Effective pair potentials between protein amino acids

P. Pliego-Pastrana and M. D. Carbajal-Tinoco

Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 14-740, 07000 México D.F., Mexico

(Received 27 February 2003; revised manuscript received 29 April 2003; published 11 July 2003)

We present an effective potential describing the interaction between pairs of residues (alanines, in this case) that belong to a protein. The effective potential is extracted from an experimental correlation function, by means of the Ornstein-Zernike equation together with a closure approximation. It is found that the most relevant features of the effective potential are consistent with the formation of two different secondary structures of proteins.

DOI: 10.1103/PhysRevE.68.011903 PACS number(s): 87.15.−v

I. INTRODUCTION

Understanding the mechanisms of protein folding is of central importance in structural biology. The main problem of protein folding is the determination of the protein's native structure based on their amino acid sequence [1]. The solution to this problem is of outstanding value in molecular biophysics and biopharmacy. Furthermore, the development of new materials will be greatly facilitated. The problem of protein folding has been studied with rather different approaches. On one hand, Ising-like models allow us to enumerate exhaustively all conformations [2]. These models can be enriched with pairwise contact potentials [3]. Vendruscolo and Domany [4] demonstrated, however, that there is no set of interresidue contact parameters able to predict alone the native fold of a protein. On the other hand, Snow *et al.* [5] recently presented a successful comparison between a numerical simulation and experimental results of the protein folding dynamics of a small protein (of 23 residues). Using an atom-based force field, they required a vast computational effort. It would be thus desirable to combine the advantages of the over mentioned approaches, i.e., simplicity and accuracy.

In this paper, we propose the use of an effective potential to describe the prominent features of the protein folding process. Let us mention that, to the best of our knowledge, this relatively new method has provided a useful characterization of quite different systems. For example, the pairwise interactions among colloidal particles [6,7], as well as the quantification of the interaction between pairs of magnetic flux lines in type-II superconductors [8]. An effective pair potential is defined as the pair potential between "particles" which reproduces the pair correlation involving these particles. The effective pair potential can be mediated by particles of other species which can be experimentally observable or even nonobservable such as solvent molecules or dissolved ions [9,10]. Although the method presented here is quite general, we focus our attention to the effective potential between pairs of alanines that belong to the same protein. This effective interaction is extracted from an experimental radial distribution function through the Ornstein-Zernike (OZ) equation [11], together with an appropriate closure relation. The resulting potential function predicts the formation of some of the most important structural motifs that can be retrieved in proteins.

II. EXPERIMENTAL CORRELATION FUNCTIONS

Our study is based on the analysis of proteins of high molecular weight, each one containing more than 2000 residues. We selected structures of great extent provided that only systems with a large number of elements are expected to attain thermodynamic equilibrium (TE) (we recall that the OZ equation is valid only for systems in TE). We obtained radial distribution functions $g(r)$ from a series of 196 structurally distinct proteins of the Protein Data Bank. Our list includes hydrolases (e.g., 1jyn, 1jz0), oxidoreductases (1cw3, 1geg, 1e7p), atpases (1cow, 1e79), and groels (1aon, 1gr5), among others. Each selected protein contains at least 50 alanines inside a volume of analysis, V. This requirement is necessary to obtain the complete relevant $g(r)$. More important, we assumed that all structures are in TE. For each protein, we determined the positions of the centroids of the N alanines located inside the sphere of volume V (the position and size of the sphere are such that big voids are minimized). The corresponding number density is then $\rho = N/V$. For the whole series, we found a variation in ρ of one decade, i.e., from 1.8×10^{-4} to 9.5×10^{-4} Å^{-3}. Pair correlation functions of individual proteins were computed on the understanding that $\rho g(r) 4\pi r^2 dr$ is the number of alanines between two concentric spheres of radii r and $r+dr$, respectively, about a central alanine [11]. The spatial resolution dr was estimated to be 0.2 Å, considering the uncertainties in centroids' coordinates.

