Prediction of catalytic residues from X-ray protein structure refinement parameters

Yen-Yi Liu[1], Chih-Chieh Chen[2,3*]

[1]Central Regional Laboratory, Center for Research, Diagnostics and Vaccine Development, Centers for Disease Control, Taichung 40855, Taiwan
Email: current788@gmail.com
[2]Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[3]Medical Science and Technology Center, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
Email: chieh@mail.nsysu.edu.tw

Abstract*—Catalytic residue investigation is important for biologists to study protein functions. In previous studies, many researchers have successfully applied features, which were from sequence- and structure-based, to predict the position of catalytic residues in proteins. A highly correlation between atomic fluctuations and the catalytic positions have ever been observed. In this study, we were trying to investigate if this information was hidden in the X-ray diffraction data. The results from our test indicated the possibility to catch-up the catalytic residues in the protein structure refinement process.*

Keywords—TLS refinement, Catalytic site prediction, B-factor.

I. INTRODUCTION

To realize the function of a protein is the major goal for structural biologists to solve the structure. However, many protein structures from structural genomic project [1] are usually functional unknown. In order to decrease spent time and money, researchers usually need to pursue theoretical methods for identifying the potential functional sites. For this demand, many in silicon methods have been designed to achieve this requirement. These methods including sequence based such as, L1pred [2] and PINGU [3], and structural based such as, Jorge Fajardo approach [4] and WCN approach [5].

Recently, Huang *et al.*, 2011 [5] provided a prediction method based on the dynamics nature of catalytic residues. They filtered out the candidate residues by ranking the crowdedness value, which they called weighted contact number (WCN), for each residue in a given protein. The occurrences for each amino acid type of catalytic residues have also been investigated. Through their approach, potential catalytic residues can be thus simply predicted from a single protein structure. The concept of Huang's method is from Lin *et al.* 2008 [6], in which article they have mentioned the WCN model can be simplified to a centroid model (CM) [7]. The CM has more than once been used as the alternative for the translation/libration/screw (TLS) model [8-13], which is frequently used in X-ray structural refinement process. Therefore, the application of the TLS model could be reasonably considered as a way to predict catalytic residues from a single structure.

The major advantage by using TLS model is its university in structural biology community. If the prediction can be finished coupled with protein structural refinement, it would be very useful for structural biologists. Here, we showed that the prediction of catalytic residues in enzymes could be achieved directly from X-ray structural refinement process by using TLS-derived B-factors.

We test our approach for a previous reported enzyme dataset [5], the results showed that the viability of our approach to locate the residues in proteins.

II. METHODOLOGY

2.1 The computed B-factor profiles based on TLS model

The atomic fluctuation derived from the TLS model [8] is defined by Sternberg *et al.* [10] as:

$$\left\langle \Delta \mathbf{r}_i \cdot \Delta \mathbf{r}_i \right\rangle = \frac{1}{3}\mathbf{tr}\left(\mathbf{T} + \mathbf{S}^{\mathrm{T}} \times \mathbf{n} - \mathbf{n} \times \mathbf{S} - \mathbf{n} \times \mathbf{L} \times \mathbf{n}\right), \quad (1)$$

where $\mathbf{Dr}_i$ refers to the atomic displacement of an atom; $\mathbf{T}$, $\mathbf{L}$, and $\mathbf{S}$ indicate the translation, libration and screw matrixes, respectively. Position relative to the origin of an atom is donated by $\mathbf{n}$. The translation matrix is built on the displacement correlations between translation vectors along three directions in the Cartesian coordinate system. The libration matrix contains the displacement correlations between rotation vectors about three Cartesian axes. Correlations between the translation and rotation vectors are used to build the screw matrix. Each of $\mathbf{T}$, $\mathbf{L}$, and $\mathbf{S}$ is 3×3 matrix, where $\mathbf{T}$ and $\mathbf{L}$ are symmetric matrixes, and $\mathbf{S}$ is usually with arbitrarily specified origin. In total, 10 TLS parameters are needed to be refined to obtain the required atomic fluctuation [10]. The computed B-factor based on Equation 1 will be referred as TLS-derived B-factor or B_{TLS} through this article.

2.2 The TLS parameter refinement

To optimize the TLS parameters, the program REFMAC [14] of CCP4 software suit was used. The coordinates and structural factor were used as input data for REFMAC. In the parameters refinement process, the values of which were iteratively altered to find a best fitness against X-ray diffraction data [15-17]. In this study, we set each protein chain as a TLS group and perform 10 cycles of the TLS refinement. The obtained TLS parameters were used for computing B_{TLS}.

