Bioinformatics, 35(16), 2019, 2738–2748
doi: 10.1093/bioinformatics/bty1021
Advance Access Publication Date: 7 January 2019
Original Paper

OXFORD

Sequence analysis

The DNA walk and its demonstration of deterministic chaos—relevance to genomic alterations in lung cancer

Blake Hewelt [1], Haiqing Li[2], Mohit Kumar Jolly[3,4], Prakash Kulkarni[1], Isa Mambetsariev[1] and Ravi Salgia[1,*]

[1]Department of Medical Oncology and Therapeutics Research and [2]Department of Computational & Quantitative Medicine, Beckman Research Institute, City of Hope, Duarte, CA 91010, USA, [3]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA and [4]Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on July 16, 2018; revised on December 5, 2018; editorial decision on December 6, 2018; accepted on January 4, 2019;

Abstract

Motivation: Advancements in cancer genetics have facilitated the development of therapies with actionable mutations. Although mutated genes have been studied extensively, their chaotic behavior has not been appreciated. Thus, in contrast to naïve DNA, mutated DNA sequences can display characteristics of unpredictability and sensitivity to the initial conditions that may be dictated by the environment, expression patterns and presence of other genomic alterations. Employing a DNA walk as a form of 2D analysis of the nucleotide sequence, we demonstrate that chaotic behavior in the sequence of a mutated gene can be predicted.

Results: Using fractal analysis for these DNA walks, we have determined the complexity and nucleotide variance of commonly observed mutated genes in non-small cell lung cancer, and their wild-type counterparts. DNA walks for wild-type genes demonstrate varying levels of chaos, with BRAF, NTRK1 and MET exhibiting greater levels of chaos than KRAS, paxillin and EGFR. Analyzing changes in chaotic properties, such as changes in periodicity and linearity, reveal that while deletion mutations indicate a notable disruption in fractal 'self-similarity', fusion mutations demonstrate bifurcations between the two genes. Our results suggest that the fractals generated by DNA walks can yield important insights into potential consequences of these mutated genes.

Availability and implementation: Introduction to Turtle graphics in Python is an open source article on learning to develop a script for Turtle graphics in Python, freely available on the web at https://docs.python.org/2/library/turtle.html. cDNA sequences were obtained through NCBI RefSeq database, an open source database that contains information on a large array of genes, such as their nucleotide and amino acid sequences, freely available at https://www.ncbi.nlm.nih.gov/refseq/. FracLac plugin for Fractal analysis in ImageJ is an open source plugin for the ImageJ program to perform fractal analysis, free to download at https://imagej.nih.gov/ij/plugins/fraclac/FLHelp/Introduction.html.

Contact: rsalgia@coh.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Lung cancer is a devastating genetic disease that has become the most commonly diagnosed form of cancer with over 1.6 million cases each year and the leading cause of cancer deaths (Herbst *et al.*, 2018). Advancements in cancer genetics have identified numerous mutations that can play important roles in disease pathogenesis. Furthermore, they have also helped discover novel, 'targeted', treatments for patients harboring these mutations in several types of cancer, especially non-small cell lung cancer (NSCLC) (Herbst *et al.*, 2018). However, while chaotic disruptions at a tissue organization level have been well-studied in cancer progression, chaos demonstrated within the sequences of the oncogenes that are causative, remains underappreciated (Soto and Sonnenschein, 2011). As we arrive at further precision medicine, understanding genetic function (or dysfunction) will help guide us even more for better outcomes for patients with lung cancer and other tumors.

Chaos refers to very high sensitivity to initial conditions, and consequent unpredictability (Oestreicher, 2007). The manifestation of chaos in genetic sequences can be studied via DNA walks—graphical representations of the genomic sequence where the steps in the 'walk' are the nucleotides in the corresponding nucleic acid sequence. Using the concepts of fractals and its associated parameters—fractal dimension (FD) and lacunarity—one can quantify the extent of fractal behavior and chaos in a given system (Lennon *et al.*, 2015). Fractals are geometric objects that are self-similar (i.e. any small part of the object is an exact replica of the same) and have a non-integer FD that describes the intrinsic shape of the object (Lennon *et al.*, 2015). Lacunarity, or pattern variance, is another metric, which describes the texture of the object at hand. The values of FD and lacunarity may be determined with the box-counting method, a method of analyzing complex shapes and patterns by analyzing the object or image in a defined grid (Lennon *et al.*, 2015). Binary box-counting determines the value of each box in the grid as 1 if a part of the image is contained within the box or 0 if the box is empty and the FD is calculated by analyzing the boxes in the grid that contain the image.

FD has been used to gain insights into diverse local and global traits of many DNA sequences, as well as represent a subset of chaotic function. This analysis has its origins in DNA walks. In this method described by Hamori and Ruskin (1983), the information content of a nucleotide sequence is mapped into a 3D space function called H curves. Each of the four nucleotides is represented by a vector in 3D space having a characteristic, but variable, orientation. As an alternative to H curves, Gates proposed the 2D DNA sequence representation, which projects C/G changes on x-axis and A/T changes on y-axis. Thus, all genetic sequences can be represented uniquely in 2D, and differences among them can be highlighted (Gates, 1985). This study selects this 2D DNA Walk schema because of its simplicity and efficiency. 2D DNA walk provides useful insights into new global sequence patterns and homologies, repeated structures, relative base abundances and probable evolutionary paths (Buldyrev *et al.*, 1993).

The general properties of nucleotide sequences can be examined using FD. Peng *et al.* (1992a,b) defined the term 'DNA Walk' as a 1D random walk model, where the walk steps up if a pyrimidine occurs at a position i along the DNA chain, whereas the walk steps down if a purine occurs there. The authors showed a long-range power law correlation in intron-containing genes and non-transcribed regulatory DNA sequences (Peng *et al.*, 1992a,b), enabling quantitative measurement of correlations among nucleotides over long distances along the DNA chain, for a given DNA/RNA sequence. Further, this method was also used to analyze the fractal landscape; for instance, fractal complexity was shown to increase during the evolution of myosin heavy chain gene family (Buldyrev *et al.*, 1993). Because of the benefits of graphic representation of the long genomic sequences, DNA walk has been wildly used in the genome-wide periodicities and in identifying large-scale, local and global properties identification (Buldyrev *et al.*, 1993; Peng *et al.*, 1992a,b; Poptsova *et al.*, 2009; Zielinski *et al.*, 2008), in detecting the origin of replication (Lobry, 1996), in comparing whole genomes (Roten *et al.*, 2002), and in uncovering the protein coding regions (Gao *et al.*, 2005).

Several recent studies have used DNA walks as the model for cancer investigation (Lennon *et al.*, 2015) and diagnostics (Namazi and Kiminezhadmalaie, 2015; Namazi *et al.*, 2015) and drug resistance analysis (Saini and Dewan, 2016). A comparison of the FD spectrum of DNA walks of the lung cancer patients' DNA with healthy individuals control group showed that the FD of cancer patients' DNA were significantly higher than that for healthy individuals (Namazi and Kiminezhadmalaie, 2015). A similar pattern was found in skin cancer (Namazi *et al.*, 2015), where the FD of DNA Walk was proposed as a measurement to identify the diseased associated mutations. These attempts largely focus on the macro level changes of sequences between wild-type and cancer cells (Namazi and Kiminezhadmalaie, 2015; Namazi *et al.*, 2015). Our study on the other hand, focuses on the impacts of the micro level mutations on the FD and lacunarity of the walk. We have created these walks for the coding regions of commonly affected genes in NSCLC—anaplastic lymphoma kinase (ALK), BRAF, epidermal growth factor receptor (EGFR), ERBB2, KRAS, MET, Neutrophic receptor kinase 1 (NTRK1), PXN and ROS1—and performed fractal analysis so that we may observe the different levels of chaos demonstrated within the sequence of these genes in their wild-type forms and how this demonstration of chaos may play a role in why these mutations occur. We later generated the DNA walks for mutations in the oncogenes EGFR, MET, ALK and KRAS that often occur in NSCLC in order to analyze the changes in chaotic properties and determine shared characteristics among the multiple mutation types assessed, such as point mutations, deletions, insertions and fusions. As such, we may find common trends in the fractal properties for each mutation and become able to discern the presence of a particular mutation type when a change is observed in these fractal images.

