Electronic structure of trypsin inhibitor from squash seeds in aqueous solution

Haoping Zheng*
Pohl Institute of Solid State Physics, Tongji University, Shanghai 200092, China
(Received 4 February 2000; revised manuscript received 25 April 2000)

The electronic structure of the trypsin inhibitor from seeds of the squash *Cucurbita maxima* (CMTI-I) in aqueous solution is obtained by *ab initio*, all-electron, full-potential calculations using the self-consistent cluster-embedding (SCCE) method. The reactive site of the inhibitor is explained theoretically, which is in agreement with the experimental results. It is shown that the coordinates of oxygen atoms in the inhibitor, determined by nuclear magnetic resonance and combination of distance geometry and dynamical simulated annealing, are systematically less accurate than that of other kinds of heavy atoms.

PACS number(s): 87.15.By, 87.10.+e

I. INTRODUCTION

In order to understand why a protein macromolecule with certain three-dimensional structure is of some special biological functions, it is essential to know the electronic structures of related protein macromolecules. Thus the calculation of electronic structures of protein macromolecules is of considerable importance to life science and drug design. However, this is a good challenge. A biological macromolecule consists of 10^3 to 10^5 atoms distributed nonperiodically. The computational effort of a free-cluster calculation based on density functional theory (DFT) scales as N^3, where the N is the number of electrons in system. In general, it is impossible to perform an *ab initio*, all-electron, full-potential free-cluster calculation for a biological macromolecule even using supercomputers.

In past several years, there has been a great deal of interest in developing so-called $O(N)$ methods for DFT, which scale linearly with number N of atoms [1–6]. One of them has been applied to large free clusters containing several hundreds of carbon atoms, but given only the total energy of the cluster [7]. The self-consistent cluster-embedding (SCCE) calculation method [8] is developed by the author. It uses a *localized* noninteracting electron model to describe systems. Five successful SCCE calculations have been performed for the crystals NiO, CoO, $LaNi_5$, Ni and hydrogen-decorated vacancies in Ni [9–12]. By using the ''divide-and-conquer'' scheme, the computational effort of a SCCE calculation may scale linearly with the number of atoms, which enables us to calculate biological macromolecules.

Squash proteinase inhibitors are the smallest known protein inhibitors of serine proteinases. They consist of 29 to 32 amino acid residues. On the basis of amino acid sequence, reactive site location and half-cystine content, they were established as a family of serine proteinase inhibitors. The trypsin inhibitor from seeds of the squash *Cucurbita maxima* (CMTI-I) is a representative member of this group of inhibitors. Despite its small size, it is a powerful inhibitor, binding to bovine β-trypsin with an association constant of $3.2 \times 10^{11}\ M^{-1}$, which is among the highest of those determined for trypsin [13–16].

The present paper represents our first effort to perform an *ab initio*, all-electron, full-potential calculation for a squash trypsin inhibitor using the SCCE method. The obtained electronic structure is well converged and meaningful. Considering the central role played by proteinase inhibitors in many physiological processes, this may help in the development of smaller peptide analogs with clinical potential. Section II gives the theoretical model. The computational procedure is described in Sec. III. The results and discussions are given in Sec. IV. Section V is conclusions.

II. THEORETICAL MODEL

The self-consistent cluster-embedding (SCCE) calculation method [8] is described briefly as follows. According to density functional theory (DFT), the ground-state energy functional of a system containing N electrons and M fixed nuclei can be written in the form [17,18] (no relativistic effect is included)

$$E_G[\rho]=T[\rho]+E_{xc}[\rho]+\int\int\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}\,d\mathbf{r}' -2\sum_{j=1}^{M}\int\frac{\rho(\mathbf{r})Z_j}{|\mathbf{r}-\mathbf{R}_j|}d\mathbf{r}+\sum_{i\neq j}^{M}\frac{Z_iZ_j}{|\mathbf{R}_i-\mathbf{R}_j|}, \quad (1)$$

where $T[\rho]$ is the total kinetic energy of N *noninteracting* (NI) electrons. When $\rho(\mathbf{r})$ is correct, $E_G[\rho]$ attains its minimum. Atomic units with energy in Rydbergs are used throughout this paper ($e^2=2$, $\hbar=1$, $2m_e=1$).

In order to solve Eq. (1), the local spin density approximation (LSDA) is used to give the form of $E_{xc}[\rho]$. The ''one-electron approximation'' is used to introduce a set of NI electrons which can be represented by a set of *stationary state* one-electron wave functions $\phi_n^\sigma(\mathbf{r})$. The total $\rho(\mathbf{r})$ and $T[\rho]$ are the sums of the charge density and kinetic energy of each NI electron, respectively,

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

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

$$T[\rho]=\sum_{\text{occupied } l}\int \phi_l^{up*}(\mathbf{r})(-\nabla^2)\phi_l^{up}(\mathbf{r})d\mathbf{r}$$

$$+\sum_{\text{occupied } m}\int \phi_m^{dn*}(\mathbf{r})(-\nabla^2)\phi_m^{dn}(\mathbf{r})d\mathbf{r}. \quad (3)$$

The single-electron Schrödinger equations are obtained by the variation of functional (1) with respect to $\phi_n^{\sigma*}(\mathbf{r})$ under the conservation rule $\delta\int\rho(\mathbf{r})d\mathbf{r}=0$, in which the $\phi_n^{\sigma}(\mathbf{r})$ are trial wave functions.

