pyDockDNA: a new web server for energy-based protein-DNA docking and scoring

Luis Ángel Rodríguez-Lumbreras[1,2], Brian Jiménez-García[1,3], Silvia Giménez-Santamarina[1,4] and Juan Fernández-Recio[1,2*]

[1]Barcelona Supercomputing Center, Barcelona, Spain

[2]Instituto de Ciencias de la Vid y del Vino (ICVV), CSIC-UR-Gobierno de La Rioja, Logroño, Spain.

[3]Zymvol Biomodeling SL, Barcelona, Spain

[4]ICMol, Universitat de València, Paterna, Spain

*** Correspondence:**
Juan Fernández-Recio
juan.fernandezrecio@icvv.es

Keywords: Structural modeling, Ab initio docking, protein-DNA interaction, Scoring function, Nucleotide parameters, Docking benchmark.

Abstract

Proteins and nucleic acids are essential biological macromolecules for cell life. Indeed, interactions between proteins and DNA regulate many biological processes such as protein synthesis, signal transduction, DNA storage, or DNA replication and repair. Despite their importance, less than 4% of total structures deposited in the Protein Data Bank (PDB) correspond to protein-DNA complexes, and very few computational methods are available to model their structure. We present here the pyDockDNA web server, which can successfully model a protein-DNA complex with a reasonable predictive success rate as benchmarked in a standard dataset of proteins in complex with DNA in B-DNA conformation. The server implements the pyDockDNA program, as a module of pyDock suite, thus including third-party programs, modules, and previously developed tools, as well as new modules and parameters to handle the DNA properly. The user is asked to enter PDB files for protein and DNA input structures (or suitable models) and select the chains to be docked. The server calculations are mainly divided into three steps: sampling by FTDOCK, scoring with new energy-based parameters and the possibility of applying external restraints. The user can select different options for these steps. The final output screen shows a 3D representation of the top 10 models and a table sorting the model according to the scoring function selected previously. All these output files can be downloaded, including the top 100 models predicted by pyDockDNA. The server can be freely accessed for academic use (https://model3dbio.csic.es/pydockdna).

1 INTRODUCTION

Proteins and nucleic acids are fundamental biological macromolecules whose functions and interactions are vital to regulating cell's life. Their interactions regulate many biological processes such

as protein synthesis, signal transduction, DNA storage, and DNA replication and repair, among others. Learning how protein and DNA interact is fundamental to fully elucidate many central biological processes and disease mechanisms, and can also support the discovery of novel therapeutic targets. Although 192,025 structures have been experimentally determined and deposited in the June 2022 release of Protein Data Bank (PDB), only 10,480 of them correspond to protein-nucleic acid complexes (this includes 6,732 protein-DNA complexes). Thus, the number of protein-DNA structures experimentally determined is clearly much smaller than the number of protein-DNA complexes that are expected to be formed in cells. This gap is partially explained by the difficulty of the experimental determination process, i.e. a very time-consuming process in the best scenarios or impossible in many cases due to limitations on the experimental techniques. For this reason, a computational approach on modelling protein-DNA interactions could be of enormous help.

Even though theoretical models of macromolecular structures are usually less accurate than direct experimental measurements, they can yield sufficient information to build a working hypothesis, complementing experimental approaches in elucidating protein-DNA interactions and guiding further experimental analyses to identify essential amino acids or nucleotide residues. From a computational point of view, there are two main approaches to model the structure of a protein-DNA complex: template-based modelling and *ab initio* docking. Template-based modelling aims to model a complex based on the structure of a homologous complex. The popularity of template-based methods has increased in the past years, especially for modelling protein-protein complexes, thanks to the development and support of many structural databases of protein interactions that can provide the required templates, such as 3D Complex (Levy et al., 2006), Dockground (Kundrotas et al., 2018), or Interactome3D (Mosca et al., 2013). However, the quality of template-based predictions clearly depends on the availability of suitable templates, not particularly high in the case of protein-DNA interactions (see for instance PDIdb (Norambuena and Melo, 2010), which makes these methods of very limited applicability. On the other hand, *ab initio* docking methods aim at predicting the three-dimensional structures of macromolecular complexes, starting from the atomic coordinates of their components. *Ab initio* docking methods do not depend on a priori in external information which makes them more useful in the actual protein-DNA context.

