PBxplore: a tool to analyze local protein structure and deformability with Protein Blocks

Jonathan Barnoud[1,2,3,4,5,†], **Hubert Santuz**[1,2,3,4,†], **Pierrick Craveur**[1,2,3,4,6], **Agnel Praveen Joseph**[1,2,3,4,7,+], **Vincent Jallu**[8], **Alexandre G. de Brevern**[1,2,3,4,*,‡], **and Pierre Poulain**[1,2,3,4,9,*,‡]

[1]INSERM, U 1134, DSIMB, F-75739 Paris, France.

[2]Univ. Paris Diderot, Sorbonne Paris Cité, UMR-S 1134, F-75739 Paris, France.

[3]Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.

[4]Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.

[5]Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 7, AG Groningen 9747, The Netherlands.

[6]The Scripps Research Institute, Department of Integrative Structural and Computational Biology, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

[7]Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

[8]INTS, Platelet Unit, F-75739 Paris, France.

[9]Mitochondria, Metals and Oxidative Stress Group, Institut Jacques Monod, UMR 7592, Univ. Paris Diderot, CNRS, Sorbonne Paris Cité, F-75205 Paris, France.

[†]These authors contributed equally to this work.

[‡]These authors contributed equally to this work.

[*]Corresponding authors: alexandre.de-brevern@inserm.fr, pierre.poulain@univ-paris-diderot.fr

ABSTRACT

Proteins are highly dynamic macromolecules. A classical way to analyze their inner flexibility is to perform molecular dynamics simulations. It provides pertinent results both for basic and applied researches. The different approaches to define and analyze their rigidity or flexibility have been established years ago. In this context, we present the advantage to use small structural prototypes, namely the Protein Blocks (PBs). PBs give a good approximation of the local structure of protein backbone. More importantly, they allow analyzes of local protein deformability which cannot be done with other methods and had been used efficiently in different applications. PBxplore is a suite of tools to analyze the dynamics of protein structures using PBs. It is able to process large amount of data such as those produced by molecular dynamics simulations. It produces various outputs with text and graphics, such as frequencies, entropy and information logo. PBxplore is available at https://github.com/pierrepo/PBxplore and is released under the open-source MIT license.

Introduction

Proteins are highly dynamic macromolecules[1,2]. To analyze their inner flexibility, computational biologists often use molecular dynamics (MD) simulations. The quantification of protein flexibility is based on various methods such as Root Mean Square Fluctuations (RMSF) that relies on multiple MD snapshots or Normal Mode Analysis (NMA) that relies on a single structure and focus on quantifying large movements.

Other interesting *in silico* approaches assess protein motions through the protein residue network[3] or dynamical correlations from MD simulations[4,5]. We can also notice the development of the MOdular NETwork Analysis (MONETA), which localizes the perturbations propagation throughout a protein structure[6].

However, a classical limitation of all analyzes of protein structures relies in their description. Protein structures are often considered as rigid bodies described by two regular states, namely the α-helices[7,8] and the β-sheets (composed of β-strands)[9], and one non-repetitive state, the coil (or loops)[10]. The use of only three states oversimplifies the description of protein structures[11]; 50 % of all residues are classified as coil, even when they encompass repeated local structures[12,13], emphasizing the lack of a more detailed description. To this aim, elaboration of small prototypes or ”structural alphabets” (SAs) has emerged. Protein Blocks (PBs)[14] is the most used structural alphabet[15–17]. They approximate conformations of protein backbones and code the local structures of proteins as one-dimensional sequences[18].

PBs are composed of 16 local prototypes designed through an unsupervised training performed on a representative non-redundant databank of protein structures[14]. PBs are labeled from *a* to *p* (see Figure 1a). PBs *m* and *d* can be described as prototypes for α-helix and central β-strand, respectively. PBs *a* to *c* primarily represent β-strand N-caps and PBs *e* and *f*, β-strand C-caps; PBs *g* to *j* are specific to coils, PBs *k* and *l* are specific to α-helix N-caps, and PBs *n* to *p* to α-helix C-caps[15]. Figure 1 illustrates how a PB sequence is assigned from a protein structure. Starting from the 3D coordinates of the barstar protein (Figure 1b), the local structure of each amino acid is compared to the 16 PB definitions (Figure 1a). The most similar protein block is assigned to the residue under consideration (the similarity metrics is explained in a latter section of the article). Eventually, assignment leads to the PB sequence (Figure 1c).

PBs are efficient to describe long protein fragments[19,20] and short loops[13,21,22]. They have also been used to analyze protein contacts[23], to propose a structural model of a transmembrane protein[15], to reconstruct globular protein structures[24], to design peptides[25] and to define binding site signatures[26], to perform local protein conformation predictions[27–31], to predict β-turns[32] and recently to understand local conformational changes due to mutations of the αIIbβ3 human integrin[33–35].

PBs are also useful to compare and superimpose protein structures with pairwise and multiple approaches[36,37], namely iPBA[38] and mulPBA[39], both currently showing best results compared to other superimposition methods. Eventually, PBs is also the only SA which has been used to predict protein structures from their sequences[40,41] and to predict protein flexibility[42,43].

Our results on biological systems such as, the DARC protein[44], the human αIIbβ3 integrin[33–35] and the KISSR1 protein[45], highlighted the usefulness of PBs to understand local deformations of protein structures. Specially, these analyzes have shown that a region considered as highly flexible through RMSF quantifications, can be seen through PBs as locally highly rigid. This unexpected behavior is explained by a local rigidity, surrounded by deformable regions[46]. The only other related approach based on SA is GSATools[47], it is specialized in the analysis of functional correlations between local and global motions, and the mechanisms of allosteric communication.

We thus propose PBxplore, a tool to analyze local protein structure and deformability using PBs. It is available at https://github.com/pierrepo/PBxplore. PBxplore can read PDB structure files[48], PDBx/mmCIF structure files[49], and MD trajectory formats from most MD engines, including Gromacs MD topology and trajectory files[50,51]. Starting from 3D protein structures, PBxplore assigns PBs sequences; computes a local measurement of entropy, a density map of PBs along the protein sequence and a WebLogo-like representation of PBs.

In this paper, we first present the principle of PBxplore, then its different tools, and finally a simple user-case with the $\beta3$ subunit of the human platelet integrin αIIb$\beta3$.

