SCIENTIFIC REPORTS

OPEN

Application of Chaotic Laws to Improve Haplotype Assembly Using Chaos Game Representation

Received: 4 September 2018
Accepted: 1 July 2019
Published online: 17 July 2019

Mohammad Hossein Olyaee[1], Alireza Khanteymoori[1] & Khosrow Khalifeh[2]

Sequence data are deposited in the form of unphased genotypes and it is not possible to directly identify the location of a particular allele on a specific parental chromosome or haplotype. This study employed nonlinear time series modeling approaches to analyze the haplotype sequences obtained from the NGS sequencing method. To evaluate the chaotic behavior of haplotypes, we analyzed their whole sequences, as well as several subsequences from distinct haplotypes, in terms of the SNP distribution on their chromosomes. This analysis utilized chaos game representation (CGR) followed by the application of two different scaling methods. It was found that chaotic behavior clearly exists in most haplotype subsequences. For testing the applicability of the proposed model, the present research determined the alleles in gap positions and positions with low coverage by using chromosome subsequences in which 10% of each subsequence's alleles are replaced by gaps. After conversion of the subsequences' CGR into the coordinate series, a Local Projection (LP) method predicted the measure of ambiguous positions in the coordinate series. It was discovered that the average reconstruction rate for all input data is more than 97%, demonstrating that applying this knowledge can effectively improve the reconstruction rate of given haplotypes.

Deposited in bioinformatics databases is a wide range of available sequence data obtained from different high throughput sequencing tools. This wealth of data, when accompanied by advances in computational methods, has revolutionized the study of genome variation under the new emerging field of systems biology.

More than 99% of human genome is identical among individuals as well as different ethnic groups. In other words, less than 1% of genetic differences are responsible for all of the observed variations among people all over the world[1].

Therefore, specifying these differences in genetic material and evaluating the distribution on the DNA sequences of different human populations may have important implications in solving various problems in biology and medicine. In line with this assumption, two leading projects, the international haplotype map (Hapmap)[2] and 1,000 genomes[3], have been pursued to characterize common patterns of human genetic variations.

Single nucleotide polymorphisms (SNPs) are the most common types of genetic variations in the human genome. SNP refers to the occurrence of different single nucleotides at specific positions in the human genome, which resulted from mutations followed by natural selection during the evolutionary time scale. The possible nucleotides define alleles for that position[4]. A SNP sequence along each chromosome is known as a haplotype. Both SNPs and haplotypes provide valuable information for assessing genetic variations in a systematic manner. Different research fields, such as disease susceptibility, drug design, and genome-wide association studies (GWASs)[5], can greatly benefit from this data.

The distribution of SNPs across genome elements has been investigated by a multitude of studies. These have illustrated that SNPs tend to be clustered across the genome elements in a deterministic manner in which the position of the each mutation is usually affected by its neighbors and the sequences of SNPs are often highly correlated with each other[6–8]. Based on this finding, several studies have proposed models to describe how SNPs clustered along the genome sequence lead to the construction of haplotypes[9–11]. In order to identify genes involved in genetic diseases, massive amounts of SNP and haplotype data were utilized by GWASs to detect highly statistically significant correlations between SNPs on the genetic materials and various numbers of phenotypes[12]. (https://ghr.nlm.nih.gov/primer/genomicresearch/gwastudies). These are essential for the prediction, diagnosis,

[1]Department of Computer Engineering, University of Zanjan, Zanjan, Iran. [2]Department of Biology, Faculty of Sciences, University of Zanjan, Zanjan, Iran. Correspondence and requests for materials should be addressed to A.K. (email: khanteymoori@znu.ac.ir)

prevention or medical therapy of diseases by contextualizing reference big data and provide the basic elements of modern personalized medicine[13–16]. Accordingly, identifying haplotypes, particularly when input fragments contain large sections of gaps without enough coverage is critical. In order to handle reading errors, low coverage, and large number of input fragment gaps, several fragment assembly algorithms have been proposed to reconstruct the haplotypes from fragments of homologous human chromosomes from a single individual[17–24]. The identification of correlation between SNPs is the key challenge of recognizing haplotype sequences. Chaos theory provides a powerful tool for discriminating between random and deterministic processes if a suitable phase space embedding can be found. Indeed, several studies have shown that the underlying information structure can be revealed by chaos theory without the reliance on the respective equations of systems dynamics[25]. Based on this assumption, Chaos theory has been increasingly applied in Life sciences for understanding the complexity of biological systems[26]. For example, many attempts have been made to explain the chaotic behavior of biological sequences[27–32]. Moreover, chaotic view point has been applied to evaluate biological signals such as electroencephalogram (EEG) signals[33–35].

Mapping protein sequences in 2D space with chaos game representation (CGR) has shown that the structural classes of proteins can be distinguished by comparing their chaotic behavior[36]. CGR is an iterative mapping algorithm which was initially developed by Jeffrey[37] for visualizing genomic sequences as chaotic systems[38]. This method can transform an input one dimensional biological sequence into an intuitive two dimensional picture[39].

This study utilized nonlinear chaotic analysis with a surrogate data test and multi-fractal analysis to determine whether haplotypes can be detected as non-random SNP sequences. Also NA12878 dataset was used in binary form containing haplotype sequences of all human chromosomes.

Since SNPs defining haplotypes are highly correlated with each other, several subsequences are extracted from each haplotype sequence and each haplotype is locally evaluated. After the CGR method transformed each subsequence to a line, the corresponding coordinate series was extracted and its chaotic behavior was evaluated by a surrogate test. For more detailed assessment, a multi-fractal spectrum of the sequences was also computed. The resulting data confirmed that haplotype sequences of representative chromosomes originate from a non-stochastic process involving the neighbor effect of its constituents.

