Electronic structures of *Ascaris* trypsin inhibitor in solution

Haoping Zheng*
Pohl Institute of Solid State Physics, Tongji University, Shanghai 200092, China
(Received 11 July 2003; published 25 November 2003)

The electronic structures of *Ascaris* trypsin inhibitor in solution are obtained by the first-principles, all-electron, *ab initio* calculation using the self-consistent cluster-embedding (SCCE) method. The inhibitor, made up of 62 amino acid residues with 912 atoms, has two three-dimensional solution structures: 1ata and 1atb. The calculated ground-state energy of structure 1atb is lower than that of structure 1ata by 6.12 eV. The active sites are determined and explained: only structure 1atb has a N terminal at residue ARG+31. This shows that the structure 1atb is the stable and active form of the inhibitor, which is in agreement with the experimental results. The calculation reveals that some parts of the inhibitor can be easily changed while the inhibitor's biological activity may be kept. This kind of information may be helpful in fighting viruses such as AIDS, SARS, and flu, since these viruses have higher variability. The calculation offers an independent theoretical estimate of the precision of structure determination.

DOI: 10.1103/PhysRevE.68.051908 PACS number(s): 87.15.Aa, 87.14.Ee, 87.15.By

I. INTRODUCTION

Ascaris lives in the hostile environment of the gut, and has developed specific mechanisms to protect itself from the host digestive enzymes. The protease inhibitors seem not to be secreted, but rather are bound on the surface of the worm, gut and other tissues, as well as on the surface of eggs and developing larvae, where they form complexes with host proteases. The presence of inactivated proteases on the surface of the eggs and larvae ensure that the migrating larvae are not perceived as foreign, thereby permitting them to evade the host's immune system as they migrate from the intestines to the liver. In addition, the *Ascaris* protease inhibitors inhibit both clot lysis by plasmin and streptokinase activated fibrinolysis of human plasma clots, suggesting that the presence of inhibitors on the larval surface clots may modify blood homeostasis during larval migration [1]. *Ascaris* trypsin inhibitor contains 912 atoms and has two three-dimensional solution structures with the atomic coordinates deposited in Protein Data Bank (PDB): one at *p*H 4.75 and another at *p*H 2.4. The electronic structures of the inhibitor, which is currently unknown, should help us in explaining the interesting properties above.

The computational study of protein molecules has until now focused mainly on their structural properties, say, protein folding calculation, while the nature of their electronic structures has received far less attention because of the difficulty in electronic structure calculation. Yet the knowledge of electronic structure is essential for understanding the properties and biological functions of a protein molecule. Furthermore, according to computational condensed matter physics and quantum chemistry, some properties may only be studied by electronic structure calculation due to limited experimental condition. In physics, the knowledge of electronic structures of crystals, impurities, surfaces, and interfaces has helped us greatly in developing many electronic devices. In chemistry, the knowledge of electronic structures of molecules enables much more effective use of chemical reactions. So the new proteomics should be promoted greatly, if we know the electronic structures of proteins, and it may give us valuable clue in the research of diagnosis and treatment, as well as in the development of new medicines.

The calculation of electronic structure of a protein, however, is almost impossible if we use traditional free-cluster method based on density functional theory, because of its incredibly huge computational effort, which is too large to be affordable by any supercomputer in the present and near future. In the past several years, there has been a great deal of interest in developing so-called $O(M)$ methods, for which the computational effort scales linearly with number of atoms M [2–11]. But to the best of my knowledge, there has been no successful first-principles, all-electron, *ab initio* calculation of any protein molecule before 1999.

The self-consistent cluster-embedding (SCCE) calculation method is developed by the author based on density functional theory [12,13]. It has been successfully applied to crystals NiO, CoO, Ni, GaN, $LaNi_5$, $LaNi_5H_7$, hydrogen-decorated vacancies in Ni, and the Ga vacancy in GaN [12,14–17]. The computational effort of a SCCE calculation can scale quasilinearly with the number of atoms, while the calculation precision is kept. This makes the electronic structure calculation of protein molecule a reality. For the first time, a successful first-principles, all-electron, *ab initio* calculation of electronic structure of a real protein molecule, trypsin inhibitor from squash seeds in aqueous solution (CMTI-I, 436 atoms), was completed in 1999 [18,19]. The active site of the inhibitor is determined and explained, and the precision of structure determination of the inhibitor is tested theoretically.

In this paper, we present the calculated electronic structures of a larger protein molecule, *Ascaris* trypsin inhibitor in solution, which contains 912 atoms and has two three-dimensional structures. Section II gives the theoretical model and computational procedure. The results and discussion are given in Sec. III.

II. THEORETICAL MODEL AND COMPUTATIONAL PROCEDURE

The SCCE method has been explained in detail in Refs. [13,16,18]. Here we give only a brief introduction. Accord-

*Email address: zhenghp@mail.tongji.edu.cn

ing to the density functional theory [20,21], the total energy of a system containing N electrons and M fixed nuclei can be written as

$$E_v[\rho]=T_{ni}[\rho]+E_{xc}[\rho]+\int\int\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' -2\sum_{j=1}^{M}\int\frac{\rho(\mathbf{r})Z_j}{|\mathbf{r}-\mathbf{R}_j|}d\mathbf{r}+\sum_{i\neq j}^{M}\frac{Z_iZ_j}{|\mathbf{R}_i-\mathbf{R}_j|}. \quad (1)$$

