β -Sheet secondary structure of the trimeric globular domain of C1q of complement and collagen types VIII and X by Fourier-transform infrared spectroscopy and averaged structure predictions

Kathryn F. SMITH,* Parvez I. HARIS,† Dennis CHAPMAN,† Kenneth B. M. REID‡ and Stephen J. PERKINS*§

*Department of Biochemistry and Chemistry and †Department of Protein and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K. and ‡MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

C1q plays a key role in the recognition of immune complexes, thereby initiating the classical pathway of complement activation. Although the triple-helix conformation of its N-terminal segment is well established, the secondary structure of the trimeric globular C-terminal domain is as yet unknown. The secondary structures of human C1q and C1q stalks and pepsin-extracted human collagen types I, III and IV (with no significant non-collagen-like structure) were studied by Fourier-transform i.r. spectroscopy in $^2\text{H}_2\text{O}$ buffers. After second-derivative calculation to resolve the fine structure of the broad amide I band, the Fourier-transform i.r. spectrum of C1q showed two major bands, one at 1637 cm^{-1} , which is a characteristic frequency for β -sheets, and one at 1661 cm^{-1} . Both major bands were also detected for C1q in H_2O buffers. Only the second major band was observed at 1655 cm^{-1} in pepsin-digested C1q which contains primarily the N-terminal triple-helix region. The Fourier-transform i.r. spectra of collagen

in $^2\text{H}_2\text{O}$ also showed a major band at 1659 cm^{-1} (and minor bands at 1632 cm^{-1} and 1682 cm^{-1}). It is concluded that the C1q globular heads contain primarily β -sheet structures. The C-terminal domains of C1q show approximately 25% sequence identity with the non-collagen-like C-terminal regions of the short-chain collagen types VIII and X. To complement the Fourier-transform-i.r. spectroscopic data, averaged Robson and Chou–Fasman structure predictions on 15 similar sequences for the globular domains of C1q and collagen types VIII and X were performed. These showed a clear pattern of ten β -strands interspersed by β -turns and/or loops. Residues thought to be important for C1q–immune complex interactions with IgG and IgM were predicted to be at a surface-exposed loop. Sequence insertions and deletions, glycosylation sites, the free cysteine residue and RGD recognition sequences were also predicted to be at surface-exposed positions.

INTRODUCTION

The C1 enzyme complex activates the classical pathway of complement (Reid, 1986). It is composed of the subunits C1q, C1r and C1s. Subunits C1r and C1s interact to form a tetramer, C1s-C1r-C1r-C1s, which makes contacts with the collagen-like stalks of C1q. Activation of C1q is initiated by the binding of C1q to immune complexes of IgG or IgM via the globular head domains. This induces a conformational change in the collagenous stalks of C1q, and this change leads to the autoactivation of C1r₂C1s₂ into the active proteinase C1r₂C1s₂ (Schumaker et al., 1986; Heinz, 1989). The activated C1 complex then cleaves components C2 and C4 in the complement cascade.

The structure of C1q has been investigated by electron microscopy and scanning tunnelling microscopy (Knobel et al., 1975; Brodsky-Doyle et al., 1976; Wells et al., 1992), neutron scattering (Perkins et al., 1984; Perkins, 1985) and sedimentation (Liberti and Paul, 1978). These show that C1q has a hexameric structure with an appearance resembling a bunch of tulips (Figure 1). Sequence studies indicate that C1q contains 18 polypeptide chains of three types, $6 \times \text{A}$, $6 \times \text{B}$ and $6 \times \text{C}$ (Reid, 1985; Sellar et al., 1991). Each chain is approximately 226 amino acids in length, and one copy of each chain forms the monomeric unit of C1q. At the N-terminus, the 81 residues in the A-, B- and C-chains form a collagen-like triple helix (Figure 1; Reid, 1986). The triple helix terminates at the C-terminus in six globular heads, each containing 136 amino acids (Sellar et al., 1991).

Although the secondary structure of the C1q collagenous region is well established as a triple helix, no information is as yet

available on the structure of the globular domain of C1q. Previously c.d. has been used to show that C1q has a collagen triple-helix structure, but it was not possible to identify the secondary structure in the C1q globular head, although the c.d. spectrum is not characteristic of the collagen triple helix (Brodsky-Doyle et al., 1976). Fourier-transform i.r. spectroscopy can be effectively used in conjunction with c.d. to investigate protein secondary structures (Haris and Chapman, 1992; Surewicz et al., 1993). In particular, Fourier-transform i.r. spectroscopy produces discrete spectral bands which can be assigned to specific secondary-structure conformations. The spectroscopic analysis can be reinforced by comparison with an averaged secondary-structure prediction calculated from a number of similar sequences. Sequence averaging reduces statistical scatter in the individual predictions and the effect of possible sequence alignment errors (Perkins et al., 1988). Human and mouse A-, B- and C-chains in C1q have now been sequenced. The C-terminal sequences of the globular domains of the family of short-chain collagen types VIII and X and precerebellin are similar in sequence to C1q (Reid and Day, 1990; Urade et al., 1991; Sellar et al., 1991; Brass et al., 1992). In contrast with the heterotrimeric sequence of C1q, collagen type X has three $\alpha 1$ chains, and collagen type VIII has two $\alpha 1$ chains and one $\alpha 2$ chain (LuValle et al., 1988; Mann et al., 1990).

In the present study the secondary structure of the globular domain of C1q is shown to be β -sheet on the basis of a comparative Fourier-transform i.r. spectroscopic analysis of C1q and the C1q stalks with three pepsin-extracted collagen types which are almost entirely triple helix in structure. Pepsin digestion

Abbreviation used: C1q head, globular domain of C1q.

