# Supplementary Materials

# To

# Open MS/MS Spectral Library Search to Identify Unanticipated Post-Translational Modifications and Increase Spectral Identification Rate

## 1 DATABASE DESCRIPTIONS

The protein sequence database used for the ISB-18mix dataset consists of the following three parts:

- **Standard proteins**: the 18 standard proteins mentioned in Ref. [1].
- **Pollution proteins**: the 17 laboratory proteins and the 36 dust/contact proteins from cRAP [2]; and the 15 pollution proteins mentioned in [2]. Note that there were 3 proteins of the 36 dust/contact proteins from cRAP replicated to the 3 of the 18 standard proteins, so the amount of the pollution proteins was 65 (17+36+15-3).
- **Database background proteins**: the 5883 proteins from the yeast database orf_trans_all.20070815 [3].

Thus, the protein sequence database contained 5966 (18+65+5883) proteins in total. The corresponding target-decoy sequence database contained 11932 proteins.

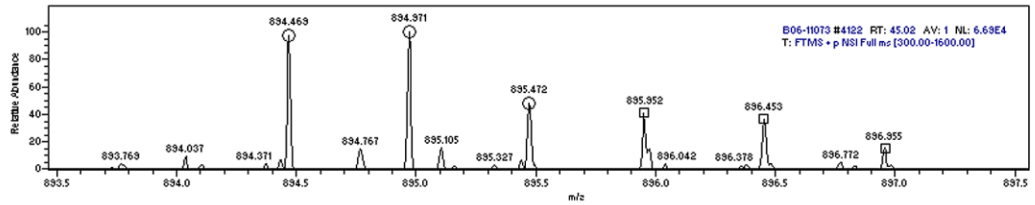
### References

[1] Klimek, J., et al. (2008) The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools, Journal of Proteome Research, 7, 96-103.

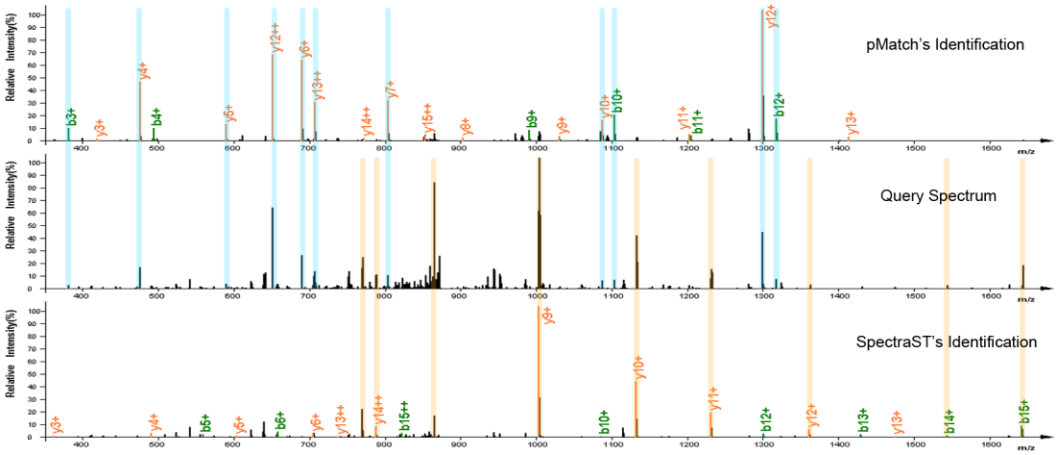
[2] http://www.thegpm.org/crap/index.html

[3] ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_protein/archive/orf_trans_all.20070815.fasta.gz