Atomic units, with the unit of energy being the Rydberg constant $e^2/2a_o$ (13.6049 eV), are used throughout this paper: $e^2=2$, $\hbar=1$, and $2m_e=1$. In deriving Eq. (1), Kohn and Sham have assumed the existence of a noninteracting electron system having the same ground-state charge density $\rho(\mathbf{r})$ as that of a real interacting system [21]. Each noninteracting electron is represented by a stationary state one-electron wave function $\phi_n^\sigma(\mathbf{r})$. So $\rho(\mathbf{r})$ and kinetic energy $T_{ni}[\rho]$ of the noninteracting system are

$$\rho(\mathbf{r})=\rho^\uparrow(\mathbf{r})+\rho^\downarrow(\mathbf{r})=\sum_{\text{occupied } l}|\phi_l^\uparrow(\mathbf{r})|^2+\sum_{\text{occupied } m}|\phi_m^\downarrow(\mathbf{r})|^2, \quad (2)$$

$$T_{ni}[\rho]=\sum_{\text{occupied } l}\int\phi_l^{\uparrow *}(\mathbf{r})(-\nabla^2)\phi_l^\uparrow(\mathbf{r})d\mathbf{r} +\sum_{\text{occupied } m}\int\phi_m^{\downarrow *}(\mathbf{r})(-\nabla^2)\phi_m^\downarrow(\mathbf{r})d\mathbf{r}. \quad (3)$$

In local spin density approximation, the exchange-correlation (xc) energy can be written as

$$E_{xc}[\rho]=\int\rho(\mathbf{r})\epsilon_{xc}(\rho^\uparrow(\mathbf{r}),\rho^\downarrow(\mathbf{r}))d\mathbf{r}. \quad (4)$$

Using formulas (2)–(4), the single-electron Schrödinger equation, i.e., the well known Kohn-Sham equation, is obtained by the variation of functional (1) with respect to $\phi_n^{\sigma *}(\mathbf{r})$ under conservation rule $\int\rho(\mathbf{r})d\mathbf{r}=N$ [21]:

$$\left\{-\nabla^2+2\int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'-2\sum_{i=1}^{M}\frac{Z_i}{|\mathbf{r}-\mathbf{R}_i|}+V_{xc}^\sigma(\mathbf{r})\right\}\phi_n^\sigma(\mathbf{r}) =\epsilon_n^\sigma\phi_n^\sigma(\mathbf{r}). \quad (5)$$

Assume $E_{xc}[\rho]$ to be exact, so functional (1) is exact. However, the variation of functional (1) is exact only if the trial one-electron wave functions $\phi_n^\sigma(\mathbf{r})$ are unconstrained. But when we solve Eq. (5), $\phi_n^\sigma(\mathbf{r})$ are constrained which implies an approximation: each $\phi_n^\sigma(\mathbf{r})$ is constrained to satisfy a certain boundary condition and is expanded into a set of finite number of bases. We can use the following two kinds of constrained trial one-electron wave functions $\phi_n^\sigma(\mathbf{r})$ (i.e., noninteracting electrons), which satisfy different boundary conditions, to describe the real system approximately.

(1) Each one-electron wave function $\phi_n^\sigma(\mathbf{r})$ is constrained to spread over the whole region occupied by the system. Under this constraint, Eq. (5) can be used for free-cluster calculation with the natural finite boundary condition $\phi_n^\sigma(\mathbf{r})\overset{|\mathbf{r}|\to\infty}{\to}0$ or for band structure calculation with periodic boundary condition.

(2) Each one-electron wave function $\phi_n^\sigma(\mathbf{r})$ is constrained to be distributed in the part of the region occupied by the system. Under this constraint, Eq. (5) is used for SCCE calculation [13]: The N localized $\phi_n^\sigma(\mathbf{r})$ are divided into k groups. $\phi_n^\sigma(\mathbf{r})$ in the ith group localize in and around the ith embedded cluster, and satisfy the ith set of special boundary conditions:

$$\phi_n^\sigma(\mathbf{r})=0 \quad \text{for } r \text{ in the core regions of surrounding atoms,} \quad (6)$$

$$\phi_n^\sigma(\mathbf{r})\to 0 \quad \text{for } r \text{ leaving the } i\text{th embedded cluster.} \quad (7)$$

For each embedded cluster, its electronic density is represented by $\rho_1(\mathbf{r})$, the rest of the system is treated as an environment with electronic density $\rho_2(\mathbf{r})$ which has a small overlap with $\rho_1(\mathbf{r})$. Because all N $\phi_n^\sigma(\mathbf{r})$ are localized, formulas (2) and (3) can be rewritten as ($N=N_1+N_2$)

$$\rho(\mathbf{r})=\sum_{\text{occupied } n\ \sigma}^{N}|\phi_n^\sigma(\mathbf{r})|^2 =\sum_{\text{occupied } n_1\ \sigma}^{N_1}|\phi_{n_1}^\sigma(\mathbf{r})|^2+\sum_{\text{occupied } n_2\ \sigma}^{N_2}|\phi_{n_2}^\sigma(\mathbf{r})|^2 \equiv\rho_1(\mathbf{r})+\rho_2(\mathbf{r}), \quad (2')$$

$$T_{ni}[\rho]=T_{ni}[\rho_1+\rho_2] =\sum_{\text{occupied } n\ \sigma}^{N}\int\phi_n^{\sigma *}(\mathbf{r})(-\nabla^2)\phi_n^\sigma(\mathbf{r})d\mathbf{r} =\sum_{\text{occupied } n_1\ \sigma}^{N_1}\int\phi_{n_1}^{\sigma *}(\mathbf{r})(-\nabla^2)\phi_{n_1}^\sigma(\mathbf{r})d\mathbf{r} +\sum_{\text{occupied } n_2\ \sigma}^{N_2}\int\phi_{n_2}^{\sigma *}(\mathbf{r})(-\nabla^2)\phi_{n_2}^\sigma(\mathbf{r})d\mathbf{r} \equiv T_{ni}[\rho_1]+T_{ni}[\rho_2]. \quad (3')$$

A protein molecule is divided into k embedded clusters. The calculation contains the following two kinds of iterations.

