



Coevolutionary Analysis Reveals a Conserved Dual Binding Interface between Extracytoplasmic Function σ Factors and Class I Anti- σ Factors

Delia Casas-Pastor,^a Angelika Diehl,^{a,b}  Georg Fritz^{a,b}

^aCenter for Synthetic Microbiology (SYNMIKRO), Philipps-University Marburg, Marburg, Germany

^bSchool of Molecular Sciences, University of Western Australia, Perth, Australia

ABSTRACT Extracytoplasmic function σ factors (ECFs) belong to the most abundant signal transduction mechanisms in bacteria. Among the diverse regulators of ECF activity, class I anti- σ factors are the most important signal transducers in response to internal and external stress conditions. Despite the conserved secondary structure of the class I anti- σ factor domain (ASDI) that binds and inhibits the ECF under noninducing conditions, the binding interface between ECFs and ASDIs is surprisingly variable between the published cocrystal structures. In this work, we provide a comprehensive computational analysis of the ASDI protein family and study the different contact themes between ECFs and ASDIs. To this end, we harness the coevolution of these diverse protein families and predict covarying amino acid residues as likely candidates of an interaction interface. As a result, we find two common binding interfaces linking the first alpha-helix of the ASDI to the DNA-binding region in the σ_4 domain of the ECF, and the fourth alpha-helix of the ASDI to the RNA polymerase (RNAP)-binding region of the σ_2 domain. The conservation of these two binding interfaces contrasts with the apparent quaternary structure diversity of the ECF/ASDI complexes, partially explaining the high specificity between cognate ECF and ASDI pairs. Furthermore, we suggest that the dual inhibition of RNAP- and DNA-binding interfaces is likely a universal feature of other ECF anti- σ factors, preventing the formation of non-functional trimeric complexes between σ /anti- σ factors and RNAP or DNA.

IMPORTANCE In the bacterial world, extracytoplasmic function σ factors (ECFs) are the most widespread family of alternative σ factors, mediating many cellular responses to environmental cues, such as stress. This work uses a computational approach to investigate how these σ factors interact with class I anti- σ factors—the most abundant regulators of ECF activity. By comprehensively classifying the anti- σ s into phylogenetic groups and by comparing this phylogeny to the one of the cognate ECFs, the study shows how these protein families have coevolved to maintain their interaction over evolutionary time. These results shed light on the common contact residues that link ECFs and anti- σ s in different phylogenetic families and set the basis for the rational design of anti- σ s to specifically target certain ECFs. This will help to prevent the cross talk between heterologous ECF/anti- σ pairs, allowing their use as orthogonal regulators for the construction of genetic circuits in synthetic biology.

KEYWORDS RNA polymerase, coevolutionary analysis, comparative genomics, computational biology, direct coupling analysis, gene regulation, transcription factors


Extracytoplasmic function σ factors (ECFs) are one of the most abundant signal transduction mechanisms in the bacterial kingdom, often mediating the cellular response to external and internal stress conditions. Although these minimalistic members of the σ^{70} family contain only the σ_2 and σ_4 domains essential for recruiting RNA

Citation Casas-Pastor D, Diehl A, Fritz G. 2020. Coevolutionary analysis reveals a conserved dual binding interface between extracytoplasmic function σ factors and class I anti- σ factors. *mSystems* 5:e00310-20. <https://doi.org/10.1128/mSystems.00310-20>.

Editor David F. Savage, University of California, Berkeley

Copyright © 2020 Casas-Pastor et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Georg Fritz, georg.fritz@uwa.edu.au.

 Co-evolution between ECF σ factors and their cognate anti- σ factors reveals that anti- σ factors use a dual binding interface to shield the σ factor from contact to RNA polymerase and DNA binding. @Fritz_lab @SMS_UWA

Received 8 April 2020

Accepted 17 July 2020

Published 4 August 2020

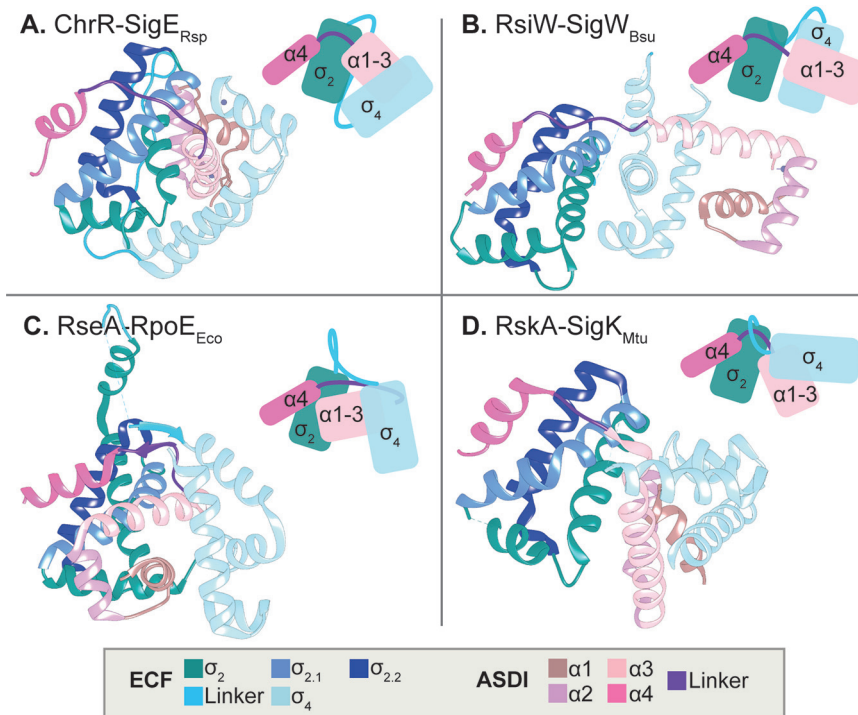


FIG 1 Structures of ECF σ factors in complex with class I anti- σ factors. ECFs are shown in shades of pink, whereas anti- σ factors appear in shades of blue. Different areas of the protein are differentially colored (see legend). Different anti- σ factors show different binding conformations. (A) SigE-ChrR from *R. sphaeroides* (PDB accession no. 2Q1Z [13]). (B) SigW-RsiW from *B. subtilis* (PDB accession no. 5WUQ [14]). (C) RpoE-RseA from *E. coli* (PDB accession no. 1OR7 [16]). (D) SigK-RskA from *M. tuberculosis* (PDB accession no. 4NQW [15]).

polymerase (RNAP) to specific promoter sequences (1), ECFs have evolved into a surprisingly diverse protein family. By now, we know 156 phylogenetic ECF groups (2, 3)—many of which feature group-specific target promoter motifs and conserved regulators of ECF activity, suggesting similar modes of signal transduction within an ECF group. Among these diverse signaling mechanisms, the most common regulators of ECF activity are so-called anti- σ factors, which, under noninducing conditions, sequester ECF into inactive complexes via their anti- σ domain (ASD). Under inducing conditions, anti- σ factors release their ECFs by various mechanisms, including anti- σ proteolysis (4–6), conformational change (7, 8), or sequestration by ECF-mimicking anti-anti- σ factors (9, 10). Given that bacteria harbor an average of 10 ECFs, and some species encode more than 100 ECFs per genome (3), the pertinent question arises how the different ECF/anti- σ factor pairs prevent massive cross talk between each other, or in other words, how do they achieve signaling specificity?

To date, three structurally distinct classes of anti- σ factors, termed classes I to III, have been described in literature (11, 12). Here, we focus on the class I anti- σ factors, which are not only the first characterized (13) but also the most abundant anti- σ s known to date (2, 3). Class I anti- σ factors are defined by their N-terminal anti- σ domain I (ASDI), which features a common secondary structure consisting of four alpha-helices: the first three (N-terminal) helices form a bundle that binds to the σ_4 domain of the ECF, and separated by a flexible linker, the fourth helix binds to the σ_2 domain. Interestingly, while this general theme has been found in all of the four crystal structures of ASDI/ECF complexes solved to date (13–16), these structures also expose a significant diversity in the binding topology between ECFs and ASDIs (Fig. 1). The most striking difference relates to the overall ECF/ASDI conformation (Fig. 1), showing that in three of the cocrystal structures, ChrR/SigE_{Rsp} (*Rhodobacter sphaeroides*), RseA/RpoE_{Eco} (*Escherichia coli*), and RskA/SigK_{Mtu} (*Mycobacterium tuberculosis*), the ASDI is sandwiched between

the σ_2 and σ_4 domains, while RsiW wraps around the two σ domains of SigW (*Bacillus subtilis*). Furthermore, while the three-helix bundle of some ASDIs require zinc coordination for ECF inhibition (ChrR_{Rsp} [13]), another structure features a zinc-binding motif but binds the ECF independently of zinc (RsiW_{B. subtilis} [14]), and others do not rely on a zinc-binding motif at all (RskA_{Mtu}, RseA_{Eco} [15, 16]). Thus, it is tempting to speculate that the divergent binding topologies between ECFs and ASDIs could be important to prevent cross talk between ECF/ASDI pairs of different ECF groups and that these conformations might be conserved within ECF groups. If so, we reasoned that protein sequences of ECF and ASDI proteins have coevolved and that ASDI protein sequences should cluster into phylogenetic groups similar to the ECF groups.

Due to the diversity in ECF/ASDI quaternary structure, we here wondered whether there is a minimal contact interface conserved across all members of the ASDI family. To predict amino acid residues involved in such conserved contact interfaces, we turn to direct coupling analysis (DCA)—a bioinformatic method that exploits evolutionary covariation to predict contacting residues (17). When two residues interact, mutations in one need to be compensated by changes in the second so as to preserve the interaction (17). The same mechanism also applies for indirect contacts; however, DCA is able to distinguish direct from indirect interactions and considers only the former for the calculation of their covariation score (17). One of the highlights of DCA is that aside from stable conformations, it can also provide information on the transient, unstable conformations that occur during the dynamic process of interaction (18).

