Normalization Result

Jialei Wang

Cornell University
Operations Research & Information Engineering
jw865@cornell.edu

March 16, 2015

What we know / assume

- Some notations: suppose a peptide i was tested in jth round for treatment k, the observed value is y_{ijk}, and the underlying true value is θ_{ik}. The physical meaning of "true value" can be the illumination caused by labeling activity. In our experiment setting, we performed 5 rounds of experiments, so $j = 1, 2, 3, 4, 5$, and we have 4 kinds of treatments: sfp ($k = 1$), sfp + AcpH ($k = 2$), AcpS ($k = 3$), AcpS + AcpH ($k = 4$).
- We assume the observed value y_{ijk} is the result of scaling and shifting of true value θ_{ik} with some noise, and the formulation is

  $$y_{ijk} = \sigma_{jk}(\theta_{ik} + \epsilon_{ijk}) + \mu_{jk}$$

  where σ_{jk} is scaling factor, and μ_{jk} is shifting factor corresponding to the photo of spot array for treatment k in jth round experiment.
- We know $\theta_{ik_1} - \theta_{ik_2} \geq 0$ if k_1 corresponds to sfp and k_2 corresponds to sfp+AcpH, or k_1 corresponds to AcpS and k_2 corresponds to AcpS+AcpH.
- We assume noise is independently identically distributed $\epsilon_{ijk} \sim \mathcal{N}(0, a^2)$

Problem Formulation

Now we can write down log-likelihood of the data:

$$-\,logL = \frac{1}{2}\Sigma_{ijk}(\frac{y_{ijk}-\mu_{jk}}{\sigma_{jk}} - \theta_{ik})^2/a^2 + \frac{N}{2}\log(2\pi a), \\ \text{s.t. } \forall i, \theta_{i1}-\theta_{i2} \geq 0, \theta_{i3}-\theta_{i4} \geq 0. \tag{1}$$

To obtain Maximum Likelihood estimate of the unknown parameters, we minimize (1) (minimization because of the "-" sign). Observe the equation, we can separate the minimization into two parts: minimize with respect to $\mu_{jk}, \sigma_{jk}, \theta_{ik}$ for denominator of the first term, and then minimize w.r.t a for the whole equation. We are not interested in a, so we focus on the first minimization. The problem becomes

$$\min_{\mu_{jk},\sigma_{jk},\theta_{ik}} \quad \Sigma_{ijk}(\frac{y_{ijk}-\mu_{jk}}{\sigma_{jk}} - \theta_{ik})^2 \tag{2}$$

$$\text{s.t.} \quad \theta_{ik_1} - \theta_{ik_2} \geq 0 \tag{3}$$

$$\theta_{ik_3} - \theta_{ik_4} \geq 0 \tag{4}$$

$$\theta_{ik_1} - \theta_{ik_2} = 0 \; \forall i \in I_1 \tag{5}$$

$$\theta_{ik_3} - \theta_{ik_4} = 0 \; \forall i \in I_2 \tag{6}$$

$$\Sigma_i \theta_i = \text{constant} \tag{7}$$

Problem Formulation

where $I_1 = \{$peptides that we know is NOT unlabeled after sfp labeling$\}$, and $I_2 = \{$peptides that we know is NOT unlabeled after AcpS labeling$\}$. Let me explain more about the constraints:

1. (3)(4) are the constraints we discussed before.
2. (5)(6) are the “anchors” that help align θ between labeling and unlabeling.
3. (7) avoids θs shrinking to zero in optimization.

This problem is in the form of quadratic program, and can be solved analytically. After solving (2), we get “true value” θ for each peptide in our dataset, then **how do we show if these numbers are meaningful?**

Rank various types of hits

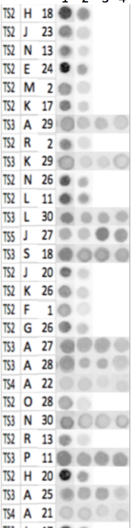
- The first usage of normalized values is to rank peptides according to their activity, and select a few of the top hits for kinetic analysis. I use the normalized value θs to provide rankings for non-specific sfp-type hits, non-specific AcpS-type hits, specific sfp-type hits and specific AcpS-type hits. Below are definitions of the hits:
    - non-specific sfp-type hits are the peptides that are sfp active and AcpH active. The quality measure of how good a peptide i being this type of hit is: $a_1\theta_{i1} + b_1(\theta_{i1} - \theta_{i2})$, a_1, b_1 are weights, the reason of using this formulation and coefficients we choose will be discussed later.
    - non-specific AcpS-type hits are AcpS active and AcpH active. Quality measure is $a_2\theta_{i3} + b_2(\theta_{i3} - \theta_{i4})$.
    - specific sfp-type hits are sfp active, AcpH active and AcpS inactive. Quality measure is $a_3\theta_{i1} + b_3(\theta_{i1} - \theta_{i2}) - c_3\theta_{i3}$.
    - specific AcpS-type hits are AcpS active, AcpH active and sfp inactive. Quality measure is $a_4\theta_{i3} + b_4(\theta_{i3} - \theta_{i4}) - c_4\theta_{i1}$.

Intuition behind the specified quality measure

The intuition behind this formulation is that, for a peptide to be sfp-specific hit for example, it needs to be sfp active, then unlabeled by AcpH, and AcpS inactive; these three individual effects can be quantified using θs we obtained in normalization, which are $\theta_{i1}, (\theta_{i1} - \theta_{i2})$ and $-\theta_{i3}$, and a natural way to combine these effects is to sum them up. We add some flexibility by adding weight to each effect. One way to assign the weights is to make three effects in the same scale. We compute standard deviation of each quantity across all peptides, and assign the weights to be the inverse of standard deviation.

