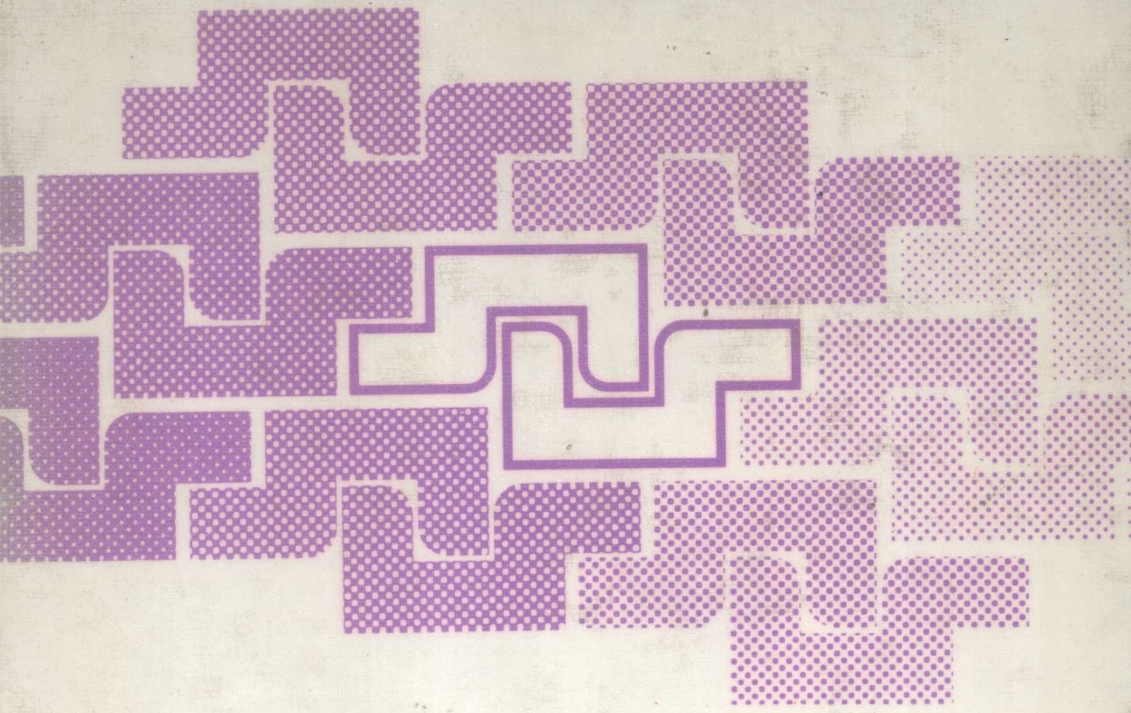


Spectroscopy of Biological Molecules

Theory and Applications – Chemistry, Physics, Biology, and Medicine

Edited by

Camille Sandorfy and Theophile Theophanides

NATO ASI Series

Series C: Mathematical and Physical Sciences Vol. 139

Spectroscopy of Biological Molecules

Theory and Applications – Chemistry, Physics, Biology, and Medicine

edited by

Camille Sandorfy

and

Theophile Theophanides

Department of Chemistry, University of Montreal,
Montreal, Quebec, Canada

D. Reidel Publishing Company

Dordrecht / Boston / Lancaster

Published in cooperation with NATO Scientific Affairs Division

Proceedings of the NATO Advanced Study Institute on
Spectroscopy of Biological Molecules
Acquafredda di Maratea, Italy
July 4 - 15, 1983

Library of Congress Cataloging in Publication Data

NATO Advanced Study Institute on Spectroscopy of Biological Molecules (1983 : Acquafredda di Maratea, Italy)
Spectroscopy of biological molecules.

(NATO ASI series. Series C, Mathematical and physical sciences ; vol. 139)
"Proceedings of the NATO Advanced Study Institute on Spectroscopy of Biological Molecules, Acquafredda di Maratea, Italy, July 4–15, 1983."–T.p. verso.
Includes index.
1. Spectroscopy–Congresses. 2. Biomolecules–Analysis–Congresses.
I. Sandorfy, Camille, 1920– . II. Theophanides, Theo M. III. Title.
IV. Series: NATO ASI series. Series C. Mathematical and physical sciences ; no. 139.
QP519.9.S36N38 1983 574.19'285 84-17862
ISBN 90-277-1849-0

Published by D. Reidel Publishing Company
P.O. Box 17, 3300 AA Dordrecht, Holland

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
190 Old Derby Street, Hingham, MA 02043, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, Holland

D. Reidel Publishing Company is a member of the Kluwer Academic Publishers Group

Printed in The Netherlands.

