Database for chicken full-length cDNAs

Yong Wang, Zhenggang Wang, Juan Li, Yajun Wang, and Frederick C. C. Leung

Department of Zoology, The University of Hong Kong, Pokfulam, Hong Kong Special Administrative Region, China

Submitted 31 May 2006; accepted in final form 1 November 2006

Wang Y, Wang Z, Li J, Wang Y, Leung FC. Database for chicken full-length cDNAs. Physiol Genomics 28: 141–145, 2007. The generation of full-length cDNA databases is essential for functional genomics studies as well as for correct annotation of species genomic sequences. Human and mouse full-length cDNA projects have provided the biomedical research community with a large amount of gene information. Recent completion of the chicken genome sequence draft now enables a similar full-length cDNA project to be initiated for this species. In this report, we introduce the development of a chicken full-length cDNA database, which will facilitate future research work in this biological system. In this project, chicken expressed sequence tags (ESTs) were aligned onto human and mouse full-length cDNAs (or open reading frames) on the basis of their similarity. More than 588,000 chicken ESTs were aligned to ~170,000 full-length human and mouse templates obtained from the NEDO, RIKEN, and MGC databases. Many of these templates have known biological functions, and their orthologous chicken genes in the EMBL database are also provided in our database, which is available at http://bioinfo.hku.hk/chicken/. We will continue to collect known chicken full-length cDNAs to update the database for public use. The cDNA alignment results presented herein and on our database will be useful for animal science and veterinary researchers wishing to clone and confirm full-length chicken cDNAs of interest.

cDNA; expressed sequence tag; alignment

The Chicken is a Model Organism of both scientific and economic value and its genome structure, gene expression, and gene function have been extensively studied in relation to evolutionary and developmental biology and genetic improvement of economically important traits. A draft of the chicken genome sequence was first released in 2004 (5). Isolation of the full-length cDNAs and their splice variants is one of the major tasks in the postgenomic era. The human and mouse full-length cDNA projects have yielded large amounts of valuable information for biomedical researchers who employ functional genomics approaches to study key biological processes (6–8). However, the limited number of characterized chicken full-length cDNAs available in public database makes it difficult to perform systematic studies on gene expression and function in target tissues of this economically significant agricultural species, highlighting the need to establish a chicken full-length cDNA database.

The usual approach for individual researchers trying to predict the sequence of a full-length cDNA is first to explore the species’ genomic DNA sequence and then use the in silico results for PCR primer design and empirical confirmation using experimental approaches. However, the current version of the chicken genome (assembly 6) still has many gaps and assembly errors, making it difficult to perform in silico analysis for PCR primer design and prohibiting a panoramic view of the structure of genes on the genome.

In this paper we demonstrate an alternative approach for predicting chicken full-length cDNAs. We did this by aligning chicken expressed sequence tags (ESTs) onto human and mouse full-length cDNAs and open reading frames (ORFs). All data were then used to build a chicken full-length cDNA database, which currently includes basic local alignment search tool (BLAST) alignment results for >588,000 chicken ESTs over ~170,000 full-length templates derived from the New Energy and Industrial Technology Development Organization (NEDO), Rikagaku Kenkyusho (RIKEN), and Mammalian Gene Collection (MGC) databases. Subsequent experimental work will help to fill the gaps and check the quality of the predicted full-length cDNAs in this chicken database.1

Materials and Methods

Alignment of chicken ESTs on human and mouse templates. Chicken ESTs (588,739) were collected from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/), including 517,727 ESTs derived from the Delaware Chick-EST project (3). Over 30,000 human full-length cDNAs were downloaded from the NCBI and NEDO databases (http://cdn.as.ins.u-tokyo.ac.jp) (8) and >100,000 mouse full-length cDNAs were obtained from the NCBI and RIKEN databases (http://genome.gsc.riken.go.jp) (6). We also obtained ~38,000 human and mouse full-length ORFs from the MGC database (http://mgc.nci.nih.gov) (1, 7). Each full-length cDNA or ORF was used as template to find homologous chicken ESTs through BLAST searches. ESTs with matching sequences >60 bp in length (E-value<threshold) were stored for display. In addition, ESTs were selected if the distance between homologous fragments (<60 bp) of these were identical to those between the matching pieces of template (Fig. 1).

