Improved Protein Residue–Residue Contact Prediction Using Image Denoising Methods

A. Villegas-Morcillo, J. A. Morales-Cordovilla,
A. M. Gomez, V. Sanchez
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain

Introduction

Motivation

- **Protein 3D structure** is closely related to its biological function.
- Generation of huge quantities of protein **amino acid sequences** from DNA sequencing processes.
- We need **computational methods** that predict the protein structure from its sequence.
- Advances in template-free modeling are motivated by **contact map representations**.
- But **Gaussian noise** is found in estimated contact maps from evolutionary couplings.

Objective

- Improve the prediction of **protein inter-residue contacts** by reducing Gaussian noise in estimated contact maps.

Experimental Framework

Datasets

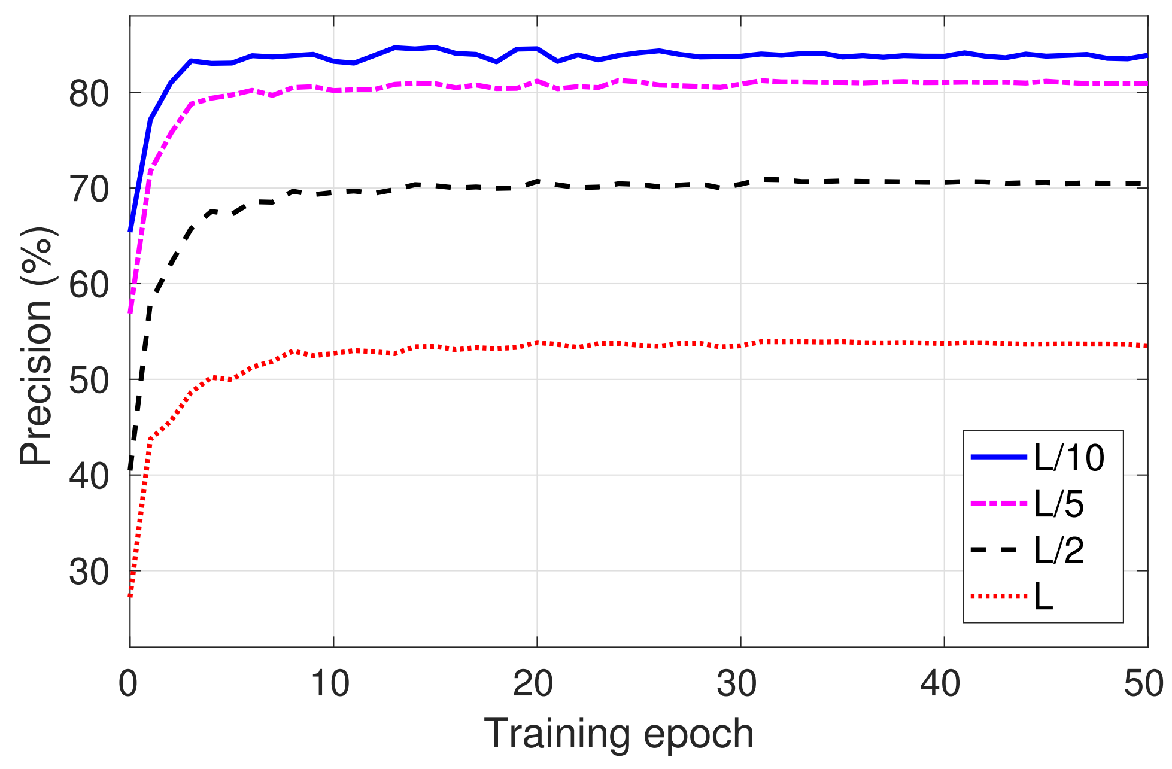
- **Test**: 150 Pfam proteins, 116 proteins from CASP10, and 103 proteins from CASP11.
- **DCNN training**: 3427 proteins in total (with 300 proteins to validate).

Evaluation Criteria

- Divide contacts in short- (6−11), medium- (12−23), and long-range (>23 amino acids).
- Compute the precision of top L/k contacts with L: sequence length and $k = \{10,5,2,1\}$.

Parameter Setting

- **K-SVD**: patches of size 5×5 and dictionaries with $K = 900$ atoms.
- **DCNN**: patches of size 35×35 (stride 10), batches of 256 and 50 epochs (early stopping at epoch 31).



Conclusions

Conclusions

- **Contact precision** values increase after applying noise reduction techniques.
- **Residual DCNN** strategy performs the best identifying more true contacts.

Future Work

- Explore other DCNN architectures.
- Study the impact of improved contacts in the prediction of the **protein 3D structure**.

Proposed Methods for Contact Map Denoising

Image Denoising Problem

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}$$

- **Y**: True contact map (PDB structure) showing residue spatial proximity (C_β distance < 8 Å).
- **X**: Estimated contact map (CCMpred method) from the protein multiple sequence alignment (MSA) using evolutionary coupling analysis.
- **Z**: Additive Gaussian noise.

