COMPUTATIONAL ENZYME DESIGN

SÍLVIA OSUNA[*,†]

[*]*Universitat de Girona, Institut de Química Computacional i Catàlisi (IQCC) I Departament de Química, carrer Maria Aurèlia Capmany 69, 17003 Girona, Spain*
[†]*ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain*

My view of the present state of research on computational enzyme design

Enzymes are typically engineered for catalytic activity, enantioselectivity, thermodynamic stability, substrate specificity, stability in non-aqueous solvents and co-solvents. Available enzyme design approaches can be classified into rational design, and Directed Evolution (DE).[1] DE is able to provide highly active tailor-made enzymes at the expense of experimentally generating and screening tens of thousands of variants. However, the high economic cost associated with DE limits the broad application of enzyme-catalyzed processes for chemical manufacture. Most importantly, it is also unknown how the introduced mutations contribute to enzyme proficiency. Different rational design approaches exist that range from multiple sequence alignments (MSA), structural evaluation of the active site pocket and available tunnels, to the application of sophisticated computational tools such as Quantum Mechanics (QM), hybrid QM and Molecular Mechanics (QM/MM), Empirical Valence Bond (EVB), Molecular Dynamics (MD), and Monte Carlo simulations.[2] One of the most popular approaches is the *inside-out* strategy based on modelling the transition state(s) (TS) of the desired transformation (defined as *theozyme*) with Quantum Mechanics (QM) and grafting this ideal arrangement into an existing protein scaffold with Rosetta.[3] These rational approaches hold the promise of providing a comprehensive understanding of the relationship between mutations and its impact into enzymatic activity, yet none of the existing computational approaches is able to generate highly proficient enzymes rivalling natural ones and those generated with DE. In my view, the low activity of rationally designed enzyme variants can be attributed to the following limitations: (1) the high complexity of enzymatic catalysis and the lack of a computational approach able to accurately consider the multiple chemical steps and associated conformational changes taking place along the catalytic itinerary,[2] (2) the need for reducing the sequence space, which is often solved by introducing mutations only in the active site pocket or entry/exit channel (as opposed to DE that introduces mutations throughout the structure),[2] (3) the lack of fast yet accurate computational screening protocols for estimating the catalytic activity.

My recent research contributions to computational enzyme design

Understanding enzymatic function requires the evaluation of the chemical steps along the mechanism, but also the exploration of the ensemble of thermally accessible conformations that enzymes adopt in solution. The ensemble of both reactive and unreactive conformations presenting different relative stabilities can be represented in the so-called free energy landscape (FEL, see Fig. 1A). We computationally reconstructed the FEL of some natural and laboratory evolution (DE) pathways using extensive MD simulations, Markov state modelling (MSM), and enhanced sampling techniques.[4-6] These studies demonstrated that increased enzymatic activity is often achieved by introducing mutations that alter the enzyme conformational ensemble. The introduced mutations located at the active site and often at distal positions induce a long-range effect that impacts the enzyme active site pocket and thus catalysis. This is achieved by favouring the catalytically productive conformational states and disfavouring the non-productive ones for the novel functionality, thus converting computational enzyme design into a population shift problem.[2] However, computational enzyme design seen as a population shift problem requires the reconstruction of a FEL for each generated variant, which is computationally too expensive for allowing the fast routine design of enzymes.[7] Most importantly the reconstructed FELs do not provide any clue on which positions either located at the active site or distal might be responsible for stabilizing a desired conformational change. We hypothesize that by using graph theory coupled to the extensive MD simulations for FEL reconstruction the existing long-range allosteric network of interactions can be revealed and used for predicting distal and active site mutations (Fig. 1C).[2, 8] To that end, we developed the Shortest Path Map (SPM) tool that relies on the construction of a graph based on the computed mean distances and correlation values obtained along MD simulations.[2, 6] SPM decreases the sequence space to a smaller number of conformationally relevant positions, and has the potential of identifying the challenging distal activity-enhancing positions. Indeed, we successfully applied SPM to identify DE mutations in retro-aldolase, monoamine oxidase, and tryptophan synthase enzymes.[2]

In a recent publication, we combined SPM and ancestral sequence reconstruction to rationally design new stand-alone tryptophan synthase B (TrpB) variants (see Fig. 1A-C).[8] Tryptophan synthase (TrpS) is a heterodimeric enzyme complex composed of two subunits: TrpA and TrpB that are allosterically connected. The tight allosteric communication between subunits involves, in the case of TrpB, open-to-closed transitions of the rigid COMM domain that forms a lid covering the active site (see Fig. 1B). The existing allosteric communication between subunits makes both TrpA and TrpB much less efficient when isolated, *i.e.* their stand-alone activity is low. [5, 8] However, the ancestral reconstruction of TrpB enzymes (LBCA TrpB) revealed a high stand-alone activity for the ancestral variants, which was lost along evolution.[9] The Arnold lab applied DE on *pf*TrpB and generated a new enzyme 0B2-*pf*TrpB that presented higher catalytic activity when isolated.[10]

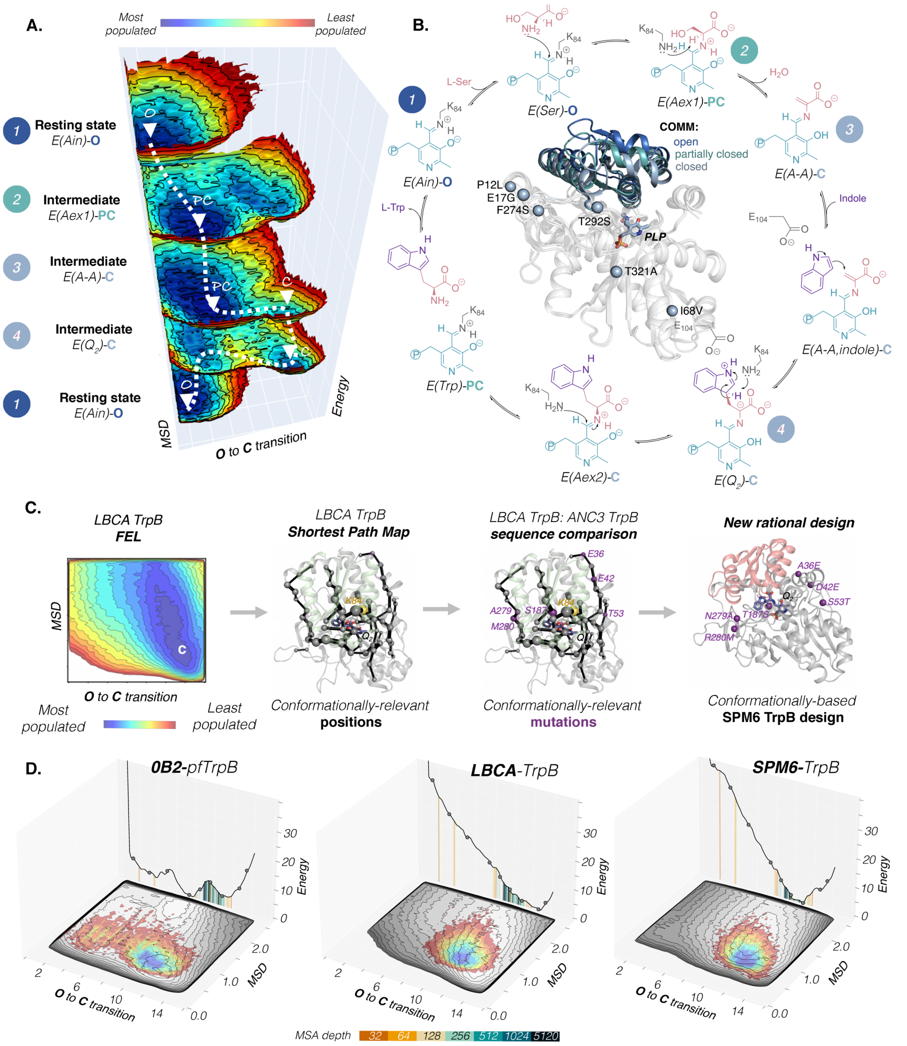

Fig. 1. **A.** Computationally reconstructed Free Energy Landscape (FEL) of the laboratory evolved 0B2-*pf*TrpB tryptophan synthase B that displays stand-alone activity (data from reference[5]) at several reaction intermediates along the catalytic itinerary (shown in panel B). TrpB adopts a different conformation of the catalytically-relevant COMM domain along the process: open (**O**, dark blue) states are adopted in the resting state E(Ain), partially closed (**PC**, teal) at the reaction intermediates E(Aex1) and E(A-A), and closed (**C**, light blue) at E(Q$_2$) states. Most stable conformations are represented in blue, whereas least stable ones in red. **B.** Reaction mechanism of TrpB and detail of the COMM domain conformation along the cycle.[11] Overlay of the COMM domain conformation as shown by X-ray data: **O** highlighted in dark blue, **PC** in teal, and **C** in light blue. **C.** The mutations introduced with DE to generate 0B2-*pf*TrpB are marked with blue spheres. Computational pipeline developed for rationally designing new stand-alone enzyme variants based on the combination of

shortest path map (SPM) and ancestral sequence reconstruction. **D.** Development of an X-ray template-based AF2 approach for estimating the conformational heterogeneity of TrpB systems.[12] AF2 predictions are represented on the 2D-FEL representation using vertical lines colored from orange to dark blue depending on the number of sequences provided in the MSA. From these AF2 structures short nanosecond timescale MD simulations were run for FEL reconstruction (shown on top of the computationally expensive FELs (in gray) obtained by means of extensive metadynamics simulations).

We computationally reconstructed the FEL of the ancestrally reconstructed LBCA TrpB, as well as the wild-type *pf*TrpS complex, isolated *pf*TrpB, and laboratory-evolved stand-alone 0B2-*pf*TrpB enzyme.[5, 8] These works elucidated the conformational ensemble that a stand-alone catalyst has to display for being efficient. We developed a rational computational protocol for achieving stand-alone activity of TrpB subunit based on the following steps (summarized in Fig. 1C): (1) reconstruction of the FEL of the ancestral LBCA TrpB displaying stand-alone activity, (2) application of the SPM methodology to detect the conformationally-relevant positions, (3) sequence comparison at the conformationally-relevant SPM positions between the reference ancestral scaffold and the target ANC3 TrpB variant that had no stand-alone activity, (4) transfer of the 6 non-conserved SPM mutations to the target ANC3 TrpB scaffold for generating the new SPM6-TrpB variant.[8] Interestingly the experimental validation of the SPM6 TrpB design indicated a 7-fold increase (in terms of k_{cat}) of stand-alone activity. Although we did not reach the isolated activity of the reference LBCA TrpB, it is worth highlighting that by testing only one single variant the fold increase in k_{cat} was similar to the 9-fold obtained by DE that required the generation and screening of more than 3000 variants.[10] This study therefore provides evidence for the potential of our SPM methodology for computational enzyme design.

The recent success of the Alphafold2 neural network (AF2) in predicting the folded structure from the primary sequence with high levels of precision has revolutionized the field of protein design.[13] Despite AF2's impressive performance, application of AF2 for understanding and engineering function directly from the obtained single *static* picture is not straightforward. However, in this direction we recently tested the applicability of AF2 for elucidating the conformational heterogeneity of several TrpB enzymes.[12] We developed a template-based AF2 approach for estimating TrpB ability to adopt multiple conformations of the catalytically relevant COMM domain, which is required for enhanced stand-alone activity. Our results revealed the potential of AF2, especially if combined with short nanosecond timescale MD simulations, for estimating the changes induced by mutation in the FEL at a rather reduced computational cost.

Outlook to future developments of research on computational enzyme design

Inspired by the AF2 approach, some deep learning techniques have also recently been developed for protein design that can potentially mitigate some of the limitations mentioned above. The combination of the convolutional neural network trRosetta[14] and Rosetta was shown to be successful for the design of new stable proteins.[15] The AlphaDesign based on AF2 was also developed to predict novel proteins.[16] These examples show the potential of deep learning techniques to generate new functional variants within the allowed biological constraints. The application of AF2 (or other deep learning strategies) to computational enzyme design for any target reaction and substrate still remains largely underdeveloped. In the near future I anticipate that many hybrid biophysical and deep learning strategies will be developed to solve the mentioned limitations and allow the fast routine rational design of efficient enzymes.

Acknowledgments

S.O. thanks the funding from the European Research Council (ERC-2015-StG-679001), Human Frontier Science Program for project grant RGP0054/2020, Spanish MINECO for project PGC2018-102192-B-I00, and Generalitat de Catalunya (2017 SGR-1707).

References and citations

[1] E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, et al., *Nat. Rev. Dis. Primers* ***1***, 46 (2021).
[2] S. Osuna, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* ***11***, e1502 (2020).
[3] G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker and K. N. Houk, *Angew. Chem. Int. Ed.* ***52***, 5700 (2013) and references cited therein.
[4] C. Curado-Carballada, F. Feixas, J. Iglesias-Fernández and S. Osuna, *Angew. Chem. Int. Ed.* ***58***, 3097 (2019).
[5] M. A. Maria-Solano, J. Iglesias-Fernández and S. Osuna, *J. Am. Chem. Soc.* ***141***, 13049 (2019).
[6] A. Romero-Rivera, M. Garcia-Borràs and S. Osuna, *ACS Catal.* ***7***, 8524 (2017).
[7] M. A. Maria-Solano, E. Serrano-Hervás, A. Romero-Rivera, J. Iglesias-Fernández and S. Osuna, *Chem. Commun.* ***54***, 6622 (2018).
[8] M. A. Maria-Solano, T. Kinateder, J. Iglesias-Fernández, R. Sterner and S. Osuna, *ACS Catal.* ***11***, 13733 (2021).
[9] M. Schupfner, K. Straub, F. Busch, R. Merkl and R. Sterner, *Proc. Nat. Acad. Sci. USA* ***117***, 346 (2020).
[10] A. R. Buller, S. Brinkmann-Chen, D. K. Romney, M. Herger, J. Murciano-Calles and F. H. Arnold, *Proc. Natl. Acad. Sci. USA* ***112***, 14599 (2015).
[11] M. F. Dunn, *Arch. Biochem. Biophys.* ***519***, 154 (2012).
[12] G. Casadevall, C. Duran, M. Estévez-Gay and S. Osuna, *Prot. Sci.*, accepted for publication (2022*).*
[13] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., *Nature* ***596***, 583 (2021).
[14] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, et al., *Nature* ***600***, 547 (2021).
[15] C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, et al., *Proc. Natl. Acad. Sci. USA* ***118***, e2017228118 (2021).

[16] M. Jendrusch, J. O. Korbel and S. K. Sadiq, *Biorxiv*, 2021.2010.2011.463937 (2021).