Design and Implementation

PBxplore is written in Python[52–54]. It is compatible with Python 2.7, and with Python 3.4 or greater. It requires the Numpy Python library for array manipulation[55], the matplotlib library for graphical representations, and the MDAnalysis library for molecular dynamics simulation files input[56]. Optionally, PBxplore functionalities can be enhanced by the installation and the use of WebLogo[57] to create sequence logos.

PBxplore is available as a set of command-line tools and as a Python module. Users less familiar with the Python programming language can use the command-line programs. These programs can be linked up together to make a structure analysis pipeline of protein flexibility. For more advanced users, PBxplore provides an API to access its core functionalities and allow creation of custom workflows.

PBxplore is released under the open-source MIT license[58]. It is available on the software development platform GitHub[59] at https://github.com/pierrepo/PBxplore. The package contains unit and regression tests and is continuously tested using Travis CI[60]. An extensive documentation is available on Read the Docs[61] at https://pbxplore.readthedocs.io.

Installation

The easiest way to install PBxplore is through the Python Package Index (PyPI):

```
pip install --user pbxplore
```

It will ensure all required dependencies are installed correctly.

Command-line Tools

A schematic description of PBxplore command line interface is provided in Figure 2. The interface is composed of three different programs: `PBassign` to assign PBs, `PBcount` to compute PBs frequency on multiple conformations, and `PBstat` to perform statistical analyses and visualization. These programs can be linked up together to make a structure analysis pipeline to study protein flexibility.

PBassign

The very first task is to assign PBs from the protein structure(s). A PB is associated to each pentapeptide included in the protein sequence. To assign a PB to a residue n, 5 residues are required (residues $n-2$, $n-1$, n, $n+1$ and $n+2$). From the structure of these 5 residues, 8 dihedral angles (ψ and ϕ) are computed, going from the ψ angle of residue $n-2$ to the ϕ angle of residue $n+2$[15]. This set of 8 dihedral angles is then compared to the reference angles set of the 16 PBs[14] using the Root Mean Square Deviation Angle (RMSDA) measure, i.e., an Euclidean distance on angles. PB with the smallest RMSDA is assigned to residue n. A dummy PB ”Z” is assigned to residues for which all 8 angles cannot be computed. Hence, the first two N-terminal and the last two C-terminal residues are always assigned to PB ”Z”.

The program `PBassign` reads one or several protein 3D structures and performs PBs assignment as one PBs sequence per input structure. `PBassign` can process multiple structures at once, either provided as individual structure files, as a directory containing many structure files or as topology and trajectory files issued from MD simulations. Note that PBxplore should be able to read any trajectory file format handled by the MDAnalysis library, yet we have tested Gromacs and CHARMM trajectories. Output PBs sequences are bundled in a single file in fasta format.

PBcount

During the course of a MD simulation, the local protein conformations can change. It is then interesting to analyze them through PB description. For that, once PBs are assigned, PBs frequencies per residue can be computed.

The program `PBcount` reads PBs sequences for different conformation of the same protein from a file in the fasta format (as outputed by `PBassign`). Many input files can be provided at once. The output data is a 2D matrix of x rows by y columns, where x is the length of the protein sequence and y is the 16 distinct PBs. A matrix element is the count of a given PB at a given position in the protein sequence.

PBstat

The number of possible conformational states covered by PBs is higher than the classical secondary structure description (16 states instead of 3). As a consequence, the amount of information produced by `PBcount` can be complex to handle. Hence, we propose three simple ways to visualize the variation of PBs which occur during a MD simulation.

The program `PBstat` reads PBs frequencies as computed by `PBcount`. It can produce three types of outputs based on the input argument(s). The first two use the matplotlib library and the last one requires the installation of the third-party tool Weblogo[57]. `PBstat` offers also two options (`--residue-min` and `--residue-max`) to define a residue frame allowing the user to quickly look at segments of interest. The three graphical representations proposed are:

- *Distribution of PBs.* This feature plots the frequency of each PB along the protein sequence. The output file could be in format .png, .jpg or .pdf. A dedicated colorblind safe color range[62] allows visualizing the distribution of PBs. For a given position in the protein sequence, blue corresponds to a null frequency when the particular PB is never met at this position and red corresponds to a frequency of 1 when the particular PB is always found at this position. It is produced with the `--map` argument.

- *Equivalent number of PBs (N_{eq}).* The N_{eq} is a statistical measurement similar to entropy[18]. It represents the average number of PBs taken by a given residue. N_{eq} is calculated as follows:

  $$N_{eq} = \exp(-\sum_{i=1}^{16} f_x \ln f_x)$$

  where f_x is the probability (or frequency) of the PB x. A N_{eq} value of 1 indicates that only a single type of PB is observed, while a value of 16 is equivalent to a random distribution, i.e. all PBs are observed with the same frequency 1/16. For example, a N_{eq} value around 5 means that, across all the PBs observed at the position of interest, 5 different PBs are mainly observed. If the N_{eq} exactly equals to 5, this means that 5 different PBs are observed in equal proportions (i.e. 1/5).

  A high N_{eq} value can be associated with a local deformability of the structure whereas a N_{eq} value close to 1 means a rigid structure. In the context of structures issued from MD simulations, the concept of deformability / rigidity is independent to the one of mobility. The distribution of PBs is produced with the `--neq` argument.

- *Logo representation of PBs frequency.* This is a WebLogo-like representation[57] of PBs sequences. The size of each PB is proportional to its frequency at a given position in the sequence. This type of representation is useful to pinpoint PBs patterns. This WebLogo-like representation is produced with the `--logo` argument.

Python Module

PBxplore is also a Python module that more advanced users can embed in their own Python script. Here is a Python 3 example that assigns PBs from the structure of the barstar ribonuclease inhibitor[63]:

```
import urllib.request
import pbxplore as pbx


# Download the pdb file
urllib.request.urlretrieve('https://files.rcsb.org/view/1BTA.pdb', '1BTA.pdb')


# The function pbx.chain_from_files() reads a list of files
# and for each one returns the chain and its name.
for chain_name, chain in pbx.chains_from_files(['1BTA.pdb']):
    # Compute phi and psi angles
    dihedrals = chain.get_phi_psi_angles()
    # Assign PBss
    pb_seq = pbx.assign(dihedrals)
    print('PBs sequence for chain {}:\n{}'.format(chain_name, pb_seq))
```

The documentation contains complete and executable Jupyter notebooks explaining how to properly use the module. It goes from the PBs assignments to the visualization of the protein deformability using the analysis functions. This allows the user to quickly understand the architecture of the module.

