JOURNAL OF PROTEOMICS 73 (2010) 1163–1175

available at www.sciencedirect.com

www.elsevier.com/locate/jprot

Meta sequence analysis of human blood peptides and their parent proteins☆

*Peter Bowden, Voitek Pendrak, Peihong Zhu, John G. Marshall**

Department of Chemistry and Biology, Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
Human
Plasma
Serum
Protein
Peptide
Database
Blood

ABSTRACT

Sequence analysis of the blood peptides and their qualities will be key to understanding the mechanisms that contribute to error in LC–ESI-MS/MS. Analysis of peptides and their proteins at the level of sequences is much more direct and informative than the comparison of disparate accession numbers. A portable database of all blood peptide and protein sequences with descriptor fields and gene ontology terms might be useful for designing immunological or MRM assays from human blood. The results of twelve studies of human blood peptides and/or proteins identified by LC–MS/MS and correlated against a disparate array of genetic libraries were parsed and matched to proteins from the human ENSEMBL, SwissProt and RefSeq databases by SQL. The reported peptide and protein sequences were organized into an SQL database with full protein sequences and up to five unique peptides in order of prevalence along with the peptide count for each protein. Structured query language or BLAST was used to acquire descriptive information in current databases. Sampling error at the level of peptides is the largest source of disparity between groups. Chi Square analysis of peptide to protein distributions confirmed the significant agreement between groups on identified proteins.

1. Introduction

There is a need to discover and assay proteins from blood. Blood likely contains the proteins of many different tissue and cell types [1] and these proteins were all first expressed as mRNAs from genes. The mRNA transcripts may be captured as cDNAs, or the parent genome sequenced, and the proteins inferred from the nucleotide sequence [2–15]. Re-arrangements at the level of nucleic acid and post translational processing and modifications give rise to an immense number of possible protein sequences expressed in different tissues and cell types [16–27]. The International Protein Index (IPI), conceived as an integrated database for proteomics experiments, was built as a complete and non-redundant (NR) compilation of the Swiss-Prot, TrEMBL, ENSEMBL and RefSeq databases [28]. RefSeq is a largely non-redundant database of human tran-

Abbreviations: ACN, acetonitrile; BLAST, basic local alignment search tool; cDNA, complementary DNA; CID, collision induced dissociation; DEAE, weak anion exchanger diethylaminoethanol resin; ESI, electrospray ionization; FL, full length; GO, gene ontology; ID, inner diameter; HPLC, high pressure liquid chromatography; LC, liquid chromatography; NP, known protein; MALDI, matrix-assisted laser desorption; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PAGE, polyacrylamide gel electrophoresis; PBS, phosphate buffered saline; QA, quaternary amine; RefSeq, Human Reference Sequence Database; RP, reversed phased chromatography using C18 resin; SCX, strong cation exchanger; SQL, structured query language; UPLC, ultra pressure liquid chromatography.

☆ Contributions: P.B., performed the calculations shown in part using the modified code of V.P. and P.Z. and; V.P., wrote the original SQL codes and approaches; P.Z., wrote the original BLAST parsing code and GO terms; J.G.M., conceived the study, supervised the work, and wrote the paper with P.B.

* *Corresponding author*. Department of Chemistry and Biology, Ryerson University, 350 Victoria Street, Toronto Ontario, Canada M5B 2K3. Tel.: +1 416 799 5000x4219; fax: +1 416 979 5044.

E-mail address: 4marshal@Ryerson.ca (J.G. Marshall).

1874-3919/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.jprot.2010.02.007

scripts together with their gene ID numbers, and gene ontology (GO) annotations [29] that were parsed into an SQL database [30]. Protein sequence databases from these three organizations differ in format and content and contain both overlapping and unique information, with some proteins differing by only one, a few, or many amino acids. In some cases there is extensive data regarding the functions of the protein sequence but the function of many proteins are unknown and some predicted proteins inferred from genomic sequences may be only hypothetical and might not be expressed in cells.

