

# On conformational changes of proteins using collective motions in torsion angles and $L_1$ regularization

*Ryabinina R. B.<sup>1</sup>, Emtsev D. I.<sup>2</sup>*

<sup>1</sup> ryabinina.rb@phystech.edu <sup>2</sup> daniil.emcev.ru@yandex.ru

<sup>1,2</sup>MIPT

Investigation of conformational transitions in proteins is an important and well-studied problem in structural bioinformatics with applications ranged from drug design to understanding hidden effects in the structural data. A related important and open question in computational structural bioinformatics is how to efficiently represent transitions between protein structures. Here, we address the problem of how a sparse subset of collective coordinates in the torsion subspace can describe functional conformational changes in proteins. The solution strategy consists in determining the change of torsion angles through the fit of the linearized change of Cartesian coordinates. However, if the fit is not regularized, the structures produced in this approach demonstrate the deviation of several Angstroms from the targets. Rescaled ridge regression (RRR) has been recently introduced to regularize multi-dimensional regressions with correlated explanatory variables. The resulting torsional conformational changes generate conformations that are much more similar to the target conformations. This approach also predicts atomic thermal fluctuations that are better correlated with the ones measured experimentally. Our goal is to find a solution of a regression problem with an  $L_1$  regularization constraint using the LASSO formulation. Not much has been done in the torsional angle subspace (internal coordinates) for this problem and nearly nothing has been done using  $L_1$  regularization.

## 1 Introduction

There is growing interest in the investigation of the intrinsic dynamic properties of proteins in their native state. They play a key role in ensuring proper functional activity, notably for catalysis, allosteric regulation or molecular recognition. Despite recent progress, the experimental studies of protein dynamics remain rather challenging, and computational methods often constitute valuable alternatives. It is often assumed that torsion angles are the natural degrees of freedom for describing protein motions [1], since bond lengths and bond angles are strongly constrained by covalent forces. Because of this reason, several computational methods have been developed to study protein dynamics in torsion angle space. There is a desperate need of systems to predict dynamic motions. Elastic network models (ENMs) [2–4] are becoming increasingly popular since they provide detailed analytic predictions of native protein dynamics at a very reasonable computational cost.

Ridge regression is one of the most common methods for regularizing fits with many variables. It relies heavily on the choice of an adequate value for the ridge (regularization) parameter. Its optimal value is generally unknown, and several criteria have been proposed for its determination. For example, the cross-validation technique is the most popular choice for it.

Inspired by successful use of the least absolute shrinkage and selection operator, we propose  $L_1$  formulations with the direction vectors reconstructed from the internal coordinates. Here we are using as a base the TNM model [1], an elastic network model whose degrees of freedom are the torsion angles of the protein backbone. So, our aim here is construct a novel approach with the  $L_1$  regularizations and compare results with the previous methods.

It is worth noting immediately that compared to the classical variable selection methods, such as subset selection, the LASSO has two advantages. First, the selection process in the LASSO is based on continuous trajectories of regression coefficients as functions of the penalty level and hence it is more stable than the subset selection methods. Second, the using of least absolute shrinkage and selection operator is computationally feasible for high-dimensional data [5–7] .

## 2 Conformational changes and normal modes in torsion angle space

We are investigating here computational models, describing the conformational changes, observed in the data bank of proteins. There were two different structures of one protein, resolved with various functional conformations. It is shown that these large functional conformational changes correlate well with the thermal dynamics predicted by the ENM [8–10]. In our work we raise the question of how well we can estimate conformational changes are described using only torsion angles by means of least absolute shrinkage and selection operator.

Further, we consider how can the change in torsion angles be determined from the observed changes of Cartesian coordinates and introduce you to the notation. First of all, here we have Cartesian coordinates of the two conformations A and B and use the symbols  $r_i^A$  and  $r_i^B$ , where  $i$  corresponds to the atom and  $r^A, r^B$  are a three-dimensional vectors. Moreover, it is important, that studied systems consist accurately of the same atoms when we compare the empirical conformational changes with the predicted through the TNM. We should not forget it, so that uninteresting rigid body degrees of freedom are defined in the same way in the empirical system and in TNM normal modes.

Here we look at  $L_1$  regularization method for finding changes in torsion angles from the observed coordinates of the two conformations

$$\Delta r = J\Delta\varphi$$

Here,  $\Delta r$  vector consists of  $3n$  coordinates, where  $n$  is the number of atoms from the extended backbone implementation of the TNM.  $\Delta\varphi$  has dimension of normal modes and  $J$  is the Jacobian matrix containing  $\frac{\Delta r_{int_i}}{\Delta\varphi_a}$ . Also, to receive formulation of problem primarily we consider ordinary least square regression

$$\min_{\Delta\varphi} (\Delta\varphi, J^\top M J \Delta\varphi) - 2(\Delta\varphi, J^\top M \Delta r).$$

That has this form due to  $J^\top M \Delta r = J^\top M J \Delta\varphi$  reformation. This method is inapplicable with correlated explanatory variables which is the components of the Jacobian matrix because it tends to lead to overfitting problems. Thus, we are considering  $L_1$  regularization approach allowing get rid of valueless features.

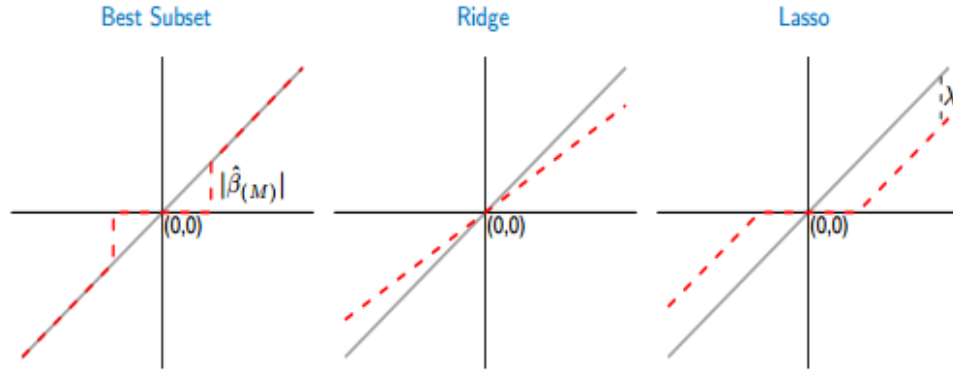
## 3 Overview of $L_1$ regularization methods of obtaining conformational changes

There are a lot of contemporary methods to work with  $L_1$  regularization. We look through several of them and present our method with best results

### 3.1 Implemented method

Least absolute shrinkage and selection operator with coordinate descent

$$\frac{1}{2n} \|\Delta r - J\Delta\varphi\|_2^2 + \alpha \|\Delta\varphi\|_1$$



**Рис. 1** Estimated coefficients at coordinate descent in Lasso

where  $n$  - number of samples,  $\Delta r$  - vector of displacements in protein structure,  $\Delta\varphi$  - torsional angle of rotated protein,  $\alpha$  - optimization hyperparameter

### 3.2 Least absolute shrinkage and selection operator

Firstly we consider simple least absolute shrinkage and selection operator which is more preferable comparing to ridge, because instead of just shrinking coefficients in ridge regression, we reduce the dimension of  $\Delta\varphi$ , selecting best-working features. We consider the problem as

$$\begin{aligned} \min_{\Delta\varphi} & (\Delta\varphi, J^T M J \Delta\varphi) - 2(\Delta\varphi, J^T M \Delta r) \\ \text{s.t.} & \sum_{j=1}^p |\Delta\varphi_j| \leq s \end{aligned}$$

That allows us use the quadratic programming approach of solving our optimization problem. For transformation constraints from  $\sum_{j=1}^p |\Delta\varphi_j| \leq s$  to  $A\Delta\varphi \leq s$  we do the following substitution

$$\Delta\varphi_i = \Delta\varphi_i^+ - \Delta\varphi_i^-, \text{ where } \Delta\varphi_i^{+,-} \geq 0 \text{ and } \Delta\varphi_i^+ - \Delta\varphi_i^- \geq 0 \quad \forall i \in 1 \dots p$$

Also, we denote by  $Q = J^T M J$  and  $c = -J^T M \Delta r$ , so there is a new formulation which could be transformed to the problem of quadratic programming with dimension of  $2p$

$$\begin{aligned} \min_{\Delta\varphi^+, \Delta\varphi^-} & \frac{1}{2} (\Delta\varphi_i^+ - \Delta\varphi_i^-)^T Q (\Delta\varphi_i^+ - \Delta\varphi_i^-) + c^T (\Delta\varphi_i^+ - \Delta\varphi_i^-) \\ \text{s.t.} & \Delta\varphi^+ + \Delta\varphi^- \leq s \\ & \Delta\varphi^{+,-} \geq 0 \end{aligned}$$

It can be rewritten in the matrix form as:

$$\min_{\Delta\varphi^+, \Delta\varphi^-} \frac{1}{2} \begin{bmatrix} \Delta\varphi_i^+ & \Delta\varphi_i^- \end{bmatrix} \begin{bmatrix} Q & -Q \\ -Q & Q \end{bmatrix} \begin{bmatrix} \Delta\varphi_i^+ \\ \Delta\varphi_i^- \end{bmatrix} + \begin{bmatrix} c^T & -c^T \end{bmatrix} \begin{bmatrix} \Delta\varphi_i^+ \\ \Delta\varphi_i^- \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} I_p & I_p \\ -I_{2p} & \end{bmatrix} \begin{bmatrix} \Delta\varphi_i^+ \\ \Delta\varphi_i^- \end{bmatrix} \leq \begin{bmatrix} s_p \\ 0_{2p} \end{bmatrix} \quad (2)$$

Where  $I_p$  is the  $p$ -dimensional unit matrix,  $s_p$  -  $p$ -dimensional vector, consisting only of the value  $s$ , and  $0_{2p}$  is a  $2p$ -dimensional zero vector

### 3.3 Elastic net regularization

Further, we analyze elastic-net regression in the application to finding conformational changes. Lets consider the next formulation with  $\alpha > 0$  and discuss disadvantages

$$\min_{\Delta\varphi} (\Delta\varphi, J^T M J \Delta\varphi) - 2(\Delta\varphi, J^T M \Delta r) + \alpha(\Delta\varphi, \Delta\varphi) + (1 - \alpha) \sum_{j=1}^p |\Delta\varphi_j|$$

The elastic net model is able to select groups of variables when they are highly correlated. It doesn't have the problem of selecting more than  $n$  predictors when  $p \gg n$ , but LASSO saturates in this situation. When there are highly correlated predictors, the LASSO tends to just pick one predictor out of the group. These disadvantages of LASSO could be overcome by the elastic net. But in our problem, it is not necessary, because we have  $p \ll n$  and the real disadvantage is the computational cost. That will be due to the inevitable cross-validation of the relative weight of  $L_1$  vs.  $L_2$  penalty,  $\alpha$ , and that increases the computational cost by the number of values in the  $\alpha$  grid. Another drawback, which can be also an advantage, is the flexibility of the estimator. Increased probability of overfitting comes with greater flexibility. It may be that the optimal  $\alpha$  for the population and for the given sample size is 0, turning the elastic net into the lasso, but it is possible to choose a different value due to chance [11].

### 3.4 Least angle regression

In addition, we review least angle regression as it produces a full piecewise linear solution path, which simplifies cross-validation. The fact is that in the following there will be discussed the choice of the regularization parameter for our first model. Despite of this point the method has considerable disadvantages. With any amount of noise in the dependent variable and with high-dimensional multicollinear independent variables, there is no reason to believe that the selected variables will have a high probability of being the actual underlying causal variables. However, since LARS relies on an iterative update of residuals, it appears to be particularly sensitive to the effects of noise [12]. Furthermore, it is in our case that we have a lot of this, so we do not use LARS as the main strategy and as the solution of finding regularization parameter.

## 4 Method

We use this formulation of Lasso from the scikit-learn v0.20.1 library. The method that showed best results and most adequate answer is about coordinate descent. The formulation as usual like Lasso. We use coordinate descent.

$$\frac{1}{2n} \|\Delta r - J \Delta\varphi\|_2^2 + \alpha \|\Delta\varphi\|_1$$

We split our dataset into 10 parts (10-fold cross validation) We also used grid search automatically tuning hyperparameters on a parameter grid to choose the parameters with maximum cross-validation score. We set initially grid with  $\alpha \in [10^{-5}, 10^{-3}]$  with 30 values. It works considerably slower than just Lasso using cross validation.

## 5 Implementation

We use LassoCV and GreedySearchCV models for predicting  $\Delta r$ .  $\Delta r$  is our target. We take 30 pairs of initial dataset and for each pair calculate RMSD (initial, final) (it is given in

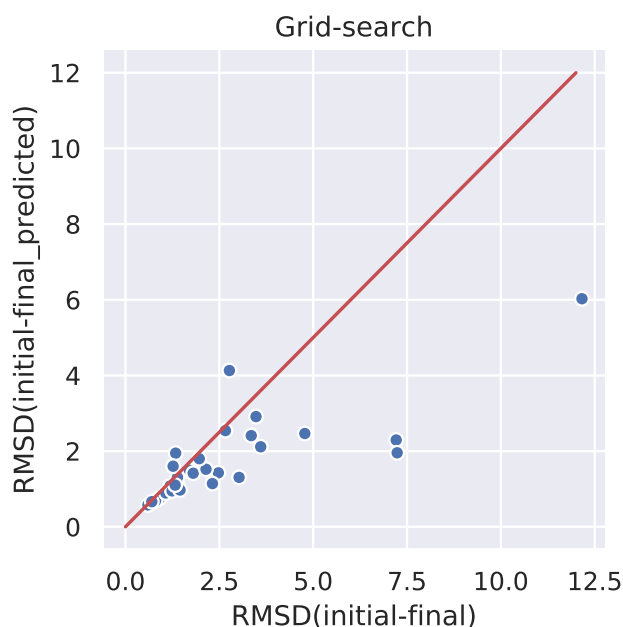
existed code) and  $\text{RMSD}(\text{initial}, \text{predicted})$ . Then we plot it on the one table to see correlation between two RMSD. The lower RMSD of predicted conformation and final conformation the more accurate the prediction so there should be strong correlation between this RMSD and the RMSD between the initial and final conformation. The larger the conformation change, the worse the reconstructed conformation. We tried to catch the pattern of correlation plotting results in two dimensional space ( $\text{RMSD}_1$ ,  $\text{RMSD}_2$ ) and checking their location relatively bisectrix. Also we estimated correlation of RMSD using linear approximation of data. We compared results of Lasso with cross validation, Lasso with grid-search cross-validation, Elastic net with cross-validation and Ridge regression.

## 6 Results

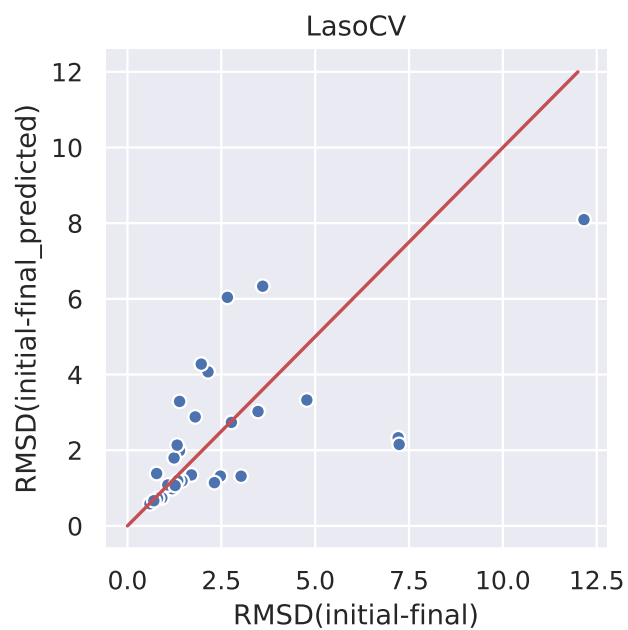
Finally, we analyze the properties of the torsional conformational changes obtained with Lasso based on grid-search and cross-validation showed reasonable results. We summarize the average RMSD between the reconstructed conformations and the target conformations for the six methods and showed, that least absolute shrinkage and selection operator model with cross-validation can work as good as Ridge regression.

LassoCV	EnCV	L=0	Ridge M	Ridge C	LassoCVGS
2.39	2.17	6.79	1.71	1.53	1.74

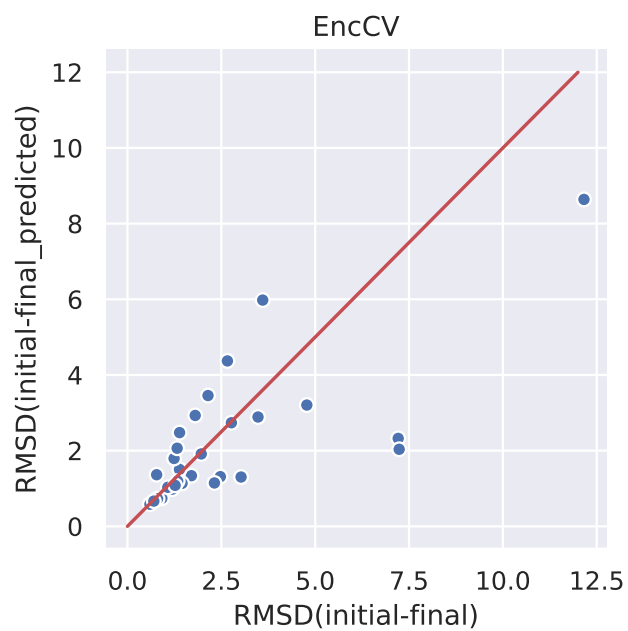
Two models: Grid search Lasso and Ridge regression type M yield accurate reconstructed structures, whose RMSD is on the average 1.7, i.e. the reconstructed conformations are approximately half the way between the initial and final conformational. As a result, sparse subset of collective coordinates in the torsion subspace can describe functional conformational changes in proteins.



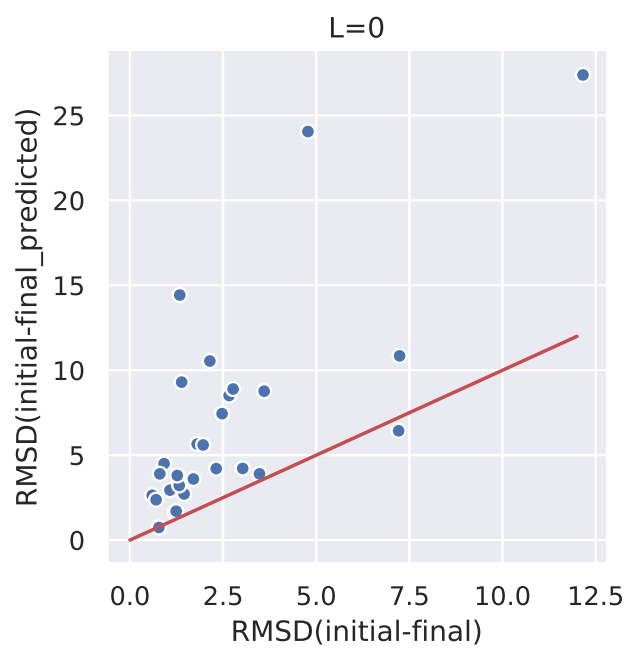
**Fig. 2** Correlation of RMSD in Grid search Lasso



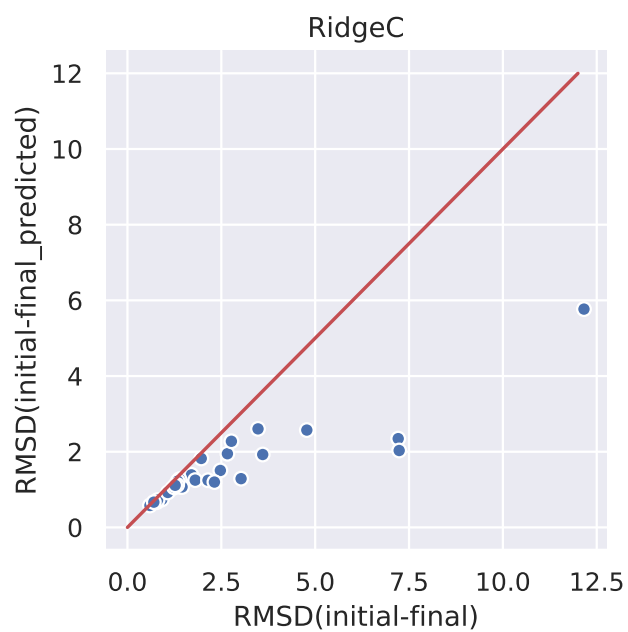
**Рис. 3** Correlation of RMSD in Lasso with cross validation



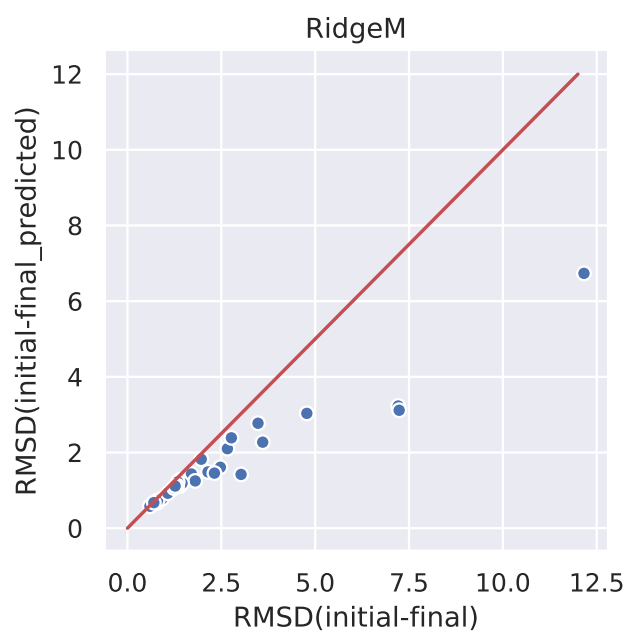
**Рис. 4** Correlation of RMSD in Elasticnet with cross validation



