HPLC enrichment/isolation of proteins for post-translational modification studies from complex mixtures

Eszter Tóth[1], Olivér Ozohanics[1], Balázs Bobály[1,2], Ágnes Gömöry[1], Anita Jekő[1], László Drahos[1], Károly Vékey[1*]

[1]Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2., Budapest 1117, Hungary

[2]Budapest University of Technology and Economics, Department of Inorganic and Analytical Chemistry, Szt. Gellért tér 4., Budapest 1111, Hungary

ABSTRACT

The paper describes a macroporous RP-HPLC method for separation and isolation/enrichment of proteins from complex mixtures. The method is robust and efficient; using 2.1 or 4.6 mm diameter columns provides sufficient material for subsequent proteomic analysis. The main advantage of the method is that most protein variants are isolated in the same fraction, as separation is not based on differences in isoelectric point. This is highly advantageous for studying complex mixtures and post-translational modifications. Examples related to glycosylation analysis are discussed in detail.

[*]To whom correspondence should be addressed:

20   Károly Vékey

21   Research Centre for Natural Sciences, Hungarian Academy of Sciences

22   H-1519 Budapest, P.O. Box 286. Hungary

23   E-mail: vekey.karoly@ttk.mta.hu

24   Phone: +36-1-382-6515

25

26   **HIGHLIGHTS:**

27   • Macroporous RP-HPLC for enrichment/isolation of proteins from complex mixtures

28   • Non-biased protein enrichment for analyzing glycosylation

29   • Sample preparation method for glycosylation analysis

30

31   **KEYWORDS:**

32   Mass spectrometry; HPLC; sample preparation; proteins; glycosylation

1   Introduction

The structural and functional diversity of proteins is, in a large part, due to post-translational modifications (PTM, like glycosylation, phosphorylation, ubiquitination). One of the most common PTM of proteins is glycosylation, which is implicated in various biological functions like transport of proteins, immune response and cell-cell communication. Relation between glycosylation of human plasma proteins and various diseases has also been shown [1, 2].

Blood plasma and other biological samples are very complex mixtures; often containing a wide range of proteins. Due to this complexity, characterization of the glycosylation pattern of plasma proteins typically requires the use of a range of sample preparation and/or isolation methods, like immunoaffinity isolation, plasma fractionation, glycoprotein enrichment etc.

Glycoprotein/glycopeptide enrichment is often performed using various lectins. This is an excellent choice for enriching glycoproteins. However, selective binding of diverse glycoforms influences enrichment, and may cause significant bias in the glycosylation pattern determined [3].

Probably the most efficient way to isolate *a single protein component* of a complex mixture (e.g. in order to determine its glycosylation pattern) is based on immunoaffinity binding. However, antibodies are often not available, they may have non-specific binding properties, and are variable among different producers [4, 5]. When analysis of glycosylation pattern (or other PTMs) of several glycoproteins is required, immunoaffinity may not be a feasible choice for protein isolation.

53   Enrichment and/or fractionation of the complex protein mixture is also a common alternative.

54   For analysis of *large number of diverse proteins* the best choice is often reducing the complexity

55   of the biological mixture by fractionation. There are many possibilities, each with advantages

56   and disadvantages. Multidimensional separation techniques (like 2D gel electrophoresis) are

57   more efficient, but are time consuming, and often not feasible for high throughput analysis [6, 7].

58   Fractionation can be performed either on the intact protein mixture, or following the digestion of

59   the proteins [8]. The latter case typically involves proteolytic enzymes (like trypsin).

60   Fractionation at the peptide level has the advantage that there are a wide variety of well-

61   established analytical methodologies for separation of small and medium size molecules. The

62   disadvantages are, that the protein digest is even more complex than the original mixture; and

63   different components of the same protein are collected in different fractions.

64   Fractionation at the protein level is also feasible [7, 9], although the analytical methodologies are

65   not so well defined. Most frequently ion exchange, size exclusion, hydrophobic interaction,

66   partition, affinity chromatography, dye-ligand chromatography and electrophoretic methods are

67   applied [10, 11]. These techniques are efficient, and often separate not only the proteins, but also

68   various protein variants: sequence variants (e.g. genetic variants); diverse PTMs; and among the

69   latter various glycoforms.

70   When the intention is the study of different PTMs of the same protein (like the glycosylation

71   pattern) collecting all variants of a given protein in one fraction is often considered highly

72   advantageous. This typically precludes the use of ion-exchange chromatography or

73   electrophoretic methods; as these separate proteins based on the number of charges (and

74   therefore glyco- or phosphoproteins with the different number of charged groups) [12]. On the

75   other hand, the use of reversed-phase HPLC is most likely advantageous from this point of view,

76   as it separates proteins with high efficiency, but the different protein variants are expected to

77   elute together. In the present paper we focus on this approach, and apply it to the study of site-

78   specific glycosylation patterns and genetic variants.

79   In order to enhance performance of HPLC, first we deplete human plasma of the most abundant

80   proteins. In the present case only human serum albumin (HSA) and immunoglobulin G (IgG) are

81   depleted, because we would like to study the glycosylation of some high abundance proteins, as

82   well. Note that, depletion of other high abundance proteins further increases the ability to

83   analyze glycosylation of low abundance components. Note also that while immunoaffinity

84   binding has its limitations, for the purpose of plasma depletion this is an excellent choice; as the

85   objective is only to remove the bulk of the most abundant plasma proteins. Reversed phase

86   HPLC of intact proteins was made feasible by the introduction of large pore size (macroporous)

87   columns. While this is not novel technology, nevertheless requires extensive optimization to

88   obtain good performance [13-15]. After optimizing HPLC performance, fractions were collected,

89   containing a few, highly enriched proteins. Site-specific glycosylation patterns were determined

90   for the individual components, using tryptic digestion and nano-HPLC-MS(/MS) analysis. Due

91   to the high specificity of MS/MS, determination of the site-specific glycosylation pattern of even

92   *several plasma proteins* is feasible.

93

2   Material and methods

2.1   Samples and chemicals

Human serotransferrin standard and human alpha-1-acid glycoprotein standard were purchased from Sigma-Aldrich[®] (St. Louis, MO, USA). Haptoglobin standard was purchased from Fluka Chemie GmbH (Sigma-Aldrich[®], Zwijndrecht, Netherlands). Human blood plasma sample was obtained from a healthy volunteer, it was divided into aliquots and stored at -20 °C.

1,4-dithio-L,D-threitol (DTT) and 2-iodoacetamide (IAA) were purchased from Fluka Chemie GmbH (Sigma-Aldrich[®], Zwijndrecht, Netherlands). RapiGest SF (lyophilized sodium-3-[(2-methyl-2-undecyl-1,3-dioxolan-4-yl)-methoxyl]-1-propane-sulfonate) was obtained from Waters (Milford, MA, USA). Mass spectrometry grade trypsin (Promega Corporation, Madison, WI, USA) and proteomics grade PNGase F from Elizabethkingia meningoseptica (Sigma-Aldrich[®], St. Louis, MO, USA) were used.

All other reagents were purchased from Sigma-Aldrich[®] (St. Louis, MO, USA).

2.2   Removal of albumin and IgG

To reduce sample complexity, two high abundance proteins: albumin and IgG were removed from human plasma with Agilent Multiple Affinity Removal Spin Cartridge HSA/IgG (Agilent Technologies, Santa Clara, CA, USA).

