Crystal Structure of Invasin: A Bacterial Integrin-Binding Protein

Zsuzsa A. Hamburger,[1] Michele S. Brown,[3] Ralph R. Isberg,[3] Pamela J. Bjorkman[1,2]*

The *Yersinia pseudotuberculosis* invasin protein promotes bacterial entry by binding to host cell integrins with higher affinity than natural substrates such as fibronectin. The 2.3 angstrom crystal structure of the invasin extracellular region reveals five domains that form a 180 angstrom rod with structural similarities to tandem fibronectin type III domains. The integrin-binding surfaces of invasin and fibronectin include similarly located key residues, but in the context of different folds and surface shapes. The structures of invasin and fibronectin provide an example of convergent evolution, in which invasin presents an optimized surface for integrin binding, in comparison with host substrates.

Many bacterial pathogens bind and enter eukaryotic cells to establish infection. *Yersinia pseudotuberculosis* and *Y. enterocolitica* are enteropathogenic Gram-negative bacteria that cause gastroenteritis when they are translocated across the intestinal epithelium at Peyer's patches by way of M cells. Translocated bacteria enter the lymphatic system and colonize the liver and spleen, where they grow mainly extracellularly (*1*). Invasin is an outer membrane protein required for efficient uptake of *Yersinia* into M cells (*2, 3*). Invasin mediates entry into eukaryotic cells by binding to members of the β_1 integrin family that lack I, or insertion, domains, such as $\alpha_3\beta_1$, $\alpha_4\beta_1$, $\alpha_5\beta_1$, $\alpha_6\beta_1$, and $\alpha_v\beta_1$ (*3*). Integrins are heterodimeric integral membrane proteins that mediate communication between the extracellular environment and the cytoskeleton by binding to cytoskeletal components and either extracellular matrix proteins or cell surface proteins (*4*). Invasin binding to β_1 integrins is thought to activate a reorganization of the host cytoskeleton to form pseudopods that envelop the bacterium (*5*). Another family of enteropathogenic bacterial proteins related to invasin, the intimins, does not appear to use integrins as its primary receptors for invasion (*6*). Instead, intimins mediate attachment of the bacteria to host cells by binding to a bacterially secreted protein Tir, which upon secretion becomes inserted into the host membrane (*6*).

Yersinia pseudotuberculosis invasin is a 986-residue protein. The NH_2-terminal ~500 amino acids, which are thought to reside in the outer membrane (*7*), are related (~36% sequence identity) to the analogous regions of intimins (*8*). The COOH-terminal 497 residues of invasin, which make up the extracellular region, can be expressed as a soluble protein (Inv497) that binds integrins and promotes uptake when attached to bacteria or beads (*9*). The shortest invasin fragment capable of binding integrins consists of the COOH-terminal 192 amino acids (*7*). This fragment is not homologous to the integrin-binding domains of fibronectin [the fibronectin type III repeats 9 and 10 (Fn-III 9–10)] (*8*), although mutagenesis studies and competition assays indicate that invasin and fibronectin bind to $\alpha_3\beta_1$ and $\alpha_5\beta_1$ integrins at the same or overlapping sites (*10*). The integrin-binding region of invasin also lacks significant sequence identity with the corresponding regions of intimins (~20% identity) (*8*). To gain insight into enteric bacterial pathogenesis and to compare the structural basis of integrin binding by invasin and Fn-III domains, we solved the crystal structure of Inv497.

Inv497 was expressed in *Escherichia coli* and purified (*9*). The structure was solved to 2.3 Å by multiple isomorphous replacement with anomalous scattering (MIRAS) (Table 1) (*11, 12*). Inv497 is a rodlike molecule with overall dimensions of ~180 Å by 30 Å by 30 Å (Fig.

Table 1. Summary of data collection and refinement statistics for Inv497. Inv497 crystals (space group $P2_1$, a = 61.1 Å, b = 50.7 Å, c = 97.9 Å, β = 98.3°; one molecule per asymmetric unit) were grown at 22°C in hanging drops by combining 1 μl of protein solution [Inv497 (5 to 10 mg/ml), 20 mM Hepes at pH 7.0, and 1 mM EDTA] with 1 μl of precipitant solution (20 mM sodium citrate at pH 5.6, 20% polyethylene glycol 4000, and 20% isopropanol). Crystals were improved by microseeding. SeMet crystals, derived from selenomethionine-substituted Inv497 protein (*9*), grew under similar conditions. For cryoprotection, 5 μl of mother liquor containing 25% isopropanol was added to the crystals immediately before transferring them to liquid nitrogen. A cryocooled xenon derivative was prepared by mounting a cryoprotected crystal in a nylon loop and subjecting it to 200 psi of xenon for 2.5 min in a xenon pressure cell (*11*). A small microfuge tube containing excess mother liquor was placed in the pressurization chamber to maintain vapor pressure and prevent cracking of the crystals. Immediately after depressurization, the crystals were transferred to liquid nitrogen. The PIP derivative was prepared by the addition of one grain of PIP to a drop containing several crystals, followed by soaking for 5 hours. Data from the native and the xenon derivative crystals were collected at −170°C at a wavelength of 0.98 Å on a MAR Research image plate detector at beam line 9-1 at SSRL. Data from the PIP and SeMet derivatives were collected at −170°C on an RAXIS IIC image plate using a Rigaku rotating anode. Statistics in parentheses refer to the highest resolution bin. Phasing, model building, and refinement were done as described (*11, 12*).

