Journal of Mathematical Chemistry 14 (1993) 141–152

The application of the extended negative factor counting method to deal with the electronic energy levels of pig insulin

Yuan-Jie Ye and János Ladik

Chair for Theoretical Chemistry and Laboratory of the National Foundation for Cancer Research, Friedrich-Alexander University Erlangen-Nürnberg, Egerlandstrasse 3, D-8520 Erlangen, Germany

Received 22 July 1992

The quantum-chemical calculation for pig insulin was done by the ENFC method in which the matrix elements were calculated in the ab initio level with the help of a minimal basis. The aqueous solution surrounding insulin was simulated by putting point charges around the residues that have electric charges. The electronic density of states (DOS) of insulin confirmed the conclusions obtained from aperiodic model peptide chains. It is shown that the frontier orbitals of the insulin were mainly localized on those residues which are involved in the expression of biological activity of insulin.

1. Introduction

Although the relationship between the structure of insulin and its activity has been the subject of investigations for decades [1–7], many details of this relation remain unclear. In previous works, many scientists have investigated the relationship between the three-dimensional conformation and biological activities of insulin molecule by standard approaches in terms of hydrogen bonds, hydrophobicity, and van der Waals forces. In early studies, it was found that the region of insulin which forms a dimer in its crystal is important in the expression of its activity [1,2]. In the review of Gammeltoft [3] the receptor binding area of insulin was considered as a hydrophobic region surrounded by a hydrophilic one. This area was extended by Liang et al. [5] to include more residues which were involved in the expression of insulin activity. As the experiments on insulins were developed, more residues were found to be involved in the expression of its activity [6,7]. All these facts indicated that the whole insulin molecule interacts with its receptor. That is, the bind-

ing to its receptor is a complicated process in which many residues of the insulin are involved. The electronic structure of insulin should be investigated to obtain more knowledge about its biological activity.

Recently, the electronic structure of proteins was investigated using periodic and aperiodic polypeptide models [8–14]. It was found that the activity of proteins and DNA may be explained partially by their electric conductivity. Bakhshi et al. [14] worked out a seven-component polypeptide to simulate a real protein. The results of this calculation should be compared to those obtained for a native protein to verify their validity.

Ye [15] has proven that the extended negative factor counting (ENFC) theorem can solve the eigenvalue problem of tridiagonal block matrices with elements corresponding to cross-links, which may be derived for the quantum-chemical calculation on a native protein molecule.

In this paper, pig insulin, one of the smallest native protein molecules, was calculated using the ab initio SCF LCAO method applying a minimal basis set for the ENFC method. The environment of an aqueous solution was simulated by putting point charges around the residues that have electric charges. The results show that there are good agreements between the active sites of insulins and the positions where the frontier orbitals are localized. Further the seven-component polypeptide promises a good approximation to model the electronic structure of real proteins.

2. Approximations

The primary sequence of pig insulin is shown in fig. 1. The geometry of the insulin [16] used in the calculation was taken from the Brookheaven Protein Data Bank. There are only non-hydrogen atoms in the original data set. All of the coordinates of the hydrogen atoms were obtained theoretically. There are two peptide chains, three disulphur bridges (cross-links), 51 amino acids, and 782 atoms including hydrogen atoms, in pig insulin. The total number of the basis functions is 2418 when a minimal basis set is used. It is impossible to calculate the whole molecule in a self-consistent scheme. From Gazdy et al. [17] it is known that for the qualitative analysis of the level distribution (DOS) the local self-consistent method is valid. That is, the Fock matrix of the whole molecule can be constructed with the help of the Fock matrices of clusters corresponding only to first neighbours' interactions. In this paper the following approximations were applied:

(i) We neglect all interactions and overlaps between two amino acid residues which are not linked by any chemical bond. In this approximation the Fock (and overlap) matrix $\mathbf{F}$ (and $\mathbf{S}$) has the following form:

