J. Chem. Sci. Vol. 124, No. 1, January 2012, pp. 149–158. © Indian Academy of Sciences.

Tailoring approach for obtaining molecular orbitals of large systems[#]

ANUJA P RAHALKAR[a,b] and SHRIDHAR R GADRE[a,*]

[a]Department of Chemistry, Indian Institute of Technology Kanpur, Kanpur 208 016, India
[b]Department of Chemistry, University of Pune, Ganeshkhind, Pune 411007, India
e-mail: gadre@iitk.ac.in

Abstract. Molecular orbitals (MO's) within Hartree-Fock (HF) theory are of vital importance as they provide preliminary information of bonding and features such as electron localization and chemical reactivity. The contemporary literature treats the Kohn–Sham orbitals within density functional theory (DFT) equivalently to the MO's obtained within HF framework. The high scaling order of *ab initio* methods is the main hurdle in obtaining the MO's for large molecular systems. With this view, an attempt is made in the present work to employ molecular tailoring approach (MTA) for obtaining the complete set of MO's including occupied and virtual orbitals, for large molecules at HF and B3LYP levels of theory. The energies of highest occupied and lowest unoccupied molecular orbitals, and hence the band gaps, are accurately estimated by MTA for most of the test cases benchmarked in this study, which include π-conjugated molecules. Typically, the root mean square errors of valence MO's are in range of 0.001 to 0.010 a.u. for all the test cases examined. MTA shows a time advantage factor of 2 to 3 over the corresponding actual calculation, for many of the systems reported.

Keywords. Molecular orbitals; orbital energies; highest occupied molecular orbital (HOMO); lowest unoccupied molecular orbital (LUMO); band gap; molecular tailoring approach; Hartree-Fock (HF) theory; density functional theory (DFT).

1. Introduction

Molecular orbital (MO) theory[1,2] provides a basic framework for understanding the chemical bonding and the reactivity patterns of a molecule. In particular, the nature of the highest occupied (HOMO) and the lowest unoccupied (LUMO) molecular orbitals is of vital importance.[3] The extent and phases of HOMO and LUMO shed light on the molecular reactivity. The energy gap between HOMO and LUMO is useful for estimating chemical properties of molecules such as conductivity, reactivity, etc.[4]

There is vast literature available on the calculations of MO's based on first principle electronic structure theory and the comparison with relevant experimental data.[5,6] *Ab initio* methods are known to be generally reliable for treating molecular systems, with a goal to obtain their structures and electronic properties. However, the main obstacle in using these methods is their high scaling order. For the basic Hartree-Fock (HF) method, the formal scaling is $O(\mathrm{N}^4)$, although for large systems the method is known to scale[7] as $O(\mathrm{N}^3)$, where N is the number of basis functions. The scaling order for more accurate post-HF methods is even higher. For instance, the Møller-Plesset second order Perturbation (MP2) theory scales as $O(\mathrm{N}^5)$, and with better correlated methods the scaling is even higher. Another popular method in molecular quantum chemistry is the density functional theory (DFT). The role played by MO's within HF theory is played by the Kohn-Sham orbitals within DFT framework. The scaling of DFT is known to be similar to that of the HF theory. The problem of high scaling order of *ab initio* methods leads to tremendously large computational requirements with increasing system size. In view of this, many fragment based methods are being developed in order to enable computations on large systems at *ab initio* level.

Christoffersen *et al.* pioneered[8] the idea of fragment-based strategies. This method was implemented for some large organic systems employing one or two floating Gaussians per electron pair or a sub-minimal basis set such as STO-2G. After many years, Yang[9] suggested another fragment-based method christened by him as divide-and-conquer (DC) method. This method was benchmarked for a tetra-peptide of glycine. A few years later, Gadre and co-workers came up with an independently developed fragmentation-based method *viz.* molecular tailoring approach (MTA).[10–18] In this approach, a large molecule is cut into a set of small overlapping fragments, on which the calculations are performed at the desired level of theory. The results of

[#]Dedicated to Prof. N Sathyamurthy on his 60th birthday
[*]For correspondence

these fragments are employed for estimating the corresponding entities of the parent large molecule. In the initial days of MTA, this strategy was employed[10–12] to acquire a good quality density matrix (DM) for the parent molecule, which was further used for enumerating one-electron properties such as molecular electrostatic potential (MESP), molecular electron density (MED), dipole moments, etc. These properties and the topography of the corresponding scalar fields were computed for a variety of molecular systems, e.g., α-tocopherol, ZSM-5 silicalite zeolite. MESP analysis was also performed on a model polypeptide at HF/6-31G** level. A few tests performed at MP2 level yielded good results in terms of the accuracy as well as efficiency. Further, the method was extended[13] to optimization of geometry by accurate gradient estimations. However, the energy estimates were only qualitative. The MTA-based energy as well as gradients were improved[14] to higher accuracy by incorporating set inclusion-exclusion principle while patching the results. Thorough benchmarking[14] has shown that the MTA-based energy estimates for spatially extended molecules are accurate to within 1.5 mH ($\sim$1 kcal/mol) as compared to their conventional counterparts. Also, the numerically significant gradients (typically greater than 10^{-3} a.u.) are produced with sufficient accuracy. This cardinality-guided approach was further extended[15] for calculating the Hessian matrix and IR spectra of large molecules at HF, DFT as well as MP2 levels of theory. The Hessian matrix synthesized by MTA is highly accurate. More explicitly, the RMS deviation in MTA-based Hessian elements was approximately 10^{-4} a.u. for all the test cases examined[15] at HF, DFT and MP2 levels of theory. Computation of Hessian being expensive and time-consuming task, MTA showed[15] high time advantage factors over the conventional calculations. In the current form, MTA can be applied only to closed shell species. However, the level of theory and basis set are no bar! MTA can be successfully applied to any class of molecules from atomic/molecular clusters[16] to biological molecules to highly conjugated molecules such as β-carotene and model of graphene[17], etc. It can also handle systems with charged centres.[14,18] In summary, MTA offers an attractive tool for handling large spatially extended systems at *ab initio* level.

