A Comprehensive Software Suite for the Analysis of cDNAs

Kazuharu Arakawa[1,2], Haruo Suzuki[1,2], Kosuke Fujishima[1,2], Kenji Fujimoto[1,2], Sho Ueda[1], Motomu Matsui[1], and Masaru Tomita[1]*

[1] *Institute for Advanced Biosciences, and* [2] *Bioinformatics Program of the Graduate School of Media and Governance, Keio University, Fujisawa 252-8520, Japan.*

We have developed a comprehensive software suite for bioinformatics research of cDNAs; it is aimed at rapid characterization of the features of genes and the proteins they code. Methods implemented include the detection of translation initiation and termination signals, statistical analysis of codon usage, comparative study of amino acid composition, comparative modeling of the structures of product proteins, prediction of alternative splice forms, and metabolic pathway reconstruction. The software package is freely available under the GNU General Public License at http://www.g-language.org/data/cdna/.

Key words: cDNA, bioinformatics, software

Introduction

The success of genome sequencing projects has resulted in exponential growth of the genome data that are available (*1*), enabling computational datamining based on the "digital" data of nucleotide sequences. Bioinformatics is therefore the key to the successful and full use of sequence data for post-genomic research. It enables us to improve our understanding of the inner workings of life by providing predictions and hypotheses on the characteristics and interplay of intracellular components (*2*). Although there are established software pipelines for genome annotation, including methods of gene identification, similarity search, domain prediction, identification of repeats, and prediction of the functions of product proteins (*3–5*), a generic toolkit for post-genomic analysis based on the annotated genome data readily applicable to automated comparative study has not yet existed.

Although there is also substantial interest in genome-wide features and non-coding regions, the primary target unit in bioinformatics sequence analysis is the gene and its product protein. This focus on the coding regions is especially important for comparative genomics, considering the presence of "neglected genomes", which results in deficits in the phylogenetic sequences of biologically relevant organisms (*6*). The majority of the genetic information from neglected species is likely to be sequenced by cost-effective methods, such as expressed sequence tag (EST) analysis or genome survey sequence (GSS) analysis, rather than by complete genome sequencing projects, and there is an annotation system specifically for this purpose named prot4EST (*7*). Therefore, the mRNA or cDNA sequence is the most pertinent basic unit for comparative genomic study. Although not cost effective, a number of genome-wide cDNA projects such as FANTOM (*8*) and H-Invitational (*9*) have enhanced the cDNA projects, providing large-scale, accurate, and comprehensive datasets.

We present here a comprehensive software suite for sequence-based bioinformatics analysis of annotated cDNAs, including tools for the detection of translation initiation and termination signals, statistical analysis of codon usage, comparative study of amino acid composition, comparative modeling of product protein structures, prediction of alternative splice forms, and comparative systems biology by metabolic pathway reconstruction. Therefore, the analyses implemented in the software package cover the levels of nucleotide sequences, codons, mRNA variants, amino acid sequences, protein structures, and metabolic pathways. Although the target unit is cDNA, the package can be used for a variety of other datasets, including ESTs, GSSs, and complete genome data.

Translation initiation and termination signals and codon usage provide insights into the expression level and translation mechanism of the gene, and are also characteristic features in comparative genomics (*10*, *11*). Although the detection of translation initiation and termination signals *de novo* requires so-

*** Corresponding author.**
E-mail: mt@sfc.keio.ac.jp

phisticated machine learning methods (*12–14*), several straightforward statistical methods are employed for annotated datasets. We have combined the position weighted matrix (PWM) method (*15*), the absolute and relative entropy method (*16*), and the information content method (*17–19*) in our software. Likewise, a number of gene parameters (126 in total), including several indices of codon usage, such as three entropy functions (H, D, and D_{syn}; ref. *20*), the effective number of codons (N_c; ref. *21*), the codon bias index (CBI; ref. *22*), the intrinsic codon deviation index (ICDI; ref. *23*), and the weighted sum of relative entropy (E_w; ref. *24*), are calculated for multivariate analysis of codon usage. Indices that require reference sets of highly expressed genes, including the codon adaptation index (CAI; ref. *25*) and the predicted expression level for characterizing predicted highly expressed (PHX) genes (*26*), are also implemented as options. Data provided by these tools will also be applicable for the fields of protein engineering and gene annotation, providing gene expression prediction and protein classification data.

