P-04

Boltzmann Machine Learning and Regularization Methods for Inferring Evolutionary Fields and Couplings from a Multiple Sequence Alignment

Sanzo Miyazawa
sanzo.miyazawa@gmail.com

at IIBMP2021, the annual meeting of the Japanese Society for Bioinformatics,
in September 27–29, 2021

Abstract

The inverse Potts problem to infer a Boltzmann distribution for homologous protein sequences from their single-site and pairwise amino acid frequencies recently attracts a great deal of attention in the studies of protein structure and evolution. We study regularization and learning methods and how to tune regularization parameters to correctly infer interactions in Boltzmann machine learning. Using L_2 regularization for fields, group L_1 for couplings is shown to be very effective for sparse couplings in comparison with L_2 and L_1. Two regularization parameters are tuned to yield equal values for both the sample and ensemble averages of evolutionary energy. Both the averages smoothly change, but their learning profiles are very different between learning (gradient-descent) methods, Adam, NAG, and modified PROP. The Adam method is modified to make an increment vector proportional to the gradient vector for sparse couplings and to use a soft-thresholding function for group L_1. It is shown by first inferring interactions from protein sequences and then from Monte Carlo samples that the fields and couplings can be well recovered, but that recovering the pairwise correlations in the resolution of a total energy is harder for the natural proteins than for the protein-like sequences. Selective temperature for folding/structural constrains in protein evolution is also estimated.

Publications:

1. IEEE/ACM IEEE/ACM Trans. Comput. Biol. Bioinform. 2020; DOI:10.1109/TCBB.2020.2993232 (arXiv:1909.05006)
2. J. Theor. Biol. 433, 21-38, 2017; DOI:10.1016/j.jtbi.2017.08.018 (arXiv:1612.09379)

1. Background

The probability distribution ($P(\boldsymbol{\sigma})$) of homologous sequences ($\boldsymbol{\sigma}$) in a protein family can be well approximated by a Boltzmann distribution (Figliuzzi et al., 2018):

$$P(\boldsymbol{\sigma}) \quad \propto \quad \exp(-\psi_N(\boldsymbol{\sigma})) \;\;,\;\; \psi_N(\boldsymbol{\sigma}) \equiv -(\sum_{i}^{L}(h_i(\sigma_i) + \sum_{j>i} J_{ij}(\sigma_i, \sigma_j))) \tag{1}$$

where $\sigma_i \in$ {amino acids, deletion}, h_i is one-body at site i, and J_{ij} is two-body interaction between sites i and j. The estimation of h and J from homologous sequences is not only useful for predicting residue contacts in protien structure but for estimating folding free energy and evolutionary fitness of proteins.

- A protein folding theory based on the random energy model (REM) indicates:

$$P(\boldsymbol{\sigma}) \quad \propto \quad P^{\text{mut}}(\boldsymbol{\sigma}) \exp(\frac{-\Delta G_{ND}(\boldsymbol{\sigma}, T)}{k_B T_s}) \quad \propto \exp(\frac{-G_N(\boldsymbol{\sigma})}{k_B T_s}) \quad \text{if } \boldsymbol{f}(\boldsymbol{\sigma}) = \text{constant} \tag{2}$$

  where $\Delta G_{ND} \equiv G_N - G_D$, G_N and G_D are the free energies of native and denatured state, T_s is the effective temperature representing the strength of selection pressure (Shakhnovich et al., 1993).

- We prove that if a mutational process in protein evolution is a reversible Markov process, the equilibrium ensemble of genes for diploid will obey a Boltzmann distribution (Miyazawa, 2017):

$$P(\boldsymbol{\sigma}) \quad \propto \quad P^{\text{mut}}(\boldsymbol{\sigma}) \exp(4N_e m(1 - 1/(2N))) \tag{3}$$

  where N_e and N are effective and actual population sizes, and m is the Malthusian fitness of a gene.