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

Figure 1 Schematic cross-section of the overall morphology of C1q in terms of a globular C-terminal protein domain attached to an N-terminal collagenous stalk

One-third of the hexamer is depicted. Amino acid totals correspond to the average of the A-, B- and C-chains in each monomer of C1q, and carbohydrate residues are not indicated. The triple helix is approximately 81 residues long in the collagenous segment. The globular domain of C1q is also referred to as the head of C1q.

is commonly employed to release collagen molecules from their cross-links in tissues (Hill and Harper, 1984). An averaged Robson and Chou–Fasman secondary-structure prediction algorithm was applied to 15 aligned similar sequences to confirm and extend the analysis of this β -sheet structure. The predicted location of amino acid residues involved in the interaction between C1q and immune complexes, glycosylation sites and free cysteine residues are discussed with respect to their positions within the globular domains of C1q.

This work was presented at the XVth International Complement Workshop, Japan, 27 September–1 October 1993 (Smith et al., 1993).

MATERIALS AND METHODS

Fourier-transform i.r. spectroscopy of C1q, the C1q stalks and collagen

C1q was isolated from human serum by a procedure involving euglobulin precipitation at pH 5.5, affinity chromatography on IgG–Sephacrose and ion-exchange chromatography on CM-cellulose [procedures reviewed by Reid (1981)]. For the spectroscopic analysis, C1q was concentrated to 2.4 mg/ml in PBS (12 mM sodium/potassium phosphate; 140 mM NaCl, pH 7.4) using Amicon concentrators with a PM-10 membrane. Alternatively, it was precipitated at low ionic strength in 10 mM sodium phosphate/10 mM NaCl, pH 7.4, and the precipitate was resuspended in distilled water at a final concentration of 3.0 mg/ml. The concentration of C1q in the suspension was measured by redissolving it in 6 M guanidinium chloride, pH 7.5, and using an A_{280} (1%, 1 cm) of 6.82 at 280 nm.

The collagen-like stalks of C1q (Figure 1) were isolated by digestion of C1q with pepsin (30:1, w/w, substrate/enzyme) in 100 mM sodium acetate buffer, pH 4.45, at 37 °C for 20 h as described by Reid (1976). After centrifugation and adjustment of the pH to 8.0, the supernatant of the digest was fractionated on a column of Sephacryl S-300 equilibrated with 100 mM acetic acid. The fraction containing the collagen-like stalks of C1q was

located by measurement of the A_{280} values, and the presence of the C1q stalks was confirmed by analysis by SDS/PAGE (Reid, 1976). The solution of the collagen-like stalk fragment (M_r 189900) was freeze-dried and then redissolved in PBS to give a concentration of 2.25 mg/ml using an A_{280} (1%, 1 cm) of 2.23 at 280 nm.

Human collagen types I, III and IV (isolated from placenta by pepsin extraction) were obtained from Sigma Chemical Co. (product codes C7774, C4407 and C7521 in that order) (Niyibizi et al., 1984; Hill and Harper, 1984; Bailey et al., 1979). The collagen samples (2.0–2.5 mg) were dissolved in either 100 μ l of PBS (in which their solubilities were low) or 200 μ l of 0.015 M or 0.001 M acetic acid [in which their solubilities were much improved (George and Veis, 1991)]. All samples to be used for Fourier-transform i.r. spectroscopy were dialysed at 6 °C with four changes of buffer over at least 36 h into PBS in 99.8% $^2\text{H}_2\text{O}$ to ensure the fullest possible exchange of $^1\text{H}_2\text{O}$, and the removal of acetate in the collagen samples where this was present. The collagen samples were centrifuged before spectroscopic analysis to remove undissolved material.

Fourier-transform i.r. spectra were recorded on a Perkin–Elmer 1750 spectrometer equipped with a TGS detector. A Perkin–Elmer model 7300 data station was used for data acquisition, storage and analysis. Samples (volumes up to 50 μ l) were placed in a thermostatically controlled Beckman FH-01 CFT microcell fitted with CaF_2 windows. The sample compartment was continuously purged with dry air to eliminate absorption by water vapour in the spectral region of interest. Spectral conditions [defined more fully in Haris et al. (1986) and Perkins et al. (1988)] were as follows: number of scans, 200; spectral resolution, 4 cm^{-1} ; sample thickness of 50 μm using a Teflon spacer; sample temperature, 20 °C. After subtraction of the background measured with the $^2\text{H}_2\text{O}$ buffer dialysate, second-derivative calculations were performed using the Perkin–Elmer DERIV function (Haris et al., 1986).

Sequence alignment and secondary-structure prediction

A total of 15 sequences were used as follows: human C1q A-chain (Sellar et al., 1991), mouse C1q A-chain (Petry et al., 1991), human C1q B-chain (Reid, 1985), mouse C1q B-chain (Petry et al., 1989), human C1q C-chain (Sellar et al., 1991), mouse C1q C-chain (Petry et al., 1992), precerebellin (Urade et al., 1991), human $\alpha 1$ type VIII collagen (Muragaki et al., 1991a), rabbit $\alpha 1$ type VIII collagen (Yamaguchi et al., 1989, 1991), human and mouse $\alpha 2$ type VIII collagen (Muragaki et al., 1991b), human $\alpha 1$ type X collagen (Riechenberger et al., 1992; Thomas et al., 1991a), mouse $\alpha 1$ type X collagen (Apte et al., 1992; Kong et al., 1993; Elima et al., 1993), bovine $\alpha 1$ type X collagen (Thomas et al., 1991b) and chicken $\alpha 1$ type X collagen (LuValle et al., 1988). The mouse $\alpha 1$ type X collagen sequence of Apte et al. (1992) differed at five residues from those of Kong et al. (1993) and Elima et al. (1993); the latter two sequences gave an improved alignment and were used in this study.

The sequences were aligned by both eye and use of the multiple-alignment program CLUSTAL V using a range of fixed- and variable-gap penalties (Higgins et al., 1992). The final sequence alignment maximized the occurrence of conserved or chemically homologous residues and minimized gaps. These groups of homologous residues are defined: tiny, G, A, S; aliphatic, I, L, V; aromatic, H, F, Y, W; positive, R, H, K; negative, D, E. Secondary-structure predictions based on the sequence alignment were performed using a Robson and Chou–Fasman methods (Garnier et al., 1978; Chou and Fasman, 1978). The 15 predictions were averaged with the FORTRAN

programs PREDRB8 and PREDCF (Perkins et al., 1988). The two programs produced the percentage frequency at which each residue was found in the helix, extended, turn or coil conformation by the Robson and Chou–Fasman methods respectively, together with graphical output, and PREDRB8 also gave the averaged total score in centinats.

