The influence of base pair tautomerism on single point mutations in aqueous DNA

Gheorghiu, A.[1], Arabi, A.A.[2], Coveney, P.V.[1,3]

[1] Centre for Computational Science, University College London, United Kingdom

[2] College of Natural Health and Sciences, Zayed University, United Arab Emirates

[3] Informatics Institute, University of Amsterdam, Netherlands

1. Introduction

The human genome consists of approximately 3 billion base pairs, stored as nucleic acid sequences. Due to its vast complexity, the genome is fragile – unsurprisingly the DNA within is susceptible to change. The mutations that occur in these DNA sequences are crucial to both natural evolution and the occurrence of genetic diseases. While some of these changes might be a consequence of exposure to high energy electromagnetic fields or other forms of radiation, mutations may also arise due to mistakes during the DNA replication process.

Although remarkably accurate, the high-fidelity DNA replication process generates base substitution errors at a rate of 10^{-4} to 10^{-5} per replicated nucleotide. However, due to various intrinsic repair mechanisms, errors in human genome replication are actually less frequent (approx. ~10^{-8} to 10^{-10} per replicated nucleotide). These replication errors, known as *point mutations*, may occur as a result of wobble base pairing, Hoogsteen (*anti-syn*) base pairing, ionisation and tautomerisation (the frequency of each is uncertain).

The purpose of this work is to determine the impact of base pair tautomerisation on the rate of single point mutations in DNA. The origin of these tautomers is an ongoing subject of investigation, extensively studied by idealised gas phase quantum models. However, these models are typically simplified to a single base pair and often produce conflicting results to one another. Unfortunately, the experimental data on base pair tautomerisation are sparse. This work studies the double proton transfer (DPT) tautomerisation pathway using a more advanced multiscale computational approach. The model begins with an experimentally resolved DNA structure which is then thermalised and sampled effectively using ensemble-based classical molecular dynamics. From there, an ensemble of reaction coordinates, transition states and consequently, the rate of tautomerisation for selected base pairs in aqueous DNA are computed using quantum-mechanics/molecular-mechanics (QM/MM).

The reported errors arise from the configurations drawn from MD simulations to the QM approximations used. Performing an ensemble of calculations accounts for the stochastic aspects of our simulations while make systematic errors easier to identify. Our work predicts the double proton transfer tautomerisation of the G:C base pair to occur via an asynchronous stepwise mechanism at forward rate of ~10^{5} s^{-1}. Overall, our results are better in agreement with experimental data than previous QM gas phase calculations.

2. Background

The base pair tautomeric forms were first proposed by Watson & Crick (1953). Löwdin (1963) later pointed out that these tautomeric states facilitate base pair mismatching and thus propagate errors during genetic replication. Recently, structural evidence of a C:A mismatch adopting a Watson-Crick geometry in an active site of DNA polymerase was resolved, providing some of the first experimental evidence for Löwdin's mutation mechanism

[1]. This is because the C:A mismatch could only adopt a Watson-Crick geometry if it were formed as a result of a tautomeric single point mutation.

Density functional theory (DFT) models are typically restricted by computational cost in their ability to include both DNA macrostructure and/or bulk explicit solvent. Previous DFT studies have shown that double proton transfer reactions within G:C base pairs is energetically more favourable than A:T tautomerism.

Coincidentally, a universal G:C to A:T DNA mutation bias was observed in *E. coli*, whereby 70% of all point mutations reduced overall G:C content [2].

Some recent models in the literature adopt a hybrid quantum-mechanics/molecular-mechanics (QM/MM) approach to study DNA. The advantage of using QM/MM methods reside in its ability to model a reaction pathway at the QM level, while treating the rest of the DNA strand and bulk solution at a classical level. Doing so, facilitates the study of experimentally accessible DNA structures in physiological conditions. We build our QM/MM model loosely on previous work [3], however, it is non-trivial to isolate the sources of error within our calculations. These include, but are not limited to, the embedding technique used and the link-atom approximations to the size of the QM region.

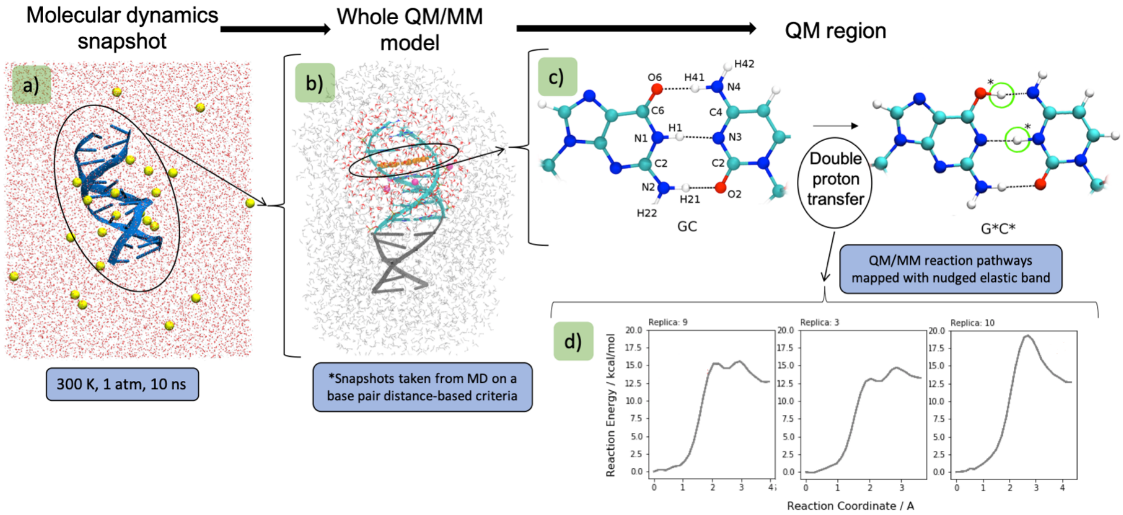

Figure 1 *A visual schematic of the workflow a) classical molecular dynamics using NAMD, b) an open boundary QM/MM model using ChemShell, c) single base pair (the QM region) tautomerisation and d) QM/MM reaction pathways showing both the stepwise and a concerted process*

3. Methods

Molecular Dynamics: The x-ray structure of the Drew-Dickerson Dodecamer (*1BNA)* was neutralised and solvated using the TIP3P water model. Using NAMD 2.12, ensemble classical molecular dynamics was then performed in periodic boundary conditions to the following protocol: DNA was restrained, and the energy of system minimised using the latest modified AMBER *parmbsc1* forcefield. The temperature of the system was incrementally raised from 50 K to 300 K over a time period of 30 ps. The restraints were then systematically removed over 1 ns, followed by an unrestricted 10 ns run. This entire process was repeated ten times, totalling 100 ns of simulation time. The molecular dynamics simulations were performed using the UCL high performance computing (HPC) facility Grace.

Choosing the QM method: Various QM approximations were benchmarked using NWChem 6.6, assessing their ability to describe base pair geometries and interaction energies.

The interaction energy between isolated base pairs was measured at the selected QM approximation and compared to highly accurate coupled-cluster reference values [4]. The most computationally efficient QM method that compared best to the reference value was then chosen for the ensuing QM/MM work. Our analysis predicted B3LYP/aug-cc-pvdz with XDM dispersion correction to describe base pair interactions accurately at a reasonable computational cost. The QM calculations were performed using the Blue Waters supercomputer at the National Center for Supercomputing Applications, USA.

QM/MM: A selection of thermalised DNA structures from the MD simulation were further modelled using QM/MM. Specifically, snapshots with the most probable average inter-nucleobase distance for the selected base pair were chosen. ChemShell 3.7 linked with NWChem 6.6 and DL-POLY was then used to perform QM/MM calculations. All of the tasks involving optimisations were performed using the DL-FIND module. The snapshot was converted to an open-boundary system to include the full DNA strand and a 15 Å solvation sphere (Fig. 1.b.). Following a partially restricted QM/MM geometry optimisation, the DPT reaction pathway was computed using the climbing image nudged elastic band method. For each step in the optimised reaction pathway, the Hessian was calculated at a thermal correction of 300 K. From these calculations, we then estimated the range of rate coefficients our models can predict using transition state theory. All ChemShell calculations were performed using the UK national supercomputer ARCHER.