Another well-established fragment-based method is the fragment molecular orbital method (FMO),[19,20] pioneered by Kitaura *et al*. Within the FMO framework, the criterion for fragmentation is based on many-body decomposition energy analysis. FMO also employs electrostatics while performing calculations on fragments.The acronym FMO is suggestive of the ability of the method to yield MO's of the molecule under study. However, the method in its popular form does not directly permit calculation of MO's. Some of the actively working groups have recently attempted to extend the FMO methodology for extracting all the MO's or at least the valence MO's for the whole molecule. Inadomi *et al.*[21] have proposed a method called FMO-MO, which is an extension of FMO method. In this method, the density matrix of the parent molecule is obtained from FMO fragments and the total Fock matrix is constructed by performing one extra SCF iteration for the whole system.[19,21] Tsuneyuki and coworkers[22] have come up with another extension of FMO which is called as FMO-LCMO. In this FMO-LCMO method, one-electron Hamiltonian is constructed from the density matrix of each fragment. In order to reduce the large dimension of one-electron Hamiltonian matrix for large biomolecules, it has been assumed that the most significant contributions to HOMO and LUMO come from the corresponding MO's of monomer fragments defined by FMO. They have tested this method for a pseudo-Glycine pentamer with a minimal STO-3G basis set. Also, this method seems to be useful only for obtaining a few MO's lying in the vicinity of the HOMO and LUMO.[19,22]

Among the other fragment based methods, Li and coworkers proposed[23] a localized molecular orbital assembler approach for HF level to obtain the density matrix for a large molecule, which is further used to get estimate for the total energy of the molecule. However, it is not tested for the quality of MO's. Other fragmentation-based or DC-type methods by Nakai *et al.*,[24] Bettens[25] and Collins[26] have not been applied for calculation of MO's and MO energies.

In the present work, we report an exploratory study, within MTA framework, for obtaining the set of all occupied as well as virtual MO's for large spatially extended molecular systems. In this methodology, the DM for the whole molecule is assembled by MTA, and it is expected to be a good quality DM. This DM is given as a guess and a single SCF iteration on the parent molecular system is performed in conventional way to obtain the complete set of MO's of the parent molecule.

2. Methodology and computational details

Molecular tailoring approach divides the large parent molecule into a set of overlapping fragments. It employs the distance criterion for fragmentation, the justification for which is offered by the 'near-sightedness' principle by Kohn.[27] The details of fragmentation and the parameters to control accuracy, etc.

are described elaborately in reference 14. The calculations are performed on the fragments and the results of fragments are used for extracting the estimate of the concerned molecular property for the parent system. Presently, the method only supports closed shell calculations. The simple set of inclusion–exclusion principle is put to use for stitching back the fragment results. For example, the energy of parent system is estimated as,

$$\begin{aligned} E &= \sum E^{f_i} - \sum E^{f_i \cap f_j} + \ldots \\ &\quad + (-1)^{k-1} \sum E^{f_i \cap f_j \cap \ldots \cap f_k}. \end{aligned} \tag{1}$$

Here, E stands for the MTA-based estimate of the energy of the parent molecule, E^{fi} denotes the energy of fragment i; $E^{fi \cap fj}$ denotes the energy of the binary overlap of fragments i and j and so on.

Similarly, the expressions for gradients and Hessian matrix elements of the parent molecule are given as,

$$\begin{aligned} \frac{\partial E}{\partial X_i} &= \sum \frac{\partial E^{f} i}{\partial X_i^{fi}} - \sum \frac{\partial E^{fi \cap fj}}{\partial X_i^{fi \cap fj}} + \ldots \\ &\quad + (-1)^{k-1} \sum \frac{\partial E^{fi \cap fj \cap \ldots \cap fk}}{\partial X_i^{fi \cap fj \cap \ldots \cap fk}} \end{aligned} \tag{2}$$

$$\begin{aligned} \mathbf{H}_{ab} &= \sum \mathbf{H}_{ab}^{f_i} - \sum \mathbf{H}_{ab}^{f_i \cap f_j} + \ldots \\ &\quad + (-1)^{k-1} \sum \mathbf{H}_{ab}^{f_i \cap f_j \cap \ldots \cap f_k} \end{aligned} \tag{3}$$

Terms in Equations (2) and (3) are respectively the first and second order derivatives of the energy terms, with respect to nuclear coordinates, with $\mathbf{H}_{ab}$ denoting the element (a, b) of the Hessian matrix.

On the same lines, as given in the early days of MTA, the expression for the density matrix (DM) is,

$$\mathbf{P}_{ab} = \sum \mathbf{P}_{ab}^{fi} - \sum \mathbf{P}_{ab}^{fi} + \ldots + (-1)^{k-1} \sum \mathbf{P}_{ab}^{f_i \cap f_j \cap \ldots \cap f_k} \tag{4}$$

where, $\mathbf{P}_{ab}$ denoting the element (a, b) of the density matrix, $\mathbf{P}_{ab}^{f_i}$ denoting the element (a, b) taken from the fragment i and so on.

This DM is then fed to GAMESS[28] and a single SCF iteration is performed to construct Fock matrix (Equation 5) and to obtain MO's following the standard Roothaan-Hall equations (Equation 6) for closed shell case.

$$\mathbf{F}_{\mu v} = \mathbf{h}_{\mu v} + \sum_{\rho\sigma} \mathbf{P}_{\rho\sigma} \left[2\,(\mu v|\rho\sigma) - (\mu\rho|\sigma v)\right] \tag{5}$$

$$\mathbf{FC} = \mathbf{SC}\,\boldsymbol{\epsilon}\,, \tag{6}$$

$\mathbf{F}_{\mu v}$ and $\mathbf{h}_{\mu v}$ denote the (μ, v)th element of the Fock matrix and the one-electron Hamiltonian matrix. Similarly, $\mathbf{P}_{\rho\sigma}$ denote the (ρ, σ)th element of the DM. Here $(\mu v|\rho\sigma)$ denotes the 2-electron integrals involving basis functions μ, v, ρ and σ. **F**, **C** and **S** are Fock-, coefficient- and overlap matrices, respectively, while $\boldsymbol{\epsilon}$ is the diagonal matrix with the orbital energies as diagonal elements. All these steps are parallelized in line with the earlier works[29] from our group.

