

A PROBABILISTIC FRAMEWORK FOR PEPTIDE AND PROTEIN QUANTITATION

ASMS 2006

Keith Richardson¹; Richard Denny¹; John Skilling²; Iain Campuzano¹; Thérèse McKenna¹; Phillip Young¹
¹Waters Corporation, Manchester, United Kingdom; ²Maximum Entropy Data Consultants Ltd, Kenmare, Ireland

INTRODUCTION

The purpose of quantitation is to measure changes in concentration of proteins or peptides across several biological or artificial conditions. Examples of commonly used conditions are time point, patient and drug. Experimentally, each condition is usually represented by several replicate acquisitions. Each acquisition is subject to systematic errors such as injection volume errors, and non-systematic effects such as counting statistics. Due to the complexity of the samples and low concentration of the components the data is sometimes incomplete, can include interferences and the assignment of data to peptides or clusters is often uncertain.

Our (Bayesian) view is that the only consistent way of expressing and combining all sources of uncertainty is to use the standard rules of probability. In order to use this approach we must specify:

- Prior probabilities—probability distributions for the model parameters, such as the peptide and protein concentrations we seek, along with any other parameters required to model the experimental situation.
- A likelihood function—a probability distribution for the data given the model parameters.

We may then invoke Bayes' theorem to find the posterior probability distribution for the model parameters,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

We demonstrate the robustness of the approach by looking at proteins with missing data, weak identifications and outlying measurements.

METHODS

Peptide Quantitation

As an example, consider the situation where data (ion counts) for a peptide are collected for two conditions with three replicate acquisitions for each condition. The assignments of the data to the peptide are not certain but each have been assigned a reliability (probability of being correct) of 0.8. The data are summarised as follows:

Replicate	Condition $i=A$		Condition $i=B$	
	Counts D_{ik}	Reliability P_{ik}	Counts D_{Bk}	Reliability P_{Bk}
$k=1$	80	0.8	160	0.8
$k=2$	80	0.8	160	0.8
$k=3$	40	0.8	160	0.8

In order to infer the peptide concentration ratio, a prior probability distribution for the peptide concentration y is specified. We choose to use an exponential distribution,

$$\Pr(y) = \frac{e^{-y/\lambda}}{\lambda}$$

where Λ is a hyper-parameter to be set in advance. Also, we must consider the probabilities of the different configurations of data which may be relevant to the peptide. These are calculated from the supplied reliabilities. In this case there are 24 distinct configurations whose multiplicities sum to $2^6=64$. The likelihood is,

$$\Pr(\text{data} | y_A, y_B, S_{A,1}, S_{A,2}, S_{A,3}, S_{B,1}, S_{B,2}, S_{B,3}) = \left[\prod_{S_{ik} \in \text{ON}} \frac{e^{-y_i} y_i^{D_{ik}}}{D_{ik}!} \right] \left[\prod_{S_{ik} \in \text{OFF}} \frac{e^{-D_{ik}/\Lambda}}{\Lambda} \right]$$

where the $\{S_{ik}\}$ is the set of binary ON/OFF values which defines the configuration. This uses a Poisson distribution as the data are ion counts.

Any data that have not been observed cannot appear in the likelihood – no interpolation or rejection of incomplete replicate sets is required. Moreover, this formulation can extract estimates of ratios and their uncertainties where only a single dataset of each condition is acquired and be extended to describe any number of conditions and replicates.

The joint probability of everything for a single configuration is,

$$\Pr(\text{data}_i, y_A, y_B, S_{A,1}, S_{A,2}, S_{A,3}, S_{B,1}, S_{B,2}, S_{B,3}) = \frac{e^{-(y_A+y_B)/\Lambda}}{\Lambda^2} \left[\prod_{S_{ik} \in \text{ON}} \frac{p_{ik} e^{-y_i} y_i^{D_{ik}}}{D_{ik}!} \right] \left[\prod_{S_{ik} \in \text{OFF}} \frac{(1-p_{ik}) e^{-D_{ik}/\Lambda}}{\Lambda} \right]$$

In this example, it is straightforward to marginalise over one of the y_i to give a probability distribution for the ratio, y_B/y_A . This is plotted for the distinct configurations in figure 1 ($\Lambda=113.3$).