Data Set	Resolution (Å)	Complete (%)*	R_{merge} (%)†	$I/\sigma I$	rms f_h/E‡
Native	2.30 (2.34–2.30)	96.3 (97.2)	5.1 (29.6)	17.9 (2.9)	—
Xenon	2.75 (2.85–2.75)	90.4 (76.3)	4.6 (10.0)	17.1 (7.9)	1.2
PIP	2.80 (2.90–2.80)	87.7 (63.8)	7.4 (23.8)	11.1 (2.8)	1.4
SeMet	3.00 (3.11–3.00)	93.1 (91.6)	15.9 (40.0)	7.4 (2.9)	1.6

Refinement statistics

Resolution (Å)	30.0–2.3	Number of nonhydrogen atoms	
Reflections in working set	24,256	Protein	3,593
Reflections in test set	1,216	Water	195
R_{free} (%)§	27.4	Citrate	13
R_{cryst} (%)‖	22.4	Nonglycine residues in most favorable region of Ramachandran plot (%) as defined (*12*)	86.6
rms deviation from ideality			
Bond lengths (Å)	0.008		
Bond angles (degrees)	1.47		

*Complete represents (number of independent reflections)/total theoretical number. †$R_{merge}(I) = [\Sigma|I(i) - \langle I(h)\rangle|/\Sigma I(i)]$, where $I(i)$ is the ith observation of the intensity of the hkl reflection and $\langle I\rangle$ is the mean intensity from multiple measurements of the hkl reflection. ‡rms f_h/E represents phasing power, where rms is root mean square, f_h is the heavy-atom structure factor amplitude, and E is the residual lack of closure error. §R_{free} is calculated over reflections in a test set not included in atomic refinement (*12*). ‖$R_{cryst}(F) = \Sigma_h||F_{obs}(h)| - |F_{calc}(h)||/\Sigma_h|F_{obs}(h)|$, where $|F_{obs}(h)|$ and $|F_{calc}(h)|$ are the observed and calculated structure factor amplitudes for the hkl reflection.

[1]Division of Biology 156-29, [2]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, USA. [3]Department of Microbiology and Molecular Biology, Howard Hughes Medical Institute, Tufts University School of Medicine, Boston, MA 02111, USA.

*To whom correspondence should be addressed. E-mail: bjorkman@cco.caltech.edu

1A), consistent with analytical ultracentrifugation analyses that suggest the fragment has an extended monomeric structure in solution (*13*). The Inv497 structure bears an overall resemblance to that of another $\alpha_5\beta_1$-binding fragment, Fn-III repeats 7 through 10 (Fn-III 7–10) (*14*), as they are both elongated molecules composed of tandem domains. The first four Inv497 domains (D1, D2, D3, and D4) are composed mainly of β structure, and the fifth domain (D5) includes α helices and β sheets. Despite only 20% sequence identity (*8*), the D3 to D5 region of Inv497 is structurally similar to a 280-residue fragment of the extracellular portion of enteropathogenic *E. coli* intimin (*14*).

The four NH_2-terminal domains of Inv497 adopt folds resembling eukaryotic members of the immunoglobulin superfamily (IgSF) (*15*), although the Inv497 domains do not share significant sequence identity with IgSF domains and lack the disulfide bond and core residues conserved in IgSF structures (*8, 15*). D1 belongs to the I2 set of the IgSF, and D2 and D3 belong to the I1 set (*15*). D4 adopts the folding topology of the C1 set of IgSF domains, a fold seen in the constant domains of antibodies, T cell receptors, and major histocompatability complex (MHC) molecules (*15*). Unlike these C1 domains, D4 of Inv497 includes a 15–amino acid insertion between strands A and B that forms two additional β strands (A″ and A‴) (Fig. 1B). D1 and D2 of the intimin fragment are also Ig-like, and the second domain includes an insertion similar to that found in Inv497 D4 (*14*).

D5 of Inv497 has a folding topology related to that of C-type lectin-like domains (CTLDs) (Fig. 1B) (*16*). This superfamily includes true C-type lectins such as mannose-binding protein (*14*) and E-selectin (*14*), which contain carbohydrate recognition domains (CRDs) that bind carbohydrates in a calcium-dependent manner, and evolutionarily related proteins such as the Ly49 family of natural killer cell receptors, which bind ligands in the absence of calcium and may not recognize carbohydrates (*16*). A characteristic feature of C-type lectin CRDs is a long stretch of extended structure including one or two calcium-binding sites, which is required for carbohydrate recognition (*16*). The COOH-terminal domains of Inv497 and intimin lack these calcium-binding loops (Fig. 1B) (*14, 16*). Inv497 is not known to bind carbohydrates (*17*); thus, the importance of the CTLD fold remains to be determined. By analogy with Ly49A, which recognizes a carbohydrate-independent epitope on its class I MHC ligand (*16*), Inv497 may recognize an unglycosylated region of integrins.

Like CTLDs and structurally related proteins such as the COOH-terminal domain of intimin (D3) (*14*), Inv497 D5 is composed of two antiparallel β sheets with interspersed α-helical and loop regions and includes a disulfide bond linking helix 1 to β strand 5 (Fig. 1B). An additional disulfide bond linking β strand 3 and the loop following strand 4 is found in CTLDs and CRDs but is absent in Inv497 D5 and intimin D3. Whereas C-type lectin CRDs contain two α helices located between the first and second β strands, the region corresponding to the second helix is replaced by a loop in Inv497 D5 (Fig. 1B) and CD94, a component of the CD94/NKG2 natural killer cell receptor (*14*). In Inv497 D5, the loop is preceded by a two-turn α helix (residues 917 to 921 and 931 to 936) interrupted by a nine-residue loop (residues 922 to 930) (Fig. 1C) (*18*). The corresponding region in intimin was not interpretable in the nuclear magnetic resonance structure (*14*).

Extensive interactions between Inv497 D4 and D5 create a superdomain that is composed of the 192 residues identified as necessary and sufficient for integrin binding (*7*). The interface between D4 and D5 is significantly larger than the interfaces between tandem IgSF domains and between the Ig-like invasin domains (D4 to D5 buried surface area is 1925 Å^2 in comparison with ~500 Å^2 for IgSF interfaces) (*19*). The D4-D5 interface is predominantly hydrophobic, although a number of hydrogen bonds are also present (Fig. 2). The interrupted helix in D5 and strands A″ and A‴ in D4 (Fig. 1B)