(i) Intracluster iteration: For each embedded cluster, Eq. (5) is calculated self-consistently; $\rho_1(\mathbf{r})$ of the embedded cluster is self-consistently changed during the iterations, while the rest of the system is served as fixed environment $\rho_2(\mathbf{r})$.

(ii) Intercluster iteration: The k embedded clusters are synchronously calculated by k CPUs. After the convergence of intracluster iterations of all k embedded clusters, the results are used for constructing new environments $\rho_2(\mathbf{r})$ for each embedded cluster, and a new intercluster iteration begins. We refer readers to the paper Refs. [13,16,18] for further details concerning the SCCE calculation.

The *Ascaris* trypsin inhibitor is a member of a family of serine protease inhibitors isolated from the helminthic worm *Ascarislumbricoides* var. *suum*. It is made up of 62 amino

acid residues and contains 912 atoms [22]. We choose the results of nuclear magnetic resonance experiments as our starting point and take the atomic coordinates of 912 atoms from Protein Data Bank: PDB 1ata (the solution structure at *p*H 4.75) and PDB 1atb (the solution structure at *p*H 2.4), both are minimized mean structures [23].

The experiments show that an inhibitor has usually several active sites where the molecule can be attacked more easily than other parts. From the viewpoint of physics, this can be understood as follows: (a) A valence electron in an inhibitor is localized, it cannot be shared by all atoms in the inhibitor. (b) An active site has two possible ways to interact: there is the highest occupied local valence electron which can easily be lost, or there is the lowest unoccupied local state which has a tendency to be occupied by an additional electron. In addition, there is no report of fractional charged amino acid residue. So in our SCCE calculation, the basic assumption is that each embedded cluster, neutral or charged, has an integer number of electrons. This guarantees localized electrons, and so the localized active sites.

We divide the *Ascaris* trypsin inhibitor into 47 embedded clusters (see columns 1 and 2 of Tables I and II). Each amino acid residue is initially treated as electrically neutral, thus the top occupied local states of 47 embedded clusters have different eigenvalues. After the first convergence, the electron transfer is made according to the eigenvalues as well as the chemical considerations: an electron is moved from amino acid residue *i* (if *i* has the highest top occupied local state) to amino acid residue *j* (if *j* has the lowest unoccupied local state), while keeping the whole molecule electrically neutral. Such an electron transfer may change the whole set of eigenvalues dramatically. If we find that the electron transfer leads to much lower unoccupied local state in residue *i* and much higher top occupied local state in residue *j*, the electron transfer is canceled. From the viewpoint of chemistry, only NH_3^+ (or NH_2^+) in residues ARG, LYS, HIS, and N terminal, and COO^- in residues GLU, ASP, and C terminal need to be tested. After several tests, it is found that the charged residues GLU^-1, ASP^-49, GLU^-58, ASP^-59, LYS^+7, LYS^+25, LYS^+34, LYS^+56, and LYS^+62 are impossible. The reasonable configuration is as follows: (2) ALA2 and GLU^-3, (3) LYS^+4, (8) GLU^-10, (12) LYS^+14, (15) GLU^-19, (24) ARG^+31, (25) GLU^-32 and CYS33, (29) ARG^+37, (30) CYS38 and GLU^-39, and (37) ARG^+48. All other 37 embedded clusters are electrically neutral (see the second columns of Tables I and II). Of course, in order to keep every embedded cluster that has an integer number of electron, a localized highest occupied molecular orbital of one cluster may be higher than a localized lowest unoccupied molecular orbital of another cluster. But as long as every real electron in the inhibitor localizes around one atom or two neighboring atoms, our results should be a good approximation to the real case, and it is reasonable for the calculated system to be in equilibrium because it has the lowest total energy under the assumption of localizability.

The numbers of optimized Gaussian bases of atoms H, C, N, O, and S are 11, 26, 29, 29 and 41, respectively, which are the same as that we used in the calculation of trypsin inhibitor CMTI-I [18]. The number of electrons and the number of Gaussian bases of each embedded cluster, which are same for two structures, are given in the third and fourth columns of both tables, respectively. We use the von Barth and Hedin form of the exchange-correlation potential [24], as parametrized by Rajagopal and co-workers [25]. The mean coordinates of heavy (nonhydrogen) atoms in each embedded cluster are calculated. The grid points, filling a sphere with radius 33.0 a.u. centered at mean coordinates of each embedded cluster, are taken for the calculation of V_{xc}. The number of grid points of a single embedded cluster is about 110 000 0–170 000 0. The optimum values of the core radii of five elementary atoms, used for surrounding atoms in the SCCE calculations, are determined: $R_C=0.7472$ a.u., $R_N=0.6955$ a.u., $R_O=0.6955$ a.u., $R_H=0.6972$ a.u., and $R_S=1.2554$ a.u. For each structure, the well converged results are obtained after 13 intercluster iterations. The electron number of any embedded cluster, remaining in the surrounding atomic core regions, is less than $0.0025e$, which shows that the special finite boundary condition (6) is satisfied with high precision, so the SCCE calculations are valid.

III. RESULTS AND DISCUSSION

The calculation gives the following information.

A. Electronic structures

The calculations give the charge densities as well as the eigenstates of both structures. We put the eigenvalues of bottom unoccupied states and top occupied states of each embedded cluster into fifth and sixth columns of both tables, respectively. The information following each eigenvalue is the result of Mulliken population analysis. For example, C:sp, in the fifth column of cluster 5 (LYS7) of structure 1ata, means that the largest part of the orbital comes from carbon *s* and *p* valence electrons. Table I shows that the embedded clusters 19, 5, 42, 8, 25, and 33 have the highest (top) occupied local states which should be the easiest positions for losing an electron; the embedded clusters 38, 45, 44, 37, 12, 39, and 47 have the lowest (bottom) unoccupied local states which should be the easiest positions for receiving an electron. In Table II, the embedded clusters 19, 5, 8, 26, 42, 13, and 33 have the highest (top) occupied local states; the embedded clusters 38, 47, 45, 12, 46, 44, and 24 have the lowest (bottom) unoccupied local states.

The energy of each embedded cluster is put in the seventh column. Their sum gives the ground-state energy $-52\,463.4459$ Ry of the structure 1ata and $-52\,463.8960$ Ry of the structure 1atb. The energy difference of 0.4501 Ry $=6.12$ eV shows that structure 1atb is more stable than structure 1ata.

B. The active sites of protein

In order to discuss the active sites, the mean coordinates of all heavy (nonhydrogen) atoms are calculated.

For structure 1ata, they are $\overline{X_a}=0.0840$ Å, $\overline{Y_a}=0.1302$ Å, $\overline{Z_a}=-0.2680$ Å. The eighth column of Table I shows the farthest heavy atoms of 62 residues, and their distances from the mean coordinates ($\overline{X_a}$, $\overline{Y_a}$, $\overline{Z_a}$). Among

TABLE I. Information of 47 embedded-clusters of *Ascaris* trypsin inhibitor (structure 1ata).

No.	Amino acid residues	No. of electrons	No. of bases	Bottom unoccupied states (Ry)	Top occupied states (Ry)	Energy (Ry)	Farthest non-H atom distance (a.u.)
1	GLU1	69	334	−0.2436 C:*p*	−0.2436 C:*p*	−947.8636	O 22.03
2	ALA2	106	503	−0.2583 C:*p*	−0.2612 O:*p*	−1624.2904	N [a] 16.85
	GLU^-3						O 19.87
3	LYS^+4	70	386	−0.2560 C:*ps*	−0.2631 C:*p*	−53.9213	N 18.17
4	CYS5	107	489	−0.2336 C:*p*	−0.2336 C:*p*	−3095.6169	O [a] 13.77
	THR6						C 15.93
5	LYS7	71	386	−0.1593 C:*sp*	−0.1593 C:*sp*	−163.5554	N 14.88
6	PRO8	52	265	−0.2518 C:*p*	−0.2855 C:*ps*	−425.8547	C 12.16
7	ASN9	60	286	−0.2541 C:*p*	−0.2764 C:*ps*	−959.6406	N 8.84
8	GLU^-10	68	312	−0.1315 C:*p*	−0.1686 C:*sp*	−1271.7942	O 9.03
9	GLN11	68	334	−0.2327 C:*ps*	−0.2528 C:*ps*	−777.8550	O 10.63
10	TRP12	98	483	−0.2440 C:*p*	−0.2537 C:*ps*	−938.6123	C 13.94
11	THR13	54	268	−0.2109 C:*p*	−0.2438 C:*ps*	−523.6224	O [a] 10.19
12	LYS^+14	70	386	−0.2981 H:*s*	−0.3067 C:*sp*	−29.5117	N 17.13
13	CYS15	83	364	−0.2059 C:*p*	−0.2059 C:*p*	−2777.3060	C 10.94
	GLY16						N [a] 9.00
14	GLY17	83	364	−0.2279 N:*ps*	−0.2279 N:*ps*	−2813.5476	C [a] 8.77
	CYS18						S 10.61
15	GLU^-19	68	312	−0.2108 C:*p*	−0.2300 O:*p*	−1360.5929	O [a] 7.53
16	GLY20	84	411	−0.2239 C:*ps*	−0.2247 C:*p*	−1082.3513	C [a] 7.31
	THR21						O 10.94
17	CYS22	91	412	−0.2732 C:*p*	−0.2732 C:*p*	−2678.0147	O [a] 13.89
	ALA23						C 14.74
18	GLN24	68	334	−0.2047 C:*p*	−0.2923 C:*p*	−847.7804	O 15.41
19	LYS25	71	386	−0.1432 H:*s*	−0.1432 H:*s*	−79.5601	N [a] 11.23
20	ILE26	62	335	−0.1902 C:*p*	−0.2608 C:*p*	−53.0884	O [a] 12.29
21	VAL27	54	287	−0.1869 C:*p*	−0.2382 C:*ps*	−80.9722	C [a] 12.09
22	PRO28	105	486	−0.2852 C:*s*	−0.2852 C:*s*	−2719.2756	C 15.70
	CYS29						C [a] 15.21
23	THR30	54	268	−0.2411 C:*p*	−0.2615 C:*s*	−619.8669	C 18.94
24	ARG^+31	84	444	−0.2553 N:*p*	−0.2772 N:*p*	−346.4991	N [a] 15.85
25	GLU^-32	121	533	−0.1886 C:*p*	−0.1886 C:*p*	−3412.7183	O 19.70
	CYS33						O [a] 13.74
26	LYS34	71	386	−0.2146 N:*s*	−0.2146 N:*s*	−30.4937	C [a] 12.44
27	PRO35	52	265	−0.2648 C:*p*	−0.2704 C:*p*	−443.1929	C 14.95
28	PRO36	52	265	−0.2653 C:*p*	−0.3202 C:*p*	−396.1256	C 15.19
29	ARG^+37	84	444	−0.2489 N:*p*	−0.2826 C:*s*	−391.1245	N 14.33
30	CYS38	121	533	−0.2009 C:*p*	−0.2009 C:*p*	−3727.9386	S 10.92
	GLU^-39						O 7.36
31	CYS40	53	221	−0.2193 S:*ps*	−0.2193 S:*ps*	−2651.0122	N [a] 3.01
32	ILE41	62	335	−0.1850 C:*p*	−0.2702 C:*p*	+140.4853	C 6.02
33	ALA42	84	411	−0.1921 C:*ps*	−0.1952 C:*p*	−1026.4444	O [a] 7.44
	SER43						O [a] 9.95
34	ALA44	68	334	−0.2481 C:*ps*	−0.2520 N:*s*	−816.9660	O [a] 9.72
	GLY45						C [a] 8.23
35	PHE46	78	391	−0.2243 C:*p*	−0.2785 C:*ps*	−509.1620	C 7.60
36	VAL47	54	287	−0.2202 C:*p*	−0.2232 C:*ps*	+6.5809	C 8.44
37	ARG^+48	84	444	−0.3095 C:*p*	−0.3407 C:*ps*	−158.4219	O [a] 9.12
38	ASP49	97	455	−0.3606 N:*s*	−0.3606 N:*s*	−1654.1593	O 13.24
	ALA50						C 14.55

TABLE I. (*Continued*).

No.	Amino acid residues	No. of electrons	No. of bases	Bottom unoccupied states (Ry)	Top occupied states (Ry)	Energy (Ry)	Farthest non-H atom distance (a.u.)
39	GLN51	98	477	−0.2963 C:p	−0.3062 N:p	−1308.7041	N 16.63
	GLY52						N [a] 11.19
40	ASN53	113	507	−0.2781 C:p	−0.2781 C:p	−3574.7735	N 11.95
	CYS54						O [a] 6.82
41	ILE55	62	335	−0.2065 C:p	−0.2145 C:p	+60.4517	C 10.70
42	LYS56	71	386	−0.1627 C:p	−0.1627 C:p	−78.2458	N 12.96
43	PHE57	78	391	−0.2534 C:p	−0.2582 C:s	−533.5937	O [a] 12.60
44	GLU58	67	312	−0.3127 N:sp	−0.3127 N:sp	−1083.1508	O 16.94
45	ASP59	112	485	−0.3427 N:p	−0.3463 C:s	−3637.1004	O 13.71
	CYS60						O [a] 15.71
46	PRO61	52	265	−0.2885 C:p	−0.3316 C:p	−359.9561	C [a] 17.63
47	LYS62	79	415	−0.2931 C:s	−0.2931 C:s	−606.6866	N 22.98

[a]The backbone atom.

them, there are 36 side-chain atoms and 26 backbone atoms (with notation a). For the six clusters (19, 5, 42, 8, 25, and 33) with the highest (top) occupied states, only cluster 25 is the possible active site because it is relatively far from the center of the inhibitor and the farthest heavy atom is side-chain oxygen in carboxyl COO^-. While among the seven clusters (38, 45, 44, 37, 12, 39, and 47) with the lowest (bottom) unoccupied local states, only cluster 47 (LYS62) is the possible active site because it is relatively far from the center of the inhibitor and its farthest heavy atom is side-chain nitrogen in NH_3^+ (see Fig. 1, upper right).

For structure 1atb, $\overline{X}_b = 0.0325$ Å, $\overline{Y}_b = 0.0036$ Å, and $\overline{Z}_b = -0.4416$ Å. The eighth column of Table II shows the farthest heavy atoms of 62 residues, and their distances from the mean coordinates ($\overline{X}_b$, $\overline{Y}_b$, $\overline{Z}_b$). Among them, there are 40 side-chain atoms and 22 backbone atoms (with notation a). For the seven clusters (19, 5, 8, 26, 42, 13, and 33) with the highest (top) occupied states, none is a possible active site because they are all close to the center of the inhibitor; while among the seven clusters (38, 47, 45, 12, 46, 44, and 24) with the lowest bottom unoccupied local states, cluster 47 (LYS62, see the *up right* of Fig. 2) and cluster 24 (ARG^+31, see Fig. 2, lower left) are possible active sites because they are relatively far from the center of the inhibitor and their farthest heavy atoms are side-chain nitrogen in NH_3^+ (LYS62) or NH_2^+ (ARG^+31) .

In both structures 1ata and 1atb, however, the residue LYS62 is electrically neutral due to its NH_3^+ of the side chain as well as COO^- at the C-terminal, so it is unlikely for LYS62 to get an additional electron. On the other hand, the eigenvalue −0.1886 Ry of the top occupied state of cluster 25 in structure 1ata is not high enough, and the electron of COO^- in residues GLU^-32 is not so easily lost. Thus the reasonable conclusion is that the structure 1ata (pH 4.75) has no active site, and the structure 1atb (pH 2.4) has only one active site which is a N terminal at residue ARG+31 with eigenvalue −0.2816 Ry of the bottom unoccupied local state (located at the lower left of Fig. 2, compare this part with that of Fig. 1). It should be the easiest position for receiving an electron which may lead to molecule structure change. These results are in agreement with the experimental results: the lower pH form of *Ascaris* trypsin inhibitor is the stable and active form of the protein [23].

