Published in final edited form as:

Curr Protoc Bioinformatics. ; 56: 13.30.1–13.30.11. doi:10.1002/cpbi.16.

The Search Engine for Multi-Proteoform Complexes: An Online Tool for the Identification and Stoichiometry Determination of Protein Complexes

Owen S. Skinner¹, Luis F. Schachner¹, and Neil L. Kelleher^{1,2}

¹Department of Chemistry, Northwestern University, Evanston, Illinois

²Department of Molecular Biosciences, Northwestern University, Evanston, Illinois

Abstract

Recent advances in top-down mass spectrometry using native electrospray now enable the analysis of intact protein complexes with relatively small sample amounts in an untargeted mode. Here, we describe how to characterize both homo- and heteropolymeric complexes with high molecular specificity using input data produced by tandem mass spectrometry of whole protein assemblies. The tool described is a “search engine for multi-proteoform complexes,” (SEMPC) and is freely-available at <http://complexsearch.kelleher.northwestern.edu>. The output is a list of candidate multi-proteoform complexes and scoring metrics, which are used to define a distinct set of one or more unique protein subunits, their overall stoichiometry in the intact complex and their pre- and post-translational modifications. Thus, we present an approach for the identification and characterization of intact protein complexes from native mass spectrometry data.

Keywords

Protein Complexes; Multi-Proteoform Complexes; Top-Down Proteomics; Native Mass Spectrometry

INTRODUCTION

The search engine for multi-proteoform complexes (SEMPC) is a web tool designed to identify and characterize multi-proteoform complexes (MPCs) (Skinner et al., 2016) using data from a three-tiered native mass spectrometry experiment (MS¹, MS², and MS³). It is freely available at <http://complexsearch.kelleher.northwestern.edu>. The web tool uses information from the UniProt knowledgebase and the CORUM database of protein complexes (Ruepp et al., 2010), and provides an output list of candidate MPCs that are ranked based on their MPC-score, a metric that indicates the level of confidence in the identification and assignment of subunit stoichiometry (Skinner et al., 2016). Subunits are also given independent metrics asserting the confidence of their identifications (i.e., E-values and C-scores).

OSS – OwenSkinner2012@u.northwestern.edu

LFS – luisschachner2015@u.northwestern.edu

NLK – n-kelleher@northwestern.edu, Phone: (847) 467-4362

In **Basic Protocol 1**, we outline the SEMPC search process for a complex that is present in the databases by using the toyocamycin nitrile hydratase (TNH) alpha subunit (UniProt accession: B6CWJ3) as an example. The TNH complex has been extensively studied by native mass spectrometry from a structural biology perspective (Blackwell et al., 2011), and was shown to be an $\alpha_2\beta_2\gamma_2$ heterohexameric complex. It was placed in the database of human complexes for validation and benchmarking purposes.

In **Alternate Protocol 1**, we describe how to search a protein complex that is not present in the CORUM database by using the “Add New Complex” feature of the SEMPC. We use the human cytoplasmic malate dehydrogenase (UniProt accession: P40925) as an example; it is annotated in UniProt as a homodimer.

BASIC PROTOCOL 1

Characterizing a Multi-Proteoform Complex with SEMPC—The SEMPC provides a simple and intuitive web interface that takes as input neutral mass values derived from any mass spectrometer and produces as an output a ranked and scored list of candidate MPCs along with their possible subunits and stoichiometries. The search process first queries a set of candidate proteoforms generated by a process called “shotgun-annotation” (Pesavento et al., 2004) from the ~20,300 human SwissProt entries. This first query determines from the input MS²- and MS³-level data the specific identity of the subunit proteoform (including known post-translational modifications, PTMs, and single-nucleotide polymorphisms, SNPs). The second search step uses the CORUM database to determine the known interacting partners of the identified subunit, and generates a list of the masses of all potential MPCs based on the most likely stoichiometries and known interacting partners. In order to determine the score (referred to here as the MPC-score) and the ranked output list, the search tool compares the theoretical masses with the user-input MS¹ mass and further takes into account the overall level of subunit characterization.

Necessary Resources

Hardware: Any computer with internet access

Software: Software capable of reading an Excel (.xlsx) file

Getting started: Access and explore the online SEMPC

1. Access the SEMPC online at <http://complexsearch.kelleher.northwestern.edu>.
2. Click “About” on the top-right in the purple title bar (Fig. XX.XX.1).
Examples of MS¹-MS³ spectra can be found on the about page, in addition to links for further information.
3. Click on the Excel icon found at the end of the Example section to download a spreadsheet (Supp_Table_5_NeutralMasses.xlsx) of neutral masses from example MPCs (Fig. XX.XX.2).

