A web-based tool for rational screening of mutants libraries using ProSAR

Magali Berland, Bernard Offmann, Isabelle Andre, Magali Remaud-Simeon, Philippe Charton

Protein Engineering, Design & Selection vol. 27 no. 10 pp. 375–381, 2014
Published online August 28, 2014 doi:10.1093/protein/gzu035

SHORT COMMUNICATION

A web-based tool for rational screening of mutants libraries using ProSAR

Magali Berland[1,2,3], Bernard Offmann[3,4,*], Isabelle André[5,6,7], Magali Remaud-Siméon[5,6,7] and Philippe Charton[1,2]

[1]DSIMB, INSERM UMR-S1134, Université de La Réunion, 97715 Saint-Denis, France, [2]Laboratory of Excellence GR-Ex, 75015 Paris, France, [3]Université de Nantes, UFIP, UMR CNRS 6286, 2 chemin de la Houssinière, 44000 Nantes, France, [4]Peaccel SAS, Saint-Denis, La Réunion, France, [5]Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, 31077 Toulouse, France, [6]CNRS, UMR 5504, 31400 Toulouse, France and [7]INRA, UMR 792 Ingénierie des Systèmes Biologiques et des Procédés, 31400 Toulouse, France

[*]To whom correspondence should be addressed.
E-mail: bernard.offmann@univ-nantes.fr

Received April 11, 2014; revised July 23, 2014; accepted July 24, 2014

Edited by Frances Arnold

In directed evolution experiments, it is at stake to have methods to screen efficiently the mutant libraries. We propose a web-based tool that implements an established *in silico* method for the rational screening of mutant libraries. The method, known as ProSAR, attempts to link sequence data to activity. The method uses statistical models trained on small experimental datasets provided by the user. These can integrate potential epistatic interactions between mutations and be used in many diverse biological contexts. It drastically improves the search for leading mutants. The tool is freely available to non-commercial users at http://bo-protscience.fr/prosar/.

Keywords: epistatic interactions between mutations/fitness landscape exploration/protein engineering/rational screening/sequence–activity relationship

Introduction

In vitro directed evolution, a process that mimics Darwinian evolution (Stemmer, 1994; Ness *et al.*, 2002), is a well-established strategy for protein optimization (Arnold and Moore, 1997; Dalby, 2011; Wang *et al.*, 2012), protein design (Jäckel *et al.*, 2008; Bloom and Arnold, 2009) and dissecting protein stability, structure and function (Yuen and Liu, 2007; Bloom and Arnold, 2009; Romero and Arnold, 2009; Erijman *et al.*, 2011). It allows the exploration of both the sequence space (Smith, 1970) and fitness landscape (Tracewell and Arnold, 2009) of proteins through iterative cycles of mutagenesis and screening by which neutral and/or beneficial mutations are accumulated (Arnold, 2009). The resulting libraries contain a few mutants displaying improved properties compared with the parental wild-type sequence that has been subjected to directed evolution (Fig. 1). Unveiling these mutants among the astronomically large sequence space to produce them experimentally is wet-lab intensive and practically limited by the screening capacities.

Cost and time for screening mutant libraries are hence at stake. It raises a key question: is there a rational way to discover the leading mutants (Fig. 1)? Here, we propose a tool that implements an *in silico* method for the rational screening of mutant libraries generated during directed evolution cycles to reduce the efforts devoted to their screening. Among the few *in silico* approaches that have been developed towards that goal (Fox *et al.*, 2003; Barak *et al.*, 2008; Damborsky and Brezovsky, 2009; Romero *et al.*, 2013), we selected a method that attempts to link the sequences of a small number of screened protein variants to their corresponding *activities* (Fox, 2005). Here, the notion of *activity* refers to the *fitness* of a protein variant for reasons of clarity and symmetry with the existing literature. This method, termed *protein sequence activity relationship* (ProSAR), assumes that phenotypical information is encoded either directly or indirectly in the amino acid sequence of the protein and explore the sequence space using statistical modeling and sophisticated machine learning algorithms (Fox, 2005). Interestingly, it requires no data derived from three-dimensional structure. This approach relies on the availability of a minimal set of experimental data. Sequence and activity data are provided to a genetic algorithm (GA) that builds partial least square (pls) regression models where the contribution of residues and epistatic coupling between residue pairs are combined. These models can be used to predict high performing variants from the library that were not sampled experimentally and thus, drive further directed evolution rounds.