2.3 Datasets of catalytic residues

The initial dataset was selected from Huang, etc. [5]. This dataset contained 760 enzyme X-ray structures, which were selected from the Catalytic Site Atlas (CAS) [18], with pairing sequence identity $\leq$ 30%. We filtered out the structures without structural factors deposited and also those with errors in the SF files. Besides, only structures with all catalytic residues on the same protein chain were selected. The finally used dataset consists 371 enzyme structures (Table 1). We will use 371-set to call this dataset through this article.

III. RESULTS AND DISCUSSION

3.1 Characteristics of profiles between zB and zB_{TLS}

From the 371-set, we observed that the catalytic residues frequently locate at the local minimum of the zB_{TLS} profile. Figure 1 shows profiles plotting for several selected examples, which are 1W1O_A, 1PFQ_A, 1THG_A and 2PGD_A. From profiles shown in Figure 1, the catalytic residues, which are presented as hollow circles, almost located on the minimum regions. Therefore, it seems that the zB_{TLS} profile could be used for easily predicting the catalytic residues. However, since the TLS-derived B-factor has been reported to have high correlation with the experimental B-factor [14, 19], the comparison of profile characteristics between zB_{TLS} and zB is necessary in order to investigate if the zB_{TLS} profile is more sensitive than the zB profile to detect catalytic residues. Figure 2 shows the profile comparisons between zB_{TLS} and zB from the same PDBs used in Figure 1. From Figure 2, larger amplitudes of zB_{TLS} profiles than zB profiles can be observed, and locations of catalytic residues in zB profiles have no obvious trends.

3.2 Comparison of frequency distributions of amino acid types between catalytic and non-catalytic residues

Figure 3A illustrates the comparison of amino acid occurrences between catalytic and non-catalytic residues in 371-set. In catalytic residues, ASP, HIS, GLU, and ARG obviously appear more frequently than which in non-catalytic residues. To represent the amino acid occurrences for catalytic residues more clearly, in Figure 3B, the occurrences for different amino acid types of catalytic residues in 371-set are shown in an increasing order. The most frequently appearing amino acid types in 371-set are ASP, HIS, GLU, ARG and LYS that are accounted for 64% of the total amino acid types. In Figure 3B, we can clearly found that for the amino acid types accounted more than 5% are charged or polar amino acids. However, in non-catalytic residues, this preference is not existed (shown in Figure 3C).

