Supplementary Methods

Homology modeling of the XylS N-terminal domain

Identification of template

The analysis was based on XylS from *Pseudomonas putida* (SwissProt (Boeckmann et al. 2003) sequence entry XYLS_PSEPU). The NCBI Conserved Domain (Marchler-Bauer et al. 2005) search tool showed that the full XylS sequence consists of two domains, a N-terminal domain of approximately 210 residues containing a Pfam AraC binding motif (E-value 2e-06), and a C-terminal domain containing a Smart HTH_ARAC helix-turn-helix motif (E-value 3e-15). This is consistent with experimental data (Kaldalu et al. 2000). The 3D structure of the AraC binding domain is known from the structure of AraC from *Escherichia coli* (ARAC_ECOLI), residues 7-170 (PDB (Berman et al. 2000) code 2ARC (Soisson et al. 1997)), and this was used as a template for building a homology model of the N-terminal domain of XylS.

Finding optimal alignment

The sequence similarity between XylS and AraC is quite low. Alignment of XylS 1-210 vs. AraC 1-170 using Blast bl2seq (Tatusova and Madden 1999) found no significant similarity. The Fasta Lalign tool (Pearson and Lipman 1988; (Huang and Miller 1991) found alignments with approximately 20% sequence identity. However, the actual alignment was very dependent upon choice of mutation matrix, and the significance estimated by PRSS (Pearson 1996) was not very strong (0.0014 and 0.98 for the two best alignments). Therefore several alternative alignment strategies were tested, in particular the BioInfoBank Meta server with 3DJury evaluation (Ginalski et al. 2003) , PSI-Blast (Altschul et al. 1997), Pfam (Finn et al. 2006), LOOP (Tobi and Elber 2000; (Meller and Elber 2001; (Teodorescu et al. 2004) and 123D+ (Alexandrov et al. 1996). The PSI-Blast alignm (Wheeler et al. 2003)ent corresponds to 15 iterations on the NCBI non-redundant database, starting with residues 1-210 of XylS.

All methods basically agree on the first part of the alignment (see **Supplementary Figure 1**), and in particular on the region from 75 to 120 in XylS, corresponding to the central part of the β-barrel in AraC. For the very first part of the alignment, corresponding to the first part of the β-barrel, 3DJury, PSI-Blast and 123D+ on one hand and Pfam and LOOPP on the other hand agree on slightly different alignments, mainly shifted by two residues. For the region from

120 to 150 in XylS, corresponding to the final strand of the β-barrel of AraC and the linkage between the barrel and the two a-helices, 3DJury and Pfam agree on the same alignment, whereas LOOPP and 123D+ indicate very different alignments. For the final part of the alignment, corresponding to the two a-helices in the AraC dimer interface, none of the methods agree. However, in particular for the last a-helix the alternative solutions are basically shifts of approximately 3-4 residues, corresponding to in-phase shifts of the α-helix.

These alternative alignments were used for making a consensus alignment for model building. Sequence libraries with representative sequences from the XylS and AraC protein families were generated with PSI-Blast and aligned with ClustalX (Thompson et al. 1997). Conserved positions were identified in these multiple alignments, and this information was added to the consensus alignment (**Supplementary Figure 1**). This information indicates that it is reasonable to have a gap of length 2 in XylS around position 70, as suggested by Pfam. The conservation pattern also confirms the alignment between 75 and 110. The alignment close to position 130 and the conservation pattern in this region indicates a gap of 3 residues in XylS just before position 120. For the last part of the alignment both the alignment alternatives and the conservation pattern is noisy. However, it seems reasonable to assume that there is a gap of unknown length in AraC, close to position 175 in XylS. Model building was used for testing these alternative gap positions.

Model building

The consensus alignment described above was used as a starting point, and SwissModel (Guex and Peitsch 1997) was used to build models of the XylS dimer, based on the AraC dimer, with several alternative gap positions. The final total energy from SwissModel was used as an indication of model quality for a given gap position. This energy will at least indicate serious problems associated with a specific model. Initially only the position of each gap was tested, using a gap length of 2 for the final gap. Several alternative positions were tested in a systematic way, using more than 50 different models, and the optimal position (giving the lowest total energy) was found for each gap. During this optimization the total energy of the model changed from -6000 to -10000 kJ/mol. The final gap positions are shown in **Supplementary Figure 1**.