In order to test the ability of the proposed model to accurately manipulate haplotype sequences, single individual haplotype (SIH) reconstruction as a complicated task in computational biology was taken into account and the knowledge of chaotic behavior was utilized to improve the rate of haplotype reconstruction. The main concern of SIH is the reconstruction of haplotypes from several input fragments originating from a given sequencing method. As mentioned earlier, sequencing errors and missing information (gaps) are the main challenges in dealing with this problem. Existing methods suffer from huge numbers of gaps as these lead to positions with low coverage and thus low confidence in attempts to identify alleles in such areas[4]. Here, the current work mapped each haplotype by CGR and extracted a coordinate series in the same way as previously described. The Local Projection (LP) method then locally estimated the trajectory in the neighborhood of each ambiguous point and, finally, each ambiguous point was determined by a projection to the resultant curve. The experimental results revealed that utilizing the knowledge of chaotic behavior can help improve the reconstruction rate and also play a complementary role in the existing methods.

Materials and Methods

In order to provide a comprehensive analysis of biological sequences, the current study applied a five-step rule, as proposed by Chou[40], in the following order: (a) provide a valid dataset to evaluate the hypothesis; (b) express biological sequences with appropriate mathematical notations while preserving all of their hidden information; (c) explain the proposed method exactly; (d) evaluate the final results; and (e) provide the source code of implementations.

Materials. The current work's dataset included the HapMap NA12878 Whole-Genome Sequence (WGS) sample for a European (i.e., CEU) female individual, also known as HG001, which is a well-known reference genome dataset containing haplotype sequences of all human chromosomes[41,42]. The reference haplotypes were the trio-phased variant calls from the GATK resource bundle[43]. They were produced by a fosmid-based technology from the HapMap sample NA12878 and filtered in 1,252,769 positions that were also covered by fragments of the NA12878 dataset.

Chaos game representation. CGR is a well-known algorithm which iteratively maps an input sequence into 2D space. This mapping leads to visualization of the input sequence in a picture. Furthermore, this procedure can reveal the hidden patterns of subsequences[28].

For sequences with four alphabets, such as DNA, the final picture takes on a square format. Each vertex equals one nucleotide, i.e. A, T, C, and G. The sequence is mapped in the area of the square with a unit length such that each nucleotide base is plotted as a point. The first point is plotted at the center of the square. Next, the first base is placed halfway between the center of the square and the vertex which corresponds to the first base. As seen in the following formula, the coordinate of the i^{th} base (b_i) is placed halfway between the $(i-1)^{th}$ point and the respective vertex (v_i).

$$b_i = 0.5 \times (b_{i-1} + v_i)a \tag{1}$$

The plot is known as the CGR of the input sequence.

Analysis of nonlinear time series. Suppose X_t is a scalar time series where $t = 1, 2, \ldots, N$. If this time series is observed from a deterministic phenomenon perspective, it can be projected into a low dimensional state space called phase space. If $Y_t = \{x_t, x_{t+\tau}, \ldots, x_{t+(m-1)\tau}\}$ is X_t in the phase space, then the phase space can be recon-

structed according to Takens' embedding theorem[44]. For this purpose, parameters τ and m, as the time delay and embedding dimension respectively, should be determined. Dimension m completely demonstrates the object and its topological features. There are several approaches, such as Average Mutual Information (AMI)[45] and False Nearest Neighbor (FNN)[46], which heuristically estimate phase space parameters based on the available data.

Lyapunov exponent. Sensitive dependence on initial conditions is one of the main properties of chaotic systems. For an m dimensional chaotic system, the Lyapunov exponent (λ) is a spectrum containing m real numbers which quantifies sensitivity to initial conditions. It should be noted that the sign of its largest measure is positive for chaotic systems and its quantity indicates the extent of the chaotic system's predictability.

Suppose $Y(0)$ and $Y_\epsilon(0)$ are two initial neighbor points in phase space, in which $\|Y(0) - Y_\epsilon(0)\| = \epsilon$. By the evolution of time, the points are separated and the average of this separation, equaling λ_{max}, is obtained according to the following equation:

$$\lambda_{max} = \lim_{t\to\infty}\lim_{\epsilon\to 0}\frac{1}{t}\ln\left(\frac{\|Y(t) - Y_\epsilon(t)\|}{\epsilon}\right) \tag{2}$$

Although calculating λ from experimental data is a difficult task, several methods have been proposed to determine the largest Lyapunov exponent[47–50]. In this study, the Eckmann's method[51] was chosen because it is one of the most practical approaches for determining the Lyapunov exponent from the experimental data[52]. The first step involves mapping X_t as a scalar time series to $Y_t = \{x_t, x_{t+\tau}, \ldots, x_{t+(m-1)\tau}\}$ by reconstructing the phase space. Suppose Y_j and $Y_{j+\tau_2}$ are two points in the phase space such that $aY_{j+\tau_2}$ is the evolution of Y_j which has been provided by a rule or map as below:

$$F(Y_j) = Y_{j+\tau_2} \tag{3}$$

In the above relation, τ_2 is the iteration step which can be selected independently from τ. In the next step, for each point Y_j, all of its neighbors is found. Suppose Y_j^r is the rth nearest neighbor of Y_j, calculating the Lyapunov exponent involves determining $\underline{D}\ F(Y_j)$ which maps all neighborhoods of vectors $Y_j^r - Y_j$ to $Y_{j+\tau_2}^r - Y_{j+\tau_2}$. It should be noted that $\underline{D}\ F(Y_j)$ is the $m \times m$ Jacobian matrix of F at Y_j.

The Lyapunov exponent is obtained by calculating the eigenvalues of the matrix $(\underline{D}\ F^K)'\underline{D}\ F^K$ where $\underline{D}\ F^K$ is computed as below:

$$\underline{D}\ F^K = \underline{D}\ F(K).\ \ \underline{D}\ F(K-1)\ldots\underline{D}\ F(1) \tag{4}$$

where K is an arbitrary integer of evaluation points, and $\underline{D}\ F(K) = \underline{D}\ F(Y_K)$.

