ANALYSIS OF HUMAN C1q BY COMBINED BOTTOM-UP AND TOP-DOWN MASS SPECTROMETRY: DETAILED MAPPING OF POST-TRANSLATIONAL MODIFICATIONS AND INSIGHTS INTO THE C1r/C1s BINDING SITES

Delphine Pflieger1,2*, Cédric Przybylski1,2*, Florence Gonnet1,2, Jean-Pierre Le Caer3, Thomas Lunardi4, Gérard J. Arlaud4, Régis Daniel1,2

1Université d’Evry-Val-d’Essonne, Laboratoire Analyse et Modélisation pour la Biologie et l’Environnement, F-91025 Evry, France
2CNRS UMR 8587, Laboratoire Analyse et Modélisation pour la Biologie et l’Environnement, F-91025 Evry, France
3CNRS UPR 2301, Equipe de Spectrométrie de Masse, Institut de Chimie des Substances Naturelles, 91198 Gif-sur-Yvette, France Cedex
4CNRS UMR 5075, Laboratoire d’Enzymologie Moléculaire, Institut de Biologie Structurale Jean-Pierre Ebel, Université Joseph Fourier, 38027 Grenoble Cedex

Running head: Mapping of post-translational modifications of C1q by MS

Address correspondence to: Delphine Pflieger and Régis Daniel, Université d’Evry-Val-d’Essonne, Laboratoire Analyse et Modélisation pour la Biologie et l’Environnement, Bd F. Mitterrand, F-91025 Evry, France. Fax: +33-1-69-47-7655; E-mail: delphine.pflieger@univ-evry.fr, regis.daniel@univ-evry.fr

*: these authors contributed equally to the manuscript
SUMMARY

C1q is a subunit of the C1 complex, a key player in innate immunity that triggers activation of the classical complement pathway. Featuring a unique structural organization and comprising a collagen-like domain with a high level of post-translational modifications, C1q represents a challenging protein assembly for structural biology. We report for the first time a comprehensive proteomic study of C1q combining bottom-up and top-down analyses. C1q was submitted to proteolytic digestion by a combination of collagenase and trypsin for bottom-up analyses; in addition to classical LC-MS/MS analyses which provided reliable identification of hydroxylated proline and lysine residues, sugar-loss-triggered MS$^3$ scans were acquired on a LTQ-Orbitrap instrument to strengthen the localization of glucosylgalactosyl disaccharide moieties on hydroxylysine residues. Top-down analyses performed on the same instrument allowed high-accuracy and high-resolution mass measurements of the intact full-length C1q polypeptide chains and the iterative fragmentation of the proteins in the MS$^n$ mode. This study illustrates the usefulness of combining the two complementary analytical approaches to obtain a detailed characterization of the post-translational modification pattern of the collagen-like domain of C1q, and highlights the structural heterogeneity of individual molecules. Most importantly, three lysine residues of the collagen-like domain, namely K59 (A chain), K61 (B chain) and K58 (C chain) were unambiguously shown to be completely unmodified. These lysine residues are located about half-way along the collagen-like fibers. They are thus fully available and in an appropriate position to interact with the C1r and C1s protease partners of C1q, and are therefore likely to play an essential role in C1 assembly.
INTRODUCTION

The multiprotein complex C1 (790 000 g.mol⁻¹), the first component of the classical complement pathway, is a key effector in host defense leading to the elimination of pathogens and altered host cells (1). C1 comprises a subunit named C1q insuring its recognition and binding functions, and a protease subunit consisting of the Ca²⁺-dependent tetramer C1s-C1r-C1r-C1s (2). The enzyme subunit requires to be properly folded within C1q to undergo activation upon binding of C1q to a target and to trigger the complement cascade. These requirements for appropriate folding and activation are fulfilled by the C1q subunit through mechanisms remaining to be deciphered. The interface between C1q and the tetramer is not fully identified (3-5). Although the C1q binding sites of C1r and C1s have been recently determined by site directed mutagenesis (6), much less is known about the interaction sites on C1q.

C1q is a multimeric glycoprotein comprising 18 chains of three different types, A, B and C (217-226 residues). These form 3 C-C and 6 A-B disulfide-bonded dimers, which assemble into six hetero-trimeric (A-B:C) subunits through the non-covalent association (A-B:C)-(C:B-A). This hexameric structure appears in electron microscopy as a bouquet of six flowers and exhibits unique structural features. It consists of a short N-terminal region (3-9 residues) containing the inter-chain disulfide bonds, prolonged by a triple-helical collagen-like region (CLR, about 81 residues) which forms a “stalk” and then diverges into six individual “stems”, each terminating in a C-terminal heterotrimeric globular “head” (or globular region, GR, about 135 residues) (Figure 1) (7, 8). Several lines of evidence indicate that the tetramer binds to the collagen-like region of C1q. The intact CLR produced by pepsin digestion of C1q competes with C1q for association with the tetramer (9), and a monoclonal antibody to CLR prevents the interaction between C1q and the tetramer (10). Indeed, the effective association of CLR with the tetramer has been observed by analytical
ultracentrifugation providing further indication that the interaction sites between C1q and the
tetramer are confined to the collagen region (11). Electron microscopy studies of the cross-
linked C1 complex showed that the tetramer is located in the cone formed by the six collagen-
like stems between the globular heads and the stalk (12, 13). Electron microscopy and neutron
scattering studies of isolated C1s-C1r-C1r-C1s indicated an extended conformation of the
tetramer in solution (14, 15). However the volume inside the C1q cone is too small to contain
the entire elongated tetramer. Based on these considerations, on neutron scattering studies of
C1 (16) and on the C1s CUB1-EGF homodimer X-ray structure, a structural model of C1 has
been proposed in which the tetramer undergoes a large conformational change into a compact
“8-shaped” structure wound in and out the C1q stems (17). This model has been recently
refined based on the identification of the C1q binding sites of C1r and C1s (6). It has been
proposed that both C1r/C1s CUB1-EGF-CUB2 heterodimers are located inside the C1q cone
and mediate ionic interactions through acidic residues contributed by the C1r and C1s CUB
modules. Such ionic interactions at the C1q/C1s-C1r-C1r-C1s interface are expected to
involve lysine residues of the collagen-like stems of C1q as suggested by the observation that
chemical modification of C1q with lysine-specific reagents inhibits C1 assembly and C1q
hemolytic activity (4, 18).

Being collagen-like, the sequences of the CLR chains contain the repeating Gly-X-Z
triplet where X is often a proline and Z is frequently a hydroxylysine or a hydroxyproline.
C1q contains 8.3 % carbohydrate (7), most of it being present in the CLR as
glucosylgalactosyl disaccharide units linked to hydroxylysine residues (19). Because of the
steric hindrance arising from the disaccharide moieties, non-glycosylated lysine residues are
more plausible candidates for ionic interactions at the C1q/ C1s-C1r-C1r-C1s interface. It was
previously reported that 82.6 % of the hydroxylysines are glycosylated (19), so that few
unmodified lysine residues are available for this purpose. The primary structures initially
reported in the 1970s did not allow the post-translational modifications (PTMs, hydroxylations, glycosylations) to be fully accurately identified along the CLR chains (20-25). Furthermore, the full-length C1q protein could not be analyzed to date by nuclear magnetic resonance (NMR) or X-ray crystallography, so that the basic residues of the C1q stems involved in the C1q/tetramer interface remain to be identified. We have previously reported the MALDI-TOF mass spectrometry analysis of the entire C1q protein (26). The aim of the present study was to identify the C1q PTMs by MS and locate them along its chains in order to obtain further insights into the lysine residues likely involved in the interaction with the C1s-C1r-C1r-C1s tetramer. For this purpose, we have performed for the first time a comprehensive proteomic study of C1q by using two very complementary approaches. On the one hand, C1q and CLR samples were digested by a combination of collagenase and trypsin, the resulting peptides being analyzed by capillary liquid chromatography and tandem mass spectrometry (nanoLC-MS/MS). On the other hand, the bottom-up approach was completed by top-down analyses, consisting of the direct infusion into the LTQ-Orbitrap instrument of the intact CLR domain and of the full-length A, B and C chains of C1q. Spectra were successively acquired in the MS mode in the Orbitrap analyzer, to obtain high-accuracy and high-resolution measurements of the intact protein masses, and in the MS^n mode, to yield iterative fragmentation spectra of the proteins. The combination of bottom-up and top-down data proved most successful in yielding a significant coverage of the three chain sequences and in deciphering the heterogeneity of the C1q covalent modifications, both in terms of proline hydroxylation and lysine glycosylation, thus revealing an unexpected level of complexity of the post-translational modifications.
EXPERIMENTAL PROCEDURES

Reagents and materials—DL-dithiothreitol (DTT), iodoacetamide (IAA), and NH₄HCO₃ were purchased from Sigma-Aldrich (St. Louis, MO, U.S.A.). Tris (2-carboxyethyl)phosphine (TCEP) was purchased from Sigma (ref C-4706) and methyl methane thiosulfonate (MMTS) was obtained from Fluka (ref 64306). HPLC gradient grade acetonitrile and normapur grade formic acid were purchased from VWR (West Chester, PE, U.S.A.). All buffers and solutions were prepared using ultra-pure water (milliQ, Millipore, Bedford, MA, USA). Type VII collagenase from Clostridium histolyticum (EC 3.4.24.3) was obtained from Sigma-Aldrich (St. Louis, MO, U.S.A.). Sequencing-grade modified porcine trypsin (EC 3.4.21.4) was purchased from Promega (Madison, WI, U.S.A).

