ELSEVIER

Physica B 279 (2000) 246–252

PHYSICA B

www.elsevier.com/locate/physb

Calculated state densities of aperiodic nucleotide base stacks

Yuan-Jie Ye[a,b], Run-Shen Chen[a,b], Alberto Martinez[b], Peter Otto[b], Janos Ladik[b,*]

[a]*Department of Protein Engineering, Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing 100101, People's Republic of China*

[b]*Institute for Theoretical Chemistry and Laboratory of National Foundation for Cancer Research, Friedrich-Alexander-University of Erlangen-Nürnberg, Egerlandstraβ 3, D-91058 Erlangen, Germany*

Received 10 November 1998; received in revised form 22 August 1999; accepted 4 October 1999

Abstract

Electronic density of states (DOS) histograms and of the nucleotide base stack regions of a segment of human oncogene (both single and double stranded, in B conformation) and of single-stranded random DNA base stack (also in B conformation), were calculated. The computations were performed with the help of the ab initio matrix block negative factor counting (NFC) method for the DOSs. The neglected effects of the sugar–phosphate chain and the water environment (with the counterions) were assessed on the basis of previous ab initio band structure calculations. Further, in the calculation of single nucleotide base stacks also basis set and correlation effects have been investigated. In the case of a single strand the level spacing widths of the allowed regions and the fundamental gap were calculated also with Clementi's double ς basis and corrected for correlation at the MP2 level. The inverse interaction method was applied for the study of Anderson localization.

Keywords: Densities of state; Electronic structure; Nucleotide base stacks

1. Introduction

Most biopolymers, like proteins, DNA, RNA, etc. are aperiodic. In several previous papers we have studied different aperiodic proteins by calculating their density of states (DOS) [1–7].

There is also an earlier DOS calculation of an aperiodic base stack in the DNA B conformation [8] with a random sequence [9].

In the case of 100 base pairs taking into account the first 200 levels in the valence bands regions (all of which originate from the highest filled orbitals, HOMOs of the free bases) we have calculated the DOSs. After that, using the inverse iteration method [10] we have determined the localized (Anderson localization) wave functions belonging to these levels.

2. Methods

In the calculations, the geometry of the bases was taken from a data set determined by X-ray diffraction on a single crystal of DNA [11]. The rotation angle between the nearest neighboring bases (base pairs) is as usual 36° and the stacking distance is 3.36 Å. In this way, we have constructed the geometry of the whole stack in which the main

* Corresponding author. Tel.: +49-9131-852-8831; fax: +49-9131-852-7736.

E-mail address: ladik@pctc.chemie.uni-erlangen.de (J. Ladik)

0921-4526/00/$ - see front matter
PII: S0921-4526(99)00792-9

features of a DNA molecule could be maintained, though the sugar–phosphate backbones were not taken into account explicitly. In this connection it should be mentioned that according to earlier band structure calculations of periodic homopolynucleotides (sugar–phosphate and always the same base) the resulting bands can always be classified as sugar–phosphate bands and as base stack bands [12–14]. The valence (highest filled) band and the conduction (lowest unfilled) band of these systems originated in all the three calculated cases [with cytosine (C), thymine (T) adenine (A) as nucleotide base] from the highest filled or lowest unfilled levels of the single bases. The highest filled or lowest unfilled sugar–phosphate bands were always below or above the valence and conduction bands, respectively. Inspecting in more detail these homopolynucleotide band structures, one could notice that they are, in all the three cases, in a very good approximation a superposition of the band structures of the sugar–phosphate chain and of the periodic base stacks. This is due to the effect of the mutual screening of the charges on the subunits of a nucleotide (~ -1.2 e on the phosphate group [15], $\approx +0.6$ e on the sugar units and ≈ -0.2 e on the bases [12–14] and $+1.0$ charge on the counterions). These alternating charges on the sugar and phosphate units *do not* change the charge distributions significantly on the base pairs (otherwise, the band structures of the homopolynucleotides could not be practically identical with those of the superposition of the sugar–phosphate chain and of the base stacks).

Concerning the effect of the water environment, an early Hartree–Fock band structure calculation [16] on a periodic C stack has shown that taking into account the effect of the water molecules (in the form of five water clusters in each plane of the C molecules based on a Monte Carlo calculation [17]), the water environment hardly influences the band structure of the C stack.

