BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 3. 369–375
Advance Access published on 25 October 2013

doi:10.1093/bib/bbt072

Sequence analysis by iterated maps, a review

Jonas S. Almeida

Submitted: 12th July 2013; Received (in revised form): 18th September 2013

Abstract

Among alignment-free methods, Iterated Maps (IMs) are on a particular extreme: they are also scale free (order free). The use of IMs for sequence analysis is also distinct from other alignment-free methodologies in being rooted in statistical mechanics instead of computational linguistics. Both of these roots go back over two decades to the use of fractal geometry in the characterization of phase-space representations. The time series analysis origin of the field is betrayed by the title of the manuscript that started this alignment-free subdomain in 1990, 'Chaos Game Representation'. The clash between the analysis of sequences as continuous series and the better established use of Markovian approaches to discrete series was almost immediate, with a defining critique published in same journal 2 years later. The rest of that decade would go by before the scale-free nature of the IM space was uncovered. The ensuing decade saw this scalability generalized for non-genomic alphabets as well as an interest in its use for graphic representation of biological sequences. Finally, in the past couple of years, in step with the emergence of BigData and MapReduce as a new computational paradigm, there is a surprising third act in the IM story. Multiple reports have described gains in computational efficiency of multiple orders of magnitude over more conventional sequence analysis methodologies. The stage appears to be now set for a recasting of IMs with a central role in processing nextgen sequencing results.

Keywords: *sequence analysis; iterated maps; chaos game; mapreduce; big data; alignment-free*

JEFFREY 1990

Over two decades ago, during a period of great excitement over the use of statistical mechanics approaches to the nascent field of Biocomplexity and Systems Biology, it was proposed in [1] 'Chaos game representation of gene structure' to apply iterated function systems (IFS) to the investigation of DNA sequences. There, HJ Jeffrey, a computer scientist with a formal background in Mathematics, borrows from work done in the preceding 15 years in non-linear dynamics studies of fractal geometries. The approach followed by that line of research was directed to the identification of IFS that produced fractal figures with desirable properties. That function is designated as a 'mapping function', or simply, as a 'map', therefore the modern 'iterated maps' (IMs) designation. Jeffrey's Chaos Game Representation, CGR, proposes an IM for genomic sequences that places each nucleotide at the edges of a unit square, and then moves a pointer half the distance to the corresponding edge [Equation (1)]. The CGR article was part of a larger movement towards understanding 'the fractal geometry of nature' [2] that had been gaining momentum since the mid 60s [3]. This was the period when the concept of self-similar Markov processes driving the identification of iterated mapping functions took shape. One might therefore interpret the CGR approach to the analysis of nucleotide sequences as part of a broader movement towards the identification of organizing mathematical principles leading to 'self-organized criticality' [4], a concept generalized and popularized

Corresponding author. Jonas S. Almeida, Division of Informatics, Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA. Tel: 205 975 3286; Fax: 205 934 5499; E-mail: jalmeida@uab.edu

Jonas S. Almeida. After a string of academic appointments in Engineering, Biostatistics and Bioinformatics Departments, in 2011 I became the inaugural director of a Division of Informatics charged with computational groundwork for Personalized Medicine. More at jonasalmeida.info.

in [5] as a ‘order at the edge of chaos’. Coincidentally, that expression borrowed from the description, published the same year as CGR’s, of cellular automata results as ‘computation at the edge of chaos’ [6].

The Sierpinski’s triangle (Figure 1) is the paradigmatic example of a self-similar set with a fractal geometry. As described in the original article, the generation of this figure from a random set was the immediate inspiration for the CGR IM [1]. Specifically, if, in an equilateral triangle, one moves a pointer half the distance to each of the edges [the CGR game, Equation (1)] in a random order, a Sierpinski’s triangle is obtained. Using the same rule for a unit square will generate a uniformly covered distribution of points. As illustrated in Figure 1, the Sierpinski triangle and CGR share the same IM [Equation (1)], even if that may not be immediately apparent in the original IFS formalism. Although not approached in the CGR article, it is also worth noting that if the CGR game is played with only two points, using a uniformly binary sequence, a uniformly random distribution of points in a line will be generated. As discussed in the CGR report, the original motivation for these games was to assess the quality of a pseudorandom generator functions in a computer [8], a significant issue in those days. As found many years later [9], the equivalent solution for polygons with more than four edges (say, protein sequences would require a polygon with 20 edges) will in fact need to use increments bigger than half to preserve essential properties of the IM representation. Interestingly, to the best of our knowledge, Jeffrey’s milestone CGR article was his only peer-reviewed foray into the analysis of Biological Sequences, with the focus of the rest of his active research record directed to other fields. This cross-disciplinary advancement of IM applications to Biological Sequence analysis is a pattern that recurs, repeatedly, in the 23 years of work reviewed here.

THE IM FUNCTION

Originally, IMs (then designated as iterated function systems, IFS) were described as a set of linear equations, one per dimension. For example, CGR x,y coordinates would be generated by $x = ax + by + e$ and $y = cx + dy + f$. However, as is immediately apparent by inspecting CGR’s IFS, each coordinate can be determined independently (b and c terms are zero). As a consequence, a more compact notation came into use over the years and will be adopted here. Equation (1) uses the modern description of the original CGR game as a procedure where an array of map positions, y, is generated from a set of Boolean values x.

$$y_0^j = \alpha; \;\; y_i^j = y_{i-1}^j + \beta\left(x_{i-1}^j - y_{i-1}^j\right); \qquad (1)$$

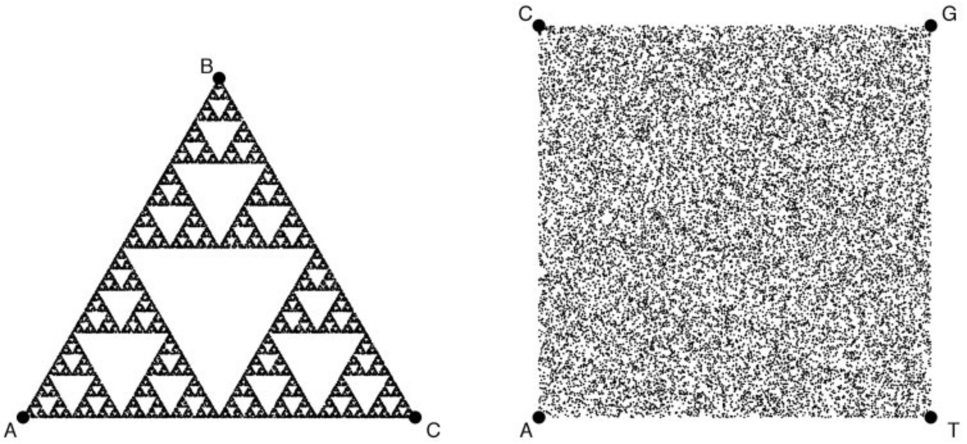

Figure 1: Sierpinski’s triangle (left) and CGR square (right), both generated by running the IM described in Equation 1 with a set of 20 000 random sets of, respectively, three symbols and four symbols - the edges (circles) of the corresponding figures. The square layout (on the right) is the one proposed in [1] to process genomic sequences, with four possible nucleotides, ACGT, one per edge of the CGR square. Any deviation from the uniform random distribution (any structure in the sequence) will produce a structure in the IM projection that is amenable to alignment-free and scale-free analysis. See Figure 1 of [7] for a graphic illustration of the process of generating CGR coordinates, applied to a gene sequence. The code used to generate this figure is available at http://bit.ly/imfig1.

In Equation (1), y contains the CGR positions, one per element of the sequence S:

$\alpha = \beta = \frac{1}{2}$, the start coordinates and the position increment factor, respectively;

$j = 1,2$, i.e. the two dimensions, m, of the CGR square;

$i = 1, \ldots, n$, n is the length of the genomic sequence S, so for each position, i,

$x_i^0 = 0$ if $S_i = A$ or C, otherwise $x_i^0 = 1$,

$x_i^1 = 0$ if $S_i = A$ or T, otherwise $x_i^1 = 1$.

which defines a unit square bound by the four nucleotides, ACGT, at positions, respectively, (0,0)(0,1)(1,1)(1,0).