The methodology for prediction and modelling of protein-protein complexes is very well established despite there are still many challenges to be addressed. Numerous protein-protein docking methods have been developed and assessed as shown in the Critical Assessment of PRediction of Interactions (CAPRI) community-wide experiment. During the past editions of the CAPRI experiment (Janin et al., 2003), targets other than protein-protein complexes were proposed: protein-RNA complex (Lensink and Wodak, 2010) (T33, T34), protein-peptide (T60-64) or protein-heparin (T57) among others. However, protein-DNA docking received limited attention from the CAPRI community and developers of computational methods. Macromolecular docking protocols that accept protein and DNA coordinates as input include FTDock (Gabb et al., 1997), GRAMM-X (Tovchigrechko and Vakser, 2006), HEX (Macindoe et al., 2010), PatchDock (Schneidman-Duhovny et al., 2005; Macindoe et al., 2010) and NPDock (Tuszynska et al., 2015), HDock (Yan et al., 2017), ClusPro (Comeau et al., 2004) and HADDOCK (Van Zundert et al., 2016) servers. From this list of tools, only NPDock and HDock were originally developed for protein-nucleic acid docking; the rest were developed as protein-protein docking tools that also accept nucleic acids coordinates, but they lack an intrinsic scoring function dedicated to assessing protein-DNA interactions. These protocols usually report high predictive rates in bound conditions, i.e. when the co-crystallized partners in a known complex structure are separated and re-docked. However, despite bound docking is useful for testing and development purposes, it does not represent realistic conditions and therefore it is of limited practical value for biology. Therefore, it is important to have available datasets to test protein-DNA docking tools in unbound conditions.

Compared to protein-protein docking, where the most recent release of the Weng's group Protein-Protein Docking Benchmark 5.5 (Vreven et al., 2015) has 257 entries, and to protein-RNA docking, where there are different reported benchmarks (Barik et al., 2012; Pérez-Cano et al., 2012; Huang and Zou, 2013; Nithin et al., 2017), for protein-DNA docking there is only one available benchmark, which contains 47 complexes (van Dijk and Bonvin, 2008). Using this benchmark, protein-DNA docking protocols report moderate success rates in unbound conditions. For instance, on a subset of 23 cases from this benchmark, HDock success rate for top 10 models (i.e. at least one near-native structure within the top 10 models) is less than 10%, while success rate for top 100 is slightly over 30% (Yan et al., 2017). NPDock reports a maximum success rate (i.e. at least one near-native conformation found in the entire prediction set) of 7/47 (15%) (Tuszynska et al., 2015). Protein-DNA docking with HADDOCK reported an excellent performance (van Dijk and Bonvin, 2010) when using restraints from the real interface. This represents a very promising approach, but in a realistic scenario, lack of knowledge on the actual complex interface might limit its application. A more recent coarse-version of HADDOCK protein-DNA docking shows similar accuracy with ~6-fold speed increase over atomistic calculations (Honorato et al., 2019). The need of new computational tools to address unbound protein-DNA docking is clear. We present here a new web server that implements the pyDockDNA protein-DNA docking and scoring protocol, as a new module of pyDock version 4 (upcoming publication). The original pyDock docking and scoring approach (Cheng et al., 2007), which showed excellent performance for the prediction of protein-protein docking (Lensink et al., 2019; Rosell et al., 2020), has been rewritten in Python 3 and extended for its application to protein-DNA docking, with new functionalities to handle the nucleic acid structures and upgraded atomic solvation parameters for a more accurate scoring of protein-DNA interactions.

2 MATERIALS AND METHODS

Data Sets: protein-DNA docking benchmark and external case studies

In order to test the new pyDockDNA docking protocol, we used a previously developed protein-DNA docking benchmark (version 1.2) (van Dijk and Bonvin, 2008). The benchmark contains bound and unbound x-ray crystallography and NMR structures for 47 protein-DNA complexes, in which DNA is in B-DNA conformation. These are classified as 'easy', 'intermediate' or 'difficult' cases, based on the interface RMSD values between the bound and unbound components of the complex.

