SCIENTIFIC REPORTS
nature research

OPEN

Characterization of partially ordered states in the intrinsically disordered N-terminal domain of p53 using millisecond molecular dynamics simulations

Pablo Herrera-Nieto[1], Adrià Pérez[1] & Gianni De Fabritiis[1,2]✉

The exploration of intrinsically disordered proteins in isolation is a crucial step to understand their complex dynamical behavior. In particular, the emergence of partially ordered states has not been explored in depth. The experimental characterization of such partially ordered states remains elusive due to their transient nature. Molecular dynamics mitigates this limitation thanks to its capability to explore biologically relevant timescales while retaining atomistic resolution. Here, millisecond unbiased molecular dynamics simulations were performed in the exemplar N-terminal region of p53. In combination with state-of-the-art Markov state models, simulations revealed the existence of several partially ordered states accounting for ~40% of the equilibrium population. Some of the most relevant states feature helical conformations similar to the bound structure of p53 to Mdm2, as well as novel β-sheet elements. This highlights the potential complexity underlying the energy surface of intrinsically disordered proteins.

Over the last decades the understanding of protein function was summarized by the *sequence-structure-function* triumvirate: protein sequences encode folds able to perform specific tasks. Intrinsically disordered proteins (IDPs) defy this principle by mediating their biological functions despite lacking a stable three-dimensional structure[1–3]. Such behavior configures a relatively flat energy surface where many isoenergetic conformations coexist[4]. This surface can be modified to a certain extent, as revealed by the shift towards certain subpopulations observed in the formation of protein-IDP[5–8] or molecule-IDP complexes[9,10]. Similarly, kinetic parameters governing the conversions amongst subpopulations can also be modified by post-translational modifications[11,12]. Thus, the energy surface of IDPs if far from being constituted exclusively by random coiled conformations, and pieces of evidence support the existence of partially ordered states[11]. The characterization of such partially ordered states is crucial to understand IDPs' function, their mechanisms of action, and their potential modulation.

The structural heterogeneity of IDPs is summarized as a collection or ensemble of conformations. They can be resolved experimentally by using nuclear magnetic resonance (NMR) or small-angle X-ray scattering data. The main limitation of IDP ensembles resolved in that way is that they focus on global averages rather than diving in particular atomic coordinates[13]. There are also many computational approximations to address this task[14]. They generally involve an initial step of conformer generation followed by a refinement step that minimizes differences between the generated library and experimental data. However, many computationally resolved ensembles can match the same experimental observations.

Molecular dynamics simulations (MD) have been extensively used over the years to navigate complex energy surfaces in other biological problems, i.e. in folding[15], protein–protein binding[16,17] and, modulation of IDPs by post-translational modifications[11] or by interacting with their folded partners[8]. In the context of IDP ensembles, MD simulations have been primarily applied as a tool for conformational generation. Nevertheless, the main goal of MD in this area would be to define reliable ensembles without the need for biasing or reweighting procedures.

[1]Computational Science Laboratory, Barcelona Biomedical Research Park (PRBB), Universitat Pompeu Fabra, C Dr Aiguader 88, 08003 Barcelona, Spain. [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain. ✉email: gianni.defabritiis@upf.edu

In this line, recent studies have employed enhanced sampling methods such as Hamiltonian replica exchange MD to define IDP ensembles matching the available experimental information[18]. In terms of aggregated time, the study run for $\sim 10\ \mu s$, while others have performed more extensive simulations, $\sim 200\ \mu s$ but in a single trajectory[19].

Current technologies allow MD simulations to reach aggregated times in the order of milliseconds[20], thus making this tool a valuable one for the exploration of biological systems at increasingly longer time scales[16]. The potential offered by high throughput MD simulations coupled with Markov State Models (MSMs)[21] analysis for the exploration of conformational landscapes of IDP has been tested in some aggregation-prone peptides[22,23]. The main advantage offered by this tandem is the possibility to address subpopulations within ensembles and to study the kinetics controlling them, rather than working with population averages. By focusing on the most relevant subpopulations and their kinetic properties, it is possible to gain insight into the emergence of partially ordered states in atomistic detail.

