The balance of flexibility and rigidity in the active site residues of hen egg white lysozyme*

Qi Jian-Xun(齐建勋)[a)b)] and Jiang Fan(江 凡)[a)†]

[a)] *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*

[b)] *Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*

(Received 3 December 2010; revised manuscript received 8 January 2011)

The crystallographic temperature factors (B factor) of individual atoms contain important information about the thermal motion of the atoms in a macromolecule. Previously the theory of flexibility of active site has been established based on the observation that the enzyme activity is sensitive to low concentration denaturing agents. It has been found that the loss of enzyme activity occurs well before the disruption of the three-dimensional structural scaffold of the enzyme. To test the theory of conformational flexibility of enzyme active site, crystal structures were perturbed by soaking in low concentration guanidine hydrochloride solutions. It was found that many lysozyme crystals tested could still diffract until the concentration of guanidine hydrochloride reached 3 M. It was also found that the B factors averaged over individually collected data sets were more accurate. Thus it suggested that accurate measurement of crystal temperature factors could be achieved for medium-high or even medium resolution crystals by averaging over multiple data sets. Furthermore, we found that the correctly predicted active sites included not only the more flexible residues, but also some more rigid residues. Both the flexible and the rigid residues in the active site played an important role in forming the active site residue network, covering the majority of the substrate binding residues. Therefore, this experimental prediction method may be useful for characterizing the binding site and the function of a protein, such as drug targeting.

Keywords: crystallographic temperature factor, denaturation, conformational flexibility, active site prediction

PACS: 87.15.hp, 61.05.C– **DOI:** 10.1088/1674-1056/20/5/058701

1. Introduction

Proteins are dynamic molecules that are in constant motion. These motions range from small conformational changes due to the local movements of backbone and side chain atoms to large scale loop or domain motions, in which parts of a protein move as a whole with respect to one another.[1] Since the structural data provide the essential framework for understanding the molecular mechanism of a protein, the study of structural flexibility is necessary to understand the mechanism of molecular functions such as molecular recognition and catalytic activity.[2−10] In addition, characterizing the protein mobility and the protein flexibility will provide valuable guidelines for structure-based drug design.[11,12]

The crystallographic temperature factors or the B factors, which are derived from the x-ray analysis of a single crystal, provide a quantitative measurement on the mobility of each atom in the structure.[13] They take into account the decrease of intensity in the diffraction due to the dynamic disorder caused by the thermal vibration of atoms and the static disorder related to the imperfect orientation of the molecules.[14,15] A quantitative expression of the B factor in the isotropic case is $A = 8\pi^2\langle u^2\rangle$, where $\langle u^2\rangle$ is the mean square displacement of the atom averaged over time and crystal lattice. While in the anisotropic case, the B factors are frequently represented by anisotropic ellipsoids with six parameters per atom.[16]

The B factors have been studied from a variety of viewpoints. It has been shown that the dynamic motion in protein crystals could be interpreted by the accurately measured B factors.[17] A study of lysozyme crystals has revealed that the residues of highest apparent motion are in the active site region, where conformational change has been observed upon substrate binding.[18] Furthermore, the applications of the B factors include predicting the protein

*Project supported by the National Natural Science Foundation of China (Grant Nos. 10674172 and 10874229).

†Corresponding author. E-mail: fjiang@aphy.iphy.ac.cn

© 2011 Chinese Physical Society and IOP Publishing Ltd

flexibility,[19−22] characterizing the active sites,[23,24] analyzing the protein thermal stability,[25,26] considering the side chain mobility within the structure,[27] and investigating the protein disordered regions.[28] However, the experimentally determined B value is not an absolute quantity, which inherently depends on the measurement of Bragg intensities of sufficiently high resolution.[29] Besides, the B factors in different structures may be different owing to the applications of different refinement procedures.[30] Typically, normalized B factors[19,21,25,26,30] are compared among different structures.

The experimental determined B factors provide valuable information on the flexibility and the mobility of the protein structures. More importantly, the local flexibility of active sites may be especially helpful in understanding the protein functions. Tsou[31−33] has demonstrated that enzyme active sites may be conformationally more flexible than the overall enzyme. Jiang[34] has also shown that there are conserved "structural codes" for protein scaffold structures, which are formed at the early stage of protein folding and therefore are usually more stable than the other parts of the structures. As a result, under mild denaturing conditions, the active site conformations could be perturbed significantly without detectable global changes in the enzyme structures.

In this study, we proposed new methods for accurately measuring the B factors and for predicting the active site residues of the hen egg white lysozyme by using partially denaturing perturbation. We show that protein crystals could be perturbed with chaotropic agents such as GdnHCl,[35,36] and still give high resolution diffraction patterns. X-ray crystallography has been used to structurally characterize the interaction between the protein and the denaturants.[37−41] Here, we perturb the crystal structure by quick soaking the crystal in a low concentration GdnHCl solution and freeze the crystal structure in an ensemble of perturbed states by subsequent flash cooling the soaked crystal.[35] It will be shown that the diffraction analysis of the frozen crystal, especially the B factors averaged over multiple crystals, are more accurate and more reliable than a single measurement from one data set. The difference in the B factors between the native and the denatured crystals can be used to explore the characteristics of the active site effectively.

2. Materials and methods

2.1. Sample preparation

Hen egg white lysozyme was purchased from Sigma-Aldrich Co. and used without further purification. Prior to the crystallization, all solutions were filtered through 0.22 mm filters and left at 4 °C in a refrigerator for a few days. Crystals of tetragonal lysozyme were grown by the hanging-drop vapor diffusion method at 293 K. The crystallization drop consisted of equal volumes of protein solution (14 mg/ml lysozyme and 1.33% NaCl in a 200-mM sodium acetate buffer at pH 4.4) and reservoir solution (4% NaCl in a 200-mM sodium acetate buffer at pH 4.4). It took about 5–8 days to yield isomorphous crystals with dimensions of ~ 0.3 mm $\times$ 0.3 mm $\times$ 0.2 mm.

The denaturant GdnHCl was SigmaUltra grade and was dissolved into a 200-mM pH 4.4 sodium acetate buffer solution containing 5% of NaCl and 15% (v/v) Glycerol. Totally four concentrations of GdnHCl were used to perturb the crystal structures: 0.5, 1, 2 and 3 M.