2. Purposes of the present study

In the Boltzman machine, which is slower but can estimate h and J more acurately than the mean field and pseudo-likelihood approximations, the cross entropy $S(\phi)$ including a regularization $R(\phi)$ is minimized by a gradient descent method;

$$S(\phi) \equiv \frac{-1}{W} \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \log P(\boldsymbol{\sigma}_N|\phi) + R \quad \text{where} \quad h_i(a_k) = \phi_i - \sum_{j(\neq i)} \sum_l \phi_{ij}(a_k, a_l) P_j(a_l)\,,\ J_{ij}(a_k, a_l) = \phi_{ij}(a_k, a_l) \qquad (4)$$

where $P_j(a_l) \equiv \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \delta_{\boldsymbol{\sigma}_{Nj}, a_l} / W$, $W \equiv \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}$, and $a_k \in \{\text{amino acids, deletion}\}$. The equilibrium distribution $P(\boldsymbol{\sigma}_N|\phi)$ is generated by the MCMC method. For protein sequences:

- Interactions J_{ij} should be sparse and significant for closely-located, interacting residue pairs in a 3D structure.
- The random energy model (REM) for proten folding indicates that the sample mean of $\psi_N(\boldsymbol{\sigma_N})$ over homologous sequences is equal to the ensemble average over the Boltzmann distribution, which may be evaluated by approximating the distribution of $\psi_N(\boldsymbol{\sigma})$ of random sequences as a Gaussian distribution, $\bar{\psi} - \delta\psi^2$, where the $\bar{\psi}$ and $\delta\psi^2$ are the mean and variance of $\psi_N(\boldsymbol{\sigma})$.

In order to correctly infer interactions, we study

1. which regularizer is better, L2-L2, L2-L1, and L2-GL1, which denote L_2 for $h_i(a)$ and L_2, L_1, and group L_1 for $J_{ij}(a, b)$, respectively; soft-thresholding functions are used for L_1 and group L_1.
2. which gradient descent method is appropriate, Adam, NAG, modified RPROP, and modified Adam invented here; in all except the first one an increment vector is proportional to a gradient vector.
3. how to tune regularization parameters, λ_1 and λ_2; both the sample and ensemble averages of evolutionary energy ψ_N should be equal to each other.

3. Results

3-1. Protein families and structures studied.

Table: Protein Families Employed.

Pfam ID	N / N_{eff} [a]	M [b] / M_{eff} [a]	L [c]	PDB ID
PF00595[†]	13814 / 4748.8	1255 / 340.0	81	1GM1-A:16-96
PF00153	54582 / 19473.9	255 / 139.8	97	2LCK-A:112-208

[†] Identical sequences are removed.

[a] The effective number of sequences, $\sum_{\sigma_N} w_{\sigma_N}$, where the sample weight w_{σ_N} for a natural sequence $\boldsymbol{\sigma}_N$ is equal to the inverse of the number of sequences that are less than 20% different from a given sequence.

[b] The number of unique sequences that include no deletion for PF00595 and no more than 2 for PF00153.

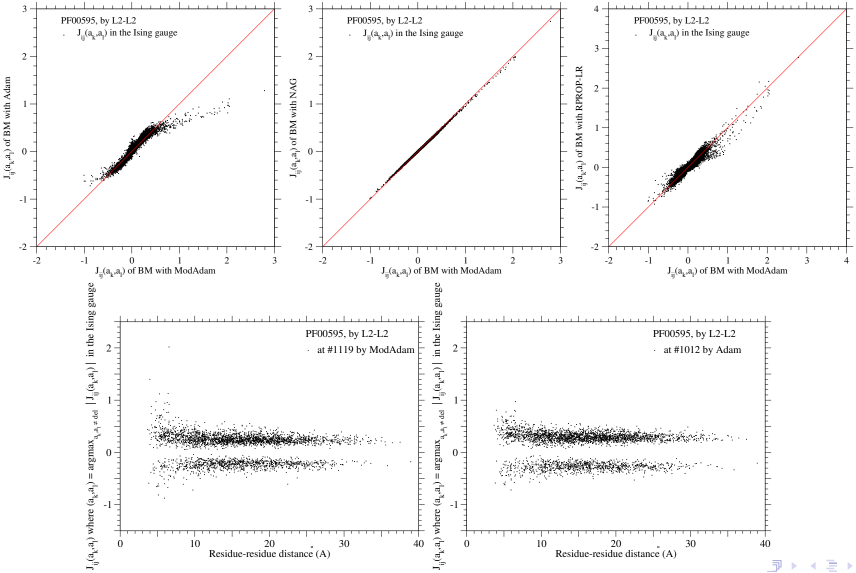
[c] The number of residues in a sequence.

3-2. Regularization Parameters and Characteristics of Boltzmann Machines[a] with the ModAdam for PF00595.

