Structural bioinformatics

Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy

Nurcan Tunçbag, Attila Gursoy* and Ozlem Keskin*

Center for Computational Biology and Bioinformatics and College of Engineering, Koc University, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey

Received on January 12, 2009; revised on March 19, 2009; accepted on April 3, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Hot spots are residues comprising only a small fraction of interfaces yet accounting for the majority of the binding energy. These residues are critical in understanding the principles of protein interactions. Experimental studies like alanine scanning mutagenesis require significant effort; therefore, there is a need for computational methods to predict hot spots in protein interfaces.

Results: We present a new intuitive efficient method to determine computational hot spots based on conservation (C), solvent accessibility (accessible surface area (ASA)) and statistical pairwise residue potentials (PP) of the interface residues. Combination of these features is examined in a comprehensive way to study their effect in hot spot detection. The predicted hot spots are observed to match with the experimental hot spots with an accuracy of 70% and a precision of 73% in Binding Interface Database (BID). Several machine learning methods are also applied to predict hot spots. Performance of our empirical approach exceeds methods. Residue occlusion from solvent in the complexes and pairwise potentials are found to be the main discriminative features in hot spot prediction.

Conclusion: Our empirical method is a simple approach in hot spot prediction yet with its high accuracy and computational effectiveness. We believe that this method provides insights for the researchers working on characterization of protein binding sites and design of specific therapeutic agents for protein interactions.

Availability: The list of training and test sets are available as Supplementary Data at http://prism.ccbb.ku.edu.tr/hotpoint/supplement.doc

Contact: agursoy@ku.edu.tr; okeskin@ku.edu.tr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Proteins function by interacting with other molecules through their interfaces. Studies on protein interfaces have revealed that energies are not uniformly distributed. Instead, there are certain critical residues called hot spots comprising only a small fraction of interfaces yet accounting for the majority of the binding energy (Bogan and Thorn, 1998; Clackson and Wells, 1995). Experimentally, a hot spot can be found by evaluating free energy change upon mutating it to an alanine, playing key roles on the stability of the protein association. Thorn and Bogan (2001) deposited hot spots from alanine scanning mutagenesis experiments, in the Alanine Scanning Energetics Database (ASEdb). Binding Interface Database (BID) (Fischer et al., 2003) presents experimentally verified hot spots at interfaces collected from literature.

Analysis of amino acid composition of hot spots shows that some residues are more favorable. The most frequent ones, Tyr, Arg and Trp, are critical due to their size and conformation in hot spots (Bogan and Thorn, 1998). In addition, Bogan and Thorn reported that hot spots are surrounded by energetically less important residues that most likely serve to occlude bulk solvent from the hot spots. Occlusion of solvent is found to be a necessary condition for highly energetic interactions. Hot spot information from experimental studies are available only for a very limited number of complexes, therefore, there is a need for computational methods to identify hot spots of protein interaction sites. (Delano, 2002). In a pioneering work, Kortemme and Baker (2002) proposed a physical model (Robetta) to detect hot spots at protein–protein interfaces accounting for energies of packing interactions, hydrogen bonds and solvation. Computational hot spots, the residues they identified computationally based on their model, show accordance with experimental hot spots in ASEdb. Similarly, Gao et al. (2004) used non-covalent interactions to estimate energetic contribution of interfacial residues to binding. They reported an 88% success rate for predicting hot spots obtained from alanine scanning mutagenesis experiments (Gao et al., 2004).

Another energy-based model developed by Serrano and co-workers (Guerois et al., 2002) was used to predict the energetic effect of mutations on protein complexes. The calculated energy change of mutations agreed well with the experimental results. Their method is applicable to hot spot predictions as well.

Molecular dynamics (MD) simulations can provide detailed analysis of protein interfaces at the atomic level for more accurate prediction of hot spots (Gonzalez-Ruiz and Gohlik, 2006). Rajamani et al. (2004) studied 11 protein complexes and found that anchoring...
residues in protein interfaces show restricted mobility and may act as hot spots. Kollman and co-workers (Huo et al., 2002) applied MD to find computational alanine scanning of 1:1 human growth hormone-receptor complex and reported a good agreement with the experimental data. Although these energy- and MD-based methods are successful to identify hot spots of individual protein complexes, they are not applicable, in practice, for large-scale hot spot predictions due to their computational cost.

The importance of conservation in protein interfaces is well studied (Caffrey et al., 2004; Grishin and Phillips, 1994; Valdar and Thornton, 2001). Residues at protein interfaces (Fraser et al., 2002) and functional sites (Panchenko et al., 2004) were observed to be mutating at a slower pace compared with the rest of the protein surface. There are several studies focusing on the detection of hot spots based on conservation: correlation between hot spot residues and structurally conserved residues were found to be remarkable (Ma et al., 2003). These hot spots are also found to be buried and tightly packed with other residues (Keskin et al., 2005) resulting in densely packed clusters of networked hot spots, called ‘hot regions’. It was shown that central residues are highly conserved in sequence alignments and non-exposed to the solvent in the protein complex and concluded that these residues either correspond to experimental hot spots or are in contact with an experimentally annotated hot spot (del Sol and O’Meara, 2005).

Hot spots in binding regions are located around crevices (Li et al., 2004). Predicted crevices using physicochemical properties and conservation of protein surfaces may correspond to binding hot spot regions (Burgoyne and Jackson, 2006). Another study has illustrated that there is a correlation between energy change and decrease in the accessible surface area (ASA) of individual residues as a consequence of complexation (Guharoy and Chakrabarti, 2005). In a recent work, solvent accessibility is combined with conservation in an empirical formula to identify hot spots computationally (Guney et al., 2008). Moreira et al. (2007) have supported that hot spots are protected from solvent by a rim region; however, they concluded that more computational analysis should be applied to elucidate this theory. Another approach to predict hot spots is graph analysis of the proteins. Brinda et al. (2002) have used graph representation of homodimeric protein complexes (residue network). Spectral analysis of the residue network identified some important residues involved in dimer stability that might correspond to hot spots. Recently, a neural network-based approach using various features of interfaces such as sequence profiles, solvent accessibility and evolutionary conservation is employed in computational hot spot prediction (Ofran and Rost, 2007). The method has advantage of using only sequence; thus, it is applicable when the structure is not available and also when the binding partner is unknown. A hybrid computational model combining decision tree (using atomic contacts, physicochemical properties and shape specificity contributions) with computational alanine scanning method is proposed to predict hot spots (Darnell et al., 2007). In a recent work, Grosdidier and Fernandez-Recco (2008) predict hot spots by using docking methods without protein complex knowledge. Their performance on a subset of Kortemme’s dataset reached a precision value of 0.78 and sensitivity 0.46.

Most machine learning (ML)-based hot spot prediction methods learn complex relations between training data and hot spots; however, it is very difficult to translate these relations into simple, intuitive rules (Ofran and Rost, 2007). Here, we present a new efficient method to determine computational hot spots in protein-protein interfaces from structure. The method is based on a few simple rules involving solvent accessibility and pair potentials of residues. Computational effectiveness of this model makes it favorable for hot spot prediction at large scale. As a result, by using only two features (ASA in complex and pair potential) we reached noteworthy accuracies both in training set and test set. Particularly, use of knowledge-based potentials between residues is found to be critical in identifying hot spots. We further performed an exhaustive comparison of our empirical method with various ML-based methods by using independent training and testing data. The empirical model, containing solvent accessibility and pair potentials, outperforms other empirical and ML-based methods with its performance values both on ASEdb and BIND, 70% and 70% accuracy, respectively.

2 METHODS

2.1 Training set

Proteins that have experimental hot spot data and available crystal structures are used in developing a scoring formula. Alanine scanning data were obtained from the ASEdb and a previously compiled dataset from Robetta. The redundancy in this dataset is removed using PISCES sequence culling server (Wang and Dunbrack, 2003) with sequence identity not more than 35% as in the procedure of Darnell et al. (2007). The interface residues whose observed binding free energies are ≥2.0 kcal/mol are considered as hot spots. Also, the interface residues whose binding free energy is <0.4 kcal/mol are labeled as non-hot spots in a similar way with Gao et al. (2004). Other residues having binding free energy between 0.4 and 2.0 are not included in the training to discriminate better. Actual training set used during two-class (hot spot and non-hot spot) prediction model construction consists of 150 residues, for which both conservation and solvent accessibility information is available, of which 58 residues are hot spots and 92 residues are non-hot spots.

2.2 Test set

A test set, used for assessing performance of proposed prediction models, is taken from BIND (Fischer et al., 2003). BIND contains binding free energy strengths of monomers. The test set is filtered for identical sequences in a similar fashion to the training set. The resulting set shrinks to 112 residues on 25 monomers (54 hot spots and 58 non-hot spots) when residues with known conservation scores and accessibility are considered. Hot spot residues are labeled as the ones with ‘strong’ interaction strengths and others are tagged as non-hot spots. The data originating from training and test sets are mutually exclusive. The list of training and test sets are available as Supplementary Material at http://prism.ccb.ku.edu/tr/hotspot supplement.doc.

2.3 Features

2.3.1 Accessibility

The ASA of each residue in monomer state and in complex state in the training and test sets are calculated by using Success (Hubbard and Thornton, 1993). These ASAs are then converted into relative accessibility:

\[
\text{relCompASA} = \frac{\text{ASA in Complex}}{\text{maxASA}} \times 100
\]

\[
\text{relASA} = \frac{[\text{ASA in Monomer}]}{[\text{ASA in Complex}]} \times 100
\]

where ‘relCompASA’ is the relative ASA in complex of i-th residue and ‘relASA’ is the relative difference ASA between complex and monomer state of i-th residue; in other words, the ASA change of the residue upon

\[
\text{relASA} = \frac{[\text{ASA in Monomer}]}{[\text{ASA in Complex}]} \times 100
\]

\[
\text{relCompASA} = \frac{\text{ASA in Complex}}{\text{maxASA}} \times 100
\]
These features are selected considering following criteria: hot spots are defined as the absolute of sum of its pair potentials: the distance between two residue centers (Bahar and Jernigan, 1996). We use the default cutoff in the hot spot predictions in all models.

$$\Delta_{ij}$$

Between two residues

$$T$$

Possible pairs of 20 amino acids) in R3D structures. Knowledge-based solvent mediated inter-residue potentials frequencies of contacts between different residues in proteins with known 3D structures. Knowledge-based solvent mediated inter-residue potentials (Keskin et al., 1998), extracted from protein interfaces, are used in this work. Although these potentials are not very different from the potentials extracted from overall proteins, subtle changes might be important to detect interface hot spot residues. In the Supplementary Material, 210 distinct potentials (all possible pairs of 20 amino acids in R3D unit (R, universal gas constant; T, temperature) for contacting residue pairs are supplied. Contact potential between two residues i and j is found as:

$$\text{Pair}(i,j) = \begin{cases} \text{contact potential of type } (i,j) & \text{if } d_{ij} < 7.0 \text{ and } \lvert i-j \rvert \geq 4 \\ 0 & \text{otherwise} \end{cases}$$

where Pair(i, j) is the contact potential of residues i and j and d_{ij} is the distance between two residue centers (Bahar and Jernigan, 1996). We extracted the neighbors around the residues whose side chain center of mass are closer than the cutoff (7.0 Å). Overall contact potential of residue i is defined as the absolute sum of its pair potentials:

$$\text{PP}_i = \sum_{j \neq i} \text{Pair}(i,j) \quad \text{for } \lvert i-j \rvert \geq 4$$

2.3.4 Computational alanine scanning (Robetta) Robetta (Kortemme and Baker, 2002; Kortemme et al., 2004) is a server that includes computational alanine scanning. Robetta server gives changes in the binding free energy (ΔΔG) values based on an atomic energy function including Lennard Jones interactions, solvation interactions and hydrogen bonding. The calculated ΔΔG is named 'Robetta' throughout our work. Robetta ≥ 1.0 kcal/mol is the default cutoff in the hot spot predictions in all models.

2.4 Determination of computational hot spots Size of the experimental hot spot data is small to be used in learning-based methods with large number of features to determine the hot spot characteristics. We prefer to construct our model incrementally first examining single features (base cases), and then improving our model by addition of other significant features. In the base models, we use only one feature, such as relative ASA in complex, relative difference ASA, conservation and pair potentials to discriminate hot and non-hot residues. These features are selected considering following criteria: hot spots are buried (Bogan and Thorn, 1998), structurally more conserved, highly packed (Keskin et al., 2005), known to be mostly of specific residue types, i.e. aromatic (Bogan and Thorn, 1998). The performance of the base models is used as lower bounds to assess the performance of our model and several ML-based prediction approaches.

(1) Base cases:

a. ScoreRobetta ≥ Score
b. relASA ≤ relASA complex
c. relCompASA ≤ relCompASA + Robetta
d. PairPotential, relASA complex

where ScoreRobetta, relASA complex, relCompASA and Robetta are thresholds, and currently the default values are set to 7, 30, 20, 18.0 and 1%, respectively. The explanation and justification for these default values are given in Section 3.

(2) Combination of two features: We have tested the performance of some possible two features: ScoreRobetta + relCompASA, ScoreRobetta + PP, relCompASA + PP, relCompASA + Robetta.

(3) Addition of a third feature:

relCompASA ≤ relCompASA + ScoreRobetta (and ScoreRobetta or Robetta ≥ Robetta), relCompASA ≤ relCompASA + PP (and ScoreRobetta or PP, relCompASA + Robetta).

Further, we have used ML techniques to predict hot spots using the training set for learning. Several algorithms are employed for classification: Decision tree (J48), decision table, support vector machine (SVM), BayesNet, Naive Bayes, RBFNetwork and Majority voting. The features for each residue (for the learning algorithm) consist of the same ones that we have used in the formulations above, relCompASA, ScoreRobetta and PP. The results and comparison of these formulations are discussed in Section 3.

3 RESULTS

3.1 Distribution of features of hot spots and non-hot spots In order to decide on the threshold values, we have prepared histograms of relative complex ASA (relCompASA), relative change in ASA upon complexation (relASA), conservation score and pair potentials for the hot spot and non-hot spot residues in ASEdb as shown in Figure 1. The mean and SDs of each feature are calculated for hot and non-hot residues. Further, t-tests are performed to determine if the difference between two distributions of hot and non-hot spots is statistically significant for each feature. For significant ones, we evaluate the formulas (in Section 2) by trying several threshold values between the two mean values.

Figure 1a shows the distribution of relCompASA. Though many of the hot spot residues have similar relCompASA values with non-hot spot residues, they have different mean values (hot spots: 11.9%, non-hot spots: 26.4%). The P-value for relcompASA is found as 4.7 × 10^{-7} (<0.05), which implies the significance between the means of the hot and non-hot distributions. There are significantly more non-hot spot residues that have relative complex ASA ≥ 20% (relCompASA ≥ 20.0). This is also consistent with previous studies indicating that hot spots are buried (Bogan and Thorn, 1998; Keskin et al., 2005; Li et al., 2004).

Figure 1b shows the distribution of change in ASA upon complexation. The means are found as 34.8 for hot spots and 26.4 for non-hot spots for relASA. This feature is also discriminative with a P-value 5 × 10^{-3} (<0.05). The threshold is determined...
We have evaluated prediction performance of our models (Supplementary Material). Significantly discriminative consistent with our histogram analysis and f-measure ($F_{1}$) (statistically significant to discriminate hot spots and non-hot spots as 20.3 and 12.7, respectively. This feature is potentials of residues. The means for hot spots and non-hot conservation in hot spot prediction.

Figure 1c shows conservation score distribution that does not have a clear distinction between hot and non-hot spots. The mean value for hot residues is 4.2 and for non-hot residues 3.9. The difference between two sets is insignificant ($P$-value = 0.22). This indicates that conservation may not be a good discriminating factor by itself. However, to check this slight difference, we select the threshold for conservation score as 7.0 ($t$-score = 7.0) and test the performance of conservation in hot spot prediction.

Figure 1d displays the histogram for knowledge-based pair potentials of residues. The means for hot spots and non-hot spots are found as 20.3 and 12.7, respectively. This feature is statistically significant to discriminate hot spots and non-hot spots ($P$-value = $5.4 \times 10^{-6}$). A threshold of 18.0 ($t$-pairPotential = 18.0) is chosen since a residue with pair potential more than 18.0 has a higher tendency to be a hot spot.

We further performed ANOVA analysis and determined the most important features to distinguish hot spots from non-hot spots. relCompASA, relASA and pair potentials were found to be significantly discriminative consistent with our histogram analysis (Supplementary Material).

### 3.2 Comparison of empirical hot spot detection formulations

We have evaluated prediction performance of our models (formulations) and assessed the success of the formulations by comparing accuracy ($A$), recall ($R$), precision ($P$), specificity ($S$) and f-measure ($F_{1}$) (described in Supplementary Material). In our study, recall and specificity bear importance, since we emphasize predicting both hot spots and non-hot spots. However, precision strikes as a key determinant in quantifying how accurate the positive predictions are. The models are comprehensively tested on an independent test set (BID), and their statistical performances are presented in Table 1. The first part of the table compares single feature models. Among them, conservation is observed to have no significant effect on its own. It gives the least successful results ($F_{1}$ scores and accuracies, 0.40 and 0.61 on training set, 0.49 and 0.54 on test set, respectively) compared to other features both on ASEdb and BID. This is expected according to the histogram which states that hot spots are marginally conserved (Fig. 1c) in line with the results of Ofran and Rost (2007). However, conservation was found to improve predictions substantially (Ofran and Rost, 2007), which is not the case in our results. Interface residues are found to be more conserved than the rest of the surface residues (Grishin and Phillips, 1994; Valdar and Thornton, 2001); further, central interface residues were more conserved than peripheral ones (Bordner and Abagyan, 2005). Caffrey et al. (2004) analyzed interfaces using surface patches, they found that the difference between the patches and the rest was even less pronounced. Here, our results suggest that sequence conservation is not a discriminative characteristic of hot spots ($P = 0.50$, $R = 0.33$ on training set; $P = 0.52$, $R = 0.46$ on test set). However, we observe that the totally conserved residues (with top-score 9 in our conservation scoring) are found to be substantially buried in the middle regions of the interfaces. On the other hand, not all buried residues are necessarily conserved.

Occlusion of a residue from solvent in complex state is indicated by a small relCompASA. Our results show that low relCompASA is critical for a residue to be a hot spot. Bogan and Thorn (1998) indicated that hot spots located near the center of the interface are a general property of the interfaces; and for a residue to be a hot spot, it must be largely protected from bulk solvent (corresponding to low relCompASA). Even if a residue was exposed to solvent prior to binding, it might lose a high percentage of its surface area and become protected from the solvent. This scenario is consistent with what Li et al. (2004) suggested: hot spots are either found on the complemented pockets or on the protruding surfaces. Complemented pockets and their corresponding protruding residues bind to each other, eventually, to protect each other from the solvent. relΔASA indicates the change in the solvent accessibility of a residue. The rationale for choosing relΔASA > 30% in our...
ASEdb and BID datasets

Two features performances on ASEdb and BID datasets

Multiple features performances on ASEdb and BID datasets

We have further tested the effect of combining features. First, we combined two features: (relCompASA + Score), (PP + Score), (relCompASA + Robetta) and (relCompASA + PP). We observe that in all these cases, adding a second feature increases the precision, specificity and accuracy but decreases the recall. In other words, fewer positive hot spot predictions are made with higher percentage of true cases; in addition, non-hot spot predictions improve with other residues as reported by Nussinov and her colleagues (Keskin et al., 2003; Chakrabarti and Janin, 2002), protein interfaces also have core and rim regions and hot spots are usually located in the cores of the interfaces. A residue in the core with favorable contacts has a higher chance to be a hot spot. This could be the reason why pair potential works well. Furthermore, using full atomistic energy parameters, Robetta has computational disadvantage for large-scale predictions. In addition, the performance differences of the two models on the two distinct datasets indicate the different nature of the two datasets. The hot spots in ASEdb are defined by a single threshold of 2 kcal/mol; however, in BID, there is no single threshold but rather hot spots are divided into strong, intermediate and neutral interactions. Thresholds change from one case to another.

formulation is to be able to find the protruding residues based on this fact. However, probably due to the small number of protruding hot spot residues, this parameter (\( P = 0.50, R = 0.55, A = 0.52, F_1 = 0.53 \)) does not perform better than relCompASA (\( P = 0.60, R = 0.67, A = 0.63, F_1 = 0.63 \)). As a result, relCompASA seems to discriminate better hot spots from non-hot spots. Both of the energetic models (the knowledge-based pair potentials and full atomistic energy terms of Robetta) seem to be quite successful to find the hot spots. Robetta’s recall, precision and accuracy are higher in ASEdb (0.66, 0.72 and 0.73, respectively), lower in BID (0.66, 0.72 and 0.73, respectively). On the other hand, pair potential performs better in BID (0.69, 0.70 and 0.71, respectively) compared to Robetta. Note that ASEdb is the training set and BID is our independent test set. As stated by Janin and his group (Bahadur et al., 2003; Chakrabarti and Janin, 2002), protein interfaces also have core and rim regions and hot spots are usually located in the cores of the interfaces. A residue in the core with favorable contacts has a higher chance to be a hot spot. This could be the reason why pair potential works well. Furthermore, using full atomistic energy parameters, Robetta has computational disadvantage for large-scale predictions. In addition, the performance differences of the two models on the two distinct datasets indicate the different nature of the two datasets. The hot spots in ASEdb are defined by a single threshold of 2 kcal/mol; however, in BID, there is no single threshold but rather hot spots are divided into strong, intermediate and neutral interactions. Thresholds change from one case to another.

We have further tested the effect of combining features. First, we combined two features: (relCompASA + Score), (PP + Score), (relCompASA + Robetta) and (relCompASA + PP). We observe that in all these cases, adding a second feature increases the precision, specificity and accuracy but decreases the recall. In other words, fewer positive hot spot predictions are made with higher percentage of true cases; in addition, non-hot spot predictions improve compared to single feature models. On ASEdb, (relCompASA + Robetta) model has the highest \( F_1 \) score (0.67); whereas on BID, (relCompASA + PP) has the highest \( F_1 \) score (0.65). Compared to the single feature performances, adding the relCompASA in the Robetta model increases the precision from 0.63 to 0.75 and in the PP model from 0.69 to 0.73 in BID. Similarly, specificity increases from 0.69 to 0.85 (for Robetta) and 0.71 to 0.79 (for PP) in BID. The model starts to pick the hot spots and non-hot spots with higher specificity and precision. Further, adding relCompASA improves the performance of pair potentials on ASEdb with respect to pair potentials only while maintaining the BID performance. Similarly, relCompASA improves performance of Robetta on BID compared to using Robetta only. Our results indicate that hot spots are mostly buried and form a network of favorable interactions with other residues as reported by Nussinov and her colleagues (Keskin et al., 2005). When conservation score is added to these two-feature models, both the precision and specificity decrease. We should note that, we tried many other features, combinations and thresholds, but listed only the high scoring ones. As a result, our prediction based on relCompASA and pair potentials demonstrated

---

**Table 1.** Performance values of various empirical prediction methods used to identify hot spots in the protein interfaces

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>( P )</th>
<th>( R )</th>
<th>( S )</th>
<th>( A )</th>
<th>( F_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single feature performances on ASEdb and BID datasets (Empirical formulas)</td>
<td>Score ( \geq ) 7.0</td>
<td>Training set</td>
<td>0.50</td>
<td>0.33</td>
<td>0.79</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.52</td>
<td>0.46</td>
<td>0.60</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>relCompASA ( \leq ) 20.0</td>
<td>Training set</td>
<td>0.55</td>
<td>0.81</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.60</td>
<td>0.67</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>relAtASA ( \geq ) 30.0</td>
<td>Training set</td>
<td>0.50</td>
<td>0.57</td>
<td>0.64</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.50</td>
<td>0.55</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Robetta ( \geq ) 2.0</td>
<td>Training set</td>
<td>0.82</td>
<td>0.47</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.64</td>
<td>0.26</td>
<td>0.86</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Robetta ( \geq ) 1.0</td>
<td>Training set</td>
<td>0.63</td>
<td>0.72</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.63</td>
<td>0.57</td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>PP ( \geq ) 18.0</td>
<td>Training set</td>
<td>0.56</td>
<td>0.55</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Two features performances on ASEdb and BID datasets</td>
<td>relCompASA ( \leq ) 20.0 and Score ( \geq ) 7.0</td>
<td>Training set</td>
<td>0.61</td>
<td>0.29</td>
<td>0.88</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.71</td>
<td>0.32</td>
<td>0.88</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>PP ( \geq ) 18.0 and Score ( \geq ) 7.0</td>
<td>Training set</td>
<td>0.57</td>
<td>0.14</td>
<td>0.94</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.75</td>
<td>0.33</td>
<td>0.90</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>relCompASA ( \leq ) 20.0 and Robetta ( \geq ) 1.0</td>
<td>Training set</td>
<td>0.72</td>
<td>0.62</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.75</td>
<td>0.50</td>
<td>0.85</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>relCompASA ( \leq ) 20.0 and PP ( \geq ) 18.0</td>
<td>Training set</td>
<td>0.64</td>
<td>0.52</td>
<td>0.82</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.73</td>
<td>0.59</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>Multiple features performances on ASEdb and BID datasets</td>
<td>relCompASA ( \leq ) 20.0 and (Score ( \geq ) 7.0 or Robetta ( \geq ) 1.0)</td>
<td>Training set</td>
<td>0.64</td>
<td>0.69</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.73</td>
<td>0.59</td>
<td>0.79</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>relCompASA ( \leq ) 20.0 and (Score ( \geq ) 7.0 or PP ( \geq ) 18.0)</td>
<td>Training set</td>
<td>0.63</td>
<td>0.67</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test set</td>
<td>0.67</td>
<td>0.63</td>
<td>0.71</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Definitions are given in Supplementary Material.
Table 2. ML-based models results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Testing</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>A</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesNet</td>
<td>10-fold</td>
<td>0.58</td>
<td>0.64</td>
<td>0.71</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>0.64</td>
<td>0.63</td>
<td>0.67</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>10-fold</td>
<td>0.57</td>
<td>0.66</td>
<td>0.69</td>
<td>0.67</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>0.63</td>
<td>0.67</td>
<td>0.64</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>RBFNetwork</td>
<td>10-fold</td>
<td>0.59</td>
<td>0.55</td>
<td>0.76</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>0.67</td>
<td>0.48</td>
<td>0.78</td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td>SVM</td>
<td>10-fold</td>
<td>0.57</td>
<td>0.36</td>
<td>0.83</td>
<td>0.65</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>0.73</td>
<td>0.44</td>
<td>0.85</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>10-fold</td>
<td>0.47</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>(J48)</td>
<td>0.65</td>
<td>0.52</td>
<td>0.74</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>Decision table</td>
<td>10-fold</td>
<td>0.58</td>
<td>0.67</td>
<td>0.71</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Test set</td>
<td>0.64</td>
<td>0.63</td>
<td>0.67</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Majority voting</td>
<td>10-fold</td>
<td>0.56</td>
<td>0.62</td>
<td>0.70</td>
<td>0.67</td>
<td>0.59</td>
</tr>
<tr>
<td>(all except SVM)</td>
<td>Test set</td>
<td>0.64</td>
<td>0.63</td>
<td>0.67</td>
<td>0.65</td>
<td>0.64</td>
</tr>
</tbody>
</table>

These are the corresponding implementations from Weka (Witten and Frank, 2005).

Table 3. Hot spot prediction performances on test set (BID)

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>A</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robetta</td>
<td>0.63</td>
<td>0.57</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>KFC</td>
<td>0.51</td>
<td>0.36</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>KFCA</td>
<td>0.53</td>
<td>0.48</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>0.72</td>
<td>0.57</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Our formula</td>
<td>0.73</td>
<td>0.59</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

0.70 accuracy and 0.73 precision, 0.59 recall and 0.79 specificity on the independent test set. It performed better than the base models and the ML-based models (discussed in the next section). We predict 32 of the hot spots correctly with 12 false positives. On the other hand, 46 of the non-hot spots are correctly classified with 22 false negatives (Table S4).

3.3 ML-based approaches

The ML methods fail to create a distinctive improvement over our proposed model. Performance of ML-based models is illustrated in Table 2 with the details of the classifiers on 10-fold cross-validation and on test set. In general, ML-based models do not exceed our empirical formula (A = 0.70). The main reason for this relative failure is probably deficiency of training data. Nevertheless, decision trees play an indispensable role in determination of relative importance of the features. We have applied decision tree for three features: relative compASA, pair potential and conservation score. The decision tree model determines pair potential as the most discriminating feature followed by relCompASA and pair potentials—perm performs better than both KFC and KFCA with its precision of 0.73 and recall of 0.59 with 0.65 F1 score compared to KFC (0.51, 0.36 and 0.42, respectively) and KFCA (0.53, 0.48 and 0.41, respectively) (Table 3). We further applied ISIS, a sequence-based approach, on BID giving following performance: P = 0.48, R = 0.70 and F1 = 0.57. Although the precision of ISIS is low, we should note that the method is not designed for hot spot prediction but rather finding binding site residues and it does not use structure information. Therefore, it is not fair to compare it with the structure-based methods.

Further, we performed linear discriminant analysis (LDA), trained on ASEdb and tested on BID, resulting in comparable performance. However, our method has advantages of presenting a simple and intuitive rule relating physical properties to hot spots. The details are given in the Supplementary Material.

When we analyze overall performances, we noticed that our results are similar to Robetta; however, it outperforms any ML-based predictions including KFC. Besides its high prediction performance, another advantage of our method over Robetta is its computational effectiveness and applicability to the large-scale datasets.

3.5 Case studies

3.5.1 Erythropoietic receptor—erythropoietin mimetic peptide

Erythropoietin (EPO) is a hormone participating in the regulation of proliferation and differentiation of immature erythroid cells. EPO mimetic peptide (EMPI) functions as a mimetic of EPO. There is a competition between EMPI (pdbID:1ebp, chainC) and EPO to bind the erythropoietic receptor (EPOR) (pdbID:1ebp, chainA) (Livnah et al., 1996). Despite the unrelated sequences of EMPI and EPO, both can bind to the EPOR stimulating biological activity.

Experimentally defined hot spots in 1ebpAC interface are F93_A, M150_A, F205_A and W13_C. In addition, T151_A, L111_C and T12_C are found experimentally to be non-hot spots (in BID). Our empirical method predicts two of the four hot spots correctly—F205_A and M150_A. Despite their high contact potentials, because F93_A and W13_C are exposed to solvent, they are predicted as non-hot spots. Also, all of the three non-hot spots are predicted correctly.

In total, five of the seven residues are correctly predicted (Fig. S1). KFC predicts all seven residues as non-hot spots. Robetta identifies M150_A and W13_C as hot spots correctly and the rest as non-hot spots.
3.5.2 Streptococcal protein G—mammalian immunoglobulin

Streptococcal protein G (pdbID: 1fcc, chain C) is a cell wall protein which binds mammalian immunoglobulin (pdbID: 1fcc, chain A) (Bjorck and Blomberg, 1987). Protein G has experimentally determined three hot, four non-hot spots in its binding site to immunoglobulin. Mutations of residues E27_C, K31_C and W43_C in immunoglobulin strongly affect its binding to protein G. These hot spots are located in the middle of the binding site of immunoglobulin to protein G and form a cluster of hot spots. Our method labels all these residues as hot spots. Also, non-hot residues are distributed at the edges of binding site and more accessible to the solvent and they have less contact to other residues. All of them are predicted as non-hot by our model (Fig. S2). Robetta and KFC perform similar. They identify E27_C and W43_C as hot spots correctly and the rest as non-hot spots.

These two cases are selected from BID randomly; however, when other cases are examined we noticed that our predictions correlate with Robetta.

4 CONCLUSION

Hot spots are residues comprising only a small fraction of interfaces yet accounting for the majority of the binding energy. We present a new efficient method to determine computational hot spots based on pair potentials and solvent accessibility of interface residues. We note that solvent occlusion is a necessary factor to define a hot spot, but not sufficient itself. Conservation does not have significant effect in hot spot prediction as a single feature. Residue occlusions from solvent and pairwise potentials are found to be the main discriminative features in hot spot prediction. The predicted hot spots are observed to match with the experimental hot spots with an accuracy of 70%. We also compared our empirical methods to several ML methods and other hot spot prediction methods. Our method outperforms them with its high performance.

We believe that the results provide insights for researchers working on characterization of protein interaction sites (Aytuna et al., 2005; Ogmen et al., 2005; Tuncbag et al., 2008). Such studies provide insights for function when clear evolutionary structural relationship between the sequences being compared exists and insights into which residues are most important in defining particular protein interface signatures. We should note that both ASEdb and BID are obsolete and there is a need to establish a hot spot database from current literature.

ACKNOWLEDGEMENTS

We thank Nir Ben-Tal’s group for providing RAS and special thanks to Emre Guney (UPF) for his helpful discussions. O.K. acknowledges TUBA Distinguished Young Investigator Award.

Funding: TUBITAK (to N.T.)

Conflict of Interest: none declared.

REFERENCES