The downloaded spreadsheet contains neutral masses from 11 example MPCs, all of which can be used with the SEMPC. The dataset includes single values for MS¹ and MS² masses, and a list of MS³ fragment ion masses.

4. Return to the website and click “Home” on the purple title bar (Fig. XX.XX.2). You will return to the input page.

Performing a search

5. Input 89419 into the MS¹ average mass field, corresponding to the intact mass of the TNH alpha subunit example found in the first column of the downloaded example spreadsheet from step 3 (Fig. XX.XX.1).

The MS¹ value is expressed in terms of an average neutral mass in daltons, and is used here because in many cases, isotopic resolution of the intact complex is unavailable. If isotopic resolution is available, you may use the weighted-average of isotope masses instead.

For heteromeric complexes, such as the TNH example discussed here, each subunit will have the same MS¹ intact MPC mass. For example, the TNH alpha, beta, and gamma examples from the spreadsheet downloaded in step 3 differ only in their MS² and MS³ masses. The data for the three TNH subunits are populated in columns C, D and E in the Excel spreadsheet.

6. Input 500 Da into the MS¹ tolerance field.

Only candidate MPCs within (+/-) the tolerance around the observed MS¹ mass are reported in the results list. Importantly, this DOES NOT affect the scores of each candidate, only the output list. The value of 500 is chosen in order to account for unannotated modifications and potential noncovalent cofactors that affect the observed complex mass, but may not be present in the database or even retained on the ejected subunit.

7. Input 21083.7 into the MS²-Subunit average mass field, corresponding to the intact mass of the TNH alpha monomer (Fig. XX.XX.1).

8. Input 100 Da into the MS² tolerance field.

Set the tolerance to a value that is much higher than would be expected from the instrument accuracy. Noncovalent adducts as well as unexpected modifications can drastically shift the mass of the ejected monomer proteoform.

9. Input the monoisotopic neutral fragment ion mass list for the TNH alpha monomer found in the spreadsheet from step 3. The first mass in this list is: 553.24907 (Fig. XX.XX.1).

10. Input 10 ppm into the MS³ tolerance field.

The input tolerance is instrument specific; however, a smaller tolerance often produces a more confident subunit identification, to the limits of the mass analyzer employed.

11. Press the “Search” Button. A browser window appears providing the current search status. The search window will update regularly and display the subunit search result.

You can close the search window and re-access it by using its uniquely assigned URL.

12. After several minutes, the results window will display the MPC search results (Fig. XX.XX.3). A single result will appear for this example.

You can access the window even after closing it on a web browser by using its uniquely assigned URL.

Navigating the Results Page

13. Explore the results page (Fig. XX.XX.3). There are nine columns which are populated with result values for the candidate MPC. The first two values, “10013” and “Toyocamycin nitrile hydratase”, indicate the ID number and complex name, respectively, found in the CORUM database. The next value, “B6CWJ3” is the UniProt accession number of the TNH alpha subunit identified from the MS² and MS³ results. In the “Stoichiometry” column, the three interacting subunits (B6CWJ3, 4, and 5, corresponding to the TNH alpha, beta, and gamma subunits) are listed with their theoretical average masses and copy number within the MPC. The TNH example shows two copies of each subunit, thus the identified MPC corresponds with an $\alpha_2\beta_2\gamma_2$ heterohexamers. The “P Score” and “E Value” columns indicate the overall high confidence of the MS³-based alpha subunit identification, at 5×10^{-51} and 2×10^{-45} respectively. The C-score of 364 shows the extremely high level of monomer proteoform characterization. Finally, the MPC-score of 81 indicates the overall high confidence in the characterization of that MPC, and is based on the difference between theoretical and observed MS¹ intact complex mass (Delta Mass) and the subunit C Score. The results can be exported into .XLSX format and downloaded by pressing the green Excel icon in the upper right corner of the window.

As noted above, the TNH complex is not present in CORUM but was previously added manually to the database. Thus the CORUM ID and CORUM complex name cannot be found in that database. However, the UniProt accession number is correct.

For more information on the four scores used to characterize the confidence of MPC identification, see the **Guidelines for Understanding Results** section.