2.2. Denaturation procedures

In order to get the protein with different crystal structures, lower concentration GdnHCl (i.e. 0–3 M) solutions were used as denaturant to perturb the crystal structures partially. There were altogether 12 crystals of almost identical dimensions, about 0.3 mm$\times$0.3 mm$\times$0.2 mm, selected and used.

The perturbed lysozyme crystals were harvested by the following procedure. First, we consecutively soaked a crystal in the cryoprotectant solutions containing increasing concentrations of Glycerol (5%, 10%, 15%) for about 3 min each. Second, we took out the crystal from the 15% (v/v) Glycerol solution and soaked it into a less than 10 ml drop containing a certain concentration of GdnHCl, for example, 0.5 M. In our experiments, the crystals were rinsed in a 2-M GndHCl for about 3 min, which is comparable to 40 s in 3-M GdnHCl. Care must be taken when perturbing the lysozyme crystal with 3-M GdnHCl, because a longer period of time (more than 1 min) would cause the crystal to shrink quickly or even break into several small pieces and dissolve into the solution quickly, which would be quite beyond the extent of so-called partially denaturing perturbation.

2.3. Data collection and structure determination

The partially denaturated crystal was quickly mounted to a goniometer head equipped with a standard cryoloop and flash cooled in a cold nitrogen-gas stream to 100 K using an Oxford cryosystem. Diffraction data were collected with an in-house X-ray source (Rigaku RA-Micro7 Desktop Rotating Anode X-Ray Generator with a Cu target at a wavelength of 1.5418 Å operating at 40 kV, 20 mA and R-AXIS IV++ imaging-plate detector). The collected intensities were indexed, integrated, corrected for the absorption, then scaled and merged using HKL2000.[42] The structures of denatured and native lysozymes were solved by the molecular replacement method using the program MOLREP[43] with 1RFP as the search model. The initial model was refined using REFMAC5,[44] and the extensive model building was performed using COOT.[45] The further rounds of refinements were performed using the phenix.refine program implemented in the PHENIX package[46] with the isotropic ADP refinement and the bulk solvent modeling. The stereochemical quality of the final model was checked using the program PROCHECK.[47] The calculation of the average B factors by residues was performed using the CCP4 module BAVERAGE.[48] The final statistics of data collection and refinement are shown in Table 1.

3. Results

3.1. Accurate measurement of the B factors

3.1.1. Statistical results with lysozyme crystals

First, 72 PDB entries of hen egg white lysozyme crystal structures containing B values were taken from an unbiased selection of the Protein Data Bank (PDB), as shown in Table 2. These 72 entries were selected for having the same amino acid sequence but different structures, i.e., the same set of atoms in different entries have different coordinates and different B factors. Secondly, structure comparisons were performed using the program LSQMAN from the Uppsala Software Factory.[49] Due to the limitation of the program that at most 26 chains or NMR models can be accommodated at one time, the 72 entries were divided equally into three groups contained 24 chains each. Within each group, the structure which had the lowest value of root mean square deviation (RMSD) in the least square superposition was considered as the central one. Corresponding to three groups, three different central structures were obtained. When the structures in the three groups were reshuffled, three different new central structures were obtained. By trial and error, in total six frequently selected central structures were found: 1lfq, 1lfu, 1jis, 1kxy, 1RFP and 1uig. When these structures were put into one group, the 1RFP structure would be the central structure. If used as the central structure for the whole set of 72 entries, the 1RFP had the smallest RMSD 0.109 Å among the six central structures. Thus it would be safe enough to select the 1RFP structure as the central structure and the reference structure for the 72 PDB entries, whose resolution was 1.75 Å.

The correlation coefficients were calculated against the 1RFP structure for the B factors. There is no clear relationship between the resolutions and the correlation coefficients, except that a good portion of the 71 structures have correlation coefficients above 0.85. It also shows that the B factors of very high resolution structures do not correlate well with the others and are not necessarily more accurate. We have found that the three structures with the highest resolutions (three data points on the far left of the Fig. 1) did not even correlate well with each other. The correlation coefficients were less than 0.60.

In order to study the relationship between the B factor of the reference structure and that of the other structures in the sample set, correlation coefficients were computed and analyzed. The residue B factors were obtained by averaging the B factors of main chain and side chain atoms. The correlation coefficients between the residue B factor of the reference structure and that of the individual structures were calculated and their distribution as a function of resolution is shown in Fig. 1. Most correlation coefficients are above 0.85. However, there is no particular linear or any other relationship between the resolution and the correlation coefficient, suggesting that an individual set of B factors is not always reliable, not even with high resolution measurements. The average of the 71 correlation coefficients is 0.8681. In contrast, the correlation coefficient between the B factor of the reference structure and that of the average structure of the remaining 71 entries is 0.9792, which is not only higher than 68 of the individual coefficients, but also indicates a fairly good consistency with the reference B factors. This trend is shown in Fig. 1. These results show that the averaged temperature factors are more accurate and reliable than that of a single measurement, even the single measurement result is from high resolution data.

Table 1. Statistics of data collection and refinement for the denatured tetragonal lysozyme crystals.

	Cryst1	Cryst2	Cryst3	Cryst4	Cryst5	Cryst6	Cryst7	Cryst8	Cryst9	Cryst10	Cryst11	Cryst12
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit-cell parameters/Å												
a	78.57	78.52	78.43	78.71	78.55	78.63	78.61	78.71	78.63	78.47	78.44	78.89
b	78.57	78.52	78.43	78.71	78.55	78.63	78.61	78.71	78.63	78.47	78.44	78.89
c	37.04	37.02	37.01	37.03	37.03	37.01	37.02	36.96	37.00	36.94	37.09	37.20
Resolution limit	35.14-1.80	35.11–1.80	39.21–1.90	39.41–1.80	39.28–1.80	39.32–1.90	39.31–1.80	39.36–2.00	39.31–2.00	39.24–1.90	39.22–2.00	39.45–2.00
	(1.86–1.80)	(1.86–1.80)	(1.97–1.90)	(1.86–1.80)	(1.86–1.80)	(1.97–1.90)	(1.86–1.80)	(2.07–2.00)	(2.07–2.00)	(1.97–1.90)	(2.07–2.00)	(2.07–2.00)
Reflections measured	275731	210188	237621	246542	279776	234629	278547	187053	208537	235402	215630	21.690
Unique reflections	11193	11186	9510	11193	10885	9579	11162	7842	8233	9048	8232	8318
Multiplicity	24.6(16.8)	18.8(12.3)	25.0(19.2)	22.0(14.8)	25.7(16.8)	24.5(18.8)	24.9(17.5)	23.9(19.2)	25.3(15.6)	24.8(13.5)	26.2(23.4)	26.3(23.4)
Emerge/%	6.1(12.2)	3.4(11.7)	6.5(24.1)	7.0(12.9)	3.9(10.2)	6.9(24.3)	8.5(17.5)	11.2(21.8)	7.5(22.1)	5.1(15.2)	8.4(16.8)	5.3(19.1)
Average/$I/\sigma(I)$	35.7(14.6)	59.5(18.5)	38.4(13.6)	32.3(13.3)	61.8(23.6)	34.0(12.9)	29.5(11.6)	22.9(10.8)	26.2(9.1)	38.4(12.2)	28.0(12.7)	39.2(15.2)
Completeness/%	99.7(99.2)	99.8(100)	99.5(100)	99.2(96.0)	97.0(100)	99.7(100)	99.3(94.7)	94.8(68.1)	99.6(99.0)	99.2(99.3)	99.7(100)	99.5(100)
$(R/R_{\rm free})$/%	18.6/23.0	18.3/22.1	18.4/19.7	20.8/25.6	18.5/22.1	18.9/20.5	19.5/24.1	20.3/24.4	19.4/22.6	21.3/23.5	20.5/23.0	19.1/23.0
rmsd bond/Å	0.007	0.002	0.002	0.003	0.005	0.004	0.005	0.002	0.003	0.002	0.002	0.001
rmsd angle/°	1.042	.0666	0.691	0.750	0.950	0.871	0.957	0.587	0.709	0.610	0.626	0.593
Ramachandran plot/%												
Favored	97.6	98.4	98.4	98.4	97.6	97.6	97.6	98.4	98.4	99.2	96.9	97.6
Allowed	2.4	1.6	1.6	1.6	2.4	2.4	2.4	1.6	1.6	0.8	3.1	2.4

Values in parentheses are for the last resolution shell. The $R = \sum |F_{\rm o} - F_{\rm c}| / \sum F_{\rm o}$ for the 95% of the reflection data used in the refinement. The $F_{\rm o}$ and $F_{\rm c}$ are the observed and the calculated structure factor amplitudes, respectively. The $R_{\rm free}$ is the equivalent of R, except that it is calculated for a randomly chosen 5% test set excluded from the refinement. The Ramachandran plot was calculated by PROCHECK. There are in total 12 independent tetragonal lysozymes crystals data sets collected after quick soaking in 4 different concentration GdnHCl solutions: Cryst1, Cryst2 and Cryst3 were soaked in the 0.5 M GdnHCl for 3 min; Cryst4, Cryst5, Cryst6 were soaked in 1M GdnHCl for 3 min; Cryst7, Cryst8, Cryst9, Cryst10 were soaked in 2 M GdnHCl for 3 min, while Cryst11 and Cryst12 were soaked in 3 M GdnHCl for about 40 s.

Table 2. PDB codes of 72 lysozyme chains used in this analysis.

PDB ID								
193lA	1bhzA	1fluA	1ioqA	1jisA	1kxy_	1lzgA	1uidA	2cdsA
194lA	1bvxA	1flwA	1iorA	1jitA	1ljnA	1lznA	1uieA	2lyoA
1akiA	1c10A	1flyA	1iosA	1jiyA	1lpiA	1n4fA	1uifA	2lz2A
1at5A	1dpwA	1fn5A	1iotA	1jj0A	1lyoA	1ps5A	1uigA	3lymA
1at6A	1dpxA	1hswA	1ir7A	1jj1A	1lzbA	1qioA	1uihA	3lyoA
1azf_	1f0wA	1hsxA	1ir8A	1jseA	1lzcA	1qtkA	1xeiA	3lztA
1b0dA	1f10A	1ieeA	1ir9A	1kxw_	1lzdA	1RFPA	1xejA	4lyoA
1bgiA	1flqA	1io5A	1jefA	1kxx_	1lzeA	1uicA	1xekA	4lztA

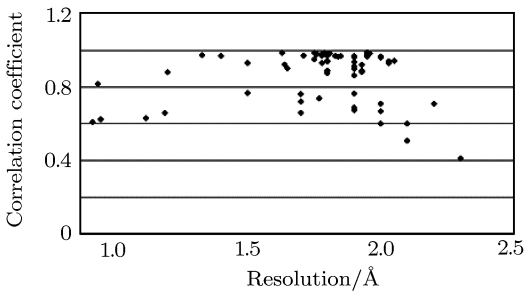


Fig. 1. Scatter plot of data resolutions and correlation coefficients.

A relative difference B factor is defined to describe more precisely the difference between the reference structure and the averaged structure. The relative difference B factor is defined as follows:

$$B_{\text{reldiff}} = 2(B_{\text{ave}} - B_{\text{ref}})/(B_{\text{ave}} + B_{\text{ref}}), \qquad (1)$$

where B_{ave} is the averaged residue B factor of the 71 structures and B_{ref} is the residue B factor of the reference structure. It is clear from the definition that the relative difference B factor is dimensionless and may be useful in evaluating the relative accuracy of the two measurements of the real B factors. A small relative difference B factor indicates that there are two stable sets of the B factors and hence there is an accurate measurement. For example, if a positive relative difference B factor occurred, there might be an under-estimation of the B factor in the 1RFP structure relative to the average B factor, while a negative one means an over-estimation. When the relative difference B factor is close to zero and stable along a stretch of residues, the region of residues in the two measurements is considered to have stable B factors.