## 2 FIGURES

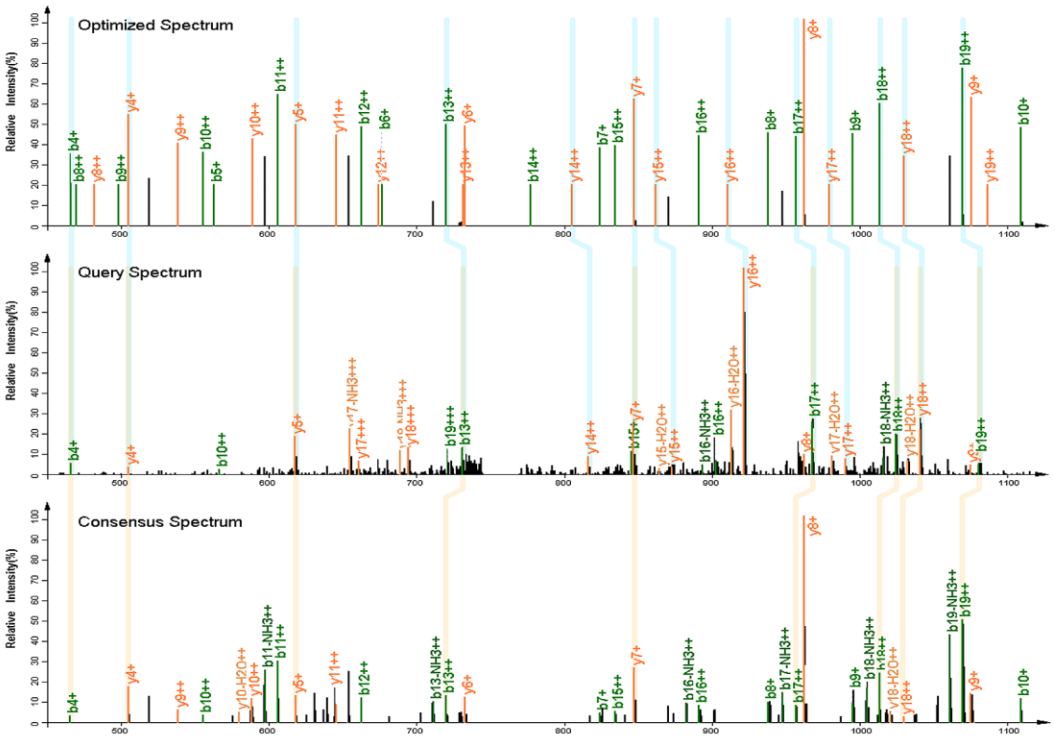



(a)

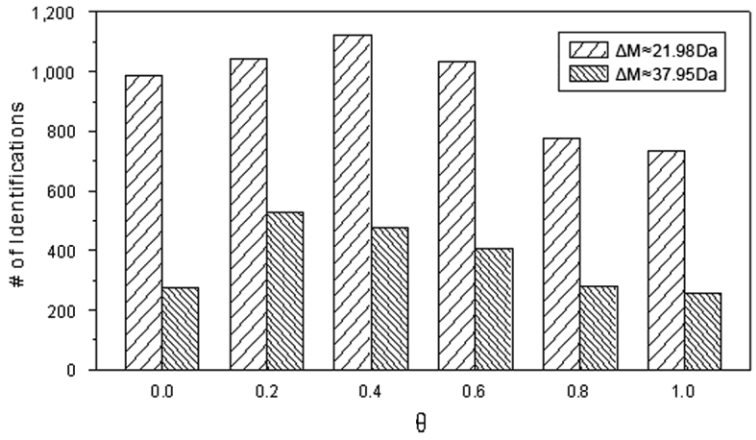



(b)