ALTERNATE PROTOCOL 1

Targeted characterization of a protein complex not present in the CORUM

database—The SEMPC relies on the CORUM and UniProt databases to identify and characterize MPCs in discovery mode. However, a complex or subunit that is not present in the CORUM database will not return a correct search result. The SEMPC provides the “Add New Complex” feature to add one or more additional targeted proteins to the search database. The added accession numbers are considered in the subunit search and as a single interaction that is added to the CORUM database for the duration of the search.

Necessary Resources

Hardware: Any computer with internet access

Software: Software capable of reading an Excel (.xlsx) file

Access the SEMPC online

1. Access the SEMPC online at <http://complexsearch.kelleher.northwestern.edu>.

Performing a search: Adding a new complex

2. Access the MS¹, MS², and MS³ data from human cytoplasmic malate dehydrogenase from Table XX.XX.1. Insert the information into the appropriate fields of the SEMPC as described in steps 5-10 of **Basic Protocol**
1. Use the same 500 Da, 100 Da, and 10 ppm tolerances for MS¹, MS², and MS³, respectively as described for the TNH complex above. Do not yet press the search button.
3. Press the “Add New Complex” button (Fig. XX.XX.1). Insert the UniProt accession number for malate dehydrogenase P40925 into the textbox that appears. The addition of a single accession number will search for only homomeric MPCs of that monomer.

Any number of accession numbers can be added to the list, and they will be considered as a single interaction and added to the CORUM database for the duration of a single search. All accession numbers must be valid UniProt accession numbers at the time of the search, but may be from any organism and either annotated or unannotated.

4. Press the “Search” Button. The Results window will appear in several minutes.

At search time, the added accession numbers are used to reference the current version of UniProt, and all of their modifications and sequence variation are used for the search. The usual CORUM and CORUM-filtered UniProt databases are also searched.

Navigating the Results Page

5. Examine the results page (Fig. XX.XX.4). Several unique result fields are populated when an “Add New Complex” MPC is returned. The CORUM ID number of –1 and the complex name of “Custom Complex” indicate that the candidate MPC is not present in the CORUM Database. Interpret the remaining values in the same way as for those found with a regular search. The example result indicates a homodimer, with each monomer N-terminally acetylated. The small delta mass of –1 and extremely high MPC score of 2,983 indicate a high level of confidence associated with this identification.

If a new complex that is already present in the CORUM database is added through the “Add New Complex” feature, the CORUM ID and name will be used instead of the custom complex name.

GUIDELINES FOR UNDERSTANDING RESULTS

The SEMPC provides as a result an ordered and scored list of candidate MPCs that most closely match the input values and set parameters. Critical to the concept of an MPC is the idea of a proteoform, which is a monomeric protein with its specific combination of isoforms, single nucleotide polymorphisms and other primary sequence mutations, and post translational modifications (Smith and Kelleher, 2013). An MPC is defined as a specific assembly of monomer proteoforms, including the overall stoichiometry and modification state of each monomer (Skinner et al., 2016). For example, a homodimeric complex with a single phosphorylated subunit would be considered a different MPC than the same complex that has both subunits phosphorylated, or one with three copies of the same subunit.

Due to the increased complexity in searching intact MPCs, several scoring metrics are used to provide confidence metrics of subunit and intact complex identification. For the subunit proteoform, the *p*-score (Meng et al., 2001) and E-value (Zamdborg et al., 2007) indicate the overall confidence in the identification. Briefly, the *p*-score relies on a Poisson model to predict the probability that an identification happened by chance; smaller scores are more confident. The E-value is the *p*-score multiplied by the database size to apply a Bonferroni correction for multiple hypothesis testing; generally, E-values of below 10^{-4} – 10^{-10} are considered confident. However, both of these metrics can vary greatly based on the data and databases used, thus we recommend a target-decoy approach for accurate false discovery rate generation (Tran et al., 2011). On the other hand, the level of subunit proteoform characterization is indicated by the C-score (LeDuc et al., 2014), where higher values are considered more confident localization of PTMs and polymorphisms within a given proteoform. Finally, the MPC-score is based on the subunit C-score and the MS¹ delta mass value, with higher scores indicating more confident results.

