Investigation of electronic structures of model polypeptide chains using genetic algorithm

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

Received 20 October 2009; revised and accepted 22 December 2009

The applications of genetic algorithm towards the investigation of electronic structures of aperiodic model polypeptide chains, first at the \textit{ab initio} Hartree-Fock level and subsequently at the quasi particle band structure level are illustrated. The calculations have been performed using Clementi’s basis sets and the effects of both minimal and double zeta basis sets, as well as the effect of change of secondary structure of the polypeptide chains, viz., the $\alpha$-helix and $\beta$-pleated sheet structures on the band gap values have been studied. The effects of electron correlation and hydration on density of states of model polypeptide chains containing glycine and alanine units have also been investigated. Optimal compositions of the most conducting polypeptide chain returned by genetic algorithm are found to be in good agreement with the results obtained from systematic search.

\textbf{Keywords:} Theoretical chemistry, Genetic algorithms, Electronic structures, Polypeptides

Protein molecules play a vital role in the living cell and are involved in virtually all cell functions. These molecules may vary in structure as well as functions. The quantum mechanical investigations of biopolymers\textsuperscript{1} have always been a difficult task. This is because of the complex structure of proteins which are composed of one or more polypeptide chains (made up from 20 amino acid residues) that for parts of their lengths may assume a large number of random arrangements or form various configurations like the $\alpha$-helix and $\beta$-pleated sheet structures pertaining to their secondary structure and folding in space. In our present work, our aim is to design model polypeptide chains with minimum band gap and maximum electronic delocalization.

There have been many advances in the theoretical “tailoring” of polymers in recent years\textsuperscript{2,3}, particularly studies involving the conduction properties, the ionization potentials, electron affinity values and the fundamental band gaps\textsuperscript{6-9} of simulated polymer chains comprising of monomer units in a known and pre-defined composition\textsuperscript{10}. Among several other techniques, one technique which is finding wide applicability is the genetic algorithm\textsuperscript{11-13} (GA) technique to design polymers with pre-specified properties\textsuperscript{14,15}. GA is a powerful tool to solve high-complexity computational problems and is categorized as a global search heuristic (heuristics are experience-based techniques that help in problem solving, learning and discovery. A heuristic method is particularly useful to rapidly come to a solution that is closest to the best possible answer or ‘optimal solution’). Genetic algorithm is a particular class of evolutionary algorithms that uses techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination). Following these ideas, this algorithm allows us to use the computer to evolve automatic solutions for diverse problems over time. Apart from computationally modeling the various phenomena occurring in nature, this algorithm helps in optimization, simulation, designing and modeling purposes in diverse fields of science and technology. Because of its many positive attributes such as ease of use, flexibility and versatility in application, genetic algorithm has gained importance in materials research very rapidly ever since its inception by John Holland in 1975. GA can be adapted to the given task and is thus, a problem specific algorithm.

The genetic operators which need to be essentially defined for a representation are: initialization, mutation, cross-over and comparison. There are two types of GA techniques used, namely, the generational GA (the entire populations are replaced with each iteration) and the steady state GA (a few members are replaced with each generation).
In our present work, we have used the latter and the reason behind our choice will become obvious in the lines to follow. The cycle of stages involved in a simple GA run include:

(a) Creation of a random population of \( n \) chromosomes in the form of bits (sequences of zeroes and ones) of specified length\(^ {16,17} \); (b) Evaluation of the fitness \( f(x) \) of each of the chromosome in the population; (c) Creation of new population by repeating the following steps until the new population is complete: (i) Selection, including selection of two parent chromosomes from a population according to their relative fitness values (the better the fitness, more chances to be selected), (ii) Cross-over, involving a crossing over between the two parents selected so as to form new off-springs, (iii) Mutation, involving the mutation of new off-spring to create new individual, (iv) Accepting or placing new off-spring in a new population, (v) Use of the newly generated population for a further run of the algorithm, (vi) If the termination criteria are fulfilled, the GA stops and returns the best solution in the current population, (vii) Return to step (b).

In our present work, we have applied the GA to investigate the electronic properties and band structures of the aperiodic model polypeptide chains consisting of glycine and alanine units, initially at the \( \text{ab initio} \) Hartree-Fock level\(^ {18} \) and subsequently, the resulting band structures have been corrected for correlation (using quasi particle band structure data). The calculations have been performed using Clementi’s basis sets and the effects of both MB (minimal basis) as well as DZ (double zeta) basis sets on the band gap values have been studied. We have also investigated the effect of change in secondary structure of the polypeptide chains, viz., the \( \alpha \)-helix and \( \beta \)-pleated sheet structures along with the effects of electron correlation and hydration on DOS (density of states) of model polypeptide chains.