**Рис. 5** Correlation of RMSD in L=0 regression

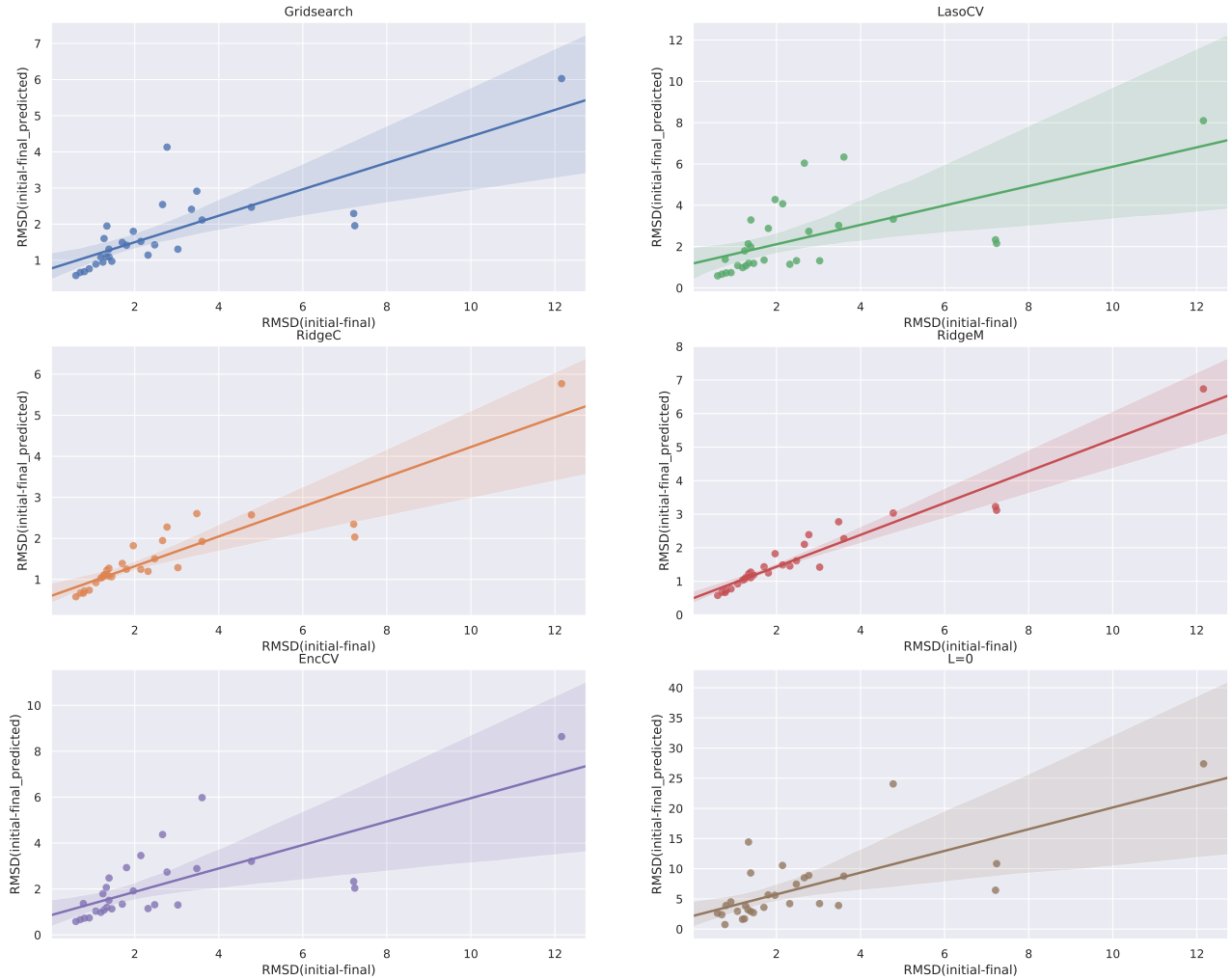


**Рис. 6** Correlation of RMSD in Ridge



**Рис. 7** Correlation of RMSD in RidgeM





**Рис. 8** Comparison of correlations of RMSD in all methods (using linear approximation).

Root mean square deviation between the conformation modelled applying the torsional differences on top of the initial conformation and the final conformation as a function of the RMSD between the initial and the final conformation for six alternative methods for obtaining the torsional differences.

## References

- [1] R. Mendez and U. Bastolla, “Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins.,” *Physical review letters*, vol. 104, p. 228103, Jun 2010.
- [2] M. M. Tirion, “Large amplitude elastic motions in proteins from a single-parameter, atomic analysis,” *Phys. Rev. Lett.*, vol. 77, pp. 1905–1908, Aug 1996.
- [3] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, “Anisotropy of fluctuation dynamics of proteins with an elastic network model,” *Biophysical Journal*, vol. 80, no. 1, pp. 505 – 515, 2001.
- [4] I. Bahar and A. Rader, “Coarse-grained normal mode analysis in structural biology,” *Current Opinion in Structural Biology*, vol. 15, no. 5, pp. 586 – 592, 2005. Carbohydrates and glycoconjugates/Biophysical methods.
- [5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [6] S. Katayama and S. Imori, “Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis,” *Journal of Multivariate Analysis*, vol. 132, pp. 138–150, 2014.
- [7] M. R. Osborne, B. Presnell, and B. A. Turlach, “On the lasso and its dual,” tech. rep., Feb. 11 1998.
- [8] F. Tama and Y. H. Sanejouand, “Conformational change of proteins arising from normal mode calculations.,” *Protein engineering*, vol. 14, pp. 1–6, Jan 2001.
- [9] L. Meireles, M. Gur, A. Bakan, and I. Bahar, “Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins.,” *Protein science : a publication of the Protein Society*, vol. 20, pp. 1645–58, Oct 2011.
- [10] H. G. Dos Santos, J. Klett, R. Méndez, and U. Bastolla, “Characterizing conformation changes in proteins through the torsional elastic response.,” *Biochimica et biophysica acta*, vol. 1834, pp. 836–46, May 2013.
- [11] H. Zou and T. Hastie, “Addendum: Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society Series B*, vol. 67, pp. 768–768, 2005.
- [12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.