The structural basis of phage display elucidated by the crystal structure of the N-terminal domains of g3p

Jacek Lubkowski[1], Frank Hennecke[2], Andreas Plückthun[2] and Alexander Wlodawer[1]

The structure of the two N-terminal domains of the gene 3 protein of filamentous phages (residues 1–217) has been solved by multiwavelength anomalous diffraction and refined at 1.46 A resolution. Each domain consists of either five or eight β-strands and a single α-helix. Despite missing sequence homology, their cores superimposed with a root-mean-square deviation of 2 Å. The domains are engaged in extensive interactions, resulting in a horseshoe shape with aliphatic amino acids and threonines lining the inside, delineating the likely binding site for the F-pilus. The glycine-rich linker connecting the domains is invisible in the otherwise highly ordered structure and may confer flexibility between the domains required during the infection process.

Filamentous bacteriophages have become, during recent years, an important tool for the identification of interacting biomolecules as well as their evolutionary improvement[1–3]. By genetic fusion to the phage minor coat protein from gene 3 (g3p), peptides and proteins become displayed on the surface of phages and may be selected from large libraries on the basis of their binding properties, whereas the genetic information is packaged in the phage genome. Selection is accomplished either by binding of phages to immobilized target molecules ('phage display'; for review see ref. 4), or by linking infectivity of phage particles to the binding of the displayed protein to a cognate ligand ('selectively infective phages (SIP)'; for review see ref. 5). The g3p protein plays a central role in the development of these techniques. It is located, most likely in five copies, at one tip of the filamentous phage particle and is involved in the infection of *Escherichia coli* cells carrying F-pili, which are encoded on the F-episome. The g3p protein of the M13, f1, or fd phage has a modular structure consisting of three domains of 67 (N1), 131 (N2), and 150 (CT) amino acids, connected by glycine-rich linkers of 19 (G1) and 39 (G2) amino acids (Fig. 1).

According to the current model, infection of the host cell is initiated by binding of the N2 domain to the tip of an F-pilus[6], which is retracted upon phage binding[7] by an as yet unknown mechanism, thereby guiding the bound phage to the bacterial envelope. At this stage, interaction of the N1 domain with the TolA protein[8–10], anchored in the cytoplasmic membrane but spanning the whole cytoplasm, triggers a completely unknown process that finally leads to entry of the phage DNA into the bacterial cytoplasm. The CT domain plays a role in phage morphogenesis and caps one end of the phage particle[11,12]. No special function has been ascribed to the glycine-rich linkers G1 and G2 besides that of passive connectors of the three domains. It has been observed, however, that they enhance the infectivity of the phage[13], probably by conferring flexibility in connecting the domains and adjusting the required distance between them.

A better knowledge of g3p structure and function in the infection process may be highly advantageous not only for a better understanding of phage biology itself, but also for improvement of the biotechnological applications of directed molecular evolution, such as the SIP technology[5]. Here we present the crystal structure of the two-domain fragment of g3p (residues 1–217, referred to as N1-N2), solved by multiwavelength anomalous diffraction (MAD) and refined at 1.46 Å resolution. This high-quality structure shows the unexpected structural similarities of both domains and delineates the mode of their interactions.

The structure of N1-N2

We have prepared the N-terminal part of g3p comprising the N1 and N2 domains, N1-N2, that are linked by their natural glycine-rich linker. This was accomplished by refolding the protein with an engineered C-terminal histidine-tag from E. coli inclusion bodies, using a redox shuffle to form the disulfide bonds, and purification by immobilized metal ion affinity chromatography (IMAC), anion exchange chromatography, and gel filtration. For determining the phases by X-ray crystallography, a selenomethionine-labeled sample was prepared in the same way.

The structure of N1-N2 was solved by using selenomethionine-labeled protein and MAD, and was subsequently refined with 1.46 Å data collected on a crystal of the native protein. The initial maps were of good quality and allowed us to trace the chain without major difficulty. Some of the more unusual aspects of the structure, such as the presence of a *cis*-proline near the C-terminal end of the protein and oxidation of one of the tryptophans, could be seen in the initial maps (Fig. 2*a,c*). These features became even clearer after refinement (Fig. 2*b,d*), attesting to the high quality of the final structure.

N1-N2 consists predominantly of β-structure (Fig. 3), also carrying one short α-helix in each domain. In the smaller N1 domain, helix α1 is located close to the N terminus and is followed by five β-strands arranged in a barrel-like motif. These β-strands, together with two additional β-strands from the N2 domain, form a seven-stranded antiparallel sheet with topolo-

[1]Macromolecular Structure Laboratory, ABL-Basic Research Program, NCI-FCRDC, P.O. Box B, Frederick, Maryland 21702, USA. [2]Biochemisches Institut, Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland.

