False Discovery Rate and Application to HIV Data with BLOSUM62

Kyurhi Kim [1] Junyong Park [2]

[1]Graduate Student, Department of Statistics, Seoul National University [2]Professor, Department of Statistics, Seoul National University

Abstract

This study aimed to enhance the identification of significant sites in sequence data by incorporating biological knowledge. We proposed two models: one based on the empirical Bayes model under independence of amino acids and the other uses pairwise associations of amino acids based on Markov random field with on the BLOSUM62 substitution matrix. These methods combined observed data with prior information from BLOSUM62, avoiding subjectivity of hyperparameter choices. Unlike Fisher's test with the BH procedure, which found no significant sites, both proposed approaches identified and improved the detection of significant sites.

Motivation

- With sparse count data, Fisher's exact test has limitations due to its discreteness and loss of information by conditioning on marginal totals.
- Also, only the configuration of counts are considered, not the types of amino acids.
- Some pairs of amino acids tend to occur more often than other pairs, so it is more reasonable to take into account such information.

Preliminaries

Local FDR

- Perform simultaneous hypothesis tests and classify the results as follows:

	Null Decision	Non-Null Decision	total
Actual Null	$N_0 - V$	V	N_0
Actual Non-Null	$N_1 - S$	S	N_1
Total	$N - R$	R	N

- FDR $:= \mathbb{E}(\frac{V}{max(R,1)})$, where $\frac{V}{max(R,1)}$ is the False Discovery Proportion.
- Efron et al.[1] proposed a two-component mixture model $f(z_i) = \pi_0 f_0(z_i) + (1-\pi_0) f_1(z_i)$.
- Local FDR is defined as follows: P(ith gene is null$|z_i$) $= \frac{\pi_0 f_0(z_i)}{f(z_i)}$.

BLOSUM62(BLOcks SUbstitution Matrix 62)

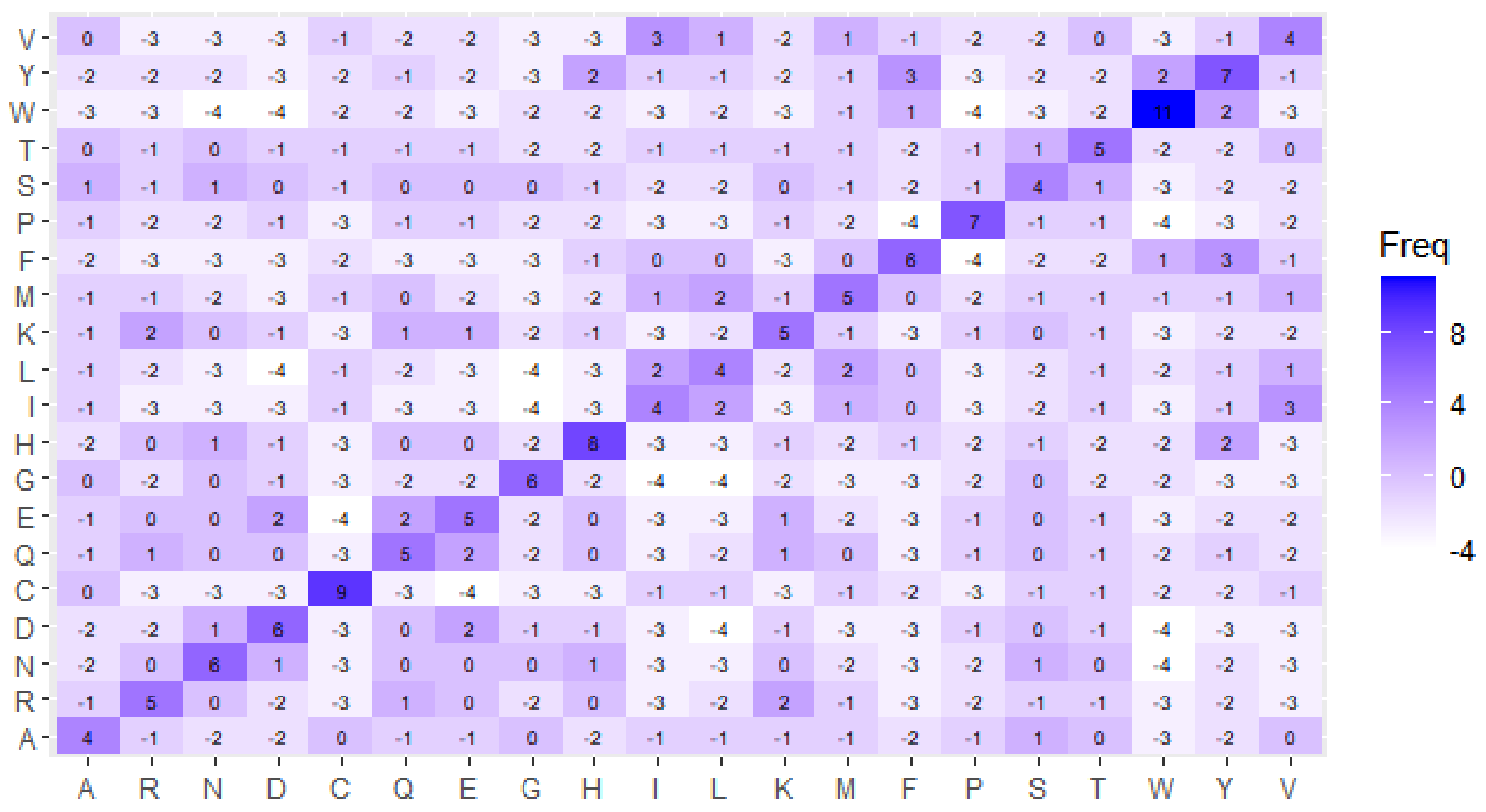


Figure 1. BLOSUM62 Substitution matrix

- BLOSUM62[2] is used to score the similarity between amino acids in protein sequences.
- Each pair of amino acids with a smaller score occurs less often than those with higher scores.

1.Model with only Independent Term

Data Descriptions

- Each site $\mathbf{x}_i \sim$ Mult$(n_1, \mathbf{p}_T)$, $\mathbf{y}_i \sim$ Mult$(n_2, \mathbf{p}_{NT})$ for T and NT group, respectively, where $\mathbf{p}_T$ and $\mathbf{p}_{NT}$ are 20 dimensional probability vectors for amino acids.
- H_{0i}: Transmitted(T) and Non-Transmitted(NT) groups are different in ith site, $i = 1, ..., K$.

Proposed Method

- Let $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$, then the marginal distribution of $\mathbf{z}_i$; $f(\mathbf{z}_i) = \pi_0 f_0(\mathbf{z}_i) + (1-\pi_0) f_1(\mathbf{z}_i)$.
- The prior distribution of $\mathbf{p}_T$ and $\mathbf{p}_{NT}$ are defined as:

$$\mathbf{p} \equiv \mathbf{p}_T \equiv \mathbf{p}_{NT} \sim Dirichlet(\alpha_1, ..., \alpha_{20}) \text{ under the null,}$$
$$\mathbf{p}_T \sim Dirichlet(\alpha_1, ..., \alpha_{20}), \mathbf{p}_{NT} \sim Dirichlet(\alpha_1, ..., \alpha_{20}) \text{ under the alternative,}$$

- Since $E(p_s) = \frac{\alpha_s}{\sum_{s=1}^{20} \alpha_s}$, use $\alpha_s = \beta q_s$ by matching the moments of p_s with q_s derived from the BLOSUM62.

Empirical Bayes for Parameter Estimation

- Empirical Bayes approach was used to obtain $\widehat{\beta}_0$ for the null, and $\widehat{\beta}_T$, $\widehat{\beta}_N$ for alternative:

$$\hat{\beta}_0 = \text{argmax}_{\beta>0} \prod_{i=1}^{K} P(z_i|\alpha) = \text{argmax}_{\beta>0} \prod_{i=1}^{K} \frac{\Gamma(\beta) \prod_{j=1}^{20} \Gamma(\beta q_j + x_{ij} + y_{ij})}{\prod_{j=1}^{20} \Gamma(\beta q_j) \Gamma(n_1 + n_2 + \beta)},$$

- To estimate π_0, we introduced a latent indicator variable e_i where it takes 1 if group T and NT are different in i-th site and 0 otherwise, and a prior distribution $Unif(0,1)$ for π_0.