Computational methods for protein structure prediction are widely used to predict 3-dimensional (3D) protein structures from primary amino acid sequences. The comparative modeling (CM) method is one of the most reliable computational structure modeling methods when a homologous protein of known structure is available in the database (*27*). In our software, a pipeline to automatically run a similarity search using the MODELLER software (*28–30*) is implemented for CM, and the software generates the output in PDB (Protein Data Bank) format. On the other hand, recent studies have revealed that about 20% to 30% of open reading frames (ORFs) in completely sequenced genomes are defined as orphan genes that are mostly species-specific and have no sequence similarity with any other ORFs (*31*). Therefore, since the importance of functional annotation by similarity-independent methods is becoming prominent, in this package we also present a tool for classifying and clustering proteins using amino acid composition. Amino acid composition has been widely used for analyzing protein evolution (*32*) and gene expression (*33*) and for predicting protein function (*34*). Our method is an effective means that provides an insight into protein function, especially for those proteins only distantly related to other ones.

Alternative splicing is a common feature in higher eukaryotes, generating proteome diversity (*35*). Most of the reports on the computational prediction of alternative splice forms require various datasets that include the complete genome sequence for clustering (*36–38*), and there is no open source software for the detection of alternative splice forms and the structure of precursor mRNA without genome data. Our tool is applicable to this challenging task.

For the purpose of system-level understanding and comparison of the genome, our package also includes a metabolic reconstruction and pathway alignment tool. Although other software packages, such as PathoLogic (*39*), realize more detailed pathway reconstruction on the basis of text-mining of the annotation, our tool achieves fast reconstruction on the basis of similarity search of the coding sequence, not relying on the completeness of the annotation. IdentiCS (*40*) and metaSHARK (*41*) provide rich features for the metabolic pathway reconstruction based on similarity searches, and these kinds of software are aimed for annotation purposes rather than quick means for comparative study. Moreover, the pathway data reconstructed by our method are directly applicable with other tools in our software for the pathway alignment for comparative study.

System and Methods

Software overview

The entire package was developed by using the Perl programming language and the generic bioinformatics workbench, G-language Genome Analysis Environment version 1.51 (G-language GAE; ref. *42*) on Fedora Core 2 Linux. The software architecture is outlined in Figure 1. Generally, in a complete genome sequence flatfile, the whole entry is contained within a single locus entry, and cDNA and EST sequences are distributed in file formats with multiple locus entries, therefore making it difficult for software to work properly in both formats. Unlike that, all tools in our software package were developed to accept and work with formats of GenBank/EMBL/FASTA data. To make the package as effective as a generic toolkit for post-genomic analysis, all tools were designed to be as automated as possible and to produce graphical outputs as well as detailed text data for intuitive interpretation. Most parameters listed in this section are configurable from the software interface. The software package, documentation, and sample output are available at http://www.g-language.org/data/cdna/. The software package is distributed under the open-source GNU General Public License.

