MCSS-based Docking and Improved Scoring of Protein-Nucleotide Complexes: I. A step forward to the Fragment-Based Design of Oligonucleotides

Nicolas Chevrollier[1] **and Fabrice Leclerc**[1,✉]

[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris Sud, Gif-sur-Yvette, F-91198, France

Abstract

Computational fragment-based approaches have been widely used in drug design and drug discovery. One of the limitations for their application is the lack of performance of the scoring functions. With the emergence of new fragment-based approaches for single-stranded RNA ligands, we propose an analysis of the docking power of an MCSS-based approach evaluated on nucleotide binding sites. Combined with a clustering of MCSS-generated poses and some state-of-the-art scoring functions, the results suggest that it could be used in the design of oligonucleotides.

MCSS | ligand binding | docking | scoring | protein-nucleotide interactions | FBD

Correspondence: ***fabrice.leclerc@i2bc.paris-saclay.fr***

Introduction

Fragment-based approaches have been widely used in ligand design with several examples of "success stories" when applied to drug design and drug discovery (1–4); from the middle of the 90's (5) until now, more than 30 fragment-based drug candidates have entered the clinic (6). Despite some hindrance related to synthetic accessibility and/or ligand-design strategies, fragment-based approaches remain very attractive while dealing in a more efficient way with chemical space, molecular complexity, probability of binding and ligand efficiency (6). After high throughput screening, fragment-based approaches represent one of the three major lead generation strategies for clinical candidates (7). Both experimental and computational approaches have been developed based on the same principles that weak-binding fragments can be converted into highly efficient ligands by covalent linking. One of the contributions to the gain of binding affinity with respect to that of the individual fragments comes from the rigid body entropic barrier which is supposed to be independent of the molecular size. This gain is optimal when there is no energy penalty associated with conformation of the linker and when the binding mode of each fragment is preserved in the ligand. Weak-binding fragments should still have enough favorable contacts to counterbalance the loss of rigid body entropy on binding. In practice, the first step is to design and build a fragment library, the second is to screen the fragments and the third to assemble them into ligands as lead compounds (8, 9).

In the case of the experimental approaches, the fragments are validated by some screening methods some of which are high throughput, e.g. by surface plasmon resonance (10). This is a critical step in the process of fragment-based design (FBD). In the case of the computational approaches, the FBD is by default a structure-based approach like in the X-ray crystallography-based screening of fragments or in any other structural biology assisted FBD (2). However, the hits obtained *in silico* are not generally validated until the end of the process leading to the assembled ligands. Very few published studies actually compare *in silico* to experimental approaches to validate virtual hits like in the screening of fragment-like inhibitors against the N^5-CAIR mutase (11). A computational screen of fragment libraries is faster and more cost-effective than in experimental approaches. However, the performance of such approaches may vary although the case of the N^5-CAIR mutase shows a good overlap between the computational and experimental approaches. The lack of accuracy of the scoring functions is often invoked for the poor performance, i.e. the difficulty to discriminate native-like poses from false binding poses (12, 13). In the absence of validation after the screening step, sub-optimal fragments may be selected that are poor binders or that would not bind at all at the targeted site. Thus, there is no guarantee we can identify the optimal fragments or those with the higher binding specificity by virtual screening.

Traditionally, the FBD approaches have been applied to the design of ligands assembled using small chemical groups selected from the fragment library which is often built based on drug-like criteria. Since the fragments library should also cover some chemical space with the diversity of chemical groups and molecular properties, a good strategy is needed to assemble the fragments. The fragment merging or linking strategies consist in connecting covalently two non-competitive fragments either by fusing some chemical bonds or by creating some additional chemical bond(s) as a spacer to link both fragments. The alternative strategy, fragment growing (or fragment evolution), is less challenging; it can be viewed as an optimization process where one fragment is modified by adding some functional group that can make favorable contacts around the primary binding site of the frag-

DRAFT

ment.

In the case of biopolymers, the chemical connectivity is well-defined and thus the assembling strategy involves solving a distance-constraint problem to join the connecting atoms of successive residues. In the early days of computational FBD approaches, different flavors were implemented to design peptide ligands where the fragments are amino acid residues or moieties (14–18). Designing biopolymers is an easier task since the synthetic accessibility of designed ligands by *in silico* FBD is still very challenging. More recently, an FBD approach was applied using multi-residues fragments corresponding to 8-mers to avoid the high degrees of freedom to manage for the conformational sampling of long peptides (19). A similar approach was applied to RNA ligands using trinucleotides to predict the binding mode of single-stranded RNAs to proteins (20, 21). In both cases, peptide or RNA ligands, the sequence of the oligomer is used as input for modeling the ligand-protein complex in order to reproduce known protein-ligand interactions. Thus, some improvements are still required to make progress towards the *de novo* fragment-based design of bound oligomers to protein targets.

The primary binding site of the RNA binding domains from RBPs generally corresponds to an interface that can accommodate k-mers with k between 4 and 10 (22–24). However, the RNA motifs that actually make contacts with the protein span a much shorter stretch of contiguous residues corresponding to 3-mers up to 5-mers in many structural families of RBPs (25). Such short RNA motifs, revealed by structural biology approaches, can be successive separated by short spacers and form longer bi-partite or tri-partite motifs that can easily extend to 10-mers (26). Very recent data obtained by high-throughput binding assays and sequencing on 78 human RBPs confirm this observation where the RNA motifs are composed of conserved 3-mers separated by spacers of 0 to 10 residues (27).

