HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2016 July 18.

Published in final edited form as:

Nat Methods. 2016 March ; 13(3): 237–240. doi:10.1038/nmeth.3731.

An informatic framework for decoding protein complexes by top-down mass spectrometry

Owen S. Skinner[1,2], **Pierre C. Havugimana**[2,3,4], **Nicole A. Haverland**[1], **Luca Fornelli**[1,3,4], **Bryan P. Early**[4], **Joseph B. Greer**[4], **Ryan T. Fellers**[4], **Kenneth R. Durbin**[5], **Luis H. F. Do Vale**[4,6], **Rafael D. Melani**[1], **Henrique S. Seckler**[1], **Micah T. Nelp**[7], **Mikhail E. Belov**[8], **Stevan R. Horning**[8], **Alexander A. Makarov**[8], **Richard D. LeDuc**[3], **Vahe Bandarian**[7], **Philip D. Compton**[1,4], and **Neil L. Kelleher**[1,3,4,5,9]

[1]Department of Chemistry, Northwestern University, Evanston, Illinois, USA

[3]Chemistry of Life Processes Institute, Northwestern University, Evanston, Illinois, USA

[4]Proteomics Center of Excellence, Northwestern University, Evanston, Illinois, USA

[5]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, USA

[6]Brazilian Center for Protein Research, University of Brasilia, Brasilia, Federal District, Brazil

[7]Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona, USA

[8]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany

Abstract

Efforts to map the human protein interactome have resulted in information about hundreds to thousands of multi-protein assemblies housed in public repositories, but the molecular characterization and stoichiometry of their protein subunits remains largely unknown. Here, we combined the CORUM and UniProt databases to create candidates for an error-tolerant search engine designed for hierarchical top-down analyses, identification, and scoring of multi-proteoform complexes by native mass spectrometry.

For over two decades, mass spectrometry (MS)-based proteomics has been a major technology driving the discovery of protein-protein interactions both in microbes[1–5] and

[9]Corresponding author: ; Email: n-kelleher@northwestern.edu

[2]These authors contributed equally to this work

COMPETING FINANCIAL INTERESTS

ProSightPC software is a commercial product for which R.T.F., R.D.L., and N.L.K. receive proceeds. M.E.B., S.R.H., and A.A.M. are employees of Thermo Fischer Scientific, the manufacturer of Q-Exactive instruments.

AUTHOR CONTRIBUTIONS

P.C.H. and N.L.K. devised the computational approach with contributions from O.S.S., R.D.L., and R.T.F. O.S.S. interpreted the data with help from P.C.H. and P.D.C. O.S.S. and P.D.C. collected native MS data. Further supporting results were collected by L.F., K.R.D., L.H.F.D.V., and R.D.M. R.D.L. developed the MPC-score. B.P.E., J.B.G, and R.T.F. implemented the search engine and scoring system. M.E.B., S.R.H., and A.A.M. participated in design of the instrument and support of initial experiments. M.T.N. and V.B. provided the TNH complex. O.S.S., P.C.H., and N.L.K. co-wrote the manuscript with critical inputs from N.A.H. and H.S.S. and contributions from all co-authors. N.L.K. initiated and supervised the project.

mammalian species[6–9]. However, the prevailing method for proteomics relies on proteolysis (*i.e.,* the "bottom-up" approach) and therefore disconnects information about combinations of sequence variation, post-translational modification (PTM) and protein-protein interactions that underlie the great diversity of cellular functions. Although large-scale top-down proteomics determines the composition of whole proteins in denaturing conditions[10], a more complete understanding of the processes driving human cell biology and disease progression requires new methods to more completely capture specific molecular states (*i.e.*, proteoform[11] composition and stoichiometry) of protein assemblies that we refer to hereafter as "multi-proteoform complexes", or MPCs (see Supplementary Table 1 for term usage). This ambitious goal calls for the combination of top-down approaches with native MS, which is especially challenging for mixtures of complexes originating from endogenous sources. However, with recent improvements in biochemical fractionation of cell extracts in native-like conditions[7,8,12] and native top-down MS experiments demonstrating the ejection and subsequent fragmentation of intact subunits[12,13,14,15], precise characterization of MPCs in discovery mode is now in the offing.

To achieve the evolutionary next stage of untargeted methods for precision mapping of MPCs, tailored computational tools and statistical models will need to process data derived from a new stream of multi-stage tandem MS[13]. Here we report the first computational methodology (databases, an algorithm, and scoring) for the identification and characterization of intact protein complexes by native MS. The framework is accessible via a web-based tool and takes as input: an intact mass (MS^1), a subunit mass (MS^2) and the product ions from fragmentation of that subunit (MS^3). The search then couples information from the UniProt Knowledgebase and the CORUM database of protein complexes to achieve error-tolerant identification and scoring of the MPCs present in a sample.

