Electronic structure of single stranded DNA base stacks

Shyam Kishor
Department of Chemistry, J.V. College, Baraut, U.P. 250 611, India

and

A K Bakhshi
Department of Chemistry, University of Delhi, Delhi 110 007, India
E-mail: akbakhshi2000@yahoo.com

Received 21 August 2002; revised 21 May 2003

The electronic density of states (DOS) curves of single stranded periodic and aperiodic DNA base stacks in B-conformation obtained with the help of simple negative factor counting (NFC) method are presented. Band structure results of polyadenine [poly (A)]; polyguanine [poly (G)]; polycytocine [poly (C)] and polythymine [poly (T)] obtained on the basis of ab initio Hartree Fock crystal orbital method using both minimal basis and double zeta basis sets have been used as input to calculate these DOS. In contrast to the periodic systems, the peaks in the DOS curves of the corresponding aperiodic systems are broader with fewer gaps in between for both levels of approximation. The calculated large values of band gap obtained for these systems rule out the possibility of intrinsic conduction in them.

Biopolymers such as DNA or proteins play a key role in life processes. In order to interpret their physical and chemical properties and on the basis of their biological functions, one needs to have a fairly good approximation of their electronic structures. DNA-B is composed of two antiparallel right handed helix of polydeoxyribonucleotides which in DNA form a double helix. In each strand the four different nucleotide bases A, T, G and C form stacks and have a non-periodic sequence though the sugar phosphate backbones in the DNA helix are periodic. The non-periodicity of the base stacks makes the direct SCF calculations of their electronic structure impossible. Since the proposal of the above double helical structure of DNA by Watson and Crick in 1953, there have been numerous theoretical investigations aimed at the determination of the electronic structure of DNA. Earlier calculations of the band structures of periodic nucleotide base stacks were performed with the help of the crystal orbital (CO) method on the different semi-empirical levels. In recent years also several methods like ab initio matrix block negative factor counting (NFC) method, DV-Xα cluster calculations, and density-matrix divide and conquer approach have been used to investigate the electronic structures of periodic and aperiodic DNA and proteins. The major difficulty in the investigation of the electronic structure of DNA lies in the treatment of its aperiodicity. Since it is not possible to perform direct SCF calculation, one has to employ approximate methods.

In this paper the simple negative factor counting method has been used to obtain the DOS of aperiodic and periodic DNA base stacks. As compared to matrix block NFC method, this method is simple and less time-consuming. This is due to the fact that the simple NFC method assumes one orbital per unit cell in contrast to matrix block NFC method which can be applied to the case of arbitrary number of orbitals per unit cell. The simple NFC method has already been shown to yield very good results for various copolymers. As a first step, we have mainly concentrated on single stranded aperiodic DNA and have determined the DOS of periodic and aperiodic sequences of the four nucleotide bases present in DNA. In the present calculations, the α and β values for the valence and the conduction bands of four nucleotide base stacks have been obtained from their band structure results, using minimal basis and double zeta basis sets obtained by the ab-initio Hartree Fock crystal orbital method.

Methods of Investigations

The electronic density of states (DOS) of the single stranded periodic and aperiodic-DNA can be obtained using the simple negative factor counting (NFC) method based on Dean's negative eigen value 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 Huckel determinant of the chain consisting of $N$ units as:

$$|H(\lambda)| = \begin{vmatrix}
\alpha_i - \lambda & 0 & \ldots & 0 \\
0 & \alpha_2 - \lambda & 0 & \ldots & 0 \\
0 & 0 & \alpha_3 - \lambda & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 \\
0 & 0 & \ldots & 0 & \beta_{N-1,N} & \alpha_N - \lambda
\end{vmatrix} = 0$$

...(1)

The diagonal Huckel 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$^{20,21}$ 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:

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

...(3)

This is, however, not possible in the case of a long chain ($N = 10^2 - 10^4$). Therefore, instead of trying to find the roots of (1), one can bring it into a diagonal form by applying a successive gaussian elimination. In this case one can write:

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

...(4)

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

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

...(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.

