Database, 2020, 1–7
doi: 10.1093/database/baaa031
Database tool

Database tool

Revenant: a database of resurrected proteins

Matias Sebastian Carletti[1,†], Alexander Miguel Monzon[1,2,†], Emilio Garcia-Rios[3], Guillermo Benitez[1], Layla Hirsh[3], Maria Silvina Fornasari[1] and Gustavo Parisi[1,*]

[1]Departamento de Ciencia y Tecnología, CONICET, Universidad Nacional de Quilmes, Roque Saenz Peña 182, Bernal, B1876BXD, Buenos Aires, Argentina, [2]Department of Biomedical Sciences, University of Padova, Viale G. Colombo 3, Padova, I-35131, Padova, Italy, and [3]Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú.

*Corresponding author: E-mail: gusparisi@gmail.com
†These authors contributed equally to this work.

Citation details: Carletti,M. S., Monzon,A. M., Garcia-Rios,E. *et al.* Revenant: a database of resurrected proteins. *Database* (2020) Vol. 2020: article ID baaa031; doi:10.1093/database/baaa031

Received 3 July 2019; Revised 6 March 2020; Accepted 31 March 2020

Abstract

Revenant is a database of resurrected proteins coming from extinct organisms. Currently, it contains a manually curated collection of 84 resurrected proteins derived from bibliographic data. Each protein is extensively annotated, including structural, biochemical and biophysical information. Revenant contains a browse capability designed as a timeline from where the different proteins can be accessed. The oldest Revenant entries are between 4200 and 3500 million years ago, while the younger entries are between 8.8 and 6.3 million years ago. These proteins have been resurrected using computational tools called ancestral sequence reconstruction techniques combined with wet-laboratory synthesis and expression. Resurrected proteins are commonly used, with a noticeable increase during the past years, to explore and test different evolutionary hypotheses such as protein stability, to explore the origin of new functions, to get biochemical insights into past metabolisms and to explore specificity and promiscuous behaviour of ancient proteins.

Database URL: http://revenant.inf.pucp.edu.pe/

Introduction

As a time machine, a combination of *in silico* and wet laboratory approaches allow the prediction of most probable sequences of proteins coming from organisms that lived millions of years ago (1). These predicted and synthesized sequences coming from extinct organisms are called resurrected proteins. Protein resurrection consists mainly in five steps (Figure 1) (2, 3). In the first one, a set of extant sequences, homologous to the ancestral protein to be studied, are aligned and used to estimate a phylogenetic tree. Phylogenetic trees are used to infer evolutionary relationships between ancestral and extant organisms. In a tree, extant organisms are represented by the terminal nodes or tips of the tree, (Figure 1), while