Orthologous genes to human and mouse templates on the chicken genome. Orthologous gene tables of chicken-human and chicken-mouse were retrieved from the European Molecular Biology Laboratory (EMBL) database (http://wwwensembl.org), including genomic locations of the genes and accession numbers of their corresponding protein sequences. In a two-step procedure described in Fig. 2, the orthologous chicken genes matching the human and mouse templates on the chicken genome were identified. In the first step, alignment positions of human and mouse templates on their genomes were found

1 The 2nd International Symposium on Animal Functional Genomics was held May 16–19, 2006 at Michigan State University in East Lansing, MI, and was organized by Jeanne Burton of Michigan State University and Guilherme J. M. Rosa of University of Wisconsin-Madison (see meeting report by Drs. Burton and Rosa, Physiol Genomics 28: 1–4, 2006).
Using a sequence of interest, one can find chicken full-length cDNAs or ORFs and ESTs. Because it is difficult to obtain target full-length cDNA or ORF by searching gene names, the BLAST search approach is recommended when the DNA sequence of a gene of interest is known.

An example of a graphic display of alignments between chicken ESTs and full-length templates of human is shown in Fig. 4. Alignments with different E-values are distinguished by colors. The positions and sequences of the similar parts between ESTs and templates are exhibited in a frame.

In addition, the database houses information on 2,004 experimentally confirmed chicken genes. In the EMBL, gene information is found under the “description” icon. This information derives from experimental evidence published in scientific papers. In our database, known chicken full-length cDNAs that currently include 2,327 identified by RIKEN groups (2) are available for download.

We found 4,658 and 7,324 EMBL chicken genes that could be linked to human full-length templates from NEDO and MGC, respectively. High similarity between the chicken genes and the templates was not expected. In this study, we used protein coverage percentage in p-match alignment to assess the homology between the chicken genes and the full-length templates. In the database, more than 40% protein coverage was observed in 58% of the identified EMBL chicken genes in alignments to their human templates in NEDO. A much higher percent (74%) of the chicken genes shows >40% protein coverage in alignments to human templates from MGC (Fig. 5). The ORFs for NEDO full-length cDNAs were predicted by software tools and therefore contained some incorrect predictions. Presence of incorrect ORFs can explain the lower protein coverage in alignments of NEDO templates. Therefore, our findings on the chicken genome for the full-length templates can serve as references for experimental design, but further experimental confirmation of them is appropriate and required.

The database described in this study has been used to predict full-length cDNAs of chicken genes. The predictions for two such cDNAs were then subjected to experimental verification, chicken STAT3 (AY641397) and chicken SMAD1 (AY953143). Two human MGC full-length ORFs, BC000627 and BC001878, were identified by a search of gene names, “STAT3” and “SMAD1”. Figures 6 and 7 show that many chicken ESTs align with the full-length ORFs of human STAT3 and SMAD1. In Fig. 6, ESTs 1–5 cover the whole ORF region of human STAT3, and thus chicken full-length cDNA of STAT3 can be predicted directly based on our alignment results. After further experiments on gene characterization, we noticed that full-length cDNA of chicken STAT3 is also 92% similar to the predicted one. In the case of SMAD1, two ESTs were not sufficient to predict the full-length cDNA (Fig. 7). However, this alignment information from the database allowed the design of specific primers to amplify the full-length chicken SMAD1 cDNA.

To conclude, the chicken full-length cDNA database described herein was developed with the aim of providing a platform for researchers that bridges bioinformatics and biology. Main sources of data included in the database were obtained from two chicken EST databases: BBSRC (http://chick.umist.ac.uk/) and Delaware Chick EST (http://www.chickest.udel.edu/), which are continually updated with EST information derived from different tissues and developmental stages of chickens. In the Delaware database, the

**RESULTS AND DISCUSSION**

The public website of the chicken full-length cDNA database developed in this study is: http://bioinfo.hku.hk/chicken/. Alignments between chicken ESTs and human and mouse templates can be obtained from this database in multiple ways. All full-length cDNAs and ORFs are presented in tables. Some of these have links to alignment information, and a few of them have links to orthologous chicken genes in the EMBL. Using search engines of the database, users are able to find the templates or interested ESTs through search queries of keywords such as accession number, organism, clone, or tissue (Fig. 3). The database also provides a BLAST search engine. Using a sequence of interest, one can find chicken full-length cDNAs or ORFs and ESTs. Because it is difficult to obtain target full-length cDNA or ORF by searching gene names, the BLAST search approach is recommended when the DNA sequence of a gene of interest is known.

An example of a graphic display of alignments between chicken ESTs and full-length templates of human is shown in Fig. 4. Alignments with different E-values are distinguished by colors. The positions and sequences of the similar parts between ESTs and templates are exhibited in a frame.

In addition, the database houses information on 2,004 experimentally confirmed chicken genes. In the EMBL, gene information is found under the “description” icon. This information derives from experimental evidence published in scientific papers. In our database, known chicken full-length cDNAs that currently include 2,327 identified by RIKEN groups (2) are available for download.