In this study, we set out to provide the first phylogenetic classification of ASDI proteins and reveal striking patterns of coevolution between these regulators and their cognate ECF σ factors. For the ECF/ASDI interaction, we used DCA to predict the residues that form the core ECF/ASDI contact. The arising sequence logos show divergent use of residues across ASDI groups, thus explaining the low binding affinity of noncognate ECF/ASDI pairs from different groups. However, the predicted interaction partners in the fourth helix of ASDI and their respective counterparts are less conserved even within the ASDI groups. This might explain how ASDI proteins maintain binding specificity even within ASDI groups. These results allow a first, *in silico* assessment of potential cross talk between two ECF/ASDI pairs without expensive *in vivo* testing, opening new ways to rationally design synthetic circuits using orthogonal ECF/ASDI pairs.

RESULTS

ASDI retrieval and classification. We focused on the class I anti- σ factors (ASDIs) as the main regulators of ECF σ factors, in order to gain a better understanding of their general binding mechanism for ECFs. Given that anti- σ factors are often coencoded with their ECF targets (1, 2, 13, 19, 20), we first set out to collect ASDIs from the genetic neighborhood of the ECF coding sequences. To this end, we focused on a set of 21,047 putative anti- σ factors identified during a recent classification effort for ECF σ factors by our group (3). To identify ASDI-containing proteins from this data set, we used hidden Markov models (HMMs) developed from a small data set of both zinc-binding and non-zinc-binding ASDIs published earlier by Staroń and colleagues (2) (see Materials and Methods for details). This step yielded 7,490 proteins, showing that ~36% of all putative anti- σ factors are ASDIs. In order to further expand the size of the ASDI sequence library, we built a new extended HMM from the ASDI domain of these sequences. We used this extended model to search for ASDIs in the genetic neighborhood of all classified ECFs identified in reference 3, using only the 33,843 ECFs from representative and reference organisms as labeled by the National Center for Biotechnology Information (NCBI). This yielded 11,939 proteins, from which we removed the ones with ASDIs shorter than 50 amino acids, since these could be divergent class II anti- σ factors (21). The final number of ASDIs retrieved by this pipeline was 10,930, of which 10,806 have a nonredundant anti- σ domain. This shows that, on average, about one-third (~32%) of the ECF σ factors contain a protein with an ASDI domain in their genetic neighborhood, suggesting that ASDIs are the most widespread regulators of

groups were named with “AS” followed by a number dependent on the ECF group with which they are found in genomic proximity. Even though ASDIs from the same clade of the ASDI tree are usually coencoded with (and thus likely regulate) members of the same ECF group, some ASDI groups have slightly divergent features and are located in different clades of the ASDI tree. Two of these ASDI groups are, for example, AS19-1 and AS19-2, which regulate members of ECF19 (Fig. 2), but they are divergent in their ASDI helix 1 (consensus motif HTLAGAYALDAL in AS19-1 versus HLDPDQLALLA in AS19-2) and helix 2 (consensus motif LDDERAAFERHL in AS19-1 versus GEPLDADERAHL in AS19-2). Given that group AS19-2 is more closely related to AS27 than AS19-1, this suggests that these groups may have independently evolved the ability to bind to ECFs of group 19.

ASDIs that regulate ECFs from the same subgroup are usually located together in the tree but split into distinct ASDI subgroups (data not shown), probably due to the larger sequence diversity of anti- σ factors than of ECFs. We observed that, even though there was a mixture of zinc-binding and non-zinc-binding ASDIs in the input data set (as indicated by the presence or absence of an “H_xC_xC” motif), both types distribute across the ASDI tree, generating ASDI groups that are mixtures of zinc- and non-zinc-binding proteins, such as AS19-1 and AS27 (Fig. 2). Exceptions are groups AS33-1 and AS33-2, whose difference is the presence or absence of the zinc-binding domain, respectively (Fig. 2, ring 3).

Additionally, we predicted the mode number of transmembrane helices (TMHs) in the different ASDI subgroups using the consensus prediction from online TopCons (22). Most of the full-length anti- σ factor sequences (~65%) are predicted to contain at least one TMH, suggesting that they are bound to the membrane, while the remaining ones (35%) are likely soluble anti- σ factors. Although the whole data set of ASDIs is composed of similar amounts of zinc- (~56%) and non-zinc-binding (~44%) proteins, we observed that among the soluble ASDIs there was an overrepresentation (~72%) of sequences with a zinc-binding motif. This is consistent with the notion that cytoplasmic ASDIs are often involved in sensing intracellular redox conditions (13, 23, 24). The membrane-bound anti- σ factors contained ~48% of sequences with a zinc-binding motif, contrasting with earlier observations that membrane-bound anti- σ factors showed an underrepresentation of zinc-binding domains (13). However, the data in this earlier work were based on a much smaller sequence data set of only 1,266 sequences (13), suggesting that this apparent bias may have been due to random sampling of the sequenced genomes at the time. Our finding of an approximately equal distribution of zinc- and non-zinc-binding motifs in the membrane-bound ASDIs indicates that the Zn-binding motif could be playing a nonsensing role, e.g., by taking a more static, structural function as is the case for RsiW in *B. subtilis* (14).

If the Zn-binding motif does not play an active sensory role, the general notion is that the ASDI domains have associated with additional protein domains that allow stimulus perception and ultimately trigger anti- σ factor release (25, 26). To assess the conservation of additional protein domains, usually located C terminal of the ASDI domain, we scanned full-length class I anti- σ factors with Pfam 31.0 models (27) as well as the extended model of the ASDI domain. When indicating the positions of these domains in the different class I anti- σ factor subgroups (Fig. 2, ring 4), we found that the protein domains associated with ASDIs are typically well conserved for ASDIs from the same group but differ between groups. This suggests that ASDIs regulating members of the same ECF group are likely sensing similar input cues, by binding directly either to the triggering molecule or to other sensory proteins. The full list of ASDIs, together with their partner ECF and their ASDI group and subgroup, can be found in Table S1.

Given the ample degree of correlation between ECF and ASDI classifications, we evaluated whether these families coevolved. For this, we calculated the Pearson correlation coefficient (PCC) of the pairwise distance matrices of ASDIs and ECFs, as described by Goh et al. (28), leading to a PCC of 0.82. In order to determine the significance of this correlation coefficient, we adopted the strategy of Dintner et al. (29)

TABLE 1 Pearson's correlation coefficient of the distances for ECF and ASDI pairs in organisms that contain RsbW-like and RpoD-like proteins, used as negative controls for lack of correlation

| | ECFs | ASDIs | RsbW | RpoD |
|-------|------|-------|------|------|
| ECFs | 1.00 | | | |
| ASDIs | 0.82 | 1.00 | | |
| RsbW | 0.56 | 0.63 | 1.00 | |
| RpoD | 0.48 | 0.50 | 0.67 | 1.00 |

and included as negative controls RsbW-like anti- σ factors and RpoD-like σ factors, which do not interact with ECFs and ASDIs, respectively: RsbW is the anti- σ factor of the alternative σ factor σ^B and a protein kinase of the anti-anti- σ factor RsbV in *Bacillus subtilis* (30), while RpoD is the housekeeping σ factor of *Escherichia coli* (31). For these negative controls, we obtained low PCCs (0.5 to 0.6), which are similar to the ones obtained by Dintner et al. for negative controls in bacterial two-component systems (29). This indicates that the PCC of 0.82 obtained for the correlation between ECFs and ASDIs is highly significant, showing that there has been strong coevolution between these protein families (Table 1).

A taxonomic analysis of the ASDI protein family (Fig. S2A) further shows that ASDI groups are often composed of sequences originating from a single bacterial phylum, e.g., for AS02 (*Proteobacteria*), AS245 (*Firmicutes*), or AS12 (*Actinobacteria*), consistent with the observation that some ECF groups are phylum specific (3). While most other ASDI groups are also typically composed of sequences from a dominant phylum, they often contain a few ASDI subgroups from other phyla, e.g., for AS11, AS26, or AS243, suggesting that these subgroups may have resulted from horizontal gene transfer. Another interesting observation is that while all ASDIs from AS12 are found in *Actinobacteria*, the regulated ECFs make up only 89% of the sequences in the ECF12 group (3). The other ECF12s are found mostly in *Bacteroidetes* (~7%) and *Proteobacteria* (~2%), and the fact that they do not feature an ASDI in their genomic neighborhood indicates that they either are regulated by orphan ASDIs or have adopted another mode of regulating ECF activity (3). In fact, when more closely examining the phylum-specific frequency of ASDIs in the genomic neighborhood of ECFs (Fig. S2B), we found that in *Bacteroidetes* only 6% of ECFs are associated with ASDIs, which is significantly lower than the average of 32% found across all phyla. This is consistent with the observation by Staroń et al. (2), who noted that the *Bacteroidetes* group of RpoE-like proteins (ECF03) also lacks a conserved anti- σ factor in their genomic context, suggesting again either that gene synteny is broken or that other modes of ECF regulation, e.g., via other anti- σ classes, are dominant in this phylum. Other phyla, in contrast, feature a strong overabundance of ASDI-associated ECFs, such as the *Gemmatimonadetes* (~79%) or the *Chloroflexi* (72%), but further studies are needed to identify the origin of this taxonomic bias.