2 Materials and methods

The DNA walk diagram is generated using a tool implemented in python. It includes three modules. The data import module reads the genomic sequence data in different formats (such as FASTA, and NCBI GenBank flat file format). The graphic generation module draws the DNA walk diagram using Python turtle library. The default 2D walk diagram encodes nucleotide as A (west), T (east), C (south) and G(north). The last module is data export module, which exports the diagram to scalable vector graphic format for downstream analysis. This tool can also highlight the different genomic regions (such as exons, introns and mutations) using different colors. For example, specific exons may be color coded to the user's specifications to easily analyze the nucleotide sequence. This allows the user to modify the DNA walk to their specification.

DNA walks for ALK, BRAF, EGFR, ERBB2, KRAS, MET, NTRK1, PXN and ROS1 coding regions were all generated using a tool implemented in Python. The cDNA sequence, based on the mRNA sequence with the 5′- and 3′-untranslated region, was

Table 1. The chromosome location and RefSeq ID for ALK, BRAF, EGFR, ERBB2, KRAS, MET, NTRK1, PXN and ROS1

Gene	Chromosome location	RefSeq ID
ALK	2p23.2-p23.1	NM_004304
BRAF	7q34	NM_004333
EGFR	7p11.2	NM_005228
ERBB2	17q12	NM_001005862
KRAS	12p12.1	NM_004985
MET	7q31.2	NM_000245
NTRK1	1q23.1	NM_001012331
PXN	12q24.23	NM_001243756
ROS1	6q22.1	NM_002944

Table 2. The fractal values of ALK, BRAF, EGFR, ERBB2, KRAS, MET, NTRK1, PXN and ROS1 DNA walks obtained by using the box-counting mono-fractal analysis in the ImageJ plugin FracLac

Gene	Dimension	Lacunarity
ALK	1.4116	0.7007
BRAF	1.504	0.6336
EGFR	1.3627	0.6239
ERBB2	1.5403	0.5683
KRAS	1.3244	0.5742
MET	1.5322	0.5978
NTRK1	1.5316	0.4732
PXN	1.2905	0.5613
ROS1	1.4181	0.8145

obtained from NCBI RefSeq database and used as the input for the Python script. Table 1 lists the RefSeq ID of each gene we used in this study. The DNA walk images were then analyzed through the FracLac plugin for the application ImageJ. Using the binary box-counting method, the FracLac plugin was then able to find the FD and lacunarity of the image.

To observe certain principles of chaos within these systems, we found the coding sequence for genes with different mutations. We generated DNA walks and analyzed the FDs for EGFR L858R, G719A, S768I and T790M point mutations, EGFR exon 19 deletion, EGFR-RAD51 fusion, the EGFR exon 20 insertion variant Val769_Asp770insMetAlaSerValAsp, artificial L858L synonymous mutations and two EGFR point mutation variants of unknown clinical significance (VUS) compared with the DNA walk for the wild-type protein. The synonymous mutations and VUS mutations have been performed to determine the impact mutations that are not actionable or proven to be cancer related have on the DNA walk. For splice site mutations and translocations, we compared the walks and dimensions of three MET exon 14 splice site variants, as well as the KIF5B-MET fusion, to that of wild-type. To further analyze the effects of translocations in these systems, we compared the walks and dimensions of three echinoderm microtubule-associated protein-like 4 EML4-ALK fusions to those of wild-type ALK. Last, we have generated and analyzed the walks for KRAS and three common codon 12 mutations associated with NSCLC to understand the effects mechanistic point mutations have on the walk of a gene that exhibits intrinsic disorder. Additional figures for all the genes described above and the Python script used to perform the DNA walks are provided in the Supplementary Material.

3 Results

3.1 DNA walks and fractal patterns of wild-type proteins

We first observed the DNA walks of EGFR, ALK receptor tyrosine kinase (ALK), MET proto-oncogene, receptor tyrosine kinase (MET), B-RAF proto-oncogene, serine/threonine kinase (BRAF), KRAS proto-oncogene GTPase (KRAS), Erb-B2 receptor tyrosine kinase (ERBB2), NTRK1, ROS proto-oncogene receptor tyrosine kinase 1 (ROS1) and paxillin (PXN) individually in their wild-type forms, and measuring the respective FD and lacunarity (Table 2). The data demonstrated below are obtained through fractal analysis of the DNA walks as a mono-fractal profile; although, multi-fractal analysis of the DNA walks may be performed through the FracLac plugin providing multiple FD values for the DNA walk. Next, we generated all of the wild-type walks together in one system (Fig. 1), to compare the levels of chaos exhibited by each gene. The principles most notably indicated by the integration of all the DNA walks in one system are the periodic coupling, and a fractal profile. At the individual level, we are able to see some dense points of attraction in the walks for the cDNA sequences of ERBB2, NTRK1 and MET in particular. These points of attraction are demonstrated by frequent overlapping in a particular orbit or point within the walk, causing what appears to be large clusters. Because the walks along a genome sequence typically represent how the frequency of each nucleotide in a given nucleotide pair changes locally, the points of attraction in these walks indicate variation in the nucleotide sequence rather than repetitive sequences, as seen in EGFR, KRAS and PXN (Namazi and Kiminezhadmalaie, 2015).

The fractal values for these genes indicate a relationship with base pair variance in the cDNA sequences themselves (Fig. 2). The FD describes the shape and complexity for each walk, allowing us to compare the chaos exhibited by the nucleotide sequences of these genes. In the case of a 2D system such as the DNA walk, FD's value will range between [1.0 and 2.0] for our data, where 1.0 implies the fractal image to be a 1D system. Lacunarity, on the other hand, is a quantitative value for the level of variance within the illustrated genetic walk, where a higher lacunarity suggests larger gaps or free space in the fractal pattern whereas a lower lacunarity translates to lower variance within the 'self-similar' image. Lacunarity is, thus, inversely related to the rate of alternating base pair patterns occurring within the sequence (Lennon *et al.*, 2015). Therefore, genes with higher lacunarity and lower FD such as PXN and KRAS demonstrate less base pair variance in their sequences than the genes with lower lacunarity and higher FD, such as MET and NTRK1.

3.2 Walks and fractal patterns for mutated oncogenes

Next, we generated random walks for mutated EGFR, ALK and MET and compared them to their wild-type counterparts. As shown in Figure 3, walks for the c.2236_2250del exon 19 deletion, Val769_Asp770insMetAlaSerValAsp exon 20 insertion, and EGFR-RAD51 fusion are illustrated where the mutations are indicated in red. The fusion with exons 4 through 10 of RAD51 occurs after exon 24 of EGFR, replacing the linear EGFR exons with the periodically dense RAD51 exons (Konduri *et al.*, 2016). Point mutations for EGFR have also been performed, indicating little change in the system due to the substitution of a single step (Table 3). The deletion of exon 19 decreases the linearity of the system and decreases the distance between two periodic points, resulting in a slightly higher FD than that of the wild-type. Similar to that of the exon 19 deletions EGFR's exon 20 insertion decreases linearity by increasing the variation of the walk's path. EGFR'S exon 20 insertion increased the periodicity in the fractal walk while also being exposed to more holes in the fractal pattern, consequently increasing both FD and