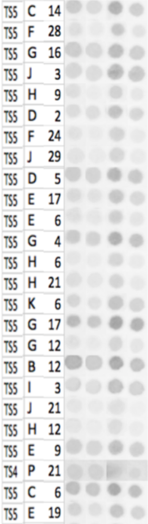
Here are screen shots of part of the ranking of peptides, to be explained in the next slide

Non-specific
sfp-type hit



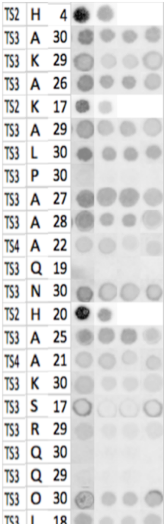
			1	2	3	4
TS2	H	18				
TS2	J	23				
TS2	N	13				
TS2	E	24				
TS2	M	2				
TS2	K	17				
TS3	A	29				
TS2	R	2				
TS3	K	29				
TS2	N	26				
TS2	L	11				
TS3	L	30				
TS5	J	27				
TS3	S	18				
TS2	J	20				
TS2	K	26				
TS2	F	1				
TS2	G	26				
TS3	A	27				
TS3	A	28				
TS4	A	22				
TS2	O	28				
TS3	N	30				
TS2	R	13				
TS3	P	11				
TS2	H	20				
TS3	A	25				
TS4	A	21				

Non-specific
AcpS-type hit



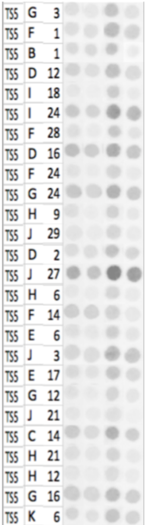
			1	2	3	4
TS5	C	14				
TS5	F	28				
TS5	G	16				
TS5	J	3				
TS5	H	9				
TS5	D	2				
TS5	F	24				
TS5	J	29				
TS5	D	5				
TS5	E	17				
TS5	E	6				
TS5	G	4				
TS5	H	6				
TS5	H	21				
TS5	K	6				
TS5	G	17				
TS5	G	12				
TS5	B	12				
TS5	I	3				
TS5	J	21				
TS5	H	12				
TS5	E	9				
TS4	P	21				
TS5	C	6				
TS5	E	19				

Specific
Sfp-type hit



			1	2	3	4
TS2	H	4				
TS3	A	30				
TS3	K	29				
TS3	A	26				
TS2	K	17				
TS3	A	29				
TS3	L	30				
TS3	P	30				
TS3	A	27				
TS3	A	28				
TS4	A	22				
TS3	Q	19				
TS3	N	30				
TS2	H	20				
TS3	A	25				
TS4	A	21				
TS3	K	30				
TS3	S	17				
TS3	R	29				
TS3	Q	30				
TS3	Q	29				
TS3	O	30				
TS3	L	18				

Specific
AcpS-type hit



			1	2	3	4
TS5	G	3				
TS5	F	1				
TS5	B	1				
TS5	D	12				
TS5	I	18				
TS5	I	24				
TS5	F	28				
TS5	D	16				
TS5	F	24				
TS5	G	24				
TS5	H	9				
TS5	J	29				
TS5	D	2				
TS5	J	27				
TS5	H	6				
TS5	F	14				
TS5	E	6				
TS5	J	3				
TS5	E	17				
TS5	G	12				
TS5	J	21				
TS5	C	14				
TS5	H	21				
TS5	H	12				
TS5	G	16				
TS5	K	6				

Here is more detail about the screen shots

- In the figure from previous slide, each row corresponds to one peptide tested in the experiments, and the information about where this peptide is tested (in which TS and which spot) is provided. Normally there are four spots in each row, and they are sfp activity, sfp+AcpH activity, AcpS activity and AcpS+AcpH activity from left to right. Because peptides in TS2 did not test for AcpS and AcpS+AcpH, their third and forth spot are blank.
- The visual representation makes it easy to verify hits. For example, for non-specific sfp-type hit, we want the spots in first column dark and second column light; for specific sfp-type hit, we want the spots in first column dark, second column light and third column light.
- There are peptides tested across different rounds of experiments, when I rank those peptides in the images, I listed all the spots where they showed up. For example, in specific sfp-type hit, first and second row corresponds to the same peptide; that's why TS2 H4 is ranked top, otherwise it does not make sense because in TS2, AcpS activity was not tested, and there is no point to talk about specific labeling.

Comments

- Sometimes the hits generated by this method are not true hits from visual inspection. For example, first row in non-specific sfp-type hit, although sfp labeling is significant, the spot after AcpH unlabeling still shows strong labeling activity. It is ranked top is because the measure we used for ranking only considers labeling and difference between labeling and unlabeling, when labeling is very strong, even the dye after unlabeling is not completely washed off, the difference is still big. It suggests that this ranking method is not 100% reliable, and that's why I present the result in the form of images, so that we can easily pick the true hits from visual inspection.
- Full ranking results are shown in two forms
  - Image: as shown partly in previous slide, there are four images, measure 1 corresponds to non-specific sfp-type hit, measure 2 corresponds to non-specific AcpS-type hit, measure 3 corresponds to specific sfp-type hit and measure 4 corresponds to specific AcpS-type hit. Because the full ranking images list all peptides, be sure to zoom in!
  - Table: records peptide sequences. The column "unique_idx" indicates if a peptide is tested in multiple spots (same index means same peptide).