Journal of Bioinformatics and Sequence Analysis Vol. 2(5), pp. 69-74, October 2010
Available online at http://www.academicjournals.org/JBSA

Full length Research Paper

UniDPlot: A software to detect weak similarities between two DNA sequences

Marc Girondot[1,2]* and Jean-Yves Sire[3]

[1]Laboratoire d'Écologie, Systématique et Évolution, UMR 8079 Centre National de la Recherche Scientifique, Université Paris Sud et ENGREF, 91405 Orsay cedex 05, France.
[2]Département de Systématique et Evolution, Muséum National d'Histoire Naturelle de Paris, 25 rue Cuvier, 75005 Paris, France.
[3]Université Pierre and Marie Curie-Paris 6, UMR 7138 "Systématique, Adaptation, Evolution", 7 quai St-Bernard, 75005 Paris, France.

Accepted 21 June, 2010

Search for DNA sequence similarity is a crucial step in many evolutionary analyses and several bioinformatic tools are available to fulfill this task. Basic Local Alignment Search Tool (BLAST) is the most commonly and highly efficient algorithm used. However, it often fails in identifying sequences showing very weak similarity. An alternative method is to use Dot Plot, but such a graphical method is not suitable for the analysis of large sequences (e.g. hundreds of kilobases) as this is now more often required in the context of genome sequencing programs. As an alternative to the classical Dot Plot method, we designed UniDPlot, which permits to search for weak similarity either between two large sequences (e.g., genome regions, ...) or between one large sequence and a short one (e.g., exons, ...). UniDPlot methodology contracts the output of the Dot Plot similarity matrix along the length of the largest sequence, while defining statistical limits of significance using a bootstrap procedure. To illustrate the efficiency of this method, we used UniDPlot to search for the fate of the gene that encodes the major enamel protein, amelogenin, in chicken. Although we showed that amelogenin was invalidated through a pseudogeneization process, we recovered the entire sequence in the chicken genome. Using UniDPlot, we have identified a pseudogene, which was not detected by classical methods. UniDPlot can be used to search for missing genes, or motifs of various sizes in different genomic contexts.

Key words: DNA sequence similarity, UniDimensional plot (UniDPlot) software, genomes.

INTRODUCTION

Search for similar sequences among genomes or within a target genome is one of the more classical tasks in Bioinformatics. Until now, many tools were developed, and one of the most commonly used algorithms is Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). Although BLAST supports many different options, it could fail to detect similarity when two evolutionary distant sequences are used (Miller, 2001). Position-Specific Iterated BLAST (PSI-BLAST) is an alternative to search for weak similarity using amino acid sequences (Altschul et al., 1997). PSI-BLAST is designed to detect relationships between the query and members of the database, when they are not detectable by standard BLAST searches. The added sensitivity of this program compared to regular BLAST is provided by the use of a profile that is automatically constructed from a multiple alignment of the highest scoring hits in the initial BLAST search. However, PSI-BLAST cannot be run with nucleotide sequences, as for instance in searching pseudogenes or regulatory sequences. Another alternative to regular BLAST is provided by Dot Plot. Originally called diagram (Gibbs and Mcintyre1, 1970), Dot Plot regroups several methods that visually compare two sequences and look for regions of close similarity. A

*Corresponding author. E-mail: marc.girondot@u-psud.fr.

dot plot is a visual representation of the similarities between two sequences. Each axis of a rectangular array represents one of the two sequences to be compared. Whenever one base in one sequence is identical in the other sequence, a dot is drawn at the corresponding position of the array. Thus, when two sequences share similarity over their entire length a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner. If two sequences only share patches of similarity this will be revealed by diagonal stretches.

Maizel and Lenk (1981) popularized Dot Plot and suggested the use of a filter to reduce the noise resulting from matches that occur by chance. As there are combinations of only four different nucleotides, the probability is high that a nucleotide matches another nucleotide in a region of the sequences with no homology. Therefore, the result does not reflect a similarity between the two sequences but the only limited number of bases permitted in DNA sequences. A large variety of filters can be used (Sonnhammer and Durbin, 1995). Maizel and Lenk (Maizel and Lenk, 1981) suggested to place a dot only when there is a significant proportion of successive matching bases. Recent advances in Dot Plot methodology involved parallelization (Mueller et al., 2006) but the visualization of the results is still a bottleneck for using the method. This is particularly well illustrated when a short sequence is to be compared to a large one as, for instance, a genomic fragment. In this case, the result of the Dot Plot will resume into a narrow black line for identical scales in both axes.