Fused-silica PicoTips (360 µm O.D., 20 µm I.D.) with a nominal tip end I.D. of 10±1.0 µm used for LC-nanoESI-LTQ-Orbitrap coupling were obtained from New Objectives (Woburn, MA, U.S.A, reference FS360-20-10-N-20-C10.5). For the analysis of whole proteins by direct nanoESI infusion, metallized tips of reference BG12-69-2-CE (New Objectives) were used.

Preparation of C1q and its collagen-like region (CLR)— Human C1q was purified as previously described (27). For comparison, a commercial preparation of human C1q was purchased from Calbiochem (La Jolla, CA, U.S.A). The collagen-like region (CLR) of C1q was prepared as previously described (28, 29).

Protein reduction and alkylation— A first series of protein samples was prepared as follows. Twenty µL of commercial and non-commercial C1q, as well as 10 µL of CLR (20 µg, 17.8 µg and 32 µg of protein, respectively) were reduced by incubation for 1 h at 60°C in the presence of 4.5 mM DTT. Samples were then alkylated by addition of 22 mM IAA and incubation for 45 min at room temperature in the dark. In view of the previously reported over-alkylation of proteins by IAA (30), a second series of C1q samples was prepared by
reduction with 5 mM TCEP at 37°C for 45 min and alkylation with 10 mM MMTS at room temperature for 15 min.

Protein digestion by collagenase and trypsin—Reduced and alkylated samples were supplemented with collagenase to obtain an enzyme/substrate ratio of 0.5 (w/w), and incubated for 3 h at 37°C. Before performing trypsin digestion of the mixture containing the C1q globular regions and collagenase, 50 mM NH₄HCO₃, pH 8.0, was added to the samples to reach a final concentration of 20 mM, in order to obtain pH conditions more suitable to the serine protease activity (pH >7.5). Two and 3 µL of trypsin at 0.4 µg/µL were then added to each C1q and CLR sample, respectively, before incubating them for 18 h at 37°C.

Reversed phase-LC-MS/MS analysis on a QqTOF instrument—A first series of LC-MS/MS analyses was performed using a Famos-Switchos-UltiMate chromatographic system (LC Packings/Dionex) coupled to a hybrid quadrupole QqTOF mass spectrometer QSTAR Pulsar i (Applied Biosystems/MDS Sciex) equipped with a nanoelectrospray source Protana XYZ manipulator (Protana, Denmark). Protein digests (1 pmol) were loaded onto a C18 precolumn (PepMap100 C18, 300 µm I.D., 5 mm length, 5 µm particle size, 100 Å porosity, Dionex) for desalting and concentrating at a flow rate of 30 µL/min in solvent A (water/acetonitrile/formic acid, 98/2/0.1, v/v/v). Peptides were then eluted from the precolumn and separated on a capillary column (Pepmap C18, 75 µm I.D., 150 mm length, 3 µm particle size, 100 Å porosity, Dionex) at 200 nL/min, using a gradient as follows: solvent A for 5 min, linear increase to 60% of solvent B (water/acetonitrile/formic acid, 10/90/0.1, v/v/v) in 50 min, then ramp to 90% B in 5 min (held 10 min) and return to 100% A in 5 min for a 15-min long re-equilibration of the columns. The auto-sampler was kept at 10°C. Peptides eluting from the column were analyzed using the information-dependent data acquisition feature in the Analyst QS software v1.1: species ionized in the nanoESI source were detected for 1 s in the MS mode and the three most intense signals associated to either
doubly or triply charged species were subsequently selected to be fragmented in the MS/MS mode for 3 s each. MS detection and MS/MS acquisitions were performed over the m/z ranges 400-1400 and 100-2000, respectively. Analyses were carried out with the dynamic exclusion of already fragmented m/z values for 3 min.

Reversed phase-LC-MS/MS analysis on a LTQ-Orbitrap instrument— A second series of LC-MS/MS analyses was realized using an UltiMate 3000 chromatographic system (Dionex) coupled to a hybrid LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Typically 200-300 fmol of digested C1q or CLR were injected onto a C18 pre-column (Dionex). After desalting for 5 min with buffer A (0.1% formic acid in water), peptides were separated on a capillary column (same reference as the one used on the QqTOF instrument), using a gradient from 100% solvent A to 60% solvent B (water/acetonitrile/formic acid, 10/90/0.1, v/v) in 60 min. The column was then further washed with 95% solvent B for 10 min. One series of MS analyses (ITMS² analyses) consisted in acquiring cycles composed of one MS scan in the Orbitrap analyzer (profile mode, Rs = 15000, m/z range 400-2000), followed by three MS/MS scans (CID fragmentation and detection in the linear ion trap analyzer, centroid mode, isolation width 2 Da) triggered on the three most intense species detected in the preceding MS scan. Singly charged species were excluded from fragmentation; dynamic exclusion of already fragmented ions was applied for 90 s, with a repeat count of 1, a repeat duration of 20 s and an exclusion mass width of +/- 5 ppm. Automatic gain control allowed accumulating up to $5 \times 10^5$ ions for FTMS scans, $10^5$ ions for FTMSⁿ scans and $10^4$ ions for ITMSⁿ scans. The maximum injection time was 100 ms for acquiring FTMS and ITMSⁿ scans. Only one microscan was acquired for each scan type, even though three were accumulated in FTMS mode in initial experiments. When testing acquisitions consisting of CID fragmentation in the ion trap analyzer and detection of the resulting fragments in the Orbitrap (OTMS² analyses), the maximum injection time was 200
ms. In experiments combining MS² and neutral-loss-triggered MS³ scans (MS²/MS³ analyses), MS² spectra were only acquired in the ion trap on the two precursor ions giving the most intense signals in MS, and MS³ was launched whenever a neutral loss (NL) of m/z 108.035, 162.053, 216.070, 324.106 and 486.158 was detected with a +/- 0.5 Da tolerance among the 8 most intense MS² fragments; MS³ was oriented towards the MS² fragment corresponding to the biggest NL. For these MS²/MS³ analyses, a precursor selection window of 2 and 5 Da was used for MS² and MS³ scans, respectively. In addition, 3 microscans were accumulated to build an MS³ scan. Acquired raw data were processed by the software Bioworks to create Mascot compatible MGF files (no grouping of MS/MS scans was allowed).

Protein identification using NanoESI direct infusion on the LTQ-Orbitrap — Five µL of reduced non-commercial C1q and of CLR were desalted on ZipTip C4 (Millipore), eluted in (water/acetonitrile/formic acid 50/50/0.5, v/v/v) to 5 pmol/µL, and loaded into a metallized nanoelectrospray needle (PicoTip emitters, reference BG12-69-2-CE-20, New Objective). A spray was obtained while working at a capillary temperature of 240°C and adjusting the voltage applied to the nanoESI tip between 1.4 and 2.4 kV. Automatic Gain Control parameters were set to 2.10^6 for FTMS, 2.10^5 for FTMS² and 10^4 for ITMSⁿ scans. Target resolution was 60,000 for FTMS and FTMS² analyses. The maximum injection time was set to 500 ms in FTMS and FTMS² and to 100 ms in ITMSⁿ scans. The precursor selection width for FTMS² fragmentation was around 3 Da, and always adjusted to find a compromise between sensitivity and the clean selection of a single isotopic distribution. The selection window for ITMSⁿ fragmentation (n≥3) was 5 Da. The spectra shown in this article correspond to the accumulation of scans over approximately one minute, yet good signal to noise ratios could be obtained within less time.
Database searches—LC-MS/MS data (.MGF files obtained from .WIFF or .RAW data) were searched using the Mascot software (www.matrixscience.com). Acquired data was compared to a homemade database named ‘mature C1q’, consisting of only five sequences extracted from the Swiss-Prot database release 54.7: porcine trypsin (P00761), collagenase (Q9X721) and the three mature protein sequences C1qA, C1qB and C1qC (accession numbers P02745, P02746 and P02747; the signal peptides 1-22, 1-25 and 1-28, respectively, were removed from the sequences stored in Swiss-Prot to obtain the mature sequences). Searches were first performed while considering that the analyzed peptides resulted from the combined use of collagenase and trypsin (collagenase + trypsin search). The sequential use of collagenase and trypsin to digest C1q was considered to produce peptides resulting from cleavages N-terminally of glycine residues and C-terminally of lysine and arginine residues. Collagenase is indeed known to specifically degrade collagen by cleaving N-terminally of GPX triplets. In a second step, searches were run by considering that trypsin may have produced half-tryptic peptides (collagenase + semi-trypsin search). In both searches, 9 missed cleavages were allowed. Finally, an error tolerant search was performed after the two previous search steps, to allow in particular identification of the N-terminal region of chain C1qB which is known to be cyclized into pyrrolidone carboxylic acid. For data obtained on the QqTOF instrument, tolerances on mass measurement of precursors and fragments were set to 50 ppm and 0.4 Da, respectively. For Orbitrap data, tolerances on mass measurements were 5 ppm (FTMS) and 0.8 Da (ITMSⁿ). Cysteine residues were considered to be fully alkylated either by IAA or MMTS; hydroxylation of lysine and proline residues, oxidation of methionine residues as well as glucosylgalactosyl modification on hydroxylysines were included as potential modifications. We modified the definition of the modification ‘glucosylgalactosyl (K)’ initially present in Mascot for neutral losses of 162.0528 mass units and 324.1056 mass units to be taken into account for scoring.
RESULTS

Bottom-up analysis of C1q and of its collagen-like regions (CLR)

Three types of samples were analyzed to determine the PTMs of C1q: laboratory-purified C1q, commercial C1q and CLR. Each of them was digested using successively collagenase and trypsin, and analyzed by LC-MS/MS. To facilitate reading, hydroxylated proline and lysine residues are noted P* and K*, respectively, and a lysine residue bearing a glucosylgalactosyl (Glc-Gal) moiety is noted K#.