Results

This section aims at giving the reader a quick tour of PBxplore features on a real-life example. We will focus on the $\beta 3$ subunit of the human platelet integrin αIIb$\beta 3$ that plays a central role in hemostasis and thrombosis. The $\beta 3$ subunit has also been reported in cases of alloimmune thrombocytopenia[64, 65]. We studied recently this protein by MD simulations (for more details, see references[33–35]).

The $\beta 3$ integrin subunit structure[66] comes from the structure of the integrin complex (PDB 3FCS[67]). Final structure has 690 residues and was used for MD simulations. All files mentioned below are available in the `demo_paper` directory from the GitHub repository (https://github.com/pierrepo/PBxplore/tree/master/demo_paper).

Protein Blocks assignment

The initial file `beta3.pdb` contains 225 structures issued from a single 50 ns MD simulation of the $\beta 3$ integrin.

```
PBassign -p beta3.pdb -o beta3
```

This instruction generates the file `beta3.PB.fasta`. It contains as many PB sequences as there are structures in the input `beta3.pdb` file.

Protein Blocks assignment is the slowest step. In this example, it took roughly 80 seconds on a laptop with a quad-core-1.6-GHz processor.

Protein Blocks frequency

```
PBcount -f beta3.PB.fasta -o beta3
```

The above command line produces the file `beta3.PB.count` that contains a 2D-matrix with 16 columns (as many as different PBs) and 690 rows (one per residue) plus one supplementary column for residue number and one supplementary row for PBs labels.

Statistical analysis

Distribution of PBs

```
PBstat -f beta3.PB.count -o beta3 --map
```

Figure 3 shows the distribution of PBs for the $\beta3$ integrin. The color scale ranges from blue (the PB is not found at this position) to red (the PB is always found at this position). The $\beta3$ protein counts 690 residues. This leads to a cluttered figure and prevents getting any details on a specific residue (Figure 3a). However, it exhibits some interesting patterns colored in red that correspond to series of neighboring residues exhibiting a fixed PB during the entire MD simulation. See for instance patterns associated to PBs *d* and *m* that reveal β-sheets and α-helices secondary structures[15].

With a large protein such as this one, it is better to look at limited segments. A focus on the PSI domain (residue 1 to 56)[33,67] of the $\beta3$ integrin was achieved with the command:

```
PBstat -f beta3.PB.count -o beta3 --map --residue-min 1 --residue-max 56
```

Figure 3b shows the PSI domain dynamics in terms of PBs. Interestingly, residue 33 is the site of the human platelet antigen (HPA)-1 alloimmune system. It is the first cause of alloimmune thrombocytopenia in Caucasian populations and a risk factor for thrombosis[64,65]. In Figure 3b, this residue occupies a stable conformation with PB *h*. Residues 33 to 35 define a stable core composed of PBs *h*-*i*-*a*. This core is found in all of the 255 conformations extracted from the MD simulation and then is considered as highly rigid. On the opposite, residue 52 is flexible as it is found associated to PBs *i*, *j*, *k* and *l* corresponding to coil and α-helix conformations.

Equivalent number of PBs

The N_{eq} is a statistical measurement similar to entropy and is related to the flexibility of a given residue. The higher is the value, the more flexible is the backbone. The N_{eq} for the PSI domain (residue 1 to 56) was obtained from the command line:

```
PBstat -f beta3.PB.count -o beta3 --neq --residue-min 1 --residue-max 56
```

The output file `beta3.PB.Neq.1-56` contains two columns, corresponding to the residue numbers and the N_{eq} values. Figure 4a represents the N_{eq} along with the PBs sequence of the PSI domain, as generated by `PBstat`. The rigid region 33-35 and the flexible residue 52 are easily spotted, with low N_{eq} values for the former and a high N_{eq} value for the latter.

An interesting point, seen in our previous studies, is that the region delimited by residues 33 to 35 was shown to be highly mobile by the RMSF analysis we performed in Jallu et al.[33] (for more details, see Materials and Methods section in Jallu et al.[33]). For comparison, RMSF and N_{eq} are represented on the same graph on Figure 4b. This high mobility was correlated with the location of this region in a loop, which globally moved a lot in our MD simulations. Here, we observe that the region 33-35 is rigid. The high values of RMSF we observed in our previous work were due to flexible residues in the vicinity of the region 33-35, probably acting as hinges (residues 32 and 36–37). Understanding the flexibility of residues 33 to 35 is important since this region defines the HPA-1 alloantigenic system involved in severe cases of alloimmune thrombocytopenia.

PBxplore allows discriminating between flexible and rigid residues; the N_{eq} is a metric of deformability and flexibility whereas RMSF quantifies mobility.

Logo representation of PBs frequency

While the N_{eq} analysis focuses on the flexibility of amino acids, the WebLogo-like representation[57] aims at identifying the diversity of PBs and their frequencies at a given position in the protein sequence. With a focus on the PSI domain, the following command line was used:

```
PBstat -f beta3.PB.count -o beta3 --logo --residue-min 1 --residue-max 56
```

Figure 5 represents PBs found at a given position. The rigid region 33-35 is composed of a succession of PBs *h*-*i*-*a* while the flexible residue 52 is associated to PBs *i*, *j*, *k* and *l*. This third representation summarized pertinent information, as shown in ref[34].

Conclusion

From our previous works[33–35, 45], we have seen the usefulness of a tool dedicated to the analysis of local protein structures and deformability with PBs. We also showed the relevance of studying molecular deformability in the scope of structures issued from molecular dynamics simulations. Thus, we propose to the community PBxplore, available at https://github.com/pierrepo/PBxplore. PBxplore is written in a modular fashion that allows embedding in any PBs related Python applications.

Software Availability

PBxplore is released under the open-source MIT license[58]. Its source code can be freely downloaded from the GitHub repository of the project: https://github.com/pierrepo/PBxplore. In addition, the present version of PBxplore (1.3.6) is also archived in the digital repository Zenodo[68].

References

1. Frauenfelder, H., Sligar, S. & Wolynes, P. The energy landscapes and motions of proteins. *Sci.* **254**, 1598–1603 (1991). DOI 10.1126/science.1749933.