Here, we present a new method that we have called UniDimensional Plot (UniDPlot). It is an adaptation of the original Dot Plot method. UniDPlot was designed to compare a short sequence to a large one, while testing the significance of the similarity obtained. In order to illustrate the usefulness of this new method, we used UniDPlot to detect a missing gene in the annotated chicken (*Gallus gallus*) genome (build 2.1, November 30, 2006).

IMPLEMENTATION

DotPlot projection

The classical dot plot algorithm uses a pairwise comparison between two sequences, and the results are presented as a dot-matrix. For a particular position in both sequences, the same base present is shown as a dot and two different bases are shown as a blank. A sliding window is often used to filter the output for better visualisation (Maizel and Lenk, 1981). For UniDPlot method, a projection of the maximum score for each position along the largest sequence against all possible positions of the shortest sequence is plotted. This creates a plot that permits to visualize regions with various values of similarity between the two sequences compared.

Substitution model

Evolutionary divergent sequences could have a known pattern of nucleotide divergence. In the classical Dot Plot algorithm, only 0 - 1 outputs are possible (Gibbs and Mcintyre1, 1970) and gray levels have been further added (Wimmer, 2007). Here, the output can use a matrix in which all base pairs have different weights. Such a matrix can be directly calculated by the software from two aligned sequences. The models implemented are: identical, transition and transversion, and complete matrix obtained by a simple comparison of the two aligned sequences.

Test of significance

The basics of significant test for dot-plot have been given by Gibbs and Mcintyrel (Gibbs and Mcintyre1, 1970). However, this procedure is applicable only for identical model of substitution without filter. Here, the expected number of maximum matches between the two sequences compared is calculated using a resampling procedure. Briefly, the same number of comparisons are done with two random sequences obtained using the observed frequencies of ATGC for each sequences applying the same substitution model. This permits to define a limit above which such a similarity has never been reached in the same number of trials. During this resampling, the mean and standard deviation of proportion of maximum identities between two sequences are also calculated. This procedure is used rather than an analytical one, due to the complexity of the model of maximum identities proportion based on a sliding window procedure.

Biological background for the test

The ability to form teeth was lost in an ancestor of all modern birds, approximately 80 - 100 million years ago. However, experiments in chicken have revealed that the oral epithelium can respond to inductive signals from mouse mesenchyme, leading to reactivation of the odontogenic pathway (Mitsiadis et al., 2003). Recently, tooth germs similar to crocodile rudimentary teeth were found in a chicken mutant (Harris et al., 2006). These "chicken teeth" did not develop further, but the question remains whether true teeth would have been obtained if the experiments were performed longer. An alternative approach to check whether or not obtaining true hens' teeth in the next future was not utopia was to look for the fate of the dental protein genes, 100 million years after tooth loss. Previous molecular attempts to localise amelogenin gene, the major protein in enamel formation, in chicken DNA were unsuccessful (Girondot and Sire, 1998). Blast searching (BLASTN) for these genes using either full length amniotic sequences or various e-primers defined from conserved regions proved to be unfruitful, even when using low search sensitivity (distant homology).

As an alternative we used gene synteny between mammals and birds to try to locate amelogenin gene in the chicken genome.

In placental mammals, AMEL is located close to the rhoGTPase activating protein 6 gene (ARHGAP6). For instance, in humans AMELX is located at position Xp22.3, between ARHGAP6 and HCCS (holocytochrome C synthetase) gene. MID1 (midline 1) and MSL3L1 (male-specific lethal 3-like 1) mark out this region. AMELX codes in antisense within the 200 kb large intron 1 of ARHGAP6, and its 5' UTR is located at approximately 40 kb far from the 5' region of ARHGAP6 exon 2. In the opossum, AMEL is similarly located, but 58 kb far from ARHGAP6 exon 2.