Algorithm 1 Metropolis Hastings within Gibbs Sampling for the Null Proportion π_0

```
for $m = 1$ to $M$ do
    Sample $\pi_0^{(m)} \sim f(\pi_0|\mathbf{e}^{(m-1)}, \mathbf{x}, \mathbf{y}, \mathbf{p}) \sim \text{Beta}(n - \sum_i e_i + 1, \sum_i e_i + 1)$
    Sample $\mathbf{e}^{(m)} \sim f(e|\pi_0^{(m)}) \sim \text{Bernoulli}(1 - \text{lfdr}(z_i))$ , where $\text{lfdr}(z_i) = \frac{\pi_0^{(m)} f_0(z_i)}{f(z_i)}$
end for
```

2. Model considering Pairwise Information

- Off-diagonal scores in the BLOSUM62 matrix represent that each pair of amino acids with a smaller score occurs less often than those with higher scores.

Proposed Model Based on Pairwise Probabilities

- For $\mathbf{z}_i$ in the HIV data, the probability distribution function is modeled as

$$P(\mathbf{z}_i|\Theta_s, \Theta_{st}, \delta) = \exp\left(\sum_{s=1}^{20} \theta_s z_{is} + \delta \sum_{t \neq s} \theta_{st} z_{is} z_{it}\right) \cdot C(\Theta),$$

- $C(\Theta)$ is a normalizing constant, δ is a tuning parameter for the magnitude of pairwise amino acids effect, and Θ_s and Θ_{st} are coefficients for independent and pairwise terms, respectively. (Similarly for $\mathbf{x}_i$ and $\mathbf{y}_i$).
- We used pseudo-likelihood $P(\mathbf{z_i}|\Theta) \approx \prod_{s=1}^{20} P(z_{is}|\mathbf{z}_{i,-s}, \Theta)$ to handle normalizing constant.

Posterior Inference

- We generated posterior samples using MCMC, and computed the local FDR given α.
  - Obtain Initial values of $\mathbf{q}_s$, $\mathbf{q}_{st}$ from BLOSUM62.
  - Give prior distribution for Bayesian inference, $\Theta_s \sim \text{Dir}(\alpha_1, ..., \alpha_{20})$, $\Theta_{st} \sim \text{Dir}(\alpha_{1,2}, ..., \alpha_{19,20})$.
  - Then marginally univariate θ_s's and θ_{st}'s follow Beta distribution as:

$$\theta_s \sim Beta(\alpha_s, \sum_{j \neq s} \alpha_j), \; \theta_{st} \sim Beta(\alpha_{st}, \sum_{j \neq \{st\}} \alpha_j)$$

  where $\alpha_s = \beta_1 q_s, \alpha_{st} = \beta_2 q_{st}$, for tuning parameter β_1, β_2.
  - Propose θ_s^* from $N(\theta_s^{(m-1)}, \sigma^2))$ and Update θ_s based on $p(\theta_s|\mathbf{z}, \Theta_{-s}) \propto p(\mathbf{z}_s|\theta_s) Pr(\theta_s|\Theta_{-s}))$.
  - Similarly, Propose θ_{st}^* from N$(\theta_{st}^{(m-1)}, \phi^2))$ and Update θ_{st} based on $p(\theta_{st}|\mathbf{z}, \Theta_{-st}) \propto p(\mathbf{z}_{st}|\theta_{st}) Pr(\theta_{st}|\Theta_{-st}))$