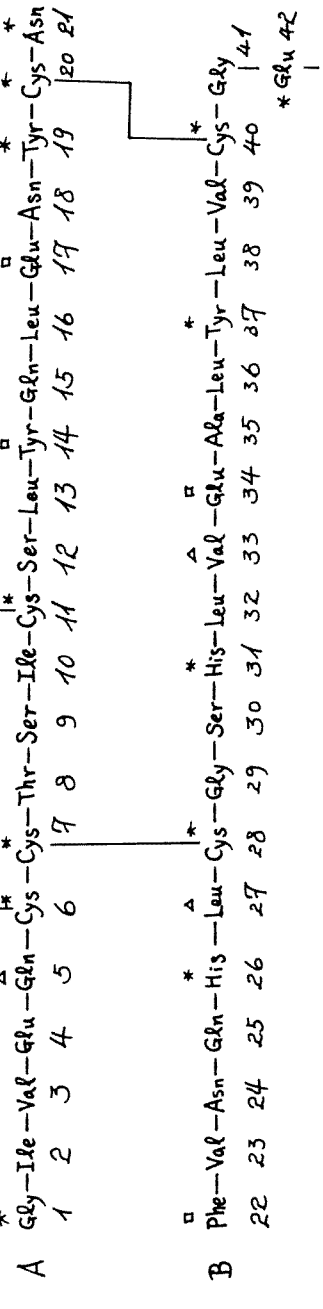

Fig. 1. The primary sequence of insulin and the relationship between its frontier orbitals and active sites. (*) Active sites with HOMOs and LUMOs; (△) active sites without HOMOs and LUMOs; (□) uncertain sites with HOMOs or LUMOs.

$$\begin{pmatrix} \mathbf{F}(1,1) & \mathbf{F}(1,2) & \mathbf{0} & & & & \mathbf{0} \\ \mathbf{F}(2,1) & \mathbf{F}(2,2) & \mathbf{F}(2,3) & & & & \\ \mathbf{0} & & \mathbf{F}(6,11) & & & & \\ & & & \mathbf{F}(7,28) & & & \\ & \mathbf{F}(11,6) & & & & & \\ & & & & \mathbf{F}(20,40) & & \\ & & \mathbf{F}(28,7) & & & & \\ & & & \mathbf{F}(40,20) & & \mathbf{0} & \\ & & & & & \mathbf{F}(50,51) & \\ \mathbf{0} & & & & \mathbf{0} & \mathbf{F}(51,50) & \mathbf{F}(51,51) \end{pmatrix},$$

in which $\mathbf{F}(n, n')$ is the matrix block for a residue when n equals n' and between residues when n does not equal n'. The numbering is counted from the beginning of chain A so that the numbering of the first residue of chain B will be 22. $\mathbf{F}(21, 22)$ and $\mathbf{F}(22, 21)$ are equal to zero because there is no chemical bond between these two residues. The corresponding generalized eigenvalue equation can be solved by the ENFC method [15].

(ii) The whole molecular system is divided into clusters of dimers that consist of two amino acid residues. At the ends of each dimer pseudo-atoms are added to simulate the chemical environment of the nearest neighbouring units, that is, an aldehyde group and an amido group are added at the N-terminal and the C-terminal of the dimer, respectively, and a methyl group and a sulphur atom are added to obtain a complete disulphur bridge and its chemical environment if the disulphur bridge is broken in the formation of the dimer (see fig. 2).

These clusters are calculated in self-consistent way to obtain the matrix elements of the whole molecular system.

(iii) To simulate the solution, point charges were put around the residues that have electric charges (see fig. 3).

It should be noted that the simulation of the environment is necessary for the quantum-chemical calculation on insulin because the conformation used originates from X-ray diffraction experiments on its crystal, which contains many water molecules. Therefore it is rather probable that the conformation of crystalline insulin is close to that in solution [18].

(iv) All the clusters were computed and the Fock (and overlap) matrix $\mathbf{F}$ (and $\mathbf{S}$) of the entire molecular system were constructed using the following formulae:

$$F_{ij}^{(\mathrm{M})}(n,n) = \frac{1}{m}\sum_{k=1}^{m} F_{ij}^{(\mathrm{c})}(n,n,k), \quad m = 1,2,3; \tag{2}$$

$$F_{ij}^{(\mathrm{M})}(n,n') = F_{ij}^{(\mathrm{c})}(n,n'), \quad n \neq n', \tag{3}$$

