On conformational changes of proteins using collective motions in torsion angle space and L1 regularization

Ryabinina R. B.[1]***, Emtsev D. I.***[2]
[1] ryabinina.rb@phystech.edu [2] daniil.emcev.ru@yandex.ru
[1 2]MIPT

Abstract

Investigation of conformations in proteins is an important and well-studied problem in bioinformatics with applications ranged from drug design to understanding hidden effects in the data. This is an important and open question in computational structural bioinformatics - how to efficiently represent transitions between protein structures. Here we address the problem of how sparse subset of collective coordinates in the torsion subspace can describe functional conformational changes in proteins. The following 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 way demonstrate the deviation of several Angstrom from 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, and they are better correlated with the thermal fluctuations of torsion angles and with the normal modes predicted by the TNM than the torsional conformational changes obtained through ordinary regression. Our goal is to find a solution of a ridge regression problem with an L1 regularization constraint using the LASSO formulation. Not much has been done in the torsional angle subspace (internal coordinates) and nearly nothing has been done using L1 regularization.

Introduction

There is growing interest in the investigation of the intrinsic dynamic properties of proteins in their native state, which play a key role in ensuring proper functional activity, notably for catalysis, allosteric regulation or molecular recognition. Despite recent progress, the experimental study of protein dynamics remains 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, 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) 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 regularising fits with many variables. It relies heavily on the choice of an adequate value for the ridge parameter but,also several criteria have been proposed, there is no solution how to systematically determine the optimal value of this parameter. There are two popular techniques, Tikhonov regularization and ridge regularization which deal with collinearity in multivariate regression. Inspired with successful use of Lasso method we proposed a LASSO formulation with the direction vectors reconstructed from the internal coordinates. We are using as a base model TNM. We received better computing using RMSD measure between the prediction and

the solution on several benchmarks
Our aim here is threefold: (1) devise a novel approach with Lasso regularization (2) apply this procedure for fitting the B-factors with predicted internal and rigid-bodymotions, in order to properly calibrate models of protein dynamics and to infer the respective amplitudes of the fluctuations dynamics and to infer the respective amplitudes of the fluctuations (3) compare results with previous methods
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 is hence more stable than subset selection methods. Second, the LASSO is computationally feasible for high-dimensional data [Osborne, Presnell and Turlach (2000a, 2000b), Efron et al. (2004)].
We use a test set of 380 non-redundant monometric proteins whose structure has been solved by X-ray crystallography, with a resolution better than 2 Γ…, extracted from the Top500 dataset used to benchmark the MolProbity program.
Theory part

$$Y = X\beta + \varepsilon \tag{1}$$

where $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$, and $X \in \mathbb{R}^{nxp}$. We can expand this to $y_i = \sum_{j=1}^{p} \beta_i X_{ij} + \varepsilon_i$, $\forall i = = 0, 1, ..., n$. Here β_j are non-random unknown parameters, X_{ij} are non-random and observable, and ε_i are random so y_i are random.
From multiple linear regression we have the coefficient estimate

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2}$$

which we can rewrite as $[(X^T X)^{-1} X^T]^{-1} \hat{\beta} = Y$.
The LASSO model can be shown in the same form as equation (3) above:

$$(Y - X\beta)^T (Y - X\beta) + \lambda |\beta|_1 \tag{3}$$

Where $|\beta|_1 = \sum_{j=1}^{p} |\beta|_j$.