Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W345–W349***
doi:10.1093/nar/gkl283

TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB

Chisato Yamasaki[1,2], Hiroaki Kawashima[1,3], Fusano Todokoro[3], Yasuhiro Imamizu[3], Makoto Ogawa[3], Motohiko Tanino[1,2], Takeshi Itoh[2,4], Takashi Gojobori[2,5,6] and Tadashi Imanishi[2,*]

[1]Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [2]Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [3]DYNACOM Co., Ltd, 643 Mobara, Mobara-shi, Chiba 297-0026, Japan, [4]Genome Research Department, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan, [5]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan, and [6]Department of Genetics, The Graduate University for Advanced Studies, 1111 Yata, Mishima Shizuoka 411-8540, Japan

Received February 14, 2006; Revised February 25, 2006; Accepted April 5, 2006

ABSTRACT

Transcriptome Auto-annotation Conducting Tool (TACT) is a newly developed web-based automated tool for conducting functional annotation of transcripts by the integration of sequence similarity searches and functional motif predictions. We developed the TACT system by integrating two kinds of similarity searches, FASTY and BLASTX, against protein sequence databases, UniProtKB (Swiss-Prot/TrEMBL) and RefSeq, and a unified motif prediction program, InterProScan, into the ORF-prediction pipeline originally designed for the 'H-Invitational' human transcriptome annotation project. This system successively applies these constituent programs to an mRNA sequence in order to predict the most plausible ORF and the function of the protein encoded. In this study, we applied the TACT system to 19 574 non-redundant human transcripts registered in H-InvDB and evaluated its predictive power by the degree of agreement with human-curated functional annotation in H-InvDB. As a result, the TACT system could assign functional description to 12 559 transcripts (64.2%), the remainder being hypothetical proteins. Furthermore, the overall agreement of functional annotation with H-InvDB, including those transcripts annotated as hypothetical proteins, was 83.9% (16 432/19 574). These results show that the TACT system is useful for functional annotation and that the prediction of ORFs and protein functions is highly accurate and close to the results of human curation. TACT is freely available at http://www.jbirc.aist.go.jp/tact/.

INTRODUCTION

Automatic prediction of functions of transcripts is extremely important and useful; it has a wide variety of applications in studies based on sequence data of the genome, cDNAs and ESTs in various species, especially in human. Studies on human transcripts have been systematically and extensively carried out to draw the outline of the human transcriptome (1–5). Some other studies have reported the functional annotation of *Mus musculus* (6) and *Arabidopsis* (7) full-length cDNAs. However, the functional annotation of those transcripts relied heavily on human curation because currently there is no freely available web server to provide the automated functional annotation.

The human transcriptome consists of protein-coding and non-protein-coding functional RNAs. Several sequence analysis techniques are available to provide insights for predicting the function of the transcripts to some extent. For example, sequence similarity search tools, such as BLASTX (8) and FASTY (9) provide the homologs of the transcripts in protein databases and motif prediction programs, such as InterProScan (10) will provide the predicted functional motifs in the protein-coding sequence (CDS) of the transcripts. However, any one of these alone is not enough to comprehensively judge and assign the function of the transcripts as

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: imanishi@jbirc.aist.go.jp

protein-coding genes. Thus, an integration of sequence analysis tools is necessary.

We have previously reported the integrative annotation of human genes by the international cooperative project entitled 'Human Full-length cDNA Annotation Invitational' (abbreviated as H-Invitational or H-Inv) and the construction of an integrative database of the human transcriptome, named H-Invitational Database (H-InvDB) (11,12). In the H-Invitational project, we collected information about human full-length cDNAs, and conducted extensive bioinformatics analyses by making full use of biological databases and computational tools and rearranging annotation by biologists (13). A standard for human curation was proposed, established and applied to annotate all the collected H-Inv cDNAs. We assigned the standardized functional annotation to 19 574 representative H-Inv proteins by human curation, based on the results of similarity search and InterProScan (11,12). In this study, we describe the newly developed transcriptome auto-annotation conducting tool (TACT), a web-based automated prediction tool for functional annotation that was originally designed for the H-Invitational project. We developed the TACT system by integrating two kinds of similarity searches, BLASTX (8) and FASTY (9), against protein sequence databases and a unified motif prediction program, InterProScan (10), into the ORF-prediction pipeline. This system successively applies these constituent programs to an mRNA or cDNA sequence in order to predict the most plausible ORF and the function as a protein. Furthermore, we applied the TACT system to 19 574 non-redundant human transcripts registered in H-InvDB, and evaluated its predictive power by the degree of agreement with the functional annotation results resulting from human curation.

