Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

An online conserved SSR discovery through cross-species comparison

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
29 May 2009
Number of times this article has been viewed

Tun-Wen Pai[1]
Chien-Ming Chen[1]
Meng-Chang Hsiao[1]
Ronshan Cheng[2]
Wen-Shyong Tzou[3]
Chin-Hua Hu[3]

[1]Department of Computer Science and Engineering; [2]Department of Aquaculture, [3]Institute of Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan, Republic of China

Abstract: Simple sequence repeats (SSRs) play important roles in gene regulation and genome evolution. Although there exist several online resources for SSR mining, most of them only extract general SSR patterns without providing functional information. Here, an online search tool, CG-SSR (Comparative Genomics SSR discovery), has been developed for discovering potential functional SSRs from vertebrate genomes through cross-species comparison. In addition to revealing SSR candidates in conserved regions among various species, it also combines accurate coordinate and functional genomics information. CG-SSR is the first comprehensive and efficient online tool for conserved SSR discovery.

Keywords: microsatellites, genome, comparative genomics, functional SSR, gene ontology, conserved region

Introduction

SSRs, also called simple tandem repeats (STRs) or microsatellites, are DNA segments composed of tandem repetitions of relatively short motifs. They are commonly and easily identified DNA sequences which consist of repeated units one to six base pairs in length.[1–4] For decades, SSRs were mainly considered to be genetic markers in DNA fingerprinting and diversity studies due to their high rate of polymorphism. Nevertheless, recent studies pointed out that SSR expansions and/or contractions in protein-coding regions bring about a gain or loss of gene function through frameshift mutations.[2–4] SSR variations in 5'UTRs could affect gene transcription and translation, whereas SSR expansions in 3'UTRs could cause transcription slippage and result in disrupting splicing and possibly disturbing cellular functions. For example, a CGG repeat pattern within the 5'UTR of the *FMR1* gene is expanded in families with fragile X syndrome. When the length of SSR exceeds 200 CGGs, mental retardation occurs due to the absence of the encoded FMR protein.[5] Another example is a CTG expansion located in the 3'UTR of a *kinase* gene involved in myotonic dystrophy type 1 (*DM1*), a multisystemic dominantly inherited disorder. *DM1* disorder affects skeletal and smooth muscle as well as the eye, heart, endocrine system, and central nervous system.[6] Furthermore, SSRs in introns affect transcription, mRNA splicing, or export to the cytoplasm, which have been shown that SSRs indeed possess a functional role as cis-regulatory elements. For example, a CA simple sequence repeat within the first 2000 bases in intron 1 enhances *egfr* transcription and involves in breast carcinogenesis.[7] These functional SSRs, SSRs with biological functions, play an important role in gene regulation.[2] Consequently, the discovery of potential functional SSRs to decipher gene regulatory networks intrigues biologists.

Correspondence: Tun-Wen Pai
Department of Computer Science and Engineering, National Taiwan Ocean University, No 2, Peining Road, Keelung, 20224, Taiwan, Republic of China
Tel +886 2 2462 2192 Ex 6618
Fax +886 2 2462 3249
Email twp@mail.ntou.edu.tw

There are many *in silico* SSR mining tools and databases such as MMDBJ,[8] Satellog,[9] MRD,[10] SSRD,[11] and EuMicroSatdb.[12] All of these mining tools are briefly described by Aishwarya and colleagues,[12] but none of them allows users to retrieve potential functional SSRs using comparative genomics. These available tools either emphasize a collection of SSRs from specific organisms or provide only limited functions for various genome comparisons. Accordingly, it is still tedious for biologists to find functional SSRs from millions of SSR candidates.

Due to functional constraints, DNA regions involved in gene regulation or genome evolution are expected to be conserved among related species. It has been shown that cross-species comparison of DNA sequences can facilitate identification of candidate regulatory elements.[13] If SSRs possess significant biological functions, they are likely to be located in conserved regions. In this report, the proposed comparative genomics SSR discovery (CG-SSR) web service comprises eleven representative vertebrate species (human, chimpanzee, orangutan, mouse, rat, opossum, rhesus, cow, dog, zebrafish, and medaka) for constructing the fundamental SSR database. It also includes another thirteen species (cat, horse, marmoset, guinea pig, platypus, chicken, lizard, *Xenopus tropicalis*, tetraodon, fugu, stickleback, lamprey, lancelet) for verification of conserved regions of the retrieved SSRs. Users can evaluate the biological significance of SSRs through cross-species conservation because such comparisons are available for the various species selected as the representative model organisms in CG-SSR. CG-SSR also combines other relevant functional genomics information such as GO (Gene Ontology),[14] InterPro,[15] and Pfam.[16] GO provides biological annotation of genes in light of their associated biological processes, cellular components, and molecular functions. The InterPro database contains identifiable features – protein families, domains, repeats and sites – found in known proteins that can be applied to novel proteins. Through hidden Markov models and multiple sequence alignments, the Pfam database collects domain information for a large quantity of protein families. Taken together, these functional genomics resources permit biologists to decipher candidate roles for SSRs in gene regulatory networks.

Methods

To construct the database for CG-SSR, vertebrate genome sequences and gene coordinates were obtained from Ensembl Genome Browser.[17] Whole genomes were scanned based on an efficient correlation method for SSR mining that is composed of two major phases including autocorrelation and overlapping adjacent phases (see Supplementary materials). In the first autocorrelation phase, the CG-SSR performs seed finding, region growing, trimming, recovering, pattern refinement, and noise filtering processes to discover all possible SSR repeats as initial candidates. For the second phase, the system verifies the overlapping records and confirms no redundant patterns by employing merging, cutting and threshold filtering processes. Figure 1 illustrates the flowchart of CG-SSR searching algorithm and a detailed description of the developed methods and examples can be found in the Supplementary materials.

The core algorithm for discovering SSR patterns from genome sequences employs autocorrelation methodology. The basic concept assumes that the target sequence contains repeat substrings with basic patterns of length N within a range from 1 to 6. If we shift the target sequence with N nucleotides to its right and compare to the original sequence, all repeat patterns can be discovered since they overlapped with their shifted sequence at least within N nucleotides continuously. Based on the observation of such overlapping, we can easily detect the repeating locations by shifting and matching the whole target sequence without knowing the nucleotide contents of repeated patterns. If the continuously matched nucleotides are longer than the shifting length N, at least one repeat with the basic pattern length N is proved to exist in the target sequence. Hence, exact matched strings can be identified and considered as the seeds of perfect SSR candidates. These seeds were then extended by region growing techniques to formulate imperfect SSR sequences. To identify imperfect SSRs with various tolerant conditions, a module applied forward region growing method to perform neighboring comparison. As long as the coordinates and contents of the perfect repeat patterns were obtained, the forward region growing processes were performed by examining the right-hand-side neighboring nucleotides and skipping noise-like nucleotides which did not belong to the perfect repeat patterns. As no extra basic unit pattern could be found by continuously extending the verification on right-hand-side neighboring nucleotides, the searching processes terminated. After the region growing processes, the boundary detection through trimming/recovering operations was applied to delimit the appropriate range of each SSR sequence. In this module, verifying pattern boundaries on both sides was achieved by left-hand-side trimming and right-hand-side recovering processes. The right-hand-side trimming processes were achieved by backward scanning from the rightmost nucleotide of the primitively searched

