Published online 10 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D807–D814***
doi:10.1093/nar/gkq1144

CIPRO 2.5: *Ciona intestinalis* protein database, a unique integrated repository of large-scale omics data, bioinformatic analyses and curated annotation, with user rating and reviewing functionality

Toshinori Endo[1,*], Keisuke Ueno[1], Kouki Yonezawa[1], Katsuhiko Mineta[1], Kohji Hotta[2], Yutaka Satou[3], Lixy Yamada[4], Michio Ogasawara[5], Hiroki Takahashi[6], Ayako Nakajima[7], Mia Nakachi[7], Mamoru Nomura[7], Junko Yaguchi[7], Yasunori Sasakura[7], Chisato Yamasaki[8], Miho Sera[8], Akiyasu C. Yoshizawa[9], Tadashi Imanishi[1,8], Hisaaki Taniguchi[9] and Kazuo Inaba[7]

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, [2]Department of Biosciences and Informatics, Keio University, Yokohama 223-8522, [3]Department of Zoology, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, [4]Sugashima Marine Biological Laboratory, Nagoya University, 429-63 Sugashima, Toba 517-0004, [5]Graduate School of Advanced Integration Science, Chiba University, Inage-Ku, Chiba 263-8522, [6]National Institute for Basic Biology, Nishigonaka 38, Myodaiji, Okazaki 444-8585, [7]Shimoda Marine Research Center, University of Tsukuba, 5-10-1 Shimoda, Shizuoka, 415-0025, [8]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064 and [9]Institute for Enzyme Research, the University of Tokushima, 3-18-15, Kuramoto-cho, Tokushima 770-8503, Japan

Received August 15, 2010; Revised October 9, 2010; Accepted October 25, 2010

ABSTRACT

The *Ciona intestinalis* protein database (CIPRO) is an integrated protein database for the tunicate species *C. intestinalis*. The database is unique in two respects: first, because of its phylogenetic position, *Ciona* is suitable model for understanding vertebrate evolution; and second, the database includes original large-scale transcriptomic and proteomic data. *Ciona intestinalis* has also been a favorite of developmental biologists. Therefore, large amounts of data exist on its development and morphology, along with a recent genome sequence and gene expression data. The CIPRO database is aimed at collecting those published data as well as providing unique information from unpublished experimental data, such as 3D expression profiling, 2D-PAGE and mass spectrometry-based large-scale analyses at various developmental stages, curated annotation data and various bioinformatic data, to facilitate research in diverse areas, including developmental, comparative and evolutionary biology. For medical and evolutionary research, homologs in humans and major model organisms are intentionally included. The current database is based on a recently developed KH model containing 36 034 unique sequences, but for higher usability it covers 89 683 all known and predicted proteins from all gene models for this species. Of these sequences, more than 10 000 proteins have been manually annotated. Furthermore, to establish a community-supported protein database, these annotations are open to evaluation by users through the CIPRO website. CIPRO 2.5 is freely accessible at http://cipro.ibio.jp/2.5.

[*] To whom correspondence should be addressed. Tel: +81 11 706 6547; Fax: +81 11 706 6546; Email: endo@ibio.jp

INTRODUCTION

The marine tunicate species *Ciona intestinalis* (Urochordata) has been an attractive research organism for developmental biology for more than a century (1,2). Because its transparent body and small number of constituting cells allow for the easy observation of its development, *C. intestinalis* had become one of the favorites of developmental biologists and thus large amounts of accumulated knowledge exist about its development and morphology (3–6). To this, the recent progress added the genome sequence and gene expression data (7–14). Furthermore, genome projects of other related species have revealed that *Ciona* is the closest to vertebrates among chordates, rather than the cephalochordates, such as amphioxus, which had been thought to be the closer relative based on morphology (10). Therefore, *C. intestinalis* turned out to be not only a good model organism for developmental biology, but also one of the most important species for understanding the origin and evolution of vertebrates.

Here we introduce the *Ciona intestinalis* protein database (CIPRO), an integrated comprehensive proteome database for this tunicate species. CIPRO is based on recently published, reliable gene models supplemented with data from other databases and also includes original experimental data, such as 2D-PAGE images combined with proteomic analyses (15,16) and 3D expression data (17) at various developmental stages and in adult tissues. In addition to the unpublished transcriptomic and proteomic data, the gene models in CIPRO have been automatically annotated based on bioinformatic data. Of these, more than 10 000 proteins have been further supplied with manually curated annotation based on expression data and biochemical and physiological knowledge. Because of the unique evolutionary position of the species and its simple body plan, the database should provide useful information not only to tunicate researchers, but also to researchers in fields such as developmental biology, evolutionary biology and medicine. Over the past 5 years, advances in comparative genomics have led to the sequencing of genomes of other marine invertebrates, including the sea urchin (18), sea anemone (19) and amphioxus (10). CIPRO is the first integrated protein database for a marine invertebrate and could therefore serve as a model for future marine invertebrate protein databases. In addition, molecular data related to homologs in humans and major model organisms have been included intentionally to facilitate medical and evolutionary research. The current CIPRO database is based on a recently developed KH model containing 36 034 unique sequences (11). However it covers 89 683 all known and predicted proteins from all gene models for this species in order to achieve higher usability for researchers using several gene models (Figure 1). All of these sequences have been automatically annotated, and more than 10 000 have been manually annotated based on large-scale transcriptomic, proteomic and bioinformatic data. In addition, we have included bioinformatic data such as 3D structural models and sequence homology data to facilitate protein comparison, as well as information about chemicals and potential antibodies that target *C. intestinalis* proteins. Finally, the CIPRO database website incorporates a functionality that enables the research community to evaluate and/or edit this information with ratings, curation and comments.