For MTA-based MO calculation, a single-point HF-level energy calculation is performed on the set of fragments. The DM's of these fragments are used for synthesizing the DM for the whole molecule as per

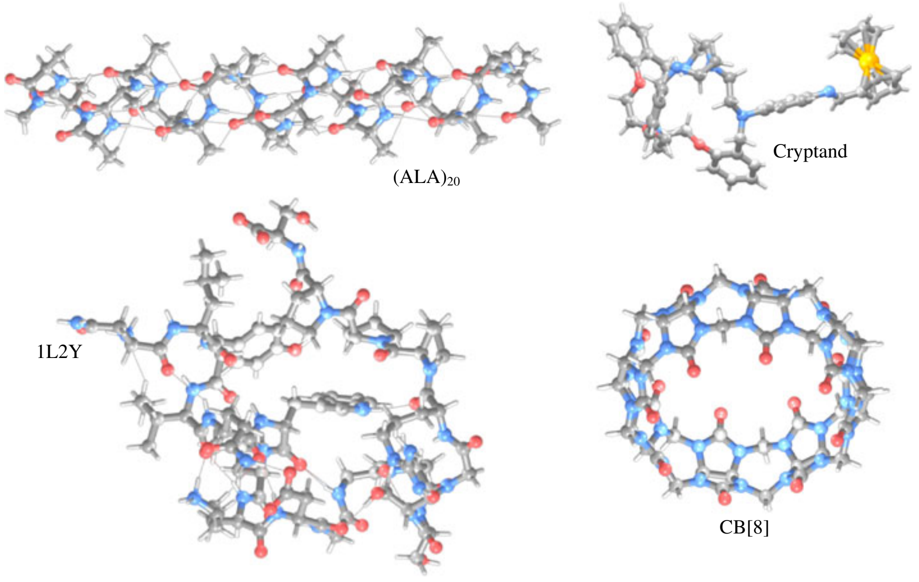


Figure 1. Geometries of helical form of polypeptide of alanine (ALA)$_{20}$, cryptand ligand containing a ferrocene unit, TrpCage protein (1L2Y) and cucurbit[8]uril (CB[8]), which form test cases at HF level of theory (see text for details).

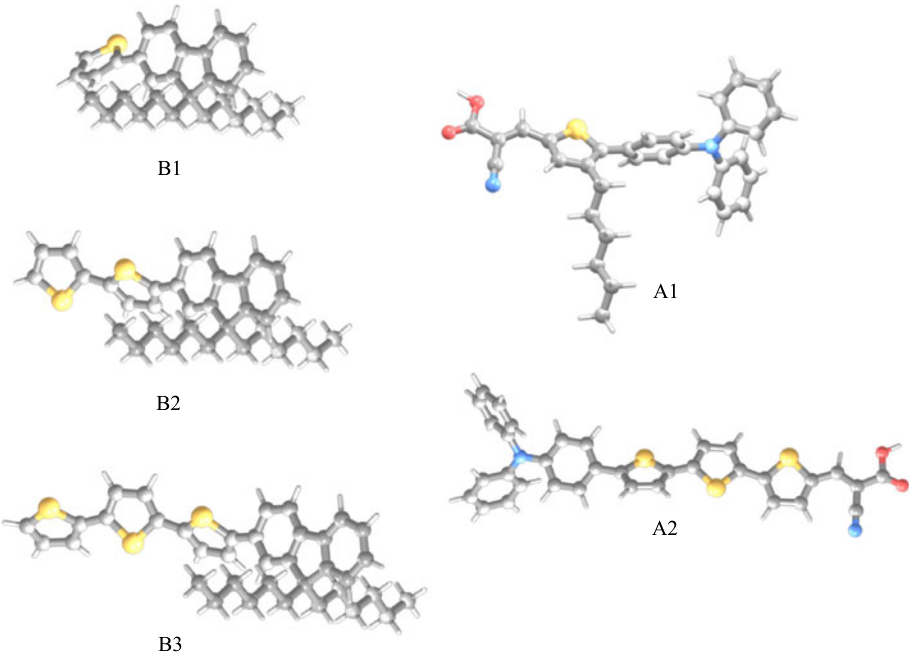


Figure 2. Conjugated organic materials from Reference 5, tested at B3LYP/ 6-311+G* level of theory. A1 and A2 belong to subclass A; B1, B2 and B3 belong to subclass B (see text for details).

Equation 4. The Fock matrix is constructed by employing Equation 5 for which all the electron repulsion integrals are to be evaluated only once. Further, the set of MO's for the parent molecule is obtained from Equation 6. Within DFT, the Kohn-Sham orbitals are treated akin to MO's within the HF theory. In this work, these two terms will be treated as equivalent, as per the contemporary literature. This method of calculating MO's is benchmarked for several structurally and chemically diverse molecular systems and clusters at HF and Becke-3-parameter-Li-Yang-Parr (B3LYP) functional within DFT framework.

The HF being the simplest method, initial testing of the above mentioned algorithm is performed at HF level. The test cases (see figure 1) at HF level include novel host cucurbit[8]uril, a cryptand ligand containing a ferrocene unit, an alanine polypeptide consisting of 20 alanine units and a small protein (1L2Y). More thorough benchmarks are performed at B3LYP level as it is well-known for producing reliable optical properties.[30] All the calculations at B3LYP level are performed employing 6−311+G* basis set, except for some systems as specified in the next section. The test cases for B3LYP are mostly π-conjugated systems such as thiophene containing molecules (*cf.* figure 2), a small model of graphene, β-carotene (figure 3). A covalently bonded cluster which has potential use in semiconductor material *viz.* $(ZnS)_{24}$ (figure 3) is also included. All the images in figures 1–4 are generated using

