Penalized MDF for Protein Movement Detection

Hiba Alawieh [1] & Nicolas Wicker [1] & Baydaa Al Ayoubi [2] & Luc Moulinier [3]

[1] *Laboratoire Paul Painlevé, Université Lille 1, 59655 Villeneuve d'Ascq, France. alawieh.hiba@gmail.com, Nicolas.Wicker@math.univ-lille1.fr*
[2] *Université Libanaise, faculté des sciences 1, département mathématiques appliquées, Al Hadath, Beyrouth, Liban. ayoubib@ul.edu.lb*
[3] *ICube/LBGI, faculté de Médecine, 67 000 Strasbourg, France. moumou@igbmc.fr*

Résumé. La structure tridimensionelle des protéines peut prendre différentes conformations qui dépendent des réactions qu'elles subissent. Plusieurs méthodes existent pour étudier ces changements conformationnels, mais une seule, appelée DynDom, est clairement consacrée à la détection des déplacements. Nous proposons une méthode alternative fondée sur l'analyse multivariée des données, appelée "Penalized Multidimensional Fitting (Penalized MDF)" basée sur le mouvement pénalisé des points afin d'approcher les distances données par une matrice de référence. L'objectif consiste à détecter les mouvements importants des acides aminés en approchant les distances d'une conformation par les distances d'une seconde conformation dont on modifie les coordonnées. Cette méthode est appliquée à deux protéines différentes.

Mots-clés. Protéines, changements conformationnels, ligand-binding, analyse multidimensionnelle, MDF, ...

Abstract. The three-dimensional structure of a given protein can take different conformations depending upon the reaction it undergoes and its substrate/cofactor/partners binding state. Various methods exist to study these conformational changes but only one, called DynDom, is clearly focused on movement detection. An alternative method is proposed, making use of multivariate data analysis, called "Penalized Multidimensional Fitting (Penalized MDF)" based on penalized points movements in order to approach the distances between points after movement to the distances given by the reference matrix. The objective is to detect the amino acids that undergo an important movement by fitting the distances of one conformation to the distances of the second one by modifying only the coordinates of the first one. This method is applied on two different proteins .

Keywords. Proteins, Conformational changes, Ligand binding, Multidimensional analysis, MDF, ...

1 Introduction

Proteins are heteropolymers that can take three-dimensional structure. These structures are flexible, highly dynamic, and their biological functions depend intimately on them. This structure deformation can be induced in various ways such as binding other molecules, enzyme catalysis reaction, etc. A novel method called Penalized Multidimensional Fitting (Penalized MDF) is presented to detect movement by using two conformations of the same protein. This multivariate analysis approach is an adaptation of Multidimensional Fitting [1]. The idea is to compare one protein conformation with another one by modifying the coordinate matrix of the first one, called target matrix, in order to approach the distances calculated on the second matrix, called reference matrix. What differentiates our method from Procrustes analysis method is that the latter compares two configurations by moving one configuration relatively to the second through a rotation, translation or scaling that moves all the points a same distance [4]. Penalization is necessary as it is clear that without it, every transformation would be possible, and then the solution would reduce to take the target matrix equal to the coordinate matrix corresponding to the reference matrix ! This of course, does not give any information on which part of the protein has moved. The main work here is then to devise a good penalization, and then to apply this methos on differents proteins.

2 Penalized MDF method

Let $X = \{X_1 | \cdots | X_n\}'$ be the $n \times p$ target matrix and $D = \{d_{ij}\}$ the $n \times n$ reference matrix calculated on an other structure of the same protein, this matrix contains the Euclidean distances between the amino acids. Besides, we note $\triangle = \{\delta_{ij}\}$ the distance matrix obtained from X after MDF.