Overall sequence coverage of chains A, B and C — The digested samples were first characterized by triplicate LC-MS/MS analyses on QqTOF and LTQ-Orbitrap instruments (MS² analyses). In the latter case, fragmentation and detection of the fragments were carried out in the linear ion trap of the hybrid instrument (ITMS² analyses). The peptide sequences identified in the three chains constitutive of C1q are shown in Figure 2. Database searches were performed assuming that collagenase would cleave N-terminally of glycine residues within the CLR regions (see Materials and Methods). Indeed, nearly all identified sequences resulted from cleavages N-terminally of glycine residues and/or C-terminally of K/R residues. It is worth mentioning that, except for glycosylated peptides (see below), the identification of a peptide was routinely validated only if a minimal continuous sequence stretch of 5 amino acids was identified thanks to y-type or b-type fragment ions. Nonetheless, the frequent occurrence of proline residues at the Z position within GXZ triplets led us to be more tolerant: some peptide identifications were accepted when series of y ions, separated by three residues, corresponded to fragments containing an N-terminal proline (31, 32).

The C1q A chain, its CLR (residues 9-87) and its globular head region (residues 88-223) were covered at 77%, 78% and 76%, respectively. The sequence 100-128 of the C1q A globular head could not be detected likely because of the presence of a N-linked oligosaccharide on residue N124 (33). The C1q B chain, its CLR (residues 6-89) and its
globular head region (residues 90-226) were covered at 90%, 100% and 83%, respectively. The ‘error tolerant’ search provided the identification of the N-terminal peptide of C1qB, 1%QLSCTGPP*8, in which glutamine Q1 was detected as a pyroglutamic acid (Figure 2, white square), in agreement with previous data (22). Finally, the C1q C chain, its CLR (residues 3-86) and its globular head region (residues 87-217) were covered at 68%, 84% and 57%, respectively. These analyses resolved reported conflicts (22, 23) and confirmed the identity of residues deduced from the cDNA sequence (34): P75, K81, C150, S156 and the sequence LIFP (218-221) in the C1q A chain, N58 and G73 in the C1q B chain, K29, P38, K44 and P56 in the C1q C chain.

Identification of hydroxylated sequences — The collagen-like region (CLR) of C1q contains the repeating triplet Gly-X-Z where Z is frequently a proline or a hydroxyproline residue. In most cases, peptides containing hydroxylated residues (but no glycosylation) fragmented in a manner similar to non-modified species, allowing confident determination of their sequence; yet, as stated above, the frequent occurrence of proline residues was detrimental to the detection of continuous y/b fragment series. As presented in Figure 2, several proline residues of the three C1q chains were detected in both a non-modified and a hydroxylated form, indicating incomplete modification. For instance, peptide 72GPMGIPGEPEEGR85 of C1q C was confidently identified in a non-modified form, as well as in singly and doubly hydroxylated forms, modifications being observed on residue P77 alone and additionally on P80. Taking into account the possible oxidation of methionine residues, this peptide was finally identified as doubly hydroxylated (on both P77 and P80) and oxidized on M74. Even though oxidation and hydroxylation yield equal mass increments, the detected MS^2 fragments ruled out the possible hydroxylation of P73.
Several peptides containing one or two glycosylated lysine residues and fully or partially hydroxylated proline residues were also identified. Even though identification of glycosylated sequences was often tedious (see section below), P23 and P35 from C1q A were clearly identified as being fully modified, whereas P32 was confidently determined to be partially modified, thanks to the detection of several overlapping peptides.

Identification of glycosylated sequences — Composition analysis has indicated that C1q contains 8.3% carbohydrate (7), most of it being present in CLR as Glc-Gal disaccharide units linked to 82.6% of the hydroxylysine residues (19). In contrast to peptides containing only hydroxylated residues, glycosylated sequences systematically provided MS² spectra dominated by sequential neutral losses of saccharide moieties (162.05 mass units), whereas fragmentation along the peptide backbone produced peaks of minor intensity, often precluding robust sequence determination and localization of the glycosylation site. We therefore tested different fragmentation/detection schemes on the LTQ-Orbitrap instrument to obtain complementary and redundant identification of glycosylated sequences and thus increase confidence in the identification of lysine residues bearing a Glc-Gal motif.

The MS² data acquired on the QqTOF and LTQ-Orbitrap instruments allowed identification of the glycosylated sequences shown in Figure 2. All sequences identified by the Mascot software are shown, provided that all fragment peaks of major intensities matched the theoretical fragments expected for the proposed sequence. We checked the relevance of hydroxylation positioning, and re-introduced uncertainty when the detected fragments were insufficient to definitely localize this modification on a specific proline or lysine residue (see for example peptide 39-59 of C1q A). Some peptides bearing one glycosylation and containing one lysine residue could be readily identified; Figure S1 shows a MS² spectrum obtained on the QqTOF instrument and unambiguously identifying peptide
14

39TGIQGLK#GDQGEP51 from C1q A. Determination was more challenging when glycosylation site(s) occurred within sequences containing more lysine residues than the number of disaccharide moieties. This is exemplified in sequences 73-92 from C1q A, 69-90 from C1q B and 42-58 from C1q C, which all contain three lysine residues and were identified as being doubly modified by Glc-Gal motifs. The Mascot search software then often provided identification of sequences corresponding to the different possible combinations of free and glycosylated lysines. For example, peptide 69-90 from C1q B was proposed to be modified either on K71 and K83 or on K83 and K90. Interestingly, glycosylated lysine residues were sometimes detected as still bearing one sugar moiety (the lysine was modified by 178.0477 mass units) during fragmentation in the linear ion trap of the LTQ-Orbitrap instrument; this behaviour was never observed using the QqTOF instrument, with which initially glycosylated lysine residues were systematically detected in a simply hydroxylated form in MS² spectra. Two spectra showing lysine residues carrying a single sugar unit during MS/MS fragmentation are provided in Supplementary Figures S2A and S2B; such fragmentation patterns helped us confirm the glycosylated nature of some lysine residues. More generally, to discriminate which lysine residues were most probably modified by Glc-Gal, we determined the best consensus resulting from the different proposed glycosylated sequences. Taking again the doubly glycosylated sequence 69-90 from C1q B as an example, the identification of peptide 69GPK#GGPGAP*GAP*80 unambiguously pointed K71 as being glycosylated (Figure 2); additionally, several MS² spectra identifying sequence 69GPK#GGPGAP*GAP*GP(KGESGDYK)90# allowed detection of the y2 ion at 310.176 mass units, indicating that the C-terminal lysine was free. We therefore concluded on the consensus sequence 69GPK#GGPGAP*GAP*GPK#GESGDYK90. This reasoning aiming to deduce the most probable modified residues in the three C1q chains from the largest number of converging sequences identified was applied systematically. The conclusions derived from
MS² analyses are collected in Figure 3. In addition, identification of glycosylated sequences was further confirmed by performing LC-MS/MS analyses of digested CLR and detecting both the precursor and the fragment ions in the Orbitrap cell (OTMS² analyses) (35). Visual inspection of the fragmentation data allowed us to select MS² spectra from glycosylated sequences by pointing multiple losses of 162.0528 mass units. Then, by simply combining the knowledge of the accurate mass of the precursor and of the number of sugar losses, we could match certain experimental spectra to theoretical sequences of C1q, obtained by *in silico* digestion of the three chains (considering cleavages N-terminally of G and C-terminally of K/R residues). The glycosylated sequences thus identified are listed in Table S1. This highlighted the fact that the knowledge of the precursor mass at 5 ppm accuracy and of its glycosylated nature was often sufficient to unequivocally identify sequences from C1q; this observation supports the identification of glycosylated sequences obtained from the previous ITMS² data acquired on the LTQ-Orbitrap instrument.

Identification of glycosylated sequences was hampered by the preferential loss of the labile disaccharide moieties upon CID fragmentation. This is usually also observed with peptides phosphorylated on S/T residues, of which a better characterization has been largely described using an additional fragmentation step (MS³) triggered on the MS² fragment corresponding to the loss of phosphoric acid (36). To more confidently identify glycosylated peptides from the C1q chains, we therefore analyzed digested C1q and CLR samples while specifying the acquisition of MS³ spectra on the MS² fragments corresponding to the heaviest detected sugar loss from 2+ and 3+ precursor species. MS³ spectra were then interpreted by considering lysines as possibly hydroxylated. The glycosylated sequences identified from these ITMS³ analyses are represented in detail in Figure 4 and merged in Figure 3. Interestingly, overlapping peptides containing lysines K78 and K81 in C1q A mostly exhibited a loss of two disaccharide moieties between MS² and MS³, but also happened to
only lose one such modification. This confirmed that these two lysines were mostly doubly glycosylated but also occurred as a glycosylated/hydroxylated pair, as previously observed in MS² analyses (Figure 2).

**Outcome of bottom-up analyses**

The analysis of digested C1q and CLR samples provided significant sequence coverage of the three C1q chains. The majority of proline residues (P) considered to be hydroxyproline in previous reports (shown in bold in Figure 3) were confirmed to be fully modified (22). Nonetheless, our study revealed a more subtle modification pattern, in which some prolines (e.g. P31, P51 and P57 from C1q A, and P20, P53 and P56 from C1q B), either recorded as modified or free, were in fact shown to be partially hydroxylated. In addition, C1q A, B and C were determined to bear (minimally) 4, 5 and 3 glycosylations, respectively, on 5-hydroxylysine residues. In particular, three previously unknown glycosylation sites were identified on K50 of C1q B, and K29 and K44 of C1q C. Besides, peptides containing K78 and K81 of C1q A were observed as either doubly or singly glycosylated; similarly, lysine K83 from C1q B was mostly observed in a glycosylated form, but was also found in a hydroxylated form in one peptide. Nevertheless, several sequence stretches with potential PTMs remained uncovered by this bottom-up approach, such as in the N-terminal end and the globular region of C1q A. Furthermore, these data revealed an additional, unanticipated level of post-translational heterogeneity. To increase the sequence coverage and address the question of the PTM heterogeneity at the whole protein level, we next completed the bottom-up analyses by acquiring top-down data.
Top-down analysis of C1q and its collagen-like-regions

LTQ-Orbitrap instruments have already been demonstrated to allow analysis of intact proteins up to 50 kDa, to determine their molecular mass with ppm accuracy and to yield sequence information thanks to iterative MS^n level analyses (37). C1q was therefore subjected to top-down characterization at the C1q A, B and C chain levels (<26 kDa); similarly, the CLR polypeptides were analyzed by iterative MS^n experiments (up to MS^4). The latter sample was obtained from C1q using pepsin, a non-specific enzyme which preferentially cleaves on the C-terminal side of aromatic residues (F, W and Y) as well as L, A, E and Q (proteolysis at pH>2), generating several possible sequences. The challenge in this study was to handle the simultaneous analysis, without prior separation by LC, of a mixture consisting of different protein chains exhibiting variable levels of hydroxylation and glycosylation.

This complexity is illustrated by the mass spectra of reduced CLR and C1q acquired by direct nanoESI infusion of the proteins on the LTQ-Orbitrap analyzer. Ionic species with charge states ranging from 7 to 17+ and from 18 to 28+ were detected during MS analysis of CLR (Figure 5A) and reduced C1q (Figure 5B), respectively. The corresponding deconvoluted spectra are provided in Figures 6A and B. The sequences that could be attributed to the signals detected during top-down analysis of C1q and CLR are listed in Table 1. More details on the MS^n data that allowed establishing these matches are provided in the next section.

Four signals, designated A-1 to A-4 in Figure 6A, were detected for the A chain in the CLR sample. A-1 and A-2 were attributed to sequences 1-97 and 1-95 from C1q A, respectively, decorated with 8 hydroxylations and 5 glycosylations. The lower-intensity signals A-3 and A-4 corresponded to the same sequence stretches, yet bearing only 4 glycosylations (-340.10 mass units). This top-down data indicated a heterogeneous level of glycosylation for the A species. Since the residue pair (K78, K81) was detected by bottom-up
analysis as being either doubly glycosylated or hydroxylated/glycosylated, one of these residues must be glycosylated in chains A-1, A-2 and only hydroxylated in A-3, A-4. Similarly, two signals, B-1 and B-2, could be attributed to sequence 1-97 from C1q B. Whereas B-1 bears 12 hydroxylations and 5 glycosylations, the minor fraction B-2 only carries 4 saccharide moieties but 13 hydroxylations (-324.1 mass units). In B-2, lysine K83 could lack glycosylation, considering that this residue was found by bottom-up analysis to be either glycosylated or hydroxylated. Finally, four signals were assigned to different sequence stretches of C1q C: polypeptides 1-90, 1-92, 1-93 and 1-94 could be detected, each being decorated with 3 glycosylations and 14 hydroxylations. The most intense species C-1, i.e. probably the most abundant CLR C1q C polypeptide, corresponded to the cleavage C-terminally of an F residue, in agreement with the cleavage specificity of pepsin. Each peak pattern detected for the CLR polypeptides additionally consisted of a series of ionic species by increments of 16 mass units, obviously indicating a variable level of hydroxylation (15.9949 mass units). The ranges of hydroxylation motifs present on each C1q chain are indicated in Table 1.

Iterative fragmentation of CLR and C1q chains

A chain characterization

Species A-1 was fragmented by MS/MS, in its 14+ (m/z 817.35) and 16+ (m/z 715.90) charge states, with detection of the resulting fragments in the Orbitrap analyzer. Globally the same fragment ions were generated in both cases. The deconvoluted MS² spectrum of the m/z 817.35 ion is provided in Figure 7. Detailed inspection of MS² fragments obtained on CLR C1q A and of the matched theoretical sequences allowed us to determine the position of some hydroxylations and glycosylations; this information is represented in Figure 3. Based on these MS² data, the calculated monoisotopic mass of 11410.4432 mass units (species A-1)
could be attributed definitely to the CLR sequence 1-97 of C1q A, modified with 5 glycosylations and 8 hydroxylations.

To validate the N-linked glycosylation at N124, we acquired an MS/MS spectrum on the ionic cluster centred on m/z = 1061.63 (26+), corresponding to the C1q A chain (Supplementary Figure S3). The detected fragments of higher charge states (y173 ions) and highest intensities at m/z 1189.7425 (18+) and m/z 1259.9043 (17+), merged into the same monoisotopic mass of 21386.24 mass units in the deconvoluted spectrum (Figure 8). This value matched the C-terminal region of C1q A starting at residue P51 and bearing 2 O-glycosylations, 4 hydroxylations and a monosialylated fucosylated biantennary N-glycan with a mass of 2059.74 mass units (theoretical mass 21387.28 mass units). In addition, the intense ion signals detected at about m/z 1173.6829 (18+) and m/z 1242.6631 (17+), corresponding to a monoisotopic mass of 21095.16 mass units, were assigned to the same C-terminal sequence, yet having lost the terminal sialic acid group (Δm -291.073 mass units). The detection of sialylated glycan chains is in agreement with a previous analysis indicating that 80% of Asn-linked glycans of C1q contain sialic acid (38). Whereas a minor N-glycan population with 2 sialic acid residues has been reported (38), we found that these glycans were substituted with a single sialic acid residue. The signal detected at m/z 883.4193 (7+) (b507+ ion) corresponded to the N-terminal sequence 1-50 (6173.879 mass units) previously observed during fragmentation of CLR C1q A, and bearing 3 glycosylations and 4 hydroxylations, and was complementary to the C-terminal fragment at 21387.24 mass units. As a whole, 5 glycosylations and 8 hydroxylations, as well as one sialylated fucosylated biantennary N-glycan of 2059.74 mass units, were thus confirmed to be present on C1q A, accounting for a mass of 27561.29 mass units. The minor fraction possessing 4 K# probably contains one of the two lysines K78 and K81 in a hydroxylated form.
B chain characterization

The MS² spectrum acquired on the ionic species around m/z = 1102.41 (10+) is provided in Figures 9A and 9B. As reported above in the bottom-up analysis, preferential fragmentation was observed at the N-terminus of proline residues. The sequence 1%QLSCTG(PP)*AIP*GIP*GIP*GTPGPD²³ could be confidently determined from the MS² spectrum, unambiguously identifying CLR C1q B. We could thus precisely match the B-1 polypeptide at mass 11008.10 mass units, with 5.1 ppm error, to the sequence stretch 1-97 of C1q B (-----KATQKIAF), bearing 5 glycosylations, 12 hydroxylations, and with a loss of NH₃ at the protein N-terminus, due to conversion of the glutamine residue into pyrrolidone carboxylic acid (-17.0265 mass units).

The ionic species around m/z 1104.02 (10+) (deconvoluted monoisotopic mass of 1102.41 mass units) was also fragmented and led to a fragmentation pattern similar to that obtained for the ion at m/z 1102.41 (10+) (Supplementary Figure S4). As deduced from the mass difference between both 10+ species, they only differed by one hydroxylation, which could be placed within the sequence stretch 20-23. Indeed, a common y₇₄ fragment was detected at 983.67 (9+), whereas fragment y₇₈ was either seen at 1024.245 (9+) or at 1026.135 (9+), thus definitely localizing the additional hydroxylation within the sequence 2⁰PGPD²³. Based on these results, we conclude that CLR C1q B molecules bear 11 to 14 hydroxylations, indicating a variable level of hydroxylation, as observed for CLR C1q A.