Correspondence should be addressed to A.W. *email: wlodawer@ncifcrf.gov*

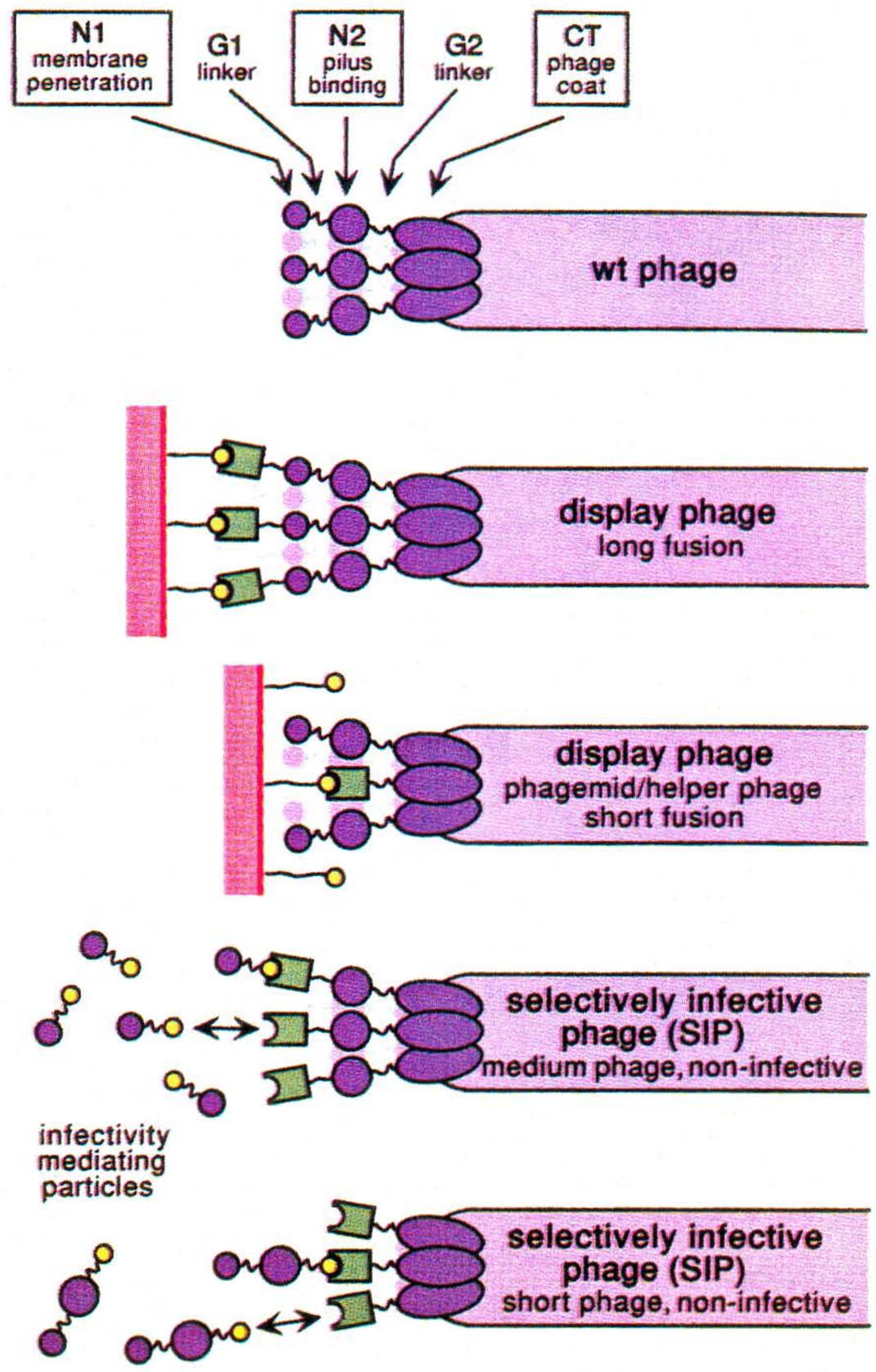

Fig. 1 Application of filamentous bacteriophages for identification of protein–ligand pairs. In wild type phages, probably five copies of g3p are located at the tip of the phage, mediating infection of the host cell. The domains and glycine-rich linkers are indicated, and labeled as used in the text. In display phages, proteins or peptides (green) are genetically fused either to the N terminus of g3p (long fusion) or to its CT domain (short fusion). In the latter case, wt g3p has to be provided by a helper phage to retain infectivity of the display phage. Identification of displayed proteins that bind to a particular ligand is accomplished by incubation of phage libraries with immobilized ligand (yellow) and subsequent elution of bound clones[4]. Selectively infective phages are noninfective due to the replacement of the N1 domain or both N-terminal domains by the displayed protein or peptide. Infectivity is restored by binding of a cognate ligand that, in turn, is covalently coupled to the missing g3p domains[5].

gy [4x,-1,-1,-1,2x,1]. The strands in the N1 domain are of different length. Strands β2 and β3 are quite short, whereas strands β4 and β5 are the longest in the whole structure and their topology departs visibly from an ideal one. Strands β3, β4 and β5 form two hairpins (hβ3/4 and hβ4/5). Two bulges[14] (residues Gly 55, Gly 42, Val 43, and Leu 37, Ile 60 and Gly 61 respectively) are present within hβ4/5, disrupting the network of hydrogen bonds typical for the β-sheet. These bulges are likely the result of several topological constraints observed within the N1 domain, introduced by the two disulfide bridges and by the interactions of β4 and β5 with other topological motifs. One disulfide bridge is formed by Cys 7 (α1) and Cys 36 (β4) and shows the conformation of a left-handed helix. Several exclusively hydrophobic interactions between α1 and β5 are also present in this region. The second disulfide bridge has the conformation of a short right-handed hook and is formed by Cys 46 and Cys 53, located close to the tip of hβ4/5. hβ4/5 integrates both the N1 and N2 domains, since β4 interacts with β6 (the first strand of the N2 domain) through several hydrogen bonds, besides interacting with β3 and β5.

The larger N2 domain consists of eight β-strands, with six of them (β7–β12) arranged within a mixed β-sheet with topology [4x,-1,-1,-1,4]. The last strand β13, interacts with the first one β6, and participates in seven-stranded sheet formed mainly by the N1 domain. These two strands extend away from the core of the N2 domain, like a finger extending from the palm. The 22-residue fragment between β6 and β7 carries several prolines and does not contain any regular α/β motifs. It is, however, very well defined and is likely responsible for stabilizing conformations of the rest of the N2 domain, mainly through hydrophobic interactions. Similarly to what was seen in the N1 domain, two strands in the N2 domain β10 and β11, are visibly longer than the rest. These and two relatively short strands, β8 and β9, form three hairpins (hβ8/9, hβ9/10, and hβ10/11). The latter hairpin stands apart from the core of N2, remaining topologically a finger. Within this hairpin, the hydrogen bond network is disrupted by one bulge (Gln 167, Thr 152, and Gly 153). In contrast, however, to the β6/β13 motif, the observed conformation of the hairpin hβ10/11 is stabilized by very few interactions that could be ascribed to the crystal packing. The absence of interactions between hβ10/11 and the rest of the molecule results in slightly elevated flexibility of this motif compared with the core of the N2 domain. One of the two *cis*-proline residues, Pro 161, is located at the tip of this hairpin. Strand β11 is followed by the only α-helix (α2) in the N2 domain. This helix interacts with the rest of the N2 domain, mostly through hydrophobic contacts. Several residues located past α2 create a series of turns, with the last one, containing Cys 188, being involved in a disulfide bridge. This connection to Cys 201 has the conformation of a right-handed hook and is the only one in the N2 domain. The loop between Cys 188 and the short strand β12 is stabilized by the above-mentioned disulfide bridge and several other interactions between the side chains, including an infrequently observed π-interaction between His 191 and Phe 199. The N2 domain terminates with a very well defined heptapeptide that contains three proline residues. One of these prolines, Pro 213, is also found in a *cis* conformation.

The N-terminal residue (Glu 91) of the N2 domain is ~19 Å apart from the C-terminal residue in the N1 domain (Pro 65). Although the current model is missing 25 residues, it is evident that the interdomain linker cannot be in an extended conformation in this structure, but must be disordered.

Comparison of the N1 and N2 domains

We found an unexpected feature when comparing the two domains of N1-N2. Despite their relatively low sequence homology (15% identity after structure-based alignment; Fig. 4*a*), N1 and N2 share a nearly identical fold (Fig. 4*b*). As calculated with the program DALI[15], the root-mean-square deviation (r.m.s.d.) for 43 Cα pairs is 2.0 Å. The equivalent topological elements (Fig. 4*c*), consist of five-stranded antiparallel subsets of two β-sheets (β1, β5(part), β4(part), β3 and β2 in the N1 domain and β7, β11(part), β10(part), β9 and β8 in the N2 domain). Careful analysis of the superposition suggests, however, that topological and possibly structural similarity extends even farther, over the entire hairpins hβ4/5 and hβ10/11, which are of identical length. In the crystal structure, the observed orientation of hβ4/5 is stabilized by a series of