Spectroscopy of Biological Molecules

Theory and Applications – Chemistry, Physics, Biology, and Medicine

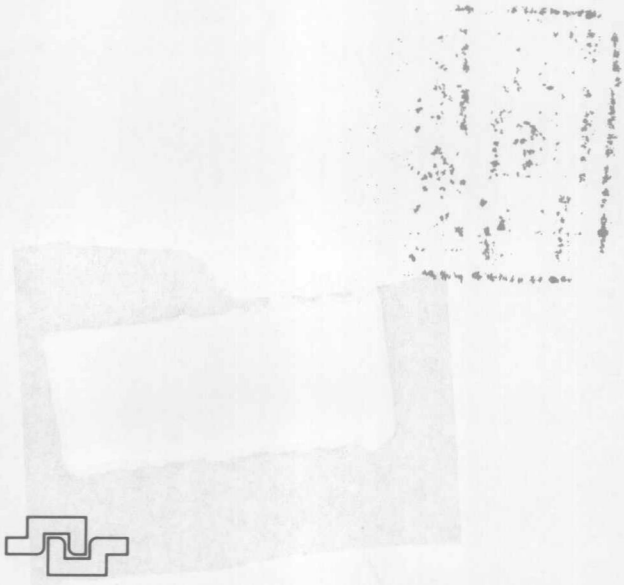
NATO ASI Series

Advanced Science Institutes Series

A series presenting the results of activities sponsored by the NATO Science Committee, which aims at the dissemination of advanced scientific and technological knowledge, with a view to strengthening links between scientific communities.

The series is published by an international board of publishers in conjunction with the NATO Scientific Affairs Division

A	Life Sciences	Plenum Publishing Corporation
B	Physics	London and New York
C	Mathematical and Physical Sciences	D. Reidel Publishing Company Dordrecht, Boston and Lancaster
D	Behavioural and Social Sciences	Martinus Nijhoff Publishers
E	Engineering and Materials Sciences	The Hague, Boston and Lancaster
F	Computer and Systems Sciences	Springer-Verlag
G	Ecological Sciences	Berlin, Heidelberg, New York and Tokyo



Series C: Mathematical and Physical Sciences Vol. 139

PREFACE

This volume contains the proceedings of the NATO-Advanced Study Institute on the "Spectroscopy of Biological Molecules", which took place on July 4-15, 1983 in Acquafredda di Maratea, Italy.

The institute concentrated on three main subjects: the structure and dymanics of DNA, proteins, and visual and plant pigments. Its timeliness has been linked to rapid advances in certain spectroscopic techniques which yielded a considerable amount of new information on the structure and interactions of biologically important molecules. Among these techniques Fourier transform infrared, resonance and surface enhanced Raman spectroscopies, Raman microscopy and microprobing, time resolved techniques, two photon and ultrafast electronic, and C-13, N-15 and P-31 NMR spectroscopies and kinetic and static IR difference spectroscopy receiced a great deal of attention at the Institute. In addition, an entirely new technique, near-millimeter-wave spectroscopy has been presented and discussed.

Two introductory quantum chemical lectures, one on the structure of water in DNA, and another on the energy bands in DNA and proteins set the stage for the experimentally oriented lectures that followed. Fundamental knowledge on hydrogen bonding was the topic of two other lectures.

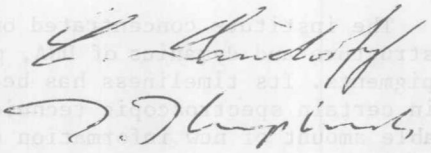
Panel discussions were held on the structure and conformations of DNA, metal-DNA adducts and proteins and on visual pigments.

Many scientists who normally attend different conferences and never meet, met at Aquafredda di Maratea. We feel, that at the end of our Institute a synthetic view emerged on the powerful spectroscopic and theoretical methods which are now available for the study of biological molecules.

The directors of this Advanced Study Institute would like to express their profound gratitude to NATO's Scientific Affairs Division for making the holding of the Institute possible. They express their heartfelt thanks to all the Lecturers and Participants for their valuable contributions and for consenting a certain financial sacrifice so that

we could have the required number of lecturers in all the sections of our broad subject matter.

Last, but not least we thank our Italian friends for their help and kind hospitality and Nature for the blue sky and sea of Italy.



Montréal, Québec, December, 1983

C. Sandorfy
T. Theophanides

TABLE OF CONTENTS

ENERGY BANDS IN DNA

Janos J. LADIK

Chair for Theoretical Chemistry, Friedrich-Alexander-University Erlangen-Nürnberg, FRG and Laboratory of National Foundation for Cancer Research at the Chair for Theoretical Chemistry, University Erlangen-Nürnberg

ABSTRACT.- Ab initio SCF LCAO band structures of homopolynucleotides are presented. In the case of a cytosine stack the effect of the surrounding water on the band structure is shown. It is mentioned in the case of polyacetylene that using a good basis set and taking into account the major part of correlation the Hartree-Fock gap reduces to a value in reasonable agreement with experiment. Therefore, one can expect that the gap for DNA is also considerably smaller than minimal basis calculations indicated and can raise the question whether DNA is an intrinsic semiconductor.

Finally, the effect of aperiodicity on the band structure of DNA is discussed on the basis of calculations using the negative factor counting (NFC) technique.

I. INTRODUCTION

Polymers play an important role as plastics. Highly conducting polymers are in the last decade objects of extensive experimental investigations being candidates for the discovery of new physical phenomena and serious attempts are made for their technical application (like batteries). Biopolymers like nucleic acids (DNA and RNA), proteins, polysacharides, lipids etc. have fundamental importance in life processes. To understand the different physical and chemical properties of polymers (which underlie in the case of biopolymers also their biological functions) one has to obtain a fair knowledge of their electronic structure.

To treat quantum mechanically DNA one has to proceed stepwise:

1.) One starts with ab initio SCF LCAO crystal orbital (CO)

C. Sandorfy and T. Theophanides (eds.), Spectroscopy of Biological Molecules, 1–13.

calculations /1/ taking a nucleotide base, a nucleotide base pair or a whole nucleotide as unit cell (periodic polynucleotides).

2.) As next step one has to consider that DNA is aperiodic, therefore one has to apply appropriate techniques /2/ to treat this compositional disorder using the results of periodic chain calculations as input (see below).

3.) After this one has to take into account the effect of the surrounding water and ions on the band structure by constructing an effective potential of the environment /3/.

4.) Finally, one should take into account (using a good basis set) also the major part of the electronic correlation /4/.

5.) Having performed for a polymer steps 1.) - 4.) one is in a good position to calculate different properties (electronic and vibrational spectra, transport and magnetic properties etc.) of it.

6.) In the case of biopolymers it is very important to treat also interactions between polymer chains (for instance the genetic regulation of a cell is mostly dependent on DNA-protein interactions in a nucleoprotein).

Since the unit cells in DNA are fairly large, the calculations performed until now belong mostly to the first step (band structure calculations of homopolynucleotides. There are a few attempts to treat aperiodicity in DNA and there exists a pilot calculation to treat the effect of the surrounding water molecules on the band structure of a cytosine stack. In this short review the results of a part of these calculations performed for DNA will be presented together with the necessary methods. For the calculation of the electronic structure of proteins which presents a still more formidable problem than that of DNA, one has to perform the same steps. Until now only the band structures of a few homopolypeptides have been computed and the density of states of mixed glycine-alanine and glycine-serine chains, respectively, have been determined (see the papers of Seel and of Day, Suhai and Ladik at reference /2/). Since, however, these calculations though being similar are less advanced than in the case of DNA, we do not discuss them in this paper.

No correlation calculations have been performed on biopolymers yet, but such computations have been successfully executed in the cases of transpolyacetylene /5/ and polydiacetylenes. The same holds for exciton spectra which have been successfully computed applying intermediate exciton theory /6/ (including correlation) for the before mentioned two chains /7/. There is an early calculation on transport properties of periodic DNA models using simple tight binding (Hückel) band structures /8/. There exist pilot calculations also for the interactions between hompolynucleotides and for polyglycine chains in different conformations /9/. Finally one should mention that in the case of a good basis set calculations of polyethylene ($(CH_2)_x$) taking into account also the major part of the correlation even the

mechanical properties of this system could be computed with results /10/ in good agreement with experiment.

The good results obtained for chains with small unit cells give rise to hopes that with even larger computers and with the improvement of the numerical techniques in the next few years the calculations on the electronic structure of DNA and proteins and on their properties will reach the same level of sophistications as those on the simple chains.

II. METHODS

II.1. Ab initio Crystal Orbital Method

If there is a translational (or more generally any periodic) symmetry in an infinite solid or polymer the infinite matrix which one obtains in any LCAO theory can be brought with the help of a simple unitary transformation into a block-diagonal form /1/. The order of these blocks is (in the ab initio case) equal to the number of basis functions in the unit cell. In this way the original hypermatrix equation splits into N+1 matrix equations if N+1 denotes the number of blocks (unit cells). Each such equation has an index which denotes the serial number of the matrix block to which the equation belongs. If $N \to \infty$ this serial number becomes continous. Physically it is the vector $\vec{k}$ (or one of its components in the one-dimensional (1D) case) of the reciprocal lattice which gives the crystal momentum /1/.

Here no attempt will be made to reproduce the derivation of the expressions of the ab initio SCF LCAO CO method (for this see /1/) only the final equations are written down (for the sake of simplicity in the 1D case). Let us express a crystal orbital in the form of a linear combination of Bloch orbitals

$$\Psi_n(k,\vec{r}) = 1/\sqrt{N+1} \sum_{j=-N/2}^{N/2} e^{ikja} \sum_{s=1}^{m} \left(c(k)_{m,s} \chi_s(\vec{r} - \vec{R}_{js}) \right) \quad (1)$$

Here N+1 is the number of unit cells, m is the number of basis functions in the unit cell, a the elementary translation, $\chi_s(\vec{r} - \vec{R}_{js})$ the sth atomic orbital (AO) centered at the Sth atom (to which orbital s belongs) in the j-th unit cell (with position vector $\vec{R}_{js}$) and n is the band index. The coefficients $c(k)_{n,s}$ can be determined from the generalized eigenvalue equation

$$\underline{\underline{F}}(k)\ \underline{c}_n(k) = \varepsilon_n(k)\ \underline{\underline{S}}(k)\ \underline{c}_n(k). \quad (2)$$

Here the overlap matrix $\underline{\underline{S}}(k)$ is the Fourier transform of the matrix blocks containing the overlap integrals between basis functions belonging to different cells,

$$\underline{\underline{S}}(k) = \sum_{q=-N/2}^{N/2} e^{ikqa}\, \underline{\underline{S}}(q); \qquad (3)$$

(for instance $\underline{\underline{S}}(o)$ contains all the overlap integrals within one cell, $\underline{\underline{S}}(1)$ the ones between the reference cell and its next neighbour and so on). Similarly one obtains /1/

$$\underline{\underline{F}}(k) = \sum_{q=-N/2}^{N/2} e^{ikqa}\, \underline{\underline{F}}(q) \qquad (4)$$

The elements of the matrices $\underline{\underline{F}}(q)$ are defined according to the detailed derivation /1/ as

$$\left[\underline{\underline{F}}(q)\right]_{r,s} = \langle \chi_r^0 | \hat{H}^N | \chi_s^q \rangle + \sum_{u,v=1}^{m} \sum_{q_1,q_2=-N/2}^{N/2} P(q_1-q_2)_{u,v} \cdot \qquad (5)$$

$$\cdot \left[\langle \chi_r^0(1)\chi_u^{q_1}(2) \left| \frac{1}{r_{12}} \right| \chi_s^q(1)\chi_v^{q_2}(2) \rangle - \frac{1}{2} \langle \chi_r^0(1)\chi_u^{q_1}(2) \left| \frac{1}{r_{12}} \right| \chi_v^{q_2}(1)\chi_s^q(2) \rangle \right]$$

Here the shorthand notation χ_s^q means $\chi_s(\vec{r} - \vec{R}_{sq})$, $s \in S$, thus the superscripts are always cell and the subscripts basis function indices and $\hat{H}^N$ stands for the one-electron part of the Fock operator of the chain. Finally the elements of the charge-bond order matrices $\underline{\underline{P}}(q_1 - q_2)$ are defined as a generalization of the original definition given by Coulson as

$$P(q_1 - q_2)_{u,v} = 2a/2\pi \int_{-\pi/a}^{\pi/a} \sum_{h=1}^{n^*} c_{h,u}(k)\, c_{h,v}(k)\, e^{ika(q_1-q_2)} dk \qquad (6)$$

where n^* is the number of filled bands.

Inspecting equ.-s (2) - (6) we see that they represent nothing else but Roothaan's SCF LCAO equations for a closed shell system /11/ generalized for an infinite chain with periodic boundary conditions. One has to solve them at a number of points in k space simultaneously to be able to construct the matrices $\underline{\underline{P}}(q_1 - q_2)$ for the next iteration step. Using appropriate numerical integration techniques usually it is enough to have 6-8 k points between o and π/a [$\varepsilon(k) = \varepsilon(-k)$] to obtain consistent results. The SCF procedure does not converge, however, in most cases so easily as in the corresponding molecules, especially if one uses a larger basis set. The main reason for this is that in practice one can go in the summations (3) and (4) only until a limited numbers of neighbours (actually to keep charge neutrality and obtain reasonably reliable results one has to go to different numbers of neighbours for different types of integrals /12/) and the error caused by this procedure is strongly amplified in cases when the matrix $\underline{\underline{S}}(k)$ has some very small eigen-

values /13/. (For the elimination of $\underline{\underline{S}}(k)$ in Equ. (2) one uses again Löwdin's symmetric orthogonalization procedure.)

Finally it should be pointed out that the formalism described here is valid not only for simple translation, but also for cases of repeated combined symmetry operations (for instance helix operation). As group theoretical considerations show it in this case 1) one has to put the nuclei into the right positions by going from one cell to the next and 2) one has to rotate correspondingly also the basis functions /14/.

II.2. Negative Factor Counting Technique for the Determination of States in a Disordered Chain

To obtain an orientation about the level distribution in aperiodic DNA and proteins one can apply the negative factor counting (NFC) method /2/ to determine the density of states in these disordered systems. According to this method if we write for the disordered chain a Hückel determinant which is tridiagonal due to the fact that only first neighbours' interactions are taken into account

$$\left|\underline{\underline{H}}(\lambda)\right| = \begin{vmatrix} \alpha_1 - \lambda & \beta_2 & 0 & 0 \ldots & 0 \\ \beta_2 & \alpha_2 - \lambda & \beta_3 & 0 & \\ 0 & \beta_3 & \alpha_3 - \lambda & \beta_4 & \\ \vdots & & & \ddots & \\ 0 & & & \beta_N & \alpha_N - \lambda \end{vmatrix} = 0, \tag{7}$$

this can be easily transformed into a didiagonal form with the help of successive Gaussian eliminations. Therefore, the determinant

$$\left|\underline{\underline{H}}(\lambda)\right| = \prod_{i=1}^{N} (\lambda_i - \lambda) \tag{8}$$

can be rewritten as

$$\left|\underline{\underline{H}}(\lambda)\right| = \prod_{i=1}^{N} \varepsilon_i(\lambda) , \tag{9}$$

where the diagonal elements of the didiagonal determinant are given by the simple recursion relation

$$\varepsilon_i(\lambda) = \alpha_1 - \lambda - \beta_i^2 / \varepsilon_{i-1}(\lambda), \quad i = 1,2,3,\ldots,N, \tag{10a}$$

$$\varepsilon_1(\lambda) = \alpha_1 - \lambda . \tag{10b}$$

Comparing equ.-s (8) and (9) it is easy to see that for a given value, the number of eigenvalues smaller than λ

($\lambda_i < \lambda$) has to be equal to the number of negative $\varepsilon_i(\lambda)$ factors in equ. (9) /2/. (Calculations of all the eigenvalues ε_i of a long chain (N = 10^4 or 10^3) is impossible, but the computation of the $\varepsilon_i(\lambda)$ factors with the help of equ-s. (10a) and (10b) is very fast.) By giving λ different values throughout the spectrum and taking the difference of the number of negative $\varepsilon_i(\lambda)$'s belonging to consecutive λ values, one can ontain a histogram for the distribution of eigenvalues (density of states) of $\underline{\underline{H}}$ for any desired accuracy.

For actual calculations one has to make a band structure calculation for each component of the disordered chain assuming that it is periodically repeated. Then the values α_i (diagonal elements of $\underline{\underline{H}}$) can be determined from the positions of the bands of the components (the middle point or weighted middle points of the bands) and the off-diagonal elements β_i from the widths of the bands. It should be pointed out that the NFC in this simple form gives only the level distribution belonging to one band (for instance, valence band or conduction band) of the disordered chain.