FEATURES OF THE CIPRO DATABASE

The CIPRO database has several unique features that reflect both the evolutionary position of the organism and the experimental omics data collected for the database. First, several bioinformatic analyses and tools have been used to highlight the relationship between *C. intestinalis* and other organisms, with special emphasis on humans. Sequence homology analysis based on BLAST, links to OMIM and other databases, and other bioinformatic analyses including 3D structural modeling results are presented for each protein entry. Second, omics data that include transcriptomic analyses such as EST analysis and DNA microarray data, proteomic data obtained with 2D-PAGE and large-scale LC/MS analyses, and 3D expression data, have been collected and presented with emphasis on developmental changes and distribution in adult tissues. Third, every sequence entry has been automatically annotated based on sequence homology and the presence of known functional domains; additionally, parts of the entries have been further annotated manually based on bioinformatic data, expression data and existing biochemical and physiological knowledge. Fourth, all of the data can be accessed via an original user-friendly interface that includes an extra capability for evaluation and refinement by community-wide users. Both registered and anonymous users can not only access all the data contained, but also evaluate and/or revise the contents to refine the whole database. We discuss the features in detail in the following sections.

SEQUENCE DATA

The database is made up of protein sequences basically from our recently developed KH model containing 36 034 unique sequences (11). To achieve high usability, however, it totally includes 89 683 non-redundant sequences derived from all gene models available for *C. intestinalis*, as will be discussed below in detail. As shown in Figure 1, it contains five publicly available proteomes of *C. intestinalis* plus PROCITS data set (20). Although the number is still too large for a proteome compared with other species, we have chosen to retain all the sequence entries for the following reasons. First, those proteomes share relatively small number of sequences and thus hardly reducible (Figure 1). Second, even if they are clustered together by the BLASTClust program so that only distinct sequences are included, the number of entries is still 70 493, or 79% of the combined unique sequence set. Third, each gene model bears a unique identifier and many of identifiers are used in the published independent experimental results. Therefore the original identifiers are convenient as reference for many cases. Finally, none of the gene models is perfect, and we

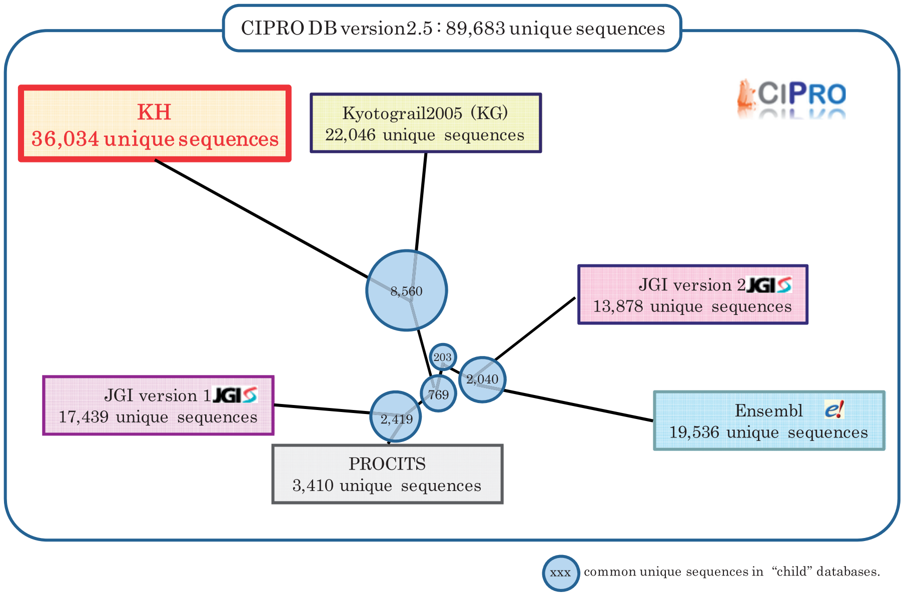


Figure 1. Similarity and the number of shared sequences among proteomes contained in CIPRO. Five proteomes and one local protein data set for *Ciona intestinalis* contained in CIPRO is shown. Branch lengths represent approximate distances between data set and their internal nodes, based on the proportion of shared sequences with their neighbors. The numbers in cyan circles show the number of sequences shared by descended nodes.

are still in the process of sorting out the entries based on manual annotation. We also expect that the capability for user-based annotation/refinement will facilitate the process by filtering out some unrealistic entries. This may reflect the uncertainty in gene prediction in the genome of this species. In addition, it might partly be explained by the existence of *trans*-splicing, which is not common for other model organisms. To examine this, integrated data representation of CIPRO, including comparative data, should be helpful for further investigation. Especially, comparison with the proteome of *C. savignyi*, a closest relative species whose whole genome sequence was determined, would silhouette the shape of true proteome of *C. intestinalis*. We therefore included the BLAST results against *C. savignyi* proteome. The target proteome is composed of known, novel and *ab initio* predicted peptides, where they are distinguishable by identifier and remark.

ORIGINAL COMPREHENSIVE USER-FRIENDLY INTERFACES

Individual protein data derived from bench experiments and bioinformatics analyses are presented in a single panel in CIPRO, as shown in Figure 2. The left side of the panel shows the basic text information, including sequence, length, deduced molecular weight, isoelectric point, summary of homology search results, domain search result, gene ontology (21), disease information for human homolog, cross references, automatic annotation, link to simple phylogeny, assignment results to the KEGG OC clusters and duplicate sequences. The right side shows the experimental results and a graphical representation of the results of bioinformatics analyses. The experimental data include 2D-PAGE images with identified spots, photographs of the cellular localization and a complex chart of mRNA and protein expression profiles based on EST, microarray and 2D-gel data. The bioinformatics analyses include cytolocalization, predicted 3D structure, predicted secondary structure with modification sites, chromosomal location of human homologs and a summary of BLAST hits. In addition to this information, a user comment section is provided so that content enrichment is possible without remodeling the system. Details for each item are described below.

ORIGINAL EXPERIMENTAL DATA

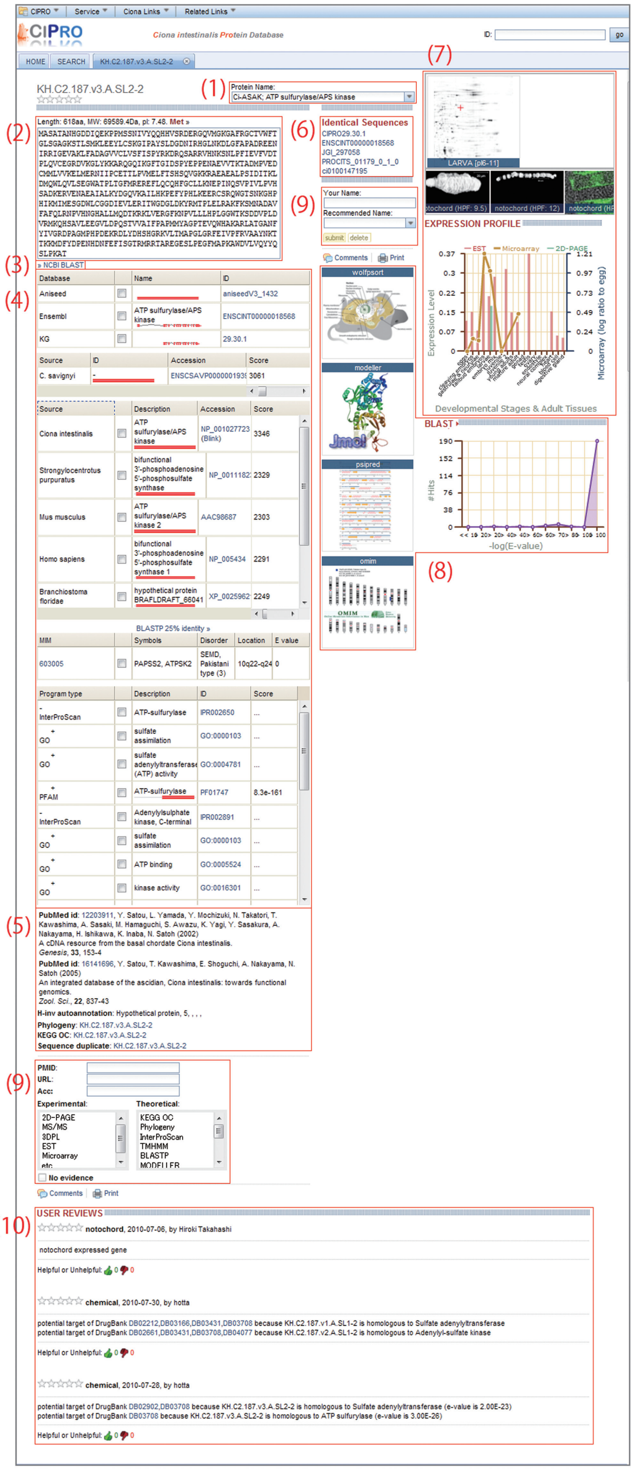
Original experimental data are shown in the right panel [Figure 2 (8)]. These data are mostly unpublished and provided by the project members.

