# Possible Arrangement of the Five Domains in Human Complement Factor I As Determined by a Combination of X-ray and Neutron Scattering and Homology Modeling

Dean Chamberlain, Christopher G. Ullman, and Stephen J. Perkins*

*Department of Biochemistry and Molecular Biology, Royal Free Campus, Royal Free and University College Medical School of University College London, Rowland Hill Street, London NW3 2PF, U.K.*

*Received March 5, 1998; Revised Manuscript Received August 10, 1998*

ABSTRACT: Human factor I is a multidomain plasma serine protease with one factor I−membrane attack complex (FIMAC) domain, one CD5 domain, two low-density lipoprotein receptor (LDLr) domains, and one serine protease (SP) domain and is essential for the regulation of complement. The domain arrangement in factor I was determined by X-ray and neutron scattering on serum-derived human factor I (sFI) and recombinant insect cell factor I (rFI). While the radii of gyration of both were the same at 4.05 nm and both had overall lengths of 14 nm, the cross-sectional radii of gyration were different at 1.70 nm for sFI and 1.57 nm for rFI. This difference was attributed to their different means of glycosylation which is complex-type for sFI and high-mannose-type for rFI. Homology models were constructed for the FIMAC, LDLr, and SP domains of factor I using related crystal structures, and CD5 was represented as a globular protein by referencing its electron microscopy dimensions. In these models, 38 of the 40 Cys residues in factor I were predicted to form internal disulfide bridges. The two remaining Cys residues at the N terminus of the FIMAC domain and at the center of the first LDLr domain were potentially not bridged. It was postulated that, if these two Cys residues were bridged to each other, the FIMAC, CD5, and LDLr-1 domains would form a compact triangular arrangement. This hypothesis was tested by automated scattering curve fit searches based on 9600 bilobal models, setting the FIMAC, CD5, and LDLr-1 domains as one lobe and the large SP domain as the other lobe. The searches gave a single small family of similar structures with a separation of 5.9 nm between the centers of the lobes which gave similar good X-ray and neutron fits for both sFI and rFI, despite the different glycosylations of sFI and rFI. These best-fit structures for factor I showed that this domain model is plausible, and suggested that the SP and the CD5 and LDLr-1 domains may present exposed surfaces in factor I whose roles are to interact separately with its substrates C3b and C4b and with cofactor proteins.

Factor I of the complement system of immune defense is a five-domain serine protease which is involved in the regulation of the C3 convertase of the classical or alternative pathways of activation (*1*, *2*). It specifically cleaves the α′ chains of C3b and C4b into smaller fragments in the presence of the cofactor proteins factor H or C4b-binding protein, respectively. Other cofactors for factor I-mediated cleavage of C3b and C4b include complement receptor type 1 (CR1, CD35) and the membrane cofactor protein (MCP, CD45). Unusually, factor I is not inhibited by any known plasma protease inhibitors, and is specific only for C3b and C4b. The importance of factor I is demonstrated in deficiencies that lead to the excessive consumption of C3 and recurrent pyogenic infections (*3*).

Human factor I contains a heavy chain with 317 residues (including 27 Cys residues) and a catalytic light chain with 244 residues (including 11 Cys residues) (*4*, *5*). An 18-residue signal sequence and a four-residue RRKR linker between the heavy and light chains are removed during processing. Three glycosylation sites are present on each of the heavy and light chains (Figure 1), giving a total molecular weight of 85 000 and an extent of glycosylation of 26% (w/w) (*6*). The four domains in the heavy chain (Figure 1) are the factor I−membrane attack complex (FIMAC)[1] domain, the CD5-type domain (also known as the scavenger receptor cysteine-rich domain), and two low-density lipoprotein receptor (LDLr-1 and -2) domains. Additional 24- and 32-residue sequences are present at the N terminus and C terminus of the heavy chain, respectively, the latter of which is strongly species-dependent (*7*, *8*). The light chain contains a serine protease (SP) domain which is disulfide-linked to the heavy chain. To appreciate the function of the five domains in factor I, we have expressed and characterized recombinant factor I (rFI) in a baculovirus system (*9*). rFI was determined to be folded correctly, but differed in its molecular weight from serum-derived factor I (sFI). This can be attributed to the replacement of the complex-type oligosaccharide chains by high-mannose-type chains in baculovirus (*10*).

* To whom correspondence and requests for reprints should be addressed.

[1] Abbreviations: sFI, serum factor I; rFI, recombinant factor I; FIMAC, factor I−membrane attack complex; LDLr, low-density lipoprotein receptor; SP, serine protease.

S0006-2960(98)00518-2 CCC: $15.00 
Published on Web 09/18/1998

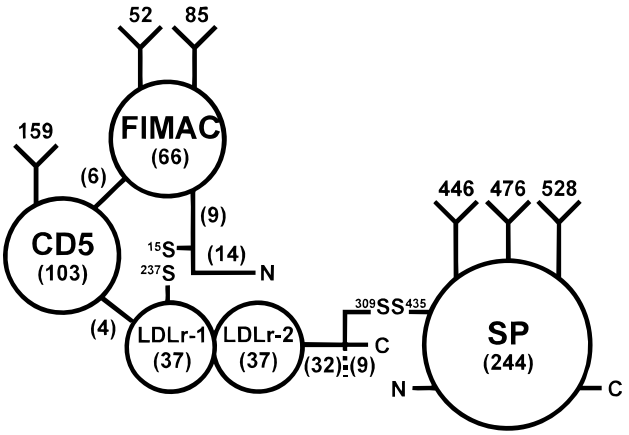


FIGURE 1: Possible domain structure of factor I. This is constructed from the FIMAC, CD5, LDLr-1, LDLr-2, and SP domains. Cys15 and Cys237 are predicted to be bridged to link the FIMAC, CD5, and LDLr-1 domains in a triangular arrangement. The heavy and light chains are linked by Cys309 and Cys435. Putative N-linked oligosaccharide chains are located at Asn52, Asn85, and Asn159 in the heavy chain and at Asn446, Asn476, and Asn528 in the light chain. There are 561 residues in processed factor I. The number of residues in each part of factor I is indicated in brackets.

The three-dimensional arrangement of the domains in factor I is poorly understood. X-ray and neutron scattering studies of factor I in solution had shown that its overall length is between 12.8 and 15 nm, and electron microscopy of factor I in vacuo stained with uranyl acetate had shown that factor I was 13 nm in length and bilobal; however, no molecular explanation of these findings had been provided (*6*, *11*). Since that time, solution scattering has been improved by the establishment of a procedure for calculating scattering curves from known crystal structures (*12*−*14*). A new automated curve-fit procedure employs constraints based on homologous atomic structures for individual domains and their known covalent connectivity to yield molecular models for domain arrangements in the intact protein (*15*−*17*; reviewed in ref *18*). This constrained modeling method can now be applied to factor I as the result of recently determined crystal structures that are relevant for the FIMAC, LDLr, and SP domains of factor I and the recently determined disulfide bridges in the globular structure of the CD5 domain (*19*−*23*). Here, homology modeling and a consideration of disulfide bridges in factor I suggested that the FIMAC, CD5, and LDLr-1 domains may form a compact triangular arrangement stabilized by a disulfide bridge between Cys15 and Cys237. Together with the SP domain, this defined two globular entities that would constitute a bilobal structure in factor I. The feasibility of such a structural model was tested using automated X-ray and neutron curve fit analyses of possible bilobal structures for sFI. Since solution scattering is sensitive to oligosaccharide conformations (*17*), and rFI contained a different oligosaccharide structure, an independent set of X-ray and neutron curve fits for rFI could be used to test the outcome of the sFI analysis. The outcome of these analyses showed that this bilobal structure for factor I was plausible, and the functional significance of such a domain structure is discussed.

## MATERIALS AND METHODS

*(a) Expression and Purification of Factor I for Scattering*. Four preparations of sFI were obtained from 0.6 L of outdated human plasma for each one (*1*, *9*). Using separate apparatus to avoid cross contamination with sFI, four preparations of purified active rFI were obtained from a recombinant baculovirus expression system using 2−3 L of culture for each one. Full details are presented elsewhere (*9*). Samples were stored frozen at −20 °C. When needed for scattering, samples were subjected to gel filtration to remove trace aggregates using a Superdex-200 Prep grade column (1.6 cm × 60 cm) (Pharmacia) and stored at 4 °C. For X-ray scattering and neutron scattering in $H_2O$ buffers, samples were dialyzed into Dulbecco's phosphate-buffered saline at pH 7.0 (137 mM NaCl, 2.7 mM KCl, 8.1 mM $Na_2HPO_4$, and 1.5 mM $KH_2PO_4$) (Sigma) together with 0.1 mM Pefabloc-SC (Pentapharm) and 0.5 mM EDTA. For neutron scattering, the samples were dialyzed as above but now using $^2H_2O$ buffers with four buffer changes over 36 h at 6 °C. Before and after data collection, samples were checked by SDS−PAGE. Fourier transform infrared and circular dichroism spectroscopy was performed on rFI and sFI to verify the similarity of their folded protein structures (*9*).

To assay for free Cys residues in factor I, using Ellman's reagent [5,5′-dithiobis(2-nitrobenzoic acid)] with 1.5 nmol of sFI in 0.1 M sodium phosphate (pH 7.3), using 0.5−10 nmol of reduced glutathione as a reference, we detected less than 0.3 free thiol group per sFI. In 5 M guanidine, similar results were obtained for unfolded sFI and rFI. In addition, the use of 2 $\mu$Ci of $^{14}$C-labeled iodoacetamide with 20 $\mu$g of denatured sFI with 8 M guanidine in 0.2 M Tris (pH 8.2) confirmed this. A second sFI sample was previously incubated with 0.5 mM iodoacetamide for 10 min at 37 °C followed by dialysis (negative control). A third sFI sample was previously incubated with 40 mM dithiothreitol for 2 h in the dark (positive control). The three samples were dialyzed extensively and counted for bound radioactivity. Significant radioactivity was present only in the dithiothreitol-treated sample.

