Mass spectrometry data processing

March 19, 2017

Spectral search

Spectral search was carried out with MaxQuant v.1.5.0.2 (search details from Erwin go here) against Ensembl 78. The following modifications were searched: methionine oxydation, phosphorylation on S, T and Y, N-term acetylation and pyro-Q on Gln and Glu. pyro-Q was considered as a sample preparation artefact and the peptides were further processed as if they were unmodified. In addition, spectra with a score < 40 were removed from further analysis. All ratios were expressed in natural logarithmic form.

SILAC pairs information was extracted from the evidence.txt file, while all MS/MS information was collected from the msms.txt file.

Ratio normalization

To normalize for biases in protein extraction efficiency between biological replicates, we substracted the median of the biological replicate in which the ratio was observed.

Isoforms

To accomodate isoforms, we assumed that all isoforms of a protein were expressed similarly. We generated an artificial database where for each gene in Ensembl 78 (identified by an ENSG id), we selected the longest isoform as reference protein (ENSP id). All peptide of other isoforms of the corresponding ENSG that were not mapped into the longest ENSP were added at the end, separated by a J amino acid (Jinho has a reference here).

Counting effective observations

In order to account for uncertainties in the MS identification of a SILAC pair $\mathcal{P}$ ratio to a (potentially modified) peptide, we computed the effective number of times a has been observed $\omega = 1$. For a modified peptide, we list all possible peptide modifications and rate their probabilities. We further correct for cofragmentation probabilities in the case where an MS^2 spectra was observed only for a light or heavy peak.

Peptide modifications

We extracted modification probabilities all MS^2 spectra were extracted from the msms.txt file. For unmodified peptides, the corresponding effective number of observation $\omega = 1$. For modified peptides, let A be an indicator matrix between the phosphosites j and all combination of potential peptides i with the correct set of modifications.

If A can be inversed, we get that

$$\omega_i = \sum_j \left(A^{-1}\right)_{ij} p_j$$

with p being the site probabilities extracted from MaxQuant.

If A is not solvable, it becomes an optimization problem:

$$\min \left[\sum_j p_j \sum_i \left(A^{-1}\right)_{ij} \omega_i \right]$$

Assessing cofragmentation

Let $\mathcal{P}$ indicate a SILAC pair and $C_{L,H}$ indicate whether co-localization has occured ($C = 1$) or not ($C = 0$) for the light and heavy peaks respectively. This is determined by having multiple sequences scored in the msms.txt file for a single Scan number and raw file. From those silac pairs where both heavy and light MS² spectra have been observed, we estimate the four probabilities:

$$P(C_L, C_H|\mathcal{P}) = \begin{cases} p^C_{11} & \text{for } C_L = 1, C_H = 1 \\ p^C_{01} & \text{for } C_L = 0, C_H = 1 \\ p^C_{10} & \text{for } C_L = 1, C_H = 0 \\ p^C_{00} & \text{for } C_L = 0, C_H = 0 \end{cases}$$

Let $\mathcal{N} = \{C_L = 0, C_H = 0\}$ (no co-localisation problems), so $P(\mathcal{N}|\mathcal{P}) = p^C_{00}$. Let also p^{top}_L, p^{top}_H be the probability for the top-scoring peptide according to the MS² analysis for L and H respectively. Let $\mathcal{S}^{top}$ indicate the case where the top scoring peptide is the same for the L and H MS² analysis. Focusing on all SILAC pairs where there is no colocalization problems we can assess by global statistics

$$P(\mathcal{S}^{top} \,|\, p^{top}_L, p^{top}_H, \mathcal{N}, \mathcal{P})$$

$$P(p^{top}_L, p^{top}_H \,|\, \mathcal{N}, \mathcal{P})$$

The effective number of observations, n^{eff}_i, for a p-peptide i in cases where both MS² spectra have been obtained is given by $n^{eff}_i = p_L(i) \cdot p_H(i)$, where $p_{L,H}(i)$ is the probability of peptide i in the L and H MS-2 analysis, respectively. If only one MS² spectra has been observed, say for the light sample, we can calculate the expected effective number of observations for the top scoring peptide i_{top} according to the observed p^{top}_L and the global statistics derived above. Specifically,

