The DynaMine webserver: predicting protein dynamics from sequence

Elisa Cilia[1,2,*], Rita Pancsa[3,4], Peter Tompa[2,3,4,5], Tom Lenaerts[1,2,6] and Wim F. Vranken[2,3,4]

[1]MLG, Computer Science Department, Université Libre de Bruxelles (ULB), Brussels, Belgium, [2]Interuniversity Institute of Bioinformatics in Brussels (IB[2]), ULB-VUB, Brussels, Belgium, [3]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels, Belgium, [4]Department of Structural Biology, VIB, Brussels, Belgium, [5]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary and [6]AI-Lab, Computer Science Department, Vrije Universiteit Brussel, Brussels, Belgium

ABSTRACT

Protein dynamics are important for understanding protein function. Unfortunately, accurate protein dynamics information is difficult to obtain: here we present the DynaMine webserver, which provides predictions for the fast backbone movements of proteins directly from their amino-acid sequence. DynaMine rapidly produces a profile describing the statistical potential for such movements at residue-level resolution. The predicted values have meaning on an absolute scale and go beyond the traditional binary classification of residues as ordered or disordered, thus allowing for direct dynamics comparisons between protein regions. Through this webserver, we provide molecular biologists with an efficient and easy to use tool for predicting the dynamical characteristics of any protein of interest, even in the absence of experimental observations. The prediction results are visualized and can be directly downloaded. The DynaMine webserver, including instructive examples describing the meaning of the profiles, is available at http://dynamine.ibsquare.be.

INTRODUCTION

Dynamics are fundamentally related to protein function (1). Especially intrinsically disordered proteins (2,3) are striking examples of the essential role played by dynamics. They fulfill key biological functions despite lacking a consistent three-dimensional structure, instead adopting an ensemble of conformations (4), so challenging the long-standing structure–function paradigm (5). Even though their amino-acid residues sample many different conformations, their sequence context remains important and can determine whether they prefer certain conformations to others (6). Understanding dynamics and disorder still poses significant

challenges, mainly because accurate protein dynamics information and its relation to conformation and function remain difficult to obtain, both experimentally and computationally.

Computationally, molecular dynamics is an excellent tool to obtain a timeline of protein dynamics (7,8), but it requires dedicated resources, while the results are variable and depend on the starting structures and calculation setup. Nuclear Magnetic Resonance (NMR) is the leading experimental technique to study dynamics and conformational states of proteins in solution at atomic resolution (1,9). Functionally important protein motions range from fast, small-scale fluctuations (sub-nanosecond timescales) to slow (microsecond timescale and above), global conformational transitions, like loop rearrangements or domain reorganizations. NMR spin relaxation experiments give information about fast local movements, but require substantial effort and the resulting data are not routinely deposited in public databases. Atomic-level NMR chemical shifts give instead an averaged picture of local fast dynamics up to the microsecond and low millisecond range (10) and are freely accessible and abundantly available for a diverse collection of proteins from fully structured to disordered (11).

Through statistical analysis, we have recently leveraged a large amount of NMR chemical shift data for proteins in solution to obtain a quantitative insight into the relationship between the amino-acid sequence and backbone dynamics (12). The DynaMine predictor developed from these data predicts the residue-level potential of a protein for backbone dynamics based on sequence information alone, as opposed to previous approaches (13,14). This opens up the vast amount of available protein sequences lacking structural information for dynamics analysis. In addition, we were able to show that exploiting this dynamics information is sufficient to predict disorder with accuracy comparable to the most sophisticated existing disorder predictors (15–17) without relying on any prior disorder annotation or structural information. This provided an independent evidence

*To whom correspondence should be addressed. Tel: +32 02 650 5868; Fax: +32 02 650 5609; Email: ecilia@ulb.ac.be

of the already anticipated link between dynamics and structural disorder ([18],[19]). Finally, in eight case studies covering a broad range of distinct structural and functional properties, we demonstrated the potential of the predictor in distinguishing regions of different structural organization ([12]).