DCA predicts two main contact interfaces between ASDIs and ECFs. Given the variability in the binding conformations in the four published ECF/ASDI cocrystal structures, we next wondered whether there exist universally conserved “core-binding interfaces” that are shared within the whole family of ASDI proteins, or whether the strong coevolution between the protein families gave rise to fundamentally different binding conformations. To identify potentially conserved contact interfaces, we sought to exploit the coevolutionary information between our ASDI data set (above) and the ECF classification (3). Specifically, we aimed at predicting amino acid residues on ASDIs and ECFs that display significant covariation, suggesting that they are in direct contact and that the mutation in one residue is balanced by a compensatory mutation in its binding partner. To this end, we applied direct coupling analysis (DCA) (17) to the full set of ASDIs and their cognate ECFs (Table S1). The results of this analysis revealed a large amount of high DCA scores within the σ_2 and σ_4 domains of the ECF σ factor and also connecting the two σ domains (Fig. 3A). This pattern matches previous DCA results in ECF σ factors (32) and is indicative of the conserved secondary and tertiary structure

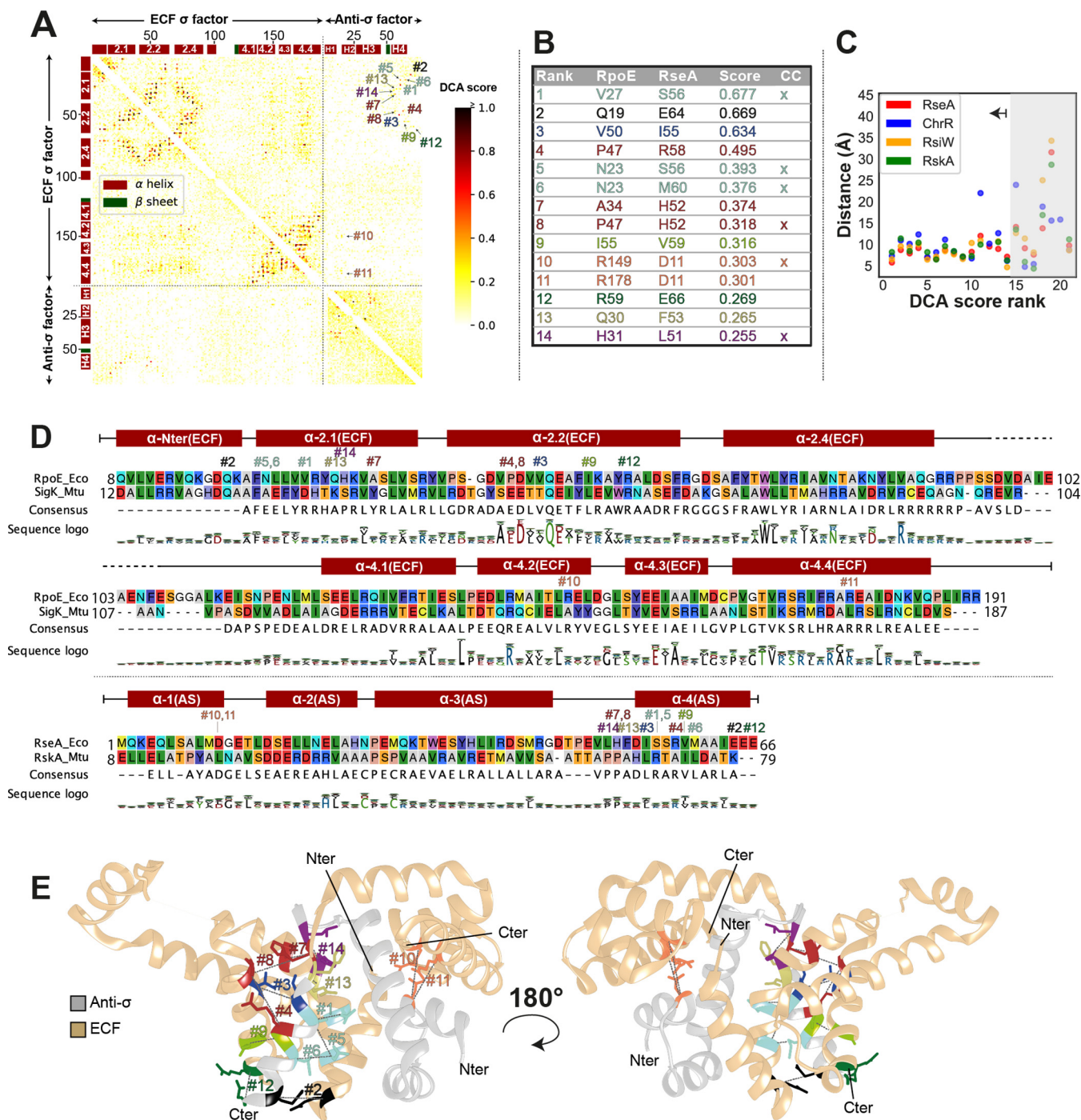


FIG 3 DCA results on the contact between ECFs and ASDIs. (A) DCA contact map. Each axis represents the concatenated protein sequences of RpoE and RseA, from *E. coli*, used as reference for the amino acid labeling. High DCA scores, indicated by darker colors, correspond to residues with a high likelihood to bind *in vivo*. The 14 highest scores (DCA score ≥ 0.255) are marked in the heatmap and labeled according to their rank. (B) Table of the 14 highest-scoring DCA predictions, mapped to the amino acid coordinates of RpoE and RseA from *E. coli*. The common contact (CC) column indicates the DCA predictions that are also common contacts observed in the four crystal structures of ECFs/ASDIs, as derived by Voronoi tessellation (Table 2). (C) Scatterplot of the top 21 DCA predictions against the distance between the alpha carbons of the predicted contacts, as derived from the four structures of ECF/ASDI complexes (Fig. 1). The top 14 predictions are in close proximity in most of the three-dimensional structures. Complexes are labeled after their anti- σ factor, where RseA corresponds to RpoE/RseA complex from *E. coli* (PDB accession no. 1OR7 [16]), ChrR to SigE/ChrR from *R. sphaeroides* (PDB accession no. 2Q1Z [13]), RsiW to SigW/RsiW from *B. subtilis* (PDB accession no. 5WUQ [14]), and RskA to SigK/RskA from *M. tuberculosis* (PDB accession no. 4NQW [15]). (D) Multiple-sequence alignment of two selected ECF/ASDI pairs, RpoE/RseA from *E. coli* and SigK/RskA from *M. tuberculosis*. Labels of the top 14 contacts indicate their position. The presence of alpha-helices and their names are depicted on top of the alignment. The sequence logo depicts the amino acid composition of the full ECF and ASDI alignments derived from 10,930 sequences, respectively. (E) Three-dimensional depiction of the top 14 predictions in the structure of RpoE/RseA complex (PDB accession no. 1OR7 [16]). ECF is colored in beige, and anti- σ factor is in gray. Predicted contacts are labeled according to their rank. N and C termini from ECF and anti- σ factor are labeled.

on this family of proteins. We also observed high scores interconnecting helices 1, 2, and 3 of the ASDI, while helix 4 shows no strong DCA coupling scores with other parts of the ASDI domain (Fig. 3A). This agrees with the cocrystal structures of ECF/ASDI complexes (Fig. 1), where helices 1, 2, and 3 form a helix bundle, which is connected to helix 4 by a flexible linker (13–16). We then focused on the predictions that link ECFs and ASDIs since these are the ones responsible for ECF inhibition. At first glance, the contact map shows several high DCA scores linking the fourth helix of the ASDI with the σ_2 domain (Fig. 3A). Under closer inspection, the top 14 interprotein contact predictions (DCA score ≥ 0.255) are located in close proximity in most of the crystal structures (Fig. 3B and C). Of those, 12 are connecting the σ_2 domain and helix 4 of the ASDI, and two (DCA#10 and #11) connect a single residue of helix 1 of the ASDI to two residues located in the σ_4 domain of the ECF (Fig. 3E). In the first case, the predicted contact area includes ECF regions 2.1 and 2.2 (Fig. 3D), whose main function is binding to the clamp helices of the β' subunit of the RNAP (33–35). Thus, it is likely that binding of ASDI's helix 4 to this area prevents ECF binding to the RNAP core, hampering ECF-dependent transcription when the anti- σ factor is present. Instead, DCA predictions #10 and #11 involve ECF helices 4.2 and 4.4 (Fig. 3D), in two residues involved in the contact with the -35 element of the promoter (33, 36). Taken together, the presence of these strong coevolutionary signals suggests that the majority of the 10,860 ASDI proteins establish contact to the ECF via these two binding interfaces, connecting the ASDI with both the σ_2 and σ_4 domains.

However, although the top 14 DCA predictions connect residues located in close 3-dimensional proximity in most of the four resolved cocrystal structures of ECFs/ASDIs, only six are direct contacts in the four crystal structures (Fig. 3B). While the other 8 “close hits” could merely be false-positive predictions, it is tempting to speculate that these residues might form close contacts in other ECF/ASDI groups, which might take slightly different binding conformations from those captured by the four structures solved to date. Alternatively, these close hits may form transient contacts during the initial recognition between ASDI and ECF. Another observation was that 19 direct contacts that are shared between the four ECF/ASDI cocrystal structures were not predicted by DCA (Table 2), suggesting either that DCA fails to predict them or that these contacts are less prevalent in the remainder of the ECF/ASDI protein families.