Results

Method 1: Emprical Bayes Model

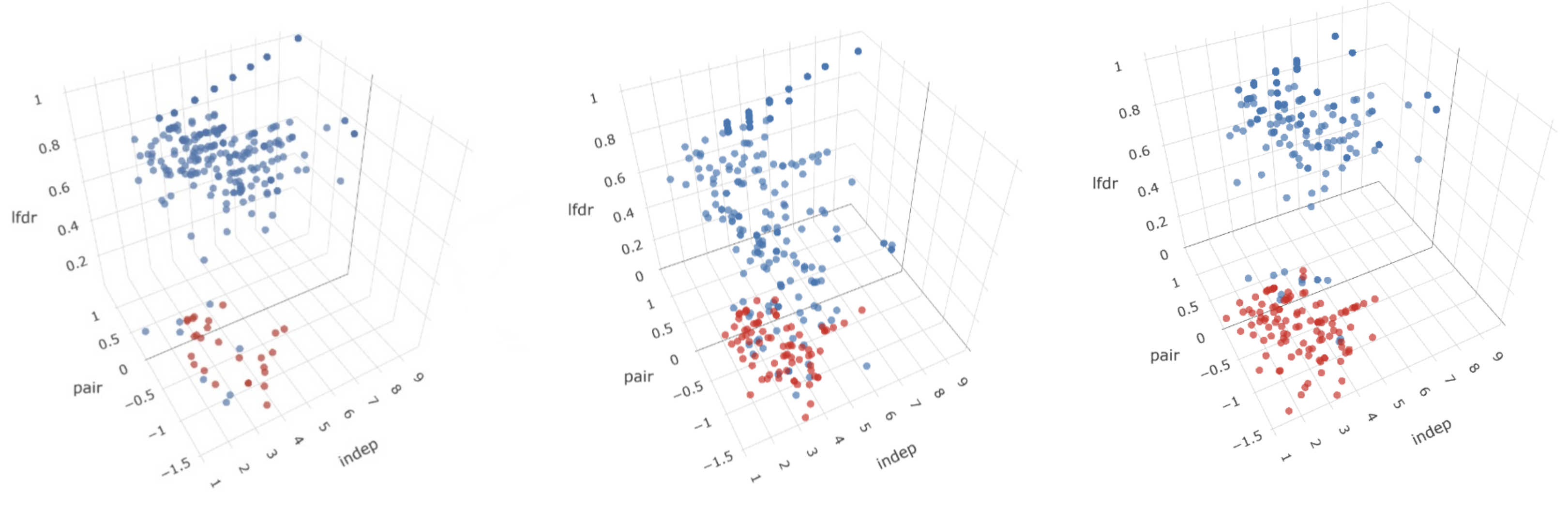
- Averaging M sampling as $lfdr(\mathbf{z}_i) \approx \frac{1}{M} \sum_{m=1}^{M} lfdr(\mathbf{z}_i|\Theta^{(m)})$, $i = 1, ..., K$, where $\Theta^{(m)} = (\Theta_s^{(m)}, \Theta_{st}^{(m)})$ are generated from the posterior distributions at mth sampling.
- Reject H_{0i} if $lfdr(\mathbf{z}_i) \leq \alpha$, where $\alpha = 0.05$.
- We obtained $\hat{\beta}_0 = 0.2596$, $\hat{\beta}_T = 0.2730$ and $\hat{\beta}_N = 0.0648$, and rejected 26 sites among 812.

Method 2: Considering Pairwise Information

lfdr (Model1)	T	NT	BLOSUM62	lfdr (Model2, $\delta = 0.5$)	rank
0.0499	I5	G3	I=4, G=6, IG=-4	2.76e-35	1
0.0499	I5	G3	I=4, G=6, IG=-4	1.30e-34	2
0.0498	N5	D3	N=6, D=6, ND=1	7.80e-29	4
0.0495	P5	H3	P=7, H=8, PH=-2	1.94e-24	5
0.0496	D5	F3	D=6, F=6, DF=-3	2.61e-18	6
0.0493	I5	M3	I=4, M=5, IM=1	3.49e-18	7
0.0498	A5	W3	A=4, W=11, AW=-3	1.18e-16	9
0.0499	G5	W3	G=6, W=11, GW=-2	1.58e-14	10
0.0496	A5	S3	A=4, S=4, AS=1	4.60e-13	11
0.0496	K5	S3	K=5, S=4, KS=0	2.90e-11	12

Table 1. Results of analysis for 10 site with lfdr less than α, and constructed with T5, N3.

- With $\delta = 0.1$, we rejected 79 sites and $\delta = 0.5$, rejected 112 sites.
- The result implies that both independent and pairwise term operate simultaneously.



(a) Model 1 (b) Model 2 with $\delta = 0.1$ (c) Model 2 with $\delta = 0.5$

Figure 2. 3d Plot (independent score vs pairwise score vs estimated local fdr) for the results of analysis

- Compared to Model 1, more rejection sites (red points) are observed in the plots of Model 2 when the scores in x, y axis are small.
- Model 2 rejects more points toward the bottom-left as pairwise terms are introduced.
- As δ increases, the impact of the pairwise term within the model becomes more significant.
- Model 2 incorporating pairwise terms assigns greater weight to rare events, thus identifying significant sites with lower scores better than the model 1.

References

[1] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[2] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.