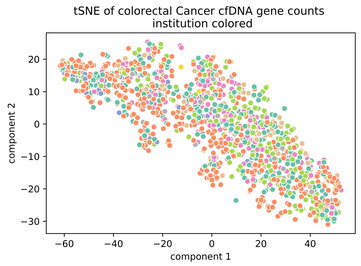
METCC: METric learning for Confounder Control
Making distance matter in high dimensional biological analysis

Kabir Manghnani, Adam M. Drake, Nathan Wan, Imran S. Haque

Freenome, Inc., South San Francisco, CA

BACKGROUND

- High-dimensional data acquired from biological experiments, such as next-generation sequencing of cfDNA in blood plasma, are subject to a number of confounding effects.
- These confounders pose a challenge when developing solutions to pattern recognition problems using biological data because they can obscure the biological signal of interest. Visualizations of data show institution-specific clustering/shifts



- As an example, previous studies found that an estimated 32% of variability in the *1000 Genomes Project* dataset can be explained by sequencing date (Leek et al., 2010).
- Prior work such as Hidden Covariates with Prior (HCP) (Mostafavi et al., 2013), used mixed effects models to adjust for confounder effects. HCP learns a Gaussian model where the observations $\hat{Y}$, follow a distribution parameterized by the sum of the true latent signal Y, known covariates FB, and unknown sources of covariates XW.
- We propose a Metric Learning based model to normalize out confounder signal.
- Metric learning methods are advantageous because the loss function requires only the variable of interest to normalize, as opposed to mixed

OBJECTIVE

We analyze:

1. the extent to which data normalized with HCP and METCC retains information about the unwanted technical effects; and
2. the performance of supervised models trained on normalized data.

METHODS

- Let X be an $n \times p$ matrix of observed biological data with n samples and p measurements. Let y be a biological variable of interest such as phenotype label or disease status.
- We seek to learn a distance function D_w parameterized by the map $g: \mathbb{R}^p \rightarrow \mathbb{R}^k$ where the distance between two samples x_i and x_j is determined by $D_w(x_i, x_j) = \|g(x_i) - g(x_j)\|_2$.
- The objective is to transform the data via g so that the variability measured between samples x_i and x_j is low when $y_i = y_j$ and high when $y_i \neq y_j$. This can be optimized using a contrastive, or Siamese, loss function proposed by Hadsell et al.
- Such approaches have been extended and shown to produce better representations when both positive and negative class sample are used for each "anchor" sample (Hoffer et al., 2014).
- The "Triplet" extension of this approach used in our experiments runs a triplet of samples (x, x^-, x^+) through D_w and defines $d_- = D_w(x, x^-)$ and $d_+ = D_w(x, x^+)$ so as to minimize $\|d_+, d_- - 1\|_2^2$ (Balntas et al., 2016).

METHODS, cont'd

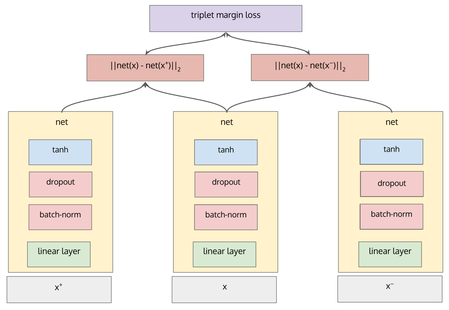
Intuition:

1. minimize distance between representations of examples that have the same label (L = 0)
2. maximize distance between representations of examples with different labels (L = 1)

$$argmin_w (1-L) D_w\left(x_i, x_j\right)^2 + (L)\, max\left(0, m - D_w\left(x_i, x_j\right)^2\right)$$

- Penalizes variance that is not correlated to disease prediction task (e.g. batch effects).

Figure 1. Network Architecture

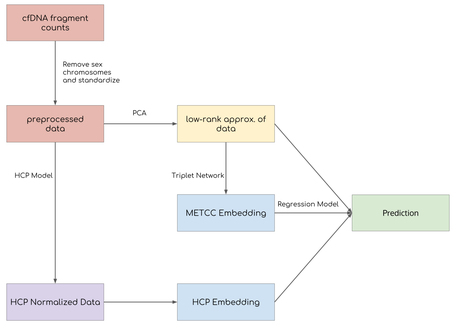


Architecture of triplet network used.

EXPERIMENTAL SETUP

- We use counts of cfDNA reads that align to CHESS gene bodies from 817 samples (Wan et al. 2018). In addition to a disease label of healthy or CRC, each sample has associated an institution where it originated, age when the blood was drawn, and batch which it was processed. We grouped age into the bins: [0-50, 50-55, 55-60, 60-75, 75-80, 80-85, 85+].
- We apply all three methods to generate a set of embeddings for each over identical folds of k-fold cross validated data (k=4). We subsequently trained classifiers to predict each of the 4 sets of labels using both K-Nearest Neighbor (KNN) and Logistic Regression (LR).