To obtain a better overview of the residues involved in the contact interfaces, we plotted the residues predicted by DCA—both in the ECF and in the ASDI—for the 12 largest ASDI groups with more than 100 sequences (Fig. 4). The resulting logos showed that contacts involving ASDI's helix 1 and the σ_4 domain (DCA#10 and #11) are generally conserved within groups but different between groups. Predictions DCA#10 and #11 feature two main types of contacts, either a charged or a hydrophobic interaction (Fig. 4). This pattern is most evident for prediction DCA#11, which tends to harbor a positive amino acid in the ECF (e.g., R178 in RpoE_{E.coli}) and a negative residue in the ASDI (e.g., D11 in RseA_{E.coli}), as found in groups ECF02, ECF12, ECF14, ECF27, ECF235, and ECF245. However, in some cases this is replaced by a hydrophobic contact, typically with leucine on both the ECF and ASDI (e.g., L177 in SigK and L18 in RskA from *Mycobacterium tuberculosis*), as found in groups ECF17, ECF18, and ECF19. In contrast to these clear-cut contact motifs predicted for helix 1, residues in helix 4 of the ASDI (all predictions except DCA#10 and #11) exhibit a weaker conservation even within most of the ASDI groups (Fig. 4). This has some exceptions, such as the prediction DCA#7, featuring a conserved contact between an aromatic residue (W or Y) on the ASDI and a proline (P) on the ECF side in groups ECF12, ECF14, ECF27, and ECF245 (Fig. 4). Together with the observation that helix 4 of the ASDI holds most of the DCA predictions (Fig. 3D), this suggests that helix 4 is in charge of further determining the specificity of the ASDIs, keeping them orthogonal from other ASDIs of the same group. Indeed, anti- σ factors that regulate ECFs from the same group have been found to be mostly orthogonal (37).

Specificity-determining positions of ASDI groups coincide with the predicted binding interfaces. Next, we asked whether the ASDI residues predicted to be in

TABLE 2 Common contacts between ECFs and ASDIs found in the four cocrystal structures by Voronoi tessellation^a

| ECF residue (mapped to RpoE _{E.coli}) | AS residue (mapped to RseA _{E.coli}) | ECF region | ASDI region | DCA prediction (rank) | ASDI SDP? (yes/no) |
|---|--|--------------|--------------|-----------------------|--------------------|
| 22 | 59 | σ 2.1 | H4 | | N |
| 22 | 60 | σ 2.1 | H4 | | N |
| 22 | 63 | σ 2.1 | H4 | | N |
| 23 | 56 | σ 2.1 | H4 | 5 | Y |
| 23 | 60 | σ 2.1 | H4 | 6 | N |
| 26 | 55 | σ 2.1 | H4 | | N |
| 26 | 56 | σ 2.1 | H4 | | Y |
| 26 | 59 | σ 2.1 | H4 | | N |
| 27 | 56 | σ 2.1 | H4 | 1 | Y |
| 31 | 51 | σ 2.1 | Linker H3-H4 | 14 | N |
| 35 | 48 | σ 2.1 | Linker H3-H4 | | N |
| 47 | 52 | σ 2.2 | Linker H3-H4 | 8 | N |
| 47 | 55 | σ 2.2 | H4 | | N |
| 51 | 55 | σ 2.2 | H4 | | N |
| 51 | 58 | σ 2.2 | H4 | | N |
| 51 | 59 | σ 2.2 | H4 | | N |
| 54 | 59 | σ 2.2 | H4 | | N |
| 58 | 63 | σ 2.2 | H4 | | N |
| 131 | 42 | σ 4.1 | H3 | | N |
| 135 | 43 | σ 4.1 | H3 | | N |
| 149 | 11 | σ 4.2 | H1 | 10 | Y |
| 150 | 10 | σ 4.2 | H1 | | N |
| 150 | 40 | σ 4.2 | H3 | | N |
| 150 | 43 | σ 4.2 | H3 | | N |
| 151 | 43 | σ 4.2 | H3 | | N |

^aThe four crystal structures analyzed correspond to SigK/RskA from *M. tuberculosis* (PDB accession no. 4NQW [15]), SigW/RsiW from *B. subtilis* (PDB accession no. 5WUQ [14]), SigE/ChrR from *R. sphaeroides* (PDB accession no. 2Q1Z [13]), and RpoE/RseA from *E. coli* (PDB accession no. 1OR7 [16]). Coordinates of the different amino acids are shown in RpoE/RseA proteins. ECF and ASDI regions where the amino acids are located are shown. For simplicity, the σ_4 domain is split into four subregions ($\sigma_{4,1}$ to $\sigma_{4,4}$) according to the presence of alpha-helices. "H" indicates alpha-helix. The rank of the DCA prediction is displayed in the second-to-last column when the interaction is predicted by DCA. If the residue is an SDP in the ASD, it is indicated in the last column. "Y" for yes is highlighted by bold in the last column.

contact with the ECF are also key residues that determine the distinction between ASDI groups. If this was the case, it would suggest that ASDI groups would be primarily distinguished by their interaction with their respective ECF. Alternatively, if ASDI groups were primarily determined by residues outside predicted contact interfaces, this would argue that interactions with potential ligands or intraprotein interactions determine protein subfamilies (38). The presence of such group-specific amino residues—so-called specificity determining positions (SDPs)—can be detected by S3det, a bioinformatic tool based in multiple correspondence analysis that finds residues associated with subfamilies of proteins (39). Using this tool, we predicted SDPs by comparing every pair of the 12 largest ASDI groups and taking only the highest-scoring SDP prediction of every ASDI group into further consideration (see Materials and Methods). As a result, we identified five SDPs, named by running numbers (SDP#1 to SDP#5) from N to C terminus: two in helix 1, one in helix 3, one in helix 4, and the last one exclusively present in group AS243 (Fig. 5A). Proteins from group AS26 did not hold any prediction, since they do not fit well into the multiple sequence alignment of the full ASDI data set—probably due to extensive differences at the sequence level (cf. Fig. 4). Similarly, AS243's SDP#5 corresponds almost exclusively to a gapped position in the alignment with the rest of the groups, as indicated by the absence (or very narrow representation in the case of AS245) of conserved residues for SPD#5 (Fig. 5B). These differences at the sequence level might reflect functional differences between standard ASDIs and ASDIs from groups 243 and 26. In favor of this hypothesis, one member of AS243, FecR from *E. coli*, is distinguished from ASDIs of other (non-AS243) groups in that its 59 N-terminal amino acids are essential for ECF activity (40). Probably due to these unique features,

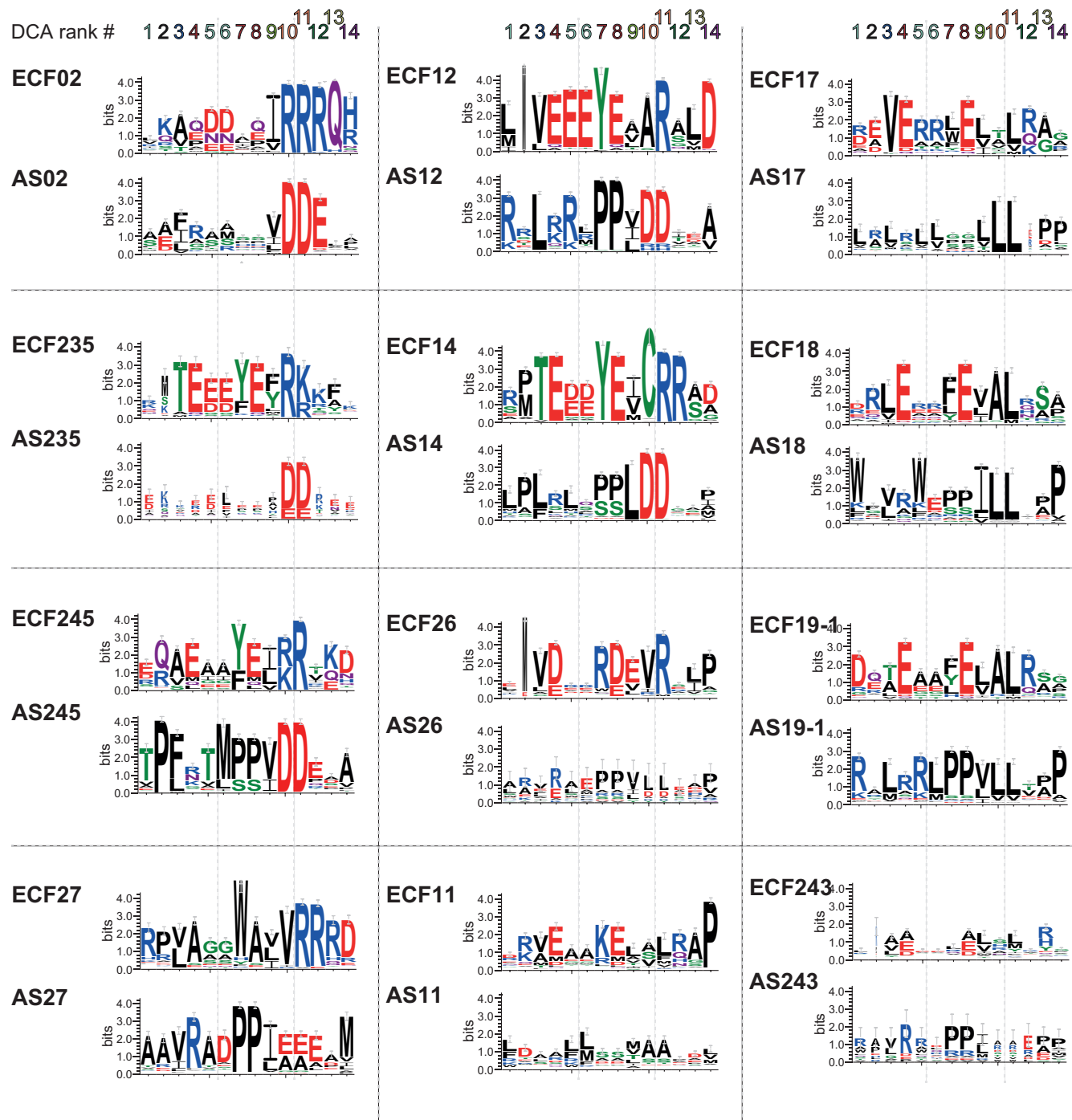


FIG 4 Sequence logos of the top 14 DCA predictions, computed for the 12 ASDI groups with more than 100 sequences. The sequence logos show the amino acid composition for the DCA-predicted contact points for both the ECF and anti- σ factor in each ECF/ASDI group. The contacts are ordered from left to right according to their DCA rank, as indicated on top. The sequence logos are manually arranged based on their similarity.