Here we make use of extensive, unbiased full-atom MD simulations and state-of-the-art MSMs to explore the structural variability of the N-terminal region of p53 in isolation. p53 is a widely studied protein, given its relation to oncogenic processes. It includes disordered sections at both N and C terminals, which interacts with various partners[24]. Of special importance is the short region comprising residues 10 to 40. Its forms a stable α -helix upon interaction with the Mdm2 protein[5] and has been the primary subject for many computational and experimental studies. The p53–Mdm2 complex has served as a template for the development of peptidomimetics drugs[25] and as the preferred benchmark for several MD studies aiming to reconstruct the binding process and the associated kinetics[16,26–29]. NMR and SAXS studies of the N-terminal region in isolation revealed a helicity profile similar to the one observed in the p53-Mdm2 complex, implying that bound conformations might also be sampled prior to binding[30].

The main results show the existence of many kinetically relevant states, accounting for $\sim 40\%$ of the equilibrium population, including high levels of secondary structural elements. In particular, simulations show the presence of an α-helix enriched states similar to the folded pose found in complex with Mdm2, as well as, a tangled interplay between β-strands formation leading to novel β-sheet enriched structures. Altogether, this illustrates the complexity of partially ordered states within the conformational space of an exemplar IDP, such as the N-terminal region of p53.

Results and discussion

Identification of secondary structure enriched states. The simulation time of the MD run totaled ~ 1.4 ms. Initially, the secondary structure of the aggregated MD was analyzed. Data showed the coexistence of both α-helix and β-strand, each one peaking at ~ 20 % in the central region of the protein (Fig. 1b). The helicity profile follows a bell-shaped distribution, while β-strand is more sparsely scattered in three groups in the proximity of residues S15, K24, and V31.

MD data was used to create an MSM based on $backbone_{C\alpha} + sidechain_{O,N}$ self distance matrix that splits the space into 11 different sets of kinetically related conformations referred to as macrostates (labeled as *M1-11*). MSM subpopulations successfully separate metastable sets of conformations enriched in each secondary structure type (Fig. 1c,d), implying that these structural elements do appear in a concerted way, rather than being the average of residue independent structural propensities.

The helicity profile displayed by the helix-enriched state matches the bound conformation of p53 when interacting with Mdm2 (Fig. 1a). It spams from residue T18 to L26, and maximum levels of helicity being found in W23—an essential amino acid for that interaction. Similar profiles arise from NMR studies[30]. Besides this state, many others also display various degrees of helicity (Fig. S1). The tendency of IDPs to acquire secondary structure profiles resembling their folded conformation has also been observed in other IDPs, and it has also been related to the binding mechanism to their partner[31] and their signaling properties[32].

For β-strand, segregation of secondary structural elements into their own states becomes especially evident in *M2*, where three β-strands—namely $\beta 1$, $\beta 2$, and $\beta 3$ from N to C terminal—are organized in an anti-parallel double-sheet (Fig. 1c), defining the partially ordered state with the most significant level of structure. Besides the secondary structure enriched macrostates aforementioned, many other states also exhibit different profiles of β -strand and α-helix (Figs. S1 and S2). This includes a number of states displaying different β-sheet arrangements, featuring only one strand, either $\beta 1$-$\beta 2$ or $\beta 2$-$\beta 3$. Altogether, this highlights the variety of possible configurations found across the conformational landscape of p53.

Kinetic characterization of the conformational landscape of p53. Population wise, partially ordered states account for a significant proportion of the equilibrium population ($\sim 40\%$, Fig. 2a). The triple-stranded macrostates, the most folded ones, have low populations ($< 1\%$), in contrast to double-stranded states like *M10*, which reach $\sim 20\%$ at equilibrium. However, the most populated state—*M11*; with $\sim 60\%$ of the population—is structurally heterogeneous, and lacks any secondary structural element or long-range contacts. Hence, partially ordered states are not energetically favored compared to the most extended configurations, and their free energies range from 0.5 to – 2.5 k cal M^{-1} (Fig. 2b). This can be visualized in more detail in the energy surface of p53 (Fig. S3). There are two well-defined minima separated by a small energy barrier. One of them is covered with the extended and the helical states (*M6,9,11*), and the other by $\beta 2$–$\beta 3$ conformations (*M10*). High energetic areas are occupied by the most structured states, like *M2*. Such profile, with many energetically similar states, fits the description of IDPs in isolation. The level of compaction of p53 follows a similar trend (Fig. S4). The most collapsed states have a radius of gyration close to the expected for a folded protein. More flexible ones combine rigid and non-rigid sections and sample extended conformations. The lack of over-compaction of the ensemble is in line with the capabilities of state-of-the-art force fields[33].