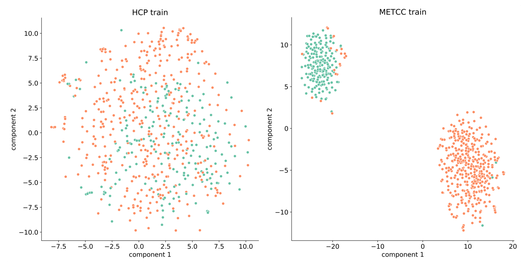
Figure 2. Embedding Generation/Evaluation Pipeline



Construction of the 3 normalizations. Choices were made to ensure equal dimensionality among them. Classifiers were trained with K=21 for KNN and a random search for regularization weight for each fold for LR.

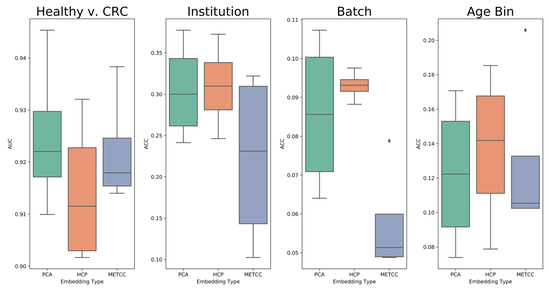
RESULTS

Figure 3. Comparison of tSNE of HCP and METCC embeddings



- We are able to train METCC embeddings where label distance is learned. Visualized with tSNE are one fold of embeddings normalized by HCP and METCC respectively.
- METCC embeddings when trained with LR achieved comparable performance.
- With KNN, disease performance far exceeded other embeddings, but performance on covariate predictions were not significantly different.

Figure 4. Disease status / Covariate Prediction Task Performance



- Performance drop in disease classification can be interpreted as a proxy measure of how much information is lost to the normalization process. Prediction of other labels can be interpreted as a measure of how much confounder information is left in the embeddings after normalization.

Table 1. Mean k-fold Prediction Task Performance (k=4)

Disease Prediction Task				
Normalization	Train AUC (KNN)	Test AUC (KNN)	Train AUC (LR)	Test AUC (LR)
PCA-only	0.86	0.78	0.98	0.93
HCP	0.90	0.82	0.98	0.91
METCC	**0.99**	**0.87**	0.99	0.92
Confounder Prediction Tasks				
Normalization	Train ACC (KNN)	Test ACC (KNN)	Train ACC (LR)	Test ACC (LR)
PCA-only (inst.)	0.53	0.24	0.87	0.31
HCP (inst.)	0.54	0.20	0.89	0.31
METCC (inst.)	0.50	0.23	**0.42**	**0.23**
PCA-only (batch)	0.29	0.06	0.82	0.08
HCP (batch)	0.19	0.04	0.88	0.09
METCC (batch)	0.22	0.07	**0.13**	**0.06**
PCA-only (age-bin)	0.37	0.12	0.71	0.12
HCP (age-bin)	0.27	0.14	0.74	0.13
METCC (age-bin)	0.31	0.12	0.26	0.13

Figure 4 and Table 1 depict classification performance. Area Under the Receiver Operating Characteristics (AUC) was used to measure disease status prediction performance. Accuracy (ACC) was used to measure institution, batch, and age-bin tasks which are not binary classification.

DATASET CONFOUNDING

Figure 5. Known covariate distribution in cfDNA dataset

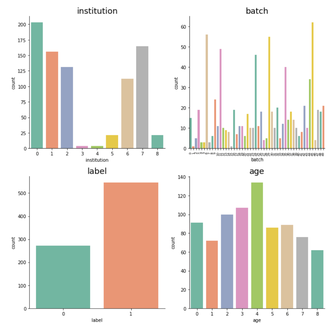
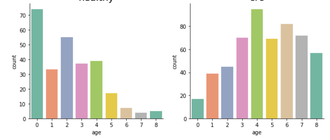


Figure 6. Disease status is confounded by age



- METCC may not lose information about age due to the fact that the disease label is a clear confounder, which was used to train METCC embeddings.

CONCLUSIONS

- METCC allows us to apply black box metric techniques to confounder normalization. This allows us to account for confounding effects without having *a priori* knowledge of these effects, labels to which can be rare or hard to procure.
- Embeddings generated with METCC can outperform embeddings normalized with a mixed effects model in a biological prediction task.
- HCP appears to retain slightly more covariate signal in the data than METCC.
- Future work intends to compare black box metric learning models to non-linear mixed effects models.
- We intend to do further exploration into the question of how best we can compare a supervised approach, like METCC that separates based on label, to embeddings generated by unsupervised techniques.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Signe Fransen, Girish Putcha and David Weinberg for their extensive suggestions, feedback, and editorial support.

REFERENCES

Hoffer et al. 2014; arxiv. 1412.6622

Balntas et al. 2016; The British Machine Vision Conference. 119.1-119.11

Hadsell et al. 2006; IEEE Computer Society Conference. (2): 1735-1742

Leek et al. 2010; *Nat Rev Genet*.11(10):733-9

Mostafavi et al. 2013; *PLoS One* 8 (7):1-10

Wan et al. 2018; *bioRxiv* https://doi.org/10.1101/478065

Presented at the 11th annual RECOMB/ISCB Conference on Regulatory & Systems Genomics, December 8-10, 2018, New York, NY USA