the observed misalignment between members of AS243 and the other ASDI groups precludes further interpretation of the SDPs in this group. In contrast, all other predicted SDPs (except SDP#5) are part of the contact interfaces with the ECF in the existing crystal structures (Fig. 5C). Conserved position D11 in *RseA*_{E.coli} predicted by DCA (Fig. 3B, DCA#10 and #11), was part of the predicted SDPs (Fig. 5A, SDP#2). Yet another SDP, S56 in helix 4 (Fig. 5A, SDP#4), was predicted by DCA (Fig. 3B, DCA#1 and #5). Predictions SDP#1 and SDP#3 connect S7 in helix 1 and Y36 in helix 3 in *RseA*_{E.coli}

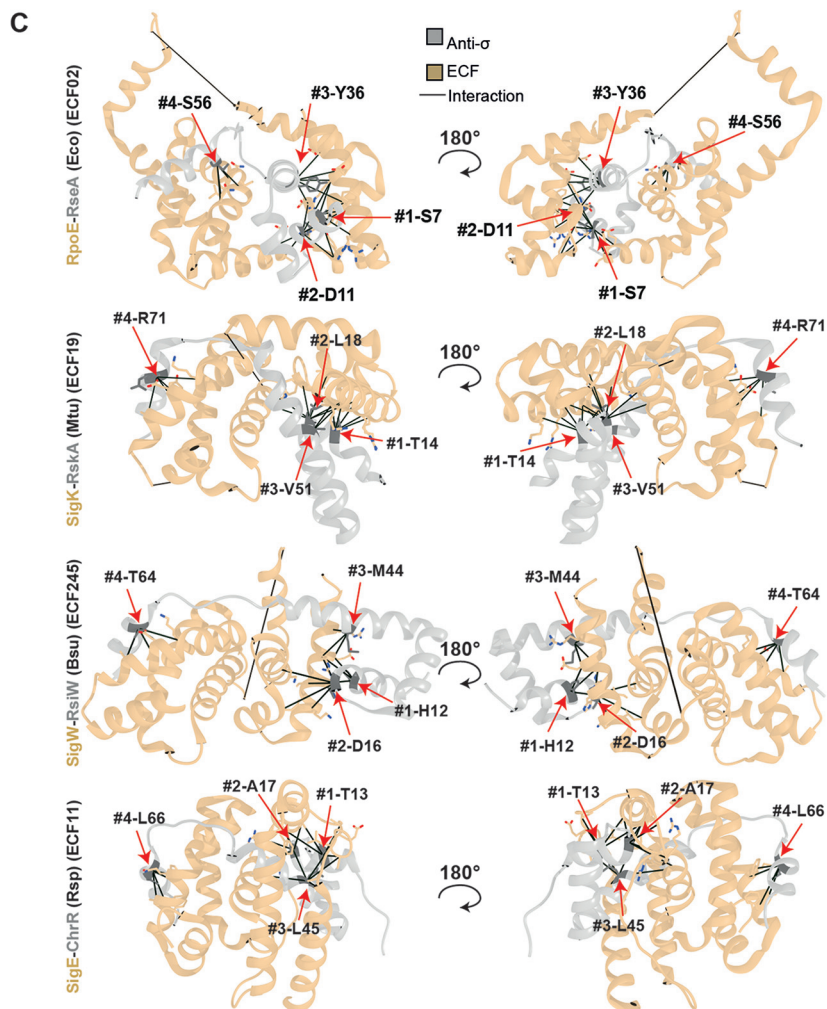
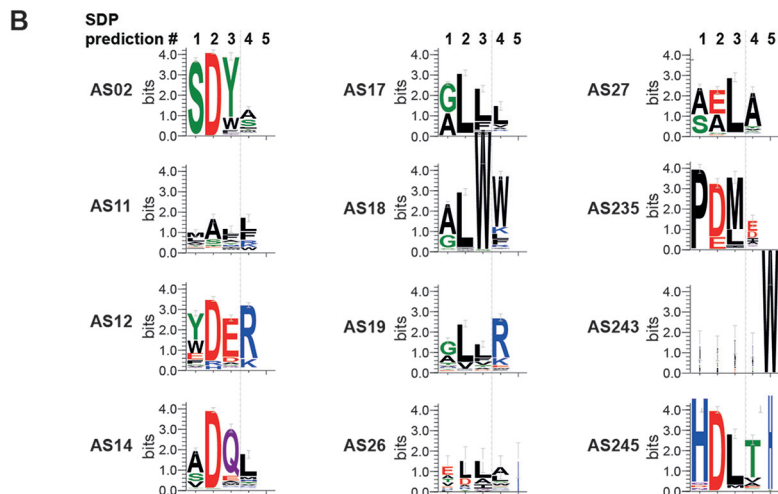
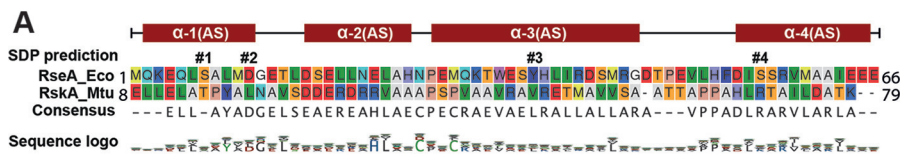


FIG 5 Description of the specificity determining positions (SDPs) that distinguish different ASDI groups. (A) Multiple-sequence alignment of the anti- σ factors RseA from *E. coli* and RskA from *M. tuberculosis* (Continued on next page)

to the σ_4 domain, usually in its last helix (Fig. 5C). Interestingly, SDPs #1, #2, and #3 form a cluster of interactions with the same area of the ECF, which usually corresponds to the last helix of the σ_4 domain, except in the SigE/ChrR structure, where the contact appears before this area (Fig. 5C). Thus, besides some exceptions in groups AS26 and AS243, these results suggest that the main characteristic that discriminates between ASDI groups is their ability to interact with the σ factors within their cognate ECF groups.

Given that these residues are conserved within phylogenetic ASDI groups, face the ECF in the solved ECF/ASDI crystal structures, and feature different amino acids in different groups, it is likely that they take part in determining specificity toward the target ECF. This is supported by the fact that most of these SDPs are also DCA predictions (Table 2).

DISCUSSION

In this study, we used a computational approach to study how the class I anti- σ factor family members interact with their cognate ECF σ factors. Based on the similarity between ECF and ASDI phylogenies, we showed that these protein families have coevolved—likely because they are in direct contact with each other—and exploited this coevolution to predict two conserved binding interfaces for the ASDI/ECF interaction. Although previous work provided insight into the cocrystal structures of individual ASDI/ECF pairs, the present work puts these case studies into a broader, evolutionary perspective, by providing the first phylogenetic classification of the class I anti- σ factor protein family. Interestingly, within the resulting AS groups—solely defined by the sequence of their ASDI domain—we observed a striking conservation of the fused protein domains. Compared to early work by Campbell et al. (13), the explosion in sequenced genomes in recent years allowed us to expand the ASDI data set from 1,266 to more than 10,000 putative ECF/ASDI pairs from NCBI reference genomes, providing a more comprehensive and phylogenetically balanced overview of the diversity of these proteins. In agreement with the work of Campbell et al. (13), we found that about one-third (~32%) of all ECFs are genomically associated with, and thus likely regulated by, ASDIs. Yet, our expanded ASDI library showed important differences compared to previous work in that (i) we find more ASDIs containing a zinc-binding motif (~56% compared to ~38% [13]); (ii) we find more cytoplasmic anti- σ factors (~35% compared to ~28% [13]); (iii) cytoplasmic anti- σ factors are still overrepresented in zinc-binding motifs, but to a smaller extent (~72% of the soluble anti- σ factors are zinc binding in our data set compared to 92% in reference 13); and (iv) membrane-bound ASDIs are not underrepresented in zinc-binding motifs as suggested in reference 13, with about half of the proteins (~48%) being zinc-binding anti- σ factors. These data suggest that ASDIs are more diverse than previously thought and argue against a functional role of the zinc-binding domain exclusively in soluble anti- σ factors. This is supported by the ASDI phylogenetic tree (Fig. 2), where zinc- and non-zinc-binding ASDI groups are mixed across the tree and sometimes even within the same group, as in the case of AS27 and AS19-1. In these mixed zinc- and non-zinc-binding groups, this suggests that the zinc-binding motif may play a structural instead of a sensory role, as shown for RsiW from *B. subtilis* (group AS245) (14).

Our analysis of DCA predictions and SDPs show that there exists a conserved, dual

FIG 5 Legend (Continued)

showing the position of the SDPs, labeled with numbers according to sequence position. Alpha-helices and their names are indicated with red boxes on the ASDI sequences. The sequence logo shows the amino acid composition of the full ASDI alignment. (B) Logo of SDPs in every ASDI group with more than 100 proteins. Positions are labeled as in panel A. (C) ASDI specificity determining positions plotted in the structure of ECF/ASDI complexes. ECFs are colored in beige, and anti- σ factors are in gray; SDPs are colored in green and labeled with their identifier as in panel A. The RpoE/RseA complex is present in *E. coli* (PDB accession no. 1OR7 [16]), the SigK/RskA complex is in *M. tuberculosis* (Mtu, PDB accession no. 4NQW [15]), SigW/RsiW is in *B. subtilis* (Bsu, PDB accession no. 5WUQ [14]), and SigE/ChrR is in *R. sphaeroides* (Rsp, PDB accession no. 2Q1Z [13]). Contacts with the ECF are represented by connector lines.

binding interface, with ASDI's helix 1 binding to the σ_4 domain and ASDI's helix 4 binding to the σ_2 domain. These results agree with crystal structures of ECF/ASDI complexes (13–16) and suggest that the contacts seen in these few examples are indeed realized across the full ECF/ASDI families. Further, our results suggest that ASDI's helix 2 is not critical for ECF binding but is important for ASDI tertiary structure. ASDI's helix 3, which is located between ECF's σ_2 and σ_4 domains in three out of four structures (13, 15, 16), harbors an SDP involved in the interaction with σ_4 domain, in similar residues as those contacted by the prediction on helix 1. This modularity of the ASDI interaction is reflected in the function of the ECF residues involved in the predictions. Contacted residues in regions 2.1 and 2.2 are mostly involved in the contact with the clamp helices of the β' subunit of the RNAP (33, 35), whereas predicted contacts in σ_4 are part of the contact interface with the -35 element of the promoter (33, 36).