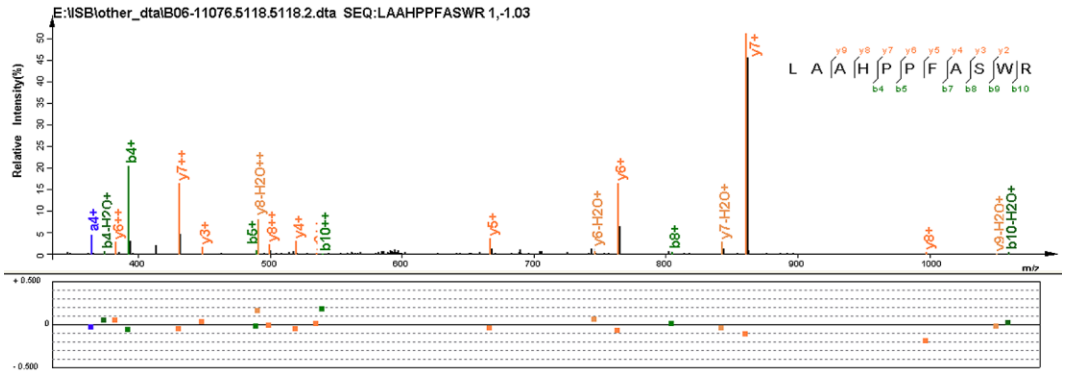
**Supplementary Fig. 1.** A co-eluted spectrum identified in the ISB-18mix dataset. Fig. (a) shows the precursor ions of the query spectrum in the corresponding MS. There are obviously two clusters of isotope peaks annotated respectively by circles and squares. In Fig. (b), the three spectra from top to bottom are: the library spectrum of pMatch's identification with peptide sequence of SYELPDGQVITIGNER, the query spectrum, and the library spectrum of SpectraST's identification with sequence of TITLEVEPSDTIENVK. These two sequences have their theoretical precursor mass fitting the m/z value of the two mono isotope peaks in Fig. (a). Most of the significant peaks from the query spectrum are hit with peaks from either library spectrum.



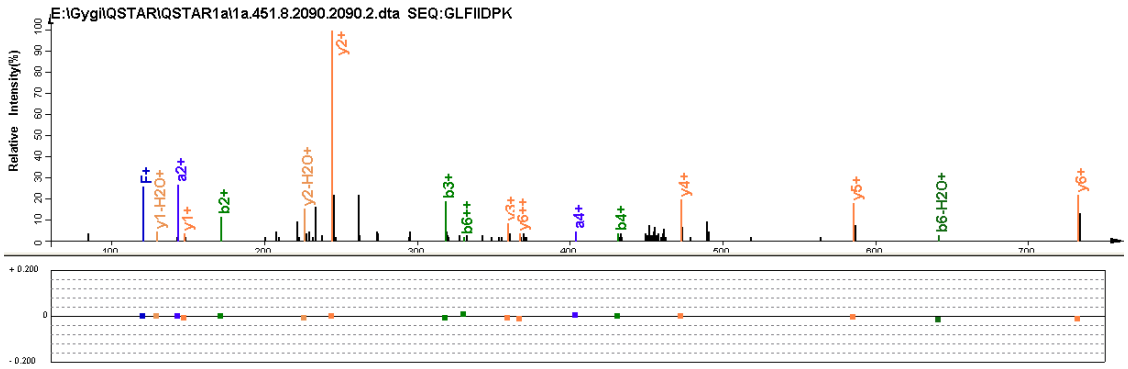

**Supplementary Fig. 2.** An example of impact on peptide's fragmentation pattern by a sodium adduct. The three spectra from top to bottom are: the pMatch's optimized library spectrum, the query spectrum with a sodium adduct, and the consensus library spectrum. The peptide sequence of the query spectrum is IITHPNFNGNTLDNDIMLIK with a sodium adduct, while both library spectra are with the same sequence but sodium-free. The consensus library spectrum does not contain any significant peaks of y14++, y15++, y16++, and y17++, while the query spectrum has those four, in particular y16++ as the most intensive peak. The optimized spectrum is armored by full sequence information, all the unobserved peaks including those four mentioned ones budded at the corresponding m/z position, with a relatively low intensity θ.



