Nucleic Acids Research, 1998, Vol. 26, No. 1 327–331

An Integrated Sequence-Structure Database incorporating matching mRNA sequence, amino acid sequence and protein three-dimensional structure data

Ivan A. Adzhubei[1,2], Alexei A. Adzhubei[1,*] and Stephen Neidle[1]

[1]CRC Biomolecular Structure Unit, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK and [2]Department of Molecular Biology, Faculty of Biology, Lomonosov Moscow State University, 119899 Moscow, Russia

Received August 22, 1997; Revised and Accepted September 22, 1997

ABSTRACT

We have constructed a non-homologous database, termed the Integrated Sequence-Structure Database (ISSD) which comprises the coding sequences of genes, amino acid sequences of the corresponding proteins, their secondary structure and φ,ψ angles assignments, and polypeptide backbone coordinates. Each protein entry in the database holds the alignment of nucleotide sequence, amino acid sequence and the PDB three-dimensional structure data. The nucleotide and amino acid sequences for each entry are selected on the basis of exact matches of the source organism and cell environment. The current version 1.0 of ISSD is available on the WWW at http://www.protein.bio.msu.su/issd/ and includes 107 non-homologous mammalian proteins, of which 80 are human proteins. The database has been used by us for the analysis of synonymous codon usage patterns in mRNA sequences showing their correlation with the three-dimensional structure features in the encoded proteins. Possible ISSD applications include optimisation of protein expression, improvement of the protein structure prediction accuracy, and analysis of evolutionary aspects of the nucleotide sequence–protein structure relationship.

INTRODUCTION AND RATIONALE

Since the genetic code is degenerate (the 61 sense codons are arranged, in the universal genetic code, into groups or families of synonyms each coding for a single amino acid residue) there is no back-translation algorithm that can, strictly speaking, yield the native nucleic coding sequence solely on the basis of amino acid sequence data. Knowledge of the exact coding sequence as part of the general set of data describing the sequence-structure relationship in proteins is thus important. The degeneracy of the genetic code, primarily connected with 'silent' nucleotide at the third codon-position, results in additional degrees of freedom in the nucleotide sequence, allowing it to carry superfluous information relevant to the encoded amino acid sequence. It has been found (1–3) that there is a correlation between patterns of positive rare codon usage bias in mRNA sequences and segments linking domains and regular secondary structure blocks in proteins. The rates of translation can be lower for inter-domain regions (1,4) and can also vary for different secondary structures (5). The bias of nucleotides in specific codon-positions, associated with the regions adjacent to α-helices and β-sheets has been reported (6) although no correlation with the rare codon usage was found in this study.

It is difficult to interpret these results since different sets of data were used for the analyses and none were published by the authors. Importantly, the organism specificity of the codon usage should be taken into account when compiling a dataset. To separate sequence-structure effects from the raw genome codon bias it is necessary to have a dataset compiled using the gene and protein data for the same organisms and, whenever possible, matching tissue/cell types. Generally available databases (e.g. GenBank, PDB) were not suitable for such analysis for several reasons. Firstly, their structural information is extremely redundant. For instance, the 'non-redundant' version of GenBank (available via NCBI Entrez) retains duplicate entries for all overlapping nucleotide sequences, even if they differ only in a few nucleotides. Protein Data Bank (Brookhaven) contains numerous structures for identical proteins or proteins with single residue substitutions. Secondly, the alignment of different levels of sequence and structure data presents a problem. Automatic generation of alignments directly from PDB or GenBank records is unreliable due to relatively high level of sequence errors and inconsistency of data formats even within a single database, either PDB or GenBank. Hence the best solution was to compile a specialised database using resources of the publicly available general purpose databases, additionally applying strict manual data consistency checks. This technique was used to construct ISSD.

The initial version of ISSD was used by us to carry out a direct analysis of the synonymous codon distribution frequencies between protein secondary structure types, which showed that the distribution is statistically non-random (7), with the codon structural preferences related to the nucleotide in the third 'silent' codon-position. Specific structural preferences for some synonymous codons at the N- or C-termini of α-helices and β-sheets were also observed.

DATABASE CONSTRUCTION

The Integrated Sequence-Structure Database was compiled according to the algorithm presented in Figure 1. The main objective of ISSD is to integrate sequence and protein three-dimensional structure data of each given protein molecule, from multiple entries

*To whom correspondence should be addressed

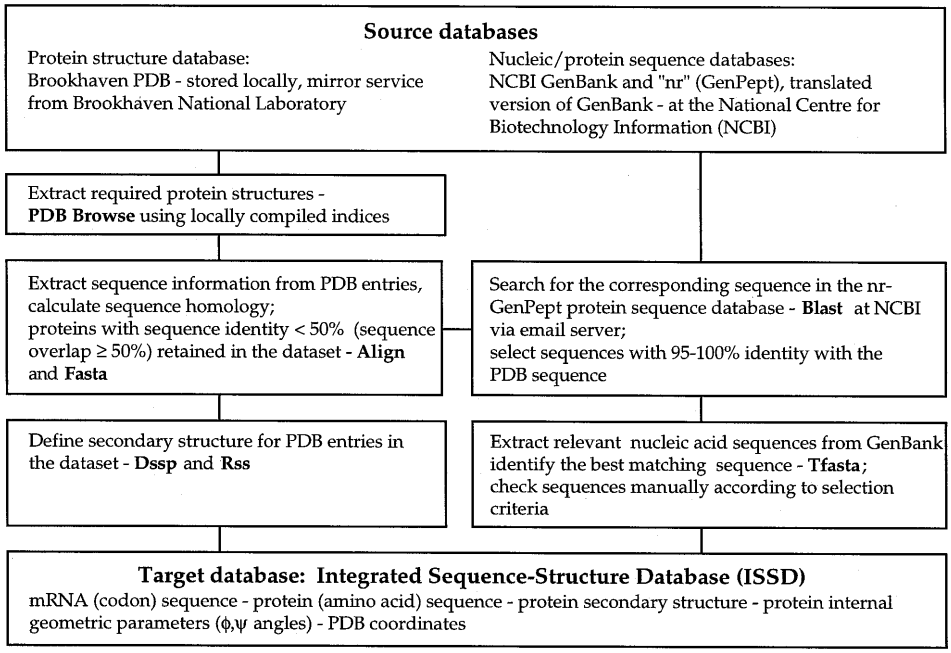

Figure 1. The Integrated Sequence-Structure Database (ISSD) compilation algorithm. Computer programs used in the database compilation are shown in bold print.

Standard a.a. families				Extended a.a. families				Reduced a.a. families				Reduced a.a. families			
	fabs	fdb	RSCU		fabs	fdb	RSCU		fabs	fdb	RSCU		fabs	fdb	RSCU
ALAfam	1419	0.065		LEUfam	1891	0.087		CYSfam	554	0.025		ASNfam	986	0.045	
AU GCU	400	0.018	1.128	LU CUU	200	0.009	0.635	CU UGU	237	0.011	0.856	NU AAU	405	0.019	0.822
AC GCC	623	0.029	1.756	LC CUC	389	0.018	1.234	CC UGC	317	0.015	1.144	NC AAC	581	0.027	1.178
AA GCA	273	0.013	0.770	LA CUA	112	0.005	0.355								
AG GCG	123	0.006	0.347	LG CUG	849	0.039	2.694	ASPfam	1175	0.054		GLNfam	841	0.039	
				La UUA	98	0.005	0.311	DU GAU	524	0.024	0.892	QA CAA	215	0.010	0.511
GLYfam	1625	0.075		Lg UUG	243	0.011	0.771	DC GAC	651	0.030	1.108	QG CAG	626	0.029	1.489
GU GGU	260	0.012	0.640												
GC GGC	612	0.028	1.506	ARGfam	1046	0.048		GLUfam	1437	0.066		TYRfam	760	0.035	
GA GGA	385	0.018	0.948	RU CGU	110	0.005	0.631	EA GAA	586	0.027	0.816	YU UAU	290	0.013	0.763
GG GGG	368	0.017	0.906	RC CGC	207	0.010	1.187	EG GAG	851	0.039	1.184	YC UAC	470	0.022	1.237
				RA CGA	111	0.005	0.637								
PROfam	980	0.045		RG CGG	157	0.007	0.901	PHEfam	951	0.044		METfam	430	0.020	
PU CCU	276	0.013	1.127	Ra AGA	227	0.010	1.302	FU UUU	389	0.018	0.818	MG AUG	430	0.020	1.000
PC CCC	356	0.016	1.453	Rg AGG	234	0.011	1.342	FC UUC	562	0.026	1.182				
PA CCA	250	0.011	1.020									TRPfam	326	0.015	
PG CCG	98	0.005	0.400	SERfam	1438	0.066		HISfam	485	0.022		WG UGG	326	0.015	1.000
				SU UCU	251	0.012	1.047	HU CAU	189	0.009	0.779				
THRfam	1303	0.060		SC UCC	371	0.017	1.548	HC CAC	296	0.014	1.221	----------			
TU ACU	297	0.014	0.912	SA UCA	153	0.007	0.638					total			
TC ACC	574	0.026	1.762	SG UCG	70	0.003	0.292	ILEfam	1117	0.051		db	21741		
TA ACA	314	0.014	0.964	Su AGU	210	0.010	0.876	IU AUU	375	0.017	1.007				
TG ACG	118	0.005	0.362	Sc AGC	383	0.018	1.598	IC AUC	607	0.028	1.630				
								IA AUA	135	0.006	0.363				
VALfam	1522	0.070													
VU GUU	260	0.012	0.683					LYSfam	1455	0.067					
VC GUC	394	0.018	1.035					KA AAA	526	0.024	0.723				
VA GUA	126	0.006	0.331					KG AAG	929	0.043	1.277				
VG GUG	742	0.034	1.950												