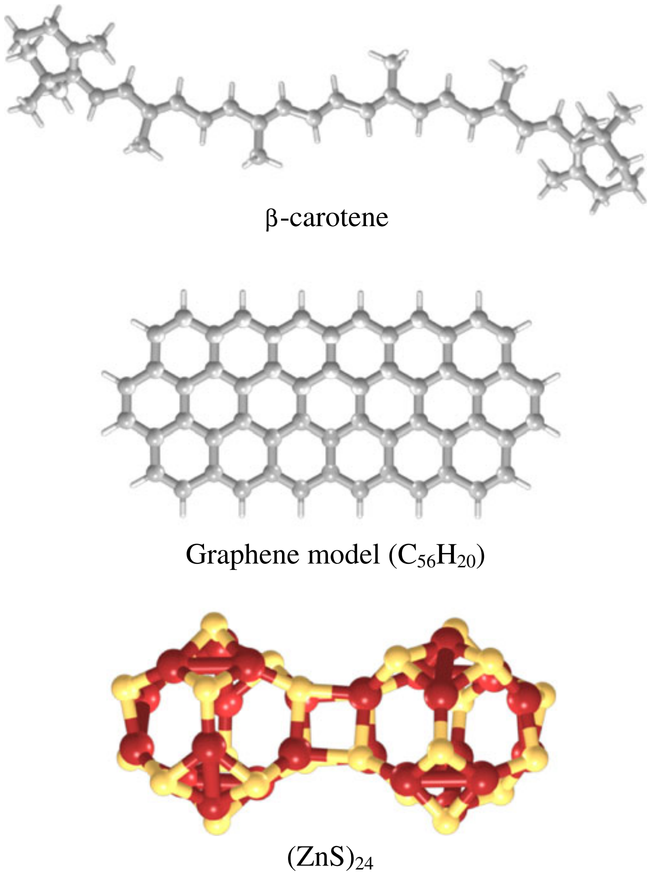


Figure 3. Conjugated organic molecules β-carotene, small model of graphene tested at B3LYP/6-311+G* and a covalently bonded cluster of $(ZnS)_{24}$ tested at B3LYP/6-31G* level of theory (see text for details).

MeTAStudio,[31] an open source programmable interface for computational chemists. All the calculations are performed on a cluster of Core 2 Quad (Q9650) @ 3.00 GHz with 8 GB RAM. For comparison purpose, the elapsed time for all the calculations is converted into time that would be taken if the calculation had been performed on one machine of the same specification.

3. Results and discussion

HF being a preliminary *ab initio* method, the initial testing of the propsed MTA-based MO methodology is performed at HF level. The test cases chosen for HF level contain about 100 to 300 atoms. Polyalanine in helical form consisting of 20 alanine units, $(ALA)_{20}$, has strong dipole effects, which makes it difficult to fragment in such a way as to produce accurate results.[18] For accurate energy estimation by MTA, this system needs to be scissored into larger fragments, and there is only little time advantage over the corresponding actual calculation. In this work, we explore the possibility to economically produce reliable MO's from MTA-based DM. It is seen from the earlier works[10–12] that the MTA-based DM is accurate enough even if the fragments are of moderate size. Thus, $(ALA)_{20}$ is cut into 5 main fragments with average size of 71 atoms. The R-goodness of the fragmentation scheme is 2.3 Å. For details of R-goodness parameter and its correlation with accuracy of MTA, see to Reference 14. Also, it has been shown earlier[18] that for a particular fragmentation scheme within MTA, the errors in MTA-based energy estimates for lower basis set such as STO-3G or 6-31G remain similar to those for higher basis sets involving polarization and diffuse functions. Thus, a pilot MTA-calculation is performed at HF/6-311G* basis set and compared with its actual counterpart. The HOMO-LUMO energies are in error by ∼0.008 a.u., the actual values being −0.278 and 0.043 a.u., respectively and the error in band gap is ∼0.016 a.u. for actual value of 0.321 a.u. (*cf.* table 1). Introducing a set of diffuse functions in the basis set, MTA and actual calculations are also performed at HF/6-311+G* level. In this actual calculation, the SCF did not converge in 100 iterations; hence, the results could not be compared.

Another system of biological interest is a small protein, 1L2Y, which consists of 20 amino acids (304 atoms) and is known to be the fastest folding protein.[32] This TrpCage protein has 5 charged centres and overall a unit positive charge. Again, a comparatively poor fragmentation scheme, with R-goodness of 3.3 Å and 5 main fragments with average size of 127 atoms, is employed for the MTA-based MO calculation. Due to poor fragmentation scheme, the MTA estimate of molecular energy is off by ∼22 mH (*cf.* table 1) and is beyond chemical accuracy. The energy estimate can be easily improved by employing better fragmentation scheme, however, the main purpose in this work is to obtain reliable MO's with minimal effort. As seen from table 1, the HOMO and LUMO energies are accurate, with the error in band gap being 0.001 a.u., the actual value of band gap being 0.317 a.u. Also, the MTA calculation turned out 3 times faster than the corresponding actual calculation on an identical hardware. On employing higher basis set, it is anticipated that there will be further improvement in the time advantage factor.

In order to have diversity in the systems used for benchmarking, a cryptand ligand containing a ferrocene unit is chosen. This calculation is performed at HF/6-31G* level of theory. Higher basis set involving diffuse functions could not be employed as the GAMESS package does not support this basis set for transition metals. For this system, a scheme of 5 main fragments with R-goodness of 3.7 Å is employed. The agreement in MTA energy with the actual energy is excellent (the error

Table 1. MTA-based total energies E, energies of HOMO and LUMO (E_{HOMO}, E_{LUMO}) and the band gap (E_{GAP}) for various systems (the actual values are in parentheses, all in a.u.) at HF/6-311+G* unless otherwise specified in square brackets. N_a and N_{BF} being the number of atoms and basis functions. T is the elapsed time in min and T_r is the ratio of time taken by MTA to that by actual calculation.

System	N_a	N_{BF}	E	E_{HOMO}	E_{LUMO}	E_{GAP}	T	T_r
$(ALA)_{20}$	212	2316	−5165.14847	−0.286	0.051	0.337	174	2.4
[6-311G*]			(−5165.21479)	(−0.278)	(0.043)	(0.321)	(420)	
$(ALA)_{20}$	212	2736	−5165.21217	−0.287	0.007	0.297	857	-
CB[8]	144	2352	−4787.43611	−0.391	0.033	0.424	321	-
1L2Y	304	2610	−7439.81092	−0.327	−0.010	0.317	400	3.0
[6-31G*]			(−7439.83268)	(−0.327)	(−0.011)	(0.316)	(1200)	
Cryptand	118	1040	−3740.48908	−0.287	0.105	0.392	57	0.9
[6-31G*]			(−3740.48868)	(−0.288)	(0.105)	(0.393)	53	

is 0.3 mH). The MTA-based LUMO energy matches exactly with the actual value of 0.105 up to 3 digits after decimal. Also, the HOMO energy and the band gap matches well with their actual counterparts, *viz.* −0.288 and 0.392 a.u., the errors being ∼ 0.001 a.u. The system being moderate in size (118 atoms), the MTA-based calculation did not give any time advantage over the actual calculation. However, this case clearly brings out the applicability of MTA-based MO method to large complexes involving transition metal ions, as band gaps for such systems are of interest to chemists.

