Published online 12 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W479–W487***
doi: 10.1093/nar/gkw416

HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering

Jaroslav Bendl[1,2,3,†], Jan Stourac[1,†], Eva Sebestova[1], Ondrej Vavra[1], Milos Musil[1,2], Jan Brezovsky[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Masaryk University, 625 00 Brno, Czech Republic, [2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic and [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received February 21, 2016; Revised April 22, 2016; Accepted May 03, 2016

ABSTRACT

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and design of smart libraries for engineering proteins' stability, catalytic activity, substrate specificity and enantioselectivity. The server integrates sequence, structural and evolutionary information obtained from 3 databases and 20 computational tools. Users are guided through the processes of selecting hot spots using four different protein engineering strategies and optimizing the resulting library's size by narrowing down a set of substitutions at individual randomized positions. The only required input is a query protein structure. The results of the calculations are mapped onto the protein's structure and visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable for mutagenesis and can automatically design appropriate codons for each implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations of protein structures and assists protein engineers with the rational design of site-specific mutations and focused libraries. It is freely available at http://loschmidt.chemi.muni.cz/hotspotwizard.

INTRODUCTION

The development of tailor-made enzymes for industrial applications is facilitated by understanding the molecular mechanisms of protein function. However, despite significant advances in recent decades, it is not yet clear how a protein's sequence encodes its function (1,2). Traditional directed evolution circumvents this problem by using repeated rounds of random mutagenesis and screening of large sequence libraries to explore the mutational landscape and find proteins with desired properties (2–5). This approach has the advantage of requiring no prior knowledge of the protein's structure or understanding of its structure–function relationships (6), but necessitates the laborious and costly screening of very large libraries (4). The efficiency of directed evolution experiments can be significantly improved by creating smaller, higher quality libraries that are more likely to yield positive results. Such 'smart' libraries can be generated by focusing mutagenesis on a limited number of 'hot spot' positions that are likely to affect the property of interest, or by selecting a limited set of substitutions (1–5).

The optimal strategy for identifying hot spots depends on the property being targeted. Catalytic properties such as activity, specificity and stereoselectivity are often related to amino acid residues that mediate substrate binding, transition-state stabilization or product release (7,8). Such residues can be identified using tools for predicting and analyzing enzyme-ligand interactions (9–11) or detecting binding pockets or access tunnels (12–14). Strategies for improving protein stability include rigidification of flexible sites, cavity-filling, tunnel engineering, consensus and ancestral mutation methods, or redesigning of surface charges (15–17). While hot spots for some of these strategies can be identified straightforwardly using a single computational tool (18), others require multi-step analyses or the use of molecular modelling methods (19). Having obtained a set of promising sites for manipulating the desired property, the next challenge is to draw up a list of allowed substitutions at individual positions. This can be done by considering the amino acid distribution at the corresponding positions in sequence homologs (20,21), by using reduced sets of amino acids with either specific desired physicochemical properties or a balanced set of these properties (22,23), or on the basis of the predicted effects of specific substitutions on the protein's properties (24,25). Finally, an appropriate degen-

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz
Correspondence may also be addressed to Jan Brezovsky. Email: brezovsky@mail.muni.cz
[†] These authors contributed equally to the work as first authors.

erate codon covering the specified set of amino acids must be selected for each targeted position. Ideally, these codons should exhibit minimal amino acid bias and minimize the frequency of premature stop codons (26). Several tools are available to facilitate this task and to calculate the size of the designed library (27).

Here, we present HotSpot Wizard 2.0, a web server for the automated identification of hot spots and design of smart libraries for engineering protein stability, enzymatic activity, substrate specificity and enantioselectivity. Compared to its predecessor (28), HotSpot Wizard 2.0 introduces several major improvements, extending the scope and quality of its analyses. It implements four different established protein engineering strategies, enabling the user to selectively target sites affecting the protein's stability and catalytic properties. Users can easily select suitable substitutions for individual hot spots based on predictions of tolerated amino acids or amino acid distributions in sequence homologs, and suitable degenerate codons for these substitutions can be designed automatically via the HotSpot Wizard interface. A new graphical user interface provides an intuitive and comprehensive overview of the results of the analysis, allowing users to think directly about the obtained designs. The resulting pipeline of twenty integrated tools and three databases represents a unique one-stop solution that makes library design accessible even to users with no prior knowledge of bioinformatics.

MATERIALS AND METHODS

The workflow of HotSpot Wizard is outlined in Figure 1. In order to explore the mutational landscape and find the most promising mutagenesis targets, a protein selected by the user is annotated using several prediction tools and databases (Phase 1). With this knowledge in hand, four protein engineering strategies are used to identify suitable hot spots for improving desired protein properties (Phase 2). Finally, suitable substitutions and appropriate degenerate codons are proposed for each selected hot spot, enabling the design of a smart library (Phase 3).