2. Bu, Z. & Callaway, D. J. Proteins MOVE! Protein dynamics and long-range allostery in cell signaling. In *Advances in Protein Chemistry and Structural Biology*, vol. 83, 163–221 (Elsevier, 2011). DOI 10.1016/B978-0-12-381262-9.00005-7.

3. Atilgan, A. R., Turgut, D. & Atilgan, C. Screened Nonbonded Interactions in Native Proteins Manipulate Optimal Paths for Robust Residue Communication. *Biophys. J.* **92**, 3052–3062 (2007). DOI 10.1529/biophysj.106.099440.

4. Ghosh, A. & Vishveshwara, S. A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci.* **104**, 15711–15716 (2007). DOI 10.1073/pnas.0704459104.

5. Dixit, A. & Verkhivker, G. M. Computational Modeling of Allosteric Communication Reveals Organizing Principles of Mutation-Induced Signaling in ABL and EGFR Kinases. *PLoS Comput. Biol.* **7**, e1002179 (2011). DOI 10.1371/journal.pcbi.1002179.

6. Laine, E., Auclair, C. & Tchertanov, L. Allosteric Communication across the Native and Mutated KIT Receptor Tyrosine Kinase. *PLoS Comput. Biol.* **8**, e1002661 (2012). DOI 10.1371/journal.pcbi.1002661.

7. Pauling, L. & Corey, R. B. Two Hydrogen-Bonded Spiral Configurations of the Polypetide Chain. *J. Am. Chem. Soc.* **72**, 5349–5349 (1950). DOI 10.1021/ja01167a545.

8. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci.* **37**, 205–211 (1951). DOI 10.1073/pnas.37.4.205.

9. Pauling, L. & Corey, R. B. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proc. Natl. Acad. Sci.* **37**, 251–256 (1951). DOI 10.1073/pnas.37.5.251.

10. Eisenberg, D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci.* **100**, 11207–11210 (2003). DOI 10.1073/pnas.2034522100.

11. Richardson, J. S. The Anatomy and Taxonomy of Protein Structure. In *Advances in Protein Chemistry*, vol. 34, 167–339 (Elsevier, 1981).

12. Thornton, J. M., Sibanda, B. L., Edwards, M. S. & Barlow, D. J. Analysis, design and modification of loop regions in proteins. *BioEssays* **8**, 63–69 (1988). DOI 10.1002/bies.950080205.

13. Fourrier, L., Benros, C. & de Brevern, A. G. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC bioinformatics* **5**, 58 (2004). DOI 10.1186/1471-2105-5-58.

14. de Brevern, A. G., Etchebest, C. & Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41**, 271–287 (2000). DOI 10.1002/1097-0134(20001115)41:3¡271::AID-PROT10¿3.0.CO;2-Z.

15. de Brevern, A. G. New assessment of a structural alphabet. *In Silico Biol.* **5**, 283–289 (2005).

16. Etchebest, C., Benros, C., Hazout, S. & de Brevern, A. G. A structural alphabet for local protein structures: Improved prediction methods. *Proteins: Struct. Funct. Bioinforma.* **59**, 810–827 (2005). DOI 10.1002/prot.20458.

17. Joseph, A. P. *et al.* A short survey on protein blocks. *Biophys. Rev.* **2**, 137–147 (2010). DOI 10.1007/s12551-010-0036-1.

18. Offmann, B., Tyagi, M. & de Brevern, A. Local Protein Structures. *Curr. Bioinforma.* **2**, 165–202 (2007). DOI 10.2174/157489307781662105.

19. de Brevern, A. G., Valadié, H., Hazout, S. & Etchebest, C. Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. *Protein Sci.* **11**, 2871–2886 (2002). DOI 10.1110/ps.0220502.

20. Bornot, A., Etchebest, C. & de Brevern, A. G. A new prediction strategy for long local protein structures using an original description. *Proteins* **76**, 570–587 (2009). DOI 10.1002/prot.22370.

21. Tyagi, M., Bornot, A., Offmann, B. & de Brevern, A. G. Protein short loop prediction in terms of a structural alphabet. *Comput. Biol. Chem.* **33**, 329–333 (2009). DOI 10.1016/j.compbiolchem.2009.06.002.

22. Tyagi, M., Bornot, A., Offmann, B. & de Brevern, A. G. Analysis of loop boundaries using different local structure assignment methods. *Protein Sci.* **18**, 1869–1881 (2009). DOI 10.1002/pro.198.

23. Faure, G., Bornot, A. & de Brevern, A. G. Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie* **90**, 626–639 (2008). DOI 10.1016/j.biochi.2007.11.007.

24. Dong, Q.-w., Wang, X.-l. & Lin, L. Methods for optimizing the structure alphabet sequences of proteins. *Comput. Biol. Medicine* **37**, 1610–1616 (2007). DOI 10.1016/j.compbiomed.2007.03.002.

25. Thomas, A. *et al.* Prediction of peptide structure: How far are we? *Proteins: Struct. Funct. Bioinforma.* **65**, 889–897 (2006). DOI 10.1002/prot.21151.

26. Dudev, M. & Lim, C. Discovering structural motifs using a structural alphabet: Application to magnesium-binding sites. *BMC Bioinforma.* **8**, 106 (2007). DOI 10.1186/1471-2105-8-106.

27. Li, Q., Zhou, C. & Liu, H. Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. *Proteins: Struct. Funct. Bioinforma.* **74**, 820–836 (2009). DOI 10.1002/prot.22191.

28. Rangwala, H., Kauffman, C. & Karypis, G. svmPRAT: SVM-based Protein Residue Annotation Toolkit. *BMC Bioinforma.* **10**, 439 (2009). DOI 10.1186/1471-2105-10-439.

29. Suresh, V., Ganesan, K. & Parthasarathy, K. A Protein Block Based Fold Recognition Method for the Annotation of Twilight Zone Sequences. *Protein Pept Lett* **20**, 249–254 (2013).

30. Suresh, V. & Parthasarathy, S. SVM-PB-Pred: SVM Based Protein Block Prediction Method Using Sequence Profiles and Secondary Structures. *Protein & Pept. Lett.* **21**, 736–742 (2014). DOI 10.2174/09298665113209990064.

