Vol.2013-BIO-34 No.2
2013/6/27

情報科学的手法によるnon-coding RNAの遺伝子発現制御の解析

林 知里[1]　權 娟大[2]　宮崎 智[2]

概要：ゲノム解析が進んだ結果，ゲノム DNA 中にはタンパク質の遺伝子以外の分子情報をコードしている領域や発現の制御に関わる領域が存在することが明らかになった．これらの領域の中で，主な機能として発現制御等に関わる RNA 分子の遺伝子をコードしている領域があり，この領域から転写された RNA は non-coding RNA（ncRNA）と呼ばれている．ncRNA は他の遺伝子の転写制御に関わりを持つと考えられているが，未だその機能が解明されていないものも多い．一方で，ある疾患に特異的に発現する ncRNA が発見されており，ncRNA が創薬ターゲットになる可能性を持っている．また，現在発見されている ncRNA の多くはタンパク質遺伝子間に存在するが，あるタンパク質遺伝子のイントロンに ncRNA が存在することが分かっている．本研究では，タンパク質遺伝子と ncRNA のゲノム上の位置関係に着目し，疾患関連遺伝子を制御する創薬ターゲットとしての ncRNA を予測することを目的とする．本稿では，イントロンに存在する ncRNA に着目した生物種間での比較結果を報告する．

キーワード：non-coding RNA，遺伝子発現，イントロン，データベース

Non-coding RNA analysis for gene expression regulation using bioinformatics tools

Hayashi Chisato[1]　Kwon Yeondae[2]　Miyazaki Satoru[2]

Abstract: Advances in genome analysis have revealed that a genomic DNA contains gene coding regions other than protein coding sequences and transcription-regulating regions. Among these regions, there exist RNA coding regions involved in transcription regulation as a primary function, RNAs transcribed from these regions are referred to as non-coding RNAs (ncRNAs). ncRNAs are known to affect regulation of transcription, but many functions are still unknown. On the other hand, since some ncRNAs specifically express in a particular disease, they may have potential for being used as drug targets. Also, while most of ncRNAs currently observed locate on intergenic regions of protein coding genes, some ncRNAs locate on intronic regions of protein coding genes. The aim of this work is to predict ncRNAs as drug targets regulating disease-associated genes based on positional relationship between protein genes and ncRNAs on a genome. In this paper, we compare ncRNAs located on intronic regions across species.

Keywords: non-coding RNA, gene expression, intron, database

1. はじめに

生体内ではセントラルドグマという一連の流れにより，DNA から RNA を経てタンパク質が生成される．生物のゲノム解析が進んだ結果，DNA 中にタンパク質以外のコードを持つ領域が存在し，その割合は高等生物になるにつれ増

[1] 東京理科大学大学院 薬学研究科 薬科学専攻
Graduate School of Pharmaceutical Sciences, Tokyo University of Science, Noda, Chiba 278-8510, Japan
[2] 東京理科大学 薬学部 生命創薬科学科
Department of Medical and Life Science, Faculty of Pharmaceutical Sciences, Tokyo University of Science

© 2013 Information Processing Society of Japan

加していて，ヒトで最大になることが分かった [1]．この領域から転写された RNA は non-coding RNA（ncRNA）と呼ばれている．ncRNA は他の遺伝子の転写制御に関与していると考えられており，近年注目されてきている [2], [3]．

ncRNA は 200 塩基を基準に分類されており，200 塩基未満の ncRNA は small ncRNA，200 塩基以上の ncRNA は long ncRNA と呼ばれている．これまでの様々な ncRNA の研究により，small ncRNA が遺伝子発現抑制を引き起こす RNA 干渉（RNA interference: RNAi）という機能を持つことが明らかになった [4]．RNAi を引き起こす ncRNA の一種である miRNA は、細胞内で mRNA を抑制することでタンパク質の産生を調節する機能を持っている [5]．miRNA は発生期の形態形成，細胞分化，アポトーシスなど細胞の発現調整などに重要な役割を果たしており，発現調節異常ががんなどの疾患と関連があることが報告されている [6], [7]．miRNA によって起こる RNAi を応用した核酸医薬品は，低分子医薬品・抗体医薬品に次ぐ新規医薬品として開発が進められている．small ncRNA の機能が明らかになる一方で，long ncRNA は未だ機能が分からないものが多く，small ncRNA だけでは説明できなかった遺伝子発現制御などの働きに long ncRNA が関わっていると考えられている [8]．さらに，ある疾患に特異的に発現している long ncRNA が発見されているため，創薬ターゲットやバイオマーカーとして long ncRNA の医療への応用が期待されている [9], [10], [11]．

また，現在発見されている ncRNA の多くは遺伝子間に存在するが，ある遺伝子のイントロンに ncRNA が存在することが分かっている．そこで，本研究では，遺伝子と ncRNA のゲノム上の位置関係に着目し，疾患関連遺伝子を制御する創薬ターゲットとしての ncRNA の予測を目的とする．

2. イントロンに存在する ncRNA

mRNA のうちエクソンがタンパク質の情報をコードしており，タンパク質の生成過程においてスプライシングで除かれるイントロンはジャンクだと考えられていた．しかし，核小体低分子 RNA（snoRNA）の中にはリボソームタンパク質やリボソームに関連するタンパク質遺伝子のイントロン中にコードされているものが見つかっており，ホスト遺伝子の mRNA 前駆体の一部として転写された後，成熟 snoRNA として機能を持つことが分かっている [12]．さらに最近の研究では，遅発性アルツハイマー病の原因遺伝子と，そのイントロンのアンチセンスに存在する ncRNA の関連性が示唆された [13]．そこで，本研究ではイントロンに存在する ncRNA に着目し，イントロン領域に存在する ncRNA の生物種間での比較を行った．

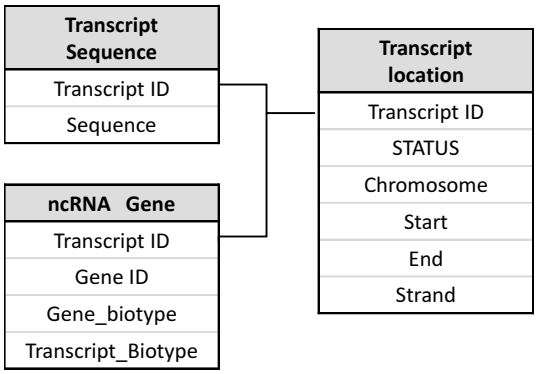

図 1 ncRNA データベース

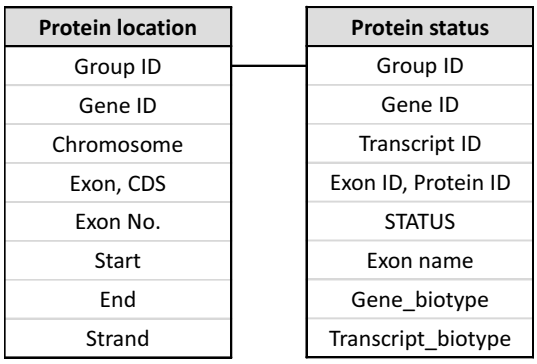

図 2 Gene データベース

3. 生物種間での intronic ncRNA の比較

本研究では，遺伝子と ncRNA のゲノム上の位置関係から，疾患関連遺伝子を制御する ncRNA の予測を目的とする．まず，本研究で用いる遺伝子および ncRNA の情報をデータベースに格納し，次に，ゲノム上の遺伝子と ncRNA の位置関係を用いて，ncRNA を遺伝子に対する位置情報で分類する．最後に，イントロンにある ncRNA を抽出し，モデル生物種間での比較を行う．

3.1 データベースの作成

遺伝子と ncRNA のゲノム上の位置関係を利用することから，遺伝子と ncRNA の両方のデータをそれぞれデータベースに格納した．また，生物種間での比較を行うために，ヒト（*H. sapiens*），マウス（*M. musculus*），ラット（*R. norvegicus*），ハエ（*D. melanogaster*），酵母（*S. cerevisiae*），線虫（*C. elegans*）の 6 種類のモデル生物のデータを格納した．

まず，Ensembl Genome データベース [14]（`http://asia.ensembl.orgindex.html`）より，ncRNA データセットおよび遺伝子情報データセット（Gene sets）を取得し，必要なデータの抽出を行った．ncRNA データベースは，塩基配列情報を格納した「Transcript Sequence」テーブル，ncRNA 転写産物のゲノム上の位置情報を格納した