Usually, the ''spread NI-electron model'' (the first kind of trial wave functions) is used: each one-electron wave function $\phi_n^{\sigma}(\mathbf{r})$ is assumed to spread over the whole region occupied by the system. The variational principle leads to the Kohn-Sham equation [18], which can be used for free-cluster calculation with a natural finite boundary condition, or for band structure calculation with a periodic boundary condition.

Here we use the ''localized NI-electron model'' (the second kind of trial wave functions): each one-electron wave function $\phi_n^{\sigma}(\mathbf{r})$ is assumed to be distributed in a part of the region occupied by the system. So the system can be divided into k embedded clusters. Each time, we calculate only a subset of one-electron eigenfunctions localized in and around an embedded cluster. This can be achieved by the self-consistent cluster-embedding (SCCE) calculations [8]. We use $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ to represent the electron charge densities located in the embedded cluster and surrounding regions (with small overlap), respectively. For N ($N=N_1+N_2$) localized and independent $\phi_n^{\sigma}(\mathbf{r})$, we have

$$\rho(\mathbf{r})=\sum_{\text{occupied } n\ \sigma}^{\mathrm{N}}|\phi_n^{\sigma}(\mathbf{r})|^2=\sum_{\text{occupied } n_1\ \sigma}^{\mathrm{N}_1}|\phi_{n_1}^{\sigma}(\mathbf{r})|^2+\sum_{\text{occupied } n_2\ \sigma}^{\mathrm{N}_2}|\phi_{n_2}^{\sigma}(\mathbf{r})|^2$$

$$\equiv\rho_1(\mathbf{r})+\rho_2(\mathbf{r}),$$

$$T[\rho]=T[\rho_1+\rho_2]=\sum_{\text{occupied } n\ \sigma}^{\mathrm{N}}\int\phi_n^{\sigma*}(\mathbf{r})(-\nabla^2)\phi_n^{\sigma}(\mathbf{r})d\mathbf{r}$$

$$=\sum_{\text{occupied } n_1\ \sigma}^{\mathrm{N}_1}\int\phi_{n_1}^{\sigma*}(\mathbf{r})(-\nabla^2)\phi_{n_1}^{\sigma}(\mathbf{r})d\mathbf{r}+\sum_{\text{occupied } n_2\ \sigma}^{\mathrm{N}_2}\int\phi_{n_2}^{\sigma*}(\mathbf{r})(-\nabla^2)\phi_{n_2}^{\sigma}(\mathbf{r})d\mathbf{r}$$

$$\equiv T[\rho_1]+T[\rho_2].$$

A *zero-value* term $\int\rho_1(\mathbf{r})V_{or}d\mathbf{r}$ is added to the right side of formula (1). For fixed $\rho_2(\mathbf{r})$, the variational principle now leads to the basic equation of the SCCE method [8]:

$$\left\{-\nabla^2+2\int\frac{\rho_1(\mathbf{r}')+\rho_2(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'-2\sum_{i=1}^{\mathrm{M}}\frac{Z_i}{|\mathbf{r}-\mathbf{R}_i|}+V_{xc}^{\sigma}(\mathbf{r})+V_{or}\right\}\phi_n^{\sigma}(\mathbf{r})=\epsilon_n^{\sigma}\phi_n^{\sigma}(\mathbf{r}), \quad (4)$$

where the $\phi_n^{\sigma}(\mathbf{r})$ represent only the cluster electrons localized in and around the embedded cluster, and satisfy the natural finite boundary condition as well as a special finite boundary condition caused by V_{or}:

$$\phi_n^{\sigma}(\mathbf{r})\big|_{\mathbf{r}\text{ is in the core regions of surrounding atoms}}=0. \quad (5)$$

For a real finite system, by calculating all k embedded clusters one by one, Eq. (4) gives a complete set of one-electron eigenfunctions of the *whole system* which makes the total energy in formula (1) minimal.

The physical reasons for boundary condition (5) are given in Ref. [8]. The validity of condition (5) and V_{or} can be understood as follows. The single-electron Schrödinger equation (4) is exactly the same as the Kohn-Sham equation except for the ''orthogonality constraint'' V_{or} which is defined as

$$V_{or}=\begin{cases}2\sum_{j=1}^{\mathrm{M}_2}\dfrac{Z_j}{|\mathbf{r}-\mathbf{R}_j|} & \text{if } \mathbf{r} \text{ is in the core regions of surrounding atoms,}\\ 0 & \text{otherwise,}\end{cases} \quad (6)$$

TABLE I. Gaussian bases.