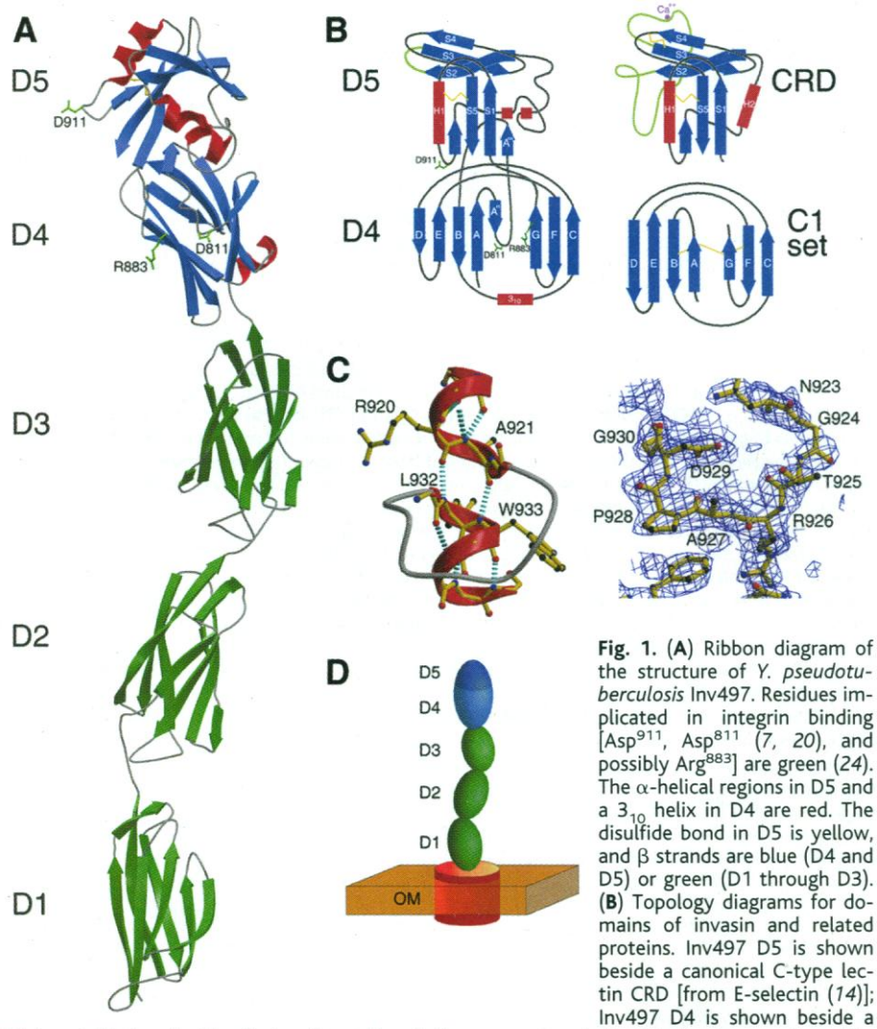

Fig. 1. (**A**) Ribbon diagram of the structure of *Y. pseudotuberculosis* Inv497. Residues implicated in integrin binding [Asp^{911}, Asp^{811} (*7, 20*), and possibly Arg^{883}] are green (*24*). The α-helical regions in D5 and a 3_{10} helix in D4 are red. The disulfide bond in D5 is yellow, and β strands are blue (D4 and D5) or green (D1 through D3). (**B**) Topology diagrams for domains of invasin and related proteins. Inv497 D5 is shown beside a canonical C-type lectin CRD [from E-selectin (*14*)]; Inv497 D4 is shown beside a C1-type IgSF domain. The β strands are blue, helices are red, and disulfide bonds are yellow. The calcium-binding loop in E-selectin (residues 54 to 89) and its truncated counterpart in Inv497 (residues 956 to 959) are green. (**C**) (left) Hydrogen bonding pattern of the interrupted helix (*18*) in D5. Main-chain atoms are shown for residues in the α helix (*24*). Side chains are shown for those residues in which main-chain atoms form hydrogen bonds (dashed light blue lines) across the break in the helix. Other side chains have been omitted for clarity. The carbon-α trace of the loop is shown in gray. Red, blue, and black balls are oxygen, nitrogen, and carbon atoms, respectively. (right) The Inv497 model (*24*) in the region of the loop (gray in left panel) of the interrupted helix superimposed on a 2.3 Å σ_A-weighted $2|F_{obs}| - |F_{calc}|$ annealed omit electron density map contoured at 1.0σ (map radius, 3.5 Å) (*12*). (**D**) Schematic model of the structure of intact invasin in which the ~500 NH_2-terminal residues reside in the *Yersinia* outer membrane (OM) (yellow) in a porin-like structure (*7*) (red), and the Inv497 portion of invasin (green and blue) projects ~180 Å from the outer membrane.

play a major role in the interaction between these two domains. In particular, a portion of the loop within the interrupted helix in D5 contacts the A‴ strand in D4 (Fig. 2). In addition, strand A‴ hydrogen bonds with strand 1 of D5, extending the second β sheet of the CTLD (Fig. 1B). The large buried surface area at the D4-D5 interface and the consequent rigidity of this portion of invasin contrasts with the flexibility between the integrin-binding portions of fibronectin, inferred from interdomain buried surface areas that are lower than average at these interfaces (Fn-III 9–10 and Fn-III 12–13) (*14, 19*). Interdomain flexibility in fibronectin was proposed to facilitate integrin binding (*15*) and is also observed in the structures of two other integrin-binding proteins, ICAM-1 (*14*) and VCAM-1 (*14*). However, invasin, which shows little or no interdomain flexibility in its integrin-binding region, binds at least five different integrins and binds $\alpha_5\beta_1$ with an affinity that is ~100 times that of fibronectin (*5, 10*). High-affinity binding of invasin is necessary for bacterial internalization, as studies have shown that bacteria coated with lower affinity ligands for $\alpha_5\beta_1$ bind, but do not penetrate, mammalian cells (*5, 10*).

