Enhancing the Coevolutionary Signal
via machine learning

Travis Hoppe, Robert Best
hoppeta@mail.nih.gov

Abstract

Analysis of coevolutionary relationships between residue pairs in the sequences of folded proteins can yield information on which residues interact in the folded state, and hence produce a contact map based only on sequence information. The distribution of contacts in folded proteins is far from random, as evident from looking at any contact map. Therefore, it should be possible to use information from neighboring contacts in order to strengthen the predictive power of coevolutionary analysis. Here, we show that application of a machine-learning algorithm to contact maps from analysis of evolutionary couplings significantly improves the precision of the derived contact map with a sacrifice in sensitivity, so that many more contacts can be predicted with confidence.

Hypothesis / Motivation

In the sequence to structure prediction problem, it has been particularly fruitful to incorporate coevolutionary data. The assumption is that co-mutating residues in multiple sequence alignment have a higher probability of being in contact in the three-dimensional structure.

We concentrate on improving one of these models, GREMLIN[1], though the methods would work on similar schemes like DCA and PSICOV. In GREMLIN the top L residue pairs are chosen from a coevolutionary scoring matrix. This works well for low values of L but the signal-to-noise ratio drowns out the weaker contacts.

We propose a straight-forward machine learning scheme to infer a contact from the *local* score pattern with the physically motivated hypothesis that the presence of a high local scores are an indicator of a contact.

Coevolutionary Analysis: GREMLIN[1]

GREMLIN optimizes the pseudolikelihood of a maximum entropy model that matches the observed single and pair frequencies from a multiple sequence alignment:

$$P(D|v,w) = \sum_{n=1}^{N}\sum_{i=1}^{L} \log P(x_i^n | x_{\hat{\imath}}^n, v, w)$$

$$P(x_i^n | x_{\hat{\imath}}^n, v, w) = \frac{1}{Z_i} \exp\left(v_i(x_i^n) + \sum_{j=1, j\neq i}^{L} w_{ij}(x_i^n, x_j^n) \right)$$

v_i encodes individual propensity of each amino acid at position i w_{ij} statistical coupling of amino acid propensities between positions i, j. This formulation models conditional distribution of the original joint distribution instead of the joint distribution itself.

Methods

- 150 globular proteins with known structure (50-275 residues each) were selected from set of non-homologous proteins covering diverse motifs and structures.
- Protein sequences were aligned using HHBLITS.
- Alignments were scored using GREMLIN.
- Scores were reduced from the original $(N, N, 21, 21)$ tensor by dropping gaps, taking the Frobenius norm and applying the Average Product Correlation (APC).
- Using a 4-fold cross-validated scheme, "images" of 5x5 were taken from the training set of GREMLIN scores to predict if a true native contact was given at the center.
- An extremely random forest (RF) was trained over the images.
- Contact maps for a given cutoff were predicted from both the GREMLIN and RF models. For reference, contact maps were also computed from the native structure.
- Coarse-grained molecular dynamics simulations were performed with each contact map to test folding.

Improved Contact Map Prediction

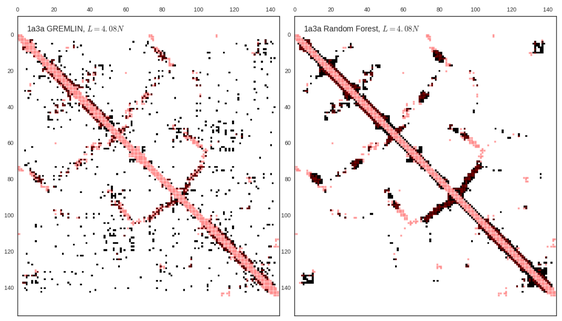

Figure 1: *Both: contact map of protein 1a31 (IIA Mannitol from E. Coli) in red; Left: original GREMLIN scoring method on in black; Right: improved RF in black.*
Using the traditional scoring method, false positives appear uniformly randomly across the contact pairs, while the RF method is more precise by utilizing local information.

