DNA Sequences Classification and Computation Scheme Based on the Symmetry Principle

Xiao-Gang Song[1] & Wen-Yuan Qiu[1]

[1] *Department of Chemistry, State Key laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, P. R. China.*

The DNA sequences containing multifarious novel symmetrical structure frequently play crucial role in how genomes work. Here we present a new scheme for understanding the structural features and potential mathematical rules of symmetrical DNA sequences using a method containing stepwise classification and recursive computation. By defining the symmetry of DNA sequences, we classify all sequences and conclude a series of recursive equations for computing the quantity of all classes of sequences existing theoretically; moreover, the symmetries of the typical sequences at different levels are analyzed. The classification and quantitative relation demonstrate that DNA sequences have recursive and nested properties. The scheme may help us better discuss the formation and the growth mechanism of DNA sequences because it has a capability of educing the information about structure and quantity of longer sequences according to that of shorter sequences by some recursive rules. Our scheme may provide a new stepping stone to the theoretical characterization, as well as structural analysis, of DNA sequences.

Advancement of DNA sequencing techniques accelerates the increase of DNA sequences data; one important challenge is to identify the biological significations of the huge amounts of DNA sequences. Rapidly accumulating evidences have indicated that DNA variations ranging from one to millions nucleotides, which include SNPs, insertions, deletions, inversions, duplications and copy-number variants etc., might cause genetic diversity and diseases[1, 2]. To explore the complex relationships of the structure-to-function in essentials, it is necessary to make some attempts and endeavours from mathematical and physical points of view. Currently, the area of DNA mathematical analysis mainly relates to some methods based on statistics, including the distribution and the correlation of nucleic acids[3, 4], complexity[5, 6], and some studies[7-10] grounded on information theory. The correlative physical studies mainly include the characterization of structural and mechanical properties, the interaction mechanism of biological macromolecule, and dynamics modeling and simulation etc.[11-16].

Simplicity and symmetry play central roles as guiding principles in nature. However, different symmetry breakings make our world show itself complex external phenomena. If the essential principles can be deeply understood and utilized in some ways, then the current patterns may be complex but derivable from relatively simple generative principles[17]. Moreover, the discovery about symmetrical elements in life is remarkable[18, 19], and then the analysis and classification of DNA sequences based on symmetry will greatly facilitate to understand the significance of DNA symmetry. In genomes, DNA mirror image, palindrome and direct repeat sequences have special functions and novel symmetries[20-23]; they are widespread and cause serious concern to us. The researches about the DNA symmetry involve the symmetries of molecular structures, bases, codons, genes and even genome on scale, as well as physical symmetry and mathematical symmetry and so forth on content[24-31].

It is clear that structural variation and symmetry are both significant to the functions of DNA sequences; but the complex relationship between such physical properties and biological functions is uncertain yet. For example, sequence structural variations lead to symmetry breaking in a certain extent; however, some breaking can result in functional diversification but others have not effect on the original function, also known as DNA polymorphisms. Moreover, some symmetrical structures have influence on the evolution of sequences[21, 22]. In the consideration of structural variation and symmetry, we stepwise classify all sequences invoking the general principles of symmetry breaking, and have derived relevant equations to compute quantity of sequences existing theoretically. The equations are defined recursively like the Fibonacci numbers; this means that if giving the initial values and recursive rules, the quantity of sequences of any length can be calculated fleetly. Another contribution in this work is the analysis of symmetry about several typical sequences at different levels. The diversified symmetries and symmetry breakings at various levels demonstrate the important physical properties of DNA sequences, and it may contribute to studying the symmetry breaking mechanism consequently understanding sequences evolution[21]. The study may be a new method for the study of the symmetry origin and the growth mechanism from part to whole of DNA symmetrical sequences, it also offer a new thinking for the theoretical characterization and structural analysis of DNA sequences.

Classification and quantity

The multifarious repeat sequences containing novel structures account for a large portion of genomes. Many researches proved that repetitive sequences always play some pivotal roles in formation, organization and regulation of genes, and are important information of biological evolution[32-34.] We attempt to understand the significance of novel symmetrical structures; hence, we classify DNA sequences step by step from the viewpoint of structures.

The principle of classification: for the sequences consisting of even (*2n*) bases, suppose that they are formed by right and left arms of equal length, we classify them according to the combinations of every two bases that locate at symmetrical sites of two arms. The combinations of two bases include 16 different cases (Fig. 1). If each base on left arm is the same as the base on the symmetrical site of right arm, e. g. sequence “AGTCC CCTGA”, then we denote the class of sequences as mirror image sequences and mark them by M, as illustrated by edge 1, 2, 3, 4 in Fig. 1. Obviously, M sequences possess mirror image symmetry, namely common bilateral symmetry.

The sequences merely illustrated by edge 5-8 are defined as the first kind of mirror conjugated sequences, and marked by M_I, e. g. sequence “AGTCC GGACT”, also called DNA palindrome. M_I sequences have not perfect bilateral symmetry, namely it is a type of symmetry breaking. However, it can be regarded as a special symmetry in which a imaginative “distorting mirror” making “A” and “T” be mirror image each other, similarly for “C” and “G”, positioned at the centre of M_I sequence instead of the “plane mirror” in M sequences. The sequences illustrated by edge 9-12 and edge 13-16 are defined as the second and the third kind of mirror conjugated sequences respectively, marked by M_{II} and M_{III}. In this way, edges1-4, 5-8, 9-12 and 13-16 illustrate four classes of different symmetries, respectively, and define four classes of symmetrical sequences.

If a sequence contains odd *(2n+1*) bases, we regard the (*n+1*)*th* base as the centre, then there are still two arms of equal length *n*. In this case, one can also classify the sequences of odd length using the method dealing with the sequences of even length. As asymmetrical bases or segments frequently occur at the centre of real symmetrical sequences[21, 22] namely central spacers, our method handling the sequences of odd length is not arbitrary. For simplicity in this study, we need to consider all sequences of even length only.

Additionally, we name the sequences possessing certain symmetry at some sites and other symmetries at the rest of sites as asymmetry sequences. For example, sequence “ACTGG GGTTA” is an asymmetry sequence because the second base “C” and the last second base “T” are not identical but the rest bases possess mirror image symmetry. In reality, the majority of symmetrical sequences in genomes often contain a varying number of bases mutations that break the whole perfect symmetry, and sometimes influence whole functions[21, 35]. As known that the mutations are inevitable in evolution, the novel structures of symmetrical sequences play vital role in maintaining the complete biological functions of sequences[21].