**Methodology**

Several studies have reported the application of GA technique for designing of polymer chains\(^ {10,19,20} \). Our purpose is to find the optimal relative concentrations of the two constituent homopolymers, A (polyglycine) (= \( x \) %) and B (polyalanine) (= \( y \) %) in the copolymer \( \text{A}_x\text{B}_y \) using the GA technique, such that the copolymer so formed has a minimum band gap value and maximum electronic delocalization. Taking \( x + y = 100 \), the value of ‘\( x \)’ hence obtained (when varied in steps of 1%) would be the optimal percentage of homopolymer A in the copolymer \( \text{A}_x\text{B}_y \). A systematic analysis of each of the possible percentages of A and B starting from 1% and going up to 99% would be computationally expensive. The GA technique (Scheme 1) is useful in finding the optimal solutions through intelligent searches with selective sampling of values in the entire configuration space.

In this work, a population consisting of five individuals (or chromosomes) was used for the sake of simplicity and hence the GA used is called a micro-genetic algorithm. Each chromosome was defined as a sequence containing 7 bits and when converted to decimal form, each chromosome represented a particular value of percentage ‘\( x \)’ (and hence \( y = 100-x \)) of homopolymer A in the copolymer. The selection scheme used was

\[
\begin{align*}
\text{Polyglycine (A)} & \\
\text{Polyalanine (B)}
\end{align*}
\]
tournament selection, in order to identify the individual possessing best fitness value. Thereafter, single point cross over was implemented taking the sixth bit position (out of 7 bits) as the point of cross over (i.e., swapping of last two bits). The mutation operator was not used because the number of chromosomes (size of population) is small and the population may become homogenous rapidly. If such a situation surfaces then a few (here four) random individuals would have to be generated again. However, if the new population consisted only of new off-springs, the best chromosome from the last population would be lost. Thus, elitism was used, i.e., at least one best solution was copied without changes to a new population so that the best solution found survived to the end of the run. This is a powerful strategy because cross-over of the fittest chromosome is more valuable than cross-over involving any other chromosome in the new population.

From GA, as a first step, once a population of chromosomes was generated defining the relative concentrations of homopolymers A and B in the copolymer; a long polymeric chain (~300 monomer units) corresponding to each of the binary chromosome (or individual) of the population was generated using random number generator functions. The Hückel determinants of these polymeric sequences were then constructed. If the aperiodic copolymer chain consisted of N units, then the Hückel determinant of order N × N would be of the following type (using tight binding approximation),

$$\begin{bmatrix}
\alpha_A & \beta & 0 & \cdots & 0 \\
\beta & \alpha_B & \beta & \cdots & 0 \\
0 & \beta & \alpha_A & \beta & \cdots \\
0 & 0 & \beta & \alpha_A & \beta \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \alpha_B
\end{bmatrix} = 0$$

where the diagonal Hückel parameters \(\alpha\) were determined on the basis of the weighted middle point of the band structure under consideration for each of the two components. The band structures were obtained as a result of \textit{ab initio} LCAO (linear combination of atomic orbitals) CO (crystal orbital) calculation for each component applying periodic boundary conditions, \(\psi_j = \sum_{j=1}^{n} c_j \varphi_j\), where \(\psi_j\) is the \(j^{th}\) molecular orbital, \(\varphi_j\) is the atomic orbital for the \(j^{th}\) atom, and \(c_j\) is the coefficient of the \(j^{th}\) atomic orbital in the \(j^{th}\) molecular orbital. Here \(r\) varies from 1 to \(n\) (the number of atomic orbitals). For off diagonal elements assuming the simple relationship,

\[\epsilon_{A}(k) = \alpha_A + 2\beta_{A,A} \cos(ka),\]

for energy dispersion, one fourth of the band width was taken if the same unit was repeated and the equation,

\[\beta_{A,B} = \frac{1}{2} (\beta_{A,A} + \beta_{B,B})\]

was applied, if a unit A was followed by a unit B.