On the basis of the above-described results one can conclude that the presence of the sugar–phosphate chains in DNA and the water environment together with the counterions influence only to a very small extent the band structure of the periodic base (pairs) stacks and therefore most probably also the level distributions of an aperiodic stack. In other words, the electronic structure of the free base or base pair stacks (both periodic or aperiodic) provides a very good approximation of the electronic structures of the same systems in DNA (especially if one takes into account the aperiodic stack of at least 50, but possibly 100 units).

Two different sequences (a) a fragment of 100 or 200 bases of the C end of a human oncogene [18] and (b) a random sequence of 100 bases constructed in the proportion $A:C:G:T = 1:1:1:1$ and without repeated bases, were calculated in their single-stranded forms. A double helical fragment was also calculated with the former sequence. The first sequence is shown in Fig. 1. In this paper the sequences are abbreviated as Sseq1 for the segment of human oncogene and Sseq2 for the random sequence without repeated bases (single chain) and as Dseq1 for the double helical segment of the human oncogene.

In the calculations, the overlapping dimer approximation was used as it was done in the calculations of proteins [19] (the stacks were partitioned into overlapping dimers). Only the 16 different dimers were calculated (in a single DNA strand one has 16 dimers along the strand, because in a XY dimer X = C, T, G or A and Y is also C, T, G or A. In the double-stranded DNA there are 10 different dimers. The Fock (overlap) matrix of the whole system can be constructed in the following way: all dimers are calculated in the same local coordinate systems defined by the first dimer. To build up the helical structure not only do the nuclei have to be rotated and translated in the direction of the z-axis (assuming that the helical axis coincides with the z-axis) by appropriate multiples of 36° and 3.36 Å, respectively. In addition the basis functions have to be rotated too, i.e. all matrix elements, in which p_x and/or p_y functions occur, have to be transformed.

CTCGA	GGGAG	GAGCC	CGGGG	CTGGG	GTACG	GAGGC	CTCTG	CACAT	CTTAG
AGTAA	AACAA	GCAGG	AGAGG	CTGGG	TGCGG	TGGCT	CATGC	CTATA	ATCCC
AGCAC	TTTAG	GAGGC	TGAGG	CGGGC	AGATC	ACCTG	AGGTC	GGGAG	TTCAA
GACCA	GCCTG	ACCAA	CAGGG	AGAAA	CCCCA	TCTTT	ACTAA	AACTA	CAAAA

Fig. 1. (a) The sequence of a single strand of a human oncogene [18] (starting at the C end) with 200 bases.

It should be mentioned that the overlapping dimer approximation gives practically the same total absolute DOS for the Fock matrix constructed in this way, than if one performs the negative factor counting (NFC) calculation on a Fock matrix constructed directly as it was checked on systems with smaller unit cells [20].

The generalized eigenvalue equations of the Fock matrix of the whole chain constructed in the above-described way can be solved by the NFC method [21–24]. The program extended negative factor counting method, (ENFC) [24], which can take into account also cross links, was used to obtain the DOS and the energy levels of the whole system. In the calculations 5377 and 5275 basis functions (using a Clementi's minimal basis [25]), respectively, were applied for the different sequences in the single-stranded chain and 10 539 basis functions were used in the case of the double-stranded chain with 100 units. The number of basis functions was 10 693 for a single-stranded and 21 088 for a double-stranded chain in the case of 200 units.

To investigate the effect of a better basis set we have performed also a double ζ calculation with Clementi's double ζ basis set [25] for the single strand for Sseq1 of 100 units and also for a double helix with the same basis, number of base pairs and sequence.

In the case of the single strand also correlation has been taken into account in the case of all the 16 dimers using the inverse Dyson equation in its diagonal approximation [26] with a Moeller–Plesset [27] self-energy and taking into account also relaxation effects [28]. For this calculation again Clementi's double ζ basis was applied. This calculation was performed by substituting the original Fock matrix of an AB dimer:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{\mathrm{AA}} & \mathbf{F}^{\mathrm{AB}} \\ \mathbf{F}^{\mathrm{AB^{tr}}} & \mathbf{F}^{\mathrm{BB}} \end{pmatrix} \tag{1}$$

by a new matrix $\tilde{\mathbf{F}}$ [29].

$$\tilde{\mathbf{F}} = \mathbf{F} + \mathbf{S}\mathbf{U}\boldsymbol{\Sigma}_{\mathrm{diag}}\mathbf{U}^{+}\mathbf{S} \tag{2}$$

(where it is assumed that the off-diagonal parts of the self-energy matrix $\boldsymbol{\Sigma}$ can be neglected [26]).

Here $\mathbf{S}$ is the overlap matrix of the whole dimer, the unitary matrix $\mathbf{U}$ is formed from the eigenvectors occurring in the subsequent equation

$$\mathbf{F}\boldsymbol{u}_1 = \varepsilon_i^{\mathrm{HF}}\mathbf{S}\boldsymbol{u}_1 \tag{3}$$

and $\boldsymbol{\Sigma}_{\mathrm{diag}}$ is the diagonal part of the self-energy matrix. If we multiply Eq. (2) on the left-hand side by $\mathbf{U}^{+}$ and on the right-hand side by $\mathbf{U}$, and take into account the normalization conditions

$$\mathbf{U}^{+}\mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{S}\mathbf{U}^{+} = \mathbf{1} \tag{4}$$

we obtain

$$\mathbf{U}^{+}\tilde{\mathbf{F}}\mathbf{U} = \mathbf{U}^{+}\mathbf{F}\mathbf{U} + \boldsymbol{\Sigma}_{\mathrm{diag}} \tag{5}$$

$$(\boldsymbol{\omega})_{i,i} = \omega_i = (\boldsymbol{\varepsilon}^{\mathrm{HF}})_{i,i} + [\boldsymbol{\Sigma}(\omega_i)]_{i,i} = \varepsilon_i^{\mathrm{HF}} + [\boldsymbol{\Sigma}(\omega_i)]_{i,i} \tag{6}$$

which is the diagonal part of the inverse Dyson equation.

One can approximate the diagonal elements of $\boldsymbol{\Sigma}$ by the MP2 expression (taking into account also relaxation [28])

$$[\Sigma^{\mathrm{MP2}}(\omega_i)]_{i,i} = \sum_{\substack{j\in occ \\ a,b\notin occ}} \frac{V_{ijab}(2V^{*}_{ijab} - V^{*}_{ijba})}{\omega_i + \varepsilon_j^{HF} - \varepsilon_a^{HF} - \varepsilon_b^{HF}} + \sum_{\substack{j\notin occ \\ a,b\in occ}} \frac{V_{ijab}(2V^{*}_{ijab} - V^{*}_{ijba})}{\omega_i + \varepsilon_j^{HF} - \varepsilon_a^{HF} - \varepsilon_b^{HF}}. \tag{7}$$

Here the star as a superscript indicates complex conjugate and V_{ijba} is the exchange integral corresponding to the matrix elements V_{ijab} which are defined as

$$V_{ijab} = \left\langle \phi_i(\boldsymbol{r}_1)\phi_a(\boldsymbol{r}_2) \left| \frac{1}{r_{12}} \right| \phi_j(\boldsymbol{r}_1)\phi_b(\boldsymbol{r}_2) \right\rangle, \tag{8}$$

where the dimer orbitals $\phi_i(\boldsymbol{r}_1)$, etc. can be written in an LCAO form

$$\phi_1(\boldsymbol{r}_1) = \sum_{t=1}^{g} c_{i,t}\chi_t(\boldsymbol{r}_1).$$

Eq. (6) is a non-linear equation which has to be solved iteratively [30] (putting $\omega_i^{(0)} = \varepsilon_i^{\mathrm{HF}}$). Therefore we have several real solutions. The physically interesting one fulfills the condition for the pole

strength

$$P_i = \left[1 - \left(\frac{\partial \sum^{MP2}(\omega)}{\partial \omega}\right)_{\omega=\omega_i}\right]^{-1} \geqslant 0.6. \quad (9)$$

In the case of molecules one can always find a dominant solution with $P_i \sim 0.8$–0.9. This is not the case for metals. In our dimer calculations we have always used Eq. (9) to find the physically significant ω_i solutions.

The Anderson localization of the different orbitals was investigated with the help of the inverse iteration method [10]. We have found the orbitals to be localized on one or two bases.

3. Results and discussion

3.1. Total density of states

Figs. 2 and 3 show the DOS of two different chain segments. The energy gaps are 8.7724 eV for the double-stranded case with the oncogene sequence (Dseq1), 10.6090 eV for the single strand with the oncogene sequence (Sseq1) and 10.7693 eV with the random sequence (Sseq2). The different sequences hardly influence the energy gaps. However, the interactions between the bases in the base pairs and between neighboring base pairs reduce the energy gap by about 2 eV. One should point out that these gaps are essentially larger than the first singlet excitation energies of the single nucleotide bases (values between 4.5 and 5.5 eV). The reason for this discrepancy is that the gap is not identical with the u.v. excitation energies. Between the valence band and the conduction band there are two exciton bands in the case of a cytosine stack in the DNA B conformation (the calculated values are between 4.7 and 5.4 eV and 5.9 and 6.2 eV, respectively [31]).

The gap values could be measured with the help of inverse photoelectron spectroscopy (the difference between the average ionization potential of the disordered nucleotide base stacks in the ground state and the ionization potential of injected electrons in the conduction band region). Such experiments are, however, unfortunately not available for the nucleotide base stacks.

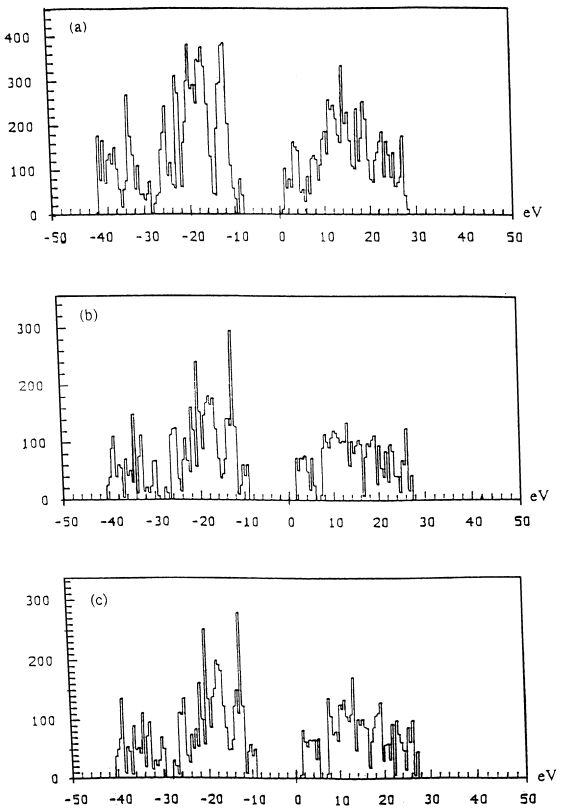

Fig. 2. The electronic density of states for (a) double-stranded DNA (100 units) with sequence 1 (human oncogene), (b) the same for a single strand of DNA with sequence 1 and (c) the same with sequence 2.

The main features of the DOS of the different systems are shown in Fig. 2 in which the energy grid is 0.45 eV. One can see that the different sequences in the single-stranded case do not influence their DOS histograms significantly. In the double helical case, the DOS curve has obviously different features to those of the single-stranded cases due to the larger number of states (higher peaks in the histograms).

In Fig. 3a and b the DOSs of a 200 base (or base pair, respectively) long stack of DNA with sequence 1 is shown.

On comparing the DOS histograms of the base pair stacks (Fig. 2a) with those of the single base stacks one sees that the fundamental gap is smaller in the latter case. This is mostly due to the different

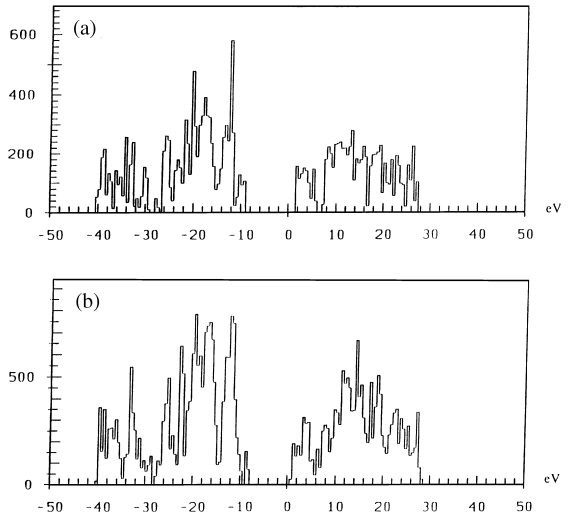

Fig. 3. Density of states of DNA with 200 units using sequence 1 (a) single-stranded chain and (b) double-stranded chain.

positions of the HOMO and LUMO levels of the four bases and not due to the H-bonded interactions between the members of the base pairs. On the other hand, the gap is nearly the same and the forms of the DOS histograms are very similar for the two different sequences in the case of a single strand both in the valence band and conduction band regions (compare Fig. 2b and c) showing that the electronic structure of an aperiodic DNA chain is rather sequence independent. Comparing further Fig. 2a with Fig. 2b (same sequence) one sees that the level distribution is not very different with two exceptions. (1) For the base pair stack one finds a somewhat stronger level distribution broadening (which is, according to our previous experience, typical for aperiodic chains [2,9]). (2) The heights of the peaks in the histograms are generally larger in the double nucleotide base stack than in the single one due to the larger (about double) number of levels falling in the same energy intervals. One can see both effects if one compares the DOSs of a base pair stack with that of a single base stack both containing 200 units and the same human oncogene sequence (compare Fig. 3a and b). One should point out that in the case of the double ζ calculations the shapes of the DOSs curves are similar to those in Fig. 2 (see Fig. 4).

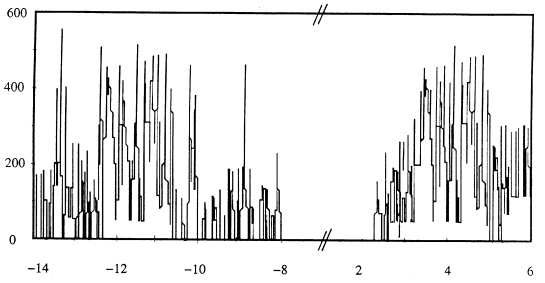

Fig. 4. The DOS of single-stranded DNA (100 units) with sequence 1 using Clementi's double ς basis (valence bands and conduction bands regions).

On the other hand, the fundamental gap especially in the correlation corrected case of the single strand (Sseq1 with 100 units; see Fig. 5) and the spacings between the peaks of the histograms in Fig. 5a and b and the whole breadth of the histograms (both in the case of the valence levels and lowest unfilled levels region) are considerably smaller. For instance, the fundamental gap for single-stranded DNA with sequ 1 using a minimal basis is 11 ev (see Fig. 2b). With the double ς basis it decreases by ~ 0.5 eV (see Fig. 4b) and with correlation by further ~ 1.3 eV. This is in agreement with the usual experience in periodic chains that the fundamental gap, the bandwidths and the other gaps (between the bands) become considerably smaller, if one uses a better basis and corrects the band structure for correlation [32].

4. Conclusion

From the previous results one can conclude that with the help pf present-day techniques one can calculate DOSs and one-electron wave functions for quite long sequences of nucleotide bases or base pairs, respectively, which we consider reliable at least if one uses a better basis and introduces also correlation corrections. We have found that the DOS curves are rather sequence independent, but of course they are strongly influenced by the level spacings (a double strand or a larger single strand has quite different DOSs than a shorter single strand).

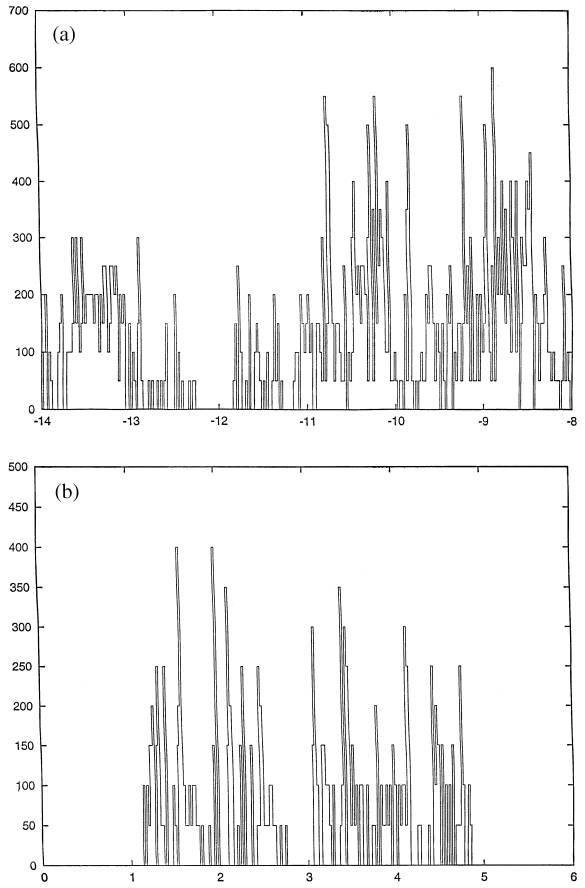

Fig. 5. The DOS of single-stranded DNA (100 units) with sequence 1 using Clementi's double ς basis. The levels were corrected for correlation at the MP2 level (a) valence bands region and (b) conduction bands region.