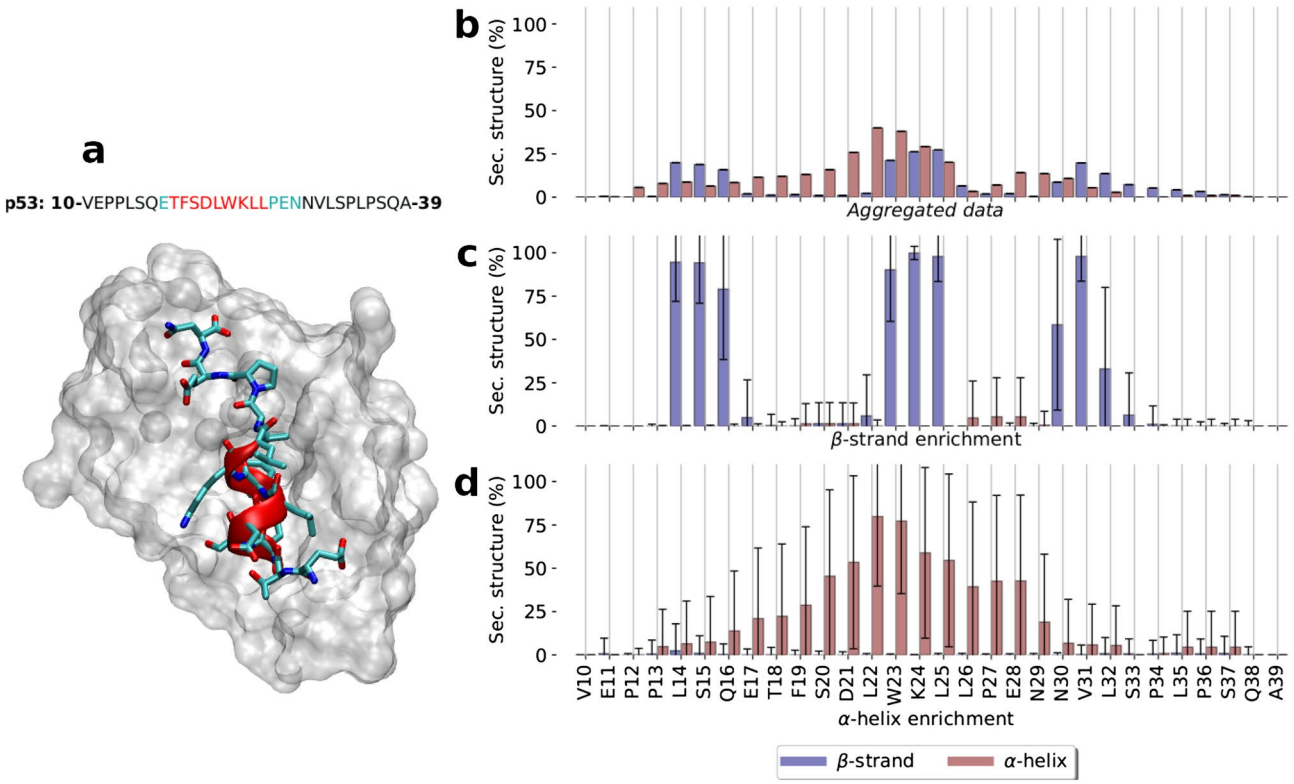

Figure 1. p53 secondary structure propensities (**a**) p53–MDM2 complex: MDM2 protein is shown as a white surface. p53 is depicted as cyan sticks and the helical region between residues 18 and 26 as red cartoon, (PDB code: *1YCR*). Above, the sequence of p53 used for the simulation is displayed: in red the helical section, in cyan the rest of the peptide found in the PDB structure, and in black the extended sequence. Secondary structure profiles derived from MD data: β-strand and α-helix profiles for the (**b**) aggregated data and for those macrostates of the MSM enriched in either (**c**) β-sheet and (**d**) α-helix.

The distinctive population profiles previously observed have an impact on the kinetic behavior of the macrostates. There is an approximately two orders of magnitude difference between maximum and minimum $mfpt_{on}$ estimations, which separate *faster/more populated (M6-10)* macrostates from *slower/less populated (M1-5)* macrostates (Fig. 2c). *Faster* macrostates comprise the helical conformation and several double-stranded states. *Slower* states, on the other hand, include triple stranded states and other low populated states. Off rates remain similar in all macrostates (Fig. 2d).

