Introduction
The gram negative bacterium *Vibrio cholerae*, the causative agent of cholera, produces cholera toxin and colonizes in the small intestine of the host. Along with the toxin, a toxin coregulated pilus (TCP) is expressed by *V. cholerae* as part of the *toxR* virulence regulon (Miller et al., 1987; Taylor et al., 1987). TCP is required for colonization of the bacterium in the intestinal mucosa of the host, shown in the infant mouse cholera model (Taylor et al., 1987; Thelin and Taylor, 1996) and human volunteer studies (Herrington et al., 1988).

TCP is a member of the type IV pili which are 1000–4000 nm long, 60–90 Å diameter fibres formed by ordered association of thousands of identical pilus subunits plus a few copies of pilus associated proteins (Rudel et al., 1995). TcpA, the major component of TCP, is a 20 kDa protein involved in bacterial adhesion to host cell surfaces (Strom and Lory, 1993). Other functions of TcpA as a CTXφ receptor (Waldor and Mekalanos, 1996) and as a coat protein of a bacteriophage VPI6 (Karaolis et al., 1999) make it more important for study. Many bacteria possess a similar type IV pilin like TcpA (Strom and Lory, 1993; Giron et al., 1997) assembled by a common cellular process (Strom et al., 1994). The crystal structure of only one type IV pilin, i.e. PilE of *Neisseria gonorrhoeae* (Parge et al., 1995) was used to build our TcpA model but subsequently another crystal structure of a truncated *Pseudomonas aeruginosa* pilin (Hayes et al., 2000) was reported.

We demonstrate that using the crystal structure of the 158-residue *N. gonorrhoeae* pilin (PilE) it is possible to build a reliable model of the *V. cholerae* TcpA consistent with the available data in the literature on TcpA.

Materials and methods
PilE crystal structure provides a basis for TcpA structure
We argue that the PilE crystal structure is an excellent guide for a major part of the TcpA structure since:

1. The hydrophathy plot is extremely similar between TcpA and PilE for the 30 N-terminal residues (Figure 3a).
2. Identical residues at positions –1 (Gly), 3 (Leu), 5 (Glu), 9 (Val) and 18 (Ala) are found in all nine type IV pilins including PilE and TcpA (Strom and Lory, 1993).
3. Single residue mutations at positions –1, 4, 9 and 20 in PilE were found to abolish autoagglutination (AA) in *N. gonorrhoeae*. Corresponding mutations in TcpA also demonstrated AA defects and were highly defective in colonization (Chiang et al., 1995).
4. In addition, both PilE and TcpA have their first cysteine at position 120 or 121, though the position of the second cysteine differs considerably. This suggests that the protein fold must be extremely similar for both pilins till approximately position 120.
5. Proteins sharing an α–β roll structure usually have related functions (Orengo and Thornton, 1993) and therefore the crystal structure (Parge et al., 1995) represents a new structural and functional class of α–β roll having wider applicability beyond *N. gonorrhoeae*.
6. A sequence alignment of classical TcpA with PilE shows only ~23% sequence similarity. This value of sequence similarity between two proteins may or may not mean the same fold. Sequence similarity is no longer considered to be a good criterion in this ‘grey’ area, as pairwise sequence comparisons have been assessed and statistical methods of identifying distant evolutionary relationships evaluated (Brenner et al., 1998). For investigating such relationships, a SCOP database search was performed. The primary sequence of classical strain of TcpA was used as a query sequence in a SCOP database search at the website http://stash.mrc-lmb.cam.ac.uk/PDB_JSL/ and this strongly indicated structural similarity with PilE with a Z-score of 776 and an E-value of 3.3×10^-48. This also showed a pilin fold as in PilE and an α–β protein.

