Identification of key residues in proteins by using their physical characters

Changjun Chen, Lin Li, and Yi Xiao*

Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China, and The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy

(Received 14 November 2005; published 26 April 2006)

Key residues in proteins are important to their stability, folding, and functions. They usually are highly conserved and can be identified by sequence or structure alignments. However, these methods can only determine the locations of key residues in sequences and structures and give less information about their physical characters. In this paper, we try to identify key residues by analyzing their inter-residue interactions. The model we study is the G_β protein domain from transducin. We show that the usual Gaussian network analysis and distance-based contact analysis have difficultly identifying the key residues in this protein, but the contact energies can do it well. We find that most key residues can be located by the lowest contact energies. This enables us to predict and analyze the key residues in other proteins. Our results suggest that contact energy analysis may provide an alternative approach to investigating the folding and stability of proteins.

DOI: 10.1103/PhysRevE.73.041926 PACS number(s): 87.15.Aa, 87.15.By

I. INTRODUCTION

It is interesting that proteins with greatly different sequences may have similar three-dimensional (3D) structures. How can this happen? It is generally believed that this is because there exist a few common residues in their sequences, which play key roles in the folding and stability of their structures. Due to these special residues, proteins with different sequences can form similar 3D structures. These special residues are called conserved residues or key residues. According to the roles in the proteins, the conserved residues can be classified into two types: Functionally and structurally- conserved residues. Obviously, the former are related to protein functions and are only distributed at the active sites of proteins. The latter are related to protein structures and are usually distributed in the cores of proteins.

For the functionally conserved residues, Buyong and coworkers studied the conservation of the residues in the protein-protein binding sites [1]. They used the multiple structure alignment (MUSTA) [2,3] to identify the conserved residues according to structural characters. They found that most of these conserved residues around the binding site are sequentially conserved, and their conservation properties correlate well with experimental enrichment of hot spots. Hot spots represent the special residues in the binding sites that have greatest binding energies. They play a critical role in the protein-protein interaction and drug discovery [4]. Therefore, the good correlation between residue conservations and hot spots in active sites shows that the investigations of functionally conserved residues are very interesting and significant.

On the other hand, it is argued that structurally conserved residues are especially important in protein stability and folding. They are even involved in the folding nucleus. According to the "nucleation-condensation" mechanism, proteins cannot fold until some definite residues in them aggregate together and form a stable folding nucleus. It is a rating limiting process, and the protein would fold fast to its native state as soon as the folding nucleus is formed [5]. Mirny and Shakhnovich did a statistical conservation analysis of nine proteins families [6]. After a comparison with the experimental results of protein engineering, they found that most residues in the folding nucleus are much more conserved than other residues in eight of the nine protein families. So investigating structurally conserved residues can really help us understand the protein folding mechanism.

Some experiments also confirmed this view. For example, Venkat and co-workers did a site-directed mutation analysis on C5 protein (the protein cofactor), which associates with RNase P [7]. They demonstrated that certain conserved residues in C5 protein are much more important than others. Even mutating a single residue of this kind in C5 protein can change the stability and substrate specificity of the RNase P holoenzyme. Their work indicates the key role of conserved residues in proteins experimentally.

Conserved residues or key residues are very important to protein folding and functions, but finding them and understanding their physical characters are still unsolved problems. Traditionally, key residues can be located by sequence or structure alignments. These methods are efficient and practical, but it is hard to get more information about how the key residues interact with others. In this paper, we try to identify the key residues in proteins and investigate their physical characters by analyzing the interresidue interactions. We find that it is difficult to determine the key residues in the G_β protein domain by the usual Gaussian network method and distance-based contact analysis. However, the contact energies calculated by an all-atom force field can identify them from other residues.

*Corresponding author. Electronic address: yxiao@mail.hust.edu.cn

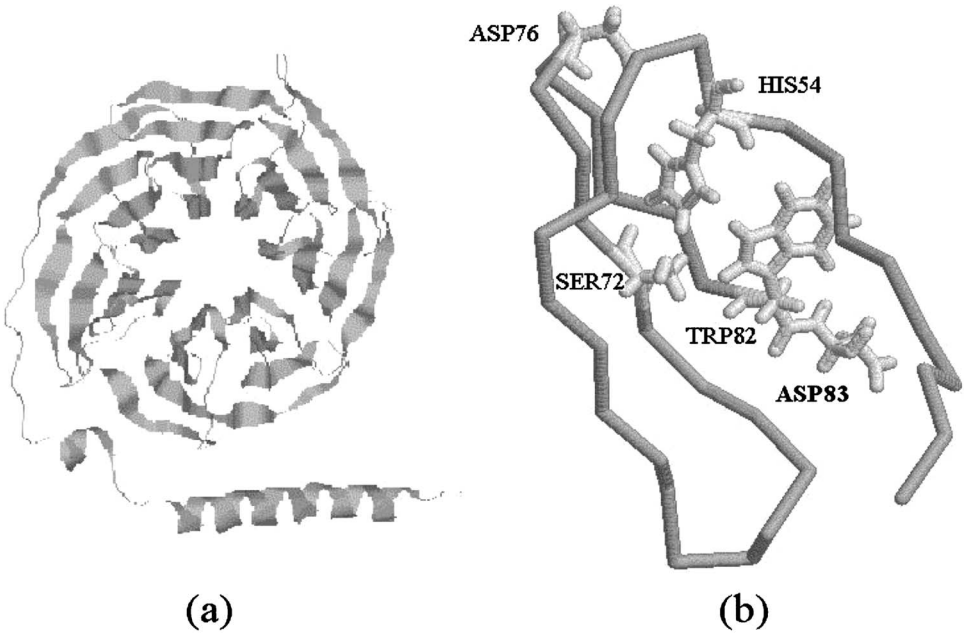

FIG. 1. (a) Ribbon diagram for the tertiary structure of the protein domain G_β from transducin (PDB id: 1tbg). (b) Five key residues in the first blade of the protein. The picture was created by the software RASMOL.

II. METHODS AND MATERIALS

In this paper, the model protein we use is the protein domain G_β from transducin (PDB id: 1tbg) [8,9]. It is a propellerlike protein, which is composed of seven blades, or called WD-repeats [Fig. 1(a)]. We selected this protein because it has a high structural symmetry and a low subsequence similarity between its subunits. Therefore, it is relatively easy to identify the conserved residues in this protein because they should also have the same symmetry. Furthermore, the key residues in this protein are well known and have already been identified by other methods. From a structure-based sequence alignment, it can be observed that there are five residues that are almost totally invariant in each repeat of the protein [Fig. 1(b)] [10–12]. These structurally conserved residues connect the outer strand of each blade to the inner three strands of the next blade, and are certainly considered as key residues critical for the structural stability of the G_β protein.

First, we use two traditional methods to analyze the physical features of residues in the protein. They are Gaussian network model (GNM) [13,14] and distance-based contact theory.

GNM is an elastic network model and can be seen as a reduced model of normal mode analysis. It treats the protein structure as many beads connected by harmonic springs. Previous works show that the mean-square fluctuations of residues calculated by GNM are in excellent agreement with experimental temperature factors [15]. So, we apply this method to find how the residues in proteins are restricted to their native positions. The mean-square fluctuation is defined as follows:

$$\langle \Delta r_i^2 \rangle = k_B T / \gamma [\Gamma^{-1}]_{ii} \tag{1}$$

where r_i is the position coordinate of the ith residue, Γ is a symmetric matrix known as Kirchhoff or connective matrix (cutoff distance is 7.5 Å), γ is the force constant of the harmonic spring, and T is the temperature.

Another method we use is contact theory [16,17]. Contact theory simply describes the interaction of each residue with others. Usually, contacts are defined by distance criteria. It is assumed that one contact is formed between two residues when the distance between their C^α atoms is less than 7.5 Å.

$$\mathcal{N}_{Ci} = \sum_{j=i, j\neq i}^{N} \delta_{ij} \quad \delta_{ij} = \begin{cases} 1 & d_{ij} \leqslant 7.5 \text{ Å} \\ 0 & d_{ij} > 7.5 \text{ Å} \end{cases}, \tag{2}$$

where $\mathcal{N}_{Ci}$ is the contact number for each residue i, and d_{ij} is the distance between residues i and residues j.

Second, we shall analyze the physical features of residues in the protein by inter-residue interactions. We redefine contacts by contact energy. We assume one contact is formed when the potential energy between the residues is lower than −0.5 kcal/mol:

$$\mathcal{N}_{Ei} = \sum_{j=i, j\neq i}^{N} c_{ij} \quad c_{ij} = \begin{cases} e_{ij} & e_{ij} \leqslant -0.5 \text{ kcal/mol} \\ 0 & e_{ij} > -0.5 \text{ kcal/mol} \end{cases}, \tag{3}$$

where $\mathcal{N}_{E_i}$ is the contact energy for each residue i, and e_{ij} is the potential energy between residues i and residues j, which is calculated by an all-atom force field. To justify the energy threshold, we calculated the contact number versus distance and energy for the G_β protein (Fig. 2). It can be seen from Fig. 2 that the contact number varies quickly with distance, while it remains almost the same when we select different "energy" thresholds less than about −0.3 kcal/mol. The contact number is not sensitive to the energy threshold. So, we

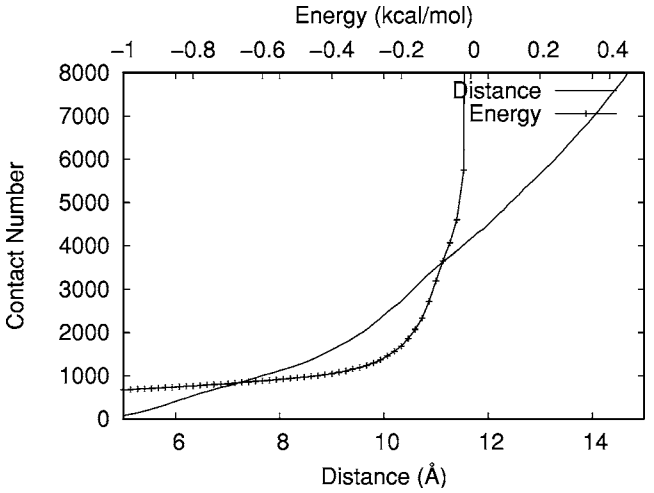

FIG. 2. Contact number versus distance and contact number versus energy for protein 1tbg.

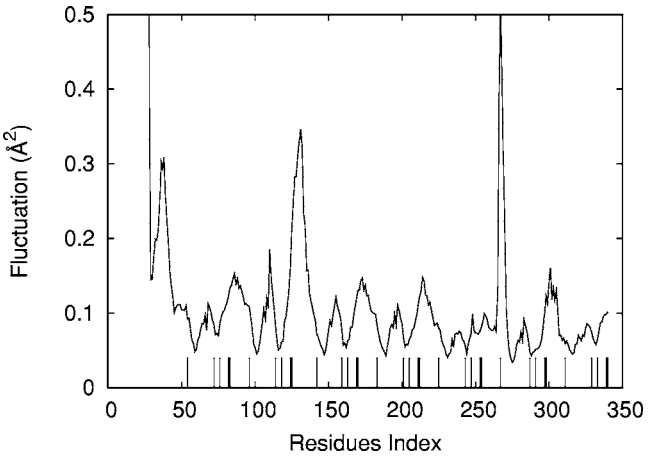

FIG. 3. Mean-square fluctuations for each residue in the slowest six modes in protein G_β from transducin (PDB id 1tbg). The vertical lines in the bottom of the picture indicate the positions of the key residues in the seven blades.

set it as −0.5 kcal/mol. Another reason we choose −0.5 kcal/mol is that the contact number calculated with this energy threshold is similar to that with a distance threshold of 7.5 Å (see Fig. 2). In fact, different energy thresholds do not change the positions of the key residues with lowest contact energies, even if the contact number changes. In our analysis, we are interested in those contacts with significantly lower energies than others. Only these key residues, with the lowest contact energies, are most important for analyzing the folding and stability of proteins.

In our calculation, we use the Generalized Born/Surface Area (GB/SA) model [18,19] as an implicit solvent model to simulate the aqueous environment. GB/SA is a reduced model from the continuum model, which treats water as a continuous medium, and there are usually three terms included in the free energy of solvation:

$$\Delta G_{\text{sol}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdw}} + \Delta G_{\text{pol}}, \tag{4}$$

where ΔG_{cav} is a solvent-solvent cavity term, corresponding to the free energy of creating a cavity of solute in the solvent continuum, ΔG_{vdw} is the free energy term representing the interactions between the solute and solvent, and ΔG_{pol} denotes electrostatic interactions between the solute and solvent. The advantage of this model is that it does not treat solvent molecules explicitly and greatly saves computational time.

The sum of the first two terms in Eq. (4) is often regarded as proportional to the solvent-accessible surface area of the solute

$$\Delta G_{\text{cav}} + \Delta G_{\text{vdw}} = \sum_{i=1}^{N} \sigma_i A_i, \tag{5}$$

where A_i is the solvent-accessible surface area of the atom i and σ_i is a special empirical parameter corresponding to atom i. Generally, all σ_i are 5.0 cal/mol/Å^2 and the solvent probe radius is 1.4 Å.

The occurrence of the last term, in Eq. (4), ΔG_{pol}, is due to the polarization of the solvent which is caused by the solute. The charge distribution of the solute directly determines that of the solvent, which in turn influences the solute reversely.

To obtain ΔG_{pol}, the most precise method is to solve the Poisson-Boltzmann (PB) equation, the result of which is very close to that of explicit water. However, it is still too slow to be applicable in normal molecular dynamics (MD) simulations. Recently, some numerical methods related to solving the PB equation have been published, which promise to be used widely.

Another attractive approach to calculate ΔG_{pol} in Eq. (4) is to use the generalized Born (GB) model proposed by Still and co-workers [18] and developed by others [19]. This model calculates ΔG_{pol} as follows:

$$\Delta G_{\text{pol}} = -166.0\left(1 - \frac{1}{\epsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}}}}, \tag{6}$$

where $D_{ij} = r_{ij}^2/4\alpha_i\alpha_j$, and r_{ij} is the distance between atom i and atom j. q_i and q_j are the charges of atom i and atom j. ϵ is the dielectric constant of the solvent. The most important, α_i, is the effective Born radius of atom i, which is related to the effective Born free energy of solvation.

In this paper, the software we use is TINKER (see: http://dasher.wustl.edu/tinker/) with CHARMM27 force field [20]. Before formal analysis, we optimize all structures with the conjugate-gradient method and the gradient tolerance is 0.2 kcal/(Å mol).

III. RESULTS AND DISCUSSIONS

First, we analyze the protein G_β with the GNM method, which can show the fluctuation behaviors of residues around the equilibrium positions. In the last seven years, GNM has been widely applied in the study of protein structures [21–26]. It has been proved that GNM can effectively high-

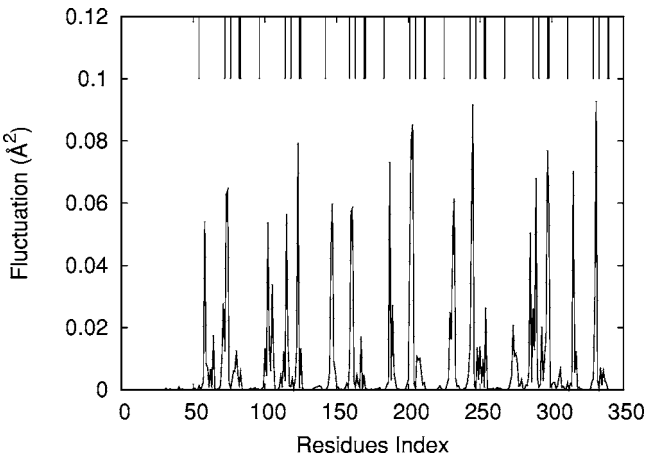

FIG. 4. Mean-square fluctuations for each residue in the fastest 20 modes in protein G_β from transducin (PDB id: 1tbg). The vertical lines at the top of the picture indicate the positions of the key residues in the seven blades.

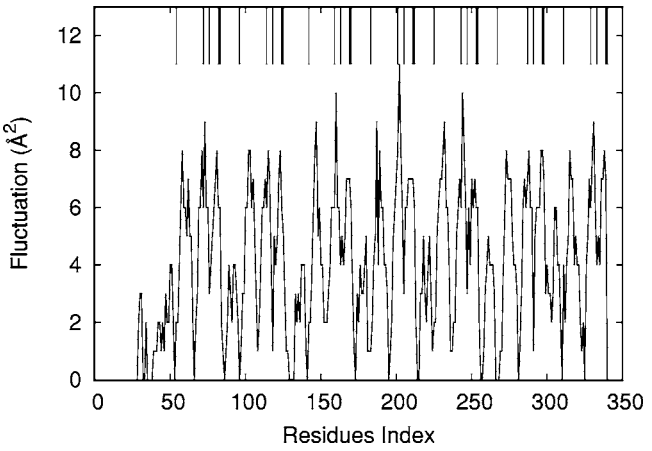

FIG. 6. Total contact number for each residue in protein G_β from transducin (PDB id: 1tbg) calculated with a distance threshold of 7.0 Å. The vertical lines in the top of the picture indicate the positions of key residues in the seven blades.

light functionally and structurally important residues with modal decomposition analysis. For the slowest (or global) modes, the motions of residues are strictly correlated to protein functions. Those residues with small square fluctuations are thought to be hinge regions between subdomains or active sites in the binding area. Moreover, the residues with large square fluctuations are thought to be recognition loop around active site. So doing slow mode analysis of GNM is a good way to determine functionally important residues. On the other hand, for the fastest modes, the fluctuation of residues is local behavior due to a detailed environment. Those residues with most rapid fluctuations are called hot residues. They are considered important to the folding of proteins. Recently, Radar and Bahar [23] applied GNM analysis to 29 proteins. They found that many residues at fast mode peaks participate in the folding nuclei [23].

In our work, we first used the GNM model to analyze the G_β protein by calculating the slowest 6 and fastest 20 modes. They are shown in Figs. 3 and 4, respectively. Figure 3 shows that the first 35 residues at the beginning of the sequence have much larger fluctuations than other residues in the main propellerlike structure because they form a long helix independently [see Fig. 1(a)]. Obviously, after the first 50 residues, there are seven low fluctuation regions, which correspond to the seven blades in the protein. This means that the residues in the blades are really restricted to their native positions. And the loops between the blades show high fluctuations. Interestingly, all low fluctuation regions have two minima that correspond to the two most stable residues, which are located at the turn of the inner two strands in each blade. The mean-square fluctuation also shows two peaks at region Blades 2 and 3 and Blades 5 and 6. They correspond to the interfaces to G_α and G_γ subunits. These would be explained in detail in the following. Overall, the slow mode minima cannot distinguish the key residues in

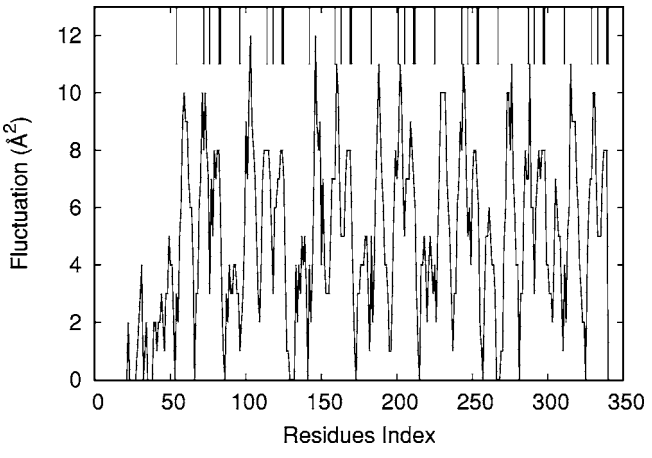

FIG. 5. Total contact number for each residue in protein G_β from transducin (PDB id: 1tbg) calculated with a distance threshold of 7.5 Å. The vertical lines in the top of the picture indicate the positions of key residues in the seven blades.

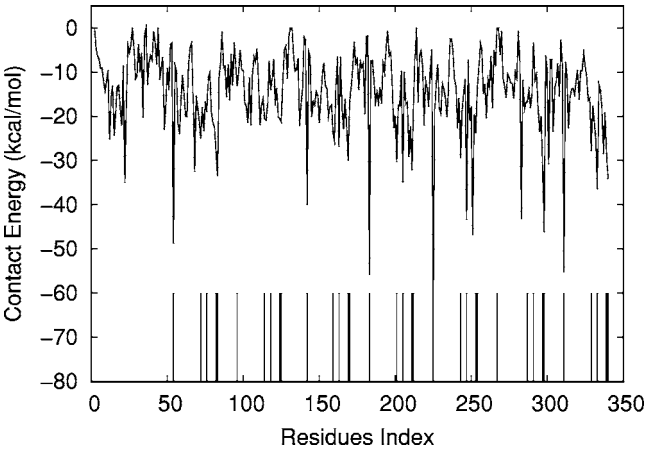

FIG. 7. Total contact energy for each residue in protein G_β from transducin (PDB id: 1tbg). The vertical lines in the bottom of the picture indicate the positions of key residues in the seven blades.

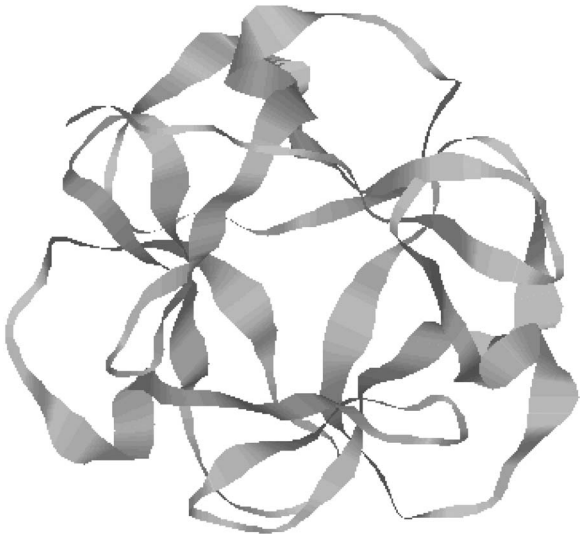

FIG. 8. Ribbon diagram for the tertiary structure of the protein hydrolase (PDB id: 1knm).