31. Zimmermann, O. & Hansmann, U. H. E. LOCUSTRA: Accurate Prediction of Local Protein Structure Using a Two-Layer Support Vector Machine Approach. *J. Chem. Inf. Model.* **48**, 1903–1908 (2008). DOI 10.1021/ci800178a.

32. Nguyen, L. A. T. *et al.* Predicting *B*eta-Turns and *B*eta-Turn Types Using a Novel Over-Sampling Approach. *J. Biomed. Sci. Eng.* **07**, 927–940 (2014). DOI 10.4236/jbise.2014.711090.

33. Jallu, V., Poulain, P., Fuchs, P. F. J., Kaplan, C. & de Brevern, A. G. Modeling and molecular dynamics of HPA-1a and -1b polymorphisms: Effects on the structure of the *b*3 subunit of the αIIbβ3 integrin. *PloS One* **7**, e47304 (2012). DOI 10.1371/journal.pone.0047304.

34. Jallu, V. *et al.* The αIIb p.Leu841Met (Cab3a+) polymorphism results in a new human platelet alloantigen involved in neonatal alloimmune thrombocytopenia. *Transfus.* **53**, 554–563 (2013). DOI 10.1111/j.1537-2995.2012.03762.x.

35. Jallu, V., Poulain, P., Fuchs, P. F. J., Kaplan, C. & de Brevern, A. G. Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit *b*3: Structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants. *Biochimie* **105**, 84–90 (2014). DOI 10.1016/j.biochi.2014.06.017.

36. Joseph, A. P., Srinivasan, N. & de Brevern, A. G. Improvement of protein structure comparison using a structural alphabet. *Biochimie* **93**, 1434–1445 (2011). DOI 10.1016/j.biochi.2011.04.010.

37. Joseph, A. P., Srinivasan, N. & de Brevern, A. G. Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies. *Biochimie* **94**, 2025–2034 (2012). DOI 10.1016/j.biochi.2012.05.028.

38. Gelly, J.-C., Joseph, A. P., Srinivasan, N. & de Brevern, A. G. iPBA: A tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res.* **39**, W18–W23 (2011). DOI 10.1093/nar/gkr333.

39. Léonard, S., Joseph, A. P., Srinivasan, N., Gelly, J.-C. & de Brevern, A. G. mulPBA: An efficient multiple protein structure alignment method based on a structural alphabet. *J. Biomol. Struct. Dyn.* **32**, 661–668 (2014). DOI 10.1080/07391102.2013.787026.

40. Ghouzam, Y., Postic, G., de Brevern, A. G. & Gelly, J.-C. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. *Bioinforma.* btv462 (2015). DOI 10.1093/bioinformatics/btv462.

41. Ghouzam, Y., Postic, G., Guerin, P.-E., de Brevern, A. G. & Gelly, J.-C. ORION: A web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci. Reports* **6** (2016). DOI 10.1038/srep28268.

42. Bornot, A., Etchebest, C. & de Brevern, A. G. Predicting protein flexibility through the prediction of local structures. *Proteins* **79**, 839–852 (2011). DOI 10.1002/prot.22922.

43. de Brevern, A. G., Bornot, A., Craveur, P., Etchebest, C. & Gelly, J.-C. PredyFlexy: Flexibility and local structure prediction from sequence. *Nucleic Acids Res.* **40**, W317–322 (2012). DOI 10.1093/nar/gks482.

44. de Brevern, A. *et al.* A structural model of a seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC). *Biochimica et Biophys. Acta (BBA) - Gen. Subj.* **1724**, 288–306 (2005). DOI 10.1016/j.bbagen.2005.05.016.

45. Chevrier, L. *et al.* PRR Repeats in the Intracellular Domain of KISS1R Are Important for Its Export to Cell Membrane. *Mol. Endocrinol.* **27**, 1004–1014 (2013). DOI 10.1210/me.2012-1386.

46. Craveur, P. *et al.* Protein flexibility in the light of structural alphabets. *Front. Mol. Biosci.* **2** (2015). DOI 10.3389/fmolb.2015.00020.

47. Pandini, A., Fornili, A., Fraternali, F. & Kleinjung, J. GSATools: Analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinforma.* **29**, 2053–2055 (2013). DOI 10.1093/bioinformatics/btt326.

48. Bernstein, F. C. *et al.* The Protein Data Bank: A computer-based archival file for macromolecular structures. *J.Mol. Biol.* **112**, 535–542 (1977).

49. Bourne, P. E. *et al.* [30] Macromolecular crystallographic information file. In *Methods in Enzymology*, vol. 277, 571–590 (Elsevier, 1997).

50. Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **7**, 306–317 (2001). DOI 10.1007/s008940100045.

51. van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J Comput. Chem* **26**, 1701–1718 (2005). DOI 10.1002/jcc.20291.

52. van Rossum, G. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam (1995).

53. Software, F. P. Python Language Reference, version 2.7. Tech. Rep. (2010).

54. Bassi, S. A primer on python for life science researchers. *PLoS Comput. Biol.* **3**, e199 (2007). DOI 10.1371/journal.pcbi.0030199.

55. Ascher, D., Dubois, P. F., Hinsen, K., James, J. H. & Oliphant, T. Numerical Python. Tech. Rep., Lawrence Livermore National Laboratory, Livermore, CA (1999).

56. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011). DOI 10.1002/jcc.21787.

57. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **14**, 1188–1190 (2004). DOI 10.1101/gr.849004.

58. Initiative, O. S. The MIT License (MIT). Tech. Rep. (2014).

59. GitHub. GitHub. https://github.com/ (2007).

60. CI, T. Travis CI. https://travis-ci.org/ (2015).

61. Holscher, E., Leifer, C. & Grace, B. Read the Docs (2010).

62. Brewer, C., Harrower, M., Sheesley, B., Woodruff, A. & Heyman, D. ColorBrewer2 (2013).

63. Lubienski, M. J., Bycroft, M., Freund, S. M. & Fersht, A. R. Three-dimensional solution structure and 13C assignments of barstar using nuclear magnetic resonance spectroscopy. *Biochem.* **33**, 8866–8877 (1994).

64. Kaplan, C. Neonatal alloimmune thrombocytopenia. In *Thrombocytopenia*, 223–244 (McCrae KR, 2006), taylor & francis group edn.

65. Kaplan, C. & Freedman, J. Platelets. In *Platelets*, 971–984 (Michelson AD, London: Academic Press, 2007).

66. Poulain, P. & de Brevern, A. G. Model of the Beta3 Subunit of Integrin alphaIIb/beta3. https://dx.doi.org/10.6084/m9.figshare.104602.v2 (2012).

67. Zhu, J. *et al.* Structure of a Complete Integrin Ectodomain in a Physiologic Resting State and Activation and Deactivation by Applied Forces. *Mol. Cell* **32**, 849–861 (2008). DOI 10.1016/j.molcel.2008.11.018.

68. Barnoud, J., Santuz, H., de Brevern, A. G. & Poulain, P. Pbxplore: V1.3.5. *Zenodo* (2017). DOI 10.5281/zenodo.546094.

69. Sevcík, J., Urbanikova, L., Dauter, Z. & Wilson, K. S. Recognition of RNase Sa by the inhibitor barstar: Structure of the complex at 1.7 A resolution. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **54**, 954–963 (1998).

70. DeLano, W. L. *The PyMOL Molecular Graphics System*, vol. Version 1.5.0.4 (Schrödinger, LLC, 2002). On World Wide Web http://www.pymol.org.

Acknowledgements

This work was supported by grants from National Institute for Blood Transfusion (INTS, France) and Lab of Excellence GR-Ex to JB, PC, SL, APJ, VJ, AGdB and PP, from the Ministry of Research (France), University Paris Diderot, Sorbonne Paris Cité (France), National Institute for Health and Medical Research (INSERM, France) to JB, PC, SL, APJ, AGdB and PP. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program ”Investissements d'avenir” of the French National Research Agency, reference ANR-11-IDEX-0005-02. AGdB acknowledges to Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grant (number 5302-2).

Author contributions

PP and AGdB conceived the project. PP, JB and HS wrote the software. AGdB, PC, APJ and VJ improved and tested the software. All authors reviewed the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Figure legends

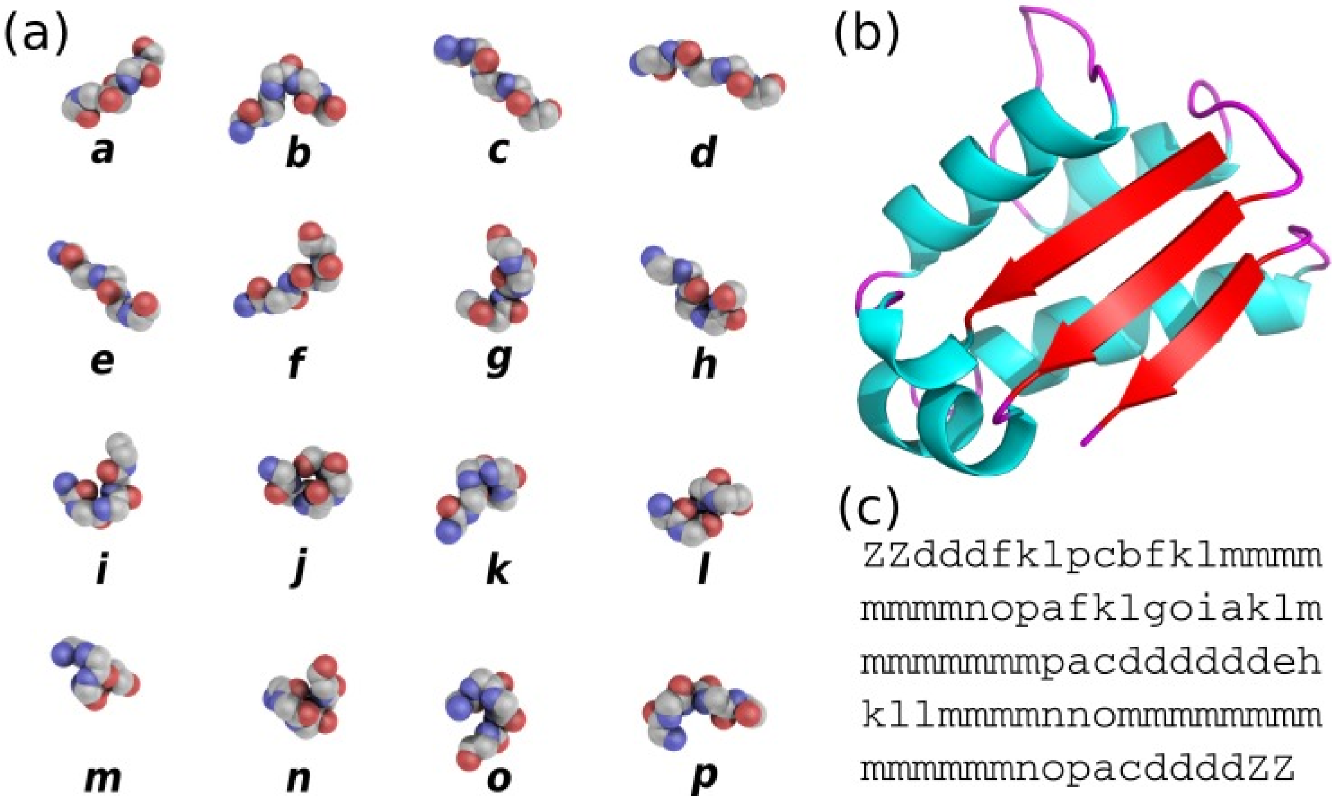

Figure 1. (a) The 16 protein blocks (PBs) represented in balls with carbon atoms in gray, oxygen atoms in red and nitrogen atoms in purple (hydrogen atoms are not represented). (b) The barstar protein (PDB ID 1AY7[69]) represented in cartoon with alpha-helices in blue, beta-strands in red and coil in pink. These representations were generated using PyMOL software[70] (c) PBs sequence obtained from PBs assignment. Z is a dummy PB meaning that no PB can be assigned to this position.