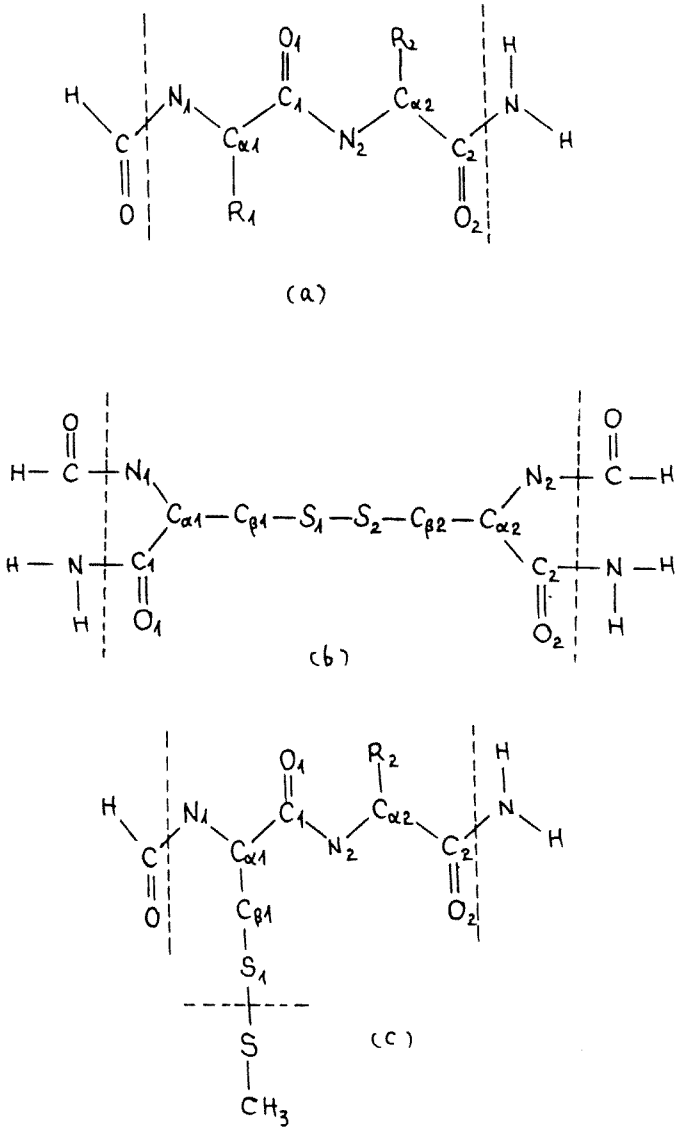

Fig. 2. The different kinds of clusters of dimers with pseudo-atoms. (a) is without cross link, (b) has a full disulfur bridge in the dimer, (c) has a broken disulfur bridge in the dimer.

where $F_{ij}^{(M)}(n, n')$ are the matrix elements of blocks of the Fock matrix of the whole molecule while $F_{ij}^{(c)}(n, n, k)$ and $F_{ij}^{(c)}(n, n')$ are those of clusters. The n and n' indicate the numberings of the residues and k indicates the clusters in which the nth residue is contained. That is, the elements of the off-diagonal blocks of the matrices are taken directly from those of the clusters, while the diagonal blocks are taken as the averages of the blocks which are corresponding to different neighbouring clusters of the dimer systems.