In addition to 8+ and 9+ fragments, the two fragmented polypeptides at m/z = 1102.41 (10+) and m/z = 1104.02 (10+) produced common 1+ and 2+ fragments, at m/z 535.3027 (2+), m/z 854.3668 (b₉⁺ ion), m/z 967.4524 (b₁₀⁺ ion), m/z 1250.6036 (b₁₃⁺ ion) and m/z 1533.75 (b₁₆⁺ ion). These MS² fragments were then selected to be fragmented in MS³. Table 2 lists the peptide sequences determined from MS³ scans, and the manually interpreted spectra are provided as Supplementary Data. Fragmentation of b ions validated the hydroxylation of
P8, P11 and P14. Residue P7 appeared as possibly hydroxylated from the MS$^2$ fragmentation spectrum obtained on m/z 1250.6036 (1+); however, this modification would be unusual, given its location at position 2 within a GXZ triplet. The doubly charged ion at m/z 535.3027 corresponded to the y$_9^{2+}$ ion, validating the C-terminal region of the CLR C1q B B-1 species as follows: $^{89}$YKATQKIAF$^{97}$.

**C chain characterization**

Due to the occurrence of F/Q/Y amino acids between residues 86-94 of the C1q C chain, several polypeptides could be expected to be generated from cleavage of this chain by pepsin. MS/MS fragmentation performed on the ions around m/z 1314.23 (8+) (monoisotopic mass 10516.843 mass units) and m/z 1258.71 (8+) (monoisotopic mass 10055.616 mass units) allowed identifying the same sequence $^1$NTGCYGIP*GM$^{10}$, based on 7+ fragments (Supplementary Figures S5 and S6, respectively), thus unambiguously attributing this species to a CLR C1q C polypeptide.

The MS/MS spectra acquired on species around m/z 1314.23 (8+), m/z 1316.38 (8+) and m/z 1256.71 (8+), m/z 1258.71 (8+) allowed us to detect peptides at charge states between 1+ and 3+, which were selected for MS$^3$ fragmentation, followed by MS$^4$ when possible (Supplementary Data). Table 2 contains the sequences thus identified. These iterative fragmentations validated the C-terminal sequences of C-1 (-----KQKFQSVF) and C-2 (-----KQKF). MS$^3$ fragmentation of the species at m/z 1256.71 and 1314.23 (containing 3 K# and 13 K* or P*) revealed hydroxylation of P71 (left unidentified by bottom-up analysis) and of P77. MS$^3$ fragmentation of the species at m/z 1258.71 and 1316.38 (containing 3 K# and 14 K* or P*) highlighted an additional hydroxylation on P80. Very interestingly, MS$^2$ fragmentation of the species at m/z 1316.38 (Supplementary Figure S7) unambiguously revealed the heterogeneous hydroxylation of biomolecules corresponding to that mass.
Indeed, this ion fragmented into two $y_{18}^{2+}$ ions, at m/z 1058.51 (corresponding to $^{77}\text{P*GEP*GEEGRYQKFQSVF}^{94}$) and m/z 1050.01 (corresponding to $^{77}(\text{PGEP})*\text{GEEGRYQKFQSVF}^{94}$), as well as two $y_{24}^{2+}$ ions containing three ($^{71}\text{P*GPMGIP*GEP*GEEGRYQKFQSVF}^{94}$) or only two hydroxylated proline residues. The detection of these two pairs of y ions, with one or two hydroxylation(s) in the sequence $^{77}\text{PGEP}^{80}$, indicated the presence of counterbalancing proline residues being either free or hydroxylated within the 1-70 stretch of the fragmented CLR C1q C chain.

**Outcome of the top-down analysis**

The top-down analysis of the collagen-like regions of C1q indicated two sites of pepsin cleavage in C1q A (after F95 and A97), one site in C1q B (after F97), and four sites in C1q C (after F90, S92, V93 and F94). Top-down analyses of CLR and intact reduced C1q allowed identification of the PTMs decorating each chain (summarized in Figure 3), some of which (e.g. the glycosylation of K10/K11 in C1q A) had not been detected by bottom-up analyses. The complexity of the hydroxylation patterns was further illustrated: in particular, fragmentation of CLR C1q C at m/z 1316.38 (Figure S7) showed that a biomolecule at a given mass can actually exhibit different combinations of hydroxylation sites. As a whole, the signals detected on intact CLR chains indicated that C1q A contains 5 K# and 6 to 9 hydroxylations (for biomolecules at detectable levels), C1q B bears 5 K# and 11 to 14 hydroxylations, whereas C1q C contains 3 K# and 12 to 15 hydroxylations.

**DISCUSSION**

We report here for the first time the proteomic analysis of the human complement protein C1q, which is composed of 18 chains from three different polypeptides, and exhibits a unique structural organisation comprising a collagen-like domain with a high level of post-
translational modifications: hydroxylations on K/P residues, Glc-Gal disaccharides on K residues, as well as a branched N-linked sugar moiety on the globular moiety of the A chains.

Clearly, identifying the lysine residues modified by Glc-Gal motifs was not trivial. It is worth mentioning that unlike for the common N-glycosylation no known enzyme allows removal of these disaccharides and that attempts to chemically eliminate these modifications led to disruption of the polypeptide chains (data not shown). The fact that the loss of sugar cycles (162.05 mass units) was favoured during MS/MS, together with the inefficient fragmentation of the peptide backbone, rendered sequence determination by the Mascot software difficult. Nonetheless, we could check from OTMS² analyses that the high mass accuracy (<5 ppm) of the LTQ-Orbitrap provided confident identification of glycosylated sequences from chains A, B and C. Additionally, we performed LC-MS/MS analyses with MS³ scans triggered on the heaviest neutral loss of sugar moieties (multiples of 324.1 mass units). This method appeared efficient in producing pairs of MS²/MS³ spectra whose precursors differed by the total number of glycosylation motifs present on the initial peptide (usually 1# or 2#). The obtained MS³ scans allowed reading a sequence in amino acids with much more confidence than the corresponding MS² scans. Such a method programming MS³ scans triggered on the heaviest sugar loss would be of general interest when studying proteins which contain a collagen-like domain exhibiting glycosylated lysine residues. This method may be more generally applicable to the study of labile oligosaccharide-type modifications.

We have combined bottom-up and top-down analyses of the C1q and CLR samples to obtain complementary information on the PTMs decorating the three C1q chains. Obtaining a really exhaustive characterization of the variable modification level of the individual chains would have required separation of the intact proteins by LC to systematically acquire MS⁹ information from each species by off-line infusion (39). Nevertheless, our combined bottom-up and top-down data were sufficient to identify P/K residues that are fully hydroxylated (or
glycosylated) in all C1q molecules (given their systematic detection in a modified form in LC-MS/MS analyses), and others that are either unmodified (or solely hydroxylated) in different biomolecules. A variable level of hydroxylation at specific P/K residues was previously described for other collagen-containing proteins (40-43). In our case, top-down analyses revealed that proteins with the same sequence can bear between N and N+4 hydroxylation motifs (with N being 6, 11 and 12 for the A, B and C C1q chains, respectively). They also showed that proteins with the same sequence and mass can exhibit different distributions of hydroxylation sites, thus highlighting a further level of PTM pattern complexity.

The determination of the primary structure of C1q initially carried out by proteolytic digestion and Edman sequencing yielded the sequences of the C1q A and B chains and that of the 94 N-terminal residues of the C1q C chain (22), and resulted in incomplete identification of the PTMs in the CLR moieties of the chains. The proteomic study reported here allowed us to cover the sequences of the whole C1q A, B, and C chains and to verify the cDNA-derived sequences, confirming that the few discrepancies noticed with the initial protein-derived sequence do not arise from polymorphism. In addition, we made the most of the analytical potentialities of the LTQ-Orbitrap instrument (acquisition of MS² and sugar-loss-based MS³ scans in bottom-up analyses, of iterative MSⁿ fragmentations on intact proteins in top-down analyses) to confidently and comprehensively identify glycosylated residues. The large majority of the identified Glc-Gal-bearing hydroxylysines appear to be fully modified (K10 or 11, K26 and K45 in C1q A; K32, K35, K50 and K71 in C1q B; K29, K44 and K47 in C1q C), yet a few were also detected as being either glycosylated or hydroxylated (K78 or K81 in C1q A and K83 in C1q B). Finally, only three lysine residues within the CLR have been systematically and unambiguously identified in a non-modified form, i.e. residues K59 in C1q A, K61 in C1q B, and K58 in C1q C. This identification is consistent with the early analyses
reported by Reid et al. (22) and represents a highly meaningful information with respect to the assembly of the C1 complex.