We employed transition path theory[34,35] to study the most relevant pathways and fluxes for macrostate interconversion. In particular, the focus was to elucidate the folding process leading from the less structured state to the triple-stranded conformation. There are three main paths (Fig. 3—central panel) involved in this process. The least transited one accounts for $\sim$ 15% of the total flow, and directly reaches the folded conformation from the extended one. On the other hand, the most transited paths involve the participation of double-stranded intermediates, with the $\beta1$–$\beta2$ structure taking $\sim$ 40% and the $\beta2$–$\beta3$ conformation being responsible for the remaining $\sim$ 30% of the flux. Additionally, other β-enriched states, such as the extended $\beta2$–$\beta3$ sheet found in *M10*, are disconnected from this network and can be directly reached from *M11* without the need of intermediates. It is interesting to point out that some conformations (such as *M3,10*), despite their structural similarity, show a $\sim$ 30 slowdown that explains the differences in stability aforementioned.

In summary, partially ordered states populating the conformational landscape are structural and kinetically diverse. States coexist at different timescales, even if they are structurally similar, such as the case of *M6,11*. These two states feature a short and an extended $\beta2$-$\beta3$ sheet but have k_{on} values of $5 \cdot 10^5$ M^{-1} s^{-1} and $1 \cdot 10^7$ M^{-1} s^{-1} respectively.

Comparison with NMR data. In order to ensure and validate MD observations, simulation data were compared against experimentally determined backbone chemical shifts (CS) for the N-terminal region of p53[36]. Experimental CS were resolved for the full-length N-terminal (residues 1–93), but only CS for residues 10 to 39 were used (to match the simulated sequence). CS allows inferring by-residue secondary structure tendencies on folded and disordered proteins. Calculations were performed using two softwares, SPARTA+[37] and SHIFTX2[38], on a set of 2 000 structures selected at random accordingly to the macrostate equilibrium probabilities. Calculated CS with both programs yielded similar results. Overall, there is a high correlation between experimental and MD calculated CS values for Cα, Cβ, and N with R^2 values of 0.98, 0.99, and 0.88 (Fig. S5). Differences between experimental and calculated CS remains within the intrinsic estimations error of each software

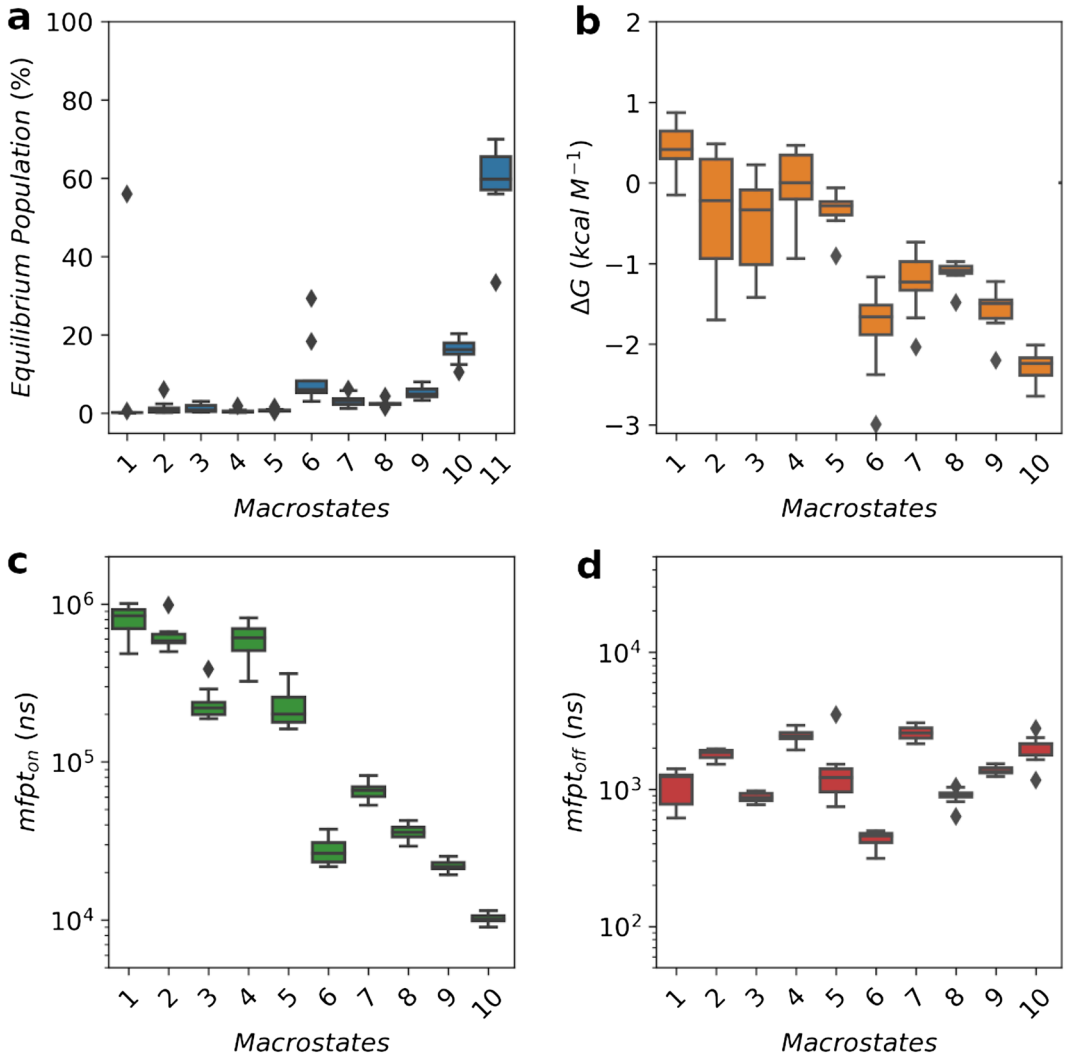

Figure 2. Estimation of (**a**) *equilibrium population*, (**b**) ΔG, (**c**) $mfpt_{on}$, (**d**) $mfpt_{off}$, for each macrostate after 10 rounds of bootstrap. Kinetics parameters for *M11* (structurally heterogeneous) are not shown as it was used as the source state for calculations.

($\sim$ 1 p.p.m for SPARTA+ and 0.4, 0.5 and 1.1 for Cα, Cβ and NH, respectively, for SHIFTX2) thus, indicating that structural rearrangements observed in MD data are in line with those determined by NMR experiments.

Conclusion

The characterization of the p53 conformational landscape using unbiased MD simulations revealed a high number of transient partially ordered states accounting for $\sim$ 40% of the equilibrium populations. Partial order arises from the formation of both α-helix and β-strand structural elements. The helical state resembles the structure acquired by p53 upon interaction with Mdm2. The MSM also showed the presence of several β-enriched states, not described before, that established long-range contacts through the arrangement of either one or two β-sheets. These processes are kinetically different, and some of the faster states are quickly accessible from the random-coiled macrostate and highly populated at equilibrium. Thus, it would be possible for some of them to play biologically relevant roles and could even provide novel strategies for the modulation of IDPs. Other computational studies with p53[39] also hinted the presence of collapsed conformations in this region. The exploration of aggregation-prone IDPs, such as the amyloid beta[22] and hIAPP[23], showed the spontaneous formation of β -hairpin metastable states, but with very small populations.

The current study provides a structural and kinetically detailed description of the conformational landscape of an IDP using MD simulations in combination with MSMs. Given the high number of short linear motifs within the human proteome, a similar pipeline could, in principle, be more extensively applied in order to investigate whether other IDPs may also share such complex behaviors. However, reaching millisecond simulation time may not prove scalable in more extensive studies with multiple targets. Novel adaptive sampling techniques[40,41], which perform a more intelligent exploration of surfaces, might mitigate this problem by reducing the computational time needed to achieve similar results. Some final considerations about the task is the computational effort