Correlation dimension method. Suppose X is a chaotic time series whose attractor has been reconstructed in phase space. The correlation dimension method is one of the most fundamental approaches for studying chaotic time series, by which its measure describes the complexity of the attractor[53]. The correlation dimension can be expressed by Equation (5):

$$C(r) = \frac{2}{(N)(N-1)}\sum_{i,j=1}^{N} H\left(r - \left\|Y_i - Y_j\right\|\right) \tag{5}$$

where N is the number of m-dimensional points on the reconstructed space, Y_i is the delay vector, r is a neighborhood, and H is the Heaviside step function. $C(r)$ is computed for a range of neighborhood sizes r and a range of embedding dimensions m. The next step plots the slopes of $C(r)$ against r on a log-linear plot. For each embedding dimension, there may be a specific curve. If these curves saturate on a common plateau, their y-value is a measure of the correlation dimension. The following describes the relationship between r and $C(r)$:

$$C(r) \propto \alpha r^{D_2} \tag{6}$$

where α is a constant value and D_2 is the correlation dimension given by Equation (7):

$$D_2 = \lim_{r\to 0}\frac{\log C(r)}{\log r} \tag{7}$$

As seen in the above formula, D_2 is estimated based on the linear region, which is found between the depopulated and saturated regions. It should be noted that the depopulated region refers to the area of the plot with no pairs of points. The saturated region includes a large value of r where $C(r)$ reaches a constant value.

It should be emphasized that the correlation dimension is suitable for situations in which the chaotic behavior of a given system is known. In other words, the correlation dimension is unable to distinguish between the stochastic and deterministic processes.

Surrogate data test. Surrogate data test is a Monte Carlo-based algorithm which can detect the chaotic behavior of an existing time series. This test supposes that the given time series is random and is provided by a stochastic process. Then, an arbitrary amount of surrogate data is generated. These data are random but preserve the statistical properties of the original data. The test starts with the hypothesis that the original time series is random. Next, a method for nonlinear time series analysis is chosen, such as that of extracting a correlation dimension, and this measure is computed for the original and surrogate time series. If the results for the original time series are completely different with those for the surrogate time series, then it can be concluded that the

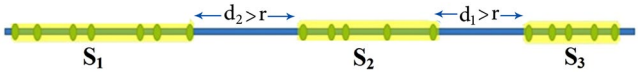

Figure 1. S_1, S_2, and S_3 are three subsequences whose SNP distances are less than predefined r.

hypothesis is not true. In other words, the original time series is related to a deterministic process. It is worth mentioning that the generated surrogate data cover most of the subset of the stochastic process class.

There are several ways to generate surrogate data, the most important of which consists of the following steps. First, Fourier transform converts the original time series to the frequency domain. Then, each element is changed by multiplication to a random phase with a unit magnitude. The resulting data are transformed back by inverse Fourier transform and, finally, randomized data with the same power, known as surrogate data, are generated.

Multi-fractal analysis. Multi-fractal refers to elements composed of several simple fractal objects. Fractal dimension cannot describe these objects' dynamic behavior. Instead, a continuous spectrum, namely the generalized fractal dimension, was developed[54]. When the attractor of a given time series is plotted in a phase space, this time series reveals chaotic behavior when the attractor is fractal or multi-fractal. Accordingly, multi-fractal analysis, as well as the surrogate data test, can help reveal the chaotic features of a given object. Up until now, several approaches have been proposed for implementing multi-fractal analysis. Fixed size box-counting is one of the most popular methods employed for solving various problems. As expressed in the following relationship, the surface of a given object is covered by several identical size ε boxes. μ is an arbitrary function which calculates the density of points (B) for each of the boxes. The partition sum of all non-empty boxes can then be calculated according to Equation (8):

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq}[\mu(B)]^q \quad q \in R \tag{8}$$

In the above relationship, q can assume any real value for discriminating the sparse from the dense regions. Equation (9) calculates the mass exponent:

$$\tau(q) = \lim_{\varepsilon\to 0}\frac{\ln Z_\varepsilon(q)}{\ln \varepsilon} \tag{9}$$

Finally, the generalized fractal dimensions are defined by the following relationships:

$$D_q = \frac{\tau(q)}{(q-1)}, \quad \text{for } q \neq 1 \tag{10}$$

$$D_q = \lim_{\varepsilon\to 0}\frac{\sum\mu(B)\ln\mu(B)}{\ln\varepsilon}, \text{ for q} = 1 \tag{11}$$

$f(\alpha)$ spectrum is used to evaluate the multi-fractal behavior of the data. Equation (12) expresses this measure:

$$f(\alpha) = q\alpha(q) - \tau(q) \tag{12}$$

Here, $\alpha(q)$ is the Lipschitz-Holder exponent which determines the singularities of a measure. This measure is related to $\tau(q)$ and is given by the following relationship:

$$\alpha(q) = \frac{d}{dq}\tau(q) \tag{13}$$

It should be noted that $f(\alpha)$ can determine the strength of multi-fractality, such that a narrower spectrum demonstrates weak multi-fractal behavior and a broader spectrum indicates stronger multi-fractality behavior.

Results and Discussions

Extracting subsequences. The analysis was performed on the full length sequences of distinctive haplotypes as well as the subsequences of haplotypes, as described below. Since the overall results of the full length analysis indicated that these sequences did not exhibit chaotic behavior, a detailed analysis of their subsequences is provided here.

As shown in Fig. 1, a number of SNPs, whose distances were less than predefined threshold r, constructed subsequence S_i. Since the extracted subsequences should have had the minimum data length required for chaos analysis, the present study assumed r equals 30,000 and selected subsequences whose lengths were greater than Thr (800) for further analysis.