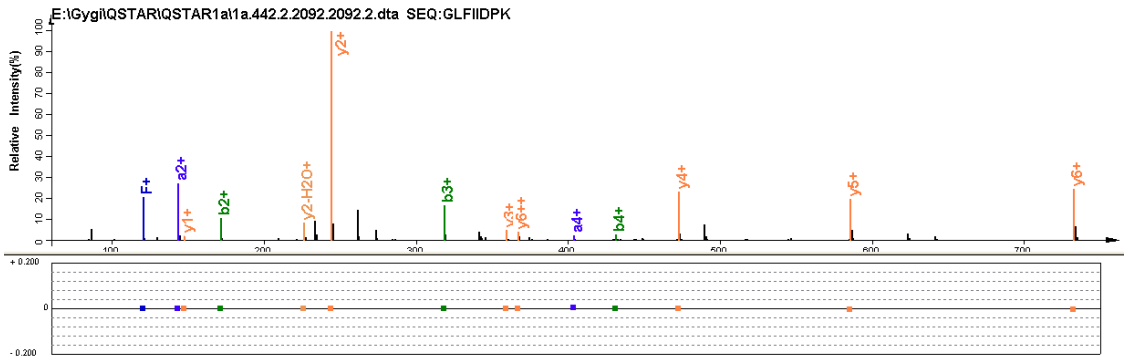

**Supplementary Fig. 3.** Histogram of two ΔM detected from the 1% FDR results of pMatch in the ISB-18mix dataset for different search parameters. The budding strategy raised the numbers of the identified sodium (ΔM approximate 21.98 Da) and calcium (ΔM approximate 37.95 Da) adducted spectra, using a proper θ value.




**Supplementary Fig. 4.** An unknown ΔM detected in the ISB-18mix dataset. The only ΔM of -1.029 Da that is unable to explain in Table 1 is identified by both engines. Unimod (www.unimod.org) has recorded a kind of PTM named lysaminoadipicsealde. This PTM is an oxidation to aminoadipic semialdehyde and has the mass of -1.031 Da, which is matching the -1.029 Da considering the bias from the instrument, but only lysine is recorded as the possible happening sites. However, pMatch identified the ΔM happened on peptide N-term. Additionally, the peak annotations in MS/MS and the corresponding MS suggest it is unlikely to be caused by co-eluted spectra.




(a)




(b)

**Supplementary Fig. 5.** Incorrectly judged precursor ion mass in the Gygi-Qstar dataset. There is a number of spectra identified with ΔM discrete from -20 Da to -3 Da. Here is an example of the tandem mass spectra. As shown, the query spectrum in Fig. (b) has its identification result with ΔM of -19.07 Da. Fig. (a) illustrates the corresponding library consensus spectrum. It is clear that there is not any noticeable mass shift between the two peak lists. Most of intensive peaks are explained with very low m/z errors.

