SMIR: a method to predict the residues involved in the core of a protein

R. Acuña[1], Z. Lacroix[1], J. Chomilier[2,3], and N. Papandreou[4]
[1]Scientific Data Management Laboratory, Arizona State University, Tempe, AZ, USA
[2]IMPMC, Sorbonne Universités, Université Pierre et Marie Curie, CNRS UMR 7590, MNHN, IRD UMR 206, Paris, France
[3]RPBS, Université Paris Diderot, Paris, France
[4]Department of Biotechnology, Agricultural University of Athens, Athens, Greece

Abstract - *Protein folding is the critical spontaneous phase when the protein gains its structural conformation. If errors occur in the process, the protein structure may fail to fold properly. We present a new method SMIR that identifies the residues involved in the protein core. A Monte Carlo algorithm is used to simulate the early steps of folding to determine the number of non-covalently bound neighbors. Residues surrounded by many others may play a role in the compactness of the protein and thus are called Most Interacting Residues (MIR). The original MIR method was updated and extended with a new smoothing method using hydrophobic-based residue analysis. SMIR is available as a web server. SMIR is free and open to all users as functionality of the Structural Prediction for pRotein fOlding UTility System (SPROUTS) at http://sprouts.rpbs.univ-paris-diderot.fr/mir.html. The new server also offers a user-friendly interface and access to previous results.*
Contact: Zoé Lacroix

Keywords: MIR, SMIR, simulation, protein, folding, lattice.

1 Introduction

Amino acids involved in inter residue contacts may play a role in the compactness of the protein and thus are called Most Interacting Residues (MIR). The MIR method was first introduced to simulate the origin of protein folding [1]. Starting from a random conformation, the folding process can be dynamically simulated in a discrete space (a lattice). Successive residues that collapse and form a local compact structure (linked to another one by an ex-tended polypeptide chain) form a fragment. The MIR method focuses exclusively on the early steps of the folding process. In its very first implementation, it aimed to delineate the fragments formed at this stage. For this reason, the method was calibrated with time limits to maximize the number of fragments before the folding process reaches a single compact domain. It assigned a score between 2 and 8 to each residue, corresponding to the mean number of non-covalent neighbors in the lattice. A high score indicates that the residue is buried, thus belongs to a fragment. A low score indicates that it is a low interacting residue, belonging to a piece of the chain which links two consecutive fragments. A correspondence between fragments and regular secondary-structure elements (SSE) was demonstrated on a set of 42 proteins, representative of various folds [1]. However, it has been shown that a pertinent analysis of globular protein structures with respect to folding properties consists in describing them as an ensemble of contiguous closed loops [2] or Tightened End Fragments (TEFs) [3]. Such description reveals that the ends of TEFs are fold elements crucial for the formation of stable structures and for navigating the very process of protein folding. Meanwhile, the MIR algorithm evolved and newer versions (including the actual presented one) aim at locating individual residues with very high mean number of neighbors (typically ≥ 6), which are called the MIRs. In the other limit, individual residues with low mean number of neighbors (typically 2) are the Least Interacting Residues (LIRs). Therefore, the residues identified as MIRs have the tendency to be buried at the early stages of the folding process. The comparison of MIR positions with the positions of the limits of closed loops, in proteins of known 3D structures, showed a statistically significant agreement. MIRs also significantly correlate with topohydrophobic positions, i.e., positions in multiple alignments of sequences of common fold occupied only by hydrophobic amino acids, and correlated to the folding nucleus [4], thereby giving a route to simulations of the protein folding process [5]. Thus, MIR is a potential method for an ab initio estimation of the residues which are important for folding and consequently, significantly sensitive to mutations.