TABLE 1
THE LIST OF THE PDB IDS OF 371-DATASET

135L:A	1BG0:A	1D6O:A	1F48:A	1H7X:C	1KDG:A	1ODT:C	1QJE:A	1TMO:A	1YBV:A
1A0I:A	1BGL:C	1D8C:A	1F6D:B	1HQC:A	1KIM:B	1OE8:B	1QK2:B	1TOX:A	1YCF:A
1A0J:C	1BJO:A	1DB3:A	1F8R:B	1HRK:A	1KNP:A	1OFD:A	1QLH:A	1TYF:I	1Z9H:B
1A26:A	1BQC:A	1DBF:C	1FC4:B	1HTO:B	1KP2:A	1OFG:F	1QMH:B	1TZ3:A	1ZE1:A
1A4I:A	1BRM:B	1DD8:B	1FCQ:A	1HZF:A	1KRA:C	1OG1:A	1QQ5:A	1U7U:A	1ZM2:F
1A4L:A	1BRW:B	1DE6:A	1FDY:C	1I19:A	1KSJ:A	1OH9:A	1QZ9:A	1U8V:C	1ZRZ:A
1A65:A	1BT1:A	1DEK:A	1FOA:A	1I1E:A	1KWS:A	1OJ4:B	1R16:A	1UAG:A	206L:A
1A8Q:A	1BTL:A	1DFO:B	1FOB:A	1I1I:P	1KYQ:B	1OK4:H	1R1J:A	1UAM:A	2A0N:A
1A95:C	1BWZ:A	1DGS:B	1FQ0:A	1I29:A	1KYW:F	1OKG:A	1R30:A	1UAQ:B	2ACE:A
1AB4:A	1BZC:A	1DIO:L	1FR2:B	1I78:B	1KZH:A	1ONR:A	1R4F:B	1UAS:A	2ADM:A
1ABR:A	1BZY:B	1DMU:A	1FR8:A	1I9A:A	1L1D:A	1OS7:B	1R4Z:A	1UCH:A	2AYH:A
1AF7:A	1C2T:A	1DNP:A	1FRO:C	1IDJ:B	1LCB:A	1OTG:C	1R6W:A	1UF7:B	2BHG:A
1AFW:B	1C3J:A	1DUP:A	1FUA:A	1IM5:A	1LCI:A	1OYG:A	1R76:A	1UK7:A	2BIF:B
1AGY:A	1C82:A	1DZR:A	1FVA:B	1IR3:A	1LJL:A	1P4R:B	1RA2:A	1ULA:A	2BKR:A
1AKD:A	1C9U:B	1E0C:A	1G0D:A	1ITQ:B	1LML:A	1P5D:X	1RBL:A	1UN1:B	2BX4:A
1AKM:A	1CB8:A	1E19:B	1G64:B	1ITX:A	1LNH:A	1PFK:A	1RHC:A	1UQR:A	2CPO:A
1AKO:A	1CF2:Q	1E2T:F	1G6T:A	1IU4:A	1LVH:A	1PFQ:B	1RHS:A	1UQT:B	2DOR:B
1AL6:A	1CG2:C	1E5Q:E	1G72:A	1J00:A	1M21:B	1PGS:A	1RK2:C	1URO:A	2ENG:A
1AMP:A	1CGK:A	1E6E:A	1G79:A	1J09:A	1M6K:A	1PIX:B	1RO7:A	1V04:A	2F61:A
1AMY:A	1CHD:A	1EB6:A	1G8F:A	1J49:B	1MLV:B	1PJ5:A	1ROZ:A	1V0E:B	2HDH:A
1APX:A	1CHK:B	1EBF:A	1G8P:A	1J53:A	1MOQ:A	1PJA:A	1RPX:C	1V0Y:A	2JCW:A
1AQ2:A	1CK7:A	1EC9:C	1G99:A	1J79:B	1MPX:C	1PJH:A	1RQL:B	1V25:B	2LIP:A
1ARZ:B	1CNS:A	1ECL:A	1GA8:A	1J7G:A	1MRQ:A	1PMI:A	1RTU:A	1W0H:A	2NAC:A
1AUG:D	1COY:A	1ECX:B	1GAL:A	1JCH:A	1MVN:A	1PS9:A	1RU4:A	1W1O:A	2NLR:A
1AUK:A	1CTN:A	1EEJ:A	1GDH:B	1JH6:A	1N20:A	1PWV:B	1S3I:A	1W2N:A	2NPX:A
1AUO:A	1CTT:A	1EF0:A	1GE7:A	1JHF:A	1NDI:A	1PXV:B	1S95:B	1WD8:A	2PFL:A
1AVQ:C	1CV2:A	1EH5:A	1GIM:A	1JM6:A	1NID:A	1PZ3:B	1S9C:B	1WNW:C	2PGD:A
1AX4:B	1CVR:A	1EHY:A	1GNS:A	1JMS:A	1NIR:A	1Q18:B	1SLL:A	1X7D:A	2PLC:A
1B02:A	1CW0:A	1EI5:A	1GOG:A	1JOF:E	1NML:A	1Q3Q:C	1SLM:A	1X9H:A	2SQC:A
1B04:A	1CZ1:A	1ELQ:B	1GP5:A	1JQN:A	1NVM:G	1Q91:A	1SML:A	1X9Y:B	2THI:A
1B57:A	1CZF:B	1ESO:A	1GQ8:A	1JS4:B	1NVT:B	1QBA:A	1SNN:A	1XGM:B	2TOH:A
1B5Q:B	1D0S:A	1EUG:A	1GQG:B	1JXH:A	1NWW:A	1QCN:A	1SZJ:R	1XIK:A	2TS1:A
1B6B:B	1D1Q:B	1EUY:A	1GT7:A	1K0W:B	1O04:E	1QF6:A	1T0U:B	1XQW:A	2YPN:A
1B7Y:A	1D2R:E	1EY2:A	1GUF:B	1K30:A	1O98:A	1QGX:A	1T7D:A	1XRS:B	3R1R:A
1B8F:A	1D2T:A	1EYP:A	1GXS:C	1K32:A	1OAC:B	1QH9:A	1TDJ:A	1XVT:A	5COX:D
1B8G:B	1D3G:A	1EZ1:A	1GZ6:A	1KC7:A	1OAS:A	1QHG:A	1THG:A	1Y9M:A	7ODC:A
1BF2:A	1D4A:B	1F2V:A	1H19:A	1KCZ:A	1OBA:A	1QHO:A	1TML:A	1YBQ:A	8TLN:E
1BFD:A									

All protein chains in the 371-dataset are solved by X-ray and pair-wise sequence identity ≤ 30%. Each of the 371-datset has been refined by TLS refinement with one TLS group per chain using REFMAC software program.

3.3 Selection of the prediction thresholds

To perform prediction, a threshold value is needed to be set first. Residues with zB_{TLS} values below this threshold (or cutoff) are defined to be the candidate catalytic residues. The threshold value was determined by performing 10-fold cross-validation for the selected set consisting of 180 PDBs with the resolution < 2.0 Å from the 371-set. The best cutoff value is found to be -0.6 with the accuracy, the sensitivity and the specificity given 70%, 71% and 79%, respectively.