Here, we present the DynaMine webserver, which incorporates this novel predictor and provides easy access to predicted dynamics profiles for protein sequences. For the webserver, we have also enabled predictions for short peptides, further validated the method, and determined ranges of predictive values where a residue is likely to be rigid, flexible, or has highly context-dependent dynamics. Through this server we aim at providing molecular biologists with an efficient and easy to use tool for estimating the dynamical characteristics of any protein of interest, even in the absence of experimental observations.

MATERIALS AND METHODS

The DynaMine webserver is implemented in Python language with the exception of JavaScript used for handling a few front-end functionalities. The webserver architecture highly decouples management of the user sessions and web interface (front-end) from the elaboration system (back-end), which includes input/output queues, scheduler and elaboration engine. In the elaboration engine reside the core DynaMine modules for prediction and results preparation. Input/output queues are accessible also to the front-end, which takes care of sending jobs to the back-end and displaying the results. Server front-end and back-end can reside on separate machines with different characteristics and performance requirements. This architecture grants the server future scalability allowing for expansion on multiple machines.

The aim of the webserver is to predict residue-level backbone dynamics from the protein sequence in the form of backbone N-H S^2 order parameters, which represent how restricted the movement of an atomic bond vector is with respect to the molecular reference frame. The values vary between 1, for fully restricted (rigid conformation), and 0, for fully random movement (highly dynamic); the DynaMine predictions represent the statistical potential for a residue to adopt certain N-H S^2 values. The S^2 estimations underlying DynaMine are based on NMR chemical shift values, and cover a timescale from femtosecond up to microseconds and low milliseconds ([10]).

The backbone N-H S^2 order parameter values resulting from the prediction (S^2_{pred}) take into consideration the local sequence environment provided by a number of residues preceding and following the target residue in the sequence. In order to extract the local sequence context of a residue, the DynaMine webserver includes a pre-processing step in which an input sequence is segmented by sliding a window of size w on each residue of the sequence. These segments provide the sequence context and hence the input features for predicting the S^2_{pred} value associated with a single residue. In the original implementation ([12]), the window w was restricted to 51 residues. In the webserver proposed here, we also provide predictions for protein sequences of shorter length, varying between 5 and 50 residues. This novel feature is provided by a collection of sibling predictors trained on the same dataset but with sliding windows w of decreasing size (from 25 to 5 in steps of 2). Every predictor consists of a linear regression model that has been retrained for the webserver on backbone S^2 values for 210 880 residues in 1952 proteins (combining previous training and test sets used in ([12])); these S^2 values (S^2_{RCI}) were estimated with the Random Coil Index (RCI) ([20]) from a carefully assembled dataset of NMR chemical shift data extracted from the BioMagResBank (BMRB) ([11]). The predictive models have been trained in the same learning setting as described in ([12]). Details on the cross-validation performance of these additional predictors are reported in the Results section.

For fully independent validation of the retrained predictors, an additional dataset of 110 sequences was compiled as described before ([12]), except that two additional filters were employed. The first one selects only BMRB entries released during 2013, which are not in our training set, and the second one checks the number of available chemical shifts per entry. If the entry contains less than 1.2 times the total number of residues C or H chemical shifts, or 0.6 times for N, the entry is discarded. This measure ensures that the RCI software can calculate good estimates of the actual S^2 value, since its performance decreases if less chemical shift data are available.

We also extracted a dataset of 1757 NMR structures from the Protein Data Bank (PDB), for which S^2_{RCI} values were available, to better qualify the twilight zone where dynamics are context-dependent (see the Results section). Based on STRIDE ([21]) and DSSP ([22],[23]) secondary structure assignments, we defined for each residue in each protein a 'unique' secondary structure code if the secondary structure assignment was fully consistent across all models in the NMR ensemble, and a 'consensus' code for the most frequently assigned secondary structure for that residue. Residues were given an 'ordered' (O) status if the same secondary structure code was uniquely assigned by either method, or if both methods assigned matching helix (H) or beta-strand (E) consensus secondary structure. This resulted in 101 570 residues with code 'O'; they are part of a well-defined secondary structure element and are unlikely to experience any conformational averaging. Residues without a uniquely assigned secondary structure code by either method, and with a non-assigned or coil (C) consensus code from both methods, were assigned as 'disordered' (D). This indicates that the structure calculation protocol could not produce a consistent answer, nor assign secondary structure for the majority of the models; 24 593 residues were assigned this 'D' status. All other 62 880 residues were considered as having undefined structure and were grouped in the gray'gray zone' (G) class.