The analysis of the DCA predictions revealed a different degree of conservation across ASDI groups, with the residues that take part in contacts between ASDI's helix 1 and ECF's σ_4 (DCA predictions #10 and #11) being conserved for most of the ECF and ASDI phylogenetic groups. Interestingly, this area, which connects D11 on the ASDI to R149 and R178 on the ECF (RseA/RpoE_{E.coli} coordinates), bears two main types of interactions, that is, hydrophobic, which usually features leucine in both ECF and ASDI (Fig. 4, groups AS17, AS18, and AS19-1), or charged, usually featuring arginine in the ECF side and aspartate in the ASDI side (Fig. 4, groups AS02, AS12, and AS14, among others). Random mutagenesis in RseA_{E.coli} (group AS02) showed that a single amino acid mutation of D11 to histidine completely inhibits RseA_{E.coli} activity (41), confirming the key role of this contact. Given their group-specific conservation and the striking polarity differences between the two binding types, we speculate that D11 defines coarse-grained specificity of ASDIs for ECFs of the same binding type, usually found in the same phylogenetic group. However, ASDIs are usually specific to their own target ECF and do not usually cross talk with members of the same group (37), indicating that there are more sources of specificity in residues that are not conserved in groups. One potential source of this specificity is the residues predicted by DCA in helix 4. These residues are generally not conserved within groups (Fig. 4) and bind the σ_2 domain in all the solved crystal structures of ASDI/ECF complexes (13–16). This lack of major conservation is extended to the predicted contacts on the ECF side, which are generally in charge of binding to the β' subunit of the RNAP.

Generality of the dual binding interface in other σ /anti- σ interactions? Paget classified anti- σ factors into two types, the ones that insert between σ_2 and σ_4 (RseA, RskA, and ChrR) and the ones that wrap around these domains (RsiW) (42). Our data show that despite these differences in binding topology, both types of ASDIs contact the two main binding interfaces described here. Moreover, a similar binding mode can be observed in the crystal structures of the ECF CnrH in complex with the class II anti- σ factor CnrY, from *Cupriavidus metallidurans* (43). The two alpha-helices of CnrY wrap around CnrH in a conformation where CnrY's first alpha-helix mimics the function of ASDI's first helix and binds to the σ_4 domain, and CnrY's second and last alpha-helix binds to the σ_2 domain in a similar manner as ASDI's fourth helix. The only crystal structure of a member of the ASDIII class of anti- σ factors, BldN, in complex with the ECF σ factor RsbN from *Streptomyces venezuelae* (12) also shows this dual binding mode. In this case, the first and second alpha-helices of BldN bind to the σ_4 domain, whereas its third and last alpha-helix binds to the regions 2.1 and 2.2 of a different RsbN molecule, similarly to ASDI's fourth helix (12). The similarity of the binding between the three types of ECF anti- σ factors is striking and contrasts with their low level of sequence similarity, which is limited to $\sim 11\%$ for RseA/BldN and $\sim 3\%$ for RseA/CnrY (using global pairwise alignments calculated by the Needleman-Wunsch algorithm implemented at EBI [44]). This explains why, even though the same regions of the anti- σ factor interact with a similar area of the ECF in the three types of ECF anti- σ factors, the specific residues that carry out the interaction with the ECF may differ

between ASD types. It is unclear why bacteria need at least three types of ASDs. On one hand, different ASDs may provide extra specificity to ECF inhibition, which could help to reduce the apparent tendency to cross talk of anti- σ factors (45). On the other hand, the three types of ASDs could have emerged from different proteins and optimized their ECF inhibition by blocking the same ECF regions through convergent evolution. Future analysis that includes all the ASDs known to date could help in understanding their evolution.

Interestingly, dual binding interfaces between σ and anti- σ factor extend beyond ECF σ factors. For instance, in *E. coli* the anti- σ factor FliM of the class 3 σ factor FliA (containing a σ_3 domain) also targets σ_2 and σ_4 regions with two different areas of the protein (46). However, the FliM inhibitory contacts are inverted relative to ECF anti- σ factors: FliM is composed of four alpha-helices, of which the first and second bind to the surface of the σ_2 domain, similarly to the fourth helix of ASDs. In FliM, the third and fourth helices are the ones that bind to σ_4 (46), similarly to the first helix of ASDs. Interestingly, FliM does not bind to FliA's σ_3 domain, strengthening the idea that the blockage of both σ_2 and σ_4 is the core of σ factor inhibition. Whether this is also the case for housekeeping σ s and their anti- σ s remains to be seen, as to date only the interaction between the anti- σ factor Rsd and a truncated version of RpoD (containing only the σ_4 domain) was studied in *E. coli* (47, 48). Thus, even though the present analysis was restricted to the interaction between ASDs and ECFs, we suggest that the dual inhibition of RNAP- and DNA-binding interfaces is likely a universal feature of other anti- σ factors, preventing formation of nonfunctional trimeric complexes between σ /anti- σ factors and RNAP or DNA.

MATERIALS AND METHODS

General bioinformatic tools. Generally, multiple sequence alignments (MSAs) were generated by Clustal Omega 1.2.3. with options `-iter = 2` and `-max-guidetree-iterations = 1` and manually curated (49). However, UPP (50) (default options) was used for alignments subjected to DCA or to 53det, since they require stable columns of equivalent residues with few gaps. Hidden Markov models (HMMs) were built using *hmmbuild* function and used for scanning libraries using *hmmsearch* function, both from HMMER suite 3.1b2 (51) and both with default parameters. For the extraction of the amino acid residue interactions between ECF and ASDI from cocrystal structures, we used Voronoi tessellation as implemented in Voronota version 1.19 (52). Protein structures were visualized using UCSF Chimera version 1.10.2 (53).

ASDI extraction. ASDIs were extracted from the genetic neighborhood (± 10 coding sequences) of a library of 46,293 ECF σ factors in their most recent classification (3). In order to minimize taxonomic bias, these ECFs were extracted from organisms tagged as representative or reference species by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), using only RefSeq entries when both RefSeq and GenBank records are available for the same genome. To identify ASDI domain-containing proteins, we first used two HMMs, one built from the zinc-binding and another from the non-zinc-binding anti- σ factors from the work of Staroń et al. (2). We selected the optimal bit score threshold for the retrieval of new ASDIs for each HMM by optimizing a receiver operator characteristic (ROC) curve using the function *roc_courve* from *sklearn.metrics* (54). Proteins that were used for the construction of each model were used as positive controls, and the remaining, non-ASDI anti- σ factors from the work of Staroń et al. (2) were used as negative controls. The resulting bit score thresholds, 0.4 for non-zinc-binding and 14.2 for zinc-binding models, were applied for the extraction of ASDIs from the set of putative anti- σ factors from reference 3. This resulted in 7,490 ASDIs, which were subsequently used for the construction of an extended HMM of the ASDI family. The thresholding bit score that best separates real ASDIs from other proteins was optimized using a ROC curve as described above, resulting in a bit score threshold of 0.2. We used the extended HMM to look for further members of the ASDI family in the genetic neighborhood of ECFs (± 10 coding sequences) from reference 3. In order to lessen the bias toward frequently sequenced organisms, we included only proteins from representative or reference genomes as labeled by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), using only RefSeq entries when both RefSeq and GenBank records are available for the same genome. This yielded 11,939 putative ASDI-containing proteins. We further curated these data, removing proteins with anti- σ domains shorter than 50 amino acids, since these could be anti- σ factors of class II (21). The area of the ASDI was defined as the envelope region of the highest-scoring hit of the extended HMM, discarding areas that are part of the transmembrane helices or extracellular. This resulted in 10,930 ASDIs, with an average length of 101 ± 33 (standard deviation) amino acids.

Clustering of ASDIs. We clustered ASDIs according to amino acid sequence similarity. Given the large number of proteins, we first grouped them into clusters or closely related sequences, the so-called subgroups. These were built with a divisive strategy, where proteins were subjected to a bisecting K-means clustering approach until the maximum k-tuple distance between any protein of the cluster was smaller than 0.6, as measured by Clustal Omega with `-distmat-out -full` and `-full-iter` flags (49, 55).

Bisecting K-means was implemented using *KMeans* function from the *sklearn.cluster* module (54). The 3,790 proteins that did not enter into any subgroup were left ungrouped. Thanks to this grouping, it was easier to see subgroups that may contain outliers that passed the HMM threshold but do not likely display anti- σ factor activity. In order to distinguish and discard these outliers from our clustering, we assessed the presence of Pfam domains (Pfam 31.0 [27]) in the anti- σ factors from each subgroup. We discarded 132 subgroups (606 proteins) where the Pfam domains indicated an unlikely anti- σ factor function (data not shown). In summary, the resulting 1,475 subgroups defined during this process contained 6,534 proteins (~60% of the starting ASDIs), with a median group size of 3 proteins and a standard deviation of 6.17 proteins. Given the low size of proteins in each subgroup, we further clustered the manually curated alignment of the consensus sequences of each subgroup into a maximum-likelihood phylogenetic tree using IQ-TREE version 1.5.5 (56) with 1,000 ultrafast bootstraps. As an outgroup of this tree, we included the anti- σ factor class II CnrY, from *Cupriavidus metallidurans*. The resulting tree was visualized in iTOL (57) and split into monophyletic ASDI groups according to the ECF group of their cognate partner. With this strategy, we defined 23 ASDI groups, of which 12 contain more than 100 proteins.