*(b) X-ray and Neutron Scattering Data Collection*. X-ray data were obtained in one beam session at Station 2.1 using a camera with a quadrant detector at the Synchrotron Radiation Source (Daresbury, U.K.). A sample−detector distance of 3.58 m was used, with beam currents of 154−184 mA and a storage ring energy of 2.0 GeV. This resulted in a usable $Q$ range of 0.1−2.3 nm$^{-1}$ ($Q = 4\pi \sin\theta/\lambda$, scattering angle of $2\theta$, wavelength of $\lambda$). Samples were measured between 1.7 and 3.3 mg/mL at 15 °C in cells with a path thickness of 1 mm with mica windows. Data acquisition times were 10 min, obtained as 10 time frames of 1 min each to confirm the absence of radiation damage. Neutron data were obtained in one session on Instrument D22 at the Institut Laue-Langevin (Grenoble, France) which is analogous to Instrument D11 (*24*). Sample−detector distances of 1.4 and 5.6 m were used. Using a $\lambda$ of 1.00 nm, a 64 cm × 64 cm detector, and a rectangular beam aperture of 7 mm × 10 mm, the two detector positions resulted in a usable $Q$ range of 0.06−2 nm$^{-1}$. Samples were measured between 0.3 and 1.4 mg/mL at 15 °C in rectangular quartz Hellma cuvettes with a path length of 2 mm for acquisition times between 10 and 30 min. Neutron data were also obtained in two sessions on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory (Didcot, U.K.), using a proton beam current of 170−185 $\mu$A to generate neutrons. On the basis of a fixed

sample−detector distance of 4.3 m, the usable $Q$ range was 0.1−2.0 nm$^{-1}$. For concentrations between 1.7 and 4.3 mg/mL, the data acquisition time was 3−4 h at a sample temperature of 15 °C. Other details, data reduction, and references are given elsewhere (*14*).

*(c) Guinier and Distance Distribution Function Analyses.* Guinier analyses at low $Q$ values gave the radius of gyration $R_G$ and the forward scattering at zero angle $I(0)$ (*25*):

$$\ln I(Q) = \ln I(0) - R_G^2 Q^2/3$$

This expression is valid in a $QR_G$ range of up to 1.5. The $R_G$ value is a measure of structural elongation if the internal inhomogeneity of scattering densities has no effect. The $I(0)/c$ values (where $c$ is the sample concentration) lead to molecular weights $M_r$. For elongated macromolecules, the mean radius of gyration of the cross section $R_{XS}$ and the mean cross-sectional intensity at zero angle $[I(Q)Q]_{Q\to 0}$ (*26*) were obtained from

$$\ln[I(Q)Q] = [\ln[I(Q)Q]]_{Q\to 0} - R_{XS}^2 Q^2/2$$

Combination of the $R_G$ and $R_{XS}$ analyses lead to triaxial dimensions (*26*). Indirect transformation of the scattering data in reciprocal space $I(Q)$ into that in real space $P(r)$ was performed using GNOM (*27*).

$$P(r) = \frac{1}{2\pi^2}\int_\infty^0 I(Q)Qr\,\sin(Qr)\,\mathrm{d}Q$$

$P(r)$ corresponds to the distribution of distances $r$ between volume elements, from which the $R_G$ and $I(0)$ values can be determined as well as the maximum dimension $L$. A range of assumed maximum lengths for sFI and rFI were tested to optimize the calculation of the $P(r)$ curve (*14*).

*(d) Homology Modeling of the Domains of Factor I.* Homology models for the FIMAC, LDLr, and SP domains were constructed using the sequence alignment of Figure 2 and INSIGHT II 95.0, BIOPOLYMER, HOMOLOGY, and DISCOVERY software (Biosym/MSI, San Diego, CA) on Silicon Graphics INDY workstations. Loops were built using the pdb_select.1995-jun-01 database derived from 349 crystal structures at 0.2 nm resolution or better (*28*, *29*). Energy refinements were based on the consistent valence force field. Iterations were made using combinations of the steepest descent and conjugate algorithms to improve the connectivity of the model and minimize bad contacts or stereochemistry. Models were stereochemically verified using PROCHECK (*30*). Solvent accessibilities were calculated using COMPARER (*31*, *32*). Electrostatic maps were calculated using DELPHI (Biosym/MSI).

The FIMAC model (residues 24−89) was constructed using residues 54−159 in the crystal structure of SPARC (Brookhaven Protein Data Bank code 1bmo; *19*, *20*). Using the rigid body fragment assembly method, nine structurally conserved regions based on α-helix and $\beta$-sheet residues and Cys residues (total of 27 residues) and eight designated loops (total of 27 residues) were defined. Three loops (11 factor I residues, 37−40, 50−53, and 79−81) that correspond to deletions were constructed using database searches. A total of 24 conformationally unassigned N-terminal residues were added to the FIMAC model using the end repair command

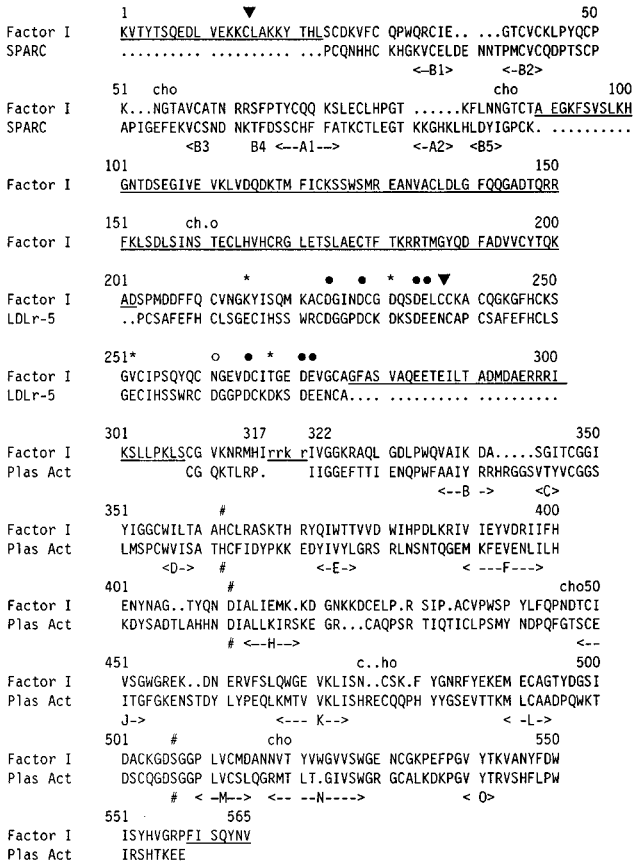


FIGURE 2: Sequence alignment used for homology modeling of factor I. The numbering is that of human factor I. The domain sequences are aligned with those corresponding to crystal structures for human SPARC (PDB code 1bmo), the fifth domain of the human low-density lipoprotein receptor (LDLr-5; PDB code 1ajj), and human plasminogen activator (Plas Act; PDB code 1lmw). Underlined sequences do not correspond to known homologous crystal structures, although the disulfide bridges in CD5 are known. Symbols above the factor I sequence denote the locations of two exposed unpaired Cys residues (▼, see the text), the four acidic residues that constitute a potential $Ca^{2+}$ binding site in the LDLr domain (● and ○ for nonconserved residues and * for other related residues), and six putative N-linked oligosaccharide sites (cho). The catalytic triad at His362, Asp411, and Ser507 is marked by #.

to represent the N terminus of factor I, while six additional residues were added at the C terminus as a designated loop based on the 1bmo structure to facilitate the connection of the FIMAC and CD5 domains (Figures 1 and 2). Since these constitute 30 residues of the 615−663 amino acid and carbohydrate residues in factor I, their contribution to the scattering curve modeling was considered to be low (see below). Energy refinements were performed at the six loop splice junctions, and then the five disulfide bridges were created. The final energy refinements were performed on the side chain atoms of mutated residues in the structurally conserved regions, the side chain atoms of both types of loop residues, and the added N- and C-terminal residues. The secondary structure backbone was retained by fixing the main chain atoms in the conserved regions, and tethering these in the loop regions. The length of this model is 4.14 nm (between Ser24 and Thr89).

The globular CD5 structure visualized by electron microscopy (*23*) was represented by the coordinates of an immunoglobulin fold since this is the same size as that of CD5. Residues 3−107 in the $V_L$ domain of human IgG1

HIL (PDB code 8fab; chain A) were used because their total was close to the 102 residues found in the CD5 superfamily (see below). Since the construction of a Debye sphere model required the correct total of amino acids within the coordinates to set the volume correctly (see below), residues in the IgG1 HIL sequence were directly replaced one by one by those in the CD5 sequence of factor I (Figure 2). In the homology modeling procedure, this structure was treated as a designated loop, and energy refinements were performed on all atoms to minimize bad contacts.