We first evaluated the feasibility of using a 3-tiered, hierarchical search by assessing the nature of the "MPC space." Equation (1) governs the expansion of MPC space and considers the number of potential incorporated proteoforms, encompassing events that 1), change the base sequence (*e.g.*, splicing and alternative start sites) and 2), occur at specific residues (*e.g.*, coding single nucleotide polymorphisms (cSNPs), and most PTMs[11]) to generate the number of possible MPCs per complex:

$$Q_x = \prod_{j=1}^{k_x} n_j \qquad \text{Equation 1}$$

Q_X = number of possible MPCs for a complex X

n_j = number of annotated proteoforms for a subunit j

k_X = number of subunits for a complex X

Given equation (1), and considering both categories of variation noted above (*e.g.*, splicing and PTMs) for the 1,644 non-redundant human complexes in CORUM[16], the total number of MPCs was approximately 2×10^{35}, making a direct search of this space computationally

unfavorable. However, a simplification of the search space can be achieved by dividing the challenge into steps (*vide infra*).

As a first approach to MPC identification, we implemented an error-tolerant search logic to probe two portions of MPC space (Fig. 1). In step 1 of the approach, two databases are created. The first is referred to as CORUM-Proteoform and contains candidate proteoforms (created by shotgun annotation[17] using features from the Swiss-Prot database) for each of the 2,239 subunits from the 1,644 human complexes in CORUM. A second database is created by using the known protein-protein interactions from CORUM coupled with isoform information from Swiss-Prot to form MPC candidates, and is termed CORUM-MPC. For improved efficiency of searching MPC-space, our current implementation populates MPC-candidates in the CORUM-MPC database "on the fly" and is limited to entries containing the hits from step 2.

In step 2 (Fig. 1), the mass of an ejected intact subunit and its fragment ions initiate an error-tolerant search against CORUM-Proteoform. This search is analogous to those performed in proteomics today[18,19], and handles the complexity of the proteoform search space. In step 3, complexes with subunits identified in step 2 are expanded into all possible isoform and stoichiometry combinations using the CORUM-MPC database. The search is performed by comparing the predicted masses of MPCs containing the step 2 subunit with the measured mass of the whole complex. In order to reduce the overall search space required, PTMs and cSNPs of the potential interacting monomers are not considered in this step. However, all modifications from the identified proteoform from step 2 are included. A specific example highlighting the benefit of the multi-step process is shown for the 14 different subunits of the human 20S proteasome (Supplementary Fig. 1). There are 144 MPC combinations considering only isoforms; however, step 2 identification of a single isoform of P28074 corresponds to a 3-fold reduction of the step 3 search space (from 144 to just 48 MPCs). Finally, in step 4 confidence scores for MPCs are calculated using a Bayesian model that takes into account the confidence of the original subunit characterization (step 2), observed MS^1 mass differences, a Gaussian likelihood distribution, and the total number of candidate MPCs with similar MS^1 masses (Supplementary Table 2). The MPC-score follows a Phred-like scale, so generally low, medium, and high scores are in the ranges of <30, 30–60, and 60–3,000, respectively. A web-based implementation of the complete informatics process is available at http://complexsearch.kelleher.northwestern.edu (Supplementary Fig. 2).

We started with the tandem MS analysis of the TNH complex (Fig. 2), previously found to be $\alpha_2\beta_2\gamma_2$ heterohexamer[20]. First, we measured the average mass of the intact complex to be 89,419 +/− 20 Da (Mean +/− SD, MS^1, Fig. 2a, determined from the most abundant charge state peaks). Following activation, the complex ejected three monomers with average masses of 21,083.7 +/− 0.7 Da, 13,607.2 +/− 0.1 Da, and 9,974.2 +/− 0.1 Da (MS^2, all isotopically resolved, Fig. 2b). A single charge state of each of these subunits was then quadrupole-isolated and fragmented (Fig. 2c). Following step 2 (Fig. 1), the MS^2 subunit masses along with those for their fragment ions were searched separately against CORUM-Proteoform. Subunits α (B6CWJ3), β (B6CWJ5), and γ (B6CWJ4) were confidently identified with E-values of 2×10^{-45}, 2×10^{-47}, and 5×10^{-32}, respectively (Fig. 2c). After steps 3 and 4 of the process were completed, each of the three subunit-based searches identified the $\alpha_2\beta_2\gamma_2$

hexamer, with an average MPC-score of 81 and a delta mass of +85 Da (Fig. 2d, Fig. 3, delta mass consistent with binding of 1–2 cobalt and oxygen atoms[20]). The second-best MPC found was the $\alpha_1\beta_4\gamma_2$, corresponding to a large delta mass of −1,223 Da and an abysmal MPC-score of 3.6×10^{-8}. The high MPC-score for the correct $\alpha_2\beta_2\gamma_2$ hexamer after each search reflects the robustness of the searching method when applied to heteromeric complexes.

