Nucl Acid Res (2012) 40:W317-22.

PredyFlexy: Flexibility and Local Structure prediction from sequence

Alexandre G. de Brevern[1,2,3,+,*], Aurélie Bornot[1,2,3,+,#], Pierrick Craveur[1,2,3]
Catherine Etchebest[1,2,3] & Jean-Christophe Gelly[1,2,3]

[1] INSERM, U665, DSIMB, F-75739 Paris, France.
[2] Univ Paris Diderot, Sorbonne Paris Cité, UMR_S 665, F-75739 Paris, France.
[3] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.
[#] Present address: AstraZeneca, Discovery Sciences, Computational Biology, Alderley Park UK.

Short title: protein flexibility prediction

* Corresponding author:
mailing address: de Brevern Alexandre G., INSERM UMR_S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail: alexandre.debrevern@univ-paris-diderot.fr
Tel: +33(1) 44 49 30 38
Fax: +33(1) 47 34 74 31

[+] Both authors contributed equally

Key words: amino acid; structural alphabet; Long Structural Prototypes; protein folds; disorder; Support Vector Machines; evolutionary information; Protein Data Bank.

Abstract

Protein structures are necessary for understanding protein function at a molecular level. Dynamics and flexibility of protein structures are also a key element of protein function. So we have proposed to look at protein flexibility using novel methods: (i) using a structural alphabet and (ii) combining classical X-ray B-factor data and Molecular Dynamics simulations.

First, we established a library composed of structural prototypes (LSPs) to describe protein structure by a limited set of recurring local structures. We developed a prediction method that proposes structural candidates in terms of LSPs and predict protein flexibility along a given sequence.

Secondly, we examine flexibility according to two different descriptors: X-ray B-factors considered as good indicators of flexibility and the root mean square fluctuations, based on molecular dynamics simulations. We then define three flexibility classes and propose a method based on the LSP prediction method for predicting flexibility along the sequence.

This method does not resort to sophisticate learning of flexibility but predicts flexibility from average flexibility of *predicted* local structures. The method is implemented in PredyFlexy web server. Results are similar to those obtained with the most recent, cutting-edge methods based on direct learning of flexibility data conducted with sophisticated algorithms.

PredyFlexy can be accessed at http://www.dsimb.inserm.fr/dsimb_tools/predyflexy.

Introduction

X-ray experiments have been valuable tools to understand the intimate relation between protein structures and biological functions. They have revealed a large diversity of well-defined folds, each being adopted by members of a given functional family. However, recent studies have shown that conformational changes are required by numerous proteins in their folded state to accomplish their function (*e.g.*, enzyme catalysis, activity modulation, macromolecular interactions, ligand binding, cell motility (1-4)). This has led to revisit the importance of dynamics and to focus on regions with peculiar flexibility properties, supposed to participate in conformational changes. Hence, determining those regions would be extremely useful to decipher and eventually control biological function. Actually, few studies have focused on flexible regions in folded ordered proteins. Studies have mainly focused on (*i*) the analysis of specific protein structures to catch and/or simulate the flexible and rigid regions and (*ii*) the sole information of the sequence to predict flexibility

In the first case, 3D-structures are required all along. B-factors available with X-ray structures were first used as the main criteria to define protein rigidity and flexibility. Nowadays, the distinction between flexible and rigid regions takes advantage of dedicated approaches for exploring dynamics. The most popular approaches consist in atomistic molecular dynamics simulations, which are available through different packages like Gromacs (5), Amber (6), NAMD (7) or Charmm (8). Principal component analyses (PCA) of the resulting data allow identifying regions involved in the different type of motions and provide relevant information about the visited conformational space. Less time-consuming methods are also available, *e.g.*, *FlexServ* (9), *ElNemo* (10) or *Nomad* (11) which perform Normal Mode Analysis (NMA) of elastic network models (ENM). Data can also be gained with Brownian dynamics (BD) and Discrete Dynamics (DMD) (9), or more specialized approaches, *e.g.*, to define hinges between domains, as *StoneHinge* (12), *HingeProt* (13), and

tCONCOORD (14), which predicts conformational flexibility based on geometrical considerations. All these methods give a large amount of data that bring quantitative information enabling precise ranking of flexible and rigid regions by highlighting as well local deformation as large domain motions.

In the second case, the prediction is based on the sole amino acid sequence. Historically, the flexibility was first predicted as Boolean, *i.e.*, rigid or flexible, using simple statistical analyses of B-factor values (15-16). In the same spirit Schlessinger and Rost (17) developed more recently, *PROFB*val, a method that improved the two - states flexibility prediction by using Artificial Neural Networks (ANNs) combined with evolutionary information. Instead of ANNs, Pan and Shen used Support Vector Regression coupled with Random Forest (18). Chen and co-workers proposed an innovative development of Logistic Regressions and colocation-based representation with multiple features to predict flexible and rigid region (19). Nuclear Magnetic Resonance data are alternative sources of information for protein dynamics. Zhang and co-workers (20) as well as Trott and co-workers (21) chose to exploit these data rather than X-ray B-factors. Zhang's group used variation of backbone torsion angles from NMR structural models whereas Trott and co-workers preferred order parameters to define the protein flexibility. Both groups performed prediction with neural networks. Galzitskaya and co-workers extend the *FoldUnfold* (22) methodology, which was originally designed to predict disorder, to the prediction of flexibility (23).

Interesting works related to protein flexibility prediction have focussed on more specific question. Hence, Hirose and co-workers used NMA to define specific motions in proteins. These motions were predicted using a Random Forest algorithm and were further used to explore protein-protein interaction (24). Hwang and co-workers focused on prediction of flexible loops and combined B-factors, dihedral angles and accessibility (25). Kuznetsov and co-workers proposed a web-server for predicting residue involved in conformational

switches in proteins. Interestingly, it can use either protein sequence or structure. The prediction from the sequence is done with Support Vector Machines (26-27).

We take advantage of the method we previously elaborated to predict local protein structures. We have described global protein structures using a limited set of recurring local structures named long structural prototypes (LSP (28)). These LSPs encompass all known local protein structures and ensure good quality 3D local approximation. We have proposed a prediction method based on evolutionary information coupled with support vector machines (SVM). This method provides with a list of five possible structural candidates for a target sequence. The prediction rate reaches 63.1%, a rather high value given the high number of structural classes (29). We use the output of this structural prediction as the input for our prediction method of flexibility.

The originality of our method lies (*i*) in the use of a combination of two descriptors for quantifying protein dynamics, *i.e.* the X-ray B-factors and the Root Mean Square Fluctuation (RMSF) computed from Molecular Dynamics, (*ii*) in the prediction of flexibility through structural prediction of LSPs (see above) and (*iii*) by considering three classes of flexibility defined by the chosen descriptors and in which LSPs were distributed. This method turns out to be rather efficient compared to the most commonly used ones. The prediction rate is slightly better than the one of PROFbval (17) that was optimized for two classes. Importantly, we also propose a confidence index for assessing the quality of the prediction rate. The method is implemented in a useful webserver (http://www.dsimb.inserm.fr/dsimb_tools/predyflexy/), PredyFlexy that is able to give different type of predictions as well a confidence index with outputs as well as flat file.

Methods

The server can be used to predict protein flexibility as well to predict local protein

structure defined by LSPs. Figure 1 explains the two main steps of the prediction methodology. At first, LSPs are predicted, and then using this prediction, protein flexibility is predicted. Prediction is defined using classical normalised B-factors ($B\text{-factor}_{Norm}$) and normalized Root Mean Square Fluctuations ($RMSF_{Norm}$) from Molecular Dynamics.

LSPs Prediction. We have proposed a library consisting of 120 overlapping structural classes of 11-residues long fragments (28). This library was constructed with an original unsupervised structural clustering method called the Hybrid Protein Model (HPM, (30)). The Hybrid Protein principle is similar to a self-organizing neural network (31-32). It was constructed as a ring of *N* neurons (here *N*=120), each representing a cluster of structurally similar 3D fragments encoded into series of Protein Blocks (PBs). PBs are a structural alphabet (33), *i.e*., a set of local protein fragments, able to provide correct approximation of protein structure. Its training strategy consisted in learning the similarities between protein structural fragments deduced from the alignment of their PBs series (34-35). Once the HPM was trained, each neuron or cluster was associated with a set of fragments representing a structural class using root mean square deviation (RMSD, (28)). For each class, a mean representative fragment, or a "local structure prototype" (LSP), was chosen. The 120 LSPs correctly approximated the local structure ensembles. The major advantage of this library is its capacity to capture the continuity between the identified recurrent local structures (29). Relevant sequence-structure relationships were also observed and further used for prediction. Briefly, LSPs prediction is based on SVM training. With the LSP prediction is provided a confidence index CI that is based on the discriminative power of the SVMs. The higher the CI, the better the prediction rate is. For more details on LSPs and their prediction, please see (36).

Protein structure datasets. A dataset of 172 X-ray high-resolution (≤ 1.5 Å) globular

protein structures was extracted from the Protein Data Bank (PDB) using the PDB-REPREDB database web service (37), which provides the user with different choices of thresholds for selecting chains of given sequence and structural similarity. The method is detailed in (38). We chose chains sharing less than 10% sequence identity and for which the Cα RMSD between aligned residues differ by at least 10 Å. Proteins composed of a single domain, not involved in a protein complex, and that did not have extensive number of contacts with ligands were considered only. A final dataset of 43 protein structures was obtained. This dataset 1 was used to calibrate thresholds for RMSF computed from Molecular Dynamics simulations using Gromacs (5). Parameters and conditions defined in ref. (39) were used for the simulations. A larger, non-redundant databank composed of 1421 X-ray structures with resolution higher than 1.5 Å, sequence identity smaller than 30% and Cα RMSDs larger than 10 Å (selected using PDB-REPRDB (37)) was used for the prediction itself (dataset 2).

Definition of protein structure flexibility classes. We extracted Cα B-factors from the PDB files of the protein structures dataset 1. For comparison purposes, the raw values were normalized for each protein using the method in (40). After removing outliers detected statistically with a median-based approach, the normalized B-factors were calculated as $\text{B-factor}_{\text{Norm}} = (\text{B-factor}_{\text{Raw}}-\mu)/\sigma$ where μ and σ stand for the mean and the standard deviation of the Cα B-factor, respectively. Flexibility of each 11-residue long overlapping fragment in the dataset was characterised by the $\text{B-factor}_{\text{Norm}}$ associated with its central Cα.

Similarly, we extracted flexibility measurements from MD simulations. Cα root mean square fluctuation (Cα RMSF) was calculated using g_rmsf GROMACS tool (5) after superimposing each snapshot structure on the initial conformation. Cα RMSF gives the mean amplitude of each Cα movement compared to a mean reference position:

$RMSF^{i}_{Norm} = \sqrt{\frac{1}{T}\sum_{t=0}^{T} (\vec{R}^{i}_{t} - \vec{R}^{i}_{ave})^2}$ where T is the production time expressed in snapshot number, $\vec{R}^{i}_{t}$ the coordinates of Cα atom *i* of structure at time *t* and $\vec{R}^{i}_{ave}$, average coordinates of Cα atom *i* over production time. Raw RMSF values were normalized for each protein. The $RMSF_{Norm}$ associated with the central Cα of each 11-residue fragment characterised the flexibility using MD.

Hence, to each fragment is associated a couple of values $Bfactor_{Norm}$ and $RMSF_{Norm}$. The three flexibility classes, rigid, intermediate and flexible, were then defined from a fine calibration of thresholds combining Cα RMSF (noted τ_F) and B-Factors (noted τ_B). The calibration was based on a backward-forward procedure targeting the optimal flexibility prediction rate. Fragments for which the couple (Cα B-Factors, Cα–RMSF) is (*i*) smaller than τ_{B1}, τ_{F1} are rigid, (*ii*) larger than τ_{B1}, τ_{F1} but smaller τ_{B2}, τ_{F2} are intermediate, and (*iii*) larger τ_{B2}, τ_{F2} are flexible.

Finally, a detailed analysis of RMSF and Cα B-Factors couples for each LSP allowed attributing a well-defined flexibility class to each of them as well as a mean $B\text{-}factor_{Norm}$ and a mean $RMSF_{Norm}$. This was obtained by (*i*) computing the propensity of fragments belonging to a LSP to be associated with a given flexibility class and (*ii*) selecting as the unique assigned class for each LSP, the class that maximizes the propensity (see ref (39) for details).

Flexibility Prediction: For a target sequence, the local structure prediction is first performed and yields the five best LSP candidates. Then the predicted flexibility class is obtained by simply calculating the rounded average of the flexibility classes of the 5 candidates. In the same way, the $Bfactor_{Norm}$ and $RMSF_{Norm}$ is predicted by averaging the mean $Bfactor_{Norm}$ and $RMSF_{Norm}$ of the 5 structural candidates. At this stage, no training on the data was performed. The prediction reflects the informativity of structural prediction from sequence for flexibility.

Discussion

The PredyFlexy method is based on the flexibility analysis of local protein structures through an appropriate combination of the B-factor of X-ray experiment and the fluctuation of residues during molecular dynamics simulations. A correlation (r^2 equals 0.68) was obtained between Cα-$Bfactor_{Norm}$ and Cα-$RMSF_{Norm}$. This value confirms that even though related, both descriptors bring different information justifying the interest to combine both measures of the flexibility. The *PredyFlexy* method led to an average, well-balanced prediction rate of 49.4% for the three defined flexibility classes, a value considerably higher than a random prediction rate. The correlation r^2 between observed and predicted values for $Bfactor_{Norm}$ and $RMSF_{Norm}$ reached 0.71 and 0.69 respectively. When outliers (5% of the values), detected by the median-based approach proposed by Smith et al (40), were excluded, correlations r^2 climbed to 0.94 and 0.96, respectively. This correlation is slightly better than the best correlation value obtained by the PONDR VSL1 prediction methods (41).

For comparison purpose, we regrouped our three flexibility classes into two classes to assess a two-class prediction. Depending on the grouping, we obtained prediction rates comparable and even better than the current methods available (17-18). Details are given in Table IV of ref (39). This confirms that LSP description is truly useful for addressing flexibility prediction.

Web Server: PredyFlexy provides a user-friendly web interface that combines predictions for local structure and flexibility. The homepage contents a short summary of the two aspects of the method. In this page, the sole input, the protein sequence, must be provided. Two possibilities are offered: the sequence, in Fasta format, may be paste in a first window frame or download from a file, the filename being given in a second window frame. This page contents additional links: “Contacts”, which refers to authors’ homepage, “About

Method", which details the methodology and its flowchart, "Download", which allows to obtain a local version of the program by sending an email for registration, "Example", which illustrates with a concrete case, the input and output of the server (see below), "DSIMB", which connects to team's homepage. In figure 1, are described the different steps that led from a protein sequence to the output results of the prediction.

Input. A single Fasta sequence must be provided (Figure 1A). A check is performed to ensure that only natural amino acids are used.

Background Step: "PredyFlexy Running". For the given sequence, a Position Substitution Sequence Matrix (PSSM) is first computed with PSI-BLAST v. 2.2.09 (42) using default parameters and Swissprot databank (43) (Figure 1B). The sequence is then divided into overlapping fragments of 11 residues long (Figure 1B), corresponding to the LSP size. In a third step, LSP prediction is done using 120 independent Support Vector Machines (libsvm-2.81) that was previously optimized for each LSP (36). This method yields for a target sequence a list of five structural candidates associated with the highest scores (Figure 1d). The prediction rate reaches 63.1%, a rather high value given the high number of structural classes (36). From this prediction, the corresponding flexibility class of the LSP is attributed. Hence, at this stage, each sequence fragment is characterized by five flexibility states, one per structural LSP in the list. Finally, the predicted flexibility state of a 11-residue sequence is a simple mean of the flexibility states for the five predicted LSPs candidates. (Figure 1e). Using so a similar approach, local B-factor$_{Norm}$ and RMSF$_{Norm}$ are predicted (Figure 1f).

Output. Once the job is finished, a window opens with the results. Results are given as a text file that can be downloaded. Results may also be visualized through different graphical outputs. The first graphs represent, the values, along the sequence, of the B-factor$_{Norm}$ (green),

the $RMSF_{Norm}$(yellow) on the left y-axis, and on the right y-axis, the confidence index (gray line). For clarity, the results are represented by blocks of 50 residues. The sequence is reported in the same graph above the x-axis. These combined representations allow the user to focus on the regions with a high confidence index, *i.e.*, larger than 15 (representing more than 25% of residues), frequently associated with regions with low flexibility. In the second part of the page, a table summarizes the results of the local structure prediction, the confidence index and the flexibility class. The lines correspond to each position along the sequence. In the two first columns are indicated the position and the corresponding amino acid (one letter). The five following columns contain the five best LSP candidates represented by their 3D structure and their corresponding number in the Hybrid Protein Model (for details, see (36)). The two last columns correspond to the confidence index value and the predicted flexibility class (0 for rigid, 1 for intermediate and 2 for flexible). The confidence index is represented by 19 discrete values ranked from 1 to 19, with the prediction confidence increasing. For a rapid visualization inspection, values for CI and flexibility classes are also represented with colors. Note that due to the LSP size, the 10 first and 10 last residues are not predicted.

The text file brings the same information (except the 3D representation) and two additional columns for the predicted B-$factor_{Norm}$ and $RMSF_{Norm}$.

Implementation. Implementation of this tool is done in Python and HTML, while the graphical plots are done using R software (44-45). The front-end use is based on html and php. Perl/cgi programs control the input while python and other programs carry out the processing behind the database search and pairwise comparisons.

Figure 2 illustrates the results of the prediction of isomerase in a region ranging from residue 100 to 150. As the confidence index is higher than 15, the regions from (a) to (d) are very well predicted, while the region (e) is not reliable with a very low confidence index (equals to 3). So, the user can be quite sure of a succession from flexible (a) to rigid (b) with a

intermediate to flexible zone (c) then come back to rigid zone (d). By looking at the distribution of predicted LSPs, the user can analyse more deeply what could be the local conformations adopted by this region, *i.e.,* a succession of short helical regions alternated by short loops going to an extended conformation.

Conclusion

Very few webservers are dedicated to the prediction of flexibility from the sole information of the sequence. We propose an original tool that combines in one run the prediction of the local structures and the associated flexibility. We also chose to predict flexibility in three classes compared to two in most studies. We also provide B-factor and RMSF prediction. Additionally, very useful and important information is provided by the confidence index. This value allows the user to assess the predictability of its sequence or region of interest. We hope the availability of our method through PredyFlexy Web server will help researchers to better understand the properties of their protein and design new experiments focusing on appropriate regions depending on their goal.

Funding

These works were supported by grants from the Ministry of Research (France), University of Paris Diderot, Sorbonne Paris Cité, National Institute for Blood Transfusion (INTS, France), Institute for Health and Medical Research (INSERM, France) to AdB, CE and JCG. AB and PC acknowledge grants from French Ministry of Research.

Figure legends

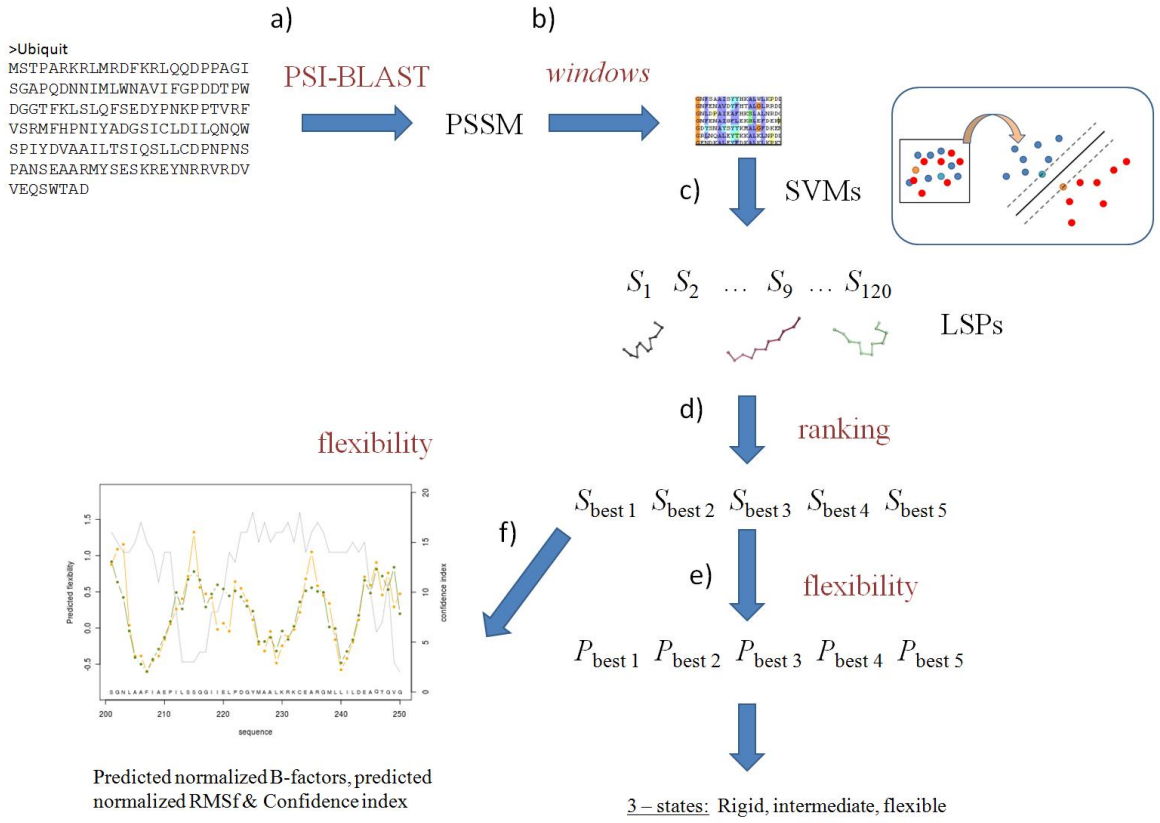

Figure 1. *The framework of PredyFlexy and underlying methods*. User must give a single sequence as input (a) a PSSM is computed using PSI-BLAST (b) and split into fragments of 11 residues. Prediction of LSPs is done using trained SVMs (c) scores are ranked and the best five are kept (d). Using these LSPs, prediction of flexibility is done in three states (rigid, intermediate and flexible) (e) are also provided predicted B-factor$_{Norm}$, RMSF$_{Norm}$ and a confidence index (f).

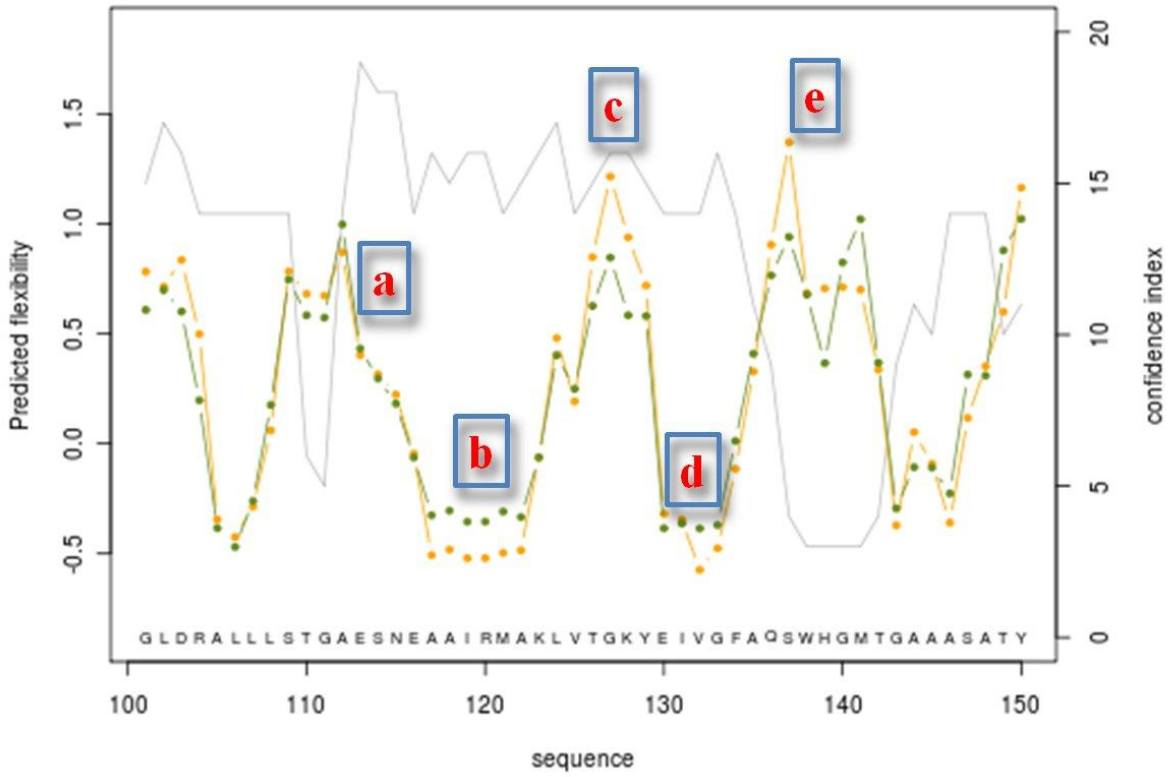

Figure 2. *Protein prediction example*. The prediction highlights the regions from residue 100 to 150. In regions (a) to (d) are located residues predicted with a high accuracy (confidence index of 15 or better, it represents flexible (a) to rigid (b) with an intermediate to a flexible zone (c) then coming back to rigid zone (d). Following region (e) is predicted as flexible, but the low confidence index (equals to 3) makes the prediction not reliable.

References

1. Beckett, D. (2009) Regulating transcription regulators via allostery and flexibility. *Proc Natl Acad Sci U S A*, **106**, 22035-22036.
http://www.ncbi.nlm.nih.gov/pubmed/20080782
http://dx.doi.org/10.1073/pnas.0912300107
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2799754

2. Hammes, G.G., Benkovic, S.J. and Hammes-Schiffer, S. (2011) Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry*, **50**, 10422-10430.
http://www.ncbi.nlm.nih.gov/pubmed/22029278
http://dx.doi.org/10.1021/bi201486f
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3226911

3. Lill, M.A. (2011) Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry*, **50**, 6157-6169.
http://www.ncbi.nlm.nih.gov/pubmed/21678954
http://dx.doi.org/10.1021/bi2004558
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3172316

4. Lin, J.H. (2011) Accommodating protein flexibility for structure-based drug design. *Curr Top Med Chem*, **11**, 171-178.
http://www.ncbi.nlm.nih.gov/pubmed/20939792

5. Lindahl, E., Hess, B. and van der Spoel, D. (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, **7**, 306-317.

6. Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.J. (2005) The Amber biomolecular simulation programs. *J Comput Chem*, **26**, 1668-1688.
http://www.ncbi.nlm.nih.gov/pubmed/16200636
http://dx.doi.org/10.1002/jcc.20290

7. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**, 1781-1802.
http://www.ncbi.nlm.nih.gov/pubmed/16222654
http://dx.doi.org/10.1002/jcc.20289

8. Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: the biomolecular simulation program. *J Comput Chem*, **30**, 1545-1614.
http://www.ncbi.nlm.nih.gov/pubmed/19444816
http://dx.doi.org/10.1002/jcc.21287

9. Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Gelpi, J.L. and Orozco, M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, **25**, 1709-1710.
http://www.ncbi.nlm.nih.gov/pubmed/19429600
http://dx.doi.org/10.1093/bioinformatics/btp304

10. Suhre, K. and Sanejouand, Y.H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*, **32**, W610-614.
http://www.ncbi.nlm.nih.gov/pubmed/15215461
http://dx.doi.org/10.1093/nar/gkh368
32/suppl_2/W610 [pii]

11. Lindahl, E., Azuara, C., Koehl, P. and Delarue, M. (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res*, **34**, W52-56.

http://www.ncbi.nlm.nih.gov/pubmed/16845062
http://dx.doi.org/34/suppl_2/W52 [pii]
10.1093/nar/gkl082

12. Keating, K.S., Flores, S.C., Gerstein, M.B. and Kuhn, L.A. (2009) StoneHinge: hinge prediction by network analysis of individual protein structures. *Protein Sci*, **18**, 359-371.
http://www.ncbi.nlm.nih.gov/pubmed/19180449
http://dx.doi.org/10.1002/pro.38
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2708048

13. Emekli, U., Schneidman-Duhovny, D., Wolfson, H.J., Nussinov, R. and Haliloglu, T. (2008) HingeProt: automated prediction of hinges in protein structures. *Proteins*, **70**, 1219-1227.
http://www.ncbi.nlm.nih.gov/pubmed/17847101
http://dx.doi.org/10.1002/prot.21613

14. Seeliger, D. and De Groot, B.L. (2009) tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules. *J Comput Chem*, **30**, 1160-1166.
http://www.ncbi.nlm.nih.gov/pubmed/18942729
http://dx.doi.org/10.1002/jcc.21127

15. Karplus, P. and Schulz, G. Prediction of chain flexibility in proteins. A tool for the selection of peptide antigens. *Naturwissenschaften*, **72**, 212-213.
http://dx.doi.org/DOI: 10.1007/BF01195768

16. Vihinen, M., Torkkila, E. and Riikonen, P. (1994) Accuracy of protein flexibility predictions. *Proteins*, **19**, 141-149.
http://www.ncbi.nlm.nih.gov/pubmed/8090708
http://dx.doi.org/10.1002/prot.340190207

17. Schlessinger, A., Yachdav, G. and Rost, B. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891-893.
http://www.ncbi.nlm.nih.gov/pubmed/16455751
http://dx.doi.org/10.1093/bioinformatics/btl032

18. Pan, X.Y. and Shen, H.B. (2009) Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept Lett*, **16**, 1447-1454.
http://www.ncbi.nlm.nih.gov/pubmed/20001907

19. Chen, K., Kurgan, L.A. and Ruan, J. (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol*, **7**, 25.
http://www.ncbi.nlm.nih.gov/pubmed/17437643
http://dx.doi.org/10.1186/1472-6807-7-25
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1863424

20. Zhang, T., Faraggi, E. and Zhou, Y. (2010) Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins*, **78**, 3353-3362.
http://www.ncbi.nlm.nih.gov/pubmed/20818661
http://dx.doi.org/10.1002/prot.22842
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2976825

21. Trott, O., Siggers, K., Rost, B. and Palmer, A.G., 3rd. (2008) Protein conformational flexibility prediction using machine learning. *J Magn Reson*, **192**, 37-47.
http://www.ncbi.nlm.nih.gov/pubmed/18313957
http://dx.doi.org/10.1016/j.jmr.2008.01.011
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2413295

22. Galzitskaya, O.V., Garbuzynskiy, S.O. and Lobanov, M.Y. (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, **22**, 2948-2949.
http://www.ncbi.nlm.nih.gov/pubmed/17021161
http://dx.doi.org/10.1093/bioinformatics/btl504

23. Mamonova, T.B., Glyakina, A.V., Kurnikova, M.G. and Galzitskaya, O.V. (2010) Flexibility and mobility in mesophilic and thermophilic homologous proteins from molecular dynamics and FoldUnfold method. *J Bioinform Comput Biol*, **8**, 377-394.
http://www.ncbi.nlm.nih.gov/pubmed/20556851

24. Hirose, S., Yokota, K., Kuroda, Y., Wako, H., Endo, S., Kanai, S. and Noguchi, T. (2010) Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC Struct Biol*, **10**, 20.
http://www.ncbi.nlm.nih.gov/pubmed/20626880
http://dx.doi.org/10.1186/1472-6807-10-20
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245509

25. Hwang, H., Vreven, T., Whitfield, T.W., Wiehe, K. and Weng, Z. (2011) A machine learning approach for the prediction of protein surface loop flexibility. *Proteins*, **79**, 2467-2474.
http://www.ncbi.nlm.nih.gov/pubmed/21633973
http://dx.doi.org/10.1002/prot.23070

26. Kuznetsov, I.B. (2008) Ordered conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data. *Proteins*, **72**, 74-87.
http://www.ncbi.nlm.nih.gov/pubmed/18186479
http://dx.doi.org/10.1002/prot.21899

27. Kuznetsov, I.B. and McDuffie, M. (2008) FlexPred: a web-server for predicting residue positions involved in conformational switches in proteins. *Bioinformation*, **3**, 134-136.
http://www.ncbi.nlm.nih.gov/pubmed/19238251
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639688

28. Benros, C., de Brevern, A.G., Etchebest, C. and Hazout, S. (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins*, **62**, 865-880.
http://www.ncbi.nlm.nih.gov/pubmed/16385557
http://dx.doi.org/10.1002/prot.20815

29. Benros, C., de Brevern, A.G. and Hazout, S. (2009) Analyzing the sequence-structure relationship of a library of local structural prototypes. *J Theor Biol*, **256**, 215-226.
http://www.ncbi.nlm.nih.gov/pubmed/18977232
http://dx.doi.org/10.1016/j.jtbi.2008.08.032

30. de Brevern, A.G. and Hazout, S. (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics*, **19**, 345-353.
http://www.ncbi.nlm.nih.gov/pubmed/12584119

31. Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern*, **43**, 59-69.

32. Kohonen, T. (2001) *Self-Organizing Maps (3rd edition)*. Springer.
33. Offmann, B., Tyagi, M. and de Brevern, A.G. (2007) Local Protein Structures. *Current Bioinformatics*, **3**, 165-202.

34. Joseph, A.P., Agarwal, G., Mahajan, S., Gelly, J.C., Swapna, L.S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadie, H. *et al.* (2010) A short survey on protein blocks. *Biophys Rev*, **2**, 137-147.
http://www.ncbi.nlm.nih.gov/pubmed/21731588
http://dx.doi.org/10.1007/s12551-010-0036-1
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3124139

35. de Brevern, A.G., Etchebest, C. and Hazout, S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**, 271-287.
http://www.ncbi.nlm.nih.gov/pubmed/11025540

36. Bornot, A., Etchebest, C. and de Brevern, A.G. (2009) A new prediction strategy for long local protein structures using an original description. *Proteins*, **76**, 570-587.
http://www.ncbi.nlm.nih.gov/pubmed/19241475
http://dx.doi.org/10.1002/prot.22370
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752736

37. Noguchi, T. and Akiyama, Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res*, **31**, 492-493.
http://www.ncbi.nlm.nih.gov/pubmed/12520060
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC165469

38. Noguchi, T., Matsuda, H. and Akiyama, Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res*, **29**, 219-220.
http://www.ncbi.nlm.nih.gov/pubmed/11125096

39. Bornot, A., Etchebest, C. and de Brevern, A.G. (2011) Predicting protein flexibility through the prediction of local structures. *Proteins*, **79**, 839-852.
http://www.ncbi.nlm.nih.gov/pubmed/21287616
http://dx.doi.org/10.1002/prot.22922

40. Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K. and Zhu, G. (2003) Improved amino acid flexibility parameters. *Protein Sci*, **12**, 1060-1072.
http://www.ncbi.nlm.nih.gov/pubmed/12717028
http://dx.doi.org/10.1110/ps.0236203
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2323876

41. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P. and Dunker, A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61 Suppl 7**, 176-182.
http://www.ncbi.nlm.nih.gov/pubmed/16187360
http://dx.doi.org/10.1002/prot.20735

42. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
http://www.ncbi.nlm.nih.gov/pubmed/9254694
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917

43. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, **39**, D214-219.
http://www.ncbi.nlm.nih.gov/pubmed/21051339
http://dx.doi.org/10.1093/nar/gkq1020
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013648

44. Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

45. R Development Core Team. (2011) In Computing, R. F. f. S. (ed.), Vienna, Austria.