The next task was to calculate the eigenvalues of energy so as to obtain the band gap and hence the DOS for the polymeric chain. For this purpose, GA was coupled with the NFC (negative factor counting) method which is an efficient technique based on Dean’s negative energy eigenvalue theorem to compute the eigenvalues of a real symmetric matrix.

**Results and Discussion**

In the present calculations, the \(\alpha\) and \(\beta\) values for the valence and conduction bands of polyglycine and polyalanine obtained from their band structure results using MB set and DZ basis set by \textit{ab initio} Hartree-Fock crystal orbital method have been used (Table 1). The values of the most important electronic property, i.e., the band gap \((E_g)\) obtained using GA technique as well as systematic search, are listed in Table 2 along with the IPN (inverse participation number) values obtained for different systems. The IPN is a measure of the level of delocalization of an MO (molecular orbital) in the polymeric chain. It can be obtained from the relationship,

\[I_j = \frac{\left(\sum_{r=1}^{n} |c_j|^2\right)^2}{\left(\sum_{r=1}^{n} |c_j|^2\right)^2}\]

using the values of the normalized coefficients as input. The normalized coefficients are actually the eigenvectors in the form of a column matrix, the elements of which are obtained after several iterative steps so as to attain self-consistency. IPN \((I_j)\) cannot be zero, as that would mean that all coefficients \((c_{jr})\) in the expression for \(I_j\) are zeroes at the same time,
which amounts to the wave function being non-existent. Also, IPN cannot be one, as that would mean that the cross/ overlap terms for the polymer chain are zeroes and therefore, there is contribution from one unit only. This would imply that the wave function is completely localized on one unit and there is no contribution from the orbital on the neighbouring unit in the chain, which is again not feasible. Thus, IPN can only take up values between 0 (maximum delocalization) and 1 (localized over only one orbital) and can neither be zero nor one. For calculation of eigenvectors, we have used IIM (inverse iteration method)\(^{25, 27}\). IIM is generally used when one already has good eigenvalues (obtained from NFC method) and only a few selected eigenvectors corresponding to the energy values of interest are required, viz., the energy levels in the upper regime of the occupied valence band and those in the lower regime of the unoccupied conduction band. Once the normalized coefficients have been obtained, the IPN can be easily calculated.

Thus, the IPN value is returned to the GA program and then the fitness function \(f(x)\) is evaluated for each chromosome (as mentioned in step (b) above) to complete the GA run. The fitness function used herein has been defined such that it takes into account both band gap value (between zero and \(\rho\)) and the corresponding IPN value (between 0.0 and 1.0) of that chromosome while giving them the same statistical weight,

\[
\frac{1}{(1/\rho) \text{ gap} + \text{IPN}}
\]

where \(\rho\) is the difference between the HO energy level (of VB) and the LU energy level (of CB).

<table>
<thead>
<tr>
<th>System</th>
<th>Secondary structure</th>
<th>Basis set</th>
<th>Band gap (eV)</th>
<th>Band gap (eV)</th>
<th>IPN</th>
<th>Optimized composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>SS</td>
<td>GA</td>
<td>SS</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(\beta)-pleated</td>
<td>MB</td>
<td>15.647</td>
<td>15.647</td>
<td>0.005347</td>
<td>A(_1)B(_93)A(_1)B(_93)</td>
</tr>
<tr>
<td>2</td>
<td>(\alpha)-helix</td>
<td>MB</td>
<td>14.743</td>
<td>14.743</td>
<td>0.007391</td>
<td>A(_1)B(_99)A(_1)B(_99)</td>
</tr>
<tr>
<td>3</td>
<td>(\beta)-pleated</td>
<td>DZ</td>
<td>14.322</td>
<td>14.322</td>
<td>0.005025</td>
<td>A(_1)B(_99)A(_1)B(_99)</td>
</tr>
<tr>
<td>4</td>
<td>(\beta)-pleated</td>
<td>DZ + W</td>
<td>12.958</td>
<td>12.958</td>
<td>0.005356</td>
<td>A(_1)B(_7)A(_1)B(_7)</td>
</tr>
<tr>
<td>5</td>
<td>(\beta)-pleated</td>
<td>DZ + QP</td>
<td>13.617</td>
<td>13.617</td>
<td>0.005237</td>
<td>A(_1)B(_3)A(_1)B(_3)</td>
</tr>
<tr>
<td>6</td>
<td>(\beta)-pleated</td>
<td>DZ + QP + W</td>
<td>11.759</td>
<td>11.759</td>
<td>0.005414</td>
<td>A(_1)B(_8)A(_1)B(_8)</td>
</tr>
</tbody>
</table>