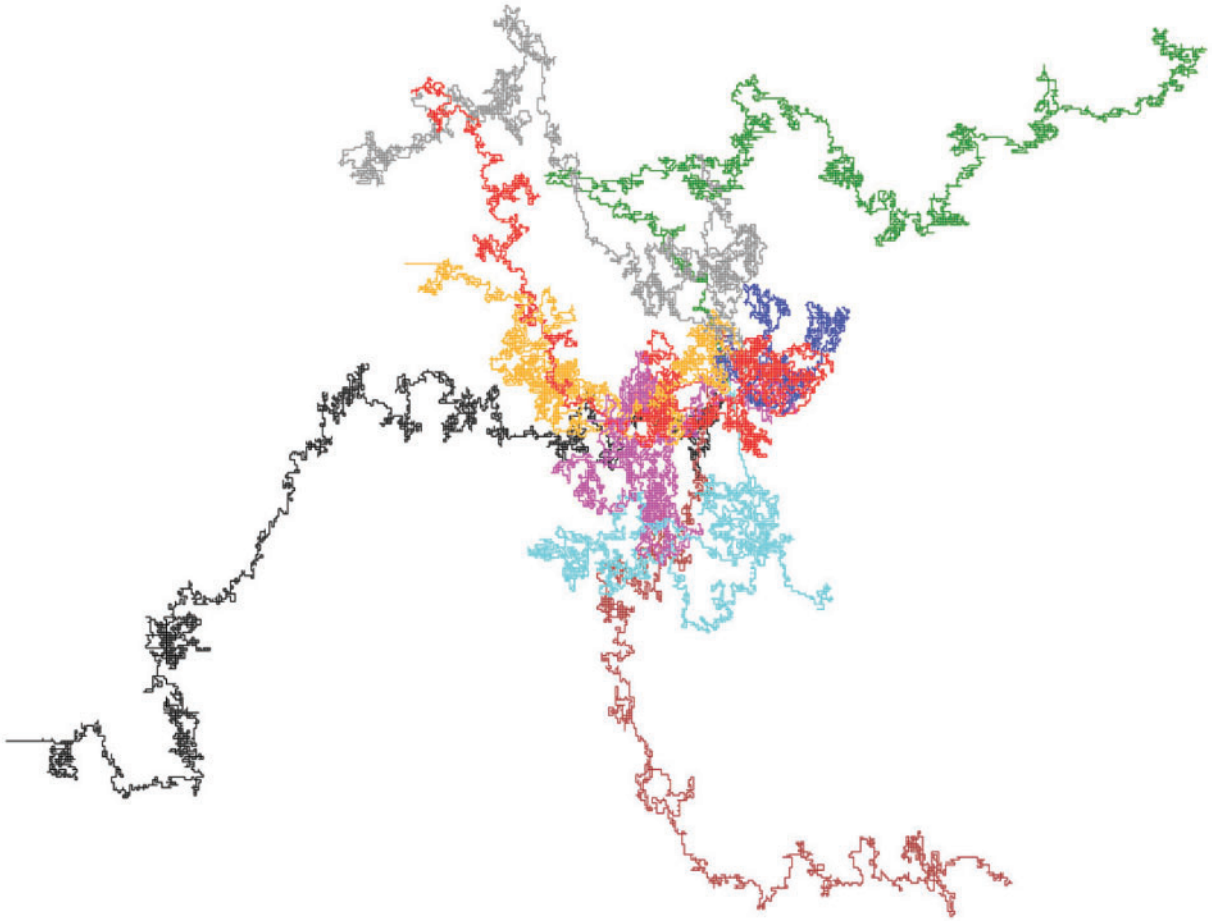

Fig. 1. cDNA walks of the following genes in a single system: ALK (grey), KRAS (green), ROS1 (red), MET (yellow), NTRK1 (blue), ERBB2 (pink), EGFR (black), BRAF (teal) and PXN (brown). The shape of the walk is determined through fractal analysis to determine the characteristics of chaos exhibited. Periodic points, indicated by arrows above, contribute to the shape of these DNA walks. NTRK1, ERBB2 and MET demonstrate dense periodicity where the steps appear to attracted to multiple foci. Meanwhile, EGFR, KRAS and PXN appear to be more linear with weaker and arbitrarily smaller periodic points throughout the walk

lacunarity for this walk. Last, the fusion of EGFR and RAD51 eliminated the exons that contain a linear base pair sequence, thus resulting in a decrease in the walk's overall linearity and increasing the base pair variance.

Using the data collected in Table 3, Figure 4 shown below is a scatter diagram was created to easily observe the FD and lacunarity for each variant of EGFR. Due to the involvement of a second gene causing a significant change in EGFR's original fractal profile, the data for the fusion between EGFR and RAD51 has been omitted from this figure in order to solely analyze the fractal profile of the EGFR gene and its mutations. The P373L point mutation and two of the L858L synonymous mutations appear as a distinct group in the figure below that demonstrates a similar FD to that of the wild-type walk, with an increased lacunarity. The miniscule change in FD for these three mutations may suggest no biologically significant changes to the protein. While the FD for abovementioned mutations were expected to indicate little change, the two other synonymous L858L variants and V441F mutation, however, demonstrate increased FD and lacunarity, with the V441F mutation demonstrating the greatest fractal changes. The c.2574G>T and c.2574G>C synonymous mutations appear as outliers since their FD does not fall within the range the other synonymous variants and P373L mutation have reported. Furthermore, we notice the EGFR L858R mutation has significantly decreased in FD and increased in lacunarity, suggesting that the sensitized point mutation reduces the gene's fractal profile. Meanwhile, we see that the S768I sensitizing mutation demonstrating a more chaotic fractal profile with the increase in FD while the lacunarity has decreased. With the limited data obtained for the EGFR insertion and deletion mutations, these mutations contribute to an increase in fractal 'self-similarity' by increasing the periodicity.

Figure 5 illustrates a MET point mutation, splice site deletion and exon 14 skipping mutation found in lung adenocarcinoma and compares them to the walk for wild-type MET. All of the mutations demonstrate changes in the walk at the Exon 14 of MET's sequence, affecting the enhanced zones illustrated in Figure 5. Unlike the point mutations in the EGFR walk, distinct changes in the MET walk occur due to its point mutation creating such a larger gap in the fractal pattern and increasing the fractal self-similarity in the surrounding space, indicated by the FD and lacunarity in Table 4. The DNA walks for both deletion mutations for MET show signs of reduced nucleotide variance as these gaps within the fractal shapes increase in size. However, the loss of MET's 14th exon entirely reduces the presence of 'self-similarity' in genetic walk, indicating the mutant gene's lack of nucleotide variance. This decreased self-similarity and increased lacunarity is indicated by the reduction of the walk's shape, subsequently providing a greater impact these gaps in the fractal pattern have on the walk's shape. A KIF5B-MET fusion walk

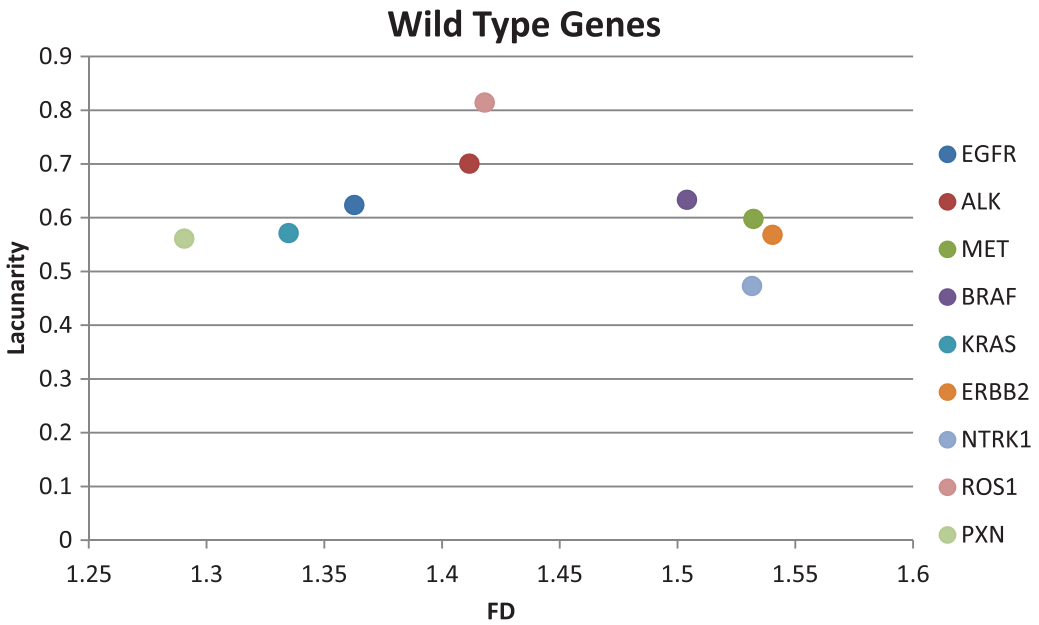

Fig. 2. Scatter diagram comparing the FD and lacunarity of wild-type EGFR, ALK, MET, BRAF, KRAS, ERBB2, NTRK1, ROS1 and PXN based on their values in Table 2