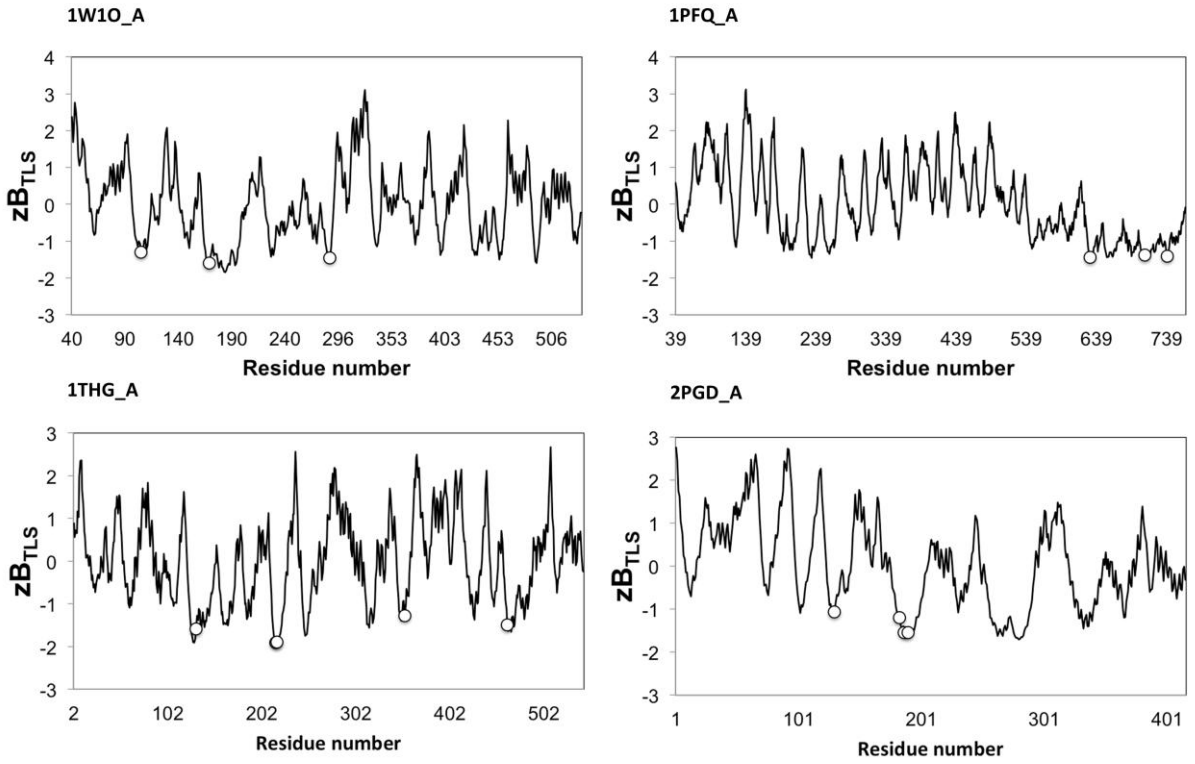

FIGURE 1. PREDICTION PROFILES BASED ON TLS METHODS FOR SEVERAL SELECTED PDBS, WHICH ARE 1W1O_A, 1PFQ_A, 1THG_A AND 2PGD_A SELECTED FROM THE 371-DATASET. FOR EACH CASE, THE CUT-OFF VALUE IS SHOWN IN DASHED LINE. HOLLOW CIRCLES INDICATES THE CATALYTIC RESIDUES.

