Computer Modeling of the Effectiveness of MS1 and MS2 for mRNA Sequence Confirmation

Clara Bern[1], Steven Broome[2], Doron Kletter[2], Gary Wilson[2], Ignat Shilov[2], Maria Basanta-Sanchez[2], Marshall Bern[2]

[1]Carlmont High School, Belmont, CA [2]Protein Metrics, LLC. Contact: bern@proteinmetrics.com

Introduction

Mass spectrometry (MS) offers a single-instrument method for verification of mRNA vaccines including primary sequence, 5' cap, poly-A tail, and modifications. Due to their size, mRNAs must be enzymatically digested for MS analysis. See Gau, Dawdy, *et al*., 2023, PMID 37270636. Challenges include:

- **Size of mRNAs**. ~4300 nucleotides and ~1.3 MDa for Covid-19 vaccines
- **Duplicate digestion pieces.** ACAAG appears three times as a fully specific RNase T1 piece in the Pfizer-BioNTech vaccine (Comirnaty).
- **Isomeric digestion pieces.** There are also isomers of ACAAG. AACAG appears three times and CAAAG appears once in Comirnaty.
- **Similar MS2 spectra.** The MS2 spectra of ACAAG and AACAG differ in only a few ions, those that fragment between the 2nd and 3rd nucleotides. And even some of these ions are ambiguous, as the ions for AC in ACAAG may also appear as internal fragments in the MS2 spectrum of AACAG.

Digestion-specific pieces of **unique mass**, different from any other digestion-specific piece, are especially advantageous, because they can be identified by MS1 alone. But how frequent are unique-mass pieces in a long mRNA? How does their frequency vary with mRNA length, piece length, and digestion efficiency? How many MS2 peaks are needed to distinguish isomers?

SCAN FOR POSTER

The authors declare no competing financial interests.

Computational Experiments

We wrote a **Python** program for the following **Monte Carlo simulation**:

1. Pick a random string *S* of length *n* over the letters A, C, G, and either U or V, assuming independent equal probability for each letter at each position.
2. Cut *S* either (a) at each G to form substrings $s_1, s_2, \ldots s_k$, ending in G with at least *minG* and at most *maxG* internal G's and mass ≤ 20 kDa, or (b) at each letter to form $s_1, s_2, \ldots s_k$, with mass at least *minM* and at most *maxM*.
3. Record the masses of $s_1, s_2, \ldots s_k$, assuming A, C, G, U, and V weigh exactly 329, 305, 345, 306, and 320 Da (for methylpseudouridine) respectively.
4. Optionally record a_2-B, a_3-B,… ions of $s_1, s_2, \ldots s_k$, where the mass of the a_j-B ion is the sum of the masses of the first *j* letters minus the base mass of the j^{th} letter, that is,135, 111, 151, 112, and 126 for A, C, G, U, and V.
5. Count the number and coverage of substrings $s_1, s_2, \ldots s_k$ with unique records.
6. Clear *S*, substrings of *S*, and records of substrings and return to step 1.

This algorithm makes some implicit assumptions:

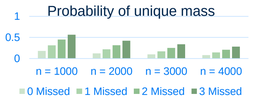
A. Mass spectrometry can give exact nominal monoisotopic mass for pieces ≤ 20 kDa.

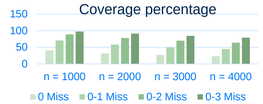
B. Specific digestion depends only upon the 5' side of the cleavage. Although the 5' assumption is not always true (see right), adding the 3' side would give similar results (with different *minG, maxG* values) for random strings.

C. a-B fragments represent MS2 analysis. We chose a-B ions, because they are unique to prefixes (not true of abcd / wxyz) and less likely to occur as internal fragments.

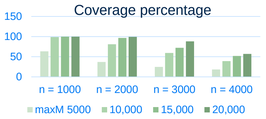
Results

MS1 only. As expected, mass uniqueness and coverage increase with missed cleavages. Numbers shown are means over 100 random mRNAs of length *n*. It takes roughly 4 missed cleavages for > 99% coverage by unique-mass pieces for 1000-mers, 5 for 2000-mers, etc. 100% coverage is rare due to clustered G's.



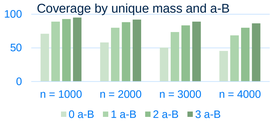


Coverage by unique **nonspecific** pieces increases with *maxM*. (*minM* is fixed at 1000.) Good coverage requires high *maxM* – 10,000 Da corresponds to ~7 missed G cleavages – due to ~4x more pieces with nonspecific digestion.

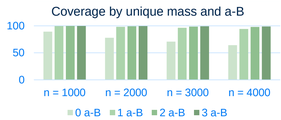


MS1 and MS2. Even a small number of a-B ions greatly increases the coverage by unique pieces. Here we show results for specific digestion with 0 to 1 and 0 to 2 missed cleavages, but the results are similar for other digestions.

0 – 1 Missed Cleavages



0 – 2 Missed Cleavages



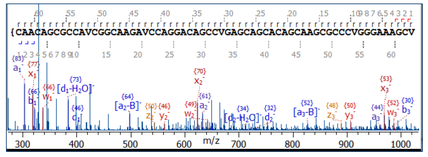
Conclusions and Discussion

The computational experiments support these conclusions:

- **MS1 alone** cannot give high coverage for mRNAs with length > 1000.
- **a-B ions** can give uniqueness in a 4000-nucleotide mRNA**.** Selected a-B ions, such as a_2-B to a_4-B, could be used in MRM or DIA methods.
- **Almost complete** coverage (say 99%) by unique digestion-specific peptides is more common than complete coverage due to subsequences like GGGG**.**
- **Missed cleavages** are helpful, but too many missed cleavages (see below) produce crowded MS1 spectra with uncertain monoisotopic mass calls and low-SNR MS2 spectra with poor fragmentation.