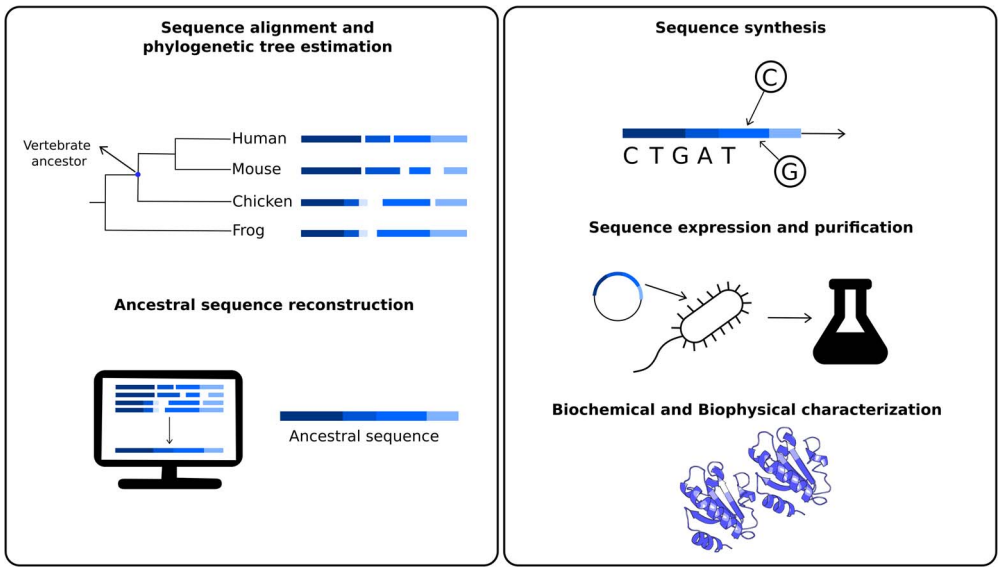

Figure 1. Schematic representation of the different steps to obtain resurrected proteins. The first step involves sequence similarity searches of a given protein to obtain a set of homologous sequences, involving the ancestral nodes to be studied. For example, one could be interested in studying biochemical properties of the studied protein in the last common ancestor for all vertebrates. Using these sequences, it is possible to estimate a phylogenetic tree to define the ancestral node to be reconstructed. In the second step, ancestral sequence reconstruction techniques are applied to estimate most probable sequences in the studied node. The third step involves the ancestral sequence synthesis. This sequence is then inserted into a vector, cloned, expressed and purified (fourth step). The fifth and final step involves a series of biochemical and biophysical characterization.

the ancestral organisms are represented by internal nodes. As an example, Figure 1 indicated the node representing the extinct organism corresponding to the last common ancestor from all extant vertebrates, or at least for those present in the set of homologous sequences considered. Using these evolutionary relationships, in the second step, it is possible to infer the most probable sequence for each of the ancestral states using the so called ancestral sequence reconstruction (ASR) techniques. These methods comprise a set of computational tools using different algorithms such as maximum parsimony (4), maximum likelihood (5–7) as well as Bayesian methods (8). In the third step, the predicted ancestral sequences could be synthesized using molecular biology techniques. If the ancestral reconstruction involves recent ancestors, site-directed mutagenesis using an extant gene can be used to obtain the ancestral sequence (9). However, in those cases where remote proteins are resurrected, gene synthesis (10) or gene fragments assembly are required to obtain the ancestral gene. Once synthesized, the gene is cloned, expressed and purified (fourth step). Then, the protein could be further experimentally characterized and studied as any other present day protein (fifth step). Most of these experiments consist in different biochemical and biophysical characterizations and also structural determination using X-ray crystallography or nuclear magnetic resonance.

The reliability of protein resurrection relies on the inference protocol used. Ambiguous estimation, dependence on tree topologies, approximations in evolutionary models and the difficulty to model indels are among the most common problems detected in resurrection (1, 2, 11–13). However, better results are obtained considering and minimizing those caveats. Furthermore, the creation of experimentally based phylogenies has contributed with a controlled system for ASR algorithms benchmarking (14).

Besides their caveats, ASR and protein resurrection are powerful strategies for testing different biological hypotheses. For example, phenotypic adaptations in dim-vision in vertebrates were recently elucidated using ancestral reconstruction and biochemical experimentation (15). Dim-vision in vertebrates is mediated with a family of proteins called rhodopsins. Slight variations in rhodopsin sequences during evolution confer the molecular basis of the spectral tuning observed in different organisms adapted to their environments (16). The authors found that 15 replacements (~3% of the average length of the rhodopsins) are essential to understand the functional adaptation. Moreover, some of these replacements occurred multiple times in vertebrate evolution strongly suggesting the existence of positive selection. Using positive selection analysis over a set of homologous proteins is a commonly used procedure to detect important positions associated with functional adaptations (17). However, this evolutionary pattern was not detected using rhodopsins evolutionary analysis, showing the importance of the use of ASR and biochemical characterization to unveil the evolution of

Figure 2. Two different browsing capabilities are available in Revenant. In the first one (top panel) proteins are listed sequentially using their RV codes. In the second browser (bottom panel) we display the Revenant proteins in an Earth's timeline showing important biological events since the origin of life.