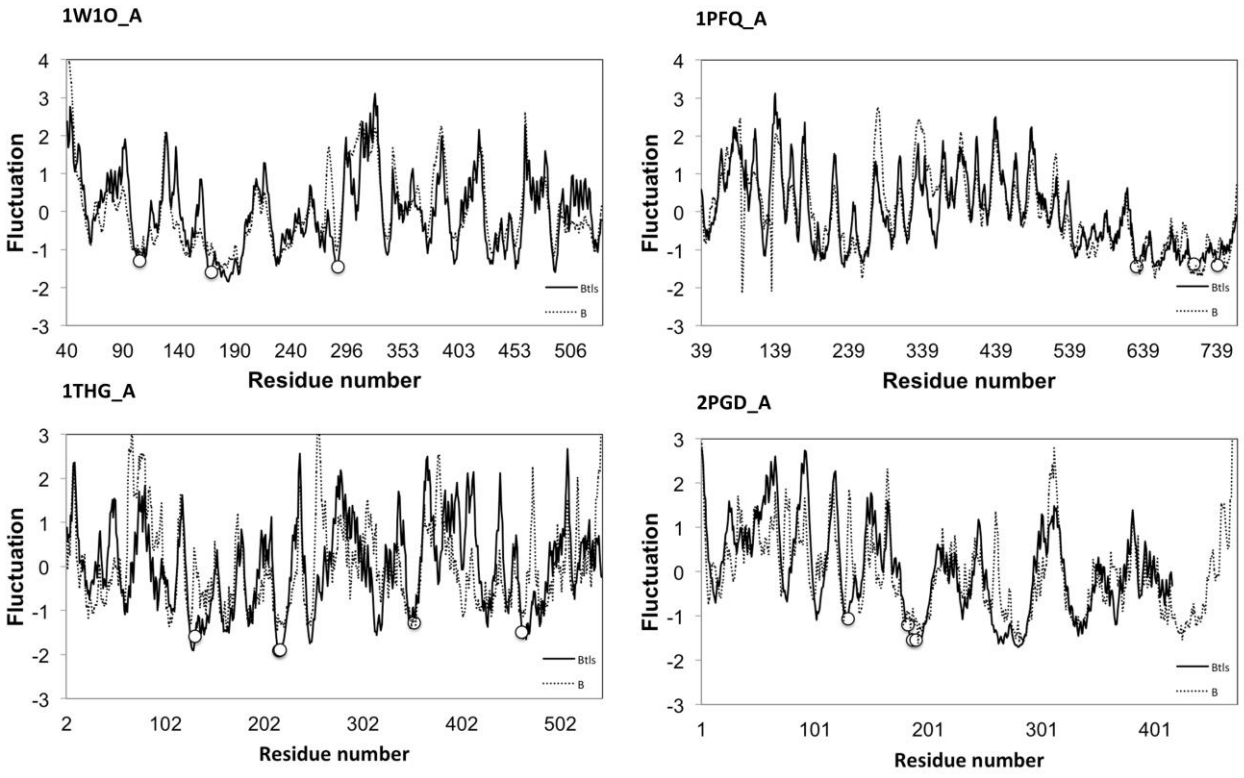

FIGURE 2. COMPARISON OF THE PROFILES AND THE CATALYTIC RESIDUE LOCATIONS BETWEEN TLS-DERIVED B-FACTOR (B_{TLS}) PROFILES AND EXPERIMENTAL B-FACTOR PROFILES FOR 1W1O_A, 1PFQ_A, 1THG_A AND 2PGD_A.