Given the important function of the C1 complex in the immune system and its role in the triggering of complement-mediated inflammation, the understanding of its assembly is a key information that, in addition, can be extended to the collectins mannan-binding lectin (MBL) and ficolins, two other important classes of pattern recognition molecules. These collectins, which trigger complement through the lectin pathway, share with C1q the ability to associate in a homologous manner with their partner proteases MBL-associated serine proteases (MASPs) through their collagen domain (44). Point mutations of recombinant MBL and ficolins have revealed the essential role of a single unmodified lysine residue in their collagen domains for the association with the MASPs (45, 46). Unlike MBL and the ficolins which are assembled from a single polypeptide chain and thus have been produced recombinantly, so far C1q could not be studied by site-directed mutagenesis to identify the CLR residues involved in the assembly of the C1 complex. Nevertheless, point mutants of C1r and C1s have been produced, providing evidence for the essential role of Asp and Glu residues within the C1r/C1s CUB modules in the interaction of the C1r/C1s tetramer with C1q, likely through ionic bonds. Given the homologous assembly of the collectins MBL/ficolins and C1q, we postulate that the free residues C1q A-K59, C1q B-K61 and C1q C-K58 are fully accessible and available for interaction with acidic residues contributed by the C1r and C1s CUB modules. In support of this proposal, the location of these three unmodified lysine residues about half-way along the CLR is in agreement with the recently proposed refined model of C1 assembly, in which the C1r/C1s tetramer is positioned inside the cone defined by the C1q stems (6). According to this model, both C1r/C1s CUB1-EGF-CUB2 heterodimers provide six binding sites distributed radially to make contacts with each of the six C1q collagen-like stems. As they are located half-way along the CLR,
unmodified lysine residues are thus in an appropriate position for making individual contacts with the CUB modules and mediating effective interaction with the C1r/C1s tetramer (Figure 10).

The MS characterization reported here is therefore fully consistent with the proposed model for the C1q-C1r/C1s tetramer assembly (6), accounting for the architecture and function of the C1 complex. Recent data reveal that the role of C1 extends beyond pathogen recognition to include implications in auto-immune diseases, ischemia-reperfusion injury, organ graft rejection and neuro-degeneration (1, 47). It is therefore generally considered that the early inhibition of the complement cascade at the C1 level would provide a therapeutic benefit (48, 49). In this context, the experimental data reported here on the interface between C1q and its C1r/C1s protease partners provide useful clues for the design of inhibitory molecules aimed at targeting the C1 complex.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the CNRS, Genopole-France, INRA and Région Ile de France for supporting this work, in particular for funding of the LTQ-Orbitrap, and warmly thank C. Gaboriaud for preparing the C1q model displayed in Figure 10.
REFERENCES

1. Lu, J., Teh, B. K., Wang, L., Wang, Y., Tan, Y. S., Lai, M. C., and Reid, K. B. M. (2008) The classical and regulatory functions of C1q in immunity and autoimmunity. *Cell Mol Immunol* 5, 9-21.

2. Gaboriaud, C., Thielens, N. M., Gregory, L. A., Rossi, V., Fontecilla-Camps, J. C., and Arlaud, G. J. (2004) Structure and activation of the C1 complex of complement: unraveling the puzzle. *Trends Immunol* 25, 368-373.

3. Busby, T. F., and Ingham, K. C. (1990) Amino-Terminal Calcium-Binding Domain of Human Complement C1s Mediates the Interaction of C1r with C1q. *Biochemistry* 29, 4613-4618.

4. Illy, C., Thielens, N. M., and Arlaud, G. J. (1993) Chemical Characterization And Location Of Ionic Interactions Involved In The Assembly Of The C1 Complex Of Human-Complement. *J Prot Chem* 12, 771-781.

5. Thielens, N. M., Illy, C., Bally, I. M., and Arlaud, G. J. (1994) Activation Of Human-Complement Serine-Proteinase C1r Is Down-Regulated By A Ca2+-Dependent Intramolecular Control That Is Released In The C1 Complex Through A Signal Transmitted By C1q. *Biochem J* 301, 509-516.

6. Bally, I., Rossi, V., Lunardi, T., Thielens, N. M., Gaboriaud, C., and Arlaud, G. J. (2009) Identification of the C1q-binding Sites of Human C1r and C1s: a refined three-dimensional model of the C1 complex of complement. *J. Biol. Chem.* 284, 19340-19348.

7. Cooper, N. R. (1985) The classical complement pathway: activation and regulation of the first complement component. *Adv. Immunol.* 37, 151-216.

8. Kishore, U., and Reid, K. B. M. (2000) C1q: structure, function, and receptors. *Immunopharmacol* 49, 159-170.

9. Reid, K. B. M., Sim, R. B., and Faiers, A. P. (1977) Inhibition Of Reconstitution Of Hemolytic Activity Of 1st Component Of Human Complement By A Pepsin-Derived Fragment Of Subcomponent C1q. *Biochem J* 161, 239-245.

10. Hsiung, L. M., Dodds, A. W., Mason, D. W., and Reid, K. B. M. (1988) A Monoclonal-Antibody To Clq Which Appears To Interact With Clr2 C1s2-Binding Site. *Febs Lett* 229, 21-24.

11. Siegel, R. C., and Schumaker, V. N. (1983) Measurement Of The Association Constants Of The Complexes Formed Between Intact Clq Or Pepsin-Treated Clq Stalks And The Unactivated Or Activated Clr2Cls2 Tetramers. *Mol Immunol* 20, 53-66.

12. Poon, P. H., Schumaker, V. N., Phillips, M. L., and Strang, C. J. (1983) Conformation and restricted segmental flexibility of C1, the first component of human complement. *J Mol Biol* 168, 563-577.

13. Strang, C. J., Siegel, R. C., Phillips, M. L., Poon, P. H., and Schumaker, V. N. (1982) Ultrastructure of the first component of human complement: electron microscopy of the crosslinked complex. *Proc Natl Acad Sci USA* 79, 586-590.

14. Tschopp, J., Villiger, W., Fuchs, H., Kilchherr, E., and Engel, J. (1980) Assembly of subcomponents C1r and C1s of first component of complement: electron microscopic and ultracentrifugal studies. *Proc Natl Acad Sci USA* 77, 7014-7018.

15. Boyd, J., Burton, D. R., Perkins, S. J., Villiers, C. L., Dwek, R. A., and Arlaud, G. J. (1983) Neutron scattering studies of the isolated Clr2Cls2 subunit of first component of human complement in solution. *Proc Natl Acad Sci USA* 80, 3769-3773.

16. Perkins, S. J., Villiers, C. L., Arlaud, G. J., Boyd, J., Burton, D. R., Colomb, M. G., and Dwek, R. A. (1984) Neutron-Scattering Studies Of Subcomponent Clq Of 1st Component Cl Of Human-Complement And Its Association With Subunit Clr2Cls2 Within Cl. *J Mol Biol* 179, 547-557.
17. Gregory, L. A., Thielens, N. M., Arlaud, G. J., Fontecilla-Camps, J. C., and Gaboriaud, C. (2003) X-ray structure of the Ca2+-binding interaction domain of C1s. *J Biol Chem* 278, 32157-32164.

18. Tissot, B., Montdargent, B., Chevolot, L., Varenne, A., Descroix, S., Gareil, P., and Daniel, R. (2003) Interaction of fucoidan with the proteins of the complement classical pathway. *Biochim Biophys Acta* 1651, 5-16.

19. Shinkai, H., and Yonemasu, K. (1979) Hydroxylysine-linked glycosides of human complement subcomponent C1q and various collagens. *Biochem J* 177, 847-852.

20. Reid, K. B. M. (1977) Amino-Acid Sequence Of N-Terminal 42 Amino-Acid Residues Of C-Chain Of Subcomponent C1q Of 1st Component Of Human Complement. *Biochem J* 161, 247-251.

21. Reid, K. B. M. (1974) Collagen-Like Amino-Acid Sequence In A Polypeptide-Chain Of Human C1q (A Subcomponent Of First Component Of Complement). *Biochem J* 141, 189-203.

22. Reid, K. B. M. (1979) Complete Amino-Acid Sequences Of The 3 Collagen-Like Regions Present In Subcomponent C1q Of The 1st Component Of Human-Complement. *Biochem J* 179, 367-371.

23. Reid, K. B. M., Gagnon, J., and Frampton, J. (1982) Completion Of The Amino-Acid-Sequences Of The A-Chain And B-Chain Of Subcomponent C1q Of The 1st Component Of Human-Complement. *Biochem J* 203, 559-569.

24. Reid, K. B. M., and Porter, R. R. (1976) Subunit Composition And Structure Of Subcomponent C1q Of 1st Component Of Human Complement. *Biochem J* 155, 19-23.

25. Reid, K. B. M., and Thompson, E. O. P. (1978) Amino-Acid Sequence Of N-Terminal 108 Amino-Acid Residues Of B-Chain Of Subcomponent C1q Of 1st Component Of Human Complement. *Biochem J* 173, 863-868.

26. Tissot, B., Gonnet, F., Iborra, A., Berthou, C., Thielens, N., Arlaud, G. J., and Daniel, R. (2005) Mass spectrometry analysis of the oligomeric C1q protein reveals the B chain as the target of trypsin cleavage and interaction with fucoidan. *Biochemistry* 44, 2602-2609.

27. Arlaud, G. J., Sim, R. B., Duplaa, A.-M., and Colomb, M. G. (1979) Differential elution of C1q, C1r and C1s from human C1 bound to immune aggregates: use in the rapid purification of C1 subcomponents. *Mol. Immunol.* 16, 445-450.