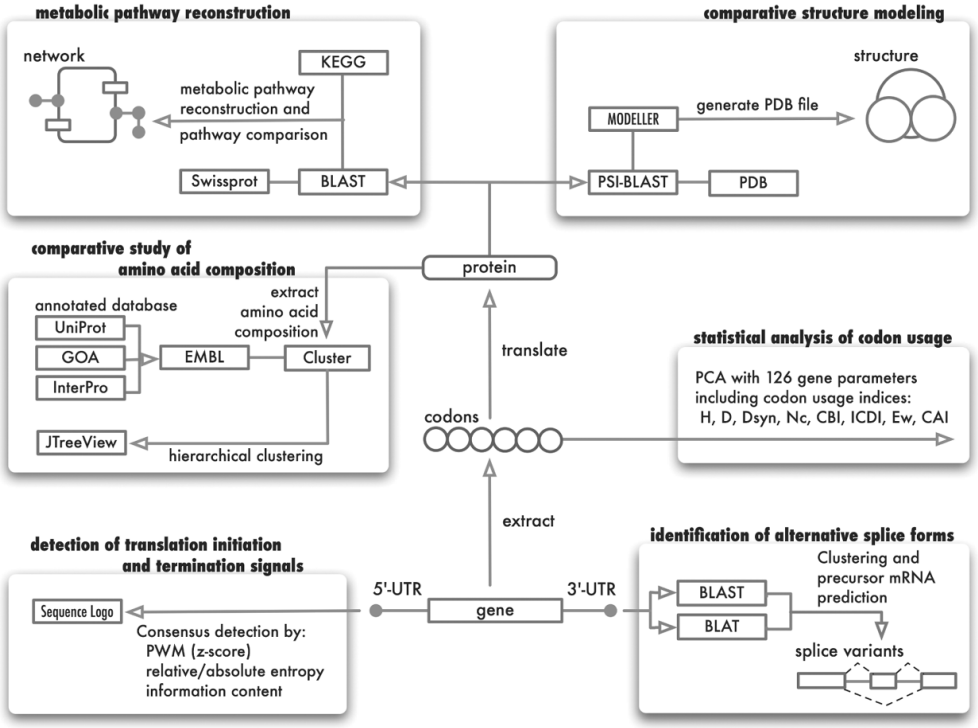

Fig. 1 Software architecture of the comprehensive software suite. The software suite includes tools for the detection of translation initiation and termination signals, statistical analysis of codon usage, comparative study of amino acid composition, comparative modeling of the structures of product proteins, prediction of alternative splice forms, and comparative systems biology by metabolic pathway reconstruction. It therefore covers the levels of nucleotide sequence, codons, mRNA variants, amino acid sequence, protein structure, and metabolic pathway. Rectangular boxes indicate external software and databases used in the package.

Detection of translation initiation and termination signals

For the detection of translation initiation and termination signals, the software goes through a pipeline to calculate PWM, absolute and relative entropy, and information content on nucleotide composition to finally display the consensus sequence with sequence logos (*43*). For all methods, the user can specify the position of the analysis to be at either the initiation or the termination site, and by default 30 bp before and after the specified site is used for the calculation of indices. Datasets with 5′-UTR or 3′-UTR sequences shorter than the length for analysis (< 30 bp by default) are omitted.

PWM is calculated for base composition, dinucleotide frequency, and trinucleotide frequency, and the statistical significance of the composition bias is assessed by calculating the standard deviation and the z-score of the most frequent element at each position. A baseline value of 0.2 is used as a cutoff value, and elements with z-scores over 1 are considered significant. For the calculation of absolute and relative entropy and information content, we have followed the methods of Schneider *et al* (*19*). For all calculated indices, the strength of consensus is displayed using the GNUplot graphing utility (http://www.gnuplot.info/). Finally, the strength of consensus and the most frequent base at each position are displayed with sequence logos to visually identify the initiation or termination signals.

Statistical analysis of codon usage

To identify the major sources of variation in codon usage among genes, the principal component analysis (PCA) of codon usage data for the cDNAs is implemented to analyze the correlations between PCA axis scores and a number of gene parameters.

Taking into account the advice of Perriere and Thioulouse (*44*) to compute in parallel multivariate analysis on different codon usage data and then to compare the results, PCA is performed on three kinds of normalized codon usage data, R1, R2, and R3, as

defined previously (*45*). In R1, each codon frequency in a gene is normalized by the total codon frequency in the gene to exclude the effect of gene length. In R2 and R3, the codon frequencies are normalized for each amino acid to exclude the effect of amino acid composition of a gene. R2 and R3 have often been termed as relative codon frequency (RF) and relative synonymous codon usage (RSCU), respectively (*44*).

A number of gene parameters (126 in total), including gene length, several indices of codon usage, amino acid usage, base usage, and dinucleotide usage, are implemented in this tool (see http://www.g-language.org/data/cdna/Additional_data_1.xls for a comprehensive listing) as follows:

1. Indices of codon usage include three entropy functions (H, D, and D_{syn}), N_c, CBI, ICDI, and E_w.

2. Indices of amino acid usage include molecular weight, mean hydropathic indices of each amino acid (*46*), and relative frequencies of selected amino acids (for example, aromatic, acidic, basic, and neutral).

3. Indices of base usage include each nucleotide content [defined as N/(A+T+G+C), where N ∈ {A, C, G, T}], G+C content [defined as (G+C)/(A+T+G+C)], and GC skew [defined as (G−C)/(G+C)] (*47*). The base usage is calculated at each codon position (first, second, and third) and for the overall gene (at all codon positions).

4. Dinucleotide usage is defined as the ratio of observed to expected dinucleotide frequencies (*48*). The dinucleotide usage is calculated for each reading frame (1–2, 2–3, and 3–1) and for the overall gene.

Correlation coefficients between each PCA axis and each gene parameter are calculated and used to rank different gene parameters. The parameter with the highest correlation coefficient is used to identify the main source of variation among genes on the PCA axis. No source will be identified if the PCA axis is better correlated with its original variable (codon) than any of the gene parameters considered (that is, the highest correlation coefficient is one of the factor loadings).

Comparative study of amino acid composition

For the study of amino acid composition using this tool, the set of amino acid sequences of a given cDNA data collection is clustered using Cluster 3.0 (*49*) with a precompiled database of annotated amino acid sequences from a wide range of species. The set is graphically displayed using JTreeView software (http://jtreeview.sourceforge.net/) for functional classification and comparative study.