COMMENTARY

Background Information

The recent development of a three-tiered native mass spectrometry approach to identifying and characterizing MPCs (Belov et al., 2013) began the development of the SEMPC as a

bioinformatic tool to facilitate data analysis. In order to better rank candidate MPCs, as well as provide a general metric of the confidence of MPC identification, we then developed the MPC-score, based on the C-score, and its similar Bayesian framework (LeDuc et al., 2014). To our knowledge, the SEMPC is the only search engine capable of identifying and characterizing intact MPCs.

Several other mass spectrometry-based proteomics techniques use enzymatic cleavage and shotgun proteomics techniques to infer protein-protein interactions. They include chemical cross-linking (Liu et al., 2015) and chromatographic co-elution (Havugimana et al., 2012), and offer much higher throughput than that provided by the SEMPC. However, because these techniques rely on protein denaturation and enzymatic digestion, they do not provide MPC stoichiometry or even subunit proteoform information.

Critical Parameters/Troubleshooting

The tolerances set for MS¹ and MS² are critical to the search result. If the values or units are not set properly, no result will be returned. At times, the values must be set much higher than the instrument's mass accuracy to account for non-covalent adducts or unannotated modifications.

ACKNOWLEDGEMENT

The authors thank Nicole Haverland, Ryan Fellers, Joseph Greer and Bryan Early for support and helpful discussions. The authors would like to thank the W. M. Keck Foundation (grant number DT061512) for enabling our team to launch this project. O.S.S. is supported by a US National Science Foundation Graduate Research Fellowship (2014171659). Ancillary support to maintain the SEMPC web-presence from the National Resource for Translational and Developmental Proteomics is also gratefully acknowledged (GM108569).

LITERATURE CITED

- Belov ME, Damoc E, Denisov E, Compton PD, Horning S, Makarov AA, Kelleher NL. From protein complexes to subunit backbone fragments: a multi-stage approach to native mass spectrometry. *Anal Chem*. 2013; 85:11163–11173. [PubMed: 24237199]
- Blackwell AE, Dodds ED, Bandarian V, Wysocki VH. Revealing the Quaternary Structure of a Heterogeneous Noncovalent Protein Complex through Surface-Induced Dissociation. *Analytical Chemistry*. 2011; 83:2862–2865. [PubMed: 21417466]
- Havugimana, Pierre C.; Hart, GT.; Nepusz, T.; Yang, H.; Turinsky, Andrei L.; Li, Z.; Wang, Peggy I.; Boutz, Daniel R.; Fong, V.; Phanse, S.; Babu, M.; Craig, Stephanie A.; Hu, P.; Wan, C.; Vlasblom, J.; Dar, V.-u.-N.; Bezginov, A.; Clark, Gregory W.; Wu, Gabriel C.; Wodak, Shoshana J.; Tillier, Elisabeth R.M.; Paccanaro, A.; Marcotte, Edward M.; Emili, A. A Census of Human Soluble Protein Complexes. *Cell*. 2012; 150:1068–1081. [PubMed: 22939629]
- LeDuc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL. The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top Down Proteomics. *Journal of proteome research*. 2014; 13:3231–3240. [PubMed: 24922115]
- Liu F, Rijkers DTS, Post H, Heck AJR. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat Meth*. 2015; 12:1179–1184.
- Meng F, Cargile BJ, Miller LM, Forbes AJ, Johnson JR, Kelleher NL. Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat Biotech*. 2001; 19:952–957.
- Pesavento JJ, Kim Y-B, Taylor GK, Kelleher NL. Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry. *J. Am. Chem. Soc*. 2004; 126:3386–3387. [PubMed: 15025441]

- Ruepp A, Waegelé B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research*. 2010; 38:D497–D501. [PubMed: 19884131]
- Skinner OS, Havugimana PC, Haverland NA, Fornelli L, Early BP, Greer JB, Fellers RT, Durbin KR, Do Vale LHF, Melani RD, Seckler HS, Nelp MT, Belov ME, Horning SR, Makarov AA, LeDuc RD, Bandarian V, Compton PD, Kelleher NL. An informatic framework for decoding protein complexes by top-down mass spectrometry. *Nat Meth*. 2016 advance online publication.
- Smith LM, Kelleher NL. Proteoform: A single term describing protein complexity. *Nat. Methods*. 2013; 10:186–187. [PubMed: 23443629]
- Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SMM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*. 2011; 480:254–U141. [PubMed: 22037311]
- Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic acids research*. 2007; 35:W701–W706. [PubMed: 17586823]