28. Sasaki, T., and Yonemasu, K. (1983) Chemical studies of the isolated collagen-like and globular fragments of complement component C1q: comparative studies on bovine and human C1q. *Biochim. Biophys. Acta* 742, 122-128.

29. Tacnet-Delorme, P., Chevallier, S., and Arland, G. J. (2001) beta-amyloid fibrils activate the C1 complex of complement under physiological conditions: Evidence for a binding site for A beta on the C1q globular regions. *Journal Of Immunology* 167, 6374-6381.

30. Boja, E. S., and Fales, H. M. (2001) Overalkylation of a protein digest with iodoacetamide. *Anal Chem* 73, 3576-3582.

31. Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A., Speed, T. P., and Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* 75, 6251-6264.

32. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R., 3rd (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* 75, 1155-1163.

33. Liu, T., Qian, W. J., Gritsenko, M. A., Camp, D. G., Monroe, M. E., Moore, R. J., and Smith, R. D. (2005) Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J. Proteome Res.* 4, 2070-2080.

34. Sellar, G. C., Blake, D. J., and Reid, K. B. M. (1991) Characterization And Organization Of The Genes Encoding The A-Chains, B-Chains And C-Chains Of Human-
Complement Subcomponent C1q - The Complete Derived Amino-Acid-Sequence Of Human C1q. Biochem J 274, 481-490.

35. Scherl, A., Shaffer, S. A., Taylor, G. K., Hernandez, P., Appel, R. D., Binz, P. A., and Goodlett, D. R. (2008) On the benefits of acquiring peptide fragment ions at high measured mass accuracy. J. Am. Soc. Mass Spectrom. 19, 891-901.

36. Ulintz, P. J., Bodenmiller, B., Andrews, P. C., Aebersold, R., and Nesvizhskii, A. I. (2008) Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. Mol Cell Proteomics 7, 71-87.

37. Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. Molecular & Cellular Proteomics 5, 949-958.

38. Mizuochi, T., Yonemasu, K., Yamashita, K., and Kobata (1978) The asparagines-linked sugar chains of subcomponent C1q of the first component of human complement. J. Biol. Chem. 253, 7404-7409.

39. Wu, S., Lourette, N. M., Tolic, N., Zhao, R., Robinson, E. W., Tolmachev, A. V., Smith, R. D., and Pasa-Tolic, L. (2009) An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. J. Proteome Res 8, 1347-1357.

40. Bos, K. J., Rucklidge, G. J., Dunbar, B., and Robins, S. P. (1999) Primary structure of the helical domain of porcine collagen X. Matrix Biol 18, 149-153.

41. Rothmann, A. B., Mortensen, H. D., Holmsovk, U., and Hojrup, P. (1997) Structural characterization of bovine collectin-43. Eur J Biochem 243, 630-635.

42. Leth-Larsen, R., Holmskov, U., and Hojrup, P. (1999) Structural characterization of human and bovine lung surfactant protein D. Biochem J 343 Pt 3, 645-652.

43. Richards, A. A., Stephens, T., Charlton, H. K., Jones, A., Macdonald, G. A., Prins, J. B., and Whitehead, J. P. (2006) Adiponectin multimerization is dependent on conserved lysines in the collagenous domain: evidence for regulation of multimerization by alterations in posttranslational modifications. Mol Endocrinol 20, 1673-1687.

44. Gaboriaud, C., Teillet, F., Gregory, L. A., Thielenls, N. M., and Arlaud, G. J. (2007) Assembly of C1 and the MBL– and ficolin–MASP complexes: Structural insights. Immunobiol 212, 279-288.

45. Teillet, F., Lacroix, M., Thiel, S., Weilguny, D., Agger, T., Arlaud, G. J., and Thielenls, N. M. (2007) Identification of the Site of Human Mannan-Binding Lectin Involved in the Interaction with Its Partner Serine Proteases: The Essential Role of Lys55. J. Immunol 178, 5710-5757.

46. Lacroix, M., Dumestre-Perard, C., Schoehn, G., Rouen, G., Cesbron, J. Y., Arlaud, G. J., and Thielenls, N. M. (2009) Residue Lys57 in the collagen-like region of human L-ficolin and its counterpart Lys47 in H-ficolin play a key role in the interaction with the mannan-binding lectin-associated serine proteases and the collectin receptor calreticulin. J. Immunol 182, 456-465.

47. Boholson, S. S., Fraser, D. A., and Tenner, A. J. (2007) Complement proteins C1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. Molecular Immunology 44, 33-43.

48. Beinrohr, L., Dobo, J., Zavodszky, P., and Gal, P. (2008) C1, MBL-MASPs and C1-inhibitor: novel approaches for targeting complement-mediated inflammation. Trends In Molecular Medicine 14, 511-521.

49. Sjoberg, A. P., Trouw, L. A., and Blom, A. M. (2009) Complement activation and inhibition: a delicate balance. Trends In Immunology 30, 83-90.

50. Gaboriaud, C., Juanhuix, J., Gruetz, A., Lacroix, M., Darnault, C., Pignol, D., Verger, D., Fontecilla-Camps, J. C., and Arlaud, G. J. (2003) The crystal structure of the globular
head of complement protein C1q provides a basis for its versatile recognition properties. *J Biol Chem* 278, 46974-46982.
**FIGURE LEGENDS**

**Figure 1.** Schematic representation of the structural organization of human C1q and of C1 assembly. (A) Model showing the collagen-like region (CLR) and globular heads of C1q, and the association with the C1r/C1s tetramer forming the C1 complex. (B) Representation of the association between the A, B and C chains leading to formation of the six ABC heterotrimeric triple helices of C1q CLR. The two following interruptions in the Gly-Y-Z triplet repeats are involved in the kink of the collagen-like region, i.e. the insertion of a threonine at position 39 in the A chain and the replacement of a glycine by an alanine at position 36 in the C chain (only the interruption in the A chain is shown).

**Figure 2.** Peptides identified by LC-MS/MS analyses of C1q successively digested with collagenase and trypsin, using QqTOF and LTQ-Orbitrap instruments. Green lines: peptides identified in both laboratory-purified and commercial C1q; black and violet lines: peptides only identified in either laboratory-purified or commercial C1q, respectively; blue lines: peptides specifically identified from the ‘collagenase+semi-trypsin’ database search. Dotted line (in C1q C): sequence GLPGPK#GEP* identified from the ‘collagenase+trypsin’ search was finally rejected in favour of sequence GPK#GEPGIP* (‘collagenase+semitrypsin’) because the latter better matched the experimental spectrum. Blue squares: hydroxylation; red squares: Glc-Gal modification; green squares: deamidation; empty square: pyrrolidone cyclisation. When precise localization of one (two) hydroxylation(s) could not be established from MS/MS spectra, an extended striped blue/white rectangle overlapping the possibly modified residues (P or K) is represented.

**Figure 3.** Summary of the post-translational modifications determined by bottom-up and top-down analyses of CLR and C1q samples. Black line: identifications from MS² data; grey line: identification from MS³ data; green line: identification from top-down data; no line: merged information on PTMs from bottom-up and top-down analyses. Fully blue squares: residue
always detected as hydroxylated; blue/white squares: residue detected either as unmodified or hydroxylated, indicating partial modification; red squares: lysine always detected in a glycosylated form; red/blue squares: lysine detected either as glycosylated or hydroxylated. ?: indicates that ambiguity remains as to whether K10 or K11 of C1q A is glycosylated.

**Figure 4.** Sequences identified from MS$^3$ scans acquired during LC-MS/MS analyses including sugar-loss-triggered MS$^3$ scans, performed on both C1q and CLR samples. Dotted line: one disaccharide lost between consecutive MS$^2$ and MS$^3$ scans; full line: two disaccharides lost. The color code is the same as in figure 1; additionally, orange lines: identifications from the CLR sample. Circled in red: lysines determined to be glycosylated from these MS$^2$/MS$^3$ analyses; circled in black: unmodified lysines.

**Figure 5.** Mass spectra of reduced CLR and C1q acquired by direct nanoESI infusion of the proteins on the LTQ-Orbitrap instrument. (a) MS spectrum obtained on reduced CLR analyzed at a protein concentration of 5 pmol/µL. (b) MS spectrum obtained on reduced C1q analyzed at a protein concentration of 5 pmol/µL.

**Figure 6.** Deconvoluted MS spectra obtained by top-down MS analysis of (a) reduced CLR and (b) reduced C1q.

**Figure 7.** Deconvoluted and deisotoped MS/MS spectrum obtained from fragmentation of the ionic species at about m/z 817.35 (14+) of CLR C1q A. # represents a Glc-Gal moiety and * a hydroxylation. The 33 most intense ions were manually interpreted, even though the deconvolution algorithm happened to be unsuccessful in determining the monoisotopic mass (errors of 1 mass unit). The most intense fragment at 5236.55 mass units corresponded to the CLR C1q A P51-A97 stretch (y$_{47}$ ion) modified by 2 glycosylations and 4 hydroxylations. Its corresponding deglycosylated form was detected at 4912.44 mass units. The complementary sequence E1-E50 (b$_{50}$ ion), bearing 3 glycosylations and 4 hydroxylations, was detected at 6173.89 mass units. Once again, its deglycosylated form was detected at 5849.78 mass units.
The b$_{50}$ ion revealed glycosylation of either K10 or K11, a modification that could not be detected by bottom-up analysis. Another couple of fragments was detected at 7439.59 mass units (y$_{65}$ ion, recalculated monoisotopic mass at 7440.59), corresponding to sequence P32-A97, bearing 3 glycosylations and 6 hydroxylations, and at 3969.84 mass units (b$_{32}$ ion) corresponding to sequence E1-E31, bearing 2 glycosylations and 2 hydroxylations.