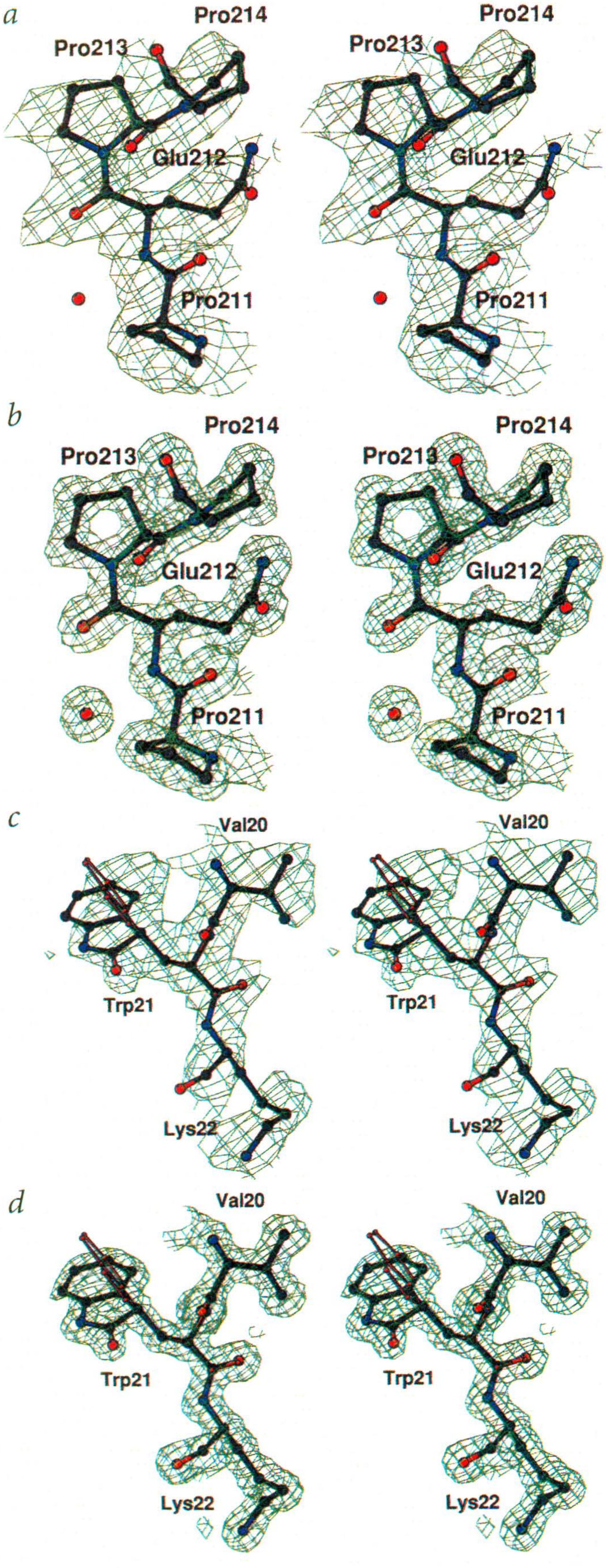

Fig. 2 Initial and final electron density maps for two regions of N1-N2, contoured at the 1.2 σ level. ***a***, Experimental map obtained with MAD phases modified using the program DM[46]. The density around Pro 213 already clearly indicates the *cis* conformation of this residue. ***b***, The same fragment shown the final $2F_o - F_c$ map. ***c,d***, The experimental and final electron density maps around oxidized tryptophan, Tro 21. It is clear even in the experimental map that the hybridization of the CG carbon is not planar, as would be expected for tryptophan, and that a non-hydrogen atom (presumably oxygen[47]) is bound to CD1. An unmodified tryptophan residue is superimposed to visualize the differences. Figure prepared with the program Bobscript, a modification of Molscript[48].

intradomain contacts, whereas no such stabilization is provided for hβ10/11. However, it is likely that both domains share common origins.

Comparison of N1-N2 with related proteins

An attempt to compare the structure of N1-N2 with all known folds found in the Protein Data Bank, using the program DALI[15], did not result in identification of any significant similarities to the two-domain construct. Some similarities have been found, however, for the individual domains. In particular, the N2 domain shows detectable similarity (Z-score 2.1) to the PDZ domain from human discs-large protein (PDB designation 1pdr)[16]. This similarity (r.m.s.d. 3.6 Å for 50 Cα pairs) covers the long strand extending from Ala 119 to Ser 128, parts of strands β8–β11, and helix α2, and thus includes a significant portion of the PDZ domain, which is considerably smaller than N2. However, two β-strands of PDZ that form part of its core do not have counterparts in N2, so the significance of this similarity should not be overstated, although it is interesting to note that PDZ, like N2, is involved in protein–protein interactions.

No similarities with a Z-score of 2.0 or more were automatically detected for the N1 domain by using DALI. The only reported correlation was with the structure of homopexin (PDB designation 1hxn)[17]. The Z-score of 1.1 cannot be considered significant, and the similarity was limited to three β-strands that did not form a domain. However, a comparison with the permuted SH3 domain (PDB designation 1tuc)[18], previously identified as similar to N1[19], showed an r.m.s.d. of 2.8 Å for 41 Cα pairs, although there is no detectable sequence homology between these proteins. As is the case with PDZ, the role of the SH3 domain is to interact with other proteins, so a structural relationship to N1 would be plausible.

Comparison with N1 in solution

The structure of a fragment comprising residues 2–67 of the N1 domain was previously reported based on NMR data[19]. This structure (PDB designation 1fgp) is an ensemble of 15 models, with no average structure. When we calculated such a structure with the program MOLMOL[20], the structure could be superimposed on the N domain with the an r.m.s.d. of 2.2 Å for 63 Cα pairs. A similar deviation was present for model 5 in the ensemble, which was the closest to the average structure that we computed. These differences are somewhat larger than those reported for other highly refined X-ray and NMR structures of proteins (for example, IL-8, ref. 21; and MCP-1, ref. 22), and are mostly influenced by the differences in conformation of five turns, as well as of the C ter-

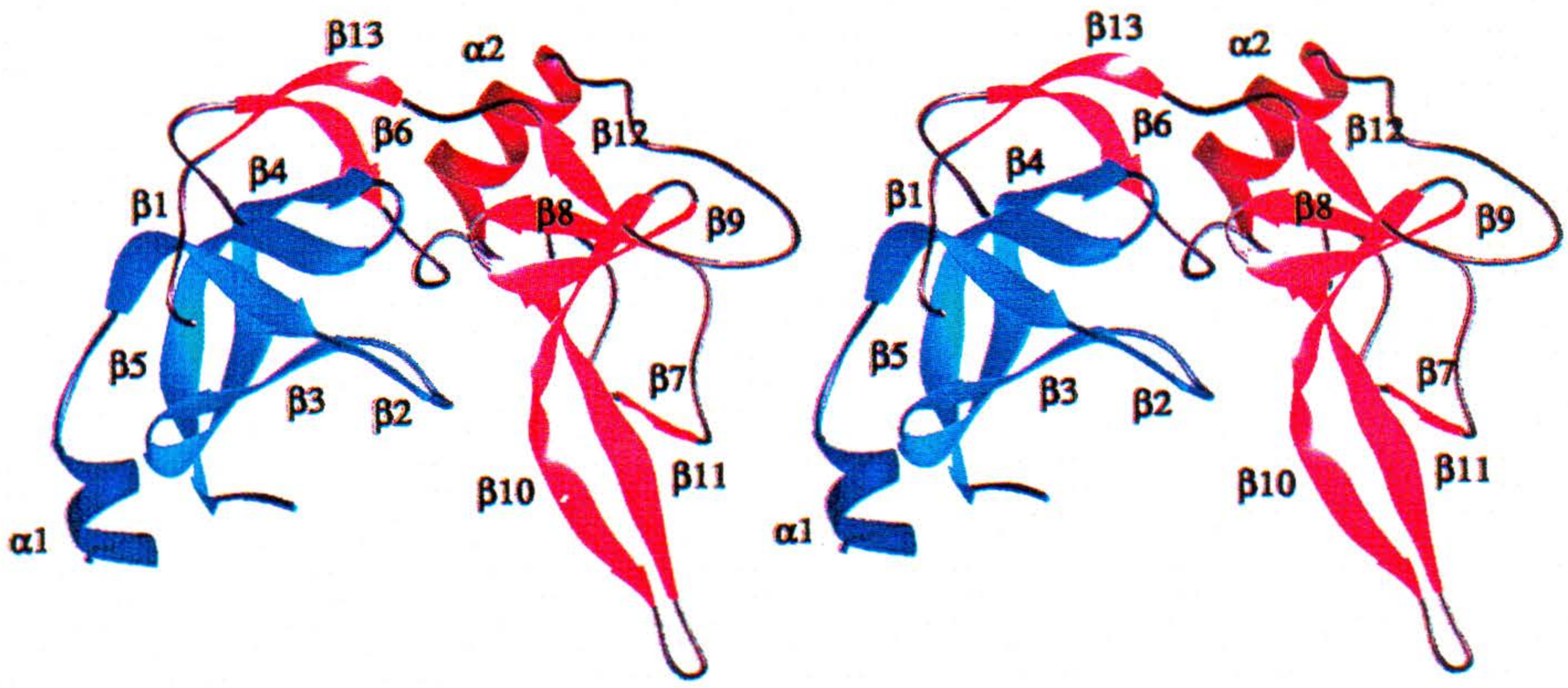

Fig. 3 Ribbon representation of the N1 domain (blue) and N2 domain (red) within one molecule of N1-N2. The proper pair of domains forming a single molecule was deduced from an analysis of the crystal contacts, as well as from a comparison of the domain interactions with those described previously using NMR techniques[10]. Figure prepared with the program Ribbons[49].