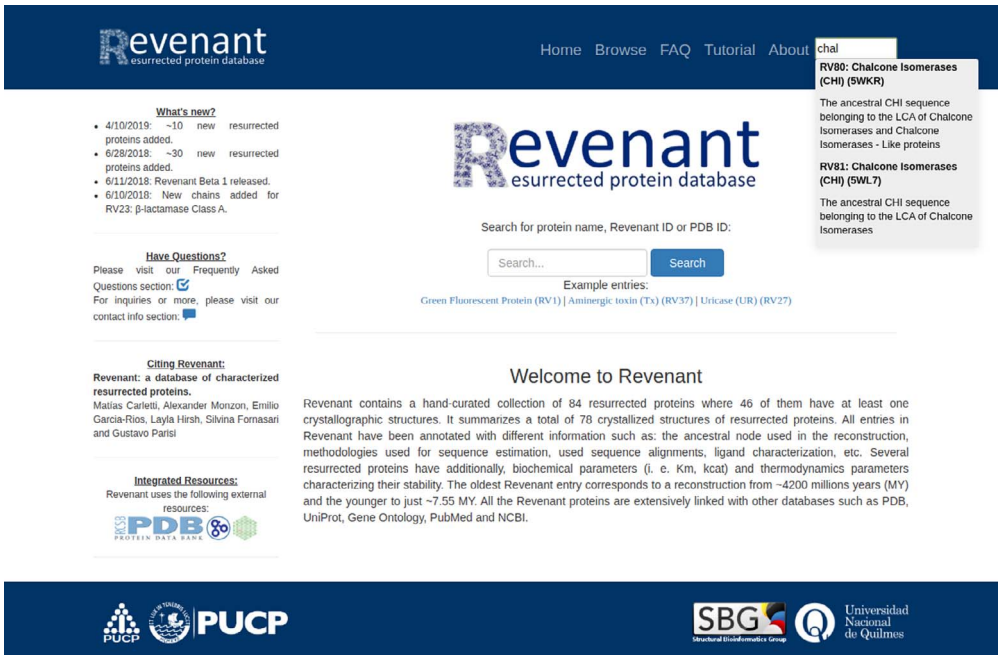

Figure 3. Screenshot of Revenant web server showing the home page and search utilities.

dim-vision in vertebrates. In a similar way, the used of ASR and resurrection techniques has had a key role to address different biological questions, such as specificity and biological activity (18), stability (19), promiscuity (20), study of alternative evolutionary histories (21), epistasis (22), evolutionary analysis of visual pigments (23), rational engineering (24) and emergence of new active sites (25), influence of evolutionary trajectories (26) and effect of duplication in functional divergence (27) just to mention a few of a large list of examples. Interestingly, as phylogenies could be calibrated with fossil records, resurrected proteins could recreate most probable states of proteins spanning very different geological times. The most challenging resurrections are about the very beginning of life on Earth (~4000 millions of years (28)).

In this work we present Revenant, the first database of resurrected proteins. It contains a manually curated collection of resurrected proteins which have been biochemically, biophysically and/or structurally characterized. Revenant proteins span several millions of years. The oldest entry corresponds to a reconstruction age between 4200 and 3500 million years ago which corresponds to the thioredoxin protein (28) (RV9, RV10 and RV11) and the younger entries between 8.8 and 6.3 million years ago corresponding to uricase (29) (RV74). Revenant proteins could display unique ancestral features. As the explained above example with the rhodopsins, experimental assays on resurrected proteins could reveal their structural arrangements, conformational diversity and dynamisms, differential stability and ligand binding affinities. These piece of evidence, along with the use of molecular phylogenies, could represent extremely useful information to test hypotheses about the origin of promiscuity, conformational epistasis, structural divergence and functional diversification grounded on a large-scale analysis. Also, the availability of a curated database as Revenant could offer a resource for evaluating the impact of evolutionary trajectories (22) on broadly used bioinformatic methods as homology modelling (30) as well as to test mechanistic evolutionary models of proteins (31, 32).