Using these refined gap positions, the same strategy was tested in order to find the optimal gap length of the final gap. However, in this case the energy values did not show any clear

trend, instead there were several local minima. Therefore a different strategy was used. The last gap mainly affects the relative position of the final a-helix, which is crucial for the protein-protein interface of the dimer. Therefore the interaction energy of the dimer interface was estimated and used as criteria. Gap lengths from 0 to 13 were tested, using three different tools (FastContact (Camacho and Zhang 2005), DComplex (Liu et al. 2004) and RPDock (Moont et al. 1999)), and with SwissModel pdb model files as input. The score values are shown in **Supplementary Figure 2**. Although there is some noise, the scores are clearly correlated and also show a clear periodicity between 3 and 4, as expected. Strong local minima are found for gap lengths around 3, 7 and 10.

For comparison a similar analysis was done on AraC, using AraC itself as template and introducing gaps of varying length into the AraC query sequence. The result (**Supplementary Figure 2**) shows that FastContact is relatively noisy, whereas both RPDock and in particular DComplex are able to identify the correct optimal gap length of zero. This may indicate that the optimal gap length for XylS is close to 3, as both RPDock and in particular DComplex have a clear minimum in this region. However, it is difficult to rule out the other gap lengths completely. Therefore a slightly different strategy was tested as well, where only the monomer was modeled, and the position of the last helix was shifted relative to the rest of the structure. The estimated interaction energies for XylS and AraC are shown (**Supplementary Figure 2**). For AraC it is again DComplex that is best able to identify the correct global minimum at gap length 0, although all methods are able to find reasonable minima when in-phase shifts of the helix are included. The data from the XylS analysis are slightly noisier compared to the data for AraC, although all three methods show a clear minimum at gap length 1. Unfortunately the score values for inter- and intra-protein contact seem to be anti-correlated. It is therefore difficult to find the optimal solution. Possibly are the XylS helices slightly twisted compared to the AraC helices, so that alternative residue positions are used for stabilizing the interaction. However, the gap length of 2 identified by 3DJury seems to be a reasonable compromise. This is also supported by Errat (Colovos and Yeates 1993), which was used to estimate model quality at all gap lengths from 0 to 13. Only the gap lengths 2, 4 and 7 have no residues above the 95% error value for the sub-sequence corresponding to the two a-helices. Gap length 4 has non-favorable dimer interaction energy. Gap length 2 and 7 are both possible solutions. However, if we assume that the loop between the two α-helices is similar in XylS and AraC, and that the consensus prediction of secondary structure is correct, then a gap length of 2 seems to be the best alternative.

The final alignment is shown in **Supplementary Figure 1**. This alignment was used for building the final 3D model of XylS. Errat plots of key steps are shown in **Supplementary Figure 3**, and the final model is shown in **Supplementary Figure 4**.

Discussion

As shown the sequence similarity between XylS and AraC is very low, and the model of XylS has to be used by care. The central part of the β-barrel is reasonably conserved, and it is probably safe to assume that this part of the model is of good quality. The model of the helices forming the dimer interface between the proteins is less well defined. This is not surprising, as this interaction is essential for selectivity and specificity in dimerization of molecules belonging to the XylS/AraC family. This is a very large family, and it is reasonable to assume that there is a strong evolutionary pressure towards diversification of contact interfaces, and therefore high degree of variation between subfamilies of sequences, in order to maintain specificity in the dimerization.

This affects the relative position of both a-helices in XylS. From **Supplementary Figure 1** we see that the predicted XylS helix starting at 157 is shifted 2 positions relative to the known AraC helix. It has been tried to shift this helix 2 positions during modeling, but this did not improve the SwissModel energy or the Errat score. Similarly there are several alternative alignments of the XylS helix starting at position 179. Here the predicted XylS helix is shorter than the known AraC helix. This will affect the model evaluation, given that the secondary structure prediction is correct, as non-helical regions will be forced into wrong secondary structure. This will e.g. affect the N-terminal part of the helix, where there is a Pro in XylS, and Pro is a known helix breaker. The length of the turn region between the helices indicated here is consistent with the length of the same region in AraC. However, the AraC turn is very tight, with three Gly in the turn region. This could indicate a slightly longer turn in XylS, which would be consistent with a shorter helix.