Phase 1: annotation of the protein

The first step in the workflow requires the user to specify the protein structure of interest, either by providing its PDB ID or by uploading a suitable PDB file. If possible, the biological assembly of the target protein is automatically generated by the MakeMultimer tool (http://watcut.uwaterloo.ca/tools/makemultimer), and information about 'essential residues' directly involved in catalysis or binding is obtained from the Catalytic Site Atlas (29) and UniProtKB/Swiss-Prot (30) databases. The DSSP algorithm (31) is then used to assign the protein's secondary structure, and its accessible surface area is computed using the Shrake and Rupley algorithm (32) with BioJava (33). The average B-factors are computed for the protein's amino acid residues (34). The raw B-factor values are accompanied by residue rankings ranging from 1–100%; rankings of 1–25%, 26–75% and 76–100% indicate high, moderate and low levels of relative structural flexibility, respectively. Protein pockets are then identified with Fpocket (35). For each chain, the pocket containing the greatest number of essential residues is identified as the catalytic pocket. If there are two or more pockets that satisfy this criterion, a decision is made according to the Fpocket score. Having identified the putative catalytic pockets, their centers of mass are determined and used as starting points to identify access tunnels with CAVER (36). Sequence homologs of the target protein are then obtained by performing a BLAST (37) search against the UniRef90 (38) database, using the target protein sequence as a query. All identified homologs are aligned with the query protein using USEARCH (39). By default, sequences whose identity with the query is below 30% or above 90% are excluded from the list of homologs. The remaining sequences are then clustered using UCLUST (39), with a 90% identity threshold to remove close homologs. The cluster representatives are sorted based on the BLAST query coverage and by default, the first 200 of them are used to create a sequence data set. A multiple sequence alignment of the resulting sequence data set is created with Clustal Omega (40) and used to (i) estimate the conservation of each position in the protein based on the Jensen–Shannon entropy (41); (ii) identify correlated positions using an ensemble of the MI (42), aMIc (43), OMES (44), SCA (45), DCA (46), McBASC (47) and ELSC (48) methods; (iii) predict the tolerated amino acids at each position in the protein sequence using RAPHYD (see Supplementary Data 1); and (iv) analyze amino acid frequencies at individual positions within the protein. The conservation scores are used to assign mutability values to individual residues. To facilitate interpretation, these values are divided into three groups: values of 1–3, 4–5 and 6–9 indicate low, moderate and high mutability, respectively.

Phase 2: identification of mutagenesis hot spots

Based on the comprehensive annotation of the target protein, four protein engineering strategies are used to identify different types of hot spots: (i) functional hot spots, (ii) stability hot spots based on structural flexibility, (iii) stability hot spots based on sequence consensus and (iv) correlated hot spots. Some examples illustrating the use of these strategies to engineer selected properties in 12 different proteins (34,49–62) are shown in Figure 2. Functional hot spots correspond to highly mutable residues located in the catalytic pockets or tunnels connecting these pockets with the bulk solvent. Residues located in close proximity to the active site have been identified as good mutagenesis targets for engineering activity, enantioselectivity and substrate specificity (52,63,64). To prevent mutagenesis at positions that are indispensable for protein function, all essential residues are designed immutable and thus excluded from the list of potential hot spots. Supplementary Data 2 shows that HotSpot Wizard provides a significantly greater proportion of viable mutants than random mutagenesis. Stability hot spots are identified by analyzing structural flexibility and sequence consensus. The former approach aims to rigidify flexible protein regions by mutating residues with high average B-factors (34). B-factor provides a metric for flexibility which is due in part to inherent flexibility of the macromolecule, but also includes stabilizing/destabilizing energy from packing in the crystal lattice. The rationale for targeting these flexible residues is that they have relatively

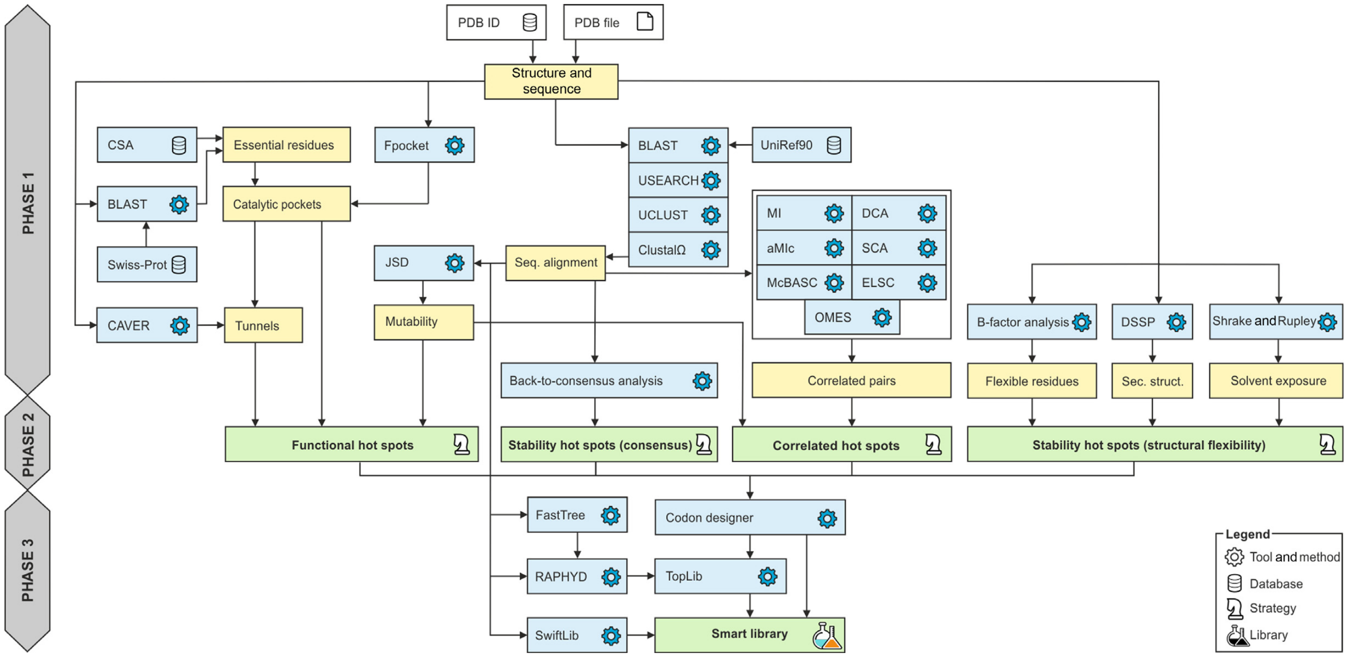


Figure 1. Workflow of HotSpot Wizard.