Although HF provides a basic framework for MO theory, it is well-known that the Kohn-Sham orbitals of B3LYP functional in DFT are more reliable for estimating band gaps and other properties of molecular orbitals. Thus, a more thorough assessment of MTA-based MO calculation is performed at B3LYP level of theory for a variety of molecules, especially π-conjugated systems. A class of organic conjugated molecules is selected as test case. Subclass A contains 2 molecules, A1 and A2, with 1 and 3 thiophene rings, respectively, attached to electron donor and acceptor groups for push-pull effect on π-electrons in conjugation (*cf.* figure 2). Subclass B contains extensive conjugation of thiophenes which are attached to long alkane side chains (*cf.* figure 2). The band gaps in such materials are important as these have applications in organic light emitting diodes. These systems are studied[5] by Truong *et al.* with shifted PM6 method for obtaining the band gaps of these systems and compared them with the respective experimental values. As evident from the table 2, the HOMO-LUMO energies and consequently, the band gaps of all the 5 systems in this class are quite accurately estimated. The errors in MO energies and band gap for these are of the order of 10^{-3} a.u. Also, as reported in Reference 5, the band gap decreases with increment in the number of thiophene rings in conjugation. Similar trends are obtained for subclass A as well as B. For instance, as reported in table 2, the band gaps for molecules B1, B2 and B3 are 0.157 a.u. (∼ 4.27 eV), 0.137 a.u. (∼ 3.73 eV) and 0.126 a.u. (∼ 3.43 eV), respectively. The corresponding experimental values (*cf.* table 2 in Reference 5) are 3.50, 3.06 and 2.81 in eV. Thus, with a scaling factor of 0.82, the MTA-based band gaps for B1 to B3 would be 3.45, 3.02 and 2.77 eV, which is matching very well with

Table 2. MTA-based total energies E, energies of HOMO and LUMO (E_{HOMO}, E_{LUMO}) and the band gap (E_{GAP}) for various systems (all in a.u.) at B3LYP/6-311+G* unless otherwise specified in square brackets. The actual values are in parentheses, N_a and N_{BF} being the number of atoms and basis functions. For explanation of T and T_r, see table 1.

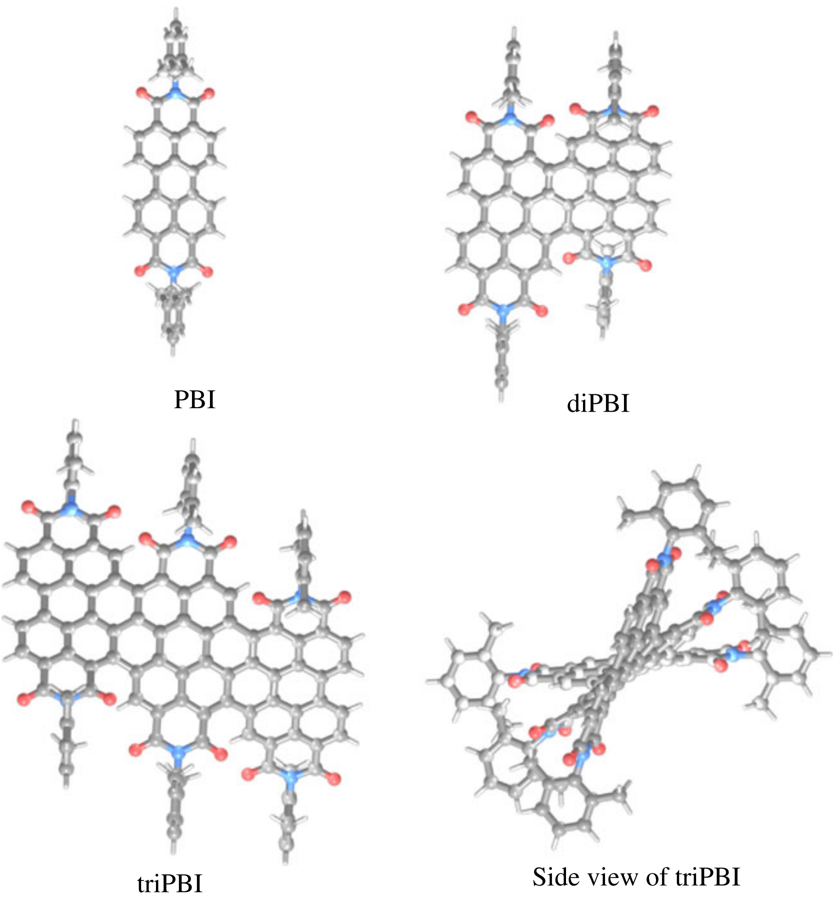
System	N_a	N_{BF}	E	E_{HOMO}	E_{LUMO}	E_{GAP}	T	T_r
A1	67	949	−1895.00017 (−1895.00047)	−0.199 (−0.200)	−0.099 (−0.102)	0.100 (0.098)	164 (452)	2.8
A2	63	1033	−2762.67610 (−2762.67662)*	−0.193 (−0.195)	−0.111 (−0.112)	0.082 (0.083)	170 (544)	3.2
B1	66	806	−1524.43339 (−1524.43541)	−0.229 (−0.211)	−0.072 (−0.053)	0.157 (0.159)	80 (300)	3.8
B2	73	935	−2076.15095 (−2076.15298)*	−0.213 (−0.202)	−0.076 (−0.061)	0.137 (0.141)	131 (420)	3.2
B3	80	1064	−2627.86813 (−2627.87018)*	−0.205 (−0.197)	−0.079 (−0.068)	0.126 (0.130)	193 (532)	2.8
β-carotene	96	1088	−1557.15076 (−1557.15137)	−0.176 (−0.178)	−0.077 (−0.079)	0.099 (0.099)	91 (200)	2.2
$C_{56}H_{20}$(S1)	76	1348	−2145.41534 (−2145.42588)	−0.136 (−0.164)	−0.0900 (−0.112)	0.046 (0.052)	214 (609)	2.9
$C_{56}H_{20}$(S2)	76	1348	−2145.49491 (−2145.42588)	−0.142 (−0.164)	−0.102 (−0.112)	0.040 (0.052)	312 (609)	2.0
PBI [6-31G]	72	466	−1948.79816 (−1948.79853)	−0.222 (−0.222)	−0.121 (−0.122)	0.101 (0.100)	14 (12)	0.9
diPBI [6-31G]	138	920	−3894.06830 (−3894.03421)	−0.223 (−0.217)	−0.140 (−0.139)	0.083 (0.078)	56 (70)	1.3
triPBI(5) [6-31G]	204	1374	−5839.25829 (−5839.26961)	−0.220 (−0.216)	0.149 (−0.145)	0.072 (0.071)	183 (228)	1.3
PBI	72	1136	−1949.77812 (−1949.77848)*	−0.229 (−0.230)	−0.130 (−0.133)	0.099 0.097	225 (-)	-
diPBI	138	2254	−3896.00091	−0.222	−0.141	0.081	3394	-
$(ZnS)_{24}$ [6-31G*]	48	1392	−52253.56729 (−52253.59269)	−0.265 (−0.264)	0.093 (0.091)	0.172 (0.173)	253 (328)	1.3

