EPFL

Exploring the sequence landscape of proteins with direct coupling analysis and pseudolikelihood maximisation

MAJEUX Zoé

Prof. DE LOS RIOS Paolo

Laboratory of Statistical Biophysics

Abstract

Misfolded proteins can lead to aggregation, resulting in neuromuscular and neurodegenerative diseases, or lysosomal dysfunction. Heat shock proteins 70 (HSP70) play a crucial role as chaperones in various protein folding processes, involving ATP hydrolysis facilitated by the J-domain binding to HSP70. Recent research have explored DnaJ domain and SIS1 protein sequences using Direct Coupling Analysis, pseudolikelihood maximization, and machine learning. However, inappropriate results necessitated a new approach, involving code modifications and optimization. These changes include a new couplings formulation, a variable number of amino acid values per position, a smaller batch size, hyperparameter tuning with different optimizers (Adam, AdamW, SGD, Adagrad, and AdaDelta), a comparison by taking the average across different models or couplings, or on the Frobenius norms. Additionally, a comparison was conducted by learning the class of the sequence or not. Furthermore, protein contact predictions were also performed for the Mitochondrial protein import protein MAS5 (gene YDJ1).

January 8, 2024

1 Introduction

The information required by a protein to attain its correct 3D conformation is encoded in its amino acid sequence [1]. Certain proteins, known as chaperones mediate and facilitate the folding of other proteins by binding to them and stabilizing unfolded or partially folded polypeptide chains [1]. Without these chaperones, such chains may become unstable and aggregate with others. Many chaperons were initially called in 1975 Heat Shock Proteins (Hsp) [3] and initially known to facilitate the refolding of proteins that have been denatured by high temperature [1]. One prominent member of these proteins is the Hsp70, which plays important roles in various proteins processes including proteins folding, new proteins assembly, misfolded proteins refolding, and degradation of proteins [4]. Hsp70 constitutes an extremely conserved family of molecular chaperones in both prokaryotic and eukaryotic cells [5]. Hsp70 is a high subjet of interest for the scientific community for about fifty years [3]. Initial studies showcased its exceptional conservation across time and species [2]. In 2005, scientists discovered that certain proteins, known as co-chaperones, could stimulate ATP activity to facilitate the folding function of Hsp70. J-domain Proteins (JDPs) are a subset of these co-chaperones, and their association with Hsp70 is critical to preventing the aggregation of non-native proteins [4].

The molecular structure of Hsp70 comprises an N-terminal Nucleotide Binding Domain (NBD) connected by a short, highly conserved, and flexible linker to a Substrate Binding Domain (SBD), followed by a disordered C-terminal portion [5]. Although structural studies date back to the 90s, revealing the NBD as two lobes with a deep cleft and subdomains contributing to ATP binding and hydrolysis [6], certain aspects remain unclear and warrant further analysis.

In the future, new sequences of the Hsp70 family or novel J-domain sequences may require investigations. A fast method to visualize their structural conformation is crucial, considering that a protein's function heavily relies on its structure. Since determining the structure from a single sequence is challenging, the approach involves utilizing evolutionary information and sequence comparison to identify coevolving pairs of amino acids—those that consistently change together [7]. By training a model to predict one amino acid based on others, a couplings analysis can determine these coevolving amino acids and provide contact predictions. The results can then be compared with predictions made by AlphaFold [8]. Modifications to a previous code, utilizing direct couplings

analysis and machine learning with pseudolikelihood maximization [9], aim to enhance performance. Previous evaluations on the JDP (length 63) were suboptimal, leading to further work with hyperparameter tuning on the JDP and the protein SIS1 (length 352) [10]. However, the network struggled to predict SIS1 contacts accurately. Here, new improvements are done: a new couplings formulation, a variable number of amino acid values per position, a smaller batch size, hyperparameter tuning with different optimizers (Adam, AdamW, SGD, Adagrad, and AdaDelta), a comparison by taking the average across different models or couplings or on the Frobenius norms, and a comparison by learning the class of the sequence or not. Additionally, predictions for the Mitochondrial protein import protein MAS5 (gene YDJ1) (length 409) have been conducted.

2 Statistical theory

2.1 Direct Couplings Analysis (DCA)

Over the course of evolution, protein sequences can undergo modifications, leading to differences in the same protein among various species. However, despite these variations, the fundamental functions and three-dimensional structures of proteins should not be lost [11]. Proteins sharing such similarities are referred to as homologous proteins and can be categorized into protein families through multiple sequence alignments (MSA) [11]. An example of homologous proteins is depicted in Figure 1, visualized using AliView [12].

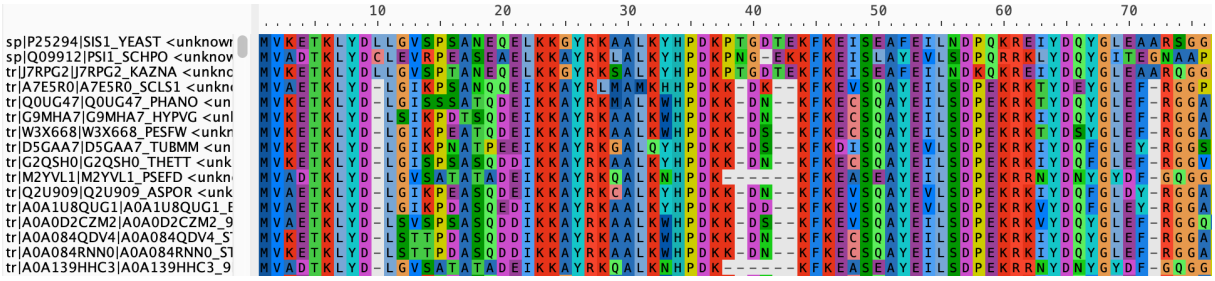


Figure 1: Crop of Aliview's results of some homologous sequences for SIS1 protein

MSA is used to represent each amino acids of L sequences of size N. This is represented by a rectangular matrix $A = \{a_m^l | i = 1, ...M, l = 1, ..., L\}$ containing L sequences (the rows) of size M. The columns m correspond to the amino acids at position m of the sequence $l \in [1, L]$ [14]. The amino acid letter is replaced by a number $k \in [1, K]$ (in general K=21 because 20 natural amino acids, 1 gap). This matrix keeps a lot of secrets

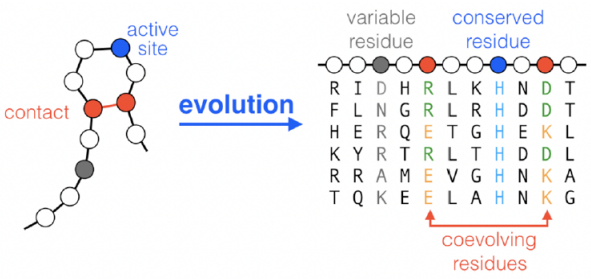


Figure 2: Information behind a MSA matrix [13]

about the protein structure as the different contacts or correlations between the amino acids of a protein belonging to this family (see Fig.2). This periodicity can be exploited by statistical models as direct coupling analysis (DCA) method that aims in providing a protein family-specific probability distribution by using the MSA (inverse statistical physics) and Potts model. First, note that given a MSA, the individual frequency, to have the amino acid k for the column m, and the pairwise frequencies, to have the pair (k,p) at columns m and n, are given by [14]:

$$\mathbf{f_m}(k) = \frac{1}{L}\sum_{l=1}^{L}\delta[a_m^l = k]$$

$$\mathbf{f_{m,n}}(k,p) = \frac{1}{L}\sum_{l=1}^{L}\delta[a_m^l = k]\delta[a_n^l = p]$$

with δ the Kronecker delta. These values are essential to predict the correlation $\mathbf{C_{m,n}}(k,p)$ for the amino acids at position m,n beeing respectively of values k and p with $k,p \in [1,K]$:

$$\mathbf{C_{m,n}}(k,p) = \mathbf{f_{m,n}}(k,p) - \mathbf{f_m}(k)\mathbf{f_n}(p)$$

Secondly, note that according to the Potts mode, the probability to have the amino acid k at position m and the probability to have the couple (k,p) at position (m,n) are given by [14]:

$$P(a_m = k) = \sum_{\mathbf{a},a_m=k} P(\mathbf{a}) = \mathbf{f_m}(k)$$

$$P(a_m = k, a_n = p) = \sum_{\mathbf{a},a_m=k,a_n=p} P(\mathbf{a}) = \mathbf{f_{m,n}}(k,p)$$

Assuming the MSA to be a sample of a Boltzmann distribution $P(\mathbf{a}) = \frac{1}{Z}\exp(-\beta H)$ and by taken the maximization of the entropy $S = -\sum_{\mathbf{a}} P(\mathbf{a})\ln P(\mathbf{a})$ gives the probability to have the sequence $\mathbf{a}$:

$$P(\mathbf{a}) = \frac{1}{Z}\exp\left(\sum_{m=1}^{M}\mathbf{h_m}(a_m) + \sum_{m=1}^{M-1}\sum_{n=m+1}^{M}\mathbf{J_{mn}}(a_m,a_n)\right)$$

with Z a normalization constant, $\mathbf{h_m} = (h_{m,1}, .., h_{m,K})^T$ a vector of fields and $\mathbf{J_{m,n}} = \begin{pmatrix} C_{mn,11} & ... & C_{mn,1K} \\ ... & ... & ... \\ C_{mn,K1} & ... & C_{mn,KK} \end{pmatrix}$ the matrix of couplings between the positions a_m^l and a_n^l for different values of amino acids $\in [1, K]$ [14]. A large $\mathbf{h_m}$ indicates a preference of position h toward the amino acid k and a large $\mathbf{J_{mn,kp}}$ indicates a high contact probability between the positions (m, n) of amino acids values (k, p) [14]. The inferring problem consists to find these quantities.

2.2 Alternative representation of MSA in one hot encoder

The MSA gives a rectangular matrix $A = \{a_m^l | i = 1, ...M, l = 1, ..., L\}$ such that each amino acid a_m^l of the sequence l takes a value $k \in [1, K]$. This matrix can be expressed as a 3 dimensional tensor of size LxMxK such that each amino acid a_m^l is now a one hot encoder vector of dimension K. This representation gives the following condition:

$$\sum_{k=1}^{K} a_{m,k}^l = 1 \qquad \forall l \in [1, L] \ and \ \forall m \in [1, M] \tag{1}$$

2.3 Inferring with pseudolikelihood maximisation

Let's recall that with pseudolikelihood approximation, the probability $P(a_m)$ to have the amino acid m is given by the product along each sequence of each probability to have the amino acid m knowing the other amino acids of the sequences

$$P(a_m) \approx \prod_{l=1}^{L} P(a_m = a_m^l | \mathbf{a}_{n \neq m} = \mathbf{a}_{n \neq m}^l) \tag{2}$$

By using that $P(A|B) = \frac{A \cap B}{P(B)}$, the probability to have the sequence amino acid $a_m$in the sequence l is given by (For the last equality, several terms cancel because the only distinction parts are for varying values a_m)

$$P(a_m = a_m^l | \mathbf{a}_{n \neq m} = \mathbf{a}_{n \neq m}^l) = \frac{P(a_m = a_m^l, \mathbf{a}_{n \neq m} = \mathbf{a}_{n \neq m}^l)}{P(\mathbf{a}_{n \neq m}^l = \mathbf{a}_{n \neq m}^l)} = \frac{P(a_m = a_m^l, \mathbf{a}_{n \neq m} = P(\mathbf{a}_{n \neq m}^l)}{\sum_{k=1}^{K} P(a_{n=k}, \mathbf{a}_{n \neq m}^l)}$$
$$= \frac{\exp\left(\mathbf{h_m}(a_m^l) + \sum_{n \neq m}^{M} \mathbf{J_{mn}}(a_m^l, a_n^l)\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{h_m}(k) + \sum_{n \neq m}^{M} \mathbf{J_{mn}}(k, a_n^l)\right)}$$

By using the definition of $\mathbf{h_m}$ and $\mathbf{J_{m,n}}$ the equation becomes:

$$P(a_m = a_m^l | \mathbf{a}_{n\neq m} = \mathbf{a}_{n\neq m}^l) = \frac{\exp\left(\mathbf{h_m}(a_m^l) + \sum_{n\neq m}^{M} \mathbf{J_{mn}}(a_m^l, a_n^l)\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{h_m}(k) + \sum_{n\neq m}^{M} \mathbf{J_{mn}}(k, a_n^l)\right)} \tag{3}$$

Using the representation in one hot shot encoder for the MSA the probability from Eq.(2) is now $P(a_{m,k})$ to have the amino acid k at position m and is given by :

$$P(a_{m,k}) \approx \prod_{l=1}^{L} P(a_{m,k} = a_{m,k}^l | \mathbf{a}_{n\neq m,k} = \mathbf{a}_{n\neq m,k}^l)$$

and by using the expression of $\mathbf{h_m}$ and $\mathbf{J_{mn}}$ the Eq.(3) becomes:

$$P(a_{m,k} = a_{m,k}^l | \mathbf{a}_{n\neq m,k} = \mathbf{a}_{n\neq m,k}^l) = \frac{\exp\left(\sum_{k=1} H_{m,k} a_{m,k}^l + \sum_{n\neq m,p} C_{mk,np} a_{m,k}^l a_{n,p}^l\right)}{\sum_{k'=1} \exp\left(H_{m,k'} + \sum_{n\neq m,p} C_{m,k',n\neq m,p} a_{n,p}^l\right)} \tag{4}$$

2.4 Cross entropy

Using Eq.(4) and the condition given by Eq.(1), the negative log likelihood becomes:

$$\mathcal{L} = -\frac{1}{M}\sum_{m=1}^{M} \log P(a_{m,k}) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L} \log\left[\frac{\exp\left(\sum_{k=1} a_{m,k}^l \left[H_{m,k} + \sum_{n\neq m,p} C_{mk,np} a_{n,p}^l\right]\right)}{\sum_{k'=1} \exp\left(H_{m,k'} + \sum_{n\neq m,p} C_{m,k',n\neq m,p} a_{n,p}^l\right)}\right]$$

$$= -\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L} \log\left[\left[\frac{\exp\left(H_{m,k} + \sum_{n\neq m,p} C_{mk,np} a_{n,p}^l\right)}{\sum_{k'=1} \exp\left(H_{m,k'} + \sum_{n\neq m,p} C_{m,k',n\neq m,p} a_{n,p}^l\right)}\right]^{\sum_{k=1} a_{m,k}^l}\right]$$

$$\mathcal{L} = -\frac{1}{M}\sum_{l,m,k} a_{m,k}^l \log\left[\mathbf{Softmax}(H_{m,k'} + \sum_{n\neq m,p} C_{m,k',n\neq m,p} a_{n,p}^l)\right] \tag{5}$$

3 Computation concepts

3.1 Architecture of the model

From Eq.(5) the architecture of the model is deduced as a linear model trained with Softmax activation and cross entropy loss such that the input and labels are the same elements $\left(a_{m,k}^l\right)$. The architecture is composed of L different classifiers that have to predict the value of an amino acid as function of all the others amino acids.

As illustrated in Fig.3 the K input neurons of each amino acid is fully connected to the K output neurons of each others amino acid (the link between the same amino acid position should be masked). The interaction coefficients are the weights learned by the network [9]. Indeed by Taylor expansion one output $z_{\alpha,\beta}((a_{m,k}))$ amino acid is given by (before Softmax activation):

$$z_{\alpha,\beta}((a_{m,k})) = W^{(1)}_{\alpha,\beta} + \sum_{m \neq n} W^{(2)}_{\alpha\beta,mk} a_{m,k} + ... \quad (6)$$

with the n-th term as the n-th order interactions. With the linear model, n=2 and by comparison with Eq.(5), the following terms are deduced: $W^{(1)}_{\alpha,\beta} = H_{\alpha,\beta}$ and $W^{(2)}_{\alpha\beta,mk} = C_{\alpha\beta,mk}$

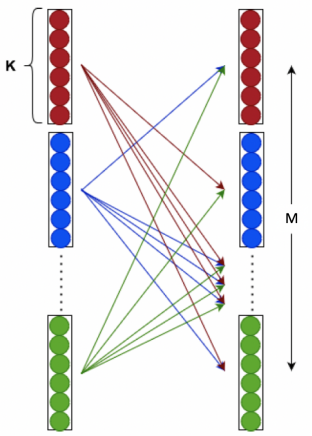


Figure 3: Linear model for a sequence with M amino acids of K possibles values [9]

3.2 Sequence reweighting

Let's remind the assumption that the MSA was a sample of a Boltzman distribution such that each sample were statistically independent. This is not the case with the data since some sequences could be more similar to others due to evolution. The inequality of similitude can be corrected by sequence re-weighting i.e by attributed lower weights for sequences more common and higher weights for sequences more rare. The similarity between two sequences can be calculated as $\text{similarity}(A^l, A^{l'}) = \sum_m \delta\{Xm^l = Xm^{l'}\}$. Let's fix a threshold of similarity S, then the weight of a sequence l is given by:

$$\omega_l = \frac{1}{\sum_{l'} \delta\{\text{similarity}(A^l, A^{l'}) > S\}}$$

And the loss from Eq.(5) becomes:

$$\mathcal{L} = -\frac{1}{M} \sum_{l,m,k} \frac{1}{\omega_l} a^l_{m,k} \log \left[\textbf{Softmax}(H_{m,k'} + \sum_{n \neq m,p} C_{m,k',n\neq m,p} a^l_{n,p}) \right] \quad (7)$$

3.3 Couplings correction

The number of parameters to learn the model is given by (a) the numbers of terms in a sequence in addition to (b) the numbers of terms given by the couplings. By considering a sequence of size M with K possibilities per amino acid, (a) is equivalent to $M \cdot K$.

Considering the pairs $(i,j) \quad \forall i \neq j$, the number of ways one can choose two elements from a sequence of size N without replacement or ordering is given by $\frac{M(M-1)}{2}$ [14]. By considering the coupling matrix size K^2, (b) is equivalent to $\frac{M(M-1)}{2} \cdot K^2$. Thus the total number of parameters is given by $M \cdot K + \frac{M(M-1)}{2} \cdot K^2$ However, with the condition from Eq.(1), the total number of parameters should be $M \cdot (K-1) + \frac{M(M-1)}{2} \cdot (K-1)^2$ which implies a model overparametred and a model that cannot converge to a solution (different parameter set can describe the same probability distribution [14]). To fix this issue, the Ising gauge is applied on the couplings such that:

$$C^{\text{New}}_{mk,np} = C_{mk,np} - \frac{1}{K}\sum_{k'}^{K} C_{mk',np} - \frac{1}{K}\sum_{p'}^{K} C_{mk,np'} + \frac{1}{K^2}\sum_{k',p'}^{K} C_{mk',np'} \tag{8}$$

This implies that the average over k or over p is null. Indeed for example:

$$\begin{aligned}\sum_{k}^{K} C^{\text{New}}_{mk,np} &= \sum_{k}^{K} C_{mk,np} - K\frac{1}{K}\sum_{k'}^{K} C_{mk',np} - \frac{1}{K}\sum_{k,p'}^{K} C_{mk,np'} + K\frac{1}{K^2}\sum_{k',p'}^{K} C_{mk',np'} \\ &= \sum_{k}^{K} C_{mk,np} - \sum_{k'}^{K} C_{mk',np} - \frac{1}{K}\sum_{k,p'}^{K} C_{mk,np'} + \frac{1}{K}\sum_{k',p'}^{K} C_{mk',np'} = 0\end{aligned}$$

3.4 The Frobenius norm

To assess the topology a last treatment to the couplings needs to be made with Frobenius norm [15]. This scalar quantity measures the coupling strength between two positions:

$$F_{m,n} = \sqrt{\sum_{k,p} C_{m,n}(k,p)^2}$$

Finaly, the contact predictions should improve with this average correction [15]:

$$F^{\text{new}}_{m,n} = F_{m,n} - \frac{\sum_r F_{m,r} \sum_r F_{r,n}}{\sum_{r,s} F_{r,s}}$$

This last equation should penalize correlations coming from phylogeny [16].

4 Modifications in comparison to previous algorithms

New functionality have been added as the choice of similitude percentage S or the choice of gap numbers accepted per sequence. Additionally, it is possible to precise which seed number is used and which optimizer (with which parameters) from PyTorch the model will use. Moreover 6, no negligible, new points are implemented:

- **New formulation for the couplings correction with Ising Gauge (Eq.(8)):** The previous one was incorrect and missed the third term.
- **Variable number of amino acids values per position:**
  For each possible position, if an amino acid $k \in [1, K]$ is never present through every sequences, all its link with others amino acids will be masked. Additionally, this amino acid will not be considered during the summation of the Ising Gauge (Eq.(8)).
- **New batch size** Previous experiments were done with a 64 batch size. However, smaller batch size like 32 should decrease the probability of over-fitting.
- **Different optimizers:**
  Since only 2 optimizers from PyTorch have been previously tested (Adam and AdamW), new ones are compared with (Adadelta, Adagrad and SDG). For each optimizer, a hyperparameters tunning is realized. Then the validation loss and contact predictions map are compared between each optimizers with the best set of parameters.
- **Three different average types from several models:**
  A new functionality allows to compute n models with the same set of parameters (except the seed number). Each model will have a different data distribution for the train, validation and test. Then the user can choose between 3 different average types: (1) An *average model* with the weights averaged from the n models. The couplings and Frobenius norm are then computed from this average model. (2) An *average couplings* from the n couplings gotten from the n models. The Frobenius norm is then computed from this average couplings. (3) An *average Frobenius* from the n Frobenius norms gotten from the n couplings.
- **Consideration of the sequences class**
  A new functionality allow the user to select the mode *with class*. In this case a new column is added to the MSA with new amino acids before to train the model on it (the one hot encoder vectors size becomes K+8):
  eukaryotes: 21:Viridiplantae, 22:Fungi, 23: Metazoa, 24:Other
  Bacteria: 25:Alphaproteobacteria, 26:Gammaproteobacteria,27:Firmicutes, 28:Other
  This column is then removed for the couplings extraction.

5 Results

5.1 Optimizer comparisons

Hyperparameters tunning have been done with each different optimizers (only with one model). Three different learning rate l_r are tested 0.01,0.005 and 0.008 with the protein SIS1, L=352, K=21, batch size=32, epochs= 50 and seed number=203. Note also that a variable number of amino acids per position is applied. The validation loss are on Fig.4 and the corresponding contact predictions plot are on Fig.5 (AdamW was not add since its loss was identical to Adam).

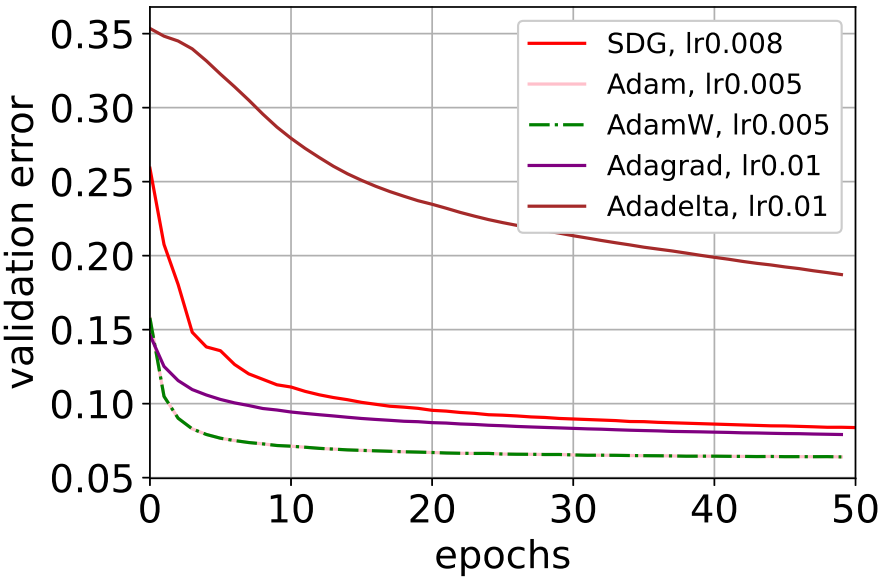
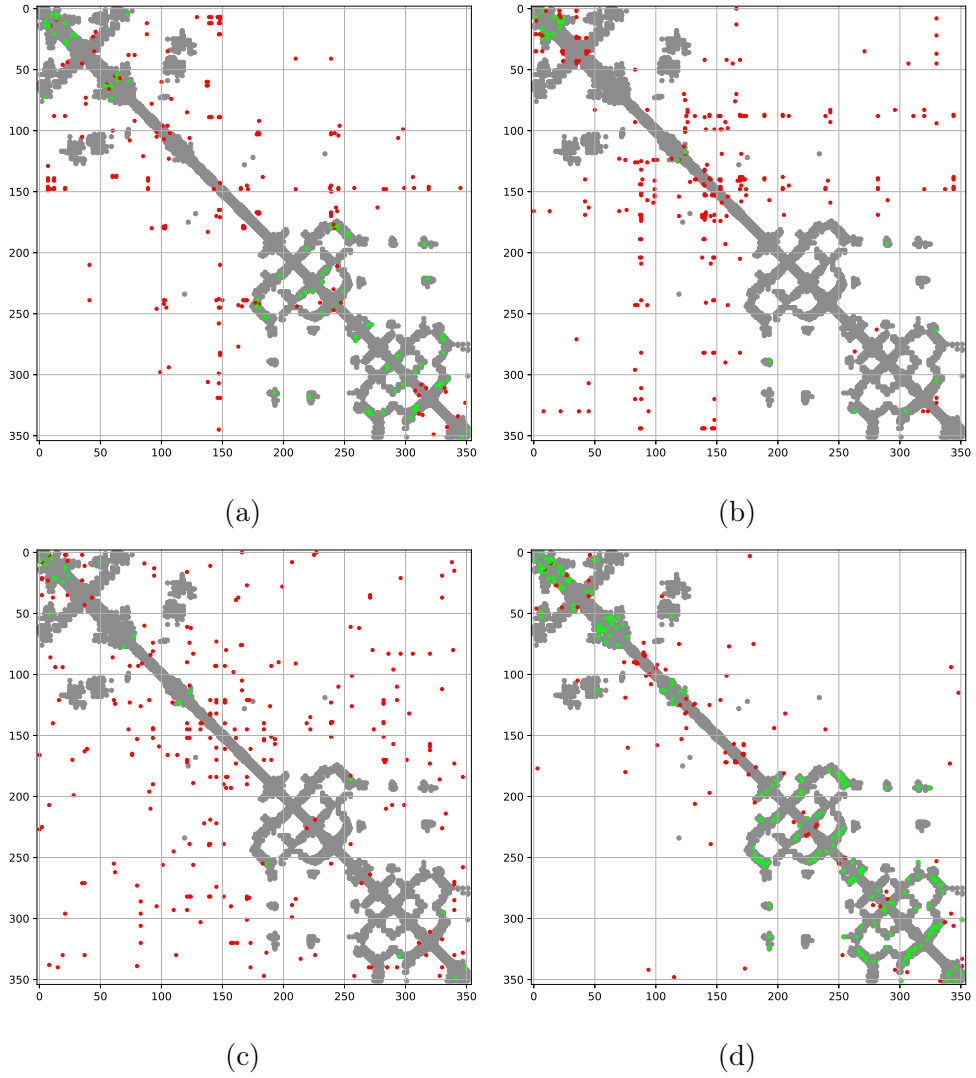


Figure 4: Validations errors for different optimizers



(a) (b)

(c) (d)

Figure 5: 150 Prediction contact of SIS1 with (a) Adadelta, lr=0.01, (b) Adagrad, lr=0.01, (c) Adam, lr=0.005 and (d) SDG, lr=0.008

5.2 Different average and couplings extractions

Four different couplings extractions are plotted (Fig.9) on three different data (MAS5, SIS1 and the JDP) with ten different model weights. Fig.6a shows lower predictions contact rates than the others. Take several models and not only one is optimal for each different data. For the JDP (Fig.6a), between $N_{pred} = 60\%L$ and $N_{pred} = 100\%L$, the precision decreases with a slope of ~ -30 with one model and ~ -15 with others cases (two times less).

For the others proteins, SIS1 and MAS5, the precision decreases almost "linearly" with Npred/L. Approximations give respectively slopes of $\sim$ -35 and -45 with 1model, $\sim$ -30 and -40, with other cases.

The three different average types (Fig. 6b, 6c, 6d) show quite similar predictions rates. However the average at the end, with the frobenius norm, show smoother curves especially for the SIS1 protein.

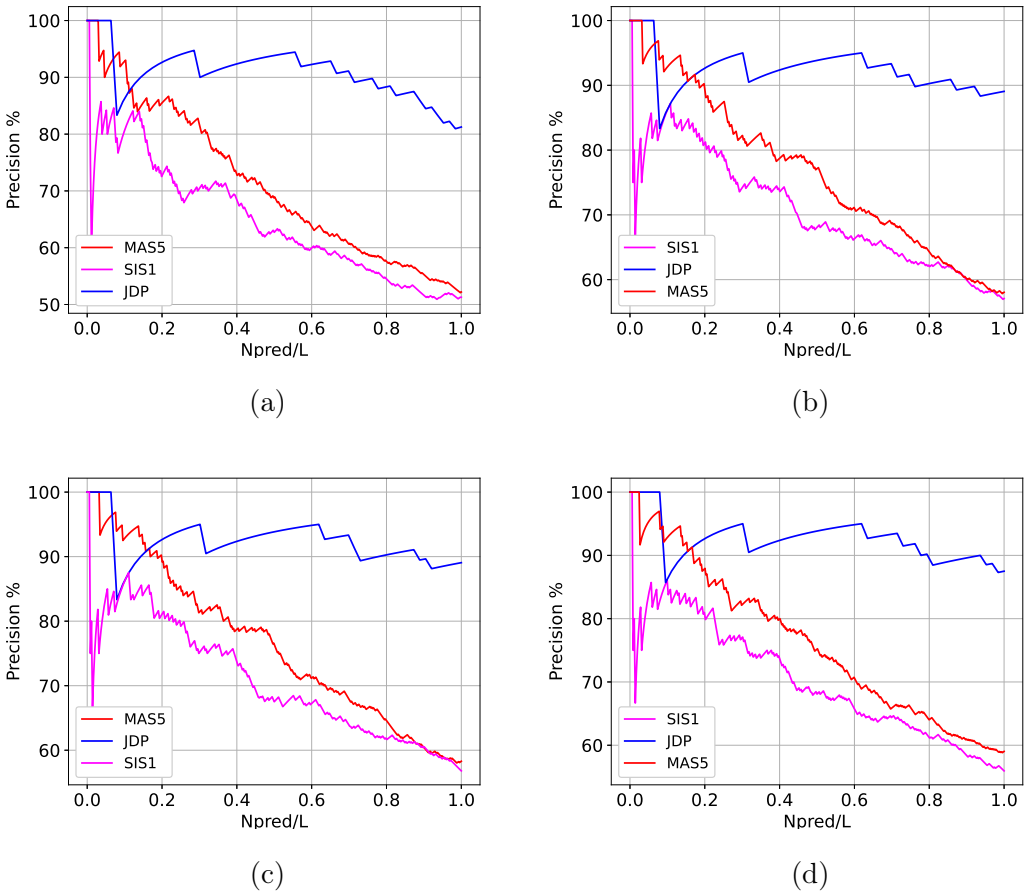


Figure 6: Rates with SDG, lr=0.008 and (a) only 1 model, or (b) average with 10 models, (c) average with 10 couplings, (d) average with 10 frobenius norms

5.3 Prediction with the class of the sequences

The accuracy as function of the number of models with class information or without are reported on Fig.7. The results show that 8 models are enough for the both cases. The case *with class* shows a continue increasing curve until reaching a plateau at 90%. The case without class shows a no-monotonic curve but that in general increases with oscillation around 87%. The contact predictions with 20 models are illustrated on Fig.8. The 3D structure from AlphaFold and the incorrect predictions labeled are illustrated on Fig.9.

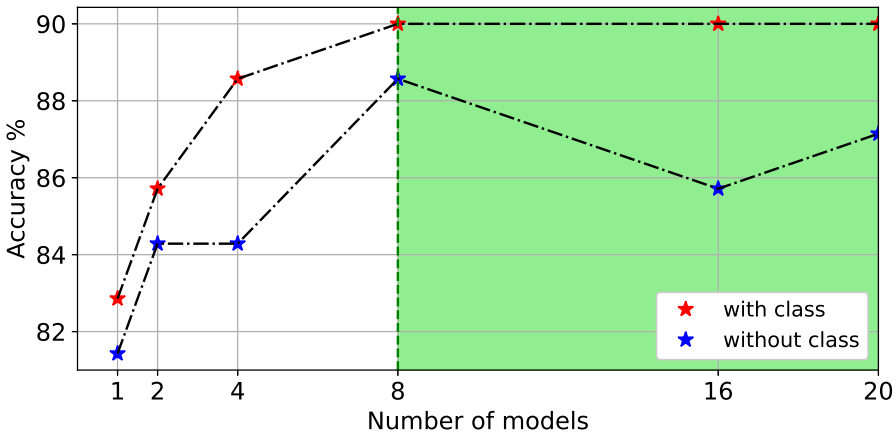


Figure 7: Accuracy for 70 contact predictions of the JDP with average on Frobenius norms and different number of models (a)

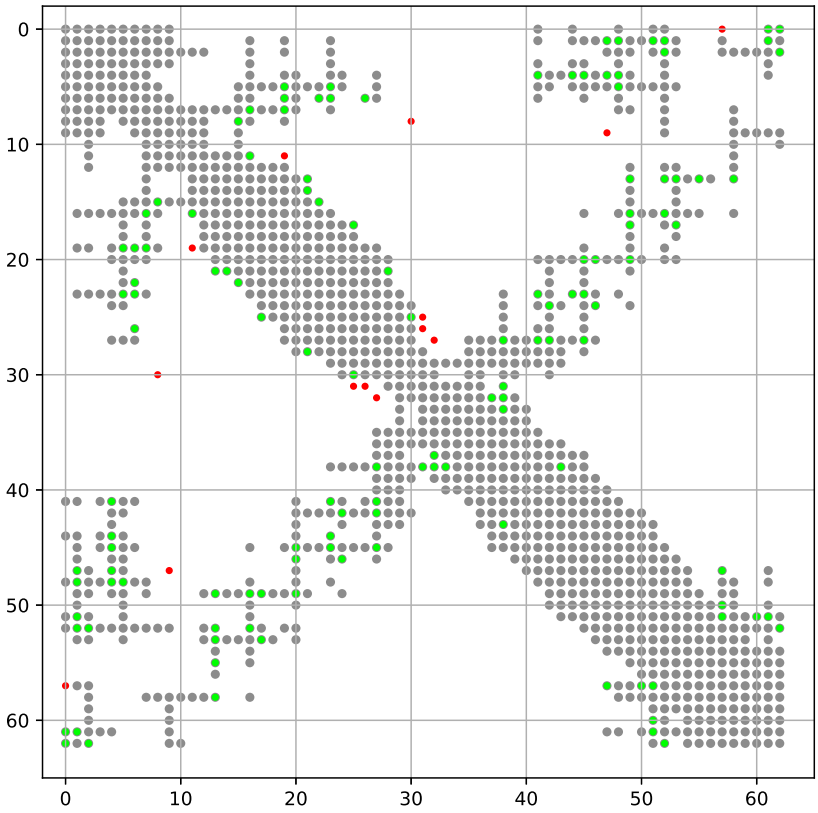


Figure 8: JDP, 20models and Frobenius norms average, with class, 70 contact predictions in 2D, 8.5Å

The wrong predictions from Fig.8 concern only predictions with lower confidence (but still high) except for the position 0 with the position 57. Additionally, it is always thin filaments of the JDP that are wrongly connected to a thicker part of the structure or

another thin element on Fig.9. They are never two thick parts of the helix presumed wrongly in contact. The smallest error concerns a distance of 9.32Å (in 3D) between positions 24 and 31, the larger error concerns a distance of 28.96Å (in 3D) between positions 8 and 30.

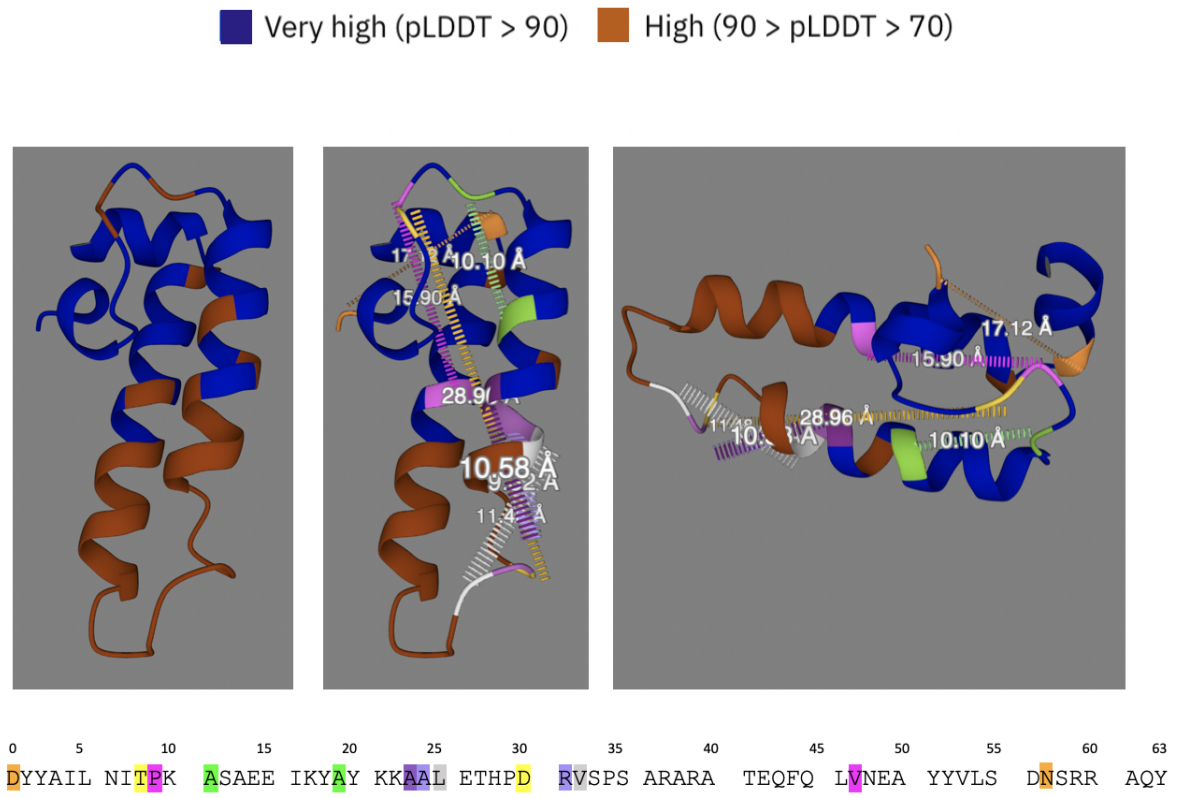
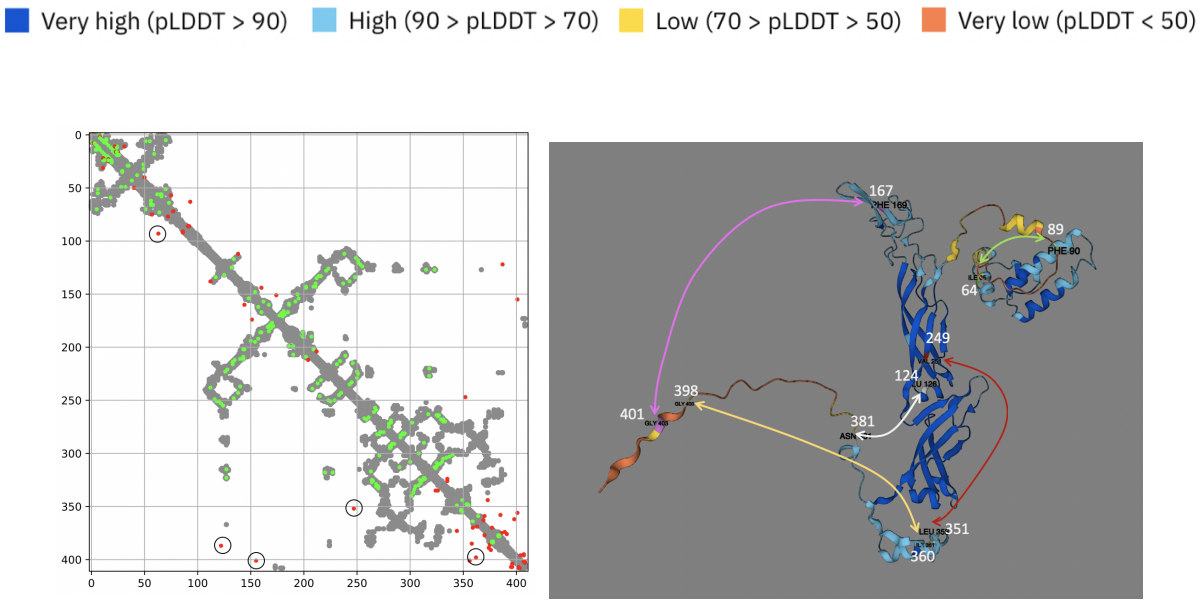


Figure 9: JDP, 20models and Frobenius norms average, with class, 70 contact predictions, 8.5Å, comparison with AlphaFold in 3D. The labeled data correspond to the wrong predictions on Fig.8

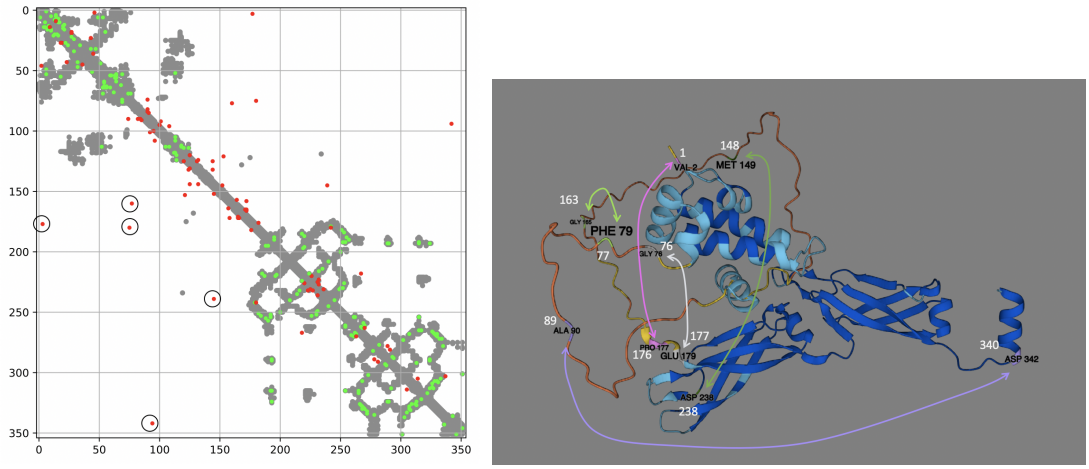
5.4 2D contact predictions for the different proteins

The different 2D contact predictions for the two different proteins MAS5 and SIS1 are illustrated respectively on Fig.10a and Fig.11a. Additionally, the 3D structure from AlphaFold are illustrated (see Fig.10b and Fig.11b). The most wrong contact (the most distant from the good cartoon, see the circles) concern, as for the JDP, thin filaments, with low predictions, with others elements of the structure.



(a) (b)

Figure 10: 170 Contact predictions for MAS5, average on 10 Frobenius norms, without class (a) 2D, (b) comparison with AlphaFold. The circles on (a) correspond to the label points on (b)



(a) (b)

Figure 11: 150 Contact predictions for SIS1, average on 10 Frobenius norms, without class (a) 2D, (b) comparison with AlphaFold. The circles on (a) correspond to the label points on (b)

6 Discussion

Observations from Fig.5 lead to the conclusion that the only effective optimizer for 50 epochs is SGD. The qualities of Adam or AdamW, known for their speed and computational low cost, may pose a problem here (similar to Adagrad). In the first epoch, these optimizers exhibit only a 15% validation error and reach a plateau after a few epochs. Such behavior could be the cause of overfitting in opposition to the slower convergence of SDG and Adadelta (see Fig.4). Note that Adadelta is very slow to converge and 50 epochs may not be sufficient. However, Fig.5a shows fewer errors than Fig.5b and Fig.5c. It's possible that this optimizer could be a viable solution with a larger number of epochs. Using multiple models positively influences the noise, as evidenced by a progression with slower decrease in precision (Fig.9). Although, the difference between the three average types doesn't impact the precision decrease, averaging the Frobenius norms affects the variability of the results, etablishing a more linear relationship between precision percentage and Npred/L, thus reducing noise. Integrating class information with JDP (Fig.7), is more optimal with monotonically increasing accuracy curves as function of the number of models until a plateau at 90%. Adding this information to the model learning prompt it to predict the class of the sequence, encouraging the identification of similarities between sequences of the same class, optimizing the solution. Wrong predictions often involve thin filaments that may have different orientation among homologous sequence. Moreover, these filaments, for SIS1 and MAS5, exhibit low confidence from AlphaFold. Future results incorporating also class information for sequences of these two proteins could yield to higher accuracy, making this investigation essential.

7 Conclusion

The utilization of a direct coupling analysis, aided by a multiple sequence alignment of various homologous sequences, enables the construction of a linear model trained with Softmax activation and cross entropy loss for amino acid prediction. Thank's to statistical concept and machine learning, the 2D contact predictions and their interpretation with the 3D structure from AlphaFold have been accomplished for three different datasets: a J-domain Protein, the protein SIS1 (gene SIS1) and the protein MAS5 (gene YDJ1). Training several models with different data partition for the train, validation and test sets

reveals an improvement in precision percentage. Additionally, providing the model with the class information of the sequence results in monotonically increasing improvement as a function of the numbers of the models before reaching a plateau at 90% correct predictions for 70 requested predictions for J-domain contact (63 sequence length). Furthermore, AlphaFold's 3D structure reveals that most of the incorrect predictions concern thin filaments that could have a different orientation between homologous sequences. Further analysis involve applying these class information for SIS1 and MAS5. The next code implementation would be to explore higher order of interactions, such as three, which can be achieved by implementing non linear or mix (linear + non linear) models.

References

[1] Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. *Protein Folding and Processing.* Available from: https://www.ncbi.nlm.nih.gov/books/NBK9843/

[2] Michael B. Evgen'ev, David G. Garbuz, Olga G. Zatsepina, 2014, *Heat shock proteins and whole body adaptation to extreme environments. Springer.*

[3] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia Mikhail Borisovich Evgen'ev, 2021, *Heat shock proteins: a history of study in Russia.*

[4] Mayer MP, Bukau B., 2005, *Hsp70 chaperones: cellular functions and molecular mechanism*

[5] M B B Gutierres, C B C Bonorino, M M Rigo, *ChaperISM: improved chaperone binding prediction using position-independent scoring matrices, Bioinformatics*, Volume 36, Issue 3, February 2020, Pages 735–741, https://doi.org/10.1093/bioinformatics/btz670

[6] Flaherty, K. M., DeLuca-Flaherty, C., and McKay, D. B., 1990, *Three- dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein.* Nature 346, 623– 628 CrossRef Medline

[7] Justas Dauparas and Haobo Wang and Avi Swartz and Peter Koo and Mor Nitzan and Sergey Ovchinnikov, 2019, *Unified framework for modeling multivariate distributions in biological sequences*

[8] AlphaFold Protein Structure Database, Available from: https://AlphaFold.ebi.ac.uk

[9] Aude Maier, 2022, *Direct Coupling Analysis for protein contact prediction using neural networks*

[10] Lisa Gennai, 2023, *TPIV Report - Laboratory of Statistical Biophysics*

[11] Figliuzzi M, Barrat-Charlaix P, Weigt M., 2018, *How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? Mol Biol Evol.*

[12] UPPSALA UNIVERSITET, AliView, Available from: https://ormbunkar.se/aliview/

[13] Reports on Progress in Physics, Simona Cocco and Christoph Feinauer and Matteo Figliuzzi and Rémi Monasson and Martin Weigt, 2018, *Inverse statistical physics of protein sequences: a key issues review*

[14] Royal Institute Of Technology, Magnus Ekeberg, 2012, *Detecting contacts in protein folds by solving the inverse Potts problem – a pseudolikelihood approach*

[15] Reports progress on Physics, Simona Cocco1, Christoph Feinauer, Matteo Figliuzzi2, Rémi Monasson and Martin Weigt, 2018, *Inverse statistical physics of protein sequences: a key issues review*

[16] Nicola Dietler, Umberto Lupo, Anne-Florence Bitbol, *Impact of phylogeny on structural contact inference from protein sequence data*, https://doi.org/10.1101/2022.09.26.509588 Now published in Journal of The Royal Society Interface doi: 10.1098/rsif.2022.0707