\(^a\)W: Polymer in presence of water, QP: the correlated quasi-particle band structure data, \(^b\)C: conduction band, V: valence band

Table 2 – Calculated gap values (in eV), inverse participation number (IPN) and the optimized compositions of the aperiodic model polypeptide chains, as obtained from genetic algorithm (GA) and systematic search (SS). [A=glycine unit, B=alanine unit]
amongst the two homopolymers A and B of the copolymer formed. Thus, the resulting value of \( f(x) \) caters to our requirements, viz., the minimum band gap and the maximum possible electronic delocalization. The population is allowed to evolve (over several iterative steps) until the convergence criteria is reached. To exit GA, one of the several known ways is to terminate/exit after a set number of iterations (or generations) when there is no more variation observed in the best fitness value (or the fittest individual). We have set this number as 15, i.e., if the best fitness value does not change over 15 generations, the GA is said to have converged to an optimal solution. Thus, the GA technique returns to us the optimized values of the relative concentrations of the two constituent homopolymers; the values obtained for various systems are listed in Table 2.

Since there is random sequencing of amino acid units (monomer units) in the polypeptide chain, the respective environments of alanine and glycine keep on changing and therefore, their energy positions (peaks) are scattered over a wide range of energy.

**Effect of basis set**

To see the dependence of the electronic DOS on the basis set, the DOS have been determined using band structure results of polyglycine and polyalanine using Clementi’s DZ basis set\(^{28}\). The results (system 3) are given in Table 2. Comparison of these results with those obtained using MB set (Table 2; system 1) shows that the fundamental band gap decreases with the use of a better basis set. This is because in a minimal (single) zeta basis set (Fig. 1(a)) only those orbitals that are occupied in the isolated atoms are considered to compose molecular orbitals, whereas, in an extended basis set, each valence orbital is supplemented by a delocalized orbital of the same quantum number. This allows more flexibility in the basis set and hence leads to lower total energy through SCF procedure. The decrease in the fundamental band gap value with the use of a better basis set (DZ) is also depicted from the fact that the VB shifts up significantly while the CB moves down (Fig. 2) leading to a net decrease in the band gap (when compared to MB set (Fig. 1(a))). Here it would be worth mentioning that the \( ab \) \( \textit{initio} \) Hartree-Fock method overestimates the band gap value due to the neglect of correlation effects and the use of MB set. With the use of better basis sets and taking correlation effects into account (discussed below), the calculated band gap value is expected to decrease and come closer to the experimental values.

**Effect of conformation**

To investigate the effect of the conformation of polypeptide chain on the electronic DOS, the DOS for both \( \alpha \) (Fig. 1(b)) and \( \beta \) (Fig. 1(a)) helices of the polypeptide chain have been calculated using Clementi’s MB set (systems 2 and 1 respectively). Band gap (Table 2) in \( \alpha \)-helix structure is found to be slightly lower than that in \( \beta \)-pleated structure because of comparatively lower VB while the CB remains almost at the same level.

**Effect of electron correlation**

It is well known that in theoretical quantum chemistry studies, to obtain reasonable results for most of the properties of the system, we need to take into account the electron correlation because one of the principal deficiencies of the \( ab \) \( \textit{initio} \) Hartree-Fock crystal orbital method is the neglect of correlation between motions of electrons with opposite spins. The result is that \( ab \) \( \textit{initio} \) Hartree-Fock method overestimates the band gap values. Correlation energy is the difference between the eigenvalue of energy obtained after solving the Schrödinger equation for the system in a given state (which takes into account all the electrons in the given state and their interaction energies) and the energy of the \( ab \) \( \textit{initio} \) Hartree-Fock limit. We have taken into consideration the correlation effects in the Möller Plesset form of the many body perturbation theory in second order via the Green’s function method.