© 2013 Information Processing Society of Japan

表 1 位置関係に基づいた ncRNA の分類結果と intronic ncRNA の割合

Organism	ncRNA 登録件数	intergenic ncRNA	intronic ncRNA	sense ncRNA	割合（%）
H. sapiens	19,505	15,962	2,910	128	14.9
M.musculus	7,886	6,418	1,345	61	17.0
R. norvegicus	4,828	3,937	837	12	17.3
D. melanogaster	1,396	1,302	8	28	0.6
C. elegans	23,872	20,116	3,549	178	14.9
S. serevisiae	413	363	13	2	3.1

表 2 intronic ncRNA の STATUS

STSTUS	*H.sapiens*	*M.musculus*	*R.norvegicus*	*D.melanogaster*	*C.elegans*	*S.cerevisiae*
miRNA	525	526	202	2	65	
sense_intronic	488	74				
sense_over-lapping	13	2				
misc_RNA	309	87	68			
lincRNA	68	12				
non_coding/ncRNA	16	1		4	3216	
rRNA	106	61	41			
snoRNA	481	356	346	1	87	8
snRNA	469	224	180		29	
tRNA				1	152	5
3prime_over-lapping_ncrna	18	2				
ncrna_host	2					
pseudogene	415					

「Transcript location」テーブル，ncRNA が転写されるゲノムの特徴を格納した「ncRNA Gene」テーブルの 3 つのテーブルからなる（図 1）.

Gene データベースは，遺伝子のゲノム上の位置情報を格納した「Protein location」テーブル，遺伝子が転写されるゲノムの特徴を格納した「Protein status」テーブルの 2 つのテーブルからなる（図 2）.

3.2 位置情報による ncRNA の分類

本研究では，Ensembl の Gene sets に登録されている遺伝子の位置情報のうち，タンパク質をコードしているとされる protein coding 転写産物の位置情報から，ゲノム上で重複する範囲を考慮して新たに遺伝子領域を決定した．得られた遺伝子領域の位置情報と ncRNA の位置情報および以下の条件を用いて ncRNA を次の 3 つに分類した（図 3）．ここで，n は任意の数を表し，塩基は配列の 5' 末端から数える．また 5' 側を上流とする．

(1) intergenic ncRNA
- 遺伝子と遺伝子の間に存在する ncRNA
- n 番目の遺伝子領域の終了塩基より ncRNA の開始点が下流で，かつ ncRNA の終了塩基より n+1 番目の遺伝子領域の開始点が下流である場合

(2) intronic ncRNA
- 遺伝子のイントロンに存在する ncRNA
- ある遺伝子において，n 番目のエクソンの終了塩基より ncRNA の開始点が上流で，かつ ncRNA の終了塩基より n+1 番目のエクソンの開始点が下流である場合

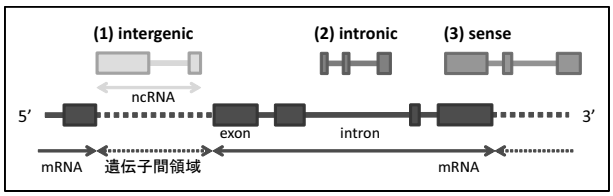

図 3 遺伝子と ncRNA の位置関係の分類

(3) sense ncRNA
- 遺伝子と範囲が重複している ncRNA
- ncRNA の開始塩基よりある遺伝子領域の終了塩基が下流で，かつその遺伝子領域の開始より ncRNA の終了塩基が下流である場合

表 1 に遺伝子と ncRNA の位置関係に基づいた分類結果および全 ncRNA に対する intronic ncRNA の割合を示す.

3.3 intronic ncRNA の STATUS

表 2 に，表 1 の intronic ncRNA の STATUS の内訳を示す.

3.4 生物種間における intronic ncRNA の配列比較

本研究では，生物種間の intronic ncRNA の配列類似度の比較を容易に行うために，表 2 の ncRNA のうち，RNAi の機能を持っている miRNA [5] に限定した．また，比較する intronic ncRNA の件数を絞るため性染色体である X 染色体に限定して解析を行った．本稿では，ヒトとマウス間