2D-PAGE images

The photo images of 2D-PAGE gels with the highlighted spots for the protein of interest are shown in the right panel of Figure 2. There may be more than one spot for the corresponding protein, suggesting possible modification or processing of the protein. We have a separate page for 2D-PAGE analysis that includes all the identified protein spots in 2D-PAGE images with quantitative data. Comparison with other developmental stages or adult tissues is also possible.

3D protein localization

The original experimental data also includes 3D protein localization (3DPL). Spatiotemporal localization images of each protein were determined by immunolocalization and GFP-fusion protein expression [Figure 2(7)]. The 3DPL data and related information (cellular localization, staining method, developmental stage, experimental



KH.C2.187.v3.A.SL2-2

Protein Name:
Ci-ASAK; ATP sulfurylase/APS kinase

Length: 618aa, MW: 69589.4Da, pI: 7.48. Met »

```
MASATANHGDDIQEKPPMSSNIVYQQHHVSRDERGQVMGKGAFRGCTVWFT
GLSGAGKSTLSMKLEEYLCSKGIPAYSLDGDNIRHGLNKDLGFAPADREEN
IRRIGEVAKLFADAGVVCLVSFISPYRKDRQSARRVHNKSNLPFIEVFVDT
PLQVCEGRDVKGLYKKARQGQIKGFTGIDSPYEPPENAEVVTKTADMPVED
CMMLVVKELMERNIIPCETTLPVMELFTSHSQVGKKRAEAEALPSIDITKL
DMQWLQVLSEGWATPLTGFMREREFLQCQHFGCLLKNEPINQSVPIVLPVH
SADKERVENAEAIALKYDGQVKAILHKPEFYPHLKEERCSRQWGTSNKGHP
HIKMIMESGDWLCGGDIEVLERITWGDGLDKYRMTPLELRAKFKSMNADAV
FAFQLRNPVHNGHALLMQDTKRKLVERGFKNPVLLLHPLGGWTKSDDVPLD
VRMKQHSAVLEEGVLDPQSTVVAIFPAPMMYAGPTEVQWHAKARLATGANF
YIVGRDPAGMPHPDEKRDLYDHSHGRKVLTMAPGLGRFEIVPFRVAAYNKT
TKKMDFYDPENHDNFEFISGTRMRRTAREGESLPEGFMAPKAWDVLVQYYQ
SLPKAT
```

» NCBI BLAST

Database		Name	ID
Aniseed	□		aniseedV3_1432
Ensembl	□	ATP sulfurylase/APS kinase	ENSCINT00000018568
KG	□		29.30.1

Source	ID	Accession	Score
C. savignyi	-	ENSCSAVP000000193	3061

Source		Description	Accession	Score
Ciona intestinalis	□	ATP sulfurylase/APS kinase	NP_001027723 (Blink)	3346
Strongylocentrotus purpuratus	□	bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthase	NP_0011182	2329
Mus musculus	□	ATP sulfurylase/APS kinase 2	AAC98687	2303
Homo sapiens	□	bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthase 1	NP_005434	2291
Branchiostoma floridae	□	hypothetical protein BRAFLDRAFT_66041	XP_0025962	2249

BLASTP 25% identity »

MIM		Symbols	Disorder	Location	E value
603005	□	PAPSS2, ATPSK2	SEMD, Pakistani type (3)	10q22-q24	0

Program type		Description	ID	Score
- InterProScan	□	ATP-sulfurylase	IPR002650	...
+ GO	□	sulfate assimilation	GO:0000103	...
+ GO	□	sulfate adenylyltransferase (ATP) activity	GO:0004781	...
+ PFAM	□	ATP-sulfurylase	PF01747	8.3e-161
- InterProScan	□	Adenylylsulphate kinase, C-terminal	IPR002891	...
+ GO	□	sulfate assimilation	GO:0000103	...
+ GO	□	ATP binding	GO:0005524	...
+ GO	□	kinase activity	GO:0016301	...

PubMed id: 12203911, Y. Satou, L. Yamada, Y. Mochizuki, N. Takatori, T. Kawashima, A. Sasaki, M. Hamaguchi, S. Awazu, K. Yagi, Y. Sasakura, A. Nakayama, H. Ishikawa, K. Inaba, N. Satoh (2002)
A cDNA resource from the basal chordate Ciona intestinalis.
Genesis, 33, 153-4

PubMed id: 16141696, Y. Satou, T. Kawashima, E. Shoguchi, A. Nakayama, N. Satoh (2005)
An integrated database of the ascidian, Ciona intestinalis: towards functional genomics.
Zool. Sci., **22**, 837-43

H-inv autoannotation: Hypothetical protein, 5, , , ,
Phylogeny: KH.C2.187.v3.A.SL2-2
KEGG OC: KH.C2.187.v3.A.SL2-2
Sequence duplicate: KH.C2.187.v3.A.SL2-2

Identical Sequences
CIPRO29.30.1
ENSCINT00000018568
JGI_297058
PROCITS_01179_0_1_0
ci0100147195

Your Name:
Recommended Name:
submit delete

PMID:
URL:
Acc:

Experimental:
2D-PAGE
MS/MS
3DPL
EST
Microarray
etc

Theoretical:
KEGG OC
Phylogeny
InterProScan
TMHMM
BLASTP
MODELLER

□ **No evidence**

USER REVIEWS

☆☆☆☆☆ **notochord**, 2010-07-06, by Hiroki Takahashi

notochord expressed gene

Helpful or Unhelpful: 0 0

☆☆☆☆☆ **chemical**, 2010-07-30, by hotta

potential target of DrugBank DB02212,DB03166,DB03431,DB03708 because KH.C2.187.v1.A.SL1-2 is homologous to Sulfate adenylyltransferase
potential target of DrugBank DB02661,DB03431,DB03708,DB04077 because KH.C2.187.v2.A.SL1-2 is homologous to Adenylyl-sulfate kinase

Helpful or Unhelpful: 0 0

☆☆☆☆☆ **chemical**, 2010-07-28, by hotta

potential target of DrugBank DB02902,DB03708 because KH.C2.187.v3.A.SL2-2 is homologous to Sulfate adenylyltransferase (e-value is 2.00E-23)
potential target of DrugBank DB03708 because KH.C2.187.v3.A.SL2-2 is homologous to ATP sulfurylase (e-value is 3.00E-26)

Helpful or Unhelpful: 0 0

Figure 2. A sample data view for protein entry KH.C2.187.v3.A.SL2-2. Sequence and functional information are shown on the left side of the window and experimental data are summarized as informative graphics in the right panel. Components indicate (1) protein short name and description, (2) amino acid length, deduced molecular weight, deduced isoelectric point, existence of stop codon and amino acid sequence, (3) link to BLAST search site at NCBI with the sequence filled in the query form, (4) homolog and motif information, (5) miscellaneous literal information including disease, automatic annotation, phylogeny, hits to KEGG Ortholog Cluster and duplicated sequences, (6) identical sequence entries, (7) experimental results, (8) graphical results of bioinformatics analyses, (9) user annotation facility and (10) user comments in which formatted text with links and pictures can be integrated.

condition, corresponding articles, etc.) are linked to the information for corresponding developmental stage of the *C. intestinalis* embryo (17). These data help users to understand the cellular and developmental functions of each protein and can also be used as control data for comparing phenotypes among mutants in knockdown or overexpression experiments.

Expression profiles for EST, microarray and quantified protein data

The graph labeled 'Expression Profile' is a summary of gene expression data from EST, microarray and 2D-PAGE protein quantification data. The raw value is displayed by mousing over the graph. By summarizing all of these data in a single chart, differences in expression between mRNA and protein are easily observable, though they may also reflect experimental fluctuations. Note that the data are based on real laboratory experiments, some columns or categories may be missing because of the absence of the data observed or obtained.

WHOLE *C. INTESTINALIS* PROTEOME DATA WITH BIOINFORMATICS ANALYSIS RESULTS

We made the database based on our recently developed KH model containing 36 034 unique sequences (11). However several sets of gene models exist for *C. intestinalis*, as mentioned above. Therefore we finally incorporated all the existing protein models available to date, including those from Kyoto Gene (KG) (22), KH (successor of KG, http://ghost.zool.kyoto-u.ac.jp/indexr1.html) (11), PROCITS (20), JGI versions 1 and 2 (http://genome.jgi-psf.org/ciona4/ciona4.home.html, http://www.broad.mit.edu/annotation/ciona/) and Ensembl (version 58.2). Identical sequences across gene models were unified to produce a total of 89 683 protein entries. The entries are accessible by all names and accession numbers in the original gene models. Automated annotation was done to these entries according to the criteria in Table 1 and as shown in Figure 2(5).

Identifiers for sequence data

The amino acid sequences in the CIPRO database are derived from all *C. intestinalis* gene models available as of April 2010. To maintain consistency and avoid confusion, the original identifiers for all gene models have been retained, with the exception of the KG2005 gene models, to which the prefix 'CIPRO' has been used instead of 'KG2005'. In some cases, genes containing more than one coding sequence (i.e. those separated by stop codons) are found in the original gene models. In the present CIPRO database, these are treated as separate sequences and marked with numerical suffixes (e.g. .1, .2 and so on). For consistency's sake, the entire sequence from the original gene model, including stop codons, is indicated with the suffix dot-zero (.0).