Database fields and contents

Each resurrected protein in Revenant represents the most probable sequence in a given node for a given phylogenetic analysis. Likewise, Revenant contains 84 entries (i.e. RV1-RV84) where 45 of them have at least one known crystallographic structure. Considering different structures of the same protein, Revenant contains a total of 78 crystallized structures of resurrected proteins. Using bibliographic information and manual curation, all entries have been annotated with different information such as the ancestral node used in the reconstruction, its estimated age, ASR methodologies used for sequence estimation, sequences and softwares used for the multiple alignments and phylogenetic estimation, structure availability and their ligand characterization and primary citation. Additionally, several entries have biochemical (i. e. Km, kcat) and biophysical parameters (i.e. melting temperature, ΔGunfolding). Furthermore, all the Revenant proteins are extensively linked with other databases such as PDB (33), UniProt (34), Gene Ontology (35) and PubMed.

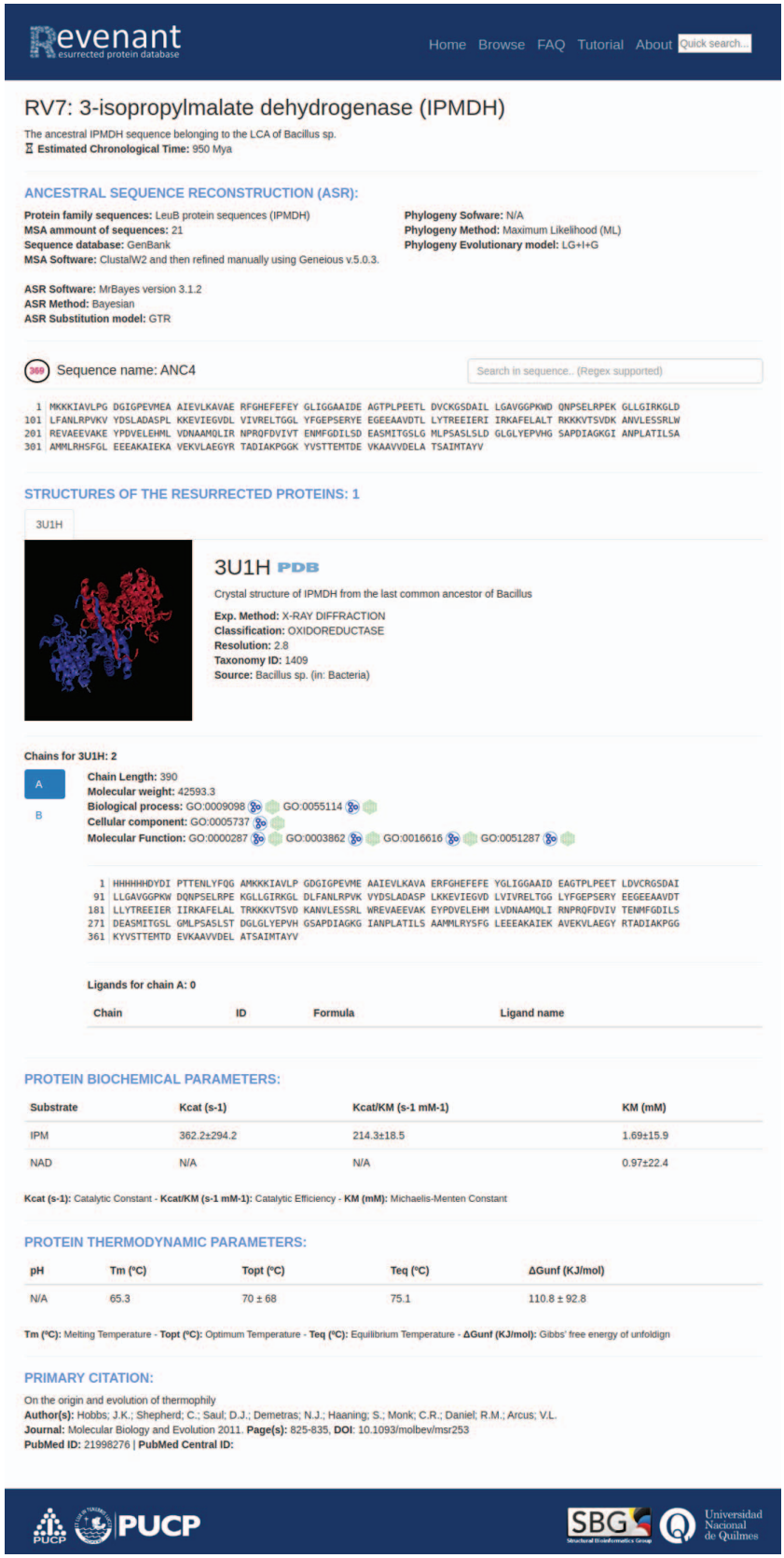

Figure 4. Main entry page. Each entry starts with a title followed by a brief explanation of the biological relevance of the resurrected protein. Additionally, each entry has fields regarding ancestral sequence reconstruction, information about their structures, biochemical and biophysical parameters and, finally, the primary citation.

Database access and user interface

Resurrected proteins in Revenant can be easily found searching by protein family name and/or PDB code. The browser contains two modes for protein search, one displays all Revenant entries as a list and the other shows a geological timeline indicating the approximate age of each Revenant entry (Figure 2).

Using the search or browse capabilities it is possible to access all Revenant entries. They are displayed along with a short description about the resurrected protein and, when it is available, the approximate age of the ancestral node (Figure 3). Further information is displayed in four different sections (Figure 4) for each entry: 'ancestral sequence reconstruction' contains all the information related to the ASR