Results and Discussion

Upon the investigation of the QM/MM electronic energies, our simulations show that the G:C double proton transfer may proceed via any of the following pathways, each with varying probabilities; *step-wise* (via two transition states & an intermediate) or *concerted* (one transition state & no intermediate), in conjunction with either *synchronous* (simultaneous) or *asynchronous* (delayed) proton exchange. Previous gas phase QM-only DNA double proton transfer studies typically predict G:C double proton transfer to occur via one (occasionally two) of the above pathways. However, based on our prior use of classical molecular dynamics the conformational landscape of the DNA base pair is explored and we observe multiple possible pathways. In G:C, we find the step-wise asynchronous pathway to occur ~75% of the time, followed by the concerted asynchronous pathway ~18% of the time. Lastly, the concerted synchronous pathway only occurs ~7% of the time. The forward free energy barrier for the step-wise asynchronous pathway (the most common) is ~ 10 kcal/mol, while the reverse free energy barrier is negative (~ -1 kcal/mol).

We predict the forward rate of G:C tautomerisation to be ~10^5 s^{-1}, which is in good agreement with current experimental data (also ~10^5 s^{-1}) [5]. Our results suggest that an improvement on previous QM-only gas phase rate predictions (10^2 - 10^4 s^{-1}) has been achieved. We find a fast reverse rate of tautomerisation ~10^{14} s^{-1} and consequently, the life time of the tautomer to be ~10 fs. During DNA replication, base pair opening is estimated to occur over several nanoseconds. Therefore, due to a mismatch in timescales between replication and tautomerisation along with the thermodynamic instability of the tautomer, we predict that tautomerisation in DNA is unlikely to have an effect on spontaneous mutations.

Our preliminary studies show that the formation of a proton transfer tautomer in an

A:T base pair is significantly less likely to occur than in G:C.

Although each QM/MM replica is standardised to a mean base pair distance criterion, the variance between each reaction coordinates for a given base pair is surprisingly large. This variety asserts the importance of performing ensemble-based simulations. It is evident that the proton transfer reaction pathway is sensitive to a plethora of other effects, not simply base pair distance. These effects are likely to include details of DNA conformation such as helical twist, stacked base pair distances and so on, as well as the arrangement of surrounding water molecules and ions.

4. Conclusion

Our work shows that tautomerism within base pairs is a complex process that is often over-simplified by QM gas-phase models. While some of our results predict similar activation energies to QM-only models, the conformational landscape of thermalised DNA is explored based on prior use of molecular dynamics. Doing so, we show that proton transfer can occur via multiple different pathways (or may not happen at all) within the same base pair. We conclude that in general, G:C tautomerism is unlikely to have a dominant effect on point mutation rates – since it is both a kinetically and thermodynamically unfavourable process. Preliminary results suggest that the occurrence of tautomerism in A:T is even less frequent than in G:C. These results comply with the universal G:C to A:T mutation bias that is observed in DNA. In all cases, the proton transfer reaction pathway is susceptible to variation, presumably depending on environmental effects such as explicit solvation and structure. We therefore conclude that various external influences may reduce (or increase) the energy barrier of base pair tautomerism to make the process influence single point mutation frequency more (or less). Future work will assess the effect of including water in the QM region in conjunction with studying base pairs at different positions in the DNA strand. Both of these factors are expected to alter the tautomerisation pathway and its probability of occurrence.

5. References

1) Wang, W., Hellinga, H.W. and Beese, L.S., 2011. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. Proceedings of the National Academy of Sciences, 108(43), 17644-17648.

2) Lee, H., Popodi, E., Tang, H. and Foster, P.L., 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences, 109(41), E2774-E2783.

3) Lu, Y., Lan, Z. and Thiel, W., 2011. Hydrogen bonding regulates the monomeric nonradiative decay of adenine in DNA strands. Angewandte Chemie International Edition, 50(30), 6864-6867.

4) Jurečka, P., Šponer, J., Černý, J. and Hobza, P., 2006. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. Physical Chemistry Chemical Physics, 8(17), 1985-1993.

5) Kimsey, I.J., Szymanski, E.S., Zahurancik, W.J., Shakya, A., Xue, Y., Chu, C.C., Sathyamoorthy, B., Suo, Z. and Al-Hashimi, H.M., 2018. Dynamic basis for dG• dT misincorporation via tautomerization and ionization. Nature, 554(7691), p.195.