C-terminal half of FlgM, is structured in cells. The driving force for solute-induced structure is likely the formation of a hydrophobic core, which is the most common characteristic of folded proteins. The other class, exemplified by the N-terminal half of FlgM, does not become structured at physiologically relevant solute concentrations. Some of these proteins may require another protein to provide a framework for structure formation. Accordingly, the dimer association involving the N-terminal half of FlgM, possibly will help in stabilizing the helical conformation necessary to carry out the biological function.

Conclusion

In the present manuscript, we updated the multiple sequence alignment of previously characterized cluster of orthologs (COG2747) and the PFAM classification (PF04316) for the FlgM family. The protein sequences annotated as 'hypothetical' could represent genuine FlgM proteins; however, their function remains to be determined experimentally. Furthermore, the phylogenetic tree of 77 protein sequences revealed the presence of at least three

Table 3 ClusPro results. DOT-# / ZDOCK-#: the name corresponds to the docking algorithm used and # is the rank of the solution. H'ij: alpha helix secondary structure, where i is the helix number in the FlgM structure (i= 1 to 4), and j is the FlgM protomer A (cyan) and B (green), respectively

					
Dimmer interface	H'_{1A}, H'_{2A}, H'_{1B}, H'_{2B}	H'_{2A}, H'_{3B}, H'_{1A}	H'_{1A}, H'_{1B}, H'_{2A}, H'_{2B}	H'_{2A}, H'_{2B}	H'_{4A}, H'_{4B}
A. Aeolicus	DOT-10, DOT-13	DOT-20	ZDOCK-1	ZDOCK-9	ZDOCK-16
FlgM	DOT-25, ZDOCK-5	DOT-22	ZDOCK-7	ZDOCK-10	ZDOCK-26
S. typhimurium	DOT-10, DOT-18	DOT-17			ZDOCK-14
FlgM	DOT-25, ZDOCK-28	DOT-26			DOT-13
		DOT-28			
		ZDOCK-10			
B. subtilis	DOT-1, ZDOCK-4	DOT-4			ZDOCK-5
FlgM		DOT-11			ZDOCK14

major sequence clades within the FlgM family. By combining the evolutionary information extracted from the multiple alignments of FlgM and the analysis of the crystal structure of *A. aeolicus* σ_{28}/FlgM complex, we proposed an essential role for the FlgM residues Val60, Lys64, Ile67, and Ala80. We also applied the ClusPro method to the crystal structure of *A. aeolicus* FlgM and the 3D models of *S. typhimurium* and *B. subtilis* homologous proteins. Our results revealed that FlgM could associate as dimer involving their N-terminal half, which in turn will help in stabilizing its helical conformation. The results presented here can be helpful for understanding how FlgM is associated as dimer, and how the flagellar export apparatus recognizes the FlgM molecules.

Acknowledgements We would like to thank Dr. Miriam Ojeda, Dr. Lila Castellanos, and Dr. Aurora Pérez-Gramatges for critical reading of the manuscript.

References

1. Chadsey MS, Hughes KT (2001) J Mol Biol 306:915–929
2. Ohnishi K, Kutsukake K, Suzuki H, Iino T (1992) Mol Microbiol 6:3149–3157
3. Chilcott GS, Hughes KT (2000) Microbiol Mol Biol Rev 64:694–708
4. Caramori T, Barailla D, Nessi C, Sacchi L, Galizzi A (1996) J Bacteriol 178:3113–3118
5. Correa NE, Barker JR, Klose KE (2004) J Bacteriol 186: 4613–4619
6. Josenhans C, Niehus E, Amersbach S, Horster A, Betz C, Drescher B, Hughes KT, Suerbaum S (2002) Mol Microbiol 43:307–322
7. Frisk A, Jyot J, Arora SK, Ramphal R (2002) J Bacteriol 184:1514–1521
8. Iyoda S, Kutsukake K (1995) Mol Gen Genet 249:417–442
9. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW (1997) Nat Struct Biol 4:285–291
10. Bertero MG, Gonzales B, Tarricone C, Ceciliani F, Galizzi A (1999) J Biol Chem 274:12103–12107
11. Helmann JD (1999) Curr Opin Microbiol 2:135–141
12. Chadsey MS, Karlinsey JE, Hughes KT (1998) EMBO J 17:3123–3136
13. Aldridge P, Hughes KT (2002) Curr Opin Microbiol 5:160–165
14. Urbauer JL, Simeonov MF, Urbauer RJ, Adelman K, Gilmore JM, Brody EN (2002) Proc Natl Acad Sci USA 99:1831–1835
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402
16. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
17. Saitou N, Nei M (1987) Mol Biol Evol 4:406–425
18. Casari G, Sander C, Valencia A (1995) Nat Struct Biol 2: 171–178
19. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Bioinformatics 17:750–751
20. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Protein Sci 9:232–241
21. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L (2003) Nucleic Acids Res 31:3804–3807
22. Kelley LA, McCallum CM, Sternberg MJ (2000) J Mol Biol 299:501–522
23. Fischer D (2000) Pac Symp Biocomput 5:116–127
24. Jones DT (1999) J Mol Biol 287:797–815
25. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R (2001) Proteins (Suppl) 5:86–91
26. Shi J, Blundell TL, Mizuguchi K (2001) J Mol Biol 310: 243–257
27. Lundström J, Rychlewski L, Bujnicki J, Elofsson A (2001) Protein Sci 10:2354–2362
28. Fischer D (2003) Proteins 51:434–441
29. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) Bioinformatics 19:1015–1018
30. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Bioinformatics 14:892–893
31. Schwede T, Kopp J, Guex N, Peitsch MC (2003) Nucleic Acids Res 31:3381–3385
32. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF (1999) J Phys Chem 103:3596–3607
33. Hooft RW, Vriend G, Sander C, Abola EE (1996) Nature 381:272
34. Sippl MJ (1993) Proteins 17:355–362
35. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) Bioinformatics 20:45–50
36. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001) Protein Eng 14: 105–113
37. Chen R, Li L, Weng Z (2003) Proteins 52:82–87
38. Sorenson MK, Ray SS, Darst SA (2004) Mol Cell 14:127–138
39. Tramontano A (1998) Methods 14:293–300
40. Jaroszewski L, Rychlewski L, Godzik A (2000) Protein Sci 9:1487–1496
41. Devos D, Valencia A (2000) Proteins 41:98–107
42. Prouty MG, Correa NE, Klose KE (2001) Mol Microbiol 39:1595–1609
43. Eilers M, Patel AB, Liu W, Smith SO (2002) Biophys J 82:2720–2736
44. Dedmon MM, Patel CN, Young GB, Pielak GJ (2002) Proc Natl Acad Sci USA 99:12681–12684
45. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) Nucleic Acids Res 25:4876–4882
46. Vriend G (1990) J Mol Graphics 8:52–56
47. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) J Comput Chem 25: 1605–1612