The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Identify SSR Regulators for Functional Gene Sets through Cross-Species Comparison

Chien-Ming Chen[1] Meng-Chang Hsiao[1] Tun-Wen Pai[1, *]
Ronshan Cheng[2] Wen-Shyong Tzou[3]
Margaret Dah-Tsyr Chang[4]

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, No. 2, Pei-Ning Rd., Keelung, Taiwan 20224, China.
[2]Department of Aquaculture, National Taiwan Ocean University, No. 2, Pei-Ning Rd., Keelung, Taiwan 20224, China.
[3]Institute of Bioscience and Biotechnology, National Taiwan Ocean University, No. 2, Pei-Ning Rd., Keelung, Taiwan 20224, China.
[4]Institute of Molecular and Cellular Biology & Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan 30013, China.

Abstract Single sequence repeats (SSRs) are DNA sequences composed of tandem repetitions of relatively short motifs. They are not only considered as genetic markers but also play an important role in gene regulatory networks, One of the greatest challenges of functional genomics. In order to identify key SSR regulators among functional gene sets, we have developed an efficient algorithm for SSR pattern retrieval and a verification mechanism based on cross-species comparison and statistical analysis. Cross-species comparison and orthologous relationship provide an added level of validation for identifying multiple genes that are likely to be regulated similarly. In addition, statistical analysis of appearance frequency rates confirms the retrieved SSRs as significant regulators from a given set of related genes. In this study, the target gene set with growth factor related genes was evaluated and several well known "CA", "CCG" and "CAG" repeat pattern were successfully identified by our proposed system. Accordingly, the novel pattern mining methods proposed to analyze large-scale genome datasets are successfully achieved for all organisms and genes of interest, and the proposed mechanism can be applied to retrieve important candidates of SSR regulators for further biological experiments.

Keywords simple sequence repeat; microsatellite; functional significance score (*FSS*); orthologous conservation score (*OCS*); gene regulatory networks; regulator

1 Introduction

Simple sequence repeats (SSRs, also called microsatellites or minisatellites) are mutation-prone DNA composed of tandem repetitions of relatively short motifs. These repeat patterns frequently appeared in both eukaryotes and prokaryotes, and can be easily observed in DNA sequences with fundamental patterns ranging from one to six base pairs in length

*twp@mail.ntou.edu.tw

[1]. For decades, SSRs were considered as genetic markers in DNA fingerprinting and diversity studies because of high polymorphism [2]. Nevertheless, recent studies pointed out that repeat-number variation affects various gene functions such as transcript stability, rates of protein folding and turnover, protein-protein interaction, aggregation and transcription rate [3]. With the development of genetic markers, single nucleotide polymorphism (SNP) is considered to replace SSR for some applications such as genome mapping [4]. Therefore, biologists emphasize more on comprehensive understanding of SSRs' biological function.

Recent studies have shown that the influence of SSRs on gene regulatory networks generally rely on the total number of repeats, motif length, flanking sequences and inclusion of variant motifs [3]. However, these properties are not sufficient to judge whether an SSR indeed possesses a biological function among millions of 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 the identification of candidate regulatory elements [5]. If these SSRs possess significant biological functions, it can be assumed that they stand a good chance of being located in conserved regions. In this study, we developed computational methods for identifying putative functional SSRs through cross-species comparison and for evaluating the statistical significance of their prevalence in a given set of functional related genes. Through the developed mechanism for statistical verification, we could expect to discover candidate SSRs as the key regulators of a specific gene sets among various species

2 Material and Method

To identify potential key SSR patterns located in the same region of orthologous genes among various species, an efficient methodology should be carefully designed for on-line analysis services. In this study, all possible SSR patterns located in different regions for all species were pre-calculated and stored in a database. To retrieve the perfect and imperfect SSRs, each chromosome in the original dataset is scanned by an auto-correlation method and a set of serial processes for precise annotation. These sequential processes include frame shifting and matching, region growing, boundary detection, boundary trimming, redundant pattern removing, and noise filtering modules. The lengths of frame shift were defined to be the same as basic unit lengths of SSRs, from mononucleotide to hexanucleotide. If the repeat condition in a sequence existed, the repeated patterns would be exactly or partially overlapped by its neighboring segments during shifting processes. By comparing the shifted sequence with the original one, the loci of repeating parts were detected. Shifting processes within 1 to 6 nucleotides were performed individually for all various genomes to primarily extract perfect SSR candidates. The exactly matched patterns were extended by region growing techniques for formulating imperfect SSR patterns, and a boundary detection module was accomplished by considering noisy nucleotides which do not belong to the repeat patterns. Once the fundamental SSR patterns were identified, a trimming operation was applied to delimit the range of each SSR by verifying the feature of compactness at both end sides. Furthermore, a splitting process for segmentation was employed to consider the cases of two overlapped SSR patterns which were usually caused by concatenating conditions of two similar repeat patterns. At

the last step of SSR searching algorithm, a noise filtering module was designed to remove SSR candidates which possess percentages of imperfect repeat contents higher than the default proportional thresholding.