Highly conserved TcpA sequences
Pairwise sequence identities between the four TcpA sequences range from 78 to 82%, with 138 invariant positions out of 197 (70%) (Figure 1). Such a high degree of sequence identity suggests a near identical three-dimensional structure (Chothia and Lesk, 1986; Brenner et al., 1998) among the strains, with hypervariable regions predicted to be mostly on the surface,
Fig. 1. Structural alignment of a PilE primary sequence with four primary sequences of TcpA associated with classical O1 (Taylor et al., 1987), El Tor O1 (Rhine and Taylor, 1994), non-O1 non-O139 (Nandi et al., 2000) and another non-O1 non-O139 strain (Novais et al., 1999) shown up to residue 137. After this residue, the two structures do not share corresponding secondary structural elements. Dots in the last three TcpA sequences indicate identity with the classical O1 strain. Residues contributing side chains for any one of the hydrophobic cores in our monomer model are marked with the symbol 'c'. The 72 core residues are mostly hydrophobic in TcpA. Residues with side chains exposed in the fibre are marked with the symbol 'e'. The next line indicates the secondary structure of TcpA in our final model and the secondary structure of PilE is given on the last line for comparison, showing that up to β6, there exists corresponding secondary structural elements in both pilin structures.

Secondary structure prediction for TcpA and PilE by the GCG package

We compared the secondary structure prediction by both Chou–Fasman and Garnier–Osguthorpe–Robson methods (Genetics Computer Group, 1995) for PilE with its secondary structure in the crystal (Parge et al., 1995). By the Chou–Fasman method, only 20 of the first 54 residues comprising as in the case of PilE (Jonsson et al., 1994; Parge et al., 1995). The variable regions 48–63, 69–85, 91–103, 111–114, 126–143 and 153–158 of PilE in various gonococcal strains (Haas and Meyer, 1986) were examined by us in the light of the PilE crystal structure: regions 56–63 and 79–85 are partly exposed, whereas 91–103, 126–143 and 153–157 are exposed and only 111–114 buried in the PilE fibres.
α1 are predicted to be α-helical, regions α2 and α3 in the crystal structure not predicted to be helical, and among the six β-strands, only β2, β4 and β5 (partly) are predicted. The accuracy of this secondary structure prediction for PilE was poor by the GCG package. Therefore, it can be expected to be poor for TcpA as well, since TcpA is structurally and functionally homologous to PilE.

Residuewise raw Chou–Fasman probabilities used to build model except for α

Though secondary structure predictions performed within the GCG package were judged to be unreliable, residuewise αα, ββ, Pt raw values obtained within the InsightII package (Biosym/MSI, 1995) were used to build our model. The region 1–54 showed higher ββ values than αα particularly for regions 1–21 and 42–45. Though region 60–70 is variable among the four TcpA sequences (Figure 1), it showed higher Ptα values compared to ββ for all four sequences, thus indicating αα in this region. Figure 2 plots ββ values of PilE and TcpA. In all four sequences, region 70–77 (ββ, and preceding loop) showed higher ββ values than Ptα. Similarly, region 87–98 (ββ) showed higher ββ values than Ptα. However, residues 104–110 (ββ) showed almost equal values of Ptα and ββ but since it is so hydrophobic it is likely to be a β-strand; also by comparison with PilE. All four TcpA sequences indicated higher ββ values than Ptα for the regions 117–121 (ββ), 122–127 (ββ) and 130–136 (ββ). Higher Ptα values than ββ were found for all four sequences for regions 138–145 (αα) and 152–158 (αα). Higher Ptα values than Ptβ were found for regions 159–165 (ββ), 175–179 (ββ) and 192–196 (ββ). Thus, except for αα and some ambiguity about 104–110, the residuewise plots of Chou–Fasman probabilities were used for the secondary structure assignment in our model.

PilE crystal structure used for building TcpA model for residues 1–137

The residuewise Chou–Fasman probabilities of TcpA and PilE indicated that up to residue 137, TcpA possesses corresponding secondary structural elements as in the PilE crystal structure. Since TcpA will be stabilized by a similar hydrophobic core as in PilE, secondary structural elements were accordingly placed so as to conserve the hydrophobic cores. Differences between our TcpA model and the PilE crystal structure within the first 137 residues can be noted from Figure 4a and the description of the model below.

Hydropathicity plots of PilE and TcpA compared

Figure 3a gives Kyte–Doolittle hydropathicity plots of both proteins obtained using the GCG package (Genetics Computer Group, 1995). The plot is extremely similar for both pilus for regions 1–30, 50–70 and 90–104. However, TcpA appears to be more hydrophobic compared to PilE (Figure 3a), particularly in regions 41–48 (the later part of αα), 70–80 (ββ and its preceding loop), 105–115 (ββ), 123–137 (ββ, ββ).