© 2013 Information Processing Society of Japan

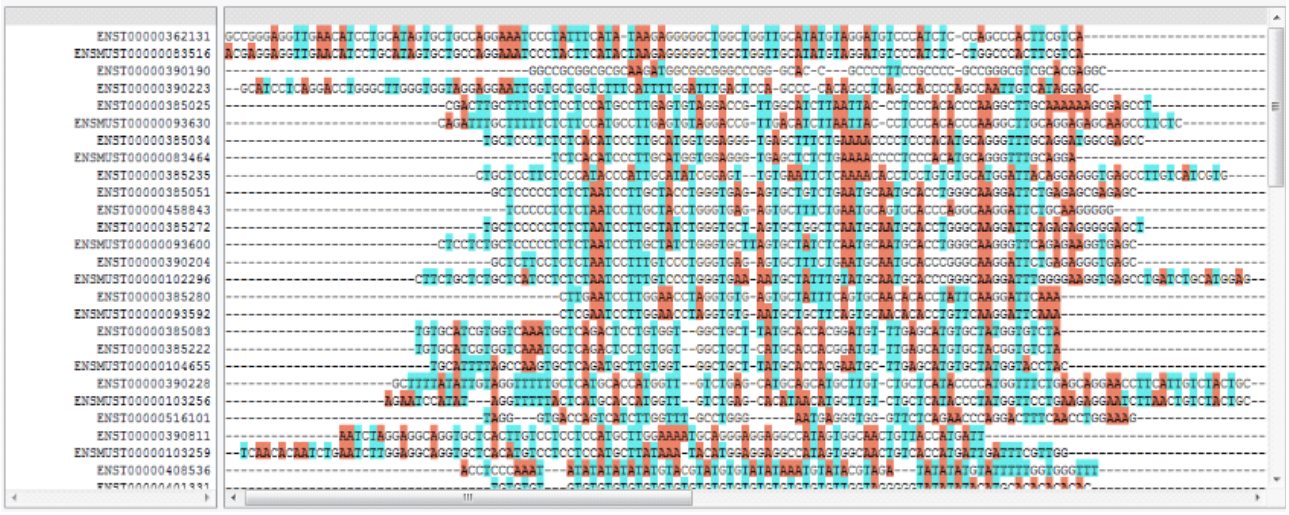

図 4 X 染色体 miRNA のマルチプルアライメント結果（抜粋）
ENST から始まるアクセッション番号の配列はヒトの配列
ENSMUST から始まるアクセッション番号の配列はマウスの配列

の比較についてのみ説明する．

まず，図 1 の ncRNA データベースより，X 染色体の intronic ncRNA のうち，ヒトとマウスの miRNA 38 件と 37 件をそれぞれ取得し，マルチプルアライメントツール ClustalX 2.0.11 [15] を用いて，上記の合計 75 件の配列に対してマルチプルアライメントを行った（図 4）．次に，マルチプルアライメント結果を用いて，近隣結合法 [16] により系統樹を作成した（図 5）．この系統樹を用いて，ヒトとマウスでペアになっている配列を比較対象として 21 ペアを選定し，相同性検索ツール bl2seq [17] を用いて，配列の一致度を求めた．表 3 に，21 ペアに対応するアクセッション番号とその配列一致度を示す．

4. 考察

4.1 イントロンに存在する ncRNA

表 1 により，イントロンに存在する ncRNA の割合は，哺乳類であるヒト，マウス，ラットの間では大きな差は見られない．ハエ，線虫，酵母の間では，線虫で最初に RNAi が発見されたこともあり，ncRNA の研究が盛んに行われているため，他の哺乳類と同程度の intronic ncRNA が登録されていると考えられる．

表 2 により，高等生物になるにつれ，イントロンに存在する ncRNA のバリエーションも増えていくことが分かる．また，miRNA のようにタンパク質の発現制御機能を持つ ncRNA がイントロンに存在していることが分かる．このことから，イントロンに存在する ncRNA が生体内で重要な機能を果たしていることが示唆された．