few contacts with neighbors, so their substitution can produce more interactions (34,54,55). In contrast, the sequence consensus protocol implements majority and frequency ratio approaches, both of which suggest mutations at positions where the wild-type amino acid differs from the most prevalent amino acid (i.e. the consensus residue) at a given position in the multiple sequence alignment. The assumption that the most common amino acid is likely to be stabilizing has proven to be very successful at creating more stable proteins (56–58,65). By default, if the consensus residue is present in at least 50% of all analyzed sequences, the corresponding position is identified as a hot spot in the majority approach. The frequency ratio approach has a less strict criterion for the consensus residue's frequency – the default value is 40%, but it must also be at least five times more frequent than the wild-type residue as a hot spot. The final strategy involves searching for coordinated changes of the amino acids at two separate positions within the protein. Such pairs of positions are referred to as correlated hot spots, and arise when one amino acid substitution has an unfavorable effect that is compensated for by a second mutation of a residue that is located in close structural proximity to the first. This second, correlated mutation typically helps to maintain protein function, stability or folding (66). Methods developed for identifying correlated pairs have revealed mutations responsible for modulating substrate specificity (67), enantioselectivity (68) and mutagenesis targets for stability engineering (69). The identification of correlated positions in HotSpot Wizard is based on an ensemble of seven prediction tools. Each tool generates a raw score for each pair of residues in the protein that measures the pair's degree of correlation. The mean and standard deviation of the degrees of correlation for all pairs of residues in the protein are then calculated and the raw scores are converted into Z-scores, which measure the number of standard deviations by which each pair's raw score deviates from the mean. Based on the work of Martin *et al*. (70), a pair is considered to be correlated if its average Z-score ≥ 3.5 and both of its positions have at least a moderate degree of mutability – by definition, highly conserved positions cannot co-evolve (71).

Phase 3: design of the smart library

The efficiency of directed evolution experiments can be improved by focusing mutagenesis on a limited number of hot spots, but also by restricting the number of allowed substitutions at individual positions using appropriate codons (20–25). For each protein engineering strategy, HotSpot Wizard provides a way to prioritize amino acids at the randomized positions (Table 1) and identifies degenerate codons encoding all desired amino acids with the minimum redundancy and the smallest possible ratio of stop codons. Alternatively, the SwiftLib tool (73) can be used to calculate optimal degenerate codons while keeping the library diversity within the specified limits (the default 10 000). Although the resulting library may not necessarily fully cover the desired set of amino acids, the probability of omitting the important amino acids is relatively low as their weights are set according to selected prioritization method (e.g. based on amino acid distributions in sequence homologs). For both approaches, the most common metrics, such as expected coverage or library size, are computed with TopLib (72).

DESCRIPTION OF THE WEB SERVER

Input

The only required input to the web server is a tertiary structure of the query protein, provided either as a PDB ID or a PDB file. The user can then choose a predefined biological