minus of N1. Whereas the coordinates were provided for the C-terminal part of the N1 domain in the NMR models (including several engineered residues, not present in the native sequence), this region was not visible at all in the X-ray structure, although it would have been possible to accommodate the NMR models for this region into the crystal packing. It is not clear, however, how meaningful was the identification of the C terminus in the NMR structures, since the coordinates were so widely divergent in different models that any particular discrepancy may not reflect significant differences between the results obtained by these two methods. The differences observed for the turns Lys 25–Thr 26–Leu 27 or Thr 13–Glu 14–Asn 15–Ser 16 may be a result of interactions with the N2 domain. These differences may thus illustrate real conformational changes of the N1 domain resulting from the separation of the N1 and N2 domains. A number of other discrepancies, mostly minor, cannot be explained at the current stage of analysis.

The N1-N2 interface

It has been shown previously[10] that similar surfaces of the N1 domain participate in adhesion either to the N2 domain of g3p or to TolA, an *E. coli* protein utilized by the phage in the infection process. Therefore, the analysis of the contacts present within the N1-N2 interface defines, on the atomic level, the regions of the N1 domain involved in phage infection. In turn, the respective areas of the N2 domain may provide some clues to the surface properties of the N1 epitope in TolA. The similarity of the region of interaction between N1 and N2 to the area that was previously reported strongly supports the assumption that the two domains described here indeed form a single molecule, which could not be delineated unambiguously due to the absence of the interdomain linker in the X-ray model.

We assessed the interdomain contacts in two ways, using different measures. In the first approach, the interface was analyzed visually and all pairs of residues that were within contact distance (at least one atom of the residue in N1 is interacting with an atom of any residue in N2 through a distance shorter than the sum of their van der Waals radii) were assigned as part of the interface. In the second approach, the interface residues were considered to be those for which the solvent-accessible areas calculated for separated domains were different from those in the two-domain N1-N2. The results from both approaches are illustrated in Fig. 4*a*. Both of the approaches that we used delineate almost the same set of residues, consistent with the results of the NMR experiments[19], except that the number of contacts seen in the crystal structure is significantly lower. It is thus most likely that the N1 and N2 domains, as shown in Fig. 3, are part of the same molecule. Additional evidence for this observation comes from the analysis of crystal packing. Only a few charged residues are located within this interface, and they are contributed exclusively by the region of the N1 domain situated between Lys 22 and Arg 29. In particular, only four residues out of 23 (Lys 22, Asp 24, Lys 25, and Asp 28) utilize their polar side chains in the interactions with the N2 domain. The N2 interface, in turn, is quite hydrophobic, and fewer residues (17) participate in contacts with the N1 domain. The charge distribution over the contact surface of each domain is shown in Fig. 5. This figure also illustrates the topological properties of the interface, showing the finger, formed by β6 and β13 in the N2 domain, surrounding the smaller N1 domain. As mentioned above, we could not confirm all of the previously reported interdomain contacts[10]. Depending on the approach, we found that either 15 or 13 residues (out of 36) that were listed as part of the N1 domain interacting with N2 did not make such contacts, and some of them (for instance, Val 58, Trp 21, Thr 41 or Cys 35) were actually located on the side of the N1 domain distant to the N2 domain. In turn, three residues forming close contacts to the N2 domain (Gly 48, Cys 53 and particularly Ser 16) were not identified before.

Crystal packing

As indicated by the moderate solvent contents (~40%), N1-N2 is relatively tightly packed in the crystals. The bulk solvent is found mainly in two topologically distinct regions. One region is shaped as a long, nearly cylindrical, straight channel formed by coaxially aligned barrels from the N1 domains and is complemented by fragments of the N2 domains. The diameter of this channel is ~11 Å when measured between Cα atoms of participating residues. The other solvent-containing region is a large spherical area (~40 Å in diameter). From the position of Pro 65 and Glu 91, the last residues to be seen before the disordered region, it is clear that the interdomain linkers are located in such spherical regions. The analysis of specific contacts between different domains also supports the identification of the N1 and N2 domains belonging to a single molecule, since the interactions between the N1 and N2 domains, other than those shown in Fig. 3, are limited to just a few contacts.

Implications for infection

The intimate association and horseshoe shape of the N1-N2 domains now suggests how the initial phases of the phage infection may proceed. Several lines of evidence point to the N2 domain as making the most important contacts with the pilus. In the absence of N2, only a low background infection is observed, which is independent of the presence of a pilus[11,12,23,24] but requires $C\alpha^{2+}$. Such a phage with only N1 may