the experimental data. Further, it is worth noting here that the chosen test molecules are of moderate size (60 to 80 atoms) and still MTA-based calculation has been ∼ 2 to 3 times faster than the corresponding actual calculation.

An orange coloured pigment, β-carotene, present in many fruits and vegetables, especially in carrots, is a highly conjugated molecule. When this molecule is subjected to MTA-based MO calculation at B3LYP/6-311+G* level of theory, it produced highly accurate estimates for energy as well as HOMO–LUMO energies and hence, the band gap. It is well-known that β-carotene shows an intense absorption ∼470 nm in uv spectra. HOMO–LUMO band gap is a very crude estimate for electronic excitation. The band gap of 0.099 a.u. (table 2) corresponds to 456 nm.

A small model of graphene is an acid test for any fragment-based method. When treated by MTA method, though the single point energy was in large error, it was successfully optimized with time advantage over the actual calculation.[17] For the present method, for obtaining MO's of such two-dimensional conjugated systems, a graphene model compound is taken up for benchmarking at B3LYP/6-311+G* level of theory. A fragmentation scheme (S1) of 5 main fragments of average size ∼33 atoms is employed. As expected, the errors in orbital energies are higher when compared to all the other systems reported in this work. More explicitly, the MTA estimate (all values are in a.u.) for HOMO energy is −0.136 while the actual value is −0.164. Similarly, for LUMO energy the error is −0.022 a.u. for MTA and actual values of 0.090 and 0.112 a.u., respectively. But due to error cancellation while calculating the band gap, the error in band gap is 0.006 a.u. This calculation showed time advantage factor of 2.9 over the actual calculation. There is a need to improve these results while retaining the time advantage. Another fragmentation scheme (S2) with slightly larger fragments shows improvement in MTA-based orbital energies at the cost of decrease in the advantage factor over the actual calculation (*cf.* table 2).

Another class of interesting systems is recently reported by Negri *et al.*,[33] in which experimental and theoretical study of construction of fully conjugated n-type grapheme nano-ribbons(GNRs) is presented. These systems have high potential for technological


PBI

diPBI

triPBI

Side view of triPBI

Figure 4. Fully conjugated n-type nanoribbons, PBI, diPBI and triPBI (in helical form). Details are given in text and Reference 33.

applications involving nanoelectronics.[33] For assessment of fragmentation scheme for the smallest of these nanoribbons (PBI), the test calculation is performed at B3LYP/6-31G level of theory. It is shown in earlier reports[18] that the performance of a given fragmentation scheme (in terms of accuracy) is almost independent of the basis set used. A fragmentation scheme for PBI consists of 3 main fragments with average size of 38 atoms. The R-goodness for this scheme is 4.3 Å. At B3LYP/6-31G level, the error in MTA-based energy estimate is very small *viz.* 0.4 mH and also, the orbital energies and consequently, the band gap energy are seen to match well with their actual counterparts within 0.001 a.u. Due to small number of basis functions, this scheme took slightly more time than the actual calculation. When this calculation is performed at B3LYP/6-311+G* level, the same fragmentation scheme shows similar accuracies, confirming the basis set independence of MTA as shown in Reference 18. Time for actual calculation of PBI at B3LYP/6-311+G* is not reported as the SCF for the whole molecule is not converged in 30 iterations, which took 280 minutes on the identical hardware on which MTA calculation is run. Thus, there is time advantage shown by MTA over the actual calculation. For the case of diPBI comprising of two units of PBI and having a twist in the conjugated fused rings (*cf.* figure 4), the error in MTA-energy estimate is beyond the chemical accuracy of 1 kcal/mol. But the HOMO, LUMO energies and the band gap are accurate enough with error of about 0.005 a.u. Also, there is little time advantage factor of 1.3 (*cf.* table 2) even when lower basis set of 6–31G is employed. The largest of these graphene nano-ribbons[33] is triPBI which is in helical form as shown in figure 4. The energy and valence orbitals of this system are also estimated with excellent accuracy by MTA. With small errors of ∼0.001–0.004 a.u. in HOMO–LUMO energies and the band gap, MTA-based calculation shows time advantage factor of 1.3 over the actual calculation.

A covalently bonded cluster of 24 ZnS units is also subjected to MTA-based evaluation of HOMO–LUMO energies at B3LYP/6-31G* level. The MTA-estimated band gap for this cluster turns out to be 0.172 a.u. with the corresponding actual value being 0.173 a.u. There is a small advantage factor of 1.3 over the actual calculation as well.

The accuracies of MTA-based energy estimates corresponding to highest occupied and lowest unoccupied orbitals are discussed so far, as these are the orbitals that determine molecular reactivity. These MO-energies also provide an estimate of the corresponding ionization potentials (IP's) via Koopmans' theorem.