The annotated database is precompiled as follows. Genome sequence data on 14 species in the EMBL database (Release 82, March 2005; ref. *50*) (Table 1) are obtained and the protein sequences are extracted. The proteins are then annotated by reference to UniProt (*51*), GOA (gene ontology annotation; ref. *52*), and InterPro (*53*). The precompiled data are distributed with the software package.

Table 1 The 14 Species in the EMBL Database Used for Data Analysis

Taxon	Species	Protein	EMBL No.
Eukaryota	*Arabidopsis thaliana*	13,647	AJ270058, AJ270060, AP000423, ATH1-2, BA000014-5
	Caenorhabditis elegans	22,623	BX284601-6
	Drosophila melanogaster	11,706	AE013599, AE014134-5, AE014296-8
	Trypanosoma brucei	1,106	AE017150, AL929608
Bacteria	*Bacillus subtilis*	4,106	AL009126
	Buchnera aphidicola	504	AE016826
	Deinococcus radiodurans	1,655	AE000513, AE001825
	Escherichia coli K12	4,254	U00096
	Helicobacter pylori	1,566	AE000511
	Streptococcus pneumoniae	2,046	AE007317
Archea	*Aeropyrum pernix*	2,694	BA000002
	Methanocaldococcus jannaschii	1,715	AE004437
	Pyrococcus furiosus	2,065	AE009950
	Sulfolobus tokodaii	2,826	BA000023

The software first calculates the amino acid frequencies of all query proteins normalized by the length of each gene to exclude the effect of gene length, then these amino acid composition data are used for clustering with the precompiled database. A hierarchical clustering method implemented with Cluster 3.0 software is used for this purpose. Users may optionally narrow down the target dataset by organism or function. Three matrix files are generated as the output, including a file for amino acid composition and two files (.cdt and .gtr) for viewing with JTreeView software. Graphical results consist of a hierarchical tree and amino acid composition for each protein, represented by color gradations.

Comparative modeling of the structure of product proteins

For CM, a pipeline is implemented to run PSI-BLAST (*54*) with the query amino acid sequence against the PDB database (*55*), and the result is used to run MODELLER (*30*). A similar pipeline was published as MODPIPE (*56*), but it is not open software and is accessible only from its web interface, MODWEB (*56*). A stand-alone form of software, as implemented in our package, is vastly superior to a web service in terms of computational time, and it is useful for private data and large sets of query sequences. We followed MODPIPE for the parameters for PSI-BLAST and for the configuration of MODELLER, but our package also converts the output file to a PDB flatfile that is readily visualized by protein structure viewers such as RASMOL (*57*).

Prediction of alternative splice forms

To take account of spliced-out exons in the prediction of precursor mRNA and exon structures without the existence of complete genome data, the cDNA sequences are first aligned by BLAT (Blast Like Alignment Tool) and BLASTN to cluster similar regions. Similar regions with identity greater than 95%, length over 40 bp, e-value less than 1.0e-10, and score over 80 are considered exon candidates and are aligned to predict the precursor mRNA structure. In each cluster, the exon candidates are aligned at the 3′-ends, and if the end positions are within 10 bp of each other, the exons are used and ordered to generate a contiguous precursor mRNA prediction. A number of the parameters described above, and several additional ones regarding the use of BLAT options such as the use of poly-A tail removal, are configurable by users. A list of clusters and a FASTA file containing the contiguous sequences of predicted precursor mRNAs are generated as a result.

Metabolic reconstruction and pathway alignment

A metabolic pathway is reconstructed by running a BLASTP search of all amino acid sequences against the UniProt database and extracting the EC (Enzyme Commission) numbers of enzymes matched with an e-value cutoff at e-25. The software automatically downloads the latest UniProt database, runs formatdb, creates FASTA queries, runs BLASTP, parses the output, queries UniProt for the EC numbers, and re-annotates the input genome and outputs the updated GenBank flatfile. Here the original annotation is not overwritten, but the new annotation is added as optional feature tags. Pathway alignment is based on this re-annotated file and compares the list of EC numbers of all proteins with the list extracted from the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (*58*). The reconstructed list is compared with the lists of the target database (EC lists of organisms, pathways, or organism-specific pathways in KEGG), and the score is calculated as follows: number of intersections − penalty × number of complements. By default the penalty is set to 0.2, and the user can specify this value for both right-side-only (a set of data only present in the query data) and left-side-only (a set of data only present in the subject database) complements. The result is displayed in a format similar to that of BLAST, showing the score, intersection, and complement lists (see http://www.g-language.org/data/cdna/Additional_data_2.txt for a sample output for a query of glycolysis in *E. coli* against the glycolysis pathways of all organisms in KEGG). Users can also produce graphical output, where the query list is visualized on the targeted organism-specific pathway image of KEGG using KEGG API (application programming interface). The database of EC lists is automatically downloaded and built by the software when it is run for the first time.

Querying the list of reconstructed pathways against the database of pathways will produce a list of pathways that are likely to be present in the queried genome, and the intersections show the putative enzymes. When queried against the database of organisms, the phylogenetic distance in terms of pathway

similarity is obtained, as quantified by the alignment score. Instead of using the reconstructed pathway, any of the database entries of KEGG can be directly used by supplying the pathway or organism entry identification.

Validation

Detection of translation initiation and termination signals

Mouse cDNA data of FANTOM (*59*) were analyzed as a test case. The results, including the PWM, graphs of the calculated indices implemented, and the sequence logo output, are supplied as additional data (http://www.g-language.org/data/cdna/Additional_data_3.doc). By our combined method, the Kozak consensus sequence (*60*) "GCCGCCACC" was detected upstream of the start codon at a statistically significant level, in agreement with previous studies (*61*).

Statistical analysis of codon usage

To identify the major sources of variation in codon usage among human genes, a PCA of the normalized codon usage data R1 (PCA-R1) for human cDNAs (*62*) was performed. The correlation coefficients between all pairs of the PCA axis scores and 126 gene parameters are given in additional data (http://www.g-language.org/data/cdna/Additional_data_1.xls). Figure 2 shows the scatter plots for the four PCA axes that account for 47.3% of the total variation in the data. The first axis is highly correlated with the G+C content at the third codon position (correlation coefficient, r = 0.98). The second and third axes are correlated with the relative frequency of acidic amino acids (r = 0.81) and the T content at the second codon position (r = 0.85), respectively. The fourth axis is better correlated with the original variable than any of the gene parameters considered; that is, the highest correlation coefficient is the factor loading for the AAG codon (r = 0.41). Thus PCA-R1 leads to the conclusion that the primary, secondary, and tertiary sources of variation in codon usage in humans are associated with the use of (G+C)-ending codons, acidic amino acids, and XTX-type codons, respectively.

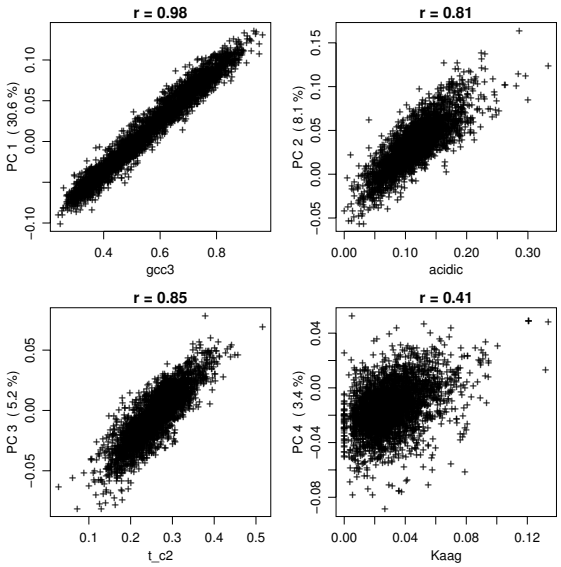

Fig. 2 Scatter plots of PCA axis scores versus other gene parameters. The principal component axes were generated from the PCA of normalized codon usage data R1 for human cDNAs. The first, second, third, and fourth axes (PC1, PC2, PC3, and PC4) have the strongest correlation with the G+C content at the third codon position (gcc3), the relative frequency of acidic amino acids (acidic), the T content at the second codon position (t_c2), and the usage of the AAG codon coding for Lys (Kaag), respectively.

In agreement with PCA-R1, the PCAs of the normalized codon usage data R2 and R3 (PCA-R2 and PCA-R3) generated a first axis that was strongly correlated with G+C content at the third codon position (data not shown). However, an interpretation of more than the first axis is quite difficult, because these axes had the strongest correlations with one of the original variables. Similar observations have also been reported in previous studies (*44*, *63*), where the factorial load of the second axis comes mainly from codons coding for Cys.

At the least, the results of these three different PCAs lead us to the same conclusion that the variability in codon usage in humans reflects a major trend associated strongly with the use of (G+C)-ending codons, as shown in previous studies (*64*).