According the manufacturer's standard protocol 50 µL plasma sample was loaded onto the cartridge and instructions were followed. In addition, at the end of the depletion steps K_2HPO_4

114   and citric acid were added to the diluted samples to prevent aggregation (30-30 mM

115   concentration in the samples), and samples were concentrated with 10 kDa centrifuge filters.

116   After filtration, volume of the depleted plasma sample was 25 µL (i.e. half of the original plasma

117   sample volume). We have checked by MS based standard proteomics experiments that (1) 66%

118   of the total protein amount was removed from the sample; (2) 80% of the total albumin amount

119   is removed from the sample.

120

2.3   RP-HPLC fractionation of depleted plasma samples

122   Proteins were isolated from depleted plasma samples using HPLC: Acquity UPLC$^{®}$ System

123   (Waters, Milford, MA, USA) with Binary Solvent Manager, Sample Manager, Column Heater

124   and TUV-detector.

125   The following macroporous HPLC columns were tested: Agilent mRP-C18 High-Recovery

126   Protein Column (4.6 × 50 mm, Agilent Technologies, Santa Clara, CA, USA); Poros R2

127   (Poly(Styrene-Divinylbenzene), 10 μm, 2.1 × 100 mm, Applied Biosystems, Foster City, CA,

128   USA); BioSuite pC18 Column 500 (7 μm, 2.0 × 150 mm, Waters, Milford, MA, USA), Acquity

129   UPLC BEH300 (C18, 1.7 μm, 2.1 × 100 mm, Waters, Milford, MA, USA) and Aeris^{TM}

130   WidePore XB-C18 (3.6 μm, 2.1 × 150 mm, Phenomenex, Torrance, CA, USA). We have

131   selected the Poros R2 column for detailed optimization based on the preliminary tests.

132   Tests were performed using standard proteins (alpha-1-acid glycoprotein, serotransferrin,

133   haptoglobin), deglycosylated standard proteins and depleted plasma sample. Deglycosylated

134   alpha-1-acid glycoprotein and deglycosylated serotransferrin were prepared the following way:

135 to 11 μL sample, containing 300 pmol protein standard, 2 μL NH_4HCO_3 (200 mM) and 10 μL

136 PNGase F (500 U/mL) were added and digested at 37 °C overnight.

137 Standard and deglycosylated standard proteins were injected in the 30 pmol – 5 nmol range,

138 depleted plasma samples were injected in 0.5-15 μL volume.

139 Optimized gradient for depleted plasma samples (Poros R2 column): The column temperature

140 was 65 °C, the flow rate was 1 mL/min. The gradient started with 20% B, kept here for 0.7 min,

141 after this followed a 15 min long gradient from 20% to 70% solvent B. The next step was

142 increasing to 95% solvent B in 0.1 min, kept there for 1.5 min (washing). Finally returning to

143 20% B in 0.1 min, and kept there for 6 min (equilibration). Solvent A was water containing 0.07

144 v/v% trifluoroacetic acid and solvent B was acetonitrile containing 0.07% v/v trifluoroacetic

145 acid.

146 Samples of 500 μL fractions were collected manually using the timer of the MassLynx software

147 and UV-detection (280 nm). To each fraction 1.3 μL NH_3 solution (25 w/w%) to neutralize TFA,

148 and 1.8 μL K_2HPO_4 (500 mM) to prevent aggregation was added and fractions were concentrated

149 to 30 μL with SpeedVac (miVac Duo Concentrator, Genevac Ltd., Ipswich, Suffolk, UK).

150

151 **2.4 In-solution digestion**

152 Samples of 30 μL volume (collected/concentrated as described above) were digested by the

153 protocol optimized for small sample volume and described before [16]. The main steps were the

154 following: 5 μL NH_4HCO_3 (200 mM) was added to the sample, proteins were unfolded and

155 reduced with 3 μL RapiGest SF (0.5 w/v%) and 1.5 μL DTT (100 mM) for 30 min at 60 °C.

8

156 Alkylation was performed adding 4 μL NH$_4$HCO$_3$ (200 mM) and 2 μL IAA (200 mM) for 30

157 min at room temperature in dark. Fractions were digested with 0.5–1.5 μL trypsin (40 μM) for

158 180 min at 37 °C. The digestion was stopped by adding 1.5 μL formic acid, followed by 30 min

159 incubation time at 37 °C. The samples were centrifuged at 13500 rpm (corresponding to 17000

160 g) for 10 min.

161

162 2.5 nano LC-MS(/MS) measurements

163 The samples prepared as described above were subjected to several nano-HPLC-MS(/MS)

164 experiments. The first objective is to determine the protein composition of the plasma fraction

165 collected; the second to identify glycoforms of the major glycoproteins in the given fraction; the

166 third is to determine the glycosylation pattern, i.e. the relative proportion of various glycoforms

167 of the same protein.

168 The digested fractions were analyzed using nanoAcquity UPLC (Waters, Milford, MA, USA)

169 coupled to a high resolution QTOF Premier mass spectrometer (Waters, Milford, MA, USA).

170 The chromatographic conditions were during all MS(/MS) measurements the following: the

171 peptides were separated on a reversed-phase analytical column (C18, 1.7 μm BEH particles, 75

172 μm i.d. × 200 mm, Waters, Milford, MA, USA). The column temperature was 55 °C. Before the

173 analytical column a Symmetry C18 trap column (180 μm i.d. × 20 mm, Waters Milford, MA,

174 USA) was used. Solvent A was water containing 0.1 v/v% formic acid and solvent B was

175 acetonitrile containing 0.1 v/v% formic acid. The gradient was the following: using 250 nL/min

176 flow rate starting with a 4 min gradient from 3% to 8% B, then a 65 min long gradient going to

177 40% solvent B. Washing was done using 450 nL/min flow rate and a 2 min long gradient from

178 40 to 75% solvent B, and kept there for 18 min. After that returning to 3% B in 2 min, and

179 equilibration was done for 18 min.

180 To determine the protein composition of the collected macroporous HPLC fractions, the

181 fractions were digested as described above and the resulted peptides were identified by data

182 dependent analysis with the most intense 3 peaks selected. The mass spectrometer operated in

183 positive electrospray ionization mode. The capillary voltage was 2.3 kV, nanoflow 1 bar, source

184 temperature 90 °C, cone voltage 35 V. The parent ion was selected in the 400–1800 m/z range,

185 MS/MS spectra were acquired in the 50-2000 m/z range. Collision gas was argon, at 4.05×10^{-3}

186 mbar. Collision energy was varied in the 7-70 eV range. Genetic variants of alpha-1-acid

187 glycoprotein were identified using these conditions, as well.

188 Glycosylation sites and major site-specific glycoforms were identified with tandem mass

189 spectrometry. The parent ion was selected in the 780-2000 m/z range, MS/MS spectra were

190 acquired in the 150-3000 m/z range. Collision energy was varied in the 5-55 eV range. Minor

191 glycoforms and relative quantitation were measured with single stage mass spectrometry in

192 extended dynamic range mode. Scans were acquired in the 500-2000 m/z range. Other

193 instrumental parameters were same as described above.

194 Site-specific glycosylation pattern of components of the fractions has been analyzed and

195 characterized by the method described earlier in detail [17].