In the current case, the relative difference B factors generally oscillate around zero with three segments more stable than that suggested by the absolute B factors of the 1RFP structure: two α helices and one β sheet (Fig. 2(a)). The first segment, the most obvious part, is the second helix (from No. 25 to 34 residues) of the reference structure, in which the relative difference B factor stably oscillates around -0.018. This is because the second α helix assembles in the inner part of the structure and the B factors measurement is more stable compared to those exposed in the solvent. The second segment, residues 88–101, corresponds to the third helix in the structure and shows the same characteristics. Similarly, the third segment is the middle β sheet ranging from residues 49 to 54 and has more stable relative difference B factors compared to the other two β sheet segments exposed to the solvent. In those regions composed of loops or exposed to the solvent, the distribution of the relative difference B factors shows a wider oscillation. From Fig. 2(a) it can be seen that big relative difference B factors can also occur in regions where absolute B factors are small. Therefore, the relative difference B factor is a more stringent test of the stability and accuracy of the B factors. More importantly, the case here indicates that the stability in the relative difference B factor means less flexibility in the 1RFP structure.

3.1.2. Perturbing the crystals with quick soaking in 0.5 M–3 M GdnHCl

In this work, 12 independent data sets of tetragonal lysozyme crystals were collected after quick soaking the crystals in four different concentration GdnHCl solutions: three data sets with 0.5 M, three data sets with 1 M, four data sets with 2 M and two data sets with 3 M. The correlation coefficients between the residue B factor of the reference structure 1RFP and that of the individual structures were cal-

culated and the results are shown in Table 3. Of the 12 correlation coefficients, the maximum is 0.8966, the minimum is 0.8310, the median is 0.8598 and the standard deviation is 0.016. The average of the 12 correlation coefficients is 0.8624. In comparison, the correlation coefficient of the B factors of the reference structure 1RFP and the averaged structure is 0.8763, which is higher than 10 of the individual correlation coefficients. The result thus demonstrates again that the averaged measurement of the crystallographic temperature factors is generally better than the individual measurement, even the individual measurement has a higher resolution. The average of the correlation coefficient matrix excluding the self correlation coefficients is 0.9540, which indicates that the independent structures after partially denaturing perturbations are still very similar.

Figure 2(b) shows the detailed distribution of the B factors in the experiment. Although only 12 data sets were obtained in 4 partially denatured conditions, it is obvious that the average B factors have the same variation pattern as that of the 1RFP structure, which is further confirmed by the correlation coefficient of 0.8763. However, it must be pointed out that the overall B factors in the experiment are smaller than that of the 1RFP structure. The average of the residue B factors in the experiment is 20.8 Å^2, comparing to 22.7 Å^2 for the reference structure. All the data sets of the partially denatured crystals were collected at 100 K, while the data of the 1RFP structure was collected at 295 K. The temperature factor is a consequence of the dynamic disorder in the crystal caused by the temperature-dependent vibration of the atoms in the structure,[14] the overall B factors will decrease accordingly at low temperatures due to less thermal vibrations, which may account for above difference. Interestingly, the relative difference B factors do offer some useful information. For example, the relative difference B factors of the segment of residues 25–34 are also stable as shown in Fig. 2(a), which indicates stable B factors and accurate measurement along this segment.

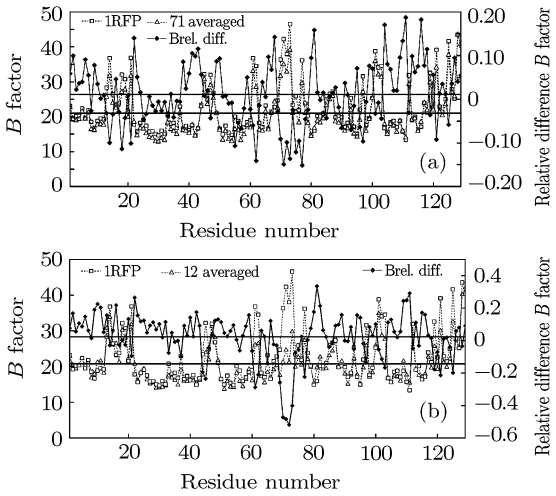


Fig. 2. Distribution of temperature factors for amino acid residues. (a) The dash lines with empty boxes are for the reference structure 1RFP, the dash lines with empty triangles are for the 71 averaged structures, the solid lines with filled diamonds are for the relative difference B factors. The two straight lines represent the average of 1RFP (lower line, 22.7 Å^2) and Brel.diff (upper line, 0.01) for the 129 amino acids respectively. (b) The dash lines with empty boxes are for the reference structure 1RFP, the dash lines with empty triangles are for the 12 averaged structures in the perturbation experiments, the solid lines with filled diamonds are for the relative difference B factors. The two straight lines represent the average of 1RFP (lower line) and Brel.diff (upper line, 0.02) for the 129 amino acids respectively.

Table 3. Correlation coefficients between different data sets.

	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8	Data9	Data10	Data11	Data12	B_12ave	1RFP
Data1	1.0000													
Data2	0.9726	1.0000												
Data3	0.9748	0.9832	1.0000											
Data4	0.9813	0.9752	0.9709	1.0000										
Data5	0.9800	0.9872	0.9761	0.9826	1.0000									
Data6	0.9776	0.9786	0.9785	0.9852	0.9809	1.0000								
Data7	0.9830	0.9620	0.9542	0.9755	0.9759	0.9648	1.0000							
Data8	0.9694	0.9447	0.9482	0.9689	0.9637	0.9635	0.9690	1.0000	1.0000					
Data9	0.9748	0.9632	0.9626	0.9837	0.9766	0.9806	0.9692	0.9776						
Data10	0.9801	0.9670	0.9683	0.9776	0.9756	0.9830	0.9760	0.9729	0.9759	1.0000				
Data11	0.9608	0.9526	0.9444	0.9594	0.9585	0.9599	0.9556	0.9517	0.9535	0.9637	1.0000			
Data12	0.9451	0.9674	0.9558	0.9476	0.9543	0.9450	0.9343	0.9292	0.9308	0.9334	0.9447	1.0000		
B_12ave	0.9905	0.9866	0.9836	0.9914	0.9915	0.9906	0.9836	0.9788	0.9866	0.9885	0.9743	0.9640	1.0000	
1RFP	0.8661	0.8728	0.8766	0.8568	0.8598	0.8683	0.8831	0.8568	0.8490	0.8592	0.8559	0.8966	0.8763	1.0000