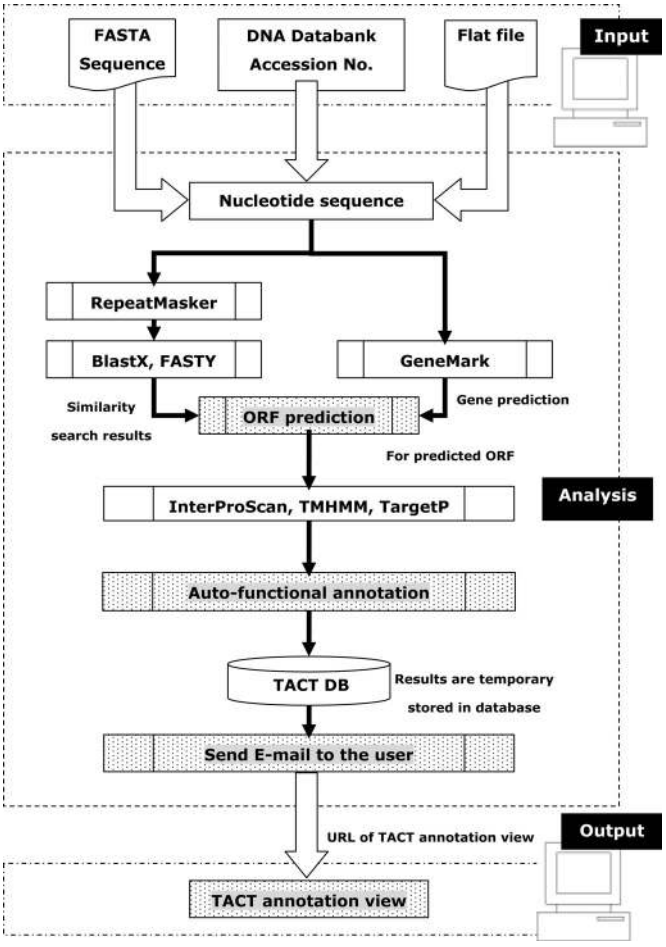

Figure 1. The TACT annotation pipeline. The flowchart illustrates the TACT computational analysis and web server interfaces. The white arrows indicate the input sequence data to TACT and output annotation data from TACT to users. The thick solid arrows indicate the data flow within the TACT server during analysis.

TACT COMPUTATIONAL PIPELINE

The TACT computational pipeline was developed by integration of two sequence similarity searches, BLASTX and FASTY, and a motif prediction by InterProScan. The computational analyses can be divided into three pipelines; sequence analysis, ORF prediction and auto-functional annotation pipelines, and is carried out as follows (Figure 1).

TACT sequence analysis pipeline

All the repetitive and low-complexity sequences in a query sequence are masked using RepeatMasker (http://www.repeatmasker.org) with Repbase. The masked sequence is then subjected to BLASTX and FASTY against UniProtKB (Swiss-Prot/TrEMBL) and RefSeq (human) protein entries. In parallel, GeneMark (14) ORF prediction is carried out.

TACT ORF prediction pipeline

Then, TACT proceeds to predict the ORF of each sequence by using a custom-made Perl script based on the similarity with UniProtKB (Swiss-Prot/TrEMBL) and RefSeq (human) protein entries and prediction by GeneMark (14). The translated region (ORF) of each sequence is predicted by one of five methods in order of priority [illustrated in Supplementary Figure 1A in our previous study (11)]: (i) prediction based on complete sequence match with known (experimentally verified) reviewed human RefSeq or UniProtKB/Swiss-Prot protein entries; (ii) prediction based on similarity with known RefSeq, UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entries; (iii) prediction based on similarity with hypothetical RefSeq, UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entries; (iv) prediction by GeneMark with probability larger than 0.5 and length longer than 80 amino acids; (v) the longest possible translation of initiation to termination codon in six frames of length greater than 80 amino acids.