196

2.6 Data evaluation

Proteins in each fraction and genetic variants of alpha-1-acid glycoprotein were identified from MS/MS measurements using ProteinLynx Global Server v.2.3 (Waters, Milford, MA, USA) and searched against v.2011_10 of SwissProt sequence database with human taxonomy using Mascot Server v.2.2 (Matrix Science, London, UK). One missed cleavage was allowed, carbamidomethyl cysteine was set as fixed modification.

MS/MS spectra corresponding to major glycopeptides were automatically evaluated by our computer software GlycoMiner v.1.13 Beta [18]. For each glycosylation site, the tryptic peptides with possible glycan structures were identified. Minor glycopeptide identification and relative glycopeptide quantitation were performed from MS measurements using the in-house developed computer software GlycoPattern v.2.0 [17]. The software calculates the retention time windows of minor glycoforms based on the known retention time of the major glycoforms. Minor glycoforms are identified based on retention time and exact mass. Internally, GlycoPattern uses an isotope model to increase the confidence of the identification. For each glycoforms the extracted ion chromatograms are generated and the chromatographic peaks are integrated.

3 Results and discussion

In the current work we have optimized a workflow for RP-HPLC fractionation of complex protein mixtures, in particular human blood plasma. The primary aim was to collect the different variants of a given protein in one fraction. Fractionation significantly simplifies sample complexity, and is particularly advantageous for analyzing protein variants, like various PTMs.

218 Still, even after fractionation, each collected fraction is a complex mixture, containing more than

219 one co-eluting protein and their variants. When all variants of a given protein are collected in one

220 fraction, this makes subsequent identification, qualitative and quantitative analysis of PTMs

221 more reliable, more robust and more sensitive. Chromatographic behavior of proteins under RP

222 conditions are expected to depend mainly on their hydrophobicity; and in most cases on their

223 peptide backbone; single point mutations or changes in the PTMs should only have a minor,

224 often negligible effect. This expectation will be studied in detail below. Note that RP-HPLC

225 fractionation is only the first step, PTM characterization typically requires several further stages

226 for sample preparation and analysis. This may be based on procedures which separate various

227 PTMs from each other (like in 2D gel electrophoresis); or on procedures (like mass

228 spectrometry), which are capable of dealing with a mixture of various PTMs.

229 Important parameters and properties of the fractionation method were tested and optimized. The

230 main steps of fractionation and analysis are the following: affinity depletion of high abundance

231 proteins; RP-HPLC fractionation of the remaining proteins; proteolysis (tryptic digestion) of the

232 collected fractions; nano-HPLC-MS(/MS) measurements; and complex data analysis to

233 determine site-specific glycosylation pattern. The complete workflow is presented in Fig. A.

234 The first issue to consider is the amount of proteins needed to be isolated/enriched. This, in turn,

235 depends on the objective and methodology of further analysis or utilization. In the present case

236 the objective is analytical characterization based on mass spectrometry. This methodology is

237 sufficiently sensitive, that using 4.6 or 2.1 mm diameter analytical HPLC (or UHPLC) columns

238 for sample preparation is adequate – although the loading capacity should be tested and

239 optimized. When larger sample amounts are needed for subsequent analysis (e.g. for most 2D gel

240 electrophoresis studies), multiple HPLC runs or the use of larger diameter column may be

241 required.

242 The whole analytical procedure starts with depletion of the blood plasma sample (removal of the

243 most abundant protein components). This is a commonly used method to simplify subsequent

244 analysis. Typically the most abundant 2, 7, 14 or 20 proteins may be removed by affinity

245 binding, using various kits. In the present study we use the simplest variant, removing albumin

246 and IgG. It is often recommended that blood plasma should be reduced and alkylated before

247 depletion and HPLC analysis. This somewhat improves chromatographic behavior (resolution

248 and shape of peaks). However, we found that following reduction and alkylation new peaks

249 appear in the chromatogram (most likely due to changes of the chromatographic contact area);

250 and a significant amount of the reduced protein precipitates from the solution. (This may partly

251 be avoided by the addition of K_2HPO_4 and citric acid to the solution; and depends significantly

252 on the individual plasma sample). To avoid these problems, we have used (and recommend)

253 depletion and subsequent RP-HPLC fractionation on the native plasma, i.e. *without* reduction

254 and alkylation.

255 The range of macroporous columns is relatively limited, nevertheless there are several types

256 available. We have made a tentative comparison of five different columns, mainly with respect to

257 semi-preparative applications (Agilent mRP-C18, BioSuite pC18, Poros R2, Acquity UPLC

258 BEH300, Aeris[TM] WidePore XB-C18). Our primary concern was their loading capacity and

259 avoiding cross-contamination of samples (i.e. a blank injected after a plasma sample should not

260 contain any proteins). Based on preliminary tests we have selected the Poros R2 column for

261 detailed optimization – but detailed comparison of these columns is outside the scope of the

262 present paper.

263   Length and slope of the gradient, trifluoroacetic acid content of solvents and flow rate have been

264   optimized for the best achievable resolution, while keeping the peak/fraction volume manageable

265   [19]. Trifluoroacetic acid content of solvents should be low; best results were obtained in the

266   0.05-0.1% range; the optimum depends on column type. Without trifluoroacetic acid the proteins

267   remain bound to the column, but higher concentration of trifluoroacetic acid decreases the pH

268   extremely, which results the precipitation of proteins and hydrolization of sugar bonds. In order

269   to avoid unnecessary dilution of the sample the flow rate was limited to 1 mL/min, and elution

270   time was 20 min. Using these parameters reasonable selectivity was obtained, even injecting as

271   much as 15 µL sample (Fig. B).

272   Using the conditions described above proteins were collected in 20, 0.5 min wide, 500 µL

273   volume fractions; from 4 to 14 min retention time based on the chromatogram shown in Fig. B.

274   Prior to subsequent sample analysis NH_3 solution and K_2HPO_4 solution were added to neutralize

275   trifluoroacetic acid and to prevent further precipitation. Then, the 500 µL fractions were

276   concentrated to 30 µL volume using vacuum centrifugation. Protein recovery of the complete

277   procedure has been tested with standards and optimized for the minimal sample loss. It was

278   found that special tubing: ethylene tetrafluoroethylene (ETFE), needle: fluorinated ethylene

279   propylene (FEP), and low protein affinity eppendorf tubes were needed to obtain over 90%

280   protein recovery for major plasma proteins.

281   Reproducibility of the chromatographic method has been tested with various protein standards

282   and plasma samples using UV-detection. Reproducibility (standard deviation) of retention times

283   is 0.15%; that of the peak areas is 1.4%. Reproducibility of the whole procedure (which includes

284   plasma depletion as well); with respect to retention times is the same (0.15%); but that of peak

285   areas is significantly worse, 9.3%. This shows that plasma depletion is less reproducible than

286   chromatographic analysis. The standard deviations shown are the averages measured for three

287   protein standards (and the same protein components of plasma), and were calculated using 3

288   replicates.

289   As described above, using macroporous RP-HPLC one can reasonably expect that different

290   variants of a given protein should elute in one chromatographic peak, i.e. the PTMs have only

291   minor influence on the retention time. As this is of major importance in the present study, we

292   have performed various tests on genetic variants and on glycoproteins to confirm this

293   assumption. In case of glycosylation, glycosylated and deglycosylated proteins; and glycoforms

294   containing acidic and neutral sugars were studied.