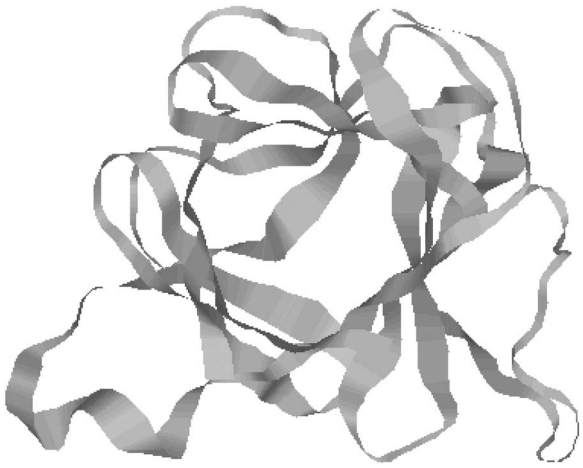

FIG. 9. Ribbon diagram for the tertiary structure of the protein cytokine (PDB id: 8i1b).

the G_β protein. For example, the key residues in the first blade are HIS54, SER72, ASP76, TRP82, and ASP83, and they have the fluctuations 0.092, 0.085, 0.086, 0.125, and 0.134 $Å^2$ respectively. These key residues maintain the 3D structure by short-range and long-range interactions with other residues. This forces some of the conserved residues to have flexibility as well as rigidity simultaneously, and so their mean-square fluctuation should be a little higher. Some conserved residues even exhibit high mobility. For example, ASP267 is a structurally conserved residue but its mean-square fluctuation can reach up to 0.51 $Å^2$, relatively higher than most of other residues. The reason may be that the GNM slow mode analysis is well known to describe the functional motion of multidomain proteins, such as the bending or rotating of subdomains along one global axis. However, the G_β protein is a single-domain protein, so it is difficult for a GNM slow mode analysis to distinguish different parts of the protein and give the functional important residues.

Now, let us turn to the GNM fast mode analysis (Fig. 4). Unlike the slow mode, Fig. 4 shows very sharp peaks. The hot residues at the peaks indicate the centers of localization of energy. These fast mode peaks have been proven to be strongly related to native state hydrogen-deuterium exchange (HX) experiments [23]. So, they are critical to protein folding. Unfortunately, the GNM fast mode analysis on this protein is not very accurate in predicting key residues, even combined with slow modes. These may be due to the reduction of GNM. GNM only treats C^α atoms and harmonic potentials in the proteins. Sometimes this is not enough to describe the residue-residue interactions. Therefore, we think adding the all-atom force field is an alternative method.

Next, we use distance-based contact theory to analyze this protein. In the past, contact theory has been widely applied to the analysis of inherent properties of the proteins, such as folding rate [27,28] and thermostability [29]. Here, we try to use it to identify the key residues. The calculated results are shown in Fig. 5. It shows that, although the contact number of residues also shows a periodic change, the residues with the largest contact number are not the structural conserved residues too. The contact number of the conserved residues varies greatly, from 4 to 10. Therefore, there is no definite relationship between the contact number and conservation of residues. We also calculated contact numbers for different distance thresholds (only show 7.0 Å in Fig. 6) and found that this does not improve the prediction power of key residues. In fact, the contact number calculated based on distance cannot describe the long-range interactions completely, such as electrostatic interactions, and so it is difficult to use it to identify the key residues in proteins.