An additional set of case studies was compiled following the criteria selection of the above described protein-DNA docking benchmark. This test set is composed of ten protein-DNA complexes, where both bound and unbound structures are available for each reference complex, and the sequences are different from those in the first protein-DNA docking benchmark. Protein-DNA complex and unbound structures were compiled from the Protein-DNA Interface Database (PDIdb) (Norambuena and Melo, 2010) and the Protein Data Bank (PDB) (Berman et al., 2000). Only complexes that meet the following conditions were considered: i) DNA sequence length larger than eight base pairs, and ii) proteins without mutations in the core of the complex interface. To find the protein unbound structures of the protein-DNA complexes selected, all the PDB entries containing only protein structures were retrieved, including structures solved by NMR. Crystallographic structures with a resolution worse than 3.0 Å were not considered. To avoid redundancy, entries with sequence similarity larger than or equal to 90% were discarded. PDBeFOLD (Krissinel and Henrick, 2004) was used to find correspondences between bound and unbound protein structures. This tool performs structural alignments between two (pairwise alignment) or more (multi-alignment) molecules using their 3-dimensional structures. The alignment

is based on the Secondary Structure Matching algorithm (Krissinel and Henrick, 2004). Alignments with a Q-score higher than 8.0, high P-score and sequence similarity around 90-100% were accepted as the corresponding unbound. Then, both bound and unbound structures for each case, were post-processed according to the protocol followed in a previously developed protein-DNA docking benchmark, for instance by checking consistency between unbound and bound coordinates in chain IDs, residue numbers and atom names (van Dijk and Bonvin, 2008). The unbound DNA models were generated by using the software 3DNA (Lu and Olson, 2003; Lu and Olson, 2008), in canonical B-DNA conformation (fiber model 4).

This additional test set (Table 1) is freely available at the "Help" section of the server (https://model3dbio.csic.es/pydockdna/info/faq_and_help#extended_bechmark).

[INSERT HERE TABLE 1]

Sampling

In this first step, the input files with the coordinates in PDB format for the structures (or models) of a protein and a DNA molecule (which can be B-DNA or any other conformation) are checked for potential format errors, missing side-chains in the protein are rebuilt with SCWRL 3.0 (Bower et al., 1997), and the electrostatics Amber94 force field (Cornell et al., 1995) is loaded, assigning the charges to the atoms. Then, rigid-body docking poses between the protein and the DNA, represented as 3D grids, are generated with a faster and parallelized version of the original FTDock (v2.0) software (Gabb et al., 1997) in which the number of cells in the grid is optimized for maximum computing efficiency (Jiménez-García et al., 2013). The molecule with the longest maximal distance between any pair of atoms is considered the receptor, that is, the fixed molecule, and the other one is the ligand or mobile molecule. By default, the program uses 0.7 Å grid cell size, 1.3 Å surface thickness, 12º rotation sampling, and keeps the best 3 poses for each rotation. For each target, a total of 10,000 docking poses were generated.

Scoring

Finally, the protein-DNA docking poses are ranked using a scoring function composed of electrostatics, desolvation and van der Waals energy. This new pyDockDNA scoring function is adapted from the previously pyDock scoring function for protein-protein docking (Grosdidier et al., 2007; Jiménez-García et al., 2013), which now includes atom types for nucleotides from Amber94 force field (Cornell et al., 1995) in order to calculate for the modelled protein-DNA complexes. The nucleotide AMBER atom types have been mapped to the previously defined atom types in pyDock within a new parameter set (*nuc.dat*).

Implementation of pyDockDNA web server

The program pyDockDNA is built as a module of the new pyDock 4.0 version (upcoming publication), thus include the same third-party programs, modules and tools from previous versions of pyDock as well as new functionalities to handle the nucleic acid structures properly. The user can select the chains to be docked, the energetic scoring function, and even include external information (from available experimental data or using predictive methods such as the DBSI server **[REF: https://doi.org/10.1093/bioinformatics/btw315]**) as residue-nucleotide distance restraints to rescore docking models as previously described for pyDockRST (Chelliah et al., 2006). The output will be a set of docking models represented in different formats: i) the 3D structure of the best-scoring 10 docking models in terms of scoring can be visualized in the output screen, ii) the PDB files for the best-

scoring 100 models can be directly downloaded, and iii) the rotation/translation vectors are provided to generate up to a total of 10,000 docking poses. A summary of the docking results can be visualized as a plot with the distribution of the different energy values obtained for all docking poses (Figure 1).

[INSERT HERE FIGURE 1]

Clustering of protein-DNA docking models in benchmarking

When testing this software (see Results) we have run several docking executions in parallel, using different initial random rotations for the input structures, and the best-scoring 100 resulting models for each individual run were merged into a single pool. To avoid redundancy in the final set, all docking orientations were clustered by pyProCT analysis software (Gil and Guallar, 2014), which implements the GROMOS clustering algorithm (Daura et al., 1999). Distance matrix is built with pyRMSD with the option "QCP OMP CALCULATOR" to compute the ligand root-mean-square deviation (L-RMSD) values for all pairs of docking orientations after their receptors were superimposed (https://github.com/victor-gil-sepulveda/pyRMSD/). The L-RMSD cut-off value 4.0 Å was used to define the clusters. For each defined cluster of models, the orientation with the lowest docking score is defined as the cluster representative.

Docking performance

We have evaluate the predicted performance of pyDockDNA in different conditions as the success rates for the obtained top *N* docking models, which is the % of benchmark cases in which a near-native (acceptable) solution is found within the top *N* docking models. A near-native solution is defined as a docking orientation model with L-RMSD ≤ 10 Å with respect to the reference structure.

3 RESULTS AND DISCUSSION

Performance of pyDockDNA evaluated on the protein-DNA docking benchmark

The pyDockDNA web server has been tested on the 47 cases of a previously reported protein-DNA docking benchmark (see Methods). It is known that using different randomly rotated input structures can slightly affect docking predictions in FFT-based docking protocols as in FTDOCK, because this can modify the mapping of the atom positions on the 3D grids (Garzon et al., 2009; Pallara et al., 2016). To check for convergence, we applied pyDockDNA to 10 different random rotations of the initial input structures for each benchmark case and computed the predictive success rates for the results obtained from each randomly rotated input structures. The results indicate even more differences in the predictive values than previously reported for protein-protein docking (Table S1). For instance, the success rates for the top 10 models ranged from 12.8% to 21.3%. Therefore, for a more robust evaluation, we merged the results of all 10 docking executions and clustered the obtained docking models to remove similar orientations (see Methods). Figure 2 shows the predictive success rates of the cluster representatives resulting from merging these 10 docking runs. The predictive success for the default pyDock scoring function (including parameters for nucleotide atoms, see Methods) are better than those obtained for the individual docking runs, which means that increasing sampling variability when using different random initial rotations, followed by redundancy removal with clustering, have improved the docking results.

[INSERT HERE FIGURE 2]

We further analyzed whether a scoring function previously developed for protein-protein docking was really optimal for protein-DNA docking, since for the latter, electrostatics energy term is expected to have a larger contribution to binding energy due to the higher overall charge of DNA molecules. Moreover, desolvation atomic parameters were previously derived for protein-protein docking in pyDock, but they were not specifically optimized here for nucleotide atoms. To analyze the role of desolvation in protein-DNA scoring, we rescored the generated docking models with the pyDockDNA scoring function but excluding desolvation energy. This greatly improved the success rates, as the curve *pyDockDNA (no desolv)* shows in Figure 2. This indeed indicates that desolvation is not really needed for the scoring of the protein-DNA docking models generated by FFT-based sampling, perhaps because the parameters have not been yet optimized for nucleotide atoms, or because electrostatics is more relevant in protein-DNA interactions than in protein-protein complexes, as above discussed. We tested other solvation parameters for protein-DNA reported in the literature (Kagawa et al., 1989), but the docking results did not improve (we are currently working on the optimization of these parameters in search of a better desolvation for protein-DNA).

In addition, we have also tried other combinations of energy terms, for instance, increasing the factor for van der Waals to 1.0 (we previously found that geometrical complementarity was very important in protein-RNA; (Pérez-Cano et al., 2016), or removing desolvation and van der Waals terms from the scoring function to test the relevance of elecrostatics scoring alone, but none of these new combined scoring functions improved the prediction rates (Figure S1).