Comparative study of amino acid composition

All amino acid sequences of the human cDNA collection (*62*) were clustered with the precompiled database. Proteins with the same function were clustered together, and within such clusters proteins from

the same organisms formed sub-clusters. Figure 3 shows the result of clustering of all human cDNA data with tRNA-synthetase in 14 organisms. Several human cDNA translated proteins (noted with blue circles in Figure 3) are included in tRNA-synthetase clusters. This classification locates several human hypothetical proteins as similar to tRNA-synthetase on the basis of amino acid composition, suggesting a functional classification of these hypothetical proteins.

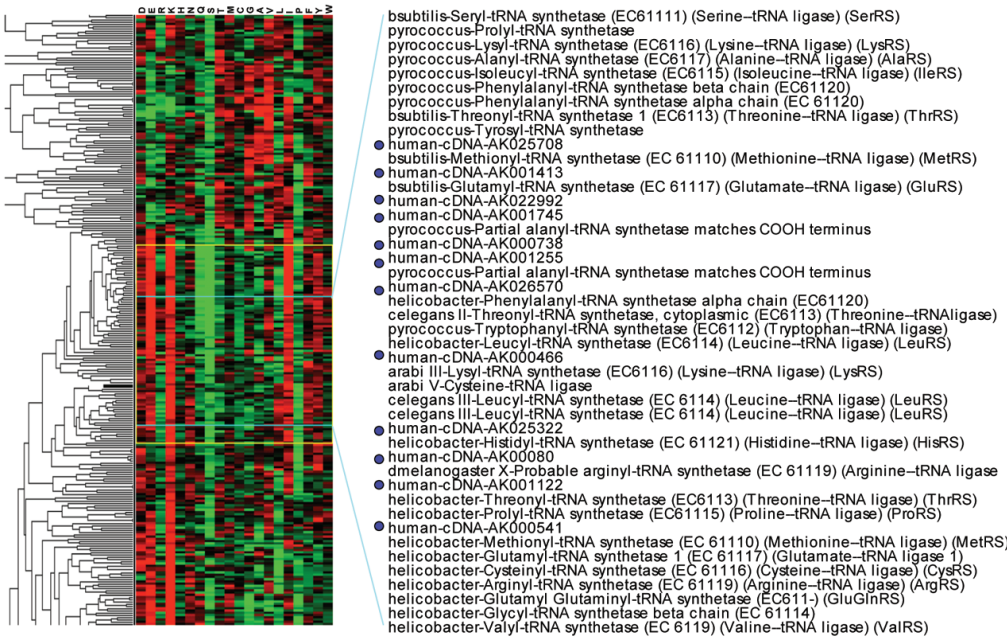

Fig. 3 Graphical overview of protein clustering results. Displayed is the clustering result for all tRNA-synthetases in 14 organisms compared with human cDNA data. Several human cDNA translated proteins (blue circles) are included in the tRNA-synthetase clusters. Upper clusters are occupied by tRNA-synthetases derived from the hyperthermophilic archaeon *Pyrococcus furiosus*, whereas the lower clusters are primarily occupied with the proteobacterium *Helicobacter pylori*.

Comparative modeling of the structures of product proteins

The CM component of our package is a wrapper around the MODELLER software; therefore, the result is identical and the validation is not applicable. Instead, as a demonstration, a sample output using the default parameters of the software was generated using the human FLJ00094 gene (GenBank Number: AK024491). The generated PDB format file of the structure is provided as additional data (http://www.g-language.org/data/cdna/Additional_data_4.pdb) that can be directly visualized in 3D with software such as RASMOL.

Prediction of alternative splice forms

Seventeen sequences (GenBank Number: AK021903, AK022756, AK024284, AK024448, AK026292, AK056232, AK056486, AK091100, AK092491, AK096570, AK097080, AK097269, AK097327, AL050019, BC003555, BC012351, and BC017762) of the first nine clusters (HIX0000001–HIX0000007, HIX0000009, HIX0000010) of H-Invitational 1.0 full-length human cDNA (*9*) were clustered using the default parameter. The tool correctly clustered the test dataset: HIX0000002, consisting of AK026292 and BC017762; HIX0000003, consisting of AK091100 and AK096570; HIX0000004, consisting of AK024448, AK056232, and AK097269; and HIX0000009, consisting of AK022756, AK024284, AK092491, AL050019, and BC003555.