In our opinion, it may be more reasonable to define the contact by the potential energies between residue pairs since this can reflect the intensity of the interactions between residues. So, we redefine the contact by potential energy, just as introduced in Sec. II. We analyze the potential energy be-

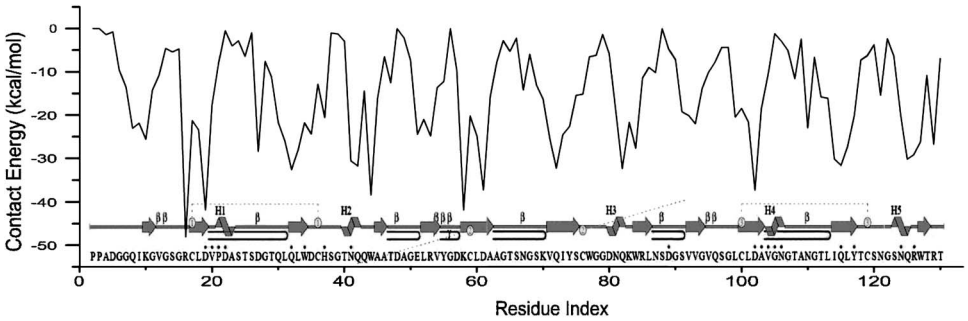

FIG. 10. Total contact energy for each residue in protein hydrolase (PDB id: 1knm).

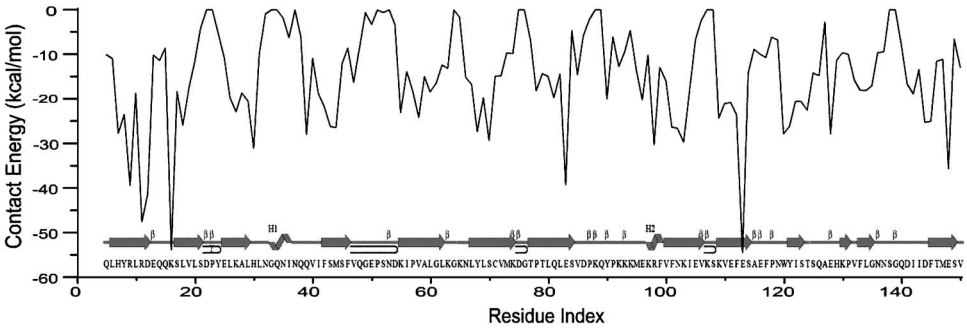

FIG. 11. Total contact energy for each residue in protein cytokine (PDB id: 8i1b).

tween all the residue pairs with an all-atom force field (CHARMM27). We assume that one contact is formed when the potential energy between the residue pair is lower than −0.5 kcal/mol. Then, the total contact energy for each residue can be calculated. We present the results in Fig. 7. It can be clearly seen that most of the structurally conserved residues have much lower contact energies than other residues, except Blade 2. The conserved residues with lowest energies are His54, His142, His183, His225, and His311 in Blades 7, 2, 3, 4, and 6, respectively. These conserved His residues are located in the loops between the blades, and each participates in a hydrogen bond network to a Ser (or Thr) in the second strand of a blade. The Ser (or Thr) in turn forms a hydrogen bond to a Trp in the third strand. This spatial arrangement of these three conserved residues preserves a network of interactions that link the neighboring blades together in proper orientation. We also find that the other two structurally conserved residues are also important. Interestingly, almost all of them are Asps and are located at similar positions, which are in the turns of the hairpins at each blade. Obviously, the role of these two conserved residues is to fix the two strands of hairpins and maintain the local structure of each blade. The analysis above indicates that all structurally conserved residues are crucial for the stability of the 3D structure of G_β. They act as joints to collect each part in the G_β protein together and allow the molecule enough rigidity but without loss of too much flexibility.

It is noted that the His residues in the loops between Blades 1 and 2 and between Blades 5 and 6 are replaced by Arg and Asp, respectively, and also do not have lower contact energies. Furthermore, the conversed residues in the inner three strands of Blade 2 also do not have lower contact energies. To understand this, it needs to be pointed out that the G_β protein is just one of the three subunits in G protein, and it combines with G_α and G_γ subunits to form a heterotrimeric $G_{\alpha\beta\gamma}$ complex. The interaction between G_α and G_β occurs at the edge of Blades 1 and 2. The interaction between G_γ and G_β occurs at the edge of the Blades 5 and 6 opposite to where G_α is bound. So, some of the residues in these blades may not form many contacts with other residues in G_β, and they may bind with residues in G_α and G_γ to help them to form stable structures. This may be the reason why the conserved residues mentioned above do not have lower contact energies.

Overall, the residues with low contact energies are much more important than other residues. They show high conservation and would exert more effects on the stability of the G_β protein. Our results show that the physical characters of the structural conserved residues may be described by their contact energies. This may help us understand the roles of the key residues in the protein structure.