295   In the case of standard (commercial) serotransferrin sample, the UV chromatogram in Fig. C

296   shows a single peak (although with a well-defined shoulder). It is the well-known behavior of

297   commercial serotransferrin standard, might be attributed to natural charge or oxidized variants

298   [11]. We have fractionated this peak into 5 narrow (0.1 min, 100 µL) fractions; digested these

299   with trypsin, and studied them in conventional nano-HPLC-MS and MS/MS experiments (details

300   are described in the Material and methods part, 2.5). The results show, that all fractions contain

301   serotransferrin; only one genetic variant was observed. We have also determined that the

302   glycosylation pattern of serotransferrin does not show changes among the various fractions.

303   Deglycosylated serotransferrin standard was injected onto the RP-HPLC column as well, and

304   found to behave very similar to the glycosylated standard: peak shape and retention time were

305   identical. We have also studied the effect of deglycosylation in the case of alpha-1-acid

306   glycoprotein, in which case ~40% of the molecular mass is due to the sugar chains. In this case

307   the retention time did change, but only by 0.5 min. This suggest, that even when most of the

308   protein surface is covered by sugar units, the retention time is predominantly determined by the

309 peptide backbone. Note also that changes in the glycosylation pattern (in contrast to complete

310 deglycosylation) of the protein has a negligible effect on retention, as described below with

311 respect to glycoform analysis.

312 We have performed an analogous study in the case of blood plasma. We have tried to follow

313 elution of serotransferrin in the plasma sample, but due to the complexity of the electrospray

314 signals the results were equivocal. To confirm the assumption that various glycoforms of

315 serotransferrin in the plasma sample indeed elute in a single chromatographic peak the following

316 test was performed. The RP-HPLC fractions were collected as described above. The fractions

317 were digested using trypsin, and the intensity of the molecular ions of the three most abundant

318 peptide fragments of serotransferrin were determined using a conventional nano-HPLC-MS

319 experiment. The relative amount of serotransferrin in each fraction was determined using label-

320 free quantitation; using the average abundance of the three most abundant unique peptide signals.

321 The result is shown in Fig. C (dots marked, scale at the right side). This clearly indicates that

322 serotransferrin is collected predominantly in one fraction only; a minor amount (ca. 2%) in a

323 neighboring fraction, while other fractions contain only traces of serotransferrin (either due to

324 peak tailing or cross-contamination).

325 Chromatographic behavior of proteins containing multiple amino acid modifications were

326 studied in the case of alpha-1-acid glycoprotein. This protein has two different genetic variants

327 (ORM1 and ORM2), and has 21 variable amino acid sites (out of 200 amino acids). After

328 fractionation we have digested the protein mixtures, and identified peptide fragments

329 characteristic for the two genetic variants. Two peptide pairs were selected for analysis,

330 ^{171}SDVVYTDWK^{179} and ^{154}EQLGEFYEALDCLR^{167} for ORM1; and ^{171}SDVMYTDWK^{179} and

331 ^{154}EQLGEFYEALDCLCIPR^{170} for ORM2. The peptides were identified in a conventional

332   proteomics experiment using nanoHPLC-MS/MS, as described in the Material and methods part.

333   All of these peptides were found in the same protein fraction, indicating that the protein variants

334   co-elute under the studied conditions.

335   Reproducibility of fraction collection was also studied. This is illustrated the case of beta-2-

336   glycoprotein_1; shown in Fig. D. Fractions were collected in three separate experiments; and the

337   relative amount of beta-2-glycoprotein_1 was determined in each fraction using label-free

338   quantitation; as described above by serotransferrin. The results show that beta-2-glycoprotein_1

339   is collected predominantly in one fraction only. Standard deviation of relative intensities in the

340   various fractions is 0.8% SD, using 3 replicates.

341   Combining the presently described fractionation method with glycoprotein analysis [17], we

342   have successfully characterized the detailed, site-specific glycosylation pattern of several plasma

343   proteins. The reproducibility of glycosylation patterns (relative abundance of various

344   glycoforms) so determined in three replicates is, on average, 15% SD. This is identical to that

345   described before for standard protein samples [17].

346   Most studied proteins gave results similar to those discussed above (one chromatographic peak).

347   One notable difference is haptoglobin; which showed a single peak having a very (8 min) long

348   tail (probably due to strong column binding; observed in the standard, using UV detection, Fig.

349   E.1). We have collected the various fractions from plasma; after digestion and nano-HPLC-

350   MS(/MS) analysis we have determined the glycosylation pattern of haptoglobin in the various

351   fractions; shown in Fig. E.2. This illustrates that, even if proteins show large tailing, this is

352   predominantly not due to separation of glycoforms. This confirms that proteins collected in one

353   fraction are suitable for determining representative glycosylation patterns.

354   We have performed a further experiment to show that RP-HPLC fractionation does not

355   compromise determination of glycosylation patterns. In other words, the glycosylation pattern

356   determined after fractionation is the same as that determined on the original sample. Due to

357   limited sensitivity, this was checked on a standard protein (alpha-1-acid glycoprotein). The

358   glycosylation pattern at site ^{56}N (corresponding to the ^{52}NEEYNK^{57} glycopeptide) was

359   determined for the commercial standard protein without fractionation, and after fractionation

360   using the methodology discussed above. The results are shown in Fig. F, and show that

361   fractionation does not change the glycosylation pattern (e.g. due to selective binding or some

362   other effects). Between the two samples 13% SD was calculated from all glycopeptides of alpha-

363   1-acid glycoprotein, using 3 replicates.

364   As described above, loading capacity was a major issue optimizing the present methodology.

365   This determines the amount of proteins isolated/fractionated by RP-HPLC for further studies.

366   Determination of PTMs is often limited by sensitivity, even if high sensitivity mass spectrometry

367   is used for analysis. In most proteomics experiments it is sufficient to identify a few (the most

368   sensitive) peptide fragments of protein to give positive identification and relative (label-free)

369   quantitation. Analyzing PTMs requires a far, 100-1000-10 000 fold larger sample amount than

370   that required for protein identification for the two main reasons: (1) Peptides containing the PTM

371   unit have often limited sensitivity. This is a particularly serious issue for glycopeptides. (2) Many

372   PTMs are minor components, e.g. phosphorylation may be present in a few %; minor (but

373   possibly biologically important) glycoforms have 1% or less relative abundance.

374   Using the Poros R2 2.1 mm diameter column allowed us to inject ca. 15 µL depleted plasma onto

375   the column, without significantly deteriorating column performance. The total protein amount in

376   this is approximately 800 µg (~13 nmol). This calculation is based on the following: Total

377 protein content of plasma is ~80 g/L; from which ~66% is removed by depletion, so ~27 g/L

378 remains in the sample. In the experiments 50 µL plasma is used, this means ~1300 µg protein in

379 the 25 µL depleted fraction. Onto the column 15 µL of 25 µL is injected, which is ~800 µg.

380 Assuming a 60 kDa average protein molecular mass, this converts to ~13 nmol total protein

381 amount injected. This means that major plasma proteins can be collected in one fraction in 20–

382 1000 pmol amount. This is typically sufficient to obtain (after proteolytic digestion) high quality

383 glycosylation patterns using nano-HPLC-MS(/MS) methodology. In order to perform 2D gel

384 electrophoresis experiments using the common Coomassie Staining typically larger protein

385 amounts are needed. This may require repeated sample fractionations, and pooling several

386 fractions for subsequent analysis.