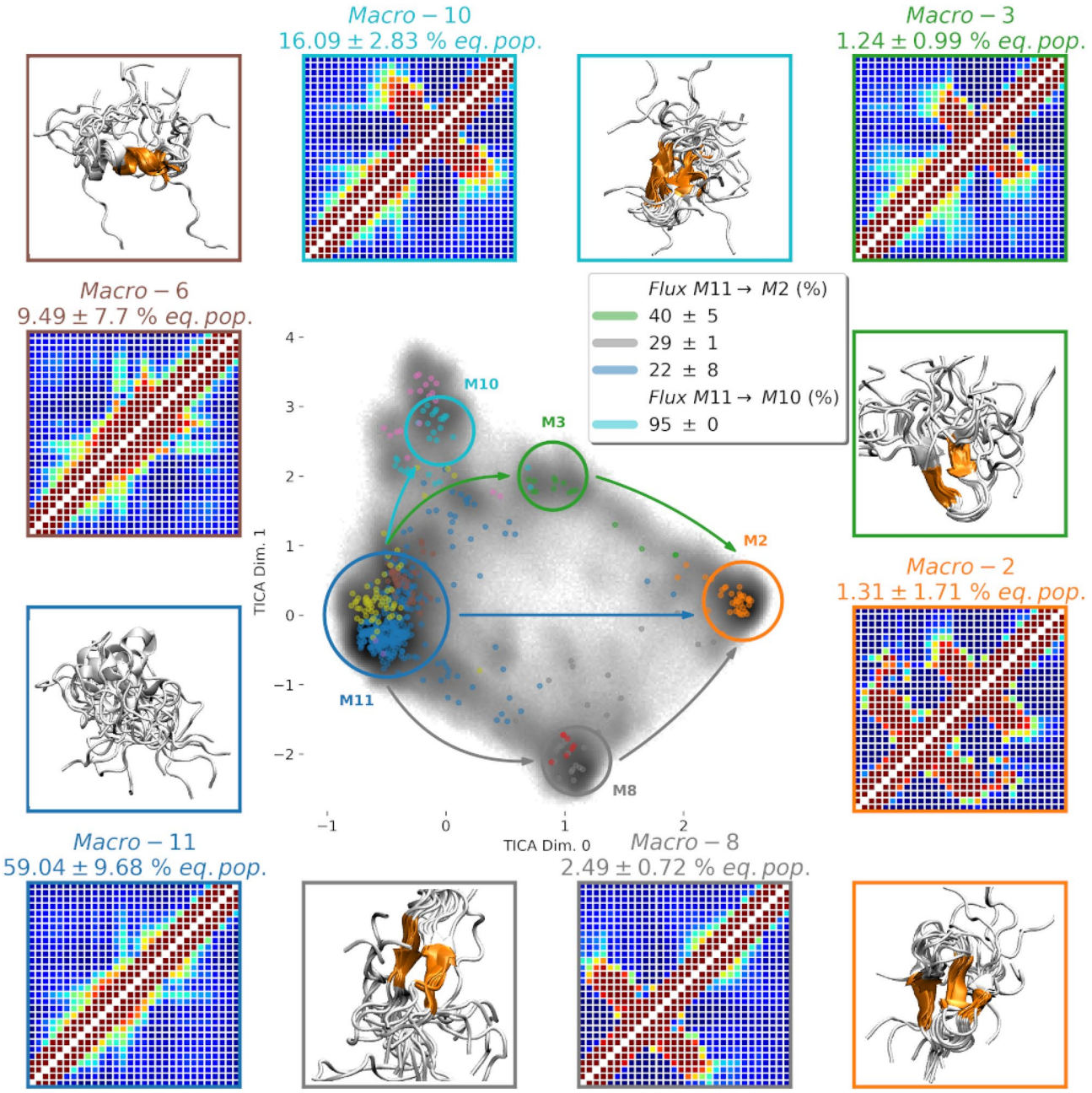

Figure 3. p53 conformational landscape. Central panel illustrates the first two dimensions of the TICA space. In grey, a 2D histogram 200 × 200 bins represents the frame count of the aggregated MD data. MSM microstates are distributed accordingly to their centers and colored with respect to their corresponding macrostate. Arrows represent the main pathways leading from the most extended macrostate (*Macro-11*) to the most β-sheet enriched ones (*Macro-2* and *Macro-10*). Correspondence between macrostate location in the central panel and side panels is color mapped. Side panels describe macrostates in terms of residue-residue contacts maps. Protein visualization is performed by superimposing 20 structures using residues highlighted in orange for structural alignment.

to simulate longer peptides as well as the need to compare with experimental data to validate partially folded states and their relaxation times. Here we use a relatively short 30 amino acids section of p53, while completely disordered domains may spam hundreds of residues.

Methods

Molecular dynamics simulation set up. In order to perform the exploration of the conformational space of p53, extensive parallel simulations were run. The selected region of p53 spammed from residue 10 to 39.

A set of 110 structures was used as initial conformations for the MD run (Fig. S8c). All systems were built with VMD[42] (version 1.9.2, http://www.ks.uiuc.edu/Research/vmd/), solvated with TIP3P[43] (each system included ~ 8,200 water molecules, resulting in a final protein concentration of ~ 6.8 mM), with a final NaCl concentration of

0.05 M. A Langevin integrator with a damping constant of 0.1 ps^{-1} was used. The integration step was set to 4 fs, with heavy hydrogen atoms scaled up to four times their natural mass. Electrostatic were computed using PME with a cutoff distance of 9 Å and grid spacing of 1 Å. Equilibration was performed at 300 K, firstly undergoing 250 steps of energy minimization followed by 0.1 ns simulations in an NVE ensemble (pressure was kept at 1 atm by using the berendesen barostat) and 2 ns in an NPT ensemble. We employed the CHARMM22* forcefield[44], a modification of the original CHARMM22 with adjusted backbone torsion potentials to produce more extended conformations. After equilibration, no proline *cis* isomers were detected.

