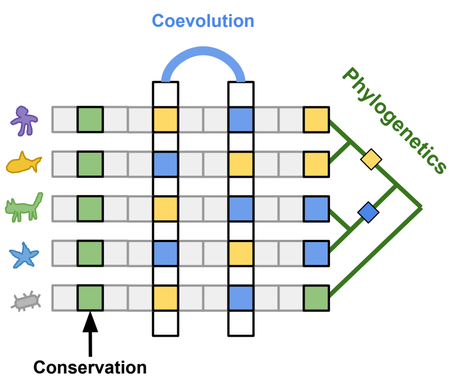
Unified framework for modeling multivariate distributions in biological sequences

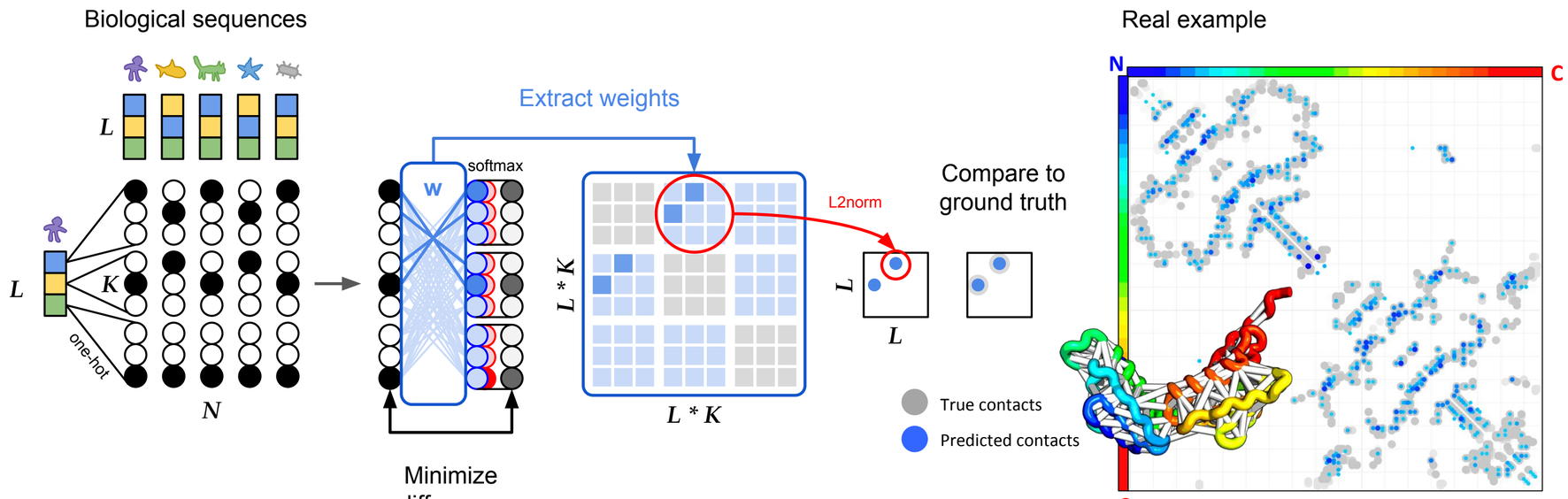
Justas Dauparas[1], Haobo Wang[1], Avi Swartz[2], Peter Koo[3], Mor Nitzan[4], Sergey Ovchinnikov[4]
[1]FAS Division of Science [2]Harvard College [3]Howard Hughes Medical Institute [4]JHDSF Program, Harvard University

Introduction

Revealing evolutionary **conserved**, **co-evolving** sites and **phylogenetics** (relationship between sequences) is a fundamental and challenging task in the protein field.



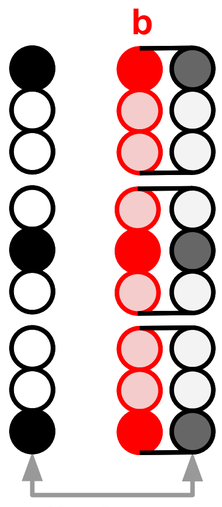
Coevolution in 1D biological sequences, captured by **weights** of a fully-connected (or dense) layer, is correlated to physical contacts in 3D structures!