Based on the above principle of classification, we consider all cases concerning combination of different symmetries and have derived the formulas to calculate the quantities of all classes of asymmetry sequences. Theoretically, the total of sequences consisting of *2n* bases equals to 4^{2n}, of which, M, M$_{\mathrm{I}}$, M$_{\mathrm{II}}$ and M$_{\mathrm{III}}$ sequences all equal to 4^n, because n bases on one arm produce 4^n different combinations and one arm is automatically determined if another arm is given under a certain symmetry. The remainder is the asymmetry sequences (Fig. 2). The amounts of the 11 classes of asymmetry sequences can be calculated by the following equations:

$$Amount\_1 = 4^n \sum_{i=1}^{n-1} C_n^i \tag{1}$$

$$Amount\_2 = 4^n \sum_{j=1}^{n-2} \sum_{i=1}^{n-j-1} C_n^j C_{n-j}^i \tag{2}$$

$$Amount\_3 = 4^n \sum_{k=1}^{(n-3)} \sum_{j=1}^{(n-k-2)} \sum_{i=1}^{(n-j-k-1)} C_n^k C_{n-k}^j C_{n-k-j}^i \tag{3}$$

where *n* denotes the single arm length, and $C_n^i = \frac{i!(n-i)!}{n!}$.

Symmetrical sequences

In the section, we study the structure and quantity of several symmetrical sequences at the different levels. Take the mirror image sequences M and the first kind of mirror conjugated sequence M$_{\mathrm{I}}$ as examples, we subdivide them, discuss their quantitative formulas and further generalize the result to another two classes of mirror conjugated sequences. Furthermore, the direct repeat sequence, which has not bilateral symmetry, is also studied similarly. In view of the similarity of structure, we discuss M and M$_{\mathrm{I}}$ sequences together in the following.

Definition 1. M sequence

Suppose sequence $S = s_1 s_2 .. s_{2n}$ of length *2n*, $s_i \in \{A, C, G, T\}$ and $1 \le i \le 2n$, it is named *M sequence* if $s_i = s_{2n-i+1}$.

Definition 2. M_I sequence

Suppose sequence $S = s_1 s_2 .. s_{2n}$ of length *2n*, $s_i \in \{A, C, G, T\}$ and $1 \le i \le 2n$, it is named M_I sequence if s_i and $s_{2n-i+1} (1 \le i \le 2n)$ are complementary.

Here the complementary relation is the normal Watson-Crick base pairing rules; namely “A” and “T” are complementary, similarly for “C” and “G”. According to the *definition 2,* sequence “CGCGAATTCGCG”, the so-called Dickerson-Drew dodecamer, is a M_I sequence that structure continues to be the subject of intense experimental and theoretical study during the past over 20 years [25, 36]. Obviously, in the above M_I sequence, the twelve bases can be segmented as three parts “CGCG”, “AATT” and “CGCG”, which are all M_I sequences. According to our investigation, the type of sequences, each segment satisfying the symmetry that the whole sequence satisfies, is considerable in M and M_I sequences. For this reason, further, we classify M and M_I sequences into M combinator and M generator, M_I combinator and M_I generator, respectively, to study their subtle structure.

Definition 3. M and M_I combinator and generator

Suppose M (M_I) sequence *S* is a combinator if and only if it can be segmented as $S = h_1 h_2 .. h_k$, $k \ge 2$, where h_i $(1 \le i \le k)$ is also M (M_I) sequence. Otherwise, *S* is a generator.

(a) GCCG CAAC | CAAC GCCG — M combinator

(b) GC GC CATG | CATG GC GC — M_I combinator

(c) ACTGC | CGTCA — M generator

(d) ACTGC | GCAGT — M_I generator

Sequence (a) is a M combinator which could be considered as a structure constructed by four linked shorter M sequences; and (b) is a M_I combinator, consists of 6 linked shorter M_I sequences. Sequences (c) and (d) are generator, they are inseparable. It is worth noting that some combinators have multiple segmentation methods. For example, M_I combinator (b) also can be constructed by linking four palindromes “GCGC”, “CATG”, “CATG”, and “GCGC” end to end, instead of the above 6 pieces, it means that the segmentation is not unique. However, we can avoid the problem of multiple segmentations by taking generators as units. In this way, the segmentation of (b) can only be 6 pieces.

Quantitatively, we calculate the amounts of M and M_I generators and combinators using computer, and the part result (*n≤18*) was listed in Table 1. By analyzing the numbers, recursive relation among these numbers is deduced as:

$$M_G(n)=\begin{cases}4 & n=1\\ 4M_G(n-1)-M_G(n/2) & n\ is\ even\\ 4M_G(n-1) & otherwise\end{cases} \qquad (4)$$

$$M_C(n)=\begin{cases}0 & n=1\\ 4M_C(n-1)+M_G(n/2) & n\ is\ even\\ 4M_C(n-1) & otherwise\end{cases} \qquad (5)$$

Given the initial values of equations when n=1, we can calculate the amounts of sequences of any length recursively. Take $M_G(n)$ as an example: if *n*=1, then $M_G(1)=4$; if *n*=2, $M_G(2)=4M_G(1)-M_G(1)=12$; if *n*=3, $M_G(3)=4M_G(2)=48$. By analogy, the amount of long sequences can be calculated recursively using the amount of shorter sequences that can be derived using that of more short sequences. For M_I sequence, because $M_{I\_G}(n)$ and $M_{I\_C}(n)$ have the same quantitative relation with $M_G(n)$ and $M_C(n)$, respectively, they can also use equation (4) and (5). Another two kinds of mirror conjugated sequences M_{II} and M_{III} both can be classified by the same method applied in M_I, and satisfy the same quantitative equations. The recursive relation is a main contribution of the study. It is effective in calculating the quantity of sequences, also is revelatory in studying DNA growth mechanism because the recursive procedure may well be similar to the growth procedure of DNA sequences.

Moreover, the recursive computational procedure means that among the numbers investigated, there exists an order that can be formulated, and containing the analogical properties that implies a type of symmetry by reason of the predictability of the recursive formulas[37].

In addition, an interesting observation is that the amount of generators is always more than that of combinators for given n, and the ratio of them rapidly tends to 2.20 with the increase of *n*. If we interestingly compare generator and combinator to “egg” and” chicken” respectively, we may could say that “eggs” are more than “chickens” and there were “eggs” before “chickens”.