The LDLr-1 and LDLr-2 domains (residues 203−276) were constructed as a double-domain structure LDLr-1/2 using residues 4−40 in the single-domain crystal structure of LDLr-5 (PDB code 1ajj; *21*, *33*). Since no gaps or insertions occurred in the sequence alignment, and the linker region between the two domains is flanked by Cys247 and Cys250 with no additional residues between them (Figure 2), both were combined into a single structure. An extended link between the two domains was created, since (as shown by Ramachandran plots) less extended or bent links would not be favored because of the resulting steric obstruction between the two domains. The N-terminal residue of the LDLr-2 domain Pro4 was changed into Ala4 to permit this connection to be made. The side chains were replaced with those in factor I. The connection with the CD5 domain was facilitated by adding six N-terminal linker residues to LDLr-1 by an end repair. Energy refinements were performed as for the FIMAC domain, all residues being considered as a structurally conserved region. The length of this double-domain model is 4.25 nm (between Ser203 and Ala276).

The SP domain (residues 309−565) was constructed from residues 1−245 in the two-chain crystal structure of the human urokinase-type plasminogen activator (PDB code 1lmw; *22*). This was based on 12 conserved regions corresponding to the 12 $\beta$-strands and the single α-helix of a SP domain (*34*) and 12 Cys residues (total of 201 residues), together with three designated loops (total of 11 residues). Nine searched loops (factor I residues 419−425, 439, 443, 458−461, 475, 476, 479−481, and 523, a total of 25 residues) corresponding to insertions and deletions were constructed from database searches. Five disulfide bridges were created. The C terminus of the heavy chain (residues 309−317) was modeled independently, and then combined with the light chain by creating the Cys309−Cys435 disulfide bridge. Energy refinements were performed at the 18 loop splice junctions, and then on the side chain and main chain residues as for the FIMAC domain except that the catalytic triad (His362, Asp411, and Ser507) was fixed in position. Seven unassigned residues were added at the C terminus by an end repair.

The oligosaccharide chains were modeled on the nine-residue structure in the Fc fragment of human IgG1 KOL (PDB code 1fc1; *35*), to which extra residues were added to generate a tetraantennary complex-type structure $NeuNAc_4$-$Gal_4Man_3GlcNAc_6$. The high-mannose-type structure was formed by an adaptation of this structure to $Man_7GlcNAc_2$. The chains were positioned at Asn residues in extended conformations from the protein surface (Figure 1).

*(e) Construction of Extended and Bilobal Domain Structures in Factor I.* The linear extended domain model for factor I was formed by positioning the long axes of the FIMAC, CD5, and LDLr-1/2 models and the N-terminal Cys309−Met315 fragment of the SP model on a common axis in arbitrary rotational orientations about this common axis. Bent domain models without steric overlap between the models were created from this by manually rotating the FIMAC and CD5 models about the N-terminal α-carbon atom of the LDLr-1/2 model as the origin.

In the bilobal domain model for factor I, one lobe was represented by a triangular domain arrangement in which the FIMAC and LDLr-1/2 domains were positioned to create a disulfide bridge between Cys15 and Cys237, and the CD5 model was inserted into the 2.7 nm gap between the α-carbon atoms at Thr89 (C terminus of the FIMAC domain) and Ser203 (N terminus of the LDLr-1/2 domains). Cartesian axes were assigned with the origin set as the α-carbon atom of Lys249 (LDLr-2); the *X*-axis was defined by the C2 atom of the GlcNAc residue on one of the four antennae of the oligosaccharide at Asn159, and the *Y*-plane was defined by the α-carbon atom of Gln258 (LDLr-2). This set the longest axis of the triangular model as the *X*-axis, which is approximately equivalent to the long axes of the FIMAC and LDLr-1/2 domains. For the SP model used to represent the other lobe, the origin was set as the C-terminal α-carbon atom of Val565 (Figure 2), and the *X*-axis and *Y*-plane were defined by the α-carbon atoms of Lys368 and Ser367, respectively. This defined the *X*-axis of the SP domain so that the three oligosaccharide chains were located in a single $Z-Y$ quadrant that was easily moved using 90° rotations about the *X*-axis.

Automated conformational searches optimized the best relative position of the two lobes in the bilobal model for factor I (*18*). Models for factor I were created using translations of the SP model as one lobe relative to the triangular model as the other lobe using an INSIGHT MSI/Biosym Command Language (BCL) macro in conjunction with Unix shell scripts. The origins of the two lobe models were positioned 9.66 nm apart, and the *X*-axes and the $X-Z$ planes of both lobe models were set parallel to each other. The SP model was translated in 20 nm × 0.5 nm steps along the *Z*-axis and in 30 nm × 0.5 nm steps along the *X*-axis for each of four 90° *X*-axis orientations of the triangular model to create 4 × 600 factor I models. The translations positioned the SP model on all sides of the triangular model as well as passing through it, while the four *X*-axis rotations of the triangular model explored the consequence of the asymmetric positioning of oligosaccharides on it. Other searches were performed for three further 90° *Y*-axis rotations of the triangular model which were equivalent to parallel *X*-axis rotations of the SP model. This gave a final total of 4 × 4 × 600 (9600) bilobal models to provide a comprehensive test of possible structures.

*(f) Scattering Curve Modeling of Factor I.* The Debye sphere models and the scattering curve calculations were derived as follows (*18*). To generate the spheres, each coordinate model was placed in a three-dimensional grid of cubes with a side length of 0.3775 nm. By use of a cutoff scheme, a sphere of the same volume as the cube was created at the center of each cube if a specified number of atoms in the coordinates were present in the cube. The cutoff was based on the constraint that the total volume of the spheres equaled that of the 561 amino acid and 102 complex-type carbohydrate residues in sFI (*36*), and this minimized the contribution of the 30 unassigned N-terminal and C-terminal

Table 1: Experimental and Modeled Scattering Analyses for sFI and rFI

| experiment | Guinier analyses[a] $R_G$ (nm) | $R_{XS}$ (nm) | GNOM analyses[b] $R_G$ (nm) |
|---|---|---|---|
| sFI (X-rays) | 4.04 ± 0.27 (4) | 1.70 ± 0.15 | 4.31 ± 0.31 |
| sFI (neutrons) | 4.00 ± 0.14 (6) | 1.51 ± 0.08 | 4.08 ± 0.21 |
| rFI (X-rays) | 4.06 ± 0.12 (6) | 1.57 ± 0.10 | 4.35 ± 0.29 |
| rFI (neutrons) | 4.18 ± 0.07 (4) | 1.22 ± 0.06 | 4.29 ± 0.12 |
| modeling | Guinier analyses $R_G$ (nm) | $R_{XS}$ (nm) | *R*-factor[b] (%) |
| linear sFI model | 6.14 | 0.88 | 16.1 |
| partly bent sFI model | 5.32 | 0.93 | 13.2 |
| half-bent sFI model | 4.75 | 1.73 | 12.2 |
| fully bent sFI model | 3.71 | 2.25 | 14.2 |
| best-fit sFI model | | | |
| X-rays | 4.10 | 1.85 | 10.2 |
| neutrons | 4.17 | 1.44 | 10.2 |
| best-fit rFI model | | | |
| X-rays | 4.01 | 1.61 | 11.4 |
| neutrons | 4.20 | 1.25 | 10.0 |

[a] The number of scattering curves measured for each sample is shown in parentheses. The *Q* range used for the $R_G$ determinations was 0.20−0.35 $nm^{-1}$ (Figure 3a,b), while that used for the $R_{XS}$ determinations was 0.45−0.81 $nm^{-1}$ (Figure 3c,d). [b] The *R*-factor goodness-of-fit parameter corresponds to X-ray data in the *Q* range between 0.20 and 2.0 $nm^{-1}$ and neutron data in the *Q* range between 0.29 and 1.8 $nm^{-1}$. These corresponded to the *Q* ranges used to calculate the *P*(*r*) curves (Figure 4).

FIMAC residues and the unavailable coordinates for the 32 C-terminal residues of the heavy chain to the scattering curve (Figure 1). The sFI models contained about 1904 spheres (102.5 $nm^3$). Models that incorporated the 54-high mannose-type carbohydrate residues in rFI were obtained using cubes with a side length of 0.371 nm and contained about 1751 spheres (89.4 $nm^3$). Next, using these Debye sphere models, the X-ray and neutron scattering curves *I*(*Q*) were calculated using the program SCT (*37*). Dry models were used for neutron curve fits since hydration shells are not visible by neutron scattering (*36*), while a hydration of 0.3 g of $H_2O$/g of glycoprotein was used for X-ray curve fits (*12*−*14*). The high carbohydrate content of factor I may affect the curve fits. As no significant differences were observed in the X-ray and neutron curve fits that would have reflected scattering density differences between the protein and carbohydrate, single-density spheres were used in modeling (*6*, *14*, *17*). For neutron fits of the D22 and LOQ data, corrections for wavelength spread or beam divergence were performed as described in ref *15*. For X-ray fits, no instrumental corrections were applied as these are considered negligible. To evaluate the models, the $R_G$ and $R_{XS}$ values were calculated from Guinier fits of the modeled curves in the same *Q* ranges used for experimental data. These values were filtered to retain only those models for which 3.9 nm ≤ $R_G$ ≤ 4.3 nm, 1.55 nm ≤ $R_{XS}$(X-rays) ≤ 1.85 nm, and 1.4 nm ≤ $R_{XS}$(neutrons) ≤ 1.6 nm (Table 1), and at least 95% of the expected total of 1810 spheres were present. After this, the filtered models were sorted and ranked using the *R*-factor goodness-of-fit parameter for curve fits, where the *R*-factor is defined by analogy with crystallography, for which *I*(0) was set to 1000 (*12*).