To our knowledge, there is to date no software available to the community that implements the ProSAR method. Here, we present the first web-based software that implements the method and discuss how it can be practically used for the rational screening of mutant libraries. We further illustrate its applicability on two different experimental datasets.

Methods

The ProSAR methodology

The ProSAR methodology (Fox, 2005) is composed of three main steps as illustrated in the Fig. 2. For the method to be applicable, the mutant library, which was constructed after a mutagenesis strategy, must contain protein variants with multiple mutations (some of them may be point mutations) and no insertion/deletion. It relies on the availability of a minimal set of experimental data, i.e. requires sequencing and phenotyping of a sample of variants from the mutant library. These sequences, along with their measured activities, form the training dataset. In this context, the term *activity* refers to any quantitative criterion representing the protein *fitness*. It can be, for example, the optimum temperature or pH, catalytic constant, affinity, time yield, product inhibition, solvent stability, substrate or product specificity, substrate conversion, regioselectivity, enantioselectivity, etc. The process starts with a *data*

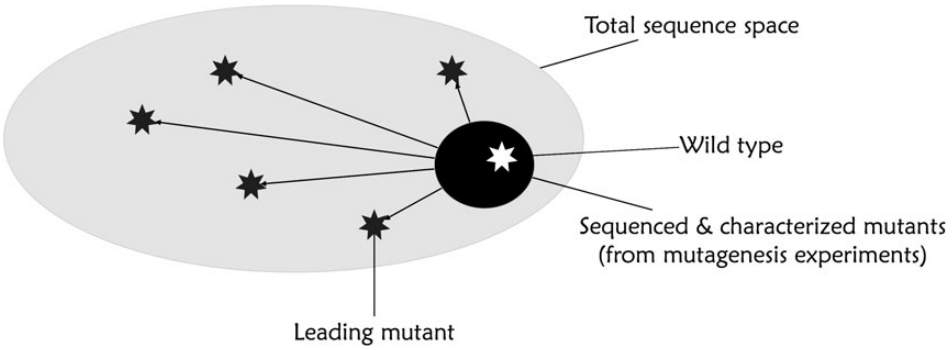


Fig. 1. The problem of rational screening of mutant libraries. From a small set of sequenced and characterized variants, *in silico* methods for rational screening aim to unveil the leading mutants from the sequence space.

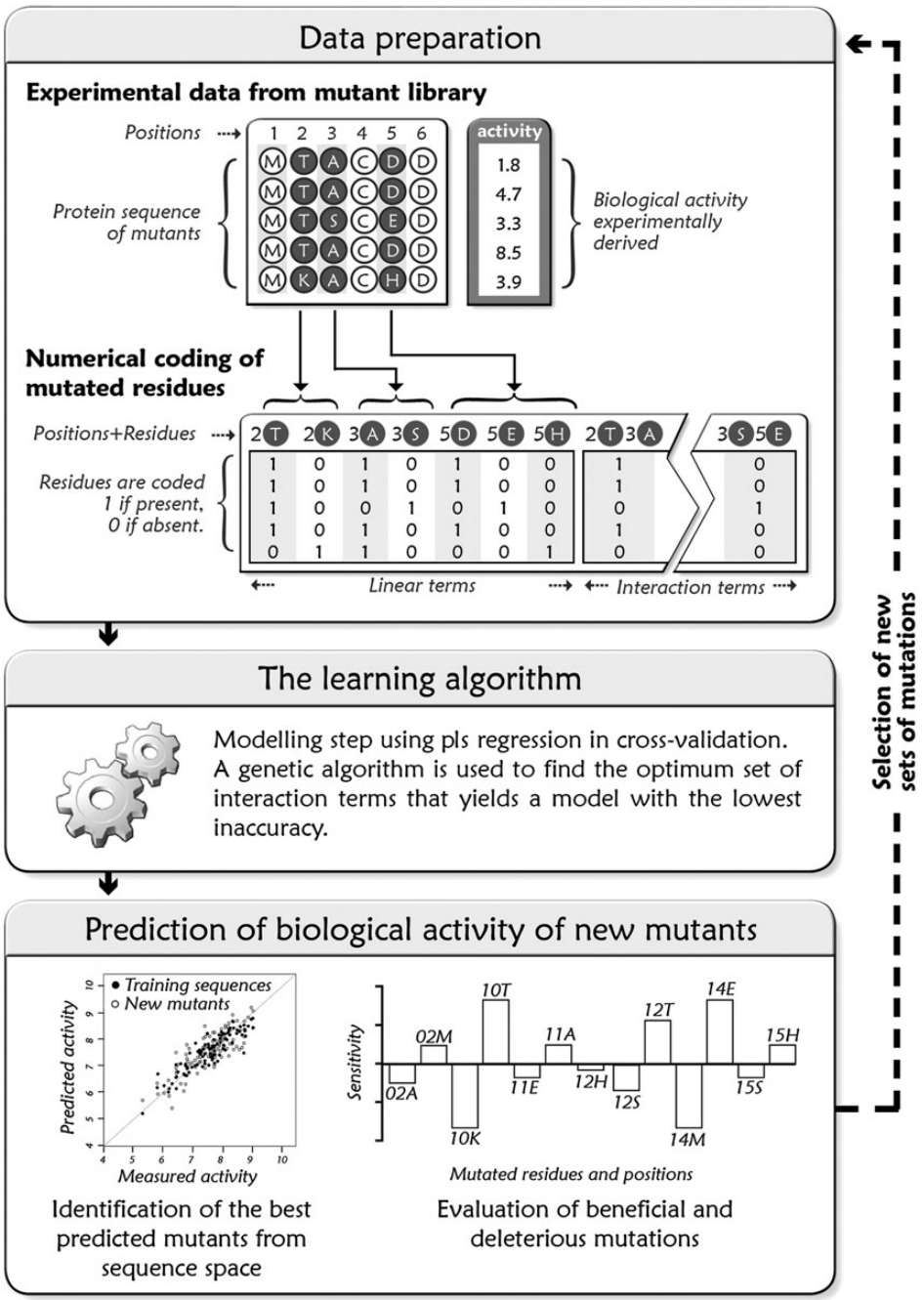


Fig. 2. Illustration of the three main steps of the implemented ProSAR method: (i) data preparation, (ii) learning process and (iii) prediction.

preparation step which consists of a numerical encoding of mutated residues and positions from a sequence alignment. Each amino acid residue is encoded as either present (1) or absent (0) for each variable position in the alignment, resulting in *linear terms*. Non-variable positions are not encoded. Epistatic coupling between residue pairs are numerically encoded as the product between two *linear terms*, i.e. between two residues at variable positions. These new terms are named *interaction terms*, or *non-linear terms*. The question of which interaction terms and how many should be considered is addressed further on.

The *learning process* step aims to set up a statistical model to link the activities to the mutations and their interactions. A *pls* regression is performed. This dimension reduction method is required for the analysis of large datasets made of many highly collinear variables (linear and interaction terms) and few observations (sequences) (Wold, 1985; de Jong *et al.*, 2001; Damborsky and Brezovsky, 2009). According to the ProSAR model, the most effective method to obtain the activity from sequence information would rely on the linear terms and only the relevant interaction terms (Fox, 2005). Taking all the interaction terms at the same time decreases the predictive power of the *pls* regression (Broadhurst *et al.*, 1997). To compensate the absence of information on the structure and on the possible interactions between amino acids, the *learning process* uses a GA) to compute an optimal subset of interaction terms. GA is a widely used search strategy to solve optimization problems (Holland, 1975; Daren, 2001; Hand *et al.*, 2001). It mimics the search efficiency observed in genetic evolution. The GA implemented in the ProSAR method consists in the computation of numerous statistical models (*pls*), each based on different interaction terms, to preferentially select the models whose inaccuracy (evaluated by cross-validation) is the lowest. The inaccuracy is calculated as the root-mean-square error between the predicted and measured activities. These details regarding algorithm descriptions were previously reported in a methodology paper (Fox, 2005).

The *prediction* step: After the convergence of the GA towards an optimal set of interaction terms, a pls regression is performed on the whole training dataset with the selected terms. This model is used to predict the activity of new sequences containing a combination of the mutations sampled in the training dataset. The evaluation of the beneficial and deleterious mutations is performed by a metric called *sensitivity*. The sensitivity determines the relative contribution to the activity of the amino acids. The sensitivity of a given residue choice X at a position i is the average change in activity between the sequences possessing the residue X at position i, and the sequences possessing another residue at this position. The sensitivity of the residue X at position i is given by:

$$S(X_i) = \frac{\sum_{p\in P, q\in Q}(a_p - a_q)}{\#P \times \#Q}$$

where P is the set of sequences that possess the residue X at position i, Q is the set of sequences that possess the residue Y ($\neq X$) at position i and a_p (respectively, a_q) is the predicted activity of protein sequence in set P (respectively, Q).

Hence, the sensitivity represents the fitness changes and the largest sensitivities then correspond to residues choices that are important for improving protein function. These residues can be fixed in the next round of directed evolution to focus the mutagenesis.

Table I. Characteristics of the experimental datasets

Dataset	n	k	NLT_{the}	Sequence space
Dextransucrase	4	[14;14;18;12]	1252	42 336
Cytochrome P450	8	3	252	6561

n is the number of mutated positions and k is the number of residues at each position. NLT_{the} is the theoretical total number of interaction terms. Some of them might be null, depending on the sequences available in the training dataset.

Experimental datasets

The performance of the ProSAR model is presented on two experimental datasets whose combinatorial features are detailed in Table I. The first dataset (Irague *et al.*, 2012) was created to investigate the relationship between the specificity of dextransucrase DSR-S (EC 2.4.1.5) mutants from *Leuconostoc mesenteroides* and the structure of their polysaccharide products, i.e. the presence of **α**(1→3) or **α**(1→6) linkages. Semi-rational engineering was carried out to generate enzyme mutants with altered specificity. Four amino acids likely to be critical for the enzyme regioselectivity were targeted from sequence and structural analysis. The measured activity is the percentage of **α**(1→3) linkage found in the synthesized polysaccharides. A total of 79 distinct mutants were screened and sequenced for this dataset. Their activity, here the **α**(1→3) percentage, ranged from 1 to 8.3%.

The second dataset (Li *et al.*, 2007; Romero *et al.*, 2013) was generated in a study of the sequence–stability–function relationship related to the cytochrome P450 family, specifically the cytochrome P450 BM3 A1, A2 and A3. The aim of this study was to improve the thermostability of cytochromes P450. New chimera biocatalysts were made up of eight consecutive fragments inherited from any of these three different parents. The measured activity is the T_{50}, i.e. the temperature at which 50% of the protein is irreversibly denatured after incubation for 10 min. The resulting dataset is composed of 242 sequences of variants and T_{50}, measurements that ranged from 39.2 to 64.4°C. Here, the variants were encoded according to the following scheme: each fragment was encoded by a letter according to its parental origin (A for parent A1, B for parent A2 and C for parent A3). For example, a variant where the second and sixth fragment comes from parent B and the others from parent A will be encoded as ABAAABAA.

Results

The prior requirement for the use of the web server is the availability of experimental data (sequences and activities) obtained in the context of a directed evolution experiment. One of the first question raised by the mining of the untapped libraries of mutants, as presented on Fig. 1, is the total size of the sequence space. This information can be obtained by using a tool implemented in the web server (toolbox menu). It is also possible to calculate the number of theoretical interaction terms from the number of mutated residue at each position (in practice, some of them might be null, depending on the mutations sampled in the sequences). The aim of these tools is to calculate the combinatorial complexity of the protein-activity system before the modeling phase.

Where are the top-ranking mutants located in the total sequence space (Fig. 1)? The ProSAR predictions answer this question in two ways, depending on the additivity of the mutations. The mutations of a protein are called additive if the activity of the protein can be predicted from the linear terms only. If the mutations are additive, there is no need to calculate interaction terms, and the prediction is straightforward from a *pls* regression model. If the mutations are not additive, the calculation of the optimal set of interaction terms requires sophisticated statistical learning methods to be used, and these methods are also implemented on the web server.

For illustration purpose, we present the application of the ProSAR method on two experimental datasets (Fig. 3). Figure 3a and b show the influence of the number of interaction terms on the model inaccuracy. The optimal number of interaction terms cannot be known in advance and depends on the available dataset since each protein-activity system is unique. For the dextransucrases (Fig. 3a), 160 interaction terms were necessary, whereas for the cytochrome P450 (Fig. 3b), 45 interaction terms were sufficient.

The observed difference between the number of interaction terms needed in the two models can be explained by the different level of additivity of either the mutations in the dextransucrase case or sequence fragments in the cytochromes P450 case. In the latter case, the fragments function as pseudo-independent structural modules that make roughly additive contributions to stability (Li *et al.*, 2007). This may explain the need for less interaction terms than for the dextransucrase. The practical consequences of these results are discussed below.

Figure 3c and d illustrate the predicted activity as a function of the measured activity in 10-fold cross-validation for a model containing the optimal number of interaction terms determined previously. The results show that in both cases, the model is reliable enough to enable the prediction of new sequences: $R^2 = 0.60$ for the dextransucrase and $R^2 = 0.94$ for the cytochrome P450. For the cytochrome P450, we compared the R^2 with the one obtained from Gaussian process models (Romero *et al.*, 2013): the R^2 was 0.90. This confirms the performance of the ProSAR method. By rendering this method available through a web server, we thus offer a powerful tool for the exploration of mutant libraries. These two examples raise a number of questions that we address in the following paragraphs.

How to take into account epistatic interaction of residue pairs?

Each studied protein-activity system has an optimal number of interaction terms to be considered. We advise to start with an initial run of the ProSAR algorithm without interaction terms to evaluate the additivity of the mutations: by observing the R^2 metric (ranging from 0 to 1), the closest it is to 1, the more additive are the mutations. This calculation is very fast compared with the calculations needed for models including interaction terms. Moreover, a linear model is by nature less prone to overfitting, and thus gives more robust models, especially in the case of few sequences available.

If the model with only linear terms has poor R^2, it means interaction terms must be integrated in the model. For this and as a consequence of results presented in Fig. 3a and b, the user is required to perform several runs of the algorithm with different number of interaction terms with the maximum number given by the NLT_{the} value computed with the toolbox available on the web server. In practice, we recommend increasing

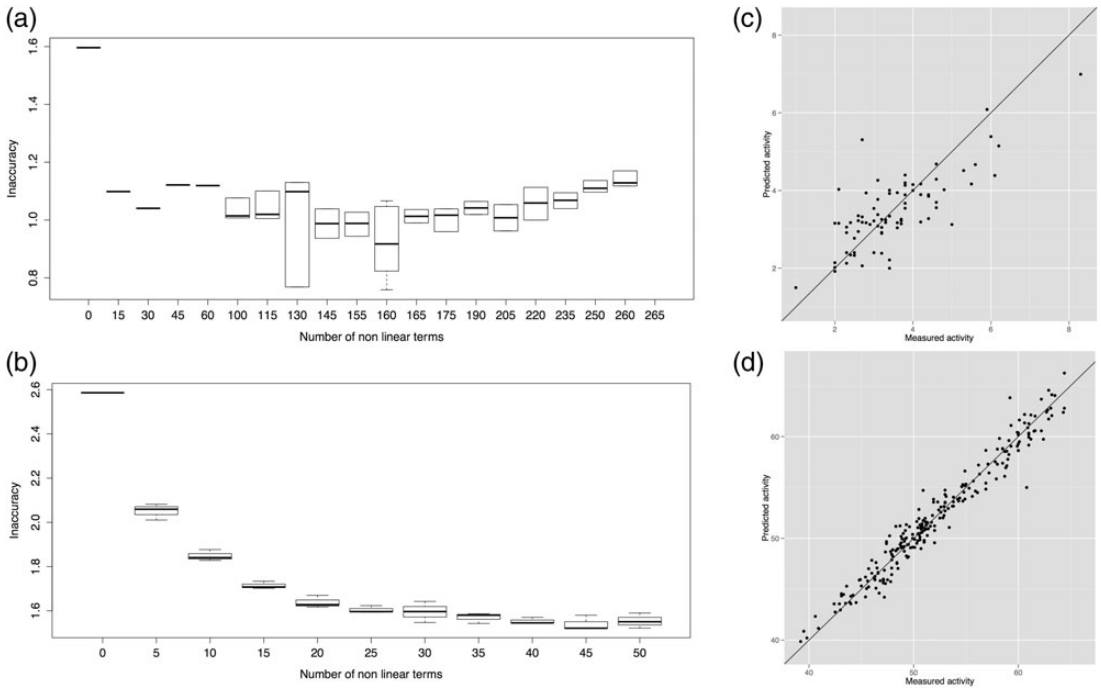


Fig. 3. Evaluation of the performance of the ProSAR method on two experimental datasets. On the left are shown the influence of the number of interaction terms on the inaccuracy (a) for the dextransucrases and (b) for the cytochrome P450. *X*-axis represents the number of interaction terms used in each simulation and *Y*-axis represents the inaccuracy of the model predictions obtained by cross-validation. Three duplicates have been performed each time. On the right are shown the correlation between measured and predicted activities for (c) for dextransucrase ($R^2 = 0.60$), and (d) cytochrome P450 ($R^2 = 0.96$). The *X*-axis represents the measured activity of the enzymes (regioselectivity for dextransucrase and thermostability for cytochrome P450) and the *Y*-axis represents the predicted activity in 10-fold cross-validation (dextransucrase: 160 non-linear terms, cytochromes P450: 45 non-linear terms).

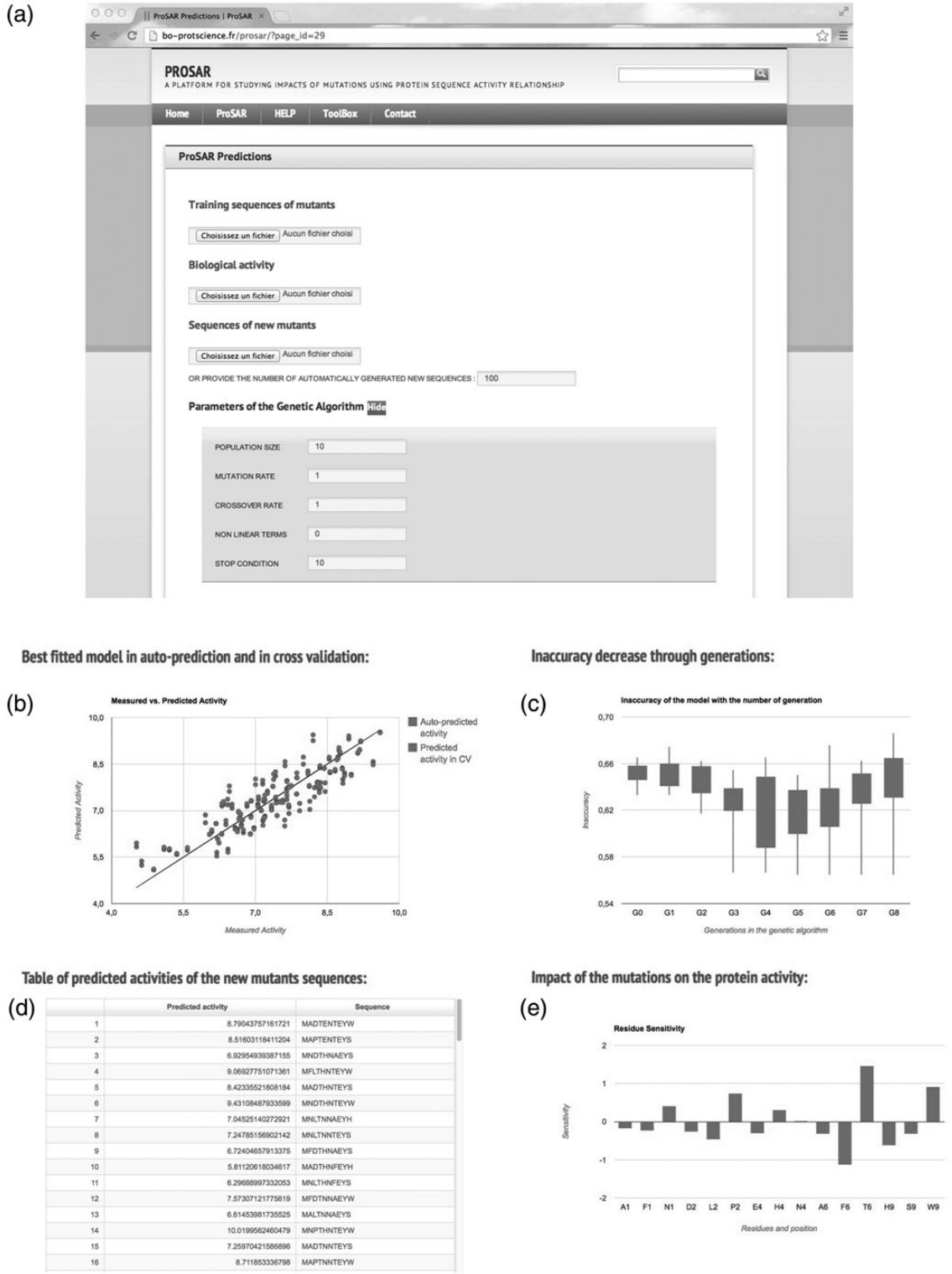
the number of interaction terms step by step until satisfactory inaccuracy and R^2 are obtained.

The screening effort required to use ProSAR

The number of sequences needed in the training dataset to establish a reliable model depends on numerous parameters:

(i) The additivity of the mutations. The more additive are the mutations, the more the combinations of mutations will be predictable and thus the less sequences will be needed in the training dataset.

(ii) The precision of the measure. The more the measure of the activity will be precise, accurate and



	Predicted activity	Sequence
1	8.79043757161721	MADTENTEYW
2	8.51603118411204	MAPTENTEYS
3	6.92954939387155	MNDTHNAEYS
4	9.06927751071361	MFLTHNTEYW
5	8.42335521808184	MADTHNTEYS
6	9.43108487933599	MNDTHNTEYW
7	7.04525140272921	MNLTNNAEYH
8	7.24785156902142	MNLTNNTEYS
9	6.72404657913375	MFDTHNAEYS
10	5.81120618034617	MADTHNFEYH
11	6.2968899733205	MNLTHNFEYS
12	7.57307121775619	MFDTNNAEYW
13	6.61453981735525	MALTNNAEYS
14	10.0199562460479	MNPTHNTEYW
15	7.25970421586896	MADTNNTEYS
16	8.711853336798	MAPTNNTEYW

Fig. 4. Screenshots of the ProSAR web server input form and output results display for a toy example. (a) ProSAR prediction input form to be completed with the sequence and activity files, and GA parameters. (b) Output of the best-fitted model. (c) Output of inaccuracy decrease through the generations for all the models represented with boxplots. (d) Output table of predicted activities of the new mutants. (e) Output of the impact of mutations on protein activity represented as barplot. The mutations with large sensitivity impact the most the activity.

reproducible, the more the method will be tolerant to the small number of sequences. The type of activity is also important, as it should give the best picture of the mechanism of the protein.

(iii) The sequence space (number of total sequences given positions and mutations). The more positions are mutated, the more sequences will be needed to capture the relative effect of each mutation. It is possible to calculate the sequence space with the toolbox available on the web server.

For a small number of variable positions (e.g. 4–8), we suggest having a training dataset composed of at least 80–100 sequences to have a proper use of the method. For much larger libraries (10 variable positions or more), more data will surely be needed. Based on our benchmarking of the method, we discourage the use of many interaction terms in the case where limited experimental data are available.

How much time does take a ProSAR modeling?

The computing time may be the major limiting factor of this approach, especially when the number of interaction terms to take into account is high. The computing time increases also with the number of generations needed for convergence, and the number of protein arrays in the algorithm. For these reasons, we suggest to increase the number of interaction terms step by step, especially when the protein-activity system has a high combinatorial complexity (see above).

Biological interpretation of the regression coefficients

The *pls* regression is a method that relies on a few latent variables which are a linear combination of the input variables to explain the change in activity. Due to this mechanism, it is not possible to give a biological meaning to the regression coefficients calculated for each variables (linear and interaction terms). However, the sensitivity metric compensate the difficult interpretation of the coefficients by indicating the relative influence of the mutations on the activity. Positive residue sensitivity means that the residue at this position provides a global positive effect on the biological activity. This metric enables to evaluate which mutations are crucial to improve the activity of the protein.

Required inputs and server outputs

The web server that implements the ProSAR method is available at http://bo-protscience.fr/prosar/. The aim of this web server is to be routinely integrated in the process of a directed evolution experiment. By virtue of its capacity to handle many diverse protein-activity systems, the ProSAR method can be quickly used through the user-friendly interface. It is freely available for academic users.

Prediction of activity change upon multiple mutations requires providing, on one hand, the list of the protein sequences, and, on the other hand, the corresponding activities measured experimentally. Example files of datasets are downloadable. Optionally, it is possible to provide a list of sequences on which the activity prediction will be performed. Otherwise, a fixed number of new sequences will be generated: they feature the possible combinations of the mutations observed in the training dataset. The parameters for the GA are filled by default to run a quick simulation. However, they are tunable to fit every dataset (Fig. 4a).

For example, in the case of the dextransucrase, we provided the 79 sequences (in FASTA format) containing single or multiple mutations as *Training sequences of mutants*. All the sequences contained the same number of residues. The corresponding biological activity (plain text file with the list of activities —one per line) were provided in the *Biological activity* field. We computed several models with different numbers of non-linear terms until we found the optimal model (i.e. with the smallest inaccuracy, Fig. 4c). The sensitivity barplot and the table of predicted activities of new sequences can help the design of new mutant by identifying relevant mutations or combination of mutations.

The server outputs are of two types:

(i) The main results are displayed online: impact of the mutations on the protein activity, table of predicted activities of the new sequences, best-fitted model in auto-prediction and in cross-validation, inaccuracy decrease through generations (Fig. 4b–e).

(ii) Simultaneously, the complete results are sent by email as a zip archive of 15 files containing among others the selected interaction terms and the *pls* regression coefficients to enable the prediction of any other mutants.

Acknowledgements

The analysis in this article was performed using both HPC resources from CCUR and CALMIP (Grant 2012-[P1215]). Technical assistance for web server setup was provided by Nicolas Fontaine from Peaccel SAS.

Competing interest: B.O. is co-founder of Peaccel SAS. This work was in part supported by Peaccel SAS.

Funding

This work was supported by the Région Réunion and the Fond Social Européen [grant no. 20100017 to M.B.] and the Région Pays de la Loire for ReMSIP project [grant no. 2012-7279-7281 to B.O.].

References

Arnold,F.H. (2009) *Cold Spring Harb. Symp. Quant. Biol.*, **74**, 41–46.
Arnold,F.H. and Moore,J.C. (1997) *Adv. Biochem. Eng. Biotechnol.*, **58**, 1–14.
Barak,Y., Nov,Y., Ackerley,D.F. and Matin,A. (2008) *ISME J*, **2**, 171–179.
Bloom,J.D. and Arnold,F.H. (2009) *Proc. Natl Acad. Sci.*, **106**(Suppl. 1), 9995–10000.
Broadhurst,D., Goodacre,R., Jones,A., Rowland,J.J. and Kell,D.B. (1997) *Anal. Chim. Acta*, **348**, 71–86.
Dalby,P.A. (2011) *Curr. Opin. Struct. Biol.*, **21**, 473–480.
Damborsky,J. and Brezovsky,J. (2009) *Curr. Opin. Chem. Biol.*, **13**, 26–34.
Daren,Z. (2001) *Comput. Chem.*, **25**, 197–204.
de Jong,S., Wise,B.M. and Ricker,N.L. (2001) *J. Chemometr.*, **15**, 85–100.
Erijman,A., Aizner,Y. and Shifman,J.M. (2011) *Biochemistry*, **50**, 602–611.
Fox,R. (2005) *J. Theor. Biol.*, **234**, 187–199.
Fox,R., Roy,A., Govindarajan,S., Minshull,J., Gustafsson,C., Jones,J.T. and Emig,R. (2003) *Protein Eng.*, **16**, 589–597.
Hand,D.J., Mannila,H. and Smyth,P. (2001) *Principles of Data Mining*, MIT Press, USA.
Holland,J.H. (1975) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, U Michigan Press, USA.
Irague,R., Rolland-Sabaté,A., Tarquis,L., Doublier,J.L., Moulis,C., Monsan,P., Remaud-Siméon,M., Potocki-Véronèse,G. and Buléon,A. (2012) *Biomacromolecules*, **13**, 187–195.
Jäckel,C., Kast,P. and Hilvert,D. (2008) *Annu. Rev. Biophys.*, **37**, 153–173.

Li,Y., Drummond,D.A., Sawayama,A.M., Snow,C.D., Bloom,J.D. and Arnold,F.H. (2007) *Nat. Biotechnol.*, **25**, 1051–1056.
Ness,J.E., Kim,S., Gottman,A., Pak,R., Krebber,A., Borchert,T.V., Govindarajan,S., Mundorff,E.C. and Minshull,J. (2002) *Nat. Biotechnol.*, **20**, 1251–1255.
Romero,P.A. and Arnold,F.H. (2009) *Nat. Rev. Mol. Cell. Biol.*, **10**, 866–876.
Romero,P.A., Krause,A. and Arnold,F.H. (2013) *Proc. Natl Acad. Sci.*, **110**, E193–E201.
Smith,J.M. (1970) *Nature*, **225**, 563–564.
Stemmer,W.P. (1994) *Proc. Natl Acad. Sci. USA.*, **91**, 10747–10751.
Tracewell,C.A. and Arnold,F.H. (2009) *Curr. Opin. Chem. Biol.*, **13**, 3–9.
Wang,M., Si,T. and Zhao,H. (2012) *Bioresour. Technol.*, **115**, 117–125.
Wold,H. (1985) *Partial Least Squares*, Encyclopedia of Statistical Sciences, Wiley Ed., USA.
Yuen,C.M. and Liu,D.R. (2007) *Nat. Methods*, **4**, 995–997.