PhD7Faster 2.0: predicting clones propagating faster from the Ph.D.-7 phage display library by coupling PseAAC and tripeptide composition

Bifang He[1,2], Heng Chen[1] and Jian Huang[2]

[1] School of Medicine, Guizhou University, Guiyang, Guizhou, China
[2] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

ABSTRACT

Selection from phage display libraries empowers isolation of high-affinity ligands for various targets. However, this method also identifies propagation-related target-unrelated peptides (PrTUPs). These false positive hits appear because of their amplification advantages. In this report, we present PhD7Faster 2.0 for predicting fast-propagating clones from the Ph.D.-7 phage display library, which was developed based on the support vector machine. Feature selection was performed against PseAAC and tripeptide composition using the incremental feature selection method. Ten-fold cross-validation results show that PhD7Faster 2.0 succeeds a decent performance with the accuracy of 81.84%, the Matthews correlation coefficient of 0.64 and the area under the ROC curve of 0.90. The permutation test with 1,000 shuffles resulted in $p < 0.001$. We implemented PhD7Faster 2.0 into a publicly accessible web tool (http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl) and constructed standalone graphical user interface and command-line versions for different systems. The standalone PhD7Faster 2.0 is able to detect PrTUPs within small datasets as well as large-scale datasets. This makes PhD7Faster 2.0 an enhanced and powerful tool for scanning and reporting faster-growing clones from the Ph.D.-7 phage display library.

Submitted 22 December 2018
Accepted 15 May 2019
Published 17 June 2019

Corresponding author
Jian Huang, hj@uestc.edu.cn

Academic editor
Tatiana Tatarinova

Additional Information and Declarations can be found on page 10

DOI 10.7717/peerj.7131

Subjects Bioinformatics, Data Mining and Machine Learning
Keywords Propagation-related TUPs, Ph.D.-7 phage display library, Predictor, Feature selection, Support vector machine

INTRODUCTION

Phage display is a high throughput and powerful screening methodology for identifying ligands for myriad target types, ranging from molecules (microRNA, protein, polysaccharide) (*He et al., 2013*; *Zhang et al., 2017*) to inorganic (gold) (*Causa et al., 2013*), organic (epoxy) (*Swaminathan & Cui, 2013*), and biological (tissue, organ) materials (*Hung et al., 2018*). Large libraries of phage-displayed peptides or proteins consist of millions to billions of variant members, which can be iteratively selected and amplified in a process referred to as biopanning (*Pande, Szewczyk & Grover, 2010*). Recently, next generation sequencing technologies have been coupled with phage display, which have substantially contributed to the analysis of output from combinatorial libraries and allowed for even faster and more robust discovery of novel ligands (Christiansen et al., 2015; Matochko et al., 2014; Ngubane et al., 2013; Rentero Rebollo et al., 2014; 't Hoen et al., 2012). The ever-increasing utility and versatility makes phage display a powerful tool in multiple research areas, such as materials science, biotechnology, pharmacology, cell biology, and diagnostics (Martins, Reis & Azevedo, 2016).

However, the phage display methodology is notorious for the enrichment of target-unrelated peptides (TUPs) (Menendez & Scott, 2005). Therefore, biopanning results are a mixture of true target binders and TUPs (Vodnik et al., 2011). These false positive TUPs have no actual affinity toward the target of interest and can fall into two categories: selection- and propagation-related TUPs (SrTUPs and PrTUPs) (Thomas, Golomb & Smith, 2010). The SrTUPs can bind to other components (plates, beads) of the screening system other than the desired target and thus creep into the output of phage display. The PrTUPs sneak into the biopanning results due to their propagation advantages, which allow them to outcompete clones with lower growth rates (Brammer et al., 2008; Matochko et al., 2014; Nguyen et al., 2014; Thomas, Golomb & Smith, 2010; Zade et al., 2017; Zygiel et al., 2017). Apparently, these TUPs may misdirect ligand discovery through biopanning and should be distinguished from actual target-binding peptides (Bakhshinejad et al., 2016). Therefore, the diagnosis of TUPs is as crucial as the identification of target binders.

Although several experimental strategies have been proposed to decrease TUP isolation during biopanning and differentiate between TUPs and true binders post-biopanning (Nguyen et al., 2014; Thomas, Golomb & Smith, 2010; Vodnik et al., 2011), TUP analysis has benefitted considerably from computational approaches. Databases (BDB (He et al., 2016a, 2018; Huang et al., 2012; Ru et al., 2010), PepBank (Shtatland et al., 2007)) and bioinformatics tools (He et al., 2016b; Huang et al., 2010; Li et al., 2017; Mandava et al., 2004; Ru et al., 2014) have been widely employed to report both SrTUPs and PrTUPs. Searching against databases for biopanning data can uncover whether query peptides have been isolated by many different targets. If so, query sequences are potential SrTUPs and PrTUPs due to lack of target specificity. For example, the peptide HAIYPRH (a typical PrTUP) has been identified by 23 completely different targets according to results of searching the BDB database. The phage displaying the peptide was later verified to have a propagation advantage owing to mutations in the regulatory region of the phage genome (Brammer et al., 2008). HWGMWSY (a SrTUP) has been isolated by 10 completely different targets according to records in the BDB database. The peptide was proved to be a plastic binder (Vodnik, Strukelj & Lunder, 2012), which resulted in this peptide repeatedly appearing in multiple reported screening experiments. SABinder (He et al., 2016b) and PSBinder (Li et al., 2017) have been designed for predicting streptavidin- and polystyrene surface-binding peptides, respectively, as they are commonly known SrTUPs. The INFO tool in the RELIC suite enables PrTUPs detection based on information content (Mandava et al., 2004), whereas PhD7Faster (PhD7Faster 1.0) based on support vector machine (SVM) allows the prediction of clones with amplification advantages from the popular commercial Ph.D.-7 phage display library (Ru et al., 2014). However, PhD7Faster 1.0 can be improved in the following three aspects. Firstly, the positive training dataset of

PhD7Faster 1.0 was selected based on the copy number of a peptide (15 or higher) after one round of amplification without consideration of the corresponding copy number in the naïve Ph.D.-7 library. Secondly, only dipeptide composition was employed to develop the classifier. Currently, many reports have demonstrated that predictors developed by combining pseudo amino acid composition (PseAAC) (*Chou, 2001*, *2005*) and tripeptide composition can achieve decent predictive performances (*Liao et al., 2011*; *Zhu et al., 2015*). Thirdly, PhD7Faster 1.0 is unable to process large datasets (e.g., next-generation sequencing data).

In this study, we develop a new predictor for identifying clones propagating faster from the Ph.D.-7 phage display library. The SVM algorithm was employed to model the predictor with the optimal feature subset after feature selection. The constructed SVM-based classifier obtained an accuracy of 81.84% in the ten-fold cross-validation. The predictor was further implemented into a web tool, called PhD7Faster 2.0, which is freely available at http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl. We also developed the standalone version of PhD7Faster 2.0 that enables the analysis of PrTUPs within large-scale datasets.

DATA AND METHODS

Benchmark datasets

The dataset used to develop the predictor was acquired from (*Matochko et al., 2014*). Derda et al. employed high-throughput sequencing technology to characterize both the naïve Ph.D.-7 phage display library and the same library after one round of amplification. By comparing the abundance of each peptide before and after amplification using Bioconductor package edgeR, 770 unique peptides were identified with significantly higher growth rate (parasitic sequences) (*Matochko et al., 2014*), which were collected into the positive training dataset. The negative dataset was composed of those peptides with the copy number of one in the amplified Ph.D.-7 phage display library. The datasets were then processed as follows: (i) peptide sequences containing ambiguous residues (such as "X", "B," and "Z") were excluded; (ii) sequences within 2 Hamming distance (h = 2, the Hamming distance between two strings of equal length is the minimum number of substitutions required to change one string into the other.) were removed. Finally, 749 peptides were retained in the positive dataset. To match the size of the positive dataset, we randomly selected 749 peptides from the negative dataset. No overlapping was found between the negative and positive datasets. Finally, the benchmark dataset was composed of 749 fast-growing peptides and 749 regular-growing peptides (See positive.fasta and negative.fasta in Supplementary Data).

PseAAC and tripeptide composition

Extracting a set of informative features is a standard and important procedure for developing predictors. Chou initially formulated the PseAAC (*Chou, 2001*, *2005*), which consists of more than 20 discrete numbers, where the top 20 represent the classical amino acid composition (AAC) of a protein sequence whereas the additional parameters incorporate some sequence-order information. PseAAC and tripeptide composition have been widely used in protein prediction related research (Chou, 2011; Lin et al., 2013). Here, they were employed to encode each peptide in the benchmark dataset.

Given a peptide P with L amino acid residues:

$$\mathrm{P} = (R_1R_2R_3R_4R_5R_6R_7\ldots R_L) \tag{1}$$

where R_i (i = 1, 2, 3 . . . L) is the residue at the ith sequence position. Accordingly, any sequence like the peptide P of Eq. (1) can be presented using a set of feature vectors with 8,000 + $n\lambda$ dimensions.

$$\mathrm{P} = \left(P_1, P_2, \cdots, P_{8{,}000}, P_{8{,}000_{+1}}, \cdots, P_{8{,}000+n\lambda}\right) \tag{2}$$

where the first 8,000 numbers P_1, P_2, . . ., $P_{8{,}000}$ reflect the effect of the conventional tripeptide composition; the remaining $n\lambda$ elements $P_{8{,}000+1}$, $P_{8{,}000+2}$, . . ., $P_{8{,}000+n\lambda}$ reflect the amphipathic sequence-order pattern. These features are calculated through the following equations:

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{8{,}000} f_i + w\sum_{j=1}^{n\lambda}\tau_j} (1 \le u \le 8{,}000) \\ \dfrac{w\tau_u}{\sum_{i=1}^{8{,}000} f_i + w\sum_{j=1}^{n\lambda}\tau_j} (8{,}000 + 1 \le u \le 8{,}000 + n\lambda) \end{cases} \tag{3}$$

where f_i (i = 1, 2, 3, . . ., 8,000) are the normalized occurrence frequencies of the 8,000 tripeptides in peptide P; w is the weight factor for the sequence-order effect; (τ_j (j = 1, 2, . . ., $n\lambda$) is the j-tier sequence-correlated factor as formulated by:

$$\begin{cases} \tau_1 = \dfrac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^{1} \\ \tau_2 = \dfrac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^{2} \\ \quad\ldots \\ \tau_{\mathrm{n}} = \dfrac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^{n} \\ \tau_{\mathrm{n}+1} = \dfrac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^{1} \\ \tau_{\mathrm{n}+2} = \dfrac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^{2} \\ \quad\ldots \\ \tau_{2\mathrm{n}} = \dfrac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^{n} \\ \quad\ldots \\ \tau_{\mathrm{n}\lambda-1} = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{n-1} \\ \tau_{\mathrm{n}\lambda} = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{n} \end{cases} \tag{4}$$

where $H^n_{i,j}$ is the physicochemical property correlation function and can be computed according to the following equation:

$$H^n_{i,j} = h^n(R_i).h^n(R_j) \tag{5}$$

where $h^n(\mathrm{R}_i)$ and $h^n(\mathrm{R}_j)$ are the values of the nth type of physicochemical property of R_i and R_j in Eq. (1), respectively. It is noteworthy that before substituting the values of all physicochemical properties into Eq. (5), they were undergone a standard conversion as described below:

$$h^k(R_i) = \frac{h^k_0(R_i) - \sum_{\alpha=1}^{20} h^k_0(R_\alpha)/20}{\sqrt{\sum_{u=1}^{20}\left(h^k_0(R_i) - \sum_{\alpha=1}^{20} h^k_0(R_\alpha)/20\right)^2}} \tag{6}$$

where R_i (i = 1, 2, . . ., 20) denotes the 20-standard amino acid in the alphabetical order of their single-letter codes. $h^k_0(R_i)$ is the initial value of the kth type of physicochemical property for amino acid residue R_i. Nine kinds of physicochemical properties, namely hydrophobicity, hydrophilicity, mass, pK1, pK2, pI, rigidity, flexibility, and irreplaceability, were considered in this report.

Feature selection

Generally, not all features make an equal contribution to the prediction system. A part of features make significant contributions, while some others make less important contributions (*Zhao et al., 2016*). Feature selection, thus, is a critical step to reduce feature dimensionality and build a highly effective prediction model (*Su et al., 2018*; *Tang, Chen & Lin, 2016*). In this work, the fselect.py program in the LIBSVM 3.23 package was applied to evaluate each feature's significance to the classification system (*Chang & Lin, 2011*). As a consequence, each feature corresponds to an *F*-score. The greater *F*-score implies the larger importance of the corresponding feature to the prediction model. We rearranged all features by *F*-scores in descending order. The incremental feature selection strategy was then utilized to determine the optimal feature subset (*He et al., 2016b*; *Li et al., 2017*), which can produce the maximal accuracy. Feature selection was conducted as follows: (i) investigating the accuracy of the first feature subset which included the feature with the largest *F*-score; (ii) examining the accuracy of the second feature subset that was generated by appending the feature with the second largest *F*-score; (iii) iterating the second step from the larger *F*-score to the smaller *F*-score until all candidate features were added. The best feature subset with the highest accuracy can be finally obtained.

Support vector machine

The SVM is a powerful supervised learning method, which has been widely applied in classification (*He et al., 2016b*; *Kang et al., 2018*; *Li et al., 2017*; *Ru et al., 2014*) and regression analysis. In this study, we utilized the LIBSVM 3.23 program (*Chang & Lin, 2011*) that could be freely available for download from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. We chose the radial basis function kernel as the kernel function. The optimal kernel width parameter γ and penalty constant C were selected by using the parameter selection tool in the LIBSVM 3.23 (*Chang & Lin, 2011*).

Performance evaluation

The 10-fold cross-validation was adopted to evaluate the predictive model in this study. Four commonly-used parameters, including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC), were employed to investigate the performance of the constructed model. These measures were expressed as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \tag{8}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{9}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{10}$$

where TP and TN denote the number of true positives and negatives, respectively. FP and FN are the number of false positives and negatives, respectively. The area under the receiver operating characteristic (ROC) curve (AUC) was also calculated as a performance measure. The AUC ranges from zero to one. The AUC of one represents a perfect prediction, 0.5 a random guess.

To estimate the statistical significance of the predictive accuracy, a permutation test with 1,000 shuffles was performed by exchanging the labels of the benchmark dataset. The 10-fold cross-validation was then conducted against the label-rearranged dataset. Thus, each permutation trial corresponds to an accuracy value. The *p*-value was calculated by the number of permutations that the Acc produced by the permuted dataset was higher than Acc based on the un-permuted dataset divided by the overall shuffle times. *p*-values of <0.05 were referred to as statistically significant.

Standalone version implementation

The standalone version of PhD7Faster 2.0 was developed with open source Qt 5.7 under the GPL & LGPLv3 licenses, which uses standard C++ for developing multiple-platform applications. Both graphical user interface (GUI) and command-line versions of PhD7Faster 2.0 were implemented. We provided different versions for Windows and Linux systems with little or no modification. All versions and source code are freely available at http://i.uestc.edu.cn/sarotup3/download.html.

RESULTS

Parameter optimization

Two important parameters: λ and w in Eq. (3) were necessary to be optimized before building the model. To obtain the best parameters, multiple experiments were performed according to the following standard:

$$\begin{cases} 1 \le \lambda \le 6 \text{ with step } \Delta = 1 \\ 0.05 \le w \le 0.70 \text{ with step } \Delta = 0.05 \end{cases} \tag{11}$$

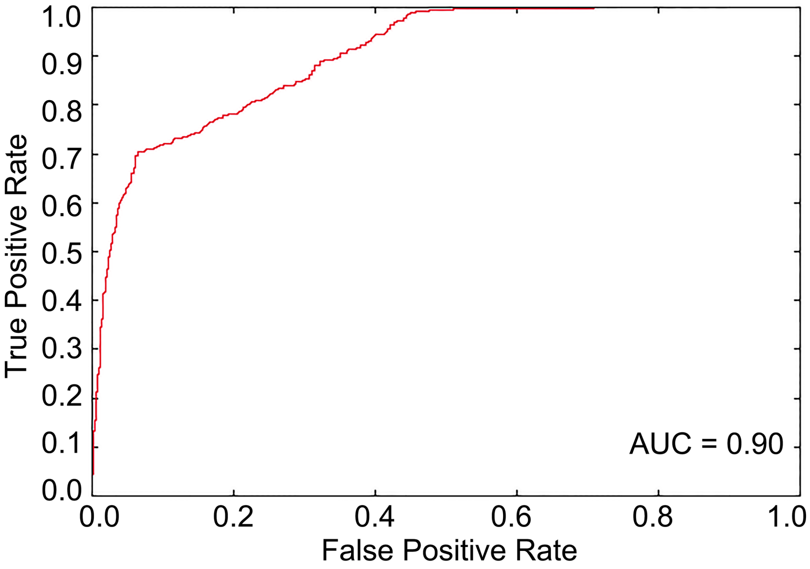


Figure 1 The ROC curve from the 10-fold cross-validation when *tp* is 0.5. The area under the ROC curve is about 0.90, which represents a decent prediction. Full-size DOI: 10.7717/peerj.7131/fig-1

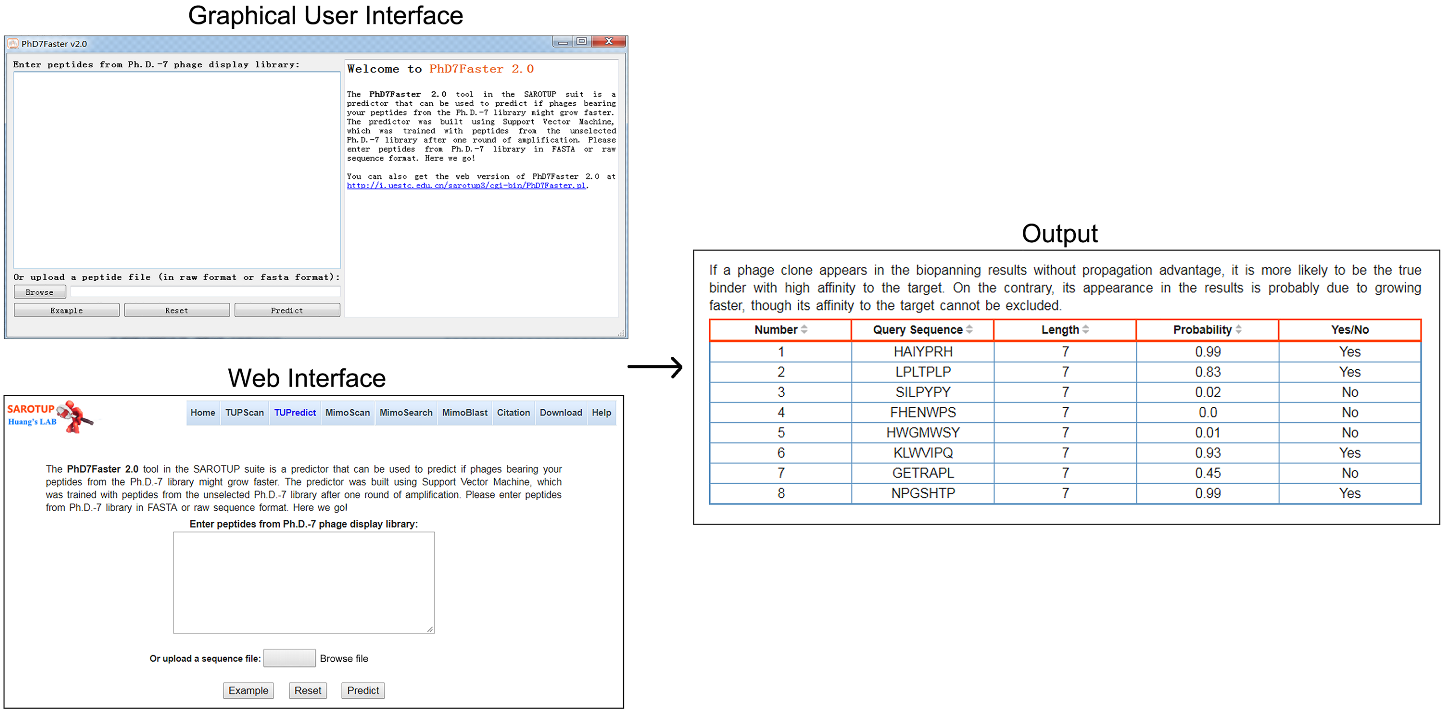
Thus, a total of 6 × 14 = 84 individual combinations were obtained. Then, we used the 10-fold cross-validation to investigate the accuracy of the model, which was built with SVM and the feature set without feature selection. $\lambda = 3$ and $w = 0.15$ produced the highest accuracy, which was considered as the best parameter combination.

Performance of PhD7Faster 2.0

The optimal feature subset with 644 features was determined through feature selection against 8,027 features including 8,000 tripeptide features and 27 PseAAC features. The SVM-based model was then trained with the optimal feature set. The results from the 10-fold cross-validation showed that the Acc of the predictive model was 81.84% with MCC of 0.64, Sn of 84.51%, and Sp of 79.17% when the threshold to distinguish between predicted positives and negatives (*tp*) was set to be 0.5. The ROC curve for model tuning is shown in Fig. 1, where the AUC is approximately 0.90. The permutation test resulted in a p-value of < 0.001. The above results indicated that PhD7Faster 2.0 achieved a promising performance.

Web and standalone versions of PhD7Faster 2.0

For the convenience of users, the SVM-based predictive model was implemented into a user-friendly web server, called PhD7Faster 2.0, which is freely available at http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl. The standalone GUI and command-line versions of PhD7Faster 2.0 for Windows and Linux systems were also provided. The interface, as well as the utilization of the GUI version, is remarkably similar to those of the web version (Fig. 2). A dataset with 20,000 peptides from the Ph.D.-7 phage display library was constructed (see testdataset.fasta in Supplementary Data). The standalone PhD7Faster 2.0 can complete analysis of the dataset within 60 s on a



Enter peptides from Ph.D.-7 phage display library:

Or upload a peptide file (in raw format or fasta format):

Welcome to PhD7Faster 2.0

The **PhD7Faster 2.0** tool in the SAROTUP suit is a predictor that can be used to predict if phages bearing your peptides from the Ph.D.-7 library might grow faster. The predictor was built using Support Vector Machine, which was trained with peptides from the unselected Ph.D.-7 library after one round of amplification. Please enter peptides from Ph.D.-7 library in FASTA or raw sequence format. Here we go!

You can also get the web version of PhD7Faster 2.0 at http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl.

The **PhD7Faster 2.0** tool in the SAROTUP suite is a predictor that can be used to predict if phages bearing your peptides from the Ph.D.-7 library might grow faster. The predictor was built using Support Vector Machine, which was trained with peptides from the unselected Ph.D.-7 library after one round of amplification. Please enter peptides from Ph.D.-7 library in FASTA or raw sequence format. Here we go!

Enter peptides from Ph.D.-7 phage display library:

Or upload a sequence file:

If a phage clone appears in the biopanning results without propagation advantage, it is more likely to be the true binder with high affinity to the target. On the contrary, its appearance in the results is probably due to growing faster, though its affinity to the target cannot be excluded.

Number	Query Sequence	Length	Probability	Yes/No
1	HAIYPRH	7	0.99	Yes
2	LPLTPLP	7	0.83	Yes
3	SILPYPY	7	0.02	No
4	FHENWPS	7	0.0	No
5	HWGMWSY	7	0.01	No
6	KLWVIPQ	7	0.93	Yes
7	GETRAPL	7	0.45	No
8	NPGSHTP	7	0.99	Yes

Figure 2 GUI version, web interface and output interface of PhD7Faster 2.0. The interface style of the GUI PhD7Faster 2.0 is consistent with that of the web server, which makes the tool user-friendly.

Full-size DOI: 10.7717/peerj.7131/fig-2

regular computer with Intel Core i3 Processor and 4GB RAM, which suggests that PhD7Faster 2.0 is highly efficient in processing massive datasets. PhD7Faster 2.0 was integrated into the SAROTUP 3.0 suite, which contains a series of computational tools to identify TUPs.

DISCUSSION

Comparison between PhDFaster 2.0 and 1.0

Parasitic sequences were identified significantly enriched in the amplified Ph.D.-7 phage display library by differential enrichment analysis of naïve and amplified Ph.D.-7 phage display libraries (*Matochko et al., 2014*). These parasitic peptides were grouped into the positive dataset of PhD7Faster 2.0. However, the positive dataset of PhD7Faster 1.0 was constructed based on threshold in copy numbers after one round of amplification in one replicate of sequencing data, irrespective of copy numbers in the naïve Ph.D.-7 library. Peptides with high abundances in both the naïve and amplified Ph.D.-7 libraries may also be selected as fast-growing sequences. Therefore, the positive training dataset of PhD7Faster 2.0 is more reliable than that of PhD7Faster 1.0.

PhD7Faster 2.0 was developed based on the combination of PseAAC and tripeptide composition, whereas only dipeptide composition was employed to build PhD7Faster 1.0. We also tried to use dipeptide composition to encode each peptide in the training dataset of PhD7Faster 2.0, but only 64% accuracy was obtained in the 10-fold cross-validation after feature selection. PseAAC coupled with tripeptide composition has been used in multiple

Table 1 **Comparison of performances of PhD7Faster 1.0 and 2.0.**

Method	Sn (%)	Sp (%)	Acc (%)	MCC
PhD7Faster 1.0	77.48	81.86	79.67	0.60
PhD7Faster 2.0	**84.51**	79.17	**81.84**	**0.64**

Note:
The measure of PhD7Faster 2.0 higher than that of PhD7Faster 1.0 is highlighted in bold.

protein prediction fields, such as predicting the subcellular localization of mycobacterial proteins (*Zhu et al., 2015*) and predicting apoptosis protein subcellular location (*Liao et al., 2011*). They contain more sequence-order information than dipeptide composition and hence can better reflect the feature of a peptide sequence. Thus, PhD7Faster 2.0 has 5% sensitivity, 2% accuracy and 0.04 MCC higher than PhD7Faster 1.0 (Table 1).

The standalone PhD7Faster 2.0 is empowered to identify PrTUPs within output of conventional phage display as well as large next-generation sequencing data, whereas PhD7Faster 1.0 can only work with small-scale data sets (several hundreds of peptides). This important improvement makes PhD7Faster 2.0 as an enhanced and powerful tool for scanning and reporting PrTUPs from the Ph.D.-7 phage display library. The emergence of PhD7Faster 2.0 highlights the significance of high throughput sequencing of different types of phage display libraries and developing bioinformatics tools for identifying PrTUPs from these libraries.

PhD7Faster 2.0 cannot predict the censorship in the Ph.D. libraries

It is possible that some peptides are likely to be censored from being displayed on the phage in the first place. The censorship of positively charged amino acids has been reported since these residues suppress proper insertion of pIII into the inner membrane of *Escherichia coli*, thus decreasing the efficiency of the assembly and extrusion of phage clones (*Peters et al., 1994*). *Rodi, Soares & Makowski (2002)* also observed that peptides of α-helix or β-sheet conformations were censored in Ph.D.-12 and Ph.D.-C7C libraries. *Steiner et al. (2006)* have shown that maturely folded proteins are displayed poorly via the Sec translocation pathway. However, this censorship is a completely different phenomenon from that of phage growing faster. Therefore, PhD7Faster 2.0 is not able to predict this censorship.

PhD7Faster 2.0 predict PrTUPs in the Ph.D.-7 library

The PrTUPs have significantly higher proliferation rates than normal-growing phage and are favored during the amplification steps. The proliferation advantage of some PrTUPs have been verified to be intrinsic to mutations in the 5′-untranslated region (UTR) of gene II in M13 phage (*Brammer et al., 2008*; *Nguyen et al., 2014*; *Zygiel et al., 2017*). *Zygiel et al. (2017)* have also described the likelihood that these mutations compensate for the replication defect afforded by the lacZα insert present in the M13 bacteriophage-based vector upon which the Ph.D.-7 (and Ph.D.-12) library was based. Thus, the particular peptide displayed (e.g., HAIYPRH, GKPMPPM, AKIDART) is merely a stowaway on a clone that propagates fast due to its gene II 5′-UTR mutation(s). In these clones, the peptide itself is completely arbitrary; it just happens to be the peptide displayed on a clone that picked up a mutation prior to or during library construction. As these mutations in the phage genome are unrelated to the displayed peptide, PhD7Faster 2.0 may not be able to predict this type of PrTUPs in the Ph.D.-7 library. In addition, Smith et al. indicated that the enhanced propagation rate of some PrTUPs may be due to the displayed peptide (*Thomas, Golomb & Smith, 2010*), and PhD7Faster 2.0 can be used to predict this type of PrTUPs in the Ph.D.-7 library. However, no direct evidence supports that displayed peptides allow the phage to propagate faster, and the biological mechanism remains to be further examined.

CONCLUSION

In this report, we propose an SVM-based tool, PhD7Faster 2.0, for predicting clones growing faster from the Ph.D.-7 phage display library. Ten-fold cross-validation results show that PhD7Faster 2.0 achieves an accuracy of 81.84% with 0.64 MCC and 0.90 AUC. The standalone version of the tool was also developed, which can predict PrTUPs within both traditional biopanning data and next generation phage display data. We also implemented a web-server for the proposed method, which can be freely accessible from http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which will lead to the improvement of this paper.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (grant number 61571095), the Fundamental Research Funds for the Central Universities of China (ZYGX2005Z006), the Sichuan Science and Technology Program (2018HH0154), the 2018 Talent Research Program of Guizhou University (702569183301 and 702570183301), the Science and Technology Plan Project of Guizhou Province of China ((2018)5781), and the China Postdoctoral Science Foundation Grant (Grant No. 2019M653369). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 61571095.
Fundamental Research Funds for the Central Universities of China: ZYGX2005Z006.
Sichuan Science and Technology Program: 2018HH0154.
2018 Talent Research Program of Guizhou University: 702569183301 and 702570183301.
Science and Technology Plan Project of Guizhou Province of China: (2018)5781.
China Postdoctoral Science Foundation Grant: 2019M653369.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Bifang He conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Heng Chen contributed reagents/materials/analysis tools, approved the final draft.
- Jian Huang conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data generated or analyzed during this study are available in the Supplementary Data. The standalone GUI is at http://i.uestc.edu.cn/sarotup3/cgi-bin/PhD7Faster.pl and the source code can be downloaded from http://i.uestc.edu.cn/sarotup3/download.html.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7131#supplemental-information.

REFERENCES

Bakhshinejad B, Zade HM, Shekarabi HS, Neman S. 2016. Phage display biopanning and isolation of target-unrelated peptides: in search of nonspecific binders hidden in a combinatorial library. *Amino Acids* **48(12)**:2699–2716 DOI 10.1007/s00726-016-2329-6.

Brammer LA, Bolduc B, Kass JL, Felice KM, Noren CJ, Hall MF. 2008. A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site. *Analytical Biochemistry* **373(1)**:88–98 DOI 10.1016/j.ab.2007.10.015.

Causa F, Della Moglie R, Iaccino E, Mimmi S, Marasco D, Scognamiglio PL, Battista E, Palmieri C, Cosenza C, Sanguigno L, Quinto I, Scala G, Netti PA. 2013. Evolutionary screening and adsorption behavior of engineered M13 bacteriophage and derived dodecapeptide for selective decoration of gold interfaces. *Journal of Colloid and Interface Science* **389(1)**:220–229 DOI 10.1016/j.jcis.2012.08.046.

Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2(3)**:27 DOI 10.1145/1961189.1961199.

Chou K-C. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43(3)**:246–255 DOI 10.1002/prot.1035.

Chou K-C. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **21(1)**:10–19 DOI 10.1093/bioinformatics/bth466.

Chou K-C. 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* **273(1)**:236–247 DOI 10.1016/j.jtbi.2010.12.024.

Christiansen A, Kringelum JV, Hansen CS, Bogh KL, Sullivan E, Patel J, Rigby NM, Eiwegger T, Szepfalusi Z, De Masi F, Nielsen M, Lund O, Dufva M. 2015. High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Scientific Reports* **5(1)**:12913 DOI 10.1038/srep12913.

He B, Chai G, Duan Y, Yan Z, Qiu L, Zhang H, Liu Z, He Q, Han K, Ru B, Guo FB, Ding H, Lin H, Wang X, Rao N, Zhou P, Huang J. 2016a. BDB: biopanning data bank. *Nucleic Acids Research* **44(D1)**:D1127–D1132 DOI 10.1093/nar/gkv1100.

He B, Jiang L, Duan Y, Chai G, Fang Y, Kang J, Yu M, Li N, Tang Z, Yao P, Wu P, Derda R, Huang J. 2018. Biopanning data bank 2018: hugging next generation phage display. *Database (Oxford)* **2018**:bay032 DOI 10.1093/database/bay032.

He B, Kang J, Ru B, Ding H, Zhou P, Huang J. 2016b. SABinder: a web service for predicting streptavidin-binding peptides. *BioMed Research International* **2016**:9175143 DOI 10.1155/2016/9175143.

He B, Mao C, Ru B, Han H, Zhou P, Huang J. 2013. Epitope mapping of metuximab on CD147 using phage display and molecular docking. *Computational and Mathematical Methods in Medicine* **2013**:983829 DOI 10.1155/2013/983829.

Huang J, Ru B, Li S, Lin H, Guo FB. 2010. SAROTUP: scanner and reporter of target-unrelated peptides. *Journal of Biomedicine and Biotechnology* **2010**:101932 DOI 10.1155/2010/101932.

Huang J, Ru B, Zhu P, Nie F, Yang J, Wang X, Dai P, Lin H, Guo FB, Rao N. 2012. MimoDB 2.0: a mimotope database and beyond. *Nucleic Acids Research* **40(D1)**:D271–D277 DOI 10.1093/nar/gkr922.

Hung LY, Fu CY, Wang CH, Chuang YJ, Tsai YC, Lo YL, Hsu PH, Chang HY, Shiesh SC, Hsu KF, Lee GB. 2018. Microfluidic platforms for rapid screening of cancer affinity reagents by using tissue samples. *Biomicrofluidics* **12(5)**:054108 DOI 10.1063/1.5050451.

Kang J, Fang Y, Yao P, Li N, Tang Q, Huang J. 2018. NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition. *Interdisciplinary Sciences: Computational Life Sciences* **11(1)**:108–114 DOI 10.1007/s12539-018-0287-2.

Li N, Kang J, Jiang L, He B, Lin H, Huang J. 2017. PSBinder: a web service for predicting polystyrene surface-binding peptides. *BioMed Research International* **2017**:5761517 DOI 10.1155/2017/5761517.

Liao B, Jiang J-B, Zeng Q-G, Zhu W. 2011. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. *Protein & Peptide Letters* **18(11)**:1086–1092 DOI 10.2174/092986611797200931.

Lin H, Ding C, Yuan L-F, Chen W, Ding H, Li Z-Q, Guo F-B, Huang J, Rao N-N. 2013. Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: approached from optimal tripeptide composition. *International Journal of Biomathematics* **6(2)**:1350003 DOI 10.1142/S1793524513500034.

Mandava S, Makowski L, Devarapalli S, Uzubell J, Rodi DJ. 2004. RELIC—a bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites. *Proteomics* **4(5)**:1439–1460 DOI 10.1002/pmic.200300680.

Martins IM, Reis RL, Azevedo HS. 2016. Phage display technology in biomaterials engineering: progress and opportunities for applications in regenerative medicine. *ACS Chemical Biology* **11(11)**:2962–2980 DOI 10.1021/acschembio.5b00717.

Matochko WL, Cory Li S, Tang SK, Derda R. 2014. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Research* **42(3)**:1784–1798 DOI 10.1093/nar/gkt1104.

Menendez A, Scott JK. 2005. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Analytical Biochemistry* **336(2)**:145–157 DOI 10.1016/j.ab.2004.09.048.

Ngubane NA, Gresh L, Ioerger TR, Sacchettini JC, Zhang YJ, Rubin EJ, Pym A, Khati M. 2013. High-throughput sequencing enhanced phage display identifies peptides that bind mycobacteria. *PLOS ONE* **8(11)**:e77844 DOI 10.1371/journal.pone.0077844.

Nguyen KT, Adamkiewicz MA, Hebert LE, Zygiel EM, Boyle HR, Martone CM, Melendez-Rios CB, Noren KA, Noren CJ, Hall MF. 2014. Identification and characterization of mutant clones with enhanced propagation rates from phage-displayed peptide libraries. *Analytical Biochemistry* **462**:35–43 DOI 10.1016/j.ab.2014.06.007.

Pande J, Szewczyk MM, Grover AK. 2010. Phage display: concept, innovations, applications and future. *Biotechnology Advances* **28(6)**:849–858 DOI 10.1016/j.biotechadv.2010.07.004.

Peters EA, Schatz PJ, Johnson SS, Dower WJ. 1994. Membrane insertion defects caused by positive charges in the early mature region of protein pIII of filamentous phage fd can be corrected by prlA suppressors. *Journal of Bacteriology* **176(14)**:4296–4305 DOI 10.1128/jb.176.14.4296-4305.1994.

Rentero Rebollo I, Sabisz M, Baeriswyl V, Heinis C. 2014. Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Research* **42(22)**:e169 DOI 10.1093/nar/gku940.

Rodi DJ, Soares AS, Makowski L. 2002. Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries. *Journal of Molecular Biology* **322(5)**:1039–1052 DOI 10.1016/S0022-2836(02)00844-6.

Ru B, Huang J, Dai P, Li S, Xia Z, Ding H, Lin H, Guo F, Wang X. 2010. MimoDB: a new repository for mimotope data derived from phage display technology. *Molecules* **15(11)**:8279–8288 DOI 10.3390/molecules15118279.

Ru B, 't Hoen PAC, Nie F, Lin H, Guo FB, Huang J. 2014. PhD7Faster: predicting clones propagating faster from the Ph.D.-7 phage display peptide library. *Journal of Bioinformatics and Computational Biology* **12(1)**:1450005 DOI 10.1142/S021972001450005X.

Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R. 2007. PepBank—a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics* **8(1)**:280 DOI 10.1186/1471-2105-8-280.

Steiner D, Forrer P, Stumpp MT, Pluckthun A. 2006. Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. *Nature Biotechnology* **24(7)**:823–831 DOI 10.1038/nbt1218.

Su Z-D, Huang Y, Zhang Z-Y, Zhao Y-W, Wang D, Chen W, Chou K-C, Lin H. 2018. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* **34(24)**:4196–4204 DOI 10.1093/bioinformatics/bty508.

Swaminathan S, Cui Y. 2013. Recognition of epoxy with phage displayed peptides. *Materials Science and Engineering: C* **33(5)**:3082–3084 DOI 10.1016/j.msec.2013.02.011.

't Hoen PA, Jirka SM, Ten Broeke BR, Schultes EA, Aguilera B, Pang KH, Heemskerk H, Aartsma-Rus A, Van Ommen GJ, Den Dunnen JT. 2012. Phage display screening without repetitious selection rounds. *Analytical Biochemistry* **421(2)**:622–631 DOI 10.1016/j.ab.2011.11.005.

Tang H, Chen W, Lin H. 2016. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Molecular BioSystems* **12(4)**:1269–1275 DOI 10.1039/c5mb00883b.

Thomas WD, Golomb M, Smith GP. 2010. Corruption of phage display libraries by target-unrelated clones: diagnosis and countermeasures. *Analytical Biochemistry* **407(2)**:237–240 DOI 10.1016/j.ab.2010.07.037.

Vodnik M, Strukelj B, Lunder M. 2012. HWGMWSY, an unanticipated polystyrene binding peptide from random phage display libraries. *Analytical Biochemistry* **424(2)**:83–86 DOI 10.1016/j.ab.2012.02.013.

Vodnik M, Zager U, Strukelj B, Lunder M. 2011. Phage display: selecting straws instead of a needle from a haystack. *Molecules* **16(1)**:790–817 DOI 10.3390/molecules16010790.

Zade HM, Keshavarz R, Shekarabi HSZ, Bakhshinejad B. 2017. Biased selection of propagation-related TUPs from phage display peptide libraries. *Amino Acids* **49(8)**:1293–1308 DOI 10.1007/s00726-017-2452-z.

Zhang Y, He B, Liu K, Ning L, Luo D, Xu K, Zhu W, Wu Z, Huang J, Xu X. 2017. A novel peptide specifically binding to VEGF receptor suppresses angiogenesis in vitro and in vivo. *Signal Transduction and Targeted Therapy* **2**:17010 DOI 10.1038/sigtrans.2017.10.

Zhao Y-W, Lai H-Y, Tang H, Chen W, Lin H. 2016. Prediction of phosphothreonine sites in human proteins by fusing different features. *Scientific Reports* **6(1)**:34817 DOI 10.1038/srep34817.

Zhu P-P, Li W-C, Zhong Z-J, Deng E-Z, Ding H, Chen W, Lin H. 2015. Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Molecular BioSystems* **11(2)**:558–563 DOI 10.1039/c4mb00645c.

Zygiel EM, Noren KA, Adamkiewicz MA, Aprile RJ, Bowditch HK, Carroll CL, Cerezo MAS, Dagher AM, Hebert CR, Hebert LE, Mahame GM, Milne SC, Silvestri KM, Sutherland SE, Sylvia AM, Taveira CN, VanValkenburgh DJ, Noren CJ, Hall MF. 2017. Various mutations compensate for a deleterious lacZalpha insert in the replication enhancer of M13 bacteriophage. *PLOS ONE* **12(4)**:e0176421 DOI 10.1371/journal.pone.0176421.