Next, we evaluated the general applicability of the search tool to homomeric complexes. In brief, we characterized the main monomer of pyruvate kinase (PK) as $-Met_{ini}$, N-terminally acetylated isoform M1 with an S->A sequence variant and a covalent β-mercaptoethanol modification localized to Cys165 at ~30% occupancy. In addition, we characterized a second interacting monomer with an endogenous cleavage followed by N-terminal acetylation and also the mercaptoethanol modification at the same stoichiometry; together these four monomeric proteoforms formed a total of 13 tetrameric MPCs (Supplementary Figs. 3, 4, Fig. 3). Unexpectedly, we observed the dimeric PGAM2 complex as a contaminant in this sample, with ~30% of the monomer observed with the C-terminal lysine clipped off; three dimeric MPCs of PGAM2 were characterized (Supplementary Fig. 5, Fig. 3). Finally, we observed the tetrameric GAPDH complex as only a single α_4 MPC, with a mass indicative of eight internal cysteine persulfide modifications (Cys-S-S-H, two per monomer, one localized to Cys152, Supplementary Fig. 6, Fig. 3). The wealth of MPCs observed from these well-characterized case studies highlights a new level of molecular precision in interrogating MPCs.

Moving to larger complexes, we applied the search process to horse ferritin, an iron-binding cage comprised of 24 L- and H-chains[21]. The MS^1-determined intact mass was 490,284 +/− 52 Da, with an isotopically-resolved MS^2 mass for the ejected subunit measured as 19,934.1 +/− 0.1 Da (Supplementary Fig. 7). Isolation and fragmentation of this subunit gave its identification as the L-chain (P02791; E-value of 4×10^{-91}), with a single proteoform characterized as: $-Met_{ini}$, N-terminally acetylated, with a previously unannotated methyl disulfide present on Cys49 at near-complete occupancy (C-score of 178). After addition of this +45.987 Da modification to the database, application of the search tool resulted in the 24-meric MPC being assigned with the correct $L_{15}H_9$ stoichiometry and a good MPC-score of 74 for the best hit (Fig. 3). Importantly, this assignment required data from only one of the two chains; the inability to eject certain monomers from heteromeric complexes is a current challenge in burgeoning field of native MS[22].

Finally, we tested the search platform on the challenging human 20S proteasome complex, known to be a 28-mer formed from two copies of each of 14 unique subunits[23]. Prior to native analysis, the complex was denatured and analyzed by an automated top-down workflow, indicating the proteoforms that made up the complex but providing no stoichiometry information (Supplementary Table 3). We measured the intact complex mass to be 725,706 +/− 319 Da, with wide charge state peaks indicating non-covalent adduction or extensive PTM heterogeneity (Supplementary Fig. 8). Four of the 14 unique subunits were ejected in the gas phase, and their masses were isotopically resolved and measured as 29,394.9 +/− 0.8 Da, 26,452.2 +/− 0.4 Da, 27,796.5 +/− 0.2 Da, and 21,902.9 +/− 0.7 Da.

Each subunit was then isolated, fragmented, and identified in step 2 as α4, α5, α7, and β6 with E-values of 5×10^{-5}, 1×10^{-13}, 3×10^{-6}, and 1×10^{-10}, respectively. The step 3 and 4 results for these four subunit-based searches produced many candidate MPCs with scores all <1 that indicate the lack of confidence in the identification due to the large number of possible MPCs with similar masses (Fig. 3, with the best hits reported in Supplementary Table 4). While the combinatorial explosion of candidate MPCs and the relatively poor quality MS^1 data makes confident identification of a single MPC difficult, the proteasome case indicates that the search method is capable of processing data from large protein assemblies and provides search results with good fidelity. Improvement of data quality and identification of additional subunits would improve the accuracy of the search method and its ability to correctly identify specific MPCs.

The computational framework reported here is the first of its kind and supports a hierarchical search based on the 3-tiers of a new tandem MS process for protein complexes. Scoring includes known protein features in UniProtKB and associations in the CORUM database to assign a level of confidence to whole protein complex identification and characterization. While protein associations are known in databases of protein complexes, the subunit stoichiometry is often not; only 7% of complexes in CORUM have protein stoichiometry specified. Even when complex stoichiometry is known, the presence of sequence variants, PTMs, and alternate endogenous cleavages can combine to produce a staggering number of MPCs. However, improved instrumentation for ejection and fragmentation of subunits, proper searching and scoring of multiple MS^3 fragment datasets simultaneously, and improved scoring for MPCs with partially-occupancy PTMs will all help in delivering precisely characterized MPCs despite the extremely large search space they represent. The end result, definition of protein complexes with absolute molecular specificity using native MS^n and optimized search tools, will illuminate the macromolecular machines that drive so many cellular functions underpinning human health and disease.