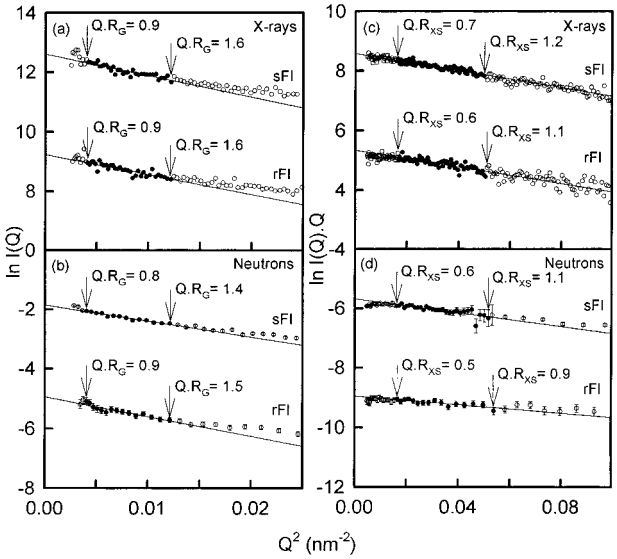


FIGURE 3: X-ray and neutron Guinier analyses of sFI and rFI. Filled circles between the indicated $QR_G$ and $QR_{XS}$ ranges show the data points used to determine the $R_G$ and $R_{XS}$ values (Table 1). Statistical error bars are shown when they are large enough to be visible. (a and c) X-ray Guinier $R_G$ and $R_{XS}$ plots are shown for sFI and rFI at concentrations of 3.3 and 2.0 mg/mL, respectively. (b and d) Neutron Guinier $R_G$ and $R_{XS}$ plots are shown for sFI and rFI in 100% $^2H_2O$ buffer at concentrations of 0.5 and 1.7 mg/mL, respectively, measured using Instruments D22 and LOQ, respectively.

## RESULTS AND DISCUSSION

*(a) X-ray and Neutron Scattering Data for sFI and rFI.* X-ray and neutron scattering were used to compare the domain structures of sFI and rFI. The X-ray data depict a hydrated structure in a high positive solute−solvent contrast, while the neutron data depict a dry structure in a high negative solute−solvent contrast. These opposite contrasts act as a control for large internal density effects that may be caused by the 26 and 15% carbohydrate contents of sFI and rFI, respectively (*17*, *36*). X-ray and neutron studies of sFI and rFI also compared the effect of the replacement of the complex-type oligosaccharides in sFI with high-mannose-type ones in rFI that is established to occur in baculovirus expression systems (*9*, *10*).

During X-ray data acquisition on Station 2.1, analyses of the 10 time frames of 1 min each revealed small time-dependent effects due to irradiation. Only the first time frame was used for data analysis to avoid this effect. X-ray data at low *Q* values on sFI measured at concentrations between 1.7 and 3.3 mg/mL yielded a mean $R_G$ value of 4.04 ± 0.27 nm from linear Guinier plots in an acceptable $QR_G$ range of 0.9−1.6 (Figure 3a and Table 1). Those for rFI measured at 2.0 mg/mL also yielded a mean X-ray $R_G$ value of 4.06 ± 0.12 nm, in good agreement with the sFI data. The X-ray data showed that the overall domain structures of sFI and rFI were similar and were unaffected by the change in oligosaccharide content.

The neutron scattering data from Instruments D22 and LOQ also resulted in linear Guinier $R_G$ plots for sFI and rFI in 100% $^2H_2O$ (Figure 3b), from which similar mean $R_G$ values of 4.00 ± 0.14 and 4.18 ± 0.07 nm, respectively, were obtained (Table 1). In molecular weight calculations based on *I*(0)/*c* values from the D22 Guinier fits, the *I*(0)/*c*

values were determined to be 0.28 ± 0.02 for sFI and 0.23 ± 0.03 for rFI on the basis of absorbance coefficients (280 nm, 1%, 1 cm) of 14 and 12, respectively (*9*). The 18% reduction in the D22 $I(0)/c$ value for rFI confirmed the expected 13% reduction in the molecular weight of rFI when the complex-type oligosaccharides were replaced by high-mannose-type ones. The mean neutron $I(0)/c$ value from the LOQ Guinier fits was 0.076 ± 0.008 for sFI and rFI relative to a polymer standard. This value corresponds to the molecular weight range of 74 500 for rFI and 85 300 for sFI by comparison with $I(0)/c$ values measured on LOQ for nine other proteins with molecular weights of 27000−254000. It is concluded that the neutron $R_G$ values validate the X-ray $R_G$ values, and that both sFI and rFI are monomeric in solution with molecular weights that are expected from their carbohydrate content.

As factor I has an elongated structure (*6*, *11*), cross-sectional analyses were performed to provide information on the mean dimensions of the two shorter axes of factor I (*26*). Linear cross-sectional X-ray and neutron Guinier $R_{XS}$ plots were obtained for both sFI and rFI in an acceptable $QR_{XS}$ range of 0.5−1.2 (Figure 3c,d). The decrease in $R_{XS}$ on going from X-rays to neutrons is consistent with the observation of both a dry structure by neutron scattering and a small contrast effect. Both the X-ray and neutron $R_{XS}$ values (Table 1) were 0.23−0.29 nm larger for sFI compared to those for rFI. This difference can be attributed to the altered oligosaccharide structures of sFI and rFI on the basis that these are predominantly located on the two shorter axes of factor I. The antennae in the branches of a typical complex-type structure ($NeuNAc_4Gal_4Man_3GlcNAc_6$) are two residues longer than those in a high-mannose-type one ($Man_7GlcNAc_2$) and contain extended $\beta(1,4)$ linkages instead of sterically bent $\alpha(1,2)$ linkages. Molecular graphics showed that the maximal dimensions of this complex-type structure are 2.8 nm × 4.2 nm, while those for the high-mannose-type structure are 2.4 nm × 1.7 nm.

The distance distribution functions $P(r)$ calculated from the entire scattering curves $I(Q)$ up to a $Q$ of 2 $nm^{-1}$ confirmed the Guinier analyses at low $Q$ values (Figure 4). The $P(r)$ analyses resulted in similar $R_G$ values of 4.08−4.35 nm. The most frequently occurring interatomic vector in sFI and rFI corresponds to the peak maximum $M$ of the $P(r)$ curve. The mean values of $M$ for the X-ray $P(r)$ curves of sFI and rFI were 3.9 ± 0.2 and 3.7 ± 0.2 nm, respectively. Those for the neutron $P(r)$ curves were smaller with values of 3.5 ± 0.2 and 2.8 ± 0.2 nm, respectively. In parallel with the slightly smaller $M$ values for rFI, both $P(r)$ curves in Figure 4 showed greater intensities at low $r$ values between 0 and 3 nm for rFI than for sFI. The greater proportion of short interatomic vectors in rFI compared to that in sFI is consistent with the shorter oligosaccharide structures present in rFI. The slight decrease in $M$ on going from X-rays to neutrons is consistent with both a change to a dry structure by neutron scattering and a small contrast effect.

The distance distribution functions $P(r)$ also provide the length $L$ of sFI and rFI. The best $P(r)$ analyses suggested that the maximum dimension $L$ for both sFI and rFI was 14 nm. Further sets of $L$ values were calculated from the Guinier analyses (*6*). From the X-ray $R_G$ and $R_{XS}$ values, $L$ was 14.2 ± 0.6 nm, and from the neutron $R_G$ and $R_{XS}$ values, $L$ was 13.5 ± 0.7 nm, both of which were in good agreement

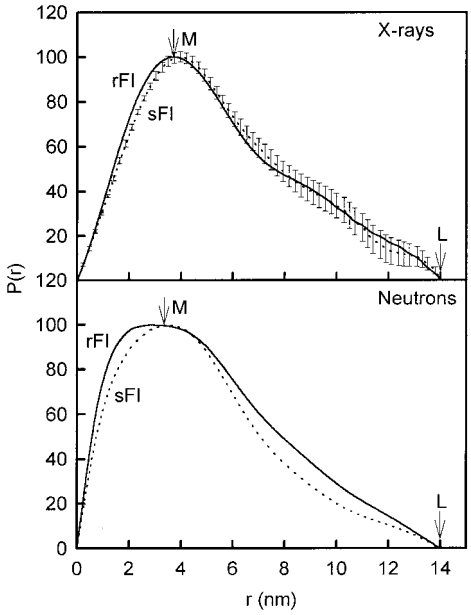


FIGURE 4: X-ray and neutron distance distribution functions $P(r)$ for sFI and rFI. The four $P(r)$ curves were calculated from the $I(Q)$ curves used in Figure 3. The dotted X-ray and neutron $P(r)$ curves correspond to sFI data. The maximum of the $P(r)$ curve is denoted by $M$, the most frequently occurring distance within sFI and rFI, and the maximum dimension is denoted by $L$. Representative error bars are shown for the X-ray $P(r)$ curve for sFI.

with the $P(r)$ analyses. This study is consistent with the previous neutron contrast variation study of sFI (*6*). The present $R_G$ values of 4.05 nm from Guinier analyses are slightly higher than the previous values, but they are now consistent with the present and previous $P(r)$ analyses.

*(b) Homology Modeling for the FIMAC Domain.* The FIMAC domain was modeled on the basis of a follistatin domain in the recent crystal structure of SPARC (*19*), since previously a distant sequence relationship had been noted between the follistatin and FIMAC sequences (*20*). The sequence identity is 27% in Figure 2. Evidence supporting a structural relationship with this crystal structure was obtained as follows.

(i) The follistatin structure is a hybrid of an N-terminal epidermal growth factor (EGF) domain with a C-terminal ovomucoid domain. The disulfide bridge pattern in the EGF domain is 1−3, 2−4, and 5−6, the first two of which occur in the follistatin structure. Sequence comparisons had already shown that this region was very similar to the EGF domain (*4*).

(ii) The consensus secondary structure prediction from 52 follistatin and FIMAC sequences gave a $\beta\beta\beta\beta\alpha\beta$ pattern (*20*). This was in agreement with the follistatin crystal structure, being 81% accurate on a residue-by-residue basis. The consensus predictions made from each of the follistatin and FIMAC sequences were also similar, in particular for the C-terminal $\beta\beta\alpha\beta$ motif corresponding to the ovomucoid domain.

(iii) Construction of the FIMAC model involved only deletions at three surface loops in the follistatin crystal structure without disruption of the structural core (Figure 2). The $\beta$-hairpin between the $\beta$-strands B1 and B2 is shortened and the sequence PIG removed to give a loop that is the same length as that in ovomucoid, and the extra $\alpha$-helix A2 containing the $Cu^{2+}$ binding site in SPARC could be

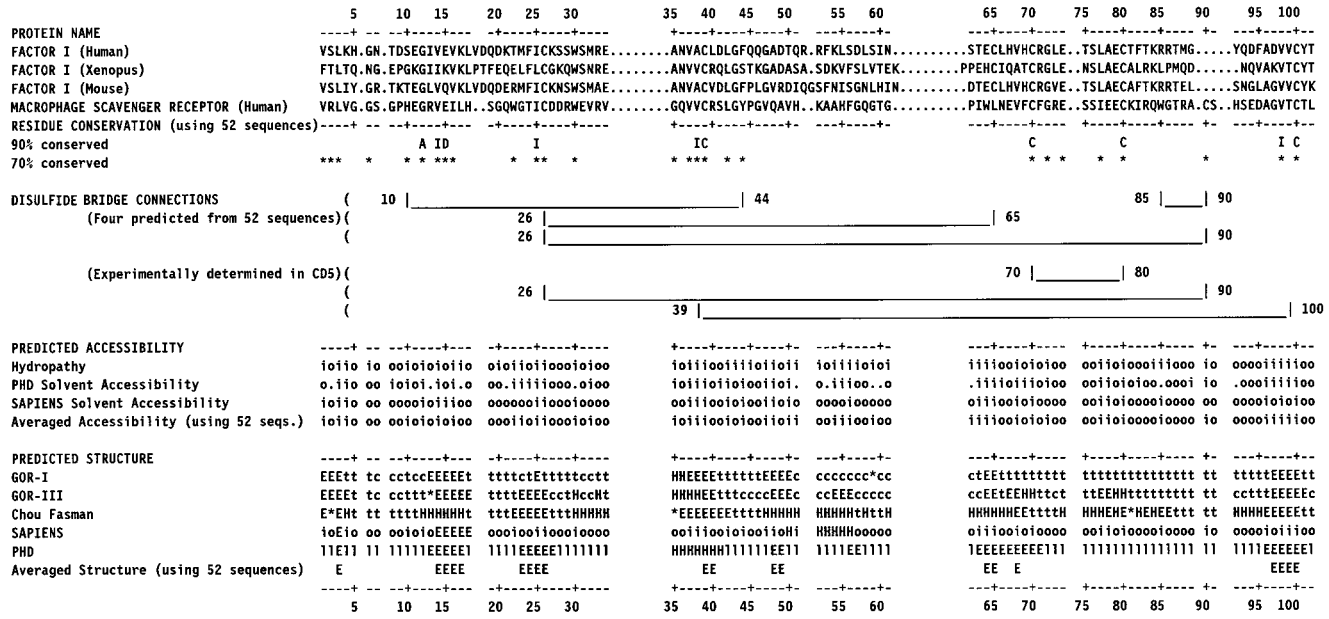


FIGURE 5: Summary of multiple sequence alignment and secondary structure analysis of 52 CD5 sequences. The full version is provided as Supporting Information, in which the sequences are identified by their accession codes on the left. The alignment shows the consensus sequence length of 102 residues which is conserved in more than 50% of the sequences, in which 10 residue types that show a 90% conservation level are identified, and 27 that show a 70% conservation level are marked with an asterisk. The outcome of three averaged hydropathy and/or accessibility and five averaged secondary structure predictions is summarized below the alignment, together with their consensus. The hydropathy and/or accessibility predictions are reported as solvent accessible (o) and solvent inaccessible (i). The consensus secondary structure prediction is based on the presence of α-helix (H) or $\beta$-strand (E) in at least three of the five predictions at each residue position. Other abbreviations are as follows: l, loop; c, coil; and t, turn.

excised (Figure 2 in ref *20*). The FIMAC model contained two semiconserved N-linked oligosaccharide sites at Asn52 and Asn84 in human, mouse, and *Xenopus* factor I, both of which were solvent-exposed as shown by accessibilities of 80% and 50% as required to support this modeling.

*(c) Modeling of the CD5 Domain*. No atomic structure is presently known for the CD5 domain. As the disulfide bridge connectivity of factor I is important for modeling its structure (see below), a multiple sequence alignment of 52 CD5 sequences was constructed to investigate this within CD5 (Figure 5; see the Supporting Information for the full alignment). The CD5 consensus length was found to be 102 residues, and an alignment was readily obtained (*38*, *39*). In this, 10 residues were conserved in more than 90% of the cases (including four Cys and five other hydrophobic residues), and 27 residues were conserved in more than 70% of the cases. In a typical CD5 sequence, between six and eight Cys residues occur at 10 positions (positions 10, 26, 39, 44, 65, 70, 80, 85, 90, and 100 in Figure 5). The conservation pattern of Cys residues in the sequence alignment showed that Cys10 and Cys44, Cys26 and Cys65, Cys26 and Cys90, and Cys85 and Cys90 were predicted to be bridged, and this left the connectivities of Cys39, Cys70, Cys80, and Cys100 unassigned (Figure 5). The recent experimental determination for the CD5 domain in the macrophage scavenger receptor showed that Cys26 and Cys90, Cys39 and Cys100, and Cys70 and Cys80 were bridged (*23*), and this resolved the unassigned disulfide bridges (Figure 5). In application to factor I, the Cys26−Cys90 bridge is replaced by one predicted between Cys26 and Cys65 because of a deletion at residues 90 and 91. The presence of this bridge in factor I is consistent with the predominantly hydrophobic nature of residue 65 found in the alignment as the result of performing three different accessibility predictions (Figure 5) (*20*, *33*). In the numbering of Figure 2, the disulfide bridges in factor I correspond to Cys136 and Cys196, Cys123 and Cys163, and Cys168 and Cys178.

Scattering curve modeling was based on the dimensions of the CD5 domain of 5.4 ± 1.0 nm × 3.5 ± 0.9 nm from electron microscopy (*23*). To identify a suitable known structure that was similar in size and could be used in scattering modeling, structure predictions were performed. Its consensus predicted secondary structure from 52 sequences using five different averaging methods (*20*, *33*) gave a content of 19% $\beta$-sheet with six $\beta$-strands and no α-helix (Figure 5). The protein fold recognition program THREADER was used to score the 52 CD5 sequences for their compatibility with 254 known protein folds. No strong matches were identified; however, the five best-scoring folds all possessed predominantly $\beta$-sheet structures. As both analyses favored $\beta$-sheet proteins, we decided to use an immunoglobulin fold structure with the same dimensions (4.4 nm × 3.4 nm) and the same volume as the CD5 domain for scattering curve modeling (*23*, *36*). A recent predicted model for this domain was based on an immunoglobulin fold, although this identification is not unequivocal at present (*39*, *40*).

*(d) Homology Modeling of the LDLr Domains*. The LDLr-1 and LDLr-2 domains in factor I were modeled on the basis of a multiple sequence alignment and one crystal and two NMR structures (*21*, *33*, *41*, *42*). The three LDLr disulfide bridges create a compact fold which is stabilized by $Ca^{2+}$ in the crystal structure. In factor I, amino acid changes in LDLr-1/2 make it unlikely that $Ca^{2+}$ binds, in agreement with functional studies (*43*). In LDLr-1, even though four acidic $Ca^{2+}$-binding residues and Asp231 are conserved, Glu215 is missing and two conserved Cys

residues are missing from positions 204 and 216 (Figure 2). In LDLr-2, a $Ca^{2+}$-binding residue is replaced by Asn261, while Glu252 and Asp268 are missing (Figure 2). Nonetheless, since the sequence length of the LDLr crystal structure coincided exactly with each of LDLr-1 and LDLr-2 in factor I, and the sequence identities with the crystal structure were high at 38 and 45%, respectively, homology models were readily constructed. The LDLr-2 model predicted 100% solvent accessibility of the N-linked oligosaccharide site at Asn244 in mouse factor I, as required (Lys244 in Figure 2). In the LDLr-1 model, it was notable that the solvent accessibilities of the adjacent Cys237 and Cys238 residues were 100 and 10%, respectively. Cys237 clearly protruded from the surface of the LDLr-1 domain and was unbridged, while Cys238 formed a buried internal link with Cys223. It was also notable that the absence of linker residues between the LDLr-1 and LDLr-2 domains meant that an extended domain arrangement had to be modeled, as otherwise a bent arrangement would result in steric conflicts between these two domains.