It is important to keep in mind the difference between protein core and nucleus. Core is a static concept, and it results from the fact that a globular protein is a micelle, with an internal phase of hydrophobic character, and an external phase of hydrophilic character, statistically. The core of a protein can be derived by a simple accessible surface area calculation or with more sophisticated methods [6]. In contrast, nucleus is a dynamic concept, it relies on a model of folding, namely the nucleation condensation model [7]. In a few words, a small set of dispersed amino acids come into

contact during the folding because of the thermal vibrations of the molecule. They are hydrophobic, and once they form such a nucleus, the rest of the structure can be formed. Among proteins sharing the same fold, part of the nucleus is conserved. Be-sides, it is now documented that nonnative contacts are necessary for the folding, and they disappear once the stability is sufficient. Figure 1 illustrates the difference between core and nucleus in the case of a fibronectin.

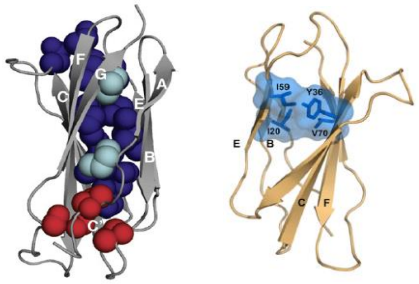

Figure 1: Difference between core (left) and nucleus (right) of the type III fibronectin [8, 9].

The knowledge of the residues constituting the folding nucleus is important for instance in the annotation of misfolding-related pathologies, but their experimental determination is not 100% secure. The role of prediction, at this moment, is a valuable complementary approach. The literature commonly admits that the number of residues involved in the folding nucleus is typically less than 10% of the sequence length, roughly one third of the hydrophobic residues. Initial MIR calculation slightly over predicts the nucleus. One guide line to improve prediction can be to produce a smoothing of the curve of NCN as a function of the sequence. This is one of the major improvements proposed with the SMIR method.

The SMIR method presented in this paper aims at improving the accuracy of MIR in the prediction of residues involved in the folding nucleus. Indeed it has been shown that MIR overestimates the folding nucleus of numerous proteins. The SMIR method is implemented and available as a server that supports the submission and the analysis of protein structures with MIR2.0 and SMIR. The server offers a dynamic interface with the display of results in a 2D graph.

2 Methods

The MIR method is an extension of previous simulations performed on cubic lattices, devoted to the complete folding of globular domains [10]. The MIR algorithm is a topological calculation, resulting from a series of energy-driven simulations of a protein backbone, where the mean number of non-covalent contacts is deduced for each residue. The analysis is performed at the early steps of folding and provides the number of Non-Covalent Neighbors (NCNs) for each residue in the sequence.

The simulation of the early steps of the folding is designed in the following manner. First, an extended initial conformation is produced for an alpha-carbon-only simplified representation of the polypeptide chain. Each alpha carbon is placed at random (while con-strained as a chain) on the nodes of a lattice. An extension of a cubic lattice, namely (2, 1, 0), originally proposed by Skolnick and Kolinski [11] is used (see Figure 2). Compared to the simple cubic lattice, it allows a wider range of backbone angles, from 64° to 143° between three contiguous alpha carbons. The number of first neighbors is also higher, 24, instead of 6. Side chains are discarded in the present simulation. Folding is produced by randomly selecting one amino acid, and submitting it to one of two available moves: end move for the N or C terminal positions, or corner move otherwise. Crankshaft move is no longer permitted with the (2, 1, 0) lattice. The new position can be occupied if it was previously empty, and the energy of the new conformation is computed by means of a statistical potential of mean force taken from the literature [12]. The Metropolis criterion is applied to accept or reject the new conformation.

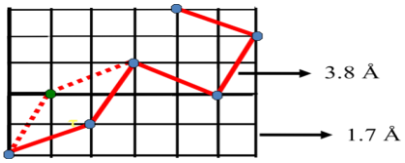

Figure 2: Details of the (2, 1, 0) lattice, with respect to the underlying cubic lattice. The dotted line indicates a possible move to a free node.