Twelve sets of publicly available proteins and/or peptides by LC–MS/MS of blood that were fractionated and analyzed by different methods were assembled together in SQL to be summarized and compared. Adkins et al. (2002) used protein A/G depletion and tryptic digestion followed by 2D polysulfoethyl/C18 of peptides via nanospray into a Thermo Decca XP with correlation by SEQUEST without enzyme limitations and searching beyond 30 aa in length to yield about 585 proteins. Tirumalai et al. (2003) used ultrafiltration with 30,000 NMWL cut off in 5% ACN followed by separation of tryptic peptides with polysulfoethyl A/C18 via nanospray into a Thermo Decca XP correlated with SEQUEST without enzyme limitations and searching beyond 30 aa in length to yield about 317 proteins. Marshall et al. (2004) used DEAE and other chromatography resins or PAGE separation of proteins with trypsin of chymotrypsin digestion prior to C18 separation, sometimes with prior QA & PS separation of peptides, into a Thermo Decca XP with correlation by SEQUEST of fully tryptic peptides of 14 aa or less yielding 650 proteins with highly stringent correlation scores [31]. Shen et al. [40] digested neat serum and separated the peptides by UPLC with C18 alone or SCX followed C18 using nanospray into a Thermo Decca XP with correlation via SEQUEST using no enzyme limitations and searching beyond 30 aa in length to yield 953 proteins with modest correlation scores. Omenn et al. [44] used depletion of albumin, IgG, IgA, IgM, transferin, haptoglobin, A1AT and/or separation of proteins by CH_2O affinity, reversed phase, PAGE, SAX, gradiflow/tca, free flow electrophoresis, IEF and/or separation of peptides by SCX and or C18 via MALDI and nanospray ESI into Paul ion traps, linear ion traps, Qq-TOF and FTICR with correlation to peptides by SEQUEST, MASCOT, PepMiner, Digger, Sonar, X!TANDEM, VIPER with various enzyme rules yielding 9303 proteins. Shen et al. [41] depleted albumin & IgG prior to digestion before UPLC separation of peptides by SCX/C18 with correlation by SEQUEST without enzyme limitation yielding a set of 2258 higher confidence and 2704 lower confidence proteins. Zhu et al. [70] used a small amount of the data that Marshall et al. [31] had previously calculated discretely (i.e. one experiment correlated individually) but instead calculated the subset jointly (i.e. many experiments contributing to the calculation of protein confidence) with a cut-off score of 2400 [32,33] to yield 2571 proteins with an estimated 5% error rate [32]. Faca et al. [49] depleted serum of albumin, IgG, IgA, transferin, haptoglobin, A1AT followed by separation of intact proteins with Mono Q, reversed phase or PAGE followed by tryptic digestion and separation of peptides by C18 with analysis via nanospray into a Thermo LTQ-FTICR and correlation revealing fully, or occasionally, semi-tryptic peptides from 2254 proteins. Zhang et al. [77] captured N-linked glycopeptides and separated them using SCX followed by C18 via nanospray into an Thermo LTQ with correlation by SEQUEST generating 523 distinct proteins. Sennels et al. [48] used a random hexapeptide library on methyacrylate beads, eluted with thiourea, urea, detergents, acids or organic solvents prior to digestion of PAGE slices and analysis via nanospray into a Thermo LTQ with correlation using SEQUEST and MASCOT yielding 1559 proteins with no peptide information. Tucholska et al. [34] compared propyl sulfate, quaternary amine, diethylaminoethanol, cibachron blue, phenol sepharose, carboxy methyl sepharose, hydroxyl apatite, heparin, concanavalin A and protein G chromatography to yield some 4396 proteins by X!TANDEM.

Mass spectra may be correlated to peptides that sometimes exist in more than one protein [33,35]. The peptide coverage from LC–MS/MS of blood may not be sufficient to correspond to only one protein, even when multiple peptides are observed, due to multiple related genes, protein domains and splice variants in Eukaryotes. Peptide fragmentation spectra from the LC–MS/MS of blood samples were correlated to disparate protein databases inferred from DNA sequences or transcripts from humans and other species [31,36–50]. A meaningful review, summary and comparison of large scale LC–MS/MS data [51] created at different times and correlated to different genetic libraries [44,52,53] cannot be undertaken without the use of computation [54,55]. In contrast to accession numbers that may change over time, the blood peptide and protein sequences themselves are immutable and portable identifiers that may be compared between libraries and datasets. One way to avoid ambiguity is to map the sequences identified to the full-length proteins of a relatively non-redundant set of transcripts. At present the most important computing tools for comparing biopolymer sequence data are Structured Query Language (SQL) [30,56] and the Basic Local Alignment of Sequence Tool (BLAST) [57]. Algorithms such as BLAST or SQL databases may be used to make unbiased comparisons of peptide or protein sequences from correlation analysis of different populations of LC–MS/MS runs [54,56,58]. SQL forms the basis of many laboratory data automation and analysis systems [56,59,60]. The use of a protein or peptide sequence as a unique identifier or "database key" has been shown to improve comparisons between independent groups [54]. An efficient method for labeling protein and peptide sequences such as the Secure Hash Algorithm [61] is ideal for internal database purposes: The term SEQUID is a standard field name representing the protein identifiers generated. Proteins or peptide sequences that are not identical between the two populations of data can thus be derived and examined in detail. Populations of LC–MS/MS runs may contain redundant peptides and proteins. To ensure that redundancy is reduced to a defined standard prior to comparison, the actual peptide or protein sequences, may be collapsed to non-redundant (NR) sets in SQL or BLAST and then summarized and compared [54]. BLAST may thus be used to collapse many proteins into a single representative protein molecule and used to compare populations of LC–MS/MS protein lists [54]. SQL, and BLAST may also be used to obtain available annotation from databases [54,56,58]. The available protein data identified from normal human blood using mass spectrometry was mapped to the RefSeq and ENSEMBL