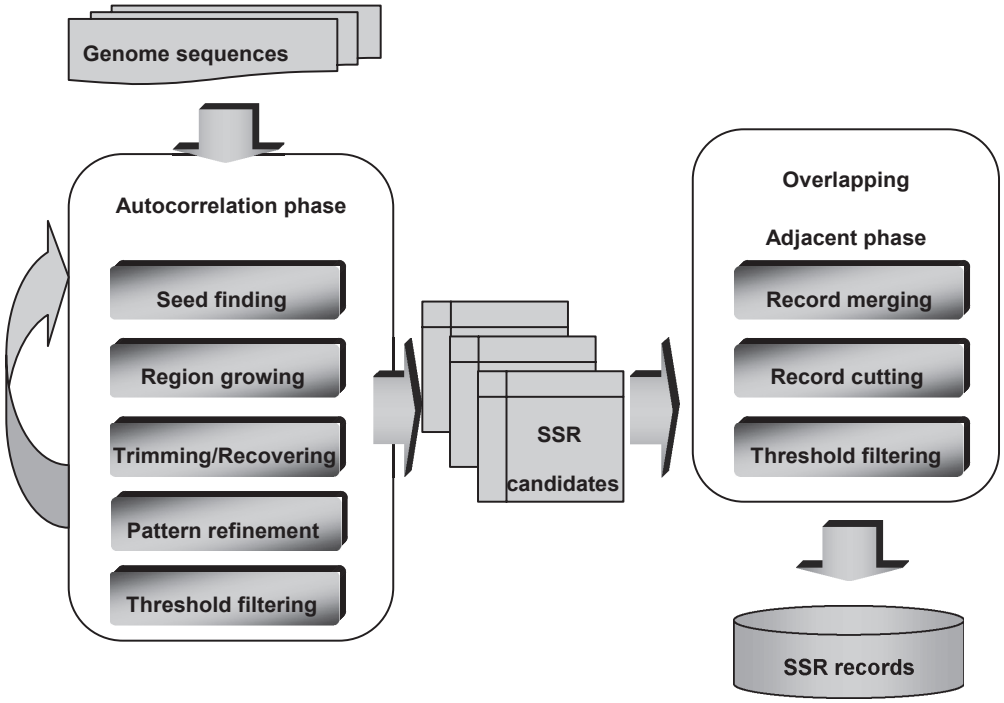


Figure 1 The flowchart of CG-SSR searching algorithm.

SSR to its last noninterrupted basic pattern. Once the last perfect basic pattern was found, the partial and noninterrupted patterns on its right-hand-side were recalled for its longest representation. For the left-hand-side of SSR string verification, this module examined only on the length of an assigned basic pattern which could not be recovered by insertion, deletion, or substitution examinations employed in the previous autocorrelation phase. In the module of pattern refinement, if a basic unit pattern is a repeated sequence itself, trivially, the pattern will be identified as SSRs within different unit lengths during autocorrelation processes. For example, a tri-nucleotide SSR string of length 60 could be identified as well as a hexa-nucleotide SSR string of length 30 from the first phase computation. Hence, for eliminating the redundant cases of this type, all self-repeating basic patterns were double checked and reduced to its smallest unit size.

After all frameshifts were completed, redundant segments were removed to eliminate possible overlapping patterns. The final noise filtering module was implemented to remove imperfect SSRs which contain impure variations higher than the defined threshold proportion. In the proposed system, two thresholding parameters of minimum length and maximum noise rate were decided in advance for verifying the qualification of a searched SSR record. The first thresholding value confirms the total length of a candidate SSR record which includes the repeats and the tolerant nucleotides, whereas the second thresholding parameter inspects the noise rate of the identified SSR records. The definition of noise rate is indicated as

$$noise\_rate = 1 - \frac{L_{patternLength} * N_{RepeatTime} + N_{IncompletePattern}}{L_{TotalLength}}$$

where $L_{PatternLength}$ represents the length of basic unit pattern, $N_{RepeatTime}$ the number of repeats of complete pattern (exact the same as the basic unit pattern), $N_{IncompletePattern}$ the number of nucleotides of incomplete patterns (partial subset of the basic unit pattern) in the SSR, and $L_{TotalLength}$ the total length of the identified SSR. Hence, a perfect SSR possesses "0" noise rate and higher noise rate represents more tolerant nucleotides appeared in the identified SSR strings. The default setting of noise rate for the initial database in CG-SSR is 0.2 which represents 20% of the total nucleotides recognized as the maximum noisy base pairs for each SSR segment.

For the second phase of overlapping adjacent verification, the SSR candidates were re-inspected based on their locations and basic unit patterns. Merging and cutting operations on neighboring SSR candidates enhanced the consecutive relationship and centralized the diverse representations. On the condition of two overlapping SSR candidates, the system firstly verified the overlapped records if they possessed

identical basic unit patterns. If two or more SSR records possessed an identical pattern, the recombination of such candidates were concatenated after evaluating the criterion of noise threshold requirements. On the other hand, if these overlapped SSR candidates of the same pattern length did not possess identical basic unit pattern, they usually resulted from the tolerant conditions. Hence, the system combined such two SSR candidates to make sure no entry was redundant. Finally, a threshold filter was applied again to satisfy the requirements of minimum SSR length and maximum noise rate. It is also true that two SSR candidates of different basic unit patterns and lengths might overlap. The ambiguity usually arose on the transistion positions of two consecutive basic patterns. In this system, two SSR strings of different lengths of basic unit patterns would be categorized and stored as two independent SSRs. Only two overlapping adjacent SSRs of the same basic pattern were merged for efficient and effective consideration.

When all of the SSR patterns in each genome were identified respectively, by comparing identical SSR patterns in the conserved regions among various species, the SSR patterns with high occurrence rates were considered as the candidates of important functional SSRs.

Results and discussion

In the retrieval processes, all perfect and imperfect SSRs were verified and annotated with accurate coordinate information of upstream, downstream, 5'UTR, 3'UTR, protein-coding, intron, and intergenic regions. Moreover, for identifying candidates of functional SSRs, the system integrated comparative genomics methods in which information from cross-species comparison was obtained from the UCSC Genome Browser[18] to provide coordinates for cross-species conserved regions. Currently, 24 species were included in the CG-SSR system for comparison. Through cross-species conservation, all conserved SSRs were identified and displayed in an explicit table. Furthermore, the biological function of each gene can refer to GO, InterPro, and Pfam resources for further comprehensive analyses. The developed CG-SSR web server is freely available at: http://cgssr.cs.ntou.edu.tw/. There are four major functions provided by the system: SSR Discovery, SSR on Transcripts, Comparative Genomics, and SSR Searching Tool.