The effect of correlation on the energy bands is shown by the DOS curves of the homopolypeptides using DZ basis set (Figs 2 & 3). The CB moves down, the VB moves up, and as a result, the quasi particle band gap becomes smaller than the \( ab \) \( \textit{initio} \) HF band gap. These results prove that with the use of better basis sets and consideration of electron correlation effects, the calculated values of band gap decrease (comparison of systems 3 and 5; the latter takes correlation correction into account) and come closer to the experimental values. For system 5, our calculations using systematic search (SS) gave values of IPN and optimized composition (Table 2) which were slightly different from those obtained from a GA run. We decided to probe further into this observation and found that corresponding to the GA optimized composition \( A_{55}B_{5} \), the band gap and IPN values are exactly the same from SS as well (13.617 eV and
Fig. 1—Electronic DOS curves of aperiodic poly(glycine, alanine) copolymer in the valence and conduction band regions in (a) β-pleated sheet, and, (b) α-helix structures, calculated using MB set for the \textit{ab initio} Hartree-Fock band structure data.

Fig. 2—Electronic DOS curves of aperiodic poly(glycine, alanine) copolymer in the valence and conduction band regions in β-pleated sheet structure, calculated using DZ basis set for the \textit{ab initio} Hartree-Fock band structure data.

Fig. 3—Electronic DOS curves of aperiodic poly(glycine, alanine) copolymer in the valence and conduction band regions in β-pleated sheet structure calculated using DZ basis set for the quasi-particle band structure data.
0.005237 respectively). However, the value of fitness function for the $A_{95}B_5$ system varied slightly in the fourth place of decimal (0.994724) from its counterpart in the $A_{99}B_1$ system (0.994933).

**Effect of hydration**

Since biomolecules like proteins and polypeptides are present in our body in an aqueous medium, therefore, electronic structure calculations on these biomolecules need investigation in the presence of solvation shell. We have shown herein, the effects of the solvation shell on the band structures of the polypeptides in terms of band shifts as well as change in band widths and the subsequent effect on the DOS plots.

The simulations of the solvent structures around biopolymers as reported by Liegener\textsuperscript{29} obtained by Monte-Carlo calculations have been used, where the band structures of the homopolypeptides have been calculated in the effective field of water molecules. In these band structure calculations, each water molecule was represented by 23 point charges contributing to the one electron matrix in the crystal orbital calculations. In Table 1, we list the various values of $\alpha$ and $\beta$ for glycine (gly) and alanine (ala). For all systems the data obtained without water interaction and with water interaction (W) are given for the valence band (V) and the conduction band (C). By looking at both the systems, (comparison between results of systems 3 & 4 and 5 & 6 and their corresponding DOS curves in Figs 2 and 4 and 3 and 5 respectively) one can easily infer that the bands are shifted by the water point charges. It can be seen that hydration has a much stronger influence on width of the DOS curves and a large downward shift can be observed in the CB in the presence of surrounding water molecules. This is obviously due to the reason that the different chemical nature of the side chain (-H in glycine and -CH\textsubscript{3} in alanine) leads to an altogether different arrangement of the surrounding water molecules in their vicinity, which has different effects on the corresponding energy bands. Thus, we can conclude that hydration also strongly influences the intrinsic conduction properties of aperiodic polypeptide chains.

In all the above cases, it was observed that the results obtained from GA as well as from systematic search are in very good agreement with each other. Hence the results obtained verify that the

![Fig. 4](image1.png)

**System 4**

*Fig. 4*— Electronic DOS curves of aperiodic poly(glycine, alanine) copolymer in the valence and conduction band regions in $\beta$-pleated sheet structure, calculated using DZ basis set for the \textit{ab initio} Hartree-Fock band structure data of the hydrated polymers.

![Fig. 5](image2.png)

**System 6**

*Fig. 5*— Electronic DOS curves of aperiodic poly(glycine, alanine) copolymer in the valence and conduction band regions in $\beta$-pleated sheet structure calculated using DZ basis set for the quasi-particle band structure data of the hydrated polymers.
methodology of GA is a very efficient tool for generating an optimal solution to the problem of theoretical “tailoring” of polypeptides chains with pre-specified conduction properties. Moreover, the GA technique has an advantage that it saves a lot of computational time and gives directly the value of optimized composition of the copolymer which possesses the minimum band gap along with maximum electronic delocalization without compromising on the accuracy of results. What GA does in a single run is reproduced by systematic search in 99 runs; thus saving a lot of computer time and memory.

In our pilot approach we have focused on binary component polypeptide chains. The studies are being extended to ternary and higher multi-component copolymer chains towards application to DNA and protein chains (which are made of a number of randomly arranged amino acids) and to the problem of tailoring of electrically conducting polymers.

Acknowledgement

One of the authors (VA) is grateful to CSIR (Council of Scientific and Industrial Research), New Delhi, for the award of Junior Research Fellowship (JRF).

References