All possible patterns of SSRs located in various regions of a gene within different lengths are identified by previously described searching algorithms. These genes and genome sequences are provided by the Ensembl archive version 48 released in December 2007 (http://www.ensembl.org/index.html). When a gene possessing more than one transcript IDs, only the ID with the smallest index number is selected and analyzed for its SSR distribution. In fact, all retrieved SSRs located in the 6 various regions in a gene including coding, Intron, 5' UTR, 3' UTR, upstream, and downstream regions are considered and annotated in the proposed system. The upstream and downstream regions are defined as the regions which are located in front of the first exon and at the rear of the last exon, according to the transcript direction, respectively. The 5' UTR region is a particular section in messenger RNA that is the leader sequence and ends just before the start codon of the coding region, and the 3' UTR region is the untranslated region in messenger RNA that follows after the coding region. Detail information of coordinates of all defined regions was also obtained from Ensembl website.

There are in total 501 distinct SSR repeat patterns from mononucleotide to hexanucleotide. In this paper, all possible SSRs in each genome are searched and the positional attributes of each SSR is annotated according to their coordinate positions. A binary counting mechanism is designed to represent the appearance condition of each pattern. If at least one SSR pattern appeared in a defined region, a bit representation of the specified pattern in this region will be set as 1, otherwise it will be set as 0. All 501 SSR patterns distributed in various regions of genes for different species are pre-analyzed. After performing SSR pattern searching and existence evaluation in each region of a gene, we can obtain a bitwise matrix within a size of 501*6. It shows the appearance condition of all SSR patterns of a gene which has been dispersed in different regions. Each element in the matrix is denoted by either 1 or 0 and called as an "SSR Distribution Matrix" of a gene. Clearly, the bitwise matrix can efficiently reflect the condition of existence for each distinct SSR pattern. There is another matrix called "SSR Background Matrix" which is obtained by taking the summation of all SSR distribution matrices from all genes of a specified species. The dimension of an SSR Background Matrix is 501*6 as well and the number in each element ranges from 0 to the gene number of the selected species.

To obtain the SSR Distribution and Background Matrices, two parameters, ranges of upstream/downstream regions and minimum length of SSRs, should be determined in advance. Firstly, different range of the upstream and downstream sequences directly affects the existing condition of SSR patterns. In this study, we set 2,000 base pairs as default setting for both upstream and downstream regions. Secondly, a minimal requirement of length for SSRs also directly reflects the final searched SSR records. The default setting of this parameter was 20 in this study which guarantied at least three repeat units for a hexanucleotide SSR, four repeat units for a pentanucleotide, and five repeat units for a tetranucleotide SSR, and *et cetera.*

In order to define important SSR patterns from a selected target gene set, two statistical indices are proposed in this paper: functional significance score (*FSS*) and orthologous conservation score (*OCS*). The *FSS* of an SSR represents the relative ratio of the appearance rate of the SSR located in the same region of a target gene set with respect to

the appearance rate of the SSR located in the same region of whole gene sequences in a genome. In another words, we first count the hit gene number of the specified functional consensus SSR from the target gene set with respect to the total number of target genes (*Fc %* = number of gene with functional consensus SSR in the target gene set / number of total genes in the target gene set). Then, we compute the percentage of hit gene number of the functional consensus SSRs in a genome (i.e. the SSR Background Matrix) with respect to the total number of genes in the specified genome (*Fb%* = number of gene with functional consensus SSR in the genome/ number of total genes in the whole genome). Therefore, the *FSS* of an SSR located in a specific region is equal to the value of *Fc %* divided by *Fb%*. A higher *FSS* score represents the larger significant representation of the searched SSR. In this paper, we only focus on the SSR patterns with *FSS* score higher than 2. To ensure the significance of an SSR, the quantity of SSR hit number in the target gene set is sorted from high to low in order, and the ranked first three functional consensus SSR patterns with high *FSS* values will be selected and highlighted to show their significance.

The second way to enhance the importance of SSRs is to consider the orthologous conservation relationship which is approached by calculating the orthologous conservation score (*OCS*) among the target and query genomes. If we define an orthologous conserved SSR as a repeat pattern which is located in the same region of orthologous genes in both target and query genomes and constrained within the specified functional gene sets, then the *OCS* is obtained by taking *Oc %* divided by *Ob%*, where *Oc %* is the percentage of the hit number of an orthologous conserved SSR within the orthologous target gene set with respect to the total number of orthologous target genes, and *Ob%* is the percentage of hit number of the orthologous conserved SSR with respect to the total number of orthologous genes from both target and query genomes. A higher OCS score represents higher significance of the identified orthologous conserved SSRs in statistical sense. In this paper, the threshold setting of 2 was applied to retrieve the important orthologous conserved SSRs. Similarly, only the first three orthologous conserved SSRs were selected based on the sorted SSR hit numbers. Figure 1 shows the data flow diagram of the proposed system in which all possible SSRs located in various regions of genes of selected genomes were pre-analyzed, and genes with functional consensus SSRs or orthologous conserved SSRs distributed in a specified functional gene set or orthologous gene set were also analyzed for comparison.

3 Result and Discussion

Based on the information of orthologous relationship, cross-species comparison, and statistical analysis on appearance frequencies, several key SSR regulators for growth factor gene set (GO: 0008083) were retrieved and shown in Table 1. These four SSR regulators have a higher occurrence than what is expected by chance alone by consulting the cross-species comparison results of human versus rhesus and human versus mouse; meanwhile, the patterns were in the form of mononucleotide (A) or dinucleotide (CA), trinucleotide (CAG) and trinucleotide (CCG) in promoter, coding sequence and 5' UTR, respectively. These SSRs could be taken as potential key regulators among growth factor activity gene regulatory networks because they showed statistical significance and high evolutional conservation.

According to recent reports, SSR expansions and/or contractions in protein-coding

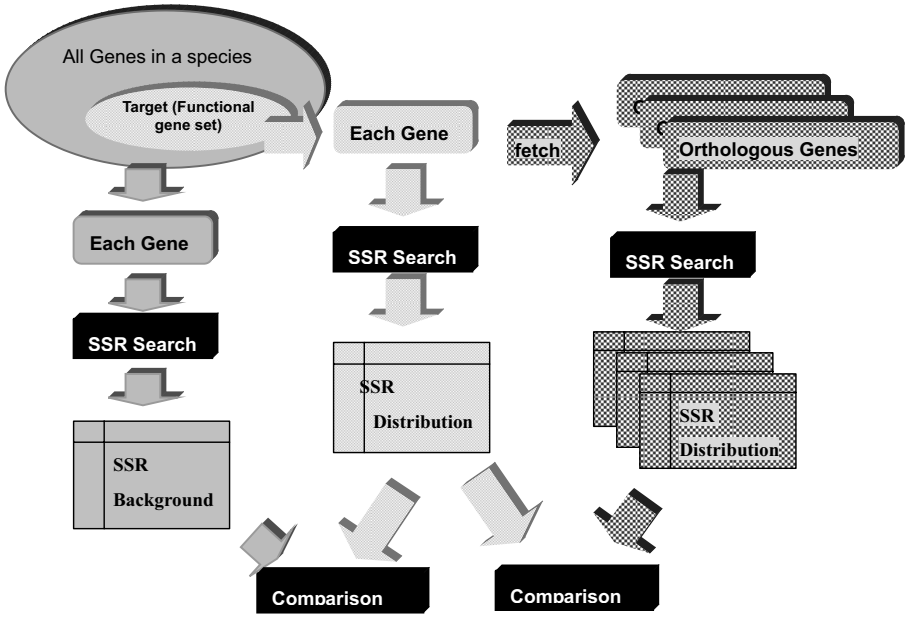

Figure 1: The data flow diagram of functional SSR identification system.

