Modeling Contaminants in AP-MS/MS Experiments

Mathieu Lavallée-Adam,[†] Philippe Cloutier,[‡] Benoit Coulombe,[‡] and Mathieu Blanchette[*,†]

McGill Centre for Bioinformatics and School of Computer Science, McGill University, Montréal, and Gene Transcription and Proteomics Laboratory, Institut de recherches cliniques de Montréal, Montréal

E-mail: blanchem@mcb.mcgill.ca

Phone: 514-398-5209. Fax: 514-398-3387

Supporting Information Available

Protein and peptide identification software information

Peaklists were created using extract_msn.exe version 2005-02-15 (Thermo Xcalibur) with the following parameters: minimum mass: 600, maximum mass: 6000, minimum number of fragment ions: 10, no grouping of MS/MS spectra was performed, and precursor charge was set to automatic. Mascot 2.2.04 (Matrix Science) was used for protein database searching with precursor-ion mass tolerance set to 10 ppm and fragment-ion mass tolerance set to 0.6 Da. The modifications allowed were carbamidomethylation and oxidation of methionine. Finally, the digestion enzyme used was trypsin and 2 missed cleavages were allowed. Database searching was performed on the human NCBI nr protein database (version 2009-04-02), which contains 10 427 007 sequences.

[*]To whom correspondence should be addressed

[†]McGill Centre for Bioinformatics and School of Computer Science, McGill University, Montréal

[‡]Gene Transcription and Proteomics Laboratory, Institut de recherches cliniques de Montréal, Montréal

Computation of corrected averages for Mascot scores

We note that in most AP-MS/MS applications, preys with Mascot scores below a certain threshold m (e.g. a fixed value $m = 20$, or the Mascot Identity Threshold[1]) are discarded and not reported, as being likely protein identification errors. In our approach, when a protein p is not reported as a possible partner of bait b, we arbitrarily set its Mascot score $M_{b,p}^{NI}$ to zero. The set of observed Mascot scores for a given prey thus follow a type I censored distribution.[2] Let B'_p be the set of control experiments for which $M_{b,p}^{NI} < m$. Assuming the uncensored $\bar{M}_p^{NI}$ follows a normal distribution, a better estimate of $\mu_{\neq b,p}$ is thus obtained from the Persson-Rootzen method:[3]

$$\mu_{\neq b,p} = \frac{1}{|B'_p|} \sum_{b \in B'_p} M_{b,p}^{NI} - \gamma_p \sigma',$$

where $\gamma_p = \phi(\lambda_{|B'_p|/|B|})|B|/|B'_p|$, ϕ is the probability density function of the standard normal distribution,

$$\sigma' = \frac{1}{2}\left[\lambda_{|B'_p|/|B|} \frac{1}{|B'_p|} \sum_{b \in B'_p} (M_{b,p}^{NI} - m) + \left\{ \left(\lambda_{|B'_p|/|B|} \frac{1}{|B'_p|} \sum_{b \in B'_p} (M_{b,p}^{NI} - m) \right)^2 + \frac{4}{|B'_p|} \sum_{b \in B'_p} (M_{b,p}^{NI} - m)^2 \right\}^{\frac{1}{2}} \right],$$

and where $\lambda_{|B'_p|/|B|}$ denotes the upper $(|B'_p|/|B|)^{th}$ quantile of the standard normal distribution. If $M_{b,p}^{NI} = 0 \; \forall b \in B$ for a given p, we set arbitrarily one $M_{b,p}^{NI}$ to be equal to $m+1$.

C^c correction factor derivation

In order to correct the C^s matrix for induced experiments noise modeling, we used the following correction:

$$C^c(i,j) = C^s(i,j) \cdot \frac{I(j)}{NI(j)}$$

The above correction was derived the following way. The matrix C^s corresponds to the joint probability of $\bar{M}_p^{NI}$ and $M_{b,p}^{NI}$ given that the data was generated from control experiments.

$$C^s(i,j) = \Pr[\bar{M}_p^{NI} = j, M_{b,p}^{NI} = i | control] = \Pr[\bar{M}_p^{NI} = j | control] \cdot \Pr[M_{b,p}^{NI} = i | \bar{M}_p^{NI} = j, control]$$

We make the assumption that the noise of a Mascot score is independent of whether the experiment was induced or not. Therefore:

$$\Pr[M_{b,p}^{NI} = i | \bar{M}_p^{NI} = j, control] = \Pr[M_{b,p}^{NI} = i | \bar{M}_p^{NI} = j]$$

Similarly, let C^c correspond to the joint probability of $\bar{M}_p^{NI}$ and $M_{b,p}^{NI}$ given that the data was generated from induced experiments.

$$C^c(i,j) = \Pr[\bar{M}_p^{NI} = j, M_{b,p}^{NI} = i | induced] = \Pr[\bar{M}_p^{NI} = j | induced] \cdot \Pr[M_{b,p}^{NI} = i | \bar{M}_p^{NI} = j]$$

Following from the above assumption,

$$C^c(i,j) = \Pr[\bar{M}_p^{NI} = j | induced] \cdot \frac{\Pr[\bar{M}_p^{NI} = j, M_{b,p}^{NI} = i | control]}{\Pr[\bar{M}_p^{NI} = j | control]}$$

which give us the correction factor for C^s in order to get C^c.

$$C^c(i,j) = C^s(i,j) \cdot \frac{\Pr[\bar{M}_p^{NI} = j | induced]}{\Pr[\bar{M}_p^{NI} = j | control]}$$

or as described in the Methods section:

$$C^c(i,j) = C^s(i,j) \cdot \frac{I(j)}{NI(j)}$$

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Perkins, D.; Pappin, D.; Creasy, D.; Cottrell, J.; et al., *Electrophoresis* **1999**, *20*, 3551–3567.

(2) Bernholtz, B. *Statistical Papers* **1977**, *18*, 2–12.

(3) Persson, T.; Rootzen, H. *Biometrika* **1977**, *64*, 123.