Table 1. Criteria for automated annotation

Category	Criteria	Notation	Unique entries
I	≥50% identity, ≥50% coverage[a]	HOMOLOGOUS TO	10 170
II	≥25% identity	SIMILAR TO	11 077
III	Found a motif or domain in databases	XXX domain containing proteins	18 927
IV	Predicted proteins with evolutionary conservation	Conserved hypothetical proteins	6372
V	Predicted proteins longer than or equal to 80 amino acids	Hypothetical proteins	28 430
VI	Predicted proteins shorter than 80 amino acids	Hypothetical short proteins	14 697

The higher category always takes precedence for the annotation of each protein.
[a]Homology to predicted proteins are not counted in this category.

BIOINFORMATICS ANALYSES IN TEXT

Figure 2 shows a screenshot of a typical protein entry. The top of the left panel [Figure 2 (2)] shows the protein information, including deduced amino acid sequence, length, calculated molecular weight, isoelectric point and protein name candidates. A link to the NCBI BLAST server is also provided with the sequence field already filled in, so that users can execute their own homology search [Figure 2 (3)].

Summary of homologs

The bottom-right corner of the panel [Figure 2(8)] shows the top hits from homology searches for each selected model organism, making the protein names in each species easily recognizable. A histogram of BLAST hits is also shown on the right panel to allow for the identification of potential protein families.

Comparative analysis data for disease association

The 'OMIM' tag shows the information for human homologs and associated disease information with a direct link to the corresponding NCBI webpage [Figure 2 (4)]. Where available, the loci of the human homologs are also shown graphically on the right panel.

Domains and motifs

A summary of the domains and motifs identified by InterProScan 4.5 (InterPro version 22.0) (23) is shown with the corresponding InterPro identifiers and definitions based on information from PFAM (24), GO, PROSITE (25), PANTHER (26) and SUPERFAMILY (27) [Figure 2(4)]. These categories are also used in the automated annotation. Any identified domains and motifs are also indicated in the box labeled 'psipred' in the right panel.

Automated annotation by H-invitational database scheme

Automated annotation was done by H-invitational database scheme (28,29), except the criteria were modified as shown in Table 1 and Figure 2 (4). For cases in which more than one reference source was available, the top-most category was applied. For example, if a protein is similar to a predicted protein and contains a motif, it was classified as category III.

Additional text information

A phylogeny with a limited number of homologs, homology to the KEGG ortholog cluster (KEGG OC, ftp://ftp.genome.jp/pub/kegg/genes/oc/oc.gz) identified by utilization of KAAS (30) and putative duplicated genes are provided as links for users to obtain further biological implication [Figure 2 (5)].

BIOINFORMATICS ANALYSES IN GRAPHICAL VIEW

One of the unique features of CIPRO is its graphical view of results from bioinformatics analyses [Figure 2 (9)]. Each icon-like picture summarizes a separate bioinformatics analysis to allow for an easy grasp of the protein character at a glance. Each component is described separately below.

Cytolocalization (labeled as 'wolfpsort')

The subcellular localization predicted by WoLF PSORT (31) is shown graphically as the color intensity of each organelle or cellular compartment. This original graphical representation was developed by us. The more intense is the color of a cellular part, the more probable it is that the protein is localized in that particular compartment. Some proteins are predicted to be localized to multiple compartments.

Transmembrane prediction (labeled as 'tmhmm')

Localization of plasma membrane and transmembrane components predictions by TMHMM 2.0c (32,33) are shown graphically by using our original software tool. This feature can be used together with other annotations (including cytolocalization and text annotation) to identify protein function such as cytokine receptors and cell adhesion molecules.

Predicted 3D structure (labeled as 'modeller')

3D structures modeled by Modeller 9v7 (34–36) are also presented in the graphical view. Clicking on the picture opens a Jmol (http://www.jmol.org/) window, allowing the user to manipulate the picture for rotation and magnification, change color to emphasize specific atoms or residues, etc.

Secondary structure and modification sites (labeled as 'psipred')

The secondary structure, possible modification sites, and domains and motifs predicted by Psipred, Netphos and InterProScan, respectively, are summarized in a single graphic picture labeled as 'psipred'. However, this label is not meant to imply that a single program was used to produce the figure. We developed a new program to generate summarized picture for the current project.

Chromosomal map location of human homologs

The picture labeled 'omim' depicts the chromosomal map location of human homologs of each protein. For multigene families, more than one location may be indicated.

Commercially available antibodies possibly targeting *C. intestinalis* proteins

Dr Di Jiang of the Sars International Centre for Marine Molecular Biology, Norway, has generously provided information about commercially available antibodies that have the potential to cross-react to *Ciona* proteins. This information was primarily obtained by homology searches with known epitopic sequences and does not guarantee that the antibody will cross-react, but it should be useful for experimental design.

COMMUNITY-WIDE CURATION CAPABILITY

To facilitate the improvement of annotation by visiting users, we have implemented a capability for users to input additional annotation and/or comments, which will then be subjected to rating by subsequent users. To aid the curation process, literature information, matched motif patterns and other related protein information are shown with links. To aid in annotation quality control, the annotator can record his/her name with the annotation. As a part of the CIPRO project, the members, mostly experimental biologists, have manually annotated more than 10 000 entries to date. During this annotation process, we found the information on specific expression patterns during development to be especially useful.