RESULTS AND DISCUSSION

Fourier-transform i.r. spectroscopy of C1q and the C1q stalks

The spectra of C1q at 2.4 mg/ml in solution (Figure 2a) and 3.0 mg/ml in suspension were measured in $^2\text{H}_2\text{O}$ buffers. The broad amide I band between 1600 and 1700 cm^{-1} in the absorbance spectrum corresponded to the C=O stretching vibration of

the polypeptide main chain. The peak maximum occurred at 1638 cm^{-1} for C1q (Figure 2a). The low intensity of absorbance in the amide II band at 1540–1550 cm^{-1} (not shown) relative to that of the amide I band at 1638 cm^{-1} shows that ^1H – ^2H exchange in C1q is high. The overlapping bands in the amide I region were resolved into their subcomponents by calculating the second-derivative spectra (Figure 2c). In order of decreasing intensity, the substructure of the amide I band for C1q showed three prominent components at 1637, 1661 and 1683 cm^{-1} . Weaker features at 1608, 1585 and 1564 cm^{-1} do not contribute to the amide I band but were also observed in the second-derivative spectra.

As a control for large ^1H – ^2H exchange or solvent effects on the band frequencies (Jackson et al., 1991; Perkins et al., 1992), attempts were made to obtain the corresponding C1q spectrum in suspension in $^1\text{H}_2\text{O}$ buffer. These were only partially successful because of the relatively low C1q concentration of 3 mg/ml. Although the spectra (not shown) were noisy, the most intense component at 1634 cm^{-1} in $^1\text{H}_2\text{O}$ is similar to that seen at 1637 cm^{-1} in $^2\text{H}_2\text{O}$. The next most intense component at 1666 cm^{-1} in $^1\text{H}_2\text{O}$ is similar to that seen at 1661 cm^{-1} in $^2\text{H}_2\text{O}$.

In order to assign the C1q bands to secondary-structure features, the C1q stalks from pepsin digestion were studied at 2.25 mg/ml (Figures 2b and 2d). This preparation contained the triple-helix region of C1q together with three to seven non-triple-helix residues at its N-terminus and additional non triple-helix residues between residues 88 and 97. The peak maximum was now observed at 1652 cm^{-1} (Figure 2b). Second-derivative analysis showed that there is only one main component at 1655 cm^{-1} , which is within 6 cm^{-1} of the 1661 cm^{-1} band seen for C1q, and this can be assigned to a triple-helix structure. In C1q, the other strong band seen at 1637 cm^{-1} (Figure 2c) can be readily assigned to the C1q globular heads. This 1637 cm^{-1} frequency is characteristic of β -strand secondary structures (Byler and Susi, 1986; Surewicz et al., 1993), and the weak peak at 1695 cm^{-1} was tentatively assigned as the antiparallel component of this β -sheet. In the C1q stalks, a minor band is seen at 1636 cm^{-1} , which may arise from triple helix and/or the non-triple-helix residues in a β -sheet conformation, and weaker features are seen at 1676 and 1687 cm^{-1} . No bands between 1650 and 1655 cm^{-1} were observed in either $^1\text{H}_2\text{O}$ or $^2\text{H}_2\text{O}$ buffers for either C1q or the C1q stalks (Figure 2). A strong band would have occurred in this range if a significant amount of α -helix was present in C1q, thus it is inferred that α -helix is absent from C1q. Direct analysis of the C1q globular heads isolated by collagenase digestion was precluded by their low solubility (up to 0.3 mg/ml), which is insufficient for Fourier-transform i.r. spectroscopy.

Fourier-transform i.r. spectroscopy of collagen types I, III and IV

To verify the observation of triple-helix structures in C1q and the C1q stalks, i.r. spectra of pepsin-extracted human collagen types I, III and IV were studied in the same $^2\text{H}_2\text{O}$ buffers (Figure 3). Collagen types I and III are two of the fibrillar collagens that constitute the major extracellular support system in vertebrate tissues. They have very similar properties (van der Rest and Garrone, 1991). Collagen type I contains two $\alpha 1(\text{I})$ chains and one $\alpha 2(\text{I})$ chain. Collagen type III is a homopolymer of three $\alpha 1(\text{III})$ chains. The amino acid sequences of collagen types I and III show that 97% and 94% respectively of residues are characteristic of collagen triple helices with glycine residues in every third position. Collagen type IV is found in basement membranes and is aggregated via terminal parts or domains. Its sequence is distinct from collagen types I and III. The two $\alpha 1(\text{IV})$ and one $\alpha 2(\text{IV})$ chains respectively contain 21 and 23 non-triplet interruptions in the triple-helical region (Brazel et al., 1987;

Figure 2 Fourier-transform i.r. spectra of the amide I region of human C1q in solution and pepsin-digested C1q

Both samples were dissolved in PBS in $^2\text{H}_2\text{O}$. (a) and (b) Absorbance spectra of C1q and C1q stalks respectively; (c) and (d) second-derivative spectra of the absorbance spectra shown in (a) and (b). The vertical scales are arbitrary.

Figure 3 Fourier-transform i.r. spectra of the amide I region of pepsin-extracted human collagen types I, III and IV

All samples were dissolved in PBS in $^2\text{H}_2\text{O}$. (a), (b) and (c) Absorbance spectra of collagen type I, type III and type IV respectively; (d), (e) and (f) second-derivative spectra of the absorbance spectra shown in (a), (b) and (c).