Figure 2. A computer-generated table listing the mean codon and amino acid frequencies for ISSD, available as a part of statistical data on the ISSD WWW site. The amino acid codon families are sorted according to the size (number of synonyms) of each family. The database frequencies fdb = fabs/total_db. The relative synonymous codon usage RSCU = fabs / (fam_fabs/fam_size), where fam_fabs is the total number of codons in an amino acid family and fam_size is the number of synonyms in that family.

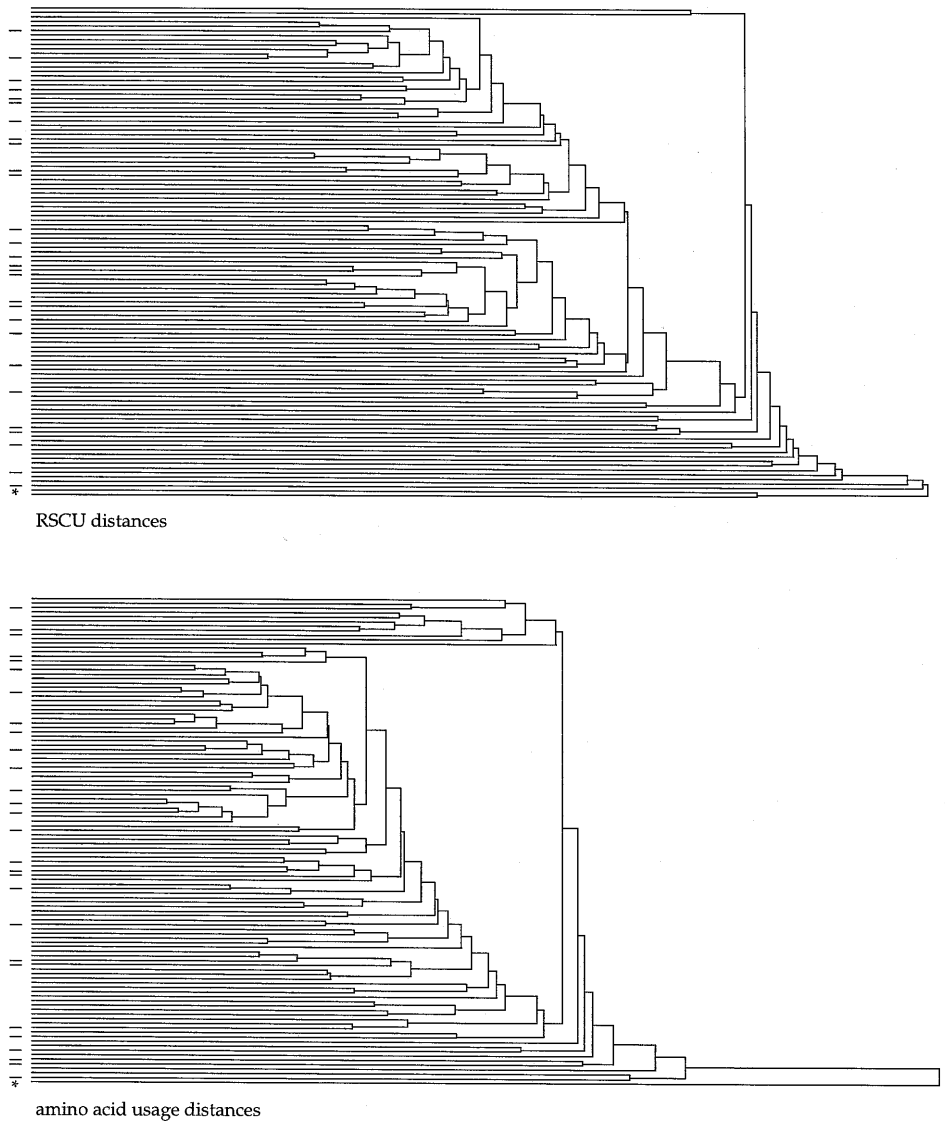

Figure 3. The dendrograms computed for the RSCU (11) and amino acid usage Euclidean distance matrices, as calculated for all entries in ISSD. Cluster analysis was performed by the program MACDENDRO (14), using the average link algorithm. The majority of entries in the current release of ISSD contain human genes/proteins. The remaining entries, marked '–', include genes/proteins of other species of mammalian higher vertebrates. An unusually high divergence in the amino acid usage from the rest of the database in defensin (PDB id 1DFN, length 30 residues), marked with an asterisk, is probably due to its short sequence.

in the source databases to a single target ISSD entry. Thus fast and convenient access is provided to the information describing all levels of protein structure, from the coding sequences of genes to three-dimensional coordinates.

The algorithm described in Figure 1 was used to establish a one-to-one correspondence between entries in different sequence and structure databases and a specific protein. It also allowed us to avoid ambiguities in sequence data that can exist in the PDB [Protein Data Bank (8)] database. A PDB entry contains SEQRES records listing the amino acid sequence of the molecule, but this sequence does not always have a one-to-one correspondence with the sequence for which the three-dimensional coordinates are given in the same entry. Older PDB entries also do not provide a reference to the relevant amino acid sequence databases. Even if such a reference is present it is still necessary to find, for the same protein, an exact alignment of its PDB-sequence and the sequence deposited in an amino acid sequence database. This is done because of potential inaccuracies in the sequence data and since discrepancies in sequence identification are possible between different databases. The PDB-sequences processed by the ISSD compilation software were extracted from ATOMS records and thus represent the exact sequences featured in the three-dimensional structure determination.

At the next step the nucleotide sequence exactly matching a given PDB-sequence has to be identified. In the ISSD compilation algorithm this is done by searching a translated version of the NCBI GenBank database, thus directly identifying all sequences with near 100% match to the initial PDB-sequence. Corresponding nucleotide sequences are then extracted from the database and the best matching sequence identified on the basis of TFASTA alignments with the PDB-sequence. At this stage of the compilation process the sequences are checked manually for a number of selection criteria. The criteria include final checks to ensure that (i) the selected

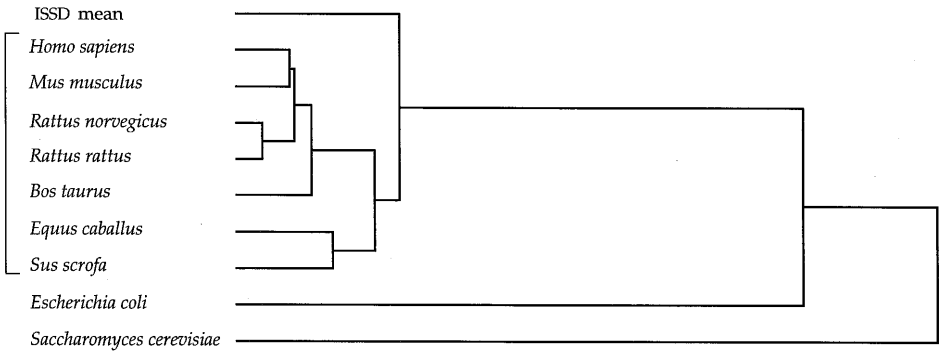

Figure 4. Dendrogram showing the relationship between the mean RSCU for ISSD and the RSCUs for the ISSD source-species (bracketed), calculated from the data independent of ISSD. The data of the codon usage in the individual species was taken from the Codon Usage Database (12). *Escherichia coli* and *S.cerevisiae* are included as reference points for the RSCU distances. The cluster analysis was performed as described in Figure 3.

