Electronic structure of model protein chains: Effect of basis set and secondary structure

Shyam Kishor
Department of Chemistry, J.Y. College, Baraut, Uttar Pradesh, India 250 611

and

A K Bakhshi*
Department of Chemistry, University of Delhi, Delhi, India 110 007
Email: akbakhshi2000@yahoo.com

Received 24 July 2003; revised 7 May 2004

*Ab initio band structure results of β-pleated polyglycine, polyalanine, and polyserine obtained using minimal basis set have been used as input to calculate the electronic density of states of different periodic and aperiodic model protein chains on the basis of negative factor counting method in tight binding approximation. The effects of the change of the basis set (double-zeta) and the secondary structure (α-helix) on the electronic density of states of these model protein chains have also been investigated.

The electronic structures of proteins are of profound interest because of their important biological functions. The quantum mechanical investigation of the electronic structure of proteins is a very challenging task due to the aperiodic arrangement of 20 amino acids in β-pleated or α-helical structures. Since it is not possible to perform direct SCF calculation, one has to employ approximate methods. Under these circumstances the electronic density of states (DOS) of a quasi one-dimensional (1D) multicomponent non-periodic protein chain can be determined using negative factor counting (NFC) technique based on Dean’s negative energy eigenvalue theorem.[1] In this paper the simple negative factor counting method has been used to obtain the electronic DOS of aperiodic and periodic model protein chains. The simple NFC method has already been shown to yield very good results for various copolymers.[2-4] In the simple NFC method one requires as an input the band structures of the periodically repeated components of the aperiodic chain under consideration. Using these band structures, a tridiagonal Hückel determinant (in the first neighbor’s interaction approximation) is constructed for a valence or a conduction band of the polymer chain. This determinant is then solved using Dean’s negative eigenvalue theorem. In the case of proteins, which consist of 20 different amino acid residues, one therefore needs to determine the band structure results of all the 20 amino acids.

In the present calculations, the α and β values for the valence and the conduction bands of polyglycine, polyalanine and polyserine in β-pleated conformation have been obtained from their band structure results obtained on the basis of the ab initio Hartree Fock crystal orbital method.[5] The geometry of the amino acid residues representing the elementary unit cell has been built up using standard bond lengths and bond angles. The secondary structure of the elementary unit cell was taken to be an antiparallel β-pleated sheet structure. To see the effect of the basis set on the electronic DOS, the DOS have been determined using band structure results[6] of polyglycine, polyalanine and polyserine using Clementis double zeta basis set along with minimal basis set. To check the dependence of DOS on the conformation of protein, DOS have been determined using band structure results[3] of polyglycine, polyalanine and polyserine for α-helix conformation using minimal basis. In the case of α-helix two peptide units form the elementary cell. The combined symmetry operation (rotation about and translation along the helix axis) is applied to generate the complete helix.

Methods of Investigations

(A) Band structure calculations
The band structure calculations of various homopolypeptides were performed on the basis of the
The geometry of the amino acid residues representing the elementary unit cell was built up using standard bond lengths and bond angles. The conformation of the elementary unit cell was taken to be the antiparallel β-pleated sheet structure.

The ab initio Hartree-Fock crystal orbital program solves the pseudo eigen value problem:

\[ F(K)C_n(K) = e_n(K)S(K)C_n(K) \]

for different K values in the Brillouin zone. F(K) and S(K) are the Fock and overlap matrices in the K representation. C_n(K) is the eigen vector and the index n denotes the band. The eigen values \( e_n(K) \) give the band structure of the polymer. Note that matrices F(K) and S(K) are complex Hermitian, \( C_n(K) \) is complex while \( e_n(K) \) is real. Since this method has been formulated and discussed many times in the literature, it is not necessary to give further details here.

In view of the large unit cells of these polymers (23 basis function for glycine, 30 for alanine and 35 for serine), all the computations were carried out using element's 7s/3p minimal basis set for the heavy atoms and four primitive Gaussians contracted to one as function for the hydrogen atoms. All the multicentre two-electron integrals larger than the threshold value of 10\(^{-8}\) a.u. were calculated and interactions up to second neighbour were taken into account.

(B) Electronic density of states

The electronic density of states of the model protein chains can be obtained using the NFC method based on Dean's negative eigenvalue theorem. This method is related to the Givens-Householder Wilkinson method for matrix diagonalization. It is based on Sturm's theorem and determines the eigenvalues of a matrix as the changes of sign of its secular determinant as a function of \( \lambda \). Evaluation of the secular determinant for a given value of \( \lambda \) is efficiently done by the Gaussian algorithm.

In this method, to obtain the DOS curves of the electronic states of a quasi 1-D chain, one writes down a Hückel determinant of the chain consisting of \( N \) units as follows:

\[
\begin{vmatrix}
\alpha_1 - \lambda & \beta_{12} & 0 & \cdots & 0 \\
\beta_{12} & \alpha_2 - \lambda & \beta_{23} & \cdots & 0 \\
0 & \beta_{23} & \alpha_3 - \lambda & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \beta_{N-1,N} \\
\end{vmatrix} = 0
\]

(1)

In this, the diagonal Hückel parameter \( \alpha_i \) is determined on the basis of the middle point of the band under consideration of each component. These bands are obtained from the ab initio Hartree-Fock crystal orbital calculations for each periodically repeated component applying periodic boundary conditions. For the off diagonal (\( \beta \)) elements assuming the simple

\[
\varepsilon_A(k) = \alpha_A + 2\beta_{AA} \cos(ka)
\]

(2)

relation for the energy dispersion, one fourth of the calculated band width is taken if the same unit is repeated and as simplest approximation \( \beta_{AB} = \frac{1}{2} (\beta_{AA} + \beta_{BB}) \) is applied, if a unit A is followed by a unit B. If one could calculate all the roots of (1) one could write

\[
|H(\lambda)| = \prod_{i=1}^{N} (\lambda_i - \lambda_i)
\]

(3)

This is, however, not possible in the case of a long chain (\( N = 10^3 - 10^5 \)). Therefore, instead of trying to find the roots of (1), we can bring it into a di-diagonal form by applying a successive gaussian elimination, i.e., in this case

\[
|H(\lambda)| = \prod_{i=1}^{N} \varepsilon_i(\lambda)
\]

(4)

where the factors \( \varepsilon_i(\lambda) \) are given by the simple recursion formula:

\[
\varepsilon_i(\lambda) = (\alpha_i - \lambda) - \beta_{i-1,i}^2 / \varepsilon_{i-1}(\lambda), \quad i = 2,3,...,N
\]

(5)

\[
\varepsilon_1(\lambda) = \alpha_1 - \lambda
\]

(6)

Since expressions (3) and (4) are equal, for any trial value of \( \lambda \) the number of eigenvalues \( \lambda_i \) smaller than \( \lambda \) has to be equal to the number of negative factors \( \varepsilon_i(\lambda) \) in (5). By giving different \( \lambda \) values throughout the whole spectrum and taking the differences of the number of negative \( \varepsilon_i(\lambda) \) values belonging to consecutive values of \( \lambda \) in the chosen grid, the distribution of eigenvalues of \( H(\lambda) \) can be
obtained to any desired accuracy. A chain length of 300 units and an energy grid of 0.05 eV have been used in our calculations.

**Results and Discussion**

**Electronic structure of protein**

In the present calculations the α and β values for the valence and the conduction bands (Table 1) of polyglycine, polyalanine and polyserine obtained from their band structure results using minimal basis set and double-zeta basis set by the ab initio Hartree Fock crystal orbital method have been used. This method is known to correctly reproduce the trends in the electronic properties.

The DOS curves of the valence band and conduction band regions of model periodic and aperiodic protein chains are shown in Figs 1-3. In the case of aperiodic poly (GAS), the random sequence of the units was generated using a random number generator programme.

The important electronic properties such as ionization potential (I.P.) (corresponding to the negative of the top of the valence band), electron affinity (E.A.) (corresponding to the bottom of the conduction band), band gap (Eg) obtained using simple NFC method are given in Table 2.

As can be seen from the DOS curves, in the case of periodic poly (GAS), both valence and conduction band regions consist of well separated, very narrow peaks of almost equal intensity (Figs 1-3). On the other hand, in aperiodic poly (GAS), both valence and conduction band regions still have many gaps but the peaks become much broader (Figs 1-3). As compared to periodic poly (GAS), in the case of aperiodic poly (GAS), the upper limit of the valence band region is unchanged and the lower limit of the conduction band region is shifted downwards in energy. The result is a slight decrease in the band gap in aperiodic chain as compared to the periodic chain. The appearance of relatively broader peaks in the case of aperiodic system as compared to those in the periodic system is not surprising. Similar broadening has been observed in the case of disordered single stranded DNA base stacks and random copolymers of heterocyclic compounds. In periodic poly (GAS), because of the large size of the unit cell (G-A-S) (three times as large as the unit cell in homopolymides), the Brillioun zone is correspondingly reduced and because of the three different types of units in the unit cell, its periodicity is broken leading to the three different peaks. Alternatively, the occurrence of these very narrow peaks may be explained on the basis of the fact that in a DOS curve of, say the valence band

---

**Table 1—Ab initio band structure results (in eV) of the homopoly-peptides**

<table>
<thead>
<tr>
<th>Basis set</th>
<th>Secondary structure</th>
<th>Poly-peptide of</th>
<th>Band</th>
<th>E_{min}</th>
<th>E_{max}</th>
<th>ΔE</th>
<th>E_{gap}</th>
<th>α_{E}</th>
<th>α_ΔE</th>
<th>p_ΔE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clementi's minimal basis</td>
<td>β-pleated Glycine</td>
<td>C</td>
<td>3.591</td>
<td>4.166</td>
<td>0.575</td>
<td>15.777</td>
<td>3.878</td>
<td>0.144</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-12.579</td>
<td>-12.186</td>
<td>0.393</td>
<td>-12.385</td>
<td>-12.058</td>
<td>0.358</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clementi's double zeta</td>
<td>β-pleated Alanine</td>
<td>C</td>
<td>3.722</td>
<td>4.395</td>
<td>0.673</td>
<td>15.776</td>
<td>12.171</td>
<td>0.193</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>V</td>
<td>-12.288</td>
<td>-12.054</td>
<td>0.234</td>
<td>-12.385</td>
<td>-12.058</td>
<td>0.358</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clementi's minimal basis</td>
<td>α-helical Glycine</td>
<td>C</td>
<td>4.265</td>
<td>4.996</td>
<td>0.731</td>
<td>15.366</td>
<td>4.630</td>
<td>0.183</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>V</td>
<td>-11.176</td>
<td>-11.101</td>
<td>0.075</td>
<td>15.366</td>
<td>4.630</td>
<td>0.183</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serine</td>
<td>C</td>
<td>3.197</td>
<td>3.712</td>
<td>0.515</td>
<td>14.700</td>
<td>3.454</td>
<td>0.129</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td>-11.158</td>
<td>-11.503</td>
<td>0.655</td>
<td>14.700</td>
<td>3.454</td>
<td>0.129</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clementi's minimal basis</td>
<td>α-helical Alanine</td>
<td>C</td>
<td>3.042</td>
<td>3.908</td>
<td>0.866</td>
<td>14.322</td>
<td>3.478</td>
<td>0.216</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td>-11.747</td>
<td>-11.280</td>
<td>0.467</td>
<td>14.322</td>
<td>3.478</td>
<td>0.216</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serine</td>
<td>C</td>
<td>2.938</td>
<td>3.094</td>
<td>0.156</td>
<td>13.530</td>
<td>3.016</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td>-10.767</td>
<td>-10.592</td>
<td>0.175</td>
<td>13.530</td>
<td>3.016</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clementi's minimal basis</td>
<td>α-helical Serine</td>
<td>C</td>
<td>3.461</td>
<td>3.934</td>
<td>0.518</td>
<td>15.363</td>
<td>3.675</td>
<td>0.129</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td>-12.056</td>
<td>-11.947</td>
<td>0.109</td>
<td>15.363</td>
<td>3.675</td>
<td>0.129</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glycine</td>
<td>C</td>
<td>3.394</td>
<td>3.704</td>
<td>0.310</td>
<td>14.741</td>
<td>3.549</td>
<td>0.077</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td>-11.393</td>
<td>-11.347</td>
<td>0.046</td>
<td>14.741</td>
<td>3.549</td>
<td>0.077</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*C and V represent conduction and valence bands respectively.

Obtained from Ref 5.