387

4   Conclusions

389 The paper demonstrates the usefulness of RP-HPLC methodology for sample preparation,

390 enrichment and (in certain cases) isolation of proteins from complex mixtures. The use of

391 macroporous stationary phases is the key to utilize HPLC for separation of macromolecules. This

392 methodology is not new [9, 20], but the availability of efficient, commercially available columns

393 is a relatively recent development. Most other protein separation methods (like capillary

394 electrophoresis or 2D gel electrophoresis) rely on isoelectric focusing, which is a very efficient

395 way for separating not only proteins, but protein variants as well. RP-HPLC separates protein

396 variants in much less degree: on the one hand this restricts the separation power (or peak

397 capacity); on the other hand this does not separate most protein variants. This was shown in case

398 of genetic variants and glycoforms, and other PTMs (like phosphoproteins) are expected to

399 behave similarly. As protein variants are not resolved, this facilitates unbiased determination of

400 the PTM distribution, which is very important in PTM studies.

401 In the present paper we have shown that various glycoforms are collected in the same fraction

402 and protein isolation in this way does not introduce bias. This was established comparing

403 glycosylation patterns with and without fractionation, and these were (within experimental error)

404 identical. Mass spectrometry (arguably the most common method for protein analysis) requires

405 only small sample amounts. Therefore, one injection onto a 2.1 or 4.6 mm diameter (analytical,

406 high efficiency) HPLC column usually provides a sufficient amount of sample for subsequent

407 analysis. The suggested macroporous RP-HPLC method is robust; which facilitates widespread

408 applications.

409

410 **5   Acknowledgement**

413

6   References

[1] C. Lebrilla, Glycan Biomarkers in Cancer Diagnosis, Journal of Analytical Science & Technology, 2 (2011) 38-45.

[2] D. Meany, D. Cha, Aberrant glycosylation associated with enzymes as cancer biomarkers, Clin Proteom, 8 (2011) 1-14.

[3] O. Ozohanics, L. Turiak, L. Drahos, K. Vekey, Comparison of glycopeptide/glycoprotein enrichment techniques, Rapid Communications in Mass Spectrometry, 26 (2012) 215-217.

[4] J.T. Koerber, N.D. Thomsen, B.T. Hannigan, W.F. Degrado, J.A. Wells, Nature-inspired design of motif-specific antibody scaffolds, Nat. Biotechnol., 31 (2013) 916-+.

[5] V. Marx, Finding the right antibody for the job, Nat. Methods, 10 (2013) 703-707.

[6] B. Cañas, C. Piñeiro, E. Calvo, D. López-Ferrer, J.M. Gallardo, Trends in sample preparation for classical and second generation proteomics, Journal of Chromatography A, 1153 (2007) 235-258.

[7] W.-H. Jin, J. Dai, S.-J. Li, Q.-C. Xia, H.-F. Zou, R. Zeng, Human Plasma Proteome Analysis by Multidimensional Chromatography Prefractionation and Linear Ion Trap Mass Spectrometry Identification, Journal of Proteome Research, 4 (2005) 613-619.

[8] S. Di Palma, M.L. Hennrich, A.J.R. Heck, S. Mohammed, Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis, Journal of Proteomics, 75 (2012) 3791-3813.

[9] N. Zolotarjova, P. Mrozinski, H. Chen, J. Martosella, Combination of affinity depletion of abundant proteins and reversed-phase fractionation in proteomic analysis of human plasma/serum, Journal of Chromatography A, 1189 (2008) 332-338.

437 [10] A. Bodzon-Kulakowska, A. Bierczynska-Krzysik, T. Dylag, A. Drabik, P. Suder, M.

438 Noga, J. Jarzebinska, J. Silberring, Methods for samples preparation in proteomic research,

439 Journal of Chromatography B, 849 (2007) 1-31.

440 [11] A. Staub, D. Guillarme, J. Schappler, J.-L. Veuthey, S. Rudaz, Intact protein analysis

441 in the biopharmaceutical field, Journal of Pharmaceutical and Biomedical Analysis, 55

442 (2011) 810-822.

443 [12] K. Ahrer, A. Jungbauer, Chromatographic and electrophoretic characterization of

444 protein variants, Journal of Chromatography B, 841 (2006) 110-122.

445 [13] S. Fekete, S. Rudaz, J. Fekete, D. Guillarme, Analysis of recombinant monoclonal

446 antibodies by RPLC: Toward a generic method development approach, Journal of

447 Pharmaceutical and Biomedical Analysis, 70 (2012) 158-168.

448 [14] S. Fekete, J.-L. Veuthey, D. Guillarme, New trends in reversed-phase liquid

449 chromatographic separations of therapeutic peptides and proteins: Theory and applications,

450 Journal of Pharmaceutical and Biomedical Analysis, 69 (2012) 9-27.

451 [15] A. Jungbauer, Chromatographic media for bioseparation, Journal of chromatography.

452 A, 1065 (2005) 3-12.

453 [16] L. Turiak, O. Ozohanics, F. Marino, L. Drahos, K. Vekey, Digestion protocol for small

454 protein amounts for nano-HPLC-MS(MS) analysis, Journal of Proteomics, 74 (2011) 942-

455 947.

456 [17] O. Ozohanics, L. Turiak, A. Puerta, K. Vekey, L. Drahos, High-performance liquid

457 chromatography coupled to mass spectrometry methodology for analyzing site-specific N-

458 glycosylation patterns, Journal of Chromatography A, 1259 (2012) 200-212.

459   [18] O. Ozohanics, J. Krenyacz, K. Ludanyi, F. Pollreisz, K. Vekey, L. Drahos,

460   GlycoMiner: a new software tool to elucidate glycopeptide composition, Rapid

461   communications in mass spectrometry : RCM, 22 (2008) 3245-3254.

462   [19] B. Bobaly, E. Toth, L. Drahos, F. Zsila, J. Visy, J. Fekete, K. Vekey, Influence of acid-

463   induced conformational variability on protein separation in reversed phase high performance

464   liquid chromatography, Journal of chromatography. A, 1325 (2014) 155-162.

465   [20] J. Martosella, N. Zolotarjova, H. Liu, G. Nicol, B.E. Boyes, Reversed-phase high-

466   performance liquid chromatographic prefractionation of immunodepleted human serum

467   proteins to enhance mass spectrometry identification of lower-abundant proteins, J Proteome

468   Res, 4 (2005) 1522-1537.

469

7 Appendices

Fig. A: Workflow for characterization of site-specific glycosylation pattern of plasma proteins.