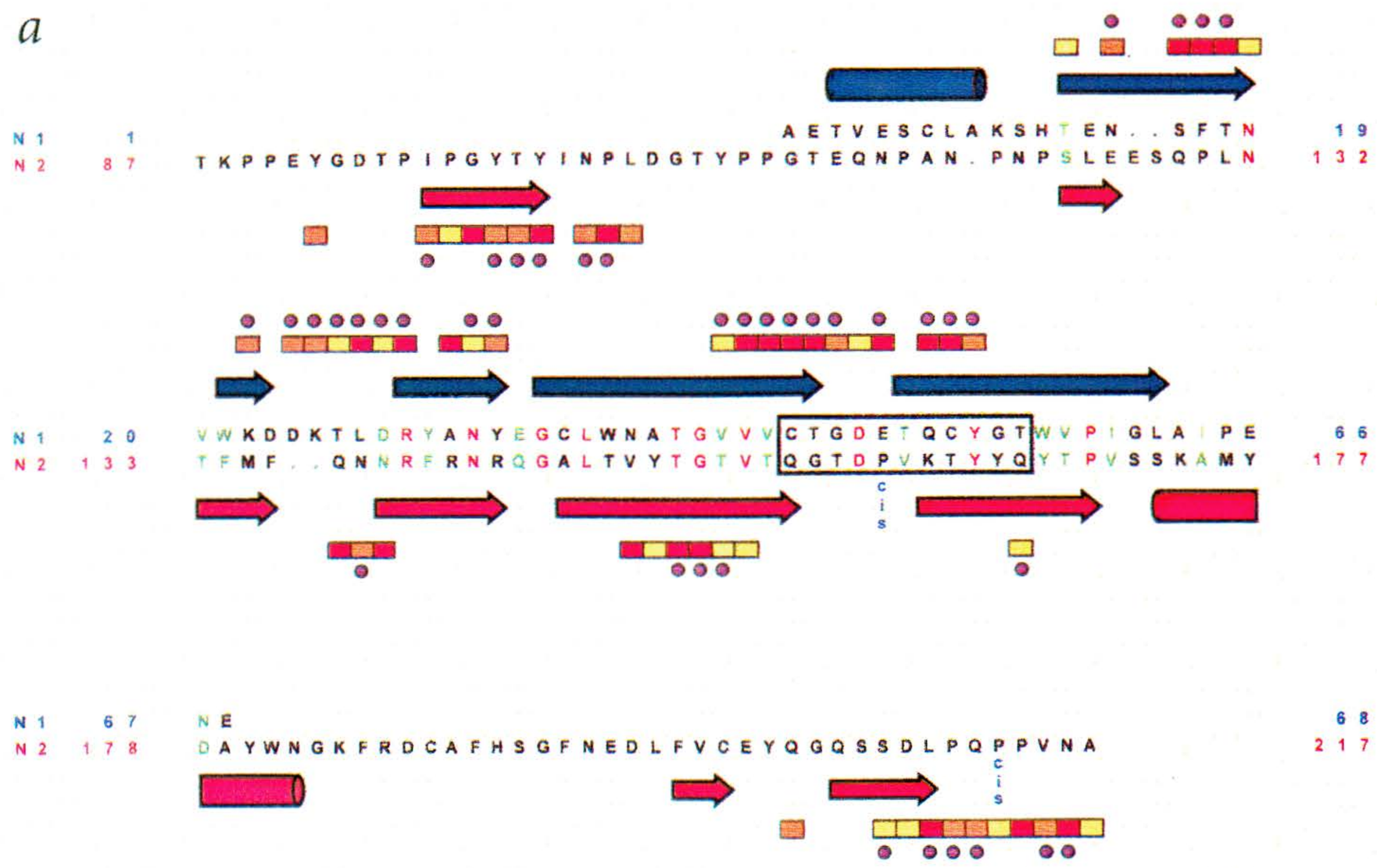

Fig. 4 Structural similarity between domains N1 and N2. ***a***, Sequence alignment of the N1 and N2 domains of g3p. Identical residues are red, whereas highly homologous ones are green. The secondary structure elements are also shown as cylinders (α-helices) and arrows (β-strands) in blue and red for the N1 and N2 domains respectively. The 11-residue fragment of both domains, shown within the box, corresponds to the sections of the hairpins hβ4/5 and hβ10/11. These sections have the same length and a moderate sequence homology, and theform a topologically identical motif. In the crystal structure, however, they have different orientations that are stabilized differently in each domain. The *cis*-proline residues in N2 are also marked. Residues participating in contacts between the N1 and N2 domains, as judged visually, are indicated with purple dots. Results of a similar analysis, base on the differences in solvent accessibility calculatedwith the program NACCESS (http://www.biochem.ucl.ac.uk/~roman/naccess/naccess.html), are shown in the form of horizontal bars. Red boxes indicate residues for which solvent accessibility in the complex is less than 30% of that in an isolated domain; orange, 30–70%; and yellow, more than 70%. ***b***, Structurally superimposed domains N1 (blue) and N2 (red) represented as tubes. The hairpins hβ4/5 and hβ10/11 are highlighted. The fragments that are structurally unique in each domain are represented by thinner tubes. The terminal residues are labeled. ***c***, A topological diagram of the N1-N2 fold with β-strands indicated by arrows and α-helices by cylinders. Elements identified in the N1 domain are shown in blue, whereas the N2 elements are red. Darker shades are used to indicate two structurally equivalent fragments in both domains. The interdomain linker is represented by a dotted line.

have once existed, and the similar fold suggestes that N2 may have arisen by gene duplication. Furthermore, the related IKe phage that is specific for N-type pili, has a g3p with homology to the N1 domain, but not to the N2 domain[25]. When the inhibition of infection of a wt phage with N1, N2, and N1-N2 was tested[24], they were found to inhibit at about 10^{-6} M, 10^{-8} M, and 5×10^{-9} M (midpoints), demonstrating the binding of N2, but also some synergistic binding of N1 and N2.

This is now understandable from the structure of N1-N2. The horseshoe-like shape of the two-domain molecule is accompanied by a large number of exposed aliphatic residues (in the N1 domain) and threonine residues (in the N2 domain), all pointing into the central channel. Thus, it is very likely that this central part of the molecule constitutes a binding site, probably for the pilus.

It is tempting to speculate that after this binding site has interacted with the pilus, and the pilus has retracted the N1 domain then becomes dislocated upon contacting TolA. Such dislocation would make N1 independently tethered. It is known that TolA, TolQ, and TolR are all needed for the infec-

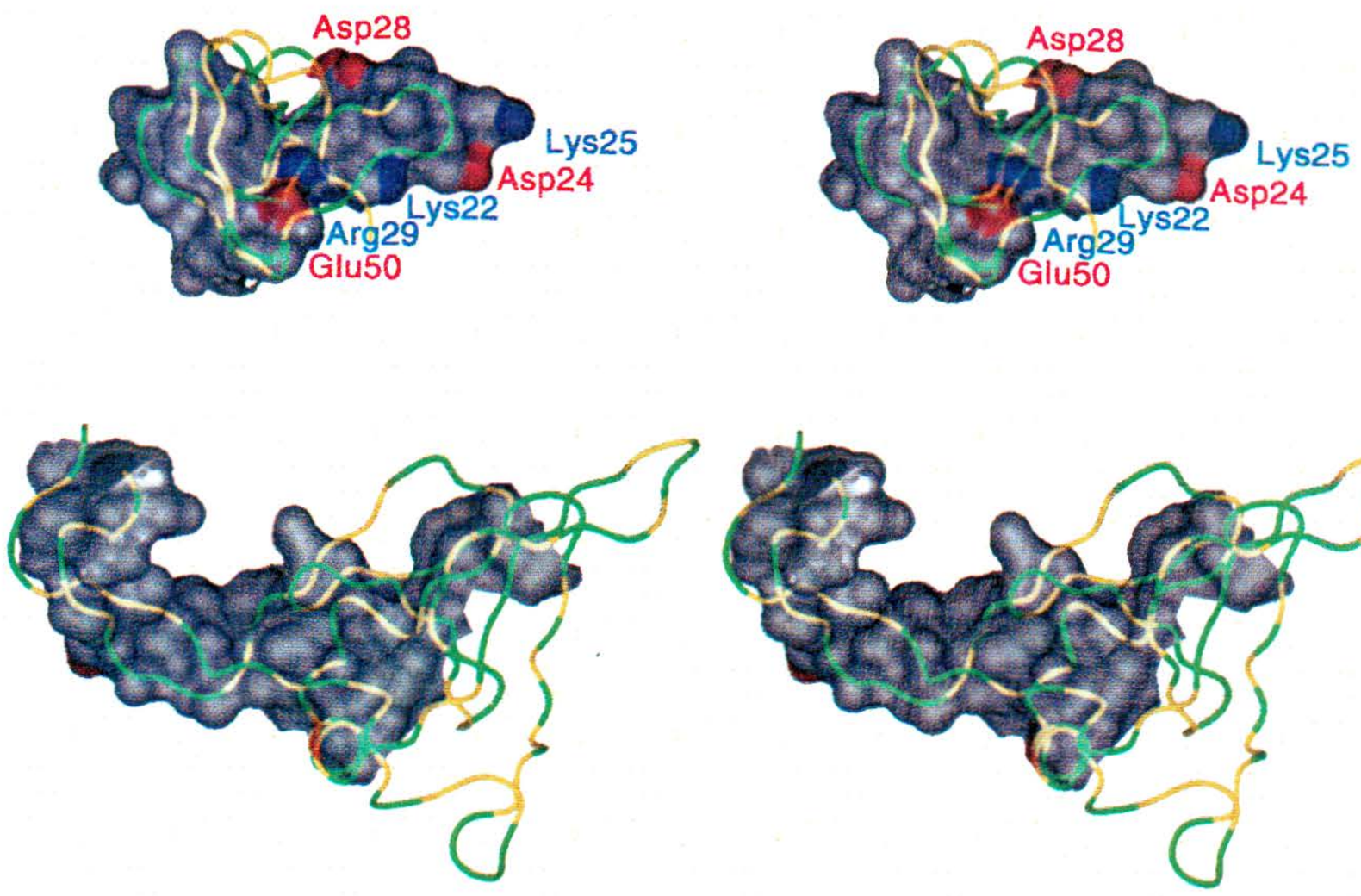

Fig. 5 The interdomain contact surfaces are shown in the semi-transparent form and related to each domain as represented by the two-colored tube. The N1 domain is shown at the top. The yellow sections of the tubes denote hydrophobic residues. The surface areas in blue correspond to positively charged residues, and in red, to negatively charged residues. In this view, the domains are separated and rotated in opposite directions to expose the contact surfaces.

tion process[8]. These proteins are assumed to be involved in the uptake of macromolecules as well as in maintaining membrane integrity[9]. Recently, the initial contact within *E. coli* was suggested to be provided by the C-terminal domain of TolA[10], which is contacted by N1.