C		N		O	
Exponent	Coefficient	Exponent	Coefficient	Exponent	Coefficient
s (8)		*s* (8)		*s* (8)	
50557.501	0.000055	74761.715	0.000050	105374.95	0.000046
7524.7856	0.000434	11123.654	0.000394	15679.240	0.000361
1694.3276	0.002316	2512.6857	0.002088	3534.5447	0.001920
472.82279	0.009872	703.77729	0.008906	987.36516	0.008206
151.71075	0.035217	225.47879	0.032081	315.97875	0.029725
53.918746	0.104184	79.615810	0.097424	111.65428	0.090452
20.659311	0.241255	30.237283	0.231728	42.699451	0.217402
8.3839760	0.383929	12.263622	0.377540	17.395596	0.368720
3.5770150	1.000000	5.2650860	1.000000	7.4383090	1.000000
1.5471180	1.000000	2.3334710	1.000000	3.2228620	1.000000
0.6130130	1.000000	0.9018560	1.000000	1.2538770	1.000000
0.2460680	1.000000	0.3583360	1.000000	0.4951550	1.000000
0.0990870	1.000000	0.1410930	1.000000	0.1916650	1.000000
0.0495435	1.000000	0.0634918	1.000000	0.0862493	1.000000
p (6)		*p* (7)		*p* (7)	
83.333155	0.001254	126.66657	0.001152	200.00000	0.000892
19.557611	0.009640	29.837389	0.009016	46.533367	0.007351
6.0803650	0.042873	9.3940380	0.040814	14.621809	0.034863
2.1793170	1.000000	3.4051040	1.000000	5.3130640	1.000000
0.8651500	1.000000	1.3500000	1.000000	2.1025250	1.000000
0.3619440	1.000000	0.5576960	1.000000	0.8502230	1.000000
0.1547400	1.000000	0.2324490	1.000000	0.3375970	1.000000
0.0654290	1.000000	0.0942640	1.000000	0.1288920	1.000000
		0.0424188	1.000000	0.0580014	1.000000

H		S			
Exponent	Coefficient	Exponent	Coefficient	Exponent	Coefficient
s (8)		*s* (9)		*p* (9)	
1170.4980	0.000050	202717.22	0.0001170	785.83222	0.0014330
173.58220	0.000580	30160.149	0.0009140	182.88200	0.0120956
38.651630	0.003180	6841.8735	0.0048116	57.574868	0.0599535
10.607200	0.013800	1914.8305	0.0203059		
		614.46376	0.0703499	21.044853	1.0000000
3.3796490	1.000000	218.31321	0.1932939	8.3062253	1.0000000
1.2025180	1.000000	83.399394	0.3753917	3.3726553	1.0000000
0.4639250	1.000000			1.3677634	1.0000000
0.1905370	1.000000	33.598513	1.0000000	0.47036433	1.0000000
0.0812406	1.000000	13.618960	1.0000000	0.19755302	1.0000000
0.0381831	1.000000	5.4517072	1.0000000	0.08889886	1.0000000
0.0189006	1.000000	2.2832059	1.0000000	0.04178246	1.0000000
		0.88690243	1.0000000		
p (1)		0.38742399	1.0000000	*d* (1)	
0.7500000	1.000000	0.16519553	1.0000000	0.6500000	1.0000000
		0.07895825	1.0000000		

where M_2 is the number of surrounding atoms. In the calculations, the V_{or} cancels the nuclear Coulomb potential in the core regions of all surrounding atoms. The cluster electrons will only feel an electron-electron positive Coulomb potential in these regions and be forced out, which leads to the boundary condition (5). It is easy to see that, as long as the boundary condition (5) is satisfied, we have $\int \rho_1(\mathbf{r}) V_{or} d\mathbf{r} = 0$. So the V_{or} in Eq. (4) has no contribution to the total energy, and the SCCE calculation is valid according to the DFT. In practice, the optimum values of core radii of surrounding atoms are determined according to two criteria: (a) there is no collapse disaster; (b) the total cluster electrons

TABLE II. Information of 20 embedded clusters. (*) The backbone atom.

Cluster no.	Amino acid residues	Number of bases	Number of grid points	Bottom unoccupied state (Ry)	Top occupied state (Ry)	Farthest heavy atom (a.u.)
1	ARG^+1	466	669513	−0.2921	−0.3027	N 30.29
2	VAL2	287	729909	−0.2902	−0.3119	C 21.93
3	CYS3	486	802945	−0.2795	−0.2833	N* 16.13
	PRO4					C 18.81
4	ARG^+5	444	714271	−0.3278	−0.3312	N 25.00
5	ILE6	335	811301	−0.1953	−0.2734	C 17.24
6	LEU7	335	783203	−0.1562	−0.2270	C 17.40
7	MET8	328	865708	−0.1934	−0.2651	C* 13.35
8	GLU^-9	533	819950	−0.1607	−0.1647	O 21.45
	CYS10					O* 15.35
9	LYS^+11	386	749471	−0.2663	−0.2766	N 24.50
10	LYS^+12	386	750831	−0.3041	−0.3246	N 24.86
11	ASP^-13	484	805325	−0.2232	−0.2248	O 14.75
	SER14					O 19.37
12	ASP^-15	485	842549	−0.1862	−0.2098	O 16.58
	CYS16					O* 13.62
13	LEU17	526	794663	−0.2018	−0.2110	C 21.40
	ALA18					C* 17.34
14	GLU^-19	533	806446	−0.2444	−0.2597	O 22.59
	CYS20					N* 11.18
15	VAL21	508	889169	−0.1782	−0.1986	C 12.47
	CYS22					O* 11.96
16	LEU23	335	841895	−0.2113	−0.2519	O* 14.12
17	GLU24	312	695040	−0.2845	−0.3006	O 25.01
18	HIS25	492	771202	−0.0692	−0.1730	C 21.73
	GLY26					N* 15.60
19	TYR27	420	809078	−0.1857	−0.2478	O 21.88
20	CYS28	393	916035	−0.2795	−0.3201	O* 6.66
	GLY29					O 10.89

remaining in the surrounding core regions are minimum. It is found that the results are not sensitive to the core radii if they are around the optimum values. In general, the boundary condition (5) is satisfied with high precision.

It should be pointed out that Wesolowski and Warshel have developed a method [19,20] in which the basic idea of partitioning the overall system in localized sybsystems is also used, but the way of obtaining the charge density of the system is different.

III. COMPUTATIONAL PROCEDURE

The complete three-dimensional structure of trypsin inhibitor from seeds of the squash *Cucurbita maxima* (CMTI-I) in aqueous solution is determined by nuclear magnetic resonance and combination of distance geometry and dynamical simulated annealing [21,22]. The above structure, however, is in complete disagreement with the models proposed by Hider *et al.* based on circular dichroism and Chou-Fasman analysis [13]. We choose the results of Ref. [21] as our starting point and take the atomic coordinates of 436 atoms from Protein Data Bank (PDB 1cti, minimized mean structure). The CMTI-I contains 29 amino acid residues. In aqueous solution, the H^+ is away from the carboxyl group of the sidechain of residues ASP and GLU. Besides, there are three disulphide bonds from six half-cystines. So the coordinates of 12 hydrogen atoms do not appear in the PDB 1cti.

Based on optimized Gaussian basis sets of H, C, N, O, and S atoms [23–27], parts of the original bases are uncontracted, several diffuse bases are inserted and two polarization functions are added (Table I). The current basis sets of C(26), N(29), O(29), H(11), and S(41) may be considered as adequate. The protein molecule CMTI-I is divided into 20 embedded clusters (Table II, the first and second columns). The third column gives the total numbers of Gaussian bases used for each embedded cluster. The mean coordinates of nonhydrogen atoms in each embedded cluster are calculated. The grid points, filling the sphere with radius 27.5 a.u. centered at mean coordinates of each embedded cluster, are taken for the calculation of V_{xc} (Table II, the fourth column). The fifth, sixth, and seventh columns of Table II will be explained in the next section.

The calculation contains two kinds of iterations: (i) Intracluster iteration. For each embedded cluster, Eq. (4) is calculated as follows: the embedded cluster is served as $\rho_1(\mathbf{r})$ which is self-consistently changed during the intracluster iterations, while the remainder of the protein is served as a

TABLE III. Core radii of five atoms.

Atoms	C	N	O	H	S
Core radius (a.u.)	0.7783	0.6734	0.7094	0.6465	1.1228

fixed surrounding environment $\rho_2(\mathbf{r})$ (see Fig. 1). (ii) Intercluster iteration. The 20 embedded clusters are synchronously calculated by 20 CPUs, respectively. After the convergence of intracluster iterations of all 20 embedded clusters, the results are used for constructing new surrounding environments $\rho_2(\mathbf{r})$ for each embedded cluster, and a new intercluster iteration begins.

After several testing calculations, the core radii of five elementary atoms, used for surrounding atoms in the SCCE calculations, are determined (see Table III). There is no collapse disaster. Because of limited computational time and up to five radii, we cannot say that the core radii are fully optimized. We believe that this has no obvious effect on the results. The well converged results are obtained after eight intercluster iterations. For each embedded cluster, the total cluster electrons remaining in the surrounding core regions are calculated and given in Table IV, which show that the special finite boundary condition (5) is satisfied with high precision.

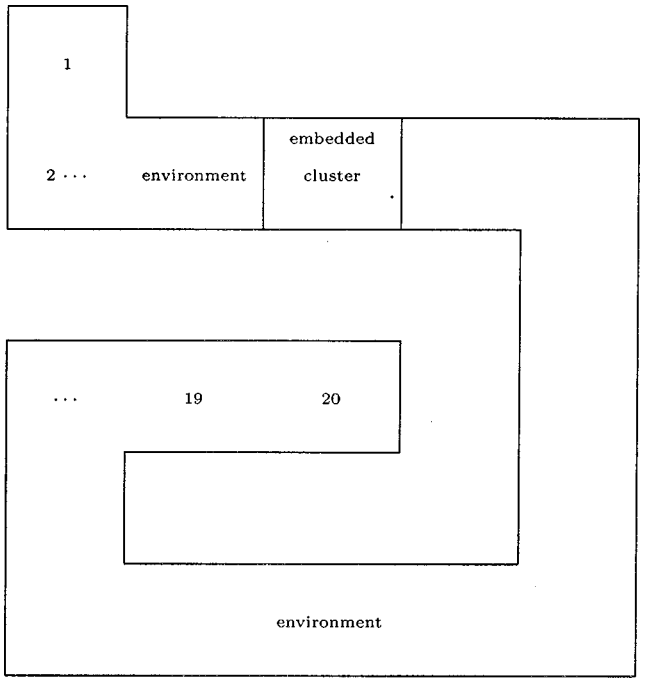

FIG. 1. Schematic diagram of the SCCE calculation: an embedded cluster and its environment in trypsin inhibitor CMTI-I.