Although it is possible to provide accurate approximations to get the effective pair potential $u^{eff}(r)$, the supplied correlation function has to be determined with enough precision to minimize errors induced by statistical noise. Thus, to improve the statistics, we averaged the results from proteins of rather close number densities. In Fig. 1, we show two functions $g(r)$ of number densities $\rho_h = (5.6\pm0.4)\times10^{-4}$ and $\rho_l = (3.5\pm0.2)\times10^{-4}$ Å^{-3}, which were obtained from the average of 42 and 49 different pair correlations, respectively. The following features can be noticed in the same figure. First, both pair correlations present a series of three well-defined peaks of decreasing height, as a function of the distance. Beyond the third peak, there is no clearly defined shape until the asymptotic value is reached. The peaks of both functions have some resemblances and some differences too. Quite remarkable, the position and shape of the corresponding peaks are very similar. On the other hand, the

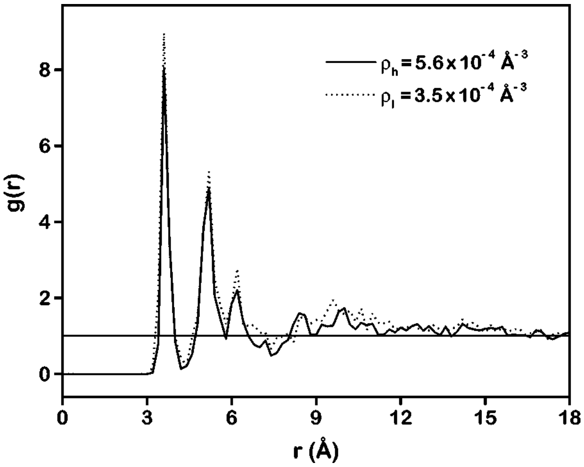

FIG. 1. Radial distribution functions between alanine centroids. The solid line corresponds to the average of a series of 42 proteins with mean number density $\rho_h = (5.6 \pm 0.4) \times 10^{-4}$ Å^{-3}, while the dotted line represents the average of 49 correlation functions with $\rho_l = (3.5 \pm 0.2) \times 10^{-4}$ Å^{-3}.

corresponding heights are not the same. Such a discrepancy could be related to a many-body effect, but this is not the case. The most dilute $g(r)$ (of number density ρ_l) has the highest peaks, contrary to the regular behavior of the known correlation functions [11]. Moreover, the mean distance between centroids of neighbor alanines, $d \equiv \rho^{-1/3}$, indicates that the correlations between pairs of alanines have the behavior of a dilute system (for all the proteins in study, $d > 10$ Å, which is a distance well beyond the third peak).

The above mentioned differences in heights could be explained in terms of statistical fluctuations. More important, if these pair correlation functions show the behavior of a dilute system, then it is possible to average the 196 functions, to a good approximation (let us recall that the whole protein is a compact object, but it has the additional contribution of the 19 remaining amino acids, so from the point of view of the alanines this is a rather dilute system). The resulting average is denoted as $g_E(r)$, and it is characterized by a mean number density $\overline{\rho} = (4.3 \pm 1.4) \times 10^{-4}$ Å^{-3}. For convenience, our results are expressed in terms of the total correlation function $h_E(r) = g_E(r) - 1$, which is also one of the functions involved in the OZ equation. As expected, the experimental $h_E(r)$ plotted in Fig. 2 preserves the main characteristics of the functions of Fig. 1, with an appreciable decrement of statistical fluctuations. Otherwise, the correlation function $h_E(r)$ could be especially useful to test and eventually improve some of the existing atomic force field models [12]. This task, however, requires extensive numerical simulations.

