Prediction of MHC Class I and II binding peptides incorporating bayesian transfer hierarchies

Ravikiran Janardhana

November 26, 2012

So far...

- Downloaded MHCBN and MHCPEP Class-I and Class-II peptide data and transformed sequences into feature vector via sparse encoding (0s of length 20 per protein (mark 1 for the current protein)).
- Baseline: Implemented Elastic Net to classify binding and non-binding sequences(15-mers of Class MHC-II). Max accuracy I could obtain was 66% when $\lambda = 0.95$ and $\alpha = 0.05$ indicating lasso worked better than ridge penalty.
- Objective function to optimize:

$$F_{joint}(\theta; D) = -\sum_{c \in L} F_{data}(D^c, \theta^c) + \alpha \sum_{c \in C} Div(\theta^c, \theta^{par(c)}) \tag{1}$$

$$F_{joint}(\theta; D) = -\{Log\ Likelihood\} + \alpha\{L1 - Distance\ between\ parameters\} \tag{2}$$

$$F_{joint}(\theta; D) = \sum_{i=1}^{k}\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j^i - \beta_0^i - X_j\beta^i)^2\} + \alpha\{\sum_{i=1}^{k-1}|\beta_0^i - \beta_0^{i+1}| + \sum_{i=1}^{k-1}|\beta^i - \beta^{i+1}|\} \tag{3}$$

For k = 2,

$$F_{joint}(\theta; D) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j^1 - \beta_0^1 - X_j\beta^1)^2\} - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j^2 - \beta_0^2 - X_j\beta^2)^2\} + \alpha\{|\beta_0^1 - \beta_0^2| + |\beta^1 - \beta^2|\} \tag{4}$$

- Is the above formulation correct ? How should I optimize ? Possibly ADMM ? The formulation looks like multiple Fused Lasso problems.
- Should I restrict the 'k' value to 2 to start off with?
- Both MHCBN and MHCPEP have mostly binding peptide sequences, there is very little non binding peptide sequences, possibly biased classification ?
- How do I generate synthetic data?