THE WEBSERVER INTERFACE

Input

On the submission page of the webserver the user can provide as input one or more protein sequences in FASTA format, either by copy-pasting them in a text field of the submission form, or by uploading a plain text file with the information. Alternatively, the user can provide a Uniprot ID, which will be used to automatically retrieve the protein se-

quence from the Uniprot database (24). The user also has the option to ask for an email notification of the results.

An example of input is provided on the submission page by clicking on the 'Example' button; this will automatically fill in the form with the FASTA sequence or Uniprot ID of the human cellular tumor antigen p53. Details on the webserver usage can also be found on the website help pages (http://dynamine.ibsquare.be/help).

Processing phase

Once a job is submitted, the user is redirected in the same browser window to a webpage reporting the status of the job (queued, running or finished), which is automatically refreshed every 10 s until job completion. Jobs submitted to DynaMine are executed in parallel on the computing cores of the server. The DynaMine scheduler implements a First Come First Served (FCFS) policy. To guarantee a fair use of the server, a maximum of 10 sequences per submission is allowed. During the calculation, each sequence is first segmented based on a sliding window of w residues. The size w depends on the length of the input sequence. The resulting segments are then provided as input features of the central target residue to the specific DynaMine predictor trained for that segment length. The predictions are then reassembled to finally build a dynamics profile of the complete amino-acid sequence. This normally happens in the order of seconds depending on the server load and the length of the submitted sequences. If the user provided an email address, she/he will be notified as soon as the results are available by an email containing a link to a page that visualizes the results.

Output

At job completion, the user is automatically redirected to the results page, which interactively displays the predictive results. This page can be bookmarked to check the results at a later time; they are stored for 1 week. The results are presented for each submitted protein in three different ways (see Figure 1): (i) as an annotated plot of the backbone dynamics profile, (ii) as a tailored graphical representation of the protein sequence, and (iii) as a detailed report for all the residues in the sequence.

The plot of the backbone dynamics profile (i) reports the amino-acid sequence on the x axis and the S^2_{pred} values on the y axis. More rigid and more flexible regions are separated by a twilight zone (gray band in Figure 2), which we define in the Results section.

In the graphical representation of the protein sequence (ii) the size and color temperature of the residues reflect their degree of backbone flexibility (see Figure 1). Colors range from red (flexible - low S^2_{pred}) to blue (ordered - high S^2_{pred}). (iii) Detailed predictions are by default shown for residues having an S^2_{pred} value greater than 0.9; moreover, two pop-up menus provide the user with the possibility to filter the predictions with a value greater or smaller than a chosen threshold, depending on whether the user needs to focus on more rigid or more flexible regions (see Figure 1).

All the outputs described above can be downloaded separately for each submitted protein from different links on the results page. These include predictions provided as a plain text file and graphical depictions also provided in high resolution (EPS format). Alternatively, the user can download a single .zip file linked at the top of the results page that contains the results for all the sequences submitted in a single job (see Figure 1).

RESULTS

Validation, performance assessment and statistics

We reassessed the performance of the webserver predictors on the training dataset described in the Materials and Methods section (1952 proteins) through 10-fold cross-validation experiments. Figure 3 shows the cross-validation results for the different DynaMine predictors trained based on different sliding windows w (x axis). The boxplots show the distributions of the Pearson's correlation (r) (Figure 3A) and the Root Mean Square Error (RMSE) (Figure 3B) over the 10 folds. The gray dot on each boxplot represents the mean r and RMSE cross-validated performance of the corresponding DynaMine predictor. The predictors trained on w smaller or equal to 25 are used to predict the dynamics of protein sequences shorter than $w+2$ residues, with the exception of sequences between 50 and 25 residues, which are predicted by the $w = 25$ predictor. We also further validated the retrained webserver predictors on unseen data for 110 new proteins for which extensive chemical shift values became available in the BMRB during 2013 (see the Materials and Methods section). By testing DynaMine on this new dataset, we obtained a performance in line with the cross-validation results (mean $r = 0.63$ and mean RMSE = 0.14).

We then explored in more detail whether the predictions work equally well on proteins with different characteristics. We first analyzed the distributions of r and RMSE computed for each sequence separately. Figure 4 shows that the median performance on this set is actually better than the mean one ($r = 0.66$ and RMSE = 0.13); nevertheless, one can spot a few sequences (outliers) for which it is very difficult to accurately predict dynamics. We investigated on this same independent dataset the presence of a bias in the DynaMine predictions for sequences that have an associated structure in PDB (25). Figure 5 shows that DynaMine has no significant bias in its predictive performance on sequences that have an associated structure in the PDB (74 of the 110) with respect to sequences that do not have one (the remaining 36). By means of Wilcoxon tests, we could not find a statistically significant difference between the performance distributions on these two separate groups of sequences (p-value equal to 0.738 when comparing the r distributions and 0.9416 when comparing the RMSE distributions).

Statistics. DynaMine has been successfully tested on all the 541 561 sequences in UniProtKB/Swiss-Prot (release October 2013) and on 9 056 346 sequences from the D^2P^2 database (26). The webserver already received in less than 3 months since its opening online, 128 unique visitors from 29 countries, 72 cities, and 258 job submissions from outside the authors' institutions (data based on Google Analytics statistics and DynaMine server logs).
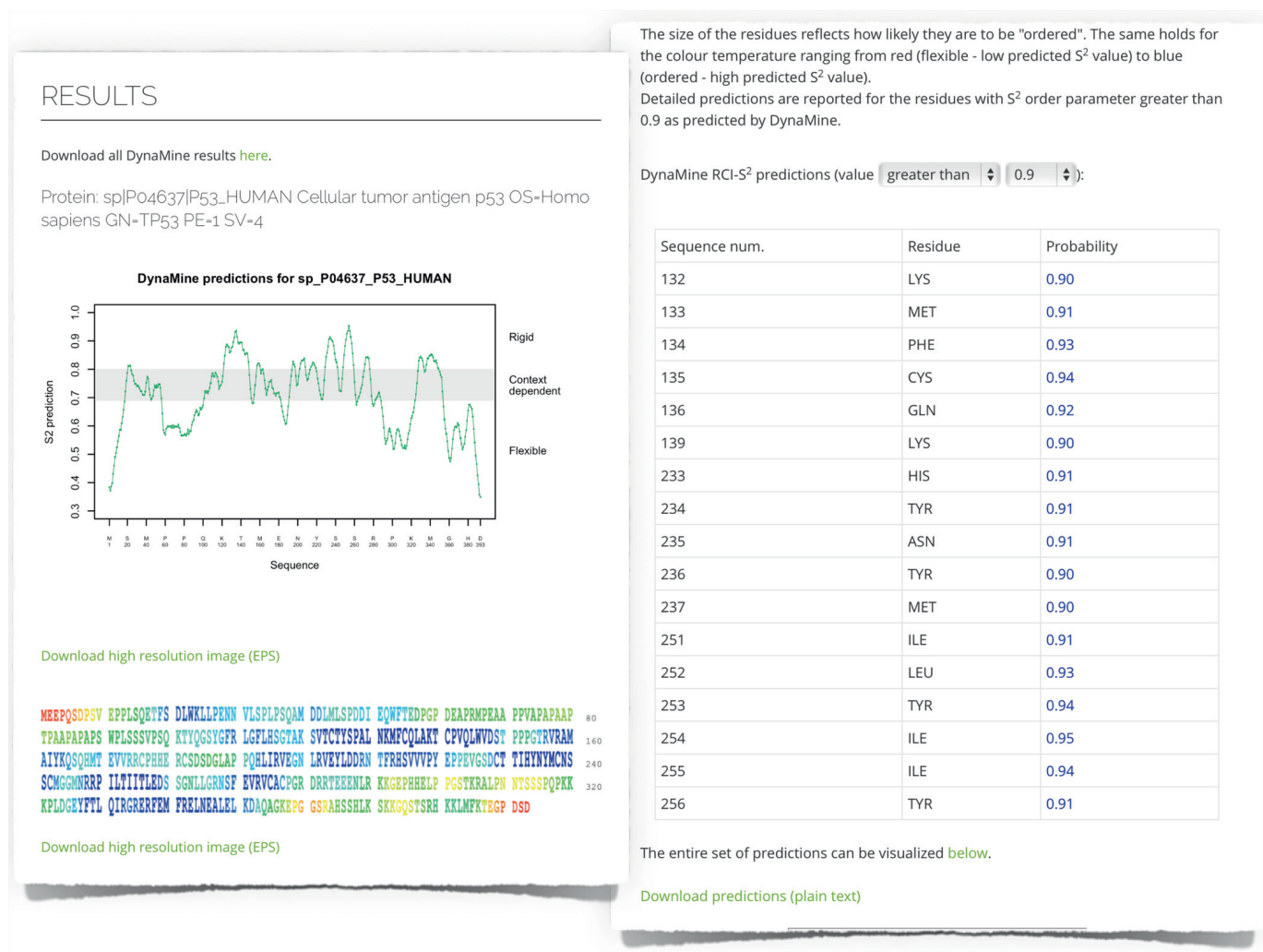
Figure 1. Output shown by the DynaMine webserver in a job results page for a sample protein sequence (human tumor suppressor p53, Uniprot ID: P04637). Graphical depictions of the predictions are shown as a dynamics profile of the amino-acid sequence and as a tailored graphical representation of the protein sequence, where the size and color temperature of the residues, ranging from red (flexible - low S^2_{pred}) to blue (ordered - high S^2_{pred}), reflect their degree of backbone flexibility. On the right-hand side, the list of residues with DynaMine predictions above a value of 0.9 is shown. Two pop-up menus are provided to filter the predictions according to the user visualization needs.
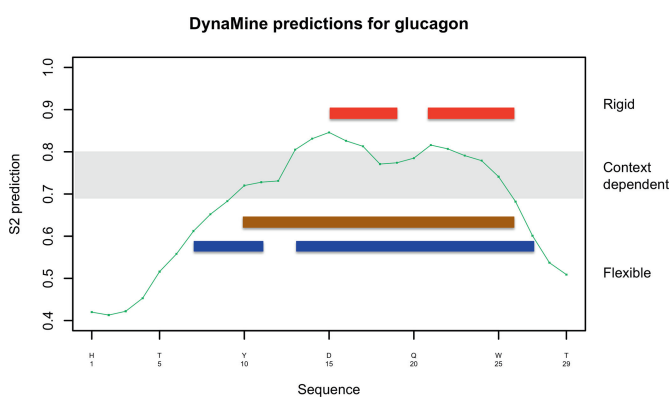

Figure 2. Prediction for the glucagon peptide as available from the webserver with added annotation for the secondary structure as seen by NMR in a lipidated analogue in water (2M5Q, red) and TFE (2M5P, blue), and by X-ray diffraction in an analogue with substitutions (1BHO, brown).

Defining the twilight zone for context-dependent dynamics

The original article (12) suggests the presence of a 'gray zone'; residues with predictions in this zone have highly context-dependent dynamics. We attempted to more accurately define this 'gray zone' based on STRIDE and DSSP secondary structure assignments for a set of NMR structures from the PDB (see the Materials and Methods section). By using per-residue assignments (built as described in the Materials and Methods section) as the expected dynamics state, we produced Receiver Operating Characteristic (ROC) curves to assess the ability of both the S^2_{RCI} and S^2_{pred} values to distinguish O from G, and G from D. The best performance point on each ROC curve was selected as the closest point to the top left corner of the plot (Table 1), and based on this analysis we currently employ the 0.80–0.69 zone as the 'gray' intermediate dynamics zone in the web plots. This classification is likely to change as we ob-
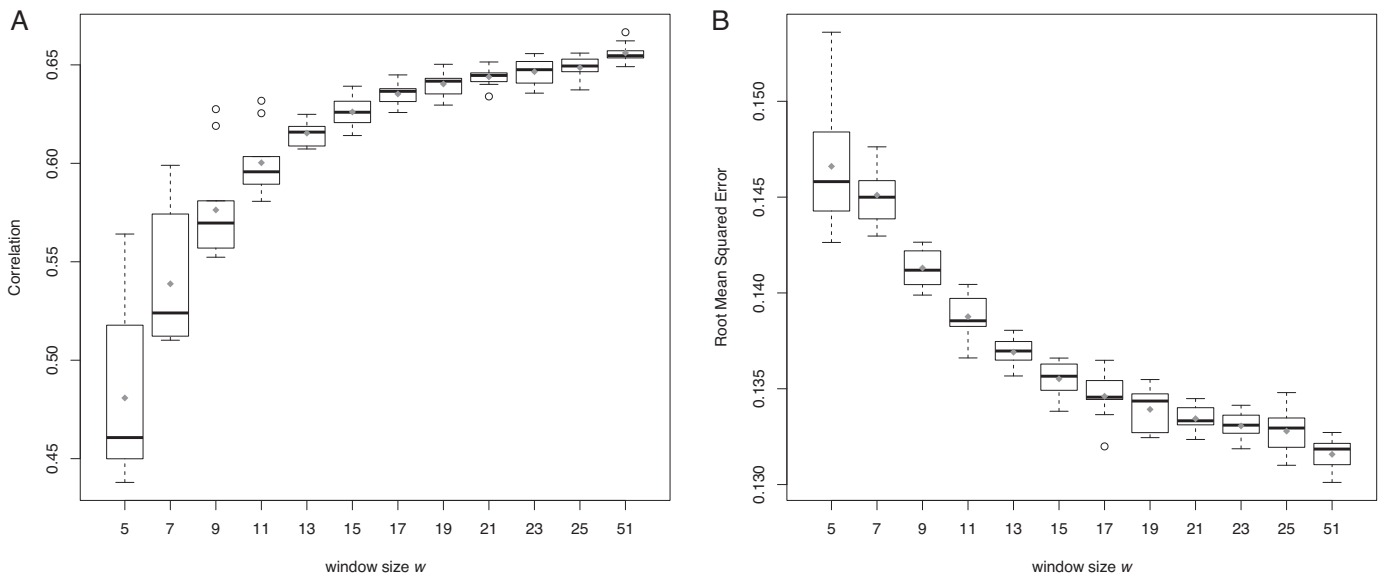
Figure 3. Boxplots showing the distributions of the Pearson's correlation (**A**) and the RMSE (**B**) over the 10 folds of the cross-validation on the webserver extended training set (1952 sequences). Mean values (average cross-validated performance) of the distributions are represented as gray dots. Each boxplot corresponds to a different DynaMine predictor trained for a specific window size *w*.
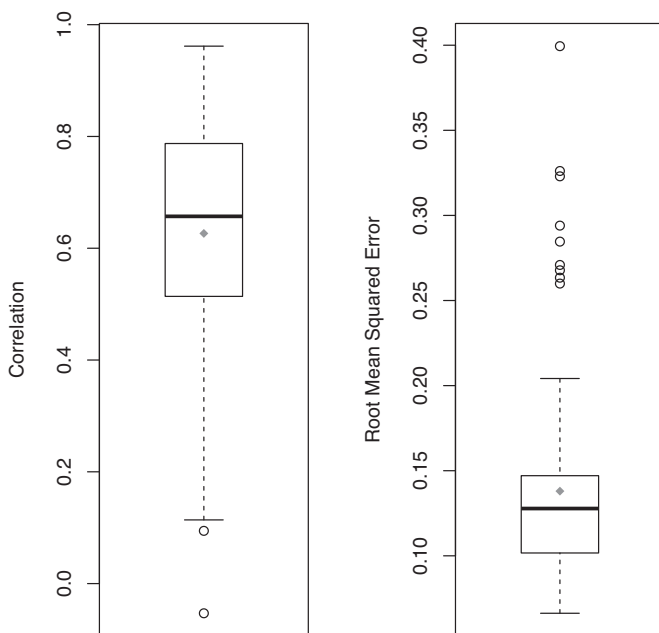

Figure 4. Boxplots of the distribution of Pearson's correlation and RMSE computed for each sequence of the independent set (110 sequences) separately. The mean of the distribution is represented with a gray dot, and the median with a black bar.

tain better data for context-dependent regions, such as for protein regions that fold upon binding.

DISCUSSION

The DynaMine webserver provides easy access to an efficient and accurate tool to predict protein backbone dynamics starting from sequence information only. Dynamics predictions are provided in the form of a dynamics pro-

file of a protein sequence. This profile can be computed even for short peptides, which tend to be more flexible and have limited structure, thanks to the inclusion of sibling predictors that were trained with varying window sizes. As a case study, we examined glucagon, a well-studied peptide hormone of 29 residues responsible for raising blood glucose levels (Figure 2). A comparison with three different structures from the PDB reveals that the predictions remain valid in this case: a lipidated analogue (red, 2M5P) in water adopts helix from residues 15–19 and 21–26, which relate well to two peaks above the 'ordered' threshold. The same analogue in the helix-inducing trifluoroethanol (TFE) co-solvent (blue, 2M5Q) adopts a much more extended helix. The lowest predicted values spanned by this helix are around 0.6, but most of the residues are encompassed in the context-dependent region. Finally, in the X-ray diffraction structure of a substituted analogue (brown, 1BHO) the helical region corresponds almost exactly to the context-dependent region. Note that in this case the crystal environment provides a changed context as compared to the training data for proteins in aqueous solution. Additional examples and case studies can be found at http://dynamine.ibsquare.be/examples. It is also worth noting that the actual dynamics observed for a protein will always be context-dependent and relate to particular (local) conditions, such as the presence of a cell membrane or ligand binding: the profiles produced by the DynaMine webserver depict the statistical potential of a sequence to adopt certain dynamics. In the provided examples and in Figure 2, we therefore illustrate a newly defined intermediate 'gray zone' of S^2_{pred} values that indicates regions of the sequence with highly context-dependent dynamic behavior, as opposed to regions highly likely to be rigid (above 0.8) or flexible (below 0.69). This underlines the potential of DynaMine in distinguishing regions of different structural organization, as already shown for proteins covering a broad range of distinct structural
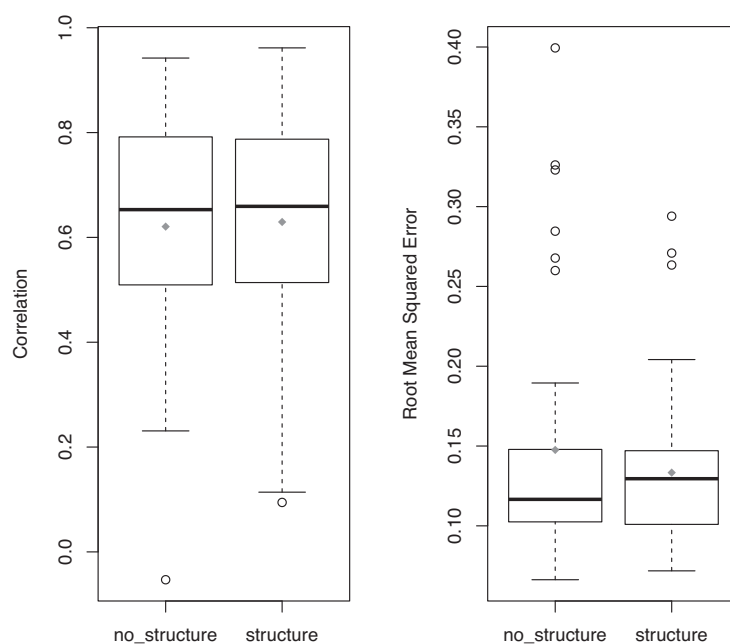
Figure 5. Boxplots of the distribution of Pearson's correlation and RMSE computed for each sequence of the independent set (110 sequences) separately and divided between entries that have an associated structure in PDB (structure) and those that do not have one (no_structure). The mean of the distribution is represented with a gray dot, and the median with a black bar.

Table 1. Cutoffs determined between the ordered and 'gray' zones of dynamics and the 'gray' and disordered zones for the S^2_{RCI} and S^2_{pred} values

	Ordered to gray	Gray to disordered	Ordered to disordered
S^2_{RCI}	0.85	0.66	0.80
S^2_{pred}	0.80	0.69	0.74

and functional properties (12). Finally, the DynaMine webserver provides visual representations of the predicted dynamics profile, which can guide its interpretation and allows direct dynamics comparison among different sequences.

FUTURE WORK

For this webserver, we plan two main future extensions. First, the extension to structure-based prediction: we plan to extract and exploit relevant structural features with the aim of improving the dynamics prediction for sequences with an associated structure. Second, we plan to extend the webserver with the possibility of comparing dynamics profiles of sequences provided in a multiple alignment. This may open the way to the analysis of the dynamics signature of protein families.

FUNDING

REFERENCES

1. Teilum,K., Olsen,J.G. and Kragelund,B.B. (2009) Functional aspects of protein flexibility. *Cell. Mol. Life Sci.*, **66**, 2231–2247.
2. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
3. Uversky,V.N. (2009) Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front. Biosci.*, **14**, 5188–5238.
4. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
5. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
6. Schweitzer-Stenner,R. (2012) Conformational propensities and residual structures in unfolded peptides and proteins. *Mol. BioSyst.*, **8**, 122–133.
7. Liang,S., Li,L., Hsu,W.-L., Pilcher,M.N., Uversky,V., Zhou,Y., Dunker,A.K. and Meroueh,S.O. (2009) Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. *Biochemistry*, **48**, 399–414.
8. de Brevern,A.G., Bornot,A., Craveur,P., Etchebest,C. and Gelly,J.C. (2012) PredyFlexy: flexibility and local structure prediction from sequence. *Nucleic Acids Res.*, **40**, W317–W322.
9. Lange,O.F., Lakomek,N.-A., Farès,C., Schröder,G.F., Walter,K.F.A., Becker,S., Meiler,J., Grubmüller,H., Griesinger,C. and de Groot,B.L. (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**, 1471–1475.
10. Berjanskii,M.V. and Wishart,D.S. (2008) Application of the random coil index to studying protein flexibility. *J. Biomol. NMR*, **40**, 31–48.
11. Ulrich,E., Akutsu,H., Doreleijers,J., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.

12. Cilia,E., Pancsa,R., Tompa,P., Lenaerts,T. and Vranken,W.F. (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.*, **4:2741**, 1-10.

13. Zhang,F. and Brüschweiler,R. (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. *J. Am. Chem. Soc.*, **124**, 12654–12655.

14. 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.

15. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

16. Ishida,T. and Kinoshita,K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, **35**, W460–W464.

17. Walsh,I., Martin,A.J., Di Domenico,T. and Tosatto,S.C. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

18. Daughdrill,G.W., Borchers,W.M. and Wu,H. (2011) Disorder predictors also predict backbone dynamics for a family of disordered proteins. *PloS One*, **6**, e29207.

19. Dyson,H.J. (2011) Expanding the proteome: disordered and alternatively folded proteins. *Q. Rev. Biophys.*, **44**, 467–518.

20. Berjanskii,M.V. and Wishart,D.S. (2007) The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Res.*, **35**, W531–W537.

21. Heinig,M. and Frishman,D. (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.*, **32**, W500–W502.

22. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

23. Joosten,R.P., te Beek,T.A., Krieger,E., Hekkelman,M.L., Hooft,R.W., Schneider,R., Sander,C. and Vriend,G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res.*, **39**, D411–D419.

24. UniProt,C. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

25. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

26. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztanyi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L. *et al.* (2013) D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.