*(e) Homology Modeling of the SP Domain.* The SP domain in factor I contains a pair of four-stranded $\beta$-sheet Greek key motifs, each adjacent to a $\beta$-hairpin with two $\beta$-strands (*34*). $\beta$-Strands B−D and F form the first Greek key with $\beta$-hairpin G−H, followed by $\beta$-strands J−M and $\beta$-hairpin N−O, respectively (Figure 2). Its modeling was based on the crystal structure of the urokinase-type plasminogen activator (*22*). This has the same six disulfide bridges as those in factor I, and its observed secondary structure calculated from the coordinates (*42*) contained 12 $\beta$-strands that agreed with those predicted for factor I. The alignment of Figure 2 showed 34% sequence identity, and homology modeling showed that all the main insertions or deletions occurred in loop regions and could be readily accommodated (Figure 2). The validity of the model was supported by the high solvent accessibilities of 60, 50, and 70% at the putative N-linked oligosaccharide sites at Asn446, Asn476, and Asn528, respectively. All three Asn residues were located on the face of the SP model opposite that containing the catalytic triad of His362, Asp411, and Ser507, and all three Asn residues were close to the attachment point of the SP domain with the heavy chain via Cys309−Cys435.

*(f) Constrained Scattering Curve Modeling of the Domain Arrangement in Factor I.* The domain arrangement of sFI and rFI in solution was assessed using curve fits of the X-ray and neutron data that were constrained by the above models for the five domains. Initially, an extended five-domain linear arrangement was investigated, in which the long axes of each domain were aligned on a common axis. Its overall length was 18 nm which is larger than the observed value of 14 nm (Figure 4). This linear model gave an $R_G$ value of 6.14 nm and an $R_{XS}$ value of 0.88 nm which deviated widely from the observed values of 4.04 and 1.70 nm, respectively (Table 1). The curve calculated from this model gave a poor fit to the experimental X-ray data at $Q$ values above 0.5 $nm^{-1}$, and the $R$-factor was high at 16.1% (Figure 6 and Table 1). Efforts to improve this by using partly bent, half-bent, and fully bent models likewise gave different $R_G$ and $R_{XS}$ values, poor curve fits at $Q$ values above 0.5 $nm^{-1}$, and high $R$-factor values (Figure 6 and Table 1). This showed that structures with extended domain arrangements similar to the linear or bent four-domain structures found in the multidomain serine proteases factor VIIa and IXa of blood coagulation were not appropriate for factor I (*45*, *46*).

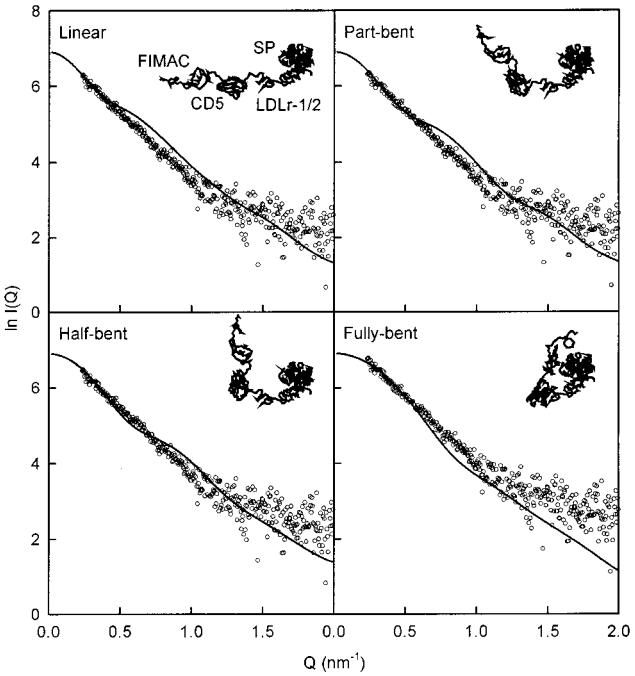


FIGURE 6: Comparison of the calculated and experimental wide-angle scattering curves $I(Q)$ for inappropriate linear, partly bent, half-bent, and fully bent extended domain models for sFI. The continuous lines represent the curves calculated from the model, and the points correspond to experimental data. Inside each panel, a schematic α-carbon view of each domain structure is shown, in all four of which the SP domain is fixed in position.

A bilobal structure had been observed in factor I by electron microscopy, although it had not been explained previously how such a structure could be formed (*11*). Here, the present homology modeling and sequence analyses for the FIMAC, CD5, LDLr, and SP domains showed that 19 disulfide bridges could be identified, and accounted for 38 of the 40 Cys residues in factor I (Figure 1). The two Cys residues remaining were Cys15 and Cys237, where Cys15 is N-terminal to the FIMAC domain, and Cys237 was found to be surface-exposed at the C terminus of the LDLr-1 model. Experimentally, the strong inhibition of factor I by dithiothreitol or 2-merceptoethanol showed that the integrity of all 20 disulfide bridges was required for activity (*43*). Here, the absence of free Cys residues in sFI was further confirmed by two experiments based on Ellman's reagent and radiolabeled iodoacetamide (Materials and Methods). It was accordingly considered plausible that Cys15 and Cys237 were disulfide bridged to create a triangular domain structure for the FIMAC, CD5, and LDLr-1 domains (Figure 1). The existence of this triangular domain structure together with the large SP domain would account for the bilobal structure seen by electron microscopy for sFI. Molecular graphics showed that such a model was readily constructed from the FIMAC, CD5, and LDLr-1 models (Materials and Methods). The formation of the model was stereochemically constrained by the predicted Cys15−Cys237 bridge, the polypeptide links between the four domains, the retention of solvent-exposed glycosylation sites at Asn52, Asn85, and Asn244, and the surface attachment of the LDLr-2 domain. These constraints permitted few structural variants in the construction of this triangular model, and this meant that the relative positions

of these domains were defined within the structural resolution of solution scattering.

The creation of the triangular and SP models now permitted the molecular modeling of a bilobal structure for sFI and rFI by a consideration of the possible arrangements between the two models, using automated curve-fit searches. The modeling tested whether better curve fits could be obtained compared to those based on the extended domain structures in Figure 6. The searches were simplified to test the major features of this bilobal model. The orientations of each of the triangular and SP models were represented by 90° rotational steps to correspond to the six sides of a cube with six different faces. This meant that $6 \times 6$ relative arrangements were possible, making a total of 36, within which four sets of four arrangements were identical for reason of symmetry (24 unique arrangements). Four more sets of four arrangements potentially correspond to similar structures if there is an approximate symmetry axis within each of the triangular and SP models in which four cube faces are similar to each other. In the triangular model, this would correspond to its longest axis. In the SP model, this corresponds to the grouping of three oligosaccharide chains on one side of a nearly spherical structure. This means that as few as 12 searches were required. It was found that a comprehensive test of bilobal models could be performed on the basis of 16 different relative 90° rotations of the two models using translational searches (Materials and Methods) with redundancy as a control. In each of these 16 searches, the SP model in a series of fixed orientations was translated in 20 nm $\times$ 30 nm $\times$ 0.5 nm steps in a common plane relative to the triangular model which was held fixed in a series of fixed orientations, and this generated 600 bilobal models per search. This tested a total of 9600 different bilobal arrangements in all.

Each of the 16 curve-fit searches for sFI was evaluated using contour maps that showed the distribution in the translational search of the number of spheres in each of the 600 models, their $R_G$ and $R_{XS}$ values, and the *R*-factor (Figure 7). The use of cutoff filters for the number of spheres in each model and the $R_G$ and $R_{XS}$ values retained only the good curve-fit structures (Materials and Methods). This showed that 12 searches gave between 10 and 22 solutions (2−4% of the total), while four searches gave no solutions. These solutions (●) were all clustered together in the region of the lowest *R*-factors in the contour maps (see the fourth panel of Figure 7). This showed that the searches had produced sensible outcomes. Interestingly, even though the 16 searches involved independent sets of translations and orientations, all the solutions from the 16 searches corresponded to a single structural family in which the SP domain was positioned close to the C terminus of the LDLr-2 domain, although the relative orientations of the triangular and SP models were different in each search. This proximity is consistent with the covalent connection between the LDLr-2 and SP domains (Figure 1). There was no evidence for any other minima. From this single family of structures, a best-fit bilobal model is presented in Figure 8 which corresponded to the smallest *R*-factor of 10.2% in the search of Figure 7 (marked with an arrow). This model gave a good curve fit over a *Q* range of up to 1.5 $nm^{-1}$ (top curve of Figure 8), which was visually much improved compared to the four fits for extended domain models in Figure 6 that only went

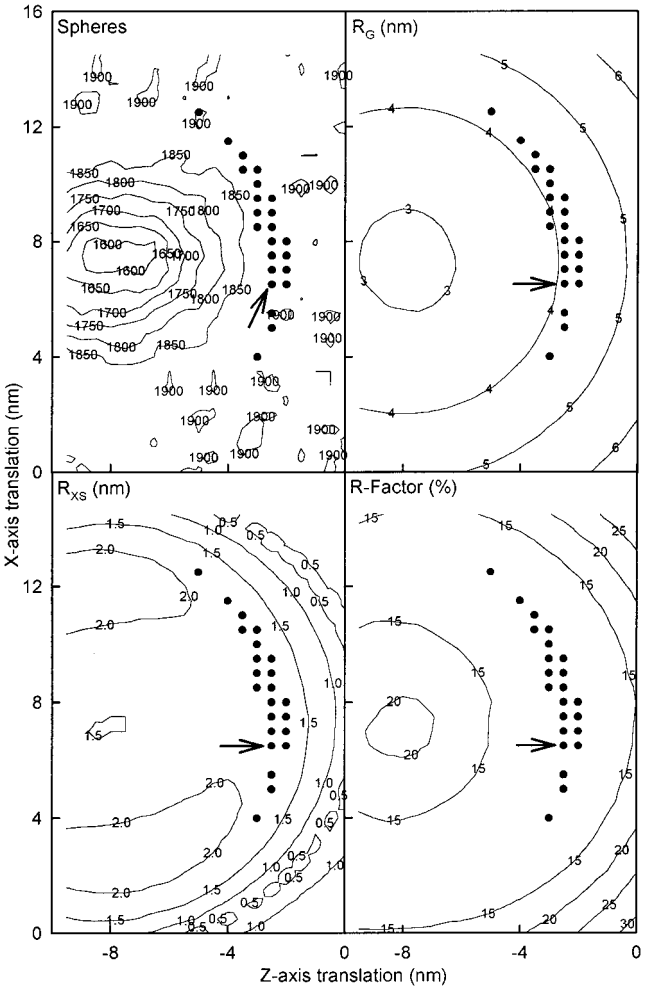


FIGURE 7: Contour maps from a curve-fit search to determine the domain structure in sFI. The contours correspond to the number of spheres and the $R_G$, $R_{XS}$, and *R*-factor values calculated for the 600 models in an automated search in the $X-Z$ plane. In the spheres panel, the pronounced minimum corresponds to the complete overlap of the triangular and SP models. The 22 good-fit models are denoted by ● to show the positions of the SP domain relative to the triangular domain after the use of filters. The best-fit model shown in Figure 8 is marked with an arrow.

up to 0.5 $nm^{-1}$. The *R*-factors were at least 2% smaller than those for the extended domain models (Table 1). The separation between the centers of mass of the heavy chain and SP domain models in sFI was 5.9 nm, and its longest dimension was 13 nm.