Methods

Here we propose a unified framework for the current models (Position-Specific Scoring Matrices, Markov Random Fields, Multivariate Gaussian models, and Autoencoders), that allows for interpretable transformations between these methods and naturally incorporates the advantages and insight gained individually in the different communities.

Position-Specific Scoring Matrix (PSSM)



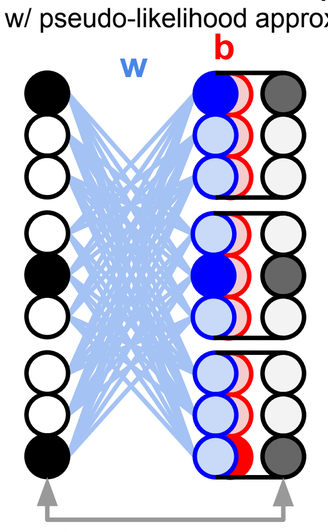
$$\mathcal{L}_{PSSM} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L} \text{cce}[\mathbf{X}, \text{softmax}(\mathbf{b})]$$

Add two-body term

(Dense Layer)

Markov Random Field (MRF)

w/ pseudo-likelihood approx.

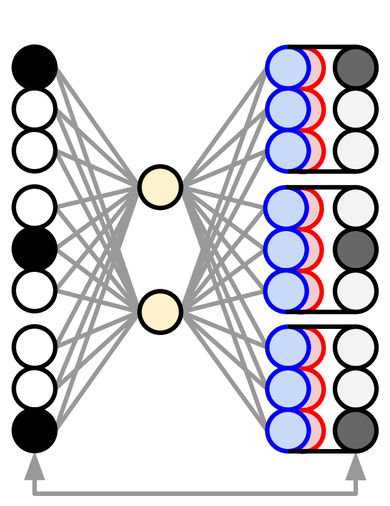


$$\mathcal{L}_{MRF} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L}\{\text{cce}[\mathbf{X}, \text{softmax}(\mathbf{H})]\}_{nl} + \lambda L2(W)$$

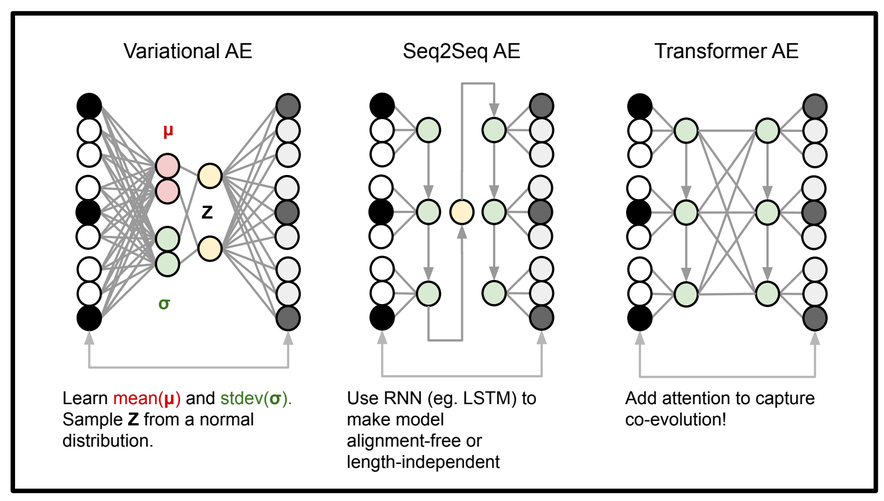
$$H_{nlk} = B_{nlk} + \sum_{r=1}^{L}\sum_{s=1}^{K} X_{nrs} W_{rslk}$$

Add hidden layers

Autoencoders (AE)



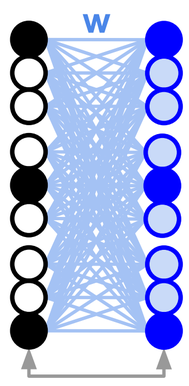
Extending the autoencoder



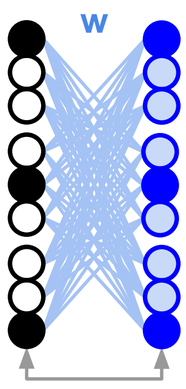
Replace softmax/CCE with MSE

Multivariate Gaussian (MG)

w/ mean-squared-error approx.



Add diagonal



$$\mathcal{L}_{MG} = \frac{1}{N}\sum_{n=1}^{N}\sum_{f=1}^{LK}\{(\mathbf{X} - \mathbf{X}W)^2\}_{nf} - 2*\text{Tr}(W) + \lambda L2(W)$$

$W = (1-\lambda)(\text{cov}(\mathbf{X}) + \lambda I)^{-1} + I$

$$\mathcal{L}_{MG} = \frac{1}{N}\sum_{n=1}^{N}\sum_{f=1}^{LK}\{(\mathbf{X}W)^2\}_{nf} - 2*\text{Tr}(W-I) + \lambda L2(W-I)$$

$W = (\text{cov}(\mathbf{X}) + \lambda I)^{-1}$

$$\mathcal{L}_{MG} = \frac{1}{N}\sum_{n=1}^{N}\sum_{f=1}^{LK}\{(\mathbf{X} - \mathbf{X}W)^2\}_{nf} + \lambda L2(W)$$

At global optima:

$W = \mathbf{c}/\text{diag}(\mathbf{c}) - I$

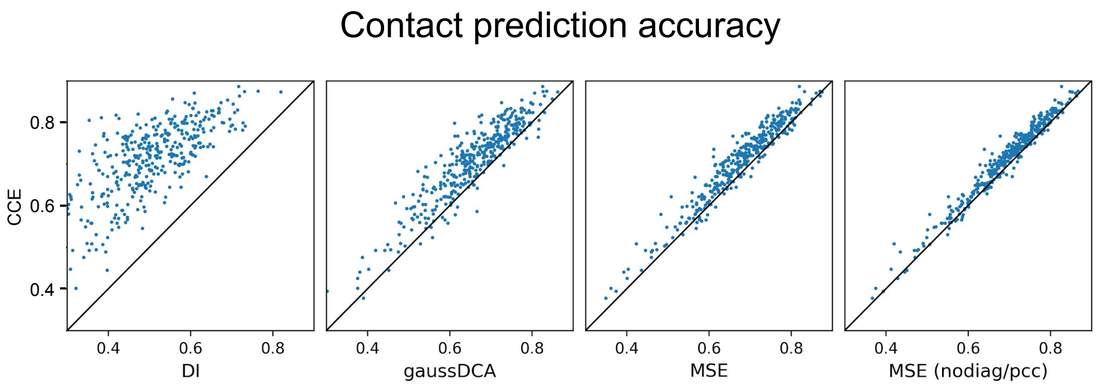
$\mathbf{c} = (\text{cov}(\mathbf{X}) + \lambda I)^{-1}$

Interesting observations:

- Pseudo-likelihood = Categorical cross-entropy (CCE)
- Multivariate Gaussian = Mean square error (MSE)
- L2 regularization = “shrinkage”
- Zeroing diagonal = “Partial correlation coefficient”

Results

Contact prediction accuracy



For contact prediction, Categorical cross entropy (CCE) is significantly more accurate than traditional inverse-covariance methods (DI, GaussDCA). By applying L2 regularization to MSE, we find the inv-cov method to be almost as accurate as CCE. Each point corresponds to a different protein.

Conclusion

- Widely used models for multivariate distributions in biological sequences can all be expressed as a single fully-connected layer, where the weights and bias of the dense layer captures the co-evolution and conservation, respectively.
- The differences in the models comes down to the loss function used, where inverse-covariance methods are effectively minimizing the mean-squared error (more appropriate for continuous data) and markov-random-field methods are minimizing the categorical cross entropy (more appropriate for categorical data).

Emails:
We are looking for collaborators/postdocs/students!

Mor Nitzan mornitzan@fas.harvard.edu
Sergey Ovchinnikov so@fas.harvard.edu

Github code

http://github.com/sokrypton/seqmodels