## 3 TABLES

**Supplementary Table 1.** Abundant ΔM detected in the ISB-18mix dataset

| ΔM (Da) | # of Spectra | | ΔM Type | Description |
|---|---|---|---|---|
| | pMatch | SpectraST | | |
| 0.984 | 1,435 | 1,691 | Modification | Deamidation (N / Q), or Co-eluted spectra |
| 21.986 | 1,044 | 1 | Modification | Sodium |
| 152.001 | 599 | 12 | Modification | CarbamidomethylDDT (C*) |
| 37.954 | 527 | 0 | Modification | Calcium |
| -17.024 | 369 | 261 | Modification | Ammonia loss (Q / N / K / C*) |
| 26.019 | 358 | 65 | Modification | Acetaldehyde +26 (N-term) |
| -18.008 | 147 | 87 | Modification | Dehydration (D / E) |
| 16.000 | 109 | 23 | Modification | Oxidation (M / W / Y / H) |
| 22.970 | 96 | 0 | Modification | Sodium, & Deamidation (N / Q) |
| -113.082 | 90 | 31 | Semi-digestion | I / L loss (N-term) |
| -212.148 | 86 | 18 | Semi-digestion | VI / VL / IV / LV loss (N-term) |
| -71.036 | 84 | 16 | Semi-digestion | A loss (N-term) |
| 1.968 | 82 | 98 | Modification | Two deamidations (N / Q), or Co-eluted spectra |
| -170.101 | 78 | 19 | Semi-digestion | AV / VA / GI / GL loss (N-term) |
| -128.095 | 70 | 0 | Semi-digestion | K loss (Missed cleaved C-term) |
| -128.057 | 68 | 8 | Semi-digestion | AG / GA / Q loss (N-term) |
| -142.074 | 61 | 0 | Semi-digestion | AA loss (N-term) |
| 5.994 | 55 | 0 | Modification | Sodium, & No oxidation (M*) |
| -116.056 | 52 | 0 | Modification | A disulfide bridge (Two C*) |
| -99.065 | 52 | 11 | Semi-digestion | V loss (N-term) |
| -101.046 | 48 | 8 | Semi-digestion | T loss (N-term) |
| -242.125 | 47 | 1 | Semi-digestion | EI / EL / IE loss (N-term) |
| 14.019 | 41 | 0 | AA substitution | V->I / L |
| -15.006 | 40 | 2 | Modification | Deamidation (N), & No oxidation (M*) |
| -0.978 | 40 | 44 | AA substitution | E->Q / D->N, or Co-eluted spectra |
| 21.958 | 40 | 0 | Modification | Calcium, & No oxidation (M*) |
| 128.101 | 40 | 0 | Missed cleavage | K added (N-term / C-term) |
| -186.079 | 37 | 0 | Semi-digestion | W loss(N-term) |
| 156.103 | 31 | 0 | Missed cleavage | R added (N-term / C-term) |
| -118.149 | 30 | 0 | Semi-digestion, & Modification | R loss (N-term / C-term), & Calcium |
| -1.029 | 30 | 16 | *?* | *Happen on N-term* |
| 152.987 | 29 | 0 | Modification | CarbamidomethylDDT (C*), & Deamidation (N) |
| -215.102 | 28 | 0 | Semi-digestion | QS / NT loss (N-term) |
| 173.984 | 27 | 0 | Modification | CarbamidomethylDDT (C*), & Sodium |
| 58.007 | 26 | 0 | AA substitution | A->E, or G->D |
| -172.048 | 25 | 0 | Semi-digestion | DG / GD loss (N-term) |
| 38.935 | 22 | 0 | Modification | Calcium, & Deamidation (N / Q) |
| -106.960 | 21 | 0 | *False positive* | - |
| -214.095 | 20 | 0 | Semi-digestion, & Modification | TP loss (N-term), & No oxidation (M*) |
| -58.002 | 20 | 0 | AA substitution | E->A, or D->G |
| -16.038 | 20 | 11 | Modification | Ammonia loss (Q / N / C*), & Deamidation (N) |

C*: Carbamidomethylated cysteine

M*: Oxidated methionine

**Supplementary Table 2.** Main search parameters for the experiments on the four extra datasets

| Dataset | Protein Sequence Database Search (pFind v2.3)[#] | | | | Library Search (pMatch v1.0) | | |
|---|---|---|---|---|---|---|---|
| | Sequence Database | Search Tolerance | | Filtration Threshold of Precursor | Search Tolerance | | θ |
| | | Precursor | Fragment | | Precursor | Fragment | |
| TAP-PSD95 | ipi.MOUSE.v3.64 | ±50 ppm | ±0.5 Th | -8 ~ +8 ppm | ±300 Da | ±0.5 Th | 0.2 |
| HUPO-14 | ipi.HUMAN.v3.65 | ±50 ppm | ±0.5 Th | -7 ~ +7 ppm | ±300 Da | ±0.5 Th | 0.2 |
| Haas-Data | orf_trans_all.20070815[*] | ±50 ppm | ±0.5 Th | -15 ~ +15 ppm | ±300 Da | ±0.5 Th | 0.2 |
| Gygi-Qstar | orf_trans_all.20070815[*] | ±1.0 Da | ±0.2 Th | -0.2 ~ +0.2 Da | ±300 Da | ±0.2 Th | 0.2 |

# For the protein sequence database searches on all of the four datasets, full tryptic specificity was applied, allowing up to two missed cleavage sites; carbamidomethylation of cysteine was specified as a fixed PTM, and oxidation of methionine as a variable one.

* ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_protein/archive/orf_trans_all.20070815.fasta.gz