regions were observed to bring about a gain or loss of gene function through frame-shift mutation [2]. Furthermore, trinucleotide repeats located in coding region were associated with various types of cancer and implicated in various neurodegenerative disorders. These studies support the idea that the enriched SSR—trinucleotide (CAG)—in coding region acts as a key regulator since growth factor activity indeed plays a crucial role in cancer biological processes. Moreover, SSRs located in promoter regions have been shown obvious influence on the transcription activity by serving as *cis*-regulatory elements [7]. Interestingly, the enriched SSR pattern, dinucleotide (CA), was identified from this study as well and the variation of such pattern indicated the great impact on the variation of mammal size [8-13]. In addition, the insulin-like growth factor 1 (IGF1) gene is a highly conserved polypeptide which regulates growth and metabolic functions in several mammal species [10-15]. In recent studies, alleles of a dinucleotide (CA) SSR appeared within different frequencies and located at the promoter regions of the IGF1 gene for variant sizes of various species are unveiled, such as human, cattle, pig, dog and horse. Significant association of the allelic form of the microsatellites with adult body size has been clearly demonstrated[10-16]. For demonstration of the precise locations of the conserved SSRs among various species, the dinucleotide (CA) SSRs located in the IGF1 promoter regions are shown in Figure 2, where the "TG" and "CA" patterns are represented as the same SSR in forward and reverse strands of DNA sequences. Due to high evolutionary conservation, the enriched (CA) SSR appears to act a significant regulator in growth factor activity regulatory networks. Recent studies also indicate that SSR variations in 5'-UTRs could affect gene transcription and translation. According to our analysis, enriched trinucleotide (CCG) is a key factor for neurodevelopmental disorders. On the other hand, SSR

Table 1: Functionally significant and orthologous conserved SSR regulators with respect to human growth factor activity gene set (GO: 0008083). Same patterns within same regions were verified through comparison and the selected key SSRs were highlighted in boldface representation. Both satisfying the threshold values of *FSS*/*OCS* and the appearance number of genes within the target set ranked in the first three positions were selected to shown in this table.

SSR pattern	Region	Functional Significance Score (*FSS*)	Rank of SSR Hit Number in Target Sets	Orthologous Conservation Score(*OCS*)	Rank of SSR Hit Number in Orthologous Conserved Target Sets
Human *vs.* Rhesus					
AC (CA)	**Upstream**	**1.54**	**2**	**2.75**	**2**
C	**Upstream**	**2.25**	**3**	**2.97**	**3**
C	5' UTR	3.28	2	5.65	1
A	Coding	4.90	1	8.58	1
CAG	**Coding**	**3.16**	**1**	**2.83**	**2**
CCG	**5' UTR**	**3.98**	**1**	**1.79**	**1**
Human *vs.* Mouse					
AC (CA)	**Upstream**	**1.54**	**2**	**2.4**	**1**
C	**Upstream**	**2.25**	**3**	**4.13**	**3**
AG	5' UTR	5.94	2	10.54	2
CCG	**5' UTR**	**3.98**	**1**	**5.27**	**1**
CAG	**Coding**	**3.16**	**1**	**2.05**	**1**
AGG	Coding	2.86	3	1.53	3
AAAT	3' UTR	8.38	2	10.54	3
AC	3' UTR	1.73	2	2.14	2

expansions in 3'-UTRs could cause transcription slippage and produce mRNA expansion. Therefore, these enriched SSRs in specific regions play as good candidates for key elements in nucleotide sequence that endows the associated gene with the specific biological functions [2].

In conclusion, through cross-comparison and statistical analysis on orthologous genes, significant SSR regulators can be retrieved by our proposed system in a very efficient and effective way. The examples shown in this paper have successfully proved our assumptions that SSR regulators of a functional gene set can be identified through the verification of appearance rates and types of allocated regions. Based on this assumption, the developed system can be applied to the retrieve key SSR regulators for any functional gene sets.

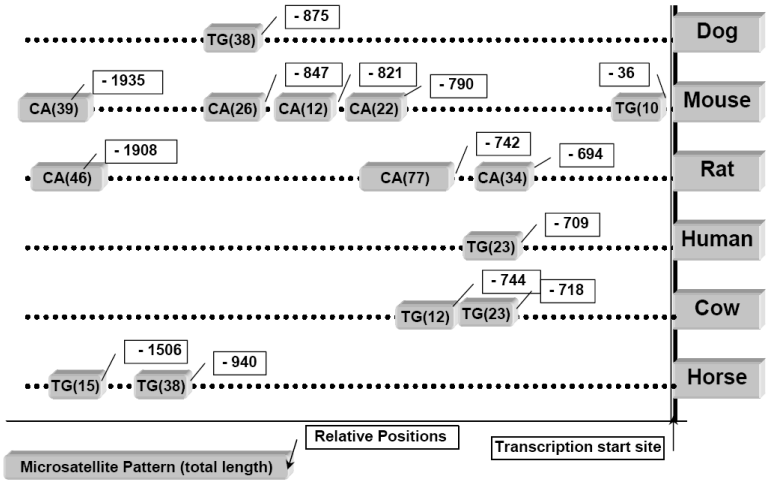

Figure 2: Alleles of a dinucleotide (CA)and (TG)SSRs found in promoter regions of IGF1 genes in several mammalian species.