Metabolic reconstruction and pathway alignment

The metabolic pathway was reconstructed using the *E. coli* K12 genome (GenBank Number: U00096). A prokaryote genome of *E. coli* was used because the KEGG entry for eukaryotes was rather incomplete and was not the best subject for this validation pur-

poses. With the default BLAST cutoff value of e-25, the software yielded 832 EC numbers in total. This list of EC numbers was aligned with the organism-specific databases of KEGG, and the top five scores came from alignments with *E. coli* K12 strains, giving the top score of 586.2 (see http://www.g-language.org/data/cdna/Additional_data_5.txt for the entire result). Among the 682 EC numbers listed in the KEGG *E. coli* K12 pathways, the metabolic reconstruction process correctly identified 635 (93.1%) ones, and 47 ones were missed. One hundred and ninety-seven enzymes not listed in KEGG were detected in the reconstructed list, but with the high threshold value, most entries retrieved the correct SwissProt entry with the BLAST search, and therefore were not false positives. Considering the simplistic method of pathway reconstruction employing only similarity searches, the number of 47 (6.9%) false negatives compared with the curated database should be acceptable for quick screening purposes, and by using the alignment tool it is easy to identify the pathways where the missed enzymes should belong.

Conclusion

We have described a comprehensive software suite for bioinformatics analysis of cDNAs. Six tools encompassing the fields of genome, proteome, and systems biology are implemented for easy and generic usage with graphical outputs for intuitive interpretation of the results. The analytical tools are effective for, but not limited to, cDNA data, because the package supports formats of GenBank/EMBL/FASTA.

Most of the tools implemented in this software package take advantage of established methods and software. They also combine several algorithms (as with translation initiation and termination signal detection), provide comprehensive pipelines for easy usage (as with comparative structure modeling), and develop novel methods (as with statistical analysis of codon usage or comparative study of amino acid composition). The tool for statistical analysis of codon usage not only calculates over 100 parameters, including indices for codon usage and nucleotide composition, but also provides a novel means of observing the primary factors of selective pressure involved in deriving the characteristic codon usage. Likewise, the tool for comparative study of amino acid composition allows comprehensive listing and interpretation of the amino acid compositions of all protein sequences, and uses these data in clustering to predict protein functions. Our validations of the tools for translation initiation and termination signal detection, statistical analysis of codon usage, and comparative study of amino acid composition are in agreement with previously reported results. Because the method used for and the results obtained with our comparative structure modeling tool are identical with those in the MODPIPE software, with the exception of additional procedures for output in PDB format for visualization, we have omitted validation of this tool. However, the provision of an open-source alternative to MODPIPE should be useful to the academic community.

Two new tools aimed at post-genomic analyses are also presented. The tool for the prediction of alternative splice forms without the complete genome sequence is obviously less accurate than methods that use the complete genome sequence, and such a comparison would not facilitate any efforts at validation. Nonetheless, our package correctly identified all clusters in the first 10 determined by the H-Invitational annotation project, and because large portions of the "neglected genomes" are likely to be made available through cost-effective EST sequencing methods that do not use the complete genome, our tool provides a means of alternative splice form detection using such data. As yet there is no reported software similar to our tool for metabolic pathway comparison; our tool is effective, as demonstrated by the validation.

Acknowledgements

This research was supported by the Japan Society for the Promotion of Science (JSPS) and a grant from the Ministry of Education, Culture, Sports, Science and Technology of Japan (The 21st Century COE Program).

References

1. Bernal, A., *et al.* 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* 29: 126-127.
2. Yu, U., *et al.* 2004. Bioinformatics in the post-genome era. *J. Biochem. Mol. Biol.* 37: 75-82.
3. Hubbard, T., *et al.* 2005. Ensembl 2005. *Nucleic Acids Res.* 33: D447-453.
4. Meyer, F., *et al.* 2003. GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* 31: 2187-2195.