Hostikka and Tryggvason, 1988). Triple helix forms 85 % of the sequence. In both type IV chains, 18 of these interruptions are aligned with each other, and the rest are small, involving only one amino acid in insertions, deletions or substitutions (Hostikka and Tryggvason, 1988). These are reflected in the higher flexibility of collagen type IV and the possible occurrence of C1q-like kinks in the triple helix as in Figure 1 (Brazel et al., 1987). The three collagen types are thus suitable for comparison with the C1q triple-helix structure.

Peptic cleavage is assumed to remove the non-triple-helix regions at the N-terminus and C-terminus of collagen types I, III and IV, but to leave intact the triple helix itself (Bailey et al., 1979; Hill and Harper, 1984; Broek et al., 1985). In the Fourier-transform i.r. absorbance spectra of pepsin-extracted collagen types I, III and IV, the maximum of the main band in $^2\text{H}_2\text{O}$ was very similar at 1653 cm^{-1} to that at 1655 cm^{-1} (Figures 3a, 3b and 3c), as seen also for the C1q stalks (Figure 2b). The low absorbance intensity in the amide II region at $1540\text{--}1550\text{ cm}^{-1}$ (not shown) indicates that $^1\text{H}\text{--}^2\text{H}$ exchange is high. In the second-derivative spectra, the most intense band appeared at $1657\text{--}1659\text{ cm}^{-1}$ in all the samples, another intense band occurred at $1632\text{--}1633\text{ cm}^{-1}$ and a relatively weak band occurred at 1682 cm^{-1} . Homopolymeric and heteropolymeric collagen triple helices thus give rise to similar band frequencies.

Figure 3 agrees with the observation of three distinct bands in the Fourier-transform i.r. spectra of rat type I collagen in $^1\text{H}_2\text{O}$

and $^2\text{H}_2\text{O}$ buffers (Lazarev et al., 1985; Jakobsen et al., 1983; Payne and Veis, 1988; George and Veis, 1991). In $^2\text{H}_2\text{O}$ buffer, three unresolved bands in non-resolution-enhanced spectra of rat collagen type I were located at 1670, 1656 and 1630 cm^{-1} (Lazarev et al., 1985). The latter two agree well with the data of Figure 3, and the frequency difference between the band reported at 1670 cm^{-1} and that at 1682 cm^{-1} in Figure 3 is attributed to its weak intensity in the Fourier-transform i.r. spectra of collagen. In H_2O buffer, the band positions are affected by the change of solvent. Two prominent bands in deconvoluted spectra of collagen type I occur again at 1660 and 1633 cm^{-1} , and new bands are seen at 1643 and 1695 cm^{-1} (Payne and Veis, 1988; George and Veis, 1991). The appearance of three amide I bands from a triple helix is attributed to the occurrence of five different environments for the three peptide CO groups, C_1O_1 , C_2O_2 and C_3O_3 in the standard collagen triplet $(\text{Gly}_1\text{-X}_2\text{-Y}_3)_n$. Note that X and Y are frequently proline and hydroxyproline imino acid residues if they are not amino acid residues, and the two X and Y residues will thus generate four of the five possible CO environments. In the standard atomic model of a triple helix, the amide or imino X-residue C_2O_2 carbonyl group forms internal hydrogen bonds with the NH group of glycine residues, whereas the glycine amide C_1O_1 and Y-residue amide or imino C_3O_3 carbonyl groups face outwards in contact with H_2O solvent molecules (Fraser et al., 1979; Okuyama et al., 1981).

All three collagen bands, most notably that at 1657–1659 cm^{-1} (Figure 3), are consistent with the Fourier-transform i.r. spectra of C1q and the C1q stalks in $^2\text{H}_2\text{O}$ (Figure 2). The weaker collagen band at 1631 cm^{-1} appears to be partially masked by the main feature in C1q at 1637 cm^{-1} and the weak feature in the C1q stalks at 1636 cm^{-1} . The amounts of proline and hydroxyproline imino acids in each of the triple-helix sequences of C1q and collagen types I, III and IV are calculated to be very similar, between 21.0 and 23.3 % of the total residues. It is thus expected that the imino acid bands assigned to 1657–1661 cm^{-1} and 1630–1637 cm^{-1} (Lazarev et al., 1985; Payne and Veis, 1988; George and Veis, 1991) should be just as prominent in the spectrum of C1q as they are in those of collagen. This well-resolved 1661 cm^{-1} band in C1q and that at 1655 cm^{-1} in the C1q stalks is consistent with the formation of triple helix along an 81-residue segment in each of the A-, B- and C-chains (Figure 1). The strong band at 1637 cm^{-1} in the C1q spectrum was wholly absent from Figure 3. This is further support for the identification of β -sheet secondary structures in the C1q globular head.

Alignment of 15 sequences for the C-terminal globular domain

Sequences similar to those in the C-terminal C1q globular domain are found also at the C-terminus of precerebellin and the C-termini of collagen types VIII and X. A total of 15 sequences were aligned both manually and automatically (Figure 4). The two methods gave similar outcomes, with the main differences occurring at points of poor sequence similarity. The consensus alignment in Figure 4 is defined on the basis of residue positions occupied in over 50 % of the sequences. This gave a total length of 134 residues in the domain. Of these, 35 and 58 residues were conserved to better than 90 % and 70 % respectively in the 15 sequences. Three highly conserved segments were located between residues 41 and 55, 93 and 106 and 124 and 133 (Figure 4). Of the 58 residues with at least 70 % conservation, 45 were hydrophobic (of which 14 were aromatic) and these may define a buried hydrophobic core. Three negatively and two positively charged residues were conserved to better than 70 %, and are probably on the protein surface. Only precerebellin has a lower overall sequence similarity to the other proteins. Precerebellin does not

Figure 4 Alignment of 15 sequences similar to the C-terminal globular domain of C1q

