Comparative Assessment of Data Sets of Protein Interaction Hot Spots Used in the Computational Method

Yunqiang Di[1,2], Changchang Wang[1,3], Huan Wu[1,2], Xinxin Yu[1,4], and Junfeng Xia[1]

[1] Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China
[2] College of Electrical Engineering and Automation, Anhui University,
Hefei, Anhui 230601, China
[3] School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China
[4] School of Life Sciences, Anhui University, Hefei, Anhui 230601, China
jfxia@ahu.edu.cn

Abstract. It seems that every biological process involves multiple protein-protein interactions. Small subsets of residues, which are called "hot spots", contribute to most of the protein-protein binding free energy. Considering its important role in the modulation of protein-protein complexes, a large number of computational methods have been proposed in the prediction of hot spots. In this work, we first collect lots of articles from 2007 to 2014 and select nine typical data sets. Then we compare the nine data sets in different aspects. We find that the maximum number of interface residues used in the previous work is 318, which can be selected as the fittest training data set used in predicting hot spots. At last, we compare and assess the features used in different works. Our result suggests that accessibility and residue conservation are critical in predicting hot spots.

Keywords: proteins-protein interaction, hot spots, computational method, training data.

1 Introduction

Protein-protein interactions play an important role in almost all biological processes such as signal transduction, transport, cellular motion, and regulatory mechanisms. Researches of residues at protein-protein interfaces has shown that only a small portion of all interface residues is actually essential for binding [1]. These residues are termed as hot spots which contribute a large fraction of the binding free energy and are crucial for preserving protein functions and maintaining the stability of protein interactions. Recent years, several studies discovered that small molecules which bound to hot spots in protein interfaces can disrupt protein-protein interactions [2]. So, identifying hot spots and revealing their mechanisms can provide promising prospect for medicinal chemistry and drug design [3-4].

Experimental methods have been used to identify hot spot residues at protein-protein interfaces. For example, alanine scanning mutagenesis has been used to

D.-S. Huang et al. (Eds.): ICIC 2014, LNBI 8590, pp. 478–486, 2014.

identify protein-protein interface hot spots [5]. Because of the high cost and low efficiency of experimental method, public databases of experimental results such as the Alanine Scanning Energetic Database (ASEdb) [6] and the Binding Interface Database (BID) [7] contain only a limited number of complexes.

Besides the experimental methods, a large number of computational methods have been proposed in the prediction of hot spots. Tuncbag *et al.* [8] constructed a web server Hotpoint to predict hot spots effectively. Darnell *et al.* [9] also provided a web server KFC to predict hot spots by using decision trees. Cho *et al.* [10] developed two feature-based predictive support vector machine (SVM) models for predicting interaction hot spots with features including weighted atom packing density, relative accessible surface area, weighted hydrophobicity and molecular interaction types. Xia *et al.* [11] introduced both a SVM model and an ensemble classifier to boost hot spots prediction accuracy. Recently, Ye *et al.* [12] used network features and microenvironment features to predict hot spots.

Although these approached have obtained good performance, there are still some problems remaining in this field. Though many features have been used in the previous studies, effective feature subsets have not been found yet. Moreover, most existing approaches use very limited data from experiment-derived databases, therefore the training data is insufficient, which may lead to pool prediction performance. Cheng *et al.* [13] also found that a rational selection of training sets had a better performance than random selection.

To assess their data sets, we compare the methods with each other and with a overlapping set of hot spots.In this paper, we collect 9 data sets about hot spots from 2007 to 2014. Firstly, we compare their training sets and analysis the same subsets they used. Then, we list the features they used and give a heuristic conclusion.

2 Datasets and Methods

(a) Datasets

We collect 2600 articles with a simple query of protein-protein interactions, hot spots prediction and computational methods on PubMed. Then we obtain 30 articles by cutting off the remaining ones whose topics are not concerning about hot spots. Finally, we select nine typical articles which are used the computational methods to predict hot spots, including APIS [11], KFC2 [14], RF [15], NFMF [12], ELM [16], KFC1 [17], MINERVA [10], DSP [18], and βACVASA [19].

Then we get data sets from the nine articles from the tables in the main text or from their supplements. The training data sets in these studies were all extracted from ASEdb [6] and the published data by Kortemme and Baker [1]. Then filtering methods were used to eliminate data redundancy by querying sequence identity. As a result, only a subset of the interface residues was chosen, and the interface residues with binding free energy ($\Delta\Delta G$) $\geq$2.0 kcal/mol are defined as hot spots [15, 17, 21]. The dataset from BID was used as test sets. BID categorizes the effect of mutations as strong, intermediate, weak or insignificant. The residues having strong interaction strengths are considered as hot spots in this study. Details of the data sets are listed in supplement Table S1-S9.

(b) Framework

As described in Fig.1. We first collect data from literature which is explained in section 2.1. Then we compare their data in different aspects such as their scale and features. We use Venn Diagrams [22] to analysis these complexes. At last we obtain the overlapping data in the nine works.

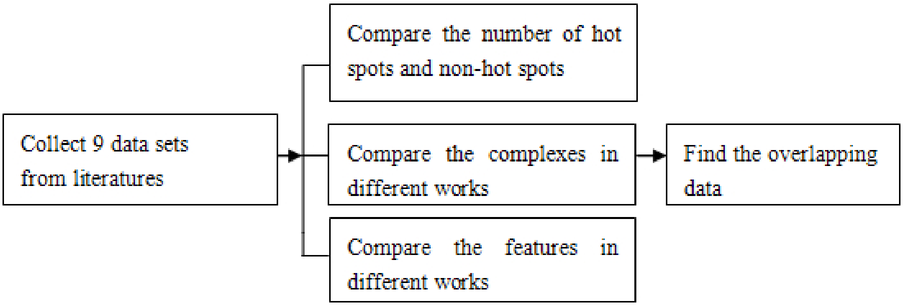
Collect 9 data sets from literatures
Compare the number of hot spots and non-hot spots
Compare the complexes in different works
Compare the features in different works
Find the overlapping data

Fig. 1. The framework of our method. Firstly, we collect 9 data sets from literatures. Then we compare the number of hot spots and non-hot spots, the complexes and features respectively. Finally, we obtain the overlapping data.

3 Results and Discussions

(a) Comparison of the Number of Training and Test Data Sets

Table 1 and 2 shows the number of training and test data sets in different works. From Table 1, we find that the number of training data sets really make a big difference. The KFC2 method has only 132 interface residues while RF, NFMF and ELM have the largest number (318) of interface residues. Then we also find that the number of hot spots and the number of non-hot spots are different in different works except NFMF, RF and ELM (the training data set of them are the same) However, the training data sets used in creating machine learning methods, such as APIS, RF, KFC1, MINERVA, DSP, and βACVASA, contain more non-hot spot residues than hot spot residues. To avoid biased predictions, the training data set of interface residues in KFC2 contains 65 hot spot residues and 67 non-hot spot residues.

From Table 2, we can see that almost all the number of test data sets is the same. Test data sets do not exist in DSP and βACVASA. The number of test data set in RF is the same as that in ELM, which is original from MINERVA. Two residues which are not in protein interfaces have been removed, so there are 125 residues in RF, not the number 127. Xia *et al.* [11] used exactly the same dataset as the one used in Cho *et al.* [10] for the purpose of comparing APIS and MINEVAR. So the number of their test data sets is also 127.

Table 1. The number of training data sets in different works

Dataset	Number of hot spots	Number of non-hot spots	Total number
APIS	62	92	154
KFC2	65	67	132
RF	77	241	318
NFMF	77	241	318
ELM	77	241	318
KFC1	60	189	249
MINERVA	119	146	265
DSP	76	145	221
βACVASA	86	148	234

Table 2. The number of test data sets in different works

Dataset	Number of hot spots	Number of non-hot spots	Total number
APIS	39	88	127
KFC2	39	87	126
RF	38	87	125
NFMF	38	86	124
ELM	38	87	125
KFC1	50	62	112
MINERVA	39	88	127
DSP	NA	NA	NA
βACVASA	NA	NA	NA

NA: Not Available

(b) Comparison of the Protein Complexes Used in the Previous Works

We list the complexes in each work. Considering the training data sets of NFMF, RF and ELM are same, so we just list the other 7 data sets in Table 3. The overlapping complexes are underlined. Then we use Venn Diagrams [22] to distinguish each other. Because the works in DSP and βACVASA do not contain test data sets, we don't use these data sets. We first divide the rest 5 data sets into 2 groups. The one group is combining the data from APIS, KFC2 and ELM, the other is combining those from the remaining two methods KFC1 and MINERVA. From Fig.2, we can see that the dataset in ELM has the widest range of complexes among the three works (APS, KFC2, and ELM). So we choose the dataset from ELM as an additional set to join into the other group and use Venn Digrams to obtain the overlapping data. Apparently, from Fig.3, we also find that the complexes used in ELM contain the widest range of the whole. To further study, we list all the training data in supplement Table S 1-7 and the same data in supplement Table S8. We find that the data set of ELM contains the maximum amount of complexes and almost cover every data set of the rest.

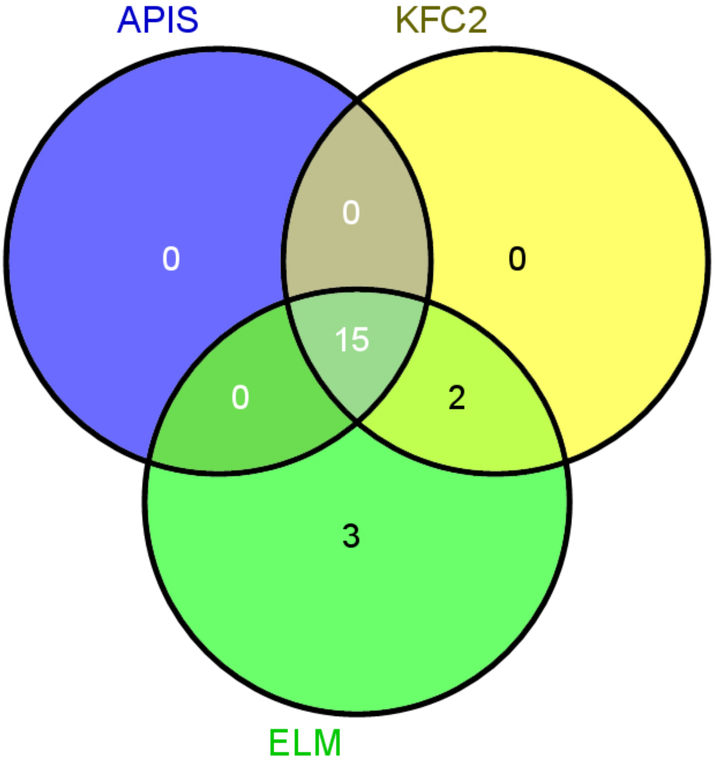

Fig. 2. The number of complexes in APIS, ELM and KFC2.The 3 complexes only exist in ELM are 1dn2, 1jck and 1jtg. 1f47 and 1nmb only exist in ELM and KFC2. The 15 complexes which all works contain are 1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1fc2, 1fcc, 1gc1, 1jrh, 1vfb, 2ptc, and 3hfm.

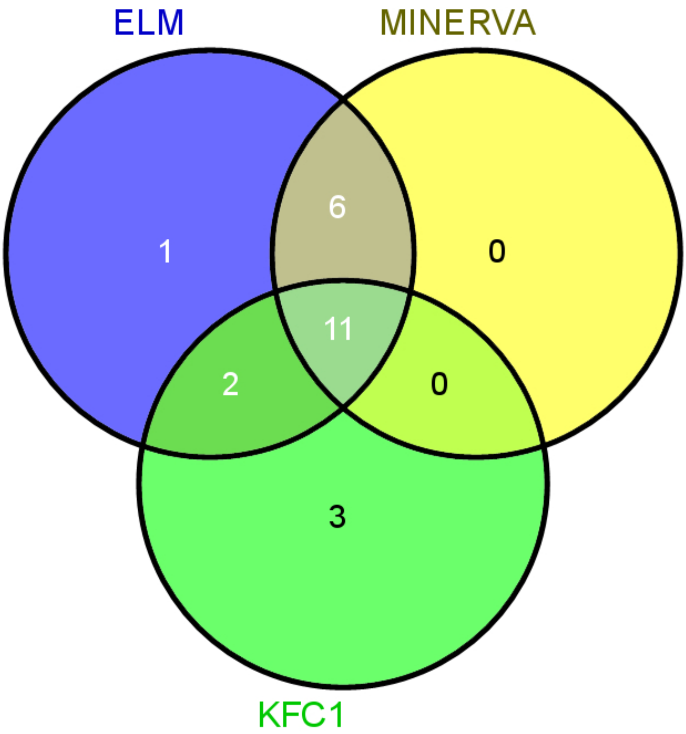

Fig. 3. The number of complexes in KFC1, ELM and MINERVA. The complex only exists in ELM is 1jck and those only in KFC1 are 1bsr, 1dx5 and 3hhr. The 2 complexes only exist both in ELM and KFC1 are 1dn2 and 1jtg and those only exist in ELM and MINERVA are 1a22, 1f47, 1fc2, 1fcc, 1jrh and 2ptc. All works contain 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1gc1, 1nmb, 1vfb, and 3hfm.

Table 3. The complexes used in the previous works

Method	Complexes
APIS	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1fc2, 1fcc, 1gc1, 1jrh, 1vfb, 2ptc, 3hfm
KFC2	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1f47, 1fc2, 1fcc, 1gc1, 1jrh, 1nmb, 1vfb, 2ptc, 3hfm
ELM	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dn2, 1dvf, 1f47, 1fc2, 1fcc, 1gc1, 1jck, 1jrh, 1jtg, 1nmb, 1vfb, 2ptc, 3hfm
KFC1	1a4y, 1ahw, 1brs, 1bsr, 1bxi, 1cbw, 1dan, 1dn2, 1dvf, 1dx5, 1gc1, 1jtg, 1nmb, 1vfb, 3hfm, 3hhr
MINERVA	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1f47, 1fc2, 1fcc, 1jrh, 1nmb, 1gc1, 1vfb, 2ptc, 3hfm
DSP	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1f47, 1fc2, 1fcc, 1gc1, 1jrh, 1jtg, 1nmb, 1vfb, 2ptc, 3hfm, 3hhr
βACVASA	1a22, 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dfj, 1dvf, 1dx5, 1f47, 1fc2, 1fcc, 1gc1, 1jck, 1jrh, 1jtg, 1nmb, 1vfb, 2ptc, 3hfm, 3hhr

The same complexes in all works are underlined.

3.1 Comparison of the Features

Table 4 gives the number of features used in different works and the methods of feature selection. And the details of features are described in supplement Table S9.We find that all models use Accessibility, Residue Conservation as features because of their importance in protein-protein interactions.

However, the computational of accessibility is a little different in the 5 works. The work in MINERVA had proved that the absolute values of solvent accessibility and surface area burial (ΔASA) had only a limited capacity to distinguish hot spots from other interface residues. To compensate that, they introduced the concept of relative surface burial (SB_r). The relative surface burial was calculated as follows:

$$SB_r(i) = \Delta ASA_i \,/\, ASA_{mono}(i) \tag{1}$$

Here $ASA_{mono}(i)$ is solvent accessibility of the *i*-th residue in a monomer.

In the work of APIS, for accessible surface area (ASA) and relative ASA (RASA), they obtained five residue attributes: total (sum of all atom values), backbone (sum of all backbone atom values), side-chain (sum of all side-chain atom values), polar (sum of all oxygen, nitrogen atom values) and non-polar (sum of all carbon atom values). The structure information was calculated by PSAIA [23]. In addition, the relative change in ASA (RcASA) was calculated as follows:

$$rel\_ASA(i) = \frac{ASA_{mono}(i) - ASA_{comp}(i)}{ASA_{mono}(i)} \tag{2}$$

Here $ASA_{comp}(i)$ is solvent accessibility of the *i*-th residue in a complex.

In the work of KFC2 and NFMF, they calculated the solvent accessible surface area using the program NACCESS [24]. In RF, they computed the relative accessible surface area (rel_ASA) of the ith residue is described in formula (2).

Table 4. The number of features used in different works

Methods	Initial number	Final number	Feature selection
MINERVA	54	12	Tree-decision
APIS	62	9	F-score
KFC2	47	14	SVM
RF	57	19	RF
NFMF	75	10	RF

4 Conclusions

In our work, we compared nine data sets from previous work. And we discuss the same training data set they all used. We think that the training data set of ELM may be the most suitable subsets to predict hot spots. In the end, we compare the features and find two features used in all works are important for protein-protein interactions. We hope that this paper can give a possible way to select training data sets and features for researchers in this field. In our future work, we will build a database that contains data both from the experimentally detected hot spots and computationally predicted hot spots.

Acknowledgments. This work is supported by the grants from the National Science Foundation of China (31301101 and 61272339), the Anhui Provincial Natural Science Foundation (1408085QF106), and the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011). The authors appreciate the constructive criticism and valuables suggestions from the anonymous reviewers.

References

1. Kortemme, T., Baker, D.: A Simple Physical Model for Binding Energy Hot Spots In Protein–Protein Complexes. Proceedings of the National Academy of Sciences 99(22), 14116–14121 (2002)
2. Walter, P., et al.: Predicting Where Small Molecules Bind at Protein-Protein Interfaces. Plos One 8(3), E58583 (2013)
3. Liu, Q., et al.: Structural Analysis of the Hot Spots in the Binding Between H1N1 HA and The 2D1 Antibody: Do Mutations of H1N1 From 1918 to 2009 Affect Much on This Binding? Bioinformatics 27(18), 2529–2536 (2011)
4. Liu, Z.-P., et al.: Bridging Protein Local Structures and Protein Functions. Amino Acids 35(3), 627–650 (2008)
5. Cunningham, B.C., Wells, J.A.: High-Resolution Epitope Mapping of Hgh-Receptor Interactions by Alanine-Scanning Mutagenesis. Science 244(4908), 1081–1085 (1989)
6. Thorn, K.S., Bogan, A.A.: Asedb: a Database of Alanine Mutations and their Effects on the Free Energy of Binding in Protein Interactions. Bioinformatics 17(3), 284–285 (2001)
7. Fischer, T., et al.: The Binding Interface Database (BID): a Compilation of Amino Acid Hot Spots in Protein Interfaces. Bioinformatics 19(11), 1453–1454 (2003)
8. Tuncbag, N., Keskin, O., Gursoy, A.: Hotpoint: Hot Spot Prediction Server for Protein Interfaces. Nucleic Acids Research 38(suppl. 2), W402–W406 (2010)
9. Darnell, S.J., Legault, L., Mitchell, J.C.: KFC Server: Interactive Forecasting of Protein Interaction Hot Spots. Nucleic Acids Research 36(suppl. 2), W265–W269 (2008)
10. Cho, K.-I., Kim, D., Lee, D.: A Feature-Based Approach to Modeling Protein–Protein Interaction Hot Spots. Nucleic Acids Research 37(8), 2672–2687 (2009)
11. Xia, J.-F., et al.: APIS: Accurate Prediction of Hot Spots in Protein Interfaces by Combining Protrusion Index with Solvent Accessibility. BMC Bioinformatics 11(1), 174 (2010)
12. Ye, L., et al.: Prediction of Hot Spots Residues in Protein–Protein Interface Using Network Feature and Microenvironment Feature. Chemometrics and Intelligent Laboratory Systems 131, 16–21 (2014)
13. Cheng, J., et al.: Training Set Selection for The Prediction of Essential Genes. Plos One 9(1), E86805 (2014)
14. Zhu, X., Mitchell, J.C.: KFC2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density, and Plasticity Features. Proteins: Structure, Function, and Bioinformatics 79(9), 2671–2683 (2011)

15. Wang, L., et al.: Prediction of Hot Spots in Protein Interfaces Using a Random Forest Model With Hybrid Features. Protein Engineering Design and Selection 25(3), 119–126 (2012)
16. Wang, L., et al.: Prediction of Hot Spots in Protein Interfaces Using Extreme Learning Machines with the Information of Spatial Neighbour Residues (2014)
17. Darnell, S.J., Page, D., Mitchell, J.C.: An Automated Decision-Tree Approach to Predicting Protein Interaction Hot Spots. Proteins: Structure, Function, and Bioinformatics 68(4), 813–823 (2007)
18. Nguyen, Q., Fablet, R., Pastor, D.: Protein Interaction Hotspot Identification Using Sequence-Based Frequency-Derived Features. IEEE Transactions on Biomedical Engineering 60(11), 2993–3002 (2013)
19. Liu, Q., et al.: Integrating Water Exclusion Theory Into ß Contacts to Predict Binding Free Energy Changes and Binding Hot Spots. BMC Bioinformatics 15(1), 57 (2014)
20. Xu, B., et al.: A Semi-Supervised Boosting SVM for Predicting Hot Spots at Protein-Protein Interfaces. BMC Systems Biology 6(suppl. 2), S6 (2012)
21. Tuncbag, N., Gursoy, A., Keskin, O.: Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy. Bioinformatics 25(12), 1513–1520 (2009)
22. Oliveros, J.C.: VENNY. An Interactive Tool for Comparing Lists with Venn Diagrams (2007)
23. Mihel, J., et al.: PSAIA–Protein Structure and Interaction Analyzer. BMC Structural Biology 8(1), 21 (2008)
24. Hubbard, S., Thornton, J.: Department of Biochemistry and Molecular Biology, University College London (1993)