Dictionary Learning for Sparse Representations

- **K-SVD** method with OMP algorithm.
- Divide **X** in patches and get sparse coding vectors:

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{\alpha}_i\|_0 \text{ subject to } \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \leq \epsilon^2$$

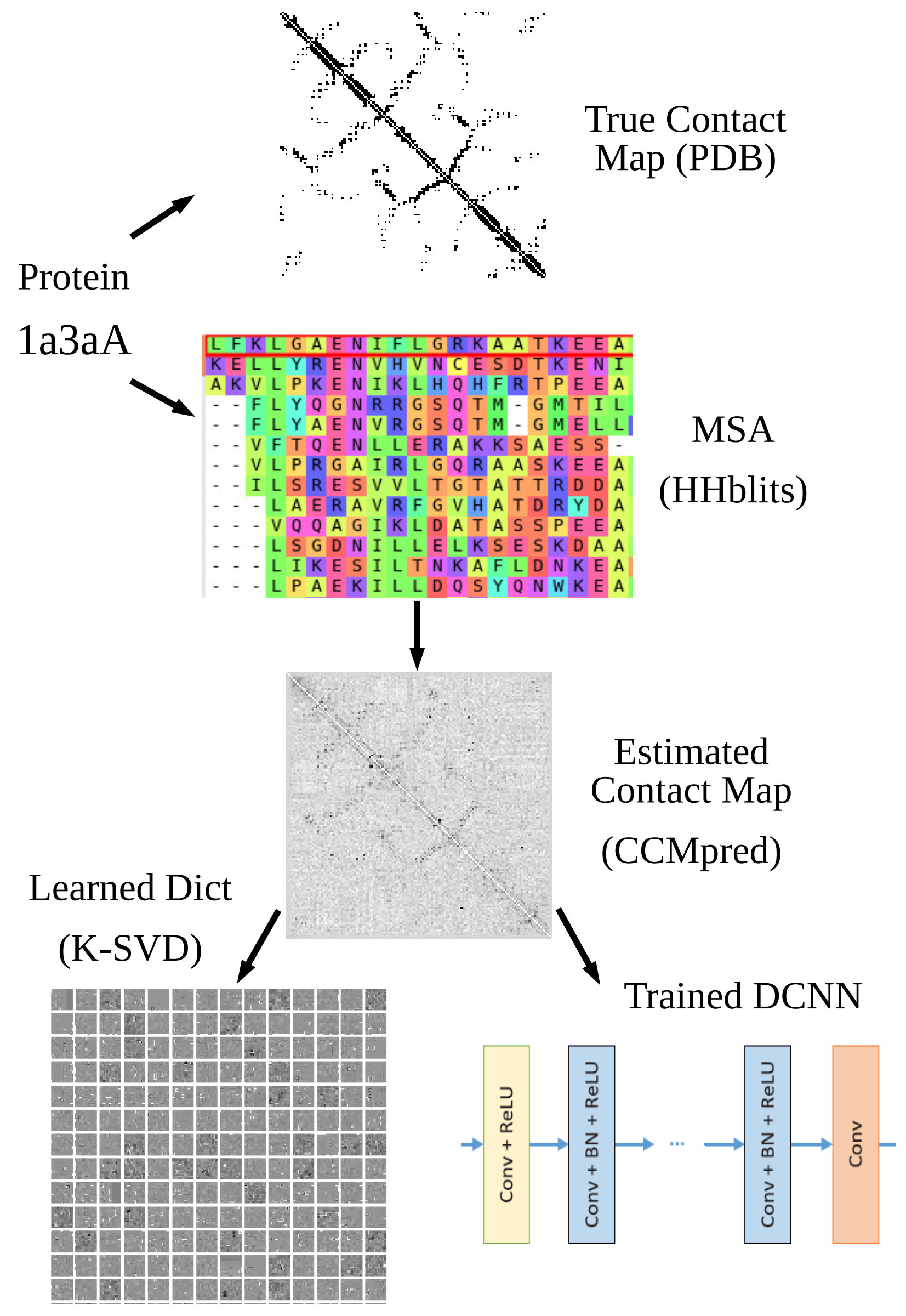
- Update patches $\hat{\mathbf{x}}_i = \mathbf{D}\hat{\boldsymbol{\alpha}}_i$ and dictionary **D** with SVD.
- Reconstruct image averaging denoised patches:

$$\hat{\mathbf{X}} = (\lambda\mathbf{X} + \textstyle\sum_i \mathbf{R}_i^T \mathbf{D}\hat{\boldsymbol{\alpha}}_i)/(\lambda\mathbf{I} + \sum_i \mathbf{R}_i^T \mathbf{R}_i)$$