3.1.3. Relationship between different B factor data sets

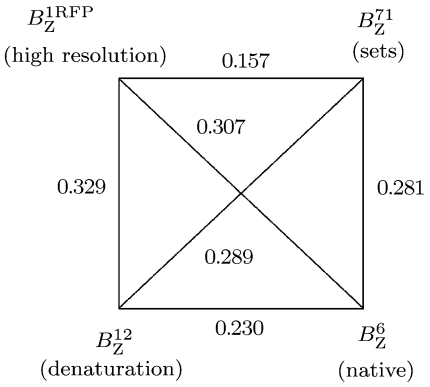
Another 6 independent data sets of native lysozyme crystals were also collected. Since the B factor data sets from the PDB files may have been handled by different refinement programs, normalized B factors are thus used for comparison and for analysis of the relationship between different data sets. The normalized B factor (Z score) is defined as follows:

$$B_{\rm Z} = \frac{B - \langle B \rangle}{\sigma(B)}, \tag{2}$$

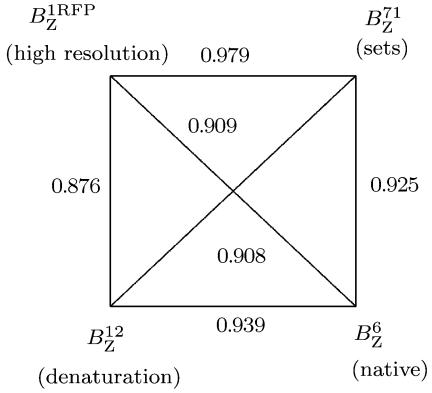
where B is the temperature factor of a certain residue in the protein sequence, $\langle B \rangle$ is the averaged temperature factor of all residues in the sequence, $\sigma(B)$ is the standard deviation of the temperature factors. By the average of multiple independent measurements, we get four sets of temperature factors: the temperature factors of the reference structure $B_{\rm Z}^{\rm 1RFP}$, those of the 71 structures $B_{\rm Z}^{71}$, those of the 12 denatured structures $B_{\rm Z}^{12}$ and those of the 6 native structures $B_{\rm Z}^{6}$. By comparing the distances defined below and the correlation coefficients between any two $B_{\rm Z}$ data sets, it is found that the distance is inversely related with the correlation coefficient. The distance between two temperature factor data sets $R_{\rm Z}$ (R-factor) is defined as follows:

$$R_{\rm Z} = \frac{\sum_j |B_{\rm Z}^{mi} - B_{\rm Z}^{ni}|}{L}, \tag{3}$$

where $B_{\rm Z}^{mi}$ is the normalized temperature factor of the i-th residue in structure m, L is the length of the protein sequence. We calculate the distances and the correlation coefficients between any two normalized B factor sets. As shown in Fig. 3, the distances of $B_{\rm Z}^{71}$, $B_{\rm Z}^{6}$ and $B_{\rm Z}^{12}$ to $B_{\rm Z}^{\rm 1RFP}$ are in the increasing order, being 0.157, 0.307 and 0.329, respectively, while their corresponding correlation coefficients are in the decreasing order, being 0.979, 0.909 and 0.876, respectively. Comparing $B_{\rm Z}^{\rm 1RFP}$ and $B_{\rm Z}^{71}$, it is found that $B_{\rm Z}^{71}$ is closer to $B_{\rm Z}^{6}$ and $B_{\rm Z}^{12}$ than to $B_{\rm Z}^{\rm 1RFP}$. This result suggests that the 71 averaged B factors are more accurate and more reliable estimations of the true B factors, because the 6 native and the 12 denatured data sets were measured in-house while all other sets were extracted from the PDB, which may come from different experimental conditions. The relationships among the four sets are the same using either the R-factor or the correlation coefficient as the measure of the distance.



(a) R-factor of $B_{\rm Z}$



(b) correlation coefficient of $B_{\rm Z}$

Fig. 3. Diagram summarizing the relationships between four sets of B factors obtained with different methods and sources. (a) The diagram uses the relative R-factor, similar to the crystallographic R-factor, to measure the distance between the data sets. (b) The diagram uses the correlation coefficient. The higher the correlation coefficient is, the smaller the distance is.

3.2. Prediction of the active site residues

The 1.8 Å resolution structure 1LZE, a complex of lysozyme with carbohydrate, is selected from the PDB. The residues within 6 Å of the carbohydrate are defined as the active site residues of 1LZE. There are in total 23 residues within 6 Å of the carbohydrate, all of which are regarded as the active site residues, as shown in Table 4. Furthermore, the two structures 1RFP and 1LZE are compared through the DaliLite program.[50] The RMSD of the main chain C_α atoms is only 0.2 Å. Therefore, it is reasonable to use these 23 residues as the active site residues of lysozyme in this study.

Table 4. The distribution of correctly predicted active site residues by $B_{\rm Z}^{12} - B_{\rm Z}^{6}$.

Condition	Residue number	B^{12}/Å^2	B^{6}/Å^2
$B_{\rm Z}^{12} - B_{\rm Z}^{6} > \langle B_{\rm Z}^{12} - B_{\rm Z}^{6}\rangle + 2\sigma$	none	none	none
$\langle B_{\rm Z}^{12} - B_{\rm Z}^{6}\rangle + 2\sigma >$ $B_{\rm Z}^{12} - B_{\rm Z}^{6} > \langle B_{\rm Z}^{12} -$ $B_{\rm Z}^{6}\rangle + \sigma$	21, 22, 81, 82, 83, 84, 85, 86, **99**, **100**, **101**, **102**, **103**	18.3 (**99**), 25.8 (**100**), 34.9 (**101**), 28.7 (**102**), 27.6 (**103**)	12.6 (**99**), 17.9 (**100**), 25.8 (**101**), 21.0 (**102**), 20.6 (**103**)
$\langle B_{\rm Z}^{12} - B_{\rm Z}^{6}\rangle - 2\sigma$ $< B_{\rm Z}^{12} - B_{\rm Z}^{6} < \langle B_{\rm Z}^{12} -$ $B_{\rm Z}^{6}\rangle - \sigma$	**43**, **44**, **45**, **48**, 49, **50**, **58**, **59**, 68, **73**, 124, 125, 126	16.2 (**43**), 17.7 (**44**), 23.9 (**45**), 28.1 (**48**), 16.3 (**50**), 16.4 (**58**), 19.2 (**59**), 29.7 (**73**)	13.6 (**43**), 18.1 (**44**), 19.7 (**45**), 24.2 (**48**), 13.7 (**50**), 11.9 (**58**), 19.2 (**59**), 23.3 (**73**)
$B_{\rm Z}^{12} - B_{\rm Z}^{6} < \langle B_{\rm Z}^{12} -$ $B_{\rm Z}^{6}\rangle - 2\sigma$	**60**, **61**, **62**, 69, 70, **71**, **72**	16.0 (**60**), 26.1 (**61**), 26.4 (**62**), 24.6 (**71**), 21.2 (**72**)	13.2 (**60**), 29.5 (**61**), 24.0 (**62**), 22.8 (**71**), 23.0 (**72**)