III. THEORETICAL BACKGROUND

Let us now provide some insights of the physical properties of the correlation function $h_E(r)$. For instance, and according to the sequence of atoms, it is possible to verify that the first peak is related to the probability of finding two centroids of alanines, which are consecutive in the amino acid sequence. In fact, $a = 3.6 \pm 0.1$ Å is the position of the first maximum, in other words, a is the most likely distance between two consecutive alanines. Because a protein is a polymeric chain, it is interesting to compare $h_E(r)$ with the correlation function describing an ideal-polymer chain [13],

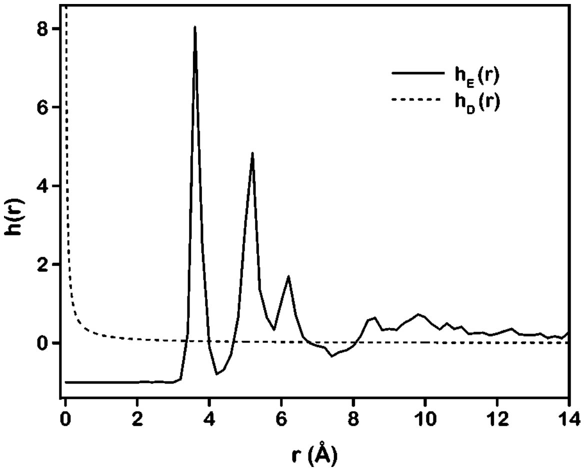

FIG. 2. Total correlation function $h_E(r)$ (solid line) obtained from the average of 196 correlation functions and characterized by $\overline{\rho} = (4.3 \pm 1.4) \times 10^{-4}$ Å^{-3}. The corresponding Debye-structure function $h_D(r)$ (see text) is also plotted (dashed line).

$$\tilde{h}_D(q) = 2[\exp(-v) + v - 1]/v^2, \tag{1}$$

where $\tilde{h}_D(q)$ is the Fourier-transformed Debye-structure function, $v \equiv R_0^2 q^2/6$, R_0 is the size of the polymer, and q is the wave vector magnitude. The parameter found in Eq. (1) is chosen to be in accord with the identified variables of $h_E(r)$, as explained in the following lines. If we suppose that the ideal chain consists of N_0 noninteracting segments of length a, then [13] $R_0 = a N_0^{1/2}$. The volume occupied by the ideal chain is approximately R_0^3, thus $\overline{\rho} \simeq N_0 / R_0^3 \simeq a^{-3} N_0^{-1/2}$. These two conditions permit the estimation of R_0, which is used to compute the corresponding Debye-structure function of Eq. (1). In the real space, the function $h_D(r)$ is also plotted in Fig. 2. As shown in this figure, $h_D(r)$ converges to zero much faster than $h_E(r)$. The comparison of both correlation functions also confirms that a structural property of folded proteins clearly differs from the ideal chain behavior [1].

The effective interaction $u^{eff}(r)$ can be understood as the pair potential that dresses an initially noninteracting chain, giving as a result a structure identical to the experimental one. From this starting point, and neglecting interprotein correlations, we suppose that the microstructure of $h_E(r)$ is determined by the OZ equation [11],

$$h_E(r) = c^{eff}(r) + \overline{\rho} \int d^3\mathbf{r}' c^{eff}(r') h_E(|\mathbf{r} - \mathbf{r}'|), \tag{2}$$

where $c^{eff}(r)$ is the effective direct correlation function thus defined. The Fourier space version of Eq. (2), namely,

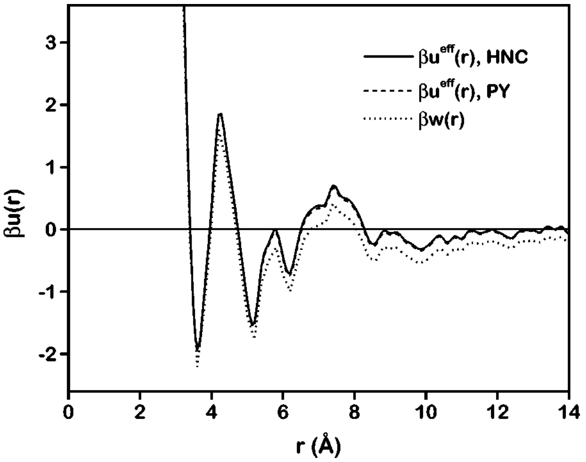

FIG. 3. Effective potential between pairs of alanines $\beta u^{eff}(r)$ obtained through the HNC (solid line) and the PY (dashed line) approximations. The potential of mean force $\beta w(r)$ is also displayed (dotted line).