The presence of a zinc-binding domain was assumed in ASDIs with a Hx₃Cx₂C sequence signature that expands over helix 2 and helix 3. Presence of transmembrane helices was assessed using the consensus prediction from online TopCons (22). The mode number of transmembrane helices was considered in order to plot the transmembrane helices for a whole subgroup of class I anti- σ factors. In this way we avoid biases caused by the extremely large number of transmembrane helices in long, divergent proteins. The position of these helices for plotting was calculated according to the average start and end positions over the anti- σ factors in a subgroup. Similarly, the position of the ASDI domain across anti- σ factors from the same subgroup was calculated according to the average start and end positions of the envelope region of the lowest E value match to the extended HMM of the ASDI family, using hmmscan from HMMER suite 3.1b2 (51). The presence of other Pfam domains in full-length class I anti- σ factors was evaluated using hmmscan function from HMMER suite 3.1b2 (51) with the library of HMMs from Pfam 31.0 (27). Pfam domains present in certain position of the MSA of the full-length anti- σ factors in more than 50% of the members of a subgroup were plotted in the ASDI tree.

ASDI/ECF coevolution. In order to evaluate the coevolution of ECFs and ASDIs, we calculated the Pearson correlation coefficient (PCC) of the distances between cognate pairs of proteins, as introduced by Goh et al. (28). The significance of this PCC was evaluated similarly to reference 29. For this purpose, the PCCs between ASDIs and ECFs and of two extra families of proteins that did not coevolve and/or interact with ECFs or ASDIs were evaluated as negative controls. In our case, these negative controls were homologs of the *E. coli* housekeeping σ factor σ^{70} (RefSeq accession no. NP_417539.1) and of the *Bacillus subtilis* anti- σ factor RsbW (RefSeq accession no. WP_061902497), since proteins for these types have never been described to interact with ASDIs or ECFs, respectively. We extracted proteins from these types using online HMMER (51) with parameters -E 1 -domE 1 -incE 0.01 -incdomE 0.03 -mx BLOSUM62 -pextend 0.4 -popen 0.02 -seqdb uniprotrefprot and mapped the hit identifiers (IDs) from UniProt to GenBank using the UniProt's ID conversion tool (58). A total of 409 genomes contained the four protein families; these are ECFs, ASDIs, RsbW, and RpoD. For each organism, we selected one of the ECF-AS factor pairs and one homolog of RsbW and RpoD. These proteins had a taxonomically diverse origin, with 39% of the proteins from *Firmicutes*, 28% from *Actinobacteria*, 11% from *Cyanobacteria*, and the rest from eight other bacterial phyla. We calculated the pairwise distance for each protein family using Clustal Omega with -full and -distmat-out flags (49). The PCC was calculated from the flattened distance matrices using *pearsonr* function from Python's *scipy.stats* resource (59).

DCA. Direct coupling analysis (DCA) was applied to the 10,930 putative ASDIs extracted during this work (see Table S1 in the supplemental material). ASDIs and their cognate ECF partners were aligned independently using UPP (50) with default parameters, and the resulting alignments were concatenated. Gaussian DCA with default parameters (60) was performed on this alignment (N = 275, M = 10,934, Meff = 965.52, theta = 0.46). The top DCA predictions were mapped into the crystal structures of RpoE/RseA from *Escherichia coli* (AS02, PDB accession no. 1OR7 [16]), SigE/ChrR from *Rhodobacter sphaeroides* (AS11, PDB accession no. 2Q1Z [13]), SigK/RskA from *Mycobacterium tuberculosis* (AS19-1, PDB accession no. 4NQW [15]), and SigW/RsiW from *Bacillus subtilis* (AS245, PDB accession no. 5WUQ [14]). Distances between predictions were calculated using the Bio.PDB module (61, 62) and Chimera (53). The 14 predictions that connected residues in close proximity (<15 Å) in most of the structures were considered true interactions.

SDPs. Specificity determining positions (SDPs) were calculated with S3det (39) on the 12 ASDI groups with more than 100 proteins and on their cognate ECFs. Aligned ASDI (or ECF) proteins were extracted from the MSA used for DCA so as to preserve the same positional mapping. S3det was executed on every pair of ASDI (or ECF) groups, resulting in a set of ranked SDP predictions for every pair of groups. We scored the SDPs associated with every group as the sum of the inverse of their ranks across the different S3det runs with contribution of the group. The highest-scoring SDP for every group was considered positive, resulting in five SDPs.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 1.5 MB.

FIG S2, TIF file, 2.7 MB.

TABLE S1, XLSX file, 2.2 MB.

ACKNOWLEDGMENTS

This work was supported by a grant from the ERA-SynBio program via the Federal Ministry of Education and Research (Germany; grant 031L0010B) and the LOEWE program of the State of Hesse (Germany).