ONLINE METHODS

Preparation of protein complexes

Recombinant toyocamycin nitrile hydratase was expressed and purified as described previously[24]. Rabbit pyruvate kinase was purchased from Roche; the phosphoglycerate mutase 2 was discovered as a contaminant in that sample. Glyceraldehyde 3-phosphate dehydrogenase was isolated from HeLa S3 cells as described previously[25]. Ferritin from horse spleen was purchased from Sigma Aldrich. Finally, human 20S proteasome was purchased from Enzo Life Sciences. All samples were desalted prior to analysis by buffer exchanging with molecular weight cutoff spin filters (Millipore) into 150 mM ammonium acetate at pH 7.

Top-down analysis of denatured 20S proteasome

The purified 20S proteasome sample was denatured with 5% acetonitrile and 0.2% formic acid (Solvent A) and run with nano-reverse-phase liquid chromatography coupled to an Orbitrap Elite (Thermo) on an 18 cm PLRP-S column using a 60-minute gradient (90-minute total run time including column wash and equilibration), where Solvent B

concentration (5% water in acetonitrile, 0.2% formic acid) was ramped from 15 to 55%. Mass-spectral data were acquired using a top-two, data-dependent acquisition with high-resolution MS^1 and MS^2 (120,000 and 60,000 resolving power at 400 m/z, respectively) scans. AGC target values were 1×10^6 for both MS^1 and MS^2. Ion fragmentation was obtained via higher-energy collisional dissociation (HCD) with NCE = 24. Denatured proteoforms were identified using ProSightPC 3.0 SP1 (Thermo).

Construction of the human multi-proteoform complex databases

Human protein complexes were downloaded from the CORUM reference database (February, 2012 version)[16] and further processed to only contain non-redundant complexes with 20 or fewer subunits. Applying the above criteria, the CORUM database was reduced from 1,671 unique complexes (2,542 proteins) to 1,644. For each complex, protein subunits were first mapped to reviewed canonical and isoform-sequence identifiers reported in the "UniProtKB/Swiss-Prot" database (March 6, 2014 version). MPC search space limited to subunit isoforms was derived by calculating the product of isoforms present in each of the 1,644 complexes, yielding a database of 9,299,540 human multi-proteoform complexes derived from 2,239 proteins (i.e., unique Swiss-Prot accession numbers). The maximum number of subunit copies considered for CORUM-MPC was 50 for complexes with ≤4 interacting gene products and 2 for those with >4. Example protein complexes not in the database were added; PGAM2 and the TNH subunits were only listed in Uniprot as unannotated TrEMBL entries.

Processing of 3-tiered data from native mass spectrometry

All native MS^1, MS^2 and MS^3 spectra were acquired on a modified Q-Exactive HF mass spectrometer[13]. The use of the term "pseudo-MS^3" in the original publication[13] indicates that the monomers are ejected in the source region, isolated, and fragmented, which is not typically considered a true MS^3 experiment. The mass values of intact complexes and ejected monomer were manually calculated from MS^1 and MS^2 spectra, respectively, with reported errors indicating the standard deviation of the mass measurement from each charge state, a metric of the experimental precision. For TNH, ferritin and 20S proteasome, spectra were first smoothed before determining MS^1 values. All MS^3 spectra were deconvoluted using Xtract software[26], or were analyzed manually and internally calibrated using mMass[27]. Unless otherwise specified, MS^1 and MS^2 were average masses and MS^3 data were input as monoisotopic values. MS^2 and MS^3 results were then subjected to a database search against the constructed CORUM-Proteoform database (Fig. 1) using absolute mass mode[28]. All masses used for the identifications can be found in Supplementary Table 5. ProSight Lite[29] was used to visualize the fragmentation of identified proteins and produce graphical fragment maps.

Scoring and ranking multi-proteoform complex (MPC) candidates

Similarly to the recent C-score approach[30] for the characterization of proteoforms, we developed the Bayesian MPC-score to striate MPC-level search results. This score is based on the observed intact mass of the complex, and the proteoform identified from the MS^2 and MS^3 spectra. To find this, the MS^2 and MS^3 data are searched against a candidate warehouse containing all protein entries in the CORUM-Proteoform database (step 2). A version of the

ProSight search algorithm is used to identify proteoforms ejected from the complex, returning a posterior probability score for each hit (see Supplementary Table 2 for descriptions of each scoring component)[30]. All candidate complexes containing at least one of the returned candidate proteins (with a *p*-score metric of identification[31] better than 1×10^{-4}) are then queried in steps 3 and 4 of the overall process depicted in Fig. 1.

The scoring model uses uniform prior probabilities for each candidate complex interrogated, though in future work these priors will be adjusted based on rules for complex formation; for example a dimer of two full length proteins might be given a higher prior probability than a heptamer of several short proteoforms. To calculate the posterior probability, only complex likelihoods were needed. To find these, two generative models were used; an MS^1 and a proteoform model.