TACT auto-functional annotation

For each sequence with predicted ORF, TACT conducts InterProScan (10). Then the automated-functional annotation for each sequence is predicted by using a custom-made Perl script applying five standards in order of priority. TACT assigns the most appropriate protein or domain ID, named as 'data source ID', to describe the functions of transcripts as proteins and classifies them according to five similarity criteria [illustrated in Supplementary Figure 2B in our previous study (11)]: (i) identical hit by BLASTX or FASTY (identity $\geq$98% and

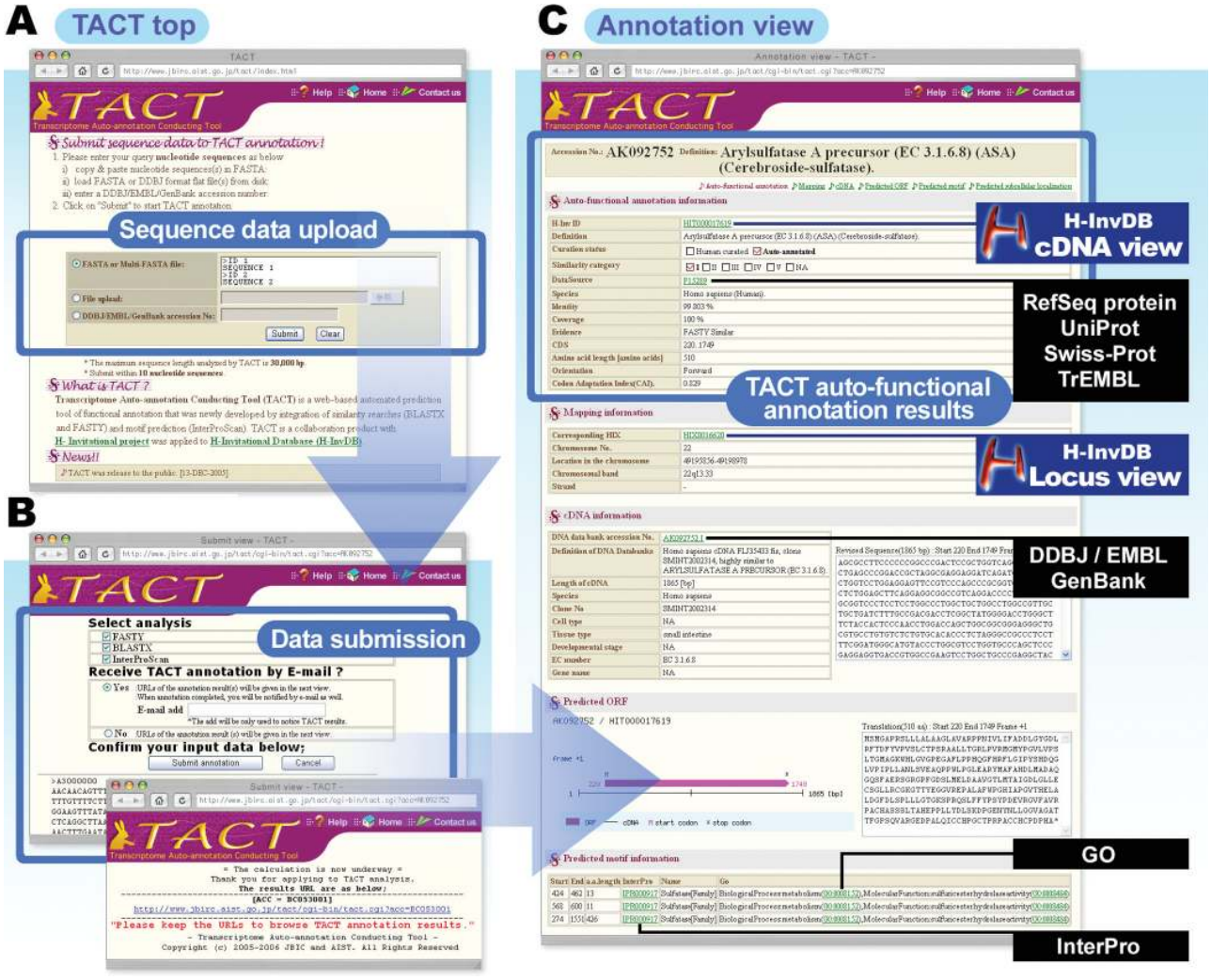

Figure 2. TACT web-based interfaces. Sample views of TACT top (**A**), data submission (**B**) and annotation view (**C**) for HIT000017619 (AK092752) are shown. The annotation view (C) shows detailed annotation information and has links to external databases as indicated. The blue arrows indicate the flows of views during the TACT analysis and black lines indicate the links to appropriate reference data in H-InvDB or external public databases.

coverage 100%) to a known human protein in reviewed RefSeq or Swiss-Prot entries (Category I); (ii) similar hit by BLASTX or FASTY (identity ≥50%) to a known protein of any species in RefSeq, UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entries (Category II); (iii) meaningful (with indication of protein functions) InterPro hit by InterProScan (Category III); (iv) similar hit by BLASTX or FASTY (identity ≥50%) to a hypothetical protein in RefSeq or UniProt entries (Category IV); (v) no data source ID assigned (Category V). Category I is solely for human transcripts, but the 'data source ID' can be assigned to the input sequence of any species without any restriction.

