Abstract—Protein functional site prediction plays a key role in understanding protein function and in protein engineering. In this work we developed a novel method using canonical correlation analysis and the conservation window technique to predict protein ligand binding sites. The method was tested with a well-known benchmark dataset and consistently outperformed the existing method Xdet, which is based on Pearson correlation, by improving the lowest and highest ranking positives for more than 18% and 22% respectively.

Keywords: Canonical Correlation Analysis, Functional Residues, Multiple Sequence Alignment, Specificity Determining Positions.

1. Introduction

Functional sites refers to regions in a three-dimensional protein structure that perform certain function, such as ligand binding. These sites are primarily composed of a cluster of amino acids, though sometimes non amino acid components (such as metal ions) may also be required for the function. Typical sites can be catalytic sites that bind substrates in enzymes, regulatory sites that bind regulatory factors, sites that bind proteins or non-protein entities such as metal ions, nucleic acids, or other ligands. One main characteristic of these functional sites is the evolutionary conservation of the sequence composition, which therefore is most commonly used for identifying them.

Identification and prediction of functional residues is an important topic in computational biology and has attracted strong interest during the past decade. Accurate prediction and characterization of functional sites can lead to distinct sequence patterns, which can be stored in public databases and searched with pattern matching techniques for identifying distant homologues and predicting function. Furthermore, identification of residues that account for protein function specificity is crucial, not only for understanding the nature of functional specificity, but also for protein engineering experiments aimed at tuning the specificity of an enzyme, regulator or transporter.

As mentioned above, functional sites are usually evolutionarily conserved in their amino acid composition. This is because some specific patterns of amino acids in the protein sequence are essential to shape up and maintain certain often well-defined three-dimensional (3D) structures that are needed for the function; drastic variation of amino acids at such sites would have caused loss of structure and function and therefore more likely succumb to selection pressure. Therefore, most methods for functional site identification utilize measures of amino acid sequence conservation in homologous sequences [1] [2], based on the assumption that functional sites are relatively conserved during evolution. However, to predict functional residues using only conservation information might cause false predictions. This is because sequence conservation reflects not only evolutionary selection at functional sites to maintain protein function, but also selection throughout the protein to maintain the stability of the folded state. Besides, protein structures can be conserved without necessarily satisfying sequence conservation. In other words, proteins evolve under selection pressure to both maintain the stability of the overall structure and biochemical function, but this pressure can be endured even allowing sequence variation, as long as the structure is maintained. Consequently, some methods have been developed for functional residue prediction based on protein structural information. These methods look for structural features frequently associated with active sites and binding sites, like low-stability regions [3], special connectivity patterns extracted from residue-residue contact networks [4], patterns in the protein surface geometry [5], energy considerations [6], and chemical properties [7]. Note that structural information, while more reliable, is typically harder to attain and thus less commonly available than sequential information for most proteins.

Pazos et al. [8] recently proposed a method for detecting functional residues that relies on neither conservation nor structural information, but rather on the functional classification of a set of proteins that are believed to share one or more functional sites. A good example of such family of proteins is the set of enzymes that catalyze the same reaction on different organisms. Pazos’ algorithm, called Xdet, takes as input a multiple sequence alignment (MSA) of the sequences in the family and predicts as functional sites those column (or positions) whose pairwise amino acid similarities maximally correlate with the proteins’ functional classification. This correlation is measured by the Pearson correlation coefficient. Because the Pearson correlation is calculated based on fixed pairings of variables, some subtle and potentially revealing correlations among residue compositions and functional specificity can be missed out.
2. Methods

2.1 Phylogeny-independent detection of functional residues

Pazos’ Xdet method [8] is intended to locate positions in a MSA which are related to the functional classification of the proteins, ideally when the functional classes can be categorized by a hierarchy, or distances between them can be defined. The idea is that, in these positions, amino acid substitutions with a more negative value in PSM or BLOSUM matrix between two proteins would be related with a high functional difference between these proteins, and vice versa. Figure 1 shows a schema of Xdet. For each position in the alignment, a matrix quantifying the amino acid changes for all pairs of proteins is constructed based on a substitution scoring scheme, such as BLOSUM. In this matrix, a given entry represents the similarity between the residues of two proteins at that position. Another matrix is constructed from an external explicit functional classification, and each entry of this matrix represents the “functional similarity” between the corresponding proteins (for the functional feature we are interested in). These two matrices are compared element-wise with a Pearson correlation coefficient $r_k$:

$$r_k = \frac{\sum_{i,j} (A(k)_{ij} - A(k)) \cdot (F_{ij} - F)}{\sqrt{\sum_{i,j} (A(k)_{ij} - A(k))^2} \cdot \sqrt{\sum_{i,j} (F_{ij} - F)^2}}$$  \hspace{1cm} (1)$$

where $A(k)_{ij}$ is the similarity between the amino acids of proteins $i$ and $j$ at position $k$ in a multiple alignment of $N$ proteins of length $L$, and $F_{ij}$ the functional similarity between proteins $i$ and $j$, and bars mean averages; $r_k$ gives the score for position $k$.

Positions with high $r_k$ values are the ones for which similarities between amino acids are correlated with the functional similarities between the corresponding proteins, and hence are predicted as the ones related with functional specificity.

2.2 Canonical Correlation Analysis (CCA)

Xdet measures the correlation between the amino acid similarity matrix and the function similarity matrix by treating them as vectors and calculating their Pearson correlation coefficient. This is a common measure of the correlation (linear dependence) between two vectors. Geometrically, the Pearson correlation coefficient can also be viewed as the cosine of the angle between the two vectors of samples drawn from the two random variables. A different measure of correlation, the canonical correlation, does not treat direction of the two vectors as fixed, but allows rotations in their principal components to find stronger correlations.

Canonical correlation analysis is a theory developed by Hotelling [9] and widely used in pattern analysis [10]. Consider two sets of variates $X$ and $Y$ (vectors of random variables) with a joint distribution, and the Pearson correlation coefficient measures the correlation as manifested in $X$ and $Y$.

$$r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$  \hspace{1cm} (2)$$

where $\text{cov}(X,Y)$ is the covariance between variates $X$ and $Y$, and $\text{var}(X)$ and $\text{var}(Y)$ are their variances. The goal of CCA is to analyze the maximum correlation if we allow the vectors $(X)$ and $(Y)$ to be rotated in the space of each set of variates

$$X \rightarrow \phi(X), Y \rightarrow \psi(Y)$$  \hspace{1cm} (3)$$

in such a way that the new coordinates maximize the correlation between the two sets of variates. This leads to the following optimization:

$$r^* = \max_{\phi,\psi} \frac{\text{cov}(\phi(X), \psi(Y))}{\sqrt{\text{var}(\phi(X)) \cdot \text{var}(\psi(Y))}}$$  \hspace{1cm} (4)$$

The maximization of the correlation $r$ in Eq. 4 boils down to choosing the orientations to rotate the vectors $X$ and $Y$ and can be solved as a generalized eigenvalue problem [10].
For any column \( k \) in the multiple sequence alignment, take the alignment similarity \( A(k)_{ij} \) as \( X \) and the functional similarity \( F_{ij} \) as \( Y \), it is straightforward to apply CCA to measure the maximized correlation \( r^*_k \) between amino acid similarity \( A(k) \) at position \( k \) and functional similarity \( F \). By allowing the correlation measure to reach beyond the fixed direction of the sampled vectors, more subtle and complex sources of mutual information may be discovered, leading to the method that is capable of finding signals in highly noisy environments – in this case signals that might have been buried in the phylogeny relationship among the sequences which is intrinsic in the MSA.

### 2.3 Conservation Window

When calculating the correlation between the amino acid similarity matrix of each column in a MSA and the functional similarity matrix, regardless of the correlation measure used, it is not uncommon to find columns with identical score but actually quite different in relation to the functional specificity. In other words, correlation itself is not sufficient to differentiate some sites. Capra and Singh [11] proposed a heuristic to unite these degenerate scores. They notice that if two columns have the same correlation score (according to any measure), one might think that the one in the area of greater conservation is likely to be of greater importance for specificity or function. The heuristic, named ConsWin (Conservation Window), incorporates the conservation of sequentially adjacent positions into the correlation score in the following way:

\[
\text{ConsWin}(c) = \lambda \cdot \text{corr}(c) + (1 - \lambda) \cdot \frac{\sum_{i \in \text{win}} \text{Cons}(c_i)}{|\text{win}|}
\]

where \( c \) is the column being considered, \( \text{win} \) is a set containing the indices of all columns in a window around, but not including, the column \( c \), and \( \text{corr}(c) \) is the correlation score for column \( c \) calculated using Eq. 1 for \( \text{Xdet} \) or using Eq. 4 for CCA. The second term is the average conservation of the window; Capra and Singh propose to use the Jensen-Shannon [12] divergence to estimate conservation \( \text{Cons}(c_i) \). This is a measure intended to quantify the similarity between probability distributions, and is defined, for a column \( c \), as:

\[
D^I_c = \alpha \cdot \text{RE}(p_c, s) + (1 - \alpha) \cdot \text{RE}(q, s)
\]

where \( s = \alpha \cdot p_c + (1 - \alpha) \cdot q \), \( \text{RE} \) is relative entropy, \( p_c \) is the column amino acid distribution, \( q \) is a background distribution (we use BLOSUM62), and \( \alpha \) is a prior weight that was set up to 0.5. The size of the window \( |\text{win}| \) will have to be experimentally optimized for best performance, so will the weight \( \lambda \). Unlike Capra and Singh, we multiply the correlation score at each position \( c \) (according to \( \text{Xdet} \) or CCA) by its corresponding conservation measure given by ConsWin\((c)\) to obtain the final score that is assigned to each column. This allows regularization taking into account background information in a broader range.

### 3. Results

We tested our new method on the dataset created and provided by Capra and Singh [11], available at http://compbio.cs.princeton.edu/specificity/. This is a dataset composed of 106 MSAs. Each MSA contains domain sequences of proteins (enzymes) from the same Pfam domain family with significant sequence identity, and the enzyme commission (EC) numbers of these sequences are the same except for the last part, which specify the substrate these enzymes catalyze. Therefore, sequences in a MSA belong to one of two different specificity groups by their specific substrate according to their fourth-level EC number. The true positive positions in each MSA are residues near ligands, which are found in two ways: for each chain (protein sequence), if a relevant ligand is present, they label as positives all chain residues with an atom within 5Å of a relevant ligand atom. Since many enzymes do not have 3D structures in complex with their substrate, catalytic sites are used as a proxy for the location of ligands and include all residues within 5Å of a catalytic site.

Therefore, for each MSA, we know what columns are the true positives, i.e. the functional residues, also called specificity determining positions (SDPs). Without using the label of each column, we can calculate its correlation coefficient according to \( \text{Xdet} \) and according to the CCA, and each score is then weighted by the ConsWin\((c)\) score. If the predictor were perfect, the weighted scores would rank the true positives at the top of the list (higher scores), and non specificity determining positions at the bottom (lower scores). A less perfect predictor would mix up the positives with negatives in the list ranked by the weighted scores. The performance for the prediction can be measured by ROC score, which is the normalized area under the curve (AUC) that plot the number of true positives as a function of false positives when moving the threshold down the ranking list.

Figure 2 (left column) compares the performance of \( \text{Xdet} \) (crosses) and CCA (bullets) by the ROC score for each MSA and averaging it over the 106 MSAs in the dataset. Results for three different \( \lambda \) are shown, namely \( \lambda = 0.1 \) (top row), \( \lambda = 0.5 \) (middle row) and \( \lambda = 0.9 \) (bottom row). The parameter \( \lambda \) balances the contributions of the functional correlation score and the conservation score to the final score of each column. High values of \( \lambda \) give more

<table>
<thead>
<tr>
<th>Method</th>
<th>Average AUC</th>
<th>Average position of lowest SDP</th>
<th>Average position of highest SDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Xdet} )</td>
<td>94.4%</td>
<td>58.04</td>
<td>5.32</td>
</tr>
<tr>
<td>( \text{CCA} )</td>
<td>95.11%</td>
<td>90.99</td>
<td>4.28</td>
</tr>
</tbody>
</table>

Table 1: Performance comparison between \( \text{Xdet} \) and CCA, when \( \lambda = 0.9 \) and \(|\text{win}| = 20\).
Fig. 2: Prediction results according to the area under the ROC curve (left column), and according to the position of positive columns (functional residues, or SDPs) in the ranking of scores assigned to all positions in the MSA (right column): lowest ranked SDPs are at the top, and highest ranked SDPs are at the bottom. Crosses correspond to scores calculated using Xdet, and bullets to scores calculated using CCA. All scores are weighted by ConsWin(c). Results for $\lambda = 0.1$ (top row), $\lambda = 0.5$ (middle row), $\lambda = 0.9$ (bottom row) and for $|\text{win}| = \{1, 2, \ldots, 20\}$ are presented. See the text for a more detailed description of these curves.

weight to the functional correlation, and low values give more importance to the conservation around the column (see Eq. 5). For each $\lambda$ in Fig. 2, we have varied the number of columns in the window, $|\text{win}|$, from 1 to 20, and calculated the AUC score for each $|\text{win}|$. For low values of $\lambda$, Xdet presents better performance than CCA, and greater window
lengths improve the prediction results. However, the best performance is obtained with $\lambda = 0.9$, in which case CCA outperforms Xdet (AUC with Xdet is 0.94, with CCA it is slightly improved to 0.95), and for this $\lambda$ the dependence on window length is not significant. These results are also shown in Table 1.

Another way to evaluate prediction performance, also used by Capra and Singh [11], is by finding the position of each positive column (functional residue, or SDP) in the ranking of the scores of all the columns in the MSA. Ideally, in a MSA with $C$ columns and $T$ true positives, the perfect predictor would rank the specificity determining columns from 1st through $T$-th, and the negative examples would be ranked $(T + 1)$-th through $C$-th. Figure 2 (right column) shows two curves for each method, Xdet (crosses) and CCA (bullets), and for each $\lambda$. The solid curve shows the position of the lowest ranked positive example, and the dotted curve shows the position of the highest ranked positive (these are averages over the 106 MSAs). The results are consistent with the AUC scores: best performance is obtained using $\lambda = 0.9$, meaning that functional correlation discriminates better than conservation (notice how when $\lambda = 0.9$ the ranking of positives is, again, independent of the conservation window length). It is clear that the CCA method outperforms the Pearson correlation coefficient in revealing the correlation between function and positions in the MSA. Using $\lambda = 0.9$, the performance improvement goes, for the lowest ranked positive, from 38.02 with Xdet to 30.80 ($\sim 18\%$ improvement) with CCA, and for the highest ranked positive, from 5.51 with Xdet to 4.27 ($\sim 22\%$ improvement) with CCA.

4. Conclusions

We have shown in this work that by applying canonical correlation analysis on vectors characterizing candidate functional sites in proteins for ligand binding and vectors characterizing known binding specificity, our new method is able to detect binding sites more accurately than the previous method that uses Pearson correlation analysis, in which the vectors are fixed in pre-aligned frames. The new method is also shown to be more amicable for regularization to take into account prior knowledge and/or background of protein sequence conservation. Also, the canonical correlation analysis provides a powerful platform that can be used to detect signals present at multiple sites in a correlated way, which cannot be detected by current methods that assume site independence. Study of multiple sites detection will be the focus of our future work.

References