ROC Curve

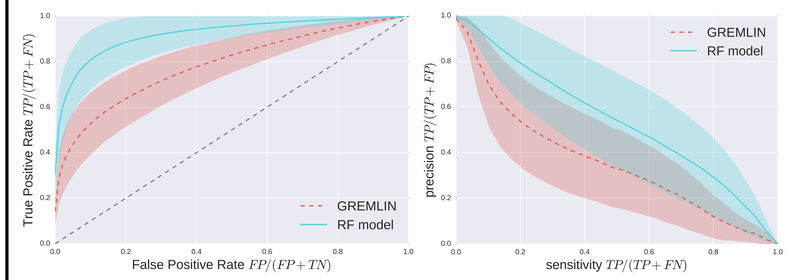

Figure 2: *Left: Receiver operating curve for GREMLIN and RF models; Right: same for sensitivity vs. precision. For all values of the sensitivity, RF-GREMLIN outperforms the traditional GREMLIN by choosing more correct contacts.*

Contact improvement is localized

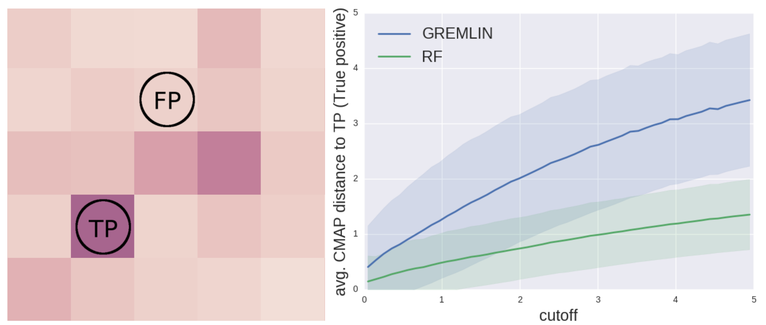

Figure 3: *Both methods give false positive (FP), but the predictions by the RF were closer to the true positives (TP) in contact map space. We measured the average city-block (L1-norm) distance from each FP to the nearest TP and averaged over all proteins contact maps predicted. Left: diagrammatic illustration of a FP to a TP, showing a distance of three. Right: Averaged L1 distance for all FPs to TPs across the dataset.*

Folding simulations

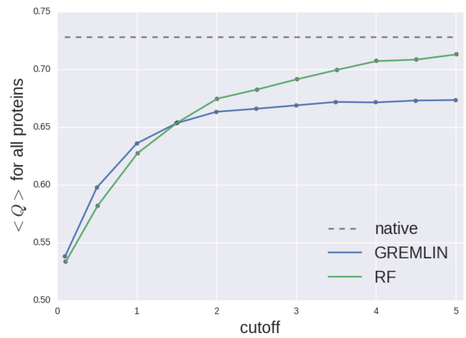

Figure 4: *To test the efficacy of the contact maps as a tool to predict the native fold of a protein (without any other prior knowledge) we employed a coarse-grained molecular-dynamics simulation. Each residue was modeled by a single C_α atom with a standard backbone potential and an attractive component if two residues were predicted to be in contact. The fraction of true native contacts formed after a short folding simulation, Q, are plotted.*

Feature selection

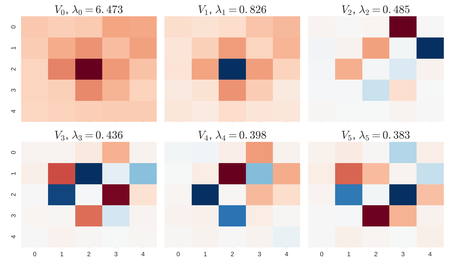

Figure 5: *To illustrate feature selection of the model, we perform singular value decomposition of terminal leaf nodes from the random forests. The dominant features (V_1, V_2) are roughly described by a Gaussian kernel and the difference between the center and the outlying values.*

References

[1] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

Biophysical Society Meeting 2016, Los Angeles CA

NIDDK