The process is stopped when roughly 10^6 to 10^7 Monte Carlo steps are reached, depending on the length of the query sequence. The full process is repeated 100 times, starting from 100 different initial conformations. The number of non-covalent neighbors (NCN) is recorded during each complete simulation. Two non-covalently bound residues are considered to interact if the distance between their respective alpha carbons does not exceed the upper limit of 5.9Å. The mean NCN is calculated at the end of the process and for all the initial conformations. The distribution of NCN along the sequence presents maxima and minima. We paid most of our attention to the maxima because we were aiming at the prediction of the core contacting residues, expected to be crucial for the formation of secondary structures [13] and whose prediction allows to determine the fold [14]. Therefore a residue i is accepted as a MIR if NCN(i) is equal or higher than 6. It results that more than 90% of the MIRs are hydrophobic (one of the six amino acids FILMVWY).

It has been demonstrated that for each protein, residues identified as MIRs constitute a non-trivial subset of the hydrophobic residues. Among families of folds (several

domains per family, similar structure, potentially different functions, and very divergent sequences) MIR occupy equivalent positions in the multiple alignments. There-fore, among families, a small number of hydrophobic positions are conserved as hydrophobic. They are compulsory for the folding to occur; they are deeply buried. For these reasons it seems reasonable to question whether they constitute the folding nucleus of the various folds. The answer is positive, as proposed by the presently available studies. They concern a very small number of families, because experimental evidence of the folding nucleus is not obvious and can show strong biases. Demonstration has been extensively proposed on two complete families, the immunoglobulins (56 structures of divergent sequences) and flavodoxins (43 structures).

One limitation of the MIR algorithm was the number of MIRs identified (by the threshold), typically around 15% of the amino acids, while the rate of amino acids expected to belong to the folding nucleus lies roughly in the range 5 to 10% This limitation also relates to the overall sharp variation in the graph. The SMIR extension addresses these issues, and it uses a Pascal triangle method to give smooth results. We also adjust the maxima that are identified in the smoothed graph to nearby (within 3 residues) hydrophobic positions, based on the accepted precision of the algorithm [15]. This is coherent with the expected accuracy for protein residue contact prediction of the contact prediction session of the Critical Assessment of protein Structure Prediction (CASP) experiments [16]. Hence, we continue to identify minima with a threshold but validate the extrema against the amino acids.

3 Results

3.1 Section and subsection headings

We model a protein as a chain of evenly spaced Cα atoms placed on a lattice [1]. We define a lattice unit (lu) to be 1.7 Å. Hence, Cα atoms are connected by vectors of the form (2,1,0), these vectors are $5^{1/2}$ lu in length which corresponds to 3.8 Å - the mean distance between adjacent Cα atoms. This results in 24 immediate neighbor positions for each point in the lattice. This represents the intersection of a 4x4x4 segmented cube with a sphere of radius 3.8 Å ($5^{1/2}$ lu) as shown in Figure 3.

The model does not take into account the presence of side chains, therefore the required separation is modeled with the 3.8 Å mini-mum distance requirement. Based on chain geometry, we limit the angle between some Cαs at position i and i+2 by requiring the distance between them to be from 4.1 to 7.2 Å (or from $6^{1/2}$ to $18^{1/2}$ lu). This corresponds to angles from 66 ° to 143 °, which is closer to the real angles in alpha and beta conformations. This is illustrated in Figure 4 where a residue i is fixed at [0, 0, 0] and all 24 possible positions for residue i+1 are represented as black vectors. There is a choice of 23 possible vectors for residue i+2. For the sake of clarity, only one position [0, 1, 2] (the green and red vectors) is shown. Red vectors are those that violate the distance (angle) restriction.