$$\tilde{c}^{eff}(q)=\frac{\tilde{h}_E(q)}{1+\bar{\rho}\tilde{h}_E(q)}, \tag{3}$$

allows the determination of $\tilde{c}^{eff}(q)$, which is transformed back to get $c^{eff}(r)$. In order to complete our scheme, it is necessary to supply an additional condition. A general expression relating $u^{eff}(r)$, $h_E(r)$, and $c^{eff}(r)$ is [11]

$$\beta u^{eff}(r)=h_E(r)-c^{eff}(r)-\ln[h_E(r)+1]+b^{eff}(r), \tag{4}$$

with $b^{eff}(r)$ being the effective bridge function, $\beta=1/k_BT$, where k_B is Boltzmann's constant, and T is the absolute temperature. In general, bridge functions are very complex and therefore some approximations have to be done at this level. Taking $b^{eff}(r)\equiv 0$ leads to the hypernetted chain (HNC) closure [11], while the Percus-Yevick (PY) expression [11]

$$\beta u^{eff}(r)=\ln\{1-c^{eff}(r)/[h_E(r)+1]\}, \tag{5}$$

is another useful approximation. In Fig. 3, we exhibit the effective potential between pairs of alanines obtained through both, the HNC and the PY closure relations. It can be noticed that both approximations conduct to basically identical results. Therefore, a unique effective pair potential will be considered from now on. As a reference, we also plot the potential of mean force [11], defined by

$$\beta w(r)\equiv -\ln[h_E(r)+1], \tag{6}$$

which is identical to $\beta u^{eff}(r)$ only in the limit $\bar{\rho}=0$. As it can be observed in Fig. 3, $\beta u^{eff}(r)$ and $\beta w(r)$ are qualitatively very similar with only a noticeable difference: the interaction $\beta w(r)$ has always lower values than $\beta u^{eff}(r)$, as a consequence, $\beta w(r)$ tends to zero more slowly than $\beta u^{eff}(r)$. Such a difference could be related to an accentuated sensitivity of the interaction potential to variations of $\bar{\rho}$, in comparison with the reported behavior of the measured correlation functions. In contrast to the oversimplified $\beta w(r)$, our calculation of $\beta u^{eff}(r)$ takes into account any possible dependence on $\bar{\rho}$, at least approximately.

IV. DISCUSSION

As shown in Fig. 3, $\beta u^{eff}(r)$ has a rather unusual form. It consists of three sharply defined potential wells and two barriers of distinct shapes, all of them have magnitude of the order of k_BT. This specific combination of barriers and wells, however, predicts the existence of two of the most important structures found in proteins. Although these structures can be reconstructed with a simplified model, systematic numerical simulations will be reported elsewhere [14]. Our model of polyalanine is based on the geometrical relations between the characteristic lengths extracted from the curve of $\beta u^{eff}(r)$. These characteristic lengths are the positions of the potential energy minima which, in principle, should lead to a global minimum. We recall that the position of the first minimum of $\beta u^{eff}(r)$ is directly related to the preferred distance between two consecutive alanines, a [of course, the maxima of $h_E(r)$ correspond to the minima of $\beta u^{eff}(r)$ or $\beta w(r)$]. The backbone of our polyalanine is defined as an initially freely jointed chain consisting of N_a links (a link connects two alanine centroids), every one of length a. Each one of the two remaining minima of $\beta u^{eff}(r)$ provides an extra condition that determines the specific structural motif, assuming that the positions of these two minima fix the distance b_χ between the alanines i and $i+2$ of the sequence.