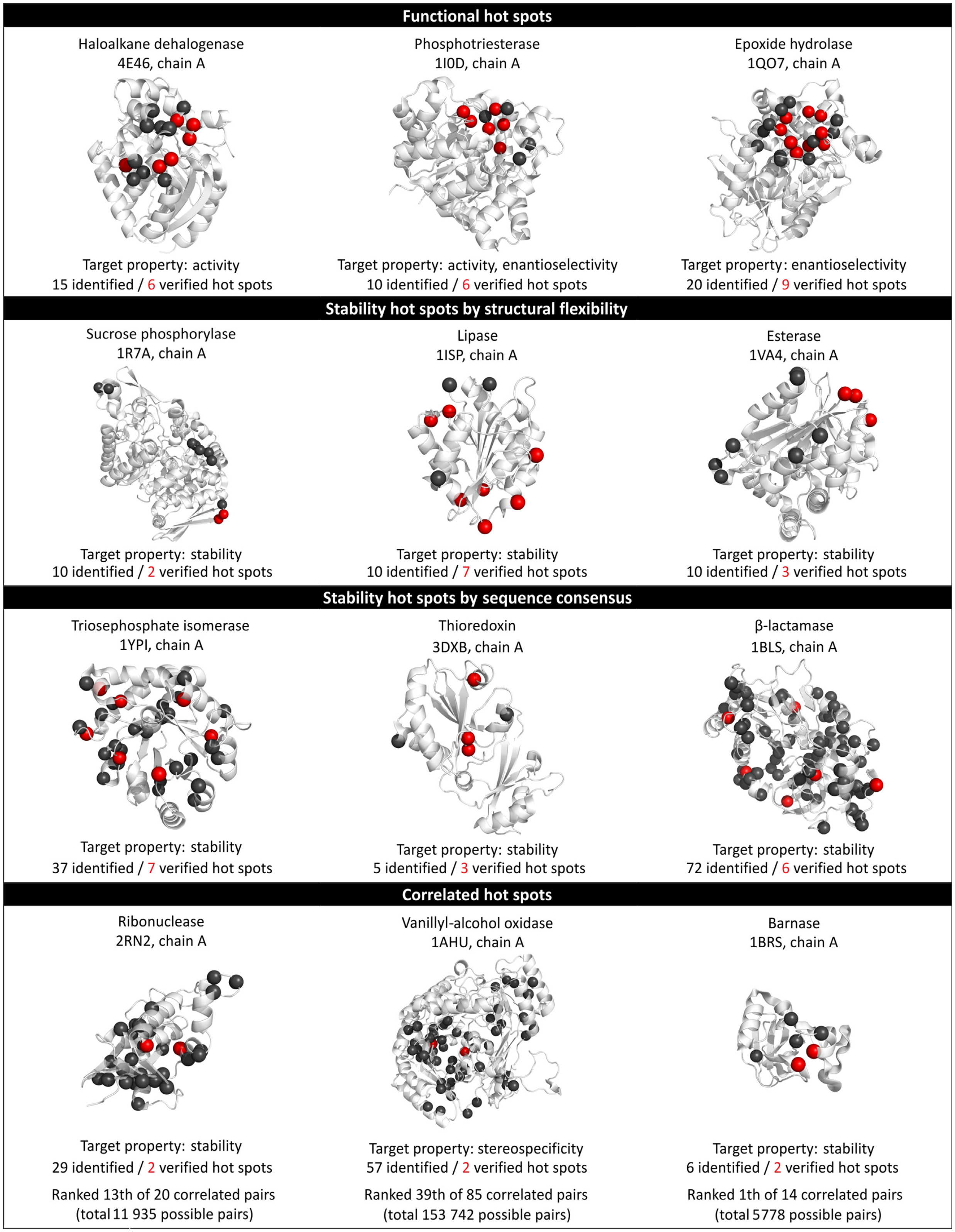

Functional hot spots
Haloalkane dehalogenase
4E46, chain A
Target property: activity
15 identified / 6 verified hot spots
Phosphotriesterase
1I0D, chain A
Target property: activity, enantioselectivity
10 identified / 6 verified hot spots
Epoxide hydrolase
1QO7, chain A
Target property: enantioselectivity
20 identified / 9 verified hot spots
Stability hot spots by structural flexibility
Sucrose phosphorylase
1R7A, chain A
Target property: stability
10 identified / 2 verified hot spots
Lipase
1ISP, chain A
Target property: stability
10 identified / 7 verified hot spots
Esterase
1VA4, chain A
Target property: stability
10 identified / 3 verified hot spots
Stability hot spots by sequence consensus
Triosephosphate isomerase
1YPI, chain A
Target property: stability
37 identified / 7 verified hot spots
Thioredoxin
3DXB, chain A
Target property: stability
5 identified / 3 verified hot spots
β-lactamase
1BLS, chain A
Target property: stability
72 identified / 6 verified hot spots
Correlated hot spots
Ribonuclease
2RN2, chain A
Target property: stability
29 identified / 2 verified hot spots
Ranked 13th of 20 correlated pairs
(total 11 935 possible pairs)
Vanillyl-alcohol oxidase
1AHU, chain A
Target property: stereospecificity
57 identified / 2 verified hot spots
Ranked 39th of 85 correlated pairs
(total 153 742 possible pairs)
Barnase
1BRS, chain A
Target property: stability
6 identified / 2 verified hot spots
Ranked 1th of 14 correlated pairs
(total 5778 possible pairs)

Figure 2. Some notable applications of the four protein engineering strategies implemented in the HotSpot Wizard web server.

Table 1. Methods for selecting substitutions at hot spot positions identified using the four different protein engineering strategies

Selection mode	Availability in strategies	Description
Amino acid frequency	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal frequency in the multiple sequence alignment
Mutational landscape	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal probability of preservation of protein function
Sequence consensus	CONS	suggests amino acid residues fulfilling the criteria of at least one of approaches implemented in sequence consensus strategy: (i) majority approach or (ii) frequency ratio approach
Correlated positions	CORREL	suggests amino acid residues fulfilling the criterion of minimal frequency of co-occurrence with some other specific residue from coupled position
Manual	ALL	manual selection of amino acid residues

FUNC – Analysis of functional hot spots; FLEX – Analysis of stability hot spots/structural flexibility approach; CONS – Analysis of stability hot spots / sequence consensus approach; CORREL – Analysis of correlated hot spots

unit generated by the MakeMultimer tool or manually select chains for which the calculation should be performed. The calculations can be configured in either basic or advanced mode. Basic mode directs the user's attention to the most important parameters, providing an overview of the identified essential residues and highlighting the main parameters involved in the identification of pockets and tunnels. The designation of essential residues is a key step in the functional strategy because these residues are excluded from the list of potential hot spots and are also used to detect catalytic pockets and access tunnels. The user should therefore inspect the automatically generated list of essential residues and correct it if necessary. If no essential residues are detected, the user should specify them manually. In basic mode, the user can specify three parameters: (i) the probe radius, which is used in pocket identification and defines the minimum radius of an alpha sphere in a pocket (default 2.8 Å); (ii) the minimum probe radius, which defines the minimum radius of a putative tunnel (default 1.4 Å); and (iii) the clustering threshold, which determines how the hierarchically clustered tunnels are cut and thus affects the number of tunnels that can be identified (default 3.5 Å). Advanced mode allows expert users to fine-tune parameters of individual calculations in the pipeline to achieve more specialized objectives.