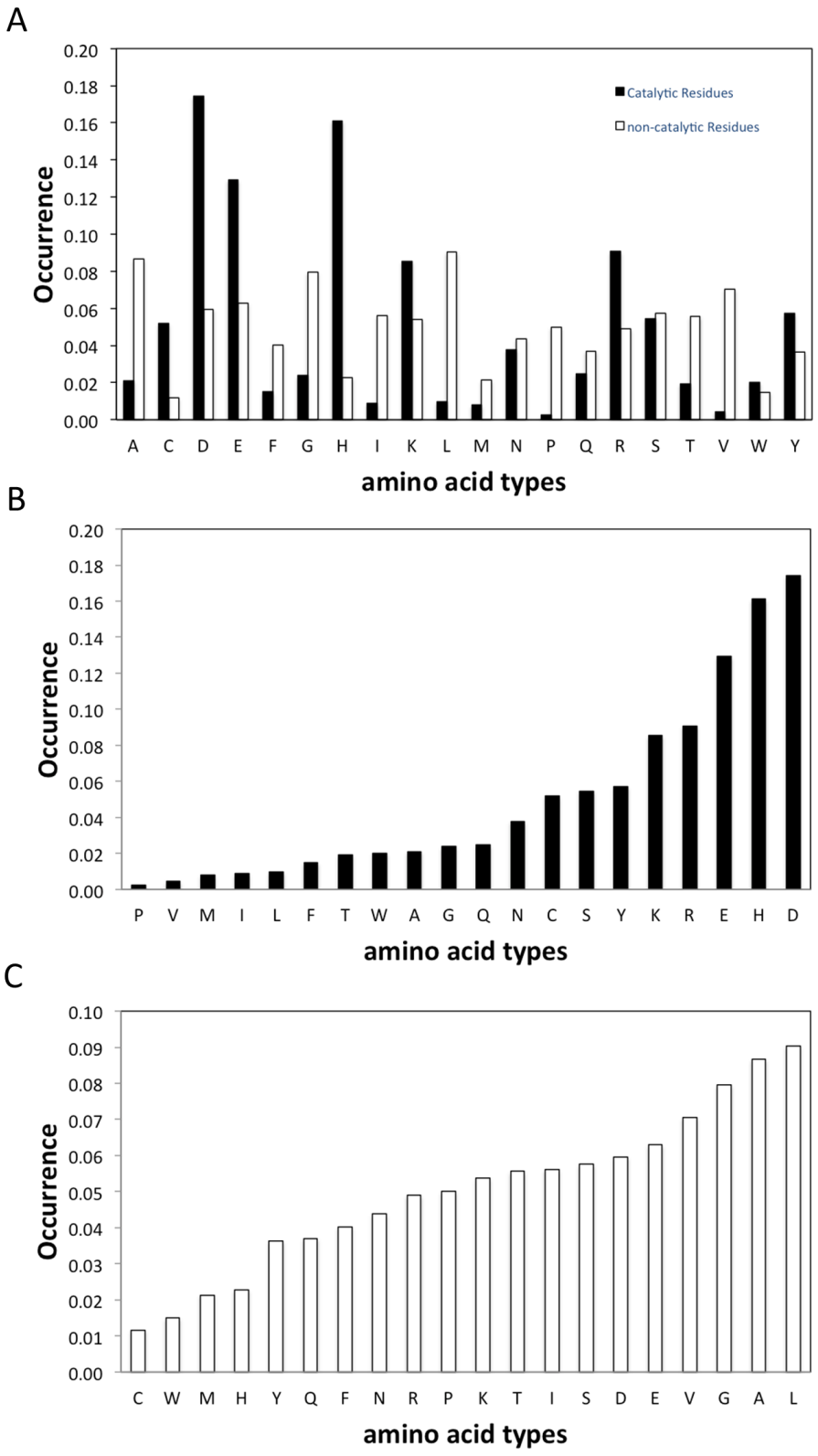

FIGURE 3. COMPARISON OF THE AMINO-ACID DISTRIBUTIONS BETWEEN CATALYTIC AND NON-CATALYTIC RESIDUES. (A) THE COMPARISON OF AMINO ACID OCCURRENCES BETWEEN CATALYTIC AND NON-CATALYTIC RESIDUES IN 371-SET. (B) THE OCCURRENCES FOR DIFFERENT AMINO ACID TYPES OF CATALYTIC RESIDUES. (C) THE AMINO-ACID PREFERENCE IN NON-CATALYTIC RESIDUES.

3.4 Resolution effect on the performance for catalytic residue prediction based on B_{TLS}

Because the computation of B_{TLS} includes the parameters fitting against X-ray diffraction data, the quality (i.e. the resolution) of the data would influence the calculated B_{TLS}. To test the influences caused by the data quality, we changed the resolution cutoff, which were > 2.0Å, < 1.9 Å, < 1.8 Å, < 1.7 Å, < 1.6 Å, < 1.5 Å, to filter out the PDBs with resolutions larger than the cutoff from the 371-set and to plot the receiver operating characteristics (ROC) curve for each resolution cutoff. The ROC curve is plotted by the true positive rate versus (TPR) the false positive rate (FPR). TPR is defined as the number of true positive predictions (i.e., correctly predicted catalytic residues) divided by the number of total positive predictions (i.e., all predicted catalytic residues), while FPR is defined as the number of false positive predictions (i.e., incorrectly predicted catalytic residues) divided by the number of total negative predictions (i.e., all predicted non-catalytic residues). Figure 4 shows the comparison of ROC curves among different resolutions cutoffs. According to the ROC curves, the performance increased along the resolution cutoffs can be observed. Therefore, we re-selected a test set with resolution better than 2.0Å from 371-set as the training data.