The MS^1 generative model compared the observed and complex theoretical mass to a truncated Gaussian distribution with a 200 Da standard deviation (chosen to model the variability of measurement and to allow for possible adducts), scaled such that 0 mass difference is assigned a value of 1.

$$likelihood = e^{-(theoretical-observed)^2 / 2*200^2}$$

If this likelihood is below 1×10^{-300}, it is set to 1×10^{-300} instead for rounding purposes. For the proteoform likelihood, if the proteoform's C-Score was 50 or above, the likelihood was set to 1, otherwise, it was set to its posterior probability from the database search used to identify the proteoform. The resulting complex likelihood is the product of the MS^1 likelihood and the proteoform likelihood. The proteoform model therefore impacts the MPC-score when the identification is below a C-score of 50; above this value, the complexes are scored more heavily on the mass difference between observed and theoretical MS^1 values.

All complex likelihoods interrogated are summed to give marginal likelihood (Supplementary Table 2). An MPC's posterior probability is its likelihood divided by this marginal likelihood. A complex's MPC-score is a transformation on its posterior probability (the same transformation as used in the C-score[30]):

$$MPC\ score = -10 * log_{10}(1 - posterior\ probaility)$$

Querying the Web-based Search Tool for Identification of Protein Complexes

The backend of the free online web-based tool is built using the ASP.NET MVC and Microsoft.NET Framework 4.5 technologies. Through the web-interface (http://complexsearch.kelleher.northwestern.edu), the user enters neutral masses determined from MS^1, MS^2, and MS^3 spectra and the respective error tolerance information in the fields provided. The user may also choose to add an additional interaction for the duration of the search, which is input as a list of UniProt accession numbers. The relevant databases (Fig. 1) are queried, followed by the calculation of statistical metrics and the results are displayed to

the user. At present, the web-tool evaluates the human protein-complex composition annotated in the adapted CORUM database and accepts data reduced from any mass spectrometer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank members of the Kelleher research group and Prof. V. Wysocki for helpful discussions and advice. O.S.S. is supported by a U. S. National Science Foundation Graduate Research Fellowship (2014171659). P.C.H. is a recipient of a Northwestern University's Chemistry of Life Processes Institute Postdoctoral Fellowship Award. L.H.F.D.V. is supported under CNPq research grant 202011/2012-7 from the Brazilian government. H.S.S is supported under the Science Without Borders scholarship 88888.075416/2013-00 from the Coordination for the Improvement of Higher Education Personnel, under the Brazilian government. This work was supported by grants from the W.M. Keck Foundation (DT061512) and the U.S. National Institutes of Health (GM067193) to N.L.K.

References

1. Babu M, et al. Nature. 2012; 489:585–9. [PubMed: 22940862]
2. Butland G, et al. Nature. 2005; 433:531–7. [PubMed: 15690043]
3. Gavin AC, et al. Nature. 2006; 440:631–6. [PubMed: 16429126]
4. Krogan NJ, et al. Nature. 2006; 440:637–43. [PubMed: 16554755]
5. Kuhner S, et al. Science. 2009; 326:1235–40. [PubMed: 19965468]
6. Guruharsha KG, et al. Cell. 2011; 147:690–703. [PubMed: 22036573]
7. Havugimana PC, et al. Cell. 2012; 150:1068–81. [PubMed: 22939629]
8. Kristensen AR, Gsponer J, Foster LJ. Nat Methods. 2012; 9:907–9. [PubMed: 22863883]
9. Malovannaya A, et al. Cell. 2011; 145:787–99. [PubMed: 21620140]
10. Tran JC, et al. Nature. 2011; 480:254–8. [PubMed: 22037311]
11. Smith LM, Kelleher NL. Nat Methods. 2013; 10:186–7. [PubMed: 23443629]
12. Skinner OS, et al. Anal Chem. 2015; 87:3032–3038. [PubMed: 25664979]
13. Belov ME, et al. Anal Chem. 2013; 85:11163–73. [PubMed: 24237199]
14. Rathore D, Dodds ED. J Am Soc Mass Spectrom. 2014; 25:1600–9. [PubMed: 25001382]
15. Dyachenko A, et al. Anal Chem. 2015; 87:6095–6102. [PubMed: 25978613]
16. Ruepp A, et al. Nucleic Acids Res. 2010; 38:D497–501. [PubMed: 19884131]
17. Pesavento JJ, et al. J Am Chem Soc. 2004; 126:3386–3387. [PubMed: 15025441]
18. Chick JM, et al. Nat Biotechnol. 2015; 33:743–749. [PubMed: 26076430]
19. Meng F, et al. Nat Biotechnol. 2001; 19:952–7. [PubMed: 11581661]
20. Blackwell AE, et al. Anal Chem. 2011; 83:2862–2865. [PubMed: 21417466]
21. Theil EC. Curr Opin Chem Biol. 2011; 15:304–311. [PubMed: 21296609]
22. Zhou M, et al. Angew Chem Int Ed Engl. 2012; 51:4336–4339. [PubMed: 22438323]
23. Loo JA, Benchaar SA, Zhang J. Mass Spectrometry. 2013; 2:S0013. [PubMed: 24349932]
24. Nelp MT, et al. Biochemistry. 2014; 53:3990–3994. [PubMed: 24914472]
25. Havugimana PC, Wong P, Emili A. J Chromatogr B Analyt Technol Biomed Life Sci. 2007; 847:54–61.
26. Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL. J Am Soc Mass Spectrom. 2005; 16:2027–38. [PubMed: 16253516]
27. Strohalm M, et al. Anal Chem. 2010; 82:4648–4651. [PubMed: 20465224]
28. LeDuc RD, et al. Nucleic Acids Res. 2004; 32:W340–5. [PubMed: 15215407]
29. Fellers RT, et al. Proteomics. 2014