MSA	regularizers	λ_1	λ_2	#Iter [b]	D_1^{KL}	D_2^{KL}	$\delta\psi^2/L$ [c]	$(\bar{\psi} - \delta\psi^2)/L$ [d]	$\overline{\psi_N}/L$ [e]	$\overline{\psi_{MC}}/L$ [f]	Precision [g]
PF00595	L2-GL1	$0.100/N_{\text{eff}}$	$=\lambda_1$	1250	0.00506	0.0709	3.23	−3.64	−3.64	−3.29	0.565
PF00595[h]	L2-GL1	$0.100/N_{\text{eff}}$	$40.0/N_{\text{eff}}$	1162[†]	0.00369	0.0759	2.75	−3.15	−3.15	−2.79	0.588
										(−2.81[i])	
MC1162[‡]	L2-GL1	$0.100/N_{\text{eff}}$	$40.0/N_{\text{eff}}$	1151	0.00283	0.0689	2.61	−2.98	−2.80	−2.63	0.500
									(−2.82[j])		
MC1162[‡]	L2-GL1	$0.891/N_{\text{eff}}$	$=\lambda_1$	1280	0.00296	0.0621	2.76	−3.14	−3.15	−2.93	0.457
MC1162[‡][h]	L2-GL1	$0.891/N_{\text{eff}}$	$12.6/N_{\text{eff}}$	1183	0.00275	0.0646	2.63	−3.00	−3.00	−2.79	0.483
									(−2.93[j])		
PF00595	L2-L1 [k]	$0.100/N_{\text{eff}}$	$=\lambda_1$	1201	0.00674	0.0747	3.19	−3.60	−3.61	−3.31	0.563
PF00595[h]	L2-L1 [k]	$0.100/N_{\text{eff}}$	$0.316/N_{\text{eff}}$	1007	0.00497	0.0736	3.08	−3.48	−3.49	−3.13	0.560
PF00595	L2-L2	$0.100/N_{\text{eff}}$	$=\lambda_1$	1047	0.00580	0.0737	3.13	−3.54	−3.55	−3.27	0.557
PF00595[h]	L2-L2	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1119	0.00387	0.0725	2.99	−3.39	−3.39	−3.04	0.551

Learning method	regularizers	λ_1	λ_2	#Iter [b]	D_1^{KL}	D_2^{KL}	$\delta\psi^2/L$ [c]	$(\bar{\psi} - \delta\psi^2)/L$ [d]	$\overline{\psi_N}/L$ [e]	$\overline{\psi_{MC}}/L$ [f]	Precision [g]
ModAdam [h]	L2-L2	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1119	0.00387	0.0725	2.99	−3.39	−3.39	−3.04	0.551
(second run) [i]				2018	0.00372	0.0696	3.12	−3.53	−3.52	−3.16	0.568
Adam [h]	L2-L2	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1012	0.00320	0.0681	3.35	−3.73	−3.59	−3.23	0.563
NAG [h]	L2-L2	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1110	0.00381	0.0724	2.94	−3.34	−3.34	−3.01	0.557
[i]				2095	0.00361	0.0690	3.08	−3.48	−3.48	−3.12	0.565
RPROP-LR [j]	L2-L2	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1052	0.00391	0.0766	2.97	−3.36	−3.28	−2.95	0.560

3-3. Adam under-estimates $J_{ij}(a_k, a_l)$, particularly for strong interactions between closely-located site pairs, because increments are the same order for all variables.



3-4. The profile of the average evolutionary energies along the learning process.

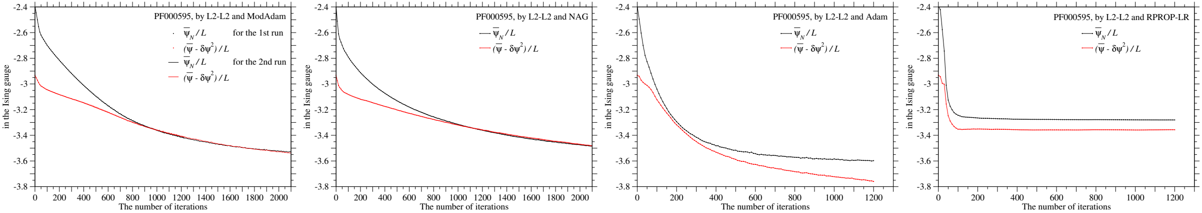


Figure: The profile of the average evolutionary energies along the learning process in the L2-L2 model by each gradient-descent method, ModAdam, NAG, Adam, and RPROP-LR from the left to right, for PF00595.

3-5. Regularization Parameters and Characteristics of Boltzmann Machines[a] with the ModAdam for PF00153.

MSA	regularizers	λ_1	λ_2	#Iter[b]	D_1^{KL}	D_2^{KL}	$\delta\psi^2/L$[c]	$(\bar\psi - \delta\psi^2)/L$[d]	$\overline{\psi_N}/L$[e]	$\overline{\psi_{MC}}/L$[f]	Precision[g]
PF00153	L2-GL1	$0.100/N_{\text{eff}}$	$=\lambda_1$	1084	0.00342	0.0264	2.71	−3.02	−3.29	−3.04	0.596
PF00153[h]	L2-GL1	$0.100/N_{\text{eff}}$	$209/N_{\text{eff}}$	1445[†]	0.00112	0.0318	2.50	−2.82	−2.83	−2.54	0.630
										(−2.51[i])	
MC1445[‡]	L2-GL1	$0.100/N_{\text{eff}}$	$209/N_{\text{eff}}$	1390	0.00151	0.0323	2.48	−2.82	−2.54	−2.52	0.630
									(−2.83[j])		
MC1445[‡]	L2-GL1	$7.94/N_{\text{eff}}$	$=\lambda_1$	1181	0.000975	0.0160	2.25	−2.57	−2.57	−2.47	0.551
MC1445[‡][h]	L2-GL1	$7.94/N_{\text{eff}}$	$20.0/N_{\text{eff}}$	1197	0.000985	0.0162	2.24	−2.55	−2.55	−2.43	0.557
									(−2.64[j])		
PF00153	L2-L1[k]	$0.100/N_{\text{eff}}$	$=\lambda_1$	1149	0.00313	0.0265	2.73	−3.05	−3.32	−3.09	0.599
PF00153[h]	L2-L1[k]	$0.100/N_{\text{eff}}$	$25.1/N_{\text{eff}}$	1208	0.00165	0.0318	2.57	−2.91	−2.91	−2.66	0.557
PF00153	L2-L2	$0.100/N_{\text{eff}}$	$=\lambda_1$	1223	0.00329	0.0264	2.76	−3.08	−3.35	−3.10	0.605
PF00153[h]	L2-L2	$0.100/N_{\text{eff}}$	$398/N_{\text{eff}}$	1066	0.00119	0.0336	2.55	−2.87	−2.86	−2.52	0.569

λ_1 and λ_2 are scaling constants of the regularizers for $\{\phi_i\}$ and $\{\phi_{ij}\}$, respectively.

D_1^{KL} and D_2^{KL} are the averages of the Kullback-Leibler divergences of the site and pairwise distributions over all residues or residue pairs.

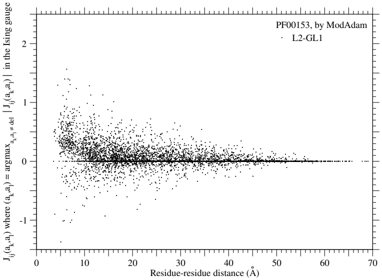
$\bar\psi$ and $\delta\psi^2$ are the mean and variance of ψ over random sequences; $\bar\psi - \delta\psi^2$ is equal to the ensemble average of $\psi(\sigma)$ in the Boltzmann distribution by the Gaussian approximation.

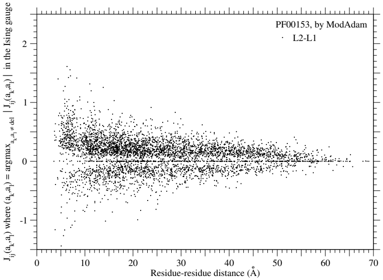
[e] The sample average of evolutionary energies per residue over the sequences with no more than 2 deletions for PF00153 and with no more than 3 for the MCMC samples; the Ising gauge is employed.

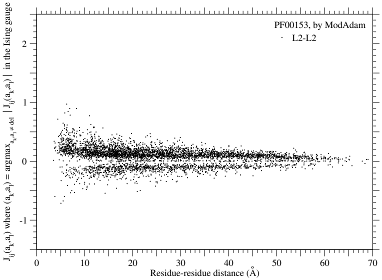
[f] The average of evolutionary energies per residue over the MCMC samples with no more than 3 deletions; the Ising gauge is employed.

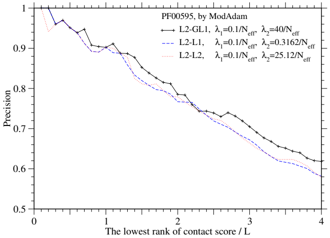
[g] Precision of contact prediction; the number of predicted contacts is 332, which is equal to the total number of closely located residue pairs within 8 Å between side-chain centers in the 3D protein structure. The corrected Frobenius norm of couplings is employed for the contact score.

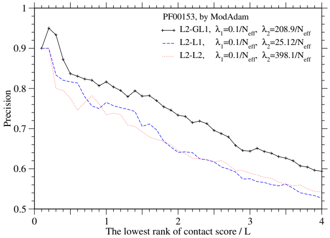
3-6. The L2-GL1 generates more reasonable values for $J_{ij}(a_k, a_l)$ than the L2-L1 and L2-L2.



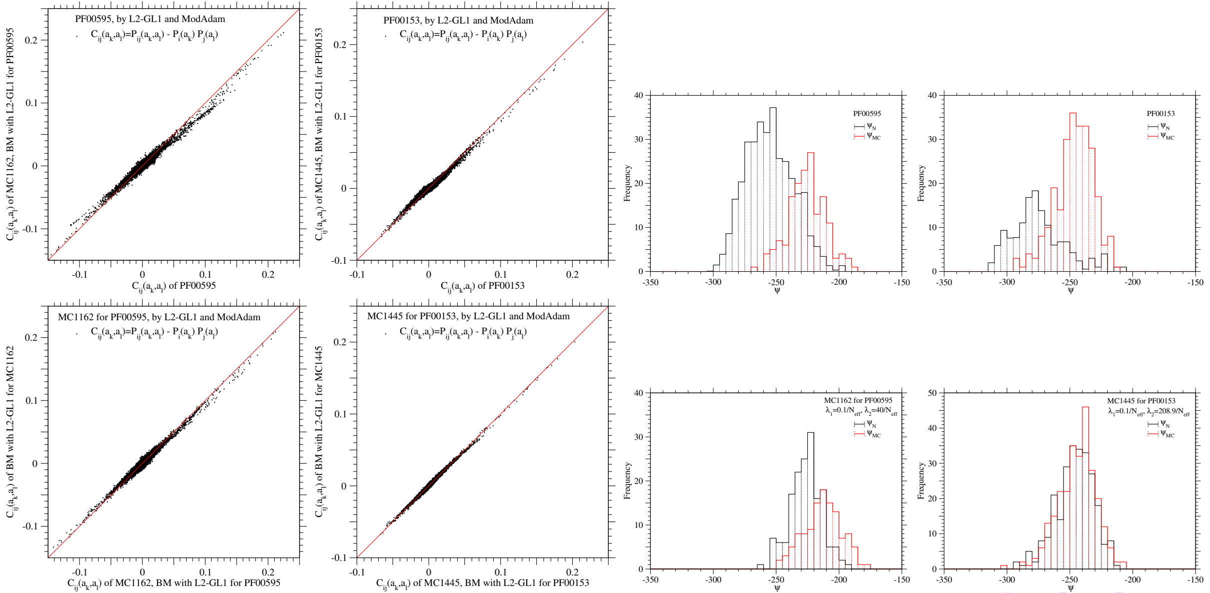




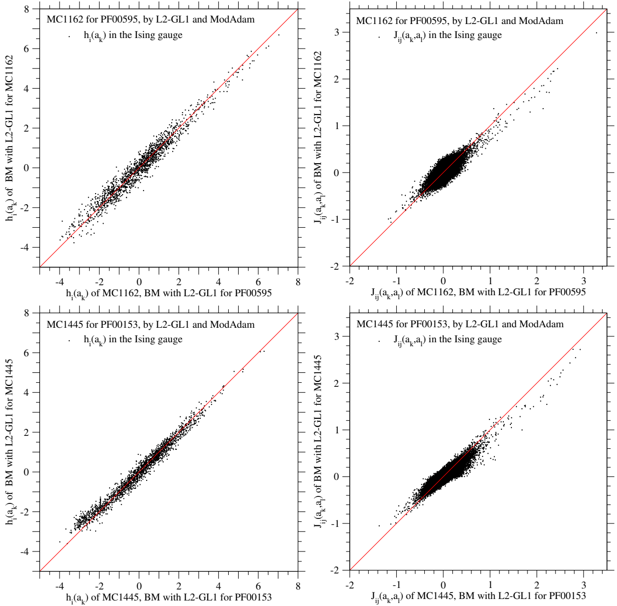




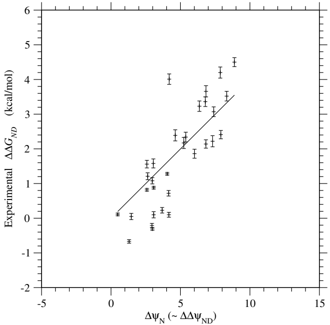
3-7. Recoverabilities of $\{C_{ij}(a_k, a_l)\}$ and Ψ_N for real protein and for protein-like sequences.



