MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **87** (2022) 193-202
ISSN: 0340-6253
doi: 10.46793/match.87-1.193B

Identifying Regions of a Gene That Are Susceptible to Mutation

Alper Bulut[a] **, Fatih Dogan**[b]

[a] *Department of Mathematics and Statistics, American University of the Middle East, Kuwait*
alper.bulut@aum.edu.kw

[b] *College of Engineering and Technology, American University of the Middle East, Kuwait*
fatih.dogan@aum.edu.kw

(Received April 18, 2021)

Abstract

In this article, the susceptibility of any gene to mutation is explained through the tumor suppressor gene TP53 using the concept of the energy of a graph. We considered the structure of TP53 gene as a weighted graph where the weights are the bond dissociation energies. We computed energies of each exon and investigated how they distributed over the bases. We observed that exons 4 and 12 have unusual energy distributions compared to other exons. The energy was found to be significantly reduced relative to other regions between the 44th and 47th codons of exon 4, and this is in line with the literature results to which these regions are subject to severe missense and frameshift mutations. The energy of exon 12 changes very rapidly at very short intervals and is not consistently distributed across the exon. The excited energy probability distribution of exon 9 is used to determine the most vulnerable region of exon 9 when subjected to any physical or chemical external influences.

1 Introduction

The genetic material DNA has been centre of attraction since its first discovery by Friedrich Miescher in 1869. He initially aimed to determine the chemical composition of cells from the leucocytes derived from the pus over the fresh surgical bandages. He managed to isolate the material inside the nuclei and realized that the substance was not

a protein but contained a high amount of phosphorus. He called this material "nuclei" because of its presence in the nuclei. Miescher also confirmed the existence of carbon, hydrogen, oxygen, and nitrogen in the structure of "nuclei", see [1]. Albrecht Kossel studied the chemical composition and properties of nucleic acids between 1885 and 1901, and he discovered that nucleic acids consist of five nitrogen bases: adenine, cytosine, guanine, thymine, and uracil, see [2]. His student Levine identified the carbohydrate portion of the nucleic acid as deoxyribose. In 1909, Jacobs and Levene claimed that the nucleic acids are built up from groups that is referred to "nucleotides" and in 1935 Levene and Tipson's report accurately shows first time the molecular structures of DNA, see [3].

In 1950, Chargaff realized that the molar ratios of total purines to total pyrimidines, and also of adenine to thymine and of guanine to cytosine are very close to 1, and he stated that "this is more than accidental" which is now known as one of the Chargaff's rule, see [4]. In 1953, Watson and Crick published the article " Molecular Structure of Nucleic Acids" in which they put forward radically different structure than the previous ones. In their models, two helical chains coiled around the same axis and two chains are held together by purine and pyrimidine bases via hydrogen bonds, see [5].

A gene is a short section of the DNA and it is responsible to make proteins in the cell. Exons and introns are nucleotide sequences within the gene such that any consecutive exons are separated by an intron. Exons carry the codes necessary for protein production and work in harmony with introns. Introns allow exons to be spliced in different combinations so that different proteins can be produced.

The main goal of this paper is to compute the graph energies of each exon located on the gene TP53. The gene TP53 is located $17p13.1$, which is the short (p) arm of chromosome 17 at position 13.1.

A graph G consists of a pair of disjoint sets (V, E), where V is the set of vertices or nodes and E is the set of edges such that $E \subseteq V \times V$. If the edges has some weights, then the graph is called weighted, otherwise is called unweighted. The adjacency matrix of an unweighted graph G is a square matrix that indicates the relations between vertices. If there exists an edge between two vertices, then this represented by 1 otherwise 0. Let A be a square matrix of size n by n in $\mathbb{R}^{n^2}$. The characteristic polynomial of A is $p(\lambda) = det(A - \lambda I)$, where I denotes the identity matrix. The roots of $p(\lambda) = det(A - \lambda I) = 0$ is called the eigenvalues of A, and the nonzero column vector v satisfying $Av = \lambda v$ is

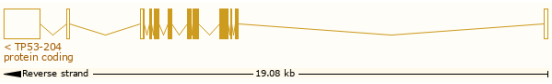
called the eigenvector of λ. The collection of all eigenvalues of the graph G is called the spectrum of G. Let A be the adjacency matrix of the graph G, the energy of graph G, $E(G)$, is defined as follow:

$$E(G) = \sum_{k=1}^{n} |\lambda_k| \tag{1}$$

The energy of a graph is defined by Gutman in [6] and it is publicly stated in a conference in 1978, then there has been a growing interest into this concept in many fields including graph theory, chemistry and recently some efforts accomplished in medicine. The historical background go back to 1930 in the concept of Hückel molecular orbital theory with the fact that the correspondence between the molecular orbital energy levels of π-electrons and the eigenvalues of the adjaceny matrices of the graph represented from the conjugated hydrocarbons, please see [7]. The concept of graph energy can be used to measure the stability of a network, for instance the research in [8] indicates that the Laplacian graph energy of the brain networks significantly lower in the Alzheimer's disease patients than in control group. The energy of "DNA similar graphs" Möbius ladder and Prism graphs have been studied in [9–11].

The TP53 is one of the most important gene that regulates the cell division through tumor suppressor protein (p53). This protein prevents the cells from growing and dividing too fast in uncontrolled way. When the DNA in a cell is damaged by different reasons, the p53 plays important role for making a decision between repairing the damaged region of the DNA or destructing the whole cell (apoptosis) to prevent duplication of damaged chromosomes [12,13]. Mutations in the evolutionarily conserved codons of TP53 are common in many types of cancers and mutational spectrum shows differences among the cancers including the colon, lung, esophagus, breast, liver, brain, reticuloendothelial tissues, and hemopoietic tissues [14]. The sequences of the TP53 gene used in our study can be easily accessed from the transcript TP53-204 (ENST00000420246.6) through the website http://asia.ensembl.org. This transcript has 12 exons (9 coding exons) and consists of 2653 base pairs. TP53 encodes the p53 protein that consists of 341 amino acids. The distributions of the exons over the gene can be represented in Figure 1, [15].

Figure 1. The distribution of Exons in the TP53

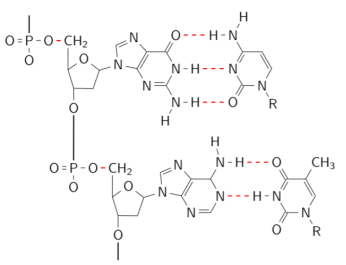


The length of each exon is shown in Table 1. The longest and shortest exons are Exon 12 and Exon 3 respectively. Exon 3 has only 22 base pairs, due to its shortness it is difficult to obtain its energy probability distribution in the physical perspective, therefore it is excluded in our study.

In many cases the edges in a graph are not identical in many perspectives. We know from the tests carried out in laboratory environments that the strengths of the bonds between the molecules that make up DNA are different. For example, the bond energy between Adenine and Thymine and similarly between Guanine and Cytosine are different. Therefore, graphical representation of a gene is only possible with weighted graphs. The weights used in this study were determined by considering the proportional values of the real values of the bonds determined experimentally and theoretically in the literature.

2 Method

Each molecule in DNA is attached via a chemical bond of varying strength. The stability of DNA depends on distribution of the energy over each bond on it. When all possible bonds are compared, the weakest connections are the hydrogen bonds across each strand and the covalent bond between $5'$ carbon of sugar and oxygen atom in the phosphate group, see Figure 2(A). We assume the rest of the bonds are too strong to be broken. This leaves the DNA to be a ladder system. Each leg is series of nucleotides connected via covalent bonds, and legs are connected to each other via hydrogen bonds. In addition, active components of the DNA, exons, are analyzed individually since each exon, by itself or in coordination with other exons, acts as a unified entity.



(a) Hydrogen and ester bonds in DNA.

(b) Graphical representation of DNA.

Figure 2. Structure of DNA (one string) and its graphical representation

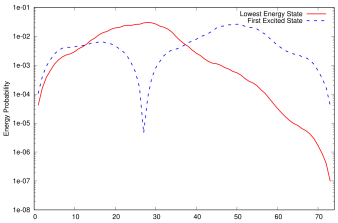


Figure 3. Energy probability distributions of Exon 9 for the lowest energy and first excited energy levels.

We analyze the stability of the exons with these constraints. The stability of the exons depends on the energy distribution of the bonds. Eigenvectors of the Hamiltionan matrix of the bonds contains information about how unified each exon is [16]. We define the energy matrix as:

$$H(i,j) = \begin{cases} B_{co} & \text{if } i = j \pm 1 \\ B_{hy} & \text{if } i = j \pm N \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where i and j are nucleotides in the exon that runs along one leg of the exon, and then the other, B_{co} is the covalent bond (ester bond) between $5'$ carbon of sugar and oxygen atom in the phosphate group, B_{hy} is the hydrogen bond strength across each step of the ladder. B_{co} is constant 92 kcal/mol for all bonds within an exon [17, 18]. As we take the exon as an isolated entity, connection to introns are omitted, therefore for the boundary of the exon, $B_{co} = 0$ (open boundary conditions). Hydrogen bond depends on the type of the nucleotide present for that bond. Between A and T, $B_{hy} = 13$ kcal/mol and between G and C, $B_{hy} = 21$ kcal/mol. Although there are small differences in the literature [19–22], the values used in this study are the average of different studies. Note that The matrix H is a symmetric real valued matrix. We used LAPACK library to diagonalize the matrix to find the eigenvalues and eigenvectors.

$$H|\Psi_k\rangle = \lambda_k|\Psi_k\rangle \tag{3}$$

The eigenvectors contain complete information on the all possible stable energy distributions on an exon. The most probable distribution is given by the ground state (k=0) (eigenvector with the lowest eigenvalue). To calculate the energy distribution on a given site i, we calculate:

$$n_i = \langle\Psi_0|\delta_i|\Psi_0\rangle \tag{4}$$

where δ_i is one for only site i, zero otherwise.

For a system of uniform distribution, open boundary conditions gives the ground state distribution of

$$n_i = \sin^2(\pi i/L) \tag{5}$$

where L is the size of the system. Any deviation from this distribution can be attributed to existence of local variation in energy distribution. Energy at the edge (i=0 or L) is zero. This represents the boundary of a unified structure. If the energy of a particular site is reduced, that site can be considered as less connected to the rest of the structure. We plot the lowest two energy levels in Figure 3. Due to orthogonality of the eigenvectors, second state has a minimum in the middle as well. This represents lowest amount of stress (additional energy) put on the exon and beyond the scope of this analysis, therefore the remaining distributions shown will be for the ground state only.

3 Findings and Results

In this section, first of all, the energies of exons present in TP53 are explained in the mathematical perspective. As is known, the energy of a graph is the sum of the absolute values of the eigenvalues of its adjacency matrix. The data derived from TP53-204 (ENST00000420246.6) shows the sequence of the bases (nucleobases) on one string of the gene, but the other string is naturally known due to pairing rules of the bases. If the number of the bases in any exon is n, then corresponding graph has $2n$ vertices and $3n-2$ edges. Moreover, the size of the adjacency matrix is $2n$ by $2n$. For instance, the length of Exon 12 is 1287, so corresponding graph has 2574 vertices and 3859 edges, and the size of the adjacency matrix is 2574 by 2574. The large matrices make it inevitable to use a software program (LAPACK library) to calculate the eigenvalues. The mathematical energies of each exon are given in Table 1.

Table 1. Energies and Average Energies of Exons in TP53.

Energies of Exons in TP53.												
Exon Number	1	2	3	4	5	6	7	8	9	10	11	12
Number of bases	105	102	22	279	184	113	110	137	74	133	107	1287
Energy (kcal)	3442.60	3342.89	712.75	9233.22	6105.17	3707.34	3608.06	4506.89	2410.18	4370.32	3509.45	42761.88
Average Energy(kcal)	32.79	32.77	32.40	33.09	33.18	32.80	32.80	32.90	32.57	32.86	32.80	33.23

The first observation from Table 1 is that the positive correlation between the length of the exon and its energy. On the other hand, the average energies (energy/the number

of the basepairs) are close to each others regardless of their lengths. Descriptive statistics reveal that the mean of the average energies is 32.85 kcal and the standard deviation of the average energies is 0.23 kcal.

We only know the total energy that stored in each exon by $E(G) = \sum_{k=1}^{n} |\lambda_k|$, whereas the main point here is how the energies are distributed over the exons. Low energy regions are more vulnerable to physical or chemical agents than the regions where the energy is densely stored. To remedy this deficiency, the problem has to be approached a little more through interdisciplinary and this is where the physics is involved.

We start with Figure 3. The main reason to share the energy probability distribution of exon 9 is to show its behaviour under stress and this can be seen in blue dash line that is the first excited state. In another words If exon 9 is subjected to a physical force, this effect will show itself from where the energy is least stored. This can be compared to the fact that when you try to break an inhomogeneous stick, it will break at its weakest place. There are 74 bases on the exon 9 and we expect the first response to any physical force between the 20th and 30th bases ...CTCCTCTCCCC..., mostly on the neighborhood of 27th base.

With exon 12 in Figure 4, we observe a similar break in lowest energy state. In literature, size and relevance of exon 12 is questioned. Exon 12 is located on the 3′-untranslated region. In our calculations, we observe a localized energy distribution for the proposed size of the exon. Within the concept of the energy distribution, energy is not continuously distributed in the exon. We attribute this to possible variation in true exon size. We have repeated the same calculation with other long exons in other genes (not shown) and found that size is not the issue.

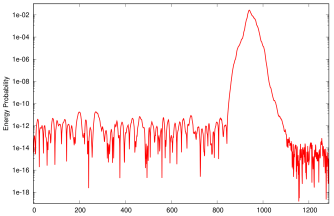


Figure 4. Energy probability distributions of Exon 12.

In Figure 5 the energy probability distributions of exons 1,2,4,5,6,7,8,10,11 are shown, as stated before we excluded exon 3 since its length is two short to obtain a meaningful

distribution.

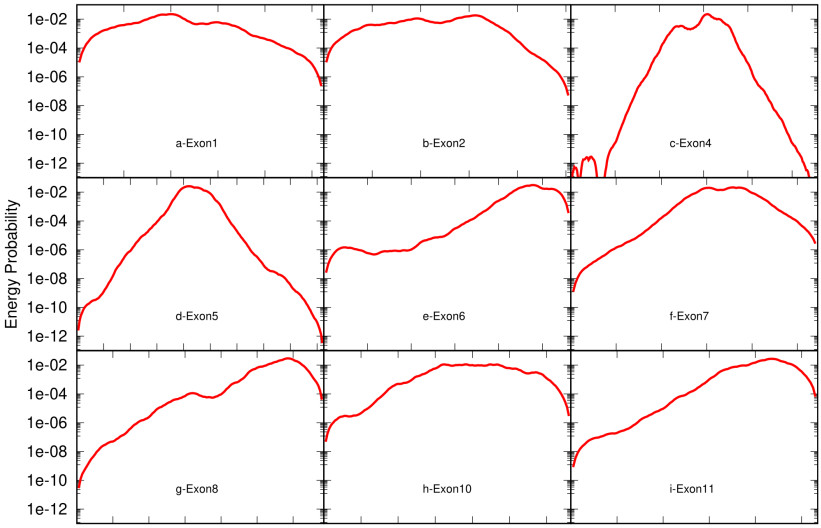


Figure 5. Energy probability distribution of Exons 1,2,4,5,6,7,8,10,11.

In each graph the horizontal axes represent the sequences of bases obtained from the transcript TP53-204 (ENST00000420246.6) and the vertical axes represent the energy probability distribution. We use logarithmic scale on the vertical axes to have better visualization. The ordering of the bases on the horizontal axes are from $5'$ to $3'$ (reverse strand), please see Figure 1. In this study, we mainly focus Exon 4 since its energy probability distribution is not uniformly distributed, as we move along the x-axis (from left to right), we see that its energy suddenly decreases between the bases 263^{rd} (the first base of codon 44) and 274^{th} (the last base of the codon 47) out of 2653. The amino acids corresponding to these codons (Methionine, Leucine, Serine, and Proline) are shown below.

$$\ldots \mathrm{G} - \overbrace{\mathrm{ATG}}^{\text{codon 44}}_{\mathrm{M}} - \overbrace{\mathrm{CTG}}^{\text{codon 45}}_{\mathrm{L}} - \overbrace{\mathrm{TCC}}^{\text{codon 46}}_{\mathrm{S}} - \overbrace{\mathrm{CCG}}^{\text{codon 47}}_{\mathrm{P}} - \mathrm{G} \ldots \tag{6}$$

Therefore; the least stable location of the gene is exon 4. If Exon 4 is exposed to any physical or chemical agents, then the mentioned region is more likely to be affected rather than other regions and this region might end up by some type of mutation. We summarized all variations, Table 2, from codon 44 to 47 through the transcript TP53-204

ENST00000420246.6. We use standard coordinates along with cDNA position number to highlight the bases. Most of the mutations in Table 2 are missense mutations. In missense mutation, a single nucleotide is replaced with another nucleotide that changes the sequencing of the codon that might end up by coding a different amino acids.

Table 2. Variations from codon 44 to codon 47 in Exon 4

	Location	cDNA position	Protein position	Alleles	Codons	Amino Acids	Consequences
atG	17:7676237	265	44	C/T	atG/atA	M/I	missense variant
Ctg	17:7676236	266	45	G/C	Ctg/Gtg	L/V	missense variant
cTg	17:7676235	267	45	A/G	cTg/cCg	L/P	missense variant
ctG	17:7676234	268	45	C/T	ctG/ctA	L/L	synonymous variant
Tcc	17:7676233	269	46	A/G	Tcc/Ccc	S/P	missense variant
tCc	17:7676232	270	46	G/T G/AC	tCc/tAc tCc/tGTc	S/Y S/CX	missense variant frameshift variant
tcC	17:7676229-7676231	271	46	GGG/C		SP/SX	frameshift variant
Ccg	17:7676230	272	47	G/A/T		P/S/T	missense variant
cCg	17:7676229	273	47	G/C	cCg/cGg	P/R	missense variant
ccG	17:7676228	274	47	C/A/T		P/P/P	synonymous variant

4 Conclusion

We have introduced a new way of analysis of nucleotide sequence in exons. Physical energy distribution is the main tool of the analysis. As an example, exon transcripts of TP53 gene has been used. Two critical outcomes have been focused. Exon 4 appears to be prone to mutation. Exon 12 as it is identified in the transcript does not appear to be a complete unified structure. Examination of other genes will be included in future studies.

References

[1] R. Dahm, Friedrich Miescher and the discovery of DNA, *Dev. Biol.* **278** (2005) 274–288.

[2] M. E. Jones, Albrecht Kossel, A biographical sketch, *Yale J. Biol. Med.* **26** (1953) 80–97.

[3] E. Frixione, L. Ruiz-Zamarripa, The "scientific catastrophe" in nucleic acids research that boosted molecular biology, *J. Biol. Chem.* **294** (2019) 2249–2255.

[4] E. Chargaff, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation, *Experientia* **6** (1950) 201–209.

[5] J. Watson, F. Crick, Molecular structure of nucleic acids, *Nature* **171** (1953) 737–738.

[6] I. Gutman, The energy of a graph, *Ber. Math. Statist. Sekt. Forschz. Graz* **103** (1978) 1–22.

[7] X. Li, Y. Shi, I. Gutman, *Graph Energy*, Springer, New York, 2012.

[8] M. Daianu, A. Mezher, N. Jahanshad, D. P. Hibar, T. M. Nir, C. R. Jack, M. W. Weiner, M. A. Bernstein, P. M. Thompson, Spectral graph theory and graph energy

metrics show evidence for the alzheimer's disease disconnection syndrome in apoe-4 risk gene carriers, *IEEE 12th Int. Symp. Biomed. Imag. (ISBI)* (2015) 458–461.

[9] A. Bulut, I. Hacioglu, The energy of all connected cubic circulant graphs, *Lin. Multilin. Algebra* **68** (2020) 679–685.

[10] A. Bulut, I. Hacioglu, Asymptotic energy of connected cubic circulant graphs, *AKCE Int. J. Graphs Comb.* **18** (2021) 25–28.

[11] A. Bulut, I. Hacioglu, K. Kaskaloglu, The minimal and maximal energies of all cubic circulant graphs, *AKCE Int. J. Graphs Comb.*, in press.

[12] A. J. Levine, p53, the cellular gatekeeper for growth and division, *Cell* **88** (1997) 323–331.

[13] A. J. Levine, M. Oren, The first 30 years of p53: growing ever more complex, *Nature Rev. Cancer* **9** (2009) 749–758.

[14] M. Hollstein, D. Sidransky, B. Vogelstein, C. C. Harris, p53 mutations in human cancers, *Science* **253** (1991) 49–53.

[15] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, J. C. Marugán, C. Cummins, C. Davidson, K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore, M. Nuhn, D. N. Oheh, A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, G. IIsley, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion, F. Cunningham, K. L. Howe, D. R. Zerbino, P. Flicek, Ensembl 2020, *Nucleic Acids Res.* **48** (2019) D682–D688.

[16] J. J. Sakurai, *Modern Quantum Mechanics*, Addison–Wesley, Reading, 1994.

[17] S. W. Benson, III - Bond energies, *J. Chem. Edu.* **42** (1965) 502–518.

[18] S. J. Blanksby, G. B. Ellison, Bond dissociation energies of organic molecules, *Acc. Chem. Res.* **36** (2003) 255–263.

[19] C. F. Guerra, F. M. Bickelhaupt, J. G. Snijders, E. J. Baerends, The nature of the hydrogen bond in DNA base pairs: the role of charge transfer and resonance assistance, *Chem. Eur. J.* **5** (1999) 3581–3594.

[20] C. F. Guerra, F. M. Bickelhaupt, J. G. Snijders, E. J. Baerends, Hydrogen bonding in DNA base pairs: Reconciliation of theory and experiment, *J. Am. Chem. Soc.* **122** (2000) 4117–4128.

[21] H. Szatylowicz, N. Sadlej-Sosnowska, Characterizing the strength of individual hydrogen bonds in DNA base pairs, *J. Chem. Inf. Model.* **50** (2010) 2151–2161.

[22] I. K. Yanson, A. B. Teplitsky, L. F. Sukhodub, Experimental studies of molecular interactions between nitrogen bases of nucleic acids, *Biopolymers* **18** (1979) 1149–1170.