How region 138–169 was built

After residue 137, PilE does not have any α-helix (Parge et al., 1995). But for TcpA, regions 138–145 and 153–158 show strong helical preference (Pα > 1.2) in all four sequences. Similarly, all four sequences show a strong preference for β-strand in region 160–167 (Pβ > 1.15). Two additional constraints were used for this region: (i) mutating Glu158 to Leu produces a kinked pilus (Sun et al., 1991), hence Glu158 must be in the interior of the fibre, interacting with αα, and (ii) ββ (163–167) is one of the most hydrophobic regions (Figure 3a) and should remain buried in the fibre; this was achieved by forming an antiparallel sheet of ββ (Figure 4a).

Placement of the two last β-strands

All four TcpA sequences indicated ββ in the region 176–180 (Pβ > 1.2) and ββ in the region 192–196 (Pβ > 1.15). From residue 169 onwards, we predict that the protein chain must travel radially outwards in the fibre since both the surface probability (Figure 3b) and antigenicity index (Figure 3c) show large peaks beginning at residue 170. Here the protein chain must not cover the loop containing Glu83, thought to be exposed (Sun et al., 1991). Hence ββ can form a parallel β-sheet with the hitherto exposed face of ββ. Next, the Cys186-containing loop runs antiparallel to ββ, which contains Cys120. The last β-strand ββ is one of the most hydrophobic regions of the protein (Figure 3a) and was therefore buried in the hydrophobic pocket between ββ and ββ within the subunit.

Refinement

The whole protein structure was energy refined in stages using the Discover program within InsightII.

Surface probability plot

In our model, αα and αα are exposed on the surface (Figure 4a), thus agreeing with peaks in surface probability for residues 22–70 (Figure 3b). The highest peak at 82–85 agrees with our model as this is an exposed loop containing Glu83 (Figures 3b and 4a). Other peaks at 96–103, 114–116, 129–131, 148–152 agree with our model as they are part of loops (Figures 3b and 4a). There is a broad peak at 169–175 agreeing with a long exposed loop (Figures 3b and 4a). Region 138–145 is hydrophobic (Figure 3a) and has low surface probability by Emin’s method (Figure 3b), but this region is exposed in our model. This is justified due to the hypervariability noticed in 135–144 (Figure 1) by analogy with PilE.

Antigenicity index plot

Fig. 3. Hydrophilicity, surface probability and antigenicity index predicted from the primary sequence of TcpA of the classical O1 are plotted in (a–c), respectively, as a function of the residue number. The hydrophilicity plot for PilE is included in (a) for comparison (dotted lines). This shows immediately that on the whole, TcpA is more hydrophobic compared to PilE. The secondary structures of both structures are also plotted between (a) and (b).

\[ \beta_2 - \beta_3, \alpha_4 - \alpha_5, \beta_7 - \beta_8 \text{ and } \beta_9 - \beta_9, \] respectively, and are expected to be antigenic in the fibre. This agrees with published data (Sun et al., 1991, 1997).

Results and discussion
TcpA structure
Our TcpA molecule has an elongated (85×42×26 Å), ladle-shaped pilin structure, similar to PilE (85×34×26 Å), the size of ladle head or the \( \alpha \)-\( \beta \) roll globular domain being larger in TcpA (Figure 4a). The TcpA model is composed of: (1) a long \( \alpha \)-helix \( \alpha_1 \) (residues 2–54; TcpA secondary structural elements denoted by a subscript number, whereas those of PilE do not have their numbers as subscripts) exactly as in PilE; (2) a second helical region \( \alpha_2 \) (residues 59–68); (3) two \( \beta \)-hairpins forming a four stranded antiparallel \( \beta \)-sheet (residues 76–95, 104–121), all with nearest neighbour (+1) \( \beta \)-strand connections; (4) the \( \beta_2 - \beta_3 \) loop connection including a single turn \( \alpha_3 \) (residues 96–103); (5) Cys120 emerges out of \( \beta_4 \) but points in the radially outward direction of the fibre, coinciding with Phe120 of PilE; (6) strands \( \beta_5 \) (residues 123–127) and \( \beta_6 \) (residues 130–136) are also disposed like in PilE; (7) two \( \alpha \)-helices \( \alpha_4 \) (residues 138–145); and \( \alpha_5 \) (residues 152–158), separated by a loop (residues 146–151); (8) \( \beta_7 \) (residues 163–168) forming two hydrogen bonds with \( \beta_4 \); (9) a long loop region (residues 169–174) travelling radially outwards in the fibre; (10) a short \( \beta \)-strand \( \beta_8 \) (residues 175–179) forming hydrogen bonds with \( \beta_6 \); (11) a loop region (residues 180–191) including Cys186 which points radially inwards in the fibre towards Cys120 and is bridged to it and (12) a final \( \beta_9 \) (residues 192–196) sandwiched between \( \beta_4 \) and \( \beta_7 \).