In addition to benchmarking the accuracies of HOMO and LUMO energies, it is worthwhile to check whether all the other MO-energies obtained via MTA faithfully mimic their actual counterparts. This test is especially meaningful for the valence orbitals lying close to HOMO and LUMO. The maximum and root mean square (RMS) errors in the valence orbital energies (which lie in range −2.0 to 2.0 a.u.) are reported in table 3 for all the test systems at HF and B3LYP levels of theory. For most of the systems, the maximum errors are well within 1% of the corresponding actual values and the RMS errors are of the order of 10^{-3} a.u. For the systems B1 to B3 which contain thiophene rings in conjugation and a model of graphene, the maximum and the RMS errors are higher *viz.* 10^{-2} a.u. pointing out the need for improvement in the methodology proposed in the present work. A pictorial representation of the comparison of MTA-estimated molecular orbital energies of some valence MO's is shown (figure 5) vis-a-vis the corresponding actual values for the cases of cryptand and graphene model. It may be noted from figure 5a that the MO energies for the case of cryptand are accurately reproduced. However, for the case of graphene, there is a constant shift in all the MO-energies (figure 5b). Due to this error cancellation, the corresponding band gap is quite well-estimated inspite of the individual MO-energies being in error. This problem of shifting of energy-levels needs to be addressed in the future studies.

Table 3. Errors in valence orbital energies within MTA-based method. Max and RMS denote the maximum and root mean square errors in MTA-estimated orbital energies as compared to the corresponding actual energies whose absolute values are greater than 2.0 a.u. These actual values are given in parentheses and basis sets in square bracktes. All the values are in a.u.

System	Max	RMS
HF		
$(ALA)_{20}$ [6-311G*]	0.010 (−1.079)	0.002
1L2Y [6-31G*]	0.008 (−1.389)	0.001
Cryptand [6-31G*]	0.001 (−1.187)	0.000
B3LYP/6-311+G*		
A1	0.004 (−0.102)	0.001
A2	0.003 (−0.033)	0.001
B1	0.028 (−0.895)	0.011
B2	0.024 (−0.901)	0.010
B3	0.023 (1.259)	0.009
$\beta-$carotene	0.003 (−0.045)	0.001
$C_{56}H_{20}$(S1)	0.029 (−0.284)	0.017
$C_{56}H_{20}$(S2)	0.022 (−0.164)	0.011
PBI[6-31G]	0.003 (1.540)	0.001
diPBI [6-31G]	0.010 (1.854)	0.003
triPBI [6-31G]	0.003 (1.849)	0.001
$(ZnS)_{24}$ [6-31G*]	0.006 (1.315)	0.002

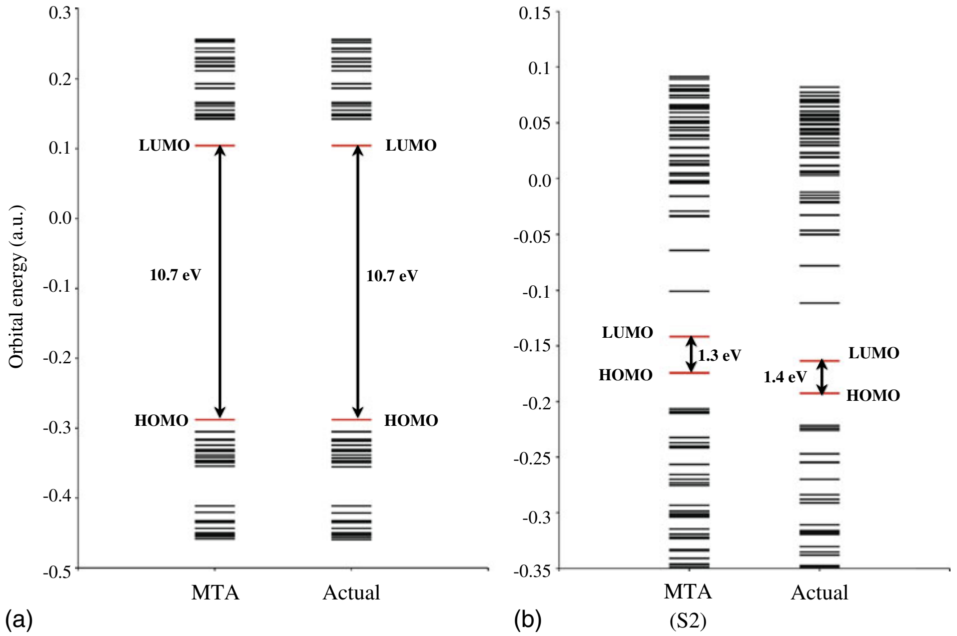


Figure 5. A comparison of MTA-based orbital energies of some valence MO's as against their actual counterparts of (**a**) cryptand ligand at HF/6-31G* and (**b**) model of graphene ($C_{56}H_{20}$) at B3LYP/6-311+G* level of theory. See 'results and discussion' section for details.

4. Conclusions

In the present study, the molecular tailoring approach is applied for obtaining the entire set of MO's for large spatially extended molecules and molecular clusters. The proposed algorithm is benchmarked for structurally and chemically diverse species including two-dimensionally conjugated molecules and covalently bonded clusters. These cases are difficult to handle at *ab initio* level due to their large size and also due to convergence problems especially when DFT is employed. Despite this, the MTA-based MO energies are seen to be quite accurately produced, for both HF as well as DFT levels of theory. The highest occupied and lowest unoccupied orbital energies and the respective band gaps (numerically of the order of 10^{-1}) are reproduced to within 0.01 a.u. for most of the cases examined.

Being a large and intricate molecule, the test case of protein 1L2Y is a challenging one. It also contains 5 charge centres with overall unit positive charge. For this system, the MTA-based MO calculation ran 3 times faster than the actual calculation at HF level of theory, the error in LUMO energy and the band gap being merely 0.001 a.u. Similarly, accuracy within few mH is achieved with advantage factor of 2.4 in the helical polyalanine case, which exhibits strong dipolar effects, making it difficult for fragmentation.

Systems consisting of thiophene rings in conjugation are of chemical interest due to their applicability in organic light emitting diodes (OLED's).[5] The band gaps of such materials are of importance as they quantify optical properties. MTA-based method has been successfully used in estimating the orbital energies and the corresponding band gaps for a class of such molecules within errors of a few mH. For all the systems in this class, the MTA-based calculation is seen to be faster than the actual one. The method is also successfully applied to two-dimensionally conjugated molecules *viz.* graphene model and graphene nano-ribbons,[33] on which the extensive contemporary research is being carried out.