30. LeDuc R, et al. J Proteome Res. 2014; 13:3231–40. [PubMed: 24922115]
31. Meng F, et al. Nat Biotechnol. 2001; 19:952–957. [PubMed: 11581661]

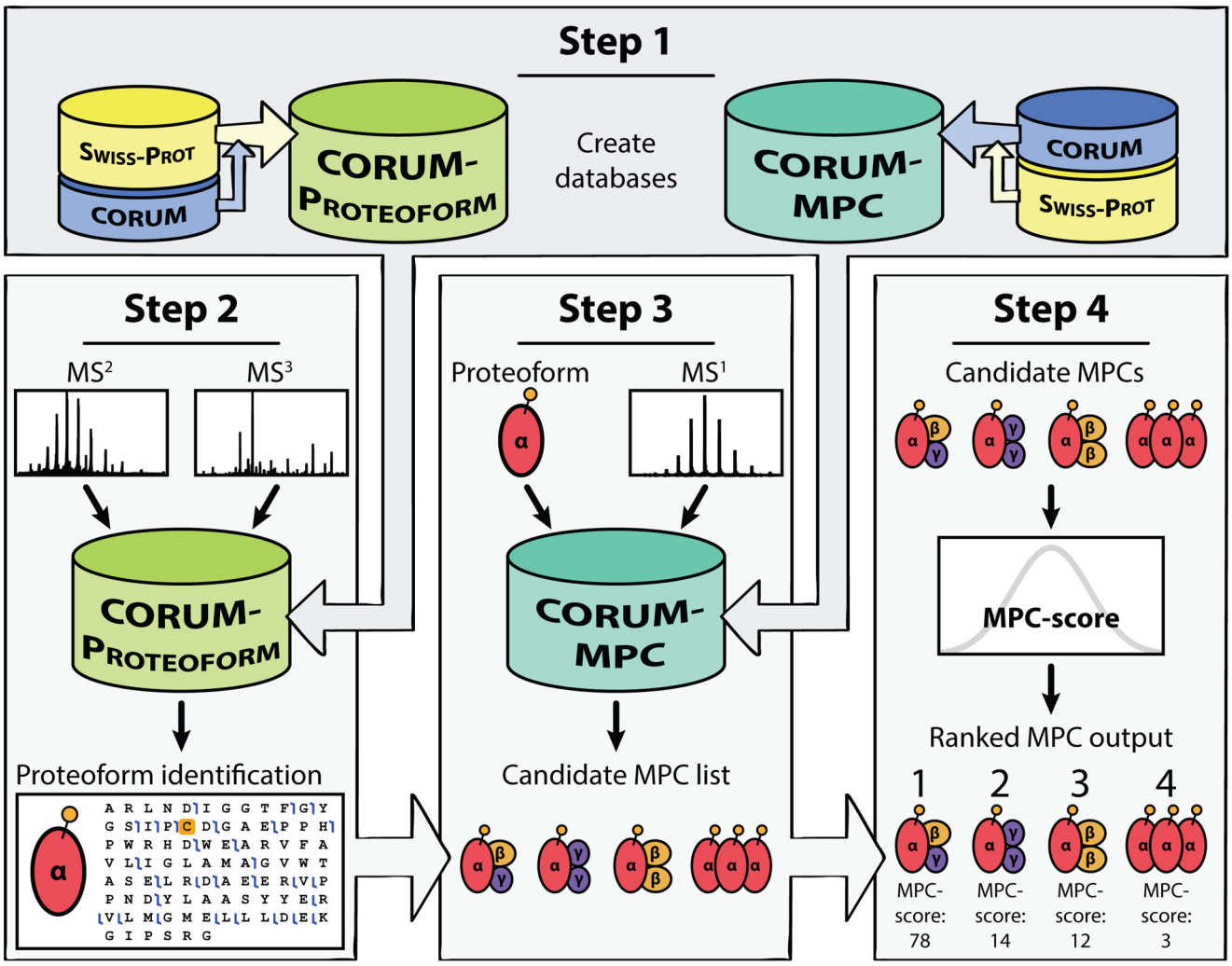

Figure 1.