In a rigid-body docking approach as pyDock, it is known that protein flexibility upon binding is perhaps the most determinant factor for docking success. To further analyze whether the docking performance of pyDockDNA is affected by the flexibility of the protein or the DNA input molecules during the complex formation, we have grouped the docking results on the protein-DNA docking benchmark according to the flexibility of the protein or the DNA, that is, based on the RMSD between the unbound molecules and the corresponding ones in the complex. Regarding protein flexibility, in order to make groups of similar size, we defined these three categories: low (unbound-bound RMSD < 1 Å), medium (1 Å ≤ unbound-bound RMSD < 3 Å) and high (unbound-bound RMSD ≥ 3 Å) flexible cases. As for DNA flexibility, we defined these three categories: low (unbound-bound RMSD < 3 Å), medium (3 Å ≤ unbound-bound RMSD < 5 Å) and high (unbound-bound RMSD ≥ 5 Å) flexible cases. The results are shown in Figure 3. We can observe that the docking predictive performance does not get worse when protein flexibility is higher (actually, for pyDockDNA with no desolvation, success rates increase when protein flexibility is medium or high). However, we can see that the docking performance for highly flexible DNA molecules is dramatically low. We should note that in this benchmark, proteins in general show smaller unbound-bound RMSD values (average 2.6 Å) than DNA (average 4.2 Å). In addition, due to the different RMSD cut-off values used for proteins and for DNA, the unbound-bound RMSD values for high flexible proteins (average 4.8 Å) are much smaller than those for DNA (average 7.8 Å), which could explain the much worse predictive rates in the group of highly flexible DNA.

[INSERT HERE FIGURE 3]

Application to external case studies

For further testing, we have applied pyDockDNA to a set of ten additional protein-DNA cases (Table 1) where the structures for the complex and the unbound protein were available at PDB, and the unbound DNA was modelled in canonical B-DNA conformation (see Methods).

For each case study, we have performed a single pyDockDNA execution on the randomly rotated unbound protein and DNA structures, a realistic scenario, since the pyDockDNA server only provides results for a docking execution (randomly rotated input structures should be provided to the server in independent executions for a more thorough docking study similar to the benchmark performance analysis above shown). Overall, we obtained predictive success rates of 10% and 30% (for the top 10 and 100 models, respectively) when using pyDockDNA scoring function, and 10% and 60% (for the top 10 and 100 models, respectively), when using pyDockDNA without desolvation. Given the small number of cases of these additional set, these values are within the expected range according to the larger docking benchmark set.

The most successful case is the complex between the DNA binding domain of Early B-cell Factor 1 (Ebf1) bound to a 22bp DNA (PDB 3MLO), where a near-native docking model (L-RMSD 3.33 Å with respect to the reference) is found with rank 5 when using pyDockDNA (no desolvation) scoring function (Figure 4A). When using pyDockDNA (including desolvation) scoring function, this docking model is ranked 6, so it is still within top 10 models. This case has low-flexible protein but high-flexible DNA.

[INSERT HERE FIGURE 4]

Another case is the complex between the catabolite gene activator protein and a 11bp DNA (PDB 1O3R), where we found an almost acceptable docking model (L-RMSD 10.76 Å with respect to the reference) with rank 5, when using pyDockDNA either including solvation or not (Figure 4B). This case has also low-flexible protein but medium-flexible DNA. If this case had been considered acceptable, the success rates for the top 10 would have been 20%. However, these percentage values are perhaps not very meaningful considering the low number of cases in this external test set. Interestingly, when using van der Waals term with weighing factor 1.0 (instead of the default factor in pyDock and pyDockDNA, that is 0.1), we find near-native solutions in 3 more cases, in addition to 3MLO: i) 5JLT (L-RMSD 7.08 Å) with rank 1 when using desolvation; ii) 2NTC (L-RMSD 7.25 Å) with rank 3 not using desolvation, and iii) 2PI0 (L-RMSD 6.63 Å) with rank 3 and 2, when using desolvation or not using it, respectively. Therefore, for half of these external case studies, we found near-native docking models within the top 10 models with pyDockDNA, using different variants of the scoring function.

In summary, we present here the pyDockDNA web server to model protein-DNA complexes, which implements a docking method based on pyDock, with new parameters for DNA. We have evaluated the performance on unbound proteins and modelled DNA molecules in canonical B-DNA conformation, using a known protein-DNA docking benchmark. The results show near 40% success rate for the top 10 models when using the pyDockDNA (no desolvation) scoring function, after merging the results from 10 docking executions using different randomly rotated initial structures, and clustering the models to remove redundant ones. The method has been applied to external case studies, with similar predictive performance.

4 AUTHOR CONTRIBUTIONS

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

L.A.R.-L. wrote the first draft, performed the analysis, optimized the energy-based parameters, implemented the final version of the server and updated the module on the standalone version of pyDock 4.0. B.J.-G. implemented the first version of the server and that of the new module on the standalone version of the pyDock software, and reviewed the draft. S.-G.-S. compiled the external case studies and validated the software. J.F.-R. devised the idea, optimized the energy-based parameters, analyzed the results, and wrote the final manuscript.

5 FUNDING

This work was supported by grant number PID2019-110167RB-I00 / AEI / 10.13039/501100011033. B.J.-G. is employed by Zymvol Biomodeling on a project which received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 801342 (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ).

6 ACKNOWLEDGMENTS

7 REFERENCES

Barik, A., Nithin, C., Manasa, P., and Bahadur, R.P. (2012). A protein-RNA docking benchmark (I): Nonredundant cases. *Proteins: Structure, Function and Bioinformatics* 80(7)**,** 1866-1871. doi: 10.1002/prot.24083.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Research* 28(1)**,** 235-242. doi: 10.1093/nar/28.1.235.

Bower, M.J., Cohen, F.E., and Dunbrack, R.L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *Journal of Molecular Biology* 267(5)**,** 1268-1282. doi: 10.1006/jmbi.1997.0926.

Chelliah, V., Blundell, T.L., and Fernandez-Recio, J. (2006). Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *J Mol Biol* 357(5)**,** 1669-1682. doi: 10.1016/j.jmb.2006.01.001.

Cheng, T.M.-K., Blundell, T.L., and Fernandez-Recio, J. (2007). pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins: Structure, Function, and Bioinformatics* 68(2)**,** 503-515. doi: 10.1002/prot.21419.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004). ClusPro: A fully automated algorithm for protein-protein docking. *Nucleic Acids Research* 32(WEB SERVER ISS.)**,** 96-99. doi: 10.1093/nar/gkh354.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., et al. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117(19)**,** 5179-5197. doi: 10.1021/ja00124a002.

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W.F., and Mark, A.E. (1999). Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* 38(1/2)**,** 236-240. doi: 10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.3.CO;2-D.

Gabb, H.a., Jackson, R.M., and Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of molecular biology* 272(1)**,** 106-120. doi: 10.1006/jmbi.1997.1203.

Garzon, J.I., Lopez-Blanco, J.R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., et al. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* 25(19)**,** 2544-2551. doi: 10.1093/bioinformatics/btp447.

Gil, V.A., and Guallar, V. (2014). PyProCT: Automated cluster analysis for structural bioinformatics. *Journal of Chemical Theory and Computation* 10(8)**,** 3236-3243. doi: 10.1021/ct500306s.

Grosdidier, S., Pons, C., Solernou, A., and Fernández-Recio, J. (2007). Prediction and scoring of docking poses with pyDock. *Proteins* 69(4)**,** 852-858. doi: 10.1002/prot.21796.

Honorato, R.V., Roel-Touris, J., and Bonvin, A. (2019). MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Front Mol Biosci* 6**,** 102. doi: 10.3389/fmolb.2019.00102.

Huang, S.Y., and Zou, X. (2013). A nonredundant structure dataset for benchmarking protein-RNA computational docking. *Journal of Computational Chemistry* 34(4)**,** 311-318. doi: 10.1002/jcc.23149.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J.E., Vajda, S., et al. (2003). CAPRI: A critical assessment of PRedicted interactions. *Proteins: Structure, Function and Genetics* 52(1)**,** 2-9. doi: 10.1002/prot.10381.

Jiménez-García, B., Pons, C., and Fernández-Recio, J. (2013). pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics* 29(13)**,** 1698-1699. doi: 10.1093/bioinformatics/btt262.

Kagawa, T.F., Stoddard, D., Zhou, G., and Ho, P.S. (1989). Quantitative Analysis of DNA Secondary Structure from Solvent-Accessible Surfaces: The B- to Z-DNA Transition as a Model. *Biochemistry* 28(16)**,** 6642-6651. doi: 10.1021/bi00442a017.

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography* 60(12 I)**,** 2256-2268. doi: 10.1107/S0907444904026460.

Kundrotas, P.J., Anishchenko, I., Dauzhenka, T., Kotthoff, I., Mnevets, D., Copeland, M.M., et al. (2018). Dockground: A comprehensive data resource for modeling of protein complexes. *Protein Sci* 27(1)**,** 172-181. doi: 10.1002/pro.3295.

Lensink, M.F., Brysbaert, G., Nadzirin, N., Velankar, S., Chaleil, R.A.G., Gerguri, T., et al. (2019). Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* 87(12)**,** 1200-1221. doi: 10.1002/prot.25838.

Lensink, M.F., and Wodak, S.J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins: Structure, Function and Bioinformatics* 78(15)**,** 3073-3084. doi: 10.1002/prot.22818.

Levy, E.D., Pereira-Leal, J.B., Chothia, C., and Teichmann, S.A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput Biol* 2(11)**,** e155. doi: 10.1371/journal.pcbi.0020155.

Lu, X.J., and Olson, W.K. (2003). 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research* 31(17)**,** 5108-5121. doi: 10.1093/nar/gkg680.

375 Lu, X.J., and Olson, W.K. (2008). 3DNA: A versatile, integrated software system for the analysis,
376 rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*
377 3(7)**,** 1213-1227. doi: 10.1038/nprot.2008.104.

378 Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.D., and Ritchie, D.W. (2010).
379 HexServer: An FFT-based protein docking server powered by graphics processors. *Nucleic*
380 *Acids Research* 38(SUPPL. 2)**,** 445-449. doi: 10.1093/nar/gkq311.

381 Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein
382 networks. *Nature Methods* 10(1)**,** 47-53. doi: 10.1038/nmeth.2289.

383 Nithin, C., Mukherjee, S., and Bahadur, R.P. (2017). A non-redundant protein-RNA docking
384 benchmark version 2.0. *Proteins* 85(2)**,** 256-267. doi: 10.1002/prot.25211.

385 Norambuena, T., and Melo, F. (2010). The Protein-DNA Interface database. *BMC bioinformatics* 11**,**
386 262-262. doi: 10.1186/1471-2105-11-262.

387 Pallara, C., Rueda, M., Abagyan, R., and Fernández-Recio, J. (2016). Conformational heterogeneity
388 of unbound proteins enhances recognition in protein-protein encounters. *Journal of Chemical*
389 *Theory and Computation* 12(7)**,** 3236-3249. doi: 10.1021/acs.jctc.6b00204.

390 Pérez-Cano, L., Jiménez-García, B., and Fernández-Recio, J. (2012). A protein-RNA docking
391 benchmark (II): Extended set from experimental and homology modeling data. *Proteins:*
392 *Structure, Function and Bioinformatics* 80(7)**,** 1872-1882. doi: 10.1002/prot.24075.

393 Pérez-Cano, L., Romero-Durana, M., and Fernández-Recio, J. (2016). Structural and energy
394 determinants in protein-RNA docking. *Methods***,** 1-8. doi: 10.1016/j.ymeth.2016.11.001.

395 Rosell, M., Rodriguez-Lumbreras, L.A., Romero-Durana, M., Jimenez-Garcia, B., Diaz, L., and
396 Fernandez-Recio, J. (2020). Integrative modeling of protein-protein interactions with pyDock
397 for the new docking challenges. *Proteins* 88(8)**,** 999-1008. doi: 10.1002/prot.25858.

398 Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and
399 SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research* 33(SUPPL. 2)**,**
400 363-367. doi: 10.1093/nar/gki481.

401 Tovchigrechko, A., and Vakser, I.A. (2006). GRAMM-X public web server for protein-protein
402 docking. *Nucleic Acids Research* 34(WEB. SERV. ISS.)**,** 310-314. doi: 10.1093/nar/gkl206.