The proposed structure is to some extent supported by published experimental data on XylS and AraC. It has been shown that the mutations L150K, L151K, N154A and L161S in AraC disrupt the dimerization (LaRonde-LeBlanc and Wolberger 2000). Similarly it has been shown that the mutations L193A and L194A in XylS affect dimerization (Ruíz et al. 2003). These mutations fall in the final a-helix in the present model, and if we increase the gap

length of the final gap to 4 or 5 L161 will overlap with L193 or L194. However, it is not easy to compare these data. Although the mutations in AraC were rationally chosen based on the known 3D structure, the mutation study showed that the interface is very strong, and all 4 mutations had to be applied in order to disrupt the dimer. No Ala scanning was performed. Therefore we do not know whether the essential residues for dimer stability actually have been found. The same is true for XylS, where only a small number of mutations have been tested, based on a short alignment between XylS and AraC. That alignment actually corresponds to a gap length of 15 in our model. It is not likely that this is a correct alignment, although it leads to a reasonable conclusion. Also, as already mentioned the dimerization domains for different sequence families are most likely different with respect to specificity. Therefore there may not be any strong conserved features that can be aligned across sequence families. However, the mutations confirm that the alignment (and the model) most likely identifies the dimerization domain correctly.

Alexandrov, N. N., R. Nussinov and R. M. Zimmer (1996). Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pacific Symposium on Biocomputing.*: 53-72.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17): 3389-3402.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**(1): 235-242.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estericher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**(1): 365-370.

Camacho, C. J. and C. Zhang (2005). FastContact: rapid estimate of contact and binding free energies. *Bioinformatics.* **21**(10): 2534-2536.

Colovos, C. and T. O. Yeates (1993). Verification of protein structures - patterns of nonbonded atomic interactions. *Protein Sci.* **2**(9): 1511-1519.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer and A. Bateman (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**((Database issue): D247-51.).

Ginalski, K., A. Elofsson, D. Fischer and L. Rychlewski (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics.* **19**(8): 1015-1018.

Guex, N. and M. C. Peitsch (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis.* **18**(15): 2714-2723.

Huang, X. Q. and W. Miller (1991). A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**(3): 337-357.

Kaldalu, N., U. Toots, V. d. Lorenzo and M. Ustav (2000). Functional domains of the TOL plasmid transcription factor XylS. *J. Bacteriol.* **182**(4): 1118-1126.

LaRonde-LeBlanc, N. and C. Wolberger (2000). Characterization of the oligomeric states of wild type and mutant AraC. *Biochemistry.* **39**: 11593-11601.

Liu, S., C. Zhang, H. Zhou and Y. Zhou (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins.* **56**(1): 93-101.

Marchler-Bauer, A., J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang and S. H. Bryant (2005). CDD: a conserved domain database for protein classification. *Nucleic Acids Res.* **33**((Database issue): D192-196).

Meller, J. and R. Elber (2001). Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins.* **45**(3): 241-261.

Moont, G., H. A. Gabb and M. J. Sternberg (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins.* **35**(3): 364-373.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**: 227-258.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci. USA.* **85**(8): 2444-2448.

Ruíz, R., S. Marqués and J. L. Ramos (2003). Leucines 193 and 194 at the N-terminal domain of the XylS protein, the positive transcriptional regulator of the TOL *meta*-cleavage pathway, are involved in dimerization. *J. Bacteriol.* **185**(10): 3036-4041.

Sletta, H., A. Nedal, T. E. V. Aune, H. Hellebust, S. Hakvåg, R. Aune, T. E. Ellingsen, S. Valla and T. Brautaset (2004). Broad-host-range plasmid pJB658 can be used for industrial-level production of a secreted host-toxic single-chain antibody fragment in *Escherichia coli*. *Appl. Env. Microbiol.* **70**(12): 7033-7039.