A problem in assigning data source IDs to transcripts is that the data sources are sometimes proteins without experimental verifications. We thus introduced a text-based judgement scheme to determine 'known proteins' and 'meaningful InterPro domain'. In practice, we avoided proteins with the following keywords that suggest proteins without experimental verification in the description: (i) hypothetical, (ii) similar to, (iii) names of cDNA clones (Rik, KIAA, FLJ, DKFZ, HSPC, MGC, CHGC and IMAGE) and (iv) IDs of InterPro domain frequent hitters. Hits to proteins with these keywords are automatically ignored.

For any sequence with no predicted ORF, the definition of the sequence is automatically annotated as 'Non-protein-coding transcript'.

Several additional sequence analyses were integrated into the TACT annotation system to provide useful reference data. For example, Gene ontology (GO) terms are assigned through the relations of InterPro IDs to GO terms. All the results of the analysis and annotated data are temporarily stored in a PostgreSQL database and are made accessible through the TACT web-based interfaces (Figure 1).

ACCURACY OF TACT

In the H-Invitational annotation project, we assigned each transcript the most appropriate protein or domain ID, named as 'data source ID', to describe the function of cDNA as a protein. The judgement was done one by one by human curation, following a standard scheme [illustrated in Supplementary Figure 4 in our previous study (11)]. For

19 574 representative H-Inv cDNAs, we conducted functional annotation by human curation using a custom-made annotation system. In this report, we evaluated the accuracies of the TACT system by the agreements between TACT annotation and human curation for 19 574 representative H-Inv cDNAs. The overall agreement was 83.9%, i.e. 16 432 TACT annotations, including those annotated as hypothetical proteins, agreed with H-Inv human curation. This result shows that by integrating three sequence analysis programs, the prediction of functional annotation becomes highly accurate and closer to the results of human curation.

In our annotation pipeline, we classified proteins into five similarity categories (Table 1): Category I proteins were defined as 'Identical to a known human protein', Category II proteins were defined as 'Similar to a known protein'; Category III proteins were defined as 'Domain-containing proteins'; Category IV proteins were defined as 'Conserved hypothetical proteins'; Category V proteins were defined as 'Hypothetical proteins'. The agreements of TACT annotation with H-InvDB for each similarity category are summarized in Table 1. In the H-Inv annotation project we checked the abstracts in PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) corresponding to the candidate protein entries for data source ID in the human curation procedure. The agreement was the lowest in Category II. For example, for HIT000012743 (AK056129), TACT auto-annotation output Q91YE4 as data source ID, so that the definition of the cDNA was 'Similar to 67 kDa polymerase-associated factor PAF67' [Category II; Similar to a mouse (Q91YE4) protein]. However, by checking the PubMed abstracts for Q91YE4, we found that the annotation should be altered to 'TPR repeat containing protein' [Category III; InterPro domain (IPR001440)-containing protein] because the function of Q91YE4 was not yet examined experimentally. This example illustrates the importance of human curation in functional annotation of transcripts. Also, exclusion of proteins described in specific PubMed entries is a possible way of improving the accuracy of TACT auto-functional annotation pipeline. The human-curated annotation was provided at the appropriate cDNA view in H-InvDB (http://www.jbirc.aist.go.jp/hinv/soup/pub_Detail.pl?hinv_id=HIT000012743).

TACT INPUT DATA

Nucleotide sequence data consisting mRNA, cDNA or EST in one of the three formats may be uploaded and submitted to the TACT system. The users may directly copy and paste FASTA-format sequence (s), upload a FASTA or DDBJ format flat file, or enter a DDBJ/EMBL/GenBank accession number (s). The maximum sequence length analyzed by TACT is 30 000 bp and the maximum number of sequences analyzed by TACT is 10. When a DDBJ/EMBL/GenBank accession number is entered as input data, then TACT obtains the DNA databank flat file through the getentry sequence retrieval system in DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/) and all the succeeding analysis will be conducted in the appropriate cascades. When an input data file is uploaded in one of the other two formats, then TACT will conduct all the analysis for the uploaded mRNA sequence data. Although it was originally developed to predict functions of protein-coding transcripts of human cDNA, the sequence data of any species can be analysed by TACT.