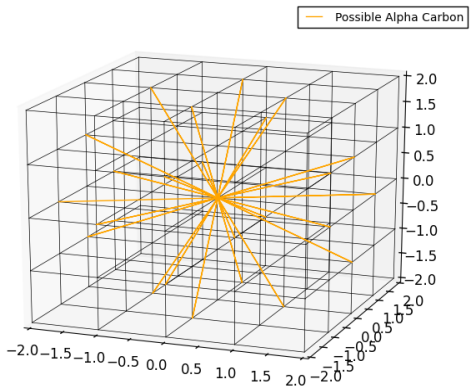

Figure 3: Vectors resulting from intersection of lattice with sphere at origin.

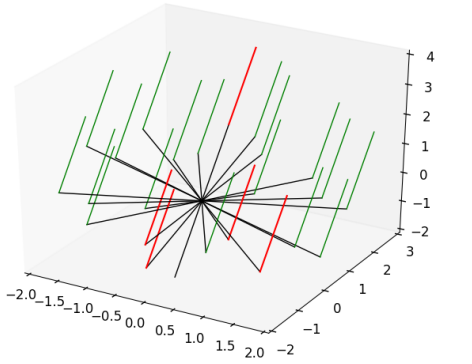

Figure 4: An example of angle restriction: vectors parallel or producing sharp corners thus violating the angle constraint are shown in red.

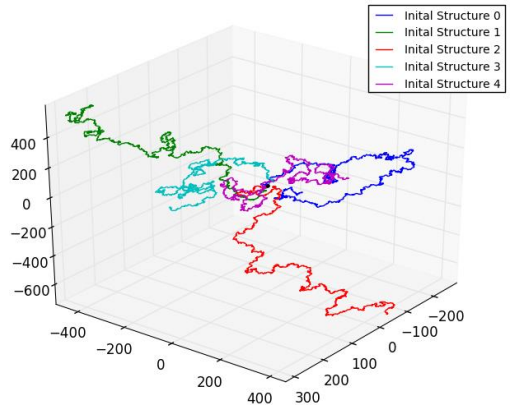

Figure 5: First five initial models.

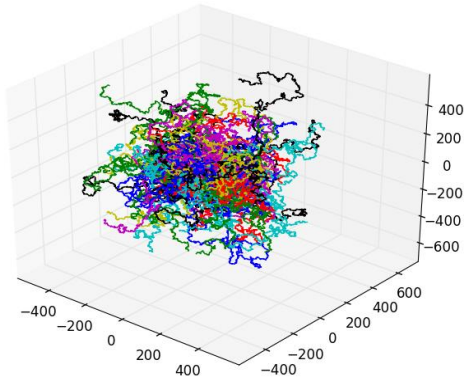

Figure 6: All initial models.

To initiate the simulations, 100 different starting models within this lattice are used. Figures 5 and 6 display respectively a sample of five and all models as a comprehensive plot. These models were computed randomly offline for chains of 1100 residues. For these models our only requirement is that they have some level of non-compactness [5]. Starting from the first residue located at position [0, 0, 0], the first n positions in the seed model will be used for an input model with n residues.

3.2 SMIR

The MIR method was first developed in 2004 [1, 5] and MIR 1.0 was first made available online as part of the RPBS server in 2005 [17]. The present SMIR server exploits MIR2.2 implemented with Fortran for server side simulation and a SMIR Javascript front end for interactive analysis. It has been found that the computation time for SMIR, once MIR results are available, is negligible on Intel Core 2 Duo based computers. The new SMIR smoothing method is implemented in Javascript with D3 [18] and has been primarily tested in Google Chrome 35. A browser based implementation allows users to retrieve this new analysis for any existing protein without the need to resubmit the entry to our submission server.

3.3 Submitting a protein

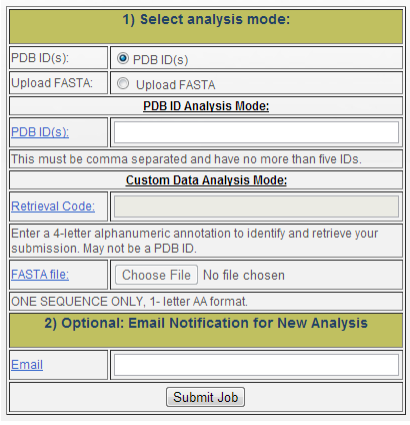

Figure 7: S/MIR Interface.

