Identification of 5 Novel Y Chromosome STR Loci and Haplotype Distribution in Chinese Han Population(Guangzhou)

ZHU Yun-liang[1,2], HUANG Yan-mei[1,3], LI Jian-jin[1], WU Xin-yao[1*]
(1. Department of Forensic Medicine, Sun Yat-sen University, Guangzhou 510080, China;
2. College of Basic Science of Medicine, Zhengzhou University, Zhenzhou 450052, China;
3. Department of Forensic Medicine, Xinxiang Medical College, Xinxiang 453000, China)

Abstract:【Objective】To screen and identify the new Y-STR loci from the Y chromosome and examine the polymorphism of these Y-STR loci.【Method】To seek and locate the position of 5 Y-STR loci, including DYS709, DYS720, DYS721, DYS722, and DYS723, and perform sequencing of these 5 Y-STR loci. Then to investigate the polymorphism in unrelated Chinese Han males.【Results】Five Y-STR loci were identified from Y chromosome sequence. By scrutinizing the physical position on Y chromosome of previously reported Y-STRs, we found that three loci were novel and two loci overlapped with two loci published only online. All loci could be male-specifically amplified with a product size ranging from 185 bp to 278 bp. After 108 males of the Chinese Han Population (Guangzhou) were examined, we found 5 DYS709, 11 DYS720 alleles, 4 DYS721 alleles, 6 DYS722 alleles, and 6 DYS723 alleles. A total of 95 haplotypes were identified, 84 of which were unique, and with a haplotype diversity of 0.997 2 ± 0.001 2 (HD ± SE).【Conclusion】This set of Y-STRs can be used as Y chromosome genetic makers in related fields.

Key words: STR typing; Y chromosome; short tandem repeats; haplotype; Chinese Han population
CLC No: Q987 **Document Mark**: A **Article ID**: 1672-3554(2010)01-0007-06

5 个新 Y-STR 基因座及其在广州地区汉族群体内的单倍型分布

朱运良[1,2], 黄艳梅[1,3], 李建金[1], 伍新尧[1*]
(1. 中山大学法医学系, 广东 广州 510080; 2. 郑州大学基础医学院, 河南 郑州 450001;
3. 新乡医学院法医学系, 河南 新乡 453000)

摘 要:【目的】从 Y 染色体 DNA 序列中筛选新的 Y-STR 基因座,调查其在汉族群体中的遗传多态性。【方法】在人类 Y 染色体 DNA 序列中查找 Y-STR 基因座,确定它们在 Y 染色体上的定位并测序,与已报道的 Y-STR 基因座比较。检测汉族(广州地区)无关男性群体中这些 Y-STR 基因座单体型。【结果】鉴别出 5 个 Y-STR 基因座(DYS709,DYS720,DYS721,DYS722 和 DYS723),其中 3 个 STR 基因座是新发现的,另 2 个基因座与在线报告的重叠。所有 5 个 STR 基因座都是男性特有的。它们的扩增产物长度介于 185 bp ~ 278 bp。在 108 例汉族无关男性个体中发现 DYS709 有 6 个,DYS720 有 11 个,DYS721 有 4 个,DYS722 有 6 个,DYS723 有 6 个等位基因。在本组群体中发现 95 种单体型,其中 84 种只出现 1 次。单体型多样性达 0.997 2 ± 0.001 2。【结论】这组 5 个 Y-STR 基因座可望应用于汉族群体的有关研究中。

关键词: STR 分型; Y 染色体; 短串联重复序列; 单体型; 汉族群体
[J SUN Yat-sen Univ(Med Sci),2010,31(1):7-12]

Y-STRs are short tandem repeats (STRs) markers on the non-recombinant part of the Y chromosome (NRY, also known as MSY, male-specific region). The MSY has no homologous regions on the X chromosome and the entire MSY is transmitted from father to son as a haplotype. Thus,