The F-pilus is made of pilin, a very hydrophobic protein consisting of 70 amino acids, that carries only five positive and three negative charges[26]. Its structure is as yet unknown, but fiber diffraction studies have suggested that the pilus has pentameric symmetry[27]. Similarly, the cylindrical surface of the filamentous phage is covered by the gene8 protein which has also been found to be arranged in fivefold symmetry[28]. Although five copies of g3p have never been seen directly in electron microscopy (EM), they have been proposed from biochemical studies[29,30]. A five-fold symmetry would be most easily consistent with the symmetry of the phage and the pilus, which would explain the fact that binding occurs only at the tip of the pilus[7] and that initial docking is so tight as to be irreversible[31]. There is no indication of oligomerization of the N1-N2 complex in the crystal structure presented here, and thus, the pentameric arrangement at the tip of the phage is most likely caused by the C-terminal domain or even the C-terminal hydrophobic anchor and/or its association with gene6 protein. Studies using EM suggest that the N-terminal domains are actually separated from the bulk of the phage[6], and we can now interpret the globule seen in EM[6] as the complex of N1 and N2. However, we cannot rule out a potential interaction of the N1-N2 complex with the CT domain that may have been destroyed during the preparation of the sample for EM.

Implications for phage display and SIP

The N terminus of N1 and the C terminus of N2 (and thus, the rest of the phage) are close together. This positioning has the remarkable implication that a peptide or protein, which is displayed fused to the CT domain (so-called 'short fusions'; Fig. 1) or fused to the N1 domain (so-called 'long fusions'; Fig. 1) is actually presented at a similar region on the phage. Both fusions are far from the inside of the horseshoe, where the interaction with the pilus may occur, which may explain why they do not interfere with the infection process.

In model experiments for developing the SIP technology β-lactamase has been inserted between N1 and N2, yielding highly infective phages[24]. In this enzyme, the N and C termini are close together, thus being ideally compatible with the ends of N1 and N2 that are also close together. In the SIP strategy, the N-terminal domains that are required for infection are decoupled from the rest of the phage to give a so-called 'adapter', and are attached to the phage only by a non-covalent cognate interaction. Therefore, infection is strictly dependent on the cognate interaction, providing a very powerful selection system. The N1-N2 adapter (Fig. 1, bottom) corresponds to the molecule that we have now crystallized, where the antigen would be fused to the C terminus (at position 217) through a flexible linker. However, an adapter consisting only of N1 is also functional, and the 'interruption' of the g3p arrangement by the protein–ligand pair is also possible[24] between N1 and N2 or between N2 and CT (Fig. 1). However, further work on the dynamics of the domains in the infection process will be required to optimize such systems.

Methods

Protein production and purification. Direct sequencing of g3p isolated from phage had shown that it begins with Ala–Glx–Thr[32] and thus its signal sequence is 18 amino acids long[33]. The sequence encoding residues 1–217 of mature g3p from M13mp18 (ref. 33) that was extended with an N-terminal methionine residue (that would become cleaved) and with the sequence Pro–Ser–Gly–His_6 at the C terminus was cloned into the vector pTFT74[34] under control of the T7 promoter and expressed in BL21(DE3)[35]. For producing the selenomethionine-labeled protein, the construct was expressed in DL41(DE3)[36], using a synthetic growth medium containing seleno-L-methionine (Se-Met)[36]. The recombinant protein was purified from cytoplasmic inclusion bodies in the presence of 8 M urea, using the coupled IMAC-AIEX protocol[37]. It was refolded by dialysis against 0.2 M Tris-HCl (pH 8.5), 0.4 M arginine, 0.2 M guanidinium hydrochloride, 0.1 M $(NH_4)_2SO_4$, 2 mM EDTA, using a redox shuffle containing 1 mM oxidized glutathione and 0.2 mM reduced glutathione[34]. After the buffer was changed to 50 mM Tris-HCl (pH 7.5), the protein was subjected to AIEX chromatography on a perfusion chromatography HQ column with NaCl gradient (0–210 mM), using a BioCAD 60 system (Perseptive Biosystems). The eluate was then subjected to gel filtration on a Sephacryl S-100 26/60 column (Pharmacia) in 50 mM Tris-HCl (pH 7.5), 150 mM NaCl. Removal of the N-terminal methionine residue, resulting in the correct N terminus of mature g3p, was verified by mass spectroscopy. The yield obtained from 8 L of E. coli culture was 150 mg of pure protein.

Crystallization and data collection. The best crystals were obtained from 5 ml of N1-N2 at ~10 mg ml^{-1} with 5 ml of precipitant (0.2 M $(NH_4)_2SO_4$, 30% (w/v) PEG-4000, 2 mM DTT, 50 mM PIPES buffer pH 6.5). The first crystals appeared after ~30 days (2 weeks in the case of Se-Met analogs) and reached their final size (average linear dimensions, 0.1 mm) during the next two to three weeks. The ability to crystallize was not affected by lyophilization of the samples. Before the crystallizations were set up, the protein was dialyzed against 50 mM PIPES buffer (pH 6.5), 2 mM DTT.

Table 1 Statistics of data collection, processing, and structure refinement for N1–N2

Crystal (resolution)	Wavelength (Å)	Number of reflections Total	Unique	Completeness (last shell)[1] A	B	R_{sym} (last shell)[1] A	B	$\langle I/\sigma_I \rangle$
#1, Se-Met	0.97934 (λ1)	153,149	21,976	99.6 (100)	99.8 (100)	0.071 (0.157)	0.069 (0.151)	13.8
(2.15 Å)	0.97927 (λ2)	152,945	21,981	99.6 (100)	99.8 (100)	0.071 (0.158)	0.068 (0.151)	13.8
	0.97915 (λ3)	151,853	21,963	99.4 (99.9)	99.7 (100)	0.072 (0.161)	0.069 (0.154)	13.5
	0.96614 (λ4)	153,230	21,966	99.5 (100)	99.7 (100)	0.074 (0.165)	0.071 (0.158)	13.3
#2, Se-Met	0.98	184,454	43,648	100. (100.)	97.9 (95.5)	0.078 (0.393)	0.074 (0.371)	11.1
(1.70 Å)			24,027[2]					
#3, native	0.98	282,603	37,511[2]	99.8 (99.2)		0.034 (0.110)		27.0
(1.46 Å)								

Refinement information (crystal #3)

Resolution (Å)	8–1.46	Average *B*	16.1
No. reflections($F > 2\sigma(F)$)	36,895	R.m.s.d.[4] bond lengths	0.009
Completeness ($F > 2\sigma(F)$)	97.6%	angles	1.584
R_{work}[3]	0.187	dihedrals	27.019
R_{free}[3] (10% of data)	0.225	impropers	1.265
No. non-hydrogen atoms	1,890		
No. water molecules	312		

Residues in Ramachandran plot (%)[4]

Most favored	87.1
Additional	12.9
Disallowed	0.0

[1] $R_{symm} = \Sigma_i |I_i - \langle I \rangle_i| / \Sigma I_i$, where i extends over all unique reflections with Bijvoet pairs merged (A) or separated (B) during scaling, and $\langle I \rangle_i$ is the mean of I_i observations. The resolution ranges defining the last shell are 2.23–2.15 Å (crystal #1), 1.76–1.70 Å (crystal #2), and 1.51–1.46 Å (crystal #3).
[2] Reflections (hkl) and (-h-k-l) are assumed to be equivalent and are merged.
[3] $R_{work/free} = \Sigma ||F_o| - |F_c|| / \Sigma |F_o|$ where F_o and F_c denote observed ($I^{1/2}$) and calculated structure factors respectively, and summation extends over all observed reflections.
[4] The Ramachandran plot was generated in PROCHECK[50] and the r.m.s. deviations from ideal target values were calculated with X-PLOR[42].