Invasin residues that are important for integrin binding include 903 to 913 (*7, 20*), which form helix 1 and the loop after it in D5. The disulfide bond between Cys^{906} and Cys^{982}, conserved in all CTLDs (Fig. 1B), is required for integrin binding (*20*), presumably because it is necessary for correct folding. Although invasin lacks an Arg-Gly-Asp (RGD) sequence, which is critical for the interaction of Fn-III 10 with integrins (*4*), an aspartate in Inv497 D5 (Asp^{911}) is required for integrin binding (*7, 20*). Like the aspartate in the Fn-III RGD sequence, Asp^{911} is located in a loop (Figs. 1A and 3B). Other host proteins, such as VCAM-1 and MAdCAM-1, which bind integrins that lack I domains, also contain a critical aspartate residue on a protruding loop (*15*). By contrast, ligands of I domain–containing integrins, such as the ICAM proteins, present their acidic integrin-binding residue in the context of a β strand rather than a loop (*15*). A second region of invasin that is ~100 amino acids from Asp^{911} contains additional residues that are implicated in integrin binding, including Asp^{811} (Figs. 1A and 3B) (*20*). This region of invasin is reminiscent of the fibronectin synergy region located in Fn-III 9, which is required for maximal $\alpha_5\beta_1$ integrin-dependent cell spreading (*21*). Invasin Asp^{811} is located in D4 between strands A″ and A‴ and lies on the same surface as Asp^{911}, separated by 32 Å (measured between carbon-α atoms). The distance between Fn-III 10 Asp^{1495} in the RGD sequence and Fn-III 9 Asp^{1373} in the synergy region is also 32 Å (*14*), although the side-chain orientation of Asp^{1373} differs from that of Asp^{811} in invasin (Fig. 3). Within the Fn-III synergy region, a critical residue for integrin binding is Arg^{1379} [32 Å from Asp^{1495} (Fig. 3B)] (*21*). The invasin synergy-like region also includes a nearby arginine, Arg^{883} [32 Å from Asp^{911} (Fig. 3B)]. The overall similarity in the relative positions of these three residues

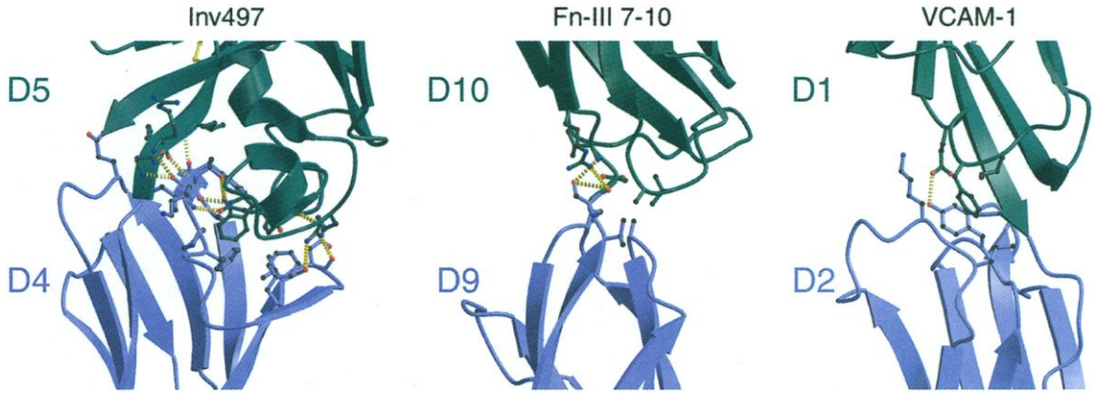

Fig. 2. Comparison of interdomain interfaces in integrin-binding regions of Inv497 (D4–D5), fibronectin type III repeats 9 and 10 (D9–D10) (*14*), and VCAM-1 (D1–D2) (*14*). Hydrogen bonds are shown as dashed yellow lines. Additional hydrogen bonds, van der Waals contacts, and a three- to fivefold larger interdomain surface area (*19*) stabilize Inv497 D4–D5 and restrict interdomain flexibility, compared to the other interfaces.

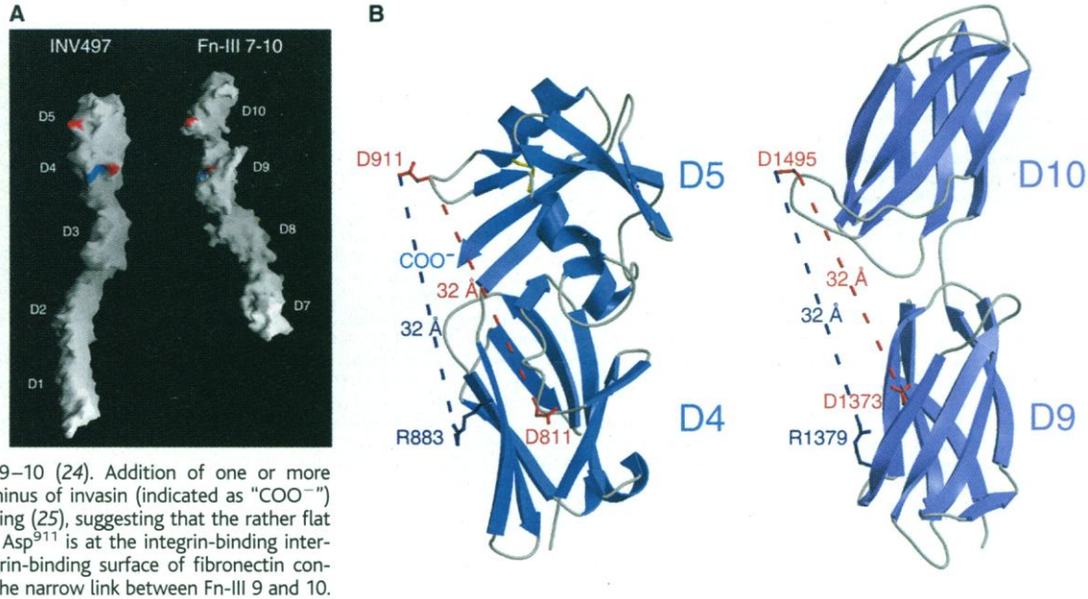

Fig. 3. Comparison of integrin-binding regions of invasin and fibronectin. Despite different folding topologies and surface structures, the relative positions of several residues implicated in interactions with integrins are similar [Asp^{811}, Asp^{911}, and Arg^{883} in Inv497; Asp^{1373}, Asp^{1495}, and Arg^{1379} in Fn-III 9 and 10; (aspartates are red; arginines are blue)]. (**A**) Surface representations (*12*) of the structures of Inv497 and Fn-III 7–10 (*14*). (**B**) Ribbon representations of Inv497 D4–D5 and Fn-III 9–10 (*24*). Addition of one or more residues to the COOH-terminus of invasin (indicated as "COO^-") interferes with integrin binding (*25*), suggesting that the rather flat region between Asp^{811} and Asp^{911} is at the integrin-binding interface. By contrast, the integrin-binding surface of fibronectin contains a cleft resulting from the narrow link between Fn-III 9 and 10.

suggests that invasin and host proteins share common integrin-binding features.