Significant differences are visible between the X-ray and neutron curves in Figure 8 because of the different instrumental geometries in use and the observation of hydrated and unhydrated structures by X-ray and neutron scattering, respectively. The large effect of this is represented by the dotted lines in Figure 8. In addition, the scattering curve of rFI is distinct from that for sFI in that an inflection is observed at 0.9 $nm^{-1}$ in the rFI curve (marked with an arrow in Figure 8) that is less pronounced in the sFI curve, which can be attributed to the much altered oligosaccharide structures present in rFI (*9*). The consequences of these differences were explored in three further searches using the neutron curve of sFI and the X-ray and neutron curves of rFI and the above 9600 models. In the neutron searches, unhydrated models were used. In the rFI searches, the oligosaccharide structures were converted to high-mannose-

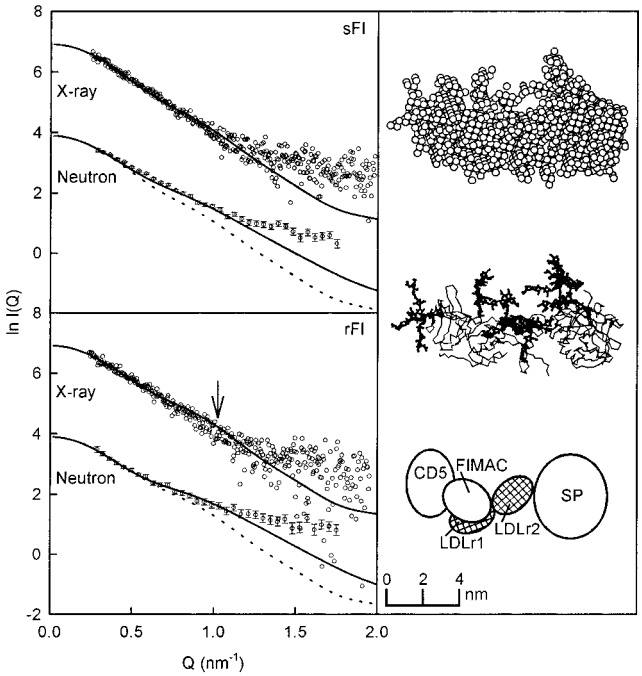


FIGURE 8: Comparison of the calculated and experimental wide-angle scattering curves $I(Q)$ for sFI and rFI. For sFI and rFI, the continuous lines represent the curves calculated from the best-fit modeled structure, while the points correspond to the X-ray and neutron experimental data as indicated. The dashed line shown with each neutron curve corresponds to the modeled X-ray curve to show the joint effect of the corrections for hydration and instrumental geometry. The pronounced inflection in the rFI curves is marked with an arrow. For the best-fit sFI structure, the sphere model, an α-carbon view, and a schematic domain outline are shown. Outline oligosaccharide structures are bold in the α-carbon view.

type (Materials and Methods). Despite these changes, all three further searches resulted in good curve fits with similar *R*-factors of 10.2, 11.4, and 10.0% (Table 1). The best-fit models from all three searches again placed the LDLr-2 and SP domains in proximity to each other like the first search, within a positional precision of 2.7 nm in the $X-Z$ plane. The best rFI model gave an unchanged X-ray $R_G$ value when compared with the best sFI model, while its X-ray $R_{XS}$ value was reduced by 0.24 nm to 1.61 nm, in agreement with the experimental value of 1.57 nm (Table 1). This difference is explained by the change from complex-type to high-mannose-type carbohydrate. Both neutron models for sFI and rFI gave reduced $R_{XS}$ values that were 0.36−0.41 nm lower than those of the X-ray models, in good agreement with the 0.19−0.35 nm smaller $R_{XS}$ values seen experimentally. This difference is explained by the observation of hydrated structures in X-ray scattering and unhydrated structures in neutron scattering. Since all four different modeling analyses consistently resulted in the same bilobal structure for factor I, this supports the conclusion that factor I possesses a bilobal domain structure in solution, as illustrated in Figure 8.

## CONCLUSIONS

The combination of neutron and X-ray scattering data, correctly sized domain structures from homology modeling, and constrained scattering curve fit analyses has revealed new insights into the domain arrangement of the multidomain protein factor I. Unlike the majority of the multidomain plasma serine proteases of the complement, coagulation, and fibrinolysis cascades, the five factor I domains do not form an extended linear arrangement, and this hypothesis was ruled out by scattering curve fits (Figure 6). Instead, good curve fits were obtained on the basis of a bilobal structure for factor I that was derived by the consideration of homology models. The combination of multiple sequence alignments for the four domain superfamilies in factor I and homology modeling for three of them identified 19 disulfide bridges from the 40 Cys residues present in factor I (*20*, *33*, *34*; Figure 5). The homology modeling is supported by the high sequence identities of 27−45% in the alignments used for the FIMAC, LDLr, and SP models (*50*). These analyses suggested that Cys15 and Cys237 were unpaired, yet no free Cys residues could be detected in our biochemical assays of factor I. The combination of this result with the earlier observation of a bilobal structure in factor I by electron microscopy (*11*) suggested that a predicted Cys15−Cys237 disulfide bridge would lead to the formation of a compact triangular assembly of the FIMAC, CD5, and LDLr-1 domains as one of the two lobes in the structure of factor I (Figure 1). The mean diameter of the SP model is 5.3 nm while that of the triangular model 4.6 nm, and both values agree well with the diameters of 5.4 and 4.9 nm for the two globular structures observed by electron microscopy (*11*). Four experimental tests of this bilobal structural model resulted in satisfactory X-ray and neutron scattering curve fits for both rFI and sFI, despite the four different sets of physical conditions, and these provide support for the bilobal model, even though strictly it is not possible to derive unique structures from scattering measurements. In conclusion, the present scattering analyses resulted in a plausible model for the domain arrangement in factor I that can now be used for the rational planning of future work, in particular disulfide bridge mapping to test the predicted Cys15−Cys237 bridge, and the synthesis of smaller fragments of factor I for structural studies. Already, on the basis of these structural analyses, we have been able to express a FIMAC domain (*51*).

The multiple sequence alignments for the four domain types in factor I (*20*, *33*, *34*; Figure 5) provide complementary information that is consistent with the scattering analyses. Summation of the Asp, Glu, Lys, and Arg residues in the three FIMAC domains of human, mouse, and *Xenopus* factor I shows that these are basic (net charge of +4 to +7). The corresponding summation for the LDLr-1 and LDLr-2 domains in factor I shows that these are acidic (net charge of −2 to −5 each). The CD5 and SP domains exhibit variable net charges but are close to neutral. The calculation of electrostatic maps using DELPHI for the homology models of the FIMAC and LDLr-1/2 domains showed that their surfaces were almost entirely basic and acidic, respectively. The opposite charges on the FIMAC and LDLr domains would enable them to attract each other, and this would be consistent with the domain arrangement in factor I shown in Figures 1 and 8.

The proposed domain model of factor I provides insights into the roles of (i) the CD5, LDLr-1, and SP domains and (ii) the oligosaccharide chains in the function of factor I. Factor I cleaves C3b or C4b at two or three sites in the presence of a cofactor, which is either the soluble proteins factor H or C4b binding protein, or the membrane-bound

proteins complement receptor type 1 or membrane cofactor protein. These cleavages control the activities of the convertase enzymes C3bBb and C4b2a within which C3b and C4b are incorporated (*1*, *2*). By analogy with the tissue factor−factor VIIa complex of blood coagulation (*46*), the cofactors may induce a conformational change in C3b, C4b, or factor I or provide a binding site to orient the SP domain toward its substrate. It is known that factor I interacts with low affinity with either factor H or C3($NH_3$) in the absence of the third member of the complex at nonoverlapping sites; however, the affinity is increased when all three components are present, and there is evidence for the binding of factor H to the heavy chain of factor I (*47*). The present molecular identification of two globular entities in a bilobal structure is consistent with these biochemical data in the sense that the triangular lobe model may interact with the cofactor while the SP lobe is able to interact with the substrate. The $R_{XS}$ and $P(r)$ analyses showed differences between sFI and rFI which can be attributed to variable carbohydrate structures at the center of factor I (*10*), and this was supported by the scattering modeling of the oligosaccharide chains at the center of factor I (Figure 8). The oligosaccharides may contribute to the structure of the interlobal region in factor I (*48*), and evidence to support this is suggested by the 55% activity of rFI when compared with that of sFI (*9*). Such a central location for the oligosaccharides implies that protein surfaces will be left exposed for functional interactions. Figure 8 suggests that these will primarily involve regions surrounding the catalytic triad on the SP domain and most of the CD5 and LDLr-1 domains in the triangular region. As the pH optimum for factor I function is maximal at pH 4−6 and decreased sharply at pH 6−7, His residues have been implicated in the protein−protein interactions involved in factor I-mediated cleavage (*47*). Interestingly, summations of the His residues in the domains of factor I showed that these are most abundant in the SP domain. Given the ionic strength dependence of the factor I-mediated cleavage of C3b (*47*), it is interesting that there are as many as nine conserved charged groups in well-aligned regions of the three CD5 sequences from factor I, several of which are not conserved in the remainder of the CD5 superfamily (Figure 5).