In chicken, ARHGAP6 (LOC418642), MID1 and MSL3L1 (LOC418641) are found close, one to another on chromosome 1, but compared to their location in humans, chicken MID1 and MSL3L1 are inverted, while HCCS is located on chicken chromosome 8 (LOC424482). In the target region, i.e. between ARHGAP6 and MID1, the Genbank prediction program indicates neither the presence of a putative candidate gene locus nor of a pseudogene.

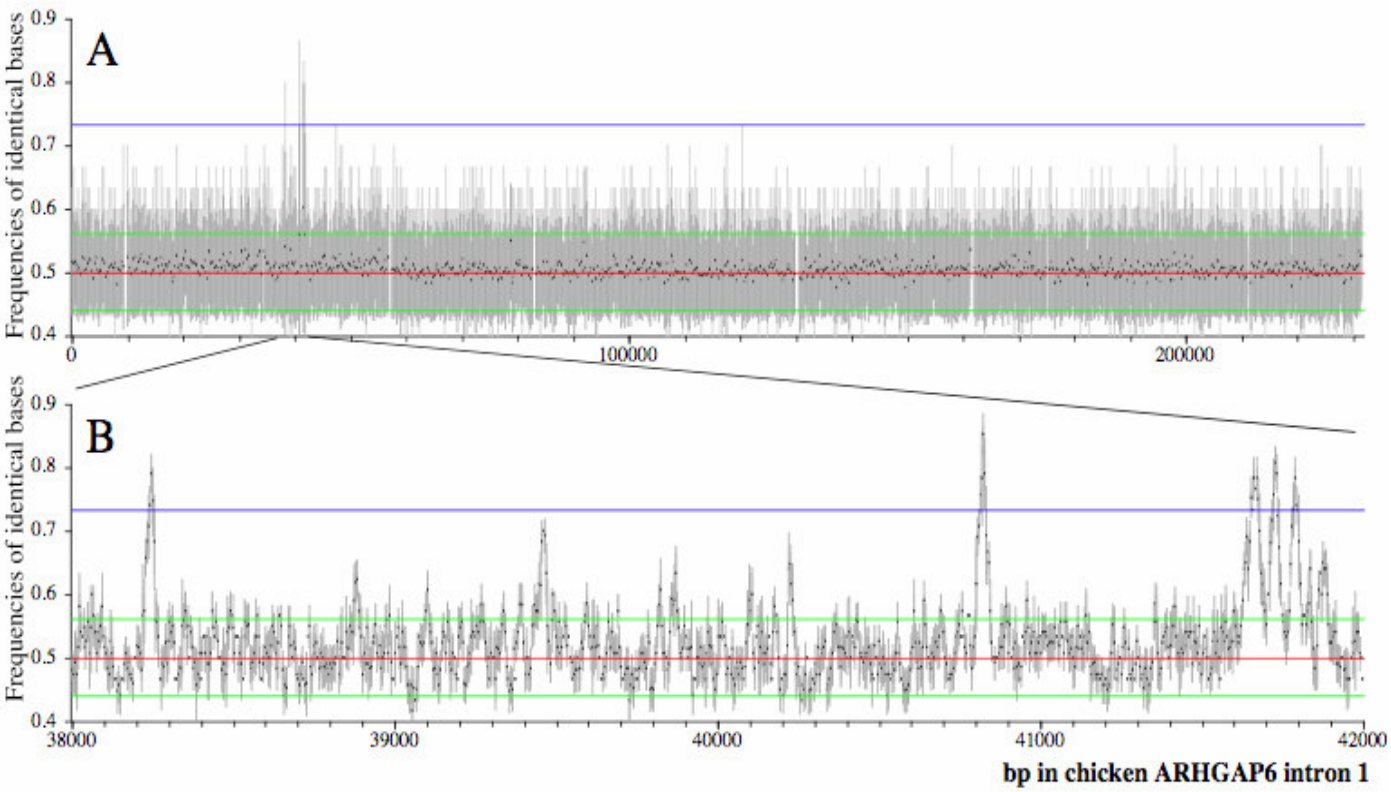

Figure 1. Similarity of crododilian amelogenin cDNA against intron 1 of chicken ARHGAP6 (A) and enlargement of the region with highest similarities (B). The blue line indicates the maximum observed similarity in random resamplings. The green lines represent twice the standard deviation of similarity around the mean value in red. The vertical grey lines indicate the maximum and minimum observed similarity at each corresponding position and each black dot is the similarity average at each position. One pixel summarized 238 bp in (A) and 4 bp in (B). The position 1 for the intron 1 of chicken ARHGAP6 is its first base.