Obtained from Ref 6.
Fig. 1—DOS curves of periodic and aperiodic polypeptides of 300 units in the β-pleated sheet conformations using minimal basis set (energy in eV, number in relative units). [G, A and S represent glycine, alanine and serine respectively].

region of a particular sequence, the energy position of the peak due to a unit, say A, is determined by two factors: (1) the position of the valence band (centre of the band) of poly (A) and (2) the electronic environment of A in the sequences. The shift in the energy position of the peak due to A in a given sequence relative to that in poly (A) shall depend upon how different the interaction of A is with its neighbours in a given sequence from its interaction with other neighbouring A’s in poly (A). In the case of periodic sequences (G-A-S) of poly (GAS), the respective environments of G, A and S remain in all positions the same. Therefore, for periodic poly (GAS) we obtain three very narrow peaks of almost equal intensity. On the other hand, in the case of a random aperiodic sequence of poly (GAS), the respective environments of A, G and S keep on changing. Therefore, their energy positions (peaks) are scattered over a much wider range of energy.

To see the effect of the basis set on the electronic DOS, the DOS have been determined using band structure results of polyglycine, polyalanine and polyserine using Clementi’s double zeta basis set. Their results are given in Table 2. Comparison of these results with those obtained using minimal basis set shows that though the fundamental band gap decreases with better basis set, the decrease is very small. Similar observation has been observed in the case of four nucleotide bases. This is because in a minimal (single zeta) basis set, only those orbitals,
which are occupied in the free atoms, are considered whereas in an extended basis set, each valence orbital is supplemented by a delocalized orbital of the same quantum number. This allows for more radial flexibility in the spatial representation of the molecular electron density and also partially corrects the deficient description of anisotropic situations by the readjustment of diffuse and localized components through the self-consistent field (SCF) procedures. If one also doubles the core orbitals, the basis set is often referred to as a double zeta basis. It needs to be pointed out here that the Hartree Fock method overestimates the band gap values due to the neglect of the correlation effects and the use of the minimal basis set. With the use of better basis sets and the consideration of correlation effects the calculated band gap values are expected to decrease and come closer to experimental values.

On the basis of the above results nothing definite can be said about the conduction properties of proteins because the gaps in the DOS curves of aperiodic poly (AGS) in its valence band region are in most cases too numerous and large to allow an effective charge transport, even if free carriers are present. To come to a more definite conclusion about the possibility of hopping conduction, one has to use better basis sets which should take into account the effect of the water and ion environment (which may
Fig. 3—DOS curves of periodic and aperiodic polypeptides of 300 units in the α-helix conformations using minimal basis set (energy in eV, number in relative units). [G, A and S represent glycine, alanine and serine respectively].

<table>
<thead>
<tr>
<th>Basis set</th>
<th>Secondary structure</th>
<th>Sequence</th>
<th>IP</th>
<th>EA</th>
<th>Eg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clementi's minimal basis</td>
<td>β-pleated</td>
<td>Periodic poly(G-A-S)</td>
<td>11.10</td>
<td>3.75</td>
<td>14.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-1 poly(G-A-S)</td>
<td>11.10</td>
<td>3.65</td>
<td>14.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-2 poly(G-A-S)</td>
<td>11.10</td>
<td>3.65</td>
<td>14.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Periodic poly(G-A-S)</td>
<td>10.65</td>
<td>2.95</td>
<td>13.60</td>
</tr>
<tr>
<td>Clementi's double zeta</td>
<td>β-pleated</td>
<td>Periodic poly(G-A-S)</td>
<td>10.60</td>
<td>2.90</td>
<td>13.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-1 poly(G-A-S)</td>
<td>10.60</td>
<td>2.85</td>
<td>13.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-2 poly(G-A-S)</td>
<td>11.35</td>
<td>2.70</td>
<td>14.05</td>
</tr>
<tr>
<td>Clementi's minimal basis</td>
<td>α-helical</td>
<td>Periodic poly(G-A-S)</td>
<td>11.35</td>
<td>2.60</td>
<td>13.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-1 poly(G-A-S)</td>
<td>11.35</td>
<td>2.60</td>
<td>13.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aperiodic-2 poly(G-A-S)</td>
<td>11.35</td>
<td>2.60</td>
<td>13.95</td>
</tr>
</tbody>
</table>
introduce extra levels in the gaps) and also take into account other factors like the interactions of polypeptides with DNA, etc. In our present study, only the interactions between the neighbouring amino acids are considered. It is quite likely that because of these additional interactions the DOS curves gets further modified leading to some further broadening of the peaks. Our present results should therefore be viewed as a first step towards the treatment of aperiodicity in real protein at the \textit{ab initio} level.

**References**