Journal of Bioinformatics and Computational Biology

PHD7FASTER: PREDICTING CLONES PROPAGATING FASTER FROM THE PH.D.-7 PHAGE DISPLAY PEPTIDE LIBRARY

BEIBEI RU[*,#], PETER A.C. 'T HOEN[†,$], FULEI NIE[*,%], HAO LIN[*, &],
FENG-BIAO GUO[*,¶] and JIAN HUANG[*,||, §]

**Center of Bioinformatics (COBI)*
Key Laboratory for NeuroInformation of Ministry of Education
University of Electronic Science and Technology of China
Chengdu 610054, P. R. China

†Center for Human and Clinical Genetics
Leiden University Medical Center
2300 RC Leiden, The Netherlands

[#]rubeibei1988@gmail.com
[$]p.a.c.hoen@lumc.nl
[%]niefulei@126.com
[&]hlin@uestc.edu.cn
[¶]fbguo@uestc.edu.cn
[||]hj@uestc.edu.cn

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Phage display can rapidly discover peptides binding to any given target; thus, it has been widely used in basic and applied research. Each round of panning consists of two basic processes: selection and amplification. However, recent studies have showed that the amplification step would decrease the diversity of phage display library due to different propagation capacity of phage clones. This may induce phages with growth advantage rather than specific affinity to appear in the final experimental results. The peptides displayed by such phages are termed propagation-related target-unrelated peptides (PrTUPs). They would mislead further analysis and research if not removed. In this paper, we describe PhD7Faster, an ensemble predictor based on support vector machine (SVM) for predicting clones with growth advantage from the Ph.D.-7 phage display peptide library. By using reduced dipeptide composition (ReDPC) as features, an accuracy (*Acc*) of 79.67% and a Matthews correlation coefficient (*MCC*) of 0.595 were achieved in 5-fold cross-validation. In addition, the SVM-based model was demonstrated to perform better than several representative machine learning algorithms. We anticipate that PhD7Faster can assist biologists to exclude potential PrTUPs and accelerate the finding of specific binders from the popular Ph.D.-7 library. The web server of PhD7Faster can be freely accessed at http://immunet.cn/sarotup/cgi-bin/PhD7Faster.pl.

Keywords: Phage display; target-unrelated peptides; PrTUPs; support vector machine; reduced dipeptide composition.

[§] Corresponding author.

1. Introduction

Phage display is a powerful technique in the identification of ligands for various targets, ranging from small molecules to whole organisms.[1, 2] Panning of phage display libraries has two essential processes. One is called the selection step, which enriches clones binding to the desired target. The other is named the amplification step, which aims to multiple the selected clones and form a secondary library. In general, a naive phage library can be narrowed to a subpopulation of binders with target-specific affinity after several rounds of panning. Then ligands of the target could be obtained through sequencing a limited number of clones, which are randomly picked out from the final library.[2] Due to its convenience, low cost, and high efficiency, phage display has been exploited to study the sites and networks of protein-protein interaction,[3-5] as well as to develop new diagnostics, therapeutics and vaccines.[6-10]

Surprisingly, it has been shown that the amplification process would also decrease the diversity of libraries.[11] Using Illumina deep-sequencing technology, Derda and co-workers found that a 10^6-scale phage library might collapse to hundreds of abundant sequences after a single round of growth in bacteria.[12] This is because phages containing distinct inserted DNA sequences possess different growth rates. According to simulations and experiments, subtle differences in growth rate could lead to big differences in clone abundances after rounds of amplification.[11] Some promising clones with affinity to the target may disappear during the amplification of phages in bacteria due to their weak growth capacities. What's worse, the final phage library might has many or even be predominated by phages propagating faster after multiple rounds of panning. Therefore, there are phages in the experimental results whose appearances are due to their growth advantage rather than specific affinity to target. The peptides displayed by such phages are called propagation-related target-unrelated peptides (PrTUPs). [13, 14] As a kind of hidden false positive hits of the phage display experiment, PrTUPs can disturb or even mislead further analysis and research if they are not removed.

Since phage display was invented in 1985, researchers have attempted to take measures to avoid the loss of library diversity and the contamination of false positive hits. Via disrupting the minus-strand replication origin by a tetracycline resistance cassette, phage display libraries based on fd-tet-derived vectors are inherently resistant to corruption by PrTUPs.[15, 16] Given the fact that amplification is the step leading to loss of diversity, some have tried to perform only one round of panning without amplification.[17, 18] This strategy is empowered by next generation sequencing (NGS) technique, which prompts to find specific binders and restrain false positive hits.[19] Others have conducted the amplification step in isolated compartments because it has been reported that phage competition occurred due to different production rates rather than total numbers of phage produced.[20, 21]

In comparison to experimental methods, computational tools may be an alternative but more convenient way to exclude PrTUPs. The program INFO calculates information content of each sequence to infer clones with growth preference.[22] TUPScan in the SAROTUP suite is capable of checking if the input peptides match the known TUP

patterns or highly suspected PrTUPs.[23] Depending on MimoDB, a database with lots of panning results and relevant background information, the MimoSearch and the MimoBlast tools can be used to find identical or highly similar peptides obtained with various targets. Some of these peptides may be selected just because of growth advantage.[24, 25]

Although the above measures can partly mitigate the disturbance from PrTUPs, they are not adequate enough in phage display field. For example, fd-tet-derived libraries can be corrupted by a new type of PrTUPs, which has a complex rearrangement that restores the minus-strand origin with retaining tetracycline resistance.[26] Amplifying phages in isolated compartments is tedious, time-consuming and not suitable for screening a library. In addition, the programs mentioned above are merely sequence analysis tools for phage-displayed peptides. They are not real predictors although they can assist researchers to identify some PrTUPs.[27]

In this study, we developed an ensemble predictor called PhD7Faster. It is based on support vector machine (SVM) and can predict clones propagating faster from the Ph.D.-7 phage display peptide library. Using reduced dipeptide composition (ReDPC), it achieved the best performance with an accuracy (*Acc*) of 79.67% and a Matthews correlation coefficient (*MCC*) of 0.595.

2. Materials and Methods

2.1. *Data sets*

The training data sets were generated from a research article published recently by ‘t Hoen *et al*.[19] They sequenced about seven millions of phage clones of the Ph.D.-7 phage display peptide library after a single round of amplification. For quality control, a custom Perl script was employed to retrieve sequences matching the pattern $(NNG/T)_7GGTGGA$, in which NNG/T is the coding scheme of displayed peptides and GGTGGA is the expected 6-nucleotide sequence following the insert. About 89% of the sequences remained after this filter step. Then the inserted DNA sequences were translated into amino acid sequence using conventional amino acid codon table except that TAG is translated as Q. This is because the amber stop codon TAG is suppressed by Glutamine (Q) in the strain used to propagate the library.

According to the manufacturer, the naive Ph.D.7 phage display peptide library with titer of 10^{13} pfu/ml contains up to 10^9 different sequences. By sequencing millions of clones, only a fraction of the entire library was analyzed. However, some peptides were found at high copy number than expected by chance. Based on Poisson distribution, the probability that a peptide appears 15 or more times in the sequenced results is 2.0×10^{-50} or less. Nevertheless, 151 peptide sequences were found at a copy number of 15 or higher, suggesting they had a significant growth preference. These peptides were collected into the positive data set. The negative data set was composed of the peptides appearing only once. However, the two data sets are extremely unbalanced: 151 peptides with growth advantage and 2,103,076 peptides without growth advantage. Down-sampling strategy was introduced to deal with the imbalanced data sets, that is, 151 peptides were randomly

chosen from the negative data set. In order to reduce random error, this process was repeated ten times. As a result, ten pairs of sub-datasets were obtained and each pair contained 151 peptides with or without growth advantage.

2.2. *Feature encoding schemes*

Extraction of a set of informative features plays a crucial role in pattern recognition. For constructing the best performance model, four types of feature encoding schemes were utilized to characterize individual peptide sequence in data sets. Amino acid composition (AAC) and dipeptide composition (DPC) have achieved accepted performances in the areas of protein bioinformatics.[28, 29] In the binary code scheme, each amino acid is represented by a 20-dimensional vector with 1 at only a specific position and 0 at all else places. Thus, a vector with 140 dimensions was used to encode a 7-mer peptide. The Bayes Feature Extraction was performed in a bi-profile manner, i.e., positive position-specific and negative position-specific profiles. These profiles were produced through calculating the frequency of each residue at each position of the peptide sequence in the positive data set and the negative data set, respectively.[30] Therefore, a 7-mer peptide was encoded by a 14-dimensional feature vector containing information on amino acid in the positive and negative spaces.

To select the optimal reduced subsets from the composition feature schemes, we conducted the following feature selection steps against AAC and DPC: (i) calculated the accuracy of each feature; (ii) added a feature to an initial null feature combination in descending order by accuracy sequentially and calculated the accuracy of each feature combination; (iii) selected the combination with the highest accuracy as the optimal reduced subset.

2.3. *Support vector machine*

Support vector machine (SVM) is a wonderful machine learning method based on statistical learning theory, which has been widely utilized in classification. The basic idea of SVM is to map the input samples onto a high dimensional space through kernel function and then to seek a separating hyperplane in this space. In this report, we used the software LibSVM3.11 to implement SVM, which is developed by Lin's lab and can be freely downloaded from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.[31] With regard to four kinds of kernel functions, preliminary examination suggested that the radial basis function (RBF) outperformed the linear function, the polynomial function and the sigmoid function. To achieve the highest training accuracy, the grid search script was applied to optimize the penalty parameter C and the kernel parameter γ.

2.4. *Performance assessments*

Five-fold cross-validation was adopted to assess the performance of the prediction model. Briefly, an original sample was randomly partitioned into 5 equal subsamples. Each single subsample was in turn used as the test data, and the remaining 4 subsamples were

retained as training data. In order to provide indicators of prediction performance, four measures: sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*) and Matthews correlation coefficient (*MCC*), were exploited and defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{FP + TN} \tag{2}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

In these equations, TP denotes the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. *MCC*, ranging from -1 to 1, is a measure of the quality of binary classifications. A value of 1 indicates a perfect prediction, 0 same as random and -1 an absolute negative correlation between observation and prediction.

2.5. *Constructing an ensemble predictor*

Ten basic predictors were built on corresponding SVM-based models trained with the 10 pairs of sub-datasets using reduced dipeptide composition (ReDPC). To abate the variance caused by any single predictor, an ensemble prediction model was constructed through voting. As shown in Fig. 1, any input was predicted by the ten basic models independently. Combining the results of the 10 models together, a peptide will be predicted to be with growth advantage in final if the average possibility was 0.5 or higher. For the convenience of scientists, we implemented the ensemble predictor as a web program called PhD7Faster.

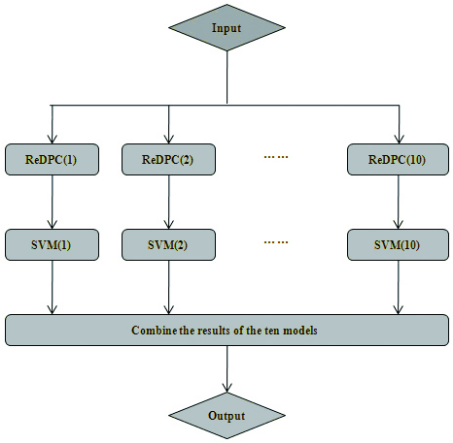

Fig. 1. Flowchart of PhD7Faster: an ensemble predictor.

3. Results and Discussion

3.1. *Sequence analysis*

The mechanism that leads to a quantity of phages propagating faster is not currently clear. We attempted to identify common motifs in sequences with growth advantage using multiple unique sequence identifier software (MUSI).[32] The results showed that ten distinct sequence patterns were identified in the 151 peptides with growth advantage. The sequence logos and proportions corresponding to each cluster were shown in Fig. 2. Remarkably, proline (P) was the most represented amino acid, appearing as a consensus amino acid in 8 out of 10 patterns. Lysine (L), serine (S) and threonine (T) belonged to the second richest amino acid and were adjacent to proline in most cases. Moreover, some amino acid pairs are also found to be overrepresented. For instance, amino acid pair PP is represented in the motifs of logo #2 and #7; LP is represented in the motifs of logo #2, #3, #7 and #9; PL is represented in the motifs of logo #7; SP is represented in the motifs of logo #4 and #6; PS is represented in the motifs of logo #1 and #9; TP is represented in the motifs of logo #5; PT is represented in the motifs of logo #3, #5 and #6.

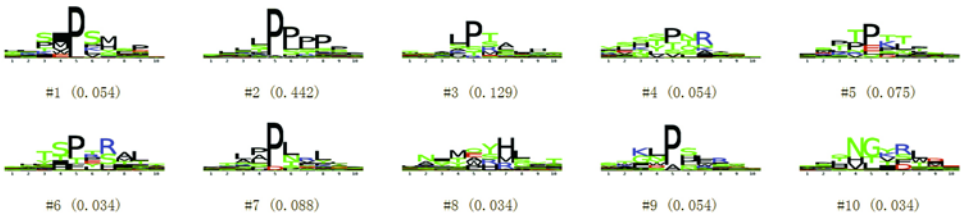

Fig. 2. Clustering analysis of the 151 peptides with growth advantage. The number in parentheses represents the sequence proportion of this cluster.

3.2. *Prediction performances of SVM with different features*

Through training SVM with different feature coding schemes, their average values of *Sn*, *Sp*, *Acc* and *MCC* based on ten pairs of sub-datasets were shown in Table 1. In terms of these four measures, Bayes Feature Extraction encoding schemes tended to produce slightly better models than the others, i.e., AAC, DPC and binary code. Subsequently, we executed feature selection technique to ascertain a subset of most informative features from the composition encoding schemes. Significant improvements were attained for AAC and DPC after reducing some redundant and irrelevant compositions. As the accuracy increased from 72.32% to 79.67% and the *MCC* increased from 0.480 to 0.595, the reduced dipeptide composition (ReDPC) even outperformed the Bayes Feature Extraction. Besides, we also found that the 400-dimensional vector was reduced to about 80-dimensional in ten pairs of sub-datasets. Amino acid pairs LP, PL, PP, PS, PT, RL, SP and TP appeared in all reduced dipeptide composition of ten sub-datasets. Interestedly, these amino acid pairs were also shown in the core of motifs detected by MUSI (see the section “Sequence analysis”).

Table 1. The prediction performances of SVM trained with different features.

Feature coding scheme	***Sn*(%)**	***Sp*(%)**	***Acc*(%)**	***MCC***
Amino acid composition (AAC)	62.52	84.64	73.58	0.491
Reduced amino acid composition (ReAAC)	66.29	83.97	75.13	0.514
Dipeptide composition (DPC)	54.90	89.74	72.32	0.480
Reduced dipeptide composition (ReDPC)	77.48	81.86	79.67	0.595
Binary code	53.97	90.26	72.12	0.476
Bayes Feature Extraction	65.96	84.97	75.46	0.522

3.3. *Prediction performances of different machine learning methods*

Based on the feature encoding schemes of Bayes Feature Extraction and ReDPC, we compared performance of our SVM based models to a number of representative machine learning algorithms. WEKA[33] was exploited to execute the predictions using different methods such as Naive Bayes, Random Forest, and Logistic function. As shown in Table 2, the SVM-based model using reduced dipeptide composition features showed best performance for prediction of peptides with growth advantage among current machine learning methods.

Table 2. The prediction performances of different machine learning methods

Methods	**Bayes Feature Extraction**		**Reduced dipeptide composition**	
	***Acc*(%)**	***MCC***	***Acc*(%)**	***MCC***
SVM	75.46	0.522	79.67	0.595
Naive Bayes	68.91	0.381	73.58	0.489
Random Forest	69.40	0.389	69.83	0.407
Decision Tree J48	67.52	0.350	69.11	0.389
RBF network	68.44	0.383	70.33	0.431
Logistic Function	71.03	0.421	70.50	0.411

3.4. *Construction of ensemble predictor*

According to the above analysis, we constructed an ensemble predictor based on SVM using ReDPC. Each basic prediction model was trained with the ten pairs of sub-datasets correspondingly. Then the ten models were integrated as an ensemble predictor via voting strategy. The ensemble predictor has been developed into a web program called PhD7Faster (http://immunet.cn/sarotup/cgi-bin/PhD7Faster.pl). Users just need to input their peptide sequences and then would get a result table with prediction details. It should be noted that the input sequences should come from the Ph.D.-7 library since the tool was trained with data from this library. With the accumulation of deep sequencing data from other libraries, we will strive to develop more predictors for identifying clones with growth advantage.

3.5. *Evaluation of PhD7Faster*

To our knowledge, there is no other next generation sequencing data on the Ph.D.-7 library at present. Therefore, an independent and comprehensive data set for evaluating the program is not available currently. However, we have tested the server using the negative data set without the 1,510 training peptides. The result showed that 31.61% of sequences were predicted to have growth advantage although they appeared only once in the approximate 7 million clones sequenced. As the false positive rate (FPR) of PhD7Faster is 18.14%, there are still 13.47% positive hits which cannot be explained simply with FPR. We infer that at least some of them do grow faster. As we mentioned previously, the primary Ph.D.-7 library is with a titer of 10^{13} pfu/ml. Thus, only 10^{-7} of the whole library was studied even if a sample at a 10^{6} scale was sequenced in this case. It is possible some peptides appeared once in this study are due to inadequate sampling.

Statistical analysis showed that 57% of unique sequences appeared only once while the remaining 43% were found twice or more times. Presumably, more copy numbers may mean a higher growth rate. To evaluate PhD7Faster further, we grouped all the qualified sequencing data based on their appearing times, i.e. copy numbers. Except the negative (only once) and the positive (15 or more times) data that have been used to train the model, the growth advantages of all peptides with a copy number from 2 to 14 were also predicted with PhD7Faster. For each group, the positive rate (i.e. ratio of peptides with growth preference to all peptides of this group) was calculated. As shown in Figure 3, there is a significant positive correlation between the copy number and the positive rate, indicating that a peptide appearing more times after amplification will more likely to be with growth advantage. The results given by PhD7Faster satisfy our assumption well, reflecting its power and reliability.

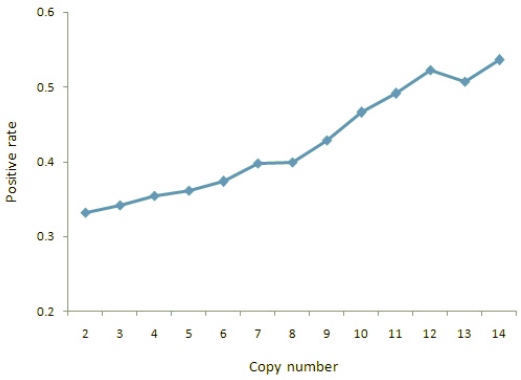

Fig. 3. The relationship between copy number and positive rate

In addition, we investigated the prediction results of PhD7Faster using all available and suitable peptides taken from the MimoDB database v4.0 as input.[24, 25] Among 311 sets of peptides from the Ph.D.-7 phage display peptide library, 3152 non-redundant

peptides were left after removing sequences with non-standard amino acids. Among them, 34.96% were predicted to be with growth preference. As most of these peptides were panned with protocols having 2 or more rounds of amplifications, it is not strange they had a little higher positive rate. The results also conform to the known facts that the Ph.D.-7 library is vulnerable to PrTUPs. Though some peptides may have target-specific affinity and growth advantage simultaneously, using PhD7Faster combined with affinity assay such as phage-ELISA will help biologists to distinguish them from PrTUPs.

4. Conclusions

The Ph.D.-7 phage display peptide library is one of the most widely used commercial libraries. It is also known for being vulnerable to PrTUPs. In this study, an ensemble predictor based on SVM employing the reduced dipeptide composition was developed to predict clones with growth advantage from the Ph.D.-7 phage display peptide library. It has been implemented as the PhD7Faster web program, which is freely accessible at http://immunet.cn/sarotup/cgi-bin/PhD7Faster.pl. This tool has an accuracy of 79.67% and *MCC* of 0.595 in 5-fold cross-validation, which is found to perform better than several other machine learning methods. Our study has provided new insights into the TUPs with growth advantage and the program may assist biologists to exclude possible PrTUPs from the Ph.D.-7 phage display peptide library.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported in part by the National Natural Science Foundation of China under the Grant 61071177 and the Program for New Century Excellent Talents in University (NCET-12-0088).

References

1. Smith GP, Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface, *Science* **228**(4705):1315-1317, 1985.
2. Smith GP, Petrenko VA, Phage Display, *Chem Rev* **97**(2):391-410, 1997.
3. Scott JK, Smith GP, Searching for peptide ligands with an epitope library, *Science* **249**(4967):386-390, 1990.
4. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CW, Fields S, Boone C, Cesareni G, A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules, *Science* **295**(5553):321-324, 2002.
5. Thom G, Cockroft AC, Buchanan AG, Candotti CJ, Cohen ES, Lowne D, Monk P, Shorrock-Hart CP, Jermutus L, Minter RR, Probing a protein-protein interaction by in vitro evolution, *Proc Natl Acad Sci U S A* **103**(20):7619-7624, 2006.
6. Riemer AB, Jensen-Jarolim E, Mimotope vaccines: epitope mimics induce anti-cancer antibodies, *Immunol Lett* **113**(1):1-5, 2007.
7. Hsiung PL, Hardy J, Friedland S, Soetikno R, Du CB, Wu AP, Sahbaie P, Crawford JM, Lowe AW, Contag CH, Wang TD, Detection of colonic dysplasia in vivo using a targeted heptapeptide and confocal microendoscopy, *Nat Med* **14**(4):454-458, 2008.

8. Ellis SE, Newlands GF, Nisbet AJ, Matthews JB, Phage-display library biopanning as a novel approach to identifying nematode vaccine antigens, *Parasite Immunol* **34**(5):285-295, 2012.
9. Moyer MW, New blood-boosting drugs aim to staunch renal anemia, *Nat Med* **18**(3):332, 2012.
10. Chan VS, Tsang HH, Tam RC, Lu L, Lau CS, B-cell-targeted therapies in systemic lupus erythematosus, *Cell Mol Immunol* **10**(2):133-142, 2013.
11. Derda R, Tang SK, Li SC, Ng S, Matochko W, Jafari MR, Diversity of phage-displayed libraries of peptides during panning and amplification, *Molecules* **16**(2):1776-1803, 2011.
12. Matochko WL, Chu K, Jin B, Lee SW, Whitesides GM, Derda R, Deep sequencing analysis of phage libraries using Illumina platform, *Methods* **58**(1):47-55, 2012.
13. Menendez A, Scott JK, The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies, *Anal Biochem* **336**(2):145-157, 2005.
14. Vodnik M, Zager U, Strukelj B, Lunder M, Phage display: selecting straws instead of a needle from a haystack, *Molecules* **16**(1):790-817, 2011.
15. Zacher AN, 3rd, Stock CA, Golden JW, 2nd, Smith GP, A new filamentous phage cloning vector: fd-tet, *Gene* **9**(1-2):127-140, 1980.
16. Smith GP, Filamentous phage assembly: morphogenetically defective mutants that do not kill the host, *Virology* **167**(1):156-165, 1988.
17. Kridel SJ, Chen E, Kotra LP, Howard EW, Mobashery S, Smith JW, Substrate hydrolysis by matrix metalloproteinase-9, *J Biol Chem* **276**(23):20572-20578, 2001.
18. Derda R, Musah S, Orner BP, Klim JR, Li L, Kiessling LL, High-throughput discovery of synthetic surfaces that support proliferation of pluripotent cells, *J Am Chem Soc* **132**(4):1289-1295, 2010.
19. 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, Phage display screening without repetitious selection rounds, *Anal Biochem* **421**(2):622-631, 2012.
20. McConnell SJ, Uveges AJ, Spinella DG, Comparison of plate versus liquid amplification of M13 phage display libraries, *Biotechniques* **18**(5):803-804, 806, 1995.
21. Derda R, Tang SK, Whitesides GM, Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets, *Angew Chem Int Ed Engl* **49**(31):5301-5304, 2010.
22. Mandava S, Makowski L, Devarapalli S, Uzubell J, Rodi DJ, RELIC--a bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites, *Proteomics* **4**(5):1439-1460, 2004.
23. Huang J, Ru B, Li S, Lin H, Guo FB, SAROTUP: scanner and reporter of target-unrelated peptides, *J Biomed Biotechnol* **2010**:101932, 2010.
24. Ru B, Huang J, Dai P, Li S, Xia Z, Ding H, Lin H, Guo F, Wang X, MimoDB: a new repository for mimotope data derived from phage display technology, *Molecules* **15**(11):8279-8288, 2010.
25. Huang J, Ru B, Zhu P, Nie F, Yang J, Wang X, Dai P, Lin H, Guo FB, Rao N, MimoDB 2.0: a mimotope database and beyond, *Nucleic Acids Res* **40**(Database issue):D271-277, 2012.
26. Thomas WD, Golomb M, Smith GP, Corruption of phage display libraries by target-unrelated clones: diagnosis and countermeasures, *Anal Biochem* **407**(2):237-240, 2010.
27. Huang J, Ru B, Dai P, Bioinformatics resources and tools for phage display, *Molecules* **16**(1):694-709, 2011.
28. Gromiha MM, Suwa M, A simple statistical method for discriminating outer membrane proteins with better accuracy, *Bioinformatics* **21**(7):961-968, 2005.
29. Chen W, Lin H, Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine, *Comput Biol Med* **42**(4):504-507, 2012.

30. Shao J, Xu D, Tsai SN, Wang Y, Ngai SM, Computational identification of protein methylation sites through bi-profile Bayes feature extraction, *PLoS One* **4**(3):e4920, 2009.
31. Chang C-C, Lin C-J, LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3):27, 2011.
32. Kim T, Tyndel MS, Huang H, Sidhu SS, Bader GD, Gfeller D, Kim PM, MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets, *Nucleic Acids Res* **40**(6):e47, 2012.
33. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter* **11**(1):10-18, 2009.