$$\begin{aligned} P(p^{top}_H, \mathcal{S}^{top}, C_H = 0|p^{top}_L, C_L = 0, \mathcal{P}) &= P(p^{top}_H, \mathcal{S}^{top} \,|\, p^{top}_L, \mathcal{N}, \mathcal{P}) P(C_H = 0|p^{top}_L, C_L = 0, \mathcal{P}) \\ &= P(\mathcal{S}^{top}|p^{top}_L, p^{top}_H, \mathcal{N}, \mathcal{P}) \cdot P(p^{top}_H|p^{top}_L, \mathcal{N}, \mathcal{P}) \cdot P(C_H = 0|p^{top}_L, C_L = 0, \mathcal{P}) \\ &= P(\mathcal{S}^{top}|p^{top}_L, p^{top}_H, \mathcal{N}, \mathcal{P}) \cdot \frac{P(p^{top}_H, p^{top}_L|\mathcal{N}, \mathcal{P})}{P(p^{top}_L|\mathcal{N}, \mathcal{P})} \frac{P(\mathcal{N}|\mathcal{P})}{P(C_L = 0|\mathcal{P})} \\ &= P(\mathcal{S}^{top}|p^{top}_L, p^{top}_H, \mathcal{N}, \mathcal{P}) \cdot \frac{P(p^{top}_H, p^{top}_L|\mathcal{N}, \mathcal{P})}{P(p^{top}_L|\mathcal{N}, \mathcal{P})} \cdot \frac{p^C_{00}}{p^C_{00} + p^C_{01}} \end{aligned}$$

Let $\tilde{P}(p^{top}_H) = P(p^{top}_H, \mathcal{S}^{top}, C_H = 0|p^{top}_L, C_L = 0, \mathcal{P})$. The expected effective number of observations is then given as

$$\langle n^{eff}_{i-top} \rangle (p^{top}_L) = p^{top}_L \cdot \langle p^{top}_H \rangle = p^{top}_L \cdot \int p^{top}_H \cdot \tilde{P}(p^{top}_H) dp^{top}_H$$

Specifically, we computed this statistics in the following bins: 0, .3, .5, .6, .7, .8, .9, .95, 1.

Bayesian variance model

Let λ be the precision of a gaussian distribution with mean μ and let

$$\Gamma(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{1}$$

be the probability density of this precision , where b is the inverse scale of the gamma-distribution and a the shape[1].

Let $D = \{x_1, \cdots, x_n\}$be a set of peptide ratio observations, we have

[1] Note that χ^2_ν is a special case of the Γ-distribution with $a = \nu/2$ and $b = \frac{1}{2}$. Furthermore, variance estimates $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ are distributed according to $s^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$

$$
\begin{aligned}
P(D|\mu,\lambda) &= \frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\lambda \sum (x_i-\mu)^2\right) \\
&= \frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\lambda \left[n(\bar{x}-\mu)^2 + (n-1)s^2\right]\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_i x_i \\
s^2 &= \frac{1}{n-1}\sum_i (x_i-\bar{x})^2
\end{aligned}
$$

So

$$
\begin{aligned}
P(\bar{x}, s^2|\mu,\lambda,n) &= P(\bar{x}|\mu,\lambda,n)P(s^2|\lambda,n) \\
&= \frac{(n\lambda)^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}n\lambda(\bar{x}-\mu)^2\right) \cdot \gamma \cdot \frac{(\gamma s^2)^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)} \exp\left(-\frac{1}{2}\gamma s^2\right), \quad \text{where} \\
\nu &= n-1 \\
\gamma &= \lambda\nu
\end{aligned} \tag{2}
$$

The latter expression shows that $s^2\gamma = s^2\nu/\sigma^2 \sim \chi^2_\nu$. Let the prior be on the form $P(\mu,\lambda) = P(\lambda|\mu)P(\mu)$. Then

$$
P(\bar{x}, s^2, \lambda|\mu, n) = P(\bar{x}, s^2|\lambda,\mu,n)P(\lambda|\mu)
$$