REFERENCES

- Lonetto MA, Brown KL, Rudd KE, Buttner MJ. 1994. Analysis of the *Streptomyces coelicolor* sigE gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions. *Proc Natl Acad Sci U S A* 91:7573–7577. <https://doi.org/10.1073/pnas.91.16.7573>.
- Staroń A, Sofia HJ, Dietrich S, Ulrich LE, Liesegang H, Mascher T. 2009. The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) σ factor protein family. *Mol Microbiol* 74:557–581. <https://doi.org/10.1111/j.1365-2958.2009.06870.x>.
- Casas-Pastor D, Müller RR, Becker A, Buttner M, Gross C, Mascher T, Goesmann A, Fritz G. 2019. Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family. *bioRxiv* <https://doi.org/10.1101/2019.12.11.873521>.
- Ades SE, Connolly LE, Alba BM, Gross CA. 1999. The *Escherichia coli* σ (E)-dependent extracytoplasmic stress response is controlled by the regulated proteolysis of an anti- σ factor. *Genes Dev* 13:2449–2461. <https://doi.org/10.1101/gad.13.18.2449>.
- Ellermeier CD, Losick R. 2006. Evidence for a novel protease governing regulated intramembrane proteolysis and resistance to antimicrobial peptides in *Bacillus subtilis*. *Genes Dev* 20:1911–1922. <https://doi.org/10.1101/gad.1440606>.
- Castro AN, Lewerke LT, Hastie JL, Ellermeier CD. 2018. Signal peptidase is necessary and sufficient for site 1 cleavage of RsiV in *Bacillus subtilis* in response to lysozyme. *J Bacteriol* 200:e00663-17. <https://doi.org/10.1128/JB.00663-17>.
- Li W, Bottrill AR, Bibb MJ, Buttner MJ, Paget MSB, Kleanthous C. 2003. The role of zinc in the disulphide stress-regulated anti-sigma factor RsrA from *Streptomyces coelicolor*. *J Mol Biol* 333:461–472. <https://doi.org/10.1016/j.jmb.2003.08.038>.
- Trepreau J, Girard E, Maillard AP, De Rosny E, Petit-Haertlein I, Kahn R, Covès J. 2011. Structural basis for metal sensing by CnrX. *J Mol Biol* 408:766–779. <https://doi.org/10.1016/j.jmb.2011.03.014>.
- Francez-Charlot A, Frunzke J, Reichen C, Ebner JZ, Gourion B, Vorholt JA. 2009. Sigma factor mimicry involved in regulation of general stress response. *Proc Natl Acad Sci U S A* 106:3467–3472. <https://doi.org/10.1073/pnas.0810291106>.
- Lourenço RF, Kohler C, Gomes SL. 2011. A two-component system, an anti-sigma factor and two paralogous ECF sigma factors are involved in the control of general stress response in *Caulobacter crescentus*. *Mol Microbiol* 80:1598–1612. <https://doi.org/10.1111/j.1365-2958.2011.07668.x>.
- Campagne S, Allain F-T, Vorholt JA. 2015. Extra cytoplasmic function sigma factors, recent structural insights into promoter recognition and regulation. *Curr Opin Struct Biol* 30:71–78. <https://doi.org/10.1016/j.sbi.2015.01.006>.
- Schumacher MA, Bush MJ, Bibb MJ, Ramos-León F, Chandra G, Zeng W, Buttner MJ. 2018. The crystal structure of the RsbN- σ BldN complex from *Streptomyces venezuelae* defines a new structural class of anti- σ factor. *Nucleic Acids Res* 46:7405–7417. <https://doi.org/10.1093/nar/gky493>.
- Campbell EA, Greenwell R, Anthony JR, Wang S, Lim L, Das K, Sofia HJ, Donohue TJ, Darst SA. 2007. A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. *Mol Cell* 27:793–805. <https://doi.org/10.1016/j.molcel.2007.07.009>.
- Devkota SR, Kwon E, Ha SC, Chang HW, Kim DY. 2017. Structural insights into the regulation of *Bacillus subtilis* SigW activity by anti-sigma RsiW. *PLoS One* 12:e0174284. <https://doi.org/10.1371/journal.pone.0174284>.
- Shukla J, Gupta R, Thakur KG, Gokhale R, Gopal B. 2014. Structural basis for the redox sensitivity of the *Mycobacterium tuberculosis* SigK-RskA σ -anti- σ complex. *Acta Crystallogr D Biol Crystallogr* 70:1026–1036. <https://doi.org/10.1107/S1399004714000121>.
- Campbell EA, Tupy JL, Gruber TM, Wang S, Sharp MM, Gross CA, Darst SA. 2003. Crystal structure of *Escherichia coli* σ E with the cytoplasmic domain of its anti- σ RseA. *Mol Cell* 11:1067–1078. [https://doi.org/10.1016/s1097-2765\(03\)00148-5](https://doi.org/10.1016/s1097-2765(03)00148-5).
- Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 106:67–72. <https://doi.org/10.1073/pnas.0805923106>.
- Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. 2012. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci U S A* 109:E1733–E1742. <https://doi.org/10.1073/pnas.1201301109>.
- Jogler C, Waldmann J, Huang X, Jogler M, Glöckner FO, Mascher T, Kolter R. 2012. Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in *Planctomycetes* by comparative genomics. *J Bacteriol* 194:6419–6430. <https://doi.org/10.1128/JB.01325-12>.
- Huang X, Pinto D, Fritz G, Mascher T. 2015. Environmental sensing in *Actinobacteria*: a comprehensive survey on the signaling capacity of this phylum. *J Bacteriol* 197:2517–2535. <https://doi.org/10.1128/JB.00176-15>.
- Sineva E, Savkina M, Ades SE. 2017. Themes and variations in gene regulation by extracytoplasmic function (ECF) sigma factors. *Curr Opin Microbiol* 36:128–137. <https://doi.org/10.1016/j.mib.2017.05.004>.
- Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* 43:W401–W407. <https://doi.org/10.1093/nar/gkv485>.
- Newman JD, Anthony JR, Donohue TJ. 2001. The importance of zinc-binding to the function of *Rhodobacter sphaeroides* ChrR as an anti-sigma factor. *J Mol Biol* 313:485–499. <https://doi.org/10.1006/jmbi.2001.5069>.
- Rajasekar KV, Zdanowski K, Yan J, Hopper JTS, Francis MLR, Seepersad C, Sharp C, Pecqueur L, Werner JM, Robinson CV, Mohammed S, Potts JR, Kleanthous C. 2016. The anti-sigma factor RsrA responds to oxidative stress by burying its hydrophobic core. *Nat Commun* 7:12194. <https://doi.org/10.1038/ncomms12194>.
- Lewerke LT, Kies PJ, Müh U, Ellermeier CD. 2018. Bacterial sensing: a putative amphipathic helix in RsiV is the switch for activating σ V in response to lysozyme. *PLoS Genet* 14:e1007527. <https://doi.org/10.1371/journal.pgen.1007527>.
- Li S, Lou X, Xu Y, Teng X, Liu R, Zhang Q, Wu W, Wang Y, Bartlam M. 2019. Structural basis for the recognition of MucA by MucB and AlgU in *Pseudomonas aeruginosa*. *FEBS J* 286:4982–4994. <https://doi.org/10.1111/febs.14995>.
- El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432. <https://doi.org/10.1093/nar/gky995>.
- Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Coevolution of proteins with their interaction partners. *J Mol Biol* 299:283–293. <https://doi.org/10.1006/jmbi.2000.3732>.
- Dintner S, Staroń A, Berchtold E, Petri T, Mascher T, Gebhard S. 2011. Coevolution of ABC transporters and two-component regulatory systems as resistance modules against antimicrobial peptides in *Firmicutes* bacteria. *J Bacteriol* 193:3851–3862. <https://doi.org/10.1128/JB.05175-11>.
- Dufour A, Haldenwang WG. 1994. Interactions between a *Bacillus subtilis* anti-sigma factor (RsbW) and its antagonist (RsbV). *J Bacteriol* 176:1813–1820. <https://doi.org/10.1128/jb.176.7.1813-1820.1994>.
- Burgess RR, Travers AA, Dunn JJ, Bautz E. 1969. Factor stimulating transcription by RNA polymerase. *Nature* 221:43–46. <https://doi.org/10.1038/221043a0>.
- Wu H, Liu Q, Casas-Pastor D, Dürr F, Mascher T, Fritz G. 2019. The role of C-terminal extensions in controlling ECF σ factor activity in the widely conserved groups ECF 41 and ECF 42. *Mol Microbiol* 112:498–514. <https://doi.org/10.1111/mmi.14261>.
- Li L, Fang C, Zhuang N, Wang T, Zhang Y. 2019. Structural basis for transcription initiation by bacterial ECF σ factors. *Nat Commun* 10:1153. <https://doi.org/10.1038/s41467-019-09096-y>.
- Wilson MJ, Lamont IL. 2006. Mutational analysis of an extracytoplasmic-

- function sigma factor to investigate its interactions with RNA polymerase and DNA. *J Bacteriol* 188:1935–1942. <https://doi.org/10.1128/JB.188.5.1935-1942.2006>.
35. Lane WJ, Darst SA. 2010. Molecular evolution of multisubunit RNA polymerases: structural analysis. *J Mol Biol* 395:686–704. <https://doi.org/10.1016/j.jmb.2009.10.063>.
 36. Lane WJ, Darst SA. 2006. The structural basis for promoter -35 element recognition by the group IV sigma factors. *PLoS Biol* 4:e269. <https://doi.org/10.1371/journal.pbio.0040269>.
 37. Rhodius VA, Segall-Shapiro TH, Sharon BD, Ghodasara A, Orlova E, Tabakh H, Burkhardt DH, Clancy K, Peterson TC, Gross CA, Voigt CA. 2013. Design of orthogonal genetic switches based on a crosstalk map of σ s, anti- σ s, and promoters. *Mol Syst Biol* 9:702. <https://doi.org/10.1038/msb.2013.58>.
 38. de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet* 14:249–261. <https://doi.org/10.1038/nrg3414>.
 39. Rausell A, Juan D, Pazos F, Valencia A. 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A* 107:1995–2000. <https://doi.org/10.1073/pnas.0908044107>.
 40. Ochs M, Angerer A, Enz S, Braun V. 1996. Surface signaling in transcriptional regulation of the ferric citrate transport system of *Escherichia coli*: mutational analysis of the alternative sigma factor FecI supports its essential role in fec transport gene transcription. *Mol Gen Genet* 250:455–465. <https://doi.org/10.1007/s004380050098>.
 41. Missiakas D, Mayer MP, Lemaire M, Georgopoulos C, Raina S. 1997. Modulation of the *Escherichia coli* sigmaE (RpoE) heat-shock transcription-factor activity by the RseA, RseB and RseC proteins. *Mol Microbiol* 24:355–371. <https://doi.org/10.1046/j.1365-2958.1997.3601713.x>.
 42. Paget MS. 2015. Bacterial sigma factors and anti-sigma factors: structure, function and distribution. *Biomolecules* 5:1245–1265. <https://doi.org/10.3390/biom5031245>.
 43. Maillard AP, Girard E, Ziani W, Petit-Härtlein I, Kahn R, Covès J. 2014. The crystal structure of the anti- σ factor CnrY in complex with the σ factor CnrH shows a new structural class of anti- σ factors targeting extracytoplasmic function σ factors. *J Mol Biol* 426:2313–2327. <https://doi.org/10.1016/j.jmb.2014.04.003>.
 44. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 47:W636–W641. <https://doi.org/10.1093/nar/gkz268>.
 45. Jamithreddy AK, Runthala A, Gopal B. 2020. Evaluation of specificity determinants in *Mycobacterium tuberculosis* σ /anti- σ factor interactions. *Biochem Biophys Res Commun* 521:900–906. <https://doi.org/10.1016/j.bbrc.2019.10.198>.
 46. Sorenson MK, Ray SS, Darst SA. 2004. Crystal structure of the flagellar σ /anti- σ complex σ 28/FlgM reveals an intact σ factor in an inactive conformation. *Mol Cell* 14:127–138. [https://doi.org/10.1016/S1097-2765\(04\)00150-9](https://doi.org/10.1016/S1097-2765(04)00150-9).
 47. Patikoglou GA, Westblade LF, Campbell EA, Lamour V, Lane WJ, Darst SA. 2007. Crystal structure of the *Escherichia coli* regulator of sigma70, Rsd, in complex with sigma70 domain 4. *J Mol Biol* 372:649–659. <https://doi.org/10.1016/j.jmb.2007.06.081>.
 48. Jishage M, Dasgupta D, Ishihama A. 2001. Mapping of the Rsd contact site on the sigma 70 subunit of *Escherichia coli* RNA polymerase. *J Bacteriol* 183:2952–2956. <https://doi.org/10.1128/JB.183.9.2952-2956.2001>.
 49. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. <https://doi.org/10.1038/msb.2011.75>.
 50. Nguyen NPD, Mirarab S, Kumar K, Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol* 16:124. <https://doi.org/10.1186/s13059-015-0688-z>.
 51. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29–W37. <https://doi.org/10.1093/nar/gkr367>.
 52. Olechnovič K, Venclovas C. 2014. Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *J Comput Chem* 35:672–681. <https://doi.org/10.1002/jcc.23538>.
 53. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612. <https://doi.org/10.1002/jcc.20084>.
 54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830.
 55. Wilbur WJ, Lipman DJ. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc Natl Acad Sci U S A* 80:726–730. <https://doi.org/10.1073/pnas.80.3.726>.
 56. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
 57. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>.
 58. Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, Wu CH. 2011. A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 27:1190–1191. <https://doi.org/10.1093/bioinformatics/btr101>.
 59. Jones E, Oliphant T, Peterson P. 2001. SciPy: open source scientific tools for Python.
 60. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. 2014. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 9:e92721. <https://doi.org/10.1371/journal.pone.0092721>.
 61. Hamelryck T, Manderick B. 2003. PDB file parser and structure class implemented in Python. *Bioinformatics* 19:2308–2310. <https://doi.org/10.1093/bioinformatics/btg299>.
 62. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon M. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.