C. The changeable parts of protein

The active site of a protein is the easiest part to get or lose an electron which may lead to molecule structure change. Now we discuss possible change of the protein without electron transfer. We use two criteria to decide whether an embedded cluster is changeable: (i) energy per electron and (ii) distance from the center of protein.

For structure 1ata, using the data of third and seventh columns, we find clusters 3, 5, 12, 19, 20, 21, 24, 26, 29, 32, 36, 37, 41, and 42 having relatively higher values of energy per electron. Consider the distance from the center of protein given in the eighth column, only the clusters 3, 5, 12, 24, and 29 are possibly changeable.

For structure 1atb, we find that clusters 3, 5, 12, 19, 20, 21, 26, 29, 32, 36, 37, 41, and 42 have relatively higher values of energy per electron. Consider the distances given in the eighth column of Table II, now only the clusters 3, 5, 12, and 29 are possibly changeable. The structure of cluster 24 (ARG^+31) is now stable without electron transfer, since its energy is reduced from −346.4991 Ry (1ata) to −532.4118 Ry (1atb).

For structure 1atb, the clusters 3, 5, 12, and 29 may be changed while the active site in ARG^+31 is kept, if the clusters' relatively higher energies can be reduced. The change can be in the structure, such as the cluster 24; or can be in the content: the residue in cluster is replaced by another residue. If we can calculate the proteins of viruses such as AIDS, SARS, and flu, such kind of information may be helpful since these viruses have higher variability.

D. The precision of atomic coordinates

It is acknowledged that there are errors in structure determination of protein molecules. Our calculation offers an in-

TABLE II. Information of 47 embedded-clusters of *Ascaris* trypsin inhibitor (structure 1atb).

No.	Amino acid residues	No. of electrons	No. of bases	Bottom unoccupied states (Ry)	Top occupied states (Ry)	Energy (Ry)	Farthest non-H atom distance (a.u.)
1	GLU1	69	334	−0.2240 C:*sp*	−0.2240 C:*sp*	−946.4225	O 21.95
2	ALA2	106	503	−0.2119 C:*sp*	−0.2180 O:*p*	−1625.4397	N [a] 16.11
	$GLU^{-}3$						O 19.32
3	$LYS^{+}4$	70	386	−0.2551 H:*s*	−0.2732 C:*sp*	−46.8194	N 18.04
4	CYS5	107	489	−0.2154 C:*p*	−0.2154 C:*p*	−3083.0463	O [a] 13.92
	THR6						C 16.58
5	LYS7	71	386	−0.1383 C:*s*	−0.1383 C:*s*	−159.3146	N 14.59
6	PRO8	52	265	−0.2300 C:*p*	−0.2543 C:*ps*	−425.0481	C 12.55
7	ASN9	60	286	−0.2478 C:*p*	−0.2727 C:*ps*	−960.3682	N 9.20
8	$GLU^{-}10$	68	312	−0.1088 C:*p*	−0.1558 C:*sp*	−1262.8632	O 9.60
9	GLN11	68	334	−0.2214 C:*p*	−0.2374 C:*ps*	−782.3434	O 10.47
10	TRP12	98	483	−0.2280 C:*p*	−0.2335 C:*ps*	−944.8205	C 13.58
11	THR13	54	268	−0.2078 C:*p*	−0.2607 C:*ps*	−519.4515	O [a] 9.67
12	$LYS^{+}14$	70	386	−0.3087 H:*s*	−0.3111 N:*s*	+0.3114	N 15.75
13	CYS15	83	364	−0.1825 C:*p*	−0.1825 C:*p*	−2802.6864	C 10.26
	GLY16						O [a] 8.62
14	GLY17	83	364	−0.2018 N:*sp*	−0.2018 N:*sp*	−2782.9524	C [a] 8.54
	CYS18						S 10.61
15	$GLU^{-}19$	68	312	−0.1653 C:*p*	−0.2148 O:*p*	−1361.0761	O [a] 7.36
16	GLY20	84	411	−0.1980 C:*ps*	−0.1999 C:*p*	−1073.5125	C [a] 7.05
	THR21						O 10.68
17	CYS22	91	412	−0.2372 C:*p*	−0.2372 C:*p*	−2677.8709	O [a] 13.78
	ALA23						C 14.47
18	GLN24	68	334	−0.1659 C:*p*	−0.2612 C:*p*	−829.2404	O 15.56
19	LYS25	71	386	−0.1201 N:*s*	−0.1201 N:*s*	−75.2499	C 10.97
20	ILE26	62	335	−0.1721 C:*p*	−0.2320 C:*p*	−44.8339	O [a] 12.11
21	VAL27	54	287	−0.1734 C:*ps*	−0.2165 C:*ps*	−77.1470	C [a] 12.06
22	PRO28	105	486	−0.2514 N:*sp*	−0.2514 N:*sp*	−2688.5762	C 15.67
	CYS29						O [a] 14.94
23	THR30	54	268	−0.2263 C:*ps*	−0.2690 C:*sp*	−621.4784	C 17.98
24	$ARG^{+}31$	84	444	−0.2816 H:*s*	−0.3060 C:*sp*	−532.4118	N 23.12
25	$GLU^{-}32$	121	533	−0.2177 C:*p*	−0.2177 C:*p*	−3444.0065	O 18.03
	CYS33						N 13.43
26	LYS34	71	386	−0.1650 N:*sp*	−0.1650 N:*sp*	−28.4694	C [a] 12.17
27	PRO35	52	265	−0.2442 C:*p*	−0.2473 C:*p*	−436.8678	C 14.74
28	PRO36	52	265	−0.2549 C:*p*	−0.3118 C:*ps*	−389.8580	C 14.69
29	$ARG^{+}37$	84	444	−0.2088 H:*s*	−0.2838 C:*ps*	−396.4221	N 14.99
30	CYS38	121	533	−0.1940 S:*p*	−0.1940 S:*p*	−3723.8125	S 10.94
	$GLU^{-}39$						O 7.90
31	CYS40	53	221	−0.2089 S:*ps*	−0.2089 S:*ps*	−2648.9175	N [a] 3.08
32	ILE41	62	335	−0.1758 C:*p*	−0.2648 C:*ps*	+147.8301	C 5.74
33	ALA42	84	411	−0.1845 C:*sp*	−0.1891 C:*p*	−1013.3123	O [a] 7.30
	SER43						O 10.14
34	ALA44	68	334	−0.2469 C:*ps*	−0.2490 N:*s*	−819.0285	O [a] 9.66
	GLY45						C [a] 8.20
35	PHE46	78	391	−0.2164 C:*p*	−0.2626 C:*ps*	−510.2287	C 7.82
36	VAL47	54	287	−0.2318 C:*p*	−0.2491 C:*ps*	+9.4732	C 8.51
37	$ARG^{+}48$	84	444	−0.2794 C:*p*	−0.3324 C:*sp*	−144.5128	O [a] 9.14
38	ASP49	97	455	−0.3475 C:*p*	−0.3475 N:*sp*	−1645.3036	O 13.40
	ALA50						C 14.63