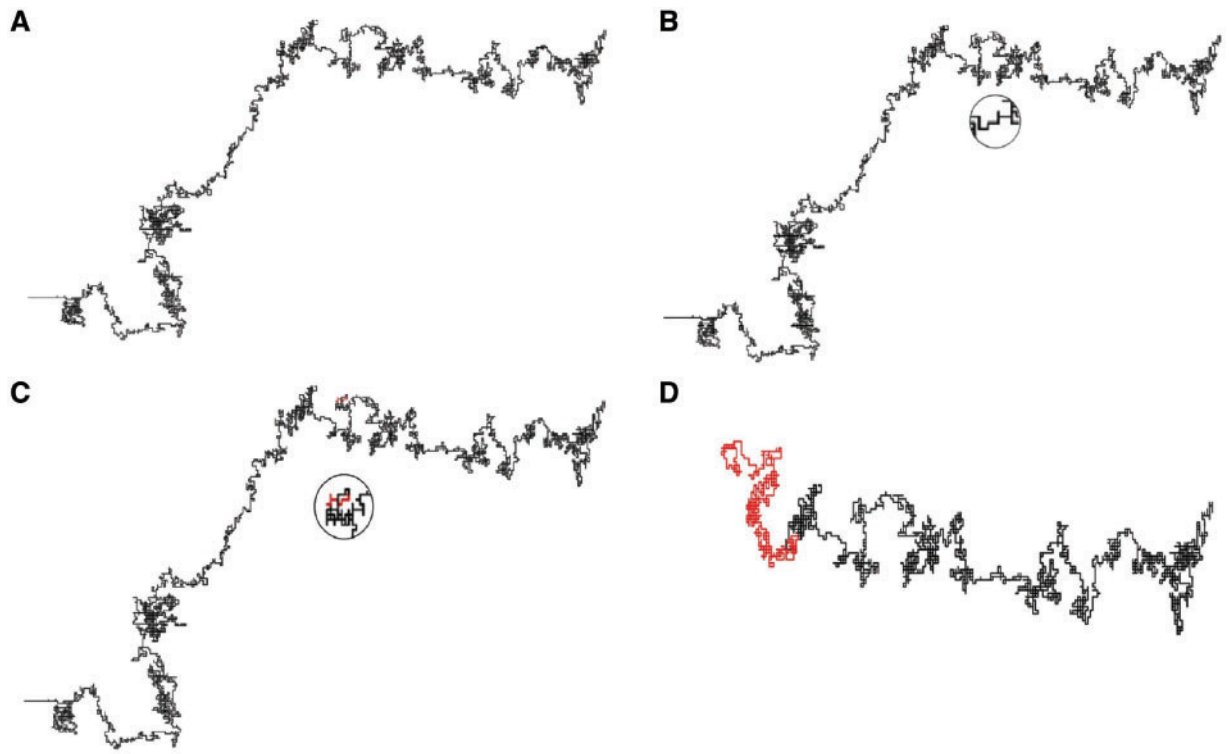

Fig. 3. Genetic walks for wild-type EGFR **(A)**, EGFRc.2236_2250del exon 19 deletion **(B)**, EGFR Val769_Asp770insMetAlaSerValAsp exon 20 insertion **(C)** and RAD51-EGFR fusion mutation where codons 1–892 of EGFR are fused to codons 75–340 of RAD51 (D). The mutations for the exon 19 deletion and exon 20 insertion are enhanced to visualize the changes of the walk. EGFR's exon 19 deletion reduces the length of the walk's path between two dense points on the north end of the walk, as highlighted in Figure 2B, which reduces the fractal's size and increases its pattern complexity. The exon 20 insertion increases the path variation in the north periodic cluster, as highlighted in Figure 2C, increasing fractal's dimension

was also generated, resulting in exon sequences from KIF5B favoring guanine and adenine followed by exons 14–21 of MET and the poly-A tail. This introduction of a linear segment in a DNA walk of high complexity and base pair variation greatly increased the lacunarity while significantly reducing the FD.

We have also compared the data for wild-type MET and the mutant counterparts that involve the gene's 14th exon in a scatter diagram (Fig. 6). Similar to EGFR, we had omitted our KIF5B-MET fusion data from the scatter plot to focus solely on the MET gene. We first notice that the exon 14 loss mutation indicated the greatest

increase in lacunarity, and that the absence of this exon caused greater holes in the fractal pattern. It is also worth noting that due to an increase in FD and decrease in lacunarity, the point mutation at codon 3007 (T > A) suggests that MET's mutated DNA has become more fractal. Although both the splice site variant deletion mutations demonstrate greater lacunarity as compared with that of the wild-type gene, the FD of these two splice site variant deletions indicates that two different groups in the gene's 14th exon have been targeted. The increase in FD in the c.2888-5_2944del variant suggests that the codon range is a rather linear dataset. On the contrary, the deletion of the codon range 3001–3021 caused the FD to decrease, suggesting that this codon region is a periodic and fractal in nature.

Table 3. The fractal values for the mutated EGFR DNA walks in comparison to their wild-type counterpart

Gene	Mutation	Dimension	Lacunarity
EGFR	Wild-type	1.3627	0.6239
EGFR	Exon 19 deletion	1.3645	0.6318
EGFR	T790M	1.3625	0.6307
EGFR	L858R	1.3614	0.6281
EGFR	G719A	1.3639	0.6266
EGFR	S768I	1.3643	0.622
EGFR	Exon 20 insertion	1.3638	0.6249
EGFR	EGFR-RAD51 Fusion	1.465	0.5417
EGFR	V441F (VUS)	1.3645	0.6271
EGFR	P373L (VUS)	1.3625	0.629
EGFR	L858L c.2572C>T	1.3629	0.627
EGFR	L858L c.2574G>T	1.3643	0.6268
EGFR	L858L c.2574G>A	1.3626	0.6277
EGFR	L858L c.2574G>C	1.3639	0.6259

Figure 7 demonstrates the wild-type ALK gene along with three common variants of the EML4–ALK fusion mutations in lung adenocarcinoma. These three fusions discussed commonly connect at the 20th exon of ALK'S genome sequence; however the coding sequence of the EML4 gene may stop at different exons of the N- terminal half, where the quantitative characteristics of the gene, such as size and fractal shape, will vary (Li *et al.*, 2014). The three fusion variants share the linear characteristics demonstrated by EML4's sequence and the reduction of the periodicity initially seen near the beginning of ALK's sequence. On observation of DNA walks and quantitative analysis of the gene's 'complexity', the ALK's fusions appear as more linear systems that grossly favor the presence of adenine, even more so than the DNA walk noted in the wild-type protein. The increased linearity of the mutated gene's sequence results in a decreased FD, as shown in Table 5. However, the lacunarity changes indicate that the gaps within the fractal are fewer or smaller, resulting in more consistent fractal patterns throughout the walk. The EML4-ALK fusion Variant B demonstrates the lowest FD and lacunarity among the four tested DNA walks for ALK, indicating that the walk for this variant exhibits the most linearity and the least chaos of the four.

We have graphically demonstrated the differences in FD and lacunarity between ALK and its translocation mutations in Figure 8 below. The three fusions demonstrate a common trend in increased linearity, as both the FD and lacunarity have decreased significantly. Notably, the EML4–ALK fusion variant B has the lowest FD and

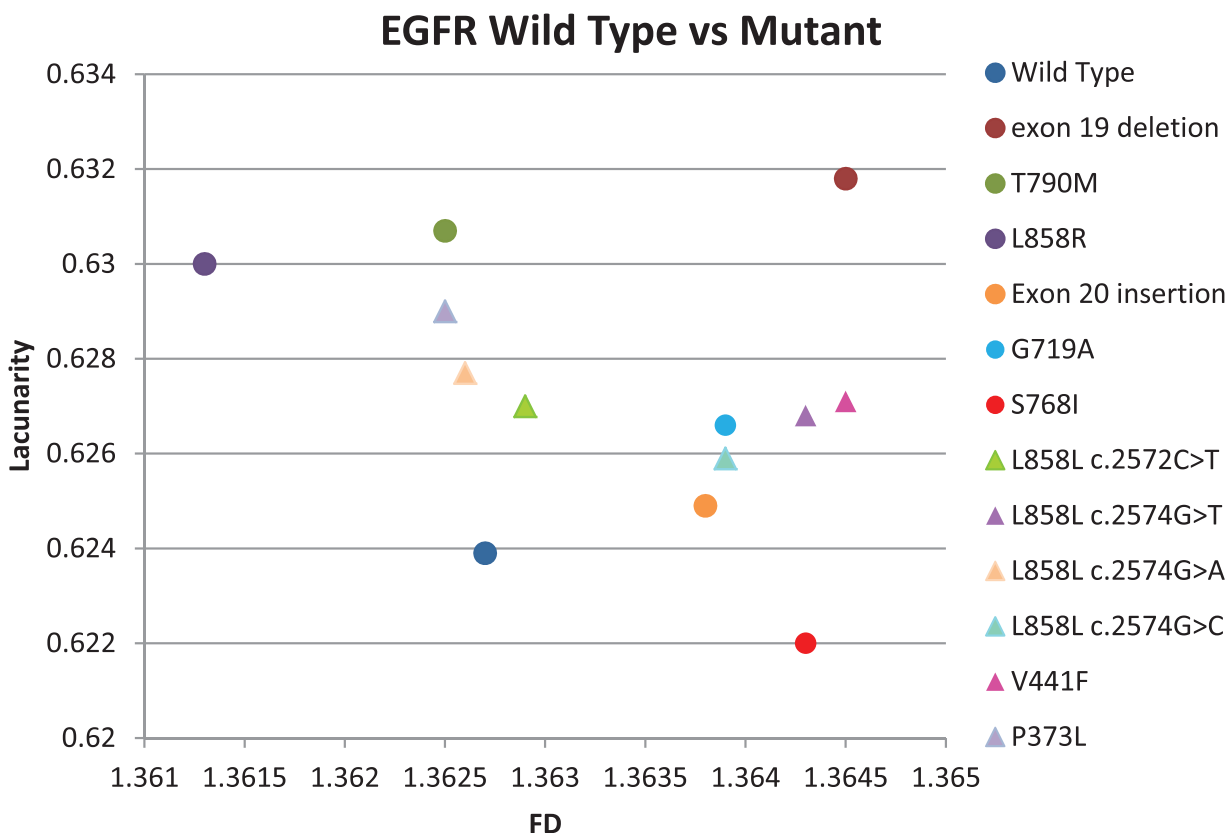

Fig. 4. Scatter diagram comparing both FD and lacunarity of wild-type EGFR with its Exon 19 deletion, T790M, L858R, Exon 20 insertion, G719A, S768I, L858L, V441F and P373L mutations. Wild-type EGFR and its clinically significant mutations are indicated by a circle while variants of unknown clinical significance are indicated by a triangle

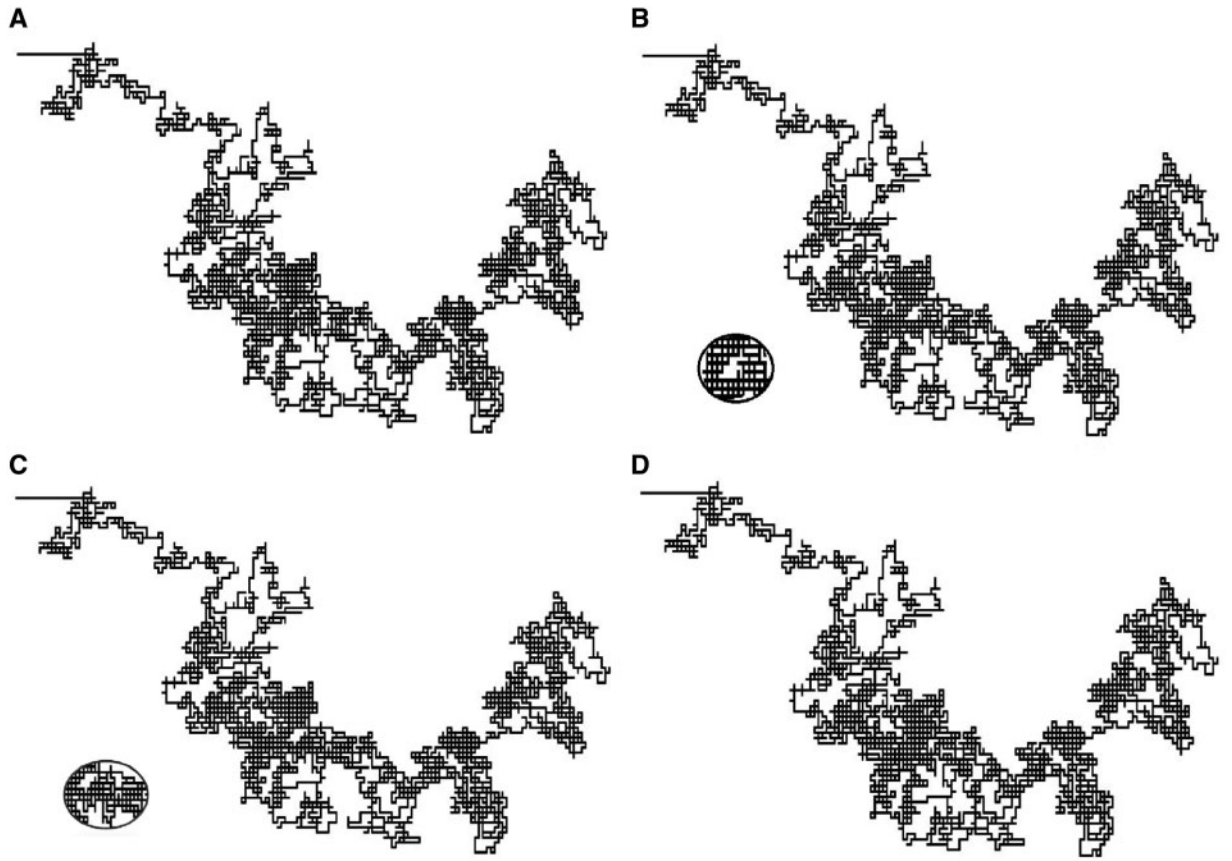

Fig. 5. Comparison of the genetic walks for MET wild-type **(A)**, MET c.3007 T > A point mutation **(B)**, MET c.3001_3021del splice site variant mutation **(C)** and MET exon 14 loss mutation **(D)**. The effects of the point mutation and deletion on the genetic walk are enhanced to visualize how the chaos driven within these systems has been changed

Table 4. The fractal values for the mutated MET DNA walks in comparison to their wild-type counterpart

Gene	Mutation	Dimension	Lacunarity
MET	Wild-type	1.5322	0.5978
MET	c.2888-5_2944del62	1.5342	0.5998
MET	c.3001_3021del	1.5303	0.6014
MET	c.3007 T>A	1.5326	0.5964
MET	Exon 14 loss	1.5316	0.6127
MET	KIF5B-MET fusion	1.2493	1.0061

lacunarity due to the absence of the periodic point seen immediately after ALK's 20th exon.

KRAS gene is a possible predictive biomarker for drug sensitivity for patients with NSCLC (Karachaliou *et al.* 2013). Its mutations are most frequently seen in exons 2 and 3 (specifically codons 12, 13 for KRAS-mutant NSCLC (Karachaliou *et al.*, 2013). Of the several point mutations that affect codon 12 of the gene's sequence, we selected three point mutations with single base pair substitution in order to determine how a single step change will affect the system at greater depth, seen in Table 6. The changes within the FD for the DNA walks of the KRAS gene indicate a minor shift toward self-similarity. The decrease in lacunarity further demonstrates greater complexity and fractal self-similarity to that of the wild-type gene. However, the demonstration of chaos exhibited by these particular point mutations may be difficult to assess with the DNA walk alone. Therefore, the graphical representation of FD and lacunarity in the form of a scatter diagram may be seen below in Figure 9, where the mutant variants of KRAS demonstrate the common trend of reduced lacunarity and increased FD.