Further, besides the recursive relation in quantity, we hope to explore the recursive relation in structure. Therefore, combinators were classified in considerable detail in the following. Considering that every combinator consists of several generators, we can classify combinators according to the numbers and lengths of the generators within them. If arm length *n*=1, combinator does not exist. If *n*=2, only four combinators constructed by two shortest generators of length 2 exist, denoting the type of combinators as“2+2”. For instance, the four sequences of M_I combinators are “ATAT”, “TATA”, “CGCG” and “GCGC”. If *n*=3, such combinators consists of three generators of length 2, denoting the type as “2+2+2”. If *n*=4, however, three different types present, including “2+2+2+2”, “2+4+2”, and “4+4”. All types of combinators of arm lengths ranging from 2 to 8 are listed in Table 2.

To characterize the types and quantity of combinators in detail, taking M combinator for instance, we denote it as $M_C(h_1,h_2,\ldots,h_j)$, $(j \geq 2)$. That means that one combinator consists of *j* generators, and the length of the *i*th generator is h_i, $(1 \leq i \leq j)$. The amount $|M_C(h_1,h_2,\ldots,h_j)|$ can be calculated as:

$$|M_C(h_1,h_2,\ldots,h_j)| = \begin{cases} M_G(h_1/2)\times M_G(h_2/2)\times\ldots\ldots\times M_G(h_{j/2}/2) & j\ is\ even \\ M_G(h_1/2)\times M_G(h_2/2)\times\ldots\ldots\times M_G(h_{j+1/2}/2) & j\ is\ odd \end{cases} \quad (6)$$

where, $M_G(h_i/2)$, the recursive equations of M generators (equation (4)), represents the amount of M generators of arm length $h_i/2$. For example, $M_C(2,2) = M_G(1) = 4$, $M_C(2,4,2) = M_G(1) \times M_G(2) = 4 \times 12 = 48$. According to the equation (6), the amount of various M combinators with different structure can be computed using the equations (4). The three types of mirror conjugated sequences M_I, M_{II} and M_{III} combinators have the same structural and quantitative properties as M combinators.

Through elaborating these specialties, we observe that generator is significant not only in quantity because its amounts are precondition of calculating the amounts of combinators, also in structure because they construct combinators completely.

Direct repeat sequence is another class of special sequences with novel symmetry, and many researches proposed that direct repeat sequence plays an important role in the control of gene expression and chromatin organization[32-34]. In the classification as described above, direct repeat sequences has not been classified due to its translational symmetry rather than the bilateral symmetry like M, M_I, M_{II} and M_{III} sequences, then we discuss its structural and quantitative properties solely.

Definition 4. Direct repeat sequence

Suppose sequence $S = s_1 s_2 .. s_n$ of length *n*, $s_i \in \{A, C, G, T\}$, $1 \le i \le n$, it is named direct repeat sequence if integer $t \ge 1$ exists such that for each $1 \le i \le n - t$ we have $s_{i+t} = s_i$.

In the definition, *t* represents the periodicity, which is the most important feature of direct repeat sequences. For example, “ACACACAC” is a direct repeat sequence, and its repetitive unit “AC” of length 2 repeats four times, then the periodicity *t*=2. We can classify direct repeat sequences of given length according to the length of repetitive unit. In the example, however, “ACAC” can be regarded as a repetitive unit that repeats two times and *t*=4. In other words, periodicity *t* is not unique sometimes for the same sequence. To ensure consistency of classification, our classification of direct repeat sequences is based on the length of the shortest repetitive unit, then the periodicity *t* equal to 2 in the above example. Consider the case of *n*=12, proper divisors of 12 are as follows: 1, 2, 3, 4 and 6, then the types of direct repeat sequences consist of $(a_1)_{12}$, $(a_1a_2)_6$, $(a_1a_2a_3)_4$, $(a_1a_2a_3a_4)_3$ and $(a_1a_2a_3a_4a_5a_6)_2$. They mean that one base repeats 12 times, *t*=1; 2 bases as a unit repeats 6 times, *t*=2; similarly, 3, 4 and 6 bases as a unit repeats 4, 3 and 2 times, respectively, *t*=3, 4 and 6. Obviously, the types of direct repeat sequences depend on the divisors of length *n*. We denote the amount of repetitive unit of length *m* as *f (m)*, it is defined recursively as follows:

$$f(m)=\begin{cases}4 & m=1\\ 4^m-\sum_i f(i) & m>1\end{cases} \quad (7)$$

where *i* indicates the proper divisor of *m*. Then the amount of direct repeat sequences of full length *n*, *R (n)*, can be calculated by accumulating its *f* (*m*):

$$R(n)=\sum_m f(m) \quad (8)$$

where *m* indicates the proper divisor of *n*.

Therefore, if *n* is a prime number, direct repeat sequences merely include four kinds of sequences: “AA…A”, “TT…T”, “CC…C” and “GG…G”. Because the proper divisor of *n* is only one, such that $R(n)= f(1)=4$. If *n* is a composite number, the types and amounts of direct repeat sequences lies on the amounts and size of the proper divisors of *n*.

Symmetry of sequences

In consideration of the principle of classification is symmetry, here we discuss the symmetries of several typical sequences, such as mirror image sequences M, first kind of mirror conjugated sequences M_I and direct repeat sequences, at the levels of bases, generators/repetitive units and molecules.

M sequences possess mirror image symmetry at the level of bases array, that is to say every two bases that locate on symmetrical sites of two arms are the same. For M combinators, it possess not only mirror image symmetry at the level of bases also both mirror image symmetry and translational symmetry at the level of generators (Fig. 3a and b). It is understandable that the symmetry of M sequences at the level of bases is determined by its definition; moreover, the symmetries of M combinators at the level of generators depends on its structural features that each generator is also read exactly the same on both directions.

M_I sequences possess the first kind of mirror conjugated symmetry (A-T, C-G) at the level of bases, that is to say every two bases that locate on symmetrical sites of two arms are complementary. M_I combinators possess only translational symmetry at the level of generators (Fig. 3c and d), and the point is different from M combinators. It is easy, similarly, to understand M_{II} and M_{III} sequences are similar to M_I sequences in structure. Therefore, M_{II} and M_{III} sequences have also translational symmetry at the level of generators besides its relevant mirror conjugated symmetry at the level of bases.

For direct repeat sequences, bases always repeatedly appear with certain periodicity that is the length of repetitive unit. Thus, direct repeat sequences have translational symmetry at the level of bases. Obviously, they possess translational symmetry at the level of repetitive units also (Fig. 3e).

It is well known that the molecular backbone of DNA strand is constructed from alternating sugar and phosphate molecule; it means the backbone has translational symmetry. If we consider bases together with backbone, all above translational symmetry still retain; however, mirror image and mirror conjugated symmetries lose.