Receiving date: 2009-07-01
Fund Items: 211 Project Special Item Construction Fund (4209008); Guangdong Science Fund (97-128)
Authors Biography: ZHU Yun-liang, Doctor of Forensic Medicine, E-mail:zhuyl@zzu.edu.cn; * Communication author: WU Xin-yao, Professor of Forensic Medicine, Tutor for Doctor, E-mail:wuxinyao@mail.sysu.edu.cn

all male offsprings of the same paternal lineage have the same haplotype unless mutation occurs. This property makes Y-STRs a useful tool in many aspects, such as in forensic science[1-2], evolution studies and population history research. However, in forensic science Y-STRs are less powerful than STRs on autosomes because the power of exclusion (PE) of Y-STRs cannot be multiplied between different loci; instead, the PE should be calculated based on haplotype frequencies defined by typed Y-STRs. In forensic practice, a large number of Y-STRs are needed in order to increase the discriminating power and to facilitate the development of multiplex PCR where usually only a small amount of DNA can be obtained. The detail information for each locus, including its repeats structure, allele nomenclature, amplification condition, mutation rate, and allele frequency distribution in different populations, are needed to render these loci useful for forensic casework. Kayser *et al*[3] surveyed human microsatellites DNA with a unit size ≥3 bp and a repeat count of ≥8 in the entire euchromatic region of the Y-chromosome and found 430 novel loci, but only 139 loci are potentially useful. The rest were not studied thoroughly because no primer pairs could be found in silico (149 loci) or the PCR could not be performed male-specifically (115 loci) or no polymorphism was observed. Prior to Kayer *et al*, 52 loci were published[2, 4-12], however, more well-characterized loci are still needed in forensic individual identification and other aspects. Here we describe in detail 5 Y-STRs on the human Y chromosome, 3 of which are novel. The other two, although reported by Kayser *et al* in online databases, had not been studied thoroughly. We have also determined the allele and haplotype frequency distribution of these 5 loci in the Chinese Han males (Guangzhou).

1 Materials and method

1.1 Population sample

A total of 108 blood samples were randomly collected from males of Chinese Han males who donated their blood to the Guangzhou Center of Transfusion in China. Informed consent had been obtained from the blood donors. The project also received approval from the Ethics Committee of the Guangzhou Center of Transfusion. A total of 4 female DNA specimens were collected from unrelated volunteers. DNA was isolated using standard organic extraction method.

1.2 Identification and PCR amplification of novel Y-STRs

We selected two clones (AC006989 and AC011289) from Yq11.221 of the human Y chromosome as our target to search for new Y-STRs loci (see the Discussion section on why these two clones have been chosen). The sequences of AC006989 and AC011289 were downloaded from GenBank and loaded into the computer program Tandem Repeats Finder[13]. We set the program to specifically output Y-STRs with a stretch of 8 or more consecutive repeats and a repeat unit of 4-6 base pairs. PCR primers were designed using Primer3 software (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi)[14] based on the sequences of the 200 bp flanking regions of these repetitive stretches. Primers were synthesized by SBS (Beijing, China). Each PCR reaction was carried out using GeneAmp PCR System 9 600 (Perkin-Elmer, USA) in a total volume of 25 μL containing 20 ng human genome DNA, 1 × Taq buffer, 1.5 mmol/L $MgCl_2$, 0.2 mmol/L each dNTP, 1 U Taq DNA polymerase (TAKARA, Dalian, China), and 300 nmol/L each primer. In a TouchDown PCR protocol, the template DNA was initially denatured at 94 ℃ for 2 min, and then followed by 10 cycles starting at 94 ℃ for 1 min, 60 ℃ for 1 min and 72 ℃ for 1 min. The annealing temperature was decreased by 0.5 ℃ per cycles then followed by 25 cycles of 94 ℃ for 1 min, 55 ℃ for 1 min and 72 ℃ for 1 min. After a final extension step at 72 ℃ for 5 min the samples were kept at 4 ℃ until separation by electrophoresis on 6% non-denaturing polyacrylamide gel and then

silver-stained. The four DNA samples from females were used as negative controls.