403 Tuszynska, I., Magnus, M., Jonak, K., Dawson, W., and Bujnicki, J.M. (2015). NPDock: A web
404 server for protein-nucleic acid docking. *Nucleic Acids Research* 43(W1)**,** W425-W430. doi:
405 10.1093/nar/gkv493.

406 van Dijk, M., and Bonvin, A.M.J.J. (2008). A protein-DNA docking benchmark. *Nucleic Acids*
407 *Research* 36(14)**,** e88-e88. doi: 10.1093/nar/gkn386.

408 van Dijk, M., and Bonvin, A.M.J.J. (2010). Pushing the limits of what is achievable in protein-DNA
409 docking: Benchmarking HADDOCK's performance. *Nucleic Acids Research* 38(17)**,** 5634-
410 5647. doi: 10.1093/nar/gkq222.

411 Van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M., Schmitz, C., Kastritis, P.L., Karaca, E., et
412 al. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of
413 Biomolecular Complexes. *Journal of Molecular Biology* 428(4)**,** 720-725. doi:
414 10.1016/j.jmb.2015.09.014.

415 Vreven, T., Moal, I.H., Vangone, A., Pierce, B.G., Kastritis, P.L., Torchala, M., et al. (2015).
416 Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark

Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* 427(19)**,** 3031-3041. doi: 10.1016/j.jmb.2015.07.016.

Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S.Y. (2017). HDOCK: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Research* 45(W1)**,** W365-W373. doi: 10.1093/nar/gkx407.

FIGURE LEGENDS

Figure 1. Schematic representation of the pyDockDNA web server main functionalities.

Figure 2. Predictive performance for the top *N=1, 5, 10, 100* models of pyDockDNA (with and without desolvation) on the protein-DNA docking benchmark.

Figure 3. Predictive performance for the top 10 models of pyDockDNA (with and without desolvation) on the protein-DNA docking benchmark when cases are grouped according to (A) protein flexibility (low: RMSD < 1 Å; medium: 1 Å ≤ RMSD < 3 Å; high: RMSD ≥ 3 Å), and (B) DNA flexibility (low: RMSD < 3 Å; medium: 3 Å ≤ RMSD < 5 Å; high: RMSD ≥ 5 Å). See more details about flexibility definition in main text.

Figure 4. Application of pyDockDNA to case studies. (A) Near-native model (in yellow) obtained by pyDockDNA docking between a modelled 22bp DNA (receptor) and Ebf1 (ligand). This model was ranked 5 with pyDockDNA (no desolvation) scoring function and has L-RMSD 3.33 Å with respect to the reference (PDB 3MLO; in red). (B) Reasonable model (in yellow) obtained by pyDockDNA docking between the catabolite gene activator protein (receptor) and a modelled 11bp DNA (ligand). This model was ranked 5 with pyDockDNA (either with desolvation or with no desolvation) scoring function and has L-RMSD 10.76 Å with respect to the reference (PDB 1O3R; in red).

TABLES

Table 1: List of case studies

PDB complex	Protein	PDB unbound protein	RMSD unbound-bound protein	DNA	RMSD unbound-bound DNA
5JLT	phage T4 MotA DNA-binding domain	1KAF	0.83[a]	22bp dsDNA	1.89
2X6V	TBX5	2X6V	0.55	11bp DNA	2.03
3POV	SOX	3FHD	1.46	19bp DNA	2.26
4UUV	ETV4 DNA-binding ETS domain	5ILU	1.24	10bp DNA	2.81
2NTC	sv40 large T antigen	2FUF	1.13[a]	21-nt PEN element of the SV40 DNA origin	2.96
2ITL	sv40 large T antigen	4NBP	5.37[a]	24-nt PEN element of the SV40 DNA origin	3.84
3MFK	Protein C-Ets1	1GVJ	5.61[a]	stromelysin-1 promoter DNA	4.34
2PI0	IRF-3	3QU6	0.76[a]	PRDIII-I region of human interferon-B promoter strand 1	4.46
1O3R	catabolite gene activator protein	4R8H	0.65	11bp DNA	4.77
3MLO	Ebf1	3LYR	0.71[a]	22bp DNA	5.11

[a] In cases with more than one protein-DNA interface in the x-ray structure, the average value is provided.