Fig. B: Depleted plasma chromatogram obtained on Poros R2 column. As the chromatogram shows reasonable selectivity was obtained after the injection of 15 μL depleted plasma sample.
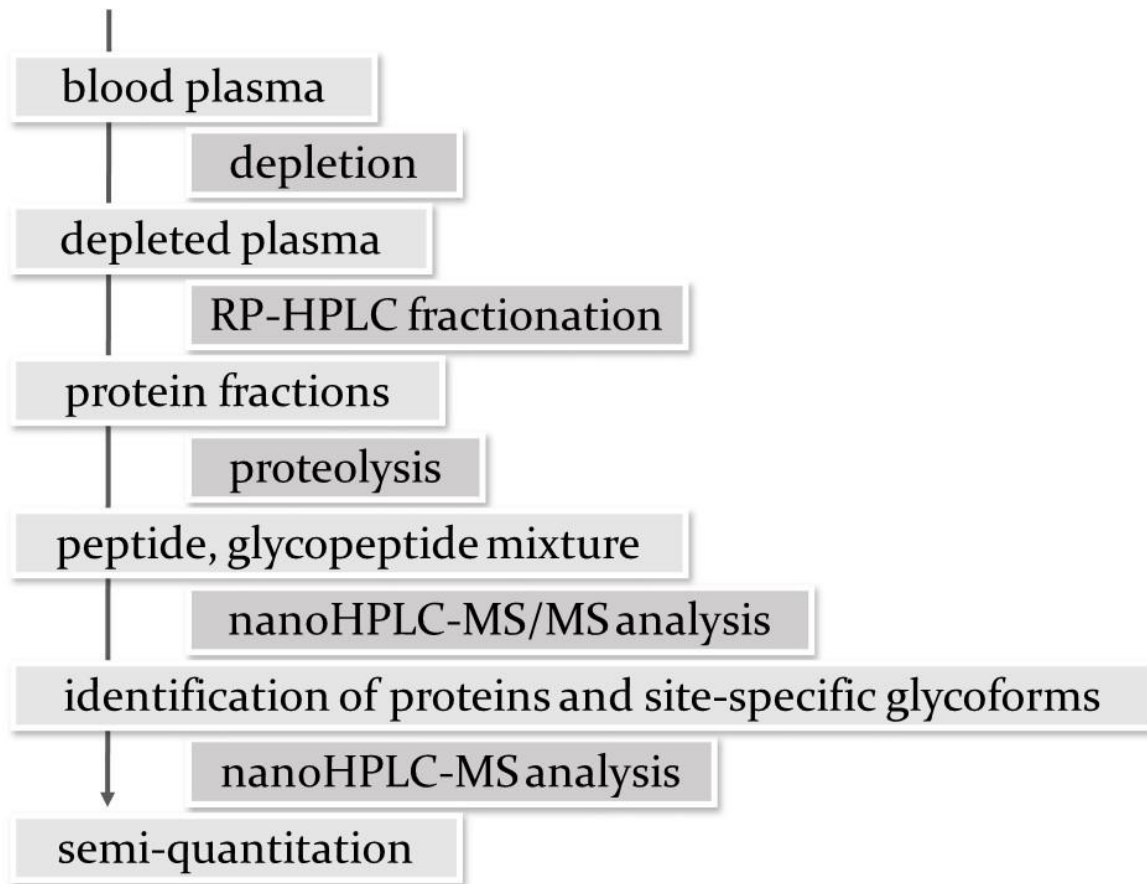
Fig. C: UV chromatogram of standard serotransferrin and distribution of plasma serotransferrin among fractions collected at different retention times. In the latter case human blood plasma was fractionated; the individual fractions were digested and studied in separate nano-HPLC-MS experiments. Label-free quantitation (based on the sum of the three most abundant peptide signals of serotransferrin) was used to determine the relative amount of serotransferrin in the various fractions, indicated by diamond shapes in the Figure. The results show that serotransferrin was indeed successfully isolated in a single chromatographic fraction from blood plasma.

Fig. D: Reproducibility of fraction collection. Protein distribution in the collected fractions is presented in case of beta-2-glycoprotein_1. 3 replicates are presented with different marker types.

Fig. E.1 and Fig. E.2: 1) UV chromatogram of standard haptoglobin and 2) glycosylation pattern of plasma haptoglobin characterized from fractions collected at different retention time. The normalized distribution of glycoforms in case of ^{184}N glycosylation site (^{179}MVSHHNLTTGATLINEQWLLTTAK^{202} glycopeptide) is presented here.

490   **Fig. F: Glycosylation pattern at the ^{56}N site (^{52}NEEYNK^{57} glycopeptide) of alpha-1-acid**

491   **glycoprotein standard without fractionation and after fractionation.** The same pattern was

492   determined after the glycoprotein standard was "fractionated" by RP-HPLC; and the fraction