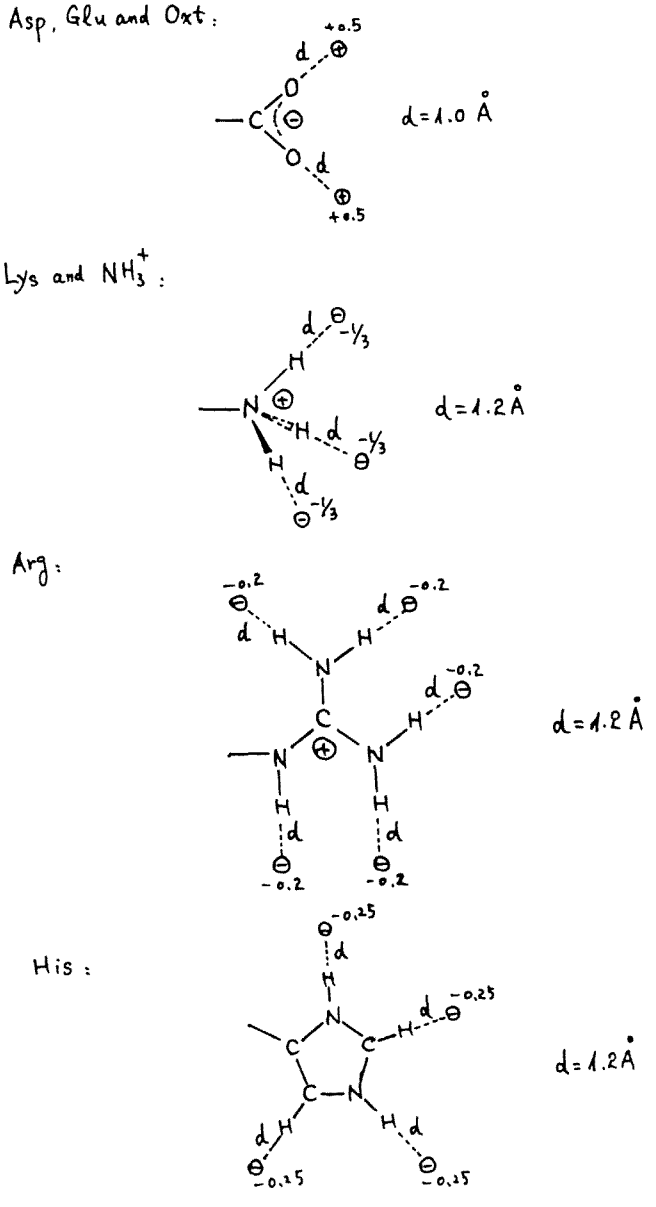

Fig. 3. The simulation of the environment of aqueous solution by putting point charges around the residues that have electronic charges. The distances of the point charges from a given residue are chosen in a way that they should interact with others in a small amount.

3. Results and discussion

In this calculation 52 dimers were computed to obtain the eigenvalue equation of the whole molecule. The density of states (DOS) of the molecule and 100 frontier orbitals (50 HOMOs and 50 LUMOs) were determined using the extended negative factor counting [15] and the inverse iteration methods [19], respectively.

3.1. DENSITY OF STATES

The electronic DOS of pig insulin is shown in fig. 4, in which (a) is for the valence and conduction band regions applying a grid of 1.0 eV, while (b) and (c) are for parts of the valence and conduction band regions using a grid of 0.05 eV, respectively.

From fig. 4 one can learn that the DOS (in number of states/eV ·2 spin) of insulin is similar to those of the aperiodic seven-component polypeptide chains [14]. The differences are that the distribution of the states is more even in insulin that in the model protein, and the energy gap of insulin (11.59 eV) is in a small amount smaller than that of the model protein (11.7 eV). This confirms the conclusion that aperiodicity depresses the gap of proteins [8–14].

From fig. 4 one can also see that in the DOS of the valence band there is a peak which corresponds to the orbitals delocalized on the peptide backbone. It is similar to that in polyacetylene which has delocalized π orbitals.

3.2. THE FRONTIER ORBITALS AND ACTIVE SITES

A part of the frontier orbitals of the insulin is shown in tables 1 and 2. From these tables one can see that the frontier orbitals of insulin are mainly localized on the aromatic side chains of the residues as well as on the cross-linking disulphur bonds. The carboxyl groups often have MOs with high lying filled energy levels. Most of the frontier orbitals are localized on the positions that are involved in the expression of the activity of insulin. From the first 18 HOMOs 13 correspond to active sites while among the first 22 LUMOs there are 16 such orbitals. The comparison between the active sites and the sites of localization of the frontier orbitals is shown in fig. 1.

From fig. 1 one finds that there is good agreement between the active sites of insulin and the sites of localization of the frontier orbitals. There are altogether 16 active sites of insulin with HOMOs or LUMOs while on four active sites no HOMOs and LUMOs are localized. There are further five residues on which frontier orbitals are localized but their role is unknown. All these imply that there should be an intrinsic relationship between the electronic structure and the activity of insulins.

Tables 1 and 2 confirm that in a real protein electron hopping happens between the localized orbitals.

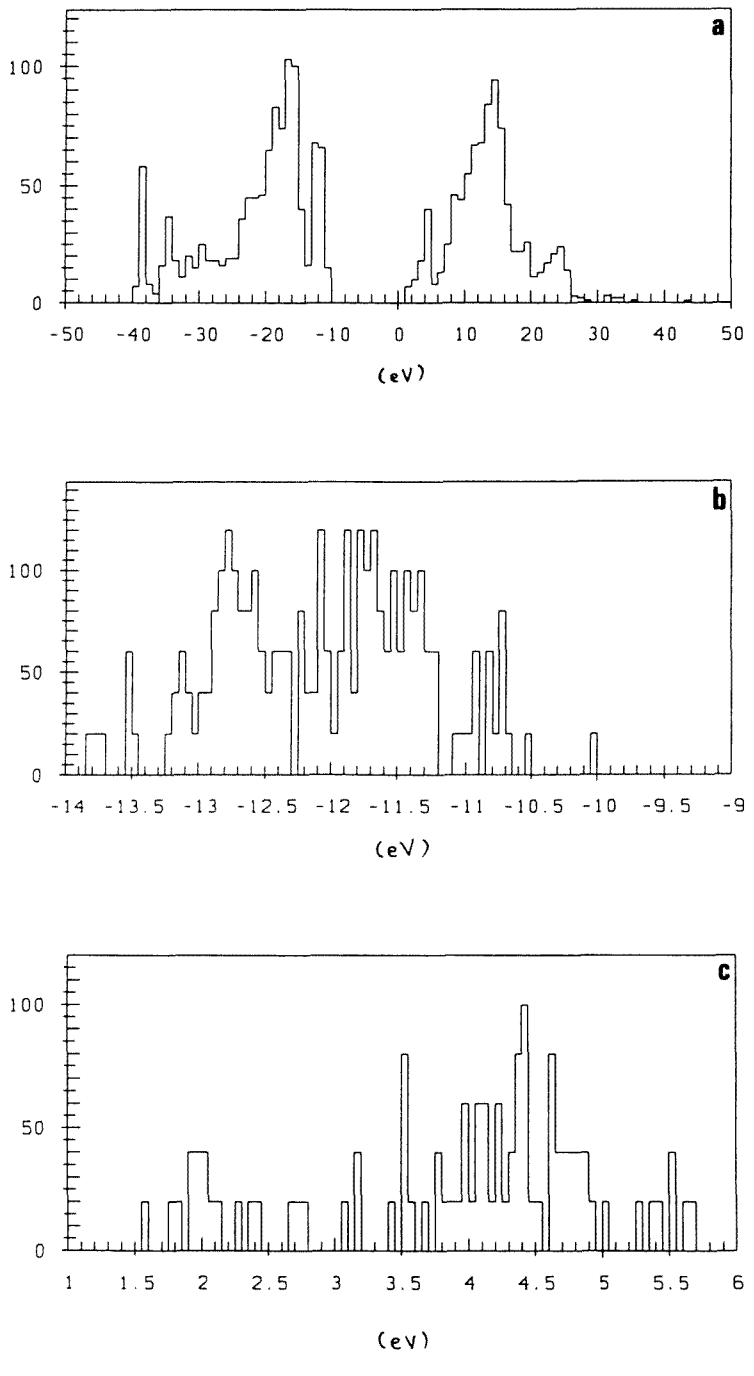

Fig. 4. The electronic DOS of insulin. (a) Both valence and conduction band regions using a grid of 1.0 eV. (b) The valence band region using a grid of 0.05 eV. (c) Conduction band region using a grid of 0.05 eV.

Table 1
The highest occupied molecular orbitals (HOMOs) of pig insulin (mol. 1).

		Energy level (eV)	Position of the wavefunction (residues)	
1538	(1)	−10.01001	Glu^{42}	*
1537	(2)	−10.51456	Lys^{50}–$\underline{Ala}^{51}$	
1536	(3)	−10.65218	Tyr^{37}	*
1535	(4)	−10.71317	Glu^{42}	*
1534	(5)	−10.71323	Ala^{51}	
1533	(6)	−10.72232	Tyr^{19}	*
1532	(7)	−10.72324	Glu^{17}	
1531	(8)	−10.76335	Cys^{6}–Cys^{11}	*
1530	(9)	−10.81308	Tyr^{47}	*
1529	(10)	−10.82336	Asn^{21}	*
1528	(11)	−10.82830	Cys^{7}–Cys^{28}	*
1527	(12)	−10.91033	Glu^{34}	
1536	(13)	−10.91965	Cys^{20}–$\underline{Asn}^{21}$	*
1525	(14)	−10.92421	His^{26}	*
1524	(15)	−10.96535	Tyr^{14}	
1523	(16)	−11.00171	Cys^{20}–Cys^{40}	*
1522	(17)	−11.06194	His^{31}	*
1521	(18)	−11.20095	Phe^{45}–$\underline{Phe}^{46}$–Tyr^{47}	*
1520	(19)	−11.20270	Val^{33}–Glu^{34}–Ala^{35}	p.b.
1519	(20)	−11.23413	–Glu^{17}–Asn^{18}–Tyr^{19}–	p.b.
1518	(21)	−11.25313	–Val^{3}–$\underline{Glu}^{4}$–Gln^{5}	p.b.
1517	(22)	−11.26617	Cys^{7}–Thr^{8}–Ser^{9}–Ile^{10}–	p.b.
1516	(23)	−11.28928	Leu^{32}–Val^{33}–Glu^{34}–Ala^{35}–Leu^{36}	p.b.
1515	(24)	−11.30259	–Leu^{16}–Glu^{17}–Asn^{18}–Tyr^{19}	p.b.

Notes:
(1) The definition of a component at the nth residue is $a(n) = \sum_{j=1}^{m_n} C_j^2(n) / \sum_{n=1}^{N} \sum_{j=1}^{m_n} C_j^2(n)$.
(2) Those residues that have $a(n) < 0.05$ are neglected.
(3) If the wavefunction belonging to a certain level is localized on several units, its main component belongs to the underlined residue. The definition of main component is $a(n) \geqslant 0.60$.
(4) *: This wavefunction corresponds to an active site.
(5) p.b.: This wavefunction is mainly distributed at the peptide bonds between residues.

Table 2
The lowest unoccupied molecular orbitals (LUMOs) of pig insulin (mol. 1).

		Energy level (eV)	Position of the wavefunction (residues)	
1539	(1′)	1.57819	Tyr^{14}	
1540	(2′)	1.79298	Tyr^{19}*	
1541	(3′)	1.82391	Tyr^{47}	*
1542	(4′)	1.92964	Phe^{22}	
1543	(5′)	1.93132	Phe^{46}	*
1544	(6′)	1.95364	Tyr^{37}	*
1545	(7′)	1.96569	Phe^{45}	*
1546	(8′)	2.00668	Phe^{22}	
1547	(9′)	2.02951	Phe^{45}	*
1548	(10′)	2.05106	Phe^{46}	*
1549	(11′)	2.10072	Tyr^{14}	
1550	(12′)	2.29470	Tyr^{47}	*
1551	(13′)	2.38962	Tyr^{19}	*
1552	(14′)	2.41894	Tyr^{37}	*
1553	(15′)	2.68373	Cys^{20}–Cys^{40}	*
1554	(16′)	2.71359	Cys^{7}–Cys^{28}	*
1555	(17′)	2.78701	Cys^{6}–Cys^{11}	*
1556	(18′)	3.06516	Gln^{15}–Leu^{16}–Glu^{17}–Asn^{18}–Tyr^{19}	p.b.
1557	(19′)	3.18243	$\underline{Gly^{1}}$–Ile^{2}	*
1558	(20′)	3.19348	$\underline{Gly^{29}}$–...–$\underline{His^{31}}$	*
1559	(21′)	3.43597	Glu^{42}–Arg^{43}–Glu^{44}	p.b.
1560	(22′)	3.51867	His^{26}	*
1561	(23′)	3.52617	Ser^{12}–Leu^{13}–Tyr^{14}–Gln^{15}–... –Glu^{17}–Asn^{18}–Tyr^{19}–... –Gly^{29}–Ser^{30}–His^{31}–	p.b.
1562	(24′)	3.52700	Ser^{12}–Leu^{13}–Tyr^{14}–Gln^{15}–... –Glu^{17}–Asn^{18}–Tyr^{19}–... –Gly^{29}–Ser^{30}–His^{31}–	p.b.

Notes:
(1) The definition of a component at the nth residue is $a(n) = \sum_{j=1}^{m_n} C_j^2(n) / \sum_{n=1}^{N} \sum_{j=1}^{m_n} C_j^2(n)$.
(2) Those residues that have $a(n) < 0.05$ are neglected.
(3) If the wavefunction belonging to a certain level is localized on several units, its main component belongs to the underlined residue. The definition of main component is $a(n) \geqslant 0.60$.
(4) *: This wavefunction corresponds to an active site.
(5) p.b.: This wavefunction is mainly distributed at the peptide bonds between residues.

4. Conclusions

In this paper the investigation of the electronic structures of pig insulin, one of the smallest native proteins, is reported. The calculation is based on the ab initio matrix block negative factor counting method with cross-links in a chain or between chains. The energy level distribution shows that the native proteins have a smaller energy gap and more even distribution of the energy levels than the model polypeptides investigated until now. This implies the native proteins may have a somewhat larger hopping conductivity than model polypeptides. The analysis of the frontier orbitals of insulin shows that there is a good agreement between the sites of localization of the frontier orbitals and the residues which play an important role in the activities of insulin. Therefore, it could be concluded that the HOMOs and LUMOs of a native protein influence its biological activity. However, it is necessary that more non-bonding interactions should be included in the quantum-chemical calculation to investigate their influences on the electronic structure of a native protein. One should point out also that more examples should be calculated for the further verification of the relationship between the role of frontier orbitals and the biological activities of proteins.

Acknowledgement

The authors would like to express their gratitute to Professor P. Otto and to Privat Dozent Dr. C.-M. Liegener for the helpful discussions.

References

[1] T. Blundell, G. Dodson and D. Hodgkin, Adv. Prot. Chem. 26 (1972) 279.
[2] D. Brandenberg and Z. Wollmer, *Insulin, Chemistry, Structure and Function of Insulin and Related Hormones* (De Gruyter, Berlin, 1980).
[3] S. Gammeltoft, Physiol. Rev. 64 (1984) 1321.
[4] E.N. Baker, T.L. Blundell, J.F. Cutfield, S.M. Cutfield, E.J. Dodson, G.G. Dodson, D.M.C. Hodgkin, R.E. Hubbard, N.W. Isaacs, C.D. Reynolds, K. Sakabe, N. Sakabe and N.M. Vijayan, Phil. Trans. Roy. Soc. 319 (1988) 369.
[5] D.C. Liang, Z.L. Wan, W.R. Chang and J.P. Zhang, XVth Congress and General Assembly of International Union of Crystallography, France 1990, C-101.
[6] G.P. Schwartz, G.T. Burke and P.G. Katsoyannis, Proc. Nat. Acad. Sci. US 84 (1987) 6408; 86 (1989) 458.
[7] S.H. Nakagawa and H.S. Tager, J. Biol. Chem. 266 (1991) 11502.
[8] J. Ladik, Int. J. Quant. Chem. 23 (1983) 1073.
[9] J. Ladik, M. Seel, P. Otto and A.K. Bakhshi, Chem. Phys. 108 (1986) 203.
[10] A.K. Bakhshi, P. Otto, J. Ladik and M. Seel, Chem. Phys. 108 (1986) 215.
[11] P. Otto, A.K. Bakhshi, J. Ladik, M. Seel and S. Chin, Chem. Phys. 108 (1986) 223.
[12] A.K. Bakhshi, J. Ladik, M. Seel and P. Otto, Chem. Phys. 108 (1986) 233.
[13] C.-M. Liegener, A. Sutjianto and J. Ladik, Chem. Phys. 145 (1990) 385.

[14] A.K. Bakhshi, P. Otto, C.-M. Liegener, E. Rehm and J. Ladik, Int. J. Quant. Chem. 38 (1990) 573.
[15] Y.-J. Ye, J. Math. Chem. 14 (1993) 121.
[16] J. Bordas, G.G. Dodson, H. Grewe, M.H.J. Koch, B. Krebs and J. Randall, Proc. Roy. Soc. B 219 (1983) 21.
[17] B. Gazdy, M. Seel and J. Ladik, Chem. Phys. 86 (1984) 41.
[18] P. Kruger, W. Strassburger, A. Wollmer, W.F. van Gunsteren and G.G. Dodson, Eur. Biophys. J. 14 (1987) 449.
[19] J.H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon Press, Oxford, 1965).