The latter additive was crucial for successful crystallization of the Se-Met variant.

The X-ray data were collected from three crystals on the X9B beamline at NSLS, Brookhaven National Laboratory (Upton, New York, USA) using a MAR345 image plate detector, at a temperature of 100 K. Before flash freezing, all crystals were transferred momentarily to the mother liquor enriched with 20% (v/v) glycerol. Crystal #1 was used for the MAD experiment[38], whereas crystals #2 and #3 were used to collect high-resolution data for the Se-Met variant and the native protein respectively. The crystals are trigonal in space group $P3_221$, with one molecule per asymmetric unit, 40% solvent content, V_m of 2.1 Å^3-dalton^{-1} (ref. 39) and unit cell dimensions $a = 48.61$ Å, $c = 153.77$ Å (crystal #1), $a = 48.56$ Å, $c = 153.20$ Å (crystal #2), and $a = 48.68$ Å, $c = 153.22$ Å (crystal #3) (see Table 1). The experiment was carried out utilizing reverse-beam geometry in data collection. The low-resolution limit for all collected data was 22 Å. Although the current high-resolution limit for the native data set is 1.46 Å, crystals of N1-N2 proved to diffract significantly farther and we plan to extend the resolution of the data. The data were processed and scaled with the HKL suite of programs[40]. Model building was done with the program O[41], which was also used later for manual corrections. The refinement and map calculations were done with X-PLOR[42].

At the time of data collection, the station X9B was not yet optimized for MAD experiments and we were not able to measure absorption spectra on either the crystal used for data collection or other crystalline samples of N1-N2. Ultimately, we obtained such spectra from crystals of an unrelated protein and thus were not able to optimize the selected wavelengths of radiation. To compensate for this shortcoming, we collected inflection point data at two energies 1 eV apart (λ1–2), in addition to data at the putative absorption peak (λ3) and at a remote high-energy point (λ4). The statistics of data collection are listed in Table 1.

Structure solution and refinement. Although the data were of good quality, the presence of either anomalous or dispersive signals could not be ascertained from scaling statistics, since the *R*-factors for anomalous differences, as reported in the program PHASES[43], varied only between 1.7% and 2.0% among the four data sets, whereas the *R*-factors between the remote data set and the other three were in the range of 1.3–1.6%. However, the locations of both selenium atoms expected in the N2 domain could be readily ascertained both by automatic Patterson superposition and by direct methods (both implemented in the program SHELX97; ref. 44). Anomalous signal was present for wavelengths 2–4, whereas the best dispersive signal was found, as expected, for a combination of remote (λ4) and edge (λ1) data. These data, together with the anomalous data for the absorption peak, were used in traditional single isomorphous/single anomalous refinement (SIRAS) in PHASES[43]. The resulting Cullis *R*-factors were 0.58 and 0.64 for 3.5 Å and 2.5 Å data respectively. Solvent flattening using the procedure of Wang[45] at 3.5 Å resulted in a map inversion *R*-factor of 0.212, a correlation coefficient of 0.890, and a mean figure of merit of 0.850 in space group $P3_121$. The corresponding indicators in space group $P3_221$ were 0.187, 0.923, and 0.857 respectively, establishing the latter to be correct. Further density modification and histogram matching, performed with the program DM (part of the CCP4 suite[46]), resulted in readily interpretable 2.5 Å maps (Fig. 2*a*,*c*).

A skeletonization procedure of the 3.5 Å SIRAS map, as implemented in the program PHASES[43], was used for the initial tracing of the chain, while detailed model building was based on a 2.5 Å density-modified map. The skeleton allowed us to recognize the structural elements of the N1 domain and adequately superimpose its NMR model. Additionally, the SIRAS map was very clear in the areas around the two Se sites, from which we began generation of the N2 domain. In all questionable areas, the original sequence of N1-N2 was substituted by a polyalanine model. The correctness of the partial model was confirmed by identification of the third disulfide bridge that connected independently generated parts of the model. Out of 190 residues built into the SIR map, the conformation of only a few had to be corrected during refinement. The refinement began at 2.5 Å resolution, and the high-resolution limit was reached in several cycles interspersed with visual map inspection. Initially, each refinement cycle employed simulated annealing, positional refinement, and overall (later substituted by a group) *B*-factor refinement. Simulated annealing was discontinued after the resolution of data extended beyond 2.0 Å, and most of the model was well refined. At this stage, isotropic *B*-factors for individual atoms were refined and water molecules were included. During the entire run of refinement, Trp 21 was substituted by an alanine, since we found it impossible to fit a standard Trp to the side chain density. The detailed conformations of residues 157–161 were also determined from the electron density omit maps, and Pro 161 was found to be present in cis conformation. The final $2F_o - F_c$ and omit maps for the side chain of Trp 21 identified the probable chemical modification of this residue as oxidation. A model of oxidized tryptophan (containing an oxindole group, 1,3-dihydro-indole-2-one) fits the

electron density very well. Although such a modification has not been described previously in a protein crystal structure, it is a common chemical modification of tryptophan[47]. The source of this modification is not known at this time.

The interdomain linker residues (66–90) and three C-terminal residues, as well as the six-His linker, are excluded from the current model because the corresponding electron density peaks could not be identified. The final model consists of 192 amino acid residues including one non-standard residue (oxidized tryptophan, Tro 21, made up of 15 non-hydrogen atoms) and two proline residues in *cis* conformation, one sulfate anion, and 312 water molecules. Two water molecules are located on a two-fold axis, with an occupancy of 0.5. Seven residues, comprising 57 non-hydrogen atoms, have been modeled in two alternate orientations with occupancies of 0.5. The average *B*-factor for non-hydrogen protein atoms is 12.8 Å^2; for water molecules, 31.3 Å^2.

Coordinates. The coordinates and structure factors have been deposited in the Protein Data Bank (accession numbers 1g3p and r1g3psf respectively).

Acknowledgments

We are indebted to F. Dyda and A. Zdanov, as well as to M. Sullivan, for their efforts to adapt the AECOM/NIH beamline X9B at NSLS for the collection of MAD data. We are grateful to C. Krebber and A. Honegger for helpful discussions and to A. Arthur for editorial assistance. Research sponsored in part by the National Cancer Institute, DHHS, under contract with ABL, and by a Swiss National Funds. FH is recipient of an EU fellowship. The contents of this publication do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Received 28 October, 1997; accepted 12 December, 1997.

1. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).
2. Phizicky, E.M. & Fields, S. Protein–protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94–123 (1995).
3. Cortese, R. *et al* Selection of biologically active peptides by phage display of random peptide libraries. *Curr. Opin. Biotechnol.* **7**, 616–621 (1996).
4. Dunn, I.S. Phage display of proteins. *Curr. Opin. Biotechnol.* **7**, 547–553 (1996).
5. Spada, S., Krebber, C. & Plückthun, A. Selectively infective phages (SIP). *Biol. Chem.* **378**, 445–456 (1997).
6. Gray, C.W., Brown, R.S. & Marvin, D.A. Adsorption complex of filamentous fd virus. *J. Mol. Biol.* **146**, 621–627 (1981).
7. Jacobson, A. Role of F pili in the penetration of bacteriophage fl. *J. Virol.* **10**, 835–843 (1972).
8. Sun, T.P. & Webster, R.E. fii, a bacterial locus required for filamentous phage infection and its relation to colicin-tolerant tolA and tolB. *J. Bacteriol.* **165**, 107–115 (1986).
9. Sun, T.P. & Webster, R.E. Nucleotide sequence of a gene cluster involved in entry of E colicins and single-stranded DNA of infecting filamentous bacteriophages into Escherichia coli. *J. Bacteriol.* **169**, 2667–2674 (1987).
10. Riechmann, L. & Holliger, P. The C-terminal domain of TolA is the coreceptor for filamentous phage infection of E. coli. *Cell* **90**, 351–360 (1997).
11. Nelson, F.K., Friedman, S.M. & Smith, G.P. Filamentous phage DNA cloning vectors: a noninfective mutant with a nonpolar deletion in gene III. *Virology* **108**, 338–350 (1981).
12. Crissman, J.W. & Smith, G.P. Gene-III protein of filamentous phages: evidence for a carboxyl-terminal domain with a role in morphogenesis. *Virology* **132**, 445–455 (1984).
13. Endemann, H., Gailus, V. & Rasched, I. Interchangeability of the adsorption proteins of bacteriophages Ff and IKe. *J. Virol.* **67**, 3332–3337 (1993).
14. Hutchinson, E.G. & Thornton, J.M. PROMOTIF — a program to identify and analyze structural motifs in proteins. *Protein Sci.* **5**, 212–220 (1996).
15. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
16. Cabral, J.H, *et al*. Crystal structure of a PDZ domain. *Nature* **382**, 649–652 (1996).
17. Faber, H.R. et al. 1.8 Å crystal structure of the C-terminal domain of rabbit serum haemopexin. *Structure* **3**, 551–559 (1995).
18. Viguera, A.R., Blanco, F.J. & Serrano, L. The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**, 670–681 (1995).
19. Holliger, P. & Riechmann, L. A conserved infection pathway for filamentous bacteriophages is suggested by the structure of the membrane penetration domain of the minor coat protein g3p from phage Fd. *Structure* **5**, 265–275 (1997).
20. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–5, 29–32 (1996).
21. Baldwin, E.T. *et al*. Crystal structure of interleukin-8: symbiosis of NMR and crystallography. *Proc. Natl. Acad. Sci. USA* **88**, 502–506 (1991).
22. Lubkowski, J. *et al*. The structure of MCP-1 in two crystal forms provides a rare example of variable quarternary interactions. *Nature Struct. Biol.* **4**, 64–69 (1997).
23. Stengele, I., Bross, P., Garces, X., Giray, J. & Rasched, I. Dissection of functional domains in phage fd adsorption protein. Discrimination between attachment and penetration sites. *J. Mol. Biol.* **212**, 143–149 (1990).
24. Krebber, C. *et al*. Selectively-infective phage (SIP): a mechanistic dissection of a novel in vivo selection for protein-ligand interactions. *J. Mol. Biol.* **268**, 607–618 (1997).
25. Peters, B.P.H., Peters, R.M., Schoenmakers, J.G.G. & Konings, R.N.H. Nucleotide sequence and genetic organizaton of the genome of the N-specific filamentous bacteriophage IKe. Comparison with the genome of the F-specific filamentous phages M13, fd and fl. *J. Mol. Biol.* **181**, 27–39 (1985).
26. Frost, L.S., Ippen-Ihler, K. & Skurray, R.A. Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiol. Rev.* **58**, 162–210 (1994).
27. Marvin, D.A. & Folkhard, W. Structure of F-pili: reassessment of the symmetry. *J. Mol. Biol.* **191**, 299–300 (1986).
28. Glucksman, M.J., Bhattacharjee, S. & Makowski, L. Three-dimensional structure of a cloning vector. X-ray diffraction studies of filamentous bacteriophage M13 at 7 Å resolution. *J. Mol. Biol.* **226**, 455–470 (1992).
29. Lin, T.C., Webster, R.E. & Konigsberg, W. Isolation and characterization of the C and D proteins coded by gene IX and gene VI in the filamentous bacteriophage fl and fd. *J. Biol. Chem.* **255,** 10331–10337 (1980).
30. Grant, R.A., Lin, T.C., Webster, R.E. & Konigsberg, W. Structure of filamentous bacteriophage: isolation, characterization, and localization of the minor coat proteins and orientation of the DNA. *Prog. Clin. Biol. Res.* **64**: 413–428 (1981).
31. Tzagoloff, H. & Pratt, D. The initial steps in infection with coliphage M13. *Virology* **24**, 372–380 (1964).
32. Goldsmith, M.E. & Konigsberg, W.H. Adsorption protein of the bacteriophage fd: isolation, molecular properties, and location in the virus. *Biochemistry* **16**, 2686–2694 (1977).
33. Beck, E. & Zink, B. Nucleotide sequence and genome organisation of filamentous bacteriophages fl and fd. *Gene* **16**, 35–58 (1981).
34. Freund, C., Ross, A., Guth, B., Plückthun, A. & Holak, T.A. Characterization of the linker peptide of the single-chain Fv fragment of an antibody by NMR spectroscopy. *FEBS Lett.* **320**, 97–100 (1993).
35. Studier, F.W. & Moffatt, B.A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130 (1986).
36. Qoronfleh, M.W. *et al*. Production of selenomethionine-labeled recombinant human neutrophil collagenase in *Escherichia coli*. *J. Biotechnol.* **39**, 119–128 (1995).
37. Plückthun, A., et al. Producing antibodies in *Escherichia coli*: from PCR to fermentation. In Antibody engineering: a practical approach (eds McCafferty, J. & Hoogenboom, H.R., Chriswell, D.J.) 203–252 (IRL Press, Oxford; 1996).
38. Hendrickson, W.A., Horton, J.R. & LeMaster, D.M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three dimensional structure. *EMBO J.* **9**, 1665–1672 (1990).
39. Matthews, B.W. Solvent content of protein crystals. *J. Mol. Biol.* 33, 491–497 (1968).
40. Otwinowski, Z. *An oscillation data processing suite for macromolecular crystallography* (Yale University, New Haven, Connecticut, USA; 1992).
41. Jones, T.A. & Kjeldgaard, M. O — the manual (Uppsala University, Uppsala, 1994).
42. Brünger, A. *X-plor version 3.1: a system for X-ray crystallography and NMR* (Yale University Press, New Haven, Connecticut, USA; 1992).
43. Furey, W. & Swaminathan, S. PHASES - A program package for the processing and analysis of diffraction data for macromolecules. *Acta Crystallogr.* 18, 73 (1990).
44. Sheldrick, G.M. Patterson superposition and ab initio phasing. *Meth. Enz.* **276**, 628–641 (1997).
45. Wang, B.C. Resolution of phase ambiguity in macromolecular crystallography. *Meth. Enz.* **115**, 90–112 (1985).
46. CCP4:SERC Collaborative computing project no. 4 (Warrington, UK; 1979).
47. Lundblad, R.L. & Noyes, C.M. Chemical modification of tryptophan. In *Chemical reagents for protein modification* (eds. Lundblad, R.L. & Noyes, C.M.) 47–71 (CRC Press, Boca Raton, Florida; 1984).
48. Kraulis, P.J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
49. Carson, M. RIBBONS 4.0. *J. Appl. Crystallogr.* **24**, 958–961 (1991).
50. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).