TABLE II. (*Continued*).

No.	Amino acid residues	No. of electrons	No. of bases	Bottom unoccupied states (Ry)	Top occupied states (Ry)	Energy (Ry)	Farthest non-H atom distance (a.u.)
39	GLN51	98	477	−0.2724 O:*p*	−0.2831 N:*p*	−1299.6778	N 17.29
	GLY52						N [a] 11.36
40	ASN53	113	507	−0.2573 C:*p*	−0.2573 C:*p*	−3557.3139	N 12.26
	CYS54						O [a] 7.38
41	ILE55	62	335	−0.1958 C:*p*	−0.2011 C:*p*	+63.7071	C 10.90
42	LYS56	71	386	−0.1678 H:*s*	−0.1678 H:*s*	−60.3267	C 12.40
43	PHE57	78	391	−0.2454 C:*p*	−0.2525 C:*sp*	−531.3582	O [a] 12.71
44	GLU58	67	312	−0.2836 C:*p*	−0.2836 C:*p*	−1081.9486	O 17.55
45	ASP59	112	485	−0.3208 N:*p*	−0.3240 C:*s*	−3628.0559	O 13.85
	CYS60						O [a] 15.93
46	PRO61	52	265	−0.2891 C:*p*	−0.2955 N:*s*	−357.4639	C [a] 17.76
47	LYS62	79	415	−0.3211 C:*sp*	−0.3211 C:*sp*	−645.3597	N 23.45

[a]The backbone atom.

dependent theoretical estimate of the precision of structure determination. Based on the Hellmann-Feynman theory, the total forces acting on each atomic nucleus are calculated after acquiring the electronic structure. In principle, the total force acting on a nucleus in equilibrium should be zero. In the free-cluster calculation and SCCE calculation, however, the charge fitting technique is used, which is designed to produce minimum error in electrostatic energy but not charge density [26]. So even for an atom in equilibrium, the calculated total force is not exactly zero but a small value. Table III gives the average total forces acting on atomic nuclei (per nuclear charge). It is shown that the total force acting on a carbon nucleus is reasonably small, which indicates that the carbon atoms are in equilibrium. Similar conclusion may be valid for sulfur atoms. But for oxygen atoms, the total force per nuclear charge is about 2.3 times as large as that of nitrogen atoms and is significantly larger than the reasonable value. The forces on hydrogen nucleus are also too large to be reasonable. Considering their light nuclei, the coordinates of hydrogen atoms in inhibitor are not accurate. But this may be caused largely by the inaccurate coordinates of heavy atoms.

The results above lead to one undoubted conclusion: The coordinates of oxygen atoms in the *Ascaris* trypsin inhibitor, determined by nuclear magnetic resonance and combination of distance geometry and dynamical simulated annealing, are less accurate than that of other kinds of non-H atoms. In other words, there is a systematical error in determining coordinates of oxygen atoms. Similar conclusion had been obtained before for trypsin inhibitor CMTI-I from squash seeds [18]. The origin for this systematical error is unclear.