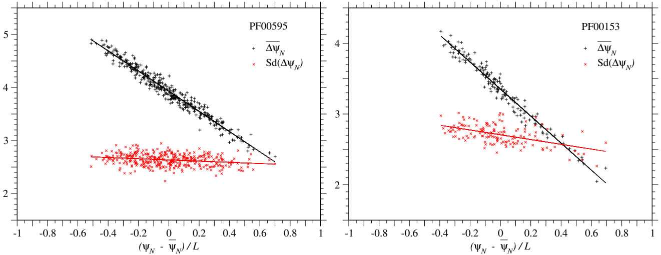
3-8. Reproducibilities of $\{h_i\}$ and $\{J_{ij}\}$ for protein-like sequences (MCMC samples).



3-9. Selective temperature T_s and constancy of the standard deviation of $T_s \Delta\Psi_N$ due to single amino acid substitutions.



$k_B T_s$ for a reference protein, PF00595.



Consistency of the standard deviation of $\Delta\psi_N$ over homologous proteins.

Pfam ID	$\overline{\psi_N}/L$ [a]	$\overline{\Delta\psi_N}$ [b]	Sd($\Delta\psi_N$) [b]	Sd(Sd($\Delta\psi_N$)) [c]	r_{ψ_N} [d]	α_{ψ_N} [d]	r_{ψ_N} [e]	α_{ψ_N} [e]	r [f]	slope [f]	$\hat{T}_s$	T_m^{exp} [g]	$\hat{T}_g$ [h]	$\hat{\omega}$
					for $\overline{\Delta\psi_N}$		for Sd($\Delta\psi_N$)			(kcal/mol)	(°K)	(°K)	(°K)	(k
PF00595	−3.15	3.94	2.64	0.113	−0.980	−1.90	−0.237	−0.113	0.920	0.400	201	313 [k]	215	1.
PF00153	−2.84	3.36	2.71	0.141	−0.981	−1.91	−0.537	−0.338			196			

4. Conclusions

The h and J learned by Boltzmann machine strongly depend on regularization methods and even gradient-descent methods as well as regularization parameters. Such hyper parameters should be optimized for maximizing some kind of accuracy in prediction. In the present case, however, the log-likelihood can be hardly evaluated even if its gradient can be easily calculated, and of course the true values for h and J are unknown for proteins. Here we employed as a criterion that J_{ij} is sparse and significant for closely-located site pairs in a protein structure, and the sample average and ensemble average of the total evolutionary energy (ψ_N) are equal to each other.

- The Adam method underestimates J_ij, particularly for strong interactions. The gradient-descent methods in which an increment vector is proportional to a gradient vector appear to be appropriate for sparse J_{ij}; they are also required for soft-threading functions for L_1 and group L_1. Here the modified Adam was chosen.
- As expected, the $L2 - GL1$ regularization is better for sparse J_{ij} than $L2 - L2$ and $L2 - L1$.
- The two regularization parameters, λ_1 for $\{\phi_i\}$ and λ_2 for $\{\phi_{ij}\}$, were determined to generate equal values for the sample and ensemble average of ψ_N estimated by the Gaussian approximation.
- $h_i(a_k)$ and $J_{ij}(a_k, a_l)$ for protein-like sequences can be well reproduced as long as they are significant. However, the distribution of ψ_N cannot be well reproduced for protein sequences but for protein-like sequences.
- $k_B T_s \mathrm{Sd}(\Delta\psi_N) \simeq \mathrm{Sd}(\Delta\Delta G_{ND})$ due to single amino acid changes is approximately constant over protein families, and can be employed to estimate T_s; $T_s \approx \{T_s \mathrm{Sd}(\Delta\psi_N)\}_{\text{ref. prot.}} / \mathrm{Sd}(\Delta\psi_N)$.
  (Refer to J. Theor. Biol. 433, 21-38, 2017; DOI:10.1016/j.jtbi.2017.08.018 (arXiv:1612.09379).)