The residue numbering in four segments shown above and below the sequences is based on residue positions that are occupied in over 50% of sequences, and every tenth residue is marked by |. Residue 1 corresponds to residue 90 in Figure 4 of Sellar et al. (1991). Above the alignment, residue conservations of greater than 70% and 90% are indicated by one and two asterisks respectively. RGD tripeptides are given in bold in the second and sixth sequence. The R and K residues thought to be important in the interaction of C1q with IgG (Marqués et al., 1993) are in bold in the human C1q chains. Beneath the alignment, putative glycosylation (Glycos) sites are underlined and indicated by g; cysteine residues are indicated likewise by c. The locations of the β -strands identified in Figure 5 are indicated by the arrowed ranges β 1 to β 10. The averaged secondary-structure predictions by the β -biased and unbiased Robson methods and the Chou-Fasman method are shown beneath the sequences (β , β -strand; α , α -helix; t, turn; c, coil). The consensus hydropathy according to the hydrophobicity scale of Eisenberg (1984) is also shown beneath the sequences (b, hydrophobic; l, hydrophilic).

have an N-terminal collagenous domain, and it is not known if this is monomeric or trimeric.

The C1q sequences are distinct from the eight collagen sequences. Residues 1, 12, 42 and 75 are positively charged in the C1q sequences, whereas their counterparts in the collagen sequences are hydrophilic and uncharged (residue 1) or are hydrophobic and aliphatic (residues 12, 42 and 75). Likewise, residues 30 and 59 in the C1q sequences are hydrophilic and uncharged, whereas their collagen counterparts are hydrophobic. In the collagen sequences, the segments found mostly between residues 80 and 92 and 115 and 123 are highly conserved, yet these have no counterparts in the C1q sequences. These differences may reflect the heterotrimeric chain association in C1q, a dimer-monomer association in collagen type VIII and a homotrimer in collagen type X.

Two distinct patterns of cysteine residues are found. In

C1q, these are Cys-45, Cys-65 and Cys-80. Cys-45 and Cys-65 probably form an intrachain disulphide bridge within each monomer, and Cys-80 is likely to contain a free SH group (Reid, 1985; Sellar et al., 1991). The collagen type VIII chains contain Cys-45 and either Cys-59 or Cys-132, whereas the collagen type-X chains contain only Cys-45. Dimers of pro- α 1(X)-type collagen have been found to be disulphide-linked, either in the globular domain as found in chicken collagen (Kwan et al., 1986) or within the triple-helix region as found in bovine collagen (Remington et al., 1984). Precerebellin contained no cysteine residues.

Secondary-structure prediction of the C1q globular domain

Averaged Robson and Chou-Fasman secondary-structure predictions were carried out for comparison with the Fourier-transform i.r. spectroscopic data. The β -biased Robson pre-

Figure 5 Consensus β -biased Robson secondary-structure prediction of the C1q globular domain

This is based on the sequence alignment of Figure 4 and the domain length of 134 residues. Secondary structures are denoted as follows: ●, β -sheet; ○, α -helix; ----, β -turn and random coil. The ten predicted β -strands are indicated above the plots as solid lines.

diction gave a total of 64 % β -sheet, 0 % α -helix, 17 % β -turn and 19 % coil, the unbiased Robson prediction gave 26 % β -sheet, 4 % α -helix, 26 % β -turn and 24 % coil and the Chou–Fasman method gave 43 % β -sheet, 21 % α -helix and 37 % β -turn. The lack of Fourier-transform i.r. spectral evidence for α -helices in C1q shows that both the β -biased and unbiased Robson predictions are compatible with these data. Quantitative integration of the i.r. bands (Lee et al., 1990) could not be carried out as no database with triple-helix structures is available. Nonetheless, assuming that the molar absorptivities of the bands at 1661 and 1637 cm^{-1} are equal and correspond solely to triple helix and β -sheet respectively, the comparable intensities of the two bands in Figure 2(a) suggest that these structures are present in comparable amounts. From the primary structure, 37.1 % of the C1q residues would be triple helix, and 38.6 % and 15.7 % of residues would be β -sheet in the β -biased and unbiased Robson predictions respectively. The β -biased prediction was thus preferred in Figure 5.

The general accuracy of the predictions is 50–60 %, but can be improved with averaging techniques or templates (Kabsch and Sander, 1983; Argos, 1989). The predictions were reproducible for each of the C1q and collagen subsets. Even with the Chou–Fasman prediction, all three predictions support the existence of an alternating pattern of ten β -strands of lengths 5–19 residues and nine β -turns/loops in the C1q globular head (Figure 4). This is most clearly seen in the β -biased Robson prediction of Figure 5. Figure 4 shows that β -strands β_3 to β_7 were well predicted by all three methods. Only part of all of β -strands β_1 , β_2 , β_8 and β_9 were predicted to be α -helix by the Chou–Fasman method and sometimes also by the unbiased Robson method. This accounts for the predicted α -helix in residues 189–200 of the B-chain of human C1q (close to our predicted β -strand β_8) by Baumann and Anderson (1990). In fact, their c.d. analysis of the peptide corresponding to residues 189–200 resulted in 3 % α -helix and 22 % β -sheet, which is supportive of the high β -sheet content at residues 98–111 in the analyses in Figures 4 and 5.

Consecutive amino acid residue side chains on a β -sheet alternate in direction above and below the plane of the β -sheet

(Creighton, 1993). Using the consensus hydrophobicity scale of Eisenberg (1984), most of the β -strands β_1 to β_{10} were found to exhibit an alternating pattern of hydrophobic (b) and hydrophilic (l) residues which ranged in length from pairs of residues (lb or bl in Figure 4) to five and seven residues (blblb and blblblb for strands β_{10} and β_4). This pattern is consistent with β -strands at a protein surface with side chains that are alternately buried or solvent-exposed. The remainder of the β -strands were hydrophobic up to five residues (bbbbb for strands β_5 and β_{10}). Three main segments of hydrophobic residues correspond to the three most conserved sequence segments between residues 41 and 55, 93 and 106 and 124 and 133 (Figure 4). These residues may define a hydrophobic core. Six of the nine predicted loops and/or turns between the strands β_1 to β_{10} were found to be primarily hydrophilic. This is consistent with their expected surface exposure.