Search Engine for Multi-Proteoform Complexes [Home](#) [About](#)

Input Complex Information for Search:

MS¹ - Complex

5 Average Mass (Da)
89419

6 Tolerance
500 ppm Da

MS² - Subunit

7 Average Mass (Da)
21083.7

8 Tolerance
100 ppm Da

MS³ - Fragment Ions

9 Monoisotopic Masses (Da)

553.2490596
622.2704215
640.2811061
655.3433271
735.3545783
753.3654634
890.5072066
943.5120606
1013.491922
1103.386375

10 Tolerance
10 ppm Da

Add New Complex **3**

Search

Northwestern
Proteomics

NATIONAL RESOURCE FOR TRANSLATIONAL
AND DEVELOPMENTAL PROTEOMICS

© 2016 - Proteomics Center of Excellence, Northwestern University

Figure XX.XX.1.

The SEMPC input page. The steps of *Basic Protocol 1* are indicated with red numerals; the “Add New Complex” feature discussed in the third step of *Alternate Protocol 1* is indicated with the blue numeral 3. Populated values in this figure correspond with those in *Basic Protocol 1*.

Figure XX.XX.2.

The SEMPC about page. The download link discussed in *Basic Protocol 1* step 3 is indicated with a red circle.

Figure XX.XX.3.
The SEMPC results page for the TNH alpha subunit search. The results are discussed in detail in *Basic Protocol 1*, step 13.

Figure XX.XX.4.
The SEMPC results page for the malate dehydrogenase complex, added via the “Add New Complex” feature. The results are discussed in more detail in *Alternate Protocol 1*, step 5.

Table XX.XX.1

Example Malate Dehydrogenase data for *Alternate Protocol 1*:

MS Tier	Mass (Da)
MS ¹	72673
MS ²	36335.7
MS ³	666.436599
	751.428621
	785.413623
	836.481947
	864.513391
	876.574628
	898.497547
	935.550284
	963.581582
	989.464004
	1011.582117
	1036.599194
	1093.620067
	1110.65043
	1138.60069
	1164.65748
	1199.603352
	1235.695482
	1292.71825
	1312.686513
	1420.775658
	1425.772097
	1450.784189
	1512.805206
	1524.841662
	1533.861224
	1604.899693
	1604.900588
	1628.773116
	1637.92599
	1767.963863
	1788.954299
	1836.985979
	1854.997889
	1950.070438
	1951.074786
	1968.084187

MS Tier	Mass (Da)
	2081.168428
	2103.503099
	2147.163724
	2229.123692
	2244.231336
	2253.134645
	2294.248972
	2331.265022
	2426.339393
	2444.350261
	2457.068493
	2460.983801
	2461.068287
	2501.371674
	2527.760137
	2989.28952
	3005.604315
	3011.559933
	3012.064116
	3012.697534
	3016.402008
	3085.530839
	3122.533409
	3200.077572
	3222.674467
	3305.752533
	3416.797944
	3433.815142
	3530.867267
	3626.94597
	3643.952374
	3731.853127
	3739.018002
	3757.040893
	3852.109422
	3870.12415
	3873.99726
	3969.192879
	4082.275486
	4195.363856
	4310.393408
	4332.371194

MS Tier	Mass (Da)
	4423.479847
	4506.514202
	4524.525542
	4546.512316
	4623.503178
	5289.849085
	5924.100553
	6039.134343
	6810.44711
	6828.460496
	8321.766628
	9287.630292
	9307.127675
	13187.70114
	13206.71526
	18585.05359
	22464.57502
	28794.3512
	29492.77396
	30141.94226
	30743.20572
	30759.20069
	30778.28113
	30872.33644
	30890.32398
	30981.69081
	30988.36495
	31027.39727
	31046.40117
	31105.37276
	31124.61878
	31142.46817
	31239.51484
	31276.56428
	31353.67569
	31377.66253
	31411.68307
	31433.65631
	31697.74692
	31735.89995
	31754.80246
	31771.80403

MS Tier	Mass (Da)
	31782.72017
	31790.72971
	31799.73668
	31812.70507
	31834.77051
	31855.75364
	31871.74448
	31890.84396
	31911.75069
	31930.70809
	31967.75967
	31987.98071
	32003.93333
	32025.91769
	32066.90305
	32083.86549
	32101.89803
	32120.86909
	32138.86847
	32212.95706
	32234.05131

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript