ISSD Version 2.0: taxonomic range extended

Ivan A. Adzhubei and Alexei A. Adzhubei[1],*

Department of Molecular Biology, Faculty of Biology, Lomonosov Moscow State University, 119899 Moscow, Russia and [1]GlaxoWellcome Experimental Research, 16 chemin des Aulx, 1228 Plan-les-Ouates, Geneva, Switzerland

Received September 28, 1998; Accepted September 30, 1998

ABSTRACT

Two more organisms from different taxonomic groups were added to a new version of the Integrated Sequence-Structure Database (ISSD). ISSD serves as an integrated source of sequence and structure information for the analysis of correlations between mRNA synonymous codon usage and three-dimensional structure of the encoded proteins. ISSD now holds 88 non-homologous *Escherichia coli* proteins and 25 yeast *Saccharomyces cerevisiae* proteins in addition to the expanded set of mammalian proteins, which includes 166 proteins (107 in ISSD Version 1.0). Comparison of ISSD sequences with organism-specific codon usage data derived from CUTG database shows that it is a representative subset of the GenBank coding sequences data. Preliminary results of the statistical analysis confirm that sequence–structure correlations observed by us earlier are also present in the upgraded ISSD (Version 2.0), including bacterial and yeast proteins. The ISSD Version 2.0 release includes an improved Web-based data search and retrieval system and is accessible via URL http://www.protein.bio.msu.su/issd/ . ISSD can be also accessed at ExPASy, URL http://www.expasy.ch/swissmod/swiss-model.html

INTRODUCTION

The Integrated Sequence-Structure Database, ISSD (1) is a specialized database designed as information resource for the analysis of sequence–structure relationships between the gene nucleotide sequences and native three-dimensional structures of the encoded proteins. The data stored in ISSD can be described as the intersection of high-resolution protein structure data extracted from PDB (2) and the complete gene coding sequences data obtained from GenBank (3). For each protein, ISSD holds full structural data for the protein backbone (secondary structure assignments, torsional angles and atomic coordinates) aligned with the codon sequence of corresponding gene (mRNA) in a layout suitable for automated analysis of the codon usage relative to protein three-dimensional structure.

The analysis of ISSD Version 1.0, which contained data for 109 mammalian proteins demonstrated (4) that synonymous codons are utilized non-randomly when coding for different types of protein secondary structure. This structure-related codon bias is a distinct phenomenon different from all other codon usage anomalies caused by genome AT/GC contents variations, overall codon bias of a particular coding frame, amino acid secondary structure preferences and other codon usage variations. However due to a limited size of the initial ISSD dataset, both the nature and scope of this phenomenon remained unclear. The recently published results (5) of a comparative analysis of correlation between synonymous codon usage and protein secondary structure for mammals (utilizing the ISSD 1.0 data) and *Escherichia coli* showed few significant correlations found in bacterial proteins in contrast with a high level of correlation observed for the mammalian proteins. An attempt to compare the prokaryote and eukaryote patterns of synonymous codon bias in secondary structure types is described in ref. 6. The datasets of *E.coli* protein structures and mRNA sequences were compared to the human datasets, showing a number of position-dependent correlations in both data sets but at a lower level than that observed by Adzhubei *et al*. (4) and Tao and Dafu (5). However, direct comparison of these results is impossible due to different proprietary data sets.

The ISSD Version 2.0 described here represents an expanded, publicly available data source consistent with the previous Version 1.0. The database aims to provide data for the concurrent examination and analysis of corresponding nucleotide coding sequences and protein three-dimensional structures.

The improvements in Version 2.0 include: (i) increase in the total number of proteins from 107 to 279 (~160%) with the total number of codons increased from 21 741 to 69 589 (~220%); (ii) addition of the two new organisms *E.coli* and *Saccharomyces cerevisiae* (baker's yeast) from different taxonomic groups; (iii) revision of all data for consistency and reliability, including the previously published ISSD 1.0 data; (iv) upgrade of the Web-based ISSD data search and retrieval interface.