The main lacuna of the present MTA-based MO calculation is the last step where a single SCF iteration for the whole molecule needs to be carried out. This is an expensive step as it involves evaluation of all the electron repulsion integrals for the parent molecule. For large systems, this step requires more computational resources. Moreover, the time for this step may equal the time taken for calculating the DM from fragments. Thus, there is a scope of further improvement in the proposed algorithm if the single SCF iteration is bypassed. One alternative way to avoid the SCF iteration is synthesizing Fock matrix from the fragment Fock matrices. This estimated Fock matrix can be subjected to diagonalization for obtaining the entire set of molecular orbitals. Such studies are underway in our laboratory.

The FMO–LCMO procedure of Tsuneyuki *et al.*[22] seems to be promising, but it cannot produce the entire set of MO's. The FMO–MO methodology by Inadomi *et al.*[21] is similar to the proposed MTA-based method. But both of these methods cannot be applied to molecules in which 'bodies' in many-body sense are not defined, as it employs FMO as the back end. For instance, FMO cannot be applied to the π-conjugated molecules, β-carotene, etc. To the best of authors' knowledge, FMO in its present form cannot be used with basis set comprising of diffuse functions. This limitation is also inherited by both of these methods. Since FMO is specialized for biomolecules, the applicability of FMO-MO and FMO-LCMO methods for these systems is promising, however, they need to be tested more thoroughly.

In summary, the molecular tailoring approach is a versatile linear scaling method for *ab initio* treatment of large molecules. The method is established for calculations of total electronic energy, Hessian matrix etc. and more importantly, geometry optimization. In the present work, the method is seen to offer a cost-effective solution for obtaining the MO's and hence, the wavefunction of the parent large molecule. With further improvement in the method it will open up a possibility to extract vital information regarding the chemical properties of large molecules and clusters at *ab initio* level of theory.

Acknowledgements

Fellowship to APR from the Council of Scientific and Industrial Research (CSIR), New Delhi and the support from Centre for Development of Advance Computing (C-DAC), Pune is gratefully acknowledged. SRG thanks the Department of Science and Technology (DST), New Delhi for the award of J C Bose Fellowship. Authors thank Professors Gejji, Truong, Negri and Mr. Sachin Yeole for geometries of some test systems reported in this work.

References

1. Mulliken R S 1951 *J. Chem. Phys.* **19** 912
2. Mulliken R S 1966 Nobel Lecture Spectroscopy, Molecular Orbitals and Chemical Bonding; See, http://nobelprize.org/nobel_prizes/chemistry/laureates/1966/mulliken-lecture.html
3. Fukui K 1982 *Science* **218** 747
4. Moliton A and Hiorns R C 2004 *Polymer Int.* **53** 1397
5. For some recent studies, see Nguyen H T and Truong T N 2010 *Chem. Phys. Lett.* **499** 263 and the references therein.
6. Koepnick B D, Lipscomb J S and Taylor D K 2010 *J. Phys. Chem.* **A114** 13228
7. Barden C and Schaefer H F, III 2000 *J. Pure Appl. Chem.* **72** 1405
8. Spangler D and Christoffersen R E 1980 *Int. J. Quant. Chem.* **17** 1075
9. a) Yang W 1991 *Phys. Rev.* **A44**, 7823 b) Zhao Q and Yang W 1995 *J. Chem. Phys.* **103** 5674
10. Gadre S R, Shirsat R N and Limaye A C 1994 *J. Phys. Chem.* **98** 9165
11. Babu K and S. R. Gadre S R 2003 *J. Comp. Chem.* **24** 484
12. Babu K, Ganesh V, Gadre S R and Ghermani N E 2004 *Theor. Chem. Acc.* **111** 255
13. Gadre S R and Ganesh V 2006 *J. Theoret. Comput. Chem.* **5** 835
14. Ganesh V, Dongare R K, Balanarayan P and Gadre S R 2006 *J. Chem. Phys.* **125** 104109
15. Rahalkar A P, Ganesh V and Gadre S R 2008 *J. Chem. Phys.* **129** 234101
16. a) Elango M, Subramanian V, Rahalkar A P, Gadre S R and Sathyamurthy N 2008 *J. Phys. Chem.* **A112** 7699 b) Mahadevi S, Rahalkar AP, Gadre S R and Sastry G N 2010 *J. Chem. Phys.* **133** 164308
17. Yeole S D and Gadre S R 2010 *J. Chem. Phys.* **132** 094102
18. Rahalkar A P, Katouda M, Gadre S R and Nagase S 2010 *J. Comput. Chem.* **31** 2405
19. Fedorov D G and Kitaura K 2009 *J. Chem. Phys.* **131** 171106
20. Fedorov D G, Ishida T, Uebayasi M and Kitaura K 2007 *J. Phys. Chem.* **A111** 2722
21. Inadomi Y, Nakano T, Kitaura K and Nagashima U 2002 *Chem. Phys. Lett.* **364** 139
22. Tsuneyuki S, Kobori T, Akagi K, Sodeyama K, Terakura K and Fukuyama H 2009 *Chem. Phys. Lett.* **476** 104
23. Li W and Li S 2005 *J. Chem. Phys.* **122** 194109
24. Akama T, Kobayashi M and Nakai H 2007 *J. Comp. Chem.* **28** 2003
25. Bettens R P A and Lee A M 2006 *J. Phys. Chem.* **A110** 8777
26. Collins M A 2007 *J. Chem. Phys.* **127** 24104
27. Kohn W 1996 *Phys. Rev. Lett.* **76** 3168
28. Schmidt M W, Baldridge K K, Boatz J A, Elbert S T, Gordon M S, Jensen J H, Koseki S, Matsunaga N, Nguyen K A, Su S J, Windus T L, Dupuis M, Montgomery J A 1993 *J. Comput. Chem.* **14** 1347. For further details: http://www.msg.ameslab.gov/GAMESS/GAMESS.html
29. a) Shirsat R N, Limaye A C and Gadre S R 1993 *J. Comput. Chem.* **14** 445 b) Limaye A C and Gadre SR 1994 *J. Chem. Phys.* **100** 1303
30. Muscat J, Wander A and Harrison N M 2001 *Chem. Phys. Lett.* **342** 397
31. The package MeTA Studio available at: http://code.google.com/p/metastudio/ See: Ganesh V2009 *J. Comput. Chem.* **30** 661
32. Qiu L, Pabit S A, Roitberg A E and Hagen S J 2002 *J. Am. Chem. Soc.* **124** 12952
33. Qian H, Negri F, Wang C and Wang Z 2008 *J. Am. Chem. Soc.* **130** 17970