In case *A*, we select the minimum of the second well, i.e., $b_A=5.2\pm 0.1$ Å. Thus, according to distances a and b_A, the two pairs of alanines $(i,i+1)$ and $(i+1,i+2)$ make an angle $\theta_A=(92.5\pm 6.0)°$. At this point, the interplay of the barriers and the second well has a very important role in the arrangement of the following alanines. An eventual trans conformation of the alanine $i+3$ would be frustrated by the second barrier, because of the high potential energy between alanines i and $i+3$. On the other hand, the alanine $i+3$ can be favorably found in a cis conformation at a distance identical to b_A of the alanine i, as a consequence of the presence of the first barrier and the second well. For the same reason, it is quite likely to find the alanine $i+4$ at the same distance b_A of the alanine i. The arrangement of amino acids just described is consistent with the 4-turn pattern defined by Kabsch and Sander [15]. Of course, at least two 4-turns are necessary to form a minimal helix [15]. Thus, defining an α helix with our parameters requires 4.1 ± 0.3 residues per turn with a pitch $\simeq b_A$.

For case *B*, the distance between alanines i and $i+2$ is now equal to the minimum of the third well, i.e., $b_B=6.2\pm 0.1$ Å. The angle between the two pairs of residues $(i,i+1)$ and $(i+1,i+2)$ is $\theta_B=(118.9\pm 9.0)°$. Contrary to case *A*, in this case the cis conformation of the residue $i+3$ is restrained by the second barrier. The residue $i+3$, however, can be found in a trans conformation of lower energy, because the distance between alanines i and $i+3$ corresponds to a distance beyond the second barrier. Therefore, the alanine $i+4$ is also likely to appear in a trans conformation with respect to its previous neighbors. Such an open arrangement of amino acids is the characteristic of β structures [15], even if the pattern just described can be identified only with a β strand [15].

V. CONCLUSIONS

In conclusion, our simplified model indicates the appearance of two characteristic lengths in the formation of α and β secondary structures. The characteristic lengths, as well as the guides to allow or forbid certain configurations, emerge from the effective pair potential that we extracted from an experimental pair correlation function. It is important to note that these characteristic lengths seem to be independent of the theoretical approach (we examined the OZ equation with two closure relations and the potential of mean force). Moreover, despite small quantitative differences, the salient features of the full interaction curves are preserved within the studied approximations. Let us finally stress the importance of the experimental correlation function which was determined without an explicit identification of the secondary structure of the alanines that belong to the proteins under analysis.

ACKNOWLEDGMENTS

The authors want to thank J. Mas-Oliva and P. González-Mozuelos for enlightening conversations. This work was supported by CONACyT under Grant No. 36464-E.

[1] *Protein Folding*, edited by T.E. Creighton (Freeman, New York, 1992).

[2] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[3] S. Miyazawa and R.L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[4] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998).

[5] C.D. Snow, H. Nguyen, V.S. Pande, and M. Gruebele, Nature (London) **420**, 102 (2002).

[6] M.D. Carbajal-Tinoco, F. Castro-Román, and J.L. Arauz-Lara, Phys. Rev. E **53**, 3745 (1996).

[7] S.H. Behrens and D.G. Grier, Phys. Rev. E **64**, 050401 (2001).

[8] C.-H. Sow, K. Harada, A. Tonomura, G. Crabtree, and D.G. Grier, Phys. Rev. Lett. **80**, 2693 (1998).

[9] P. González-Mozuelos and M.D. Carbajal-Tinoco, J. Chem. Phys. **109**, 11074 (1998); M.D. Carbajal-Tinoco and P. González-Mozuelos, *ibid.* **117**, 2344 (2002).

[10] N. Bagatella-Flores and P. González-Mozuelos, J. Chem. Phys. **117**, 6133 (2002).

[11] J.-P. Hansen and I.R. McDonald, *Theory of Simple Liquids*, 2nd. ed. (Academic Press, London, 1986).

[12] S.J. Weiner *et al.*, J. Am. Chem. Soc. **106**, 765 (1984); L. Nillson and M. Karplus, J. Comput. Chem. **4**, 187 (1986); W.L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).

[13] G. Strobl, *The Physics of Polymers*, 2nd ed. (Springer, Berlin, 1997).

[14] P. Pliego-Pastrana and M.D. Carbajal-Tinoco (unpublished).

[15] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).