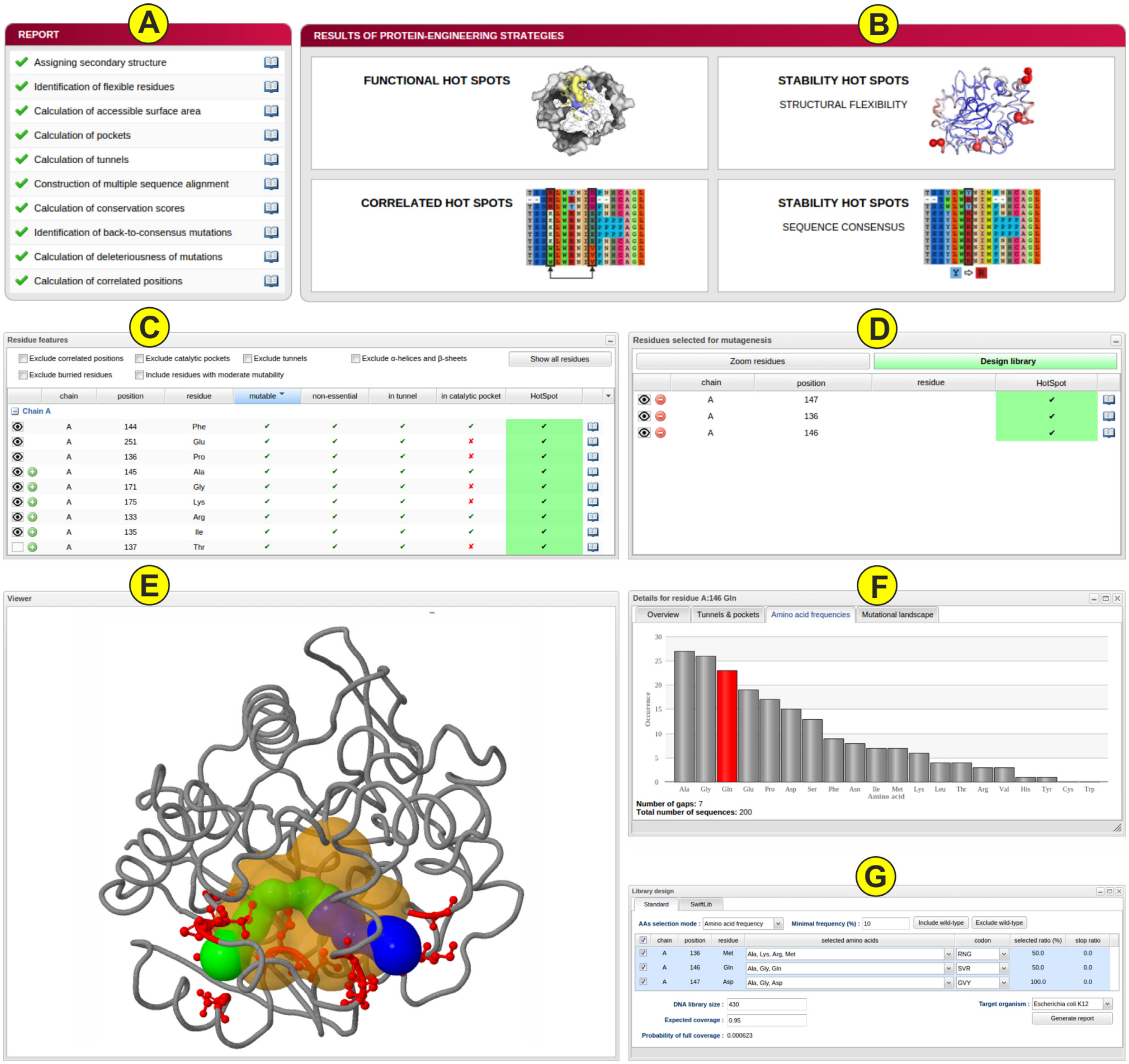
Output

Upon submission, a unique identifier is assigned to each job to track the calculation. The 'Results browser' panel provides information on the status of individual steps in the computational pipeline (Figure 3A). Once the job is finished, the navigation panel provides links to the results obtained using each of the four different protein engineering strategies (Figure 3B). The result pages for each strategy are all organized in the same way, which is described below.

Residue features. The 'Residue features' panel lists all of the identified hot spots together with information relevant to the selected protein engineering strategy (Figure 3C). Several checkboxes can be found at the top of this panel, allowing users to reduce the list of hot spots by applying additional criteria such as excluding buried residues, correlated positions or residues forming a catalytic pocket. The 'Show all residues' button enables users to inspect any residue of the target protein and possibly select hot spots based on their own criteria. Importantly, a pop-up window containing detailed information about a given residue is displayed after clicking the 'book' icon in the last column of the table. Users can visualize individual residues within the protein structure by selecting the 'eye' icon in the first column, and can add residues to the list of mutagenesis hot spots by clicking the 'plus' icon in the second column. All selected mutagenesis hot spots listed in the 'Residues selected for mutagenesis' panel (Figure 3D) can be used for designing a smart library by clicking the 'Design library' button.

Residue details. The information in the 'Residue details' panel is organized into several tabs (Figure 3F): (i) 'Overview', which provides basic information on the residue's characteristics such as its mutability, average B-factor and secondary structure; (ii) 'Annotations', describing the residue's function (only available for essential residues); (iii) 'Tunnels and Pockets', which lists the pockets and/or tunnels of which the residue is a part; (iv) 'Sequence consensus', listing potential consensus mutations for a given position; (v) 'Amino acid frequencies', providing the distribution of amino acids in the corresponding column of the multiple sequence alignment; (vi) 'Mutational landscape', quantifying the probability of preservation of protein function for individual substitutions at a given site; and (vii) 'Correlated positions', listing all positions correlated with the site in question.

Design of smart library. The 'Library design' panel allows the user to select a set of substitutions and design degenerate codons for systematic mutagenesis of the selected positions (Figure 3G). An automatic method for prioritizing amino acids suitable for the chosen protein engineering strategy will be pre-selected. The panel contains two tabs, each corresponding to one library optimization mode. In the 'Standard mode', users can manually define their own set of required substitutions for individual positions if they so desire. After any change in the list of amino acids, HotSpot Wizard automatically identifies the most suitable codons covering all desired amino acids with the lowest possible redundancy, and the library size corresponding to the specified expected coverage. The parameters of the library can be modified interchangeably, allowing the user to adjust the final library based on its size or preferred degree of its coverage. In the 'SwiftLib mode', users specify the maximum acceptable library diversity and the method reports the op-



REPORT

- ✓ Assigning secondary structure
- ✓ Identification of flexible residues
- ✓ Calculation of accessible surface area
- ✓ Calculation of pockets
- ✓ Calculation of tunnels
- ✓ Construction of multiple sequence alignment
- ✓ Calculation of conservation scores
- ✓ Identification of back-to-consensus mutations
- ✓ Calculation of deleteriousness of mutations
- ✓ Calculation of correlated positions

Residue features

□ Exclude correlated positions □ Exclude catalytic pockets □ Exclude tunnels □ Exclude α-helices and β-sheets [Show all residues]

□ Exclude burried residues □ Include residues with moderate mutability

chain	position	residue	mutable	non-essential	in tunnel	in catalytic pocket	HotSpot
Chain A							
A	144	Phe	✓	✓	✓	✓	✓
A	251	Glu	✓	✓	✓	✗	✓
A	136	Pro	✓	✓	✓	✗	✓
A	145	Ala	✓	✓	✓	✓	✓
A	171	Gly	✓	✓	✓	✗	✓
A	175	Lys	✓	✓	✓	✗	✓
A	133	Arg	✓	✓	✓	✓	✓
A	135	Ile	✓	✓	✓	✓	✓
A	137	Thr	✓	✓	✓	✗	✓

Residues selected for mutagenesis

[Zoom residues] [Design library]

chain	position	residue	HotSpot
A	147		✓
A	136		✓
A	146		✓

Number of gaps: 7
Total number of sequences: 200

Library design

Standard | SwiftLib

AAs selection mode : Amino acid frequency **Minimal frequency (%) :** 10 [Include wild-type] [Exclude wild-type]

☑	chain	position	residue	selected amino acids	codon	selected ratio (%)	stop ratio
☑	A	136	Met	Ala, Lys, Arg, Met	RNG	50.0	0.0
☑	A	146	Gln	Ala, Gly, Gln	SVR	50.0	0.0
☑	A	147	Asp	Ala, Gly, Asp	GVY	100.0	0.0

DNA library size : 430
Expected coverage : 0.95
Probability of full coverage : 0.000623
Target organism : Escherichia coli K12 [Generate report]

Figure 3. HotSpot Wizard's graphical user interface, showing results obtained for the haloalkane dehalogenase LinB (PDB ID: 1CV2). (**A**) The 'Report' panel shows the status of the calculations in the individual steps of the computational pipeline. (**B**) Results obtained using the four protein engineering strategies. (**C**) The 'Residue features' panel, which provides an overview of the identified hot spots. (**D**) The 'Residues selected for mutagenesis' panel, which presents a user-adjustable list of residues representing targets for mutagenesis. (**E**) The JSmol viewer allows interactive visualization of the protein and the identified tunnels and pockets. (**F**) The 'Residue details' pop-up window, which provides comprehensive information on the residue's annotations, organized under several tabs. (**G**) The 'Library design' panel, which shows the list of substitutions and appropriate codons for randomization of selected positions.

timal combination of codons with the minimal redundancy of amino acids. However, this efficiency is often achieved at the price of omitting some of desired amino acids with lower weights. The initial amino acid weights derived from the selected prioritization scheme can be changed by selecting the 'Edit amino acid weights'. Additionally, users can request multiple solutions and thus inspect also the solutions which are considered as less optimal by the method, but may better meet the users' needs. Finally, users can generate a nucleotide sequence from the designed amino acid sequence based on the codon usage of selected organism (default is *Escherichia coli*) with the European Molecular Biology Open Software Suite (EMBOSS) Backtranseq tool (74).

Protein visualization. The protein of interest is interactively visualized in the web browser using the JSmol applet (http://wiki.jmol.org/index.php/JSmol). Users can display individual amino acid residues as well as identified tunnels and pockets (Figure 3E). The hot spot residues are colored in red, residues in tunnels and pockets in yellow and all other residues in grey.

Structural features. The main characteristics of all pockets and access tunnels are presented in the 'Pockets' and 'Tunnels' panels, respectively. These panels allow users to visualize individual pockets and tunnels in the structure and to open a pop-up window showing a list of all the residues comprising the chosen structural feature.

CONCLUSIONS AND OUTLOOK

HotSpot Wizard 2.0 is a web server for the automatic identification of hot spots and the design of site-specific mutations and mutant libraries for engineering protein stability, catalytic activity, substrate specificity and enantioselectivity. The server provides a unified interface allowing users to apply four well-established protein engineering strategies that combine structural, functional and evolutionary information to identify suitable positions for mutagenesis. Moreover, HotSpot Wizard integrates several schemes for automatic prioritization of mutations and codon optimization for selected hot spot positions to facilitate the design of smart libraries. The automation of the multi-step procedure makes the process of library design accessible to users without expertise in bioinformatics because it eliminates the need to select, install and evaluate tools, optimize their parameters, perform conversions between different data formats, and interpret intermediate results.

In the future, we plan to implement a protocol for structure prediction based on homology modeling, extending the applicability of HotSpot Wizard to proteins for which no experimental structure is yet available. Additionally, we aim to assess other established protein engineering strategies and, if they prove suitable, to develop new modules so they can be added to the server's portfolio of methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to express many thanks to Dr Antonin Pavelka (Masaryk University, Brno, Czech Republic) for valuable discussions and Dr Yuval Nov (University of Haifa, Haifa, Israel) for kindly providing the source code of TopLib. Uwe Bornscheuer (University Greifswald), Marco Fraaije (Groningen University) and Moshe Goldsmith (Weizmann Institute of Science) are sincerely acknowledged for constructive comments on the tool. MetaCentrum and CERIT-SC are acknowledged for providing access to computing and storage facilities [LM2015085 and LM2015042].

FUNDING

Ministry of Education of the Czech Republic [LO1214, LQ1605, LM2015055 and LM2015047]; Grant Agency of the Czech Republic [GA16-06096S]; European Commission REGPOT [316345]; Horizon 2020 Research Infrastructure ELIXIR-EXCELERATE [676559]; Brno University of Technology [FIT-S-14-2299 to M.M.]. Funding for open access charge: Grants from Czech Ministry of Education [LO1214, LQ1605, LM2015055 and LM2015047].
Conflict of interest statement. None declared.

REFERENCES

1. Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.
2. Currin,A., Swainston,N., Day,P.J. and Kell,D.B. (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.*, **44**, 1172–1239.
3. Cheng,F., Zhu,L. and Schwaneberg,U. (2015) Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun. (Camb.)*, **51**, 9760–9772.
4. Lutz,S. (2010) Beyond directed evolution–semi-rational protein engineering and design. *Curr. Opin. Biotechnol.*, **21**, 734–743.
5. Acevedo-Rocha,C.G., Reetz,M.T. and Nov,Y. (2015) Economical analysis of saturation mutagenesis experiments. *Sci. Rep.*, **5**, 10654.
6. Lo Surdo,P., Walsh,M.A. and Sollazzo,M. (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.*, **11**, 382–383.
7. Denard,C.A., Ren,H. and Zhao,H. (2015) Improving and repurposing biocatalysts via directed evolution. *Curr. Opin. Chem. Biol.*, **25**, 55–64.
8. Bornscheuer,U.T., Huisman,G.W., Kazlauskas,R.J., Lutz,S., Moore,J.C. and Robins,K. (2012) Engineering the third wave of biocatalysis. *Nature*, **485**, 185–194.
9. Xie,Z.-R. and Hwang,M.-J. (2015) Methods for predicting protein-ligand binding sites. *Methods Mol. Biol.*, **1215**, 383–398.
10. Yuan,Y., Pei,J. and Lai,L. (2013) Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.*, **19**, 2326–2333.
11. Lavecchia,A. and Di Giovanni,C. (2013) Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.*, **20**, 2839–2860.
12. Sebestova,E., Bendl,J., Brezovsky,J. and Damborsky,J. (2014) Computational tools for designing smart libraries. *Methods Mol. Biol.*, **1179**, 291–314.
13. Brezovsky,J., Chovancova,E., Gora,A., Pavelka,A., Biedermannova,L. and Damborsky,J. (2013) Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol. Adv.*, **31**, 38–49.
14. Zhang,Z., Li,Y., Lin,B., Schroeder,M. and Huang,B. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.
15. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534–6565.
16. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
17. Yu,H. and Huang,H. (2014) Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol. Adv.*, **32**, 308–315.
18. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.
19. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
20. Reetz,M.T. and Wu,S. (2008) Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun. (Camb)*, **43**, 5499–5501.
21. Jochens,H. and Bornscheuer,U.T. (2010) Natural diversity to guide focused directed evolution. *Chembiochem*, **11**, 1861–1866.

22. Pines,G., Pines,A., Garst,A.D., Zeitoun,R.I., Lynch,S.A. and Gill,R.T. (2015) Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.*, **4**, 604–614.
23. Reetz,M.T., Kahakeaw,D. and Lohmer,R. (2008) Addressing the numbers problem in directed evolution. *Chembiochem*, **9**, 1797–1804.
24. Goldsmith,M. and Tawfik,D.S. (2013) Enzyme engineering by targeted libraries. *Methods Enzymol.*, **523**, 257–283.
25. Chaparro-Riggers,J.F., Polizzi,K.M. and Bommarius,A.S. (2007) Better library design: data-driven protein engineering. *Biotechnol. J.*, **2**, 180–191.
26. Gaytán,P., Contreras-Zambrano,C., Ortiz-Alvarado,M., Morales-Pablos,A. and Yáñez,J. (2009) TrimerDimer: an oligonucleotide-based saturation mutagenesis approach that removes redundant and stop codons. *Nucleic Acids Res.*, **37**, e125.
27. Nov,Y. (2014) Probabilistic methods in directed evolution: library size, mutation rate, and diversity. *Methods Mol. Biol.*, **1179**, 261–278.
28. Pavelka,A., Chovancova,E. and Damborsky,J. (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.*, **37**, W376–W383.
29. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
30. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
31. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
32. Shrake,A. and Rupley,J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79**, 351–371.
33. Prlić,A., Yates,A., Bliven,S.E., Rose,P.W., Jacobsen,J., Troshin,P.V., Chapman,M., Gao,J., Koh,C.H., Foisy,S. *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.
34. Reetz,M.T., Carballeira,J.D. and Vogel,A. (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed Engl.*, **45**, 7745–7751.
35. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
36. Chovancova,E., Pavelka,A., Benes,P., Strnad,O., Brezovsky,J., Kozlikova,B., Gora,A., Sustr,V., Klvana,M., Medek,P. *et al.* (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8**, e1002708.
37. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
38. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
39. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
40. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
41. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
42. Korber,B.T., Farber,R.M., Wolpert,D.H. and Lapedes,A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7176–7180.
43. Lee,B.-C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.
44. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
45. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
46. Weigt,M., White,R.A., Szurmant,H., Hoch,J.A. and Hwa,T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 67–72.
47. Olmea,O., Rost,B. and Valencia,A. (1999) Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, **293**, 1221–1239.
48. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.
49. Pavlova,M., Klvana,M., Prokop,Z., Chaloupkova,R., Banas,P., Otyepka,M., Wade,R.C., Tsuda,M., Nagata,Y. and Damborsky,J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, **5**, 727–733.
50. Gopal,S., Rastogi,V., Ashman,W. and Mulbry,W. (2000) Mutagenesis of organophosphorus hydrolase to enhance hydrolysis of the nerve agent VX. *Biochem. Biophys. Res. Commun.*, **279**, 516–519.
51. Watkins,L.M., Mahoney,H.J., McCulloch,J.K. and Raushel,F.M. (1997) Augmented hydrolysis of diisopropyl fluorophosphate in engineered mutants of phosphotriesterase. *J. Biol. Chem.*, **272**, 25596–25601.
52. Reetz,M.T., Wang,L.-W. and Bocola,M. (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew. Chem. Int. Ed Engl.*, **45**, 1236–1241.
53. Reetz,M.T., Torre,C., Eipper,A., Lohmer,R., Hermes,M., Brunner,B., Maichele,A., Bocola,M., Arand,M., Cronin,A. *et al.* (2004) Enhancing the enantioselectivity of an epoxide hydrolase by directed evolution. *Org. Lett.*, **6**, 177–180.
54. Cerdobbel,A., De Winter,K., Aerts,D., Kuipers,R., Joosten,H.-J., Soetaert,W. and Desmet,T. (2011) Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. *Protein Eng. Des. Sel.*, **24**, 829–834.
55. Jochens,H., Aerts,D. and Bornscheuer,U.T. (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng. Des. Sel.*, **23**, 903–909.
56. Sullivan,B.J., Nguyen,T., Durani,V., Mathur,D., Rojas,S., Thomas,M., Syu,T. and Magliery,T.J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, **420**, 384–399.
57. Pey,A.L., Rodriguez-Larrea,D., Bomke,S., Dammers,S., Godoy-Ruiz,R., Garcia-Mira,M.M. and Sanchez-Ruiz,J.M. (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
58. Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng. Des. Sel.*, **17**, 787–793.
59. Akasako,A., Haruki,M., Oobatake,M. and Kanaya,S. (1997) Conformational stabilities of Escherichia coli RNase HI variants with a series of amino acid substitutions at a cavity within the hydrophobic core. *J. Biol. Chem.*, **272**, 18686–18693.
60. van den Heuvel,R.H.H., Fraaije,M.W., Ferrer,M., Mattevi,A. and van Berkel,W.J.H. (2000) Inversion of stereospecificity of vanillyl-alcohol oxidase. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9455–9460.
61. Killick,T.R., Freund,S.M. and Fersht,A.R. (1998) Real-time NMR studies on folding of mutants of barnase and chymotrypsin inhibitor 2. *FEBS Lett.*, **423**, 110–112.
62. Encell,L.P., Friedman Ohana,R., Zimmerman,K., Otto,P., Vidugiris,G., Wood,M.G., Los,G.V., McDougall,M.G., Zimprich,C., Karassina,N. *et al.* (2012) Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. *Curr. Chem. Genomics*, **6**, 55–71.
63. Reetz,M.T., Bocola,M., Carballeira,J.D., Zha,D. and Vogel,A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed Engl.*, **44**, 4192–4196.
64. Morley,K.L. and Kazlauskas,R.J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.*, **23**, 231–237.
65. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P.G.M. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.

66. de Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
67. Kuipers,R.K.P., Joosten,H.-J., Verwiel,E., Paans,S., Akerboom,J., van der Oost,J., Leferink,N.G.H., van Berkel,W.J.H., Vriend,G. and Schaap,P.J. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins*, **76**, 608–616.
68. Nobili,A., Tao,Y., Pavlidis,I.V., van den Bergh,T., Joosten,H.-J., Tan,T. and Bornscheuer,U.T. (2015) Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. *Chembiochem*, **16**, 805–810.
69. Wang,C., Huang,R., He,B. and Du,Q. (2012) Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis. *BMC Bioinformatics*, **13**, 263.
70. Martin,L.C., Gloor,G.B., Dunn,S.D. and Wahl,L.M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
71. Fodor,A.A. and Aldrich,R.W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, **56**, 211–221.
72. Nov,Y. (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl. Environ. Microbiol.*, **78**, 258–262.
73. Jacobs,T.M., Yumerefendi,H., Kuhlman,B. and Leaver-Fay,A. (2015) SwiftLib: rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.*, **43**, e34.
74. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.