Supplemental Materials

Text S1

To characterize the roles of amino acids in the PTMs and their counterparts under the unified framework in the feature space, we theoretically devised bi-profile coding scheme under Bayes framework, which is simultaneously inclusive of positive (PTMs) and negative (non-PTMs) information and called bi-profile Bayes (Shao et al. 2009).

Consider an unlabeled sample $S=\{s_1,s_2,s_3,\cdots,s_n\}$ which denotes short linear sequence in our case, where each $s_j\ (j=0,1,\cdots,n)$ stands for one amino acid and n represents the length of peptide sequence, i.e. the size of sliding window in this study. S belongs to one of two categories C_1 or C_{-1}, where C_1 and C_{-1} represent PTM sites (positive data) and non-PTM sites (negative data), respectively. According to Bayes' rule, the posterior probability of S for these two categories can be given by

$$P(c_1|S)=\frac{P(S|c_1)P(c_1)}{P(S)} \tag{1}$$

$$P(c_{-1}|S)=\frac{P(S|c_{-1})P(c_{-1})}{P(S)} \tag{2}$$

where $P(c_1)$ and $P(c_{-1})$ denote the prior probability for each category. Assume that $s_j\ (j=1,2,3,\cdots,n)$ are mutually independent, Formula (1) and (2) can be rewritten as

$$P(S|c_1)=\prod_{j=1}^{n}P(s_j|c_1) \tag{3}$$

$$P(S|c_{-1})=\prod_{j=1}^{n}P(s_j|c_{-1}) \tag{4}$$

By the above, Formulas (1) and (2) can be further reformulated as

$$\log(P(c_1|S))=\sum_{j=1}^{n}\log(P(s_j|c_1))-\log(P(S))+C_1 \tag{5}$$

$$\log(P(c_{-1}|S))=\sum_{j=1}^{n}\log(P(s_j|c_{-1}))-\log(P(S))+C_2 \tag{6}$$

where $C_1=\log(P(c_1))$ and $C_2=\log(P(c_{-1}))$. Thus, the decision function can be represented by formula (7)

$$f(S)=\mathrm{sgn}(\log(P(c_1|S))-\log(P(c_{-1}|S))) \tag{7}$$

Assume that prior distribution of category is uniform, namely $P(c_1)=P(c_{-1})$. Formula (7) can be rewritten as

$$f(S)=\mathrm{sgn}(\sum_{j=1}^{n}\log(P(a_j|c_1))-\sum_{j=1}^{n}\log(P(a_j|c_{-1}))) \tag{8}$$

Formula (8) can further be formulated as

$$f(S)=\mathrm{sgn}(\vec{w}\bullet\vec{p}) \tag{9}$$

where $\vec{w}=(w_1,w_1,\cdots,w_n,w_{n+1},\cdots,w_{2n})$ is weigh vector, $\vec{p}=(p_1,p_2,\cdots,p_n,p_{n+1},\cdots,p_{2n})$ is the posterior probability vector. With respect to training sample S , $f(S)=1$ corresponds to class C_1 and $f(S)=-1$ to class C_{-1}. In the present project, $p_1,p_2,\cdots,p_n$ represents the posterior probability of each amino acid at each position in positive short linear sequence datasets (category C_1) (positive feature space)and $p_{n+1},\cdots,p_{2n}$

represents that in negative short linear sequence datasets (category C_{-1}) (negative feature space) , which we call Bi-profile. The posterior probability can be estimated by the different profile of each amino acid at each position in training datasets, such as the occurrence, the degree of enrichment, which we define as position-specific profile.

Under Bayes framework, any short linear sequence (PTM site or non-PTM site in our case) can be represented by a bi-profile vector $\vec{p} = (p_1, p_2, \cdots, p_n, p_{n+1}, \cdots, p_{2n})$, weighting the biological roles of each residue at each position simultaneously in the feature spaces of both PTMs and non-PTMs. Theoretically, bi-profile Bayes accounts for the issue regarding encoding an unlabeled sample in a uniform manner, in the case of which the prediction process satisfies the law of total probability. Typically, one can train one classifier through encoding the positive and negative in respective feature spaces since the labels for training samples are known, and then the prediction decision is made through comparison of the output for unlabeled sample encoded in the positive feature space to that encoded in the negative feature space (saying the output probability for some unlabeled sample belonging to positive samples is more than that belonging to negative samples, the final decision will be made that this unlabeled sample is predicted as positive sample). As described above, bi-profile Bayes proposed is based on the assumption that all of the variables are mutually independent in the vector $\vec{p} = (p_1, p_2, \cdots, p_n, p_{n+1}, \cdots, p_{2n})$. Furthermore, the training samples might exhibit linear or non-linear distribution. Hence, to capture the possible correlations among variables (patterns) and

distinguish the positive patterns (PTM sites in our case) from the counterparts (non-PTM sites), machine learning approaches have been extensively applied such as Support Vector Machines (SVMs).

Reference:

Shao J, Xu D, Tsai SN, Wang Y, Ngai SM (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PLoS One 4: e4920.