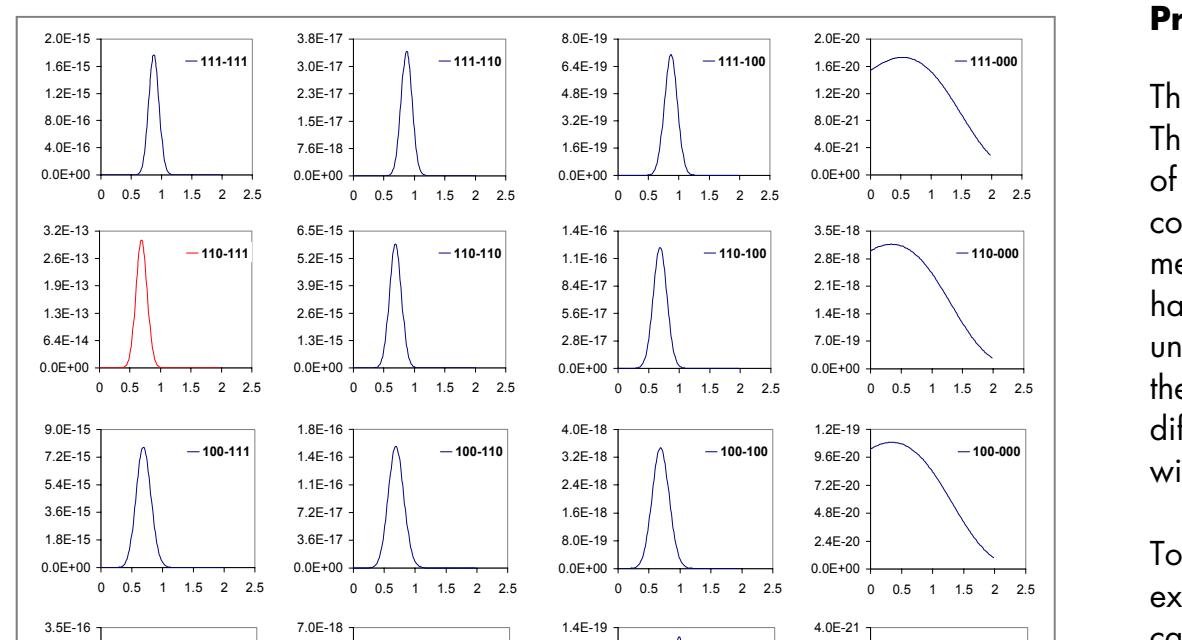


Figure 1. Probability versus log ratio for the 24 distinct configurations of switch ON/OFF states. Note that the scaling of the probability axes varies from plot to plot. The dominant configuration is 111-111 (log ratio $\sim \log 2 = 0.693$) shown in red. Its nearest subordinate configuration with a distinct maximum is 001-111 (log ratio $\sim \log 4 = 1.386$), shown in green.

The sum over all the configurations is shown in figure 2. This is the total joint probability of data and (log) ratio. It should be noted that no attempt was made to reject “outliers” – unfavourable configurations are automatically down-weighted in terms of probability. The dominant configuration is 110-111 (i.e. $S_{A,3}$ is OFF, all other switches are ON) as shown by the peak centred at about $\log 2 = 0.693$. The subsidiary peak at around $\log 4 = 1.386$ is mainly due to the 001-111 configuration.