Additionally, we have observed the existence of overlapping among five classes of symmetrical sequences (M, M_I, M_{II}, M_{III}, and direct repeat sequences); the relationship among them is shown in Fig. 4. For instance, sequence “ACCA ACCA ACCA ACCA” is not only a direct repeat sequence also a mirror image sequence. In details, the direct repeat sequence as a whole possesses mirror image symmetry if its repetitive unit possesses mirror image symmetry, and we denote the type of sequences as DM, and the amount |DM| can be calculated as:

$$|DM| = \sum_i M_G(n[i]/2) \quad (9)$$

where $M_G(n[i]/2)$, the recursive equations of M generators (equation (4)), represents the amount of M generators of arm length $n[i]/2$; and $n[i]$ represents the array composed of the even proper divisors of *n*. Similarly, the direct repeat sequences also possess mirror conjugated symmetry if repetitive units possess certain mirror conjugated symmetry, and its amount $|DM_{I}|$, $|DM_{II}|$ and $|DM_{III}|$ both equal to |DM|.

Conclusions

Our study should be a new step towards the mathematical and physical cognition of DNA sequences. Nevertheless, the study is the first screen for the understanding of how symmetry plays key roles in DNA classification and theoretical computation and, as such, it serves as a guild for the future exploration of DNA growth mechanism and symmetry breaking principles. In the study, three main contributions are as follows.
(1) We classify all sequences in detail based on symmetry, the stepwise refined classification contribute to the understanding of DNA sequences structure from small scale to large scale. (2) Moreover, we perform theoretically computation about sequences quantity, and have found a series of recursive equations that are often very useful when confronting a complex computation. Using our classification principle and recursive equations, one can fleetly and effectively calculate the quantity and structure of DNA sequences of any length existing theoretically. Classifying mirror image sequences M and mirror conjugated sequences M_{I}, M_{II} and M_{III} into generators and combinators, an observation is that the ratio of generators to combinators tends to 2.20 with the increase of length *n*. From the standpoint of structure, they can be

compared interestingly as “egg” and “chicken” respectively, and “eggs” are more than “chickens” and there were “eggs” before “chickens”. (3) The results about symmetry study show that M sequences possess translational, mirror image and both translational and mirror image symmetries at the level of molecules, bases and generators, respectively. M_I, M_{II} and M_{III} are asymmetrical at the level of molecules; and possess the symmetries corresponding to their definitions at the level of bases; however, at the level of generators, possess translational symmetry. Direct repeat sequences possess translational symmetry at all three levels, molecules, bases and repetitive units.

In reality, the majority of DNA symmetrical sequences do not satisfy a certain perfect symmetry; namely they are asymmetry or symmetry breaking in a certain extent. The phenomenon is consistent with that the symmetry breakings frequently present in symmetrical things whatever they are natural or artificial. Therefore, that exploring the mechanism of symmetry breakings in DNA sequences presents new challenges and opportunities to researchers[18, 21, 24]. So far, although our work have not be directly contacted to a certain biological signification, it is hopeful to provide a new insight into DNA analysis since it is well known that nature's rhythms are often linked to symmetry and simplicity[17-19, 37].

Reference

1. Inoue, K. & Lupski, J. R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3,** 199-242(2002).

2. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14,** 417-422(1998).

3. Peng, C. K. et al. Long-Range Correlations in Nucleotide-Sequences. *Nature* **356,** 168-170 (1992).

4. Luo, L. F., Lee, W. J., Jia, L. J., Ji, F. M. & Tsai, L. Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E.* **58,** 861-871(1998).

5. Troyanskaya, O.G. et al. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics* **18,** 679-688(2002).

6. Orlov, Y. L. & Potapov, V.N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res.* **32,** W628-W633(2004).

7. Zhou, L. Q., Yu, Z. G., Deng, J. Q., Anh, V. & Long, S. C. A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation. *J. Theor. Biol.* **232,** 559-567(2005).

8. Voss, R.F. Evolution of long-range fractal correlations and 1/F Noise in DNA-Base Sequences. *Phys. Rev. Lett.* **68,** 3805-3808(1992).

9. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C. & Marcourt, L. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theor. Biol.* **206,** 323-326(2000).

10. Arneodo, A., Bacry, E., Graves, P. V. & Muzy, J. F. Characterizing long-range correlations in DNA-sequences from wavelet analysis. *Phys. Rev. Lett.* **74,** 3293-3296(1995).

11. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature* **389,** 251-260(1997).

12. Marko, J. F. DNA under high tension: Overstretching, undertwisting, and relaxation dynamics. *Phy. Rev. E.* **57,** 2134-2149(1998).

13. Smith, S.B., Cui, Y.J. & Bustamante, C. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271,** 795-799(1996).

14. Brower-Toland, B. D. et al. Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA. *Proc. Natl. Acad. Sci. USA* **99,** 1960-1965(2002).

15. Evans, E. & Ritchie, K. Dynamic strength of molecular adhesion bonds. *Biophy. J.* **72,** 1541-1555(1997).

16. Schwalb, N. K. & Temps, F. Base sequence and higher-order structure induce the complex excited-state dynamics in DNA. *Science* **322,** 243-245(2008).

17. Stewart, I. *Nature's Numbers: The Unreal Reality of Mathematics* (Basic Books, New York, 1996).

18. Qiu, W. Y. in *Chemical Topology—Application and Techniques, Mathematical Chemistry Series* (eds. Bonchev, D. & Rouvray, D. H.) 175-237 (Gordon and Breach Science Publishers, Amsterdam, 2000).

19. Froggatt, C. D. & Nielsen, H. *Origin of Symmetries* (World Scientific, Singapore, 1991).

20. Urata, H., Ogura, E., Shinohara, K., Ueda, Y. & Akagi, M. Synthesis and Properties of Mirror-Image DNA. *Nucleic Acids Res.* **20,** 3325-3332(1992).

21. Rozen, S. et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423,** 873-876(2003).

22. Lewis, S. M. & Cote, A. G. Palindromes and genomic stress fractures: Bracing and repairing the damage. *DNA Repair* **5,** 1146-1160(2006).

23. Holste, D., Grosse, I., Beirer, S., Schieg, P. & Herzel, H. Repeats and correlations in human DNA sequences. *Phys. Rev. E.* **67,** 061913(2003).

24. Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett.* **555,** 579-582(2003).

25. Johansson, E., Parkinson, G. & Neidle, S. A new crystal form for the dodecamer C-G-C-G-A-A-T-T-C-G-C-G: Symmetry effects on sequence-dependent DNA structure. *J. Mol. Biol.* **300,** 551-561(2000).