Basis set and correlation effects decrease again considerably the level spacings and with it the fundamental gap (the difference of the computed lower limit of the conduction band and the upper limit of the valence band) and the gaps between the different regions of allowed levels.

Further, in a more realistic calculation of aperiodic DNA one should take into account also the sugar–phosphae backbone with its ions and water environment though one would not expect too large effects on the DOSs (see Introduction). The effect of water environment would most probably increase somewhat the average level spacing as in the case of a C stack band structure calculation [16].

Acknowledgements

We are indebted to Mrs. Y. Jiang and to Professer. A.K. Bakhshi for their help and for the fruitful discussions. We are grateful to the "National Natural Scientific Foundation of China", The Alexander-von-Humboldt Foundation, the German Academic Exchange (DAAD) and the "Deutsche Forschungsgemeinschaft" for the financial support which made it possible to perform this joint research in Erlangen.

References

[1] J. Ladik, M. Seel, P. Otto, A.K. Bakhshi, Chem. Phys. 108 (1986) 203.
[2] A.K. Bakhshi, J. Ladik, M. Seel, P. Otto, Chem. Phys. 108 (1986) 233.
[3] Y.-J. Ye, J. Ladik, Phys. Rev. B 48 (1993) 5120.
[4] Y.-J. Ye, J. Ladik, Phys. Rev. B 51 (1995) 13091.
[5] Y.-J. Ye, J. Ladik, Int. J. Quantum Chem. 52 (1994) 491.
[6] Y. Jiang, Y.-J. Ye, R.-S. Chen, Biophys. Chem. 59 (1996) 95.
[7] Y.-J. Ye, J. Ladik, Physiol. Chem. Phys. Med. NMR 28 (1996) 123.
[8] J.D. Watson, F.H.C. Crick, Nature 171 (1953) 737.
[9] A.K. Bakshi, P. Otto, J. Ladik, M. Seel, Chem. Phys. 108 (1986) 215.
[10] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon, Oxford, 1965.
[11] R.A. Dickerson, H.R. Drew, B.N. Conner, R.M. Wing, A.V. Fratini, M.L. Kopka, Science 216 (1982) 475.
[12] J. Ladik, S. Suhai, Phys. Lett. 77A (1980) 25.
[13] P. Otto, E. Clementi, J. Ladik, J. Chem. Phys. 8 (1980) 454.
[14] E. Clementi, G. Corongiu, Int. J. Quantum Chem. QBS9 (1982) 213.
[15] Y.-J. Ye, unpublished results.
[16] P. Otto, J. Ladik, G. Corongiu, S. Suhai, W. Förner, J. Chem. Phys. 77 (1981) 5026.
[17] G. Corongiu, E. Clementi, Biopolymers 20 (1981) 551.
[18] G.I. Bell, R. Piclet, J. Writter, Nucleic Acid Res. 8 (1980) 4091.
[19] Y.-J. Ye, J. Ladik, J. Math. Chem. 14 (1993) 141.
[20] B. Gazdy, M. Seel, J. Ladik, Chem. Phys. 86 (1984) 41.
[21] P. Dean, J.L. Martin, Proc. Roy. Soc. A 259 (1960) 409.
[22] P. Dean, Rev. Mod. Phys. 44 (1972) 127.
[23] R.S. Day, F. Martino, Chem. Phys. Lett. 84 (1981) 86.
[24] Y.-J. Ye, J. Math. Chem. 14 (1993) 121.
[25] L. Gianolo, E. Clementi, Gazz. Chim. Ital. 110 (1980) 179.

[26] L.S. Cederbaum, W. Domcke, Adv. Chem. Phys. 36 (1977) 205.
[27] C. Moeller, M.S. Plesset, Phys. Rev. 46 (1934) 618.
[28] N.A. Ostlund, A. Szabo, Modern Quantum Chemistry, McMillan, New York, 1982, p. 398.
[29] C.-M. Liegener, Chem. Phys. 133 (1989) 173.
[30] I. Palmer, J. Ladik, J. Comp. Chem. 15 (1994) 814.
[31] S. Suhai, Int. J. Quantum Chem. 11 (1984) 223.
[32] J.J. Ladik, Quantum Theory of Polymers as Solids, Plenum, New York, 1988 (Chapter V).