Sequence analysis

Prediction of zinc-binding sites in proteins from sequence

Nanjiang Shu, Tuping Zhou and Sven Hovmöller*

Structural Chemistry, Arrhenius Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden

ABSTRACT

Motivation: Motivated by the abundance, importance and unique functionality of zinc, both biologically and physiologically, we have developed an improved method for the prediction of zinc-binding sites in proteins from their amino acid sequences.

Results: By combining support vector machine (SVM) and homology-based predictions, our method predicts zinc-binding Cys, His, Asp and Glu with 75% precision (86% for Cys and His only) at 50% recall according to a 5-fold cross-validation on a non-redundant set of protein chains from the Protein Data Bank (PDB) (2727 chains, 235 of which bind zinc). Consequently, our method predicts zinc-binding Cys and His with 10% higher precision at different recall levels compared to a recently published method when tested on the same dataset.

Availability: The program is available for download at www.fos.su.se/~nanjiang/zincpred/download/

Contact: svenh@struc.su.se

Supplementary information: All Supplementary Data can be accessed at www.fos.su.se/~nanjiang/zincpred/suppliment

1 INTRODUCTION

About one-third of proteins in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977) contain metals, and it is estimated that approximately the same proportion of all proteins are metalloproteins (Holm *et al.*, 1996). Metal atoms are critical to the function, structure and stability of proteins. Zinc is the second (after iron) most abundant metal found in eukaryotic organisms (Coleman, 1992). Zinc plays important roles, mainly catalytic and structural, in many biological functions. For example, zinc ions serve as powerful electrophilic catalysts in many hydrolases and lyases (McCall *et al.*, 2000). Zinc-binding stabilizes the folded conformations of domains so that the protein can function properly (Berg and Shi, 1996), e.g. zinc-finger proteins. The biological roles of zinc have been extensively reviewed (Brewer *et al.*, 1983; Stefanidou *et al.*, 2006). The accurate prediction of zinc-binding sites is not only important for function annotation of proteins but also helpful for three-dimensional structure prediction.

Metal-binding sites have been predicted based on structural information (Gregory *et al.*, 1993; Sodhi *et al.*, 2004). Predictions from protein sequences only have received less attention. Early approaches can be found in the work of Nakata *et al.* (1995), in which they tried to predict zinc-finger

DNA-binding proteins with a neural network. Those approaches were limited by the scarcity of data at that time and the method was applicable only to certain types of zinc-binding proteins. Andreini *et al.* (2004) used a regular expression matching method supplied by PHI-BLAST (Zhang *et al.*, 1998) to explore copper-binding patterns. They showed that the confidence for a scanned pattern to be copper-binding was >90% when the percentage of identical amino acids aligned around the metal-binding pattern by PHI-BLAST was >20% with respect to the protein domain length. However, the success rate for zinc-binding sites was not available. A breakthrough for predicting zinc-binding sites from sequences was done by Menchetti *et al.* (2006) and Passerini *et al.* (2007). In their work, zinc-binding residues were predicted by a local predictor and a gated predictor using support vector machines (SVM). For the local predictor, all Cys and His (CH) were selected. Feature vectors which represented a window of residues centered at selected CH were encoded by the position specific substitution matrices (PSSM) from PSI-BLAST (Altschul *et al.*, 1997). For the gated predictor, residue pairs were picked out by scanning amino acid sequences with a semi-pattern [CH]x(0–7)[CH] (C is cysteine and H is histidine, x(0–7) stands for a consecutive substring of any amino acids with a length from 0 to 7). These selected residue pairs were encoded similarly as in the local predictor. The gated predictor combined the predictions of the local predictor and the semi-pattern predictor by a gating network. Their method predicted zinc-binding Cys and His with 60% precision at 60% recall based on a 5-fold cross-validation (Menchetti *et al.*, 2006). For the less common zinc-binding residues Asp and Glu, the result was less satisfactory. Passerini *et al.* (2006) described a method to predict metal-binding Cys and His based on a two-stage machine learning approach. The first step was similar to the local predictor in Menchetti *et al.* (2006). Individual Cys and His were encoded into feature vectors using PSSMs and global descriptors such as protein length and amino acid composition. These feature vectors were then classified by SVM. After that, a three layer bi-directional recurrent neural network (BRNN) was used to further distinguish metal-binding and non-metal-binding Cys and His. For zinc-binding Cys and His, SVM-BRNN predicted with 75% precision at 50% recall.

The rapidly increasing number of high-quality structures deposited in PDB and the availability of PSI-BLAST, which provides a reliable multiple sequence alignment among protein families, encouraged us to predict zinc-binding proteins from sequences on a database scale, using the evolutionary

---

*To whom correspondence should be addressed.

information. We have developed an improved method for predicting zinc-binding sites from sequences, focusing on four amino acids Cys, His, Asp and Glu (CHDE), since these four amino acids account for about 96% of all zinc-binding residues (Table 1). The method was tested on the same non-redundant set of PDB chains as used in Passerini *et al.* (2006).

2 METHODS

Our method consists of an SVM-based predictor and a homology-based predictor. For the SVM-based predictor, CHDEs were selected in both training set and test set and were encoded into single-site vectors and pair-based vectors which represented a window of residues centered at each selected CHDE or a pair of selected CHDEs respectively (see Section 2.2). SVM was then used to train feature vectors of the training set and to make the prediction on the test set. The publicly available Gist SVM package (version 2.1.1) (Pavlidis *et al.*, 2004) was used to implement SVM. The radial basis kernel was used. SVM predictions on individual selected residues were obtained by combining the predictions using single-site vectors and pair-based vectors with a gating network. For the homology-based predictor, each target chain in the test set was searched against the training set for remote homologues using a segment matching method (see Section 2.4). Predictions of zinc-binding residues were made based on predicted homologues in the training set. The final predictions were obtained by a consensus of SVM predictions and homology-based predictions. The whole prediction procedure is illustrated in Figure 1.

2.1 PSSM and conservation level

PSSM profiles were obtained by running PSI-BLAST (version 2.2.13) against the NCBI nr database (version April 2006) for three iterations with an E-value threshold of 0.001. The conservation score for residue k in the sequence $ConScore_k$ was calculated as follows,

$$ConScore_k = \frac{M_{k,a(k)} - MIN\_M_{a(k)}}{MAX\_M_{a(k)} - MIN\_M_{a(k)}} \quad (1)$$

$$MIN\_M_j = \min_{\substack{\text{over all residues} \\ \text{with } a(k)=j}} (M_{k,j}), \quad j = A, R, N \ldots, V(20 \text{ amino acids}) \quad (1.1)$$

$$MAX\_M_j = \max_{\substack{\text{over all residues} \\ \text{with } a(k)=j}} (M_{k,j}), \quad j = A, R, N \ldots, V(20 \text{ amino acids}) \quad (1.2)$$

where $a(k)$ was the amino acid type (one of the 20 amino acids) of residue k, $M_{k,a(k)}$ was the log-odd score of the profile at position k (the profile at each sequence position contains 20 items corresponding to 20 amino acids) on amino acid $a(k)$, MAX_M_j and MIN_M_j were the maximum and minimum log-odd scores of profiles for all residues with amino acid type j on amino acid j. The conservation score ranged from 0 (un-conserved) to 1 (most highly conserved). For example, for Cys, $MAX\_M_C = 12$ and $MIN\_M_C = -5$. Then, for a Cys with the log-odd score on C (cysteine) equaling 7, the conservation score is calculated as 0.71.
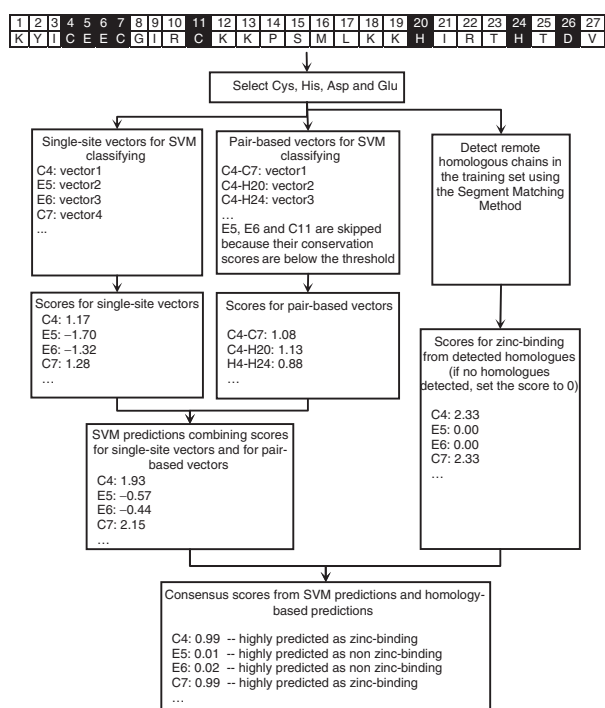
2.2 SVM-based predictor

A feature vector represented the conservativity and physicochemical properties of selected amino acids which were either zinc-binding or not. SVM were used to classify these feature vectors as either positive (zinc-binding) or negative (non-zinc-binding). A single-site vector represented a window of residues centered at each selected CHDE. SVM predictions using single-site vectors resulted in a basic prediction for all CHDEs. Biologically significant Zn atoms (i.e. Zn3 and Zn4) are

Table 1. Number of residues bound to each type of Zn atom

	C	H	D	E	Others	Subtotal	No. of Zn atoms	No. of chains
Zn1[a]	1	10	9	10	3	33	34	19
Zn2[a]	3	32	15	26	7	83	45	37
Zn3[a]	25	134	54	30	7	250	89	73
Zn4[a]	499	190	41	24	15	769	205	148
Zn5[a]	7	1	0	0	2	10	2	2
Co-cat Zn[b]	46	59	38	22	10	175	67	35
Subtotal	535	366	116	85	24	1136	375	235
Subtotal[c]	531	325	92	51	24	1023	295	210

[a]Zn1, Zn2, Zn3, Zn4 and Zn5 are Zn atoms binding to 1, 2, 3, 4 and 5 amino acid residues, respectively. [b]Co-catalytic Zn: Zn atoms bind to 3, 4 or 5 amino acids and are bridged to another metal atom(s) via side chain atoms or water molecules. [c]Subtotal for Zn3, Zn4, Zn5 and Co-Catalytic Zn.

Fig. 1. Flowchart for the overall prediction method. SVM predictions and homology-based predictions are combined into the final consensus prediction.

bound to 3 or 4 residues, and these residues are expected to be correlated. Pair-based vectors interpreted the correlation between these residues by taking a pair of selected CHDEs as centered residues in a window. The SVM predictions using single-site vectors and pair-based vectors were combined by a gating network described in Menchetti *et al.* (2006). The SVM outputs had been converted into conditional probabilities using a sigmoid function suggested in Platt (2000) in order that they could be combined by the gating network. The sigmoid function was defined as

$$P(Y = 1|x) = \frac{1}{1 + \exp(A^*f(x) + B)} \quad (2)$$

where x was the SVM input of each test instance, $f(x)$ denoted the margin of test instance x, $P(Y = 1|x)$ was the probability of zinc-binding

Table 2. Comparison of zinc-binding residues (those bind to Zn3, Zn4, Zn5 and Co-catalytic zinc) selected by the pair selecting method based on highly conserved CHDEs and semi-pattern [CHDE]x(0,7)[CHDE]

Method	No. of selected residues	No. of selected zinc-binding residues	Recall (%)	Precision (%)
Conserved CHDEs[a]	12 969	869	87.0	6.7
Semi-pattern[b]	59 642	611	61.2	1.0

[a]Residues with conservation score (described in Methods) ≥0.75 were considered as highly conserved. Residues in the pair were separated by less than 150 residues in sequence. [b]Semi-pattern [CHDE]x(0,7)[CHDE] means C, H, D or E followed by C, H, D or E within 0 to 7 residues according to Menchetti *et al.* (2006).

prediction, and *A* and *B* were the slope and offset respectively to be learned [by a 3-fold cross-validation method suggested in Platt (2000)] from the training set for the sigmoid function. Empirically, one could use $A = -2.0$, $B = -0.5$ for single-site vectors and $A = -4.0$, $B = -0.5$ for pair-based vectors. In order that the scores of SVM predictions by the gating network can be combined linearly with the scores of homology-based predictions, we converted conditional probabilities back to SVM functional margins by

$$f(x) = \frac{\ln((1 - p)/p) - B}{A} \qquad (3)$$

where $f(x)$ was the SVM prediction score for each test instance x, p was the conditional probability for zinc-binding by the gating network, and *A* and *B* were the same as used in the sigmoid function. Equation (3) is actually an inverse function to Equation (2).

To select the residue pairs, Menchetti *et al.* (2006) employed the [CHDE]x(0–7)[CHDE] semi-pattern (meaning C, H, D or E followed by C, H, D or E within 0–7 residues) to pick out such pairs. However, that method missed about 40% of all zinc-binding CHDEs (Table 2). We selected residue pairs by first picking out all CHDEs with conservation scores ≥0.75 and then picking out all pairs of these conserved CHDEs which are separated by less than 150 residues in sequence. Our method identified more zinc-binding CHDEs, with less selected residues (Table 2). Details about the encoding of the single-site vectors and pair-based vectors can be found in the Supplementary Data.

2.3 Homology-based predictor

Generally speaking, if a protein has a homologous protein that is zinc-binding, the probability that this protein is also zinc-binding is much higher, especially when this protein has a sequence pattern similar to the homologous zinc-binding protein. Therefore, if a chain in the training set was found with a similarity score (see Section 2.4.2) higher than a certain threshold, e.g. 25.0, information such as metal binding sites and disulfide-bonds was utilized to calculate the zinc-binding score for selected CHDEs of the target chain. If the matched region in the detected chain was

(a) Non-zinc-binding, set a negative score −ZS to all selected CHDEs

(b) Disulfide-bonded, set a negative score −ZS to all Cys.

(c) Zinc-binding, set a positive score ZS to residue groups that best match patterns of residues bound to a Zn atom.

ZS was defined as ZS = SS * scale, where SS was the homology score (described in Section 2.4.2) for the predicted remote homologue. Scale was defined as $\sqrt{N}/50$, where N was the number of homologues predicted for the target chain. Scale was divided by 50 so that the average ZS score derived from the predicted homologues was the same as the average score

derived from SVM predictions. The zinc-binding scores of individual residues of a target chain were calculated as the average of similarity scores derived from all predicted homologues to that chain.

2.4 Segment matching method for finding remote homologues

Homologues were predicted by a segment matching method. It contains two steps.

2.4.1 Segment matching For each chain in the test set, a sliding window of nine residues was searched against all nine-residue segments in the training set (note that the window here was different from the window used in SVM feature vector encoding, the former was centered at any residue in the sequence, while the latter was centered at a selected CHDE). The similarity between this nine-residue segment and its corresponding segment in the training set was defined as

$$\text{Score}(\alpha, \beta) = \sum_{n=1}^{9} \left(\sum_{i=1}^{20} \left(\alpha_{ni} \log\left(\frac{\beta_{ni}}{P_i}\right) + \beta_{ni} \log\left(\frac{\alpha_{ni}}{P_i}\right) \right) \right) \qquad (4)$$

where α and β were profiles of a nine-residue segment in the test set and a corresponding segment in the training set, P was the background frequency for 20 amino acids. The profile-profile score between two residue positions is the same as the PICASSO3 score as suggested by Mittelman *et al.* (2003).

2.4.2 Finding remote homologues For each segment of a chain in the test set, up to 100 fragments from the training set with the highest similarity scores as defined in Equation (4) were kept, including the PDB code and segment position for each of these closest matches. If a homologous protein chain exists in the training set, many fragments from this chain tend to have high-similarity scores with the corresponding fragments in the target chain, which means that the PDB code of this chain will appear frequently in the list of matched segments. We made use of this for finding remote homologues in the following way. For those protein chains that appeared often (typically >10% of the number of residues of the sequence) in the list, a dot plot was made. This diagram was drawn by setting the *X* values as positions of the central residues of the nine-residue segment of the target chain in the test set and the *Y* values as positions of corresponding segments in the matched list (see Fig. 2A). The dots on such diagrams tend to be clustered into consecutive lines. This is not surprising, since two adjacent nine residue segments have eight residues in common, and thus will often pick up two adjacent nine residue segments from the same protein. When such a diagram is composed of just a few long consecutive lines of dots, it is very likely that the matched protein is homologous to the target chain in the test set. A homology score was derived from the length and linearity of the pattern divided by the average sequence length of these two chains. The algorithm for deriving the homology score is summarized as follows,

Initializing the score on the dot plot:
For each diagonal on the dot plot

(1) Set the score of each dot as the number of consecutive dots

(2) Record the position of the diagonal with the top 10 longest consecutive dot segments

Calculating the score of each diagonal:
For each diagonal with the top 10 longest consecutive dot segments
The homology score is calculated as the sum of

(1) The scores of all dots on the diagonal and

(2) The scores of all dots on the other diagonals divided by the distance to that diagonal
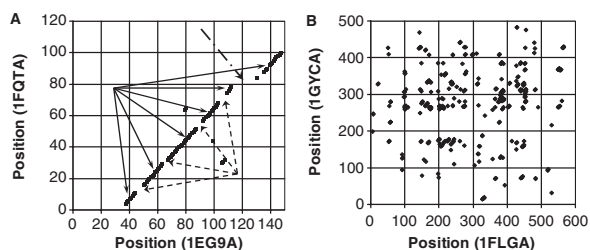
Fig. 2. Dot plots for the sequence number of the central residue of matched nine-residue segments between (**A**) 1EG9A (449 amino acids) and 1FQTA (112 amino acids) and (**B**) 1FLGA (582 amino acids) and 1GYCA (499 amino acids). Each dot on the dot plots represents a pair of nine-residue long fragments with high similarity score as defined in Equation (4). In the dot plot of (A), a high homology score (score = 41, when cutting the homology score at threshold 25, 86% of all predicted homologues are real homologues according to the SCOP (Andreeva *et al.*, 2004), see Section 3.2) was predicted from consecutively and linearly distributed dots. Intuitively, one would also expect homology between 1FQTA [1:100] and 1EG9A [40:150], and in fact, they are homologous according to the SCOP definition (these two domains belong to the same SCOP superfamily, the ISP domain). The sequence identity between these two domains is barely 15%, showing the success of the Segment Matching Method in remote homology detection. The consecutive dot segments as marked by solid arrows predicted the conserved homologous regions between 1FQTA and 1EG9A. Gaps marked by dashed arrows predicted the varied regions and the gap marked by the dashed-dot arrow predicted an insertion in the chain 1FQTA or deletion in 1EG9A in that region of the sequence. On the other hand, in the dot plot of (B), the dots are not forming long lines as in (A), though there are also a great number of dots. Low homology score (score = 6) was predicted between 1FLGA and 1GYCA meaning that they are not likely to be homologous. In fact, 1FLGA and 1GYCA are not homologous according to the SCOP definition (they belong to different folds in the SCOP, b.70.1.1 for 1FLGA and b.6.1.3 for 1GYCA).

Calculating the homology score for the whole chain:
Set the highest homology score of all diagonals as the homology score of the whole chain

The homology scores for all predicted homologues were normalized to [0, 100] by setting the lowest homology score to 0 and the highest to 100. For multi-domain sequences, the range of the homologous part was also predicted from the extent of these linear and consecutive patterns. In some cases many high-scoring nine-residue segments were detected from a single chain, but they were rather randomly spread out over the dot plot (see Fig. 2B). Such cases we did not predict as likely homologues.

2.5 Performance measurement

The Precision was defined as $TP/(TP + FP)$, where TP (true positives) referred to the number of correctly identified positive examples (zinc-binding residues or proteins); FP (false positive) was the number of negative examples (residues or proteins predicted to bind zinc, although they do not bind zinc according to PDB) that were incorrectly predicted as positive. The Recall was defined as $TP/(TP + FN)$, where FN (false negative) was the number of positive examples that were incorrectly predicted as negative. In this work, the negative examples are far more than the positive examples. The negative to positive ratios are 26:1 and 93:1 for CH and CHDE, respectively. For such an unbalanced dataset, receiver operating characteristic (ROC) curves can present an overly optimistic view of the performance of a method (Davis and Goadrich, 2006). The Recall–Precision curve, in which one plots the Precision

against the Recall, has been proposed as an alternative to the ROC curve in dealing with datasets with great unbalance in the class distribution (Zhang *et al.*, 2004). The Area Under the Recall-Precision Curve (AURPC) was used in our method for both model selection and performance measurement. AURPC was calculated by a method proposed by Davis and Goadrich (2006).

2.6 Data and statistics

We used a non-redundant set of PDB containing 2727 protein sequences as used in Passerini *et al.* (2006) to test our method. This dataset was culled at zero HSSP (Homology derived Secondary Structure of Proteins) distance by uniqueProt (Mika and Rost, 2003). These 2727 chains contain 564 444 residues, including 9202 Cys, 13 663 His, 32 466 Asp and 38 299 Glu. Out of these 2727 chains, 731 bind to at least one metal atom. Among these 731 chains, 235 bind to at least one Zn atom (see Supplementary Data). The total number of residues binding to zinc was just over 1000, as shown in Table 1. Metal atoms were considered binding to the protein chain if there were any nitrogen, oxygen or sulfur atom of the residues on the chain located within 3.0 Å to the metal atom [for specific distances between metal and atoms of the residue see Harding (2004)]. Some residues that have atoms (e.g. in chains 1I3QA, 1I3QL, 1IRXA, 1CVRA and 1PGUA) located between 3.0 and 3.5 Å to Zn atoms were manually curated as binding to zinc (see Supplementary Data). Accordingly, the number of Zn atoms and zinc-binding Cys and His slightly differed from Passerini *et al.* (2006).

Most Zn atoms (78%, see Table 1) bind to three or four residues (called Zn3 and Zn4, we annotated Zn atoms coordinated by m residues as Znm), i.e. 90% of all zinc-binding Cys, His, Asp and Glu are Zn3 or Zn4 binding. Zn atoms that bind to four residues but having no bound water molecules are considered as structural zinc, while those binding to three residues are generally catalytic zinc (Auld, 2001). An example of a protein containing both Zn3 and Zn4 is alcohol dehydrogenase, PDB code 2OHX (Al-Karadaghi *et al.*, 1994) (see Supplementary Data for the illustration of the structure). Many Zn3 and Zn4 atoms have other metal atoms nearby, bridged by a side-chain atom or a water molecule. The activities of zinc-enzymes require these bridged metal atoms working together. Such Zn atoms are called co-catalytic zinc. The Zn atoms which bind to only one or two residues are generally located on the surface of proteins. They are most probably bound to the proteins during crystallization (McPherson, 1999) but have no biological function. We focused here on predicting biologically important zinc-binding sites, i.e. structural (Zn4), catalytic (Zn3) and co-catalytic zinc-binding sites. Inter-chain Zn atoms, e.g. Zn atoms that bind to two residues in one chain and one residue in another chain, and one Zn5 atom (see Table 1) were also included. There were in total 295 such Zn atoms, bound to 531 Cys, 325 His, 92 Asp and 51 Glu (Table 1).

3 RESULTS AND DISCUSSIONS

3.1 Comparison with other methods

We evaluated the prediction of zinc-binding sites on both protein level and residue level through a 5-fold cross-validation just as in Passerini *et al.* (2006). A zinc-binding protein chain was considered as correctly predicted if there was at least one zinc-binding residue correctly predicted. The overall best results on the residue level and protein level for CH were obtained for $k = 12$ and $w = 5$ (k is the length of extension along the amino acid sequence on both sides of the centered residue in a window and w is the constant used when encoding the pair-based vectors, see Supplementary Data for more details), giving an average AURPC of 0.723 and 0.701, respectively. AURPC over 5-fold cross-validation for final predictions (Fig. 3A and B) show

that our predictions for Cys and His are about 10% higher in precision at different levels of recall than those of Passerini *et al.* (2006). Note that Passerini *et al.* predicted not only Zn3 and Zn4 binding residues, but also Zn2 binding residues. The inclusion of Zn2 binding residues tends to lower the prediction accuracy since many of them might not be biologically bound. Nevertheless, the inclusion of Zn2 as was done in Passerini *et al.* would not have a great impact on the prediction accuracy since Zn2 binding Cys and His take up only <4% of all zinc-binding Cys and His (see Table 1). In addition, the predictor of Passerini *et al.* is motivated in predicting several different transitional metal binding sites, and Zn binding site is only one of them. As a consequence, their predictor might not be specifically optimized for the prediction of zinc-binding sites. Note also that in Passerini *et al.* (2006), positive examples are proteins containing zinc-binding sites and negative examples are non-metalloproteins. In Section 3.4, we showed that residues binding to metals other than zinc, e.g. iron, were sometimes falsely predicted as zinc-binding. The exclusion of non-zinc metalloproteins from the negative examples tends to simplify the zinc-binding prediction and thus yields overoptimistic results. In our study, positive examples are zinc-binding CHDEs (999 residues in 208 chains) or zinc-binding CHs (856 residues in 199 chains) and the negative examples are all the rest of CHDEs (92 643 residues) or CHs (22 020 residues). Despite all these differences, the outperformance of our method to that of Passerini *et al.* is significant.

Passerini *et al.* (2007) evaluated their zinc-binding residue prediction method on a dataset containing 2428 chains. In their 5-fold cross-validation, no two chains having a domain belonging to the same SCOP (Andreeva *et al.*, 2004) super-family exist in the same cross-validation fold. In such a cross-validation procedure, homology-based prediction is not valid since no homologues exist. We tested our SVM predictor also on this dataset and with the same 5-fold separation. The SVM predictions of our method for CH on residue level achieved an AURPC of 0.617 (Fig. 3C). Our method outperformed that of Passerini *et al.* (2007) who obtained an AURPC of 0.500 when tested on the same dataset and the same cross-validation separation. Note that the predictor of Passerini *et al.* (2007) predicted Zn2 binding residues as well. However, most of the Zn2 atoms which were annotated as interface Zn atoms in Passerini *et al.*, (2007) were included in Zn3 or Zn4 by our method since they do bind to 3 or 4 residues, but on several different chains. Therefore, the prediction results of our method and that of Passerini *et al.* (2007) are comparable.

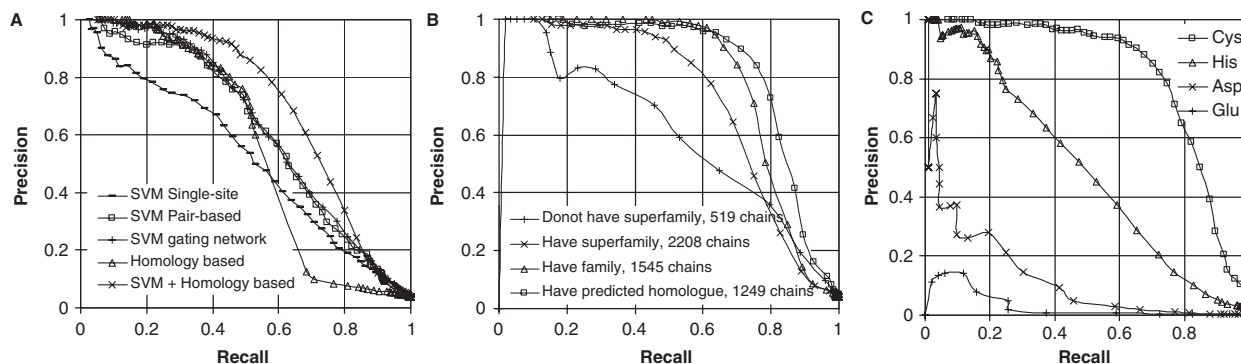## 3.2 SVM predictions versus homology-based predictions

SVM predictions using pair-based vectors performed better than single-site vectors in Recall–Precision curves on both residue level and protein level, especially at the low recall/high-precision region (Fig. 4A). This outperformance may be due to the potential correlation between zinc-binding residues as represented in pair-based vectors. The SVM predictions combining the predictions from single-site vectors and pair-based vectors were not better than predictions using only pair-based vectors. The function of the gating network was just to add predictions using single-site vectors on residues that cannot be predicted
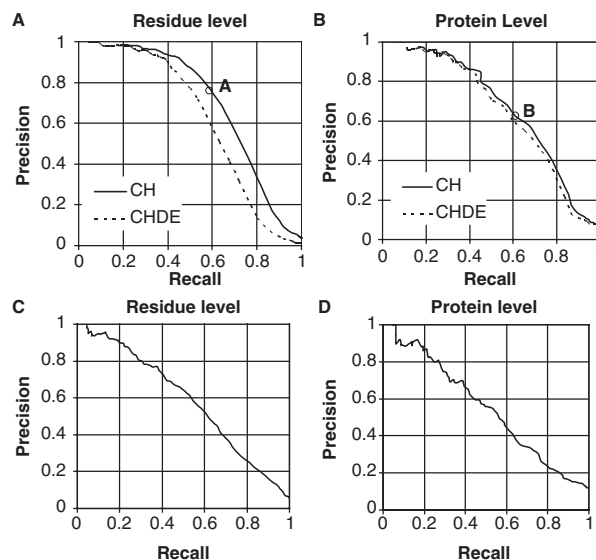
Fig. 3. Recall-Precision curves of 5-fold cross-validation of (**A**), (**B**) the final consensus predictions for CH and CHDE in the dataset of Passerini *et al.* (2006) (containing 2727 chains) and (**C**), (**D**) the SVM predictions using a gating network for CH in the dataset of Passerini *et al.* (2007) (containing 2428 chains) on residue level and protein level. At the point A, 60% of all zinc-binding Cys and His are found and 76% of the predicted zinc-binding Cys and His actually bind to zinc. At the point B, 60% of protein chains that have Cys and His binding to zinc are found and 63% of the protein chains predicted to have Cys and His binding to zinc actually bind to zinc.

using pair-based vectors (e.g. 13% of all zinc-binding CHDEs were missed by the pair selecting method based on highly conserved CHDEs). Note that the Recall–Precision curve of the SVM prediction (Fig. 4A) is quite similar to that of Passerini *et al.* (2006). It seems a simple SVM prediction using pair-based vectors did as well as a more sophisticated BRNN procedure as used by Passerini *et al.* (2006) in predicting zinc-binding sites. On the other hand, the outperformance of our method seems to come mainly from the homology-based predictor.

However, homology-based predictor alone predicted zinc-binding residues with even lower accuracy than the SVM predictor (Fig. 4A). The AURPC for CH on residue level were 0.579 and 0.650 for homology-based and SVM-based predictions, respectively. Then, why did the consensus, using both the homology-based predictor and the SVM predictor, produce better results than either one of the methods, namely AURPC 0.723? When cutting the scores of SVM predictions at 0.3 [see Equation (3)] threshold, zinc-binding Cys and His were predicted with 56% precision at 60% recall, yet still 406 Cys and His were actually non-zinc-binding. However, the homology-based predictor predicted these 406 Cys and His with low scores. The average score of SVM predictions for these 406 residues was 0.518 while it was only −0.435 for homology-based predictions. It means for these 406 residues, the inaccurate predictions by the SVM predictor were compensated by the more accurate predictions by the homology-based predictor. On the other hand, when cutting the scores of homology-based predictions at a threshold of 0.3, zinc-binding Cys and His were predicted with 78% precision at 47% recall, yet 115 Cys and His were

Fig. 4. Recall-Precision curves of the five-fold cross-validation on residue level of (**A**) predictions for CH by SVM predictor using single-site vectors, pair-based vectors and a gating network, and homology-based predictor and final consensus predictor of all 2727 chains, (**B**) the final consensus predictions for CH in different subsets of 2727 chains and (**C**) the final consensus predictions for Cys, His, Asp and Glu separately of all 2727 chains.

false positives. The average homology-based prediction score for these 115 residues was 0.666 but only 0.079 for SVM predictions. The inaccuracy in homology-based predictions was compensated by the SVM predictions. In conclusion, the consensus using both SVM predictor and homology-based predictor resulted in an overall significantly better result.

Although the dataset we used was a non-redundant set of PDB which did not have any two sequences with HSSP distance more than 0 (equivalent to amino acid sequence identity <20% when sequence length is 300), distant homologues still existed within the dataset. About 80% of all chains in this dataset (2208 out of 2727) had at least one chain within the same SCOP superfamily and 57% (1545 out of 2727) in the same SCOP family in the training set. It significantly helped the prediction of zinc-binding if there was another protein from the same SCOP family or superfamily in the dataset. The AURPC for CH on the residue level for 2208 chains which had homologues on the superfamily level was 0.740, and as high as 0.813 for the 1545 chains which had family level homologues (Fig. 4B). Our segment matching method detected these remote homologues successfully. When cutting the homology score at a threshold of 25.0, 4796 homologues were predicted for 1249 chains (many query chains have more than one predicted homologues). Out of these 4796 predicted homologues, 4126, i.e. 86% are indeed homologous to their corresponding query chains according to the SCOP (within the same SCOP superfamily). Out of those 1249 chains, 1194 (i.e. 96%) of them have at least one predicted homologue that really is a homologue according to the SCOP, and they cover 54% (1194 out of 2208) of all chains that have homologues in their corresponding training set. More details of the segment matching method as well as benchmarking will be described in a coming article.

On the other hand, there were 519 proteins without any homologue in the training set, as defined by the SCOP. For those, the AURPC was only 0.621, which was in accordance with another one of our predictions based on the dataset of Passerini *et al.* (2007) with homologues removed in the cross-validation separation (see Section 3.1).

Now, look again at the Recall–Precision curve of the final prediction of CH for the 1249 chains that each has at least one homologue (from the training set) predicted by our Segment Matching Method. The AURPC for those 1249 chains was 0.854. Zinc-binding Cys and His were predicted with 90% precision at 70% recall (Fig. 4B). This encouraging result means that if a chain has homologues predicted in the training set (not necessarily confirmed to be homologues), zinc-binding Cys and His can be predicted with 90% precision at 70% recall compared with the overall performance for all 2727 chains (56% precision at 70% recall, Fig. 3A). With more and more protein structures deposited in PDB, >65% of the newly added proteins are estimated to have at least one homologue in the SCOP domain database (Ekman *et al.*, 2005). All such proteins can now be predicted at great accuracy for zinc-binding sites.

3.3 Predictions for different residues

Of the four amino acids that bind to zinc, Cys, His, Asp and Glu, Cys was predicted with the highest accuracy. At the 60% recall level, the precision of Cys was 93%, but it was only 35, 3.0 and 1.0% for His, Asp and Glu (Fig. 4C). This is mainly because of the higher percentage of Cys that binds to zinc. The proportion of each amino acid that binds to zinc is Cys: 5.8%, His: 2.4%, Asp: 0.28% and Glu: 0.13%. The precision ratio for Cys, His, Asp and Glu at 60% recall, divided by their corresponding proportion of zinc-binding residues out of all residues was 16:15:11:8. When looked upon in this way, Cys and His were only slightly better predicted than Asp and Glu. This slightly better performance may be due to the fact that zinc-binding Cys and His are more conserved than zinc-binding Asp and Glu. For example, almost 97% of the Cys and 90% of the His that bind to zinc have conservation score over 0.7, but only about 60% of the zinc-binding Asp and 31% of the zinc-binding Glu are so conserved (see Supplementary Data). The introduction of Asp and Glu did not improve the overall prediction accuracy. The prediction results on Asp and Glu alone might not be useful for biologists given the current prediction accuracy. However, the introduction of Asp and Glu might help when the whole pattern of the zinc-binding site needs to be predicted.

3.4 False positives predicted with high confidence

We looked in detail into several cases where our zinc prediction failed. Eighteen proteins were predicted as zinc-binding with

>90% confidence, yet do not bind to zinc according to PDB. Most of these false positive chains bind to iron. Some have a series of closely located Cys or His binding to iron–sulfur clusters (1B25A and 1E7PB), iron oxides (1E5DA) or heme groups (1BVB). Others have four Cys tetragonally bound to Fe^{2+} (1B13A and 1B71A). In 1B13A and 1B71A, iron-binding sites are actually quite similar to zinc-binding sites and at such sites the Fe^{2+} could possibly be substituted by Zn^{2+}. Three chains (1CW0A, 1M65A and 1B71A) were grouped as false positives although the proteins bind to zinc but were not predicted at exactly the correct positions. For example, in 1CW0A, four residues Cys66, His71, Cys73 and Cys117 that bind to zinc are all highly conserved (conservation score >0.9). However, there is another conserved histidine nearby: His69. Our procedure is not capable of determining which one of His69 and His71 that actually binds to zinc, or perhaps both His69 and His71 bind, so that only two of those three Cys bind. Two chains (1JR8A and 1D0GR) are disulfide-bonded. Disulfide-bonded Cys were sometimes mis-predicted as zinc-binding since many of them are also highly conserved and closely located. One chain (1AW6) binds to cadmium while according to the literature (Baleja *et al.*, 1997) and the SCOP classification (Zn2/Cys6 DNA-binding Domain), it is actually a zinc-binding protein. We have classified 1AW6 as a zinc-binding protein.

Seven chains (1J6OA, 1K5KA, 1NJQA, 1NBFA, 1QXFA, 1BHI and 1N5GA) are neither metal-binding nor disulfide-bonded according to PDB. Some protein chains indeed have the three or four putatively zinc-binding residues closely located in 3D space (see Supplementary Data for an example of 1J6OA). 1QXFA (Herve du Penhoat *et al.*, 2004), 1BHI (Nagadoi *et al.*, 1999) and 1N5GA (Evanics *et al.*, 2003) are actually recorded as zinc-binding proteins according to the literature. These three proteins were also classified as zinc-binding in this study. The reason for the absence of zinc in the PDB files might be that the proteins were purified in zinc-free buffers. This shows that our method for predicting zinc-binding proteins is so powerful that it can find 'errors' in PDB.

3.5 False negatives predicted with low confidence

Some proteins that bind to zinc in PDB were strongly predicted as non-zinc-binding. For example, 11 protein chains (1A0B, 1B0NA, 1B55A, 1EC5A, 1EVKA, 1F35A, 1GL4A, 1JI3A, 1JKEA, 1PGUA and 1UDVA) that bind to zinc according to PDB were predicted at confidences lower than 10%. Four of them (1A0B, 1B0NA, 1F35A and 1JKEA) were crystallized in a condition with high-zinc concentrations (5 mM ~ 300 mM). It is likely that these proteins do not bind Zn *in vivo*. For example, 1B0NA is actually not a zinc-binding protein *in vivo* (Lewis *et al.*, 1998).

3.6 Anticipating an upper limit for zinc-binding prediction

When predicting zinc-binding sites from amino acid sequences, we have implicitly assumed that the sequences fully determine whether a protein binds zinc or not. However, proteins may exist under different conditions *in vivo*, e.g. binding or not binding to zinc. There are also many protein structures deposited in PDB, after experiments such as residue mutations, metal substitution as well as removing and adding of metals. As a consequence, there are proteins in PDB which natively bind to zinc but Zn atoms have been lost or substituted during experiments

(e.g. 1QXFA, 1BHI and 1AW6). There are also proteins in PDB which do not bind to zinc *in vivo*, but do bind zinc when soaked with high concentrations of zinc during crystallization (e.g. 1B0NA mentioned above). Luckily, most of these soaked Zn atoms are not strongly bound and usually bind to only one or two residues located on the surface of the protein. Such cases are easy to identify. More serious problems are caused by protein structures that have lost their Zn atom(s) or have had Zn atom(s) substituted by other metals during purification or crystallization, or the reverse case, where zinc has substituted other metals that bind *in vivo*. All such cases limit the outcome of the prediction of zinc-binding from sequences.

3.6.1 Effect of possible errors in PDB on the prediction We have recovered some obviously mis-annotated zinc-binding proteins in PDB by investigating the literature and the SCOP (for a complete list see Supplementary Data) manually. However, zinc-binding states for most proteins are still based entirely on PDB data. How far then can we trust the zinc-binding prediction, based on PDB data with an unknown (but with all likelihood very small) number of errors?

A comprehensive verification of zinc-binding states for all proteins in the PDB is outside the scope of this article. We are not aware of any manually curated database for metal-binding sites in proteins. Nevertheless, by artificially introducing errors about zinc-binding in PDB and carrying out prediction on such datasets with different error rates, we could estimate the effect of incorrectness in PDB on the performance of predictions. For this purpose, we randomly assigned a certain fraction of residues that bind zinc to non-zinc-binding. After that, the same procedure as was applied on the original PDB was carried out on this manually modified 'PDB'. As expected, the performance of the prediction decreases as the error rate increases, as shown in Figure 5. For example, the precision for CH on the residue level dropped by about 5% at 50% recall when 10% of the zinc-binding residues were annotated as non-zinc-binding. It is reasonable to assume the error rate in PDB regarding zinc-binding to be well under 10%. Then, the prediction for CH on the residue level based on the current PDB cannot be more than 5% worse in precision at 50% recall level, compared to a prediction based on a putative perfect 'PDB', according to the trend in Figure 5.

3.7 Predicting the whole pattern of zinc-binding sites

In this work, we focused only on predicting the zinc-binding state of individual residues, i.e., whether a Cys, His, Asp or Glu binds to zinc or not. However, the prediction of the whole zinc-binding site, especially the prediction of which three or four residues that bind to the same Zn atom, might also be very interesting. The accurate prediction of the whole zinc-binding pattern will be of great help to the engineering of metal-binding sites in proteins and it will also be a great help for 3D structure prediction, since it reduces the freedom in protein structure prediction enormously. Nevertheless, we can already see some success in the prediction of the whole zinc-binding pattern in this work. Cutting at 60% confidence level (for Cys, His, Asp and Glu), 38% (113/295) of all Zn3 and Zn4 binding sites were exactly predicted, and this number increased to 54% (159/295) when one residue tolerance was allowed (i.e. the whole Zn3 or
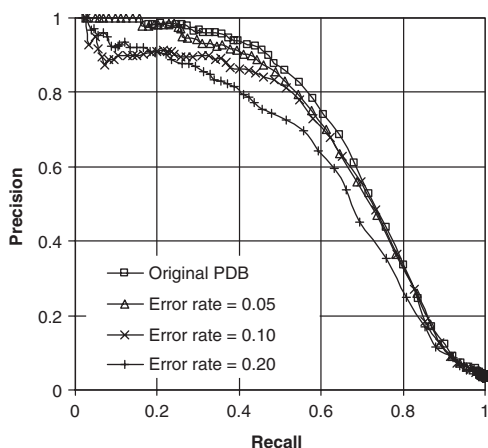
Fig. 5. Recall–Precision curves of the final consensus prediction for CH on residue level with different rates of artificially introduced errors upon zinc-binding. Error rate = 0.10 means 10% (i.e. 102) of randomly selected zinc-binding residues are annotated as non-zinc-binding. When a zinc-binding residue is selected to be annotated as non-zinc-binding, all residues binding to the same Zn atom as this one are annotated as non-zinc-binding.

Zn4 binding pattern were considered correctly predicted when two or three residues respectively, were correctly predicted). Our future goal is to predict the whole zinc-binding pattern.

4 CONCLUSION

We presented a method to predict zinc-binding sites from amino acid sequences by combining SVM predictions and homology-based predictions. The method predicted Cys, His, Asp and Glu with 75% precision (86% for Cys and His only) at 50% recall level, when tested on a non-redundant set of PDB containing 2727 unique protein chains. The success rate was even higher if homologues were predicted: for Cys, His, Asp and Glu with 76% precision (90% for Cys and His only) at the 70% recall level. The predictions were so reliable that some occasional putative errors of PDB regarding zinc-binding were found. We would expect the use of our method for predicting zinc-binding residues as a useful tool to check, for example, whether amino acids of a little-characterized protein are actually involved in binding zinc.

Conflict of Interest: none declared.

REFERENCES

Al-Karadaghi,S. *et al*. (1994) Refined crystal structure of liver alcohol dehydrogenase-NADH complex at 1.8 Å resolution. *Acta Crystallogr. D Biol. Crystallogr*., **50**, 793–807.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*., **25**, 3389–3402.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res*., **32**, D226–D229.

Andreini,C. *et al*. (2004) A hint to search for metalloproteins in gene banks. *Bioinformatics*, **20**, 1373–1380.

Auld,D.S. (2001) Zinc coordination sphere in biochemical zinc sites. *Biometals*, **14**, 271–313.

Baleja,J.D. *et al*. (1997) Refined solution structure of the DNA-binding domain of GAL4 and use of 3J(113Cd,1H) in structure determination. *J. Biomol. NMR*, **10**, 397–401.

Berg,J.M. and Shi,Y. (1996) The galvanization of biology: a growing appreciation for the roles of zinc. *Science*, **271**, 1081–1085.

Bernstein,F.C. *et al*. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542. Available at http://www.rcsb.org/pdb.

Brewer,G.J. *et al*. (1983) Biological roles of ionic zinc. *Prog. Clin. Biol. Res*, **129**, 35–51.

Coleman,J.E. (1992) Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins. *Annu. Rev. Biochem*., **61**, 897–946.

Davis,J. and Goadrich,M. (2006) The relationship between Precision–Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. ACM Press, Pittsburgh, Pennsylvania.

Ekman,D. *et al*. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol*., **348**, 231–243.

Evanics,F. *et al*. (2003) Nuclear magnetic resonance structures of the zinc finger domain of human DNA polymerase-alpha. *Biochim. Biophys. Acta*., **1651**, 163–171.

Gregory,D.S. *et al*. (1993) The prediction and characterization of metal binding sites in proteins. *Protein Eng*., **6**, 29–35.

Harding,M.M. (2004) The architecture of metal coordination groups in proteins. *Acta. Crystallogr. D Biol. Crystallogr*., **60**, 849–859.

Herve du Penhoat,C. *et al*. (2004) The NMR solution structure of the 30S ribosomal protein S27e encoded in gene RS27_ARCFU of Archaeoglobus fulgidis reveals a novel protein fold. *Protein Sci*., **13**, 1407–1416.

Holm,R.H. *et al*. (1996) Structural and functional aspects of metal sites in biology. *Chem. Rev*., **96**, 2239–2314.

Lewis,R.J. *et al*. (1998) An evolutionary link between sporulation and prophage induction in the structure of a repressor:anti-repressor complex. *J. Mol. Biol*., **283**, 907–912.

McCall,K.A. *et al*. (2000) Function and mechanism of zinc metalloenzymes. *J. Nutr*., **130**, 1437S–1446S.

McPherson,A. (1999) *Crystallization of Biological Macromolecules*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Menchetti,S. *et al*. (2006) Improving prediction of zinc binding sites by modeling the linkage between residues close in sequence. In Apostolico,A., Guerra,C., Istrail,S., Pevzner,P. and Waterman,M. (eds.) *Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology*. Springer, Berlin, Heidelberg, Venice, Italy, pp. 309–320.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucl. Acids Res*., **31**, 3789–3791.

Mittelman,D. *et al*. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.

Nagadoi,A. *et al*. (1999) Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. *J. Mol. Biol*., **287**, 593–607.

Nakata,K. (1995) Prediction of zinc finger DNA binding protein. *Comput. Appl. Biosci*., **11**, 125–131.

Passerini,A. *et al*. (2007) Predicting zinc binding at the proteome level. *BMC Bioinformatics*, **8**, 39.

Passerini,A. *et al*. (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins*, **65**, 305–316.

Pavlidis,P. *et al*. (2004) Support vector machine classification on the web. *Bioinformatics*, **20**, 586–587.

Platt,J.C. (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola,A., Schölkopf,P.B. and Schuurmans,D. (eds.) *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, pp. 61–74.

Sodhi,J.S. *et al*. (2004) Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol*., **342**, 307–320.

Stefanidou,M. *et al*. (2006) Zinc: a multipurpose trace element. *Arch. Toxicol*., **80**, 1–9.

Zhang,J. *et al*. (2004) Learning rules from highly unbalanced data sets. Data Mining, 2004. In *ICDM '04. Fourth IEEE International Conferenc*. AOL Inc., Dulles, VA, USA, pp. 571–574.

Zhang,Z. *et al*. (1998) Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res*., **26**, 3986–3990.