The interface in Figure 7 supports the submission of a PDB ID, a list of PDB IDs, or a FASTA file. In the latter case, the user will also enter a 4-letter alphanumeric code to identify the submission and later retrieve the results. The submission of an email address is optional. Should one be submitted, it will be only used for the purpose of informing the user of the availability of the results in the database with a reminder of the code. After submission, the server returns a SMIR status window (see Figure 8). Here the window displays the status for five proteins of PDB codes: 1AMM, 1DX5, 1I5I, 1QUC, 1ZAC. The top of the status windows lists the PDB ID(s) that have already been analyzed by MIR. Each different PDB is listed with a bullet point (e.g., 1AMM, 1DX5, 1I5I). If a protein has more than one chain, each available chain will be listed on that PDB's line and enclosed with parentheses (e.g., 1DX5(A), 1DX5(I), etc). The middle part of the window lists codes which are not valid retrieval PDB codes (e.g., 1QUC). The last part consists of the proteins that will be submitted to the server (e.g., 1ZAC). In this case, the PDB ID(s) will be added to the server queue for processing.

Figure 8: S/MIR status.

Each protein submitted to the server is displayed in the status window with the retrieval link to access the data once the execution is completed. If an email address was entered on the previous screen, a notification with a link will be sent upon completion. The proteins listed in the top of the SMIR status window are immediately viewable with a 2D graph (see Figure 9). If a PDB ID is not in the list of available proteins, it will be automatically submitted for analysis. Once the user's protein is ready to be analyzed, the server downloads the information associated with that PDB ID from the Protein Data Bank and runs MIR. After execution, the user may use the retrieval link or return to MIR query mode and enter that PDB ID to access the SMIR results. Additionally, the information that was generated for the new PDB ID is now available to other re-searchers for further use.

The graphical representation illustrated in Figure 9 is composed of three areas: legend for the MIR interface (top left), 2D display graph (top right) and data download (bottom). On recent browsers such as Chrome 24 or newer, the data can be downloaded with a CVS file. They can alternately be copied and pasted from a text box.

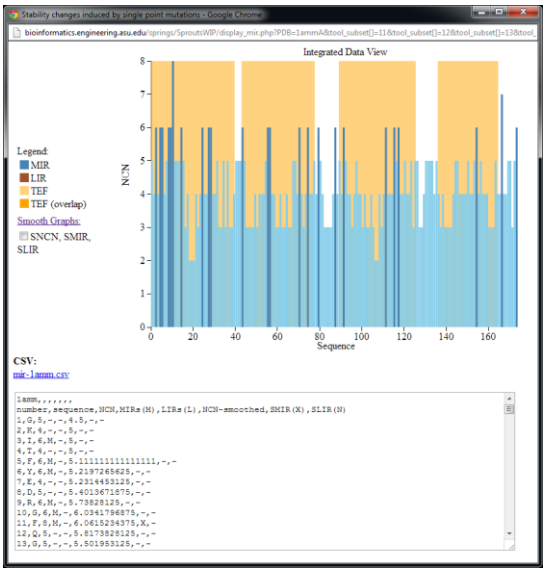

Figure 9: MIR Results.

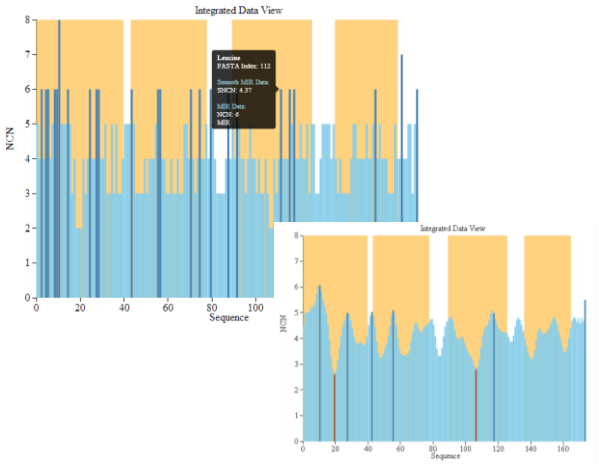

Figure 10: SMIR activated (dynamic window).