By applying these cut-offs, different numbers of subsequences were extracted for each chromosome. Table 1 presents the total number of subsequences and those with lengths greater than the threshold value.

Chaos game representation of haplotype sequences. CGR is an iterative mapping algorithm which can provide a visualize form for a biological sequence. Detailed examination of the obtained picture can reveal the chaotic behavior of a system in terms of the local patterns of the sequence[38]. In order to quantitatively assess the output of CGR, a coordinate series is extracted containing all positions of the CGR picture. A typical CGR

Chromosome	#All Subsequences	#Subsequences with a Length $>Thr$
1	874	38
2	914	41
3	697	39
4	751	39
5	594	40
6	567	35
7	577	35
8	552	34
9	418	23
10	443	25
11	474	29
12	464	30
13	304	22
14	336	17
15	346	10
16	271	16
17	346	11
18	245	17
19	177	7
20	198	10
21	111	5
22	101	5

Table 1. Subsequence information extracted from all chromosomes.

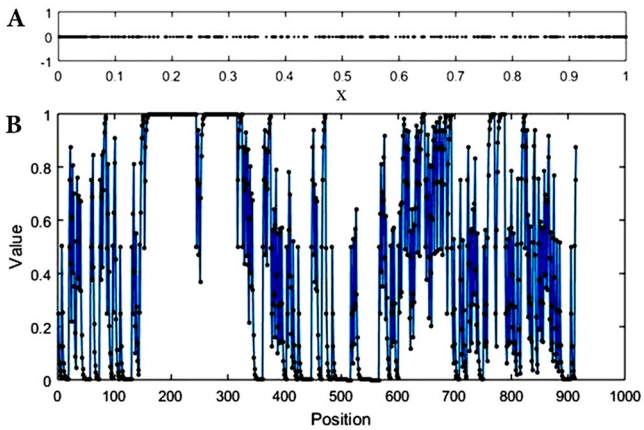

Figure 2. Binary CGR of the first extracted subsequence (**A**) and its extracted coordinate series (**B**) from the first subsequence of Chromosome 1. Since the input sequences consist of two alphabets, the obtained pictures resemble dotted lines.

picture and its corresponding coordinate series for the first subsequence of Chromosome 1's haplotype are shown in Fig. 2'sA,B, respectively. It is worth mentioning that individual values of input sequences correspond to unique points in the CGR picture and vice versa.

Since all information is preserved in the CGR plot, as well as in its respective coordinate series, the resulting data of the coordinate series assessment can be attributed to its related CGR picture.

Surrogate test. In order to examine the nonlinear properties of the original input data, the statistical surrogate data testing method was applied on individual coordinate series. The following procedure prepared the surrogate data. For each coordinate series, the embedding dimension and time delay were first determined and 10 surrogate coordinate series were then generated according to the method reviewed earlier. After this, the correlation dimension was computed for the original and its related surrogate coordinate series.

Table 2 presents the typical results of the surrogate test, including the correlation dimension values for the original coordinate series along with the minimum and maximum surrogates for all Chromosome 1 haplotype subsequences. The last column contains the values of the largest Lyapunov exponent (LLE) of each subsequence.

	Original	Surrogate Min	Surrogate Max	LLE
S_1	0.36627	0.39204	0.58594	0.33360
S_2	0.13629	0.04635	0.05898	0.13629
S_3	0.43052	0.71659	0.51951	0.34368
S_4	0.26051	0.27251	0.54544	−0.05953
S_5	0.35052	0.40979	0.57804	0.64071
S_6	0.10796	0.10934	0.16225	0.26067
S_7	0.25108	0.29168	0.38618	0.69047
S_8	0.23778	0.25782	0.47653	0.24444
S_9	0.35097	0.35113	0.62711	0.03415
S_{10}	0.27321	0.26391	0.55738	0.47606
S_{11}	0.31865	0.36589	1.00820	−0.05989
S_{12}	0.10715	0.12004	0.15777	0.83138
S_{13}	0.39225	0.45739	0.67238	0.30744
S_{14}	0.47752	0.58567	0.75620	0.29638
S_{15}	0.43977	0.49288	0.81505	0.03177
S_{16}	0.18912	0.19767	0.30535	−0.03274
S_{17}	0.53972	0.60917	0.95226	−0.10400
S_{18}	0.27841	0.28615	0.42475	0.35364
S_{19}	0.45505	0.50642	0.65790	0.33360
S_{20}	0.47605	0.55843	0.90008	−0.10132
S_{21}	0.18262	0.20794	0.39612	0.01682
S_{22}	0.11989	0.13120	0.19777	0.83163
S_{23}	0.10483	0.10412	0.13882	0.37704
S_{24}	0.54638	0.29920	0.54770	1.13682
S_{25}	0.07839	0.08140	0.12628	0.97159
S_{26}	0.09959	0.09959	0.13728	0.60269
S_{27}	0.13594	0.13708	0.25053	0.86333
S_{28}	0.15408	0.16693	0.25944	0.76313
S_{29}	0.43850	0.5150	0.83745	0.13629
S_{30}	0.13553	0.13878	0.30403	0.18735
S_{31}	0.09739	0.09993	0.18582	−0.03322
S_{32}	0.14369	0.15750	0.22236	0.53379
S_{33}	0.24935	0.28596	0.64447	0.43406
S_{34}	0.16918	0.18759	0.25632	0.60329
S_{35}	0.44667	0.47550	0.71674	0.35169
S_{36}	0.30748	0.31783	0.72264	0.13042
S_{37}	0.07777	0.07777	0.10733	−0.02983
S_{38}	0.31624	0.39668	0.93261	0.34368

Table 2. Results of the surrogate test and LLE measures for the extracted subsequences of Chromosome 1's haplotype.

According to Table 2's data, the null hypothesis of the surrogate test was rejected for most of the subsequences and the sensitivity for the initial condition was confirmed for subsequences containing positive LLE. Thus, Table 2's data demonstrates that 74% of Chromosome 1's haplotype subsequences exhibited chaotic features. The surrogate test and LLE computation were also carried out for the other chromosome haplotypes. As shown in Fig. 3, some percentages of subsequences in other haplotypes involved chaotic features. The resulting data indicate that most of the extracted subsequences exhibited chaotic behavior. Since mapping from sequences to coordinate series preserves all the information, the obtained results indicate that these subsequences originated from deterministic behavior.

Multi-fractal analysis. To confirm the above results, whose findings observe chaotic behavior in coordinate series extracted from subsequences, the present study applied multi-fractal analysis for CGRs. Multi-fractal analysis is another method for examining chaotic behavior. For investigating the scaling behavior of the CGR picture, CGRs were covered by several size ε boxes. In line with this assumption, if ε was equal to $\frac{1}{8}$, for instance, the CGR image was covered by eight boxes. Here, the input sequences had two alphabets and so the resulting CGR was a dotted line. With this method of analysis and according to the density of points in each box, the multi-fractal parameters, including D_q, α, and $f(\alpha)$, were calculated. Figure 4 provides the results of this calculation for the first subsequence of Chromosome 1's haplotype.

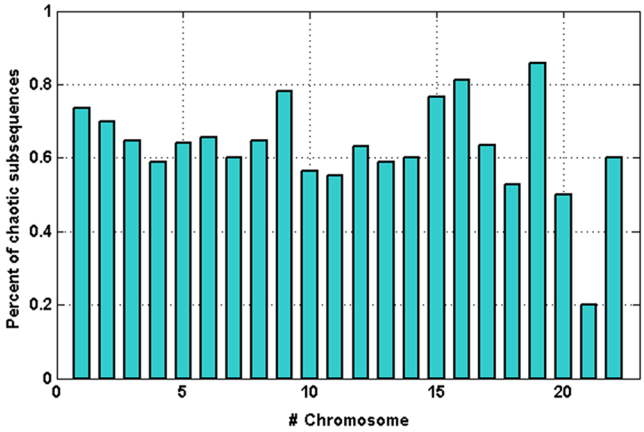

Figure 3. Percent of chaotic subsequences found throughout all chromosome haplotypes.

As presented in Fig. 4, the values of D_q, as different fractal dimensions for various q measures, as well as those of $f(\alpha)$, confirm different fractal dimensions. This reveals multi-fractality of the CGR picture. Therefore, the statistical self-similarity of the CGR should be described by a spectrum of fractal dimensions. Figure 5 shows the existence of multi-fractal property demonstrating that the subsequences originated from a deterministic process. It should be noted that multi-fractal curves reveal structural information which is hidden in the original subsequences.

Altogether, these findings indicate that the establishment of full length haplotype sequences can yield new features under the laws of stochastic processes. In addition, these results show that it is possible to extract some other features of subsequences for the purpose of evaluating their similarity and clustering state in a given haplotype by employing chaos theory assumptions. In particular, chaotic analysis reveals the deterministic nature of haplotype sequences. In some problems, such as haplotype assembly, in which the amount of noise and coverage rate of input fragments can affect the quality of reconstructed haplotypes, it is possible to rectify the achieved haplotypes via some features, such as the correlation between neighboring SNPs.

Exploiting CGR for haplotype reconstruction. To improve the reconstruction rate in the single individual haplotype (SIH) problem, one can explore how to use the chaotic feature in the extracted subsequences. In diploid organisms, such as humans, chromosomes are in pairs inherited from the father and mother respectively. As seen in Fig. 6, SIH involves several input fragments, known as *reads* (r_i), which are attained from a defined region on a pair chromosomes, based on a sequencing read technology. A set of reads can be represented by matrix $M \times N$, namely R, where each element r_{ij} belongs to {0, 1, '−'}. It is worth to noting that coverage refers to the number of reads that cover a certain position. For example, as can be seen in Fig. 6, coverage of the first column equals with 4, because it has been covered by r_1, r_5, r_7, and r_9. The haplotype assembly attempts to reconstruct haplotypes h_1 and h_2, such that these sequences are the most compatible with the input fragments. As mentioned earlier, the existing approaches show a low performance when matrix R involves columns with insufficient coverage. In these columns, there is not sufficient data to determine the measure of alleles with high confidence. Moreover, perhaps there are some positions which are not covered by any input fragments. In such cases, these positions will remain ambiguous and be represented by gaps.

The results in previous sections show the route of chaos in several extracted haplotype subsequences. These findings reveal a dependency among SNPs, which can serve as promising knowledge for determining the measure of alleles in ambiguous positions. For this purpose, it is necessary to first provide test sequences with gap positions. These sequences were obtained by corrupting the evaluated subsequences in a substitution of 10% of the individual subsequences for gaps. Next, each corrupted subsequence was mapped by CGR and its corresponding coordinate series was extracted as mentioned before. When dealing with gap measures, the algorithm was restarted and the next point was added between the center of line and the obtained allele. In the next step, the LP method estimated the allele measures in the gap positions. In deterministic chaotic flows, LP is generally applied for noise reduction. Since chaotic attractors are limited in their phase space, each divergence can be interpreted as noise. Regarding to this fact, the phase space is reconstructed. Afterwards, LP enhances the trajectories of the attractor locally. In particular, for each point, a set of its neighbors is found and a local curve which approximates them is determined. Finally, the considered point is updated by projecting to the resultant curve. Readers interested in the details of the LP algorithm may refer to an excellent book by Kantz and Schreiber[55].

In this problem, projection was only limited to the points which indicated gap positions. It should be noted that these points were initially set by the average of their neighbors' measures. Next, a set of neighbors containing $2N + 1$ points was constructed, which consisted of the considered point as well as N points before and N points after the considered point.

Figure 7 demonstrates a part of the coordinate series extracted from the first subsequence of Chromosome 1, which is fitted by the LP method. The star signs indicate positions with ambiguous measures projected to the fitted curve. To specify the allele measure of a gap position in the haplotype, the next step converts the value of the projected point to 0 or 1 by comparisons with the threshold.

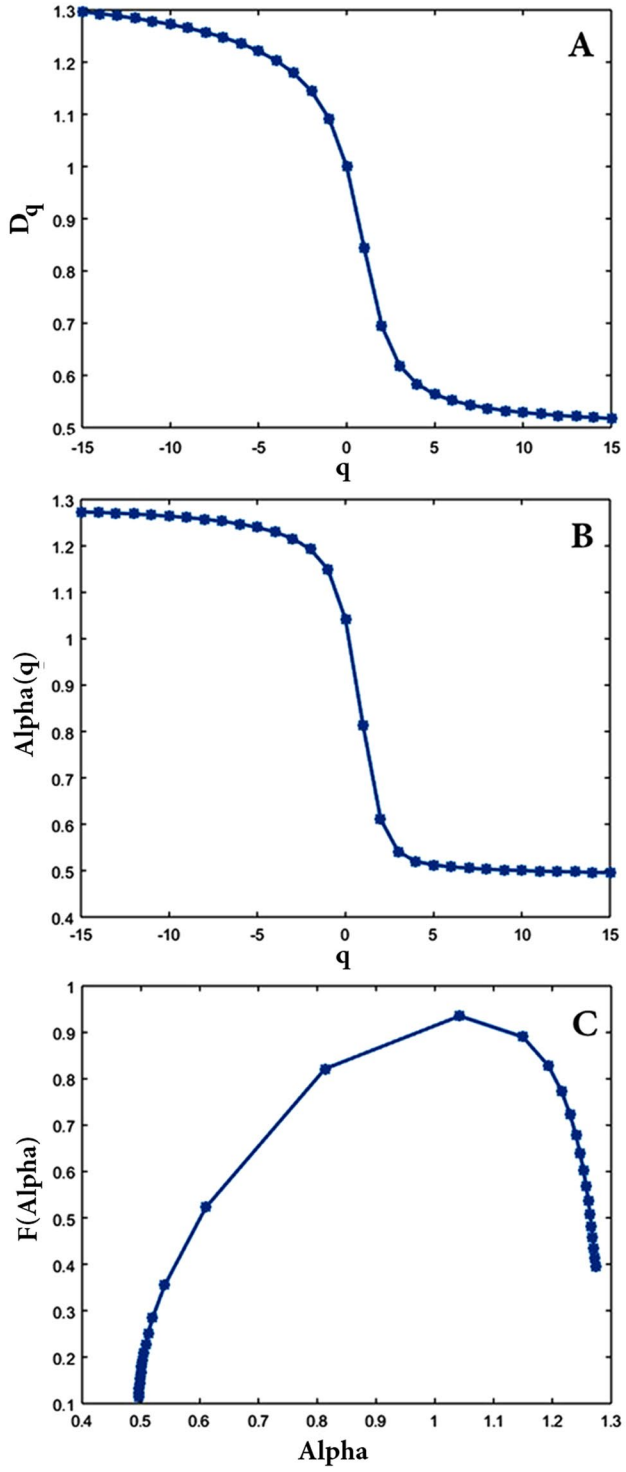

Figure 4. Three curves related to the first subsequence of Chromosome 1's haplotype. (**A**) represents D_q, (**B**) is related to spectra α, and (**C**) represents $f(\alpha)$.

$$h(i) = \begin{cases} 0 & if\ cs(i) \leq 0.5, \\ 1 & Otherwise, \end{cases} \tag{14}$$

where $cs(i)$ refers to the ith ambiguous measure projected to the fitted curve and $h(i)$ is the output for the ith position with a gap measure.

The proposed method was employed for all obtained coordinate series from all subsequences of all chromosomes. The reconstructed subsequences were compared with the original subsequences. Figure 8 depicts the

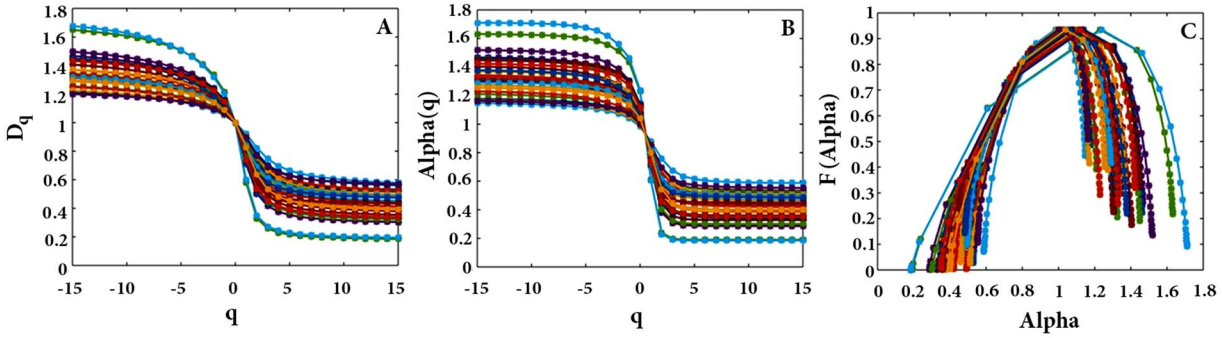

Figure 5. Multi-fractal curves of all subsequences related to Chromosome 1's haplotype (**A**) dimension spectra, (**B**) specturm α, and (**C**) $f(\alpha)$. All of the selected subsequences have positive LLE and passed the surrogate test.

$$R = \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{bmatrix} 0 & - & 1 & - & - & - & 0 & - & - & - & - \\ - & 0 & - & - & - & - & 0 & - & - & - & 1 \\ - & - & 1 & - & - & - & - & 1 & 1 & - & - \\ - & - & 1 & - & - & - & - & 1 & 1 & 0 & - \\ 1 & - & - & - & 0 & - & 1 & - & - & - & 0 \\ - & - & - & - & 0 & - & - & 1 & 1 & 0 & - \\ 1 & 0 & - & - & 0 & - & - & - & - & - & 0 \\ - & 0 & - & - & - & - & 0 & - & 1 & - & - \\ 0 & - & 1 & - & 1 & - & - & - & - & 0 & - \end{bmatrix} \end{array}$$

Haplotype assembly

$$\begin{array}{lcccccccccccc} h_1 = & 1 & 0 & 1 & - & 0 & - & 1 & 1 & 1 & 0 & 0 \\ h_2 = & 0 & 1 & 1 & - & 1 & - & 0 & 0 & 1 & 1 & 1 \end{array}$$

Figure 6. Example of an SIH problem. Matrix R contains input fragments. h_1 and h_2 are the reconstructed haplotypes. Positions 4 and 6 are ambiguous because they are not covered by any input fragments.

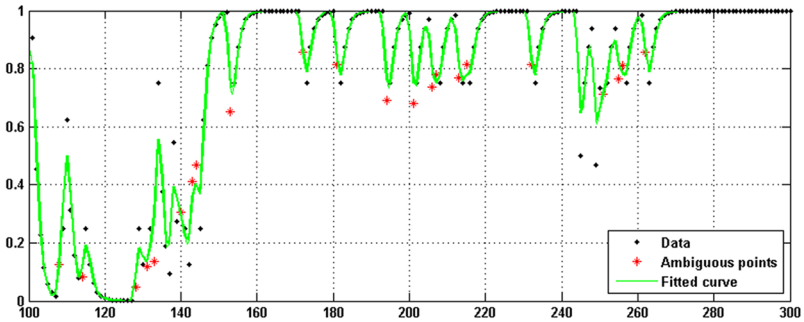

Figure 7. A part of the extracted coordinate series from Chromosome 1's subsequence (Data), the fitted curve which is the output of the proposed method (Green Curve), and the star points indicating positions with ambiguous values which are determined by a projection to the fitted curve.

percentage of improvements. As seen in Fig. 8, the boxplot demonstrates the deviation of the reconstruction rate for all subsequences belonging to a specified chromosome. The results demonstrate that the rate of reconstruction for all subsequences was more than 97% overall.

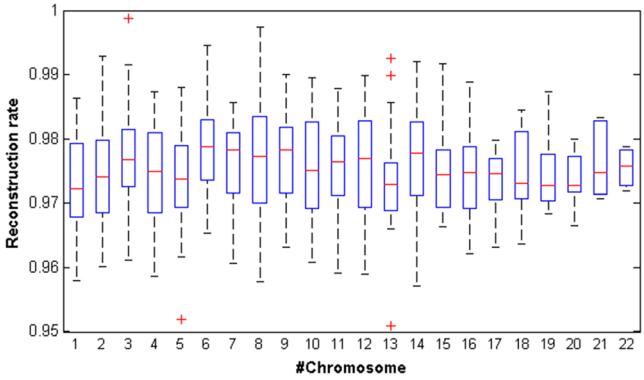

Figure 8. The deviation of reconstruction rates for all extracted subsequences from all chromosomes.

Conclusion

The current study investigated the chaotic behavior of haplotype sequences by considering the distribution of SNPs and mapping them with the CGR algorithm. The application of surrogate test and multi-fractal analysis procedure on a haplotype dataset demonstrated that the full length of chromosomes did not exhibit chaotic behavior. However, it was found that various numbers of subsequences throughout all haplotypes showed a deterministic nature. According to these findings, the laws of chaotic and stochastic processes can be employed for modeling haplotype sequences in a size-dependent manner. Moreover, the present study applied this knowledge to improve the reconstruction rate in the haplotype assembly problem. These promising results demonstrate that chaotic viewpoint can be effectively utilized to determine alleles in gap positions or low coverage positions. Finally, the source code used in the current work is available from the author upon request.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–934 (2001).
2. Gibbs, R. A. *et al.* The international HapMap project. *Nature* **426**, 789–796 (2003).
3. Consortium, G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).
4. Rhee, J.-K. *et al.* Survey of computational haplotype determination methods for single individual. *Genes & Genomics* **38**, 1–12 (2016).
5. Schaid, D. J. Evaluating associations of haplotypes with traits. *Genetic epidemiology* **27**, 348–364 (2004).
6. Ding, X. *et al.* Detecting SNP Combinations Discriminating Human Populations From HapMap Data. *IEEE transactions on nanobioscience* **14**, 220–228 (2015).
7. Koboldt, D. C., Miller, R. D. & Kwok, P. Y. Distribution of human SNPs and its effect on high-throughput genotyping. *Human mutation* **27**, 249–254 (2006).
8. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome research* **15**, 1222–1231 (2005).
9. Lee, C.-Y. A model for the clustered distribution of SNPs in the human genome. *Computational Biology and Chemistry* **64**, 94–98 (2016).
10. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
11. Amos, W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20091757 (2010).
12. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).
13. Glusman, G., Cox, H. C. & Roach, J. C. Whole-genome haplotyping approaches and genomic medicine. *Genome medicine* **6**, 73 (2014).
14. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415 (2010).
15. Kalyanasundaram, A., Gerhard, G. S. & Skelding, K. A. Genomics, haplotypes and cardiovascular disease (2007).
16. Chanock, S. J. *et al.* Replicating genotype–phenotype associations. *Nature* **447**, 655 (2007).
17. Olyaee, M.-H. & Khanteymoori, A. AROHap: An effective algorithm for single individual haplotype reconstruction based on asexual reproduction optimization. *Computational biology and chemistry* **72**, 1–10 (2018).
18. Si, H., Vikalo, H. & Vishwanath, S. Information-Theoretic Analysis of Haplotype Assembly. *IEEE Transactions on Information Theory* (2017).
19. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research* **27**, 801–812 (2017).
20. Das, S. & Vikalo, H. Optimal Haplotype Assembly via a Branch-and-Bound Algorithm. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **3**, 1–12 (2017).
21. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology* **32**, 261 (2014).
22. Aguiar, D., Wong, W. S. & Istrail, S. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. 3 (NIH Public Access).
23. Genovese, L. M., Geraci, F. & Pellegrini, M. SpeedHap: an accurate heuristic for the single individual SNP haplotyping problem with many gaps, high reading error rate and low coverage. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **5**, 492–502 (2008).

24. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
25. Garcia, S. P. & Almeida, J. S. Nearest neighbor embedding with different time delays. *Physical Review E* **71**, 037204 (2005).
26. Dokoumetzidis, A., Iliadis, A. & Macheras, P. Nonlinear dynamics and chaos theory: concepts and applications relevant to pharmacodynamics. *Pharmaceutical research* **18**, 415–426 (2001).
27. Anitas, E. M. & Slyamov, A. Structural characterization of chaos game fractals using small-angle scattering analysis. *PloS one* **12**, e0181385 (2017).
28. Almeida, J. S. Sequence analysis by iterated maps, a review. *Briefings in bioinformatics* **15**, 369–375 (2014).
29. Pandit, A., Dasanna, A. K. & Sinha, S. Multifractal analysis of HIV-1 genomes. *Molecular phylogenetics and evolution* **62**, 756–763 (2012).
30. Yang, J.-Y., Yu, Z.-G. & Anh, V. Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids. *Chaos, Solitons & Fractals* **40**, 607–620 (2009).
31. Deschavanne, P. & Tuffery, P. Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie* **90**, 615–625 (2008).
32. Joseph, J. & Sasikumar, R. Chaos game representation for comparison of whole genomes. *BMC bioinformatics* **7**, 243 (2006).
33. Güler, N. F., Übeyli, E. D. & Güler, I. Recurrent neural networks employing Lyapunov exponents for EEG signals classification. *Expert systems with applications* **29**, 506–514 (2005).
34. Jeong, J. *et al*. Nonlinear analysis of the EEG of schizophrenics with optimal embedding dimension. *Medical engineering & physics* **20**, 669–676 (1998).
35. Übeyli, E. D. Lyapunov exponents/probabilistic neural networks for analysis of EEG signals. *Expert Systems with Applications* **37**, 985–992 (2010).
36. Olyaee, M. H., Yaghoubi, A. & Yaghoobi, M. Predicting protein structural classes based on complex networks and recurrence analysis. *Journal of theoretical biology* **404**, 375–382 (2016).
37. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Research* **18**, 2163–2170 (1990).
38. Xiaohui, N., Feng, S., Xuehai, H., Jingbo, X. & Nana, L. Predicting the protein solubility by integrating chaos games representation and entropy in information theory. *Expert Systems with Applications* **41**, 1672–1679 (2014).
39. Hueso, M., Cruzado, J., Torras, J. & Navarro, E. ALUminating the path of atherosclerosis progression: chaos theory suggests a role for Alu repeats in the development of atherosclerotic vascular disease. *International journal of molecular sciences* **19**, 1734 (2018).
40. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology* **273**, 236–247 (2011).
41. Huang, J. *et al*. A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* **6**, 1–9 (2017).
42. Duitama, J. *et al*. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic acids research* **40**, 2041–2053 (2011).
43. DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491 (2011).
44. Takens, F. Detecting strange attractors in turbulence. *Lecture notes in mathematics* **898**, 366–381 (1981).
45. Fraser, A. M. & Swinney, H. L. Independent coordinates for strange attractors from mutual information. *Physical review A* **33**, 1134 (1986).
46. Kennel, M. B., Brown, R. & Abarbanel, H. D. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A* **45**, 3403 (1992).
47. Kim, B. J. & Choe, G. H. High precision numerical estimation of the largest Lyapunov exponent. *Communications in Nonlinear Science and Numerical Simulation* **15**, 1378–1384 (2010).
48. Wolf, A., Swift, J. B., Swinney, H. L. & Vastano, J. A. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena* **16**, 285–317 (1985).
49. Kantz, H. A robust method to estimate the maximal Lyapunov exponent of a time series. *Physics letters A* **185**, 77–87 (1994).
50. Rosenstein, M. T., Collins, J. J. & De Luca, C. J. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena* **65**, 117–134 (1993).
51. Eckmann, J.-P., Kamphorst, S. O., Ruelle, D. & Ciliberto, S. Liapunov exponents from time series. *Physical Review A* **34**, 4971 (1986).
52. Skokos, C. In *Dynamics of Small Solar System Bodies and Exoplanets* 63–135 (Springer, 2010).
53. Ding, M., Grebogi, C., Ott, E., Sauer, T. & Yorke, J. A. Estimating correlation dimension from a chaotic time series: when does plateau onset occur? *Physica D: Nonlinear Phenomena* **69**, 404–424 (1993).
54. Salat, H., Murcio, R. & Arcaute, E. Multifractal methodology. *Physica A: Statistical Mechanics and its Applications* (2017).
55. Kantz, H. & Schreiber, T. *Nonlinear time series analysis*. Vol. 7 (Cambridge university press, 2004).

Author Contributions

A.R.K., M.H.O. and K.K. designed the research, K.K. and M.H.O. collected data, M.H.O. and A.R.K. wrote and performed computer programs, A.R.K., M.H.O. and K.K. analyzed and interpreted the results, M.H.O. and K.K. wrote the first version of the manuscript, A.R.K. and K.K. revised and edited the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.