The posterior distribution for the precision is then given by

$$
\begin{aligned}
P(\lambda|\mu,s^2,n) &= \frac{P(s^2|\lambda,n)P(\lambda|\mu)}{P(s^2|n)} \\
&= \frac{1}{\Gamma(a_s)} b_s^{a_s} \lambda^{a_s-1} \exp(-b_s\lambda), \quad \text{where} \\
a_s &= \frac{\nu}{2} + a(\mu) \\
b_s &= b(\mu) + s^2 \cdot \frac{\nu}{2}
\end{aligned} \tag{3}
$$

The marginalization wrt. λ becomes,

$$
\begin{aligned}
P(\bar{x}|\mu,s^2,n) &= \int P(\bar{x}|\mu,\lambda,n)P(\lambda|\mu,s^2,n)d\lambda \\
&= \frac{\sqrt{n}\cdot b_s^{a_s}}{\sqrt{2\pi}\cdot\Gamma(a_s)} \int \lambda^{a_s-1/2} \exp\left(-\frac{1}{2}n\lambda(\bar{x}-\mu)^2 - b_s\lambda\right) d\lambda \\
P(\bar{x}|\mu,s^2,n) &= \sqrt{\frac{n}{2b_s}} \cdot \frac{\Gamma(a_s/2)}{\sqrt{\pi}\cdot\Gamma(a_s/2-\frac{1}{2})} \left(1 + \frac{(\bar{x}-\mu)^2\cdot n}{2b_s}\right)^{-a_s-1/2}
\end{aligned} \tag{4}
$$

We can get this in to a more convenient form by defining

$$
\begin{aligned}
\tilde{t} &= \frac{(\bar{x}-\mu)\sqrt{n}}{\sqrt{2b_s}} \\
\nu_s &= \nu + 2a(\mu)
\end{aligned}
$$

Then

$$
\begin{aligned}
P(\tilde{t}|\mu, s^2, n) &= \frac{1}{B\left(\nu_s - \frac{1}{2}; \frac{1}{2}\right)} \left(1 + \tilde{t}^2\right)^{-\frac{1}{2}(\nu_s+1)} \\
&= \frac{\Gamma(\nu_s)}{\sqrt{\pi} \cdot \Gamma(\nu_s - \frac{1}{2})} \left(1 + \tilde{t}^2\right)^{-\frac{1}{2}(\nu_s+1)}
\end{aligned} \tag{5}
$$

This is the non-standardized $t-$distribution, which can be used also for $n = 1$, in which case b_s and ν_s simply revert to their prior values $b(\mu)$ and $2a(\mu)$. Note that in the limit $\nu \gg 2\max\left\{a, \frac{b}{s^2}\right\}$, $\tilde{t}^2 \to \frac{1}{\nu}t^2$, where $t = \frac{(\bar{x}-\mu)\sqrt{n}}{s}$. Consequently, this non-standardised t-distribution (due to the regularisation from the prior) becomes equal to the standard t-distribution. The first two moments of this distribution is

$$
\begin{aligned}
\langle \tilde{t} \rangle &= 0 \\
\langle \tilde{t}^2 \rangle &= \frac{1}{\nu_s - 2}
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\langle \bar{x} \rangle &= \mu \\
Var(\bar{x}) &= \frac{1}{n} \cdot \frac{2b + s^2\nu}{\nu_s - 2} = \frac{2b + s^2(n-1)}{n \cdot (n - 3 + 2a)}
\end{aligned}
$$

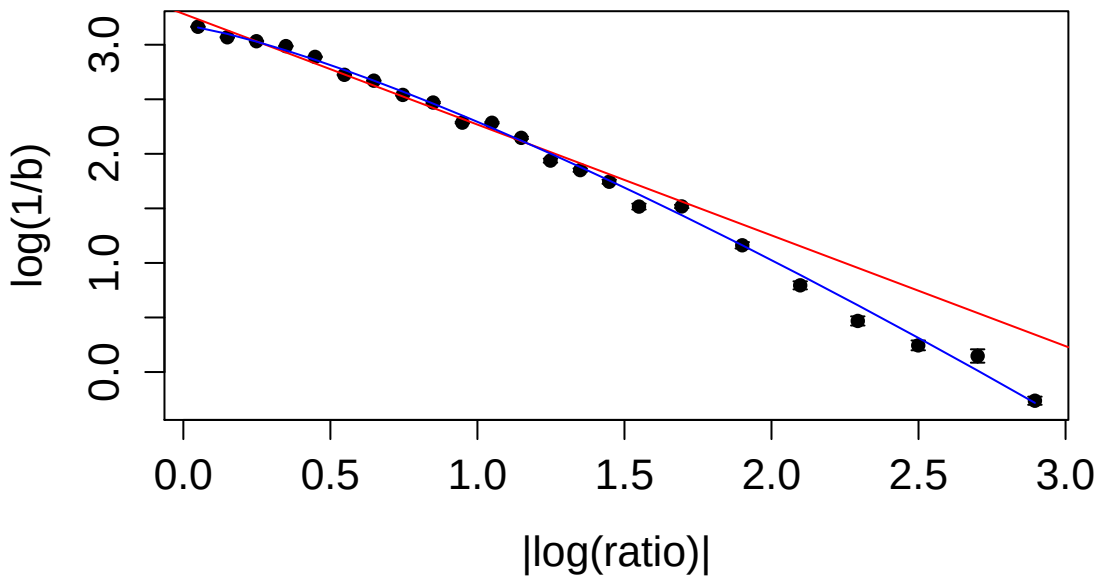
We can generalize this result to the case, where we are not certain which observation x_i (MS[1] peak) is associated with a particular peptide. Let $D = \{x_1, \cdots, x_n\}$ be a set of SILAC ratios all of which have some final probabilities $\omega_i > 0$ of being associated with a particular peptide sequence . Then we shall interpret n as the *effective* number of observations and use

$$
\begin{aligned}
n &= \sum_i \omega_i \\
\bar{x} &= \frac{1}{n} \sum_i \omega_i x_i \\
\nu &= n - 1 \\
\nu \cdot s^2 &= \sum_i \omega_i (x_i - \bar{x})^2
\end{aligned}
$$

which in turn defines the new effective posterior Γ-parameters, a_s and b_s according to Eq. (3).

Empirically, MS data (log-ratio) has precision distribution with a fixed shape-parameter $a \approx 1$, irrespective of the measurement value (log-ratio), thus corresponding to a pure exponential distribution in all cases. The scale of the precision, $1/b$, is on the other hand strongly correlated with the log-ratio measurement.

$$
\begin{aligned}
\frac{1}{b} &\simeq A \cdot \exp(-B \cdot x^{\nu}), \quad x = |\log(\text{ratio})|, \quad B \approx 0.88, \quad \nu \approx 1.28 \quad \text{(blue)} \\
\frac{1}{b} &\simeq A \exp(-B \cdot x), \quad B \simeq 1 \quad \text{(red)}
\end{aligned}
$$



The red-curve is the expected behavior signifying Variance $\propto$ Mean, since $\frac{1}{b} \sim$ precision. The fact that the characteristic variance increases at higher mean-values indicates a type of 'saturation' effect towards high and/or low values of intensities. For any specific protein (or phospho-site), the fact that $a \simeq 1$ implies that the second moment diverges in the absence of specific variance-information. If a variance of a specific protein measurement is V, based on N observation, the posterior precision distribution for λ for this protein is given by

$$\begin{aligned} \lambda &\sim \Gamma(\cdot|a_N, b_N) \\ a_N &= a + \frac{N-1}{2} \\ b_N &= b + \frac{N-1}{2} V \end{aligned}$$

Thus, any ratio-observation with $N > 1$ alleviates the divergence of the second moment.

Monte Carlo Simulation of concentration and occupancy ratios

In the previous section we have described an accurate likelihood function for log-ratios, x_i of individual peptides in terms of the number of observations, n_i and the Γ-parameters b_i and a_i. We need to discuss how to aggregate this information into occupancy ratios and concentration ratios for phosphorylatable sites and proteins, respectively. Let s denote a given phosphorylation site, $\mathcal{S}_i$ the set of phospho-sites for sequence i, S the total number of modifyable sites and $o_s = \frac{p_s}{c}$ the occupansion ratio on site s. Assuming o_s's to be mutual independent, the expected log-concentration ratio, μ_i, is given by