The 23 active sites are: GLU**35**, ASN**44**, ARG**45**, ASN**46**, ASP**52**, LEU**56**, GLN**57**, ILE**58**, ASN**59**, SER**60**, TRP**62**, TRP**63**, ARG**73**, LEU**75**, ILE**98**, ASP**101**, GLY**102**, ASN**103**, ASN**106**, ALA**107**, TRP**108**, VAL**109**, ALA**110**. The correctly predicted active residue numbers are in boldface. The numbers in the parentheses are the corresponding residue numbers for the B values. The average of B^{12} is 20.76 Å^2, while the average of B^{6} is 16.14 Å^2.

It is generally thought that the active sites of an enzyme tend to be more flexible and more sensitive to partial denaturation and thus may be more likely to associate with high B factors. By the definition of the normalized B factor, the difference B factors $B_{\rm Z}^{71} - B_{\rm Z}^{\rm 1RFP}$, $B_{\rm Z}^{12} - B_{\rm Z}^{\rm 1RFP}$, $B_{\rm Z}^{6} - B_{\rm Z}^{\rm 1RFP}$ and $B_{\rm Z}^{12} - B_{\rm Z}^{6}$ are calculated and identified. All the residues with high B factors (whose Z-scores are about one or more σ in the four data sets) are then compared with the active site region, allowing plus or minus 2 residues in position along the sequence. It should be pointed out that the two N terminal residues and the two C terminal residues are excluded from the data set statistics because of their much higher flexibility. It is clear that high B factor residues do not cover the whole active region. However, when low B factor residues (whose Z-scores are less than $-\sigma$) are included, most of the active region is covered. Thus all the residues, whose B factor Z-scores are one σ more or one σ less than the average value, are predicted as the active site residues. Take the difference B factor $B_{\rm Z}^{12} - B_{\rm Z}^{6}$ as an example, the predicted active site residue numbers and their corresponding refined temperature factors are shown in Table 4. The $\langle B_{\rm Z}^{1} - B_{\rm Z}^{6}\rangle$ is the averaged $(B_{\rm Z}^{12} - B_{\rm Z}^{6})$. It is found that the predicted active site residues include not only the residues with high B factors, but also some residues with low B factors. Thus the predicted active site residues could be divided in to two groups: the flexible residues (corresponding to $B_{\rm Z}^{12} - B_{\rm Z}^{6} > \langle B_{\rm Z}^{12} - B_{\rm Z}^{6}\rangle + \sigma$) whose B factors rise after partial denaturation and the more rigid residues (corresponding to $B_{\rm Z}^{12} - B_{\rm Z}^{6} < \langle B_{\rm Z}^{12} - B_{\rm Z}^{6}\rangle - \sigma$) whose B factors remain small or even decrease after partial denaturation.

In order to evaluate the prediction efficiency of different data sets, a series of evaluation parameters are defined as follows:

$$\text{Specificity} = \frac{\text{The number of correct predictions}}{\text{The number of active site residues}} \times 100\%, \quad (4)$$

$$\text{Coverage} = \frac{\text{The number of total predictions}}{\text{The total number of residues}} \times 100\%, \quad (5)$$

$$\text{Sensitivity} = \frac{\text{The number of correct predictions}}{\text{The number of total predictions}} \times 100\%. \quad (6)$$

Table 5 lists the prediction performance obtained by using different data sets of B factors. The first three sets are predictions based on difference B factors using 1RFP as the reference set, while the last row is a difference set between the 12 denatured set (averaged over 12 sets) and the 6 native set (averaged over 6 sets). The difference B factor $B_{\rm Z}^{12} - B_{\rm Z}^{\rm 1RFP}$ has the lowest prediction specificity, giving only 39%. Comparatively, the difference B factor $B_{\rm Z}^{12} - B_{\rm Z}^{6}$ has the highest prediction specificity of 78%. The other two data sets $B_{\rm Z}^{71} - B_{\rm Z}^{\rm 1RFP}$ and $B_{\rm Z}^{6} - B_{\rm Z}^{\rm 1RFP}$give equally

prediction specificity of 61%. For the prediction coverage, the four data sets show the same characteristics as the prediction sensitivity. The $B_{\rm Z}^{12} - B_{\rm Z}^{6}$ and the $B_{\rm Z}^{6} - B_{\rm Z}^{1{\rm RFP}}$ give the highest and lowest prediction coverage respectively. Too high or too low coverage will definitely affect the prediction specificity (or accuracy). In terms of prediction sensitivity, the difference B factors $B_{\rm Z}^{12} - B_{\rm Z}^{1{\rm RFP}}$ and $B_{\rm Z}^{6} - B_{\rm Z}^{1{\rm RFP}}$ have the lowest prediction sensitivity, all less than 50%. Whereas the difference B factors $B_{\rm Z}^{71} - B_{\rm Z}^{1{\rm RFP}}$ and $B_{\rm Z}^{12} - B_{\rm Z}^{6}$ have the highest efficiencies, giving 50% and 55% respectively. These results suggest that using the difference B factors and the data sets from partial denaturation can increase the prediction performance considerably. Figure 4 shows the distribution of the difference B factors $B_{\rm Z}^{12} - B_{\rm Z}^{6}$, where all the residues with the difference B factors higher than the average plus σ or lower than the average minus σ are the predicted active residues.

Table 5. The prediction efficiency statistics.

Data set	Specificity /%	Coverage /%	Sensitivity /%
$B_{\rm Z}^{71} - B_{\rm Z}^{1{\rm RFP}}$	61	22	50
$B_{\rm Z}^{6} - B_{\rm Z}^{1{\rm RFP}}$	61	22	48
$B_{\rm Z}^{12} - B_{\rm Z}^{1{\rm RFP}}$	39	17	41
$B_{\rm Z}^{12} - B_{\rm Z}^{6}$	78	26	55

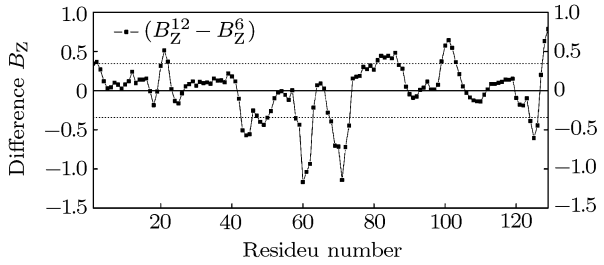


Fig. 4. Distribution of the difference B factors between the 12 denatured set and the 6 native set. The $B_{\rm Z}$ means the B factors are normalized.

4. Discussion

To identify and to characterize protein–protein interactions and their networks, a wide spectrum of experimental methods have been developed. These methods include yeast two-hybrid (Y2H),[51,52] tandem affinity purification (TAP)[53] and high-throughput mass spectroscopy protein complex identification (HMS-PCI).[54] Protein interaction databases [55,56] have also been developed to integrate protein interaction information from disparate sources. X-ray crystallography has been the most prolific technique for the structural analysis of protein and protein complex in terms of accuracy and resolution, but the number of solved protein complexes remains small.[57] So there is a need to develop other methods for analyzing protein interactions. In this work, based on the theory of the conformational flexibility of the active site regions,[31−33] we have developed a new method for characterizing the protein interaction sites by perturbing the crystal structure using quick soaking at various low concentration GdnHCl solutions. It has been shown that the partial denaturation by low concentration denaturing agents do not destroy the lysozyme crystals completely, although it lowers the diffraction resolution slightly, from 1.7 to 2.0 Å. This supports our hypothesis that the overall structure or the folding scaffold of a protein is preserved during the partial denaturation.

In addition, it has been shown that the averaged B factors from multiple measurements are indeed more accurate and more reliable than a single measurement. This result is consistent with the basic principles of measurement errors in experiment. Therefore, this method is particular useful when no high resolution diffracting crystals are available and accurate determination of B factors are still needed, although repeated measurements using multiple data sets are necessary. The accurately measured B factors would also benefit the calculation of phase information in molecular replacement.[58,59]

Combining the above two new experimental methods, we have extended the idea that the active site residues are more sensitive to denaturing agents and have designed a new method for predicting the active site residues by using the difference B factors between the native and the partial denaturated data sets. This method offers a new experimental way to predict the interaction sites of a protein. Such information is crucial in understanding the biological functions of a protein.

Another very interesting result of our study is that the active site residues are composed of not only more flexible (B factors rise after denaturation) residues but also more rigid (B factors remain stable or even decrease after denaturation) residues. It seems that the active site residues of lysozyme should be more flexible or more mobile than the whole molecule on average and consequently would be more sensitive

to the perturbation by denaturants or by physical factors such as temperature. Take the B values of the native and the denatured data sets as an example, we have found that both the absolute B values and the normalized B values of some active site residues are higher than their corresponding average values for the whole molecule (Table 4). This conforms to the results of Artymiuk[18] that the active site residues are more flexible than the non-active site residues. However, it must be pointed out that while some of the active site residues have raised their B values, the others have remained or even reduced their B values and therefore have become more rigid in the process of denaturation. Yuan[24] and Bartlett[60] have shown that the B values of active site residues are on average lower than that of the non-active residues. The appearance of both relatively flexible and rigid residues in the active site is likely linked to functions.[61,62] So the active site residues may be connected through a network that covers the most part of the substrate binding sites with a dynamic pattern of flexibility and rigidity.

It must be pointed out that our method was only tested with the lysozyme crystals. In the future, we will carry out tests with other crystals. In the absence of comparable methods, we cannot evaluate the performance of our prediction method. Nevertheless, for a protein of unknown function, the information of its potential active site residues could be very useful and may be obtained by the difference of the normalized B factors between the native and the denaturated crystals. It is expected that our method may also be useful for drug discovery, such as the evaluation of drug targets.

5. Conclusions

A method for accurately measuring the crystallographic temperature factors has been described in which GdnHCl is used as a chemical denaturant to partially perturb the structure of a crystal. The multiple independent data sets are used to calculate the average B factors for the crystal structure. The results show that the crystallographic temperature factors measured in this way may be as accurate as a single measurement obtained at a higher resolution. The method is further explored by developing a method for predicting the active site residues with a high accuracy. This prediction method may be useful for characterizing the binding site and the function of a protein.

Acknowledgements

We thank professor Qian Min-Xie for the useful suggestions and fruitful discussions during the project.