SSR Discovery

It provides all loci of perfect and imperfect SSRs on a designated chromosome of a specified genome. Users can allocate precise locations of SSRs by selecting a target species, the number and range of chromosome, parameter of minimum length, and/or specific patterns of interested SSRs.

SSR on Transcripts

It collects all genes which possess perfect and imperfect SSRs found on their exon regions. Users can allocate precise locations of SSRs by selecting a target species, the number of chromosome, and the transcript IDs. (All transcript IDs were defined in Ensembl release 48, Dec. 2007).

Comparative Genomics

It provides a tool for searching all perfect and imperfect SSRs on conserved regions or orthologous genes among various species. From the query keywords or transcript IDs, the system provides precise and complete information of SSRs including coordinates, lengths, basic unit patterns, regions, flanking sequences, and corresponding primers.

SSR Searching Tool

It provides an online SSR searching tool for discovering all possible SSRs located in the uploaded multiple DNA sequences with respect to required parameter settings.

The detail guidelines of each subsystem can be obtained by clicking on the "tips" icon in each web interface. These guidelines provide helpful information to a user who is looking for a particular point of interest in SSR discovery. In conclusion, abundant resources for functional genomics, cross-species comparison, accurate coordinate annotation, flanking sequences and corresponding primers have been well integrated to provide applicable information for users. In this study, the numbers of retrieved imperfect and perfect SSRs ($\geq$10 base pairs with repeated unit lengths of 1–6 base pairs) by CG-SSR and the number of identifiable gene characteristics and protein families from several well known databases for eleven representative vertebrate species are listed in the Table 1. All detailed results are available online at the CG-SSR website. Figure 2 shows the statistical distributions of various repeated unit patterns for eleven model species. As can be seen from the diagram, similarly distributed proportions of different unit lengths of SSR patterns occur for all species and the most amounts of perfect and imperfect SSR repeats is the dinucleotide patterns with the default parameter settings of minimum length of 10 base pairs and noise rate of 20%.

To illustrate the practical applications of CG-SSR for identifying potential functional SSRs through cross-species comparison, several well known functional SSRs were collected and shown in Table 2. According to comparative

Table 1 The total number of verified SSRs in the CG-SSR database, and the number of identifiable gene characteristics and protein families from GO, InterPro, and Pfam databases for 11 representative vertebrate species

Species	Human	Chimpanzee	Orangutan	Rhesus	Cow	Dog	Mouse	Rat	Opossum	Medaka	Zebrafish
Items											
SSRs	30,364,358	29,092,001	28,722,387	27,768,093	22,157,544	26,435,804	27,814,234	26,195,617	38,281,261	5,963,611	15,060,006
Genes	55,183	37,006	24,231	40,431	27,194	29,275	43,620	37,591	32,908	22,447	31,922
GO records	226,591	30,208	25,102	26,828	49,510	92,880	232,824	142,334	20,973	2,564	49,227
InterPro records	109,750	75,180	54,405	81,741	63,294	61,523	95,177	79,520	89,516	50,423	81,063
Pfam records	56,531	39,782	28,390	42,304	33,534	31,788	49,207	40,762	46,012	30,816	41,756
Orthologous genes	182,722	166,630	160,263	173,377	160,224	168,371	198,505	193,363	181,554	144,039	157,329
Paralogous genes	87,397	59,186	51,748	123,096	71,592	72,070	161,852	175,790	74,944	0[#]	216,612
Comparative genomics species	23	11	7	6	3	5	22	12	0[$]	7	5
Conserved region records	18,666,679	11,567,566	5,981,764	7,929,321	3,558,513	8,239,233	17,373,326	12,208,685	0[$]	1,612,839	1,266,789

Notes: [#]Information of paralogous gene was not available from Ensembl Release 49, Mar. 2008; [$]Information of conserved region for Opossum was not available from UCSC, 2008.
Abbreviations: GO, gene ontology; SSRs, simple sequence repeats; UCSC, University of California, Santa Cruz.

genomics rules, functional elements are likely to be located in conserved regions. Notably, several functional SSRs discovered by our system were located in coding regions, UTR regions, and even in intron regions. The identification of functional SSRs in conserved regions of several species is evidence of their common features.

Taking *ATM* gene (ataxia-telengiectasia mutated gene) as an example (Ensembl transcript ID: ENST00000278616), 118 SSRs longer than 20 nucleotides were found. The comparative genomics mechanism provides an efficient way to select potential functional SSRs from among the 118 candidates. For instance, a T repeat exhibited a high degree of conservation among 17 species although it is located in an intron which is generally considered as a non-functional region. Interestingly, such a specific intronic mutation of T repeat has been reported to cause aberrant splicing and abnormal transcription in colon tumors.[19] Similarly, some CAG repeats in coding regions can be found and proved as functional SSRsin the *HD* (Ensembl transcript ID: ENST00000355072), *DRPLA* (ATN1, Ensembl transcript ID: ENST00000356654), *SCA1* (ATXN1, Ensembl transcript ID: ENST00000244769), *SCA2* (ATXN1, Ensembl transcript ID: ENST00000377611), *SCA3* (ATXN3, Ensembl transcript ID: ENST00000340660), and *SCA6* (CACNA1A, Ensembl transcript ID: ENST00000325084). To extract the highly conserved SSRs and verify as the potential functional motifs, users can input the Ensembl transcript ID directly or type the keywords in the query textbox under the "Comparative Genome" website. If an abbreviated keyword cannot be found, users can try the entire gene name to retrieve its corresponding transcript IDs from the specified gene set. Consequently, exactly and partially matched genes were listed in a table. Therefore, one can click on the retrieved transcript IDs and check/uncheck the checkbox of "intron" attribute before sending the query. According to the parameter settings, those retrieved SSRs will be listed in ascending order by chromosome number. To exploit potential functional SSRs of a gene, users are suggested to select SSRs in accordance with the number of conserved species. For example, from the first six genes listed in Table 2, the CAG repeats appeared in shifted or complementary patterns are considered as potential functional motifs because they were highly conserved in 12 to 22 species from CG-SSR. Indeed, these CAG repeat expansions in coding regions were demonstrated to bring about various neuronal diseases.[2] These findings suggest that cross-species comparison can be used to identify potential functional SSRs in both coding and noncoding regions.