**Figure 8.** Deconvoluted and deisotoped MS/MS spectrum acquired on the whole C1q A chain. Detection of complementary N- and C-terminal regions from C1q A is indicated. Inset: zoom on the loss of one sialic acid group.

**Figure 9.** MS$^2$ spectrum acquired on the ionic species at about m/z 1102.41 (10+) detected during top-down MS analysis of reduced CLR. Fragmentation was performed in the LIT and detection of fragments was carried out in the Orbitrap cell.

**Figure 10.** Three-dimensional model of human C1q highlighting the position of unmodified lysine residues. The model (2) was assembled as described previously (50). The C1q chains are coloured dark blue (A), green (B), and light blue (C). The positions of the side chains of Lys A59, Lys B61, and Lys C58 (unmodified), and of Lys B65 (hydroxylated) are shown.
Table 1: Sequences and PTMs attributed to the ionic signals detected during top-down analyses of reduced C1q and CLR chains.

? indicates that the C1q B chain was only detected by a weak signal at a single mass.

| Sample | Species name | Detected masses (mass units) | Matched sequence stretch | Modifications | Mass error (exp-th)/th (ppm) | Variable numbers of hydroxylations |
|--------|--------------|------------------------------|--------------------------|---------------|------------------------------|----------------------------------|
| CLR    |              |                              |                          |               |                              |                                  |
| A-1    |              | 11410.54                     | 1-97 from C1q A          | 8*, 5#        | 8.6                          | 6* to 9*                         |
| A-2    |              | 11252.45                     | 1-95 from C1q A          | 8*, 5#        | 6.9                          | 6* to 9*                         |
| A-3    |              | 11070.40                     | 1-97 from C1q A          | 8*, 4#        | 5.3                          | 6* to 9*                         |
| A-4    |              | 10912.35                     | 1-95 from C1q A          | 8*, 4#        | 7.2                          | 6* to 9*                         |
| B-1    |              | 11008.10                     | 1-97 from C1q B          | 12*, 5#, -NH₃ | 5.0                          | 11* to 14*                       |
| B-2    |              | 10684.02                     | 1-97 from C1q B          | 13*, 4#, -NH₃ | 7.5                          | 11* to 14*                       |
| C-1    |              | 10516.91                     | 1-94 from C1q C          | 14*, 3#       | 6.6                          | 12* to 15*                       |
| C-2    |              | 10055.62                     | 1-90 from C1q C          | 14*, 3#       | 0                            | 12* to 15*                       |
| C-3    |              | 10369.75                     | 1-93 from C1q C          | 14*, 3#       | -2.9                         | 12* to 15*                       |
| C-4    |              | 10270.82                     | 1-92 from C1q C          | 14*, 3#       | 10                           | 12* to 15*                       |
| C1q    | A            | 27561.24                     | C1q A                    | 8*, 5#, 1 N-glycan | 2.5                       | 7* to 9*                         |
| B      |              | 25602.34                     | C1q B                    | 12*, 5#, -NH₃ | 4.1                          | ?                                |
| C      |              | 24042.64                     | C1q C                    | 14*, 3#       | 2.1                          | 12* to 15*                       |
Table 2: Peptide sequences identified from iterative MS^n analyses of reduced CLR C1q B and CLR C1q C, by direct infusion on the LTQ-Orbitrap instrument. (a) The m/z value of the isotope of highest intensity in the pattern selected for fragmentation is provided. %: indicates loss of NH$_3$ by the N-terminal glutamine.

| Chain | Precursor ion for FTMS$^2$ m/z (a) | Precursor ion for ITMS$^3$ m/z | Identified sequences | Further ITMS$^4$ |
|-------|-----------------------------------|--------------------------------|---------------------|----------------|
| B     | 1102.41 (10+)                     | 854.3670 (1+)                  | $^{1}\%$QLSCTGPP*A  | /               |
|   | 1104.02 (10+)                     | 967.4504 (1+)                  | $^{1}\%$QLSCTGPP*A1 | /               |
|   |                                  | 1250.6014 (1+)                 | $^{1}\%$QLSCTGPP*PAIP*GI | /               |
|   |                                  | 1533.7533 (1+)                 | $^{1}\%$QLSCTG(PP)*AIP*GIP*GI | /               |
|   |                                  | 535.3027 (2+)                  | $^{88}$YKATQKIAF$^{57}$ | /               |
| C     | 1256.71 (8+)                      | 1103.5312 (2+)                 | $^{77}$PGP*GIP*GEPGEEGRYKQKF$^{90}$ | /               |
|   |                                  | 736.0229 (3+)                  | $^{77}$P*GPMGIP*GEPGEEGRYKQKF$^{90}$ | /               |
|   |                                  | 819.3975 (2+)                  | $^{77}$P*GEPGEEGRYKQKF$^{90}$ | /               |
|   | 1258.71 (8+)                      | 1111.5269 (2+)                 | $^{77}$P*GPMGIP*GEP*GEEGRYKQKF$^{90}$ | /               |
|   |                                  | 741.3538 (3+)                  | $^{77}$P*GPMGIP*GEP*GEEGRYKQKF$^{90}$ | /               |
|   |                                  | 827.3938 (2+)                  | $^{77}$P*GEP*GEEGRYKQKF$^{90}$ | /               |
|   | 1314.23 (8+)                      | 1050.0091 (2+)                 | $^{77}$P*GEPGEEGRYKQKFSVF$^{94}$ | /               |
|   |                                  | 889.7638 (3+)                  | $^{77}$P*GPMGIP*GEPGEEGRYKQKFSVF$^{94}$ | 399.17 |
|   | 1316.38 (8+)                      | 1058.0063 (2+)                 | $^{77}$P*GEP*GEEGRYKQKFSVF$^{94}$ | /               |
|   |                                  | 895.0957 (3+)                  | $^{77}$P*GPMGIP*GEP*GEEGRYKQKFSVF$^{94}$ | 569.33 |
Figure 1

A

GLOBULAR HEAD

C1q + C1s-C1r-C1r-C1s → C1

B

Collagen-like region residues 9-87

Globular head A chain

Globular head B chain

Globular head C chain

Hinge region

GIRIGIQG
Figure 2

C1QA

MEGPRGWLVL CVLAISSLASM VTEDLCRAPD GKKGEAGRPG RRGRPGLKGE QGEPCAPGIR
c1qC

MDVGPSLPH LGLKLLLLL LLLDRGQA

2
12
22
32

GCGIPGMPG LPGAPGKDGY DGLPGPKGEP

42
52
62
72
82
92

GIPAIKPGIRG PKGOKGEPEGL PGHPKNGPM GPPGMPGVPG PMGIPGEPEG EGRYQKLFQ

102
112
122
132
142
152

VFTVTRQTHQ PPAPNSLIRF NAVLTPQGD YDSTGKFTC KVPGLYFVY HASHTANLVC

162
172
182
192
202
212

LLYRSGKVV TFCGHTSKTN QVNSGGVLLR LVQGEETWLA VNDYIDMVG QGSDVFSGFE

---

LLFPD
Figure 3

C1QA

MEGPRGWLVLCVLAISLASM VTEDLCRAPD GKKGEAGRGPG RRGRPGKGLKGE QGEPPAGPGIR

TGIGQGLKGDQ GEPGPSGNGP KVGYPGPSGP LGARGIKPGIK GTKGSPGNIK DQPRPFSNAI
C1QB

MKIPWSGQPV LMLLLLLGIL DISOQQLSCT GPFAIPGIPG IPGTPGPQDGQ PGTPGKGEK

GLPGLAGDHG EFGKKGDPQ PGNPGKVGPK GPMGPKGGPG APGAPGPKGGE SGDYKATQKI
C1QC

MDVGPSLPH LGLKLLLLLL LPLRQANT GCYGPMPG LPGAPGDGY DGPLPGKGE

GPAIPGIRG PKQKGEPGL PGHPKNPGPM GPMGPGVPG PMGIPEGE PG EGRYKQKFQS
Figure 4

C1QA_HUMAN

MEGPRGWLVL CVLAISLASM VTEDLCRAFD GKKGEAGRPG RRGRPGLKGE QGEPGAPGIR

TGIQGLKGDQ GEPGPSGNPG KVGYPGPSGP LGARGIPGKGTKGPSGNIK DQPRPAFSAIR
Figure 5
Figure 6B

![Mass spectrum graph with labeled peaks A, B, and C.
Peaks have m/z values as follows:
- A: 25602.34, 26097.51, 26653.87, 27205.29
- B: 24042.64
- C: 27561.24]
Figure 8

N-terminal part of C1q A

C-terminal part of C1q A

Loss of 1 sialic acid
($\Delta m = 291.07$)

[Graph and data points indicating mass-to-charge ratios (m/z) and relative abundances]
Figure 9A
Figure 9B
Figure 10