26. Gavish, M., Peled, A. & Chor, B. Genetic code symmetry and efficient design of GC-constrained coding sequences. *Bioinformatics* **23,** E57-E63(2007).

27. Zhang, C.T. A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol.* **187,** 297-306(1997).

28. Wilhelm, T. & Nikolajewa, S. A new classification scheme of the genetic code. *J. Mol. Evol.* **59,** 598-605(2004).

29. Nikolajewa, S., Friedel, M., Beyer, A. & Wilhelm, T. The new classification scheme of the genetic code, its early evolution, and tRNA usage. *J. Bioinform. Comput. Biol.* **4,** 609-620(2006).

30. Arques, D. G. & Michel, C. J. A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182,** 45-58(1996).

31. Arques, D. G., Fallot, J. P. & Michel, C. J. An evolutionary model of a complementary circular code. *J. Theor. Biol.* **185,** 241-253(1997).

32. International Human genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921(2001).

33. Johns, D. R., Rutledge, S. L., Stine, O. C. & Hurko, O. Directly repeated sequences associated with pathogenic mitochondrial DNA deletions. *Proc. Natl. Acad. Sci. USA* **86,** 8059-8062(1989).

34. Brahmachari, S. K. et al. Simple repetitive sequences in the genome-structure and functional-significance. *Electrophoresis* **16,** 1705-1714(1995).

35. Yen, P. H., Chai, N. N. & Salido, E. C. The human DAZ genes a putative male infertility factor on the Y chromosome,are highly polymorphic in the DAZ repeat regions. *Mamm. Genome.* **8,** 756-759(1997).

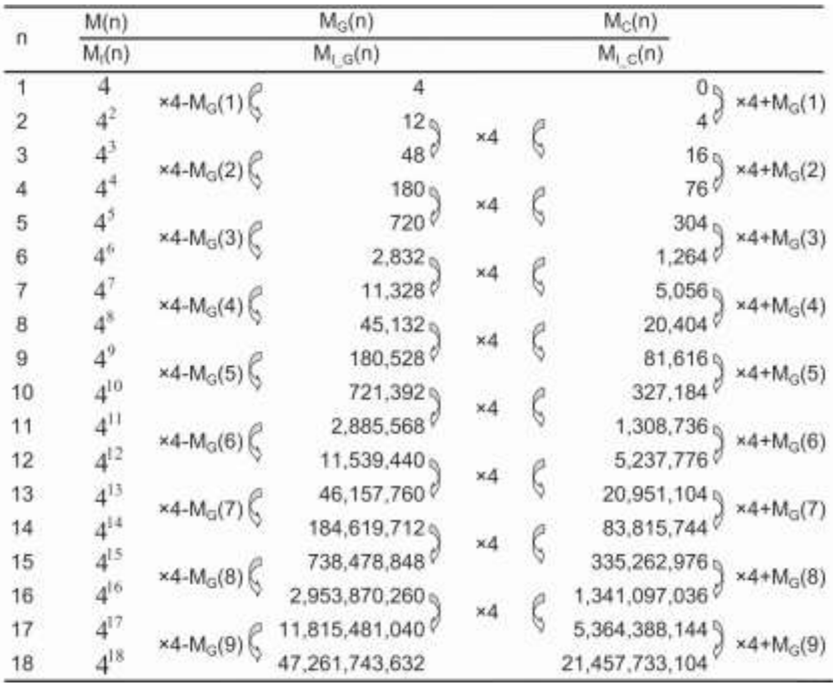
36. Drew, H. R. & Dicksson, R. E. Structure of a B-DNA dodecamer. III. Geometry of hydration, *J. Mol. Biol.* **151,** 535-556(1981).

37. Rosen, J. Symmetry Rules: *How Science and Nature Are Founded on Symmetry* (Springer-Verlag, Berlin Heidelberg, 2008).

Acknowledgements This work was supported by the National Natural Science Foundation of China.

Author Information Reprints and permissions information is available at npg.nature.com/ reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.-Y. Q. (wyqiu@lzu.edu.cn).

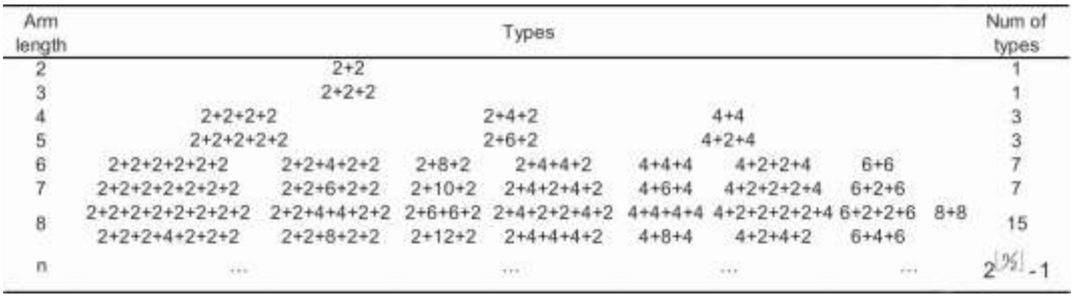
Table 1 Amounts of M and M_I generators and combinators, n≤18.



n	$M(n)$		$M_G(n)$		$M_C(n)$	
	$M_I(n)$		$M_{I\_G}(n)$		$M_{I\_C}(n)$	
1	4	×4-$M_G(1)$	4		0	×4+$M_G(1)$
2	4^2		12	×4	4	
3	4^3	×4-$M_G(2)$	48		16	×4+$M_G(2)$
4	4^4		180	×4	76	
5	4^5	×4-$M_G(3)$	720		304	×4+$M_G(3)$
6	4^6		2,832	×4	1,264	
7	4^7	×4-$M_G(4)$	11,328		5,056	×4+$M_G(4)$
8	4^8		45,132	×4	20,404	
9	4^9	×4-$M_G(5)$	180,528		81,616	×4+$M_G(5)$
10	4^{10}		721,392	×4	327,184	
11	4^{11}	×4-$M_G(6)$	2,885,568		1,308,736	×4+$M_G(6)$
12	4^{12}		11,539,440	×4	5,237,776	
13	4^{13}	×4-$M_G(7)$	46,157,760		20,951,104	×4+$M_G(7)$
14	4^{14}		184,619,712	×4	83,815,744	
15	4^{15}	×4-$M_G(8)$	738,478,848		335,262,976	×4+$M_G(8)$
16	4^{16}		2,953,870,260	×4	1,341,097,036	
17	4^{17}	×4-$M_G(9)$	11,815,481,040		5,364,388,144	×4+$M_G(9)$
18	4^{18}		47,261,743,632		21,457,733,104	

Here $M(n)$, $M_G(n)$ and $M_C(n)$ represent the amounts of M sequences, M generators and combinators, respectively. Likewise, $M_I(n)$, $M_{I\_G}(n)$ and $M_{I\_C}(n)$ represent the amounts of M_I sequences, M_I generators and combinators, respectively. Where, "n" represents the arm length of M and M_I sequences. The arrows and the side expressions illustrate the relations between the above and the below numbers. The quantitative relation of $M(n)$, $M_G(n)$ and $M_C(n)$ are the same as $M_I(n)$, $M_{I\_G}(n)$ and $M_{I\_C}(n)$, respectively.