1.3 Nomenclature, genotypes, and statistics

The name of each STRs locus identified in our study was assigned by GDB (Genome Data Base, www.gdb.org). Referring to the human Y-STRs physical map in which all 417 Y-STRs loci (including these five) deposited in GDB are sequentially positioned along the Y chromosome[15], we determined the novelty of each locus found. The alleles of each locus were named according to the recommendations of DNA Commission of the International Society of Forensic Genetics (ISFG)[16]. Two clones of chimpanzee Y chromosme sequences, AC151722 and AC151722,were retrieved by BLAST searches of each locus sequence against GenBank. The chimpanzee homologous to DYS709 was on AC151722, and the chimpanzee homologous to other four loci, DYS720, DYS721, DYS722, and DYS723, were on AC147710. We then used chimpanzee homologous sequences as the reference to identify non-variant stretches within each locus. Those non-variant stretches that were not adjacent to variable repetitive motifs and had three or fewer units were not included in the allele nomenclature, in order to allow designing primers that produce shorter PCR products which would still lie within our allele nomenclature.

We prepared allelic ladders by mixing PCR products of different lengths (presenting different alleles). The frequencies of alleles/haplotypes were obtained by simple allele/haplotype counting. Gene-diversity or haplotype-diversity (GD or HD) values were calculated according to the following equation[17].

$$GD(HD)=\frac{n(1-\sum p^2)}{n-1}$$

Standard errors were calculated according to the following equation[17].

$$SE=\sqrt{\frac{2\sum p^3-2(\sum p^2)^2}{n}}$$

Where n denotes the number of individuals, p denotes the frequency of the allele (or haplotype).

2 Results

2.1 Identification and amplification of these new Y-STRs

A total of six microsatellite loci were identified within a region of 238.2 kb on the human Y chromosome genome sequence. These loci have an array of 8 or more consecutive perfect repeats; five of these loci have tetranucleotide repeats, and one has pentanucloetide repeats. Based on the STRs physical map of the human Y chromosome, we concluded that two tetranucleotide repeats, DYS720 and DYS723, and the pentanucleotide repeat DYS721 are novel. DYS709 is the same locus as DYS516, identified by Kayser *et al*[3] (25 bp shorter than DYS709) and DYS722 is the same locus as DYS517 (35 bp smaller). DYS516 and DYS517 were merely published in online databases. We describe these two loci here (also see the discussion section). All of these five loci showed that they were unique in the Y chromosome by yielding a single PCR product, and no PCR products were observed when four female DNAs were used as templates. DYS390, which is located on AC006989, was also identified by the Tandem Repeat Founder program. This locus is used broadly in forensic science now and thus is not described here.

2.2 Allele nomenclature

In order to exclude non-variant repetitive blocks, which should not be included in the allele nomenclature, we compared the repetitive structures of each locus with their homologous sequence in chimpanzee DNA. The homologues retrieved from GenBank include two clones, AC151722 and AC147710 (DYS709 was on AC151722 and others were on AC147710). The details of comparisons are illustrated in Fig.1.

2.3 Allele and the haplotype frequency distribution in Chinese Han males (Guangzhou)

By examining the Y chromosomal DNA samples from 108 males of the Chinese Han males