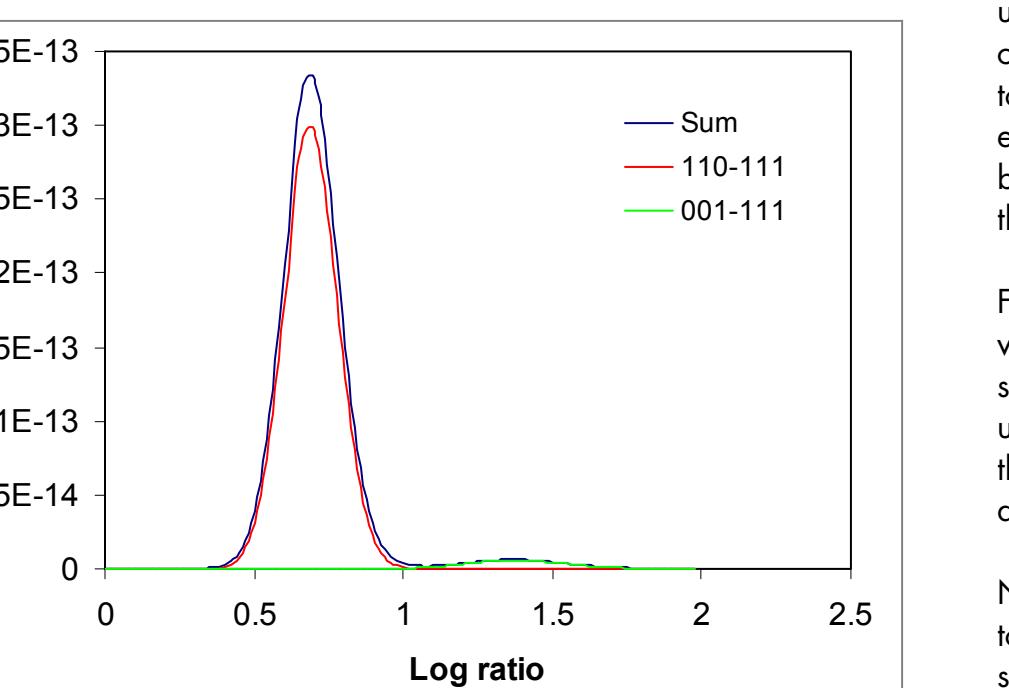


Figure 2. The multiplicity-weighted sum of the configurations shown in figure 1. The contribution from the dominant 110-111 configuration is shown in red and that from the 001-111 configuration is shown in green.

Protein Quantitation

The treatment of proteins is very similar to that described for peptides. The data collected for a protein will generally consist of an incomplete set of measurements of ion counts for several peptides across the available conditions and replicates. We expect a peptide to have similar measured ion counts across replicates within a condition. On the other hand, different peptides may be produced with different efficiencies under enzymatic digestion, and they will ionise differently depending on their amino acid composition. We therefore expect to see significantly different ion counts for peptides belonging to the same protein even within a condition.

To capture this behaviour, our model of the experimental situation is extended to include an extra degree of freedom z_j for each peptide j that captures its relative digestion and ionisation characteristics. The predicted ion count for peptide j in condition i becomes $D_{ijk} = y_i z_j$. Rather than attempting to estimate the z_j , we simply assign each one an exponential prior with a mean of unity, allowing the y_i to set the scale of the data through their dependence on Λ .

Monte Carlo Methods

The direct summation over the different configurations of switch states used in the peptide quantitation example quickly becomes unwieldy as the number of data to be considered increases. We therefore make use of Monte Carlo methods to approximate the direct summation. We have investigated the use of two such techniques: Gibbs sampling¹ and nested sampling² for the problem of protein quantitation. Our implementation of nested sampling employed a Gaussian likelihood function as opposed to the Poisson form used in the above example and the Gibbs sampling implementation. The Gaussian form allows us easily to enforce the likelihood constraints of nested sampling and gives us the flexibility to incorporate uncertainties in addition to those prescribed by counting (Poisson) statistics such as detector dead-time effects.

Normalisation and Noise

Small but noticeable systematic errors can occur during sample handling, enzymatic digestion of samples and introduction of the samples into the LC/MS equipment. This results in ion counts that are systematically high or low in each replicate. It is beneficial to detect these variations and correct for them as part of the quantitation process. This correction may use internal standards that are spiked into each sample at the same concentration or endogenous proteins whose concentrations are believed to be constant. Alternatively, normalisation can be performed using the entire set of detected peptides so long as the biological perturbation being studied is not strong enough to affect more than a small fraction of the proteins identified.

For similar reasons, random errors can be introduced which result in variations in ion counts that exceed those expected from counting statistics alone. In order for us to incorporate these extra sources of uncertainty in our probabilistic analysis we determine a noise level during the normalisation process. This noise level is used to soften the observed data before quantitation.

