There is increasing evidence that DNA can support a considerable degree of charge transport along the strand by hopping of holes from one base to another, and that this charge transport may be relevant to DNA regulation, damage detection and repair. A surprisingly useful amount of insight can be gained from the construction of simple tight-binding models of charge transport, which can be investigated using the transfer-matrix method. The data thus obtained indicate a correlation between DNA charge-transport properties and the locations of cancerous mutation. We review models for DNA charge transport and their extension to include more physically realistic diagonal-hopping terms.

Keywords: Keyword1; keyword2; keyword3.

1. Introduction

The question of whether DNA conducts electric charges is intriguing to physicists and biologists alike. Soon after Watson and Crick discovered the double-helix structure of DNA [1], Eley and Spivey were the first to suggest that DNA could serve as an electronic conductor [2]. In particular, the notion of a molecular wire was thought to apply to the DNA double helix because of its $\pi$-electron system of bases.
stacked upon each other. The suggestion that electron transfer/transport in DNA might be biologically important has triggered a series of recent experimental and theoretical investigations, for example [3–10].

In the field of nanotechnology, DNA has been suggested as a material for molecular electronics [11–14]. DNA might serve as a wire, transistor, switch or rectifier depending on its electronic properties [9, 15, 16]. Biologically, processes that may involve electron transfer along DNA strands include the function of DNA damage response enzymes, transcription factors or polymerase co-factors, all of which play important roles in the cell [17]. Indeed there is direct evidence [18] that MutY — a DNA base excision repair enzyme with an [4Fe4S]$^+$ cluster of undetermined function — takes part in some kind of electron transfer as part of the DNA repair process [19, 20]. This seems consistent with studies in which an electric current is passed through DNA revealing that damaged regions have significantly different electronic behaviour than healthy ones [18]. There is also evidence [21] that the regulation of the p53 gene, the so-called “guardian of the genome”, may involve electron transfer along the gene.

Convenient tight-binding model for DNA are usually constructed as follows: one assume one or two central conduction channels in which individual sites represent a base-pair or individual bases, respectively. These are interconnected and sometimes further linked to upper and lower sites, representing the backbone, but are mostly not interconnected along the backbone. Every link between sites implies the presence of a hopping amplitude.

Quasi-1D models incorporating these aspects have been recently introduced in Refs. 11, 22, building on earlier, even simpler 1D models [10, 23–26]. For these models, electronic transport properties have been investigated in terms of localisation lengths [22, 23, 27], crudely speaking the length over which electrons travel, as well as transmission [25] and current-voltage characteristics [11]. Various types of disorder, including random potentials, have been employed to account for different real environments and temperatures [10]. It has been found that random and A-DNA have localisation lengths allowing for electron motion among a few dozen base pairs only. However, poly(dG)-poly(dC) and also telomeric-DNA have much larger electron localization lengths [11]. In Ref. 22, a novel enhancement of localisation lengths has been observed at particular energies for an increasing binary backbone disorder. While keeping the number of parameters small, these models have been able to reproduce the wide-gap structure observed in much more accurate quantum chemical calculations of short DNA strands [11–14, 28]. Useful information about the strength of the charge transport and hence the spatial extent of electronic states along a DNA strand can be obtained, which are surprisingly close to studies of range dependence of electron transfer [3, 4, 6–8, 18, 29].

These results indicate that the transport properties of a DNA sequence depend not only on its overall composition, but also the detailed order of base pairs, i.e. the genetic sequence. The biological significance of such variations, however, are
still unclear. Very recently, by using single and double strands tight-binding models with parameters fitted from ab initio calculations [11, 30], the charge-transport (CT) changes owing to cancerous and non-cancerous point mutations have been statistically studied for the p53 gene [31]. We find that anomalously small changes in charge transfer efficiency tend to coincide with cancerous mutations. In contrast, non-cancerous mutations result, on average, in much larger changes of the CT properties. This may well be relevant to way in which carcinogenic mutations avoid the DNA damage/repair processes and hence lead to carcinogenesis.

In the present paper, we will continue the investigations in Refs. 22, 32, 31 by (i) introducing what appears to be the most appropriate tight-binding model of DNA and (ii) studying its transport characteristics for a cancer-related gene, RB1.

2. Diagonal-ladder model of DNA

In Figure 1 we show a schematic model for charge transport in DNA, detailing the set of on-site and hopping terms that are used in the construction of the tight-binding Hamiltonian [33]

\[
H = \sum_i \epsilon_i c_i^\dagger c_i + \sum_{i,j} t_{ij} c_i^\dagger c_j + t_{ji} c_j^\dagger c_i,
\]

where each site \(i\), a nucleotide base or backbone phosphate, has energy \(\epsilon_i\) and interacts with its near neighbours \(j\) \((i \neq j)\) with a Hamiltonian hopping interaction of \(t_{ij}\). This is a form of the Anderson model [34]. Such a model may conveniently be studied using the transfer-matrix method [35] to extract localisation lengths, \(\lambda(E)\), and transmission coefficients, \(T(E)\), as a function of the energy of the injected carrier.

Recent work [22, 31, 32] has shown that such models can usefully be applied to study biologically significant phenomena such as the occurrence of carcinogenic mutations. These studies have used a relatively simple set of parameters where the onsite energies for hole transport are taken to be the ionisation energies [36] \(\epsilon_G = 7.75\ eV, \epsilon_C = 8.87\ eV, \epsilon_A = 8.24\ eV, \epsilon_T = 9.14\ eV\). The hopping terms, based loosely on the results of ab-initio calculations [11, 30], were taken to be 0.35 \(eV\) between like base pairs and 0.17 \(eV\) between unlike. The interchain hopping term \(t_{\perp} (= t_{12}\ in\ Figure\ 1)\) was taken to be 0.1 \(eV\), an unphysically large figure, and no diagonal hopping terms (that is, between a base on one chain and a base one step up or down the other chain) were included at all.

It is known from electronic structure calculations (for example Ref. 37) that the dominant set of orbital overlaps for hole transfer are those between purines (guanine, adenine) in adjacent base pairs, including diagonal overlaps. The use of a large interchain hopping term may thus be seen as a compensation for the lack of diagonal terms in the model. A model including diagonal terms allows the use of a more physical, small \(t_{\perp}\) term, and is fully consistent with the “G sites/A bridges” model of hole transport in which holes are considered to be localised on guanine
Fig. 1. Schematic model for charge transport in DNA. The nucleobases are given as (dark grey) circles (pyrimidines) and ellipses (purines). Electronic pathways are shown as solid lines, and (light grey) circles denote the sugar-phosphate backbone sites. The diagram shows effective pathways for transport along a many channel model. The varying strengths of hopping elements are indicated by varying line thickness. The diagonal hopping elements are $t_{55}$ for transfer along the diagonal connecting two 5' ends and $t_{33}$ for the diagonal connecting two 3' ends. Note that diagonal hopping between purines is favored, and between pyrimidines is disfavored, by the larger size of the purines.

bases, the lowest-energy site, and to pass over A sites during hopping transport. We have therefore extended the models used in [22, 31, 32] to include the diagonal hopping terms.

The introduction of diagonal elements leads to a potential pitfall that must be avoided in constructing the transfer matrices. We can write the overall $4 \times 4$ transfer matrix $T_i$ between slices $i$ and $i + 1$ in block form as follows:

$$
T_i = \begin{pmatrix}
V_i & \tau_{i-1}^{-1} \\
1 & 0
\end{pmatrix},
$$

where the $2 \times 2$ matrix $V_i$ contains all the terms involving site energies, perpendicular hopping and the input energy $E$, 1 and 0 are $2 \times 2$ unit and null matrices, and $\tau_{i}$ is a $2 \times 2$ matrix containing the hopping elements between slice $i$ and slice $i + 1$. Ironically, the $\tau$ matrices are diagonal in the absence of diagonal hopping, while in
the presence of diagonal hopping $\tau_i$ develops off-diagonal terms:

$$
\tau_i = \begin{pmatrix}
-t_{i1} & -t_{i55} \\
-t_{i33} & -t_{i2}
\end{pmatrix},
$$

(3)

where the $t_{i55}$ term is the hopping on the 5'–5' diagonal and the $t_{i33}$ term is the hopping on the 3'–3' diagonal.

The upper right block then depends on the inverse of $\tau_{i-1}$. Any choice of the diagonal hopping terms which leads to a zero-valued determinant for a $\tau$ matrix will “break” the transfer-matrix method, as it is then impossible to form the inverse of $\tau_{i-1}$ and hence to form $T_i$. This is an example of a well-known problem with the TM method [38]; the matrices involved can become singular and the method fails even when the physics of the situation is quite well-defined. An obvious case in which $\tau_i$ becomes singular is when the linear and diagonal hopping terms are equal; for example, setting all the $t_{iX}$ terms equal to 1 makes $\tau_i$ the matrix

$$
\tau_{\text{singular}} = \begin{pmatrix}
1 & 1 \\
1 & 1
\end{pmatrix}.
$$

(4)

We note that the inclusion of diagonal terms is the minimal level of theory required to take account of the helicity of dsDNA; that is, the handedness of the helix can in principle be modelled by a systematic difference between the $t_{i55}$ and $t_{i33}$ parameter sets. In this study, however, we do not explore the issue of helicity.

Introduction of the diagonal terms introduces a large number of new parameters into the model as in principle $t_{iNN}$ could differ for each possible pair of bases and for the two diagonal directions. We have chosen a simple set of diagonal terms to reflect the geometry of the base pairs. The larger purine bases can achieve a considerable degree of electronic overlap when in a diagonal configuration, as indicated schematically in Figure 1. We have therefore assigned diagonal hopping elements of 0.1 eV for purine-purine transfer, 0.01 eV for purine-pyrimidine and 0.001 eV for pyrimidine-pyrimidine, and have suppressed the hopping term across the hydrogen bond, $t_\perp$, to 0.005 eV.

3. Comparison with previous results

In Figure 2 we show a comparison of $T(E)$ computed using the ladder model (LM) as in 31 and using the diagonal ladder model (DL) for a short length of telomeric DNA (four repeats of a ttaggg motif). The diagonal model has a much reduced $t_\perp$ term but includes explicit diagonal hopping. It is clear from the data that the two models give broadly similar $T(E)$, and that overall the diagonal model gives higher $T(E)$ in several energy ranges (for example 8 to 8.2 eV and 9.5 to 10 eV). This is consistent with our argument that the large perpendicular hopping term in the ladder model was a proxy for the more physical diagonal hopping model.
4. Results for SBS mutations in retinoblastoma

We illustrate the application of our diagonal-hopping model by studying the charge transport properties of the retinoblastoma (RB1) gene, using data obtained from the Retinoblastoma Genetics Home database [39]. We examine the charge-transport properties of sections of the DNA sequence in the vicinity of known mutations. Our approach is to compare the change in charge transport between the normal and mutated gene sequences, measured as the mean-square change in $T(E)$ integrated over a range of $E$ sufficient to include all significant transport. This method is described in more detail in [31].

The database contains information on 378 sites where single base substitutions (SBSs) have been observed in patients suffering from retinoblastoma, and on 110 sites where neutral SBSs have been observed with no phenotypic effect. The smaller size of the neutral set is more likely to reflect lower rates of detection than the genuine prevalence of neutral mutations; evidently, the majority of people with neutral mutations will never have their RB1 gene sequence recorded. Unlike the case of p53 [31], the database does not contain information on the frequency with which mutations at different sites are observed, and we cannot examine correlations between CT properties and mutation frequency. Instead, we compare the statistics
of the cancerous and neutral mutations and also the statistics of a large set of 1000 randomly-generated fictitious mutations.

We note that all of the data following was generated by selecting a location in the RB1 sequence from the list of cancerous, neutral or random sites, and generating DNA sequences of length 21 by selecting ten base pairs in each direction from the chosen site. Four DNA sequences were generated for each site, one for the reference sequence and three for the possible SBSs at the site. We used our diagonal-ladder transfer-matrix CT model to extract Lyapunov exponents $\gamma(E)$ for energies in the range 7 to 11 eV, with a spacing in $E$ of 0.005 eV, from which we could extract a transmittance $T(E)$. The change in $T(E)$ for a given mutation was then quantified by integrating the square of the difference of $T(E)$ between the normal and mutated sequences over our energy range. Since our interest here is in relative rankings we have not normalised this change.

For cancerous and neutral mutations we extracted a ranking between 1 and 3 for the observed mutation by comparing its change in $\Delta T(E)$ to that of the other two possible mutations at the site. In this case rank 1 indicates that the observed mutation causes a larger $\Delta T(E)$, according to our model, than either of the other two possible mutations; rank 3, on the other hand, indicates a smaller change. For the randomly generated set of sites we ranked each of the possible mutations at that site, giving rank 1 to the largest $\Delta T(E)$ as before.

In our earlier studies [31] we considered the overall probability for a cancerous mutation to display a lower $\Delta T(E)$ than the other possible mutations. We did not, however, disaggregate the results according to the type of the SBS. In this study, we divide up our data according to the twelve different possible SBSs and consider the distribution of rankings for each type. This reveals considerable systematic variation and indicates that future studies should include disaggregated data.

In Figure 3 we plot the average $\Delta T(E)$ for a given cancerous mutation versus the number of such mutations found in the database. There are twelve points, as there are twelve possible SBS mutations: A to C (AC), AG, AT, CA, CG, CT, GA, GC, GT, TA, TC and TG. It does not appear that there is a correlation. We should recall, however, that we do not have data on the prevalence of each mutation, only the number of different sites where it has been recorded.

In Figure 4 we show a histogram of the number of sites where cancerous SBSs are observed, with rankings according to the change in $T(E)$. It appears that we see three different sets. Three types of mutation (CT, GA and GT) are particularly common, and in two cases (CT and GA) the mutation is commonly ranked third, indicating lowest change in $T(E)$. In the case of GT, however, higher ranks are more common. A set of five mutations are less common (AG, GC, TA, TC and TG) with no clear pattern in their rankings. The remaining four mutations are uncommon (AC, AT, CA and CG). It is interesting to note that among this uncommon set, it is rare for the mutation to be ranked third in $\Delta T(E)$, consistent with the idea that mutations leading to a large change in $T(E)$ are more likely to be caught by DNA repair mechanisms.
Fig. 3. Average $\Delta T(E)$ for a given cancerous mutation versus the number of such mutations found in the database. It does not appear that there is a correlation.

In Figure 5 we show a histogram of the number of sites where neutral SBSs are observed. This distribution is highly nonuniform with most of the observed neutral mutations being purine-for-purine and pyrimidine-for-pyrimidine substitutions AG, CT, GA and TC. It is visible that the AG and TC mutations are particularly likely to be ranked third in $\Delta T(E)$. We should probably not over-interpret this, however.

Finally, in Figure 6 we show data for a set of randomly generated mutations. The differing frequencies reflect that prevalence of each base in the RB1 sequence. As in the case of the neutral mutations it seems that the purine-for-purine and pyrimidine-for-pyrimidine substitutions AG, CT, GA and TC are particularly likely to be ranked third in $\Delta T(E)$.

5. Conclusions

A review of DNA electronic structure calculations and tight-binding models indicates that a physically realistic picture of DNA charge transport requires the inclusion of diagonal hopping terms, allowing the model to reflect the favored purine-to-purine hole transfer. We show that a simple model including these terms behaves comparably to an earlier ladder model in which a large perpendicular hopping term stands proxy for the diagonal hopping. An examination of the CT properties of a set of mutations in the retinoblastoma gene RB1 indicates some interesting features, for example that the least frequently observed mutations appear more likely to have higher ranks in $\Delta T(E)$. We suggest that data on mutations and CT should
if possible be disaggregated by the type of mutation so as not to obscure interesting features of the data.

The overall likelihood of a mutation leading to a cancer will be a convolution of the likelihood of a given mutation occurring, of its escaping the notice of DNA repair mechanisms, and of its having some effect when the gene is expressed. Despite this complexity, investigations of CT properties can be informative in at least two possibly independent ways. Firstly, insofar as CT properties are directly involved in DNA damage detection and repair and DNA regulation, unusual CT properties in a sequence will mark it out as a potential trouble spot. Secondly and more abstractly, CT models are in a sense probes of the statistics of the DNA sequence, as we are extracting a calculated property which depends non-linearly on the sequence of base pairs. Therefore, if certain kinds of DNA sequence are more vulnerable to mutation and damage for either physical or chemical reasons, we may find correlations between our model properties and the properties of the DNA — even if the actual property we are probing is not directly related to the charge transport phenomena which inspired the model. In this second case, we may find that the set of parameters which are most informative for biology may diverge from those which most accurately represent the physics of DNA charge transport.

Acknowledgments SAW and RAR gratefully acknowledge the Leverhulme Trust (grant F/00 215/AH) for financial support. CTS thankfully acknowledges the National Center for High-Performance Computing in providing resources under the
Fig. 5. Number of sites with known neutral mutations and their rankings by $\Delta T(E)$. Ranks 1, 2, 3 are shown in grey, white and black.

Fig. 6. Randomly generated mutations and their rankings by $\Delta T(E)$. Ranks 1, 2, 3 are shown in grey, white and black. The differing frequencies reflect that prevalence of each base in the RB1 sequence.
national project,”Taiwan Knowledge Innovation National Grid”, and the financial support from National Science Council, Taiwan (NSC 97-2112-M-029-002-MY3).