TACT WEB-BASED INTERFACES

TACT web-based interfaces consist of the TACT top page, the TACT data submission and the TACT annotation view. In the TACT top page, general information about TACT and data input facilities are provided (Figure 2A). In the TACT data submission, the selections of the analysis options are provided (Figure 2B). It is recommended that users select all the options to achieve the highest accuracy. After selecting the options, users can check the input sequence data and the progress of the analysis just after submitting the annotation. By entering an e-mail address, the URL for the annotation results will be reported to users by e-mail when all the analysis has been completed. The TACT annotation view shows all the annotation of the input sequence. It consists of six sections: auto-functional annotation information section; mapping information section; cDNA information section; predicted ORF section; predicted motif information section and predicted subcellular localization section. The data include the functional description as protein-coding transcripts, similarity category, predicted ORF, translation, predicted functional motifs by InterProScan, GO and input nucleotide sequence in DNA databank (Figure 2C). If the DNA accession number of the input data are already recorded in H-InvDB, then the location in the human genome, and corresponding H-Invitational transcripts (HIT) and H-Invitational cluster (HIX) IDs with hyperlinks to H-InvDB will also be provided. As shown in the sample Annotation view in Figure 2, this view also links to many external public databases including DDBJ/EMBL/GenBank, RefSeq, UniProtKB/Swiss-Prot, InterPro and GO.

Table 1. The degree of agreement of TACT annotation and human curation

Similarity category	Description	No. of H-Inv proteins examined	No. of correctly predicted proteins by TACT	The agreement of TACT annotation and human curation (%)
I	Identical to a known human protein. (Identity ≥98% and coverage =100%)	5313	4735	89.1
II	Similar to a known protein of any species. (Identity ≥50%)	5859	3469	59.2
III	InterPro domain-containing protein	1387	1320	95.2
IV	Conserved hypothetical protein	1309	1265	98.9
V	Hypothetical protein	5706	5643	96.6
Total		19 574	16 432	83.9

CONCLUSION

In this study, we showed that by integrating three sequence analysis programs, the prediction of ORFs and protein functions becomes highly accurate and closer to the results of human curation. Because of its accuracy and usefulness, TACT will be an indispensable tool to predict the function of transcripts. The TACT system can always provide the latest reference information because protein and motif databases are updated regularly and frequently. This unique system for conducting functional annotation may have a wide range of application in transcriptome studies of human and other species. TACT is freely available at http://www.jbirc.aist.go.jp/tact/.

ACKNOWLEDGEMENTS

The authors thank Mr Ryo Aono for graphical design of the interfaces and Mr Tomohiro Endo for the technical support. The authors acknowledge all the members of the H-Invitational consortium, especially the staffs of JBIRC and DDBJ for construction of H-InvDB. This research is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC). Funding to pay the Open Access publication charges for this article was provided by JBIC.

Conflict of interest statement. None declared.

REFERENCES

1. Yudate,H.T., Suwa,M., Irie,R., Matsui,H., Nishikawa,T., Nakamura,Y., Yamaguchi,D., Peng,Z.Z., Yamamoto,T., Nagai,K. *et al*. (2001) HUNT: launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res*., **29**, 185–188.
2. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,J., Glassl,S., Ansorge,W., Bocher,M., Blocker,H., Bauersachs,S., Blum,H. *et al*. (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res*., **11**, 422–435.
3. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al*. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
4. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genet*., **36**, 40–45.
5. Hu,R.-M., Han,Z.G., SongmH,D., Pengm,Y.D., Huang,Q.H., Ren,S.X., Gu,Y.J., Huang,C.H., Li,Y.B., Jiang,C.L. *et al*. (2000) Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. *Proc. Natl Acad. Sci. USA*, **97**, 9543–9548.
6. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
7. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
9. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
10. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
11. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*., **2**, 856–875.
12. Yamasaki,C., Koyanagi,K.O., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al*. (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
13. Cyranoski,D. (2002) Geneticists lay foundations for human transcriptome database. *Nature*, **419**, 3–4.
14. Borodovsky,M., McIninch,J.D., Koonin,E.V., Rudd,K.E., Medigue,C., Danchin,A. *et al*. (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res*., **23**, 3554–3562.