Results and Discussion

Electronic structure of DNA

In the present calculations the $\alpha$ and $\beta$ values for the valence and the conduction bands (Tables 1 and 2) of four nucleotide bases 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$^4$. This method is

Table 1—Ab-initio band structure results (in eV) of the four nucleotide base stacks (Clementi's minimal basis)$^a$

<table>
<thead>
<tr>
<th>Poly</th>
<th>Band</th>
<th>$E_{\text{min}}$</th>
<th>$E_{\text{max}}$</th>
<th>$\Delta E$</th>
<th>$E_{\text{exp}}$</th>
<th>$\alpha_i = \frac{E_{\text{max}} + E_{\text{min}}}{2}$</th>
<th>$\beta_i = \frac{\Delta E}{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenine</td>
<td>C</td>
<td>2.309</td>
<td>2.413</td>
<td>0.104</td>
<td>12.275</td>
<td>2.361</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-10.106</td>
<td>-9.986</td>
<td>0.140</td>
<td>-10.036</td>
<td>-10.036</td>
<td>0.035</td>
</tr>
<tr>
<td>Cytosine</td>
<td>C</td>
<td>1.802</td>
<td>2.212</td>
<td>0.410</td>
<td>11.207</td>
<td>2.007</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-9.423</td>
<td>-9.405</td>
<td>0.018</td>
<td>-9.414</td>
<td>-9.414</td>
<td>0.004</td>
</tr>
<tr>
<td>Guanine</td>
<td>C</td>
<td>2.672</td>
<td>2.808</td>
<td>0.136</td>
<td>11.447</td>
<td>2.740</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-9.312</td>
<td>-8.775</td>
<td>0.537</td>
<td>-9.043</td>
<td>-9.043</td>
<td>0.134</td>
</tr>
<tr>
<td>Thymine</td>
<td>C</td>
<td>1.408</td>
<td>1.885</td>
<td>0.477</td>
<td>11.994</td>
<td>1.646</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-11.047</td>
<td>-10.586</td>
<td>0.461</td>
<td>-10.816</td>
<td>-10.816</td>
<td>0.115</td>
</tr>
</tbody>
</table>

$^a$ Obtained from ref. 4.

$^b$ C and V represent conduction and valence bands respectively.
Table 2—*Ab-initio* band structure results (in eV) of four nucleotide base stack's (Clementi’s double zeta basis)*

<table>
<thead>
<tr>
<th>Poly</th>
<th>Band</th>
<th>$E_{\text{min}}$</th>
<th>$E_{\text{max}}$</th>
<th>$\Delta E$</th>
<th>$E_{\text{pp}}$</th>
<th>$\alpha_i = \frac{E_{\text{min}} + E_{\text{max}}}{2}$</th>
<th>$\beta_i = \frac{\Delta E}{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytosine</td>
<td>C</td>
<td>2.652</td>
<td>3.200</td>
<td>0.548</td>
<td>11.448</td>
<td>2.926</td>
<td>0.137</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-8.917</td>
<td>-8.796</td>
<td>0.121</td>
<td>-8.856</td>
<td>-8.856</td>
<td>0.030</td>
</tr>
<tr>
<td>Thymine</td>
<td>C</td>
<td>2.397</td>
<td>2.594</td>
<td>0.197</td>
<td>11.790</td>
<td>2.495</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-10.115</td>
<td>-9.393</td>
<td>0.722</td>
<td>-9.754</td>
<td>-9.754</td>
<td>0.180</td>
</tr>
<tr>
<td>Adenine</td>
<td>C</td>
<td>3.458</td>
<td>3.583</td>
<td>0.125</td>
<td>11.930</td>
<td>3.520</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-8.691</td>
<td>-8.472</td>
<td>0.219</td>
<td>-8.581</td>
<td>-8.581</td>
<td>0.055</td>
</tr>
<tr>
<td>Guanine</td>
<td>C</td>
<td>2.354</td>
<td>3.949</td>
<td>1.595</td>
<td>10.113</td>
<td>3.151</td>
<td>0.399</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>-8.352</td>
<td>-7.759</td>
<td>0.593</td>
<td>-8.055</td>
<td>-8.055</td>
<td>0.148</td>
</tr>
</tbody>
</table>

*obtained from ref. 4.

*"C and V represent conduction and valence bands respectively.

**Fig. 1**—DOS curves of the single stranded periodic DNA of 300 units using minimal basis set (energy in eV, number N in relative units)
known to correctly reproduce the trends in the electronic properties.\textsuperscript{22-28}

The DOS curves of the valence band and conduction band regions of single stranded periodic DNA having different sequences and aperiodic DNA are shown in Figs (1-3). In the case of aperiodic poly (AGCT), the random sequence of the units was generated using a random number generator programme.\textsuperscript{29}

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), bandgap (Eg) obtained using simple NFC method are given in Tables 3 and 4.

As can be seen from the DOS curves, in the case of periodic poly (AGCT) both valence and conduction band regions consist of well-separated very narrow peaks of almost equal intensity (Fig. 1). On the other hand, in aperiodic poly (AGCT) both valence and conduction band regions still have many gaps but the peaks become much broader (Fig. 2). As compared to periodic poly (AGCT), in the case of aperiodic poly (AGCT) the upper limit of the valence band region is shifted upwards and the lower limit of the conduction band region is shifted downwards in energy. The result is a slight decrease in the bandgap in aperiodic case as compared to the periodic case. Similar behaviour was observed for both minimal basis and double zeta basis (Fig. 3). The appearance of relatively broader peaks in the case of aperiodic

![Fig. 2—DOS curves of the single stranded aperiodic DNA of 300 units using minimal basis set (energy in eV, number N in relative units)](image-url)

Table 3—Calculated electronic properties (in eV) of single stranded periodic and aperiodic DNA base stacks using minimal basis

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Sequence</th>
<th>Unit</th>
<th>I.P</th>
<th>E.A</th>
<th>Bandgap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(A-T-G-C)</td>
<td>300</td>
<td>9.00</td>
<td>1.65</td>
<td>10.65</td>
</tr>
<tr>
<td>2</td>
<td>(A-G-C-T)</td>
<td>1000</td>
<td>9.00</td>
<td>1.65</td>
<td>10.65</td>
</tr>
<tr>
<td>3</td>
<td>(A-G-T-C)</td>
<td>300</td>
<td>9.00</td>
<td>1.65</td>
<td>10.65</td>
</tr>
<tr>
<td>4</td>
<td>(A-C-T-G)</td>
<td>300</td>
<td>9.00</td>
<td>1.65</td>
<td>10.65</td>
</tr>
<tr>
<td>5</td>
<td>Aperiodic-1</td>
<td>300</td>
<td>8.80</td>
<td>1.50</td>
<td>10.30</td>
</tr>
<tr>
<td>6</td>
<td>Aperiodic-2</td>
<td>1000</td>
<td>8.75</td>
<td>1.45</td>
<td>10.20</td>
</tr>
</tbody>
</table>

Table 4—Calculated electronic properties (in eV) of single stranded periodic and aperiodic DNA base stacks using double zeta basis

