Final report

on the NKFIH supported project no. 104672
entitled „*Characterization of the excitation and electron transport properties of DNA by quantum chemical investigation of its building blocks*”

The aim of the project was to investigate electronic properties of the DNA molecule by high level quantum chemical methods. We promised to investigate transport properties, in particular electron transfer (charge transfer) and develop methods which are accurate but also cost effective and allow the treatment of the building blocks of the DNA molecule.

Already at the beginning of the project, I have received the opportunity to write a “perspective” article in the prestigious *International Journal of Quantum Chemistry* about the DNA related research, and how the theoretical, in particular the high-level quantum chemical calculations can contribute to this field. It was shown that charge transfer processes and the eventual conductivity of DNA is a hot question where theoretical methods could bring substantial new knowledge. This paper [1] nicely sets the scene for the present project.

First let me summarize the results on charge transfer problems. For modeling the charge transport in molecules, the theoretically most relevant quantity is the *charge transfer integral* which is used by various theoretical models. The accuracy of this important property depends on several factors. In two papers we have investigated different ways of calculating this key quantity.

The goal of paper [2] was to identify the performance of various approximate approaches applied for the calculation of charge transfer integral which differ in internal simplifications of the applied formalism. For an unbiased investigation of these approximations, high level equation-of-motion coupled cluster theory was used for the electronic structure calculations to avoid errors from this source. The calculations have been performed on the ethylene dimer as one of the simplest model systems. By studying different spatial perturbations, it was shown that while both *energy split in dimer* and *fragment charge difference* methods are equivalent with the exact formulation for symmetrical displacements, they are less efficient when describing transfer integral along the asymmetric alteration coordinate. Since the “exact” scheme was found computationally expensive, we examined the possibility to obtain the asymmetric fluctuation of the transfer integral by a Taylor expansion along the coordinate space. By exploring the efficiency of this novel approach, we showed that the Taylor expansion scheme represents an attractive alternative to the “exact” calculations due to a substantial reduction of computational costs, when a considerably large region of the potential energy surface is of interest. Moreover, we showed that the Taylor expansion scheme, irrespective of the dimer symmetry, is very accurate for the entire range of geometry fluctuations that cover the space the molecule accesses at room temperature.

In paper [3] we showed that the coupled cluster (CC) and MBPT(2) approaches are the methods of choice for performing accurate calculations of this quantity, in contrast to some recently published results. We demonstrated that a proper treatment of the involved ionized states, achieved by applying the continuum-orbital strategy, reduces the error of the transfer integrals by one order of magnitude, which in the case of the CC2 method corresponds to a lowering of the mean relative unsigned error (MRUE) from 39.9 to 3.8%. Moreover, we extended the application of the continuum-orbital strategy to the TDDFT method and showed that it leads to a dramatic improvement of the description of ionized states compared to the conventional TDDFT approach, characterized by lowering of MRUE from 209.0 to 24.5%.

To understand the mechanism of excited state electron transfer, it is important to characterize excited states from the point of view of its spatial extend. In paper [4] DNA building blocks consisting of up to four nucleobases have been investigated using the EOM-CCSD and CC2-LR methods. It was found that excited states of single-stranded poly-A systems are highly delocalized, especially the spectroscopically bright states, where delocalization over up to four fragments can be observed. In the case of poly-T systems, the states are somewhat less delocalized, extending to maximally about three fragments. A single A:T Watson–Crick pair has highly localized states, while delocalization over base pairs can be observed for some excited states of the (A)2:(T)2 system, but intrastrand delocalization is more pronounced in this case, as well. As for the characteristics of the simulated UV absorption spectra, a significant decrease of intensity can be observed in the case of single strands with increasing chain length; this is due to the stacking interactions and is in accordance with previous results. On the other hand, the breaking of H-bonds between the two strands does not alter the spectral intensity considerably, it only causes a redshift of the absorption band, thus it is unable to explain the experimentally observed DNA hyperchromism on its own, and stacking interactions need to be considered for the description of this effect as well.

The method development part of the performed research can be subdivided into two topics: one is systematic benchmark studies of the available methods and the "real" derivation of new variants. Clearly, the former is very important for two reasons: it sets goals for the required accuracy and provide a framework to guide and test the accuracy of the different approximate methods.

As of the methodological development of the electronic structure methods, already in 2013, a paper has been published which includes basic data on the excited states of nucleobases dimers obtained at the EOM-CCSD(T) level. It was concluded that that popular methods used to describe such large systems (ADC(2), CC2) are less accurate than expected earlier, and in fact underestimate the excitation energy of some states considerably (paper [5]). This article was selected as the most read article of the *Journal of Physical Chemistry A* in June 2013.

It turned out that the conclusions of this paper contradict some new and comprehensive benchmark studies in the literature where for larger systems, like nucleobases, not all reference data was available. Therefore, we have also decided to perform systematic test calculations since our tools allowed us to calculate the necessary data at a higher level of theory than in previous studies. In papers [6] and [7] on the list we were not only able to perform "standard" single-double level calculation like CCSD and CC2 for about 150 excited electronic states of 28 molecules including also nucleobases, but we were also able to include triples effect at CCSD(T) and CC3 levels. Another addition of this paper compared to previous studies was that we have also included systematic comparison for transition moments, including also high level calculations. In particular, the extension of the standard benchmark set with complete data at the CC3 level got particular attention: the citations received for paper [6] in the first two years after publication brought us into the top 5% on our filed and allowed us to successfully apply for a KH-17 grant. These results show that both CC2 and CCSD are quite accurate and the difference to CC3 excitations energies is typically not larger than 0.2−0.3 eV. While the mean deviation of the CC2 results is close to zero, CCSD systematically overshoots the CC3 results by about 0.2 eV. The standard deviation is, however, somewhat smaller for CCSD, that is, the latter method provides more systematic results. Still, only a few cases could be identified were the absolute value of the error is over 0.3 eV in case of CC2. The results are even better for CCSD, with the exception of uracil, where surprisingly large error of the excitation energies have been found for two of the four lowest n−π* transitions. As of the comparison of LR (Linear Response) and EOM (Equation of Motion) style oscillator strengths are concerned, the former is more accurate at both CC2 and CCSD levels, but the difference between them is only 1−2% in case of CCSD. The error of the CC2 oscillator strength are substantially larger than that of CCSD but qualitatively still correct.

In paper [8] these benchmark studies have been extended with the use of diffuse functions. Diffuse functions allow the description of Rydberg states as well; due to the interaction of these states with the valence ones, an accurate description of the later often requires that also the Rydberg states are accurately described. In addition, Rydberg states are also important on their own in some applications. We have presented a comprehensive statistical analysis on the accuracy of various excited state Coupled Cluster methods, accentuating the effect of diffuse basis sets on vertical excitation energies of valence and Rydberg-type states. Many popular approximate doubles and triples methods were benchmarked with basis sets up to aug-cc-pVTZ, with high level EOM- CCSDT results used as reference. The results reveal a serious deficiency of CC2 linear response and CIS(D) techniques in the description of Rydberg states, a feature not shown by the EOM-CCSD(2) and EOM-CCSD variants. The CC3 theory proves to be an accurate choice among the iterative approximate triples methods, while the novel perturbation-based CCSD(T)(a)* variant turns out to be the best way to include the effect of triple excitations in a non-iterative way.

