nature structural & molecular biology

Crystal structure of a monomeric retroviral protease solved by protein folding game players

Firas Khatib[1], Frank DiMaio[1], Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper[2], Maciej Kazmierczyk[3], Miroslaw Gilski[3,4], Szymon Krzywda[3], Helena Zabranska[5], Iva Pichova[5], James Thompson[1], Zoran Popović[2], Mariusz Jaskolski[3,4] & David Baker[1,6]

Following the failure of a wide range of attempts to solve the crystal structure of M-PMV retroviral protease by molecular replacement, we challenged players of the protein folding game Foldit to produce accurate models of the protein. Remarkably, Foldit players were able to generate models of sufficient quality for successful molecular replacement and subsequent structure determination. The refined structure provides new insights for the design of antiretroviral drugs.

© 2011 Nature America, Inc. All rights reserved.

Foldit is a multiplayer online game that enlists players worldwide to solve difficult protein-structure prediction problems. Foldit players leverage human three-dimensional problem-solving skills to interact with protein structures using direct manipulation tools and algorithms from the Rosetta structure prediction methodology[1]. Players collaborate with teammates while competing with other players to obtain the highest-scoring (lowest-energy) models. In proof-of-concept tests, Foldit players—most of whom have little or no background in biochemistry—were able to solve protein structure refinement problems in which backbone rearrangement was necessary to correctly bury hydrophobic residues[2]. Here we report Foldit player successes in real-world modeling problems with more complex deviations from native structures, leading to the solution of a long-standing protein crystal structure problem.

Many real-world protein modeling problems are amenable to comparative modeling starting from the structures of homologous proteins. To make use of homology modeling techniques in Foldit, we introduced a new capability called the Alignment Tool, which allows players to manually move alignments and thread their sequence onto the structures of known homologs (**Supplementary Fig. 1**). Players are able to combine different regions from multiple templates into a single hybrid structure (partial threading) and load in previously saved solutions as templates to hybridize with their current models.

Our aim was for Foldit players to use these new tools to solve real-world problems; the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment was an ideal venue in which to test this. CASP is a biennial experiment in protein structure prediction methods in which the amino acid sequences of structures that are close to being experimentally determined—referred to as CASP targets—are posted to allow groups from around the world to predict the native structure (http://predictioncenter.org/casp9/). Each group taking part in CASP is allowed to submit five different predictions for each sequence. Foldit participated as an independent group during CASP9 and made predictions for the targets with fewer than 165 residues that the CASP organizers did not indicate as oligomeric. For targets with homologs of known structure—the Template-Based Modeling category—Foldit players were given different alignments to templates predicted by the HHpred server[3] via the new Alignment Tool. Despite these new additions to the game, the performance of Foldit players over all CASP9 Template-Based Modeling targets was not as good as those of the best-performing methods, which made better use of information from homologous structures; extensive energy minimization used by Foldit players tended to perturb peripheral portions of the chain away from the conformations present in homologs.

For prediction problems for which there were no identifiable homologous protein structures—the CASP9 Free Modeling category—Foldit players were given the five Rosetta Server CASP9 submissions (which were publicly available to other prediction groups) as starting points, along with the Alignment Tool. Here all five starting models were available, allowing players to use partial threading to combine different features of the Rosetta models. In this Free Modeling

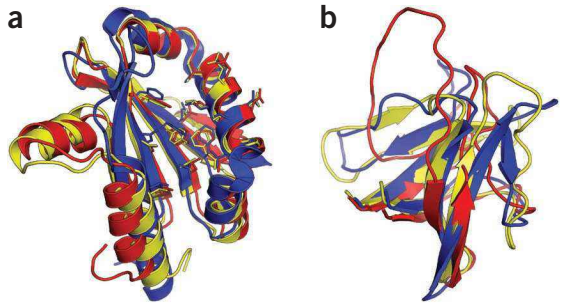

Figure 1 Successful CASP9 predictions by the Foldit Void Crushers Group. (**a**) Starting from the fourth-ranked Rosetta Server model (red) for CASP9 target T0581, the Foldit Void Crushers Group (yellow) generated a model that was closer to the crystal structure later determined (blue). (**b**) Starting from a modified Rosetta model built using the Alignment Tool (red), the Foldit Void Crushers Group generated a model (yellow) considerably closer to the later determined crystal structure (blue). Images were produced using PyMOL software (http://www.pymol.org).

[1]Department of Biochemistry, University of Washington, Seattle, Washington, USA. [2]Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA. [3]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland. [4]Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. [5]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Poznan, Czech Republic. [6]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to D.B. (dabaker@u.washington.edu).

Received 27 May; accepted 8 July; published online 18 September 2011; doi:10.1038/nsmb.2119

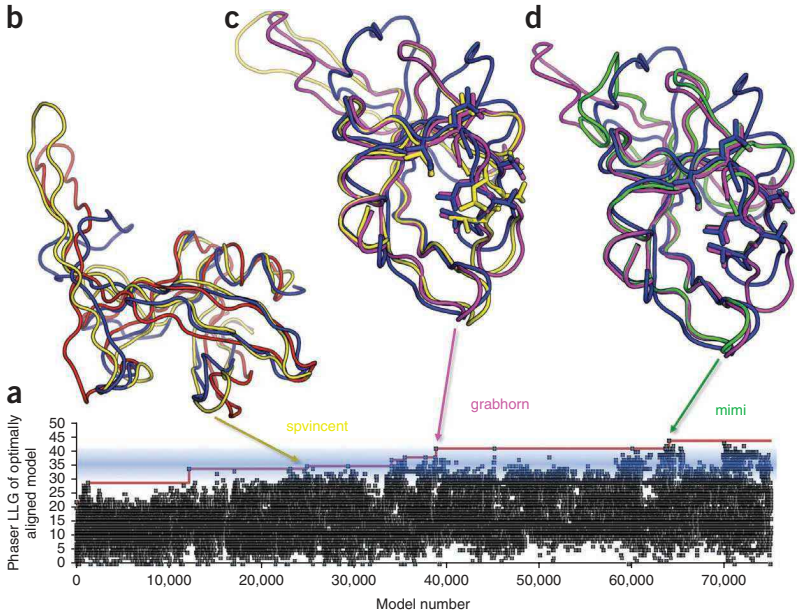

Figure 2 M-PMV retroviral protease structure improvement by the Foldit Contenders Group. (**a**) Progress of structure refinement over the first 16 d of game play. The *x* axis shows progression in time, and the *y* axis shows the Phaser log-likelihood (LLG) of each model in a near-native orientation. To identify a solution as correct by molecular replacement using Phaser, the model must have an LLG better than the best random models. The distribution of these best random predictions is indicated by the intensity of the pale blue band. (Because almost all the models are too poor to allow correct placement in the unit cell, Phaser LLGs are calculated after optimal superposition of each model onto the solved crystal structure and rigid-body optimization.) (**b**) Starting from a quite inaccurate NMR model (red), Foldit player spvincent generated a model (yellow) considerably more similar to the later determined crystal structure (blue) in the β-strand region. (**c**) Starting from spvincent's model, Foldit player grabhorn generated a model (magenta) considerably closer to the crystal structure with notable improvement of side-chain conformations in the hydrophobic core. (**d**) Foldit player mimi made additional improvements (in the loop region at the top left) and generated a model (green) of sufficient accuracy to provide an unambiguous molecular replacement solution which allowed rapid determination of the ultimate crystal structure (blue).

© 2011 Nature America, Inc. All rights reserved.

category, some of the shortcomings of the Foldit predictions became clear. The main problem was a lack of diversity in the conformational space explored by Foldit players because the starting models were already minimized with the same Rosetta energy function used by Foldit. This made it very difficult for Foldit players to get out of these local minima, and the only way for the players to improve their Foldit scores was to make very small changes ('tunneling' to the nearest local minimum) to the starting structures. However, this tunneling did lead to one of the most spectacular successes in the CASP9 experiment.

De novo structure prediction remains an exceptionally challenging problem, and very few predictions with atomic accuracy have been made in the history of CASP. For CASP9 target T0581, starting from an extended chain, the Rosetta Server, which carried out a large-scale search for the lowest-energy structure using computing power from Rosetta@home volunteers (http://boinc.bakerlab.org/rosetta/), produced a remarkably accurate model (**Fig. 1a**; compare red and blue). However, the server ranked this model fourth out of the five submissions. The Foldit Void Crushers team correctly selected this near-native model and further improved it by accurately moving the terminal helix, producing the best model for this target of any group and one of the best overall predictions at CASP9 (ref. 4) (**Fig. 1a**; compare yellow and blue). Thus, in a situation where one model out of several is in a near-native conformation, Foldit players can recognize it and improve it to become the best model. Unfortunately for the other Free Modeling targets, there were no similarly outstanding Rosetta Server starting models, so Foldit players simply tunneled to the nearest incorrect local minima.

The CASP9 Refinement category provides groups with the best predictions made for selected targets and challenges them to improve the predictions further. Foldit participated in the Refinement category for all non-oligomeric targets. Many refinement targets at CASP9 were models created using Rosetta, resulting in a similar tunneling problem as with the initial Free Modeling predictions. Using the lessons learned from those targets, we tried presenting problems in a way that encouraged Foldit players to make more dramatic changes to the starting models. We experimented with only allowing the Foldit score to count if the r.m.s. deviation with the starting model was greater than a certain threshold, but—even with these conditions—players found it difficult to make improvements because almost all perturbations increased the very low starting Rosetta energy.

For the very last CASP9 refinement target, TR624, Foldit players struggled with the same problem: the starting model was Rosetta optimized, and very few players were able to satisfy the r.m.s. deviation conditions while at the same time finding lower energies (**Supplementary Fig. 2a**). We therefore decided to repost the puzzle after perturbing the structure out of its energy minimum. We used the Alignment Tool to align the regions the CASP organizers identified as correct, and we threaded the sequence onto the correctly aligned portions of the starting structure (**Supplementary Fig. 2b**). The unaligned portions were rebuilt randomly, initially with a poor energy, encouraging diversification of models in the incorrect regions, while maintaining the favorable interactions in regions known to be

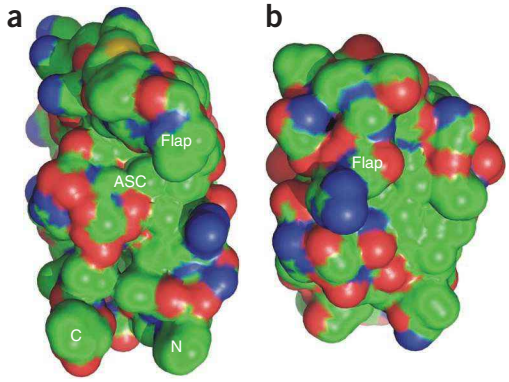

Figure 3 CPK representation of retropepsin surface. (**a**) The surface of HIV-1 PR protomer extracted from the dimeric molecule (PDB 3hvp), as seen from the direction of the removed dimerization partner. (**b**) M-PMV PR monomer shown in the same orientation and scale. In this view, the N and C termini (missing in M-PMV PR) are at the bottom, and the flap loop is at the top. The active-site cavity (ASC) is clearly seen between the flap and the body of the HIV-1 PR molecule. In M-PMV PR, the cavity is completely covered by the curled flap.

near-native. Using this modified model as a start to a new puzzle, Foldit players were able to sample closer to the native fold by rebuilding several incorrect loops (**Fig. 1b**, compare yellow and blue). The submitted top-scoring Foldit prediction by the Void Crushers Group for this puzzle also used the Alignment Tool to partially thread one of their previous solutions and was the best-ranked model for the most difficult refinement challenge in CASP9, TR624 (ref. 5) (**Supplementary Fig. 2c**,**d**).

The most important problem solved by Foldit players to date came after CASP9 and involves the Mason-Pfizer monkey virus (M-PMV) retroviral protease. Retroviral proteases (PRs) have critical roles in viral maturation and proliferation and are the focus of intensive antiretroviral drug development work[6]. All previously determined crystal structures of retroviral proteases show the biologically active homodimeric form[7]; prevention of PR dimerization has been proposed as a mechanism for disruption of PR activity[8] and a drug design avenue for antiretrovirals. The PR of M-PMV, a simian AIDS–causing monkey virus, crystallizes as a monomer, but despite the availability of several crystal forms, researchers have for over a decade been unable to solve the structure by molecular replacement (MR) using either homodimer-derived models or an NMR structure of the protein monomer[9]. A recent approach using density- and energy-guided structure refinement[10] was also unsuccessful at determining a solution, despite a success rate of over 50% using this method in cases where similarly good homologous template structures were available. Of the failures described in reference 10, M-PMV PR was the most suitable for Foldit as it is a monomer and fairly small (114 residues).

To determine whether human intuition could succeed where automated methods had failed, we challenged Foldit players to build accurate models of M-PMV PR starting from the NMR coordinates (which had failed in MR tests; see **Supplementary Discussion** and **Supplementary Fig. 3**). When the 3-week competition concluded, we screened the top-scoring Foldit models using Phaser[11] to determine whether any were of sufficient quality for MR. Remarkably, despite the complete failure of all previous approaches, several solutions by the Foldit Contenders Group produced phase estimates that were good enough to allow a rapid solution of the crystal structure.

We provided Foldit players with the ten different NMR models, which all scored poorly using Rosetta's energy function, so that players would not be trapped in local energy minima, and we included all ten NMR models as templates in the Alignment Tool. The improvement in model accuracy by the Foldit Contenders Group is illustrated in **Figure 2a**. As gameplay progressed (*x* axis), model accuracy and suitability for MR—as assessed by the Phaser log-likelihood gain (LLG) (*y* axis)—increased, with several notable jumps (arrows). **Figure 2b–d** illustrates some of the breakthrough models produced by Foldit players. Foldit player spvincent (yellow) used partial threading with the Alignment Tool and quickly improved the starting NMR model (red) to have much better agreement with the crystal structure later determined (blue). Another teammate, grabhorn, was able to improve spvincent's model (magenta), particularly in the core of the protein, and another teammate, mimi, was able to generate an even more accurate model (green) by correctly tucking in the loop at the top left. The LLG of this model was high enough to allow its unambiguous identification as a likely MR solution among the vast number of Foldit models, and standard autobuilding and structure refinement methods showed within hours that the solution was almost certainly correct. Using the Foldit solution, the final refined structure was completed a few days later (**Supplementary Table 1**).

Of particular interest in this monomeric retropepsin structure is the molecular surface that normally forms the dimer interface in homodimeric retroviral protease molecules. There is a considerable backbone rearrangement in this area in which a flap curls over the half-active site in the monomer and the N- and C termini are completely disordered (**Fig. 3**). These features provide opportunities for the design of antiretroviral drugs, including anti-HIV drugs; compounds that bind to the surface formed in the monomer, but not the dimer, should shift the equilibrium in favor of the former, which is catalytically inactive.

The critical role of Foldit players in the solution of the M-PMV PR structure shows the power of online games to channel human intuition and three-dimensional pattern-matching skills to solve challenging scientific problems. Although much attention has recently been given to the potential of crowdsourcing and game playing, this is the first instance that we are aware of in which online gamers solved a longstanding scientific problem. These results indicate the potential for integrating video games into the real-world scientific process: the ingenuity of game players is a formidable force that, if properly directed, can be used to solve a wide range of scientific problems.

Accession codes. Protein Data Bank: Coordinates of the protein atoms, together with structure factors, for the monomeric M-PMV retroviral protease have been deposited under accession number 3SQF.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS
We would like to acknowledge the members of the Foldit team for their help designing and developing the game and all the Foldit players and Rosetta@home volunteers who have made this work possible. This work was supported by the Center for Game Science at the University of Washington, US Defense Advanced Research Projects Agency (DARPA) grant N00173-08-1-G025, the DARPA PDP program, US National Science Foundation grants IIS0811902 Work supported in part by grants 1M0508 and Z40550506 from the Czech Ministry of Education to I.P., the Howard Hughes Medical Institute (D.B.) and Microsoft Corp. This material is based upon work supported by the National Science Foundation under grant no. 0906026.

AUTHOR CONTRIBUTIONS
F.K., F.D., S.C., J.T., Z.P. and D.B. contributed to the development and analysis of Foldit and to the writing of the manuscript; the F.C.G. and F.V.C.G. contributed through their gameplay, which generated the results for this manuscript; M.K. grew the crystals and collected X-ray diffraction data; M.G. processed X-ray data and analyzed the structure; S.K. refined the structure; H.Z. cloned, expressed and purified the protein; I.P. designed and coordinated the biochemical experiments, and contributed to writing the manuscript; M.J. coordinated the crystallographic study, analyzed the results and contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D. *Methods Enzymol.* **383**, 66–93 (2004).
2. Cooper, S. *et al. Nature* **466**, 756–760 (2010).
3. Söding, J., Biegert, A. & Lupas, A.N. *Nucleic Acids Res.* **33**, W244–W248 (2005).
4. Kinch, L. *et al.* CASP9 assessment of free modeling target predictions. *Proteins* (in the press).
5. MacCallum, J.L. *et al.* Assessment of protein structure refinement in CASP9. *Proteins* published online, doi:10.1002/prot.23131 (30 August 2011).
6. Mastrolorenzo, A., Rusconi, S., Scozzafava, A., Barbaro, G. & Supuran, C.T. *Curr. Med. Chem.* **14**, 2734–2748 (2007).
7. Wlodawer, A. & Gustchina, A. *Biochim. Biophys. Acta* **1477**, 16–34 (2000).
8. Koh, Y. *et al. J. Biol. Chem.* **282**, 28709–28720 (2007).
9. Veverka, V. *et al. J. Mol. Biol.* **333**, 771–780 (2003).
10. DiMaio, F. *et al. Nature* **473**, 540–543 (2011).
11. McCoy, A.J. *et al. J. Appl. Cryst.* **40**, 658–674 (2007).

© 2011 Nature America, Inc. All rights reserved.