IV. RESULTS AND DISCUSSIONS

The experiments show that an inhibitor usually has several reactive sites where the molecule can be attacked much easier than other parts. This can be understood physically as follows: (a) A valence electron in an inhibitor is localized; it cannot be shared by all atoms in the inhibitor. (b) A reactive site has two possible ways to interact: there is the highest occupied local valence electron which is easy to remove, or there is the lowest unoccupied local state which has a tendency to be occupied by an additional electron. Apparently, the *localized* noninteracting (NI) electron model used by the SCCE calculation is suitable to describe the inhibitor.

In our ''divide-and-conquer'' SCCE calculation, each NI electron is localized in and around an embedded cluster. So no amino acid residue (neutral or charged) is of fractional charge number. Each amino acid residue is initially treated as electrically neutral, thus the top occupied local states of 20 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 highest top occupied local state) to amino acid residue j (if j has 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 a much lower unoccupied local state in residue i and a much higher top occupied local state in residue j, the electron transfer is canceled. In the viewpoint of chemistry, only the NH_3^+ (or NH_2^+) in residues ARG, LYS, and HIS, the COO^- in residues GLU and ASP are tested. In general, we must determine the configuration by comparing the energy differences of all kinds of configurations. Limited by computational time, however, we are unable to try all possible electron transfers. After several tests, it is shown that the charged residues GLU^-24, GLY^-29, HIS^+25 and $ARG^{++}1$ are impossible. A reasonable configuration is as follows: (1) ARG^+1, (4) ARG^+5, (9) LYS^+11, (10) LYS^+12; (8) GLU^-9 and CYS10, (11) ASP^-13 and SER14, (12) ASP^-15 and CYS16, (14) GLU^-19 and

TABLE IV. Cluster electrons remaining in the cores of surrounding atoms.

Cluster no.	1	2	3	4	5	6	7
No. of electrons	0.0009	0.0029	0.0034	0.0022	0.0029	0.0024	0.0027
Cluster no.	8	9	10	11	12	13	14
No. of electrons	0.0028	0.0023	0.0028	0.0026	0.0034	0.0024	0.0032
Cluster no.	15	16	17	18	19	20	
No. of electrons	0.0026	0.0028	0.0022	0.0028	0.0026	0.0018	

TABLE V. Part of eigenvalues of 20 embedded clusters.

Clusters	1	2	3	4	5	6	7
Residues	ARG^+1	VAL2	CYS3 PRO4	ARG^+5	ILE6	LEU7	MET8
Unoccupied	−0.2921 C:p+s	−0.2902 C:p+s	−0.2795 C:p+S:p	−0.3278 C:p+N:p	−0.1953 C:p+O:p	−0.1562 C:p+O:p	−0.1934 C:p+O:p
Occupied	−0.3027 C:p+s	−0.3119 C:s+p	−0.2833 C:p+S:p	−0.3312 C:s+N:p	−0.2734 C:p+H:s	−0.2270 N:p+C:p	−0.2651 C:s+N:p
Eigenvalues (Ry)	−0.5223	−0.5169	−0.2996	−0.5177	−0.4230	−0.4353	−0.3116
	−0.5916	−0.6382	−0.3740	−0.5651	−0.5684	−0.5342	−0.4131
	−0.6894	−0.6586	−0.4630	−0.6219	−0.5763	−0.5552	−0.4478
	−0.7309	−0.6856	−0.5020	−0.7098	−0.5944	−0.5849	−0.5519
Clusters	8	9	10	11	12	13	14
Residues	GLU^-9 CYS10	LYS^+11	LYS^+12	ASP^-13 SER14	ASP^-15 CYS16	LEU17 ALA18	GLU^-19 CYS20
Unoccupied	−0.1607 C:p+S:p	−0.2663 H:s+C:p	−0.3041 N:s+H:s	−0.2232 H:S+C:p	−0.1862 C:p+O:p	−0.2018 C:p+s	−0.2444 C:p+O:p
Occupied	−0.1647 C:p+N:s	−0.2766 C:p+N:s	−0.3246 C:s+N:s	−0.2248 C:s+p	−0.2098 C:p+N:p	−0.2110 C:p+N:p	−0.2597 N:p+C:s
Eigenvalues (Ry)	−0.1738	−0.4863	−0.4721	−0.2776	−0.2131	−0.4236	−0.2688
	−0.1825	−0.6693	−0.6682	−0.2925	−0.2884	−0.4495	−0.2735
	−0.1984	−0.7155	−0.7065	−0.3084	−0.3052	−0.4806	−0.2780
	−0.2250	−0.7212	−0.7267	−0.4261	−0.3137	−0.5431	−0.2855
Clusters	15	16	17	18	19	20	
Residues	VAL21 CYS22	LEU23	GLU24	HIS25 GLY26	TYR27	CYS28 GLY29	
Unoccupied	−0.1782 C:p+S:p	−0.2113 C:p+H:s	−0.2845 C:p+O:p	−0.0692 N:p+H:s	−0.1857 C:p+O:p	−0.2795 S:p+C:p	
Occupied	−0.1986 C:p+N:p	−0.2519 C:s+N:p	−0.3006 C:p+O:p	−0.1730 C:p+O:p	−0.2478 C:p+N:s	−0.3201 O:p+C:p	
Eigenvalues (Ry)	−0.1998	−0.3885	−0.3025	−0.2306	−0.3180	−0.3296	
	−0.3652	−0.5333	−0.3292	−0.3669	−0.3497	−0.3363	
	−0.4144	−0.5482	−0.3335	−0.3803	−0.3898	−0.3813	
	−0.4187	−0.5615	−0.5596	−0.4492	−0.4709	−0.4166	

CYS20; all other 12 embedded clusters are electrically neutral (see Table II, second column).

After eight intercluster iterations, well converged results are obtained. Table V gives part of the eigenvalues of each embedded cluster. The second and fourth rows in Table V show the results of Mulliken population analysis. For example, the ''C:s+N:p'' below the top occupied local state of cluster 4 means that the state is mainly occupied by carbon *s* and nitrogen *p* valence electrons. For convenience, we put the eigenvalues of bottom unoccupied states and top occupied states of 20 embedded-clusters into fifth and sixth columns of Table I, respectively. The embedded clusters 4, 10, and 1 have lowest bottom unoccupied local states. They should be the easiest positions for receiving an electron. The embedded clusters 8, 18, and 15 have highest top occupied local states. They should be the easiest positions for removing an electron.

In order to discuss the reactive sites, the mean coordinates of all heavy (nonhydrogen) atoms in CMTI-I are calculated: $\overline{X}=0.7978$ Å, $\overline{Y}=0.2159$ Å, $\overline{Z}=1.1366$ Å. The seventh column of Table I gives the farthest heavy atoms and their distances from the mean coordinates $(\overline{X}, \overline{Y}, \overline{Z})$. Among them, there are 19 sidechain atoms and 10 backbone atoms (with asterisk notation). The three clusters (8, 18, and 15) with highest top occupied states cannot be reactive sites for two

TABLE VI. Total forces acted on atoms.

Atom	C	N	O	S	H
Total force per nuclear charge (a.u.)	0.0041 to 0.1202	0.0394 to 0.2420	0.2870 to 0.4261	0.0478 to 0.1177	0.2380 to 0.5229

TABLE VII. Total forces acted on atoms N, O, and S.

Cluster no. / Residues	Atom's no. in residue and name / Total force per nuclear charge (a.u.)						
1	1 N	4 O	8 Ne	10 Nh1	11 Nh2		
ARG^+	0.0941	0.3470	0.1396	0.0772	0.0735		
2	1 N	4 O					
VAL	0.1750	0.3732					
3	1 N	4 O	6 Sg		1' N	4' O	
CYS PRO	0.2384	0.3696	0.1071		0.0866	0.3779	
4	1 N	4 O	8 Ne	10 Nh1	11 Nh2		
ARG^+	0.2019	0.3839	0.1444	0.0545	0.0793		
5	1 N	4 O					
ILE	0.2324	0.4055					
6	1 N	4 O					
LEU	0.1956	0.3861					
7	1 N	4 O	7 Sd				
MET	0.2009	0.3454	0.1177				
8	1 N	4 O	8 Oe1	9 Oe2	1' N	4' O	6' Sg
GLU^- CYS	0.2348	0.3541	0.4152	0.3925	0.1575	0.3552	0.0478
9	1 N	4 O	9 Nz				
LYS^+	0.1623	0.3986	0.0743				
10	1 N	4 O	9 Nz				
LYS^+	0.1921	0.3616	0.0765				
11	1 N	4 O	7 Od1	8 Od2	1' N	4' O	6' Og
ASP^- SER	0.1976	0.3574	0.3920	0.4053	0.1408	0.3919	0.2870
12	1 N	4 O	7 Od1	8 Od2	1' N	4' O	6' Sg
ASP^- CYS	0.1912	0.3389	0.3965	0.4176	0.1013	0.3808	0.1003
13	1 N	4 O			1' N	4' O	
LEU ALA	0.1734	0.3607			0.0793	0.3535	
14	1 N	4 O	8 Oe1	9 Oe2	1' N	4' O	6' Sg
GLU^- CYS	0.1869	0.3522	0.4069	0.4126	0.1040	0.3687	0.0856
15	1 N	4 O			1' N	4' O	6' Sg
VAL CYS	0.2377	0.3169			0.1468	0.3700	0.0529
16	1 N	4 O					
LEU	0.1527	0.3603					
17	1 N	4 O	8 Oe1	9 Oe2			
GLU	0.1791	0.4149	0.4211	0.4102			
18	1 N	4 O	7 Nd1	10 Ne2	1' N	4' O	
HIS GLY	0.2010	0.3649	0.0394	0.2420	0.1536	0.4261	
19	1 N	4 O	12 Oh				
TYR	0.1901	0.3874	0.3037				
20	1 N	4 O	6 Sg		1' N	4' O	5' Oxt
CYS GLY	0.1701	0.3631	0.0858		0.1502	0.3961	0.4041

reasons: (Ai) they are relatively close to the center of the inhibitor; (Aii) the farthest heavy atoms of clusters 18 and 15 are not sidechain oxygen atoms. The three clusters (4, 10, and 1) with lowest bottom unoccupied local states are possible reactive sites because of two reasons: (Bi) they are relatively far from the center of the inhibitor and their sidechains are orientated away from the center; (Bii) the farthest heavy atoms are sidechain nitrogen in NH_3^+ (or NH_2^+). Among them, the ARG^+5 of cluster 4 is the most likely reactive site because it has the lowest bottom unoccupied local state. This is in agreement with the experimental results: the trypsin inhibitor from squash seeds (CMTI-I) in aqueous solution has an N-terminus at residue ARG5 [13–15. For the first time, the reactive site is demonstrated by theoretical calculations. Besides, the calculated electronic structure may help to explain the unusual properties of CMTI-I: it strongly inhibits the human Hageman factor fragment (HF_f, β-factor XII_a) and bovine β-trypsin, but is lacking in inhibition of human plasma, human urinary, porcine pancreatic kallikreins, human α-thrombin, and bovine α-chymotrypsin [16].