We searched for sequence similarity in the target region with UniDPlot software, using crocodilian AMEL sequence.

RESULTS AND DISCUSSION

The exons of the entire coding sequence of crocodilian AMEL were separated with stretches of 30 X, which is the size of the window that will be used. This permits to ensure that the search will not be confused by artefactual adjacent bases that are normally separated by introns.

When running on entire intron 1 of chicken ARHGAP6 (231,857 bp), UniDPlot revealed several successive hits located approximately 40,000 pb far from ARHGAP6 exon 2 (Figure 1). The similarity was higher than observed for random resamplings of sequences (blue line). When this region was enlarged, three significant peaks were observed (that is above the blue line) and an additional peak was observed in this region, just below the blue line. Such an organisation is compatible with the known structure of the crocodilian AMEL gene that is four coding exons.

In order to confirm this organization, similarity search was performed using crocodilian AMEL exons 2, 3, 5 and 6, separately. Indeed, the first exon is non-coding, exon 4 does not exist in crocodilian AMEL and exon 7 could not be used (3 coding bases only).

The results indicate clearly that the organization of these exons in the chicken genome is similar to the expected one if they belong to AMEL (Figure 2). These sequences were aligned with the crocodilian AMEL (Figure 3). The chicken AMEL gene is a pseudogene, due to an insertion of four bases in the sequence of the first translated exon 2 (signal peptide).

Conclusions

The method proposed here permits to find significant similarity that has been overlooked by automatic procedure used to annotate the *Gallus* genome. The reason is probably due to both the divergence between sequences but also the fact that the bird gene is now a pseudogene and cannot be automatically translated. More generally, this software can be used to search for missing genes in distant organisms and in comparing highly derived sequences. The methodology presented here could have a large range of use, to find missing or