We found 4,658 and 7,324 EMBL chicken genes that could be linked to human full-length templates from NEDO and MGC, respectively. High similarity between the chicken genes and the templates was not expected. In this study, we used protein coverage percentage in p-match alignment to assess the homology between the chicken genes and the full-length templates. More than 40% protein coverage was observed in 58% of the identified EMBL chicken genes in alignments to their human templates in NEDO. A much higher percent (74%) of the chicken genes shows >40% protein coverage in alignments to human templates from MGC (Fig. 5). The ORFs for NEDO full-length cDNAs were predicted by software tools and therefore contained some incorrect predictions. Presence of incorrect ORFs can explain the lower protein coverage in alignments of NEDO templates. Therefore, our findings on the chicken genome for the full-length templates can serve as references for experimental design, but further experimental confirmation of them is appropriate and required.

The database described in this study has been used to predict full-length cDNAs of chicken genes. The predictions for two such cDNAs were then subjected to experimental verification, chicken STAT3 (AY641397) and chicken SMAD1 (AY953143). Two human MGC full-length ORFs, BC000627 and BC001878, were identified by a search of gene names, “STAT3” and “SMAD1”. Figures 6 and 7 show that many chicken ESTs align with the full-length ORFs of human STAT3 and SMAD1. In Fig. 6, ESTs 1–5 cover the whole ORF region of human STAT3, and thus chicken full-length cDNA of STAT3 can be predicted directly based on our alignment results. After further experiments on gene characterization, we noticed that full-length cDNA of chicken STAT3 is also 92% similar to the predicted one. In the case of SMAD1, two ESTs were not sufficient to predict the full-length cDNA (Fig. 7). However, this alignment information from the database allowed the design of specific primers to amplify the full-length chicken SMAD1 cDNA.

To conclude, the chicken full-length cDNA database described herein was developed with the aim of providing a platform for researchers that bridges bioinformatics and biology. Main sources of data included in the database were obtained from two chicken EST databases: BBSRC (http://chick.umist.ac.uk/) and Delaware Chick EST (http://www.chickest.udel.edu/), which are continually updated with EST information derived from different tissues and developmental stages of chickens. In the Delaware database, the
chicken ESTs from different tissues are even assembled based on similarity, helping to accurately annotate the chicken genome and develop ESTs for microarray spotting (3). Biological researchers now will be responsible to examine the predicted cDNAs and return feedback as to their accuracy to the database. Full-length cDNAs confirmed in this way can then be used to detect gene expression differences and splice variants in different chicken tissues and over different developmental stages. In this way, the database will ultimately provide reliable, accurate gene annotations for the chicken genome. With these virtues, databases such as this could attract more researchers to select chicken as a model species for their studies.

At present, 34 eukaryotic genome-sequencing projects have been finished (http://www.genomesonline.org/). Apart from the human and mouse genomes, on which redundant sequencing works were done, most of the genomes remain at assembly 5 or 6 (5x-6x coverage), including the chicken genome. However, such draft genomes have proven extremely informative, allowing us to profile most basic molecular features contained inside them. Even so, draft genomes are not sufficient for comprehensive studies on genes, but it will be economically unfeasible to fully sequencing these genomes as was done with...
Fig. 4. Alignments of chicken ESTs on a full-length cDNA. The colors of the bars designate E-values of the sequence alignments.

Fig. 5. Chicken gene coverage percentages for protein alignments on orthologous human templates. The pie charts show percentages of chicken genes that fall into 5 ranges of coverage among results of peptide alignments on orthologous human templates (A: NEDO human full-length cDNAs; B: MGC-Human full-length ORFs). The coverage percentage was measured as length of homologous parts to that of the whole peptide sequence.

Fig. 6. Chicken ESTs on human MGC full-length ORF BC000627. BC000627 is the full-length ORF for the human STAT3 gene. Chicken ESTs with GenBank IDs of 25752517, 25486940, 53895356, 53895346, and 25924503 were labeled with numbers 1–5, respectively. Alignment and linkage of these ESTs allowed prediction of the corresponding full-length cDNA for chicken STAT3.
human and mouse. Thus, our full-length cDNA database contributes a valuable resource to chicken researchers that may also serve as a model for other economically and societally valuable animal species.

ACKNOWLEDGMENTS

We thank Tommy Lam and staff in computer centre of the University of Hong Kong for technical support in database construction. We appreciate critical reading of this manuscript by Drs. Rajesh Jeewon, Jeanne Burton, and Guilherme Rosa.

GRANTS

This work was partly funded and supported by Research Grant Council of the Hong Kong Government, HKU7345/03M; HKU Faculty of Science Research Development Fund; and The University of Hong Kong Research Development Fund for Strategic Research Theme on Genomics, Proteomics and Bio-informatics Grant 10206152-11222-302-01.

REFERENCES