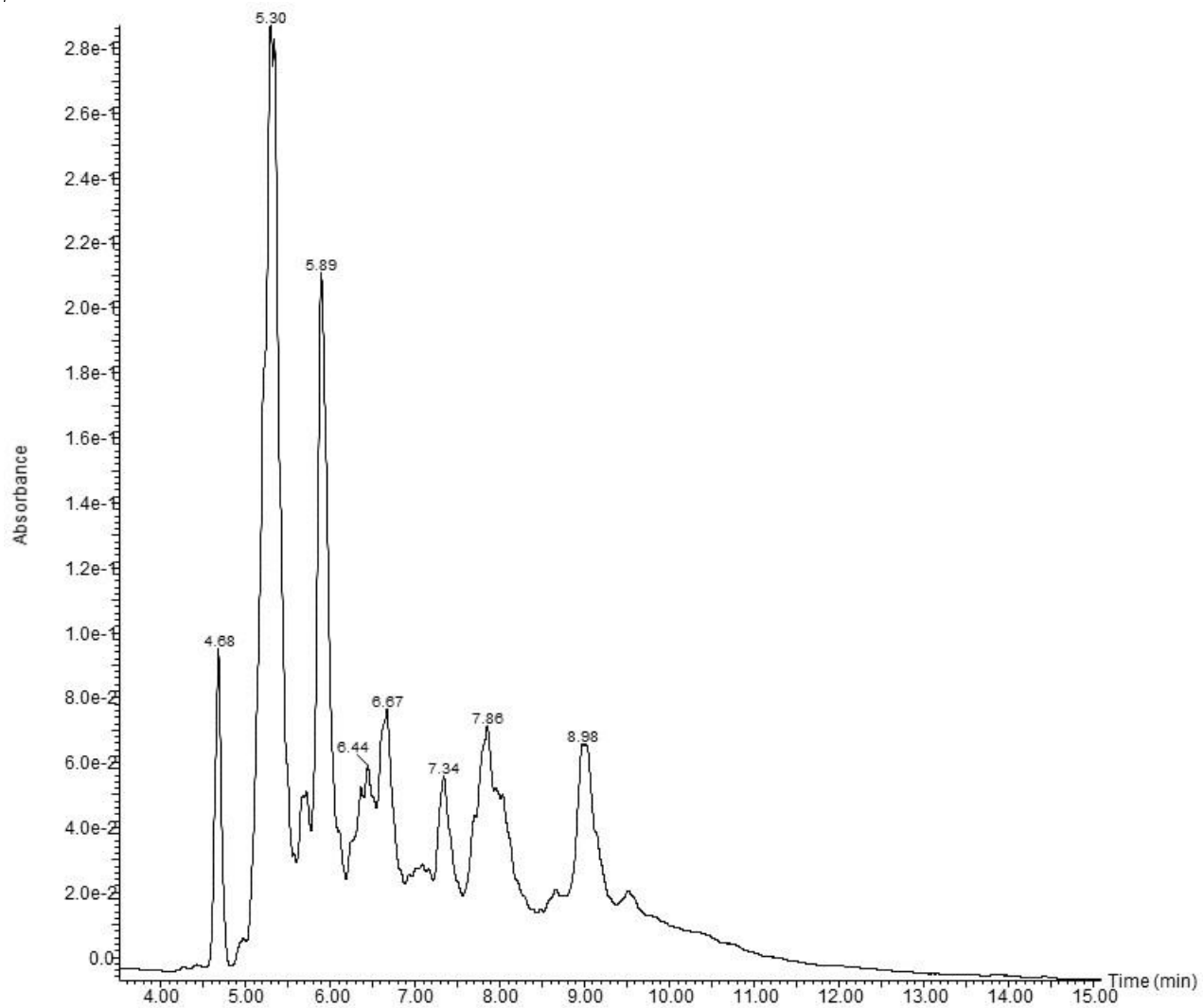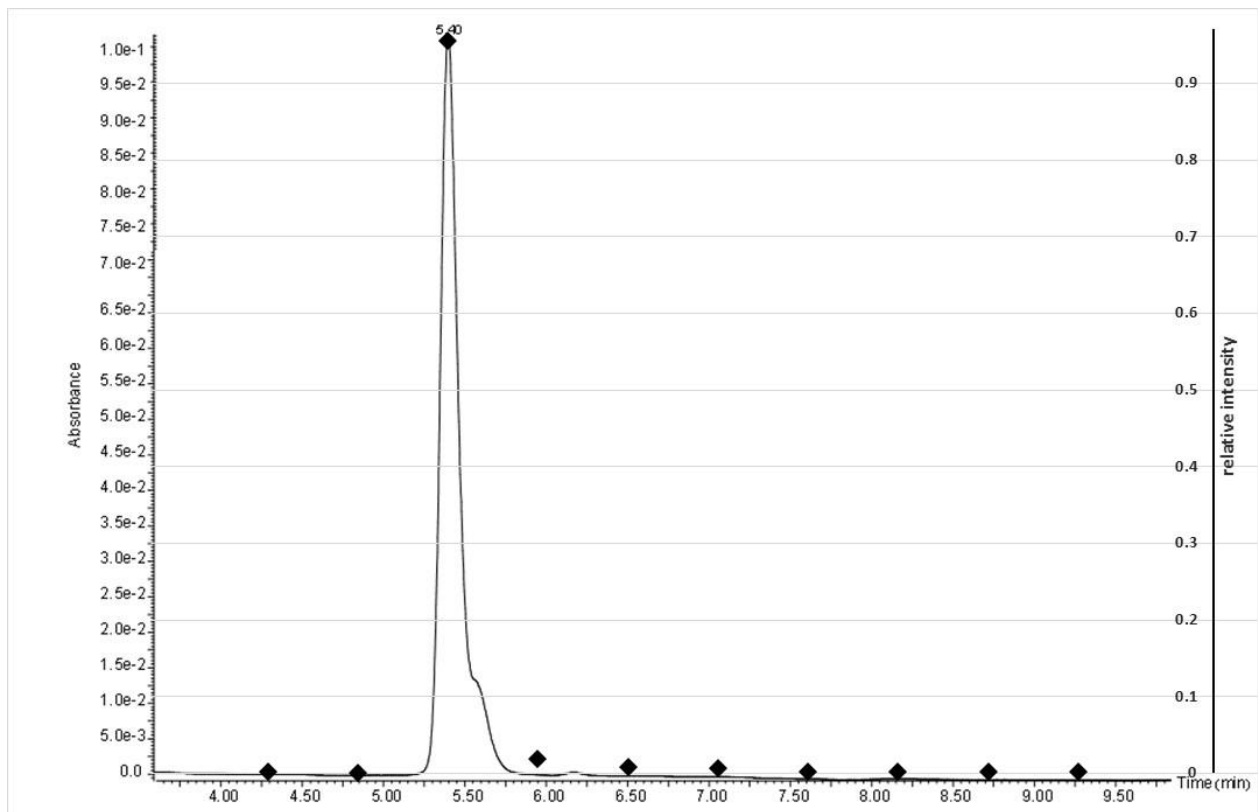493   collected was analyzed in the same way.