Soisson, S., B. MacDougall-Shackleton, R. Schleif and C. Wolberger (1997). Structural basis for ligand-regulated oligomerization of AraC. *Science.* **278**(5311): 421-425.

Tatusova, T. A. and T. L. Madden (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**(2): 247-250.

Teodorescu, O., T. Galor, J. Pillardy and R. Elber (2004). Enriching the sequence substitution matrix by structural information. *Proteins.* **54**(1): 41-48.

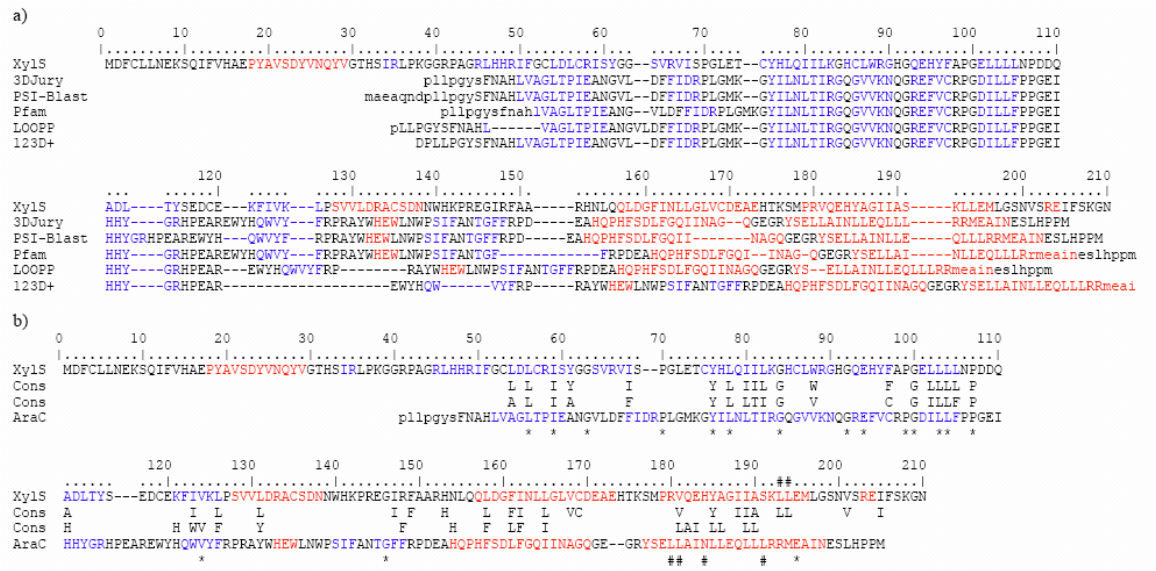
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). The CLUSTAL_X window interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**(24): 4876-4882.

Tobi, D. and R. Elber (2000). Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins.* **41**(1): 40-46.

Wheeler, D. L., D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova and L. Wagner (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**(1): 28-33.

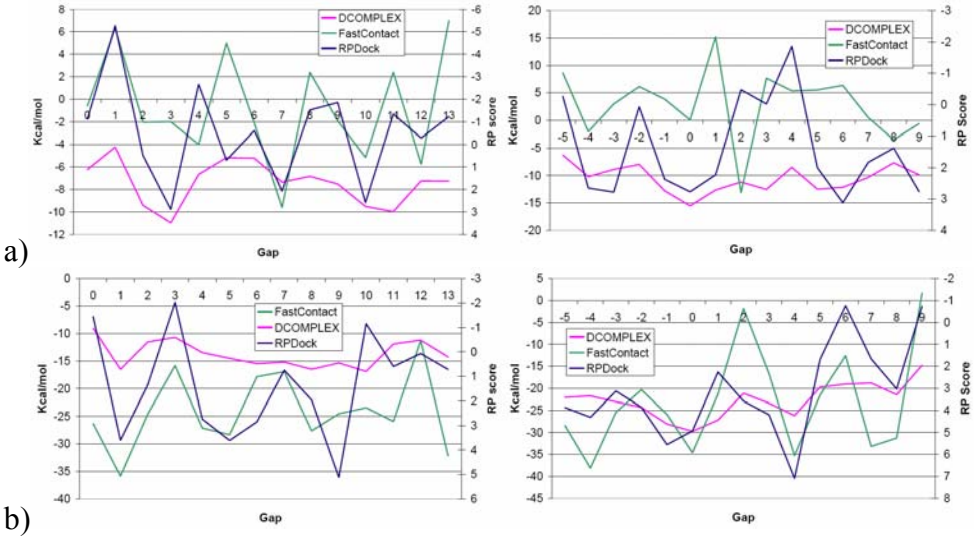
Winther-Larsen, H. C., J. M. Blatny, B. Valand, T. Brautaset and S. Valla (2000). *Pm* promoter expression mutants and their use in broad-host-range RK2 plasmid vectors. *Metab. Eng.* **2**: 92-103.