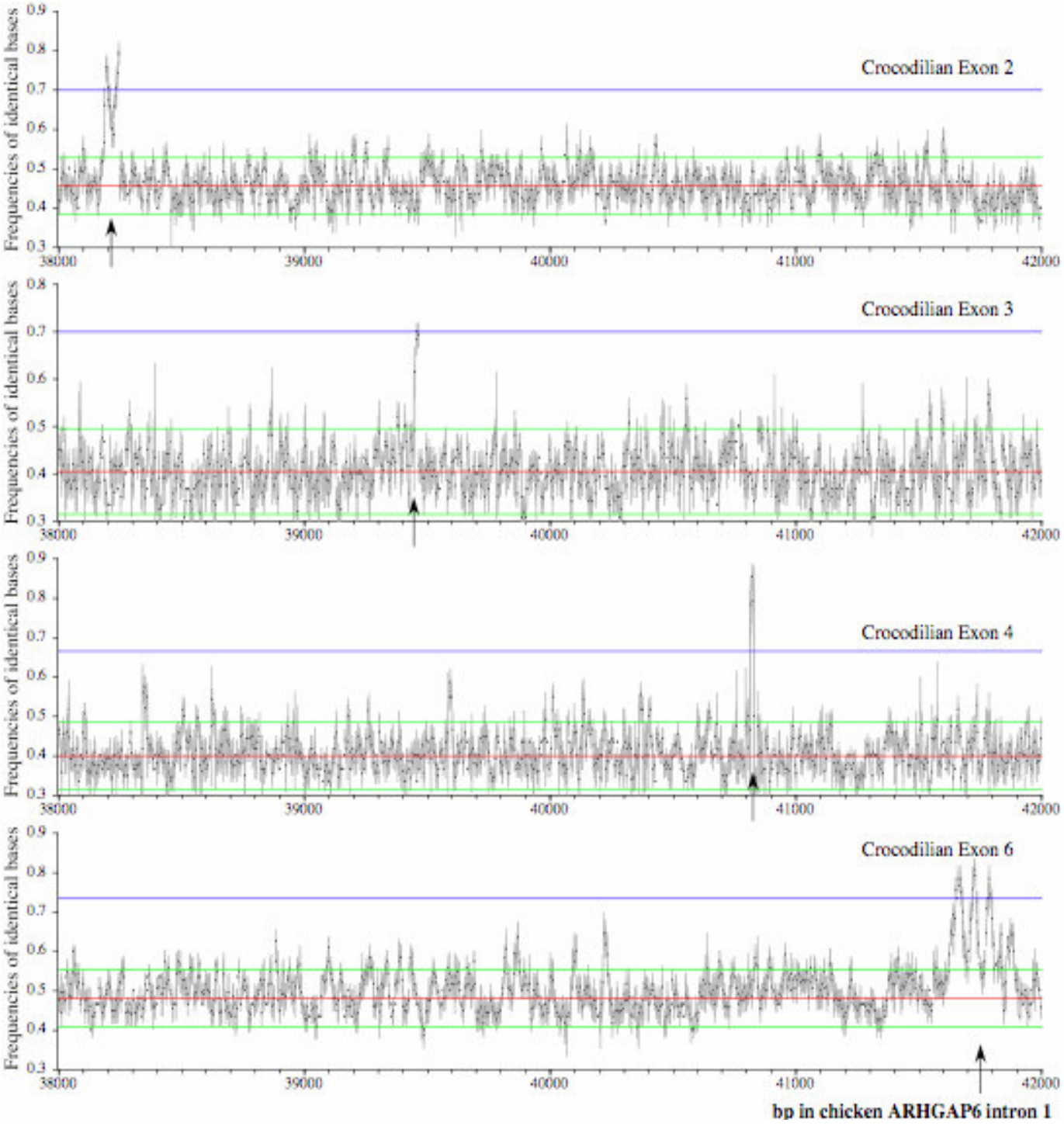

Figure 2. Similarity of the exons of crododilian amelogenin cDNA against intron 1 of chicken ARHGAP6.

duplicated exons or tracks of old insertion of retrovirus for example (Jamain et al., 2002).

UniDPlot can also be used as a combination with various other tools available to visualize alignments (Edwards et al., 2003; Jareborg and Durbin, 2000; Mayor et al., 2000). It proves to be easy to use and permits to get answer in a few minutes, even when using gigabases piece of DNA. At the present stage, no filter for low complexity regions is available and such a region must be removed prior to the analysis to search for similarities or, else, false positive hits could be infrequent. An alternative is to check *a posteriori* for the region in which