4.2 intronic ncRNA の配列比較

3.4 節で得られた 21 ペアの配列一致度が 93.3 ± 4.1%

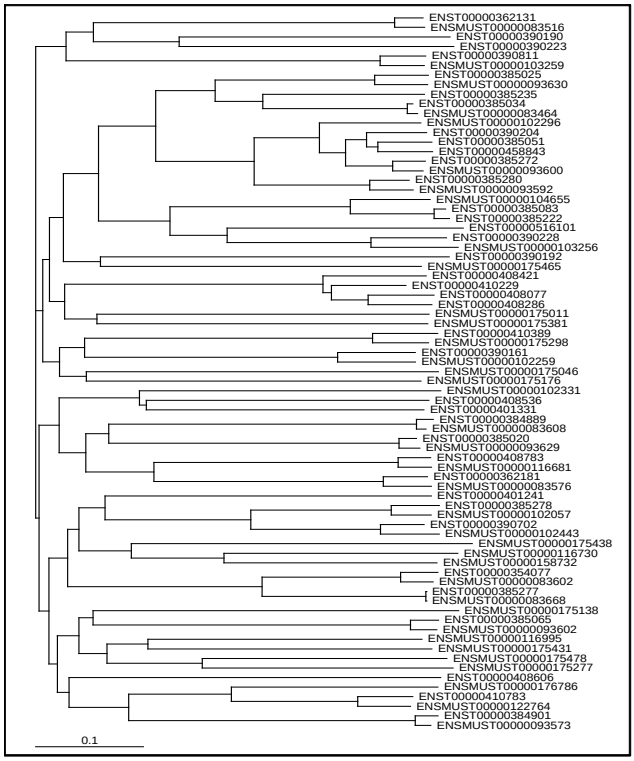

図 5 ヒトとマウスの *X* 染色体上にある intronic miRNA の系統樹

(Mean ± SD) であることから（表 3），これらの塩基配列は生物種間で保存されていると考えられる．このことから，ncRNA は生物種間で保存され，かつ生体内で重要な機能を果たしている可能性が示唆された．

また，図 5 の系統樹より，ヒト miRNA 配列同士またはマウス miRNA 同士が集まっている node があることが分かる．これらは進化の過程で得られた可能性があり，中でも生物種をまたいで同じ機能を持つ ncRNA がない場合，生物種間での疾患メカニズムの解析および医薬品の副作用

© 2013 Information Processing Society of Japan

表 3 系統樹に基づく 21 ペアの配列一致度

ヒト	マウス	配列一致度（%）
ENST00000362131	ENSMUST000000083516	96
ENST00000385034	ENSMUST000000083463	99
ENST00000385025	ENSMUST000000093630	92
ENST00000385272	ENSMUST000000093600	93
ENST00000390204	ENSMUST000000102296	88
ENST00000385280	ENSMUST000000093592	92
ENST00000390228	ENSMUST000000103256	88
ENST00000390811	ENSMUST000000103259	91
ENST00000390192	ENSMUST000000175438	99
ENST00000385065	ENSMUST000000093602	96
ENST00000410783	ENSMUST000000122764	91
ENST00000384901	ENSMUST000000093573	96
ENST00000385278	ENSMUST000000102057	95
ENST00000390702	ENSMUST000000102443	90
ENST00000354077	ENSMUST000000083602	94
ENST00000385277	ENSMUST000000083668	100
ENST00000390161	ENSMUST000000102259	83
ENST00000408783	ENSMUST000000116681	93
ENST00000362181	ENSMUST000000083576	91
ENST00000385020	ENSMUST000000093629	96
ENST00000384889	ENSMUST000000083608	96

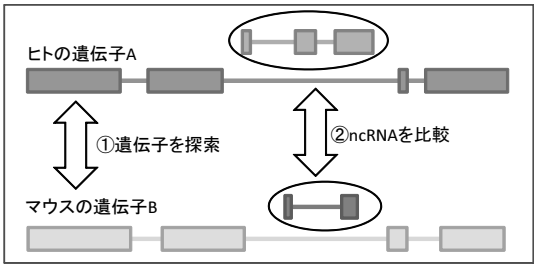

図 6 ホスト遺伝子を用いた intronic ncRNA の生物種間比較法

発現の違いの解明の糸口になるかもしれない．

5. まとめと今後の課題

本研究では，疾患関連遺伝子を制御する創薬ターゲットとしての ncRNA を予測することを目的として，モデル生物の intronic ncRNA の配列の比較を行った．まず，遺伝子と ncRNA のデータベースを作成し，ゲノム上の遺伝子に対する ncRNA の位置関係を用いてイントロンにある ncRNA を取得した．得られた ncRNA 配列を用いて，配列のアライメント，系統樹の作成，配列一致度の計算などを行い，ncRNA 配列の比較を行った．比較に用いた intronic ncRNA 配列の生物種間での類似度が高いことから，これらの配列は生物種間で保存されており，生体内である特定の機能を果たしていると予想される．