Production runs of 1 μs were performed at 310 K using the distributed computing project GPUGrid[45] using the ACEMD engine[46] (included in HTMD).

Markov state model analysis. Production runs generated a total of 1.337 trajectories (each equilibrated system was used at least 10 times) of 1 μs each, similarly to[16]. Thus, the production runs accounted for an aggregated simulation time of $\sim$ 1.4 ms in order to maximize the exploration of the conformational space. All MD data analyses were performed using HTMD[47] (version 1.22.0 https://github.com/Acellera/htmd).

MD data was used to build a MSM. The analysis was performed by featurizing atomic coordinates as the self-distance matrix between C_α and side chains nitrogen and oxygen atoms. Next, time independent component analysis method (TICA[48]) reduced data dimensionality at a fixed lag time of 20 frames. The parameters for the last building stages were selected based on the generalized matrix Rayleigh quotient (GMRQ) scores (Figs. S6 and S7). A final number of 9 TICA dimension and 600 clusters were selected. Data clusters were defined using the MiniBatchKMeans algorithm[49].

Microstates were fused at a lag time of 120 ns, following the implied time scales plot (Fig. S8a) into 11 macrostates (using the PCCA+ algorithm[50] and based on the discretization of the TICA space shown in Fig. S8e). The Chapman–Kolmogrov test (Fig. S9) confirmed that parameters yield a markovian model. Finally, transition path theory[34,35] was used to calculate fluxes between states.

For every measure, the error was estimated by creating 10 independent bootstrap replicas of the MSM using a random set containing 80% of the trajectories.

Chemical shift calculations. Calculations of MD derived chemical shifts were performed using 2,000 frames distributed amongst macrostates based on their equilibrium probability. The biological magnetic resonance data bank entry *17760*[36] was used to obtain experimental chemical shift data for the N-terminal region of p53. The experiment was performed with the full-length N-terminal (residues 1 to 93), but for the comparison we used the section 10–39. Two different softwares were used: SPARTA+[37] (version 2.90, http://spin.niddk.nih.gov/bax/software/SPARTA+/index.html) and SHIFTX2 (version 1.09, http://www.shiftx2.ca/)[38].

Received: 8 April 2020; Accepted: 8 July 2020
Published online: 24 July 2020

References

1. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
2. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197 (2005).
3. Van Der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
4. Burger, V., Gurry, T. & Stultz, C. Intrinsically disordered proteins: Where computation meets experiment. *Polymers* **6**, 2684–2719 (2014).
5. Kussie, P. H. *et al.* Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274**, 948–953 (1996).
6. Russo, A. A., Jeffrey, P. D., Patten, A. K., Massagué, J. & Pavletich, N. P. Crystal structure of the p27Kip1 cyclin-dependent-kinase inibitor bound to the cyclin A–Cdk2 complex. *Nature* **382**, 325 (1996).
7. Zor, T., De Guzman, R. N., Dyson, H. J. & Wright, P. E. Solution structure of the KIX domain of CBP bound to the transactivation domain of c-Myb. *J. Mol. Biol.* **337**, 521–534 (2004).
8. Chong, S.-H., Im, H. & Ham, S. Explicit characterization of the free energy landscape of pkid–kix coupled folding and binding. *ACS Central Sci.* **5**, 1342–1351 (2019).
9. Iconaru, L. I. *et al.* Discovery of small molecules that inhibit the disordered protein, p27 kip1. *Sci. Rep.* **5**, 15686 (2015).
10. Ban, D., Iconaru, L. I., Ramanathan, A., Zuo, J. & Kriwacki, R. W. A small molecule causes a population shift in the conformational landscape of an intrinsically disordered protein. *J. Am. Chem. Soc.* **139**, 13692–13700 (2017).
11. Stanley, N., Esteban-Martín, S. & De Fabritiis, G. Kinetic modulation of a disordered protein domain by phosphorylation. *Nat. Commun.* **5**, 5272 (2014).
12. Bah, A. & Forman-Kay, J. D. Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* **291**, 6696–6705 (2016).
13. Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* **51**, 2224–2231 (2012).
14. Fisher, C. K. & Stultz, C. M. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **21**, 426–431 (2011).
15. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
16. Paul, F. *et al.* Protein–peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nat. Commun.* **8**, 1095 (2017).
17. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling. *Nat. Chem.* **9**, 1005 (2017).
18. Shrestha, U. R. *et al.* Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Nat. Acad. Sci.* **116**, 20446–20452 (2019).

19. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* **134**, 3787–3791 (2012).
20. Martinez-Rosell, G., Giorgino, T., Harvey, M. J. & de Fabritiis, G. Drug discovery and molecular dynamics: Methods, applications and perspective beyond the second timescale. *Curr. Top. Med. Chem.* **17**, 2617–2625 (2017).
21. Prinz, J.-H. *et al.* Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
22. Lin, Y.-S., Bowman, G. R., Beauchamp, K. A. & Pande, V. S. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer. *Biophys. J .* **102**, 315–324 (2012).
23. Qiao, Q., Bowman, G. R. & Huang, X. Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation. *J. Am. Chem. Soc.* **135**, 16092–16101 (2013).
24. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta (BBA) Proteins Proteomics* **1804**, 1231–1264 (2010).
25. Vassilev, L. T. *et al.* In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science* **303**, 844–848 (2004).
26. Zwier, M. C. *et al.* Efficient atomistic simulation of pathways and calculation of rate constants for a protein–peptide binding process: Application to the MDM2 protein and an intrinsically disordered p53 peptide. *J. Phys. Chem. Lett.* **7**, 3440–3445 (2016).
27. Morrone, J. A., Perez, A., MacCallum, J. & Dill, K. A. Computed binding of peptides to proteins with meld-accelerated molecular dynamics. *J. Chem. Theory Comput.* **13**, 870–876 (2017).
28. Zhou, G., Pantelopulos, G. A., Mukherjee, S. & Voelz, V. A. Bridging microscopic and macroscopic mechanisms of p53–MDM2 binding with kinetic network models. *Biophys. J .* **113**, 785–793 (2017).
29. Tran, D. P. & Kitao, A. Kinetic selection and relaxation of the intrinsically disordered region of a protein upon binding. *J. Chem. Theory Comput.* **16**, 2835–2845 (2020).
30. Wells, M. *et al.* Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proc. Nat. Acad. Sci.* **105**, 5762–5767 (2008).
31. Arai, M., Sugase, K., Dyson, H. J. & Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Nat. Acad. Sci.* **112**, 9614–9619 (2015).
32. Borcherds, W. *et al.* Disorder and residual helicity alter p53–mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* **10**, 1000–1002 (2014).
33. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Nat. Acad. Sci.* **115**, E4758–E4766 (2018).
34. Weinan, E. & Vanden-Eijnden, E. Towards a theory of transition paths. *J. Stat. Phys.* **123**, 503 (2006).
35. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Nat. Acad. Sci.* **106**, 19011–19016 (2009).
36. Wong, T. S. *et al.* Biophysical characterizations of human mitochondrial transcription factor a and its binding to tumor suppressor p53. *Nucleic Acids Res.* **37**, 6765–6783 (2009).
37. Shen, Y. & Bax, A. Sparta+: A modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
38. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. Shiftx2: Significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50**, 43 (2011).
39. Terakawa, T. & Takada, S. Multiscale ensemble modeling of intrinsically disordered proteins: p53 n-terminal domain. *Biophys. J .* **101**, 1450–1458 (2011).
40. Doerr, S. & De Fabritiis, G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.* **10**, 2064–2069 (2014).
41. Zimmerman, M. I. & Bowman, G. R. Fast conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
42. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
43. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
44. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization?. *Biophys. J .* **100**, L47–L49 (2011).
45. Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P. & De Fabritiis, G. High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.* **50**, 397–403 (2010).
46. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
47. Doerr, S., Harvey, M., Noé, F. & De Fabritiis, G. HTMD: High-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
48. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 07B604_1 (2013).
49. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Röblitz, S. & Weber, M. Fuzzy spectral clustering by pcca+: Application to markov state models and data classification. *Adv. Data Anal. Classif.* **7**, 147–179 (2013).

Acknowledgements

The authors thank volunteers at GPUGRID.net for contributing with computational resources and Acellera for funding. G.D.F. acknowledges support from MINECO (Unidad de Excelencia María de Maeztu MDM-2014-0370 and BIO2017-82628-P) and FEDER. This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 823712 (CompBioMed2 Project).

Author Contributions

G.D.F. conceived the experiments, P.H.N and A.P. analyzed the results. P.H.N wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69322-2.

Correspondence and requests for materials should be addressed to G.F.