Table 2 Classification of combinators



Arm length	Types								Num of types
2		2+2							1
3		2+2+2							1
4	2+2+2+2			2+4+2		4+4			3
5	2+2+2+2+2			2+6+2		4+2+4			3
6	2+2+2+2+2+2	2+2+4+2+2	2+8+2	2+4+4+2	4+4+4	4+2+2+4	6+6		7
7	2+2+2+2+2+2+2	2+2+6+2+2	2+10+2	2+4+2+4+2	4+6+4	4+2+2+2+4	6+2+6		7
8	2+2+2+2+2+2+2+2 2+2+2+4+2+2+2	2+2+4+4+2+2 2+2+8+2+2	2+6+6+2 2+12+2	2+4+2+2+4+2 2+4+4+4+2	4+4+4+4 4+8+4	4+2+2+2+2+4 4+2+4+2	6+2+2+6 6+4+6	8+8	15
n	...			...		...	...		$2^{\lfloor n/2 \rfloor}-1$

If the arm length of combinator equals to *n*, the number of types equals to $2^{\lfloor n/2 \rfloor}-1$, where $\lfloor n/2 \rfloor$ indicates the largest integer at most *n/2*.

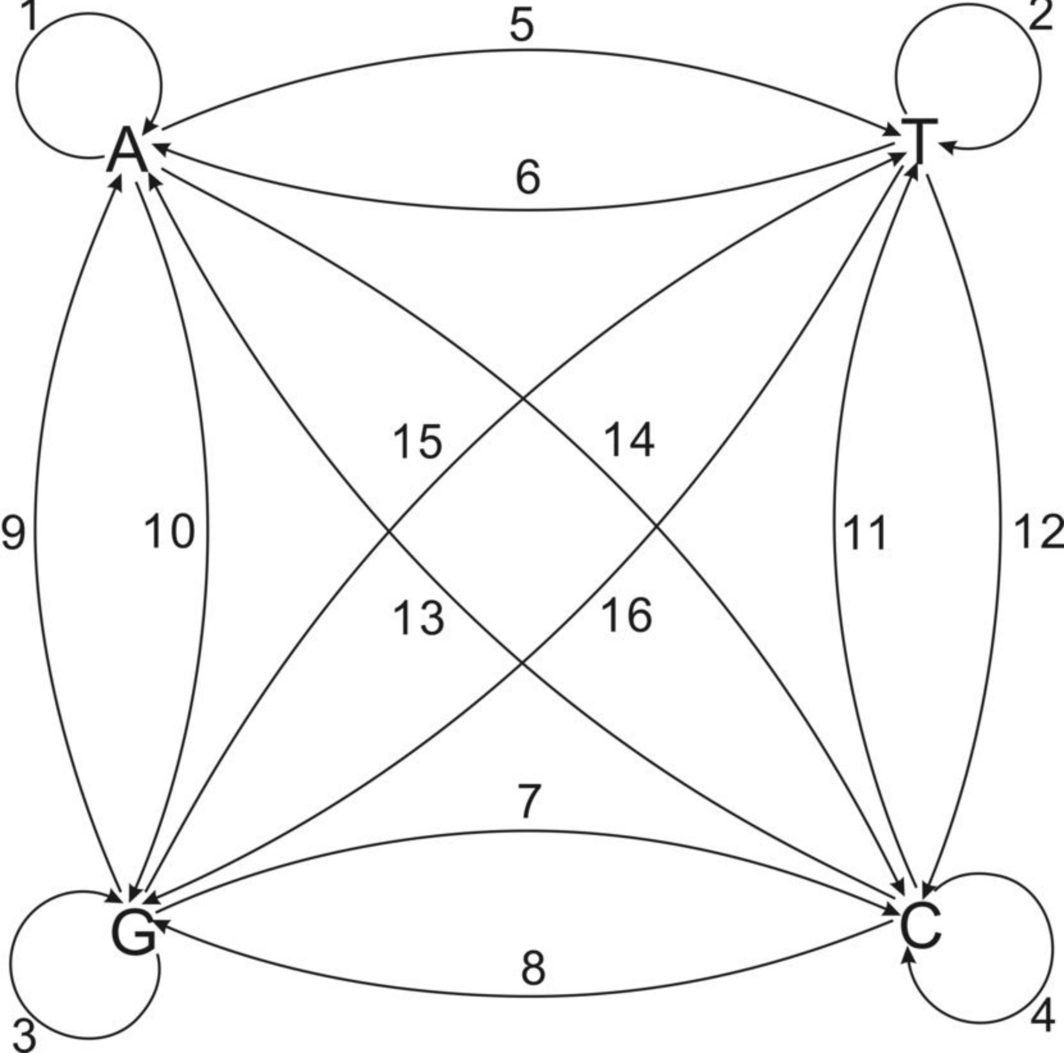
Figure 1. Combination of two bases that locate on two symmetrical sites of two arms. Four nodes indicate bases A, C, T and G, 16 different combinations of two bases are shown by 16 directed edges. Two bases linked by an edge locate at two symmetrical sites of two arms. For instance, the edge denoted as 5 represents the base A linked tail of the edge locates at left arm, and the base T linked head of the edge locates at the symmetrical site of right arm.

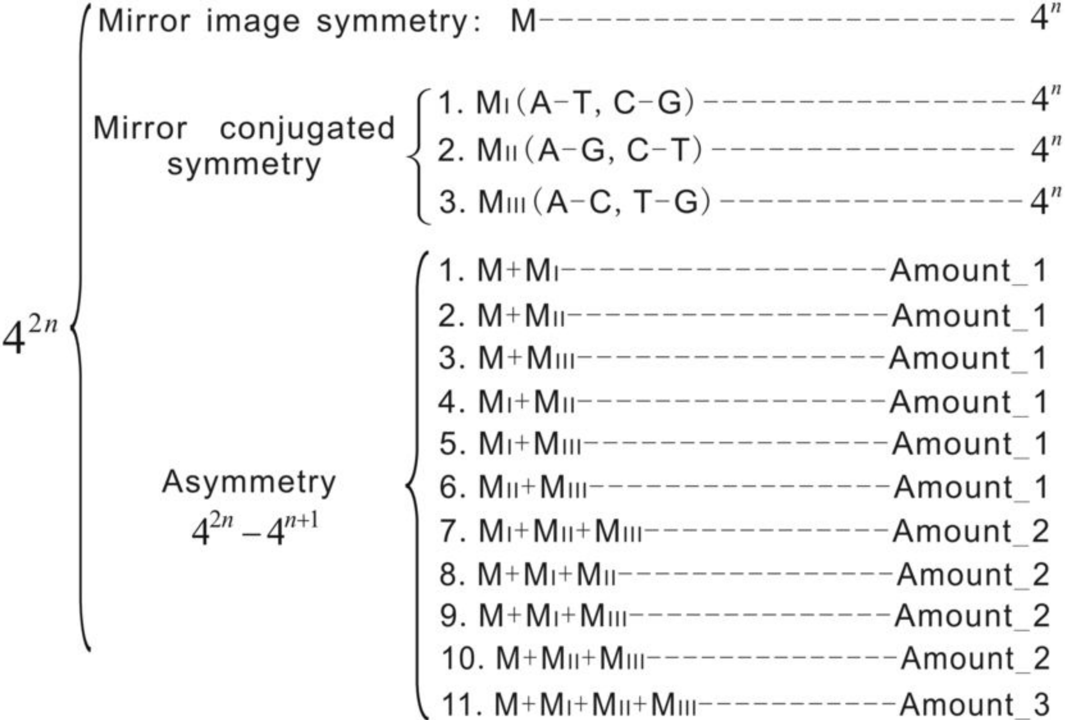
Figure 2. Classifications and quantities of all sequences. There are four kinds of symmetry sequences M, M_I, M_{II}, M_{III}, and 11 kinds of asymmetries sequences. For instance, M_I+M_{II} represent the sequences that possess the first kind of mirror conjugated symmetry at some sites and the second kind of mirror conjugated symmetry at the rest sites. Amount_1, Amount_2 and Amount_3 (equation (1), (2) and (3)) represent the amounts of the asymmetry sequences assembling two, three and four kinds of symmetries, respectively.

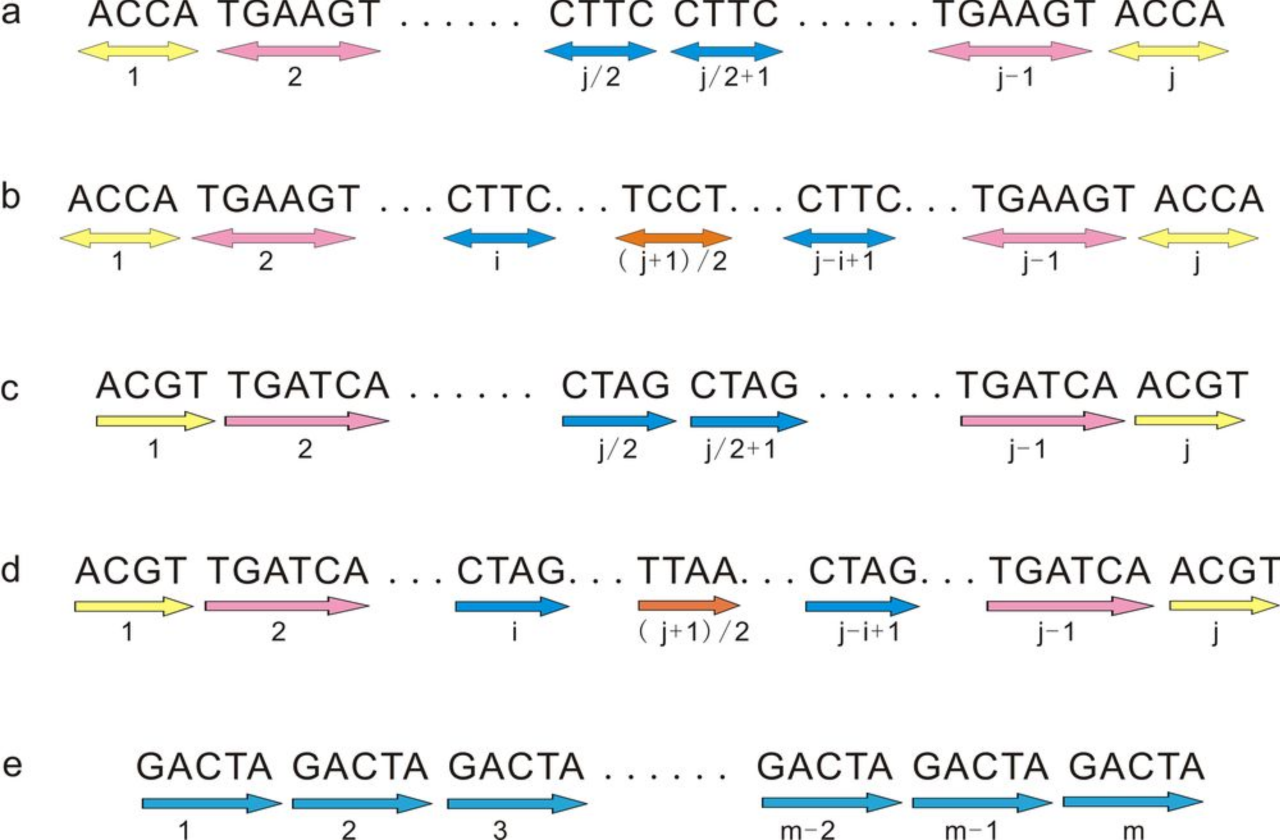
Figure 3. Symmetries of M combinators, M_I combinators and direct repeat sequences. The arrows dyed the same colour represent the same generators/repetitive units, direction of arrows are employed to demonstrate symmetry. The concolorous arrows demonstrate mirror image symmetry if their direction are reverse, translational symmetry if their direction are the same, both of the above two kinds of symmetries if arrows are bidirectional. **a** and **b**, M combinators constructed by even generators and odd generators, respectively. **c** and **d**, M_I combinators constructed by even generators and odd generators respectively. For the two kinds of combinators, their generators distribute symmetrically if the amount of generators is even, in contrast, one generator locates on the centre and the rest distribute symmetrically if the

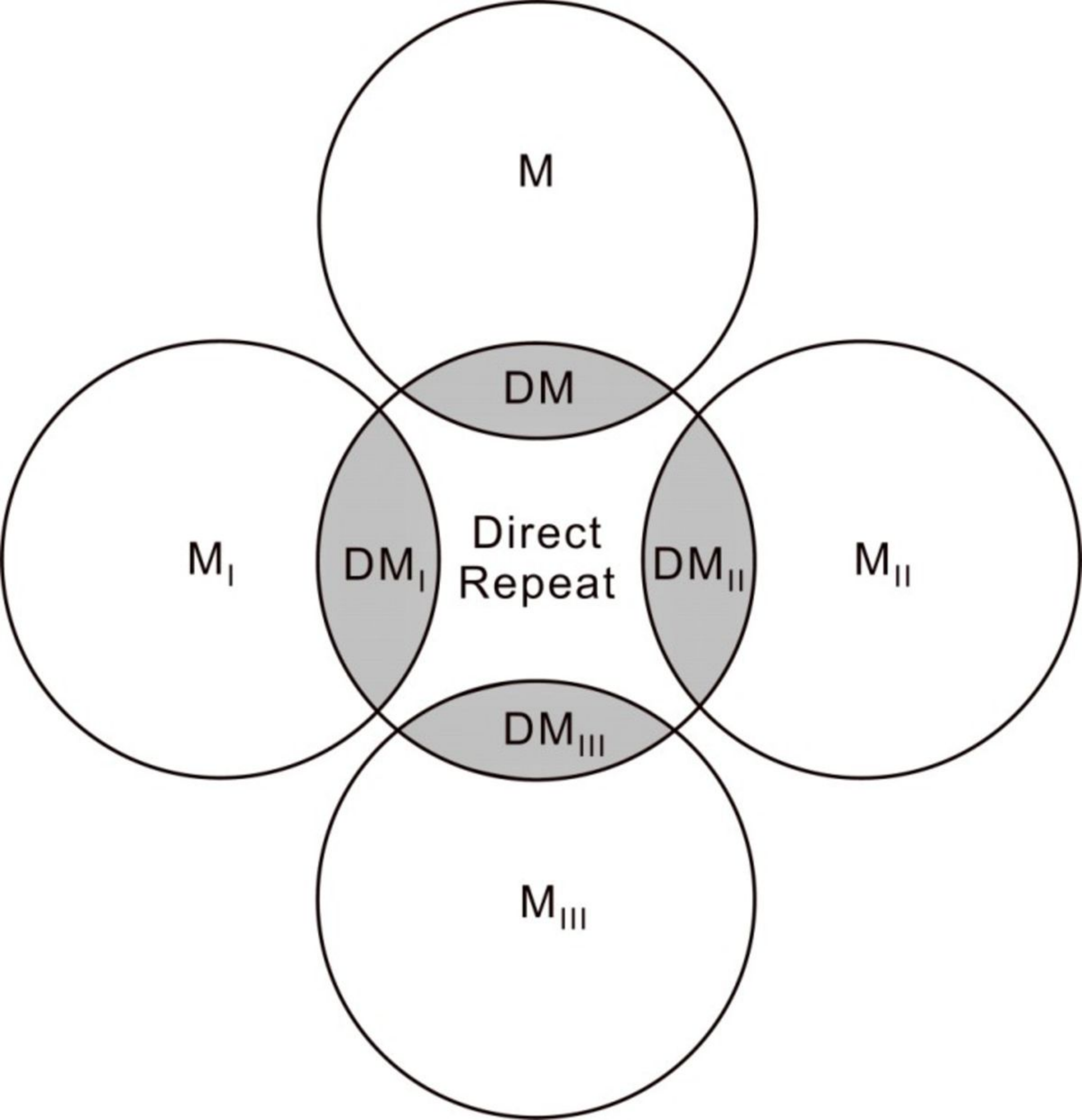
amount of generators is odd. **e**, a direct repeat sequence, where repetitive unit is GACTA, *t*=5, and the repetitive times equals to *m*.

Figure 4. Relationships among M, $\mathbf{M_I}$, $\mathbf{M_{II}}$, $\mathbf{M_{III}}$ and direct repeat sequences. We denote by DM, DM_I, DM_{II} and DM_{III} the intersections between direct repeat sequences with M, M_I, M_{II} and M_{III}, respectively. Note that M, M_I, M_{II} and M_{III} sequences have not intersections each other because of their different symmetry.


1
5
2
A
T
6
15
14
9
10
11
12
13
16
7
G
C
8
3
4


4^{2n}
Mirror image symmetry: M ---------- 4^n
Mirror conjugated symmetry
1. M_{I}(A–T, C–G) ---------- 4^n
2. M_{II}(A–G, C–T) ---------- 4^n
3. M_{III}(A–C, T–G) ---------- 4^n
Asymmetry
$4^{2n}-4^{n+1}$
1. $M+M_{I}$ ---------- Amount_1
2. $M+M_{II}$ ---------- Amount_1
3. $M+M_{III}$ ---------- Amount_1
4. $M_{I}+M_{II}$ ---------- Amount_1
5. $M_{I}+M_{III}$ ---------- Amount_1
6. $M_{II}+M_{III}$ ---------- Amount_1
7. $M_{I}+M_{II}+M_{III}$ ---------- Amount_2
8. $M+M_{I}+M_{II}$ ---------- Amount_2
9. $M+M_{I}+M_{III}$ ---------- Amount_2
10. $M+M_{II}+M_{III}$ ---------- Amount_2
11. $M+M_{I}+M_{II}+M_{III}$ ---------- Amount_3


a
ACCA TGAAGT . . . . . . CTTC CTTC . . . . . . TGAAGT ACCA
1
2
j/2
j/2+1
j−1
j
b
ACCA TGAAGT . . . CTTC. . . TCCT. . . CTTC. . . TGAAGT ACCA
1
2
i
(j+1)/2
j−i+1
j−1
j
c
ACGT TGATCA . . . . . . CTAG CTAG . . . . . . TGATCA ACGT
1
2
j/2
j/2+1
j−1
j
d
ACGT TGATCA . . . CTAG. . . TTAA. . . CTAG. . . TGATCA ACGT
1
2
i
(j+1)/2
j−i+1
j−1
j
e
GACTA GACTA GACTA . . . . . . GACTA GACTA GACTA
1
2
3
m−2
m−1
m


M
DM
M$_I$
DM$_I$
Direct Repeat
DM$_{II}$
M$_{II}$
DM$_{III}$
M$_{III}$