しかし，比較に用いた miRNA 配列は比較的配列長が短い配列であることから，選定したペア配列が染色体上に頻繁に存在する配列である場合，偶然一致したものである可能性を排除できない．そこで，3.4 節のアライメントによって得られたペア配列について，その ncRNA をイントロンに含んでいる遺伝子をホスト遺伝子とみなし，生物種間でホスト遺伝子の配列比較を行っている（図 6）．

また，がんの進行に関連があるとされている配列長が長い lincRNA（long intergenic RNA）[8] でも同様の解析を行っている．lincRNA 自体の塩基配列を直接比較することは困難であるため，ホスト遺伝子の比較により間接的に生物種間で同じ機能を持つ ncRNA の特定およびその規則性の解析を試みる．

参考文献

[1] Taft, R.J. and Mattick, J.S.: Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences, *Genome Biology*, Vol.5, No.P1, pp.1–24 (2003).
[2] Shiomi, H.: Non-coding RNAs as spatiotemporal regulators of genome network, *Experimental Medicine*, Vol.49, pp.2503–2509 (2004).
[3] Mattick, J.S.: RNA regulation: a new genetics?, *Nature*, Vol.5, pp.316–323 (2004).
[4] Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C.: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature*, Vol.391, pp.806–811 (1997).
[5] Krol, J., Loedige, I. and Filipowicz, W.: The widespread regulation of microRNA biogenesis, function and decay, *Nat Rev Genet*, Vol.11, pp.597–610 (2010).
[6] Suzuki, H.I., Yamagata, K., Sugimoto, K., Iwamoto, T., Kato, S. and Miyazono, K.: Modulation of MicroRNA processing by p53, *Nature*, Vol.460, pp.529–533 (2009).
[7] Suzuki, H.I., Arase, M., Matsuyama, H., Choi, Y.L., Ueno, T., Mano, H., Sugimoto, K. and Miyazono, K.: MCPIP1 ribonuclease antagonizes Dicer and terminates microRNA biogenesis through precursor microRNA degradation, *Molecular Cell*, Vol.44, Issue 3, pp.424–436 (2011).
[8] Tsai, M.C., Spitale, R.C. and Chang, H.Y.: Long intergenic noncoding RNAs: New lincs in cancer progression, *Cancer Res.*, Vo.71, pp.3–7 (2011).
[9] Gibb, E.A., Brown, C.J. and Lam, W.L.: The functional role of long non-coding RNA in human carcinomas, *Molecular Cancer*, Vol.10, No.38, pp.1–17 (2011).
[10] Mitra, S.A., Mitra, A.P. and Triche, T.J.: A central role for long non-coding RNA in cancer, *Front Genet*, Vol.3, Article 17, pp.1–9 (2012).
[11] Liu, Q., Huang, J., Zhou, N., Zhang, Z., Zhang, A., Lu, Z., Wu, F. and Mo, Y.Y.: LncRNA loc285194 is a p53-regulated tumor suppressor, *Nucleic Acids Res.*, Vol.41, No.9, pp.4976–4987 (2013).
[12] Tycowski, K.T., Shu, M.D. and Steitz, J.A.: A mammalian gene with introns instead of exons generating stable RNA products, *Nature*, Vol.379, pp.464–466 (1996).
[13] Eleonora, C., Massone, S., Penna, I., Nizzari, M., Gigoni, A., Dieci, G., Russo, C., Florio, T., Cancedda, R. and Pagano, A.: An intronic ncRNA-dependent regulation of SORL1 expression affecting A β formation is upregulated in post-mortem Alzheimer's disease brain samples, *Disease Model and Mechanisms*, Vol.6, pp.424–433 (2013).
[14] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., *et al.*: The Ensembl Genome data project,

© 2013 Information Processing Society of Japan

Nucleic Acids Res., Vol.30, pp.38–41 (2002).

[15] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G.: Clustal W and Clustal X version 2.0, *Bioinformatics*, Vol.23, Issue 21, pp.2947–2948 (2007).

[16] Saitou, N. and Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, Vol.4, No.4, pp.406–425 (1987).

[17] Tatusova, T.A. and Madden, T.L.: BLAST 2 Sequence, a new tool for comparing protein and nucleotide sequences, *FEMS MICROBIOL LETT.*, Vol.174, No.2, pp.247–250 (1999).

© 2013 Information Processing Society of Japan