5. Scharf, M., *et al.* 1994. GeneQuiz: a workbench for sequence analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 348-353.
6. Blaxter, M. 2002. Opinion piece. Genome sequencing: time to widen our horizons. *Brief. Funct. Genomic Proteomic* 1: 7-9.
7. Wasmuth, J.D. and Blaxter, M.L. 2004. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* 5: 187.
8. Kawai, J., *et al.* 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* 409: 685-690.
9. Imanishi, T., *et al.* 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2: e162.
10. Ermolaeva, M.D. 2001. Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* 3: 91-97.
11. Kozak, M. 1996. Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* 7: 563-574.
12. Zien, A., *et al.* 2000. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 16: 799-807.
13. Meinicke, P., *et al.* 2004. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics* 5: 169.
14. Liu, H., *et al.* 2004. Using amino acid patterns to accurately predict translation initiation sites. *In Silico Biol.* 4: 255-269.
15. Locker, J., *et al.* 2002. Definition and prediction of the full range of transcription factor binding sites—the hepatocyte nuclear factor 1 dimeric site. *Nucleic Acids Res.* 30: 3809-3817.
16. Osada, Y., *et al.* 1999. Analysis of base-pairing potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes. *Bioinformatics* 15: 578-581.
17. Ozawa, Y., *et al.* 2002. Comprehensive sequence analysis of translation termination sites in various eukaryotes. *Gene* 300: 79-87.
18. Ozawa, Y., *et al.* 2003. Comparative study of translation termination sites and release factors (RF1 and RF2) in procaryotes. *J. Mol. Evol.* 56: 665-672.
19. Schneider, T.D., *et al.* 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415-431.
20. Konopka, A. 1984. Is the information content of DNA evolutionarily significant? *J. Theor. Biol.* 107: 697-704.
21. Wright, F. 1990. The "effective number of codons" used in a gene. *Gene* 87: 23-29.
22. Morton, B.R. 1993. Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. *J. Mol. Evol.* 37: 273-280.
23. Freire-Picos, M.A., *et al.* 1994. Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. *Gene* 139: 43-49.
24. Suzuki, H., *et al.* 2004. The "weighted sum of relative entropy": a new index for synonymous codon usage bias. *Gene* 335: 19-23.
25. Sharp, P.M. and Li, W.H. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15: 1281-1295.
26. Karlin, S. and Mrazek, J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* 182: 5238-5250.
27. Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* 294: 93-96.
28. Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234: 779-815.
29. Fiser, A., *et al.* 2000. Modeling of loops in protein structures. *Protein Sci.* 9: 1753-1773.
30. Fiser, A. and Sali, A. 2003. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374: 461-491.
31. Siew, N. and Fischer, D. 2003. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53: 241-251.
32. Jordan, I.K., *et al.* 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* 433: 633-638.
33. Raghava, G.P. and Han, J.H. 2005. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics* 6: 59.
34. Cai, Y.D. and Lin, S.L. 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta* 1648: 127-133.
35. Lareau, L.F., *et al.* 2004. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* 14: 273-282.
36. Modrek, B., *et al.* 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29: 2850-2859.
37. Burke, J., *et al.* 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8: 276-290.
38. Boue, S., *et al.* 2002. Theoretical analysis of alternative splice forms using computational methods. *Bioinformatics* 18: S65-73.
39. Karp, P.D., *et al.* 2002. The Pathway Tools software. *Bioinformatics* 18: S225-232.
40. Sun, J. and Zeng, A.P. 2004. IdentiCS—identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics* 5: 112.

41. Pinney, J.W., *et al.* 2005. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res.* 33: 1399-1409.
42. Arakawa, K., *et al.* 2003. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* 19: 305-306.
43. Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097-6100.
44. Perriere, G. and Thioulouse, J. 2002. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.* 30: 4548-4555.
45. Kanaya, S., *et al.* 1996. Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage. *Comput. Appl. Biosci.* 12: 213-225.
46. Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132.
47. Lobry, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13: 660-665.
48. Yew, T.D., *et al.* 2004. Base usage and dinucleotide frequency of infectious bursal disease virus. *Virus Genes* 28: 41-53.
49. Eisen, M.B., *et al.* 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.
50. Brooksbank, C., *et al.* 2003. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* 31: 43-50.
51. Bairoch, A., *et al.* 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33: D154-159.
52. Camon, E., *et al.* 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* 32: D262-266.
53. Mulder, N.J., *et al.* 2005. InterPro, progress and status in 2005. *Nucleic Acids Res.* 33: D201-205.
54. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
55. Deshpande, N., *et al.* 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* 33: D233-237.
56. Eswar, N., *et al.* 2003. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* 31: 3375-3380.
57. Sayle, R.A. and Milner-White, E.J. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* 20: 374.
58. Kanehisa, M., *et al.* 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32: D277-280.
59. Okazaki, Y., *et al.* 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420: 563-573.
60. Kozak, M. 1987. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15: 8125-8148.
61. Sakai, H., *et al.* 2001. Correlation between sequence conservation of the 5′ untranslated region and codon usage bias in *Mus musculus* genes. *Gene* 276: 101-105.
62. Ota, T., *et al.* 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* 36: 40-45.
63. Zavala, A., *et al.* 2002. Trends in codon and amino acid usage in *Thermotoga maritima*. *J. Mol. Evol.* 54: 563-568.
64. Kanaya, S., *et al.* 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* 53: 290-298.