The domain modeling also provides functional insight about the FIMAC domain in factor I. Follistatin domains in connective tissues can act as spacers (*19*, *49)*. The molecular domain arrangement in factor I suggests that this may constitute the most likely function of the FIMAC domain to link the CD5 and LDLr-1 domains. An alternative viewpoint is a proposal that the FIMAC domain may act as a protease inhibitor in view of its structural similarity with ovomucoid. Despite the intriguing presence of the SP domain in factor I (*19*, *20*, *49*), this is considered to be unlikely from the scattering curve fits. The molecular arrangement of the domains in the best curve-fit factor I model places the FIMAC and SP domains far apart (Figure 8). If the FIMAC and SP domains were placed in proximity to each other, poor curve fits were obtained (Figure 6). Inhibition has not been detected in functional assays using osteonectin or factor I (*19*, *49*; C. G. Ullman and S. J. Perkins, unpublished results). In addition, the FIMAC domain is masked by two oligosaccharide sites and the LDLr-1 and LDLr-2 domains in the best-fit model, and this would hinder possible interaction with the SP domain. In fact, one of these two glycosylation sites is located in the ovomucoid-like scissile loop of the homology model for the FIMAC domain that would block its interaction with a potential target if it were to be a protease inhibitor.

## ACKNOWLEDGMENT

We thank Dr. R. B. Sim for useful discussions, the BBSRC for an Earmarked Studentship (D.C.), and the Wellcome Trust for grant support (C.G.U.). We also thank the Synchrotron Radiation Source, the Rutherford-Appleton Laboratory, and the Institut-Laue-Langevin for access to X-ray and neutron facilities and Dr. E. Towns-Andrews, Mrs. S. Slawson, and Mr. A. Gleeson (SRS), Dr. R. K. Heenan and Dr. S. M. King (ISIS), and Dr. P. A. Timmins (ILL) for invaluable instrumental support.

## SUPPORTING INFORMATION AVAILABLE

Full version of Figure 5 (1 page). Ordering information is given on any current masthead page.

## REFERENCES

1. Sim, R. B., Day, A. J., Moffatt, B. E., and Fontaine, M. (1993) *Methods Enzymol. 223*, 13−35.
2. Law, S. K. A., and Reid, K. B. M. (1995) *Complement*, 2nd ed., IRL Press, Oxford, U.K.
3. Vyse, T. J., Morley, B. J., Bartók, I., Theodoridis, E. L., Davies, K. A., Webster, D. B., and Walport, M. J. (1996) *J. Clin. Invest. 97*, 925−933.
4. Catterall, C. F., Lyons, A., Sim, R. B., Day, A. J., and Harris, T. J. R. (1987) *Biochem. J. 242*, 849−856.
5. Goldberger, G., Bruns, G. A. P., Rits, M., Edge, M. D., and Kwiatkowski, D. J. (1987) *J. Biol. Chem. 262*, 10065−10071.
6. Perkins, S. J., Smith, K. F., and Sim, R. B. (1993) *Biochem. J. 295*, 101−108.
7. Kunnath-Muglia, L. M., Chang, G. H., Sim, R. B., Day, A. J., and Ezekowitz, R. A. (1993) *Mol. Immunol. 30*, 1249−1256.
8. Minta, J. O., Wong, M. J., Kozak, C. A., Kunnath-Muglia, L. M., and Goldberger, G. (1996) *Mol. Immunol. 33*, 101−112.
9. Ullman, C. G., Chamberlain, D., Ansari, A., Emery, V. C., Haris, P. I., Sim, R. B., and Perkins, S. J. (1998) *Mol. Immunol.* (in press).
10. Jarvis, D. J., and Finn, E. E. (1995) *Virology 212*, 500−511.
11. DiScipio, R. G. (1992) *J. Immunol. 149*, 2592−2599.
12. Smith, K. F., Harrison, R. A., and Perkins, S. J. (1990) *Biochem. J. 267*, 203−212.
13. Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E., and Sim, R. B. (1993) *Biochem. J. 295*, 87−99.
14. Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys, M. B., and Perkins, S. J. (1997) *J. Mol. Biol. 272*, 408−422.
15. Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A., and Perkins, S. J. (1995) *Biochem. J. 311*, 283−291.
16. Beavil, A. J., Young, R. J., Sutton, B. J., and Perkins, S. J. (1995) *Biochemistry 34*, 14449−14461.
17. Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A., and Perkins, S. J. (1996) *J. Mol. Biol. 259*, 718−736.
18. Perkins, S. J., Ashton, A. W., Boehm, M. K., and Chamberlain, D. (1998) *Int. J. Biol. Macromol. 22*, 1−16.
19. Hohenester, E., Maurer, P., and Timpl, R. (1997) *EMBO J. 16*, 3778−3786.
20. Ullman, C. G., and Perkins, S. J. (1997) *Biochem. J. 326*, 939−941.
21. Fass, D., Blacklow, S., Kim, P. S., and Berger, J. M. (1997) *Nature 388*, 691−693.

22. Spraggon, G., Phillips, C., Nowak, U. K., Ponting, C. P., Saunders, D., Dobson, C. M., Stuart, D. I., and Jones, E. Y. (1995) *Structure 3*, 681−691.
23. Resnick, D., Chatterton, J. E., Schwatz, K., Slayter, H., and Krieger, M. (1996) *J. Biol. Chem. 271*, 26924−26930.
24. Lindner, P., May, R. P., and Timmins, P. A. (1992) *Physica B 180*, 967−972.
25. Glatter, O., and Kratky, O., Eds. (1982) *Small-angle X-ray Scattering*, Academic Press, New York.
26. Hjelm, R. P., Jr. (1985) *J. Appl. Crystallogr. 18*, 452−460.
27. Semenyuk, A. V., and Svergun, D. I. (1991) *J. Appl. Crystallogr. 24*, 537−540.
28. Hobohm, U., and Sander, C. (1994) *Protein Sci. 3*, 522−524.
29. Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992) *Protein Sci. 1*, 409−417.
30. Laskowski, R. A., McArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *J. Appl. Crystallogr. 26*, 283−291.
31. Lee, B., and Richards, F. M. (1971) *J. Mol. Biol. 55*, 379−400.
32. Šali, A., and Blundell, T. L. (1990) *J. Mol. Biol. 212*, 403−428.
33. Ullman, C. G., Haris, P. I., Smith, K. F., Sim, R. B., Emery, V. C., and Perkins, S. J. (1995) *FEBS Lett. 371*, 199−203.
34. Perkins, S. J., and Smith, K. F. (1993) *Biochem. J. 295*, 109−114.
35. Deisenhofer, J. (1981) *Biochemistry 20*, 2361−2370.
36. Perkins, S. J. (1986) *Eur. J. Biochem. 157*, 169−180.
37. Perkins, S. J., and Weiss, H. (1983) *J. Mol. Biol. 168*, 847−866.
38. Resnick, D., Pearson, A., and Krieger, M. (1994) *Trends Biochem. Sci. 19*, 5−8.
39. Aruffo, A., Bowen, M. A., Patel, D. D., Haynes, B. F., Starling, G. C., Gebe, J. A., and Bajorath, J. (1997) *Immunol. Today 18*, 498−504.
40. Williams, A. F. (1987) *Immunol. Today 8*, 298−303.
41. Daly, N. L., Djordjevic, J. T., Kroon, P. A., and Smith, R. (1995) *Biochemistry 34*, 14474−14481.
42. Daly, N. L., Scanlon, M. J., Djordjevic, J. T., Kroon, P. A., and Smith, R. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 6334−6338.
43. Crossley, L. G. (1980) *Methods Enzymol. 80*, 112−124.
44. Kabsch, W., and Sander, C. (1983) *Biopolymers 22*, 2577−2637.
45. Brandstetter, H., Bauer, M., Huber, R., Lollar, P., and Bode, W. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 9796−9800.
46. Banner, D. W., D'Arcy, A., Chene, C., Winkler, F. D., Guha, A., Konigsberg, W. H., Nemerson, Y., and Kirchofer, D. (1996) *Nature 380*, 41−46.
47. Soames, C. J., and Sim, R. B. (1997) *Biochem. J. 326*, 553−561.
48. Varki, A. (1993) *Glycobiology 3*, 97−130.
49. Lane, T. F., and Sage, E. H. (1994) *FASEB J. 8*, 163−173.
50. Martin, A. C. R., MacArthur, M. W., and Thornton, J. M. (1997) *Proteins: Struct., Funct., Genet. Suppl. 1*, 14−28.
51. Hinshelwood, J., Ullman, C. G., Chamberlain, D., and Perkins, S. J. (1998) *Mol. Immunol.* (in press).

BI9805184