The structure analysis is consistent with the sequence alignment. The largest insertion/deletion sites of two residues or more at residues 18/19, 71/72 and 107/110 are all at surface regions predicted to be turn or coil. The three putative N-glycosylation sites in C1q, precerebellin and collagen type X (Figure 4) are at predicted hydrophilic turn or coil structures, which is consistent with the surface location of carbohydrates in proteins. Cys-80 in C1q [thought to contain the free-SH group involved in covalent adducts with IgG (Martin et al., 1990)] is located at a hydrophilic site that is shared with Asp-80 in the collagen sequences. Cys-45 and Cys-65 in C1q are predicted to be at hydrophobic positions on β -strands β_3 and β_5 , which would be consistent with a disulphide bridge between them. There are two RGD motifs in mouse C1q that are known to be recognition sequences in ligands of some integrins. These occur in β -bend or β -turn conformations (Reed et al., 1988), and this is consistent with the prediction for residues 100–102.

CONCLUSIONS

The C-terminal globular domain of C1q and collagen types VIII and X is important both for the correct folding and alignment of the triple helix (Rosenbloom et al., 1976; Engel and Prockop, 1991) and for protein–protein recognition events (see below). Folding occurs in the C- to N-terminus direction (Bächinger et al., 1981). Deletion of the C-terminal domain of collagen type X suggested that the domain is important in the initiation and maintenance of the correct assembly of the protein (Kwan et al., 1991). The three highly conserved clusters of amino acids at residues 41–54, 93–106 and 124–133 (in particular the hydrophobic residues) may be important for this protein folding. The Fourier-transform i.r. spectroscopic data and the prediction analyses indicate that β -sheet secondary structures are formed (Figures 2 and 5). Note that the same basic secondary structure is expected to be formed by all three chains in one collagen trimer, irrespective of whether it is a heterotrimer or a homotrimer.

Ionic interactions between the C1q heads and its substrates

Protein–protein interactions between the heads of C1q and other proteins are mediated by ionic interactions. In the interactions with IgG and IgM, the globular domain of C1q binds to the $\text{C}\gamma_2$ domain of IgG and the $\text{C}\mu_3$ domain of IgM after these antibodies have formed immune complexes with an activating surface, thereby activating complement. The IgG- and IgM-binding site on the C1q globular domains is strongly electrostatic (Hughes-Jones, 1977; Poon et al., 1985). The IgG residues involved in the binding to C1q are the charged residues, Glu-318, Lys-320 and

Lys-322 (Duncan and Winter, 1988). In IgG, the aromatic residue Tyr-319 has also been implicated (McCall and Easterbrook-Smith, 1989). In IgM, Glu-418 and His-420 on the C μ 3 domain have been proposed to constitute the C1q-binding site (Perkins et al., 1991; Painter, 1993). Other examples of ionic interactions involving the globular heads of C1q are those with rat adipocyte membranes (Comis and Easterbrook-Smith, 1985), human diploid fibroblast subtype (Bordin and Page, 1989), human spectrin (Comis and Easterbrook-Smith, 1986) and human fibrinogen and fibrin (Entwhistle and Furcht, 1988).

In the C1q heads, charged residues outside the highly conserved core region of aromatic residues will be involved in these ionic interactions. Although the binding site for IgG and IgM in C1q has not yet been unequivocally established, the location of proposed sites can be assessed from the structure analysis of Figure 5. Easterbrook-Smith (1983) and Comis and Easterbrook-Smith (1985) showed that histidine and arginine residues are essential. Such positively charged groups on C1q can interact with Glu-318 in IgG and Glu-418 in IgM.

Marqués et al. (1993) have variously implicated Arg-24, Arg-39, Arg-70, Arg-72 and Arg-74 in the A-, B- and C-chains of human C1q in the binding site for immune complexes (highlighted in Figure 4), but they exhibited a preference for the Arg-70/Arg-72/Arg-74 sites. Arg-70, Arg-72 and Arg-74 are predicted to occur at an exposed loop region of variable size and low conservation in the 15 sequences, and Arg-24 and Arg-39 occur in a better defined part of the predicted structure. Figure 4 shows that Arg-24, Arg-39, Arg-72 and Arg-74 are conserved in the corresponding mouse C1q sequences, thus supporting this identification.

Wines and Easterbrook-Smith (1990) also showed that lysine residues close to the intrachain Cys-Cys bridge in the C1q head are close to the C1q-binding site (although they have not been postulated to be directly involved in C1q binding). These are Lys-41/Lys-45 and Lys-75/Lys-85 and are highlighted in Figure 4. Lys-41/Lys-45 occur in the predicted β -strand β 3 at hydrophilic locations, and Lys-75/Lys-85 occur in the predicted β -strand β 6 and the loop following this.

Duncan and Winter (1988) have speculated that Glu-107-Asp-109-Lys-111 might form a possible IgG-binding sequence in the C1q A-chain; however, Lys-111 is not conserved in the mouse C1q A-chain. Baumann and Anderson (1990) have proposed that residues 99-111 in the B-chain (EQGENVFLQAT) form an IgG-binding site; however, this is unlikely as there are no positively charged residues in this peptide, and it is not conserved in the mouse C1q sequence.

Occurrence of C1q head-like sequences in collagen types VIII and X

The globular protein structure of the C-terminal domain in collagen types VIII and X is involved in its function. Type VIII collagen is the major component of Descemet's membrane which is a basement membrane separating corneal endothelial cells from corneal stroma. Type X collagen is formed by hypertrophic chondrocytes during endochondral ossification and is involved in cartilage degradation. Both collagen types form fibrils or nets which can be part of a sheet-like structure. The C-terminal domain may be important in the self-aggregation of collagen type VIII to form stacks of hexagonal collagen lattices and in the interaction of collagen type X with other macromolecules such as proteoglycans (Mayne and Burgeson, 1987; Sawada et al., 1990; van der Rest and Garrone, 1991; Kwan et al., 1991; Chen et al., 1992). Collagen type X contains approximately 15 Ca²⁺-binding

sites per α 1 chain. As the C-terminal domain is rich in acidic amino acids, this domain may form the binding site for Ca²⁺ (Kirsch and von der Mark, 1991).