THE FIRST DECADE, 1990–2000: FOUNDATION

The year the CGR article [1] was published, 1990, was also the year when BLAST, the Basic Local Alignment Search Tool [10] came out. The availability of a convenient tool to retrieve DNA sequences according to their similarity to a probe sequence was a game changer in molecular biology methodology, as reflected by close to 50 000 citations to date (Google Scholar). At a more fundamental level, the success of BLAST generated a significant drive in biological sequence analysis methods development towards exploring and advancing the probabilistic models associated with sequence alignment. This was also a key moment in the development of Bioinformatics as a discipline, to the extent that Durbin's 'Biological sequence analysis: probabilistic models of proteins and nucleic acids' [11] was for many years the main textbook in graduate courses on the topic. In other words, in the 1990–2000 period, the study of biological sequence analysis was overwhelmingly the study of the Markov Chain Models (MCM), soon extended by Hidden Markov Models to model the succession of Biological units in genomic and proteomic sequences [12]. This context helps understand why the first probabilistic study of the application of IM to biological sequences followed a narrow Markov model perspective [13], and remains to this day its most severe critique. That study, published in the same journal as the original CGR article, concluded that the distribution of positions in the CGR space could be described in terms of oligomer frequencies (word statistics) and, therefore, 'there is no justification for ascribing their patterns to anything other than' that [13]. The author reached this conclusion by consecutively dividing the CGR plane in the middle, and each time, observing that the position frequency could be ascribed to a cell in a MCM transition probability table. Intriguingly, that study ignored the possibility of a different division rule, which would have uncovered what is probably the most exciting property of IMs: that it generates truly scale-free representations of transition (with an arbitrary Markov order). That observation would have to wait for the end of the decade to be reported [7], even if by 1999 there were already signs that order-free (scale free) MCM using IMs might be a distinct possibility [14]. However, the work described in the latter report was soon directed to applications to machine learning [15, 16], and the possibility of scale-free MCM remained ignored in the field of sequence analysis for another 2 years. In summary, a review of the literature in this period suggests that Goldman's critique [13], as was its explicit intention, significantly cooled the interest in using IMs to study biological sequences any deeper than the word statistics that derive from its graphical representation.

While IMs as a sequence modeling approach were essentially on ice during the 1990's, there was a profusion of studies making good use of the graphic representation it generates. The association between oligomeric nucleotide frequencies and evolutionary processes [17] were approached by several authors because CGR representations, as in [18], simply offered a convenient procedure to produce them. That practical advantage established a bridge between IMs and what was already the foundation [19] for what would latter be designated as alignment-based methods [20]. Retracing these developments, the trail of CGR-based studies of genomic sequences starts with fundamental studies like [21], just before [13] closes that door, and continues with several genomic signature results such as [22] and others, captured by the contemporary review of novel graphical representations of DNA sequences in [23]. This trail extends to the present day, with an increased emphasis on scalability and computational efficiency as in [24–27]. This was a period of growing interest in graphical bioinformatics and a wide variety of approaches have since been proposed, as illustrated by the extent of the recent review of graphical representation of proteins [28]. Although CGR is now recognized among the milestone developments in graphical bioinformatics [29], it is interesting to note the broader context where this early work took place. Resisting a novel numerical method with fundamental significance on the basis of lack of fundamental significance of the method itself, was then a recurring pattern in the development of the mathematics of fractal geometry [30]. This early period in IM development was a time when Experimental Mathematics had not yet gained full recognition [31].

This first decade of using IMs to study biological sequences saw another two important developments. One was an interest in using it to generate entropy profiles. This was initiated by [32] in 1993, quickly followed by [33] the following year. The understanding of scalability of IM representations more than a decade later, rekindled an interest in using it for entropic profiling, as in [34]. This important topic is separately reviewed in this special issue by the author of that study. The second development was a desire to generalize the 2D graphics of the CGR approach beyond the genomic sequences and the simpler domain of a four-nucleotide alphabet. A number of variations on the IM/CGR theme, such as [35], were pursued by different authors in this period. The description of this challenge as one of using IMs to efficiently inscribe non-overlapping polygons for alphabets longer than 4 (nucleotides) was clearly laid out by [36]. His solution, and the challenge it created (because it did not produce CGR as a solution for four unit alphabets), would remain unaddressed for another 15 years.

THE SECOND DECADE, 2001–2010: DISCOVERY

The previous decade had seen key questions, and objections, being raised. This second decade not only saw them answered but in the process discovered IM's distinctive features as a mathematical tool, opening a window into a novel mathematical representation of transition. As is often the case, the process was largely driven by newcomers who, unintentionally, stumble into a new field. As many others, the author was attracted to this field by the intriguing visual representation of full genomes in reports such as [22], [37] or [38]. As Figure 1 in that report makes elegantly clear, the distribution of positions in the CGR plane provides, in a single representation, oligomeric frequencies (frequencies of nucleotides, dinucleotides, trinucleotides, . . .) with multiple lengths. For someone not aware of the argument that that is indeed all that is there to see [13], it suggests something more comprehensive than the fixed order probability tables of MCM. It actually suggests the very opposite, that those transition matrices, of all degrees, are fully accounted for in a CGR map [14], and that, indeed, there may be a lot more to discover in the use of IMs to represent symbolic sequences. All it takes for that critical leap forward is then to divide the CGR plane into a number of quadrants, *q*, where *log*$_2$*(q)* is not an integer [7]. This procedure will resolve arbitrary Markov orders, including, unsettlingly, non-integer fractal orders. The critical advancement, as illustrated in the next paragraph, is that operations usually reserved to numerical results can be applied also to symbolic sequences.

After the leap into fractal degree is taken, as explored in [7] for multiple genes of the threonine operon of *Escherichia coli*, a succession of discoveries is unleashed by the bijective mapping between the sequence's native symbolic form and its numerical representation by the IMs. For example, it suddenly becomes possible to assess how much a gene is like the genome that contains it (the similarity of *thrA* in *E. coli* is 7.7 nucleotides), or even what is its skewness (for the same gene, the average sequence differs from the median sequence by 0.13 nucleotides). Those numbers were obtained by noting the order statistics nature of the IM representation (Figure 9 in [7]), and taking the logarithm of its maximum coordinate distance, as described in Equation (2), where *j* indicates each dimension of the map, as per the notation defined for the previous equation.

$$d(y_A,y_B) = -log_2\left(max\left(\left[\left|y^{\mathrm{j}}_{\mathrm{A}} - y^{\mathrm{j}}_{\mathrm{B}}\right|\right]\right)\right) \tag{2}$$

These results, at a time when sequence alignment was the staple of bioinformatics methods, represented what can only be described as a radically alternative route for sequence analysis. Word counting and alignment methods had a shared trajectory that can, arguably, be stretched all the way to [19], the same year CGR was proposed. The shared foundation suggested that something important might have been amiss in alignment-based approaches to justify the subsequent divergence. The two approaches diverged quickly in the ensuing decade, as reviewed by other contributions to this issue. As a consequence, a newcomer to sequence analysis in this second decade of IM work would have a different perspective of the methodologies at hand. One such newcomer, starting her PhD thesis by reviewing the state of the art, saw ample justification to designate them collectively as being 'alignment-free' [20], a designation that is now widely used. Eventually, that would be interpreted, even by those at the root of alignment methods [39], as free not only from the constraints of conventional MCM but also free from the computational inefficiencies of using dynamic programming to align sequences.

A number of foundational questions were addressed in this second decade. Optimal 2D packing rules for IM that generates Sierpinski's triangle and the GCR cube as solutions were found [9], resolving a challenge dating back to 1994 [36]. The scaling of entropy measures was approached in [34]. The use of alignment-free methods, typically associated with word statistics methods as in [40], quickly found application in a myriad of applications, sometimes even as a complement to conventional alignment methods, as in [41]. It is noteworthy that, by the beginning of this second decade, oligomer frequency-based methods (not reviewed here) had advanced substantially and had produced a wealth of alignment-free sequence distance metrics on their own [42]. Similarly, their statistical foundations, as in [43], were by then well established. As concerns the IM space itself, this was, on the contrary, still a decade of discovery.