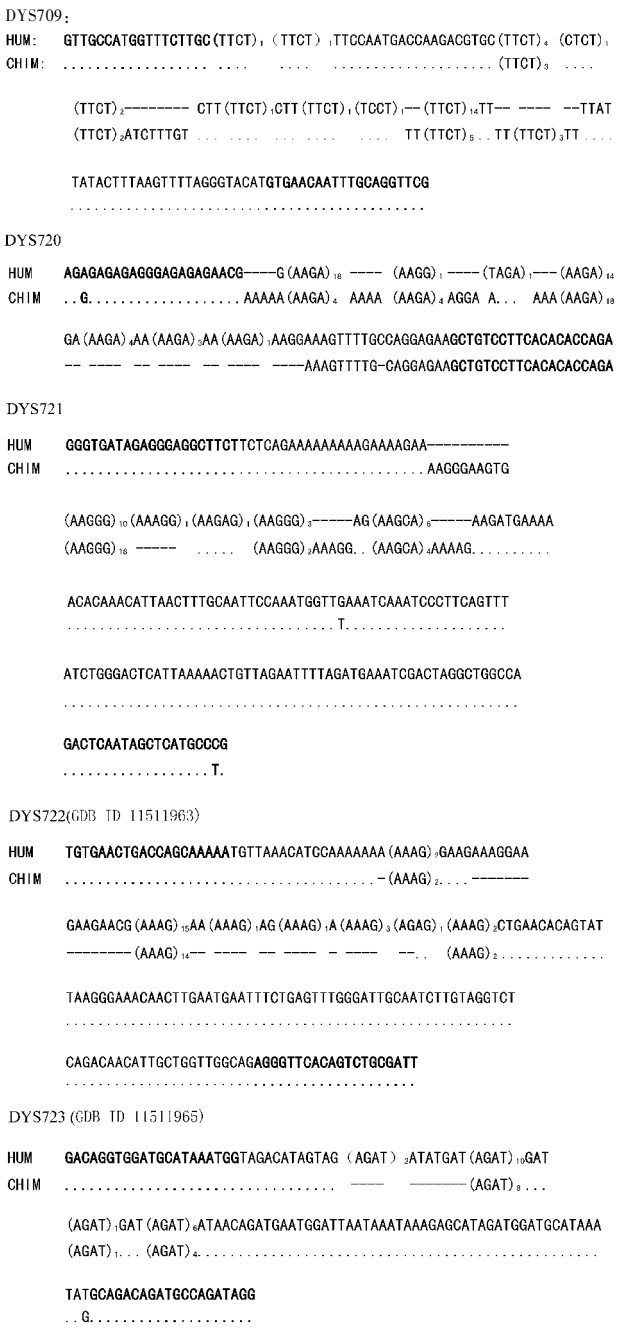

Fig.1 The sequence data and repeat structures of five Y-STRs

Primer sequences were in bold. "." denoted the same base and "-" presented deletion. HUM: human; CHIM: chimpanzee. Each locus nomenclature should be as follows (m and n denoted repeat numbers that were variable among different alleles).

DYS709: The first repeat block, $(TTCT)_2$(The first two bases were in the primer region), was 20 bp apart from the second block, and had the same repeat number as in chimpanzees, so it should be excluded from the allele nomenclature. The repetitive structure should be described as: $(TTCT)_4(CTCT)_1(TTCT)_2N_3(TTCT)_1N_3(TTCT)_1(TCCT)_1(TTCT)_m$. Therefore the allele name should be the sum of $4+1+2+1+1+1+m$.

DYS720: There were two highly variable blocks in this locus and between these two blocks there were two mutated units, AAGG and TAGA. Other repeats were adjacent to the second variant stretch and should be included in the allele nomenclature. The repetitive structure was: $(AAGA)_m(AAGG)_1(TAGA)_1(AAGA)_nN_2(AAGA)_4N_2(AAGA)_3N_2(AAGA)_1$. The allele name was thus the sum of $m+1+1+n+4+3+1$.

DYS721: This locus is a pentanucleotide repeat. Two mutated units, AAAGG and AAGAG(in chimpanzees only AAGAG), separated the first from the second repeat stretches. The second variable repeat stretch had a different repeat unit of AAGCA. The repetitive structure was: $(AAGGG)_m(AAAGG)_1(AAGAG)_1(AAGGG)_3N_2(AAGCA)_n$. So the allele name of this locus should be presented as $m+1+1+3+n$.

DYS722: The first block had two repeat units both in humans and chimpanzees and was not adjacent to the main variant repeat stretch (at a distance of 19 bp), so it should not be included in allele name. The structure of main repeat blocks was different between human and chimpanzee. The repetitive structure for the allele nomenclature in human was: $(AAAG)_mN_2(AAAG)_1N_2(AAAG)_1N_1(AAAG)_3(AGAG)_1(AAAG)_2$. The allele should be named after $m+1+1+3+1+2$.

DYS723: The first block did not exist in chimpanzees, so we suggested it should be included in the allele nomenclature. The repetitive structure was: $(AGAT)_2N_7(AGAT)_mN_3(AGAT)_1N_3(AGAT)_6$. The allele name thus was $2+m+1+6$.

(Guangzhou), we found DYS709 5 alleles, 11 DYS720 alleles, 4 DYS721 alleles, 6 DYS722 alleles, and 6 DYS723 alleles. The gene diversities (GD ± SE) were of 0.703 9 ± 0.015 8, 0.868 2 ± 0.008 8, 0.456 3 ± 0.033 8, 0.621 7 ± 0.029 4, and 0.627 1 ± 0.022 8, respectively (Table 1). A total of 95 haplotypes were found in our samples and among these haplotypes, 84 were unique. There were 10 haplotypes each of which was shared by two males. The most frequent haplotype is shared by 4 males and includes alleles 25 (DYS709), 39(DYS720), 18 (DYS721), 21(DYS722), 20 (DYS723). The haplotype diversity (HD± SE) of this system was 0.997 2 ± 0.001 2. The detail information of haplotypes can be obtained from authors upon request.

3 Discussion

There are three classes of euchromatic sequences on the MSY: X-transposed, X-degenerate, and

Table 1　The alleles and their frequencies of 5 Y-STRs in Chinese Han male population

Locus	Allele	Size (bp)	Number of Chr observed	Frequency	
DYS709	22	186	2	0.018 5	G.D: 0.703 9
	23	190	16	0.148 1	S.E: 0.015 8
	24	194	44	0.407 4	
	25	198	33	0.305 6	
	26	202	13	0.120 4	
DYS720	34	207	2	0.018 5	G.D: 0.868 1
	35	211	4	0.037 0	S.E: 0.008 8
	36	215	3	0.027 8	
	37	219	7	0.064 8	
	38	223	14	0.129 6	
	39	227	21	0.194 4	
	40	231	23	0.213 0	
	41	235	8	0.074 1	
	42	239	9	0.083 3	
	43	243	10	0.092 6	
	44	247	7	0.064 8	
DYS721	17	271	9	0.083 3	G.D: 0.456 3
	18	276	76	0.703 7	S.E: 0.033 8
	19	281	22	0.203 7	
	20	286	1	0.009 3	
DYS722	19	258	1	0.009 3	G.D: 0.621 7
	20	262	13	0.120 4	S.E: 0.029 4
	21	266	21	0.194 4	
	22	270	61	0.564 8	
	23	274	4	0.037 0	
	24	278	8	0.074 1	
DYS723	18	190	2	0.018 5	G.D: 0.627 1
	19	194	13	0.120 4	S.E: 0.022 8
	20	198	56	0.518 5	
	21	202	32	0.296 3	
	22	206	4	0.037 0	
	23	210	1	0.009 3	

ampliconic[18]. We originally intended to find single copy loci in the human Y chromosome sequence, because single copy loci are preferable to multi-copy loci in many fields, such as in forensic practice to determine the number of contributors to a forensic mixture of DNA in a stain and in human evolution studies to calculate the time to the most recent common ancestry (TMRCA). But we found that STRs from the ampliconic section tend to be multi-copy and STR from the X-transposed section is difficult to be amplified male-specifically. By roughly examining the position of reported Y-STRs on the Y chromosome, we found that the density of loci on the Yq11.221 region was low (before 2004, only DYS390 had been reported). We thus chose two random clones, AC006989 and AC011289, from this X-degenerate region as our target. Microsatellites with a unit size of 4–6 bp and repeat count of ≥8 were selected by Tandem Repeats Finder software. Loci selected by this criterion would have a high probability of showing variation and would be free from "stuttering" in STR typing. From these two clones, we identified six microsatellite loci. To determine the novelty of the loci identified, we examined the STRs physical map of the human Y chromosome, which sequentially positioned all Y-STRs loci, along the Y chromosome. We found that three of the six loci overlapped with other loci. DYS708 (15712964–15713207) overlapped with DYS390 (15713015–15713228) which is broadly used now in forensic science. The other two loci overlapped with Kayser's finding: DYS709 (15521019–15521216) was the same locus as DYS516 (15521044–15521216) and DYS722 (15716340–15716613) was the same locus as DYS517 (15716316–15716557). The remaining three loci, DYS720, DYS721, and DYS723, were novel. As the software and the criteria used to identify STR loci by us were almost the same as Kayser did, DYS720, DYS721 and DYS723 must also have been discovered by Kayser. The reason that they did not study these three loci further might be that no suitable primer pairs in silico could be found. This indicated that more Y-STRs could be identified by re-screening Y chromosome sequence. DYS516 is classified as no polymorphic locus by Kayser because no variation was found in 8 DNA samples from eight different Y-SNP haplogroups. We found that DYS709, the same locus as DYS516, was polymorphic in Chinese Han males (Guangzhou). DYS516 and DYS517 were published online-only and the nomenclature of these two loci was not

reported based on the recommendations of DNA Commission of the International Society of Forensic Genetics (ISFG), we therefore presented the allele nomenclature of these loci in this paper in order to facilitate the result comparison between different authors.

Of the five loci analyzed, four are tetranucleotide repeats and one is pentanucleotide repeat. The GD of the locus with the pentanucleotide repeat was the lowest, although it contained two variant blocks. Because of having two highly variable blocks, DYS720 has a high value of GD (0.8681), higher than most single-copy Y-STR loci. The HD of this set of Y-STRs was 0.997 2 ± 0.001 2 (HD ± SE). The lengths of PCR products of these STR loci ranged from 185 bp (Allele 19 of DYS709) to 278 bp (Allele 23 of DYS722).

Further study of these new markers should be done in order to make use of them in forensic science and other fields, for example, to develop multiplex PCR with other Y STRs, to investigate their mutation rates in different populations, etc.

To sum up, these new Y-STR loci are useful tools for forensic sciences because of their high HD, discriminating power and power of exclusion and could also be used in human evolution studies.

References:

[1] Jobling MA, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing [J]. Int J Legal Med, 1997,110(3):118-124.

[2] Kayser M, Caglia A, Corach D, et al. Evaluation of Y-chromosomal STRs: a multicenter study [J]. Int J Legal Med, 1997,110(3):125-133,141-129.

[3] Kayser M, Kittler R, Erler A, et al. A comprehensive survey of human Y-chromosomal microsatellites [J]. Am J Hum Genet, 2004,74(6):1183-1197.

[4] Iida R, Tsubota E, Matsuki T. Identification and characterization of two novel human polymorphic STRs on the Y chromosome [J]. Int J Legal Med, 2001, 115(1):54-56.

[5] Chen H, Lowther W, Avramopoulos D, et al. Homologous loci DXYS156X and DXYS156Y contain a polymorphic pentanucleotide repeat (TAAAA)n and map to human X and Y chromosomes [J]. Hum Mutat, 1994,4(3):208-211.

[6] Mathias N, Bayes M, Tyler-Smith C. Highly informative compound haplotypes for the human Y chromosome [J]. Hum Mol Genet, 1994,3(1):115-123.

[7] Jobling MA, Samara V, Pandya A, et al. Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males [J]. Hum Mol Genet, 1996,5(11):1767-1775.

[8] White PS, Tatum OL, Deaven LL, et al. New, male-specific microsatellite markers from the human Y chromosome [J]. Genomics, 1999,57(3):433-437.

[9] Ayub Q, Mohyuddin A, Qamar R, et al. Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information [J]. Nucleic Acids Res, 2000,28(2):e8.

[10] Iida R, Tsubota E, Sawazaki K, et al. Characterization and haplotype analysis of the polymorphic Y-STRs DYS443, DYS444 and DYS445 in a Japanese population [J]. Int J Legal Med, 2002,116(3):191-194.

[11] Bosch E, Lee AC, Calafell F, et al. High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions [J]. Forensic Sci Int, 2002,125 (1):42-51.

[12] Redd AJ, Agellon AB, Kearney VA, et al. Forensic value of 14 novel STRs on the human Y chromosome [J]. Forensic Sci Int, 2002,130(2-3):97-111.

[13] Benson G. Tandem repeats finder: a program to analyze DNA sequences [J]. Nucleic Acids Res, 1999,27(2):573-580.

[14] Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers [J]. Methods Mol Biol, 2000,132(9):365-386.

[15] Hanson EK, Ballantyne J. Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications [J]. Leg Med (Tokyo), 2006,8 (2):110-120.

[16] Gusmao L, Butler JM, Carracedo A, et al. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis [J]. Forensic Sci Int, 2006,157(2-3):187-197.

[17] Nei M. Molecular Evolutionary Genetics [M]. New York: Columbia University Press, 1987:176-187.

[18] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes [J]. Nature, 2003,423(6942):825-837.

(编辑 王晓鹰)