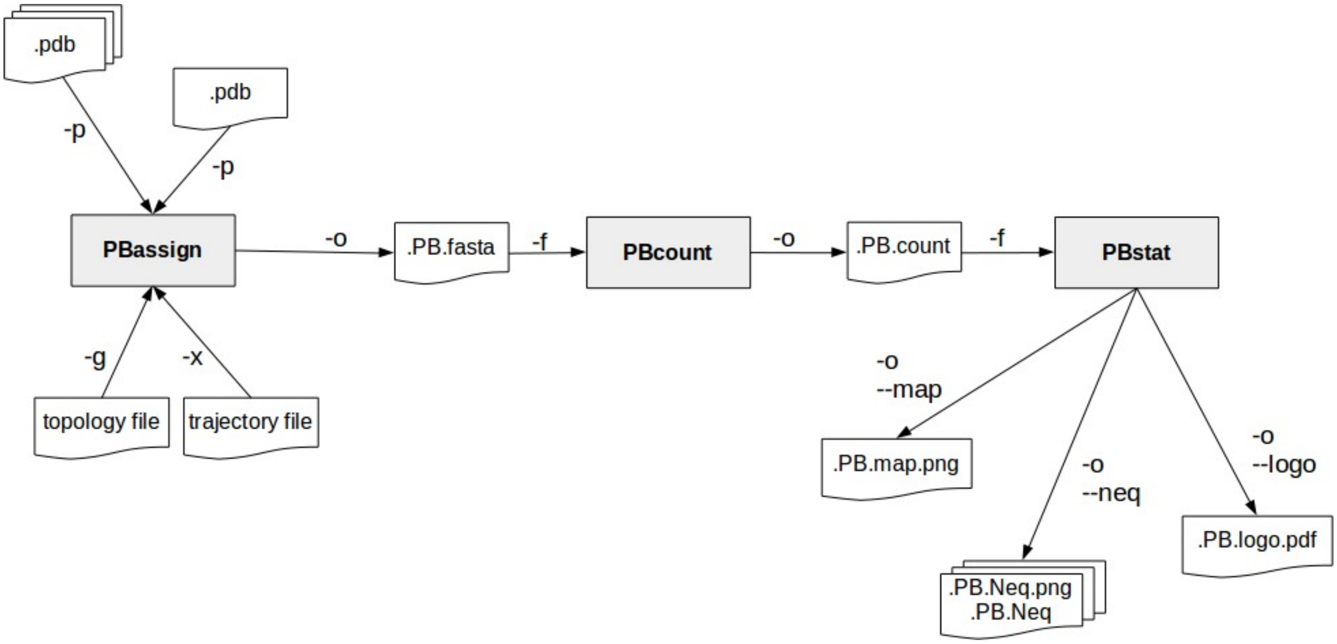

Figure 2. PBxplore is based on 3 programs that can be chained to build a structure analysis pipeline. Main input file types (.pdb, MD trajectory, MD topology), output files (.fasta, .png, .Neq, .pdf) and parameters (beginning with a single or double dash) are indicated.

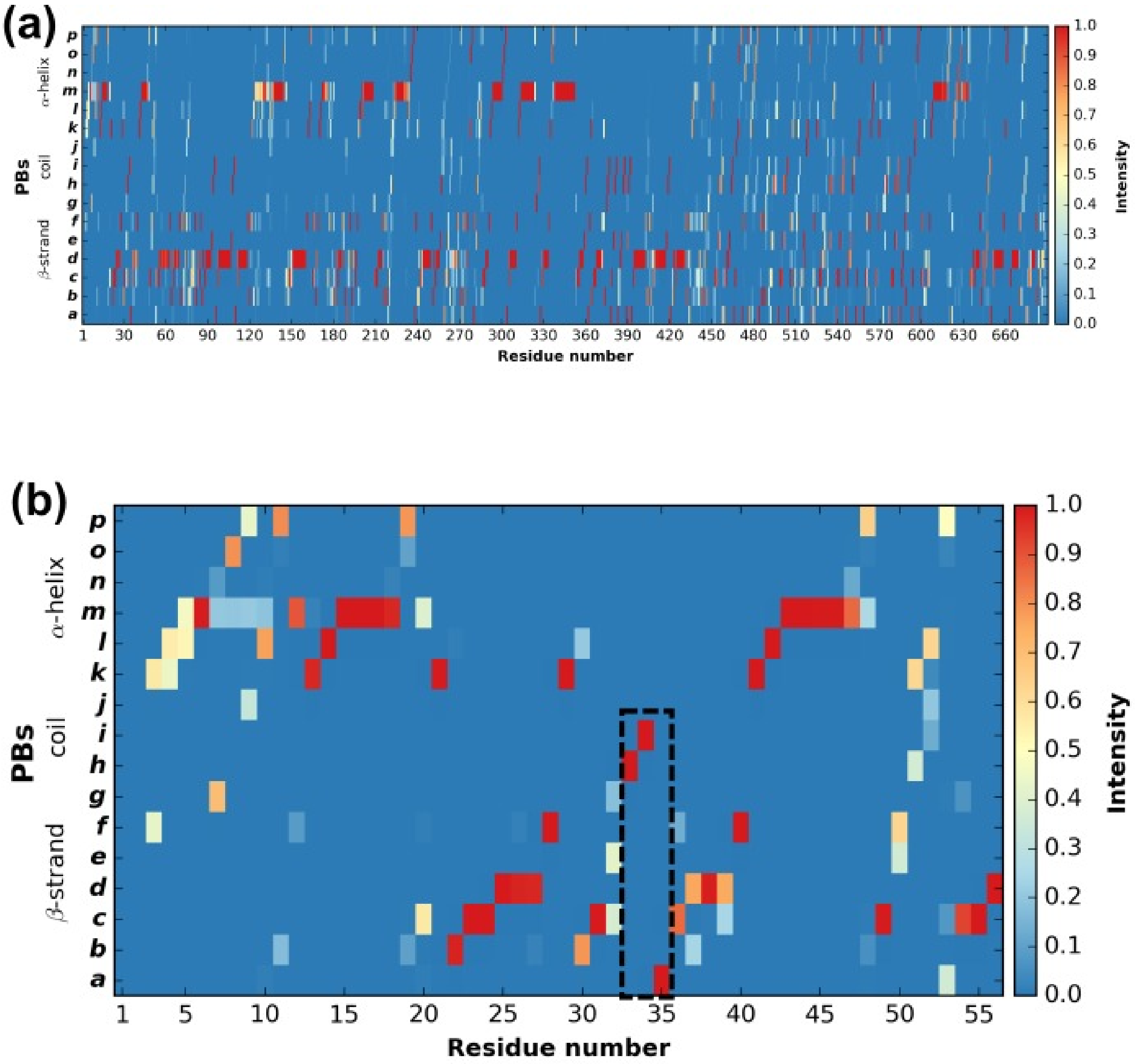

Figure 3. Distribution of PBs for the $\beta3$ integrin along the protein sequence. On the x-axis are found the 690 position residues and on the y-axis the 16 consecutive PBs from a to p (the two first and two last positions associated to ”Z” have no assignment). (a) For the entire protein. (b) For the PSI domain only (residues 1 to 56). The dashed zone pinpoints residue 33 to 35.