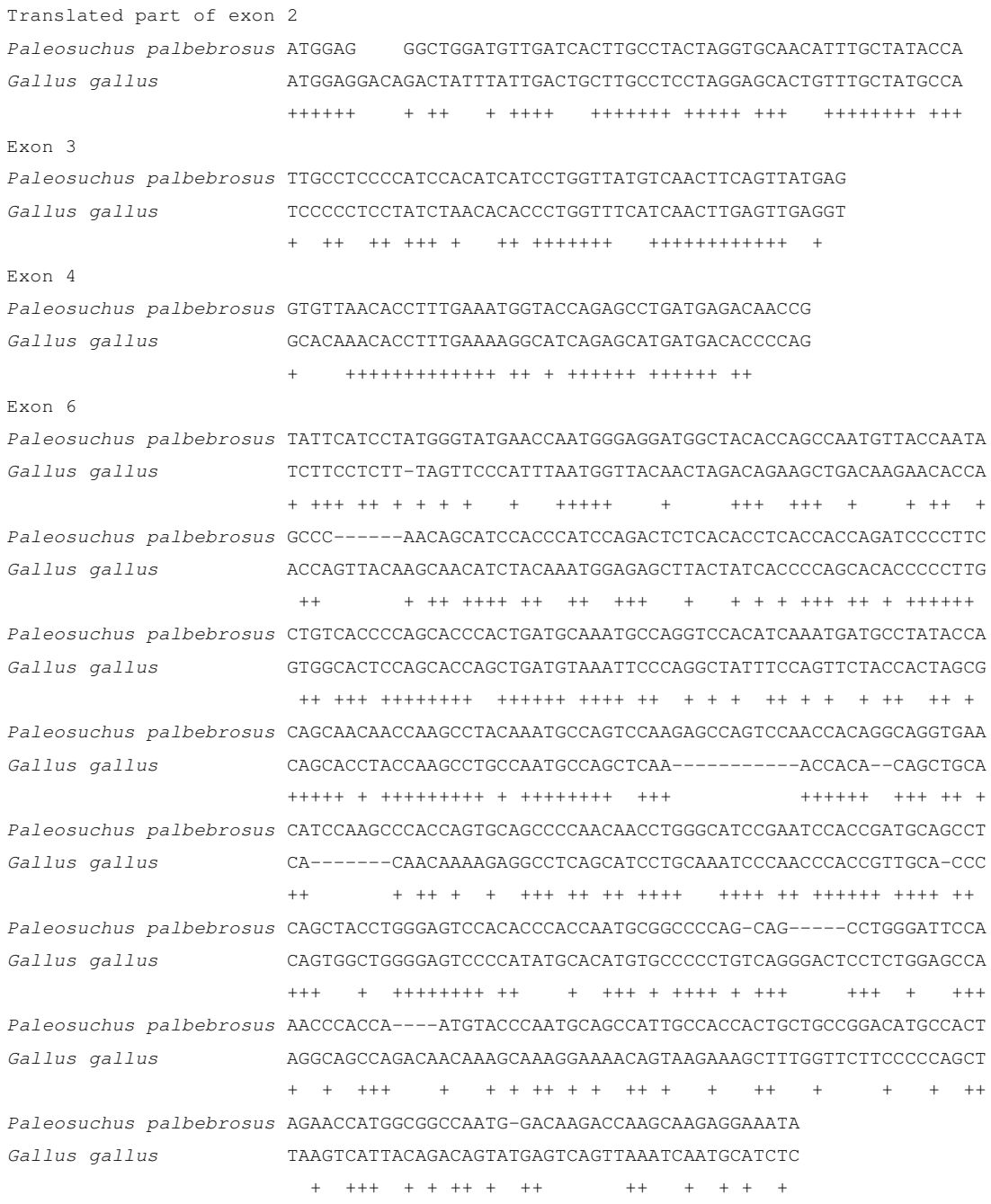

Figure 3. Alignment of crocodilian AMEL and orthologous pseugogene AMEL in chicken.

similarities have been detected.

The software is available at the bioinformatics.org web site: http://www.bioinformatics.org/unidplot/wiki/

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman. DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Edwards YJK, Carver TJ, Vavouri T, Frith M, Bishop MJ, Elgar G (2003) Theatre: a software tool for detailed comparative analysis and visualization of genomic sequence. Nucleic Acids Res. 31: 3510-3517.

Gibbs AJ, Mcintyre1 GA (1970) The Diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur. J. Biochem. 16: 1-10.

Girondot M, Sire J-Y (1998) Evolution of the amelogenin gene in toothed and tooth-less vertebrates. Eur. J. Oral Sci. 106 (1): 501-508.

Harris MP, Hasso SM, Ferguson MWJ, Fallon JF (2006). The

development of archosaurian first-generation teeth in a chicken mutant. Current Biology, 16: 371-377.
Jamain S, Girondot M, Leroy P, Clergue M, Quach H, Fellous M, Bourgeron T (2002). Transduction of the human gene AHCP by endogenous retrovirus during primate evolution. Genomics 78: 38-45.
Jareborg N, Durbin R (2000). Alfresco- A workbench for comparative genomic sequence analysis. Genome Res., 10: 1148-1157.
Maizel JV, Lenk RP (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci. U. S. A. 78: 7665-7669.
Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics, 16: 1046-1047.
Miller W (2001). Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics, 17: 391-397.
Mitsiadis TA, Chéraud Y, Sharpe P, Fontaine-Pérus J (2003). Development of teeth in chick embryos after mouse neural crest transplantations. Proc. Natl. Acad. Sci. U. S. A. 100: 6541-6545.
Mueller C, Dalkilic M, Lumsdaine A (2006). High-performance direct pairwise comparison of large genomic sequences IEEE Trans. Parallel Distrib. Syst., 17: 764-772.
Sonnhammer ELL, Durbin R (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene, 167: 1-10.