4 Discussion

Performing a DNA walk on a particular wild-type gene provides information regarding that particular gene's linearity and fractal self-similarity. A DNA walk on the gene's mutated counterpart implicated in cancer progression may demonstrate characteristic changes in chaos that may be associated to changes in the protein's function. One may analyze the differences in the fractal walk between the mutant gene of interest and its wild-type counterpart, as one observes the changes in the periodicity and the spaces between the walk's points of attraction, as well as the deviation from the walk's original path caused by the mutation. This form of analysis can be easily done by comparing the DNA walks side-by-side or by superimposition. Furthermore, FD and lacunarity are useful tools in analyzing dimensionless systems such as the DNA walk. They allow us to quantitatively compare DNA walks based on their measured linearity, as well as the sequence patterns that exist within the walks. The data presented here provides a framework to observe the changes in the fractal patterns of the nucleotide sequences of a wild-type and corresponding mutated gene(s), and qualitatively assess what these fractal changes mean functionally.

Although our dataset is limited to few common lung adenocarcinoma mutations, we can see the trends present in DNA walks with

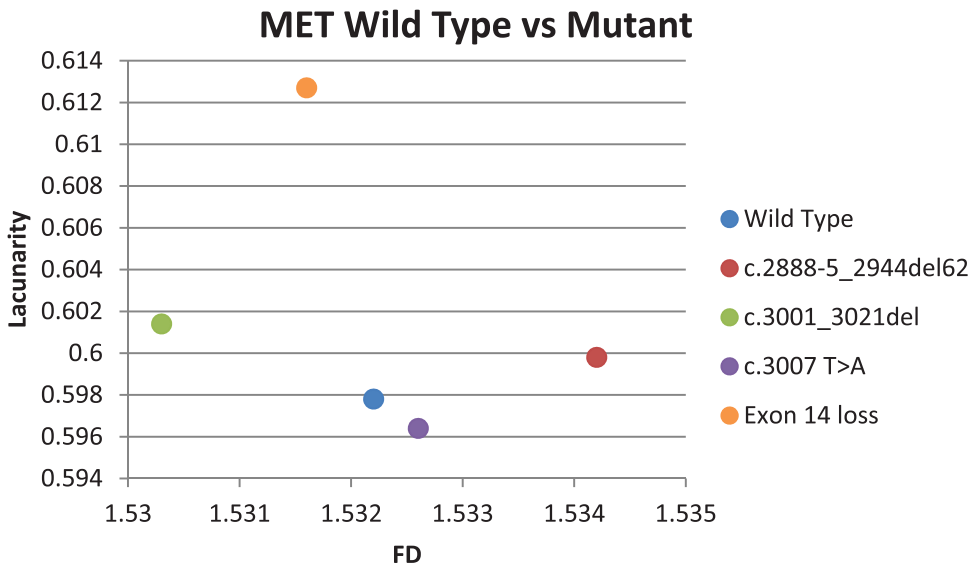

Fig. 6. Scatter diagram comparing both FD and lacunarity of wild-type MET with its mutant counterparts

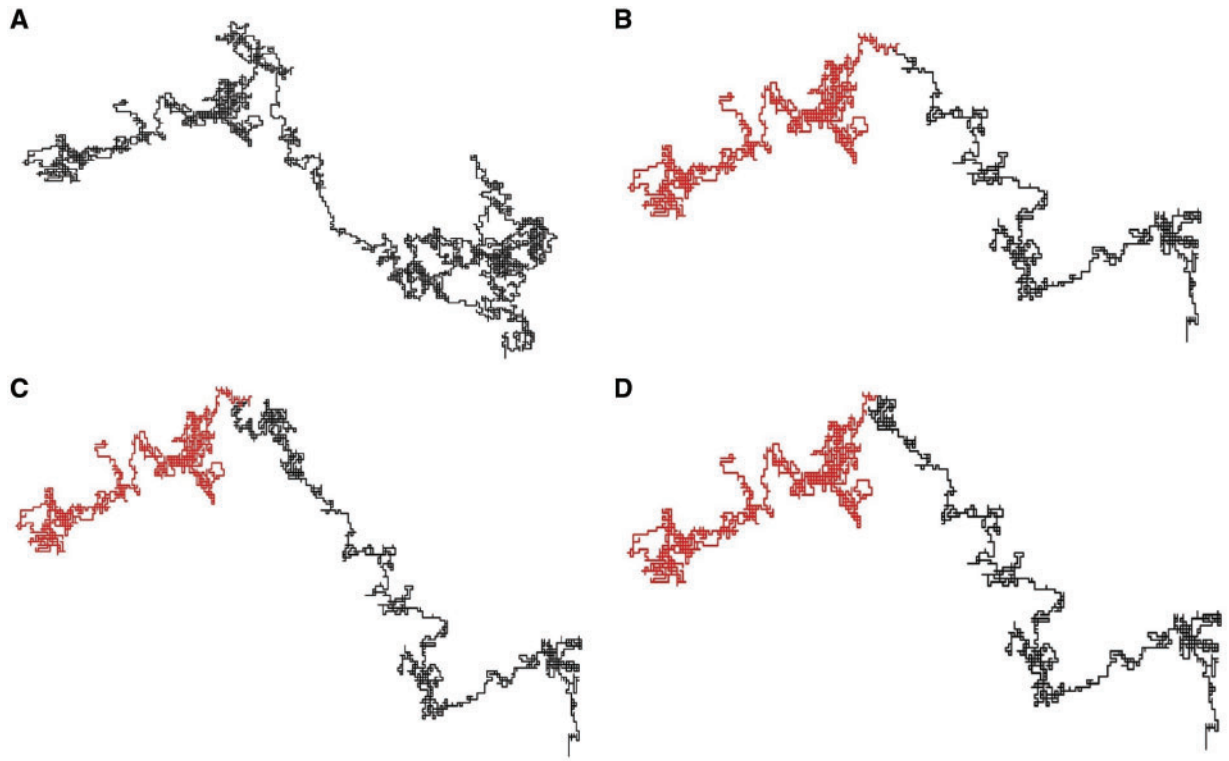

Fig. 7. Genetic walks for wild-type ALK **(A)**, EML4-ALK fusion variant A where the fusion mutation takes place at exon 13 of EML4 and exon 20 of ALK **(B)**, variant B of EML4-ALK fusion where the exon 20 of EML4 is fused with exon 20 of ALK **(C)** and Variant C of the fusion where codon 569 of EML4 is connected to codon 1078 of ALK **(D)**. ALK is illustrated in red while EML4 is illustrated in black for the mutant walks. The introduction of EML4's nucleotide sequence causes an increase in overall linearity

Table 5. The fractal values for the DNA walks for three EML4-ALK fusion variants in comparison to wild-type ALK

Gene	Mutation	Dimension	Lacunarity
ALK	Wild-type	1.4116	0.7007
ALK	EML4-ALK fusion A	1.3776	0.6474
ALK	EML4-ALK fusion B	1.339	0.6287
ALK	EML4-ALK fusion C	1.3817	0.6395

deletion and insertion mutations and gene fusions. Deletion mutations appear to disrupt the fractal patterns by removing a consistent pattern within the sequence or a string of nucleotides to create a more complex system. Therefore, this disruption of the fractal images causes an alteration in nucleotide variance and an increase in lacunarity. Periodicity is described to be less dense, the system's complexity becomes altered, and the sequence becomes less organized. Although we do not have sufficient data for insertion

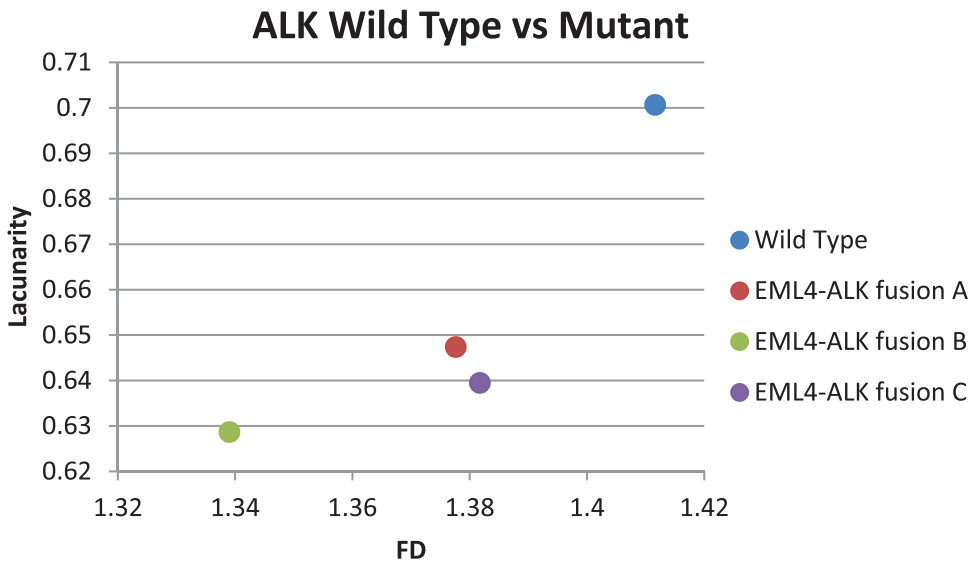

Fig. 8. Scatter diagram comparing the FD and lacunarity of ALK with the three EML4-ALK fusion variants described in Figure 7

Table 6. The fractal values for the DNA walks of KRAS with mutations in the 12th codon in comparison to wild-type KRAS

Gene	Mutation	Dimension	Lacunarity
KRAS	Wild-type	1.3244	0.5742
KRAS	G12C	1.3278	0.5669
KRAS	G12D	1.327	0.5692
KRAS	G12V	1.3286	0.5668

mutations, EGFR's exon 20 insertion increases periodicity in the walk. This increased periodicity indicates greater nucleotide variance within the mutated gene. Fusions, however, appear to make one of the two wild-type systems involved less complex as it combines one gene with less nucleotide variance (e.g. EML4 and KIF5B) with another that appears to have a relatively higher nucleotide variance (e.g. MET and ALK). This change is indicated by a bifurcation in the DNA walk of the mutated gene where, in one regime, it is characterized by relatively high complexity and periodicity, while the other demonstrates relatively high linearity and thus low complexity. This visual cue in the DNA walks for proteins with fusion mutations may indicate the existence of a change that is common for such mutations.

EGFR mutations are indicated in approximately 20% of lung adenocarcinoma cases (Dogan *et al.*, 2012), with activating mutations sensitive to TKI's such as gefitinib and erlotinib comprising of over 90% of these cases (Lynch *et al.*, 2004). The most prevalent of these TKI-sensitive mutations are the deletion mutations in exon 19 and the point mutation L858R in exon 21, with in-frame insertion mutations in exon 20 also being noted (Dogan *et al.*, 2012). EGFR mutant lung cancers begin to develop resistance to first generation TKI's such as erlotinib through the acquired T790M point mutation (Yu *et al.*, 2014). With advancements in targeted medicine, the third generation EGFR TKI osimertinib is introduced, indicating response for cases with the T790M resistance mutation (Cross *et al.*, 2014). The variety of actionable mutations in EGFR is useful in analyzing the chaotic behavior that exists in this gene. With further research in such genes that exhibit a wide variety of sensitizing mutations in cancers—from point mutations to deletions or insertion—such as EGFR, we can determine how particular mutations affect the DNA walks of these genes and possibly understand the set of mechanistic changes that these mutations may have in common.

Dysregulation of MET through amplifications, overexpression and somatic alterations have been indicated in NSCLC (Sattler and Salgia, 2016), with a variety of somatic alterations in MET's exon 14 being indicated as a therapeutic target (Frampton *et al.*, 2015). The exon 14 splice site alterations in MET, including point mutations, deletions and insertion–deletion (indel) mutations, affect the juxtamembrane domain and potentially lead to exon 14 skipping (Cortot *et al.*, 2017). MET's diverse exon 14 splice site alterations are useful in determining what the alterations with shared mechanistic transformations may have in common in terms of chaos, as well as potentially predicting whether an unknown somatic alteration in MET's 14th exon is capable of indicating exon 14 skipping.

ALK is notable for its translocation mutations in lung adenocarcinoma that can be targeted with therapy such as crizotinib and alectinib (Costa *et al.*, 2018). In particular, an inversion in the 2p chromosome develops a fusion between the EML4 gene and ALK genes (Li *et al.*, 2014). Multiple variants of the EML4–ALK fusions have been noted in primary lung cancers based on the EML4 exons involved in the N-terminal half (Li *et al.*, 2014). Analyzing EML4-ALK fusion mutations through DNA walks allows us to determine the shifts in its chaotic properties in order to predict mechanistic changes shared among other fusion mutations—such as ROS1, BET and ERG—that occur in cancer. Furthermore, analyzing the variance in the fusion locations may also provide insight on the biological nuances that exist with each mutation.

Limitations in the above mentioned quantitative analysis occur for DNA walks for oncogenes with substitution mutations. The changes these mutations have on the DNA walk's fractal properties cannot be fully appreciated through visual observation alone. Although the paths of the EGFR DNA walks do not appear to vary significantly, the changes in FD and lacunarity indicate that the degree in which these point mutations change the fractal profile is similar to that of deletion and insertion mutations. We may also see a common shift in the fractal properties for the mutated KRAS sequences; however, these quantitative shifts may not translate to a noticeable change in the DNA walk visually. It is likely that actionable point mutations in KRAS may develop changes in the system

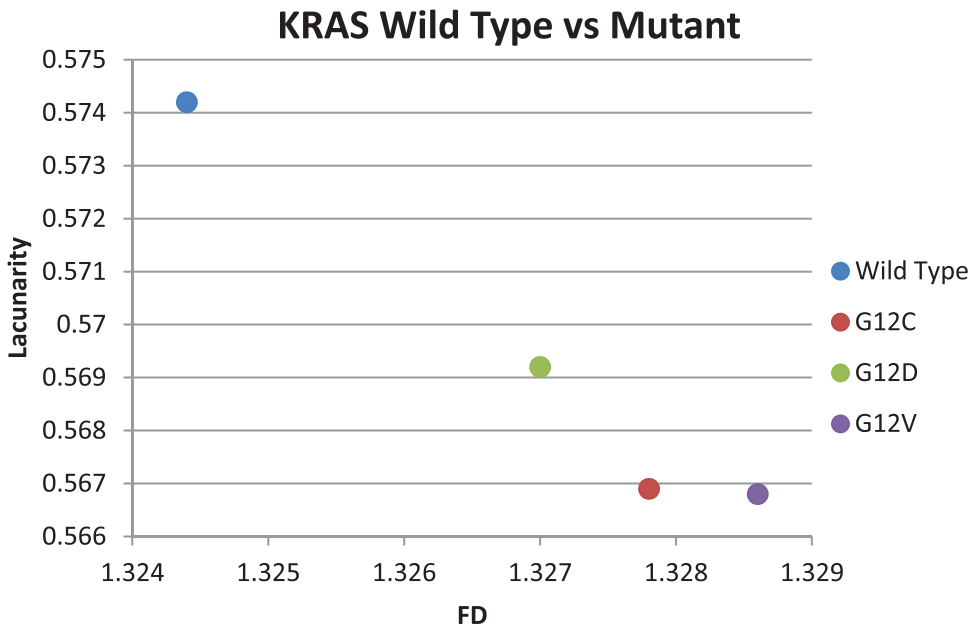

Fig. 9. Scatter diagram comparing the FD and lacunarity of wild-type KRAS with G12C, G12D and G12V codon 12 mutations

beyond the scope of this 2D analysis of DNA walks; therefore, observing protein structure to determine changes in folding or binding site may be a more viable option to analyze a point mutation's demonstration of chaos. Interestingly, a biophysical study that investigated the effect of point mutations on the structure of KRAS employing UV photodisassociation mass spectrometry (Cammarata *et al.*, 2016) revealed that different downstream effects occur due to differences in the long-range conformational or dynamic effects specific to each point mutation. Similar conclusions were also found in molecular dynamics simulations (Vatansever *et al.*, 2016). Finally, it may be worth pointing out that the short region of the C-terminal region of KRAS (Residues 167–188/189, also called the hypervariable region) is intrinsically disordered in nature (Nussinov *et al.*, 2017), i.e. it lacks a rigid 3D structure and can exist as conformational ensemble; hence, affording high flexibility. Thus, despite lack of structure, intrinsic disorder in KRAS allows intramolecular interactions spanning long distances, supports hinge motions, promotes anchoring in membranes, permits segments to fulfill multiple roles, and is crucial for activation mechanisms and intensified oncogenic signaling (Nussinov *et al.*, 2017). Such dynamic conformational fluctuations may not be efficiently captured in random walks. Furthermore, some proteins, though highly ordered, appear to reside at the brink of thermodynamic stability; even subtle changes can tip them to switch from one stable fold to another and there by acquire new functional capabilities (Bryan and Orban, 2010; He *et al.*, 2012; Kulkarni *et al.*, 2018). DNA walks may not be well suited to capture these propensities either. Thus, integrating the aspects of chaos from a change in nucleotide sequence with the disorder observed in protein conformations can help outline various aspects of noise in biological regulation.

Regardless of whether a mutated genetic walk favors order or disorder, even a wild-type case is quite likely to exhibit sensitive dependence to initial conditions. Multiple isoforms of a given gene exist to regulate cellular function, sometimes even in a diametrically opposite way (Preca *et al.*, 2015). Thus, owing to inherent cellular stochasticity, every cell is likely to have different levels of isoforms of a given gene. Similarly, post-translational modifications may alter the dynamics of a regulatory biochemical network. Hence, even a wild-type gene cannot be expected to be converted into the same DNA walk for every cell, just as a mutant gene with multiple isoforms like EGFR exon 20 insertions or EML4-ALK fusions should not be considered equal. A single nucleotide change in the sequence may potentially result in an arbitrarily different result based on its fractal properties. An intron-containing DNA sequence has been described to have a long-range dependence where the rate of decay for the correlation of a dataset can indicated by the self-similarity presented in the form of a DNA walk; however, this is not apparent when observing the fractal walks of complementary DNA sequences of a given gene (Peng *et al.*, 1992a,b). This is especially true for systems that demonstrate low self-similarity and FD (He, 2018).

Understanding the differences in the fractal images generated by walks along nucleotide sequences of wild-type genes will be useful for cancer genomics. We are able to quickly determine the level of base pair variation by analyzing the walk's shape and lacunarity. This information may lead to a deeper understanding of intrinsically disordered proteins as well as switch fold proteins, as we use machine learning to recognize particular patterns in DNA walks that may be associated with intrinsic disorder. Indeed, with ~50% of the human proteome estimated to encode IDP's, and up to 4% of the human proteome encode proteins that switch folds (Cammarata *et al.*, 2016; Nussinov *et al.*, 2017; Porter and Looger (2018) Vatansever *et al.*, 2016), a more rigorous analysis of these sequences using DNA walks may prove a worthwhile endeavor. With further research on the clinical significance of the coding regions for each gene, we may be able to determine how specific mutations affect pathways controlled by these genes. Ultimately, this approach is likely to provide the ability to generate a DNA walk of an oncogene with an unknown mutation and accurately predict the impact on its associated pathway, based on fractal analysis and observation of the chaos in the walk. This predictive ability could potentially be used to characterize drug responsiveness as well as potential mechanisms of resistance.

Funding

This work was supported by the National Cancer Institute of the National Institutes of Health under Grants No. [P30CA033572 and 1U54CA209978-01A1]. M.K.J. was supported by a Gulf Coast Consortia on the Computational Cancer Biology Training Program (CPRIT) [RP170593].

Conflict of Interest: none declared.

References

Bryan,P.N. and Orban,J. (2010) Proteins that switch folds. *Curr. Opin. Struc. Biol.*, **20**, 482–488.

Buldyrev,S.V. *et al.* (1993) Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family. *Biophys. J.*, **65**, 2673–2679.

Cammarata,M.B. *et al.* (2016) Impact of G12 mutations on the structure of K-ras probed by ultraviolet photodissociation mass spectrometry. *J. Am. Chem. Soc.*, **138**, 13187–13196.

Cortot,A.B. *et al.*(2017) Exon 14 deleted met receptor as a new biomarker and target in cancers. *J. Natl. Cancer Inst.*, **109**, djw262.

Costa,R.B. *et al.* (2018) Systematic review and meta-analysis of selected toxicities of approved ALK inhibitors in metastatic non-small cell lung cancer. *Oncotarget*, **9**, 22137–22146.

Cross,D.A. *et al.* (2014) AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov.*, **4**, 1046–1061.

Dogan,S. *et al.* (2012) Molecular epidemiology of EGFR and KRAS mutations in 3, 026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin. Cancer Res.*, **18**, 6169–6177.

Frampton,G.M. *et al.* (2015) Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. *Cancer*, **5**, 850–859.

Gao,J. *et al.* (2005) Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. *J. Biomed. Biotechnol.*, 139–146.

Gates,M.A. (1985) Simpler DNA sequence representations. *Nature*, **316**, 219.

Hamori,E.R. and Ruskin,J. (1983) H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.*, **258**, 1318–1327.

He,Z.. (2018) Integer-dimensional fractals of nonlinear dynamics, control mechanisms, and physical implications. *Scientific Report*, **8**, 10324.

He,Z. *et al.* (2012) Mutational tipping points for switching protein folds and functions. *Structure*, **20**, 283–291.

Herbst,R.S. *et al.* (2018) The biology and management of non-small cell lung cancer. *Nature*, **553**, 446–454.

Karachaliou,N. *et al.* (2013) KRAS mutations in lung cancer. *Clin. Lung Cancer*, **14**, 205–214.

Konduri,K. *et al.* (2016) EGFR fusions as novel therapeutic targets in lung cancer. *Cancer Discov.*, **6**, 601–611.

Kulkarni,P. *et al.* (2018) Structural metamorphism and polymorphism in proteins on the brink of thermodynamic stability. *Protein science : a publication of the Protein Society*.

Lennon,F.E. *et al.* (2015) Lung cancer-a fractal viewpoint. *Nat. Rev. Clin. Oncol.*, **12**, 664–675.

Li,T. *et al.* (2014) Large-scale screening and molecular characterization of EML4-ALK fusion variants in archival non-small-cell lung cancer tumor specimens using quantitative reverse transcription polymerase chain reaction assays. *J. Thorac. Oncol.*, **9**, 18–25.

Lobry,J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins. *Biochimie*, **78**, 323–326.

Lynch,T.J. *et al.* (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.*, **350**, 2129–2139.

Namazi,H. *et al.* (2015) Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA. *Oncotarget*, **6**, 42623–42631.

Namazi,H.K. and Kiminezhadmalaie, M. (2015) Diagnosis of lung cancer by fractal analysis of damaged. *Comput. Math. Methods Med.*, **2015**, 242695.

Nussinov,R. *et al.* (2017) Intrinsic protein disorder in oncogenic KRAS signaling, Cellular and molecular life sciences CMLS, **74**, 3245–3261.

Oestreicher,C. (2007) A history of chaos theory. *Dialogues Clin. Neurosci.*, **9**, 279–289.

Peng,C.K. *et al.* (1992a) Fractal landscape analysis of DNA walks. *Physica A*, **191**, 25–29.

Peng,C.-K. *et al.* (1992b) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.

Poptsova,M.S. *et al.* (2009) Hidden chromosome symmetry: in silico transformation reveals symmetry. *PLoS One*, **4**, e6396.

Porter,L.L. and Looger,L.L. *et al.* (2018) Extant fold-switching proteins are widespread. In: *Proceedings of the National Academy of Sciences of the United States of America*, **115**, pp. 5968–5973.

Preca,B.T. *et al.* (2015) A self-enforcing CD44s/ZEB1 feedback loop maintains EMT and stemness properties in cancer cells. *Int. J. Cancer.*, **137**, 2566–2577.

Roten,C.A. *et al.* (2002) Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res.*, **30**, 142–144.

Saini,S.D. and Dewan,L. (2016) Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *SpringerPlus*, **5**, 64.

Sattler,M.S. and Salgia,R. (2016) MET in the driver's seat: exon 14 skipping mutations as actionable targets in lung cancer. *J. Thorac. Oncol.*, **11**, 1381–1383.

Soto,A.S. and Sonnenschein,C. (2011) The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory. *BioEssays*, **33**, 332–340.

Vatansever,S. *et al.* (2016) Intrinsic K-Ras dynamics: a novel molecular dynamics data analysis method shows causality between residue pair motions. *Sci. Rep.*, **6**, 37012.

Yu,H.A. *et al.* (2014) Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing. *Ann. Oncol.*, **25**, 423–428.

Zielinski,J.S. *et al.* (2008) Time-dependent ARMA modeling of genomic sequences. *BMC Bioinformatics*, **9**(Suppl. 9), 14.