494

blood plasma

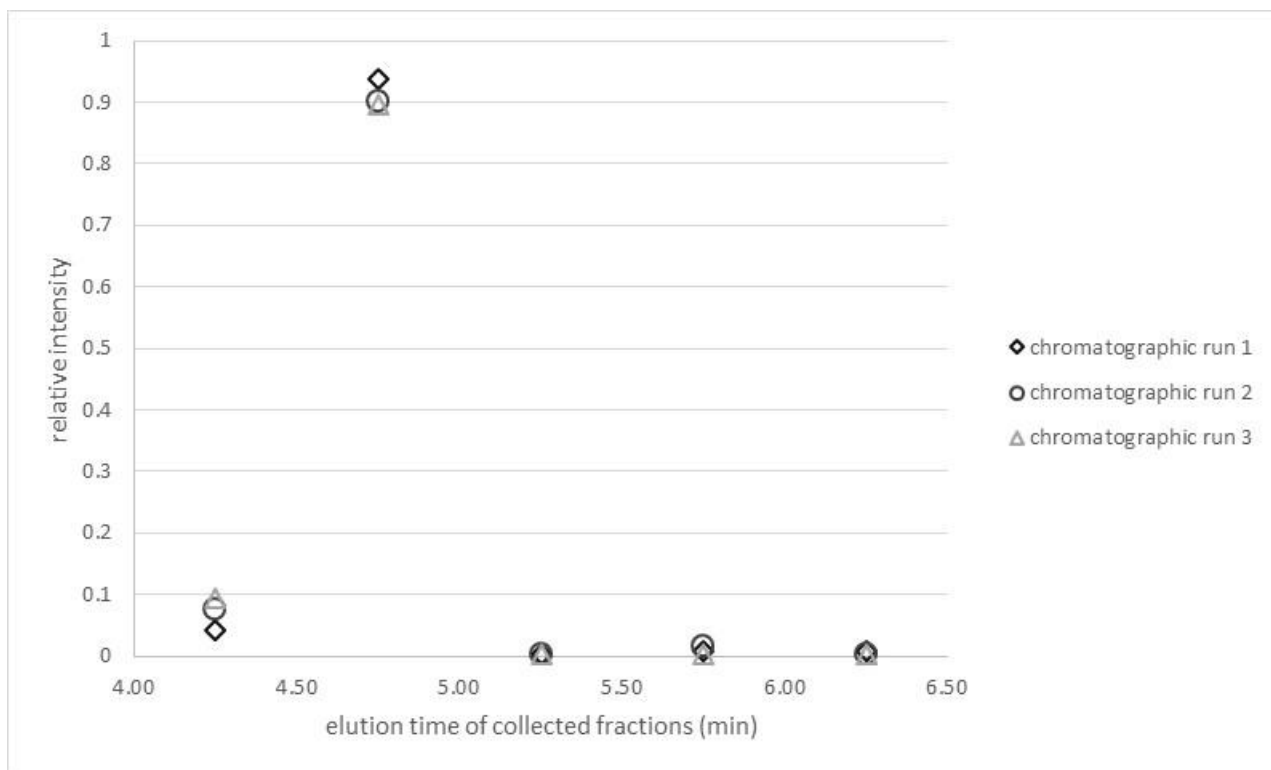depletion

depleted plasma

RP-HPLC fractionation

protein fractions

proteolysis

peptide, glycopeptide mixture

nanoHPLC-MS/MS analysis

identification of proteins and site-specific glycoforms

nanoHPLC-MS analysis

semi-quantitation

495

496   Fig. A

497

498
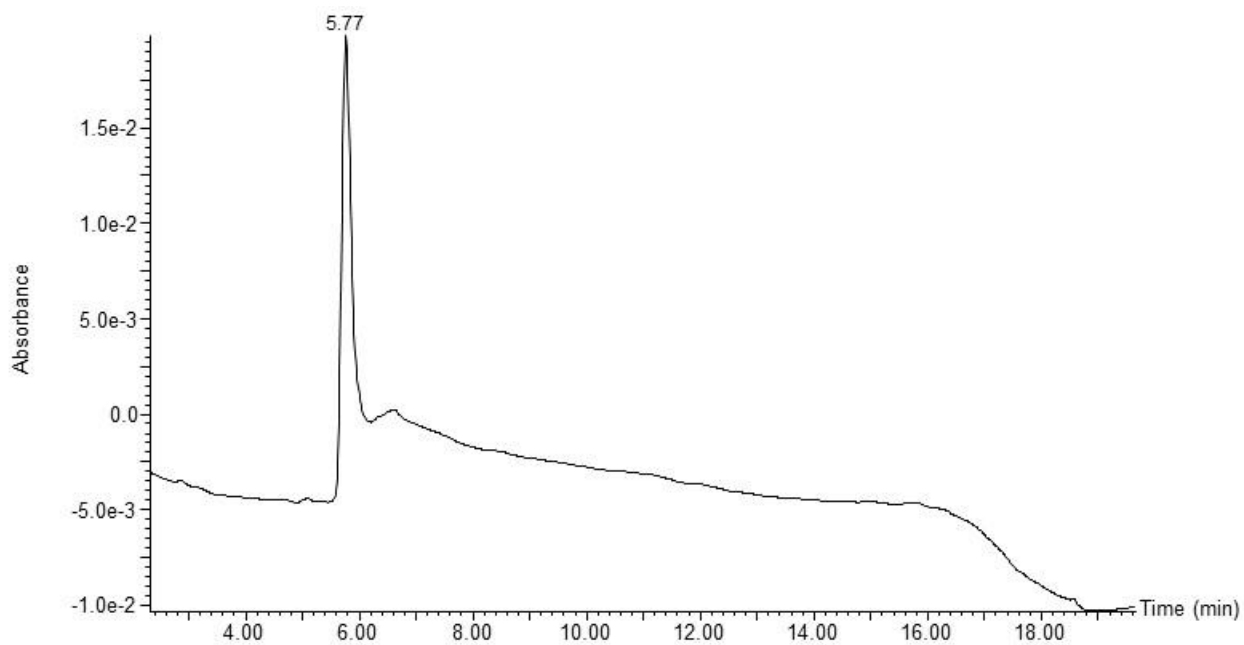
499   Fig. B
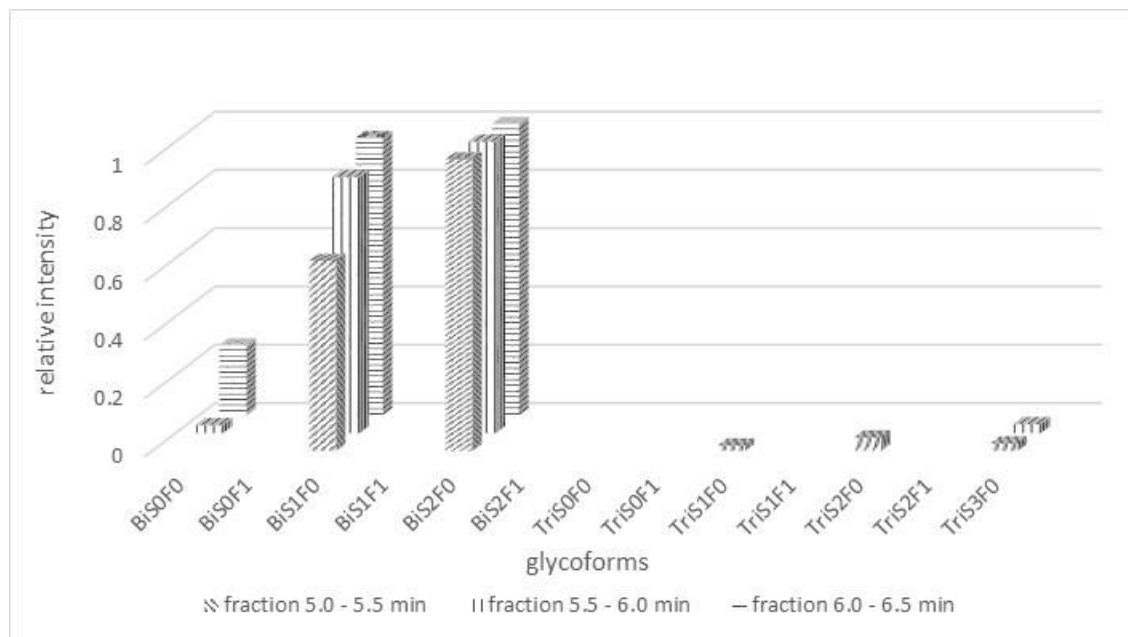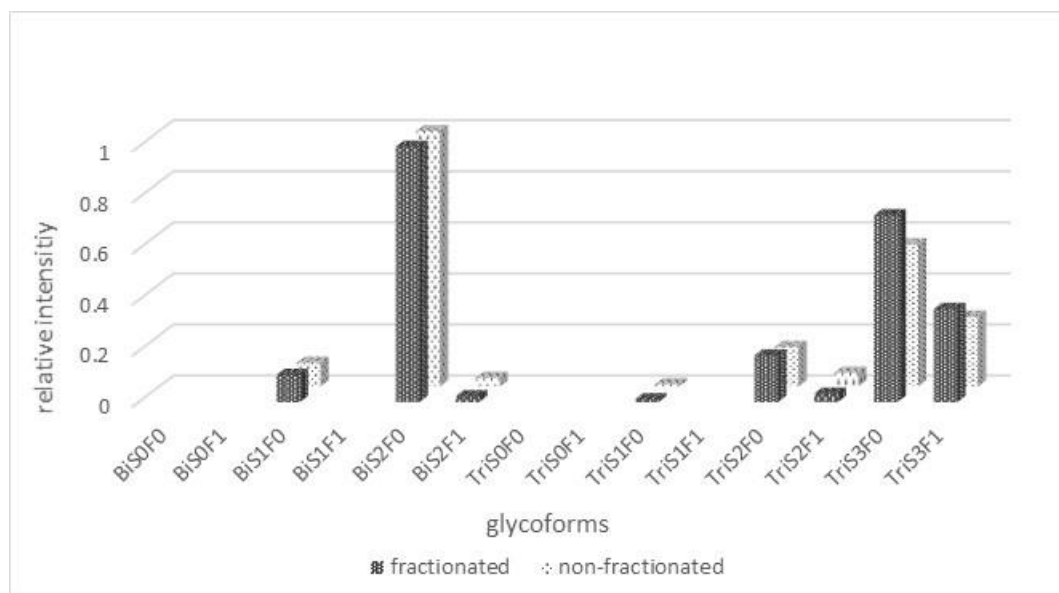
500

501   Fig. C

502

503

Fig. D

505

506   Fig. E.1

507

508   Fig. E.2

509

510

511   Fig. F