An Identification of Secondary Structure Conserved Elements in Mammalian Syntenic Regions

Toutai Mituyama
Computational Biology Research Center
AIST
Tokyo, Japan
Mituyama-toutai@aist.go.jp

***Abstract*--3,361 putative RNA secondary structure conserved regions were identified from mammalian syntenies by using leading-edge RNA sequence data analysis tools. The identified candidates are found to be concentrated in transcription factor binding sites of various genes, which infer secondary structure conserved elements play certain roles for transcription regulation.**

Keywords-component; ncRNA;bioinformatics; secondary structure;synteny;transcription factor binding site

I. INTRODUCTION

Identification of novel noncoding elements in mammalian genomes has been actively studied for many years now. As soon as the first version of a large scale genome comparison of mammalian genomes had been released, researchers discovered many conserved noncoding elements (CNEs) which has been attracting number of research efforts to find a novel class of regulatory genes. One of the most distinct examples is the ultra-conserved elements [1] which have fewer base substitutions than coding regions [2]. Other conserved elements have particular base substitution patterns. The most major conserved elements in a genome are coding regions, which have a well-known substitution pattern of codon frames which shows higher base substitution rate at the third base position of a codon frame than the other positions. Capturing the base substitution pattern is useful to identify coding genes which fail database search or code short peptides [3]. Naturally, noncoding conserved elements do not have the codon frame substitution patterns. Instead, some classes of CNEs have base-pair consistent substitution patterns where a couple of bases have coordinated substitution patterns which are consistent in terms of base-pair formation [4](Fig. 1). Such a complex substitution pattern can be observed from loci where structural RNAs such as rRNA, tRNA, snoRNA, and miRNA are conserved [5]. Theoretically, capturing such a base-pair conserving substitution pattern can help identifying secondary structure conserved elements (SCEs) from NCEs, and it was actually practiced by [4] and [5]. The former identified 3377 SCEs from human/mouse genomes and the latter identified 47510 SCEs from 17 vertebrate genomes. Another method RNAz [6] utilized an original scoring system which evaluates minimum free energy (MFE) of putative secondary structures of a given set of homologous genomic sequences, which identified 35,985 SCEs from human, mouse, rat, and dog genomes. Even though the three efforts identified many SCE candidates, most of them have been waiting to be annotated. This study reports that many of SCEs coincide with transcription factor binding motifs and binding sites. Further identification was performed to enlist even more SCEs in mammalian genomes.

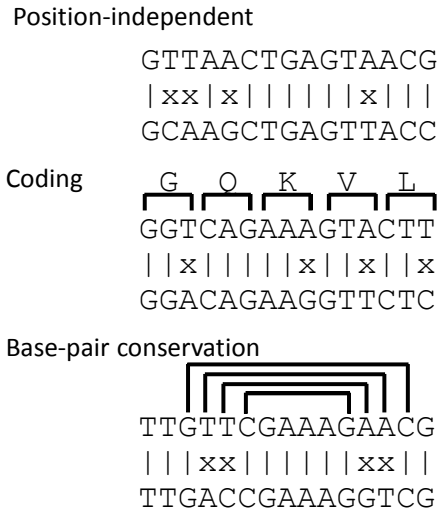

Figure 1. Three distinct base substitution patterns found in a non-coding region, coding region and a base-pair conserved region. The coding region show frequent substitutions at third base position of a codon frame. The base-pair conserved region has coordinated base substitutions consistent in terms of base-pair formation.

II. SCES AND TFBS

UCSC Genome Browser [7] features TFBS Conserved track which is a computational result of searching conserved motifs in Transfac Matrix Database (v7.0) provided by Biobase (http://www.gene-regulation.com/pub/databases.html). This track represents potential transcription factor binding sites (TFBS) and surprisingly coincides with SCEs found by QRNA, EvoFold, and RNAz in large part (Table 1). This fact infers close association between TFBS and SCEs. In addition to this, a large-scale analysis of transcription factor binding sites (publicly available as Txn Factor ChIP Track (TFT) of UCSC Genome Browser) which uses a combined method of

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

immuno-precipitation and next-generation sequencing (NGS) technology added indispensable information for annotating noncoding elements. We also found that many of the identified SCEs coincide with TFT. Since more than 41% (656,564 items out of 1,582,526) of TFT coincide with TFBS Conserved track, coinciding both of TFBS Conserved and TFT is reasonable. However, percentiles shown in "(%)2" column are lower than the expected number of coincidence (41%). If SCEs are randomly distributed over TFT, the percentiles become close to 41%. If SCEs have a positive preference to TFT over Transfac, they would have been higher than 41%. In this case, SCEs show negative preference to TFT. However, there are some procedural glitches in the previous methods. EvoFold and RNAz used multiple-alignment of genomic sequences provided by UCSC Genome Browser. Regular genome comparison computes alignment of genomic nucleotide sequence data in a position-independent manner. It fails to detect base-pair conserved regions with high rate of substitutions due to misalignment caused by the alignment method. This study deal with this problem by using recently released CentroidAlign [8] a fast and accurate aligner for structured RNAs, which is expected to capture more SCEs than the previous methods.

Table 1. SCEs identified by three previous studies and coinciding transcription factor binding motifs (Transfac) and binding sites captured by ChIP-seq (TFT). Percentiles next to Transfac column are fraction of SCEs coincide with "Transfac motifs and ones next to TFT column are fraction of Transfac-associated SCEs coincide with TFT.

Program	Total	Transfac	(%)	TFT	(%)2
QRNA	3,377	2,801	83%	936	33%
EvoFold	47,510	32,667	69%	8,607	26%
RNAz	35,984	31,096	86%	6,835	22%
Original	3,361	2,444	73%	1,248	51%

III. ORIGINAL IDENTIFICATION

Here we performed our original identification of SCEs by using CentroidAlign. Our identification was done on syntenies shared among four mammalian genomes i.e. human, mouse, rat, ant dog. Here synteny means a set of contiguous genomic sequences containing at least two adjacent genes that shared among multiple species (Fig. 2). Thusly a synteny covers exons, introns and intergenic regions. The advantage of using syntenies is to avoid false positives caused by artifacts of a genome comparison. For example, a pseudo-gene can be a major factor of "false conservation." A pseudo-gene which is not conserved across other species has homology to its parent gene which is conserved across other species. Thusly an ordinary genome comparison which is not sensitive to pseudo-genes finds highly homologous regions between the pseudo-gene and its parent gene in the other species genome as if the pseudo-gene is conserved across multiple species which leads to a false positive result (Fig. 3). The previous methods did not assess false positives of genome comparison including an example described here.

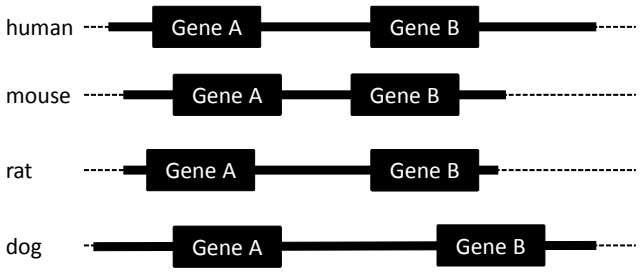

Figure 2. Syntenies are contiguous genomic sequences that contain at least two adjacent genes and are shared among multiple species. A synteny is represented as a thick black line.

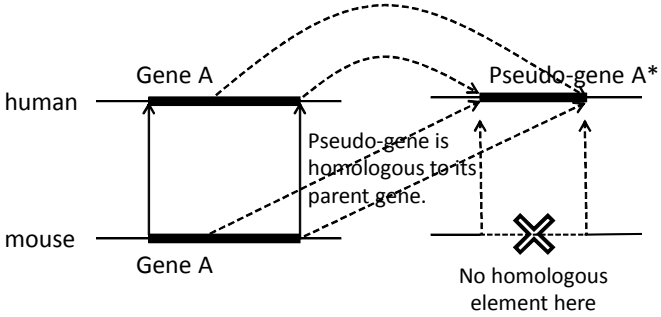

Figure 3. An example of false positives generated by an ordinary genome comparison. A pseudo-gene A* has no corresponding homologous elements in a mouse genome but it has homology to its parent gene which is conserved in a mouse genome. A genome comparison mistakenly detects pseudo-gene A* as a conserved element which is actually not conserved.

A. Dataset

Four mammalian genomes were used: human (hg18), mouse (mm9), rat (rn4), and dog (cf2). Homologous gene information was obtained from NCBI HomoloGene database (http://www.ncbi.nlm.nih.gov/HomoloGene/). Initial genomic sequence associations were obtained from UCSC Genome Browser.

B. Pipeline

Our pipeline enumerated 4,208 syntenies shared across the four mammalian genomes. The interrogation consists of following procedures. Repetitive elements in the intergenic regions were masked and aligned with secondary structure-sensitive fast multiple alignment tool CentroidAlign which computes the best alignment and secondary structure simultaneously for a given set of multiple nucleotide sequences. Although it is one of the best performing tools for the structural alignment, we had to limit the length of the input sequence under 6,000 nt in order to avoid excessive computational cost. The number of intergenic regions span less than 6,000 bp was 1,072. After the alignment, we used RNAz which discriminate an input multiple-alignment if it contains conserved secondary structure. Since RNAz allows up to 400 nt for the input alignment length, we interrogated the intergenic alignments with fixed window scanning of RNAz. We scanned the input alignments with two window

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

widths: 360 nt and 120 nt, and a sliding distance 60 nt. We merged overlapping scanning windows which were alarmed positive which means that a candidate has the p-score larger than 0.9 where the p-score represents a statistical significance and ranges 0 to 1. Finally, we obtained 3,361 positive windows with potential conserved secondary structures. Among them, 2,444 candidates coincide with Transfac motifs and 1,248 candidates coincide with TFT (Table 1). The number of total candidates of our method is comparable to QRNA and it has less coincidence with Transfac motifs. However, it acquired more TFT coincidence. Meanings of these results are discussed in the next section.

IV. DISCUSSION

Table 2 shows a comparison result of TFT coinciding candidates of four methods. Our method share relatively small number of TFT coinciding candidates with other methods. The original result has 934 novel candidates coinciding TFT but not included by the other methods. One example of the novel candidates is shown in Fig.4. With these results obtained from this study, conserved base-pairs are more closely related to TFBS.

Figure 4 shows three candidates (labeled as SYNTFT-Z-E-Q track shown at the top lane of the screen) found in the intergenic region of HOXC12 (right) and HOXC13 (left) where several clusters of TFBS were detected (shown in TFT track which is below SYNTFT-Z-E-Q track). The three candidates coincide with three clusters of TFBS. These candidates were solely detected by our original method.

-	QRNA	EvoFold	RNAz	Original
QRNA	936	180	351	23
EvoFold		8,607	1,000	48
RNAz			6,835	253
Original				1,248

Table 2. A comparison of TFT coinciding candidates of four methods. The original method shares small number of candidates with other methods.

[1] G. Bejerano, M, Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler, "Ultraconserved Elements in the Human Genome," Science, vol. 304, pp. 1321-1325, May 2004.

[2] S. Katzman, A. D. Kern, G. Bejerano, G. Fewell, L. F. Richard, K. Wilson, S. R. Salama, D. Haussler, "Human Genome Ultraconserved Elements Are Ultraselected," Science, vol. 317, pp. 915, August 2007.

[3] K. Kikuchi, M. Fukuda, T. Ito, M. Inoue, T. Yokoi, S. Chiku, T. Mitsuyama, K. Asai, T. Hirose, and Y. Aizawa, "Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation," Nucleic Acids Res. Vol. 37. Pp.4987-5000, August, 2009.

[4] E. Rivas, and S. R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," BMC Bioinformatics, vol. 2, pp.8, October 2001.

[5] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, and W. Miller, and D. Haussler, "Identification and Classification of Conserved RNA Secondary Structures in the Human Genome," PLoS Comp. Biol. Vol. 2, pp.0251-0262, April 2006.

[6] S. Washietl, I. L. Hofacker, M. Lukasser, A. Huttenhofer, and P. F. Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome," Nature Biotech. Vol. 23, pp.1383-1390, November 2005.

[7] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," Genome Res. Vol.12, pp.996-1006, May 2002.

[8] M. Hamada, K. Sato, H. Kiryu, T. Mituyama, and K. Asai, "CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score," Bioinformaitcs, vol. 25, pp.3236-3243, October 2009.

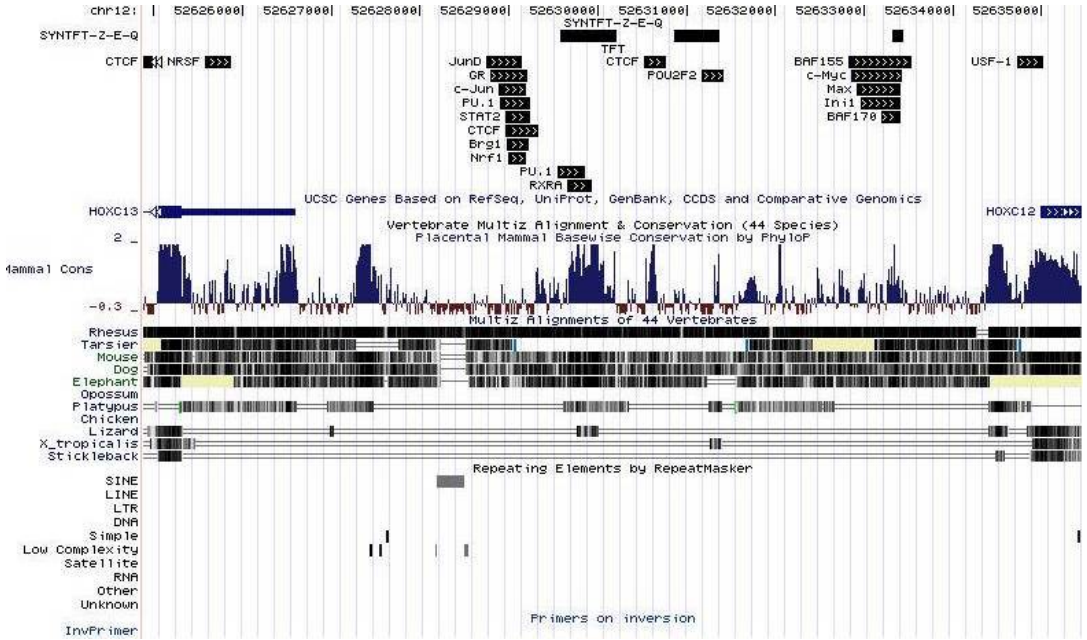

Figure 4. The intergenic region of HOXC12 and HOXC13 has five TFBS. Among them, three TFBS are associated to SCEs solely detected by this study.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE