ABSTRACT

Motivation: Half-sphere exposure (HSE) is a newly developed two-dimensional solvent exposure measure. By conceptually separating an amino acid’s sphere in a protein structure into two half spheres which represent its distinct spatial neighborhoods in the upward and downward directions, the HSE-up and HSE-down measures show superior performance compared with other measures such as accessible surface area, residue depth and contact number. However, currently there is no existing method for the prediction of HSE measures from sequence data.

Results: In this article, we propose a novel approach to predict the HSE measures and infer residue contact numbers using the predicted HSE values, based on a well-prepared non-homologous protein structure dataset. In particular, we employ support vector regression (SVR) to quantify the relationship between HSE measures and protein sequences and evaluate its prediction performance. We extensively explore five sequence-encoding schemes to examine their effects on the prediction performance. Our method could achieve the correlation coefficients of 0.72 and 0.68 between the predicted and observed HSE-up and HSE-down measures, respectively. Moreover, contact number can be accurately predicted by the summation of the predicted HSE-up and HSE-down values, which has further enlarged the application of this method. The successful application of SVR approach in this study suggests that it should be more useful in quantifying the protein sequence–structure relationship and predicting the structural property profiles from protein sequences.

Availability: The prediction webserver and supplementary materials are accessible at http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/hse/.

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

A central problem in structural biology is to predict protein threedimensional structure from primary sequence (Baker and Sali, 2001). To this end, an intermediate but useful approach is to predict protein structural properties such as secondary structure and solvent accessibility or exposure, which simplifies this prediction task by projecting the protein structures onto one-dimensional, namely, strings of residue-wise structural assignments (Kinjo and Nishikawa, 2005; Kinjo et al., 2005; Rost and Sander, 1993, 1994; Rost et al., 2004; Song and Burrage, 2006; Yuan and Huang, 2004). In this regard, solvent exposure measures describe to what extent a residue in a protein interacts with its surrounding solvent molecules and hence could provide important information for understanding and predicting many aspects of protein structure and function (Hamelryck, 2005; Yuan and Huang, 2004) and for identifying uncharacterized evolutionary mechanism of novel protein folds from existing folds (Cordes et al., 1999). In other investigations, the solvent accessibility has been successfully utilized to improve the prediction of flexible or rigid residues (Schlessinger et al., 2006) and DNA-binding residues (Ofran et al., 2007) in proteins. Therefore, the knowledge of solvent exposure is of great biological importance, which is not only useful for predicting structural and functional features of proteins and predicting the three-dimensional structures of proteins, but also helpful for our deep understanding of the sequence–structure–function relationship.

Over the years, several solvent exposure measures have been developed, for example, solvent accessible surface area (ASA) (Connolly, 1983; Miller et al., 1987), relative accessible surface area (rASA) (Rost and Sander, 1994), residue depth (RD) (Chakravarty and Varadarajan, 1999) and contact number (CN) (Nishikawa and Ooi, 1980; Pollastri et al., 2001). Despite their contributive knowledge provided by these solvent exposure measures, they have intrinsic drawbacks. For example, it is impossible to apply ASA measure to determine to what extent a residue is buried, or it is difficult to distinguish a deeply buried residue from a partially buried residue, while for these two kinds of residues their ASA values would be zeros or close to zeros. In the case of RD measure, it is difficult to compare residues with different sizes and calculating RD suffers from high computational complexity and inefficiency. While in the case of the CN measure, it could only provide a rather coarsegrained and insensitive illustration of a residue’s solvent exposure, in comparison with ASA and RD.

In this context, half-sphere exposure (HSE), as a new kind of two-dimensional solvent exposure measure (Hamelryck, 2005), is of particular interest in this study. Compared with other solvent exposure measures, HSE has a superior performance with respect to protein stability, conservation among different folds, computational speed and predictability (Hamelryck, 2005). HSE separates a residue’s sphere into two half spheres: HSE-up corresponds to the upper sphere in the direction of the chain side of the residue, while HSE-down points to the lower sphere in the direction of the opposite side. As the two half spheres specified by HSE-up and HSE-down are distinct in terms of geometry and energy, they possess different interesting properties that are related to the characterization...
of the residue’s spatial neighborhood. Compared with other solvent exposure measures such as ASA and RD, calculation of the HSE does not rely on a full-atom model, making it easier to be applied in protein structure modeling and prediction analysis, based on the simplified models. While compared with CN, HSE can provide more detailed descriptions of a residue’s local environment, as HSE captures local residues in a residue’s side chain and it’s opposite directions. All these features make HSE likely to be applied in a wider range of fold recognition, structure prediction and modeling simulations. However, it is not clear so far to what extent HSE can be predicted from protein sequences. In this article, we propose a novel approach to quantify the HSE-sequence relationship and predict HSE measures from primary sequences alone based on support vector regression (SVR). To realize the above goal, we have created a publicly available webserver called HSEpred to facilitate the HSE as well as CN prediction. This webserver allows users to perform rapid exploratory analysis of protein sequences of their interest. It allows users to submit a protein sequence in the FASTA format and select one of the three models derived from three sequence-encoding schemes to predict the HSE-up, HSE-down and CN values for all residues in the query sequence.

2 METHODS

2.1 Datasets
We prepared a high-quality dataset of 632 protein chains using PDB-REPDDB database (Noguchi and Akiyama, 2003) derived from the RCSB Protein Data Bank (Bernstein et al., 1977). All structures were solved by X-ray crystallography with resolution ≤ 2.0 Å and R-factor ≤ 0.2. All protein chains contain at least 80 amino acids or longer, and the pair-wise sequence identity is < 25%. These selection criteria are adopted to ensure that a high-quality dataset can be obtained, which will serves as a reliable basis for building the SVR models that could enable HSEpred to provide accurate HSE and CN predictions.

There are totally 159 533 amino acid residues in this dataset. The protein chain names, amino acid sequences, the 4-fold cross-validation list, and the calculated CN, HSE-up, HSE-down values for all residues in this dataset can be found in the Supplementary Material available at http://sunflower.kuicr.kyoto-u.ac.jp/~sjphsu/.

2.2 HSE measure
Hamelryck (2005) introduced the concept of HSE, a new two-dimensional measure of a residue’s solvent exposure (Hamelryck, 2005). HSE measure divides a residue’s spatial sphere into two equal parts: HSE-up and HSE-down. The former corresponds to the upper sphere on the side chain of the residue, and the latter refers to the lower sphere on the opposite side.

In the present study, a residue’s HSE-up measure is defined as the number of Ca atoms in its upper half-sphere, which contains the Ca-Cβ vector. Likewise, HSE-down is defined as the number of Ca atoms in the other lower half-sphere.

To calculate the HSE-up and HSE-down measures for all the residues in our dataset, we set the sphere radius $r_2 = 13\AA$ that was previously adopted by Hamelryck (2005) and used the huepro program in Biopython’s Bio.PDB module (http://www.biopython.org). Three steps are involved in the HSE calculation: the first step is to identify all Ca atoms within a sphere radius $r_2$ of a residue’s Ca atom; the second step is to construct a plane that is perpendicular to a given Ca-Cβ vector and goes across the centered residue’s Ca atom and equally divide the sphere into two half-spheres in the upward and downward directions; the third step is then to calculate the numbers of Ca atoms in both the upper and lower half spheres, which correspond to the values of HSE-up and HSE-down (Hamelryck, 2005). The huepro program calculates the HSE-up and HSE-down values for all residues in a PDB file and the calculated results will be written out in this PDB file’s B factor records.

2.3 Normalization of HSE measures
HSE-up and HSE-down values were normalized using the following formula before being input into the SVR:

$$y_i = \frac{y_i - \bar{y}}{\text{SD}}$$

where $y_i$ is the normalized HSE value of $i$ residue, $\bar{y}$ is the raw HSE value, $\bar{y}$ is the mean raw HSE value, SD is the standard deviation.

We first predicted the normalized HSE-up and HSE-down values from protein sequences, and then recovered the absolute HSE-up and HSE-down values from their predicted normalized values using the above equation. This normalization step can simplify the data handling process and enable the comparison of the predicted properties at the same scale.

2.4 SVR
Support vector machine (SVM) is an efficient machine learning technique based on statistical learning theory (Vapnik, 1998). SVM can usually perform better than other machine learning algorithms owing to its excellent capacity and ability to control error without causing overfitting to the data. It has been increasingly used in many aspects of bioinformatics, such as microarray data analysis (Brown et al., 2000), protein subcellular localization prediction (Hua and Sun, 2001), protein citrulline isomerization prediction (Song et al., 2006), protein fold recognition (Chen and Kurgan, 2007; Cheng and Baldu, 2006), single amino acid polymorphism identification (Ye et al., 2007), functionally flexible region (On et al., 2006), nucleosome positioning signal (Peckham et al., 2007) and protein–protein interaction (Bradford and Westhead, 2005; Shen et al., 2007) and disorder region prediction (Ishida and Kinoshita, 2007).

In practice, SVM has two practical modes: support vector classification (SVC) and SVR. In comparison with SVC, SVR has an outstanding ability in predicting the raw property values of the testing samples and it is especially effective when the input data is characterized by high dimension and non-linear function. Recently SVR has been attracting more attention and has been applied in predicting protein ASA (Yuan and Huang, 2004), CN (Ishida et al., 2006; Yuan, 2005), residue-wise contact order (RWCO) (Song and Burrage, 2006), disulfide connectivity prediction (Song et al., 2007), gene-expression level (Raghava and Han, 2005) and peptide-MHC binding affinities (Wan et al., 2006). In this study, we describe its application to predict HSE-up and HSE-down values from protein sequences only. We used the SVM light package developed by Joachims (1999) for the SVR implementation. We selected radial basis kernel (RBF kernel) function at $\epsilon < 0.01$, $\gamma > 0.01$ and $C = 5.0$ to build the SVR models for HSE-up and HSE-down. This parameter set has been previously shown to yield the best performance in the studies of ASA (Yuan and Huang, 2004), CN (Yuan, 2005), RWCO (Song and Burrage, 2006) and disulfide connectivity (Song et al., 2007).

2.5 Sequence-encoding schemes
The sequence features used to build the SVR models were divided into global (fixed values for a protein) and local (sequence descriptors describing the local sequence environment of each residue within a protein, which varied from residue to residue). Global features comprised 20 amino acid compositions ('AA'), sequence weight ('W') and sequence length ('L'), which described general protein characteristics. Local features included the position-specific scoring matrix (PSSM) in the form of PSI-BLAST (Altschul et al., 1997) profile ('LS') and the predicted secondary structure ('SS') information by PSIPRED (Jones, 1999).
We calculated the HSE-up and HSE-down measures for each residue in our dataset and showed their distributions according to five different radius cutoffs (Fig. 1). On one hand, HSE-up and HSE-down show different distributions, implying that they correspond to distinct spatial regions with different properties, which is easy to understand as HSE-up describes the extent of a residue’s solvent exposure in the direction of its side chain, while HSE-down illustrates the degree of its solvent exposure in the opposite direction of its side chain (Hamelryck, 2005). On the other hand, for both HSE-up and HSE-down, distributions with larger radius cutoffs (12, 13 and 14 Å) are more close to normal distributions. Note that other radius cutoffs such as 8 Å are also commonly used to define inter-residue interactions in the context of protein folding and stability (Gromiha and Selvaraj, 2004). However, since previous work has indicated that CNs defined with larger radius cutoffs (from 12 Å to 14 Å) are more useful in protein fold recognition and structure prediction (Karchin et al., 2004; Yuan, 2005), we set up the radius \( r_2 = 13 \) Å in the following analysis, which is also consistent with Hamelryck (2005) work.

We further plotted their two-dimensional histogram (Fig. 2). The distribution of HSE-up differs from that of HSE-down. Overall, there are two most densely aggregated regions. HSE-up has much narrower range of values with the range from 12 to 28, especially for the region close to the x-axis, which is in contrast with the much wider range of HSE-down ranging from 0 to 32.

We next studied the distribution of HSE-up and HSE-down measures according to the secondary structures. For this purpose, we extracted the secondary structure annotation for each residue in our dataset using the DSSP program (Kabsch and Sander, 1983), which assigns each residue’s secondary structure to one of the following eight classes: α-helix (H), 3_10 helix (G), β-helix (I), β-strand (E), β-bridge (B), Coil (C), L or space, Turn (T) and Bend (S) (Crooks and Brenner, 2004). We used the common CK mapping (Chandonia and Karplus, 1995) to further classify them into three classes: α-helix (H→H), β-strand (E→E) and other irregular or unstructured elements (all others→C). This distribution is displayed in Figure 3. For the current dataset, residues with the corresponding HSE measures.

Fig. 1. The distributions of HSE-up and HSE-down measures at five different radius thresholds. The five radius \( r_2 \) cutoffs are selected as 8Å, 10Å, 12Å, 13Å and 14Å, which are represented by dotted, dashed, dot-and-dashed, solid and short-dashed lines, respectively.

Fig. 2. Two-dimensional distribution of the HSE measures for all residues in our dataset. The x-axis and y-axis indicate the HSE-up and HSE-down measures, respectively. The values in the color legend denote the numbers of residues with the corresponding HSE measures.
Fig. 3. The distributions of the HSE-up and HSE-down measures according to three secondary structures: α-helix (H), β-strand (E) and coil (C).

account for 40.8, 26.8 and 32.4%, respectively. It can be found that β-strand residues tend to have larger HSE values and coiled residues are inclined to locate with smaller HSE values, while the distribution of α-helix residues remain modest in between. Additionally, in the case of HSE-up, it has a large proportion of zero- or near-zero-valued coiled residues. In the case of the HSE-down measure, its respective distribution shapes based on α-helix, β-strand and coil classifications were found to be highly similar, despite the higher percentage peak value of the distribution based on α-helix classification. The reason for why the coil distribution for HSE-up is different from HSE-down might be that HSE-up and HSE-down correspond to distinct spatial regions in terms of the geometry, and the residue contact densities in the upper half sphere are significantly lower than that in the down half sphere.

3.2 The correlations between HSE and other structure-based exposure measures

In this analysis, we calculated the correlation coefficients between the HSE and other structure-based parameters, such as CN, RD, ASA, rASA and RWCO, to investigate their interconnections (Supplementary Table 1). The results revealed several points: First, there are not strong correlations with a correlation coefficient (CC) of 0.09 between HSE-up and HSE-down measures, which means that the distribution of the number of Cα atoms in the upper half sphere has no relationship with the number of Cα atoms in the down half sphere (Hamelryck, 2005). The implication of this finding is that HSE-up and HSE-down provide distinct yet complementary information in regards to the description of a residue’s spatial environment. Second, ASA and rASA are most strongly correlated with CC = 0.93, which is not surprising as a residue’s rASA is simply the normalization of its ASA using the maximum ASA for that residue type. Third, CN has a strong negative correlation with ASA as indicated by the CC of -0.70, which is understandable as residues with larger ASAs tend to have fewer contacting residues in its structure space. Finally, as expected, both of HSE-up and HSE-down exhibit significant correlations with CN, with the CCs of 0.81 and 0.66, respectively. Given that CN can be computed by summing HSE-up and HSE-down, it is conceivable that HSE-up and HSE-down would have significant correlations with CN.

In order to investigate the relationships between HSE-up, HSE-down and ASA, we obtained the ASA values for all residues in our dataset using the DSSP program and calculated the mean values and SDs of ASAs for HSE-up and HSE-down, whose results are shown in Figure 4. A significant negative correlation with a CC of −0.76 can be observed between HSE-up and ASA, while there is no strong correlation between HSE-down and ASA (Fig. 4).

3.3 Predicting HSE using PSI-BLAST profiles

In this section, we focused on predicting HSE-up and HSE-down values from protein amino acid sequences. To quantify the relationship between HSE measures and protein sequence and to predict them based on sequence information only, we used the SVR approach to solve this problem. As discussed in the Section 2, we used the normalized HSE values instead of the absolute ones to build the SVR models based on the training datasets and then applied the built models to predict HSE values for the testing datasets. Finally, we transformed the predicted normalized HSE values into their predicted raw values, based on the mean raw HSE values and the SDs. The predicted CN of a residue is simply computed as the summation of its predicted HSE-up and HSE-down values, according to the definition.

Based on structural risk minimization principle, SVR can reduce the overfitting problem by minimizing the generalization error. In this study, we performed 4-fold cross-validation tests to carry out an objective evaluation of the SVR approach, whose prediction results have indicated that the overfitting problem is not severe. Two measures CC and RMSE were used to evaluate the prediction performance (For more details, see Supplementary Material). The average results for HSE-up, HSE-down and CN are tabulated in Table 1. Specifically, for HSE-up, the SVR based on ‘LS’ could predict its profiles with the CC of 0.69 between the predicted and observed HSE-up values, and the RMSE_raw of 6.81, respectively.
All results were evaluated using 4-fold cross-validation method.

For HSE-down, the SVR predictor based on ‘LS’ could predict its values with CC = 0.65 and RMSE\_raw = 5.62, respectively.

Moreover, by summing up the predicted HSE-up and HSE-down values, our SVR approach could predict CN with CC = 0.72 and RMSE\_raw = 8.57, respectively, which provides an accurate prediction for CN. In addition, for the encoding scheme ‘LS’, only the position-specific scoring matrices in the form of PSI-BLAST profiles served as the input to the SVR. Hence, such prediction results substantiate the effectiveness of using the PSSMs stored in the PSI-BLAST profiles to accurately predict the HSE values from protein sequence. As previous studies have indicated, the important evolutionary information hidden in the PSSM could provide better prediction performance compared with the single sequence alone (Chen and Kurgan, 2007; Ishida and Kinoshita, 2007; Song and Burrage, 2006; Song et al., 2007; Yuan, 2005).

As previous studies have indicated (Kinjo et al., 2005; Schlessinger et al., 2006; Song et al., 2007; Yuan, 2005), incorporating global sequence features might be helpful for improving the prediction accuracy. To achieve this, we utilized three global sequence descriptors, i.e., 20 amino acid compositions (‘AA’), protein sequence weight (‘W’) and sequence length (‘L’). For ‘W’ and ‘L’ descriptors, we encoded them into the SVR after the normalization using their mean sequence weights or lengths and SDs based on our dataset. We employed five different sequence encoding schemes, i.e., local sequence in the form of PSI-BLAST profiles (‘LS’), predicted secondary structure information by PSIPRED (‘SS’), local sequence plus predicted secondary structure (‘LS+SS’), local sequence plus predicted secondary structure and amino acid composition (‘LS+SS+AA’), and local sequence plus predicted secondary structure coupled with amino acid composition, sequence weight and sequence length (‘LS+SS+AA+W+L’). The prediction results for these encoding schemes are also summarized in Table1.

As expected, using ‘LS+SS+AA’, we achieved a slightly improved performance of RMSE\_raw = 6.66 for HSE-up and RMSE\_raw = 5.47 for HSE-down, respectively, although CC and RMSE\_norm remain at the same level. However, when combining sequence weight and sequence length information, ‘LS+SS+AA+W+L’ could predict HSE-up with CC of 0.72, RMSE\_norm of 0.70 and RMSE\_raw of 6.59, and HSE-down with CC of 0.68, RMSE\_norm of 0.74 and RMSE\_raw of 5.43, respectively, which is a more significant improvement compared to ‘LS+SS+AA’. These observations suggest that including either amino acid composition (‘AA’) or sequence weight (‘W’) or sequence length (‘L’) could yield the better prediction performance compared with local sequence alone, which coincides well with the previously reported importance of ‘W’ on the prediction performance of CN and RWCO (Song and Burrage, 2006; Yuan, 2005). Additionally, these results also indicate that protein size (represented by ‘W’ and ‘L’) is a very important factor that has more significant influence on the prediction performance than ‘AA’ in predicting HSE values, which is conceivable because residues in larger proteins may be slightly less exposed than in smaller ones and as a global descriptor protein size can globally determine the environment where its residues are located.

To further explore this protein-size effect, we plotted the CC and RMSE of each protein against its corresponding sequence length, as shown in Supplementary Figure1. We can see that most of the predicted proteins have CCs larger than 0.45 and RMSEs less than 6 in both cases of HSE-up and HSE-down, while some badly predicted proteins are also observed, especially for those with sequence lengths ranging from 100 to 400. These results imply that smaller proteins are less accurately predicted, owing to the underrepresentation problem when building the SVR models.

### Table 1. Prediction performance in terms of CC and RMSE using five different sequence encoding schemes

<table>
<thead>
<tr>
<th>Measures</th>
<th>Encoding schemes</th>
<th>CC</th>
<th>RMSE_norm</th>
<th>RMSE_raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSE-up</td>
<td>LS</td>
<td>0.69</td>
<td>0.73</td>
<td>6.81</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>0.42</td>
<td>0.92</td>
<td>8.57</td>
</tr>
<tr>
<td></td>
<td>LS+SS</td>
<td>0.71</td>
<td>0.71</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA</td>
<td>0.71</td>
<td>0.71</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA+W+L</td>
<td>0.72</td>
<td>0.70</td>
<td>6.59</td>
</tr>
<tr>
<td>HSE-down</td>
<td>LS</td>
<td>0.65</td>
<td>0.77</td>
<td>5.62</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>0.44</td>
<td>0.90</td>
<td>6.60</td>
</tr>
<tr>
<td></td>
<td>LS+SS</td>
<td>0.67</td>
<td>0.75</td>
<td>5.49</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA</td>
<td>0.67</td>
<td>0.75</td>
<td>5.47</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA+W+L</td>
<td>0.68</td>
<td>0.74</td>
<td>5.43</td>
</tr>
<tr>
<td>CN</td>
<td>LS</td>
<td>0.72</td>
<td>–</td>
<td>8.57</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>0.50</td>
<td>–</td>
<td>10.78</td>
</tr>
<tr>
<td></td>
<td>LS+SS</td>
<td>0.74</td>
<td>–</td>
<td>8.38</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA</td>
<td>0.74</td>
<td>–</td>
<td>8.37</td>
</tr>
<tr>
<td></td>
<td>LS+SS+AA+W+L</td>
<td>0.76</td>
<td>–</td>
<td>8.16</td>
</tr>
</tbody>
</table>

All results were evaluated using 4-fold cross-validation method.

*RMSE\_norm; *RMSE\_raw.

### 3.5 Incorporating global sequence information significantly improves the prediction performance

As previous studies have indicated (Kinjo et al., 2005; Schlessinger et al., 2006; Song et al., 2007; Schlessinger et al., 2006; Song et al., 2007; Yuan, 2005), incorporating global sequence features might be helpful for improving the prediction accuracy. To achieve this, we utilized three global sequence descriptors, i.e., 20 amino acid compositions (‘AA’), protein sequence weight (‘W’) and sequence length (‘L’). For ‘W’ and ‘L’ descriptors, we encoded them into the SVR after the normalization using their mean sequence weights or lengths and SDs based on our dataset. We employed five different sequence encoding schemes, i.e., local sequence in the form of PSI-BLAST profiles (‘LS’), predicted secondary structure information by PSIPRED (‘SS’), local sequence plus predicted secondary structure (‘LS+SS’), local sequence plus predicted secondary structure and amino acid composition (‘LS+SS+AA’), and local sequence plus predicted secondary structure coupled with amino acid composition, sequence weight and sequence length (‘LS+SS+AA+W+L’). The prediction results for these encoding schemes are also summarized in Table1.

As expected, using ‘LS+SS+AA’, we achieved a slightly improved performance of RMSE\_raw = 6.66 for HSE-up and RMSE\_raw = 5.47 for HSE-down, respectively, although CC and RMSE\_norm remain at the same level. However, when combining sequence weight and sequence length information, ‘LS+SS+AA+W+L’ could predict HSE-up with CC of 0.72, RMSE\_norm of 0.70 and RMSE\_raw of 6.59, and HSE-down with CC of 0.68, RMSE\_norm of 0.74 and RMSE\_raw of 5.43, respectively, which is a more significant improvement compared to ‘LS+SS+AA’. These observations suggest that including either amino acid composition (‘AA’) or sequence weight (‘W’) or sequence length (‘L’) could yield the better prediction performance compared with local sequence alone, which coincides well with the previously reported importance of ‘W’ on the prediction performance of CN and RWCO (Song and Burrage, 2006; Yuan, 2005). Additionally, these results also indicate that protein size (represented by ‘W’ and ‘L’) is a very important factor that has more significant influence on the prediction performance than ‘AA’ in predicting HSE values, which is conceivable because residues in larger proteins may be slightly less exposed than in smaller ones and as a global descriptor protein size can globally determine the environment where its residues are located.

To further explore this protein-size effect, we plotted the CC and RMSE of each protein against its corresponding sequence length, as shown in Supplementary Figure1. We can see that most of the predicted proteins have CCs larger than 0.45 and RMSEs less than 6 in both cases of HSE-up and HSE-down, while some badly predicted proteins are also observed, especially for those with sequence lengths ranging from 100 to 400. These results imply that smaller proteins are less accurately predicted, owing to the underrepresentation problem when building the SVR models.

### 3.6 Analysis of the mean absolute errors

We next explored the mean absolute errors (MAEs) in different ranges of HSE and CN according to different secondary structures (Supplementary Table2). The overall percentages of residues with conformation annotations of the α-helix, β-strand and coil are 40.8, 26.8 and 32.4%, respectively. First, the MAEs will increase with the increasing values of HSE-up and HSE-down, as shown in Supplementary Table2. The results imply that smaller proteins are less accurately predicted, owing to the underrepresentation problem when building the SVR models.
Yuan, 2005) and tested these methods using the current dataset. Three observations can be made from this figure. First, the marginal regions are less adequately represented when feeding SVR models. Second, residues with HSE values in the range of 0–10 or 30–40) are less accurately predicted, as they have larger MAEs. It might be that the underrepresentation of these residues makes them less likely to be adequately represented when building SVR models.

The overall distributions of CC and RMSE of the tested proteins for the five sequence encoding schemes are presented in Supplementary Figure 2. In the case of HSE-up, the peak values of CC and RMSE are very close to 0.76 and 6, respectively, which can be regarded as the upper limits of the prediction performance of the encoding schemes employed here. Analogously, the peak values of CC and RMSE in the case of HSE-down are 0.72 and 5, respectively. All the distributions of CC and RMSE for HSE-up and HSE-down, taken together, suggest that the sequence-encoding scheme ‘LS+SS+AA+W+L’ leads to the best performance.

We also plotted the MAEs for all residues in the dataset with different HSE-up and HSE-down values, as given in Figure 5. Three observations can be made from this figure. First, the ‘LS+SS+AA+W+L’ encoding scheme leads to the least MAE for the majority of the regions in Figure 5 and hence provides the best prediction performance compared with the other sequence-encoding schemes. Second, residues with HSE-up = 13 and with HSE-down = 18 are predicted with the least MAEs. It may be that these residues are more adequately represented after being input into SVR models, as they have relatively larger number of samples in the current dataset. Third, residues with larger HSE values (>36) or smaller HSE values (<5) have larger MAEs and are hence worst predicted. Similarly, it may be that the residues located in the marginal regions are less adequately represented when feeding into SVR models.

3.7 Comparison with other methods

On one hand, since this study represents the first attempt to predict the HSE values from sequences, the objective comparison with other HSE-predicted methods is not available. On the other hand, the prediction comparison is meaningful only provided that it is performed using the same datasets and the same performance evaluation measures (Ofran et al., 2007). Therefore, in order to compare the performance of our method with other approaches, we implemented methods that were previously employed to predict CN with the PSSM profiles extracted using the local window size of 15 residues. Yuan (2005) used the SVR to predict CN with the PSSM profiles extracted using the local window size of 15 residues. Yuan (2005) also used the SVR approach to predict the CN values from sequence using the local PSSM profiles as well as AA and W information. As seen in Table 2, the CC of HSEPRED is 0.02 higher than that of Yuan’s method and 0.04 higher than that of Ishida’s approach, while the RMSE is 0.23 and 0.50 smaller than these methods. These results indicate that HSEPRED provides better prediction performance compared with the other two methods.

3.8 Case study

To better understand the CC and RMSE measures, we presented two prediction examples and showed their predicted HSE and CN profiles with the structural mapping of the MAE values on their 3-dimensional structures. This kind of figure shows to what extent the predicted and observed HSE and CN values match each other, providing more intuitive observation of the prediction performance. The first example is the Escherichia coli peptide Deformylase (PDB: 1xeo) bound to Formate, an enzyme that catalyzes the deformylation of nascent polypeptides generated during protein synthesis (Jain et al., 2005). It is well predicted with CC = 0.83 and RMSE = 3.43 for HSE-up, and with CC = 0.73 and RMSE = 3.97 for HSE-down, respectively. By summing up predicted HSE-up and HSE-down, CN can be predicted with CC = 0.88 and RMSE = 5.85, respectively. Most predicted values of this protein are in good agreement with the X-ray crystal structures.
Predict half-sphere exposure

Fig. 6. The predicted and observed HSE-up, HSE-down and CN profiles for the *E.coli* peptide Deformylase (PDF) bound to Formate (PDB code: 1xeo, chain A). For each subfigure, the predicted profile is at the left and the structural mapping with gradient from the best (red) to the worst prediction (blue) is at the right. The predicted and observed HSE and CN values are represented by dashed red and solid blue lines, respectively.

agreement with their observed HSE and CN values, with the exception that a small segment from residue positions 121 to 130 are poorly predicted in the case of HSE-up (Fig. 6A). While in the case of HSE-down, there exist two regions that were badly predicted: one is from residue positions 87 to 92 and the other is from residues 131 to 137 (Fig. 6B). For the CN measure, the region from residue positions 122 to 132 was worst predicted (Fig. 6C). We can readily see that the majority of the regions are colored by red, except that only small fragments including those at the tail of the helix and in the coiled region are colored by light blue, which again demonstrates that this protein is well predicted. At the tail-end of the helices, local sequence information of smaller window size less than 15 residues would be fed into the SVR model and consequently their representation is inadequate, which will in turn influence the prediction performance. In addition, coiled regions are also badly predicted. This might be that coiled residues that have no regular secondary structures are characterized by a variety of sequence features, thus making them less efficiently represented and difficult for the SVR to capture their intrinsic properties.

The second example is the *Bacillus subtilis* YfhH protein (PDB: 1sf9), a putative transcriptional regulator. In contrast, this protein is poorly predicted with CC of 0.62 and RMSE of 5.06 for HSE-up and with CC of 0.49 and RMSE of 5.07 for HSE-down, respectively. In all the three cases of HSE-up, HSE-down and CN, the worst predicted regions are from residue position 1 to 13, position 38 to 57 and position 88 to 98 (See Supplementary Fig. 3A, B and C). It is also clear that the HSE and CN values in the beginning region from residue 1 to 12 are strongly overpredicted. To conclude, these two examples presented here provide us a better understanding of the CC and RMSE measures, i.e. the higher the CC and the smaller the RMSE are, the better the prediction performance is.

3.9 Web server implementation

To facilitate the prediction of the HSE-up, HSE-down and CN measures from protein primary sequences, we have implemented an automated web server of our SVR approach called HSEpred, which is freely accessible at http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/hse/.
HSEpred has a user-friendly interface and only requires as input the FASTA format of the query sequence. Moreover, users can have three optional SVR models to designate as the prediction models, which are built based on the sequence-encoding schemes ‘LS’, ‘LS+S’ and ‘LS+S+A+W’, respectively. After the prediction is completed, users will immediately receive an e-mail containing the prediction result including the detailed residue positions, the predicted HSE-up, HSE-down and CN values as well as their predicted profile plots.

4 DISCUSSION

Precisely predicting amino acid solvent exposure bears great biological significance in protein structure and function prediction in that such information gives detailed description about the degree to which a residue interacts with other solvent molecules and its particular spatial arrangement with respect to other neighboring residues. Owing to this, researches into protein folding mechanism and rational protein drug design necessitates the prior knowledge of solvent exposure. Besides, active sites of a protein are often located on its surface, solvent exposure measures evaluating to what extent a residue is buried or exposed provide useful information for determining its functional roles. Hence, reliable prediction of solvent exposure from primary sequence could provide valuable insights for understanding and identifying protein sequence-structure-function relationship.

However, traditional solvent exposure measures like ASA, RD and CN have own limitations. As a new solvent exposure, while keeping strong correlations with ASA and CN, HSE has several attractive advantages that enable it to outperform other measures and make it more likely to be widely applied in the studies of protein-structure prediction and modeling analysis in the future, such as conservation within protein folds, applicability based on simplified model, amino acid dependency and predictability (Hamelryck, 2005). Indeed, a recent study has established that it is possible to reconstruct the backbone of small proteins solely from the HSE vectors of the native structures and that HSE-optimized structures are generally better than CN-optimized structures in terms of the RMSD and the angle correlation with the native structures (Palaszewski et al., 2006).

In this work, we proposed a new method for predicting HSE based on protein sequences only, which has been demonstrated to achieve high prediction accuracy in terms of CC and RMSE. As this is the first method to predict HSE measures from protein sequence, we provide a CN prediction comparison with other approaches. By summing up predicted HSE-up and HSE-down values, our method could provide much better prediction accuracy compared with other approaches (Yahida et al., 2006; Yuan, 2005) based on the current dataset. In addition, the results also indicate that taking advantage of both global sequence and local sequence information is beneficial to the prediction performance improvement. Moreover, we show that protein size in terms of ‘W’ and ‘L’ is a significant determinant of prediction performance, which is remarkable considering that ‘AA’ is a 20-dimensional vector while ‘W+L’ is only a two-dimensional vector. Using protein size information can lead to better prediction accuracy than using the amino acid composition, indicating that the HSE prediction performance depends considerably on the global protein size and, to a lesser extent, on its global amino acid composition.

Nevertheless, how to further improve the prediction accuracy will continue to be a challenging task, just like many problems in structural bioinformatics. There are several possible ways that may help to further improve the prediction performance in the future studies. First, with the more availability of PDB structures that are determined with better resolutions, using high-quality dataset will be helpful. Second, combining other informative sequence features, such as predicted solvent accessibility profiles (Ofran et al., 2007; Schlessinger et al., 2006), might help to improve the prediction performance. Third, efforts on how to effectively represent the under-represented proteins with lower sequence weights or lengths should be helpful. Increasing the ratio of these proteins in the dataset is likely to contribute to the performance improvement.

As a consequence of large-scale structural genomics projects, more sequenced data will be available in the near future. In this case, calculating solvent exposure from primary sequence could provide valuable insights for understanding and identifying protein sequence-structure-function relationship and functional properties, and should be useful in protein structure modeling, prediction and drug design.

5 CONCLUSION

In this study, we proposed a novel approach to predict the HSE measures from protein sequences based on SVR. Two local sequence descriptors (PSMs in the form of PSI-BLAST profiles and predicted secondary structure by PSIPRED) and three global sequence descriptors (amino acid compositions, sequence weights and sequence length) are utilized as the input to the SVR models. We extensively investigated five different sequence-encoding schemes to examine their different effects on the prediction performance. The prediction results illustrate the effectiveness of the proposed method for accurately predicting HSE values from the sequences. The successful application of the SVR approach demonstrates its predictive power in quantifying the sequence-structure relationship and estimating the protein structural property profiles from amino acid sequences. With the growing number of sequence data as the result of large-scale structural genomics projects, we anticipate that our method could be especially useful in analyzing the genome and proteome sequences where no structural data are available.

ACKNOWLEDGEMENTS

**Funding:** J.S. would like to thank the Japan Society for the Promotion of Science (JSPS) for financially supporting this research via the JSPS Postdoctoral Fellowship for Foreign Researchers. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

**Conflict of Interest:** none declared.
REFERENCES