Financial support from the Wellcome Trust is gratefully acknowledged. We thank Mr. J. O. Beal and Mr. A. S. Nealis for excellent computational support.

REFERENCES

- Apte, S. S., Seldin, M. F., Hayashi, M. and Olsen, B. R. (1992) *Eur. J. Biochem.* **206**, 217-224
- Argos, P. (1989) in *Protein Structure: A Practical Approach* (Creighton, T. E., ed.), pp. 169-189, IRL Press, Oxford
- Bächinger, H. P., Fessler, L. I., Timpl, R. and Fessler, J. H. (1981) *J. Biol. Chem.* **256**, 13193-13199
- Bailey, A. J., Sims, T. J., Duanne, V. C. and Light, N. D. (1979) *FEBS Lett.* **99**, 361-366
- Baumann, M. A. and Anderson, B. E. (1990) *J. Biol. Chem.* **265**, 18414-18422
- Bordin, S. and Page, R. C. (1989) *Mol. Immunol.* **26**, 677-685
- Brass, A., Kadler, K. E., Thomas, J. T., Grant, M. E. and Boot-Handford, R. P. (1992) *FEBS Lett.* **303**, 126-128
- Brazel, D., Oberbäumer, I., Dieringer, H., Babel, W., Glanville, R. W. and Deutzmann, R. (1987) *Eur. J. Biochem.* **168**, 529-536
- Brodsky-Doyle, B., Leonard, K. R. and Reid, K. B. M. (1976) *Biochem. J.* **159**, 279-286
- Broek, D. L., Madri, J., Eikenberry, E. F. and Brodsky, B. (1985) *J. Biol. Chem.* **260**, 555-562
- Byler, D. M. and Susi, H. (1986) *Biopolymers* **25**, 469-487
- Chen, Q., Linsenmayer, C., Gu, H., Schmid, T. M. and Linsenmayer, T. F. (1992) *J. Cell Biol.* **117**, 687-694
- Chou, P. Y. and Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 145-148
- Comis, A. and Easterbrook-Smith, S. B. (1985) *Mol. Immunol.* **22**, 857-861
- Comis, A. and Easterbrook-Smith, S. B. (1986) *Biochim. Biophys. Acta* **870**, 426-431
- Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties*, 2nd edn., W. H. Freeman and Co., New York
- Duncan, A. R. and Winter, G. (1988) *Nature (London)* **332**, 738-740
- Easterbrook-Smith, S. B. (1983) *Bioscience reports* **3**, 135-140
- Eisenberg, D. (1984) *Annu. Rev. Biochem.* **53**, 595-623
- Elima, K., Eerola, I., Rosati, R., Metsäranta, M., Garofalo, S., Perälä, M., de Crombrughe, B. and Vuorio, E. (1993) *Biochem. J.* **289**, 247-253
- Engel, J. and Prockop, D. J. (1991) *Annu. Rev. Biophys. Chem.* **20**, 137-152
- Entwhistle, R. A. and Furcht, L. T. (1988) *Biochemistry* **27**, 507-512
- Fraser, R. D. B., MacRae, T. P. and Suzuki, E. (1979) *J. Mol. Biol.* **129**, 463-481
- Garnier, J., Osguthorpe, D. J. and Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120
- George, A. and Veis, A. (1991) *Biochemistry* **30**, 2372-2377
- Haris, P. I. and Chapman, D. (1992) *Trends Biochem. Sci.* **17**, 328-333
- Haris, P. I., Lee, D. C. and Chapman, D. (1986) *Biochim. Biophys. Acta* **874**, 255-265
- Heinz, H. P. (1989) *Behring Inst. Mitt.* **84**, 20-31
- Higgins, D. G., Bleasby, A. J. and Fuchs, R. (1992) *Comput. Applic. Biosci.* **8**, 189-191
- Hill, R. J. and Harper, E. (1984) *Anal. Biochem.* **141**, 83-93
- Hostikka, S. L. and Tryggvason, K. (1988) *J. Biol. Chem.* **263**, 19488-19493
- Hughes-Jones, N. C. (1977) *Immunology* **32**, 191-198
- Jackson, M., Haris, P. I. and Chapman, D. (1991) *Biochemistry* **30**, 9681-9686
- Jakobsen, R. J., Brown, L. L., Hutson, T. B., Fink, D. J. and Veis, A. (1983) *Science* **220**, 1288-1290
- Kabsch, W. and Sander, C. (1983) *FEBS Lett.* **155**, 179-182
- Kirsch, T. and von der Mark, K. (1991) *FEBS Lett.* **294**, 149-152
- Knobel, H. R., Villiger, W. and Isliker, H. (1975) *Eur. J. Immunol.* **5**, 78-81
- Kong, R. Y. C., Kwan, K. M., Lau, E. T., Thomas, J. T., Boot-Handford, R. P., Grant, M. E. and Cheah, K. S. E. (1993) *Eur. J. Biochem.* **213**, 99-111
- Kwan, A. P. L., Sear, C. H. J. and Grant, M. E. (1986) *FEBS Lett.* **206**, 267-272
- Kwan, A. P. L., Cummings, C. E., Chapman, J. A. and Grant, M. E. (1991) *J. Cell Biol.* **114**, 597-604
- Lazarev, Yu. A., Grishkovsky, B. A. and Khromova, T. B. (1985) *Biopolymers* **24**, 1449-1478
- Lee, D. C., Haris, P. I., Chapman, D. and Mitchell, R. C. (1990) *Biochemistry* **29**, 9185-9193
- Liberi, P. A. and Paul, S. M. (1978) *Biochemistry* **17**, 1952-1958
- LuValle, P., Ninomiya, Y., Rosenblum, N. D. and Olsen, B. R. (1988) *J. Biol. Chem.* **263**, 18378-18385
- Mann, K., Jander, R., Korsching, E., Kühn, K. and Rauterberg, J. (1990) *FEBS Lett.* **273**, 168-177
- Marqués, G., Antón, L. C., Barrio, E., Sánchez, A., Ruiz, S., Gavilanes, F. and Vivanco, F. (1993) *J. Biol. Chem.* **268**, 10393-10402
- Martin, H., Kaul, M. and Loos, M. (1990) *Eur. J. Immunol.* **20**, 1641-1645
- Mayne, R. and Burgeson, R. E. (eds.) (1987) *Structure and Function of Collagen Types*, Academic Press, Orlando