Acknowledgements

This work is supported by the Center for Marine Bioscience and Biotechnology (CMBB) in National Taiwan Ocean University, Keelung, Taiwan, and the National Science Council in Taiwan, R. O. C. (NSC96-2627-B-019-003 to T.-W. Pai and NSC97-2627-B-007-017 to M. D.-T. Chang), and Chang-Gung Memorial Hospital-National Tsing Hua University (Joint Research Grant CGTH96-T12) to M. D.-T. Chang.

References

[1] Y.Kashi and D.G.King, Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22 (2006) 253-259.

[2] Y.C.Li, A.B.Korol, T.Fahima, and E.Nevo, Microsatellites within genes: structure, function, and evolution. Mol.Biol.Evol. 21 (2004) 991-1007.

[3] J.W.Fondon, III, E.A.Hammock, A.J.Hannan, and D.G.King, Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci.2008).

[4] H.Fan and J.Y.Chu, A brief review of short tandem repeat mutation. Genomics Proteomics.Bioinformatics. 5 (2007) 7-14.

[5] E.H.Margulies and E.Birney, Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nat.Rev.Genet. 9 (2008) 303-313.

[6] M.Ashburner, C.A.Ball, J.A.Blake, D.Botstein, H.Butler, J.M.Cherry, A.P.Davis, K.Dolinski, S.S.Dwight, J.T.Eppig, M.A.Harris, D.P.Hill, L.Issel-Tarver, A.Kasarskis, S.Lewis, J.C.Matese, J.E.Richardson, M.Ringwald, G.M.Rubin, and G.Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. 25 (2000) 25-29.

[7] A.R.Iglesias, E.Kindlund, M.Tammi, and C.Wadelius, Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. Gene 341 (2004) 149-165.

[8] S.Hadjiyannakis, H.Zheng, G.N.Hendy, and C.G.Goodyer, GT repeat polymorphism in the 5' flanking region of the human growth hormone receptor gene. Mol.Cell Probes 15 (2001) 239-242.

[9] C.S.Hale, W.O.Herring, H.Shibuya, M.C.Lucy, D.B.Lubahn, D.H.Keisler, and G.S.Johnson, Decreased growth in angus steers with a short TG-microsatellite allele in the P1 promoter of the growth hormone receptor gene. J.Anim Sci. 78 (2000) 2099-2104.

[10] H.Jernstrom, T.Sandberg, E.Bageman, A.Borg, and H.Olsson, Insulin-like growth factor-1 genotype predicts breast volume after pregnancy and hormonal contraception and is associated with circulating insulin-like growth factor-1 levels: implications for risk of early-onset breast cancer in young women from hereditary breast cancer families. Int.J.Gynecol.Cancer 16 Suppl 2 (2006) 497.

[11] A.R.Caetano and A.T.Bowling, Characterization of a microsatellite in the promoter region of the IGF1 gene in domestic horses and other equids. Genome 41 (1998) 70-73.

[12] R.A.Curi, H.N.Oliveira, A.C.Silveira, and C.R.Lopes, Effects of polymorphic microsatellites in the regulatory region of IGF1 and GHR on growth and carcass traits in beef cattle. Anim Genet. 36 (2005) 58-62.

[13] N.B.Sutter, C.D.Bustamante, K.Chase, M.M.Gray, K.Zhao, L.Zhu, B.Padhukasahasram, E.Karlins, S.Davis, P.G.Jones, P.Quignon, G.S.Johnson, H.G.Parker, N.Fretwell, D.S.Mosher, D.F.Lawler, E.Satyaraj, M.Nordborg, K.G.Lark, R.K.Wayne, and E.A.Ostrander, A single IGF1 allele is a major determinant of small size in dogs. Science 316 (2007) 112-115.

[14] X.F.Zhao, N.Y.Xu, X.X.Hu, and N.Li, [Effects of microsatellite in the regulatory region of IGF1 on growth traits in Jinhua swine]. Yi.Chuan 29 (2007) 206-210.

[15] K.Iida, C.J.Rosen, C.ckert-Bicknell, and M.O.Thorner, Genetic differences in the IGF-I gene among inbred strains of mice with different serum IGF-I levels. J.Endocrinol. 186 (2005) 481-489.

[16] J.F.Taylor, L.L.Coutinho, K.L.Herring, D.S.Gallagher, Jr., R.A.Brenneman, N.Burney, J.O.Sanders, J.W.Turner, S.B.Smith, R.K.Miller, J.W.Savell, and S.K.Davis, Candidate gene analysis of GH1 for effects on growth and carcass composition of cattle. Anim Genet. 29 (1998) 194-201.