New Expi293 suite of products for structural biology, inducible expression, and protein labeling

Learn more

gibco
by Thermo Fisher Scientific

Accelerated Communication

Protein Science
DOI10.1002/pro.3757

A computational method for design of connected catalytic networks in proteins

Brian D. Weitzner[1,2,†,‡], Yakov Kipnis[1,2,‡], A. Gerard Daniel[1,2,‡], Donald Hilvert[3], David Baker[1,2,4,*]

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
[2]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA.
[3]Laboratory of Organic Chemistry, ETH Zurich, 8093 Zurich, Switzerland.
[4]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

***Corresponding Author:** dabaker@uw.edu

†Current address: Lyell Immunopharma, Seattle, WA 98109, USA.

‡These authors contributed equally to this work.

Funding Sources
WRF (BDF, AGD), HHMI (YK), the Swiss National Science Foundation (DH), HHMI (DB).

ORCID

Received: Jul 27, 2019;Revised: Oct 21, 2019;Accepted: Oct 21, 2019

Brian D. Weitzner: 0000-0002-1909-0961
Yakov Kipnis: 0000-0002-3057-4916
A. Gerard Daniel: NA
Donald Hilvert: 0000-0002-3941-621X
David Baker: 0000-0001-7896-6217

Abstract

Computational design of new active sites has generally proceeded by geometrically defining interactions between the reaction transition state(s) and surrounding side-chain functional groups which maximize transition-state stabilization, and then searching for sites in protein scaffolds where the specified side-chain–transition-state interactions can be realized. A limitation of this approach is that the interactions between the side chains themselves are not constrained. An extensive connected hydrogen bond network involving the catalytic residues was observed in a designed retroaldolase following directed evolution. Such connected networks could increase catalytic activity by preorganizing active site residues in catalytically competent orientations, and enabling concerted interactions between side chains during catalysis, for example proton shuffling. We developed a method for designing active sites in which the catalytic side chains, in addition to making interactions with the transition state, are also involved in extensive hydrogen bond networks. Because of the added constraint of hydrogen-bond connectivity between the catalytic side chains, to find solutions, a wider range of interactions between these side chains and the transition state must be considered. Our new method starts from a ChemDraw-like 2D representation of the transition state with hydrogen-bond donors, acceptors, and covalent interaction sites indicated, and all placements of side-chain functional groups that make the indicated interactions with the transition state, and are fully connected in a single hydrogen-bond network are systematically enumerated. The RosettaMatch method can then be used to identify realizations of these fully-connected active sites in protein scaffolds. The

method generates many fully-connected active site solutions for a set of model reactions that are promising starting points for the design of fully-preorganized enzyme catalysts.

The goal of de novo enzyme design is to create protein catalysts for any chemical reaction of interest.[1-5] Several approaches have been developed to generate new enzyme active sites by searching for placements of catalytically competent side chain constellations in selected protein scaffolds or curated subsets of Protein Data Bank containing up to several thousand protein structures.[6-11] Rosetta computational enzyme design calculations have proceeded by first generating an ideal active site, or theozyme, consisting of the reaction transition state surrounded by side-chain functional groups positioned so as to maximize transition-state stabilization. RosettaMatch is then used to search for geometrically compatible placements of these ideal active sites in protein scaffolds.[12] While directed evolution has succeeded in maturing computational designs to have activities comparable to native enzymes,[13-19] the activities of the original computational designs have generally been quite low. Achieving high catalytic activity directly from computation is an outstanding current challenge.

A route to increasing the activity of computational enzyme designs is suggested by the crystal structure of the optimized aldolase RA95.5-8F which, with a k_{cat}/K_m of ~ 33,800 ± 4,200 M^{-1} s^{-1}, is 200,000-fold more active than the original RA95.0 design [Fig. 1(A), 1(B)].[15] In this structure [Fig. 1(C)], the catalytic residues form an extensive hydrogen bond network (a catalytic quartet). In contrast, in the starting computational design, as in most such designs, there are only a small number of interactions between the designed catalytic residues. High catalytic residue connectivity has the advantage of allowing concerted transitions, such as proton shuffling, during catalysis and in preorganizing the active site residues in catalytically competent conformations.

Indeed, highly connected catalytic side chain networks are frequently observed in native enzymes. A limitation of the RosettaMatch method, which focuses on the geometry of side-chain interactions with the transition state, is that the extent of interaction of catalytic residues with each other cannot be directly specified.

We set out to develop a computational method to directly generate such connected catalytic networks. RosettaMatch starts from a geometric description of the optimal interaction geometries between side-chain functional groups and the transition state (from QM calculations or chemical intuition); to find fully-connected active sites it is necessary to allow a wider range of functional group–transition state geometries. We reasoned that potential losses in activity from less optimal placement of individual functional groups relative to the transition state could be more than compensated by the advantages of catalytic network connectivity. The new method, HBNetGen, starts from a ChemDraw-like specification of the hydrogen bonding and partial covalent interactions between side-chain functional groups and the reaction transition state (rather than the more detailed geometric information in the original RosettaMatch), and between the catalytic side chains themselves (much like textbook schematic depictions of enzyme active sites). The transition state is placed on a grid, and side-chain functional groups are placed around the transition state according to the active site specification. The grid enables complete enumeration of compatible functional group placements at a resolution set by the (user specified) grid spacing. At a grid spacing of 0.2 Å, there are typically ~10^3–10^4 functional group placements for each individual interaction in the active site specification. Next, for each side-

chain–side-chain hydrogen bond in the connected active site specification, each pair of functional group placements for the two side chains is queried for the presence of the hydrogen bond, and placements which do not make the required interaction with any of the enumerated placements for the second side chain are discarded. Catalytic networks are then completed by joining the pairs of hydrogen-bonding residues, and fully-connected networks which make all the specified interactions between catalytic side chains and transition state and between the catalytic side chains are output.

To test HBNetGen, we selected seven examples of fully-connected catalytic networks in the PDB and ran HBNetGen on these crystal structures and that of the evolved retroaldolase, starting from the backbone coordinates and a ChemDraw description of the connectivity of the desired active site. For direct comparison with the crystal structures, the inhibitor or reaction product was used in these tests rather than the computed reaction transition state. The results are summarized in Figure 2: for each ligand, the left-most panel shows the two-dimensional description of the network as a ChemDraw image, the center panel shows an example of a complete network at the functional group level with the number of generated networks indicated below, and the right-most panel shows a full-side-chain representation of the network with the number of unique backbone positions for the selected side-chain identities indicated below. The number of fully connected network solutions ranges from 7,267 for the alanine phosphonate network to 1,534,996 for the retroaldolase network.

To identify placements of the connected networks in sets of protein scaffolds, we adapted the previously described RosettaMatch method to input the rigid-body transformations for each residue in each network expressed in an hierarchical XML file: a RosettaMatch calculation using the sidechain conformations from a single HBNetGen network will by construction only identify fully connected active sites. However, even with the efficiency of RosettaMatch, searching for placements of hundreds of thousands of designed connected active sites in protein scaffolds, each with a unique set of side-chain conformations, is not computationally feasible for any but the simplest active sites. To make this problem tractable, we experimented with pooling the side chain conformations from all HBNetGen networks, and then using RosettaMatch to find combinations of side chains making the specified interactions with the transition state. This solution has the advantage of ensuring that each rotamer at each position is only considered once during the calculation, which considerably increases computational efficiency, but has the consequence that side-chain conformations from different networks can be mixed during matching. We speculated that side-chain rotamer conformations in the same HBNetGen network would have a higher likelihood of being placed in the same active site by RosettaMatch because of their geometric complementarity, leading to recovery of many of the input networks. However, fully connected networks were recovered infrequently with this approach (see next paragraph). To modulate the extent of network re-mixing, we experimented with clustering the set of networks on the coordinates of the central functional groups; this allows smooth interpolation between the case where each network is treated independently (cluster size 1), and

that where all networks are combined (one large cluster). A larger number of smaller clusters will produce fewer models with incomplete networks but will increase the runtime because a separate calculation must be performed for each cluster; we found that using ~250 clusters balanced performance and the run-time increase.

To evaluate the performance of HBNetGen-guided match generation, we matched the set of networks inspired by the evolved retro-aldolase (RA95.5-8F) active site, with the additional criterion that each side chain must interact with the ligand directly (Fig. S3), into a set of ~6000 ligand-binding scaffolds.[20] We modeled the transition state for carbinolamine formation with a partial covalent bond to the nucleophilic lysine, and generated positions of the remaining three residues using HBNetGen, which produced 4,858 networks. Clustering produced 222 clusters comprising 4,526 networks. We compared three approaches: (1) the original RosettaMatch residue-based method that considers each residue completely independently using functional group geometries based solely on interaction geometry with the substrate; (2) RosettaMatch using all 4,858 sidechain conformations from all networks identified by HBNetGen in one simulation; and (3) independent RosettaMatch calculations using sidechain conformations from each of 239 clusters of HBNetGen networks (thus retaining the primary side-chain dependencies determined by HBNetGen). For each resulting active site placement, we determined whether the desired connectivity of the catalytic side chains was achieved. The third approach (using the 239 networks resulting from clustering) generated 98 fully-connected active site solutions from 3031 matches, the second approach (using all side chain arrangements from all 4,858 networks)

yielded 3991 matches, but fewer (47) connected networks, and control RosettaMatch calculations identified only one fully connected network (Table I; in all cases, most of the matches involve only partial networks, as expected because RosettaMatch considers each residue individually in order to make the search computationally tractable). The smaller number of fully connected networks found in the 2nd approach compared to the 3rd approach is due to pruning within RosettaMatch to reduce the combinatorial complexity of the sampling problem (see Supporting Information). Figure 1(D-F) shows fully-connected matches of the RA95.5-8F-inspired theozyme in three different scaffolds. The networked residues are shown in magenta sticks, the ligand is shown in cyan sticks, and hydrogen bonds are indicated with black, dashed lines. While all the matches have the connectivity of the input 2D active site depiction, the three-dimensional realizations are quite different.

HBNetGen explicitly generates fully-connected active sites that are likely to be more preorganized than previous de novo designed catalytic sites. It allows exploration of different catalytic site specifications (at the ChemDraw level), completely independent of a particular protein backbone. This capability enables determination of the extent to which different sites can be realized in three dimensions with full hydrogen bonded connectivity, and investigation, again independent of any protein backbone, of whether the active site configurations found in nature were favored because of the transition state stabilization they provide or because of the connectivity of the catalytic side chains. It is likely that algorithms for finding matches to the HBNetGen connected sites in actual protein structures can be developed that are more efficient

than the simple RosettaMatch implementation described here which breaks up the networks for computational tractability. Experimental characterization of HBNetGen fully connected active sites should provide insight into the contribution of preorganization and side-chain connectivity to catalysis.

References

1. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329:309-313.

2. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF III, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. Science 319:1387-1391.

3. Röthlisberger D, , Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190-195.

4. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci USA 98:14274-14279.

5. Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN, Mayo SL (2012) Iterative approach to computational enzyme design. Proc Natl Acad Sci USA 109:3790-3795.

6. Fazelinia H, Cirino PC, Maranas CD (2009) OptGraft: A computational procedure for transferring a binding site onto an existing protein scaffold. Protein Sci 18:180-195.

7. Zhang C, Lai L (2012) AutoMatch: target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. Proteins 80:1078-1094.

8. Lassila JK, Privett HK, Allen BD, Mayo SL (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proc Natl Acad Sci USA 103:16710-16715.

9. Malisi C, Kohlbacher O, Höcker B (2009) Automated scaffold selection for enzyme design. Proteins 77:74-83.

10. Zhu X, Lai L (2009) A novel method for enzyme design. J Comput Chem 30:256-267.

11. Lei Y, Luo W, Zhu Y (2011) A matching algorithm for catalytic residue site selection in computational enzyme design. Protein Sci 20:1566-1575.

12. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, Baker D (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15:2785-2794.

13. Giger L, Caner S, Obexer R, Kast P, Baker D, Ban N, Hilvert D (2013) Evolution of a designed retro-aldolase leads to complete active site remodeling. Nat Chem Biol 9:494-498.

14. Khersonsky O, Kiss G, Röthlisberger D, Dym O, Albeck S, Houk KN, Baker D, Tawfik DS (2012) Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. Proc Natl Acad Sci USA 109:10358-10363.

15. Obexer R, Godina A, Garrabou X, Mittl PR, Baker D, Griffiths AD, Hilvert D (2017) Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. Nat Chem 9:50-56.

16. Althoff EA, Wang L, Jiang L, Giger L, Lassila JK, Wang Z, Smith M, Hari S, Kast P, Herschlag D, Hilvert D, Baker D (2012) Robust design and optimization of retroaldol enzymes. Protein Sci 21:717-726.

17. Blomberg R, Kries H, Pinkas DM, Mittl PR, Grötter MG, Privett HK, Mayo SL, Hilvert D (2013) Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 503:418-421.

18. Preiswerk N, Beck T, Schulz JD, Milovnik P, Mayer C, Siegel JB, Baker D, Hilvert D (2014) Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. Proc Natl Acad Sci USA 111:8013-8018.

19. Studer S, Hansen DA, Pianowski ZL, Mittl PRE, Debon A, Guffy SL, Der BS, Kuhlman B, Hilvert D (2018) Evolution of a highly active and enantiospecific metalloenzyme from short peptides. Science 362:1285-1288.

20. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Binding MOAD, a high-quality protein-ligand database. Nucleic Acids Res 36:D674-678.

Notes

The authors declare no competing financial interests.

Acknowledgements

This work was supported by the Washington Research Foundation (BDW, AGD), the Howard Hughes Medical Institute (YK, DB), the defense threat reduction agency (DB), and the Swiss National Science Foundation (DH).

Tables

Table 1: Placement of HBNetGen-designed RA95.5-8F inspired catalytic networks into protein scaffolds.

Simulation ID	Number of incomplete networks	Number of complete networks	Completion rate
Residue-based[a]	7455	1	0.013%
Network-based[b] (all)	3991	47	1.18%
Network-based[b] (clustered)	3031	98	3.23%

[a] Side-chain geometries from rigid-body transformation to ligand for each residue separately (default RosettaMatch settings)

[b] Side-chain geometries from networks produced by HBNetGen, tested with and without clustering. Both types of simulations are subjected to *post facto* optimization to account for the effects of discrete rotamer-sampling and ligand-position binning used by the Matcher. Use of clustered, network-derived constraints enriches the yield of complete networks roughly 240-fold.

FIGURE LEGENDS

Figure 1: Comparison of the RA95.0 and RA95.5-8F active sites suggests the importance of catalytic residue connectivity. (A) A 2D representation of the RA95.0 theozyme, which consists of Lys210, Glu53, plus a water molecule. (B) Directed evolution of RA95.0 yielded the 200,000-fold more active RA95.5-8F, whose catalytic groups include a repositioned lysine, Lys83, along with Tyr180, Tyr51 and Asn110. The three-dimensional representation of the crystallographic coordinates of RA95.5-8F are shown in (C) in the context of the complete scaffold, with the networked residues shown in magenta sticks, the ligand in cyan sticks, and hydrogen bonds indicated with black, dashed lines. Analysis of the crystal structure reveals an extensive network of hydrogen bonds between the catalytic side chains, preorganizing the active site. We hypothesize active site preorganization is the basis for the observed increase in activity. (D)-(F) Three different 3D realizations of the RA95.5-8F-inspired active site generated by HBNetGen. (d) A nonspecific lipid transfer protein (scaffold PDB accession code 1bwo); (E) thymidine kinase (1w4r); and (F) FKBP-like domain (1c9h).

Figure 2: Generation of HBNetGen three dimensional active sites from 2D ChemDraw wiring diagrams. Eight candidate networks surrounding ligands were selected from the PDB: a) (4R,5R)-3-amino-4,5-di- hydroxy-cyclohexene-1-carboxylate; b) 1,3-dihydroxyacetone-phosphate; c) arginine; d) {1-[(3-hydroxy-methyl-5-phosphonooxy-methyl-pyridin-4-ylmethyl)-amino]-ethyl}-phosphonic acid; e) 3-O-α-d-mannopyranosyl-α-d-mannopyranose; f) maltose; g) the open form of penicillin G; and h) the retro-aldol intermediate found in RA95.5-8F. Each row shows: (left) a 2D representation of the interactions that form a complete network; (middle) the lowest-energy network configuration at the functional-group level, with the total number of network configurations indicated below; and (right) a full side-chain representation of a network configuration, with total number of full side-chain realizations for the network configuration (dependent on the number of rotatable bonds for the constituent sidechains) indicated below. The total number of full side-chain 3D realizations of the 2D connected reaction schematic on the left is the production of the numbers in the second and third columns: (number of 3D placements of functional groups) x (number of sidechain placements per functional group placement).