approach used for a given reconstruction. It also shows the reconstructed sequence and its name. The 'structures of the resurrected proteins' section summarizes the information about available structures, ligands, chains and biological function. The third section contains information about protein biochemical parameters like k_{cat} and affinity constants (such as K_M) for given ligands as well as thermodynamic parameters (such as T_{opt}, T_m, and ΔGunfolding). Actually, ~23% of the entries in Revenant contain physicochemical information. The fourth and last section shows information about primary citation where the protein was resurrected. Revenant website also contains Frequently Asked Questions (FAQ) and tutorial sections to allow non-expert users to easily explore the database.

Implementation

Revenant database was designed with microservices architecture. Two main elements of the system are the presentation and data components. The presentation elements exchange data using a RESTful API and the JavaScript Object Notation. The Java programming language and Spring framework leverage the data component implementation. MySQL is used for data storage and the ReactJS framework is used for presentation. Revenant offers users both graphical web interface access and RESTful web services from http://revenant.inf.pucp.edu.pe/.

Conclusions

Revenant database offers a well-curated, updated and annotated collection of resurrected proteins. We think that Revenant can be used to explore the fascinating world of the increasing examples of resurrected proteins and their use to illuminate interesting biological and evolutionary questions (36). Furthermore, our database of ancient proteins could also be a source of sequence, structure, conformational diversity and biochemical data to test further biological hypothesis and to develop new tools related with structural bioinformatics, 3D protein modelling and protein evolution.

Acknowledgements

G.P. and M.S.F. are CONICET researchers and M.C and G. B are CONICET PhD Fellows.

Funding

Research programme 'MSCA Seal of Excellence @UniPD' (to A.M.M.); Universidad Nacional de Quilmes (grant PUNQ 1402/15); ANCyT (PICT-2014 3430).

Conflict of interest. None declared.