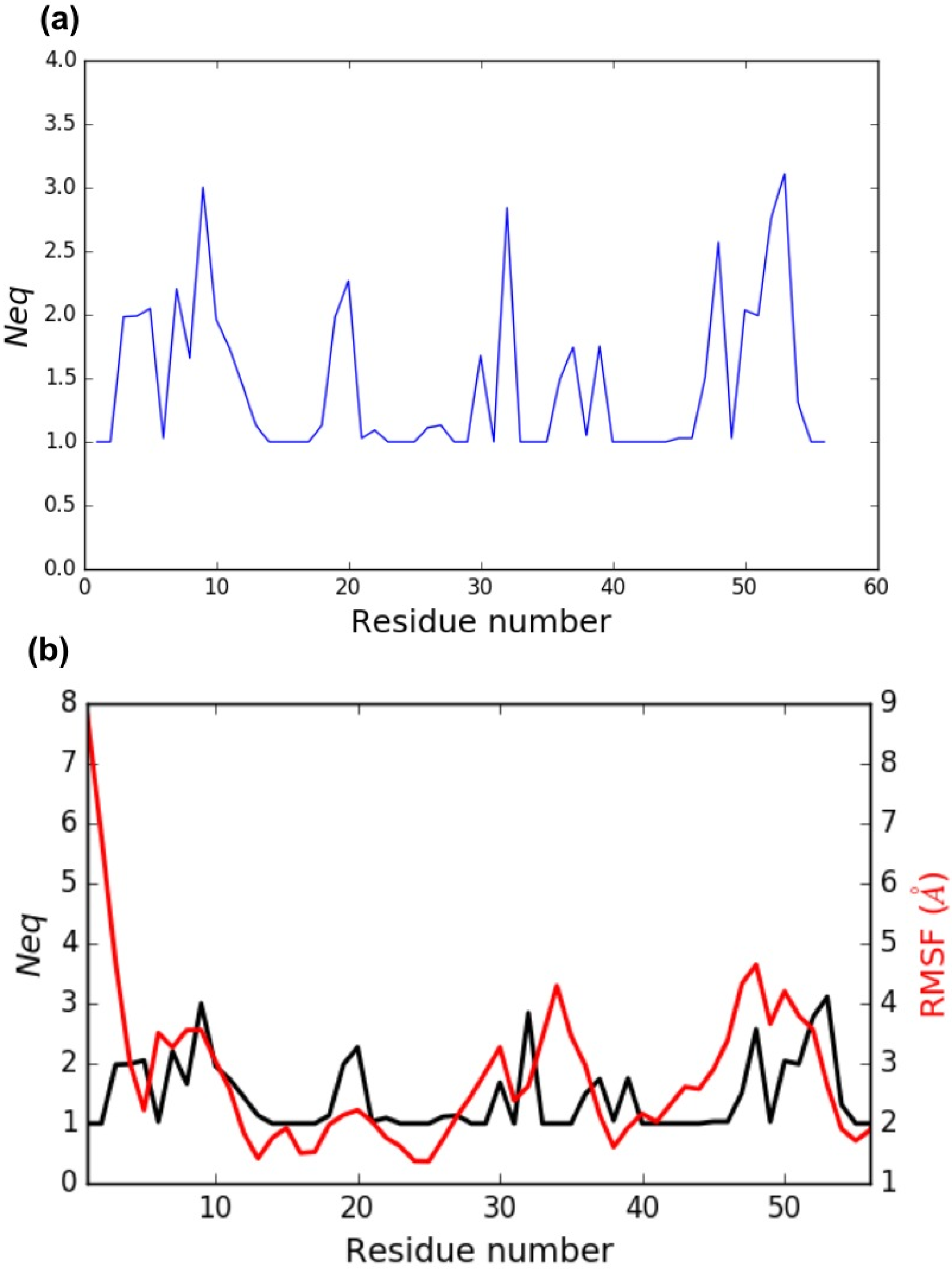

Figure 4. (a) N_{eq} versus residue number for the PSI domain (residues 1 to 56). (b) Comparison between RMSF and N_{eq}.

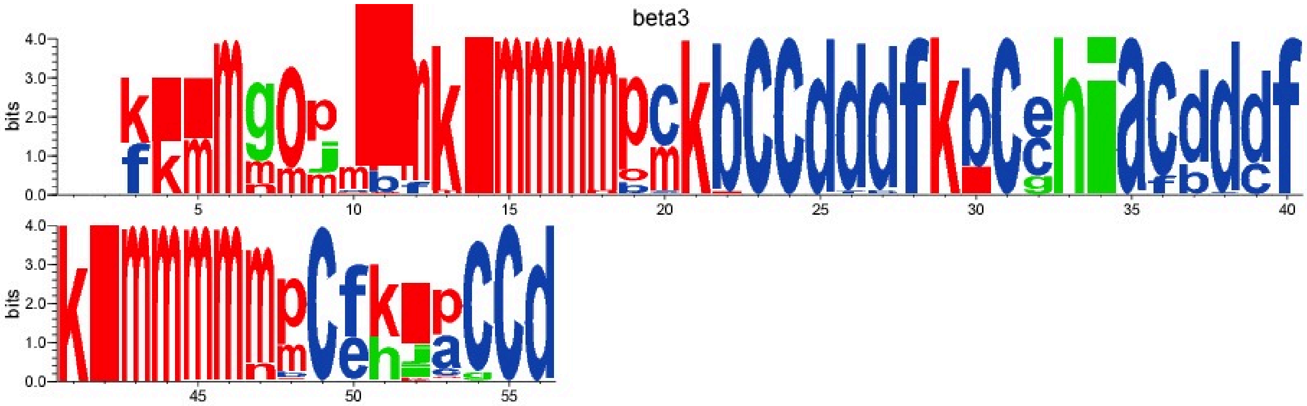

Figure 5. WebLogo-like representation of PBs for the PSI domain of the β3 integrin. PBs in red roughly correspond to α-helices, PBs in blue to β-sheets and PBs in green to coil.