Supplementary Figure 1. Alignment of XylS and AraC



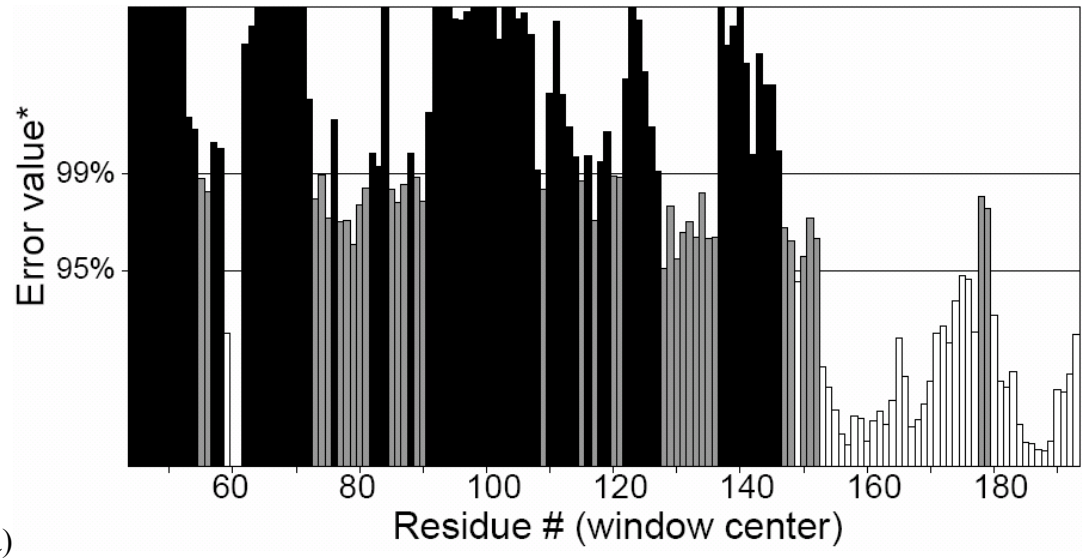
a) Pairwise alignment of XylS and AraC for different alignment strategies. The alignment is numbered according to XylS, and secondary structure is indicated with blue (β-strand) and red (α-helix). The predicted secondary structure for XylS is from the BioInfoBank Meta server. b) Final consensus alignment for XylS and AraC used for model building. Conserved positions for each subfamily are shown between the sequences. Mutations discussed in the text are indicated with #.