The 2D graph shown in Figure 10 displays MIR results in blue. Dark blue vertical bars indicate which residues are MIRs while dark red bars indicate LIRs (note that no LIR was shown in Figure 9). All bars plot NCN count at a position on the vertical axis. When browsing on the graph with the mouse, a black popup information box displays the amino acid name, its exact position in the protein (with respect to the FASTA file the protein is associated with), the number of NCN and the MIR status. The orange regions in the background indicate TEFs [1]. TEFs overlap on slightly darker orange areas. The SMIR method is activated with a checkbox. When SMIR is selected the 2D graph will show dynamically how MIR predictions (see top left of Figure 10) are replaced by SMIR predictions (see bottom right of Figure 10). When in smooth mode (i.e., when SMIR is selected), the dark blue and dark red bars indicate SMIRS and SLIRS (smoothed LIR, which are minima in the NCN curves) respectively.

3.4 Use Case and Discussion

We chose as an example, a case where the folding nucleus has been extensively studied. It is the TNf3 (PDB code 1ten), reported by the group of Jane Clarke [8]. This test case is interesting to tease the limits of our algorithm because a very small set of amino acids, four over roughly one hundred, is necessary to produce the so called Greek key topology of the native state. These four residues are highly conserved among the immunoglobulin superfamily, slightly less in the fibronectine type III superfamily. Φ-value analysis [19] have been experimentally determined, giving raise to the four amino acids forming the nucleus: L20, Y36, I59 and V70 [20]. MIR calculation illustrated in Figure 11 produces a high number of positions, 14 altogether.