DCNN Training with Residual Learning

- Deep architecture: 17 layers applying 64 convolutional filters of size 3×3 and ReLU activations.
- Optimization: Adam algorithm and batch-normalization.
- Train a residual mapping $\mathcal{R}(\mathbf{X}) \approx \mathbf{Z}$ and then calculate $\hat{\mathbf{X}} = \mathbf{X} - \mathcal{R}(\mathbf{X})$. Loss function:

$$l(\boldsymbol{\Theta}) = \tfrac{1}{2B} \textstyle\sum_{i=1}^{B} \|\mathcal{R}(\mathbf{x}_i;\boldsymbol{\Theta}) - (\mathbf{x}_i - \mathbf{y}_i)\|_F^2$$



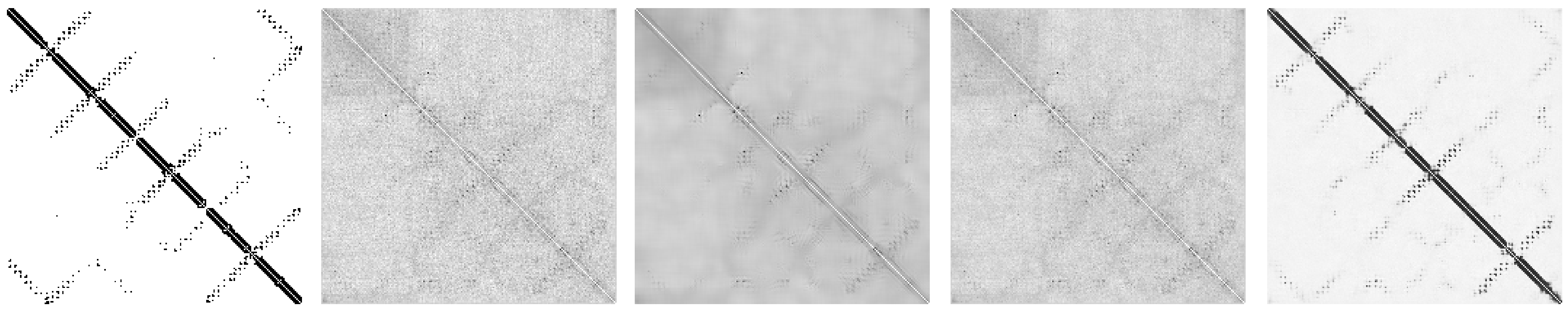
Results

- **Contact precision** values for short-, medium- and long-range for the evaluated methods on the three test datasets.

Test dataset	Method	Short-range				Medium-range				Long-range			
		L/10	*L/5*	*L/2*	*L*	*L/10*	*L/5*	*L/2*	*L*	*L/10*	*L/5*	*L/2*	*L*
150 Pfam proteins	Baseline	56.1	40.2	23.0	15.5	64.2	49.8	29.0	18.4	78.1	71.0	50.5	33.7
	R_2C filter	51.3	36.8	22.0	15.2	64.5	49.4	29.9	19.2	78.9	70.1	51.2	35.8
	K-SVD OMP	55.7	39.7	24.1	16.3	67.2	53.0	32.2	20.8	79.8	73.0	54.0	38.4
	DCNN	**77.6**	**65.2**	**40.9**	**25.0**	**80.5**	**71.0**	**48.3**	**30.3**	**89.6**	**85.3**	**72.1**	**54.5**
116 CASP10 proteins	Baseline	41.5	31.2	19.4	13.5	53.1	41.9	26.3	18.1	53.7	47.8	34.4	23.1
	R_2C filter	41.3	30.6	19.8	14.3	54.0	42.5	27.8	19.1	57.1	51.1	37.3	26.4
	K-SVD OMP	43.1	32.2	20.6	14.7	55.6	43.8	29.8	20.4	56.3	49.7	38.2	27.1
	DCNN	**58.4**	**48.5**	**31.9**	**20.7**	**65.8**	**58.1**	**43.0**	**29.8**	**67.3**	**63.2**	**50.6**	**37.6**
103 CASP11 proteins	Baseline	32.9	23.9	15.3	11.3	38.0	28.5	17.8	12.5	47.4	40.2	28.9	20.1
	R_2C filter	31.4	22.8	14.6	11.5	39.7	29.6	19.2	13.8	50.0	42.5	30.7	22.3
	K-SVD OMP	32.8	23.4	15.4	12.1	40.4	31.7	20.6	14.3	48.5	42.4	31.4	22.7
	DCNN	**48.1**	**38.8**	**26.1**	**17.6**	**53.7**	**46.6**	**31.9**	**21.2**	**56.5**	**53.2**	**43.0**	**32.6**

- Baseline: estimated contact maps using **CCMpred**.
- Comparison with other contact map denoising method: **R_2C filter**.
- Worse-than-baseline results are marked in red, and the best results are in **boldface**.

- Example: resulting contact maps for protein domain **T0682-D1** (CASP10 dataset).



a) True (PDB) b) Estimated (CCMpred) c) Denoised (R_2C) d) Denoised (K-SVD OMP) e) Denoised (DCNN)

Contact Information

Amelia Villegas-Morcillo
E-mail: melyvm@correo.ugr.es
Signal Processing, Multimedia Transmission and Speech/Audio Technologies Group, **SigMAT**
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain