Journal of the Korean Physical Society, Vol. 70, No. 9, May 2017, pp. 876∼879

Protein Contact Prediction by Using Information Theory

Jae-Young Byeon and Julian Lee*
Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

(Received 20 March 2017, in final form 7 April 2017)

We develop a novel method for predicting the inter-residue contacts of a protein from evolutionary information obtained from the alignment of multiple sequences. Our method is based on information theory, where we use conditional mutual information so that the spurious correlations coming from indirect effects are removed. The benchmark test shows better performance than the previous method using mutual information does, suggesting the potential of the new method.

PACS numbers: 87.14.Ee, 02.50.Cw, 02.50.-r, 02.50.Tt
Keywords: Protein structure prediction, Contact prediction
DOI: 10.3938/jkps.70.876

I. INTRODUCTION

Predicting the three-dimensional structure of a protein solely from the sequence information is an important unsolved problem in computational biophysics [1–3]. Information on the local structure, such as the inter-residue contacts, provides constraints on the three-dimensional structure, thus facilitating the prediction of the full three-dimensional structure [3–7]. Consequently, many efforts have been made to predict the inter-residue contacts, which is a more tractable problem than that of predicting the full three-dimensional structure [8–16]. Many methods predict contacts from the co-mutation pattern obtained from an alignment of the family of sequences homologous to the query sequence. The idea is that the pair of amino acid residues that are in spatial contact with each other tends to have a correlation in the mutation pattern.

The most rigorous measure for the mutual dependence between two random variables X and Y comes from information theory, which is the mutual information (MI) $I(X,Y)$ defined as

$$I(X,Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}, \qquad (1)$$

where $P_{XY}(x,y)$ is the joint probability of X and Y taking the values of x and y simultaneously, and $P_X(x)(P_Y(y))$ is the marginal probability that $X(Y)$ takes the value of $x(y)$. Several contact prediction methods based on MI have been developed [10,11].

The performance of early contact-prediction methods, including those based in MI, have been hindered by the existence of a co-mutation between two residues that are not in direct contact, but are connected by a chain of contacts (Fig. 1). Obviously, such residues influence each other; therefore, a simple measure of the mutual dependence such as MI would lead to a proliferation of false positives. Several ideas for eliminating such indirect effects have been implemented in recent contact-prediction methods. One such method is PSICOV, which uses a partial correlation coefficient [13]. The partial correlation coefficient is derived from the Pearson correlation coefficient by filtering out indirect effects. However, such correlation coefficients can be used only for linear correlation, and little theoretical justification exists for using it to describe the correlation of amino acids. Another class of methods uses entropy maximization to infer the probability distribution of residue contacts [14–16]. Although theoretically appealing, these methods have relatively large computational costs, and most rely heavily on various approximations.

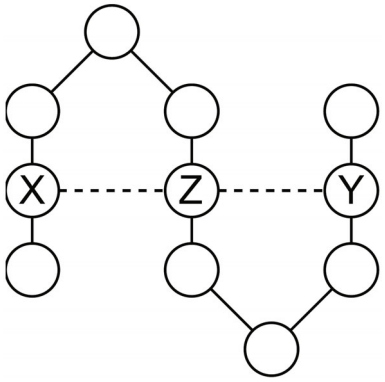

Fig. 1. Example of an indirect effect. The residues X and Y influence each other because of their common contacts to the residue Z, but they are not in direct contact.

*E-mail: jul@ssu.ac.kr; Fax: +82-2-824-4383

In fact, a variation of mutual information, called conditional mutual information, where the indirect effects are removed, exists [17,18]. The conditional mutual information (CMI) $I(X,Y|Z)$ is a measure of the correlation between a pair of variables X and Y after their mutual influence mediated by another variable Z is removed. It is given by

$$I(X,Y|Z) = \sum_{x,y,z} P_{XYZ}(x,y,z) \log \frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}, \tag{2}$$

where $P_{X|Z}(x|z) \equiv P_{XZ}(x,z)/P_Z(z)$ is a conditional probability. Similar definition applies for the other conditional probabilities. In the context of protein-contact prediction, X and Y are the amino acids appearing at a pair of positions. The variable Z represents the sequence pattern of positions other than X and Y. In contrast to mutual information, conditional mutual information has not been used for protein-contact prediction so far. We show that the performance of the prediction obtained using conditional mutual information is better than that obtained using mutual information, showing the potential for the newly developed method.

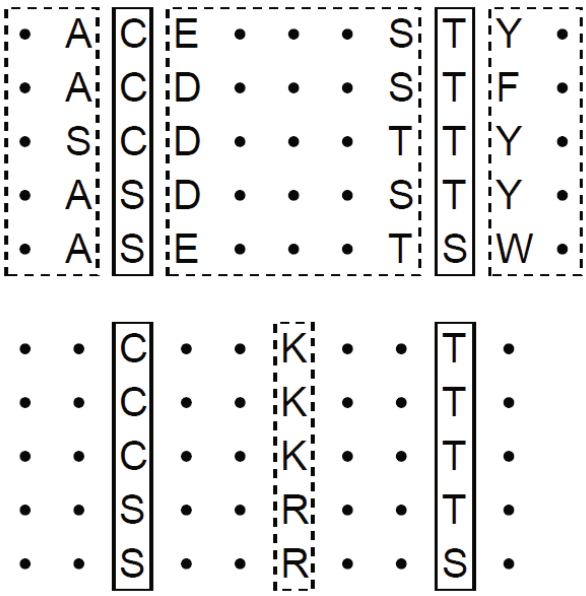

Fig. 2. The variable Z in mutual conditional information should be the pattern of the alignment after excluding the pair of positions X and Y, as shown on the top. There are 21^{L-2} possibilities, which are highly underrepresented in the actual data. Therefore, we choose the position with the highest influence on the interaction between X and Y, as the variable Z, as shown on the bottom.

II. THE METHOD

A pair of protein residues is defined to be in contact if the distance between their C_β atoms is less than 8 Å. For a contact prediction method based on a co-mutation pattern, the input to the algorithm is not the single query sequence, but an alignment of homologous protein sequences. The length of the alignment is set to the query sequence length. Gaps in the alignment can be present, so the gap is treated as the 21st amino acid. The probability in Eq. (1) or Eq. (2) can be estimated by using the frequencies of occurrences of amino acids at the given position in the sequence alignment. For example, $P_{X_i}(x)$ can be obtained from the frequency $\hat{P}_{X_i}(x) \equiv N_i(x)/\sum_{\tilde{x}} N_i(\tilde{x})$, where $N_i(x)$ is the number of sequences for which the entry on position i is amino acid x.

One problem in this estimate is due to the finite number of sequences in the alignment; some amino acids may not appear at all although their actual probabilities may not be zero. This leads to an indeterminate number in the formulas for mutual information and conditional mutual information because zeros appear both in the numerator and the denominator. In order to alleviate this problem, we add one to each $N(x)$. This is called the pseudocount [13,15]. Another problem is that some of the sequences may be overrepresented in such an alignment, leading to a bias. Therefore, removing redundant sequences from such an alignment is crucial. The criterion for such a redundancy is the similarity cutoff r. Any two sequences where the fraction of identical amino acid residues is greater than r are regarded as similar. For a given sequence, if n other similar sequences exist, then the number of occurrences of this sequence is counted as $w = 1/(n+1)$ instead of one. This procedure prevents multiple counting of homologous sequences.

The criterion for the similarity of the sequences, r, was determined following a previous method of contact prediction, PSICOV [13]. First, the mean fraction of identical residues $\bar{r}$ between each pair of sequences in the alignment was computed. Then, the cutoff was set as $r = \max[0.4, 1 - 0.12/\bar{r}]$. This ensures that if the sequences in the alignment are quite homologous to each other overall, a more lenient criterion for sequence similarity is used so that excessive removal of sequences is prevented. However, any sequences with more than 40% sequence identity are considered as similar. The probability of an amino acid is then estimated from the modified frequency after including the pseudocount and the sequence reweighting:

$$\hat{P}_X(x) \equiv (w_X(x)+1)/\sum_{\tilde{x}}(w_X(\tilde{x})+1). \tag{3}$$

If the indirect effect due to residues other than the pair of interest is to be filtered out, Z in Eq. (2) should be set as the pattern of sequences excluding the pair under consideration. If the length of the query sequence is L, then the possible number of such patterns is 21^{L-2}, which is much larger than the number of sequences contained

in any multiple-sequence alignment. Therefore, $P(Z)$ cannot be estimated with a reasonable accuracy from a multiple alignment if Z includes the effect of all $L-2$ residues. Therefore, instead of filtering out the indirect effect completely, we decided to eliminate the effect from the single residue Z that contributed most significantly to the coupling between the pair X and Y (Fig. 2).

MI estimated by aligning a finite number of sequences, has been argued to have a background noise. Such a non-zero background random MI, S_{ij}, has been modelled as [11]

$$S_{ij} \equiv \frac{S_{i-}S_{-j}}{S_{--}}, \tag{4}$$

where

$$S_{i-} \equiv \sum_j S_{ij}, \quad S_{-j} \equiv \sum_i S_{ij}, \quad S_{--} \equiv \sum_{i,j} S_{ij} \tag{5}$$

and $S_{ij} \equiv I(X_i, X_j)$ is the score function for the pair of positions i and j, which is the mutual information of the amino acids X_i and X_j at those positions.

After subtracting the background MI, the modified score function P_{ij} is given as

$$PC_{ij} = S_{ij} - \frac{S_{i-}S_{-j}}{S_{--}}. \tag{6}$$

The correction term in Eq. (6) has been called the average production correction (APC). The contact-prediction method that uses MI modified by the APC has been shown to exhibit better performance than those using MI without APC [11]. APC has also been introduced in PSICOV [13]. In a similar vein, we also test the CMI with and without APC. We compare four score functions: MI and CMI, each of them with and without APC.

The test set was selected from 146 sequence families in Pfam [19]. Biological monomers with single copies of Pfam domains, with a crystallographic-structure resolution $\leq$ 1.9 Å, and with at least 1000 sequences in the alignment, were selected.

III. RESULTS

The accuracy of the prediction method, defined as the ratio of the correct pair to that of predicted pairs in contact, can be obtained by comparing the prediction results with the actual contacts in the experimental structure. A prediction method ranks each pair of positions with score functions such as MI or CMI, so one has to choose a finite number $N_{\rm pair}$ of pairs with the highest scores as the predicted contacting pairs. The value of $N_{\rm pair}$ has to be decided based on whether the sensitivity or the specificity is more important, which in turn, depends on the applications. Here, we will avoid the issue of fixing $N_{\rm pair}$ and simply examine the accuracy of our method as a function of $N_{\rm pair}$.

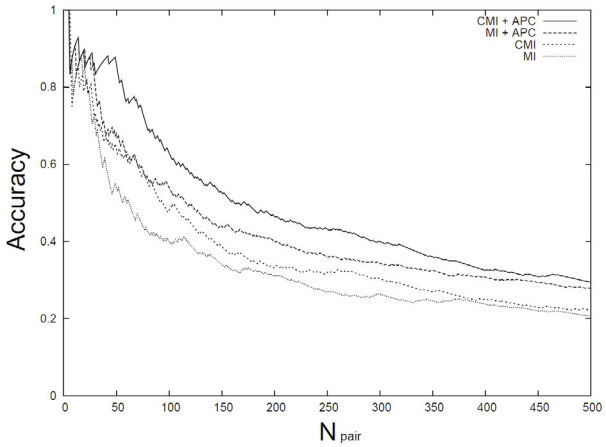

Fig. 3. Prediction accuracy of the four score functions for a protein 1JBE as functions of $N_{\rm pair}$. The results obtained by using MI or CMI are shown, with and without APC corrections.

As an example, we plot the prediction accuracies of the four score functions for a protein wild-type ApoCheY (PDB ID:1JBE, chain length 126) against $N_{\rm pair}$ (Fig. 3). As expected, although some oscillations are seen, the accuracy tends to deteriorate as $N_{\rm pair}$ increases due to the increasing number false positives. We clearly see that the graph for the CMI lies above that for the mutual information, clearly indicating that filtering out the indirect effect increases the accuracy of the prediction regardless of the number of pairs used for the prediction. The results with and without APC are also compared, and we see that APC clearly enhances the accuracy.

Similar results were obtained for a large-scale benchmark test on 146 sequences. The number of contacts is expected to increase as the sequence length L increases; therefore, choosing a size-dependent threshold $N_{\rm pair}$ for selecting the top pairs is more reasonable. Usually, $N_{\rm pair}$ is chosen to be proportional to the chain length. Therefore, we plotted the average accuracy of these benchmark tests as a function of $N_{\rm pair}/L$ (Fig. 4). The order of accuracy is the same as the result above, with the performance of CMI being better than that of MI and the performance using APC being better than that using MI or CMI alone.

Because predicting the contacts of residues that are separated far away along the sequence is more nontrivial, we also assessed the average prediction accuracy for residue pairs whose distances along the sequence were more than 8, 11, and 20. We again see that the predictions obtained using CMI with APC are the most accurate, as shown in Table 1 for several values of $N_{\rm pair}/L$.

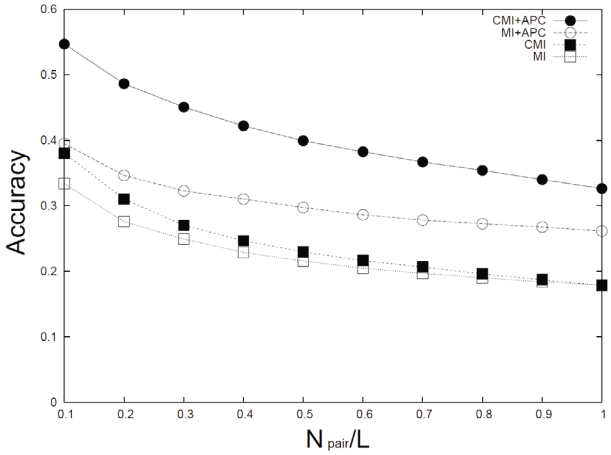

Fig. 4. Average prediction accuracy of the four score functions for 146 proteins from Pfam as functions of $N_{\rm pair}/L$. The results obtained by using MI or CMI are shown with and without APC corrections.

Table 1. Average prediction accuracy of the non-local contacts for a test set of 146 proteins for several values of $N_{\rm pair}/L$. The results obtained by using MI or CMI are shown with and without APC corrections.

	$\|i-j\|>8$			$\|i-j\|>11$			$\|i-j\|>20$		
$N_{\rm pair}/L$	1/4	1/2	1	1/4	1/2	1	1/4	1/2	1
MI	0.28	0.22	0.17	0.28	0.22	0.17	0.26	0.20	0.15
CMI	0.28	0.21	0.16	0.27	0.20	0.15	0.24	0.16	0.13
MI + APC	0.43	0.35	0.28	0.45	0.37	0.29	0.42	0.34	0.26
CMI + APC	0.53	0.43	0.33	0.54	0.43	0.33	0.51	0.41	0.30

IV. CONCLUSION

We developed a novel method for protein contact prediction based on information theory, where conditional mutual information, instead of mutual information, was used for ranking the residue pairs for the probability of contact. The results show that filtering out the indirect effect by using conditional mutual information, indeed, improves the accuracy of the method compared to that using mutual information. Although the results are quite promising, some room for improvement still exists. For example, in computing the conditional mutual information, we removed the effect of a single residue that contributes most significantly to the coupling between the pair of residues being examined, but there may be an optimal set of sequence patterns that is to be removed, which we will have to figure out. Also, the optimal procedure of sequence reweighting should be developed. If our novel method is to be a self-contained prediction method, the number of top pair to be selected, $N_{\rm pair}$, should be fixed. Finally, the performance of the new method will have to be compared with those of other prediction methods.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea, funded by the Ministry of Education, Science, and Technology (NRF-2014R1A1A2058188).

REFERENCES

[1] Y. Zhang and J. Skolnick, Proc. Natl. Acad. Sci. USA **101**, 7594 (2004).
[2] A. Fiser and A. Sali, Methods Enzymol. **374**, 461 (2003).
[3] S. Ovchinnikov, D. E. Kim, R. Y-R. Wang, Y. Liu, F. DiMaio and D. Baker, Proteins **84 (S1)**, 67 (2016).
[4] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina and C. Sander, PLos One **6**, e28766 (2011).
[5] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **109**, 10340 (2012).
[6] T. Nugent and D. T. Jones, Proc. Natl. Acad. Sci. USA **109**, E1540 (2012).
[7] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander and D. S. Marks, Cell **149**, 1607 (2012).
[8] U. Göbel, C. Sander, R. Schneider and A. Valencia, Proteins **18**, 309 (1994).
[9] S. W. Lockless and R. Ranganathan, Science **286**, 295 (1999).
[10] D. K. Chiu and T. Kolodziejczak, Bioinformatics **7**, 347 (1991).
[11] S. D. Dunn, L. M. Wahl, and G. B. Gloor, Bioinformatics **24**, 333 (2008).
[12] A. A. Fodor and R. W. Aldrich, Proteins **56**, 211 (2004).
[13] D. T. Jones, D. W. A. Buchan, D. Cozzetto and M. Pontil, Bioinformatics **28**, 184 (2012).
[14] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch and T. Hwa, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).
[15] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).
[16] M. Ekeberg, Cecilia Lövkvist, Y. Lan, M. Weigt and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[17] R. L. Dobrushin, Dokl. Akad. Nauk. **126**, 474 (1959).
[18] A. D Wyner, Information and Control **38**, 51 (1978).
[19] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, Nucleic Acids Res. **40**, D290 (2012).