The transmembrane regions of outer membrane proteins of known structure are β barrels, as represented by the structures of porins (*7*). Assuming that the membrane-associated region of invasin is also a β barrel (*7*), the structure of intact invasin may resemble the model shown in Fig. 1D, in which the cell-binding region projects ~180 Å away from the bacterial surface, ideally positioned to contact host cell integrins. Similarities between invasin and fibronectin demonstrate convergent evolution of common integrin-binding properties. However, the integrin-binding surface of invasin does not include a cleft, as found on the binding surface of fibronectin (Fig. 3); thus, invasin may bind integrins with a larger interface. Together with the restricted orientation of the invasin integrin-binding domains, a larger binding interface provides a plausible explanation for the increased integrin-binding affinity of invasin as compared with fibronectin. Differences between the integrin-binding properties of invasin and fibronectin illustrate how a bacterial pathogen is able to efficiently compete with host proteins to establish contact and subsequent infection, thereby exploiting a host receptor for its own purposes.

References and Notes

1. J. C. Pepe and V. L. Miller, *Infect. Agents Dis.* **2**, 236 (1993); A. Marra and R. R. Isberg, *Infect. Immun.* **65**, 3412 (1997); I. B. Autenrieth and R. Firsching, *J. Med. Microbiol.* **44**, 285 (1996).
2. R. R. Isberg, D. L. Voorhis, S. Falkow, *Cell* **50**, 769 (1987).
3. R. R. Isberg and J. M. Leong, *ibid.* **60**, 861 (1990); R. R. Isberg and G. Tran Van Nhieu, *Trends Microbiol.* **2**, 10 (1994).
4. R. O. Hynes, *Cell* **69**, 11 (1992).
5. G. Tran Van Nhieu, E. S. Krukonis, A. A. Reszka, A. F. Horwitz, R. R. Isberg, *J Biol. Chem.* **271**, 7665 (1996); G. Tran Van Nhieu and R. R. Isberg, *EMBO J.* **12**, 1887 (1993).
6. H. Liu, L. Magoun, J. M. Leong, *Infect. Immun.* **67**, 2045 (1999); B. Kenny *et al.*, *Cell* **91**, 511 (1997).
7. J. M. Leong, R. S. Fournier, R. R. Isberg, *EMBO J.* **9**, 1979 (1990); M. J. Worley, I. Stojiljkovic, F. Heffron, *Mol. Microbiol.* **29**, 1471 (1998). Using a neural network algorithm that predicts the membrane topology of integral outer membrane proteins [K. Diederichs, J. Freigang, S. Umhau, K. Zeth, J. Breed, *Protein Sci.* **7**, 2413 (1998)], we predict that residues 142 to 494 contain 22 β strands. Thus, the structure of this portion of invasin may resemble those of the membrane-spanning portions of porins (16 to 18 β strands) [reviewed by B. K. Jap and P. J. Walian, *Physiol. Rev.* **76**, 1073 (1996)] or FhuA (22 β strands) [K. P. Locher *et al.*, *Cell* **95**, 771 (1998); A. D. Ferguson, E. Hofmann, J. W. Coulton, K. Diederichs, W. Welte, *Science* **282**, 2215 (1998)].
8. J. Yu and J. B. Kaber, *Mol. Microbiol.* **6**, 411 (1992). A sequence identity of 30% has been established as the threshold for guaranteed three-dimensional similarity [C. Chothia and A. M. Lesk, *EMBO J.* **5**, 823 (1986)]. Length-dependent sequence identity thresholds are discussed in work by R. A. Abagyan and S. Batalov, *J. Mol. Biol.* **273**, 355 (1997), and references therein. By these criteria, the cell-binding regions of invasin and intimin do not share significant sequence identity, and individual invasin domains do not share significant sequence similarity with Fn-III, IgSF, CRD, CTLD, or CTLD-related proteins.
9. Protein expression: The COOH-terminal 497–amino acid region of invasin was produced in *E. coli* fused to maltose-binding protein, cleaved by factor Xa, and purified as previously described [J. M. Leong, P. E. Morrissey, A. Marra, R. R. Isberg, *EMBO J.* **14**, 422 (1995)]. NH_2-terminal sequencing indicated that the first residue of the cleaved protein was Ser^{495}, five residues shorter than the predicted site of cleavage (R. R. Isberg, unpublished data). A selenomethionine (SeMet)-substituted version of Inv497 was produced following the method of W. A. Hendrickson, J. R. Horton, D. M. LeMaster, *EMBO J.* **9**, 1665 (1990), and purified under the same conditions as the native protein. Amino acid composition analysis showed ~100% replacement of the eight methionines by SeMet (M. S. Brown and R. R. Isberg, data not shown).
10. G. Tran Van Nhieu and R. R. Isberg, *J. Biol. Chem.* **266**, 24367 (1991); E. S. Krukonis, P. Dersch, J. A. Eble, R. R. Isberg, *ibid.* **273**, 31837 (1998); J. A. Eble *et al.*, *Biochemistry* **37**, 10945 (1998); Y. Takada, J. Ylanne, D. Mandelman, W. Puzon, M. H. Ginsberg, *J. Cell Biol.* **119**, 913 (1992).
11. Phasing and model building: A cryocooled xenon derivative was prepared with the apparatus described by S. M. Soltis, M. H. B. Stowell, M. C. Wiener, G. N. Phillips, D. C. Rees, *J. Appl. Crystallogr.* **30**, 190 (1997). Data were processed and scaled with the HKL package [Z. Otwinowski and W. Minor, *Methods Enzymol.* **276**, 307 (1997)]. Heavy-atom refinement and phasing were performed with the program SHARP [E. De La Fortelle and G. Bricogne, *Methods Enzymol.* **276**, 472 (1997)]. Difference Patterson maps for the xenon and di-μ-iodobis(ethylenediamine) diplatinum nitrate (PIP) derivatives were interpreted with XTALVIEW [D. E. McRee, *Practical Protein Crystallography* (Academic Press, San Diego, CA, 1993)], and one xenon, three platinum, and two iodine sites were refined with SHARP. An initial MIRAS electron density map was calculated to 3.6 Å and solvent flattened with Solomon [J. P. Abrahams and A. G. W. Leslie, *Acta Crystallogr.* **D52**, 30 (1996)] as implemented in SHARP. A skeleton of the map [G. J. Kleywegt and T. A. Jones, *Acta Crystallogr.* **D52**, 826 (1997)] served as a starting point for model building with the program O (*22*). The initial electron density map revealed the Ig-like domain structures of the first four domains, but a definitive assignment of side chains was not possible, and the connectivity in D5 was ambiguous. Using MIRAS phases, we found eight selenium sites in a difference Fourier map calculated for the SeMet derivative, which allowed identification of methionines that were used as markers for the assignment of the rest of the sequence. After including the SeMet sites in heavy-atom positional refinement using SHARP, we calculated an improved solvent-flattened MIRAS electron density map to 3.2 Å resolution with a mean figure of merit of 0.509.
12. Refinement and structure analysis: After rigid body refinement of the five domains of Inv497, refinement was carried out with the simulated annealing and energy minimization protocols in the program CNS [A. T. Brünger *et al.*, *Acta Crystallogr.* **D54**, 905 (1998)], using bulk solvent and anisotropic temperature-factor corrections (B11 = 9.590, B22 = −19.707, B33 = 10.117, B12 = 0.000, B13 = −3.944, and B23 = 0.000) and protocols that minimized R_{free} [A. T. Brünger, *Nature* **355**, 472 (1992)]. In each round of model building, a combination of σ_A-weighted [R. J. Read, *Acta Crystallogr.* **A42**, 140 (1986)] $2|F_{obs}| - |F_{calc}|$ and $|F_{obs}| - |F_{calc}|$ maps (calculated with model phases combined with experimental phases or model phases alone) and simulated annealing omit maps [A. Hodel, S.-H. Kim, A. T. Brünger, *Acta Crystallogr.* **A48**, 851 (1992)] were used. In later rounds of refinement, water molecules were built into peaks greater than 3σ in $|F_{obs}| - |F_{calc}|$ maps. The eight NH_2-terminal residues were not visible in electron density maps; thus, the final model includes residues 503 to 986 of the Inv497 construct (*9*), one citrate, and 195 water molecules, with an overall B factor of 43.0 $Å^2$. Several regions include residues with real space correlation values (*22*) below 1σ from the mean (residues 531 to 534, 582 to 586, 647 to 650, 676 to 679, 779 to 780, 892 to 899, 955 to 957, and 969 to 977). Ramachandran plot statistics (Table 1) are as defined by PROCHECK [R. A. Laskowski, M. W. McArthur, D. S. Moss, J. M. Thornton, *J. Appl. Crystallogr.* **26**, 283 (1993)]. Figures were made with MOLSCRIPT [P. J. Kraulis, *J. Appl. Crystallogr.* **24**, 946 (1991)] and RASTER-3D [E. A. Merritt and M. E. P. Murphy, *Acta Crystallogr.* **D50**, 869 (1994)]. Molecular surfaces were generated with GRASP [A. Nicholls, R. Bharadwaj, B. Honig, *Biophys. J.* **64**, A166 (1993)].
13. Equilibrium analytical ultracentrifugation analyses establish that Inv497 is monomeric at micromolar concentrations in solution [P. Dersch and R. R. Isberg, *EMBO J.* **18**, 1199 (1999)]. Sedimentation velocity analytical ultracentrifugation experiments suggest that Inv497 is elongated in solution [X.-D. Su *et al.*, *Science* **281**, 991 (1998)]; thus, the extended conformation does not result from crystal packing forces.
14. Protein structures: Fn-III 7–10 [Protein Data Bank (PDB) code 1FNF] [D. J. Leahy, I. Aukhil, H. P. Erickson, *Cell* **84**, 155 (1996)]; Fn-III 12–14 (PDB code 1FNH) [A. Sharma, J. A. Askari, M. J. Humphries, E. Y. Jones, D. I. Stuart, *EMBO J.* **18**, 1468 (1999)]; intimin (coordinates obtained from S. Matthews) [G. Kelly *et al.*, *Nature Struct. Biol.* **6**, 313 (1999)]; mannose-binding protein (PDB code 1RTM) [W. I. Weis and K. Drickamer, *Structure* **2**, 1227 (1994)]; E-selectin (PDB code 1ESL) [B. J. Graves *et al.*, *Nature* **367**, 532 (1994)]; CD94 (coordinates obtained from P. D. Sun) [J. C. Boyington *et al.*, *Immunity* **10**, 75 (1999)]; VCAM-1 (PDB code 1VSC) [E. Y. Jones *et al.*, *Nature* **373**, 539 (1995); J. Wang *et al.*, *Proc Natl. Acad. Sci. U.S.A.* **92**, 5714 (1995)]; and ICAM-1 (PDB code 1IC1) [(*23*); J. Bella, P. R. Kolatkar, C. W. Marlor, J. M. Greve, M. G. Rossmann, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4140 (1998)].
15. IgSF domains were previously classified into V, C1, C2, I1, and I2 sets on the basis of similarities in sequence and structure [Y. Harpaz and C. Chothia, *J. Mol. Biol.* **238**, 528 (1994); (*23*)]. The V and C1 sets are similar to antibody variable and constant domains, respectively. The V set consists of two β sheets: one containing β strands ABED and the other containing strands A'GFCC'C". The C1 set contains an ABED and a GFC sheet. The two sheets of the C2 set are ABE and GFCC'. The I set domains are intermediate between the V and C1 sets. The I1 set contains ABED and A'GFC sheets, and the I2 set contains ABE and A'GFCC' sheets. D1 through D4 of Inv497 adopt folding topologies that resemble IgSF domains but lack the core residues and disulfide bonds conserved in IgSF members.
16. The CTLDs in natural killer cell receptors share many features of the C-type lectin CRD fold, but differ substantially from canonical C-type lectin domains in their ligand-binding characteristics because they lack most of the calcium-coordinating residues that are critical for carbohydrate recognition in CRDs [W. I. Weis, M. E. Taylor, K. Drickamer, *Immunol. Rev.* **163**, 19 (1998)]. Ly49A, a natural killer cell receptor, recognizes a carbohydrate-independent epitope on its class I MHC ligand [N. Matsumoto, R. K. Ribaudo, J. P. Abastado, D. H. Margulies, W. M. Yokoyama, *Immunity* **8**, 245 (1998)]. Other proteins such as the TSG-6 Link module (*14*), intimin D3 (*14*), and invasin D5 are not related by obvious sequence similarity to CTLDs and CRDs [for example, they do not contain the characteristic "WIGL" sequence (*24*) in β strand 2 or the "inner" disulfide bond linking strand 3 to the loop following strand 4 (Fig. 1B)], but they share a similar folding topology. The TSG-6 Link module lacks the calcium-binding loops present in CRDs but is believed to bind hyaluronan, using an exposed patch of hydrophobic and charged residues (*14*). The CTLD-related domain of intimin contains an analogous patch of residues (Lys^{227}, Tyr^{228}, Tyr^{230}, and Tyr^{231}) (*14*). The corresponding residues of Inv497 are not conserved with intimin (Thr^{931}, Leu^{932}, Gly^{934}, and Glu^{935}).
17. Although direct binding of invasin to a carbohydrate has not been demonstrated, high concentrations of *N*-acetylneuraminic acid (median inhibitory concentration of 20 mM) inhibit mammalian cell adhesion to immobilized invasin. A variety of other acetylated sugars showed no such inhibition (R. R. Isberg, unpublished results).
18. Interrupted α helices have been observed in other protein structures, including subtilisin [reviewed by J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981)] and fibritin [Y. Tao, S. V. Strelkov, V. V. Mesyanzhinov, M. G. Rossmann, *Structure* **5**, 789 (1997)].