DATABASE COMPILATION

The algorithm used to compile ISSD Version 1.0 (1) was retained in Version 2.0. However the implementation software was completely rewritten in order to improve efficiency and reliability. A new compilation package written in Perl5 consists of the two modules instead of a number of separate utilities. The first module performs remote GenBank scanning via the NCBI BLAST server and locates putative coding gene sequences to match the protein sequences extracted from PDB records. A user can select several highest-scoring matches for further processing.

*To whom correspondence should be addressed. Tel: +41 22 884 8640; Fax: +41 22 884 8650; Email: aaa75196@ggr.co.uk

All tasks, e.g. BLAST query formatting, and results parsing are carried out interactively, the user is presented with a compact menu interface listing the most important parameters of the query. The second module is used to download the necessary GenBank and PDB files from remote servers and allows the user to make final gene/protein assignments, utilizing the same interface as at the first stage. When the user confirms a selection, a range of ISSD records is generated automatically from the list of matches selected by the user. The quality and efficiency of ISSD compilation is substantially increased in Version 2.0 due to extensive data format validation performed by the algorithm at several critical points, the two-step menu-driven user interface and the automated remote and local data processing. The new ISSD maintenance software is fully portable and runs on any Unix and Unix compatible platform with Perl5 interpreter and a direct Internet connection. It was tested successfully on Intel-based personal computers under IBM OS/2 Warp 4 (with EMX runtime system) and on the Silicon Graphics workstations under SGI IRIX.

DATABASE CONTENTS

The differences in data held in ISSD Versions 1.0 and 2.0 are summarized in Table 1. Each of the three major organism-specific subsets of ISSD Version 2.0 (mammals, *E.coli* and yeast) was compiled from the unique structures of non-homologous proteins. The same approach was used in Version 1.0, which contained only mammalian proteins. None of the polypeptide chains with the sequence identity levels above 50% were included in the database. Table 2 shows the distribution of protein sequence identity levels for organism-specific subsets of ISSD 2.0.

The CUTG database (7) was used as a reference to estimate how representative is the codon usage data for nucleotide sequences in ISSD relative to the corresponding total organism-specific codon usage. The CUTG database holds codon usage data for coding sequences of genes from all organisms and taxonomic groups represented in GenBank. Since synonymous codon bias is primarily manifested in nucleotide bases in the third (silent) codon position, G+C bias was calculated for the third bases in ISSD sequence data. The results correlate with corresponding parameters calculated from CUTG (Table 3), with a minor deviation in the *S.cerevisiae* subset possibly due to its smaller size. Cluster analysis of the mean relative synonymous codon usage (RSCU) distances also showed close correlation between the ISSD and CUTG RSCU parameters for relevant organism groups (Fig. 1).

Table 1. Quantitative comparison of the ISSD Versions 1.0 and 2.0 data

Organism	Proteins[a]			Codons[b]		
	ISSD 1.0	ISSD 2.0	Increase %	ISSD 1.0	ISSD 2.0	Increase %
Homo sapiens	80	105	31	16 095	22 070	37
Mus musculus	8	15	88	1449	2589	79
Bos taurus	7	14	100	1180	4053	243
Rattus sp.	7	21	200	1417	4244	200
Sus scrofa	4	9	125	1228	2746	124
Equus caballus	1	2	100	372	545	47
Mammals subtotal	**107**	**166**	**55**	**21 741**	**36 247**	**67**
Escherichia coli	–	88	–	–	25 374	–
Saccharomyces cerevisiae	–	25	–	–	7968	–
Total	**107**	**279**	**161**	**21 741**	**69 589**	**220**

[a]Number of PDB structures, some include several polypeptide chains, Table 2.
[b]Only valid codons successfully aligned with the corresponding residues are included.

Table 2. Results of analysis[a] of the pairwise alignment matrix for proteins in the three organism-specific groups in ISSD Version 2.0

Organism group	Number of unique polypeptide chains	Percentage of alignments above the identity level of	
		30%	40%
Mammals[b]	189	0.21%	0.08%
Escherichia coli	91	0.02%	none
Saccharomyces cerevisiae	25	none	none

[a]The description of pairwise alignment procedure is given in ref. 1.
[b]See Table 1 for the list of organisms.

Table 3. G+C bases in the third codon position, ISSD Version 2.0 compared to CUTG (7) database generated from GenBank

Organism group	G+C percentage in third codon position	
	ISSD 2.0	CUTG
Mammals[a]	61.45	60.70[b]
Escherichia coli	56.21	55.39
Saccharomyces cerevisiae	43.91	38.00

[a]See Table 1 for the list of organisms.
[b]Recalculated using CUTG data for the organisms in ISSD ‘mammals’ subset.

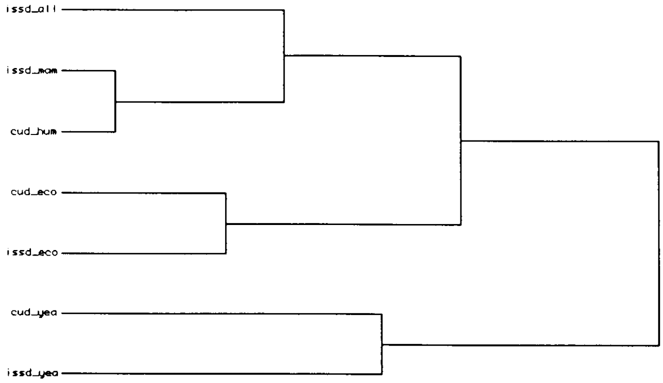

Figure 1. Comparison dendrogram of the mean RSCU distances for ISSD 2.0 and its subsets. Issd_all, the full ISSD 2.0 dataset; issd_mam, issd_eco and issd_yea, the ISSD 2.0 mammals, *E.coli* and *S.cerevisiae* organism groups respectively. Cud_hum, cud_eco and cud_yea, the CUTG human, *E.coli* and *S.cerevisiae* organism groups, respectively. The dendrogram shows a high level of correlation between RSCUs of the same or similar organism groups in ISSD and CUTG thus confirming the representative character of the ISSD dataset. The technique used for cluster analysis is described in ref. 1.

The most interesting question is whether the data of updated ISSD 2.0 can be used to reproduce the same type of synonymous codon distribution bias in different secondary structure types as observed for a smaller subset of mammalian proteins (4). Figure 2 shows the results of χ^2 analysis applied to the three organism-specific subsets of ISSD 2.0, using the technique described by Adzhubei *et al*. (4). The 'mammals' and '*E.coli*' subsets show substantial non-random synonymous codon bias for a number of codon families. Five codon families have highly significant bias ($P < 0.05$) in both 'mammals' and '*E.coli*' subsets (Fig. 2a and b). The 'yeast' subset has a less prominent structure-related codon bias (Fig. 2c), which can be explained by a small number of proteins in the subset. Only the Ile family shows significant deviation from the random distribution in this subset. The codon families that display structure-related bias are different for these three taxonomic groups, with only the Gly family showing significant bias in both 'mammals' and '*E.coli*'.

These results support the view that structure-related synonymous codon bias is a general phenomenon found in all major taxonomic groups of organisms. On the other hand, the character of this bias is highly species-specific, with eukaryotes and prokaryotes clearly demonstrating different patterns. It is still unclear if these dissimilarities are connected with evolutionary differences in the translation apparatus, i.e. translation accuracy, or cotranslational protein folding optimization, or differences in the intrinsic genome features, i.e. selection against mutational pressure. ISSD can serve as a valuable source of data for further studies of this phenomenon.

DATABASE ACCESS

ISSD Version 2.0 is available on BioProt Web-server at the Department of Molecular Biology, Lomonosov Moscow State University, via URL http://www.protein.bio.msu.su/issd/ . ISSD can be also accessed on the ExPASy server, URL http://www.expasy.ch/swissmod/swiss-model.html . When using ISSD please cite this article.

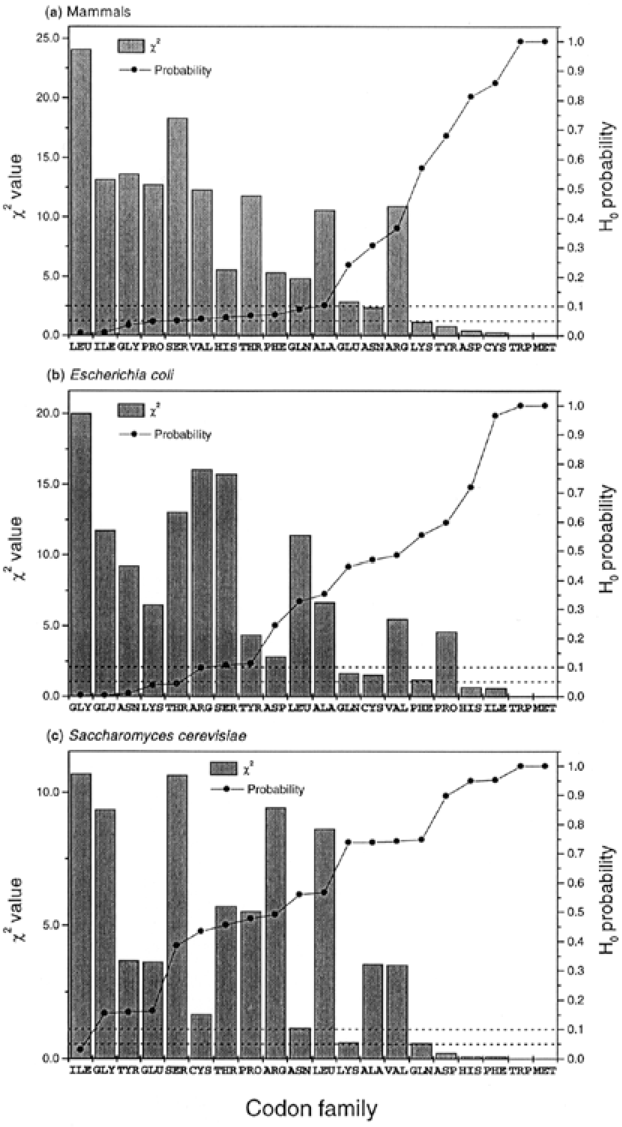

Figure 2. χ^2 analysis results, showing deviation from random synonymous codon usage in the secondary structure types (H_0 hypothesis) for proteins from different organism groups in ISSD 2.0. (**a**) Mammals; (**b**) *E.coli*; (**c**) *S.cerevisiae*. Bar columns represent the sum of χ^2 for deviations from the expected distribution of synonymous codons in 20 codon families; closed circles show probability of the H_0 hypothesis for the corresponding codon families. Dotted lines mark significance levels for the H_0 hypothesis rejection at the probability levels 0.05 and 0.10. The analysis technique used here is described in ref. 4.

The database is released together with an improved version of the Web-based data search and retrieval service. New features include the search option with several indexed keywords instead of only the organism name, combined searches using Boolean operators, and a browse or batch-download of the selected ISSD entries. The full ISSD Version 2.0 can be downloaded using the FTP protocol. Comments on the database can be sent to issd@protein.bio.msu.su

The upgrade of the ISSD Web site also includes a full description of the database format and a short scientific background article with the latest bibliographic citations and links to abstracts where available.

ACKNOWLEDGEMENTS

The authors would like to thank Nadeene Parker for help in updating ISSD. This work was supported in part by the Russian Foundation for Basic Research grant 97-04-49827 and by GlaxoWellcome.

REFERENCES

1 Adzhubei,I.A., Adzhubei,A.A. and Neidle,S. (1998) *Nucleic Acids Res.*, **26**, 327–331.
2 Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) In Carter,C.W.,Jr. and Sweet,R.M. (eds), *Methods in Enzymology*, Academic Press, San Diego, Vol. **277**, pp. 556–571.
3 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
4 Adzhubei,A.A., Adzhubei,I.A., Krasheninnikov,I.A. and Neidle,S. (1996) *FEBS Lett.*, **399**, 78–82.
5 Tao,X. and Dafu,D. (1998) *FEBS Lett.*, **434**, 93–96.
6 Oresic,M. and Shalloway,D. (1998) *J. Mol. Biol.*, **281**, 31–48.
7 Nakamura,Y., Gojobori,T. and Ikemura,T. (1998) *Nucleic Acids Res.*, **26**, 334.