Based on the above conclusion, in paper [9] search for a better method for Rydberg states has been carried out. Popular second order techniques like Linear Response CC2 (CC2-LR), Partitioned Equation-of-Motion MBPT(2) (P-EOM-MBPT(2)), or Equation-of-Motion CCSD(2) (EOM-CCSD(2)) often produce results that are controversial and are ill-balanced with their accuracy on valence and Rydberg type states. Therefore, we have connected the theory of these methods and, to investigate the origin of their different behavior, establish a series of intermediate variants. The accuracy of these on excitation energies of singlet valence and Rydberg electronic states has been benchmarked on a large sample against high-accuracy Linear Response CC3 references. The results revealed the role of individual terms of the second order similarity transformed Hamiltonian, and the reason for the bad performance of CC2-LR in the description of Rydberg states. We also clarified the importance of the T_1 transformation employed in the CC2 procedure, which is found to be very small for vertical excitation energies.

Since neither variant included in this study gave satisfactory results for all type of excitation energy, we have continued the search for a possible new variant. The idea is that we should include not only the orbital energies in the diagonal elements of the DD block of the transformed Hamiltonian like in CC2, but also those contributions which involve other first order terms. These are included in e.g. EOM-CCSD(2), but there, at the same time, also the corresponding off-diagonal elements are used. Unfortunately, a strange bug in our code prevented us to fully test this promising variant thus the corresponding paper can be written in the near future only.

The test calculations mentioned above deal with the vertical excitation energy. The fundamental question, however, is, whether tests including just excitation energies provide enough information on the overall accuracy of the methods. In applications, one often considers the (full or local) potential energy surface or properties obtained from it. For example, to compare the excitation spectrum with experimental observation, vibrational effects need to be included in the simulation in even the simplest cases. Even more obvious is the need for the potential energy surface if one is interested in the processes taking place after the excitation. The accuracy of vertical excitation energies alone does not give reliable information about the quality of the associated potential energy surface. Thus, in order to investigate the performance of various methods on excited state surfaces, benchmarks need to go beyond the comparison of vertical excitation energies. In our most recent study we have extended the scope of excited state benchmark calculations to include the above aspects and performed systematic calculation on quantities which represent the quality of the excited state surfaces like equilibrium geometries, Franck-Condon forces, and shape of the steepest decent curves. The general conclusion is that the investigated methods (CC2, CCSD, CCSD(T), CC3, CCSDT) perform very differently for the various properties. Although one can obtain good results with second order methods – in particular for the vertical excitation energies of valence states – they are only certain triples methods which can be trusted in general. The paper from this study will be submitted shortly [11] and have been presented as an invited talk at WATOC-2017 (http://www.watoc2017.com/scientific-program.html)

In cooperation with French colleagues, I have also contributed to the development of DFT methods. DFT methods represent an alternative to the Coupled-Cluster methods and might offer a cost-effective treatment of such large systems like DNA building blocks. In paper [10] we have derived and implemented analytic gradients for the "long range" version of DFT, an extension of the popular theory, which is necessary e.g. to describe charge transfer states.

Beside the systematic studies presented above, we have also performed some conceptual, application-like studies which investigate important elements of proper treatment of DNA building blocks.

In paper [12], very important conclusion has been obtained on the excited states of adenine molecule. It turned out that it is only the S_1 state of adenine which possesses an equilibrium geometry, while the potential energy surface of all other states are connected with this very minimum by conical intersection between the surfaces of different states. This founding is a very important one from conceptual point of view and changes our thinking about excited states and the corresponding potential energy surfaces.

In paper [13] a detailed quantum chemical investigation was undertaken to obtain the structure and energetics of cytosine hydrates Cyt nH_2O, with n = 1 to 7. The MP2(fc)/aug-cc-pVDZ level was used as the standard, with some DFT (B3LYP) and coupled cluster calculations. In a systematic search for microhydrated forms of cytosine, we have found that several structures have not yet been reported in the literature. The energies of different isomers, as well as binding energies are compared. When predicting the stability of a complex, we suggested a new scheme where the water molecules are extracted from a finite model of bulk water. Finally, based on energetic data, we suggested a rational definition of the first hydration shell; with this definition, it contains just six water molecules.