Several hydrophobic cores stabilize the TcpA structure
Within the ladle head, a hydrophobic core exists between the latter part of the long helix \( \alpha_1 \) (residues 30–54) and the four-stranded antiparallel sheet. Residues contributing side chains in any of four hydrophobic cores are marked by the symbol ‘c’ in Figure 1. Though there is some variation in this region in \( \alpha_1 \) or the four-stranded sheet, viable mutations always preserve the nature and size of the amino acid (Figure 1).

A second hydrophobic core exists between the sheets (\( \beta_3, \beta_3 \)) and (\( \beta_3, \beta_3 \)). A third hydrophobic core is formed between \( \beta_4 \) and \( \beta_6 \), the loop containing Cys186. A fourth hydrophobic core is formed between \( \beta_6 \), \( \beta_4 \), \( \beta_7 \) and \( \beta_9 \). Side chains participating in this core are also similarly conserved (Figure 1).

Consistency of structural database with the disulfide
In PilE, Cys121 emanates from a \( \beta \)-strand and Cys151 from a loop antiparallel to it (Parge et al., 1995). This property has
been maintained in our TcpA model. Well over half of the backbone strands are in extended conformation at the ends of disulfides (Richardson, 1981). Though a small fraction of backbone conformations at the end of disulfides are part of \( \beta \)-sheet, it is extremely rare to have a disulfide bridging two \( \beta \)-strands. Since Cys120 in TcpA is part of \( \beta_4 \) by analogy with
the PilE structure, this demands that Cys186 be part of a loop as we have in our model. Moreover, the Pβ values near residue 186 were low in the Chou–Fasman prediction.

**Protective epitopes in TcpA agree with model**

From a collection of monoclonal antibodies that recognize the native structure of TcpA, Mab16.1 and Mab169.1 were used to localize the corresponding epitopes (Sun et al., 1991). Mab16.1 recognized peptide 157–182 (rather, 168–182, a loop + β8, since 157–167 is buried in the fibre), but not peptides 144–168 or 174–199. Mab16.1 recognized 174–199 (β8 + a loop including Cys186) but not 157–182 or 144–168. Mab169.1 recognizes the folded structure of TcpA as it does not interact with the reduced form (Sun et al., 1991). It was found to be highly protective when used for immunization in the infant mouse model and Mab16.1 was ~50% protective. In the pilin monomer as well as in the fibre generated from our model, both regions recognized by these two monoclonal antibodies are on the surface.

Six peptides corresponding to regions 1–26, 24–43, 77–90, 145–168, 157–182 and 174–199 were used to immunize rabbits (Sun et al., 1997). The last four peptides conferred significant protection against V. cholerae. In our model, all these four regions are either partly or fully exposed in the fibre, hence antigenic. The last two peptides are recognized as the epitopes of Mab16.1 and Mab16.1, respectively. Region 77–90 includes β1 (one side of which is exposed) and the β1–β2 loop containing Glu83. Region 145–168 includes the loop 146–151, which in our model is on the fibre surface.

**Agreement of model with V8 protease sites**

V8 protease cuts TcpA after E83 only, though there are potential proteolytic sites at E151 and E158; these two sites are not accessible to the enzyme (Sun et al., 1997). E158 interacts with αt and is partly buried in the subunit, hence inaccessible. E151 is partly buried by 190–192 in the subunit and also forms a hydrogen bond with the adjacent N152 side chain. Only E83 is in an exposed loop and therefore V8 cuts here. Thus, the model explains why only one of the three potential sites can be cleaved by V8 protease.