Supplementary Figure 2. Score values for alternative alignments

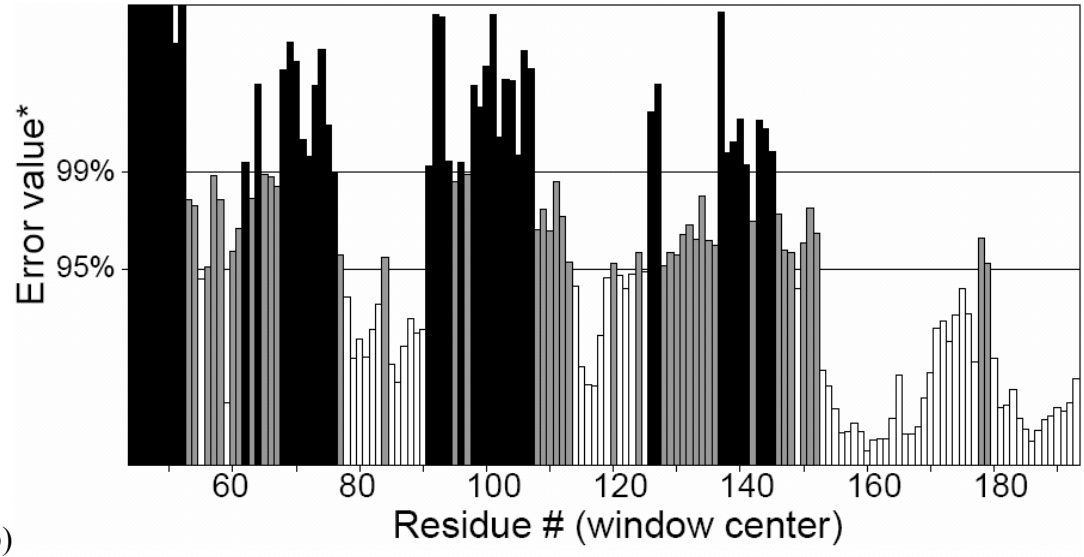


Score values from DComplex, Fastcontact (kcal/mol) and RPDock (RP score). a) Interaction score for the XylS dimer (left) and the AraC dimer (right) at different gap lengths. b) Interaction score for the final α-helix vs. the rest of the XylS monomer (left) and same for the AraC monomer (right). See Supplementary Methods text for details.

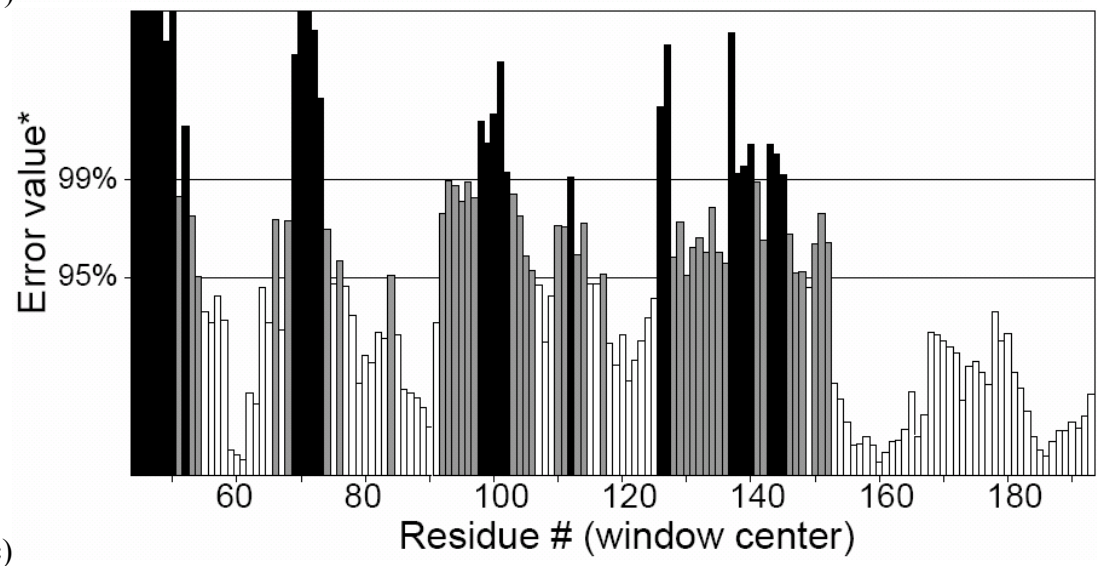
Supplementary Figure 3. Errat scores



b)