Two previous studies have been carried out using FBD approaches to model ssRNA-protein interactions. A method based on a coarse-grained model (RNA-LIM) was developed to model the structure of an ssRNA at the protein surface (28). However, its application is restricted to the RNA binding region surface and the simplified representation of the nucleotides makes impossible to distinguish between different nucleotide orientations. The more recent and advanced method was tested on a set of RBPs with RRM or Pumilio domains and could generate near-native models of RNA-protein complexes with good precision (RMSD $\leq$ 2Å) in most cases for chains up to 12-mers (20, 21). However, the scoring function still lacks the accuracy to be able to discriminate near-native poses in a robust way.

In this study, we use an updated version of MCSS for the screening of nucleotides. We examine the ability to identify and score native poses using the default MCSS scoring function on an extended benchmark of protein-nucleotide complexes. Four alternative scoring functions are evaluated; the performances of the five functions are compared. A clustering of the MCSS-generated poses is also proposed to select a fewer number of poses from the perspective of its application to the design of oligonucleotides.

Results & Discussion

Most of the docking methods and their scoring functions have been tested on different benchmarks. These benchmarks have been designed for some specific families of ligands including RNA ligands (29–31). However, the RNA-protein benchmarks include large RNAs (tRNA, rRNA, ribozyme, etc) where single-stranded RNAs are poorly represented and mostly present in the context of single-stranded regions connected to double-stranded regions. Building the benchmark from a subset of RBPs binding ssRNAs would select optimal and sub-optimal binding sites corresponding to spacer regions. In order to avoid such bias, we built a benchmark based on the protein-nucleotide complexes currently available in the Protein Data Bank (RCSB PDB (32)). A previous protein-nucleotide benchmark with 62 complexes was used to evaluate the docking power of three methods: AutoDock (4.2.3), GOLD (5.1), and MOLSDOCK (33). However, the benchmark is largely outdated with only 40% of complexes with an atomic resolution less than 2.0Å and thus not representative anymore of the structural data currently available. On the other hand, it was tested under biased conditions: the docked region was restricted to the native ligand pose (5Å^3) and the high-occupancy water molecules of the binding site were preserved within a rigid receptor. In this study, we use an updated and representative dataset of high-resolution protein-nucleotide complexes in which only nucleotide monophosphate are included (see "Protein-nucleotide Benchmark" and Methods). The nucleotides are docked in an extended region (17Å^3) around the binding site where the water molecules were removed and the residues in contact with the ligand optimized (see "MCSS Calculations" and Methods).

Protein-nucleotide Benchmark. The protein-nucleotide benchmark includes a non-redundant set of 120 complexes which are associated with 14 different molecular functions. Despite the over-representation of proteins binding AMP in the 3D structures available in PDB, all the 4 nucleotides are represented; the three other nucleotides are distributed almost equally (Sup. Note 1). The selection criteria retained to build the benchmark are detailed in Methods. The analysis of the 120 nucleotide binding sites based on different molecular and energy descriptors shows that the benchmark covers a large diversity of features which reflect that of the binding modes (Fig. 1).

MCSS Calculations. Several phosphate group models were used in the MCSS calculations to determine the optimal parameters for mapping nucleotides at the protein surface. We used five different phosphate models that differ by the valence and charge of the phosphate group (Fig. 2). All the partial charges on the phosphate groups are derived from a CHARMM parameter set which was derived based on the Manning's theory of counterion condensation to account for

DRAFT

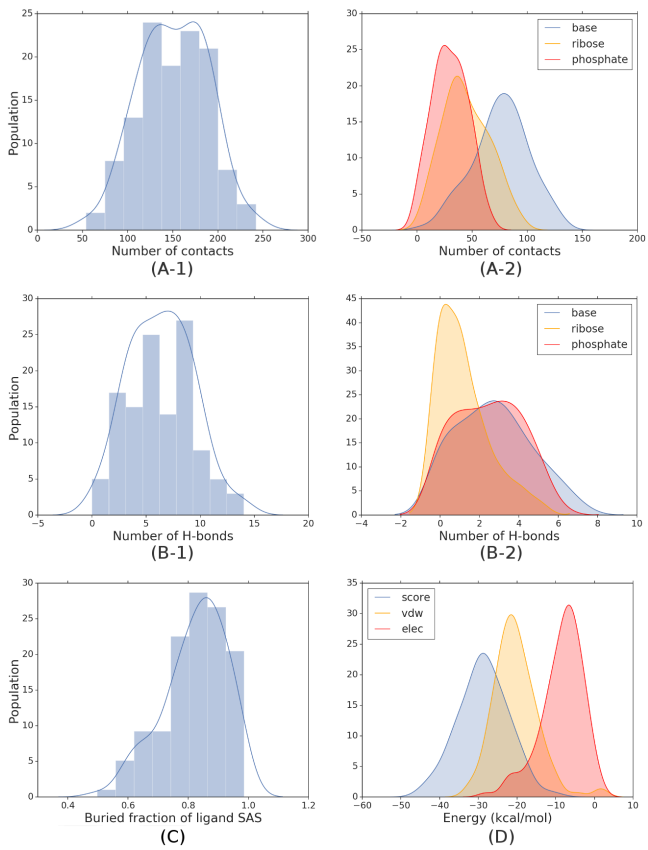

Fig. 1. Molecular and energy descriptors of the nucleotide binding sites from the benchmark of 120 complexes. A-1.: histogram of the number of contacts; B-1.: histogram of the number of H-bonds; C. histogram of the buried fraction of ligand (solvent-accessible surface); A-2.: smoothed histogram of the number of contacts with a nucleotide moiety decomposition: base, ribose, phosphate; B-2.: smoothed histogram of the number of H-bonds with a nucleotide moiety decomposition; D.: smoothed histogram of the energy score. The molecular descriptors associated with the atomic contacts and H-bonds are calculated by BINANA (34); the energy terms are calculated by CHARMM according to the MCSS scoring function.

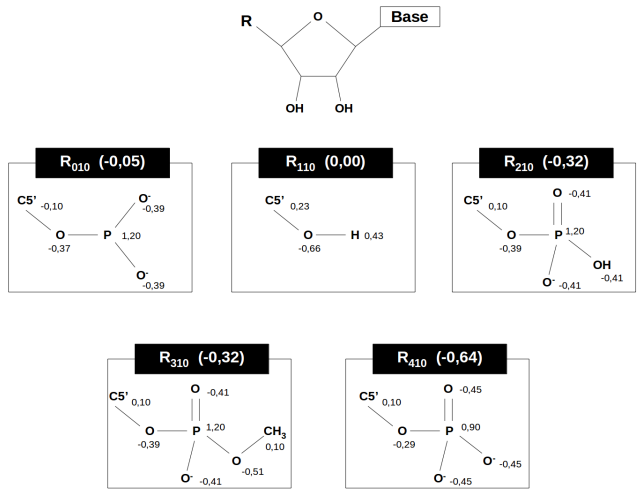

Fig. 2. Phosphate group models corresponding to different patched nucleotides. The partial charges on the phosphate groups are derived from a modified set of CHARMM22 and updated CHARMM27 parameters (35). The default charge for a full valence state of the phosphate group is -0.32 (R210, R310). A doubled net charge is assigned for the ionization state of a R-PO4^{2-} group.

the partial neutralization of the negative charges of polyelectrolytes in solution (36). The net charge on the phosphate group is scaled down according to the implicit solvent model previously used in MCSS calculations performed on nucleic acids (35).

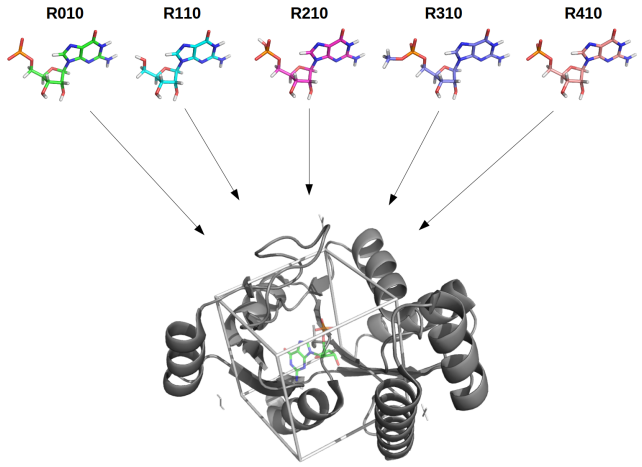

Fig. 3. Schematic description of the series of MCSS calculations performed on each protein target. The chemical structure of each 5' patched nucleotide is indicated: R010, R110, R210, R310, R410. The protein target is represented in cartoon mode with the indication of the cubic box corresponding to the explored region.

The five types of nucleotides were mapped at the protein surface (Fig. 3) and several thousand poses have been generated for each of them (see Methods). All the patched nucleotides gave an equivalent number of poses around 3000-3500 except for R110 corresponding to a nucleoside with only a bit more than 2000 poses generated (Fig. 4). As a nucleoside, R110 has a smaller size and tends to have less contacts with the protein targets. The fraction of native poses between the patched nucleotides is equivalent except for R310 which exhibits more native poses (Fig. 5). There is no significant difference depending on the charge of the phosphate group: the more charged nucleotide, R410 (-0.64), contains about the same ratio of native poses with respect to R010 or R210 (-0.32). R310 generated a higher fraction of native poses.

Scoring Nucleotide Binding. The success rate for the identification of a native pose is given for a range from top1 to top100 ranks (Fig. 6). The differences between R110 and the other patched nucleotides are more significant in particular for the success rate for top1 which is less than 20% but higher than 20% and close to 30% for some of the other patched nucleotides (R310, R410).

A clustering was applied to select the more representative poses (see Methods). The clustering leads to a decrease of the total number of poses (from more than 3000 in average to around 500) and of that of native poses as well (data not shown). However, there is an increase of the success rate for all the patched nucleotides in the top1 to top100, especially from top5 (Fig. 7). The higher success rate for the top1 is obtained with R310. In the following analyses, the results obtained with R310 are used for further comparisons.

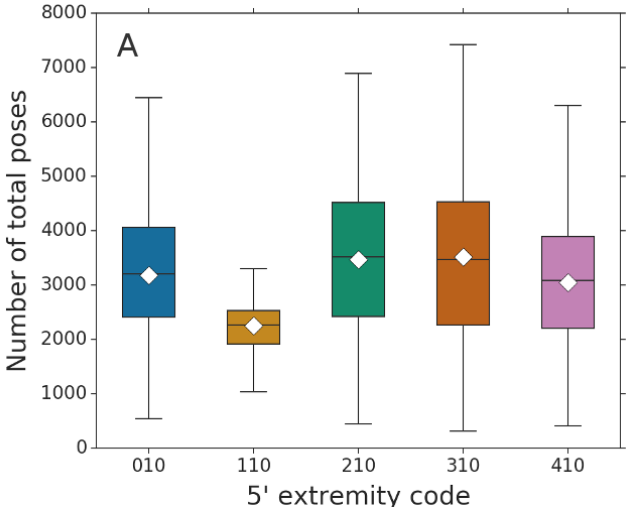

Fig. 4. Boxplot representation of the number of poses generated for the 120 protein-nucleotide complexes for each 5' patched nucleotide. A black line represents the median value, the box limits represent the first and third quartile. A diamond symbol indicates the average value.

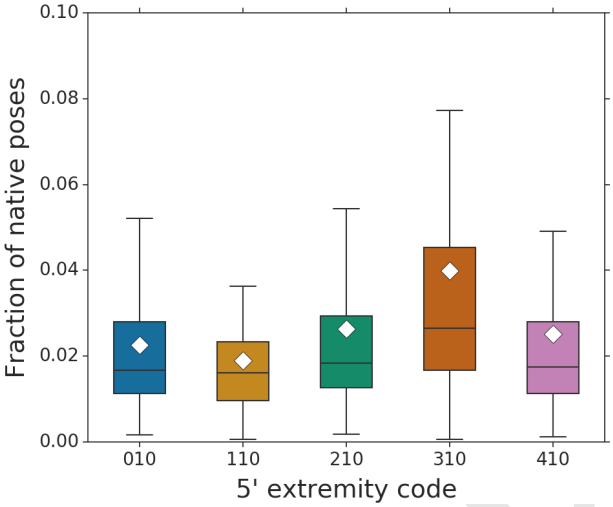

Fig. 5. Boxplot representation of the fraction of native poses generated for the 120 protein-nucleotide complexes for each 5' patched nucleotide (see legend from Fig. 4.

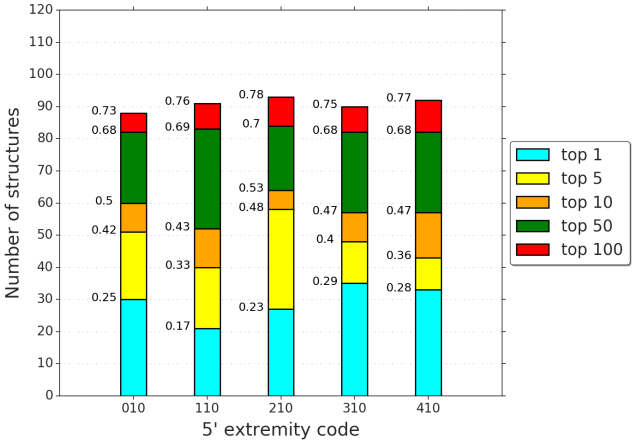

Fig. 6. Stacked histogram representation of the top n ranked native poses generated for the 120 protein-nucleotide complexes for each 5' patched nucleotide.

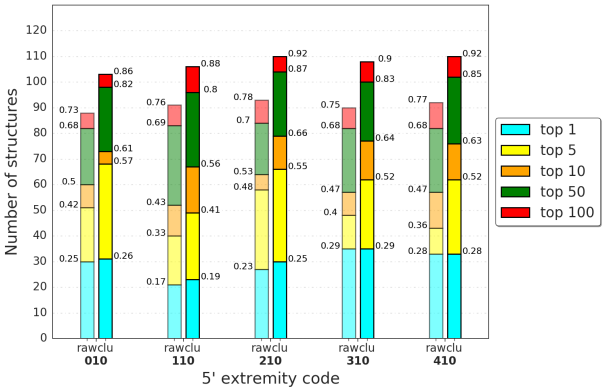

Fig. 7. Stacked histogram representation of the top n ranked native poses generated after clustering. raw: no clustering; clu: clustering based on a RMSD criterion of 2.0Å.

Alternative scoring functions. To further improve the scoring of protein-nucleotide interactions, additional scoring functions that were developed more recently were also tested: Autodock Vina score (37), Vinardo (38), $\Delta_{vina}RF_{20}$ (39), and ITscorePR (40). All the listed scoring functions are trained on protein-ligand complexes except ITscorePR which was specifically developed for protein-RNA interactions.

Autodock Vina is a well-known docking method used for virtual screening; the associated scoring function is pretty robust having regularly been used in the comparative assessment of scoring functions (CASF) challenges (41). Vinardo and $\Delta_{vina}RF_{20}$ were both derived from Vina and also tested in the CASF-2013 challenge. Vinardo was optimized and validated on large datasets (38). It was tested in particular on the DUD library that contains, among other proteins, kinases with nucleotide ligands or nucleotide analogs (42). $\Delta_{vina}RF_{20}$ was derived more recently from Vina with a new parametrization based on random forest. The performance of $\Delta_{vina}RF_{20}$ was superior to that of Vina when tested on the CASF-2007 and CASF-2013 challenges benchmarks. Finally, ITscorePR was included since it has been specifically developed for protein-RNA interactions.

The scores calculated with all the scoring functions: Autodock Vina score (37), Vinardo (38), $\Delta_{vina}RF_{20}$ (39), and ITscorePR (40), except MCSS (35) correspond to single-point calculations on the MCSS-generated poses. The performances of the five scoring functions were compared with and without clustering. All the scoring functions show pretty similar performances except ITscorePR that clearly underperforms (Fig. 8).

When no clustering is applied, Vinardo and $\Delta_{vina}RF_{20}$ generate pretty similar scores, Vinardo performing a little better especially in the top1 and top5. On the other hand, Vina and MCSS perform in a similar way from the top1 to top100 with lower scores compared to Vinardo and $\Delta_{vina}RF_{20}$. When the clustering is included: Vinardo, MCSS and $\Delta_{vina}RF_{20}$ perform with similar scores, Vina performing a bit lower especially from the top10 to top100. As observed previously for MCSS, the clustering procedure improves significantly the success rate from top5 to top100.

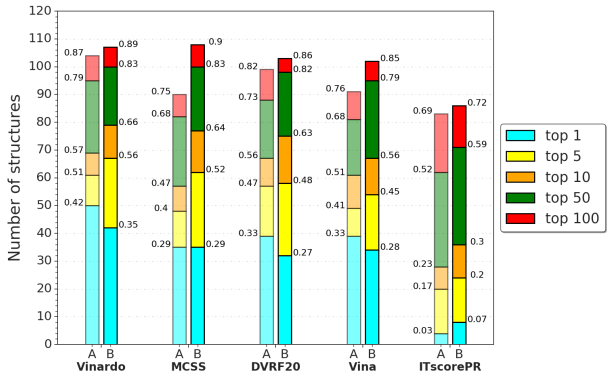

Fig. 8. Stacked histogram representation of the native poses in the top1 to top100 as scored by Vinardo, MCSS, $\Delta_{vina}RF_{20}$, Vina, and ITscorePR. A. no clustering; B. clustering.

Conclusions

MCSS was evaluated for the docking of nucleotides from analyses on a benchmark of 120 protein complexes. Different phosphate models were tested to optimize the success rate for the identification of native poses. A clustering procedure was set up that allows an increase of the success rates. Alternative scoring functions tested in the CASF challenges or developed to score protein-RNA interactions were evaluated to identify the more high-performance scoring functions. When combined with the clustering protocol, Vinardo, MCSS, and $\Delta_{vina}RF_{20}$ were found, in that order, as the best scoring functions. Assuming that the binding region of the protein can be defined within a 17Å^3 cubic box, one may expect some success rates of more than 60% for the identification of native poses in the top10. These results are encouraging from the perspective of application to a fragment-based design strategy for oligonucleotides to be validated on protein-RNA complexes.

ACKNOWLEDGEMENTS
This research was supported by the French Ministry of Higher Education, Research and Innovation.

Bibliography

1. Philip J Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery*, 6(3):211–219, March 2007. doi: 10.1038/nrd2220.
2. Christopher W Murray. The rise of fragment-based drug discovery. *Nature Chemistry*, 1(3): 187–192, June 2009. doi: 10.1038/nchem.217.
3. Monya Baker. Fragment-based lead discovery grows up. *Nature Reviews Drug Discovery*, 12(1):5–7, January 2013. doi: 10.1038/nrd3926.
4. Amanda J Price, Steven Howard, and Benjamin D Cons. Fragment-based drug discovery and its application to challenging drug targets. *Essays in biochemistry*, 61(5):475–484, November 2017. doi: 10.1042/EBC20170029.
5. S B Shuker, P J Hajduk, R P Meadows, and S W Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science (New York, NY)*, 274(5292):1531–1534, 1996. doi: 10.1126/science.274.5292.1531.
6. Daniel A Erlanson, Stephen W Fesik, Roderick E Hubbard, Wolfgang Jahnke, and Harren Jhoti. Twenty years on: the impact of fragments on drug discovery. *Nature Publishing Group*, 15(9):605–619, September 2016. doi: 10.1038/nrd.2016.109.
7. Dean G Brown and Jonas Boström. Where Do Recent Small Molecule Clinical Development Candidates Come From? *Journal Of Medicinal Chemistry*, page acs.jmedchem.8b00675, July 2018. doi: 10.1021/acs.jmedchem.8b00675.
8. Vincent Zoete, Aurélien Grosdidier, and Olivier Michielin. Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of cellular and molecular medicine*, 13(2):238–248, February 2009. doi: 10.1111/j.1582-4934.2008.00665.x.
9. Laurent Hoffer, Jean-Paul Renaud, and Dragos Horvath. Fragment-based drug design: computational & experimental state of the art. *Combinatorial Chemistry & High Throughput Screening*, 14(6):500–520, July 2011.
10. Jacob Robson-Tull. Biophysical screening in fragment-based drug design: a brief overview. *Bioscience Horizons: The International Journal of Student Research*, 11(9):1324, January 2018. doi: 10.1093/biohorizons/hzy015.
11. Tian Zhu, Hyun Lee, Hao Lei, Christopher Jones, Kavankumar Patel, Michael E Johnson, and Kirk E Hevener. Fragment-based drug discovery using a multidomain, parallel MD-MM/PBSA screening protocol. *Journal Of Chemical Information And Modeling*, 53(3):560–572, March 2013. doi: 10.1021/ci300502h.
12. A Kumar, A Voet, and K Y J Zhang. Fragment based drug design: from experimental to computational approaches. *Current medicinal chemistry*, 19(30):5128–5147, 2012.
13. Anthony E Klon. *Fragment-based methods in drug discovery*. February 2015. ISBN 9781493924868. doi: 10.1007/978-1-4939-2486-8.
14. J Singh, J Saldanha, and J M Thornton. A novel method for the modelling of peptide ligands to their receptors. *Protein engineering*, 4(3):251–261, February 1991.
15. A Caflisch, A Miranker, and M Karplus. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *Journal Of Medicinal Chemistry*, 36(15):2142–2167, July 1993.
16. C D Elkin, H J Zuccola, J M Hogle, and D Joseph-McCarthy. Computational design of D-peptide inhibitors of hepatitis delta antigen dimerization. *Journal of Computer-Aided Molecular Design*, 14(8):705–718, November 2000.
17. J Zeng, T Nheu, A Zorzet, B Catimel, E Nice, H Maruta, A W Burgess, and H R Treutlein. Design of inhibitors of Ras–Raf interaction using a computational combinatorial algorithm. *Protein engineering*, 14(1):39–45, January 2001.
18. S S So and M Karplus. Evaluation of designed ligands by a multiple screening method: application to glycogen phosphorylase inhibitors constructed with a variety of approaches. *Journal of Computer-Aided Molecular Design*, 15(7):613–647, July 2001.
19. Jun-min Liao, Yeng-Tseng Wang, and Chen-lung Steve Lin. A fragment-based docking simulation for investigating peptide-protein bindings. *Physical chemistry chemical physics : PCCP*, 19(16):10436–10442, April 2017. doi: 10.1039/c6cp07136h.
20. Isaure Chauvot de Beauchene, Sjoerd J de Vries, and Martin Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, page gkw328, April 2016. doi: 10.1093/nar/gkw328.
21. Isaure Chauvot de Beauchene, Sjoerd J de Vries, and Martin Zacharias. Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS computational biology*, 12(1):e1004697, January 2016. doi: 10.1371/journal.pcbi.1004697.
22. Zachary T Campbell, Devesh Bhimsaria, Cary T Valley, Jose A Rodriguez-Martinez, Elena Menichelli, James R Williamson, Aseem Z Ansari, and Marvin Wickens. Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Reports*, 1(5): 570–581, May 2012. doi: 10.1016/j.celrep.2012.04.003.
23. Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H Matzat, Ryan K Dale, Sarah A Smith, Christopher A Yarosh, Seth M Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S Laishram, Mei Qiao, Howard D Lipshitz, Fabio Piano, Anita H Corbett, Russ P Carstens, Brendan J Frey, Richard A Anderson, Kristen W Lynch, Luiz O F Penalva, Elissa P Lei, Andrew G Fraser, Benjamin J Blencowe, Quaid D Morris, and Timothy R Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, July 2013. doi: 10.1038/nature12311.
24. Girolamo Giudice, Fátima Sánchez-Cabo, Carlos Torroja, and Enrique Lara-Pezzi. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database : the journal of biological databases and curation*, 2016, 2016. doi: 10.1093/database/baw035.
25. Sigrid D Auweter, Florian C Oberstrass, and Frédéric H-T Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17): 4943–4959, October 2006. doi: 10.1093/nar/gkl620.
26. Tariq Afroz, Zuzana Cienikova, Antoine Cléry, and Frédéric H-T Allain. One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence. *Methods in enzymology*, 558: 235–278, 2015. doi: 10.1016/bs.mie.2015.01.015.
27. Daniel Dominguez, Peter Freese, Maria S Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J Lambert, Eric L Van Nostrand, Gabriel A Pratt, Gene W Yeo, Brenton R Graveley, and Christopher B Burge. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, 70(5):854–867.e9, June 2018. doi: 10.1016/j.molcel.2018.05.001.
28. Damien Hall, Songling Li, Kazuo Yamashita, Ryuzo Azuma, John A Carver, and Daron M Standley. RNA–LIM: A novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical Biochemistry*, 472: 52–61, March 2015. doi: 10.1016/j.ab.2014.11.004.
29. Laura Pérez-Cano, Brian Jiménez-García, and Juan Fernández-Recio. A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1872–1882, July 2012. doi: 10.1002/prot.24075.
30. Chandran Nithin, Sunandan Mukherjee, and Ranjit Prasad Bahadur. A non-redundant protein-RNA docking benchmark version 2.0. *Proteins: Structure, Function, and Bioinformatics*, 85(2):256–267, February 2017. doi: 10.1002/prot.25211.
31. Chandran Nithin, Pritha Ghosh, and Janusz M Bujnicki. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes*, 9(9):432, August 2018. doi: 10.3390/genes9090432.
32. H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.
33. Shankaran Nehru Viji, Nagarajan Balaji, and Namasivayam Gautham. Molecular docking studies of protein-nucleotide complexes using MOLSDOCK (mutually orthogonal Latin squares DOCK). *Journal of Molecular Modeling*, 18(8):3705–3722, August 2012. doi: 10.1007/s00894-012-1369-4.
34. Jacob D Durrant and J Andrew McCammon. BINANA: a novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics and Modelling*, 29(6):888–893, April 2011. doi: 10.1016/j.jmgm.2011.01.004.
35. Fabrice Leclerc and Martin Karplus. MCSS-based predictions of RNA binding sites. *Theo-*

DRAFT

retical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(1):131–137, February 1999. doi: 10.1007/s002140050419.

36. B Tidor, K K Irikura, B R Brooks, and M Karplus. Dynamics of DNA oligomers. *Journal of biomolecular structure & dynamics*, 1(1):231–252, October 1983. doi: 10.1080/07391102.1983.10507437.
37. Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, January 2010. doi: 10.1002/jcc.21334.
38. Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLOS ONE*, 11(5):e0155183–18, May 2016. doi: 10.1371/journal.pone.0155183.
39. Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3): 169–177, January 2017. doi: 10.1002/jcc.24667.
40. Zheng Huang, Min Zhang, Shawn D Burton, Levon N Katsakhyan, and Haitao Ji. Targeting the Tcf4 G13ANDE17 binding site to selectively disrupt β-catenin/T-cell factor protein-protein interactions. *ACS chemical biology*, 9(1):193–201, January 2014. doi: 10.1021/cb400795x.
41. Thomas Gaillard. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *Journal Of Chemical Information And Modeling*, 58(8):1697–1706, August 2018. doi: 10.1021/acs.jcim.8b00312.
42. Niu Huang, Brian K Shoichet, and John J Irwin. Benchmarking sets for molecular docking. *Journal Of Medicinal Chemistry*, 49(23):6789–6801, November 2006. doi: 10.1021/jm0608356.
43. Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524.
44. Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11): 1859–1865, August 2008. doi: 10.1002/jcc.20945.
45. G C P van Zundert, J P G L M Rodrigues, M Trellet, C Schmitz, P L Kastritis, E Karaca, A S J Melquiond, M van Dijk, S J de Vries, and A M J J Bonvin. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4):720–725, February 2016. doi: 10.1016/j.jmb.2015.09.014.
46. Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F Van Gunsteren, and Alan E Mark. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int. Ed*, 38(1-2):236–240, January 1999. doi: 10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M.

Methods

Protein-nucleotide Benchmark. The PDB was filtered out to select a set of protein-nucleotide complexes based on different structural criteria associated with the atomic resolution and the structural similarity. A first query was carried out to find protein complexes with each of the four nucleotides as ligands and annotated in the PDB by the following labels: AMP, C5P, 5GP, U5P. An additional criterion based on a cut-off value of 2Å resolution was also used to select only high-resolution X-ray structures. The resulting complexes were then clustered according to their sequence similarities in order to remove the redundancy. If any chain in the protein of a complex has at least 30% sequence identity with a chain in the protein from another complex, the two complexes were grouped into the same cluster. The crystal structure with the best resolution in each cluster was selected as the cluster's representative. The 188 complexes thus selected by pulling down the results from the four queries (AMP-bound: 122, C5P-bound: 18, 5GP-bound: 21, U5P-bound: 27) were then manually curated to retain those that exhibit a known binding preference for the crystallized ligand. This feature was established based on the literature and/or the annotation of the protein, e.g. a C nucleotide for CMP-kinase, etc. After curation, the dataset was reduced to 131 complexes. An additional curation was performed to eliminate some potential redundancy associated with the presence of identical binding sites for different types of nucleotides. The followed procedure consists of superimposing all the protein structures using the program TM-align (43) and review all the structures that are similar based on the TM-score (TM-score $\geq$ 0,8). Two binding sites was considered non-redundant if they differ by only one amino acid residue in direct contact with the ligand. According to this criterion, only one complex was removed from the dataset in the case of the proteins corresponding to the PDB IDs: 3DXG (U5P ligand) and 3DJX (C5P ligand); the latter complex was conserved in the dataset to compensate for the minor under-representation of C5P. The full procedure ends up with a dataset of 130 protein-nucleotide complexes.

The binding features of the 130 protein-nucleotide complexes were characterized by the number of contacts between the protein and its ligand, the fraction of buried surface area, the number of H-bonds in the binding site and the energy of interaction as calculated by the MCSS scoring function (see MCSS) Sup. Note 1. The contacts are calculated using the program BINANA (34). The full tables including the molecular features of the protein-nucleotide complexes are provided in the supplementary materials (Sup. Note 1).

MCSS. All the proteins were prepared using the CHARMM-GUI interface (44) to convert the PDB files into CRD and PSF formats. After removal of all heteroatoms, hydrogens were added to the protein using the HBUILD command from CHARMM. Histidine residues have been considered as neutral. The protein targets were then submitted to an energy minimization (tolerance gradient of 0.1 kcal/mol/Å^2). The average deviation between the experimental structure and the minimized structure is around 0.5Å.

The nucleotide library of fragments include multiple conformations, 5' and 3' patches (see MCSS documentation: https://www.mcss.cnrs.fr/MCSSDOC/Welcome.html. The initial default conformation used in the calculations is a C3'-endo/anti ribonucleotide. A set of five different patches on the 5' end was used in the current study with this nucleotide conformation: R010, R110, R210, R310, R410. Each binding region was defined by a 17Å^3 cubic box centered on the ligand centroid (Fig. 3). MCSS sample files are provided for the input and nonbonded parameters (Sup. Note 2).

The poses generated by MCSS were submitted to a clustering procedure as a postprocessing based on a hierarchical classification implemented in the HADDOCK program (45). In this implementation (46), the pose with the highest number of neighbors within a given RMSD cutoff is first identified. This pose and its neighbors constitute the first cluster. All the members of this initial cluster are then removed and the next pose with the highest number of neighbors and its neighbors are thus selected to define the second cluster. The procedure is repeated until the entire set of poses is exhausted. The RMSD cutoff used was defined at 2Å. For each resulting cluster, the pose with the best energy was choosen as a representative.

Ten protein-nucleotide complexes (PDB IDs: 1HXP, 2CFM, 2Q4H, 3L9W, 3REX 4OKE, 4XBA, 5ERS, 5M45 and 5DJH) were excluded from post-docking analyses due to 3 main reasons: (1) no native pose (RMSD $\leq$ 2.0Å) could be generated because of a nucleotide binding site too buried to be accessible (PDB ID: 5M45, 5DJH). (2) no native pose could be identified because of a huge deviation (RMSD > 2.0Å) of the

DRAFT

crystallized ligand minimized within the optimized protein binding site (PDB IDs: 1HXP, 2CFM, 4OKE, 4XBA, 5ERS). (3) the native poses identified showed highly unfavorable energies indicating the presence of steric clashes between the nucleotide and the minimized binding site (PDB IDs: 2CFM, 2Q4H, 3L9W, 3REX, 4XBA, 5ERS).

The MCSS software may be obtained after signing a license agreement upon request to Martin Karplus (marci@tammy.harvard.edu). The source code can be obtained from a Git repository on the I2BC software forge upon registration (https://forge.i2bc.paris-saclay.fr).

DRAFT

Supplementary Note 1: Benchmark of 120 protein-nucleotide complexes

Supplementary Table 1 (Table-S1.csv): list of PDB IDs including the ligand ID, the atomic resolution, functional classification, and EC number.

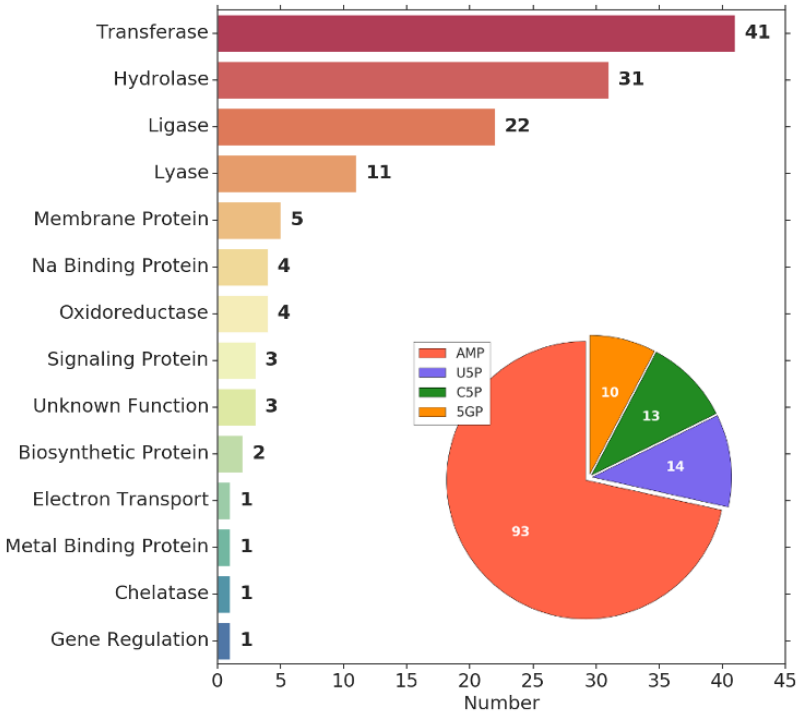

Fig. 9. Distribution of molecular functions and nucleotide types in the protein-nucleotide benchmark.

Supplementary Table 2 (Table-S2.csv): calculations of the BINANA features (number of contacts, number of H-bonds, buried fraction of ligand, etc)
Supplementary Table 3 (Table-S3.csv): calculations of the NACCESS surface terms for the fraction of buried surface of the ligand

Supplementary Note 2: MCSS

MCSS input sample (Data-S1.txt)
MCSS nonbonded parameters sample (Data-S2.txt)
Supplementary Table 4 (Table-S4.csv): MCSS score (including its VdW and elec terms) and RMSD values for each protein-nucleotide complex

DRAFT

Supplementary Note 3: Clustering of MCSS generated poses

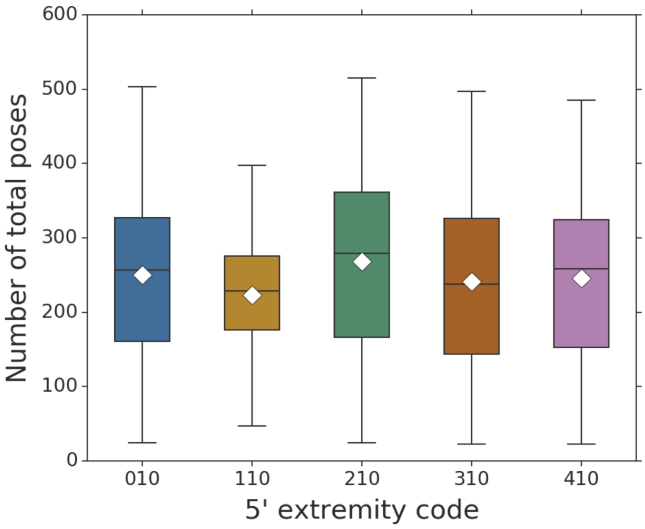

Fig. 10. Boxplot representation of the number of poses selected by the after clustering. The default boxplot representation was used with a diamond indicating the average value.

DRAFT