**Consistency with mutagenesis data**

Our model is consistent with the site-directed mutagenesis data of Sun et al. (Sun et al., 1997) and Kirn et al. (Kirn et al., 1997, 2000) as explained below. These studies describe the morphology, AA, colonization and CTX phage sensitivity in TcpA due to 18 different single amino acid substitutions.

**Abnormal fold mutants.** C120S or C186S mutant does not form pilin since TcpA does not fold properly due to lack of any of its two cysteines, as a disulphide bridge cannot be formed.

**No fibre mutant.** K165A mutant produced no fibre at all and is therefore remarkable. Therefore, K165 is an interior residue in the fibre. In our model, the K165 side chain interacts with main chain carbonyls of αδ of a neighbouring molecule, thus aiding fibre formation. Alanine at 165 cannot retain that favourable interaction stabilizing the fibre.

**AA defective mutants.** D129A, D146A, D149A, H181A and E183A mutants are AA defective. All five side chains are exposed in our fibre model, since AA has to do with interfilament interaction.

**Normal AA mutants.** K137A, E158A, K172A, K184A and K187A are normal regarding AA. The surface charge of K137 is somehow not essential for AA since El Tor has Q137, thus lacking a positively charged residue here (Figure 1). E158A is also a natural mutation with normal AA; it is buried in the fibre. K172 is on a surface loop but the K172 Nζ atom forms a hydrogen bond with D113 O in our model, thus really not a side chain affecting AA. K184A is normal in AA since, the K184 side chain is buried in the subunit, with a hydrogen bond between K184 Nζ and T125 Oγ atoms in our model. The K187 side chain is in a pocket near the disulphide bridge and is partly buried in the subunit. So changing it does not alter surface property remarkably.

**Improved AA mutants.** P169A and D175A mutations improved AA, hence they must either be on the fibre surface or affect the surface residues somehow. P169 is in the interior of the fibre but the loop following this residue can be affected by this mutation, whereas D175A is on the surface of the fibre.

**Mutants affecting fibre formation through αt.** The K121A mutant forms long bundles—fibre formation is obviously affected by this mutation, suggesting K121 is in the interior of the fibre. In our model, the K121 side chain points radially inwards and interacts with N41 O61 in αt. A smaller side chain at 121 should lead to a more compact ladle head in TcpA, affecting fibre formation. Though E158A is a normal mutation found in V. cholerae strains normal in AA, E158L gives kinked, unraveled pilus further confirming E158 must be in the interior of the fibre. The E158sup-1 mutant represents a spontaneous revertant (Kirn et al., 1997) whose location was within αt. Examining αt, we concluded that only R26 can provide a positively charged side chain making a salt bridge with an E158 side chain. In our model, E158 and R26 side chains interact. The E158L mutation displaces α slightly modifying fibre formation.

**Comparison with P. aeruginosa pilin**

After the TcpA model was built by us using the PilE structure, a crystal structure of P. aeruginosa type IVA (PAK) pilin with a partially truncated N-terminal helix (residues 29–144) was reported (Hayes et al., 2000). The presence of φβ-roller structure in PAK pilin is consistent with our prediction that this class of pilins should share this fold. Though the PAK pilin structure is smaller than PilE due to the truncation and smaller number of residues, it affirmed some of our predictions regarding TcpA: (i) there is variability among the pilins in the secondary structural elements immediately following the long N-terminal helix before the start of αt (Figure 1); (ii) the structure of the β2–β3 loop is more or less conserved among type IV pilins (Figure 4a) [figure 2 of Hayes et al. (Hayes et al., 2000)]; (iii) the first Cys residue emanates from β1 and the spatial position of the disulfide is conserved (Figure 4a); (iv) there is variation among pilins in the number of secondary structure elements near the C-terminus, including the second Cys (Figure 1); and (v) the size of the ladle head or globular domain increases with the total number of residues in the pilin (Figure 4a) [figure 2 of Hayes et al. (Hayes et al., 2000)].
n + 1 to the conserved negatively charged side chain from Glu5 of subunit n. For this interaction to be preserved in TcpA, we further argue that the pitch of the TcpA fibre must be equal to that of PilE and the extra residues in TcpA must be accommodated by increasing the diameter of the filament originating from a larger globular domain. Assuming TcpA fibre has equal density as that of PilE, the average diameter for the TcpA filament calculates to be 81 Å including side chain atoms (for 197 residues as opposed to 158), matching the actual average diameter in the fibre generated from our model. It is within the normal range (Giron et al., 1997) for these pilins.