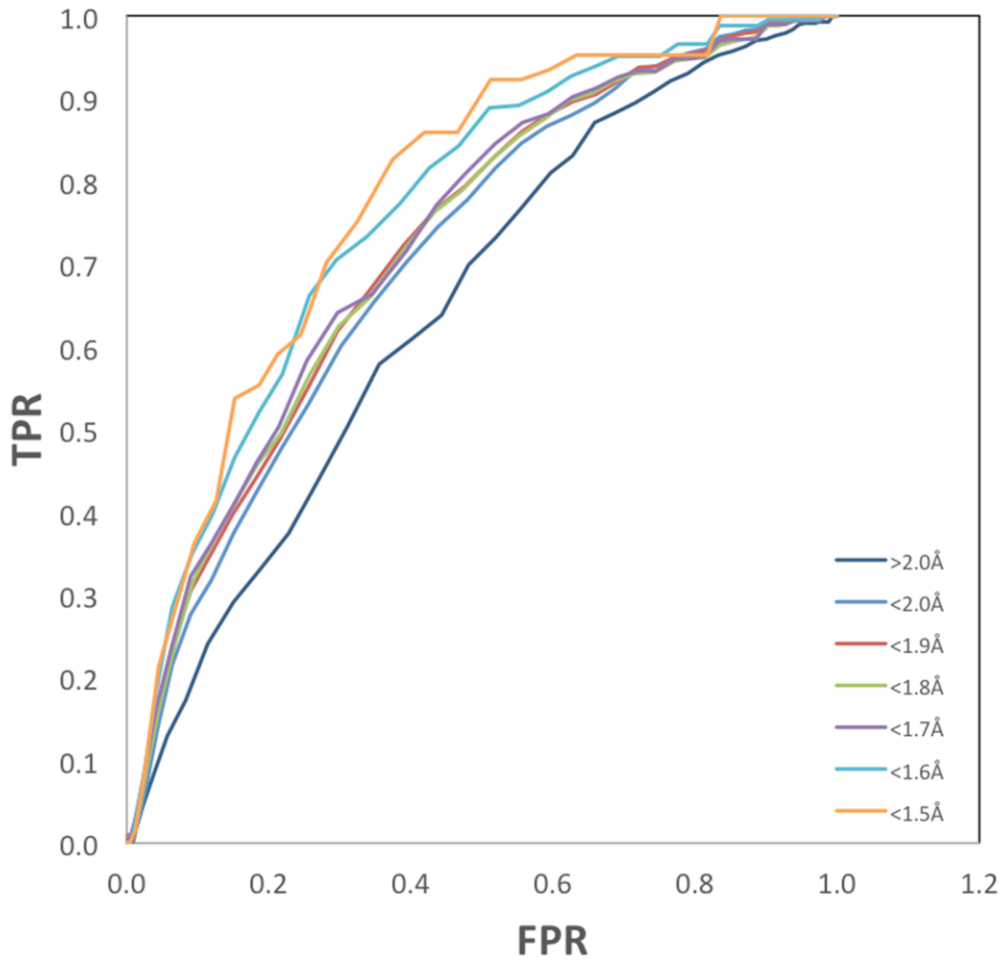

FIGURE 4. ROC CURVES SHOWING THE PREDICTION PERFORMANCES UNDER DIFFERENT RESOLUTION CUT-OFFS.

3.5 Comparison with other methods

Because our prediction method is based on atomic fluctuations, hence, to evaluate the performance, we compare the B_{TLS}-profiling method with other atomic fluctuation based method. Tile now, to my best knowledge, the weighted contact number (WCN) model is the best one based on atomic fluctuations for prediction catalytic residues. In Figure 5, we compared the prediction accuracies calculated for each PDB based on TLS and WCN models for the 371-set. The average accuracies are both around 0.70 under the cutoff setting -0.6.

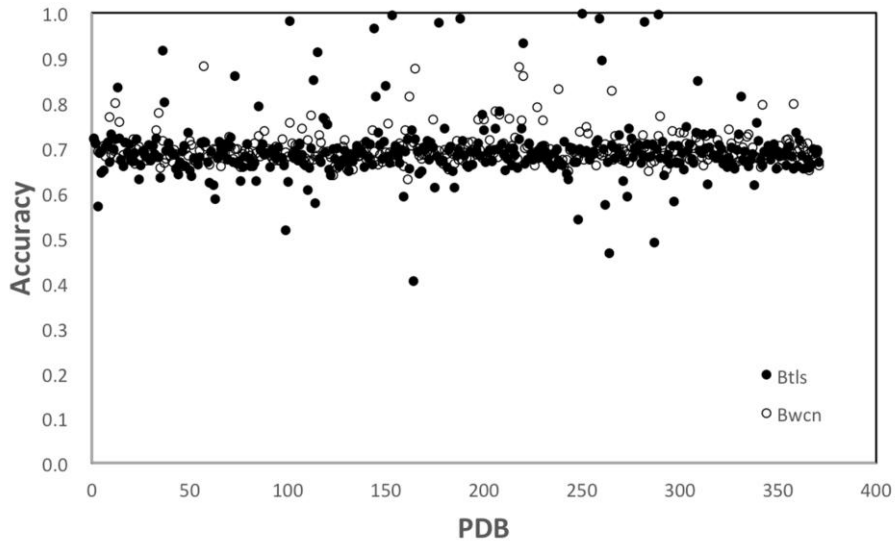

FIGURE 5. SCATTER PLOTS OF THE COMPUTED FLUCTUATIONS SHOWING THE COMPARISON BETWEEN TLS AND WCN MODELS.

IV. CONCLUSION

In this study, we used TLS model to simulate the atomic fluctuations for investigating if the signature of catalytic residues is hidden in the X-ray diffraction data. Through several designed experiments, we found TLS-derived B-factors had ability to predict the positions of catalytic residues in proteins under a cut-off value -0.6. It's not only reveal the viability for using TLS model for catalytic residue identification, but also a hint to show that the signatures of the catalytic residues might be hidden in X-ray diffraction data. Therefore, further studies are needed to clarify this point.

COMPETING INTERESTS

The authors report no conflict of interest.

ACKNOWLEDGEMENTS

This study was mainly supported by grant from 'Medical Science and Technology Center of Aiming for the Top University Program' of National Sun Yat-sen University and Ministry of Education, Taiwan.

AUTHOR CONTRIBUTIONS

Design of this method: YYL. Evaluation performance: YYL. Data analysis and discussion: YYL and CCC. Manuscript preparation: YYL and CCC.

REFERENCES

[1] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, pp. 93-6, Oct 5 2001.

[2] Y. Dou, J. Wang, J. Yang, and C. Zhang, "L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier," *PLoS One*, vol. 7, p. e35666, 2012.

[3] P. P. Pai, S. S. Ranjani, and S. Mondal, "PINGU: PredIction of eNzyme catalytic residues usinG seqUence information," *PLoS One*, vol. 10, p. e0135122, 2015.

[4] J. E. Fajardo and A. Fiser, "Protein structure based prediction of catalytic residues," *BMC Bioinformatics*, vol. 14, p. 63, 2013.

[5] S. W. Huang, S. H. Yu, C. H. Shih, H. W. Guan, T. T. Huang, and J. K. Hwang, "On the relationship between catalytic residues and their protein contact number," *Curr Protein Pept Sci*, vol. 12, pp. 574-9, Sep 2011.

[6] C. P. Lin, S. W. Huang, Y. L. Lai, S. C. Yen, C. H. Shih, C. H. Lu, *et al.*, "Deriving protein dynamical properties from weighted protein contact number," *Proteins*, vol. 72, pp. 929-35, Aug 15 2008.

[7] C. H. Shih, S. W. Huang, S. C. Yen, Y. L. Lai, S. H. Yu, and J. K. Hwang, "A simple way to compute protein dynamics without a mechanical model," *Proteins*, vol. 68, pp. 34-8, Jul 1 2007.

[8] V. Schomaker and K. N. Trueblood, "On the rigid-body motion of molecules in crystals," *Acta Crystallographica Section B,* vol. 24, pp. 63-76, 1968.

[9] D. Cruickshank, "The analysis of the anisotropic thermal motion of molecules in crystals," *Acta Crystallographica,* vol. 9, pp. 754-756, 1956.

[10] M. J. Sternberg, D. E. Grace, and D. C. Phillips, "Dynamic information from protein crystallography. An analysis of temperature factors from refinement of the hen egg-white lysozyme structure," *J Mol Biol,* vol. 130, pp. 231-52, May 25 1979.

[11] P. J. Artymiuk, C. C. Blake, D. E. Grace, S. J. Oatley, D. C. Phillips, and M. J. Sternberg, "Crystallographic studies of the dynamic properties of lysozyme," *Nature,* vol. 280, pp. 563-8, Aug 16 1979.

[12] D. C. Phillips, "Crystallographic studies of movement within proteins," *Biochem Soc Symp,* pp. 1-15, 1981.

[13] G. A. Petsko and D. Ringe, "Fluctuations in protein structure from X-ray diffraction," *Annu Rev Biophys Bioeng,* vol. 13, pp. 331-71, 1984.

[14] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, "Refinement of Macromolecular Structures by the Maximum-Likelihood Method," *Acta Crystallographica Section D,* vol. 53, pp. 240-255, 1997.

[15] B. Howlin, S. A. Butler, D. S. Moss, G. W. Harris, and H. P. C. Driessen, "Tlsanl - Tls Parameter-Analysis Program for Segmented Anisotropic Refinement of Macromolecular Structures," *Journal of Applied Crystallography,* vol. 26, pp. 622-624, Aug 1 1993.

[16] H. Driessen, M. I. J. Haneef, G. W. Harris, and B. Howlin, "Restrain - Restrained Structure-Factor Least-Squares Refinement Program for Macromolecular Structures," *Journal of Applied Crystallography,* vol. 22, pp. 510-516, Oct 1 1989.

[17] M. D. Winn, M. N. Isupov, and G. N. Murshudov, "Use of TLS parameters to model anisotropic displacements in macromolecular refinement," *Acta Crystallographica Section D-Biological Crystallography,* vol. 57, pp. 122-133, Jan 2001.

[18] C. T. Porter, "The Catalytic Site Atlas: a resource of catalytic sites and residues identi®ed in enzymes using structural data," *Nucleic Acids Research,* 2004.

[19] J. Kuriyan and W. I. Weis, "Rigid protein motion as a model for crystallographic temperature factors," *Proc Natl Acad Sci U S A,* vol. 88, pp. 2773-7, Apr 1 1991.