After retrieving functional SSR candidates, CG-SSR provides functional genomics resources—GO, InterPro,

Table 2 Illustrations of practical applications of CG-SSR in identifying potential functional SSRs through cross-species comparison. Taking human species as an example, several well known functional SSRs could be retrieved and annotated[4]

Gene	EnsemblTranscript ID	SSR motif	Repeat length (bps)
HD	ENST00000355072	CTG(CAG)	26
ATN1 (*DRPLA*)	ENST00000356654	CAG	53
ATXN1 (*SCA1*)	ENST00000244769	GCA(CAG)	91
ATXN2 (*SCA2*)	ENST00000377611	CAG	71
ATXN3 (*SCA3*)	ENST00000340660	CAG	46
CACNA1A (*SCA6*)	ENST00000325084	CAG	40
AR (Androgen receptor)	ENST00000374690	AGC(CAG)	67
PABPN1 (Poly(A)-binding protein 2)	ENST00000397276	GCG	37
WISP2 (Signal transduction genes)	ENST00000396767	T(A)	22
CALM_HUMAN	ENST00000356978	AGC(CAG)	21
FMR1 (Fragile X mental retardation-1)	ENST00000370475	GCG(CGG)	67
AFF3 (*AF4/FMR2* family member 3)	ENST00000317233	GCG(GCC)	26
DMPK (dystrophia myotonin protein kinase)	ENST00000291270	CTG	62
EGFR (Epidermal growth factor receptor)	ENST00000275493	CA	51
ATM (Serine-protein kinase *ATM*)	ENST00000278616	T	22
ATXN10 (Spinocerebellar ataxia type SCA10)	ENST00000252934	AGAAT(ATTCT)	74
FXN (Frataxin, mitochondrial precursor Friedreich ataxia)	ENST00000377270	CTT(GAA)	18

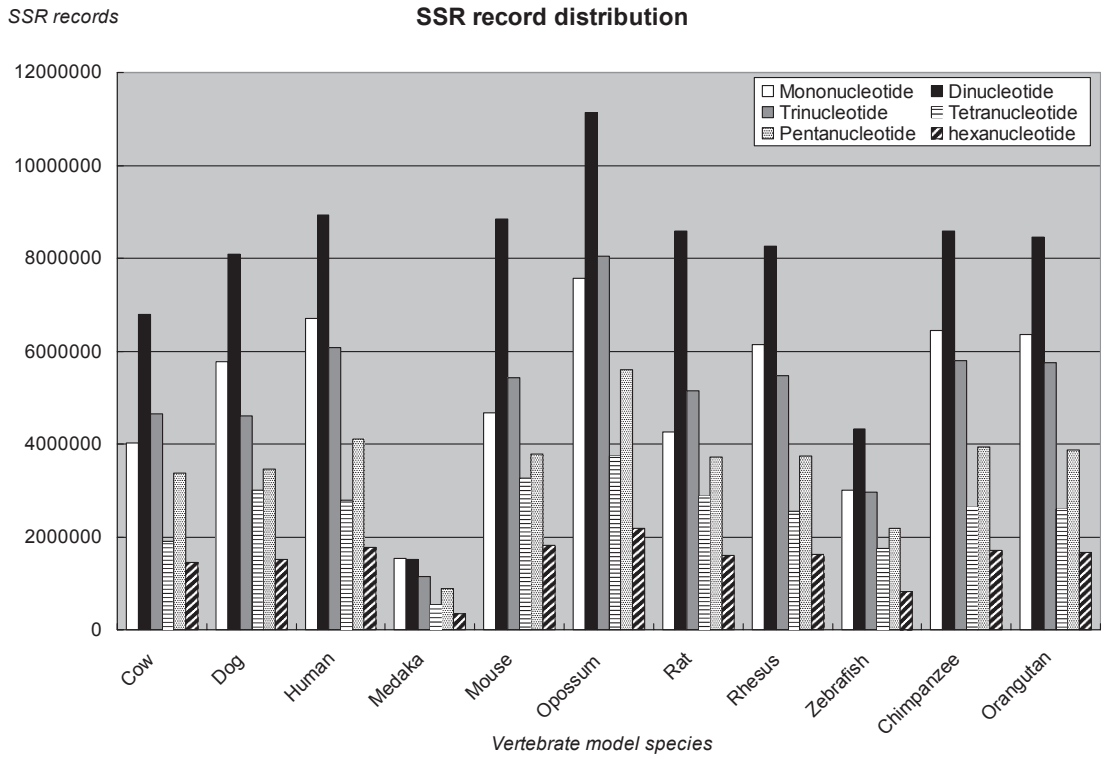


Figure 2 Distributions of various repeated unit patterns from mononucleotide to hexanucleotide for 11 vertebrate model species. The number of SSR records was identified based on the parameters of a minimum length of 10 base pairs and a maximum noise rate of 20%.

Table 2 (*Continued*)

Region	# of Conserved species	SSR related biological function	References
Coding	22	Expansion causes Huntington's disease (*HD*)	Zoghbi and Orr (2000)[20]
Coding	21	Causes dentatorubropallidoluysian atrophy (*DRPLA*)	Nakamura and colleagues (2001)[21]
Coding	20	Causes spinocerebellar ataxias	Manto (2005)[22]
Coding	13	Causes spinocerebellar ataxias	Manto (2005)[22]
Coding	12	Causes spinocerebellar ataxias	Manto (2005)[22]
Coding	15	Causes spinocerebellar ataxias	Manto (2005)[22]
Coding	11	Shorter repeat increases hepatitis B virus (*HBV*) – related hepatocellular carcinoma risk	Yu and colleagues (2001; 2002)[23,24]
Coding	14	Oculopharyngeal muscular dystrophy	Brais and colleagues (1998)[25]
Coding	8	Tumor-suppressive function	Markowitz and colleagues (1995)[26]
5'UTR	15	Required for *hCALM1* full expression,	Toutenhoofd and colleagues (1998)[27]
5'UTR	11	(CGG)40–200 related in fragile-X-like cognitive/ psychosocial impairment	Franke and colleagues (1998)[28]
5'UTR	8	Reduced *FMR2* causing abnormal neuronal gene regulation	Cummings and Zoghbi (2000)[29]
3'UTR	13	Expansion causes *DM1* disease	Ranum and Day (2002)[6]
Intron	10	CA repeat enhances *egfr* transcription and involved in breast carcinogenesis	Tidow and colleagues (2003)[7]
Intron	17	Shortening repeat tract leads to aberrant splicing and abnormal transcription in colon tumor cells	Ejima and colleagues (2000)[19]
Intron	8	Expansion leads to change of function and results in *SCA10* disease	Matsuura and colleagues (2000)[30]
Intron	7	GAA expansion inhibits *FRDA* expression or interferes mRNA formation and lead to *FRDA* disease	Ohshima and colleagues (1998)[31] Sakamoto and colleagues (2001)[32]

and Pfam—to help users ascertain the potential biological function of each SSR. Using the involvement of *WISP2* in signal transduction as an example (Ensembl transcript ID: ENST00000396767), information provided by GO implicates *WISP2* in cell growth. Interestingly, experimental verification suggests that this A repeat indeed has tumor suppressor function.[4] Users can efficiently use these functional genomics resources to obtain clues about putative roles of SSRs in gene regulatory networks.