The association of several subunits is shown in Figure 4b. The regions including α2 and the following loop (65–74) and β3 are involved in lateral contacts between adjacent monomers (subunits n and n + 1) in the fibre, thus explaining the result obtained for the K165A mutation (Kirm et al., 1997, 2000). The K165 N¢ atom hydrogen bonds to G72 O of an adjacent subunit in our model. In addition, several hydrophobic side-chain–side-chain interactions are observed in the inter-subunit interface: P169 with L69', P191 with G72', F192 with F96' and V194 with V74'. Positions 69, 72, 74, 165, 191, 192 and 194 are totally conserved in the four sequences (Figure 1). Position 96 allows either F or W, similar side chains, and V194 with V74

The regions including α2 and the following loop (65–74) and β3 are involved in lateral contacts between adjacent monomers (subunits n and n + 1) in the fibre, thus explaining the result obtained for the K165A mutation (Kirm et al., 1997, 2000). The K165 N¢ atom hydrogen bonds to G72 O of an adjacent subunit in our model. In addition, several hydrophobic side-chain–side-chain interactions are observed in the inter-subunit interface: P169 with L69', P191 with G72', F192 with F96' and V194 with V74'. Positions 69, 72, 74, 165, 191, 192 and 194 are totally conserved in the four sequences (Figure 1). Position 96 allows either F or W, similar side chains, and V194 with V74

Variation of molecular surfaces in various strains

Most of the hypervariable regions in TcpA are exposed in the fibre: 62–68, 92–96, 135–144, 170–184; only 152–158 is not fully exposed. Residues from the hypervariable regions have side chains on the fibre surface. The antigenic surfaces vary considerably within the TcpA strains, thus escaping immune detection. Further, TcpA is known to act as a receptor of CTXφ, a filamentous phage that is known to infect V.cholerae (Waldor and Mekalanos, 1996). A recent study (Boyd et al., 2000) has shown diversity in CTX core region sequences, particularly in OrfU, a phage coat protein likely to mediate its attachment to the host bacterium. Thus, the variability in TcpA sequences is expected to play a crucial role in the specific recognition of the phage by the bacterium through TcpA–OrfU interaction. However, the limited nature (up to 22% level) of variation noted amongst TcpA sequences of various V.cholerae strains would suggest that further changes in these sequences may not be advantageous to the organism which needs to preserve the structure of TcpA and its fibre forming contacts for the maintenance of its function.

Predictions from model

It is evident from the model that out of 72 core amino acids (marked by ‘e’ in Figure 1), comprising primarily of hydrophobic residues, 48 are invariant. Further, alterations in the remaining 24 positions are, more or less, restricted within the category of ‘homologous’ changes. Therefore, residues involved in the formation of the core region appear to be quite critical for the monomeric structure of the protein which is likely to be disrupted by certain mutations of these residues. Although similar changes within the surface exposed residues 1–31, or those involved in the two types of subunit–subunit interactions (e.g. n with n + 1 and n with n + 5) may not disrupt the monomer structure itself, they might affect the pilus formation capacity of the monomers. Side chains exposed in the pilus fibre (marked by ‘e’ in Figure 1) are primarily located in the hypervariable regions mentioned and are likely to account for the antigenic variations among the strains (Jonson et al., 1991; Rhine and Taylor, 1994; Novais et al., 1999; Nandi et al., 2000). It is also predictable that mutations in these positions would affect the following properties of the organism that are shown to be TCP mediated, e.g. the ability to autoagglutinate through bacterium–bacterium interaction, the ability to interact with CTXφ and/or the ability to colonize through bacterium–host cell receptor (putative) interaction.

The model has been constructed basing it on PilE as far as correspondence between secondary structural elements was observed. The remaining portion of TcpA is constrained to assume our proposed structure due to the existence of several biochemical data in the literature. Further, the existence of the disulfide bridge constrains the possible structures.

Acknowledgements

The authors thank the Department of Biotechnology, Government of India for supporting a molecular modelling facility at our Institute. The work was partly supported by the Council of Scientific and Industrial Research, Government of India.

References


R.Chattopadhyaya and A.C.Ghose


Received June 12, 2001; revised December 14, 2001; accepted January 4, 2002