- McCall, M. N. and Easterbrook-Smith, S. B. (1989) *Biochem. J.* **257**, 845–851
- Muragaki, Y., Mattei, M.-G., Yamaguchi, N., Olsen, B. R. and Ninomiya, Y. (1991a) *Eur. J. Biochem.* **197**, 615–622
- Muragaki, Y., Jacenko, O., Apte, S., Mattei, M.-G., Ninomiya, Y. and Olsen, B. R. (1991b) *J. Biol. Chem.* **266**, 7721–7727
- Niyibizi, C., Fietzek, P. P. and van der Rest, M. (1984) *J. Biol. Chem.* **259**, 14170–14174
- Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M. and Kakudo, M. (1981) *J. Mol. Biol.* **152**, 427–443
- Painter, R. H. (1993) *Behring Inst. Mitt.* **93**, 131–137
- Payne, K. J. and Veis, A. (1988) *Biopolymers* **27**, 1749–1760
- Perkins, S. J. (1985) *Biochem. J.* **228**, 13–26
- Perkins, S. J., Villiers, C. L., Arlaud, G. J., Boyd, J., Burton, D. R., Colomb, M. G. and Dwek, R. A. (1984) *J. Mol. Biol.* **179**, 547–557
- Perkins, S. J., Haris, P. I., Sim, R. B. and Chapman, D. (1988) *Biochemistry* **27**, 4004–4012
- Perkins, S. J., Nealis, A. S., Sutton, B. J. and Feinstein, A. (1991) *J. Mol. Biol.* **221**, 1345–1366
- Perkins, S. J., Smith, K. F., Nealis, A. S., Haris, P. I., Chapman, D., Bauer, C. J. and Harrison, R. A. (1992) *J. Mol. Biol.* **228**, 1235–1254
- Petry, F., Reid, K. B. M. and Loos, M. (1989) *FEBS Lett.* **258**, 89–93
- Petry, F., Reid, K. B. M. and Loos, M. (1991) *J. Immunol.* **147**, 3988–3993
- Petry, F., Reid, K. B. M. and Loos, M. (1992) *Eur. J. Biochem.* **209**, 129–134
- Poon, P. H., Phillips, M. L. and Schumaker, V. N. (1985) *J. Biol. Chem.* **260**, 9357–9365
- Reed, J., Hull, W. E., von der Lieth, C.-W., Kübler, D., Suhai, S. and Kinzel, V. (1988) *Eur. J. Biochem.* **178**, 141–154
- Reichenberger, E., Beier, F., LuValle, P., Olsen, B. R., von der Mark, K. and Bertling, W. M. (1992) *FEBS Lett.* **311**, 305–310
- Reid, K. B. M. (1976) *Biochem. J.* **155**, 5–17
- Reid, K. B. M. (1981) *Methods Enzymol.* **80**, 16–25
- Reid, K. B. M. (1985) *Biochem. J.* **231**, 729–735
- Reid, K. B. M. (1986) *Essays Biochem.* **22**, 27–68
- Reid, K. B. M. and Day, A. J. (1990) *Immunol. Today* **11**, 387–388
- Remington, M. C., Bashey, R. I., Brighton, C. T. and Jimenez, S. A. (1984) *Biochem. J.* **224**, 227–233
- Rosenbloom, J., Endo, R. and Harsch, M. (1976) *J. Biol. Chem.* **251**, 2070–2076
- Sawada, H., Konomi, H. and Hirose, K. (1990) *J. Cell Biol.* **110**, 219–227
- Schumaker, V. N., Hanson, D. C., Kilchherr, E., Phillips, M. L. and Poon, P. H. (1986) *Mol. Immunol.* **23**, 557–565
- Sellar, G. C., Blake, D. J. and Reid, K. B. M. (1991) *Biochem. J.* **274**, 481–490
- Smith, K. F., Haris, P. I., Chapman, D., Reid, K. B. M. and Perkins, S. J. (1993) *Mol. Immunol.* **30**, 52–52 (abstract)
- Surewicz, W. K., Mantsch, H. H. and Chapman, D. (1993) *Biochemistry* **32**, 389–394
- Thomas, J. T., Cresswell, C. J., Rash, B., Nicolai, H., Jones, T., Solomon, E., Grant, M. E. and Boot-Handford, R. P. (1991a) *Biochem. J.* **280**, 617–623
- Thomas, J. T., Kwan, A. P. L., Grant, M. E. and Boot-Handford, R. P. (1991b) *Biochem. J.* **273**, 141–148
- Urade, Y., Oberdick, J., Molinar-Rode, R. and Morgan, J. I. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1069–1073
- van der Rest, M. and Garrone, R. (1991) *FASEB J.* **5**, 2814–2823
- Wells, T. N. C., Stedman, M. and Leatherbarrow, R. J. (1992) *Ultramicroscopy* **42–44**, 1200–1203
- Wines, B. D. and Easterbrook-Smith, S. B. (1990) *Mol. Immunol.* **27**, 221–226
- Yamaguchi, N., Benya, P. D., van der Rest, M. and Ninomiya, Y. (1989) *J. Biol. Chem.* **264**, 16022–16029
- Yamaguchi, N., Mayne, R. and Ninomiya, Y. (1991) *J. Biol. Chem.* **266**, 4508–4513