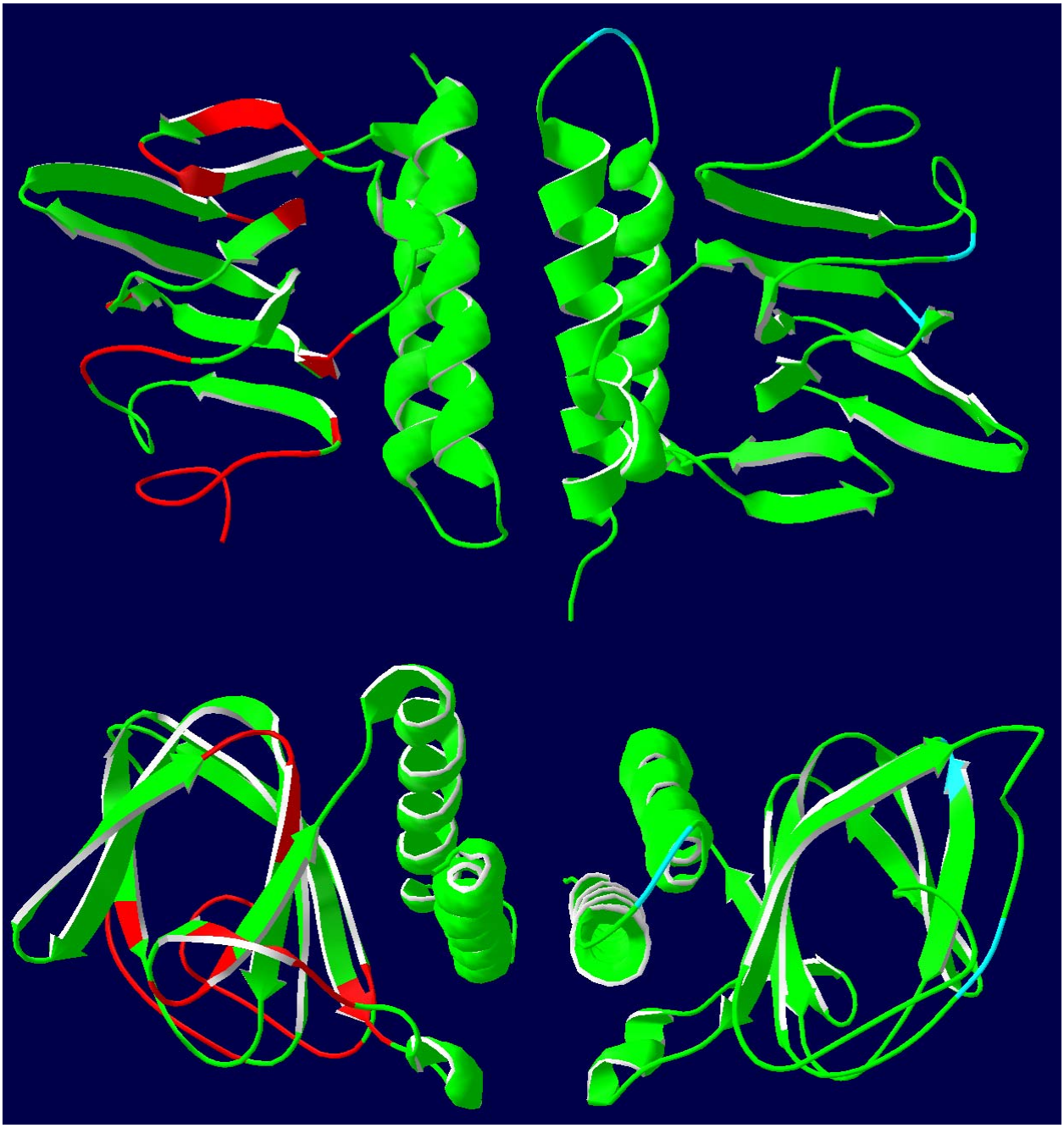


c)



Errat scores for a) initial XylS monomer model, b) initial XylS dimer model, and c) XylS dimer model after optimisation of gap positions, corresponding to the final alignment in Figure 1. For the dimer model only the first chain is shown.

Supplementary Figure 4. Final model



3D representations of the final model, in two different orientations. The problematic regions from the Errat score in Figure 3 are indicated with red in the left part of the structure, showing that this mainly affects the extreme parts of the β-barrel. The gap positions are indicated with cyan in the right part of the structure.