Till now, the electronic structures of two proteins have been obtained by first-principles, all-electron, *ab initio* calculations. The calculation of the third protein will be finished soon. This demonstrates that the calculation of electronic structure of a protein molecule is meaningful, and has be-

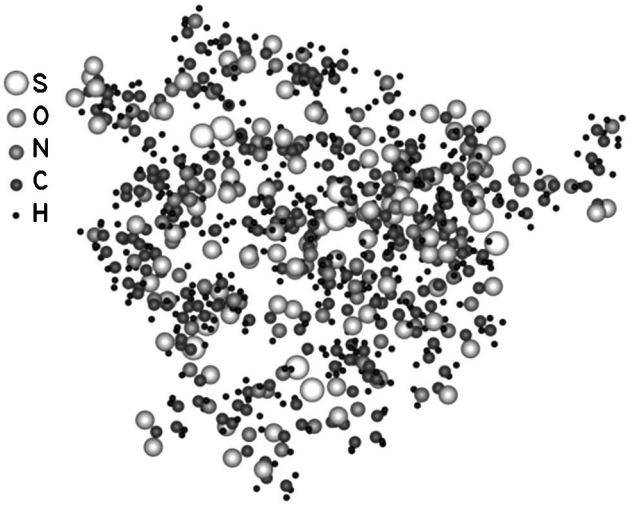

FIG. 1. The structure 1ata of *Ascaris* trypsin inhibitor.

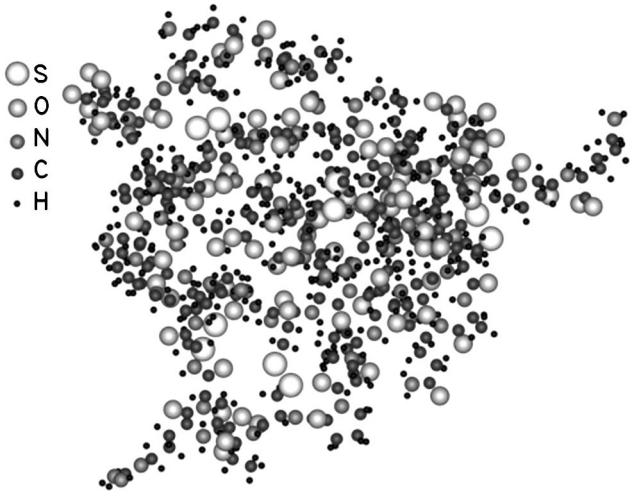

FIG. 2. The structure 1atb of *Ascaris* trypsin inhibitor.

TABLE III. Total forces acted on atoms.

	Average total force per nuclear charge (a.u.)				
Atom	C	N	O	S	H
1ata structure	0.0549	0.1675	0.3855	0.0668	0.3285
1atb structure	0.0575	0.1665	0.3850	0.0654	0.3286

come a reality with the SCCE method. In principle, there is no problem in calculating much larger protein such as HIV-1 protease. It is hoped that the progress will give us valuable clue in the research of diagnosis and treatment, as well as in the development of new medicines, especially for diseases AIDS, SARS, and flu, since these viruses have higher variability.

ACKNOWLEDGMENTS

I acknowledge a grant from the Science and Technology Development Foundation of Shanghai (Grant No. 00JC14051). This work was also supported by the Computer Network Information Center of Chinese Academy of Science in Beijing. The calculations were performed on Dawning2000-II supercomputer which are the clusters of Unix workstations developed by the National Center of Intelligent Computer, China.

[1] G.P.M. Crawford, D.J. Howse, and D.I. Grove, J. Parasitol. **68**, 1044 (1982).
[2] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991); Phys. Rev. A **44**, 7823 (1991).
[3] P. Cortona, Phys. Rev. B **44**, 8454 (1991).
[4] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[5] F. Mauri, G. Galli, and R. Car, Phys. Rev. B **47**, 9973 (1993).
[6] X.-P. Li, R.W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).
[7] T.A. Wesolowski and A. Warshel, J. Phys. Chem. **97**, 8050 (1993).
[8] P. Ordejon, D.A. Drabold, R.M. Martin, and M.P. Grumbach, Phys. Rev. B **51**, 1456 (1995).
[9] W. Yang and Tai-sung Lee, J. Chem. Phys. **103**, 5674 (1995).
[10] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
[11] R. Baer and M. Head-Gordon, Phys. Rev. Lett. **79**, 3962 (1997).
[12] Haoping Zheng, Phys. Rev. B **48**, 14868 (1993).
[13] Haoping Zheng, Phys. Lett. A **226**, 223 (1997); **231**, 453 (1997).
[14] Haoping Zheng, Physica B **212**, 125 (1995).
[15] H. Zheng, B.K. Rao, S.N. Khanna, and P. Jena, Phys. Rev. B **55**, 4174 (1997).
[16] H. Zheng, Y. Wang, and G. Ma, Eur. Phys. J. B **29**, 61 (2002).
[17] J. He and H. Zheng, Acta Phys. Sin. **51**, 2580 (2002).
[18] Haoping Zheng, Phys. Rev. E **62**, 5500 (2000).
[19] Haoping Zheng, Prog. Phys. **20**, 291 (2000).
[20] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[21] W. Kohn and L.J. Sham, Phys. Rev. **140**, A1133 (1965).
[22] A.M. Gronenborn, M. Nilges, R.J. Peanasky, and G.M. Clore, Biochemistry **29**, 183 (1990).
[23] B.L. Grasberger, G.M. Clore, and A.M. Gronenborn, Structure (London) **2**, 669 (1994).
[24] U. von Barth and L. Hedin, J. Phys. C **5**, 1629 (1972).
[25] A.K. Rajagopal, S. Singhal, and J. Kimball (unpublished) as quoted by A.K. Rajagopal, in *Advances in Chemical Physics*, edited by G.I. Prigogine and S.A. Rice (Wiley, New York, 1979), Vol. 41, p. 59.
[26] H. Sambe and R.H. Felton, J. Chem. Phys. **62**, 1122 (1975).