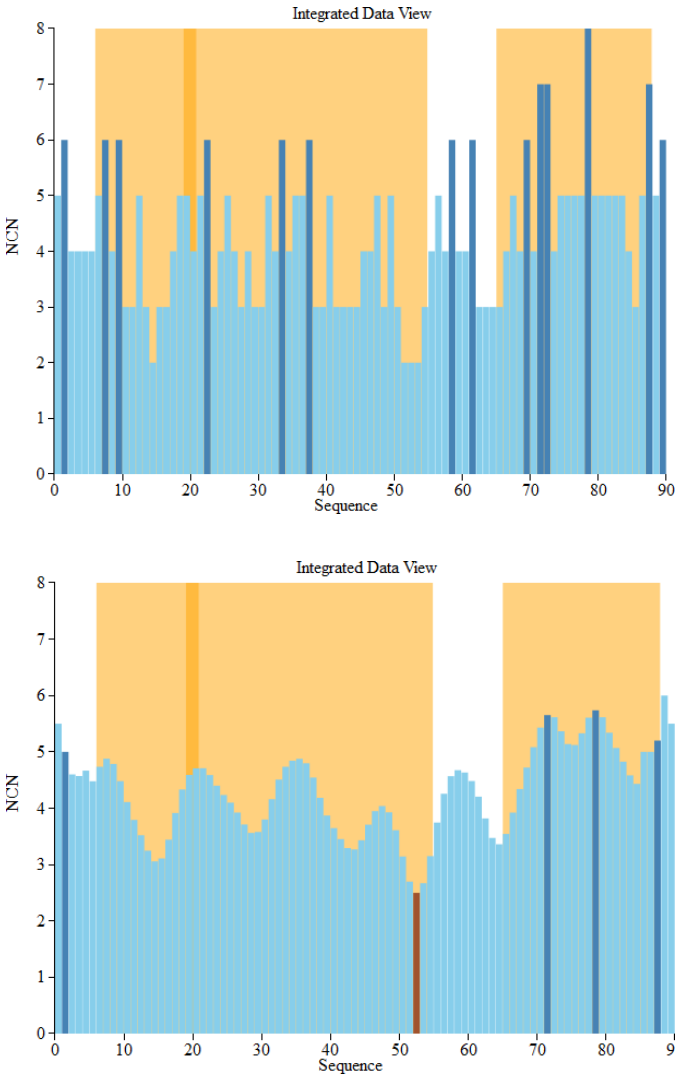

Figure 11: MIR (top) and SMIR (bottom) results for 1ten

The smoothing procedure proposed by SMIR gives a shorter list of four residues: L2, L72, M79, and F88. If one admit a window of ±2 AA, L72 is close to the list of residues involved experimentally in the folding. Although not determined to be SMIRs, L20, Y36, and I59 form maximums

in our smoothed results (refinement of SMIR thresholds may help in specific cases). This is an encouraging result for such a crude "toy model" using the only information of the sequence as an input. Actually, we do not believe the precision on positions can reasonably be more than ±2 AA.

4 Conclusions

Based on previous work [1, 5] we have presented the fundamental MIR algorithm and a method for increasing the readability and accuracy of residue interaction data. Our contribution over the previous MIR implementations is twofold: we have presented SMIR, an algorithm involving Pascal Triangle smoothing and hydrophobic residue analysis to calculate smoothed data. We have also implemented this algorithm in a new dynamic 2D graphical interface. Users may now view the smoothed MIR data for all proteins already existing in the SPROUTS database without needing to resubmit the protein for processing. These contributions refine the MIR technique so as to make MIR results more intuitive and useful to the scientific community.

One practical aspect of the prediction of MIR that can be important for wet biologists can be in the cases where they are faced with the production of inclusion bodies during the process of expression and purification. One of the ways used to circumvent this difficulty is to practice random mutations. The use of this server can be a suggestion not to mutate some positions suspected to be important for the structure, and consequently for the function, precisely the MIR. MIR and SMIR methods are also integrated in the SPROUTS workflow where they can be compared with stability analysis [21].

SMIR is hosted at the Université Paris Diderot on the server of the Ressource Parisienne en Bioinformatique Structurale (RPBS). RPBS provides scientists with a large range of resources devoted to the analysis of protein structure [17].

5 Acknowledgements

We acknowledge Pierre Tufféry for his help on using the RPBS resources, Dirk Stratmann for exciting discussions on benchmarks and method comparison and integration, and Elodie Duprat for shar-ing her results on the beta/gamma-crystallin superfamily. Mathieu Lonquety and Christophe Legendre contributed to the SPROUTS database where SMIR results are stored, and Fayez Hadji tested a preliminary version of the server. They are all thanked for their help. We also wish to acknowledge our collaborators at ASU: Rida Bazzi who is working with us on issues related to scientific workflow updates, Antonia Papandreou-Suppappola and Anna Malin who have worked on an alternative MIR method, and Banu Ozkan for evaluating SPROUTS functionalities and discussing future im-provement.

Funding: This work was partially supported by the National Science Foundation (grants IIS 0431174, IIS 0551444, IIS 0612273, IIS 0738906, IIS 0832551, IIS 0944126, and CNS 0849980) and by an invitation of the Université Pierre et Marie Curie.

Conflict of interest statement: Any opinion, finding, and conclusion or recommendation expressed in this material are those of the au-thors and do not necessarily reflect the views of the National Science Foundation.

6 References

[1] Chomilier, J., Lamarine, M., Mornon, J.P., Torres, J.H., Eliopoulos, E. and Papandre-ou, N. (2004) Analysis of fragments induced by simulated lattice protein fold-ing. Comptes Rendus Biologies, 327, 431-443.

[2] Berezovsky, I.N., Grosberg, A.Y. and Trifonov, E.N. (2000) Closed loops of nearly standard size: common basic element of protein structure. Febs Letters, 466, 283-286.

[3] Lamarine, M., Mornon, J.P., Berezovsky, I.N. and Chomilier, J. (2001) Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? Cellular and Molecular Life Sciences, 58, 492-498.

[4] Poupon, A. and Mornon, J.P. (1998) Populations of hydrophobic amino acids within protein globular domains: Identification of conserved "topohydrophobic" positions. Proteins-Structure Function and Genetics, 33, 329-342.

[5] Papandreou, N., Berezovsky, I.N., Lopes, A., Eliopoulos, E. and Chomilier, J. (2004) Universal positions in globular proteins - From observation to simulation. European Journal of Biochemistry, 271, 4762-4768.

[6] Bottini S., A Bernini, De Chiara, M, D Garlaschelli, O Spiga, M Dioguardi, E Van-nuccini, A Tramontano, N Niccolai 2013. ProCoCoA: a quantitative approach for analyzing protein core composition. Comput Biol Chem 43 :29-34.

[7] Itzhaki L.S., Otzen D.E., Fersht A.R. (1995) The structure of the transition state for folding of chmotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding, J. Mol. Biol. 25: 260-288.

[8] Lappalainen, I., Hurley, M.G. and Clarke, J. (2008) Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. Journal of Molecular Biology, 375, 547-559.

[9] Billings K., Best R., Rutherford T., Clake J. Crosstalk between the protein surface and hydrophobic core in a swapped fibronection type III domain, JMB 375 (2008) 560-571.

[10] Papandreou, N., Kanehisa, M. and Chomilier, J. (1998) Folding of the human protein FKBP. Lattice Monte-Carlo simulations. Comptes Rendus De L'Académie Des Sciences Série Iii-Sciences De La Vie-Life Sciences, 321, 835-843.

[11] Skolnick, J. and Kolinski, A. (1991) Dynamic Monte Carlo Simulations of a New Lattice Model of Globular Protein Folding, Structure and Dynamics. Journal of Molecular Biology, 221, 499-531.

[12] Miyazawa, S. and Jernigan, R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of Molecular Biology, 256, 623-644.

[13] Kister A., I. Gelfand (2009). Finding of residues crucial for supersecondary structure formation. PNAS 106: 18996-19000.

[14] Jones, D., Buchan, D., Cozzetto D., Ponti, M. (2012). PSICOV: precise structural contact prediction using spase inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184-190.

[15] Chomilier,J., Lonquety, M., Papandreou, N. and Berezovsky, I. (2006) Towards the prediction of residues involved in the folding nucleus of proteins. In Proc. DIMACS Workshop on Sequence, Structure and System Approaches to Predict Protein Function, May 3-5, 2006, Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) Center, Rutgers University. http://dimacs.rutgers.edu/Workshops/ProteinFunction/slides/chomilier.pdf

[16] Eickholt, J. and J. Cheng (2013). "A study and benchmark of DNcon: a method for protein residue contact prediction using deep networks." BMC Bioinformatics 14(Suppl): 512.

[17] Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J. et al. (2005) RPBS: a web resource for structural bioinformatics. Nucleic Acids Research, 33, W44-W49.

[18] Bostock, M., Ogievetsky, V. and Heer, J. (2011) D-3: Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics, 17, 2301-2309.

[19] Fersht A. and Sato S. (2004) Φ-value analysis and the nature of protein folding transition states, Proceedings Natl. Acad. Sci. USA, 101: 7976-7981.

[20] Hamill S., Steward A., Clarke J. (2000) The folding of an immunoglobulin like Greek key protein is defined by a common core nucleus and regions constrained by to-pology, J. Mol. Biol., 297:165-178.

[21] Acuña, R., Lacroix, Z. and Chomilier, J. (2014) SPROUTS 2.0: a workflow to predict protein stability upon point mutation, submitted to ECCB 2014.