In the case of disordered quasi-1D systems the NFC method can also be applied for the case of an arbitrary number of orbitals per site either in an ab initio form /2/ or in a semi-empirical, for instance, extended Hückel form /15/ (for the derivation see Day and Martino's paper in ref. /2/). In this case one has a matrix block instead of each diagonal element α_i and off-diagonal element β_{i+1}, respectively, in equ. (7) $\alpha_i - \lambda \rightarrow \underline{\underline{A}}_i - \lambda \underline{\underline{S}}_i$, $\beta_{i+1} \rightarrow \underline{\underline{B}}_{i+1} - \lambda \underline{\underline{S}}_{i+1}$). To construct these matrix blocks one has to perform ab initio MO calculations for the different units to obtain the diagonal blocks and cluster calculations for the off-diagonal ones. For instance in the case of a binary diordered chain in the first neighbours' interactions approximation one has to compute the AA, BB, BA and BB clusters. Having constructed the supermatrix of the disordered chain one can obtain again very easily a histogram for the density of states for all the electrons (or valence electrons) of a long disordered chain in any desired accuracy. After obtaining the level distribution one can generate also a wave function for any particular energy level using the inverse iteration technique /16/.

II.3. Mean Field Treatment of the Effect of the Environment on the Electronic Strcuture of Chains

To take into account the effect of environment on the band structure of a periodic chain or on the level distribution of a disordered chain one can build into the one-electron part of the Fock operator of the chain the effective potential V_{eff} of the environment, $\hat{H}^N = \hat{H}^N + V_{eff}$. To generate V_{eff} one has to know the positions of the molecules (in the case of DNA water molecules and ions) in the environment of the chain. Having this information one lets to interact the systems building up the environment. If

one takes into accunt besides the electrostatic interactions also their mutual polarizations (at least at the Hartree-Fock level), one obtains the so-called mutually consistent charge distributions of the constituent systems. To achieve this one can apply the mutually consistent field (MCF) method /17/ developed in Erlangen. (The same method can be used also for the calculation of interactions between infinite chains /18/). Having the MCF charge distribution of each molecule and/or ion in the environment of a unit cell one obtains for V_{eff} using the classical expression /3/

$$V_{eff}(\vec{r}) = \sum_{q=-N}^{N} \sum_{i=1}^{M} \int \frac{\tilde{\rho}_i^q(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}' \qquad (11)$$

Here N stands again for the number of unit cells, M denotes the number of subsystems surrounding the reference cell and $\tilde{\rho}_i^q(\vec{r})$ is the MCF electronic density of the i-th subsystem in the environment of the q-th cell.

II.4. Calculation of Correlation in a Linear Chain

To calculate the correlation energy per unit cell for the ground state of a polymer (either conductor or insulator) in principle one can use any size consistent method (perturbation theory, coupled cluster expansion, electron pair theories, etc.). To keep the calculations in a manageable size one can use a simple trigonometric series in k consisting of a few terms for the LCAO coefficients occurring in the Bloch functions. In this way one can put everything what is k-dependent before the integration and therefore one has to perform the two-electron integrals over the atomic orbitals only once. On the other hand, since the calculations of the k-dependent prefactor is very fast one can use a dense grid for the three independent k-values occurring in the matrix elements

$$\langle IJ \,||\, AB \rangle \equiv \langle \Psi_I(\vec{r}_1)\, \Psi_J(\vec{r}_2) | 1/r_{12}(1-\hat{P}_{12}) | \Psi_A(\vec{r}_1)\, \Psi_B(\vec{r}_2) \rangle \qquad (12)$$

This type of matrix elements, in which the composite indices I,J.. denote a band index, a k-value and the spin occur in anyone of the above mentioned methods suitable for the correlation calculations.

As first step second order Moeller-Plesset perturbation theory (MP2) /19/ can be applied which has provided (using a 4-31 G basis set) for a hydrogen chain about 70 per cent /4/ and for polyacetylene (using a 4-31 G** basis) about 75 per cent of the correlation energy.

Correlation calculations can be applied not only for the more precise calculation of the total energy per unit cell (making geometry optimization possible) but they can be used also for the