19. The following surface areas buried between domains in the Inv497, Fn-III 7–10, Fn-III 12–14, and VCAM-1 structures (*14*) were calculated with XPLOR (*12*) with a 1.4 Å probe radius: Inv497 D1–D2, 411 Å^2; Inv497 D2–D3, 454 Å^2; Inv497 D3–D4, 564 Å^2; Inv497 D4–D5, 1925 Å^2; Fn-III 7–8, 608 Å^2; Fn-III 8–9, 481 Å^2; Fn-III 9–10, 342 Å^2; Fn-III 12–13, 450 Å^2; Fn-III 13–14, 696 Å^2; and VCAM-1 D1–D2 (molecule B), 696 Å^2.
20. J. M. Leong, P. E. Morrissey, R. R. Isberg, *J. Biol. Chem.* **268**, 20524 (1993); L. H. Saltman, Y. Lu, E. M. Zaharias, R. R. Isberg, *ibid.* **271**, 23438 (1996).
21. R. D. Bowditch *et al.*, *ibid.* **269**, 10856 (1994); S.-I. Aota, M. Nomizu, K. M. Yamada, *ibid.*, p. 24756; T. P. Ugarova *et al.*, *Biochemistry* **34**, 4457 (1995).
22. T. A. Jones and M. Kjeldgaard, *Methods Enzymol.* **277**, 173 (1997).
23. J. M. Casasnovas, T. Stehle, J. Liu, J. Wang, T. A. Springer, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4134 (1998).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; G, Gly; L, Leu; N, Asn; P, Pro; R, Arg; T, Thr; and W, Trp.
25. R. R. Isberg, Y. Yang, D. L. Voorhis, *J. Biol. Chem.* **268**, 15840 (1993).
26. We thank S. M. Soltis and the staff at the Stanford Synchrotron Radiation Laboratory (SSRL) for help with xenon derivatization and data collection; M. J. Bennett, A. J. Chirino, L. M. Sánchez, D. E. Vaughn, and A. P. Yeh for discussions and help with crystallographic software; S. Matthews for intimin coordinates; P. D. Sun for CD94 coordinates; W. I. Weis for helpful discussions about C-type lectin structures; and W. I Weis, J. M. Leong, and members of the Bjorkman lab for critical reading of the manuscript. Inv497 coordinates have been deposited in the PDB (PDB code 1CWV).

16 June 1999; accepted 1 September 1999

Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families

Steve W. Lockless and Rama Ranganathan*

For mapping energetic interactions in proteins, a technique was developed that uses evolutionary data for a protein family to measure statistical interactions between amino acid positions. For the PDZ domain family, this analysis predicted a set of energetically coupled positions for a binding site residue that includes unexpected long-range interactions. Mutational studies confirm these predictions, demonstrating that the statistical energy function is a good indicator of thermodynamic coupling in proteins. Sets of interacting residues form connected pathways through the protein fold that may be the basis for efficient energy conduction within proteins.

Many cellular processes depend on the sequential establishment of protein-protein interactions that underlies the propagation of information through a signaling system. The interaction of one protein with another can be thought of as an energetic perturbation to each binding surface that distributes through the three-dimensional structure to cause specific changes in protein function (*1*). The structural basis for this process is largely unknown, but large-scale mutagenesis has begun to define some basic principles of energy parsing in proteins. Studies of the interaction of human growth hormone with its receptor show that binding energy is not smoothly distributed over the interaction surface; instead, a few residues comprising only a small fraction of the interaction surface account for most of the free-energy change (*2*). Similarly, high-affinity interaction of K^+ channel pores with peptide scorpion toxins buries ~15 residues on the toxin molecule, but most of the binding energy depends on only two amino acid positions (*3, 4*). Thus, protein interaction surfaces contain functional epitopes or hot spots of binding energy that are generally not predictable from the atomic structure.