In summary, the twelve papers published together with the one in preparation largely fulfill the goals of the project. Some corrections during the turn of the execution have been naturally undertaken, but the results obtain brought us closer to the understanding of the electronic structure of DNA building blocks and in particular answered several methodological questions.

References:

[1] Szalay, P. G. (2013). Can Coupled-Cluster Methods Be Used to Describe Excited States of the Building Blocks of DNA? *International Journal of Quantum Chemistry*, 113(14), 1821–1827. http://doi.org/10.1002/qua.24392

[2] Pershin, A., and Szalay, P. G. (2015). Development of highly accurate approximate scheme for computing the charge transfer integral. *The Journal of Chemical Physics*, 143. http://doi.org/10.1063/1.4928053

[3] Pershin, A., and Szalay, P. G. (2015). Improving the Accuracy of the Charge Transfer Integrals Obtained by Coupled Cluster Theory, MBPT (2), and TDDFT. *Journal of Chemical Theory and Computation,* 11, 5705–5711. http://doi.org/10.1021/acs.jctc.5b00837

[4] Benda, Z., and Szalay, P. G. (2016). Characterization of the Excited States of DNA Building Blocks: a Coupled Cluster Computational Study. *Physical Chemistry Chemical Physics*, 18, 23596–23606. http://doi.org/10.1039/C6CP02969H

[5] Szalay, P. G., Watson, T. J., Jr., Perera, A., Lotrich, V. F., and Bartlett, R. J. (2013). Benchmark Studies on the Building Blocks of DNA. 3. Watson–Crick and Stacked Base Pairs. *The Journal of Physical Chemistry A*, 117(15), 3149–3157. http://doi.org/10.1021/jp3100975

[6] Kánnár, D., and Szalay, P. G. (2014). Benchmarking Coupled Cluster Methods on Valence Singlet Excited States. *Journal of Chemical Theory and Computation*, 10(9), 3757–3765. http://doi.org/10.1021/ct500495n

[7] Kánnár, D., and Szalay, P. G. (2014). Benchmarking coupled cluster methods on singlet excited states of nucleobases. *Journal of Molecular Modeling*, 20(11), 1–8. http://doi.org/10.1007/s00894-014-2503-2

[8] Kánnár, D., Tajti, A., and Szalay, P. G. (2017). Accuracy of Coupled Cluster excitation energies in diffuse basis sets. *Journal of Chemical Theory and Computation,* 13, 202–209. http://doi.org/10.1021/acs.jctc.6b00875

[9] Tajti, A., and Szalay, P. G. (2016). Investigation of the Impact of Different Terms in the Second Order Hamiltonian on Excitation Energies of Valence and Rydberg States. *Journal of Chemical Theory and Computation,* 12, 5477–5482. http://doi.org/10.1021/acs.jctc.6b00723

[10] Mussard, B., Szalay, P. G., and Angyan, J. G. (2014). Analytical Energy Gradients in Range-Separated Hybrid Density Functional Theory with Random Phase Approximation. *Journal of Chemical Theory and Computation*, 10(5), 1968–1979. http://doi.org/10.1021/ct401044h

[11] Tajti, A., Stanton, J.F.,and Szalay, P.G. (2018). Accuracy of Coupled Cluster excited state potential energy surfaces, *Journal of Chemical Theory and Computation,* to be submitted

[12] Benda, Z., and Szalay, P. G. (2014). Details of the Excited State Potential Energy Surfaces of Adenine by Coupled-Cluster Techniques. *The Journal of Physical Chemistry A*, 118(32), 6197–6207. http://doi.org/10.1021/jp505331s

[13] Fogarasi, G., and Szalay, P. G. (2015). Quantum chemical MP2 results on some hydrates of cytosine: binding sites, energies and the first hydration shell. *Physical Chemistry Chemical Physics*, 17, 29880–29890. http://doi.org/10.1039/C5CP04563K