Finally, we apply contact energy analysis to two other proteins, hydrolase (PDBid 1knm) (Fig. 8) and cytokine (PDBid 8i1b) (Fig. 9), and identify the key residues in them. The structures of these two proteins are all with three-fold symmetry and mostly built by β-strands. Just as above, we plot the contact energies of residues for these two proteins (Figs. 10 and 11). We find that the first six key residues in protein 1 knm are ARG16, ASP19, TRP44, LYS58, ASP61, and ASP102 (Fig. 8) and in protein 8i1b are ARG9, ARG11, ASP12, LYS16, GLU83, and GLU113 (Fig. 9). It is noted that almost all of the residues that correspond to lowest contact energies are located at the β-strands. Another interesting thing is that many residues with lower contact energies stay at the start or the end of the β-strands.

IV. CONCLUSION

In this work, we studied the physical characters of the key residues in proteins. For the G_β subunit of the G protein from transducin, we found that it is difficult to identify the key residues by using the usual Gaussian network analysis and distance-based contact analysis. They cannot give a correct description of the key residues in this protein. However, we found that the contact energies calculated by an all-atom force field could characterize the key residues very well. The residues with lowest contact energies are in good agreement with the structurally conserved residues identified previously. The results suggest that the residues with lowest contact energies may be considered as structurally important residues. This makes it possible to predict the key residues in other proteins. Our results show that contact energy analysis may provide an alternative or complementary approach to investigating the folding and stability of proteins, except the traditional GNM analysis.

ACKNOWLEDGMENTS

This work is supported by the NSFC under Grant Nos. 30525037 and 30470412 and the Foundation of the Ministry of Education of China.

[1] M. Buyong, E. Tal, W. Haim, and N. Ruth, Proc. Natl. Acad. Sci. U.S.A. **100**, 5772 (2003).
[2] N. Leibowitz, Z. Fligelman, R. Nussinov, and H. Wolfson, Proteins: Struct., Funct., Bioinf. **43**, 235 (2001).
[3] N. Leibowitz, R. Nussinov, and H. Wolfson, J. Comput. Biol. **8**, 93 (2001).
[4] W. L. DeLano, Curr. Opin. Struct. Biol. **12**, 14 (2002).
[5] E. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96 (1996).
[6] L. Mirny and E. Shakhnovich, J. Mol. Biol. **308**, 123 (2001).
[7] G. Venkat, D. B. Andreas, L. David, and A. Sidney, J. Mol. Biol. **267**, 818 (1997).
[8] H. M. Berman, J. Westbrook, Z. Feng, *et al.* Nucleic Acids Res. **28**, 235 (2000).
[9] F. Bernstein, T. Koetzel, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[10] E. J. Neer, C. J. Schmidt, R. Nambudripad, and T. F. Smith, Nature (London) **371**, 297 (1994).
[11] T. F. Smith, C. G. Gaitatzes, K. Saxena, and E. J. Neer, Trends Biochem. Sci. **24**, 181 (1999).
[12] J. Sondek, A. Bohm, D. G. Lambright, H. E. Hamm, and P. B. Sigler, Nature (London) **379**, 369 (1996).
[13] T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[14] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998).
[15] S. Kundu, J. S. Melton, D. C. Sorensen, and Jr., G. N. Phillips, Biophys. J. **83**, 723 (2002).
[16] I. Bahar and R. L. Jernigan, J. Mol. Biol. **266**, 195 (1997).
[17] M. M. Gromiha and S. Selvaraj, Prog. Biophys. Mol. Biol. **86**, 235 (2004).
[18] V. C. Still, A. Tempezvk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).
[19] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, J. Phys. Chem. A **101**, 3005 (1997).
[20] A. D. MacKerell, S. Fischer, *et al.*, J. Phys. Chem. B **102**, 3586 (1998).
[21] P. Doruker, A. R. Atilgan, and I. Bahar, Proteins: Struct., Funct., Genet. **40**, 512 (2000).
[22] C. Chennubhotla, A. J. Rader, L. Yang, and I. Bahar, Phys. Biol. **2**, 173 (2005).
[23] A. J. Rader and I. Bahar, Polymer **45**, 659 (2004).
[24] B. Isin, P. Doruker, and I. Bahar, Biophys. J. **82**, 569 (2002).
[25] Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan, J. Struct. Biol. **147**, 302 (2004).
[26] R. L. Jernigan, M. C. Demirel, and I. Bahar, Int. J. Quantum Chem. **75**, 301 (1999).
[27] M. M. Gromiha and S. Selvaraj, J. Mol. Biol. **310**, 27 (2001).
[28] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[29] M. M. Gromiha, Biophys. Chem. **91**, 71 (2001).