USEFUL SEARCH FUNCTIONS

The search function can be used to find keywords in any field, including protein name, annotator name, the number of annotations, category for the automated annotation, deduced amino acid length, calculated and observed molecular weights, isoelectric point, homolog name with specifiable expectation value threshold, expressed tissue and/or developmental stage and provided data type. The last one is especially useful for finding particular data sets that contain information of interest. The search can also be done with combinations of parameters. BLAST search and fragment mass search functions are also available. Search results are downloadable in CSV format.

PROTEIN NAMES IN REGULAR ANNOTATION

Protein names were annotated with the abbreviated name (ANAME), followed by a semicolon and the descriptive name (DNAME) with annotation category, as follows:

> ALDH4A1; HOMOLOGOUS TO delta-1-pyrroline-5-carboxylate dehydrogenase, mitochondrial.

For cases in which more than one name exists for a homolog, each name is listed with comma separators. If only DNAMEs are available, a semicolon and a space are placed in front of the line. For partial sequences, a comma and the keyword 'partial' are suffixed. Referred information sources were checked upon annotation. When more than one reference source was available, the topmost category was applied. For example, if a protein was similar to a predicted protein and it contained a motif, it was noted as category III. If experimental information is used as evidence, it is noted in the comment field, not in the annotation.

Because a standard nomenclature for *C. intestinalis* proteins has not yet been proposed, some gene names have a prefix of 'ci-' (for *C. intestinalis*), whereas others do not. Considering the nomenclature consensus that exists for other model organisms, we think it is important to start discussing a standard nomenclature for this species. In this context, we should point out that the CIPRO database will also serve as the thesaurus for *Ciona* protein names. In the current database, the established names are retained, but the ci- prefix is omitted for new protein names.

SOFTWARE REQUIREMENTS AND USER MANUAL

The CIPRO website is best viewed with Firefox, Google Chrome or Internet Explorer with Java and JavaScript functionalities turned on. It uses the Dojo library including Ajax functionality via Google API for fast worldwide access. Safari browser prior to version 5.0 do not display the content properly because of their incompatibility with the Dojo library functions. The system for 2D-PAGE data management was originally based on the MAKE2DB tool with extensive modifications. The user manual is supplied both in English and Japanese in the Help/FAQ menu.

ACKNOWLEDGEMENTS

We appreciate the anonymous reviewer for helpful suggestions. We thank Dr Di Jiang for the antibody data and Dr Kaoru Azumi for the microarray data. We are also grateful to Drs Alu Konno, Zhu Lihong, Akiko Hozumi, Kogiku Shiba and Daisuke Shibata for providing experimental data, Ms Feifei Zhang and Ms Yu Yoshida for documentation, usability test over browsers and suggestions on the web interface improvement, and Ms Rie Tsuchiya and Ms Rie Oshima for assistance with the official management of the project.

FUNDING

Institute for Bioinformatics Research and Development (BIRD); Japan Sciences and Technology Agency (JST); Ministry of Education, Culture, Sports, Science and Technology, Japan, Grants-in-Aid (2D-PAGE data) (No. 22112004 to K.I., partial). Funding for open access charge: Institute for Bioinformatics Research and Development (BIRD); Japan Sciences and Technology Agency (JST).

Conflict of interest statement. None declared.