The discoveries of novel properties of the IM procedure during this period have a relatively small, and recurring, list of authors. This is intriguing since some of those studies have nevertheless been abundantly cited. For example, the discovery of scale-independent measures in [7] was followed by its generalization for any symbolic sequence in [44], where both a new IM similarity metric (i.e. scale independent, not based on oligomeric frequencies) and a statistical framework are uncovered. However, these new properties appear to find application not in sequence comparison, but in sequence classification (see for example [45]). This pull of IM applications towards machine learning bears some similarities to the same pattern observed in the use of scale free MCM, as discussed previously. The fact is that a careful analysis of IM-based applications—including some of our own [46] or even the density kernel described in [47]—is really that of the oligomeric frequencies, ultimately falling prey of Goldman's objection. The third act of IM applications to biological sequence analysis would have to wait, as is often the case, for a compelling and unmet need where existing solutions simply fall short.

THE THIRD DECADE, 2011: SCALING UP TO BIG DATA

The need for a new computational framework to handle the flood of sequence data produced by medical genomics [48, 49] was becoming clear by the end of the previous decade [50]: the need for solutions that rely on code distribution rather than data transfer. In a broader sense, not only the abundance of sequence data but the emergence of the Web as a global data space [51] set the stage for this third act of using IMs to analyze Biological sequences. Ironically, in a reversal of previous cross-pollination between disciplines, the generic abstraction that describes opportunities for distributed computing, first took root in the machine-learning community [52], before being commoditized as a cloud computing framework by Google [53]. Within 2 years, not only did it underlie the most important high-throughput sequence analysis frameworks [54, 55], but it was herald as the generic solution for computational biology in the Big Data era [56].

Ever since [46] and [24], there had been multiple reports highlighting the computational efficiency associated with the use of IMs in sequence analysis. These gains had become particularly impressive in applications where the use of alignment methods is notoriously problematic, such as in metagenomics [27]. The explanation for this computational advantage is now much better understood, as illustrated by the use of CGR as hash function within the Rabin-Karp algorithm [25]. Finally, the CGR iterated mapping procedure, as well of its generalization for any symbolic sequence, was found to be amenable to (distributable) Map Reduce decomposition [26]. The extent to which this result is ahead of the (in memory) dynamic programming of alignment methods, and the extent to which the field has advanced beyond the original CGR procedure, can be appreciated in the measure of sequence similarity in Equation (3). Its inspection will reveal that the length, L, of the shared sub-segment between two sequences can now be calculated directly, and solely, from the corresponding pair of IM coordinates. This is the antithesis of alignment as a computational application and is achieved entirely outside the MCM paradigm that entails it.

$$
\begin{aligned}
L(\gamma_A,\gamma_B) &= L\left(\overrightarrow{\gamma_A},\overrightarrow{\gamma_B}\right)+L\left(\overleftarrow{\gamma_A},\overleftarrow{\gamma_B}\right)-1 \\
L(\overline{\gamma_A},\overline{\gamma_B}) &= \left|\begin{array}{l} k=0 \\ while\left(round\left(\overline{\gamma_A}\cdot 2^k\right)==round(\overline{\gamma_B}\cdot 2^k)\right)\{k=k+1\} \\ return\,k \end{array}\right.
\end{aligned}
\tag{3}
$$

In this equation, A and B indicate the two sequence positions being compared (in the same sequence or in distinct sequences) and the arrows indicate the direction of the iteration in the calculation of γ in Equation (1). The value of L, the length of the longest shared segment, can then be determined solely from the coordinates of positions A and B, γ_A and γ_B, by incrementing an integer variable k until the rounding condition is no longer satisfied. The computational notation used in this equation is the same adopted in the original report [26]. It can be graphically described as a box counting procedure [13] in reverse: instead of halving the size of the box and counting the coordinates that fall within, the coordinates, γ_A and γ_B, are doubled until the resulting positions being compared fall outside a shared unit box. For a live resolution of this equation with arbitrary sequences see http://usm.github.io. The use of box counting procedures with IM coordinates is also a reminder of the deeper statistical mechanics roots of this procedure. The resulting sequence representation is amenable to methods that are more commonly associated with the extraction of fractal dimensions from phase-space representations [38, 57].

CONCLUSIONS

The use of IMs in sequence analysis has been object of intense study and application for the past 23 years. This field was initially driven, and objected, by the appeal of a single representation of multiple oligomeric frequencies. As the computational limitations of alignment-based methods mounted, IMs gained popularity as a convenient procedure to precompute those frequencies. However, during this process, the startling discovery was made that the IM representation is also a scale-free representation of transition. The Big Data era, and the commoditization of distributed computing via Map Reduce schemes, therefore opens a third act in the evolution of IM applications to sequence analysis. The procedure was found to be naturally decomposable into Map Reduce components, with IM-based measures of sequence similarity benefiting from the same implementation efficiencies as Big Data word counting workflows. Each of the three acts in the development of IM methods, and its applications to Biological Sequence analysis, has been defined by surprising advancements in bridging between symbolic and numeric data types. In conclusion, even if the fractal figures it produces are what attract most researchers to IM methods, there is clearly a lot more in this field than what meets the eye.

Key Points

- Iterated Maps are procedures rooted in non-linear dynamics that represent strings as fractal phase-space diagrams.
- They are unique among alignment-free methods for also being scale free, i.e. the representation can be resolved for arbitrary Markov orders including, which is nothing short of unsettling, non-integer orders.
- They are currently the most efficient procedure to determine oligomeric frequencies.
- Its scale-free representation of transition is inherently distributable and appears to be a natural fit to Big Data sequence processing operations.

FUNDING

Center for Clinical and Translational Sciences of the University of Alabama at Birmingham under contract no. 5UL1 RR025777–03 from NIH National Center for Research Resources (in part).

References

1. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res* 1990;**18**:2163–70.
2. Mandelbrot BB. *The Fractal Geometry of Nature*. New York: Macmillan, 1983.
3. Mandelbrot B. How long is the coast of britain? Statistical self-similarity and fractional dimension. *Science* 1967;**156**:636–38.
4. Bak P, Tang C, Wiesenfeld K. Self-organized criticality. *Phys Rev A* 1988;**38**:364–74.
5. Kauffman S. *The Origins of Order: Self Organization and Selection in Evolution*. New York: Oxford University Press, 1993.
6. Langton CG. Computation at the edge of chaos: phase transitions and emergent computation. *Phys D Nonlinear Phenomena* 1990;**42**:12–37.
7. Almeida JS, Carrico JA, Maretzek A, *et al*. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* 2001;**17**:429–37.
8. Park SK, Miller KW. Random number generators: good ones are hard to find. *Commun ACM* 1988;**31**:1192–201.
9. Almeida JS, Vinga S. Biological sequences as pictures: a generic two dimensional solution for iterated maps. *BMC Bioinformatics* 2009;**10**:100.
10. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
11. Durbin R. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.
12. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
13. Goldman N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res* 1993;**21**: 2487–91.
14. Tino P. Spatial representation of symbolic sequences through iterative function systems, systems, man and cybernetics, Part A: systems and humans. *IEEE Trans Neural Netw* 1999;**29**:386–93.

15. Tino P, Cernansky M, Benuskova L. Markovian architectural bias of recurrent neural networks. *IEEE Trans Neural Netw* 2004;**15**:6–15.
16. Tino P, Dorffner G. Predicting the future of discrete sequences from fractal representations of the past. *Mach Learn* 2001;**45**:187–217.
17. Schbath S, Prum B, de Turckheim E. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* 1995;**2**:417–37.
18. Hill KA, Schisler NJ, Singh SM. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J Mol Evol* 1992;**35**:261–9.
19. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 1990;**87**:2264–8.
20. Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics* 2003;**19**:513–23.
21. Dutta C, Das J. Mathematical characterization of Chaos Game Representation. New algorithms for nucleotide sequence analysis. *J Mol Biol* 1992;**228**:715–19.
22. Deschavanne PJ, Giron A, Vilain J, *et al*. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* 1999;**16**: 1391–9.
23. Roy A, Raychaudhury C, Nandy A. Novel techniques of graphical representation and analysis of DNA sequences—a review. *J. Biosci* 1998;**23**:55–71.
24. Joseph J, Sasikumar R. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics* 2006;**7**:243.
25. Vinga S, Carvalho AM, Francisco AP, *et al*. Pattern matching through Chaos game representation: bridging numerical and discrete data structures for biological sequence analysis. *Algorithms Mol Biol* 2012;**7**:10.
26. Almeida JS, Gruneberg A, Maass W, *et al*. Fractal MapReduce decomposition of sequence alignment. *Algorithms Mol Biol* 2012;**7**:12.
27. Swain MT. Fast Comparison of Microbial Genomes Using the Chaos Games Representation for Metagenomic Applications. *Procedia Comput Sci* 2013;**18**:1372–81.
28. Randic M, Zupan J, Balaban AT, *et al*. Graphical representation of proteins. *Chem Rev* 2011;**111**:790–862.
29. Randić M, Novič M, Plavšić D. Milestones in graphical bioinformatics. *Int J Quant Chem* 2013;**113**:2413–46.
30. Barnsley MF, Frame M, Howe R, *et al*. The influence of Benoît B. Mandelbrot on mathematics. *Not Am Math Soc* 2012;**59**:1208.
31. Borwein JM, Bailey DH. *Mathematics by Experiment: Plausible Reasoning in the 21st Century*. Natick, MA: AK Peters, 2004.
32. Oliver JL, Bernaola-Galvan P, Guerrero-Garcia J, *et al*. Entropic profiles of DNA sequences through chaos-game-derived images. *J Theor Biol* 1993;**160**:457–70.
33. Román-Roldán R, Bernaola-Galván P, Oliver JL. Entropic feature for sequence pattern through iterated function systems. *Pattern Recogn Lett* 1994;**15**:567–73.
34. Vinga S, Almeida JS. Local Renyi entropic profiles of DNA sequences. *BMC Bioinformatics* 2007;**8**:393.
35. Pleißner KP, Wernisch L, Oswald H, *et al*. Representation of amino acid sequences as two-dimensional point patterns. *Electrophoresis* 1997;**18**:2709–13.
36. Fiser A, Tusnady GE, Simon I. Chaos game representation of protein structures. *J Mol Graph* 1994;**12**:302–304, 295.
37. Hao Bl, Lee HC, Zhang Sy. Fractals related to long DNA sequences and complete genomes. *Chaos Solitons Fractals* 2000;**11**:825–36.
38. Hao BL. Fractals from genomes–exact solutions of a biology-inspired problem. *Phys A Stat Mech Appl* 2000;**282**:225–46.
39. Behnam E, Waterman MS, Smith AD. A geometric interpretation for local alignment-free sequence comparison. *J Comput Biol* 2013;**20**:471–85.
40. Sims GE, Jun SR, Wu GA, *et al*. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 2009;**106**:2677–82.
41. Kantorovitz MR, Robinson GE, Sinha S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* 2007;**23**:i249–55.
42. Wu TJ, Hsieh YC, Li LA. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics* 2001;**57**:441–8.
43. Pham TD, Zuegg J. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* 2004;**20**:3455–61.
44. Almeida JS, Vinga S. Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics* 2002;**3**:6.
45. Osman MH, Liong CY, Hashim I. Hybrid learning algorithm in neural network system for enzyme classification. *Int. J. Advance. Soft Comput. Appl* 2010;**2**:209–20.
46. Schwacke J, Almeida JS. Efficient Boolean implementation of universal sequence maps (bUSM). *BMC Bioinformatics* 2002;**3**:28.
47. Almeida JS, Vinga S. Computing distribution of scale independent motifs in biological sequences. *Algorithms Mol Biol* 2006;**1**:18.
48. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet* 2011;**52**:413–35.
49. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. *Mol Syst Biol* 2013;**9**: 640.
50. Bell G, Hey T, Szalay A. Computer science. Beyond the data deluge. *Science* 2009;**323**:1297–8.
51. Heath T, Bizer C. Linked data: evolving the web into a global data space. *Synthesis Lect Semantic Web Theory Technol* 2011;**1**:1–136.
52. Chu C, Kim SK, Lin YA, *et al*. Map-reduce for machine learning on multicore. *Advances in neural information processing systems* 2007;**19**:281.
53. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM* 2008;**51**:107–13.
54. McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**: 1297–303.
55. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 2009;**25**:1363–9.
56. Schadt EE, Linderman MD, Sorenson J, *et al*. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat Rev Genet* 2011;**12**:224.
57. Tiňo P. Multifractal properties of Hao's geometric representations of DNA sequences. *Phys A Stat Mech Appl* 2002; **304**:480–94.