It is acknowledged that there are errors in structure determination of biological macromolecules. Our calculation offers an independent theoretical testimony to the precision of structure determination. Based on the Hellmann-Feynman theory, the total forces acted on each atomic nucleus are

TABLE VIII. Total forces acted on atoms C and H. (*) This is the mean value of 0.3262, 0.3114, and 0.3342.

		Total force per nuclear charge (a.u.)			
	Amino	C		H	
Cluster no.	acid residues	Largest in backbone	Sidechain	To N in backbone	Else
1	ARG^+1	0.0784	0.0280 to 0.0590	0.3239*	0.2806 to 0.3946
2	VAL2	0.1202	0.0248 to 0.0463	0.4703	0.2513 to 0.3517
3	CYS3	0.0746	0.0200 to 0.0415	0.4613	0.2756 to 0.3718
	PRO4	0.0928		——	
4	ARG^+5	0.0591	0.0260 to 0.0390	0.4444	0.2583 to 0.3579
5	ILE6	0.0707	0.0272 to 0.0436	0.4299	0.2629 to 0.3315
6	LEU7	0.0708	0.0202 to 0.0380	0.4144	0.2395 to 0.3560
7	MET8	0.0953	0.0176 to 0.0260	0.4346	0.2798 to 0.3487
8	GLU^-9	0.1086	0.0196 to 0.0661	0.4750	0.2382 to 0.3925
	CYS10	0.0958		0.4345	
9	LYS^+11	0.0668	0.0208 to 0.0416	0.4406	0.2794 to 0.3893
10	LYS^+12	0.0939	0.0175 to 0.0741	0.4280	0.2829 to 0.3764
11	ASP^-13	0.0914	0.0302 to 0.0459	0.4225	0.2568 to 0.4100
	SER14	0.0523		0.4429	
12	ASP^-15	0.0734	0.0371 to 0.0485	0.4543	0.2698 to 0.3897
	CYS16	0.0725		0.4412	
13	LEU17	0.0777	0.0158 to 0.0421	0.4039	0.2436 to 0.3594
	ALA18	0.0852		0.3933	
14	GLU^-19	0.0592	0.0041 to 0.0575	0.4464	0.2458 to 0.3840
	CYS20	0.0835		0.4794	
15	VAL21	0.0732	0.0225 to 0.0399	0.4573	0.2409 to 0.3775
	CYS22	0.0783		0.4357	
16	LEU23	0.1092	0.0168 to 0.0553	0.4758	0.2490 to 0.3532
17	GLU24	0.0594	0.0185 to 0.0459	0.4757	0.2380 to 0.3696
18	HIS25	0.0706	0.0212 to 0.1078	0.4862	0.2731 to 0.4387
	GLY26	0.0505		0.4477	
19	TYR27	0.0948	0.0091 to 0.0675	0.4345	0.2756 to 0.3790
20	CYS28	0.0678	0.0200 to 0.0200	0.5229	0.2893 to 0.4021
	GLY29	0.0424		0.4034	

calculated after the acquirement of electronic structure. In principle, the total force acted on a nucleus in equilibrium should be zero. In the SCCE calculations (and free-cluster calculations), however, the charge fitting technique is used which is designed to produce minimum error in electrostatic energy but not charge density [28]. So even for an atom in equilibrium, the calculated total force is not exactly zero but a small value. Tables VI, VII, and VIII give the total forces (per nuclear charge) acted on atomic nuclei whose names follow the rules of PDB 1cti. The three tables show (i) the total force acted on a carbon nucleus is reasonably small (mean value is about 0.045 a.u.), which indicates that the carbon atoms are in equilibrium. Similar conclusion may be valid for sulphur atoms whose mean value is about 0.08 a.u. (ii) For oxygen atoms, no matter where they are, the total force per nuclear charge is 0.2870–0.4261 a.u., which is about twice as large as that of nitrogen atoms and is significantly larger than the reasonable value. (iii) The total forces acted on a hydrogen nucleus are 0.2380–0.5229 a.u. Table VIII shows that the forces on hydrogen nucleus connecting to backbone nitrogen atoms are larger than that on other hydrogen nucleus. 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 inhibitor, determined by nuclear magnetic resonance and combination of distance geometry and dynamical simulated annealing, are less accurate than that of other kinds of heavy atoms. In other words, there is a systematical error in determining coordinates of oxygen atoms.

A brief discussion on calculation approaches may be helpful now.

(1) A simple estimation shows that an all-electron, full-potential *free-cluster* calculation for CMTI-I would require 12 GB main memory and 9450 GB hard disk if the same Gaussian basis sets were used. One iteration would take about 1500 hours if a computer containing one CPU were used. It is likely that such a free-cluster calculation would be difficult to converge because the CMTI-I contains so many atoms and has neither translation symmetry nor point symmetry.