References

1. Gumulya,Y. and Gillam,E.M.J. (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the "retro" approach to protein engineering. *Biochem. J.*, **474**, 1–19.
2. Thornton,J.W. (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.*, **5**, 366–375.
3. Merkl,R. and Sterner,R. (2016) Ancestral protein reconstruction: techniques and applications. *Biol. Chem.*, **397**, 1–21.
4. Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406.
5. Yang,Z., Kumar,S. and Nei,M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–1650.
6. Pupko,T., Pe'er,I., Shamir,R. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.
7. Koshi,J.M. and Goldstein,R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.*, **42**, 313–320.
8. Schultz,T.R., Cocroft,R.B. and Churchill,G.A. (1996) The reconstruction of ancestral character states. *Evolution*, **50**, 504–511.
9. Stackhouse,J., Presnell,S.R., McGeehan,G.M. *et al.* (1990) The ribonuclease from an extinct bovid ruminant. *FEBS Lett.*, **262**, 104–106.
10. Dillon,P.J. and Rosen,C.A. (1990) A rapid method for the construction of synthetic genes using the polymerase chain reaction. *BioTechniques*, **9**, 298–300.
11. Groussin,M., Hobbs,J.K., Szöllősi,G.J. *et al.* (2015) Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Mol. Biol. Evol.*, **32**, 13–22.
12. Bar-Rogovsky,H., Stern,A., Penn,O. *et al.* (2015) Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Eng. Des. Sel.*, **28**, 507–518.
13. Williams,P.D., Pollock,D.D., Blackburne,B.P. *et al.* (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol.*, **2**, e69.
14. Randall,R.N., Radford,C.E., Roof,K.A. *et al.* (2016) An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.*, **7**, 12847.

15. Yokoyama,S., Tada,T., Zhang,H. *et al.* (2008) Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 13480–13485.
16. Yokoyama,S. (2000) Molecular evolution of vertebrate visual pigments. *Prog. Retin. Eye Res.*, **19**, 385–419.
17. Smith,M.D., Wertheim,J.O., Weaver,S. *et al.* (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.*, **32**, 1342–1353.
18. Eick,G.N., Colucci,J.K., Harms,M.J. *et al.* (2012) Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet.*, **8**, e1003072.
19. Gaucher,E.A., Govindarajan,S. and Ganesh,O.K. (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, **451**, 704–707.
20. Risso,V.A., Gavira,J.A., Mejia-Carmona,D.F. *et al.* (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.*, **135**, 2899–2902.
21. Starr,T.N., Picton,L.K. and Thornton,J.W. (2017) Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, **549**, 409–413.
22. Ortlund,E.A., Bridgham,J.T., Redinbo,M.R. *et al.* (2007) Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, **317**, 1544–1548.
23. Chang,B.S.W. (2003) Ancestral gene reconstruction and synthesis of ancient rhodopsins in the laboratory. *Integr. Comp. Biol.*, **43**, 500–507.
24. Blanchet,G., Alili,D., Protte,A. *et al.* (2017) Ancestral protein resurrection and engineering opportunities of the mamba aminergic toxins. *Sci. Rep.*, **7**, 2701.
25. Risso,V.A., Martinez-Rodriguez,S., Candel,A.M. *et al.* (2017) De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.*, **8**, 16113.
26. Risso,V.A., Manssour-Triedo,F., Delgado-Delgado,A. *et al.* (2015) Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.*, **32**, 440–455.
27. Voordeckers,K., Brown,C.A., Vanneste,K. *et al.* (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.*, **10**, e1001446.
28. Perez-Jimenez,R., Inglés-Prieto,A., Zhao,Z.-M. *et al.* (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.*, **18**, 592–596.
29. Kratzer,J.T., Lanaspa,M.A., Murphy,M.N. *et al.* (2014) Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 3763–3768.
30. Bordoli,L., Kiefer,F., Arnold,K. *et al.* (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.*, **4**, 1–13.
31. Bastolla,U., Roman,H.E. and Vendruscolo,M. (1999) Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.*, **200**, 49–64.
32. Parisi,G. and Echave,J. (2001) Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.*, **18**, 750–756.
33. Berman,H.M., Westbrook,J., Feng,Z. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
34. Wu,C.H., Apweiler,R., Bairoch,A. *et al.* (2006) The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
35. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*, **25**, 25–29.
36. Harms,M.J. and Thornton,J.W. (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, **14**, 559–571.