RNase 4 cuts at U|V.A and U|V.G, but also misses many cleavage sites and produces mostly 10-mers to 50-mers. This 63-mer has unique mass among sequences V.xxxV.x, and a2-B and a3-B show that the sequence starts with CAA.

Data from Wolf *et al*., *Nucl Acids Res.*, 2022, PMID 35871301

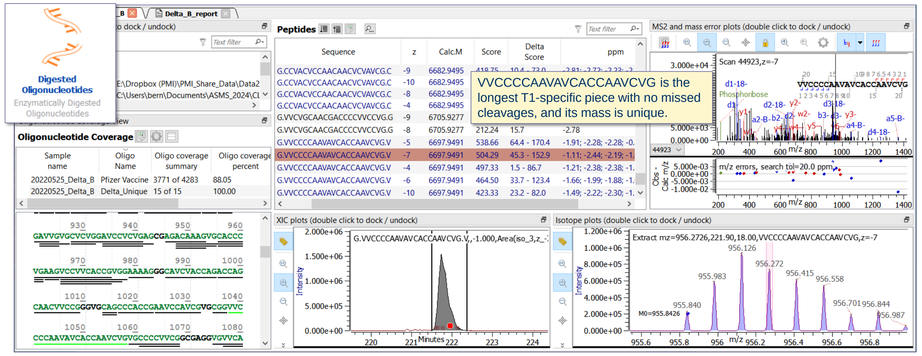


Software for Sequence Confirmation

RNase T1 digest of Pfizer-BioNTech vaccine. Data from Gau, Dawdy, *et al*., *Sci Rep*., 2023, PMID 37270636

PMI's **Digested Oligonucleotides** workflow imitates peptide mapping of protein therapeutics:

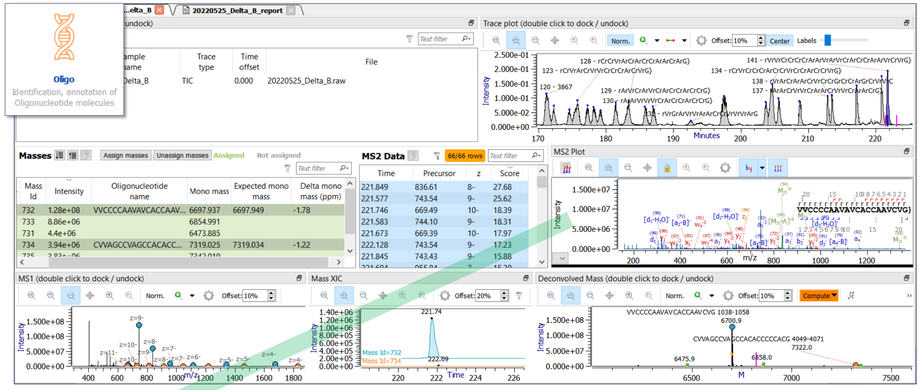
- FASTA contains the full mRNA for *in silico* digestion. V = methyl-pseudouridine, which poses less problems than U or Ψ (1 Da from C)
- Database search matches MS2 scans to pieces of the full mRNA.
- The GUI includes TIC or UV, XICs, annotated MS2, and MS1 peak profile.
- We found Delta Score, ppm error, and manual curation more effective than target / decoy FDR for confirming OSMs (oligo spectrum matches).



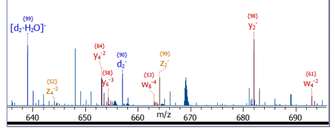
VVCCCCAAVAVCACCAAVCVG is the longest T1-specific piece with no missed cleavages, and its mass is unique.

PMI's **Oligo** workflow was designed for small synthetic oligos, but can be used for digested mRNAs as well.

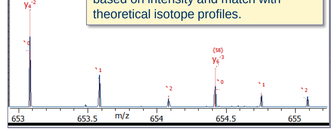
- FASTA contains pieces of the mRNA.
- Three letter code allows a mix of bases, sugars (r = ribose, d = deoxyribose, m = methylated, etc.), and 5' and 3' caps. Here } means 3' phosphate.
- The GUI includes TIC or UV trace, "Mass XIC" (all charges combined), m/z, and neutral mass deconvolution..
- User selects MS2 scans to sum and annotate.
- Summed profile-mode scans give better fragmentation coverage than centroided single scans (see above).
- PMI workflows are **complementary**. **Oligo** adds TIC or UV, both MS1 and MS2 scan summing, and charge deconvolution.



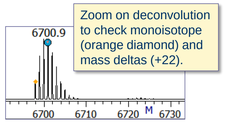
Zoom on MS2 to check peak assignments. The colored stick at monoisotopic *m/z* shows intensity integrated over isotopes.



MS2 annotation of profile-mode spectra gives individual peak scores (fe.g., {84}) based on intensity and match with theoretical isotope profiles.



Zoom on deconvolution to check monoisotope (orange diamond) and mass deltas (+22).



The low-abundance mass at 7322 Da is a missed-cleavage unique-mass piece with supporting MS2, shown in both workflows.