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Sequence</th>
<th>Unit</th>
<th>I.P</th>
<th>E.A</th>
<th>Bandgap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(A-T-G-C)</td>
<td>300</td>
<td>8.00</td>
<td>2.45</td>
<td>10.45</td>
</tr>
<tr>
<td>2</td>
<td>(A-G-C-T)</td>
<td>1000</td>
<td>8.00</td>
<td>2.45</td>
<td>10.45</td>
</tr>
<tr>
<td>3</td>
<td>(A-G-T-C)</td>
<td>300</td>
<td>8.00</td>
<td>2.45</td>
<td>10.45</td>
</tr>
<tr>
<td>4</td>
<td>(A-C-T-G)</td>
<td>300</td>
<td>8.00</td>
<td>2.45</td>
<td>10.45</td>
</tr>
<tr>
<td>5</td>
<td>Aperiodic-1</td>
<td>300</td>
<td>7.8</td>
<td>2.30</td>
<td>10.10</td>
</tr>
<tr>
<td>6</td>
<td>Aperiodic-2</td>
<td>1000</td>
<td>7.75</td>
<td>2.35</td>
<td>10.10</td>
</tr>
</tbody>
</table>
system than in the periodic system is not surprising. Similar broadening has been observed in the case of disordered polypeptide chains\textsuperscript{30,31} and random copolymers of heterocyclic compounds\textsuperscript{15-17}. In periodic poly (AGCT), because of the large size of the unit cell (A-G-C-T) (four times as large as the units cell in homopolynucleotides) the Brillouin zone is correspondingly reduced and because of the four different types of unit in the unit cell, its periodicity is broken leading to the four 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 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 is the interaction of A with its neighbours in a given sequence from its interaction with other neighbouring A's in poly (A). In the case of periodic sequences (-A-G-C-T-) of poly (AGCT) the respective environments of A, G, C and T remain in all positions the same. Therefore, for periodic poly (AGCT) we obtain four very narrow peaks of almost equal intensity. On the other hand, in the case of a random aperiodic
sequence of poly (AGCT), the respective environments of A, G, C and T 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 band structure calculations, bandgaps have been calculated using Clementi's double zeta basis set. Their results are given in Table 4. Comparison of these results with those obtained using minimal basis set (Table 3) shows that though the fundamental bandgap decreases with better basis set, the decrease is very small. It needs to be pointed out here that the Hartree Fock method overestimates the bandgap 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 bandgap values are expected to decrease and come closer to experimental values.

To check the dependence of the DOS on the sequence of units and chain length, both valence and conduction band regions of periodic and aperiodic DNA were studied with four different arrangements in periodic and two different random sequences of 300 and 1000 units. Chain lengths of 300 and 1000 units produce similar results in case of periodic DNA. Small difference is observed in case of aperiodic sequence and periodic sequence.

On the basis of the results obtained until now, not too much definite can be said about the conduction properties of DNA because the gaps in the DOS curves of aperiodic poly (AGCT) 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, should take into account the effect of the water and ion environment (which may introduce extra levels in the gaps) and also take into account other factors like the interactions of DNA with polypeptides etc. In our present study, only the interactions between the nucleotide bases in the direction of the helix are considered. In real DNA, on the other hand, there are two antiparallel strands of nucleotide bases in the inner part of the double helix. The allowed base pairs in the two right handed antiparallel stands are A-T and G-C. It is quite likely that because of these additional interactions between the bases the structure of 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 DNA at the \textit{ab initio} level.

**Conclusions**

In this paper we have studied electronic density of states (DOS) curves of single stranded periodic and aperiodic DNA using both minimal basis set and double zeta basis sets with the help of simple negative factor counting method. By the use of better basis set the calculated bandgap values were found to decrease. Regarding dependence of DOS on the chain length, 300 units is found to be good representation of chain length. To check dependence of the DOS on the sequence, four different arrangements in periodic and two different random sequences were studied and were found to be nearly identical. The calculated large values of bandgap obtained for these systems rule out the possibility of intrinsic conduction in them. Similar conclusions were also drawn by Bakhshi et al.\textsuperscript{3} using complex and time-consuming matrix block NFC method. Hence we conclude that simple NFC method is an efficient tool for calculating the DOS of periodic and aperiodic biopolymers such as DNA.

**References**