$$\begin{aligned} \exp(\mu_i) &= \exp(c) \prod_{s \in \mathcal{S}_i} \left(\frac{o'_s}{o_s}\right)^{t_{is}} \left(\frac{1-o'_s}{1-o_s}\right)^{1-t_{is}}, \\ \mu_i(c, \bar{o}, \bar{o}') &= c + \sum_s I_{is} \left[\log\left(\frac{1-o'_s}{1-o_s}\right) + t_{is}\left(\log\left(\frac{o'_s}{1-o'_s}\right) - \log\left(\frac{o_s}{1-o_s}\right)\right)\right] \end{aligned}$$

where c is the protein log-concentration ratio between the primed- and unprimed cell-line, $\bar{o}$ and $\bar{o}'$ are the collection of occupancy ratios for the unprimed and primed celle-line respectively, and t_{is} are indicator variables signifying whether

s-site is modulated in peptide i (1) or not (0). Similarly, $I_{is} = 1$ if phospho-site s is covered by peptide i and zero otherwise. Let $z = (\bar{x}_1, \cdots, \bar{x}_I)$ be the measured log-concentration ratios for the I peptides belonging to the given protein, where each $\bar{x}_i$ represent the mean value taken over n_i repeats and with effective posterior Γ-parameters of a_i and b_i and let $\bar{\mu} = (\mu_1, \cdots, \mu_I)$ be the collection of expected values. Then $P(\bar{z}|\bar{\mu}) = \prod_i P(z_i|\mu_i)$ where each factor is a non-standardized t-distribution. Define for convenience

$$\begin{aligned} \tilde{\nu_i} &= \frac{1}{2}(\nu_i + 1) \\ \gamma_i &= \frac{n_i}{2b_i} \\ K_i &= 1 + \gamma_i(\bar{x}_i - \mu_i)^2. \end{aligned}$$

Consequently the likelihood function is given as ,

$$\begin{aligned} L(c, \bar{o}, \bar{o}') &= \log P(\bar{z}|\bar{\mu}(c, \bar{o}, \bar{o}')) \\ &= \sum_{i=1}^{I} \left(-\tilde{\nu_i} \log [K_i] + \frac{1}{2} \log(\gamma_i) \right) \\ L(c, \bar{o}, \bar{o}') &= \sum_{i=1}^{I} \left(\tilde{\nu_i} \log \left[1 + \gamma_i \left(\bar{x}_i - c - \sum_s I_{is} \left[\log \left(\frac{1 - o'_s}{1 - o_s} \right) + t_{is} \left(\log \left(\frac{o'_s}{1 - o'_s} \right) - \log \left(\frac{o_s}{1 - o_s} \right) \right) \right] \right)^2 \right] + \frac{1}{2} \log(\gamma_i) \right) \end{aligned}$$

We optimized the function for each protein separately with bayesian Monte Carlo simulations, 100000 iterations were performed per parameter. We used a Jeffrey prior on o_s with$\alpha_1 = \alpha_2 = \frac{1}{2}$ and a an exponential prior with $\lambda = 2$ on c.

$$L(c, \bar{o}, \bar{o}') + P(c, \bar{o}, \bar{o}') = L(c, \bar{o}, \bar{o}') + Be(o_s|\alpha_1, \alpha_2) + Be(o_s|\alpha_1, \alpha_2) - 2\,|c|$$

Moves

2% of the proposed moves are drawn from the prior distribution of the parameters as described above. Standard moves were as

$$c_{t+1} = c_t + \mathcal{N}(\lambda = 0, \sigma = 0.05)$$

for the concentration and

$$o_{s,t+1} = o_{s,t} + \mathcal{N}(\lambda = 0, \sigma = k)$$

for occupancies, with a standard deviation corrected to propose smaller moves when o_s approaches 0 or 1.

$$k = \frac{1}{|\frac{1}{100} Be(o_s|\alpha_1, \alpha_2) \left(\frac{1}{2(1-o)} - \frac{1}{2o} \right)| + \frac{1}{0.05}}$$

In addition, all moves that would result in $o_s < 10^{-5}$ or $o_s > 1 - 10^{-5}$were automatically rejected.