Computational platform and workflow for characterization of human multi-proteoform complexes (MPCs). In step 1, two databases are created, the "CORUM-Proteoform" database, (which contains Swiss-Prot entries also present in the CORUM database that are combinatorially expanded into candidate proteoforms), and the CORUM-MPC (which contains candidate MPCs from all subunit combinations in CORUM and their known isoforms contained within Swiss-Prot). In step 2 of the workflow, a proteoform is retrieved using the mass values from MS^2 (subunit) and MS^3 (backbone fragment ions) by searching the CORUM-Proteoform database. In step 3, the identified proteoform and the MS^1 (intact complex) mass value are used to search against the CORUM-MPC database and generate a candidate MPC list. In step 4, a MPC-score is calculated for each member of the candidate MPC list by incorporating MS^1 intact mass information and the quality of proteoform characterization for the subunit ejected from the complex.

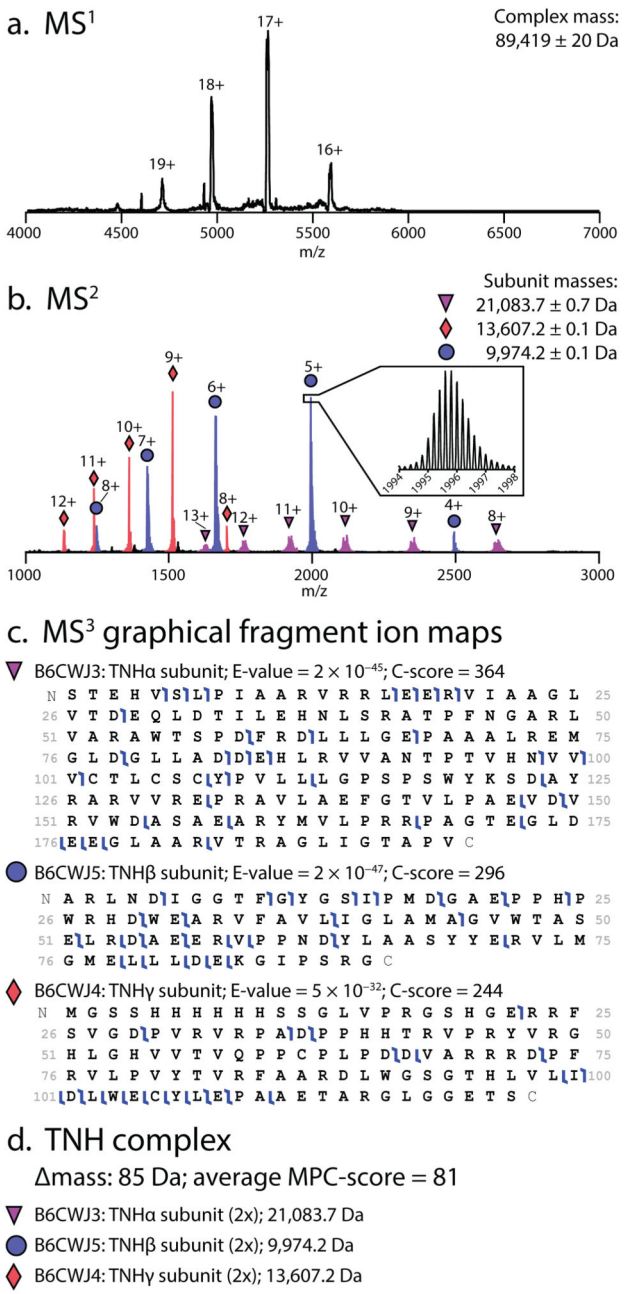

Figure 2.

Characterization of toyocamycin nitrile hydratase (TNH), a hexameric multi-proteoform complex purified to homogeneity. The 3-tiered approach to tandem top-down mass spectrometry for the TNH complex is illustrated with the **(a)** MS^1, **(b)** MS^2 and **(c)** MS^3 fragment ion maps of each of the ejected proteoforms. Masses are reported as the average +/− the S.D. of the mass measured from each of the most abundant charge states. **(d)** The identity of a specific multi-proteoform complex ($\alpha_2\beta_2\gamma_2$) obtained from database searching that combines information from MS^1, MS^2, and three MS^3 spectra is shown below the subunit graphical fragment maps.