In summary, CG-SSR comprises accurate coordinate, cross-species comparison, and functional genomics resources. It is a comprehensive, efficient and user-friendly online tool for identifying conserved SSRs as potential functional motifs in vertebrates.

Acknowledgments

This work was supported by the Center for Marine Bioscience and Biotechnology (CMBB) at the National Taiwan Ocean University, Keelung, Taiwan (97529002H1), and the National Science Council in Taiwan, R. O. C. (NSC96-2627-B-019-003 to T.-W. Pai).

Disclosure

This SSR is free for use and can be located at the following URL: http://cgssr.cs.ntou.edu.tw/. Supplementary information is available at the following URL: http://cgssr.cs.ntou.edu.tw/si_v2/. The authors report no conflicts of interest in this work.

References

1. Charlesworth B, Sniegowaski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*. 1994;371:215–220.
2. Fondon JW 3rd, Hammock EA, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci.* 2008;31:328–334.
3. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 2006;22:253–259.
4. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol.* 2004;21:991–1007.
5. Kenneson A, Zhang F, Hagedorn CH, Warren ST. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Hum Mol Genet.* 2001;10:1449–1454.
6. Ranum LPW, Day JW. Dominantly inherited, noncoding microsatellite expansion disorders. *Curr Opin Genet Dev*. 2002;12:266–271.
7. Tidow N, Boecker A, Schmidt H, et al. Distinct amplification of an untranslated regulatory sequence in the *egfr* gene contributes to early steps in breast cancer development. *Cancer Res.* 2003;63:1172–1178.

8. Sakai T, Miura I, Yamada-Ishibashi S, et al. Update of mouse microsatellite database of Japan (MMDBJ), *Exp Anim*. 2004;53: 151–154.
9. Missirlis PI, Mead CL, Butland SL, et al. Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics*. 2005;6:145.
10. Subramanian S, Madgula VM, George R, et al. MRD: a microsatellite repeats database for prokaryotic and eukaryotic genomes. *Genome Biol.* 2002;3(12):PREPRINT0011.
11. Subramanian S, Madgula VM, George R, Kumar S, Pandit MW, Singh L. SSRD: simple sequence repeats database of the human genome. *Comp Funct Genomics.* 2003;4:342–345.
12. Aishwarya V, Grover A, Sharma PC. EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*. 2007;8:225.
13. Margulies EH, Birney E. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet.* 2008;9:303–313.
14. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25:25–29.
15. Mulder NJ, Apweiler R, Attwood TK, et al. New developments in the InterPro database. *Nucleic Acids Res.* 2007;35:D224–D228.
16. Finn RD, Mistry J, Schuster-Böckler B, et al. Pfam: clans, web tools and services. *Nucleic Acids Res.* 2006;34:D247–D251.
17. Flicek P, Aken BL, Beal K, et al. Ensembl 2008. *Nucleic Acids Res.* 2008;36:D707–D714.
18. Karolchik D, Kuhn RM, Baertsch R, et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* 2008;36:D773–D779.
19. Ejima Y, Yang L, Sasaki MS. Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumor cell lines: a novel mutation target of microsatellite instability. *Int J Cancer*. 2000;86:262–268.
20. Zoghbi HY, Orr HT. Glutamine repeats and neurodegeneration. *Annu Rev Neurosci.* 2000;23:217–237.
21. Nakamura K, Jeong SY, Uchihara T, et al. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum Mol Genet*. 2001;10:1441–1448.
22. Manto MU. The wide spectrum of spinocerebellar ataxias (SCAs). *Cerebellum*. 2005;4:2–6.
23. Yu MW, Yang YC, Yang SY, et al. Hormonal markers and hepatitis B virus–related hepatocellular carcinoma risk: a nested casecontrol study among men. *J Natl Cancer Inst.* 2001;93:1644–1651.
24. Yu MW, Yang YC, Yang SY, et al. Androgen receptor exon 1 CAG repeat length and risk of hepatocellular carcinoma in women. *Hepatology*. 2002;36:156–163.
25. Brais B, Bouchard JP, Xie YG, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet.* 1998;18:164–167.
26. Markowitz S, Wang J, Myeroff L, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*. 1995;268:1336–1338.
27. Toutenhoofd SL, Garcia F, Zacharias DA, Wilson RA, Strehler EE. Minimum CAG repeat in the human calmodulin-1 gene 5′ untranslated region is required for full expression *Biochim Biophys Acta*. 1998;1398:315–320.
28. Franke P, Leboyer M, Gänsicke M, et al. Genotype-phenotype relationship in female carriers of the premutation and full mutation of FMR-1. *Psychiatry Res.* 1998;80:113–127.
29. Cummings CJ, Zoghbi HY. Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet.* 2000;1:281–328.
30. Matsuura T, Yamagata T, Burgess DL, et al. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet*. 2000;26:191–194.
31. Ohshima K, Montermini L, Wells RD, Pandolfo M. Inhibitory effects of expanded GAA.TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication in vivo. *J Biol Chem.* 1998;273:14588–14595.
32. Sakamoto N, Ohshima K, Montermini L, Pandolfo M, Wells RD. Sticky DNA, a self-associated complex formed at long GAA*TTC repeats in intron 1 of the frataxin gene, inhibits transcription. *J Biol Chem.* 2001;276:27171–27177.

Supplementary information
How does CG-SSR searching algorithm work?

Definition of SSR

Simple sequence repeats (SSRs), also called microsatellites, are nucleotide segments with basic repeat pattern of 1–6 base pairs in length. The searching algorithm of CG-SSR is designed for discovering all perfect and imperfect (with tolerant) SSRs from various genomic sequences efficiently and effectively.

Algorithm description

Figure 1 depicts the flowchart of CG-SSR searching algorithms which is composed of two major phases: (I) autocorrelation and (II) overlapping adjacent phases. The first autocorrelation phase including seed finding, region growing, trimming/recovering, refining, and threshold filtering processes discovers all possible SSR patterns as fundamental candidates. The second overlapping adjacent phase including record merging, record cutting, and threshold filtering processes verifies overlapped segments and confirms no redundant perfect/imperfect SSR patterns. Details of each procedure of these two phases are described in the following sections:

Autocorrelation phase

Autocorrelation seed finding

The core algorithm for discovering SSR patterns from genome sequences employs autocorrelation methodology. The basic concept assumes that the target sequence contains repeat substrings with basic patterns of length N within a range from 1 to 6. If we shift the target sequence with N nucleotides to its right and compare to its original the original sequence, all repeat patterns can be discovered since they overlapped with their shifted sequence at least within N nucleotides continuously. Based on the observation of such overlapping, we can easily detect the repeating locations by shifting and matching the whole target sequence without knowing the nucleotide contents of repeated patterns. If the continuously matched nucleotides are longer than the shifting length N, at least one repeat with the basic pattern length N is proved to exist in the target sequence. Therefore, the context of the basic pattern can be extracted directly from the target sequence where the repeats appear. The program performs frame shifting processes from length 1 to 6 base pairs to identify mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSRs, respectively. Figure 2 illustrates an example

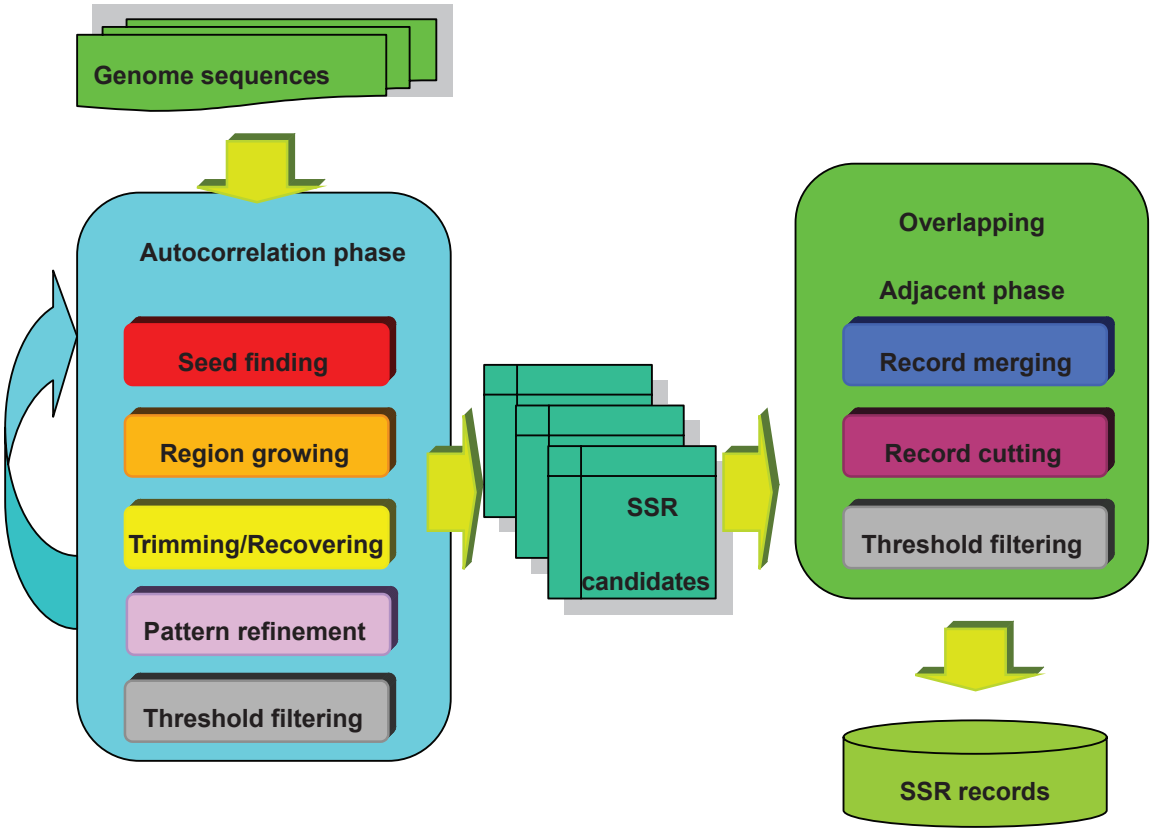


Figure 1 The flowchart of CG-SSR searching algorithm.

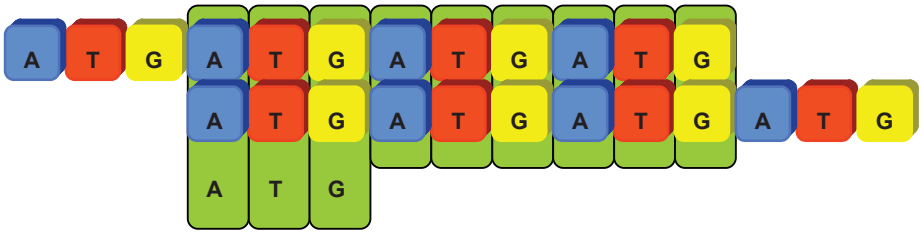


Figure 2 An example of autocorrelation process.

of autocorrelation processes for identifying seeds of repeat patterns of length N = 3.

The number of continuously matched nucleotides reveals the exact length of perfect SSR repeats. However, these discontinuous perfect patterns were interrupted due to various tolerant conditions, therefore categorized as imperfect SSRs. In order to identify both perfect and imperfect SSRs, the proposed algorithms firstly considered the coordinates of perfect repeats and employed the location information as input for the following modules.

Neighboring comparison from region growing

To identify imperfect SSRs with various tolerant conditions, the proposed algorithm applied forward region growing method to enable neighboring comparison. As long as the coordinates and contents of the perfect repeat patterns were obtained, the forward region growing processes were performed by examining the right-hand-side neighboring nucleotides and skipping noise-like nucleotides which did not belong to the perfect repeat patterns. The searching processes stopped until no extra basic unit pattern could be found by continuously extending the verification on right-hand-side neighboring nucleotides.

There are three types of sequence tolerance for imperfect SSRs shown in Figure 3: insertion, deletion, and substitution. The insertion case is categorized into two different types: the inserted nucleotides located between two basic unit patterns or appeared inside a basic unit pattern. Deletions and substitutions, apparently recognized by analyzing the length and

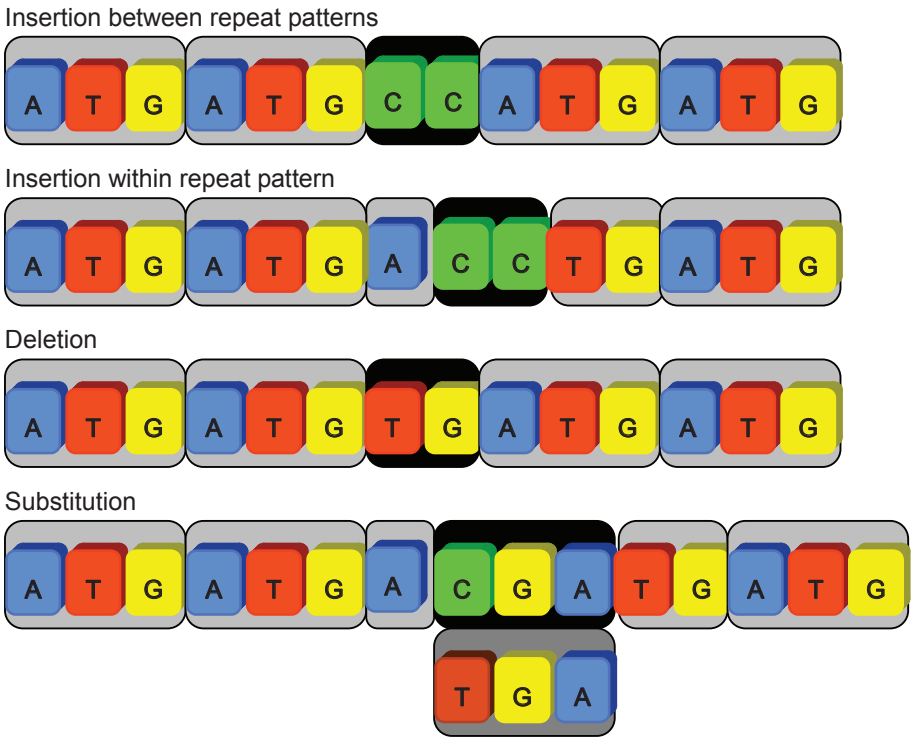


Figure 3 All tolerant cases can be considered as special cases of insertions. The substitution and deletion cases were viewed as inserting tolerant segments.

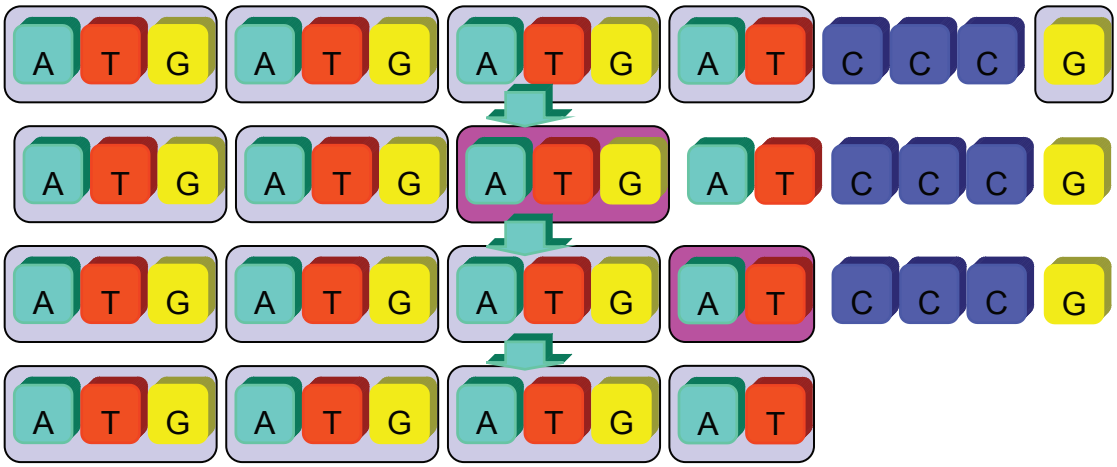


Figure 4a The right-hand-side trimming process verifies the right-hand-side of SSR to guarantee a complete, noninterrupted representation.

contents of insertion based on string comparison, can be considered as the insertion in different forms.

Right trimming and left recovering processes

In this module, verifications on both sides include left-hand-side trimming and right-hand-side recovering processes. Both processes are described as follows:

Right-hand-side trimming

Once an SSR sequence was allocated from seed finding and forward region growing, a trimming process was applied to examine its right-hand-side of extended SSR records for a guaranteed representation of noninterrupted pattern. This process was achieved by backward scanning from the right end of the previously searched SSR to its last noninterrupted basic pattern. Once the last perfect basic pattern was found, the following, partial, and noninterrupted patterns on its right side were recalled for its longest representation, such as the "AT" substring on the right end of the designated SSR in Figure 4a.

Left-hand-side recovering:

For the left-hand-side of SSR string verification, the system examined only the length of an assigned basic pattern which could not be recovered by insertion, deletion or substitution examinations employed in the previous autocorrelation phase. An example of left-hand-side recovering process was shown in Figure 4b. The "ATG" pattern in the leftmost position could not be recognized through the first autocorrelation module, but it could be identified after the verification of left-hand-side recovering process.

Threshold filtering

Two thresholding parameters were applied to verify if an SSR record was qualified for the requirements. One was the parameter of minimum length and the other was the maximum noise rate. The first thresholding filter checked the total

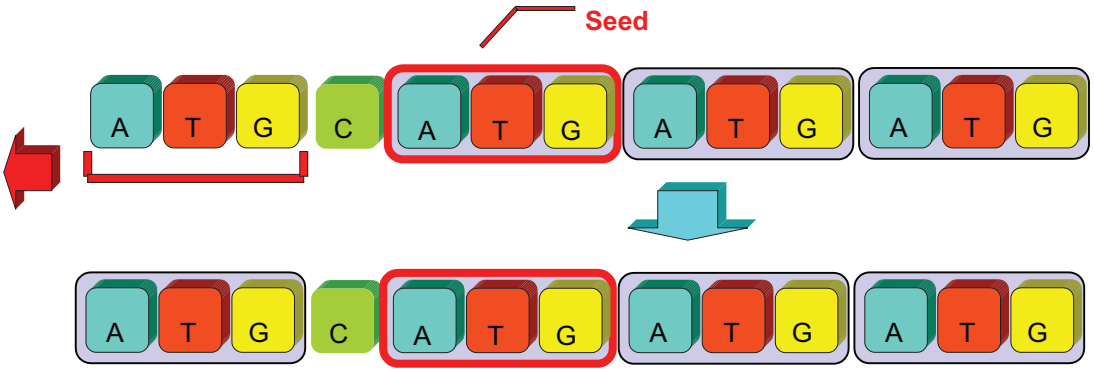
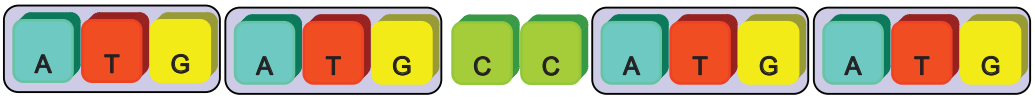


Figure 4b An example of left-hand-side recovering processes.



$$noise\_rate = 1 - \frac{3*4}{14} = 0.143$$

$$noise\_rate = 1 - \frac{3*3+2}{13} = 0.154$$

Figure 5 Examples of noise rate calculation.

length of a candidate SSR record which included the repeat contents and the tolerant nucleotides. The default argument in the system was 10 nucleotides.

The second thresholding parameter inspected the noise rate of an identified SSR record. The definition of noise rate is

$$noise\_rate = 1 - \frac{L_{patternLength} * N_{RepeatTime} + N_{IncompletePattern}}{L_{TotalLength}}$$

where $L_{PatternLength}$ represents the length of basic pattern; $N_{RepeatTime}$ represents the number of repeats of complete pattern; $N_{IncompletePattern}$ denotes the number of nucleotides of incomplete patterns in the SSR: $L_{TotalLength}$ is the total length of an identified SSR.

A perfect repeat SSR possesses "0" noise rate, and higher noise rate represents more tolerant nucleotides appeared in an identified SSR strings. The default setting is 0.2 which represents 20% of the total nucleotides recognized as noisy base pairs for the selected SSR string. Examples of noise rate calculation are shown in Figure 5.

Pattern refinement

If a basic unit pattern is a repeat sequence itself, trivially, the pattern will be identified as SSRs with different lengths during autocorrelation processes. For example, a tri-nucleotide SSR string of length 60 could be identified as well as a hexa-nucleotide SSR string of length 30 from the first phase computation. Hence, for eliminating the redundant cases of this type, all self-repeating basic patterns were double checked and reduced to its smallest unit size. In Figure 6, an example of self-repeating SSR was shown and the basic unit pattern "TG" was considered as the fundamental SSR pattern with the smallest unit size for its final representation. Therefore, the SSR string was categorized as a di-nucleotide SSR record instead of a tetra-nucleotide or a hexa-nucleotide.

To achieve robust performance of CG-SSR system, the program skipped the mono-nucleotide SSR seed-finding module. Frameshifting of one nucleotide caused overwhelming false alarm results and wasted too much time on following evaluation processes. However, the mononucleotide SSR could be retrieved by performing from frameshifting of two to six nucleotides and verified by self-repeating analysis. The mononucleotide SSRs could be successfully and efficiently identified after this refinement processes.

It was noticedthat a shifting operation applied on an identified SSR sequence formulated another new repeat sequence within a different basic unit pattern. It is shown in Figure 7

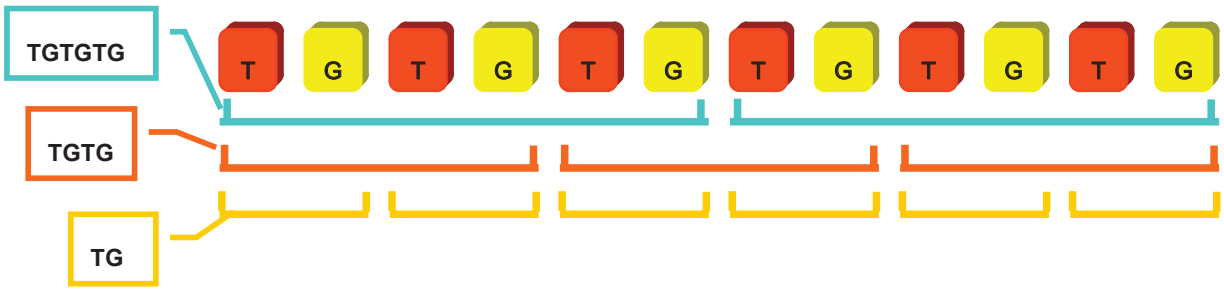


Figure 6 Self-repeating basic pattern was verified in the proposed system.

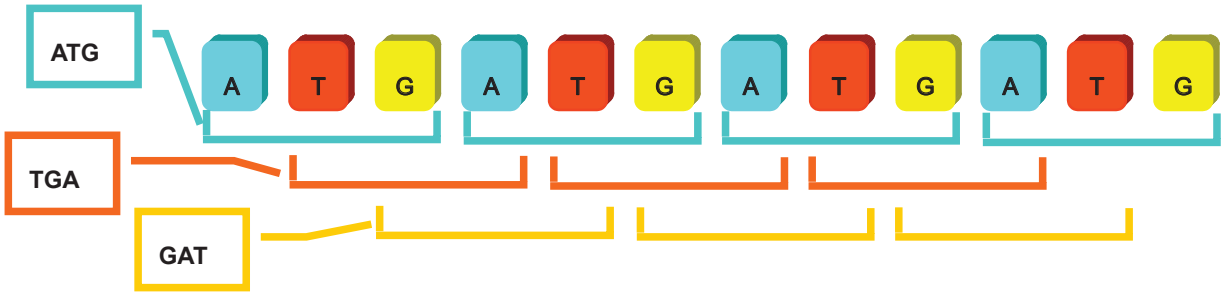


Figure 7 Different basic unit patterns due to shifting.

as an example, "ATG", "TGA", and "GAT" were different basic unit patterns for three SSR sequences. However, it should be defined that these three basic unit patterns were considered as an identical pattern as long as their contexts can be exactly matched after shifting operations.

Overlapping adjacent phase

In this phase, the SSR candidates were re-inspected based on their locations and basic unit patterns. Merging and cutting operations on neighboring SSR candidates enhance the consecutive relationship and centralize the diverse representations. The proposed filtering operations are described as follows:

Record merging and cutting

On the condition of two overlapping SSR candidates, the system firstly verified the overlapped records if they possessed identical basic unit patterns. If two or more SSR records possessed an identical pattern, the recombination of such candidates were concatenated after evaluating the criterion of noise threshold requirements. On the other hand, if these overlapped SSR candidates of the same pattern length did not possess identical basic unit pattern, they usually resulted from the tolerant conditions. Hence, the system combined such two SSR candidates to make sure no entry was redundant. Finally, a threshold filter was applied again to satisfy the requirements of minimum SSR length and maximum noise rate.

It is also true that two SSR candidates of different basic unit patterns and lengths might overlap. The ambiguity usually arose on the transistion positions of two consecutive basic patterns. In Figure 8, an example of SSR transistion from a tetra-nucleotide SSR to a tri-nucleotide SSR was shown, and the concatenation resulted from the tolerant conditions between two SSR strings. In such system, two SSR strings of different lengths of basic unit patterns would be categorized and stored as two independent SSRs. Only two overlapping adjacent SSRs of the same basic pattern were merged for efficient and effective consideration.

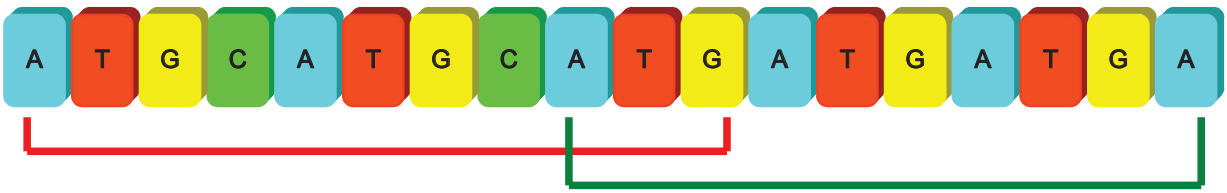


Figure 8 Ambiguity between two basic unit patterns within two different lengths.