References

[1] Gerstein M and Krebs W 1998 *Nucleic Acids Res.* **26** 4280
[2] Demchenko A P 2001 *J. Mol. Recognit.* **14** 42
[3] Liu J, Tan H and Rost B 2002 *J. Mol. Biol.* **322** 53
[4] Palmer A G 2001 *Annu. Rev. Biophs. Biomol. Struct.* **30** 129
[5] Tainer J A, Getzoff E D, Alexander H, Houghten R A, Olson A J, Lerner R A and Hendrickson W A 1984 *Nature* **312** 127
[6] Dunker A K, Brown C J, Lawson J D, Iakoucheva L M and Obradovic Z 2002 *Biochemistry-Us* **41** 6573
[7] Daniel R M, Dunn R V, Finney J L and Smith J C 2003 *Annu. Rev. Biophs. Biomol. Struct.* **32** 69
[8] Dyson H J and Wright P E 2005 *Nat. Rev. Mol. Cell Bio.* **6** 197
[9] Tompa P 2005 *FEBS Lett.* **579** 3346
[10] Dodson G and Verma C S 2006 *Cell Mol. Life Sci.* **63** 207
[11] Ahmed A, Kazemi S and Gohlke H 2007 *Frontiers in Drug Design & Discovery: Structure-Based Drug Design in the 21st Century* **3** 455
[12] Das K, Lewi P J, Hughes S H and Arnold E 2005 *Prog. Biophys. Mol. Biol.* **88** 209
[13] Trueblood K N, Burgi H B, Burzlaff H, Dunitz J D, Gramaccioli C M, Schulz H H, Shmueli U and Abrahams S C 1996 *Acta Cryst.* A **52** 770
[14] Drenth J 1999 *Principles of Protein X-ray Crystallography* (New York:Springer Verlag) p81
[15] Rossmann M G and Arnold E 2006 *International Tables for Crystallography* (Dordrecht: Kluwer academic publishers) p. 369
[16] McRee D 1999 *Practical Protein Crystallography* (New York: Academic) p. 96
[17] Sternberg M J E, Grace D E P and Phillips D C 1979 *J. Mol. Biol.* **130** 231
[18] Artymiuk P J, Blake C C F, Grace D E P, Oatley S J, Phillips D C and Sternberg M J E 1979 *Nature* **280** 563
[19] Karplus P A and Schulz G E 1985 *Naturwissenschaften* **72** 212
[20] Vihinen M, Torkkila E and Riikonen P 1994 *Proteins* **19** 141
[21] Smith D K, Radivojac P, Obradovic Z, Dunker A K and Zhu G 2003 *Protein Sci.* **12** 1060
[22] Schlessinger A and Rost B 2005 *Proteins-Structure Function and Bioinformatics* **61** 115
[23] Carugo O and Argos P 1998 *Proteins-Structure Function and Bioinformatics* **31** 201

[24] Yuan Z, Zhao J and Wang Z X 2003 *Protein Eng.* **16** 109
[25] Vihinen M 1987 *Protein Eng.* **1** 477
[26] Parthasarathy S and Murthy M R N 2000 *Protein Eng.* **13** 9
[27] Carugo O and Argos P 1997 *Protein Eng.* **10** 777
[28] Radivojac P, Obradovic Z, Smith D K, Zhu G, Vucetic S, Brown C J, Lawson J D and Dunker A K 2004 *Protein Sci.* **13** 71
[29] Kuhs W F 1992 *Acta Cryst.* A **48** 80
[30] Tronrud D E 1996 *J. Appl. Cryst.* **29** 100
[31] Tsou C L 1986 *Trends Biochem. Sci.* **11** 427
[32] Tsou C L 1993 *Science* **262** 380
[33] Tsou C L 1998 *Biochemistry-Moscow* **63** 253
[34] Jiang F and Li N 2007 *Chin. Phys.* **16** 392
[35] Dill K A and Shortle D 1991 *Annu. Rev. Biochem.* **60** 795
[36] Tanford C 1968 *Advan. Prot. Chem.* **23** 121
[37] Pike A C W and Acharya K R 1994 *Protein Sci.* **3** 706
[38] Dunbar J, Yennawar H P, Banerjee S, Luo J B and Farber G K 1997 *Protein Sci.* **6** 1727
[39] Ratnaparkhi G S and Varadarajan R 1999 *Proteins* **36** 282
[40] Mande S C and Sobhia M E 2000 *Protein Eng.* **13** 133
[41] Salem M, Mauguen Y and Prange T 2006 *Bba-Proteins Proteom* **1764** 903
[42] Otwinowski Z and Minor W 1997 *Methods Enzymol.* **276** 307
[43] Vagin A and Teplyakov A 1997 *J. Appl. Cryst.* **30** 1022
[44] Winn M D, Isupov M N and Murshudov G N 2001 *Acta Cryst.* D **57** 122
[45] Emsley P and Cowtan K 2004 *Acta Cryst.* D **60** 2126
[46] Adams P D, Grosse-Kunstleve R W, Hung L W, Ioerger T R, McCoy A J, Moriarty N W, Read R J, Sacchettini J C, Sauter N K and Terwilliger T C 2002 *Acta Cryst.* D **58** 1948
[47] Laskowski R A, Macarthur M W, Moss D S and Thornton J M 1993 *J. Appl. Cryst.* **26** 283
[48] Dodson E, Winn M and Ralph A 1997 *Methods Enzymol.* **277** 620
[49] Kleywegt G J 1996 *Acta Cryst.* D **52** 842
[50] Holm L and Park J 2000 *Bioinformatics* **16** 566
[51] Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S and Sakaki Y 2000 *Proc. Natl. Acad. Sci.* **97** 1143
[52] Uetz P, Giot L, Cagney G, Mansfield T A, Judson R S, Knight J R, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S and Rothberg J M 2000 *Nature* **403** 623
[53] Gavin A C, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J M, Michon A M, Cruciat C M, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch M, Bastuck S, Huhse B, Leutwein C, Heurtier M A, Copley R R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G and Superti-Furga G 2002 *Nature* **415** 141
[54] Ho Y, Gruhler A, Heilbut A, Bader G D, Moore L, Adams S L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems A R, Sassi H, Nielsen P A, Rasmussen K J, Andersen J R. Johansen L E, Hansen L H, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen, Sorensen B D, Matthiesen J, Hendrickson R C, Gleeson F, Pawson T, Moran M F, Durocher D, Mann M, Hogue C W Figeys D and Tyers M 2002 *Nature* **415** 180
[55] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G 2002 *FEBS Lett.* **513** 135
[56] Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D and Apweiler R 2004 *Nucleic Acids Res.* **32** D452
[57] Tong A H, Evangelista M, Parsons A B, Xu H, Bader G D, Page N, Robinson M, Raghibizadeh S, Hogue C W, Bussey H andrews B, Tyers M and Boone C 2001 *Science* **294** 2364
[58] Jiang F and Ding W 2010 *Chin. Phys.* B **19** 106101
[59] Zhang T, Wu L J, Gu Y X, Zheng C D and Fan H F 2010 *Chin. Phys.* B **19** 096101
[60] Bartlett G J, Porter C T, Borkakoti N and Thornton J M 2002 *J. Mol. Biol.* **324** 105
[61] Cui Q and Karplus M 2008 *Protein Sci.* **17** 1295
[62] Schlessinger A, Yachdav G and Rost B 2006 *Bioinformatics* **22** 891