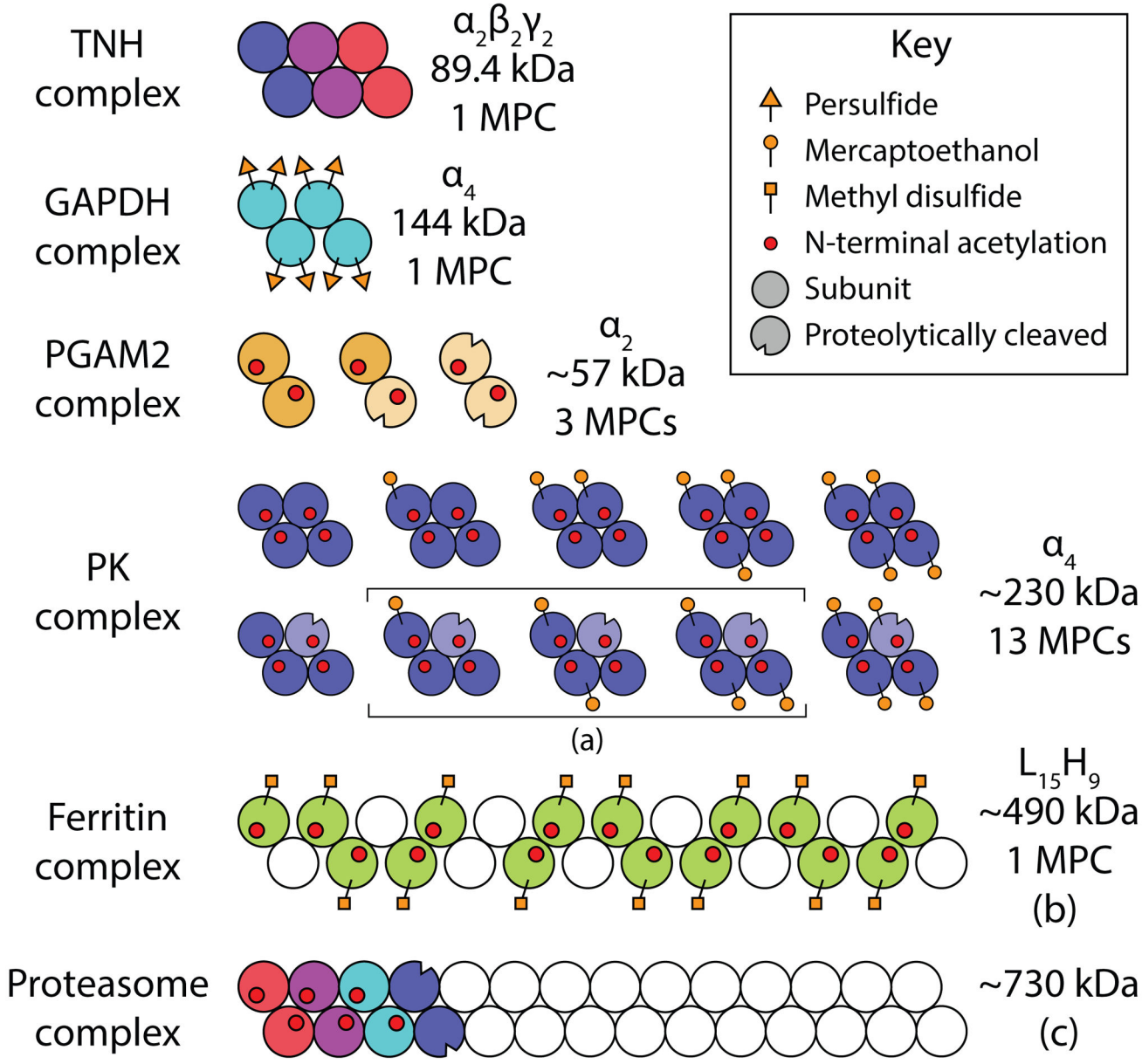

Figure 3.

A summary of the MPCs identified in this study. The complex name, intact mass, and number of observed MPCs is noted next to each complex, with modifications and endogenous cleavages on subunits specified in the key in the upper right. (a) Additional combinatorial MPCs were detected for pyruvate kinase (see Supplementary Fig. 4 and the Supplementary Discussion for greater detail). (b) Broad charge states in the case of ferritin indicate a distribution of bound iron (in the thousands) and therefore the possibility of additional MPCs lying beneath this molecular polydispersity. (c) The overall stoichiometry of the human 20S proteasome is known to be $(\alpha\text{-}1)_2(\alpha\text{-}2)_2(\alpha\text{-}3)_2(\alpha\text{-}4)_2(\alpha\text{-}5)_2(\alpha\text{-}6)_2(\alpha\text{-}7)_2(\beta\text{-}1)_2(\beta\text{-}2)_2(\beta\text{-}3)_2(\beta\text{-}4)_2(\beta\text{-}5)_2(\beta\text{-}6)_2(\beta\text{-}7)_2$.