In addition, a large body of evidence suggests that the change in free energy at a protein interaction surface propagates through the tertiary structure in a seemingly arbitrary manner. Studies addressing mechanisms of substrate specificity in serine proteases show that many sites distantly positioned from the active site contribute to a determination of the energetics of catalytic residues (*5*). The conversion of trypsin to chymotrypsin specificity required a large set of simultaneous mutations, many at unexpected positions. Similarly, mutations introduced during maturation of antibody specificity have been shown to occur at sites that are distant in tertiary structure from the antigen-binding site, despite substantial increases in binding energy (*6*).

An important step in understanding the problem of energy distribution in proteins is the full-scale mapping of energetic coupling between amino acid positions. Thermodynamic mutant cycle analysis (*3, 7*), a technique that measures the energetic interaction of two mutations, provides a direct method to systematically probe such relations of protein sites. However, practical considerations limit this technique to small-scale studies, precluding a full mapping of all energetic interactions on a complete protein. We report a study that uses evolutionary data for a protein family to measure energetic coupling between positions on a multiple sequence alignment (MSA).

Evolution of a protein fold is the result of large-scale random mutagenesis, with selection constraints imposed by function. The theory described below is based on two hypotheses that derive from the empirical observation of sequence evolution. The lack of evolutionary constraint at one position should cause the distribution of observed amino acids at that position in the MSA to approach their mean abundance in all proteins, and deviances from the mean values should quantitatively represent conservation. In addition, the functional coupling of two positions, even if distantly positioned in the structure, should mutually constrain evolution at the two positions, and these should be represented in the statistical coupling of the underlying amino acid distributions (*8*).

Two definitions guide the development of statistical parameters used in our analysis: (i) Conservation at a given site in a MSA is defined as the overall deviance of amino acid frequencies at that site from their mean values, and (ii) statistical coupling of two sites, i and j, is defined as the degree to which amino acid frequencies at site i change in response to a perturbation of frequencies at another site, j. This definition of coupling does not require that the overall conservation of site i change upon perturbation at j, but only that the amino acid population be rearranged. Therefore, we describe a site by a vector of 20 binomial probabilities of individual amino acid frequencies instead of the scalar multinomial probability of the overall amino acid distribution (*9*). This approach uniquely represents all changes in an amino acid distribution regardless of conservation at a given site.

For an evolutionarily well sampled MSA, where additional sequences do not significantly change the distribution at sites, the probability of any amino acid x at site i relative to that at another site, j, is related to the statistical free energy separating sites i and j for amino acid x ($\Delta G^x_{i\rightarrow j}$) by the Boltzmann distribution (*10*)

$$\frac{P^x_i}{P^x_j} = e^{\frac{\Delta G^x_{i\rightarrow j}}{kT^*}} \quad (1)$$

where kT^* is an arbitrary energy unit (*11*). The probability of any amino acid x at site i (P^x_i) is given by the binomial probability of the observed number of x amino acids, given its mean frequency in all proteins (Fig. 1A) (*12*). The full distribution of amino acids at a site i can then be characterized by a 20-element vector of P^x_i for all x ($\overrightarrow{P^x_i}$). If we take

Howard Hughes Medical Institute and Department of Pharmacology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75235-9050, USA.

*To whom correspondence should be addressed. E-mail: rama@chop.swmed.edu

References and Notes

[7] **Siderophore-Mediated Iron Transport: Crystal Structure of FhuA with Bound Lipopolysaccharide**
Andrew D. Ferguson; Eckhard Hofmann; James W. Coulton; Kay Diederichs; Wolfram Welte
Science, New Series, Vol. 282, No. 5397. (Dec. 18, 1998), pp. 2215-2220.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819981218%293%3A282%3A5397%3C2215%3ASITCSO%3E2.0.CO%3B2-U

[13] **Crystal Structure of Hemolin: A Horseshoe Shape with Implications for Homophilic Adhesion**
Xiao-Dong Su; Louis N. Gastinel; Daniel E. Vaughn; Ingrid Faye; Pak Poon; Pamela J. Bjorkman
Science, New Series, Vol. 281, No. 5379. (Aug. 14, 1998), pp. 991-995.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819980814%293%3A281%3A5379%3C991%3ACSOHAH%3E2.0.CO%3B2-Y

[14] **The Crystal Structure of an N-Terminal Two-Domain Fragment of Vascular Cell Adhesion Molecule 1 (VCAM-1): A Cyclic Peptide Based on the Domain 1 C-D Loop can Inhibit VCAM-1-α4 Integrin Interaction**
Jia-Huai Wang; R. Blake Pepinsky; Thilo Stehle; Jin-huan Liu; Michael Karpusas; Beth Browning; Laurelle Osborn
Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 12. (Jun. 6, 1995), pp. 5714-5718.
Stable URL:
http://links.jstor.org/sici?sici=0027-8424%2819950606%2992%3A12%3C5714%3ATCSOAN%3E2.0.CO%3B2-D

LINKED CITATIONS

[14] **The Structure of the Two Amino-Terminal Domains of Human ICAM-1 Suggests How it Functions as a Rhinovirus Receptor and as an LFA-1 Integrin Ligand**

Jordi Bella; Prasanna R. Kolatkar; Christopher W. Marlor; Jeffrey M. Greve; Michael G. Rossmann

Proceedings of the National Academy of Sciences of the United States of America, Vol. 95, No. 8. (Apr. 14, 1998), pp. 4140-4145.

Stable URL:

http://links.jstor.org/sici?sici=0027-8424%2819980414%2995%3A8%3C4140%3ATSOTTA%3E2.0.CO%3B2-A

[23] **A Dimeric Crystal Structure for the N-Terminal Two Domains of Intercellular Adhesion Molecule-1**

Jose M. Casasnovas; Thilo Stehle; Jin-Huan Liu; Jia-Huai Wang; Timothy A. Springer

Proceedings of the National Academy of Sciences of the United States of America, Vol. 95, No. 8. (Apr. 14, 1998), pp. 4134-4139.

Stable URL:

http://links.jstor.org/sici?sici=0027-8424%2819980414%2995%3A8%3C4134%3AADCSFT%3E2.0.CO%3B2-8