(2) In the SCCE calculation, a NI electron is assumed to be distributed in a part (an embedded cluster) of the region

occupied by the inhibitor. So each embedded cluster, neutral or charged, has an integer number of electrons, which leads to a localized reactive site. This is indeed the key of the success of this approach. In free-cluster calculation, each NI electron is assumed to spread over the whole region occupied by the inhibitor, which means no localized reactive site.

(3) The DFT says that the correct charge density $\rho(\mathbf{r})$ corresponds uniquely to true ground-state energy which is minimum. This is all right. But when the total energy $E_G[\rho]$ is written as Eq. (1), the one-electron approximation has already been used because the $T[\rho]$ in Eq. (1) is the total kinetic energy of a set of *NI electrons* (see Ref. [18]). The $E_G[\rho]$ is obtained by variation of functional (1) with respect to $\phi_n^{\sigma *}(\mathbf{r})$ [not to $\rho^{\sigma}(\mathbf{r})$]. So according to variational principle, the calculated energy will be close to the true ground-state energy, only if the NI electrons [the trial wave functions $\phi_n^{\sigma}(\mathbf{r})$] resemble the real electrons. We believe that the *localized* real valence electrons of protein macromolecules can be best described by *localized* NI electrons.

V. CONCLUSIONS

For the first time, *ab initio*, all-electron, full-potential calculations are performed for a biological macromolecule. The electronic structure of trypsin inhibitor from squash seeds (CMTI-I) in aqueous solution is obtained by the SCCE calculation using the ''divide-and-conquer'' scheme. The positions of localized orbitals, both top occupied and bottom unoccupied, are determined. This gives an explanation of reactive sites of inhibitor, and is in agreement with the experimental results. The calculation offers an independent theoretical testimony to the precision of structure determination. It is shown that the coordinates of oxygen atoms in inhibitor, determined by nuclear magnetic resonance and combination of distance geometry and dynamical simulated annealing, are systematically less accurate than that of other kinds of heavy atoms. The work demonstrates that the electronic structure calculations of a biological macromolecule is meaningful and has become reality with the SCCE method. It is hoped that the progress will shed some new light on quantum biology.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 39770189), and by the National High Performance Computing Center in Beijing and Shanghai, China. The calculations were performed on a Dawning2000-I, which is a cluster of Unix workstations developed by National Center of Intelligent Computer, China.

[1] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991); W. Yang and T.-S. Lee, J. Chem. Phys. **103**, 5674 (1995).
[2] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[3] F. Mauri *et al.*, Phys. Rev. B **47**, 9973 (1993).
[4] X.P. Li *et al.*, Phys. Rev. B **47**, 10 891 (1993).
[5] P. Ordejon *et al.* Phys. Rev. B **51**, 1456 (1995).
[6] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
[7] D. York, J. Lu, and W. Yang, Phys. Rev. B **49**, 8526 (1994).
[8] Haoping Zheng, Phys. Lett. A **226**, 223 (1997); **231**, 453 (1997).
[9] Haoping Zheng, Phys. Rev. B **48**, 14 868 (1993).
[10] Haoping Zheng, Physica B **212**, 125 (1995).
[11] H. Zheng, B.K. Rao, S.N. Khanna, and P. Jena, Phys. Rev. B **55**, 4174 (1997).
[12] H. Zheng, Y. Wang, and G. Ma (unpublished).
[13] R.C. Hider, A.F. Drake, I.E.G. Morrison, G. Kupryszewski, and T. Wilusz, Int. J. Pept. Protein Res. **30**, 397 (1987).
[14] M. Wieczorek *et al.*, Biochem. Biophys. Res. Commun. **126**, 646 (1985).
[15] T. Wilusz *et al.*, Hoppe Seyler's Z. Physiol. Chem. **364**, 93 (1983).
[16] Y. Hojima, J.V. Pierce, and J.J. Pisano, Biochemistry **21**, 3741 (1982).
[17] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[18] W. Kohn and L.J. Sham, Phys. Rev. **140**, A1133 (1965).
[19] T.A. Wesolowski and A. Warshel, J. Phys. Chem. **97**, 8050 (1993).
[20] T.A. Wesolowski, J. Chem. Phys. **106**, 8516 (1997).
[21] T.A. Holak, D. Gondol, J. Otlewski, and T. Wilusz, J. Mol. Biol. **210**, 635 (1989).
[22] T.A. Holak, W. Bode, R. Huber, J. Otlewski, and T. Wilusz, J. Mol. Biol. **210**, 649 (1989).
[23] F.B. van Duijneveldt, IBM J. Res. Dev. **945**, 16437 (1971).
[24] G.L. Lie and E. Clementi, J. Chem. Phys. **60**, 1275 (1974).
[25] R.A. Poirier, R. Daudel, P.G. Mezey, and I.G. Csizmadia, Int. J. Quantum Chem. **21**, 799 (1982).
[26] S. Huzinaga, J. Chem. Phys. **42**, 1293 (1965).
[27] R. Poirier, R. Kari, and I.G. Csizmadia, *Handbook of Gaussian Basis Sets* (Elsevier, New York, 1985).
[28] H. Sambe and R.H. Felton, J. Chem. Phys. **62**, 1122 (1975).