Normalisation can be treated as another quantitation problem. We wish to investigate probabilistically the variation in response of the internal standards across all replicates. Each measurement is assigned reliability less than unity so that ambiguity can be dealt with automatically in the manner described above. This also means that when normalisation is performed using the entire set of peptides, those belonging to proteins that do change will tend to be switched off. During the exploration, we accumulate a relative strength for the standards in each replicate. These strengths are used to correct the remaining data before the quantitation step.

RESULTS

A solution containing 4 µg of an *Escherichia coli* tryptic digest was spiked with MassPREPTM protein digest standards to give two final solutions, Mix 1 and 2. Mix 1 contained yeast alcohol dehydrogenase I (ADH), rabbit glycogen phosphorylase, muscle form (phosB), yeast enolase I (Enolase) and bovine serum albumin precursor (BSA), all at a concentration of 250 fmol/µL. Mix 2 contained the same proteins at the following concentrations: 250 fmol/µL, 125 fmol/µL, 500 fmol/µL, 2 pmol/µL respectively.

Three replicates of each sample were acquired on a Waters[®] Q-ToFTM Premier, operating in V-opticsTM, positive ion mode. Data were acquired utilising alternating low and elevated energy scans (MS and MS²). Samples were introduced using a Waters nanoACQUITY[™] UPLC System. The raw data were processed and searched using the Waters Protein Expression System Informatics software.

The probabilistic databank search was performed for each of the six experiments. This produced a list of protein identifications with associated peptides. Each peptide was accompanied by an identification probability $0 \leq P_{ijk} \leq 1$. Table 1 shows the peptides identified from the protein phosB in each experiment.

Quantitation was performed twice, first using Gibbs sampling, as implemented in the Waters Protein Expression System Informatics suite³, then using nested sampling. The results for the Mix 1 digest standards are given in tables 2 and 3.

No results are given for ADH as it has been used as an internal standard. The Gibbs sampling results for all proteins, including the digest standards, are summarised in figure 3. The error bars define the full 95% confidence interval.

Peptide	Mix 1		Mix 2	
	Count	95% CI	Count	95% CI
APNDFLNK	33809	34437	18717	19275
ARPEFTLPVHFYGR	23775	24039	14825	19604
DFNVGGYIQAVLDR	18358	17772	3935	15356
FAAYLER	15168	15075	13699	16898
GYNQAEYDR	5524	6017	6339	2590
HLQIYEINQR	12903	16288	10131	12420
IGEEYISLDQLR	27784	29106	11923	13048
IVYLSLEFYMGR	35592	34281	31340	28578
	82874	79447	40029	
LITAQDVVNHDGVDR	76154	75232	71568	34587
LLSVVDDEAFIR			15320	33559
MSLVVEGAK			24120	39190
NLAENISR			67683	5671
QIEQLSSGFSPK	78305	76863	74420	38355
TNFDQFPDK	35891	32993	32146	16555
TNGITPR	50294	52056	49694	20966
VAAFPGDVDR	50633	47561	48031	25241
VFADYEYVK			26929	26266
VHNPNSLFDVQVK			38695	37124
VIFLENYR	58491	55896	57764	30655
VLVDLER	19976	35655	33936	14701
VLVPNDNFEGK	91779	91409	87282	11289
YEFGIFNOK	20605	19092	21219	11830

Table 1. Ion count measurements for peptides assigned to phosB. The counts reported have been normalised using ADH as an internal standard. Green shading indicates an identification probability of 0.95 or above, yellow denotes 0.5–0.95 and red denotes a probability of less than 0.5. A blank space corresponds to a missing identification.

Protein	Ratio R	Log Ratio	95% CI	Pr(R>1)
phosB	0.50	-0.70	±0.02	0.0
Enolase	2.01	0.70	±0.05	1.0
BSA	7.85	2.06	±0.03	1.0

Table 2. Quantitation results using Gibbs sampling. The nominal ratios are 0.5, 2.0 and 8.0. The confidence interval applies to the log ratio.

Protein	Ratio R	Log Ratio	95% CI	Pr(R>1)
phosB	0.50	-0.70	±0.02	0.0
Enolase	2.01	0.70	±0.06	1.0
BSA	7.85	2.06	±0.02	1.0

Table 3. Quantitation results using nested sampling. The nominal ratios are 0.5, 2.0 and 8.0. The confidence interval applies to the log ratio.