REFERENCES

1. Conklin,E.G. (1905) The organization and cell lineage of the ascidian egg. *J. Acad. Nat. Sci., Philadelphia.*, **13**, 1–119.
2. Satoh,N., Satou,Y., Davidson,B. and Levine,M. (2003) Ciona intestinalis: an emerging model for whole-genome analyses. *Trends Genet.*, **19**, 376–381.
3. Dedecker,P., Hotta,J., Flors,C., Sliwa,M., Uji-i,H., Roeffaers,M.B., Ando,R., Mizuno,H., Miyawaki,A. and Hofkens,J. (2007) Subdiffraction imaging through the selective donut-mode depletion of thermally stable photoswitchable fluorophores: numerical analysis and application to the fluorescent protein Dronpa. *J. Am. Chem. Soc.*, **129**, 16132–16141.
4. Satou,Y., Imai,K.S. and Satoh,N. (2001) Action of morpholinos in Ciona embryos. *Genesis*, **30**, 103–106.
5. Inaba,K., Nomura,M., Nakajima,A. and Hozumi,A. (2007) Functional proteomics in Ciona intestinalis: a breakthrough in the exploration of the molecular and cellular mechanism of ascidian development. *Dev. Dyn.*, **236**, 1782–1789.
6. Yamada,L., Shoguchi,E., Wada,S., Kobayashi,K., Mochizuki,Y., Satou,Y. and Satoh,N. (2003) Morpholino-based gene knockdown screen of novel genes with developmental function in Ciona intestinalis. *Development*, **130**, 6485–6495.
7. Yamada,L., Kobayashi,K., Satou,Y. and Satoh,N. (2005) Microarray analysis of localization of maternal transcripts in eggs and early embryos of the ascidian, Ciona intestinalis. *Dev. Biol.*, **284**, 536–550.
8. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
9. Satou,Y., Kawashima,T., Kohara,Y. and Satoh,N. (2003) Large scale EST analyses in Ciona intestinalis: its application as Northern blot analyses. *Dev. Genes Evol.*, **213**, 314–318.
10. Putnam,N.H., Butts,T., Ferrier,D.E., Furlong,R.F., Hellsten,U., Kawashima,T., Robinson-Rechavi,M., Shoguchi,E., Terry,A., Yu,J.K. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.
11. Satou,Y., Mineta,K., Ogasawara,M., Sasakura,Y., Shoguchi,E., Ueno,K., Yamada,L., Matsumoto,J., Wasserscheid,J., Dewar,K. *et al.* (2008) Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biol.*, **9**, R152.
12. Sasakura,Y., Awazu,S., Chiba,S. and Satoh,N. (2003) Germ-line transgenesis of the Tc1/mariner superfamily transposon Minos in Ciona intestinalis. *Proc. Natl Acad. Sci. USA*, **100**, 7726–7730.
13. Shoguchi,E., Kawashima,T., Satou,Y., Hamaguchi,M., Sin,I.T., Kohara,Y., Putnam,N., Rokhsar,D.S. and Satoh,N. (2006) Chromosomal mapping of 170 BAC clones in the ascidian Ciona intestinalis. *Genome Res.*, **16**, 297–303.
14. Azumi,K., Takahashi,H., Miki,Y., Fujie,M., Usami,T., Ishikawa,H., Kitayama,A., Satou,Y., Ueno,N. and Satoh,N. (2003) Construction of a cDNA microarray derived from the ascidian Ciona intestinalis. *Zoolog. Sci.*, **20**, 1223–1229.
15. Hozumi,A., Padma,P., Toda,T., Ide,H. and Inaba,K. (2008) Molecular characterization of axonemal proteins and signaling molecules responsible for chemoattractant-induced sperm activation in Ciona intestinalis. *Cell Motil. Cytoskeleton*, **65**, 249–267.
16. Nomura,M., Nakajima,A. and Inaba,K. (2009) Proteomic profiles of embryonic development in the ascidian Ciona intestinalis. *Dev. Biol.*, **325**, 468–481.
17. Hotta,K., Mitsuhara,K., Takahashi,H., Inaba,K., Oka,K., Gojobori,T. and Ikeo,K. (2007) A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. *Dev. Dyn.*, **236**, 1790–1805.
18. Sodergren,E., Weinstock,G.M., Davidson,E.H., Cameron,R.A., Gibbs,R.A., Angerer,R.C., Angerer,L.M., Arnone,M.I., Burgess,D.R., Burke,R.D. *et al.* (2006) The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, **314**, 941–952.
19. Putnam,N.H., Srivastava,M., Hellsten,U., Dirks,B., Chapman,J., Salamov,A., Terry,A., Shapiro,H., Lindquist,E., Kapitonov,V.V. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.
20. Hozumi,A., Satouh,Y., Ishibe,D., Kaizu,M., Konno,A., Ushimaru,Y., Toda,T. and Inaba,K. (2004) Local database and the search program for proteomic analysis of sperm proteins in the ascidian Ciona intestinalis. *Biochem. Biophys. Res. Commun.*, **319**, 1241–1246.
21. The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
22. Satou,Y., Yamada,L., Mochizuki,Y., Takatori,N., Kawashima,T., Sasaki,A., Hamaguchi,M., Awazu,S., Yagi,K., Sasakura,Y. *et al.* (2002) A cDNA resource from the basal chordate Ciona intestinalis. *Genesis*, **33**, 153–154.
23. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
24. Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
25. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
26. Mi,H., Dong,Q., Muruganujan,A., Gaudet,P., Lewis,S. and Thomas,P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, D204–D210.
27. Wilson,D., Pethica,R., Zhou,Y., Talbot,C., Vogel,C., Madera,M., Chothia,C. and Gough,J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.
28. Yamasaki,C., Murakami,K., Takeda,J., Sato,Y., Noda,A., Sakate,R., Habara,T., Nakaoka,H., Todokoro,F., Matsuya,A. *et al.* (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res.*, **38**, D626–D632.
29. Yamasaki,C., Kawashima,H., Todokoro,F., Imamizu,Y., Ogawa,M., Tanino,M., Itoh,T., Gojobori,T. and Imanishi,T. (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *Nucleic Acids Res.*, **34**, W345–W349.
30. Moriya,Y., Itoh,M., Okuda,S., Yoshizawa,A.C. and Kanehisa,M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.
31. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
32. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

33. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
34. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
35. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
36. Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.