The MDF method allows us to modify the target matrix in order to minimize the difference between the reference matrix and the novel distance matrix computed on the modified target matrix. The idea behind MDF is to minimize the mean square error:

$$E = \sum_{1 \leq i < j \leq n} (d_{ij} - \delta_{ij})^2 \text{ where } \delta_{ij} = d(f(X_i), f(X_j)) \text{ and } f(X_i) = X_i + L_i.$$

under some constraints. For all $i \in 1, \ldots, n$, the vector $L_i = (l_{i1}, l_{i2}, \ldots, l_{ip})$ denotes the displacements for the i^{th} point.

Here, no constraint is needed but to avoid unnecessary movements, a penality term is added leading to the following optimization problem:

$$O : \begin{cases} \min \sum_{1 \leq i < j \leq n} (||X_i + L_i - X_j - L_j||_2 - d_{ij})^2 + \lambda \sum_{i=1}^{n} \text{pen}(L_i) \text{ with } L_i \in \mathbb{R}^p \end{cases}$$

The parameter λ is a positive regularization parameter that controls the trade-off between the approximation of the reference matrix by the distance matrix computed on the modified matrix and the use of a parsimonious number of displacements. To have interesting results, it is clear that having a good penalization is important.

2.1 Choice of a penalty function

The chosen norm can be used in many ways. We have only considered two cases, either having simply $\|.\|_2$ or having two homogeneous terms by taking $\|.\|_2^2$. Using $\|.\|_2^2$, we obtain, with less points moving, a larger penalty than with $\|.\|_2$. This result is not interesting for our parsimony needs. Therefore, we will use henceforth $\|.\|_2$ as penalty term.

2.2 Choice of parameter λ

We have already seen that the value of λ is crucial for obtaining good results. In this section, we want to find the best value for λ. First, in lemma 2.1 we show that there are at least two points moving in different directions. Then, using this fact we derive bounds on λ.

LEMMA 2.1 *The solution of problem O is such that there is a fixed point or at least two points moving in different directions provided that $\lambda > 0$.*

LEMMA 2.2 *In one dimension, if the solution of O is such that $\forall i = 1, \ldots, n$, $|l_i| > 0$, for two points i and j moving in opposite directions, the following bound holds:*
$\lambda \geq n(\delta_{ij} - d_{ij})$.

LEMMA 2.3 *In one dimension, if for two points i and j moving in opposite directions, the parameter λ is such that $\lambda < n(d_{ij}^0 - d_{ij} - \epsilon)$ then $\exists$ k, such as $|l_k| > \epsilon$, where d_{ij}^0 is the initial distance computed on the target matrix.*

2.3 Penalization by combining $\|.\|_2$ and $\|.\|_0$

We consider in this section, the combined penalty term as $\sum_{i=1}^{n} (\gamma \|l_i\|_2 + (1-\gamma)\|l_i\|_0)$ weighted by a parameter λ, where $\gamma \in [0, 1]$. We call the function $\gamma\|l_i\|_2 + (1-\gamma)\|l_i\|_0$ the elastic net penalty by analogy with the well-known elastic net. The ℓ_0 norm penalizes the number of nonzero movements. The expression E is:

$$E = \sum_{1 \leq i < j \leq n} (d_{ij} - \|X_i + L_i - X_j - L_j\|_2)^2 + \lambda \sum_{i=1}^{n} (\gamma \|l_i\|_2 + (1-\gamma)\|l_i\|_0)$$

3 Application

In this section, penalized MDF has been applied to two different proteins to detect significative movements in their tridimensional structure. MDF needs a reference matrix and a coordinate matrix. For two different structures of the same protein, the coordinate matrix is given by the C_α coordinates of one structure and the reference matrix is given by the Euclidean distances between the amino acids of the second structure. The optimization problem is a non-linear optimization problem. The Nlopt library [7] has been used to solve it by using DIRECT-L algorithm for global optimization and SBPLX algorithm for local optimization. For the choice of parameter λ, recall that by lemma 2.2 we have $\lambda \geq n(\delta_{ij} - d_{ij})$ for any two points i and j moving in opposite directions. Besides, by lemma 2.3, $\exists i, j$ such that $\lambda < n(d^0_{ij} - d_{ij} - \epsilon)$ as otherwise all $|l_k|$ would be inferior to ϵ. So, we see that the order of magnitude of λ is n times a small gap value that we take equal to 0.5Å for the present application. Concerning parameter γ, we use the value 0.5 which gave the best results. To have a threshold to determine if a movement is significative, we compare the computed movement after MDF with the standard deviation σ_i for each point i. For this, we use the known B-factor of each atom i which indicates the true static or dynamic mobility of an atom [6] given by: $B_i = 8\pi^2 d^2_{mi}$, to infer the mean movement d_{mi} of atom i, $\forall i = 1, \ldots, n$. Besides, $d^2_{mi} = E(\|X_i - \mu_i\|^2)$ with $X_i \rightsquigarrow \mathcal{N}(\mu_i; \sigma^2_i I_3)$ and μ_i the mean coordinates for each atom i. Then, $d^2_{mi} = \sigma^2_i E(\frac{\|X_i - \mu_i\|^2}{\sigma^2_i}) = 3\sigma^2_i$ as $\frac{\|X_i - \mu_i\|^2}{\sigma^2_i} \rightsquigarrow \chi^2_3$. Thus, $\sigma_i = \frac{d_{mi}}{\sqrt{3}}$. We suppose that the value $2\sigma_i$ is high enough to detect important movements. Penalized MDF has been applied to two proteins: human estrogen nuclear receptor (ER) and FhuA. For each protein, we compare our results with those obtained by DynDom.

3.1 Human estrogen receptor protein

ER is a Nuclear estrogen receptor composed of several functional domains that serve specific roles. Many experiments demonstrate that their C-terminal Helix (H12) is more flexible without ligand. Penalized MDF has been used to compare the conformation with and without ligand. Results are presented in table 1 for different values of γ. Figure 1 indicates that an important movement occurs at the end of the sequence (the amino acids number 214 to 231) and smaller movements at others regions. The dotted curve in Figure 1 shows the standard deviation σ_i for each individual, the important movements are detected at amino acids $26, 27, 28$, and 214 to 231. This result is confirmed by Anke and others who note that the position of this helix depends on the presence or not of a ligand [5]. Concerning positions $26, 27, 28$, they correspond to the sequence "SEA" which is apical of helix H3. Sumbayev and others explain in [8] the movements of this helix.

γ	0	0.1	0.3	0.5	0.8	1
$\sum_{1 \leq i < j \leq n} (d_{ij} - \delta_{ij})^2$	6287	6809	8913	9615	9741	9423
Nb of movements	108	102	83	76	71	99
Nb of a.a moved	61	58	48	43	36	52

Table 1: Using combined norm penalization reduce the number of amino acids that move.

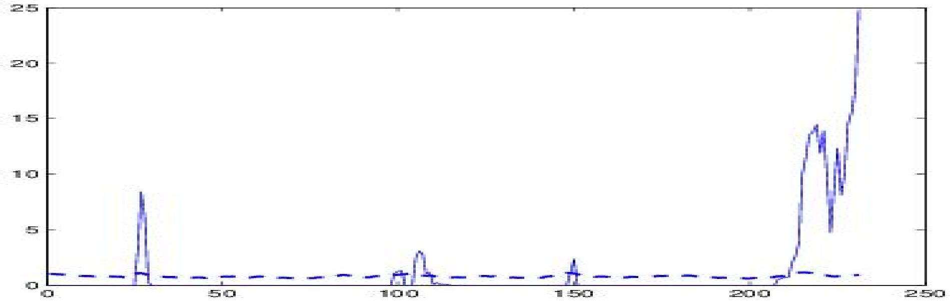

Figure 1: Distances between initial coordinates and modified coordinates for each amino acid in Human estrogen receptor (solid curve). The dotted curve corresponds to the $2\sigma_i$ threshold . The amino acids 26, 27, 28 and $214-231$ are considered as important movements.

3.2 FhuA protein

FhuA is an outer membrane receptor protein of *Escherichia coli* bacteries. X-ray analysis at 2.7Å resolution reveals two distinct conformations in the presence and absence of ferrichrome. Penalized MDF has been applied to compare the two conformations. The results are given in table 2.

γ	0	0.1	0.3	0.5	0.8	1
$\sum_{1\leq i<j\leq n}(d_{ij}-\delta_{ij})^2$	57087	60538	65632	67398	67860	68024
Nb of movements	37	29	23	21	26	44
Nb of a.a moved	25	18	12	10	13	29

Table 2: Results for FhuA

Figure 2 depicts important distances between the modified and initial coordinates for the 10 first amino acids. This result is confirmed by biology [2] et [3]. The N-terminus has moved after ligand binding.

No domain movement are detected by using DynDom for ER and FhuA proteins.

4 Conclusion

The purpose of penalized MDF is to modify only the amino acid coordinates that have significantly moved and fix the others. Penalization term and penalization parameters are crucial in the process of obtaining good results. This involves the choice of a penalty coefficient λ which is related to the minimum displacement.

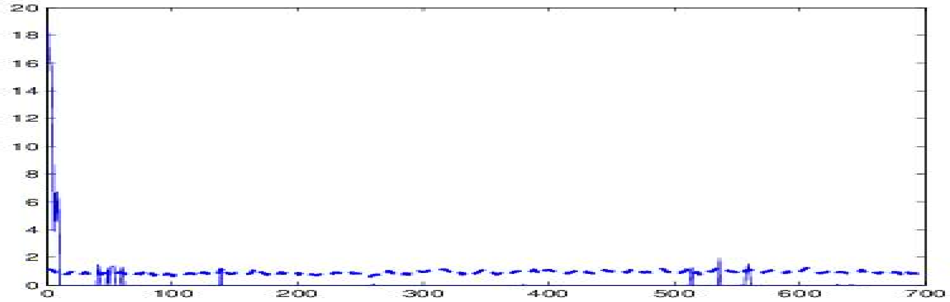

Figure 2: For FhuA, important movements are located in the N-terminus, which is confirmed by the biological literature.

Penalized MDF has been applied to two different proteins in order to find the residues that were affected by the interaction with other molecules. Further research is however needed to simplify the optimization problem and reduce the costs.

Bibliography

[1] Berge, C., Froloff, N., Kalathur, RK., Maumy, M., Poch, O., Raffelsberger, W., Wicker, N. (2010) Multidimensional fitting for multivariate data analysis, *Journal of Computational Biology*, **17**, 723–732.
[2] Faraldo-Gomez J.D., Smith G.R., Sansom M.S. (2003) Molecular dynamics simulations of the bacterial outer membrane protein FhuA: a comparative study of the ferrichrome-free and bound states, *Biophysical journal*, **85**, 1406-1420.
[3] Locher, K.P, Rees B, Koebnik R, Mitschler A, Moulinier L, Rosenbusch J.P, Moras D (1998) Transmembrane Signaling across the Ligand-Gated FhuA Receptor: Crystal Structures of Free and Ferrichrome-Bound States Reveal Allosteric Changes, *Cell press*, **95**, 771–778.
[4] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, New York.
[5] Mueller-Fahrnow, A., and Egner, U. (1999) Ligand-binding domain of estrogen receptors, *Current Opinion in Biotechnology*, **10**, 550–556.
[6] Rhodes, G. Judging the Quality of Macromolecular Models A Glossary of Terms from Crystallography NMR and Homology Modeling, http://spdbv.vital-it.ch/TheMolecularLevel/ModQual/.
[7] Steven G. Johnson, The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt
[8] Sumbayev, V.V, Bonefeld-Jorgensen, E.C., Wind, T, Andreasen, P.A. (2005) A novel pesticide-induced conformational state of the oestrogenreceptor ligand-binding domain, detected by conformation-specificpeptide binding, *FEBS Letters*, **579**, 541–548.