nucleotide sequence is of the same source/organism with the PDB-sequence, (ii) the source tissue/organ/stage of development is the same or as close as possible, (iii) the sequences do not represent mutants.

Protein secondary structure assignments were made by the program DSSP (9) and include the following types: 'H', α-helix; 'B', β-bridge; 'E', β-ladder (participating in a β-sheet), 'G', 3_{10}-helix; 'T', turn with hydrogen bond; 'S', bend; 'I', π-helix (very rare). The structure type 'P', polyproline II-type helix, was assigned by the RSS program (10). The PDB coordinates are given as they appear in the corresponding PDB entries.

DATABASE DESCRIPTION

The currently available version (Alpha 1.0) of ISSD is a regularly updated, non-homologous database (identity <50% for any pair of aligned sequences in the database), and is restricted to mammalian genes/proteins (as shown in Fig. 4). The actual sequence identity levels in the non-homologous version of ISSD were calculated using the program ALIGN to generate a total non-redundant set of global pairwise alignments. For the amino acid sequences 99.5% of the pairwise alignments in the database have identity <30%. Global pairwise alignments were also calculated for codon sequences in the database and showed no pairs with identity >30%. Structures in the database have resolution better than or equal to 2.5 Å.

The database is accessible via the HTTP protocol on the WWW. Individual entries, indexed by the organism name, and statistical data (Fig. 2) for the codon and amino acid usage in the database can be viewed using a general purpose Web browser. An entry (protein) in the ISSD includes headers giving short description of a protein (name, source organism, structure resolution) and hyperlinked references to the entries in the sequence and structural databases. The main data in an ISSD entry consists of the nucleotide coding sequence (codons) aligned with the amino acid sequence and, for each equivalent codon–amino acid, structural parameters of the backbone. Structural parameters include secondary structure assignments, accessibility, ϕ,ψ angles of the peptide groups and PDB coordinates of the polypeptide backbone. The ISSD format is designed specifically to facilitate computer processing of the database records, the full format description is given at the ISSD WWW site.

The ISSD dataset looks small in comparison with the vast amount of nucleotide sequence data available and rapidly growing number of experimentally solved protein structures. However, one should note the already mentioned excessive redundancy of the structural databases. Also, only non-homologous proteins with high resolution structures and available corresponding nucleotide sequences of the genes from the same organism were selected. An upgraded version of ISSD is in preparation and will substantially increase the number of mammalian proteins included in the database. The datasets for other species are also being prepared, with priority given to *Escherichia coli* and *Saccharomyces cerevisiae*. Two other versions of ISSD are planned: (i) a non-homologous dataset that will combine both mammalian and other genes/proteins, (ii) a dataset that will include all sequences for which the three-dimensional structures are available, classified by species, with the first release covering human genes/proteins.

Codon and amino acid usage statistics are used to monitor the database uniformity. Statistics calculated for the version 1.0 ISSD dataset include relative synonymous codon usage (RSCU) distances (11) and amino acid usage distances between sequences (calculated in a similar manner). Cluster analysis shows a relatively high diversity in synonymous codon usage for individual genes although no clusters separated by significantly high RSCU distances could be identified according to the criteria introduced earlier (11) (Fig. 3). Notably, the genes that belong to the same species do not form separate clusters but are distributed diffusely. A computer-generated listing of the non-redundant ISSD mean codon and amino acid frequencies is given in Figure 2. The ISSD mean synonymous codon usage fairly reflects codon usage in the source species as calculated from independent data of the Codon Usage Database (12) (Fig. 4). Our results are in agreement with the previously reported observations of markedly different codon usage patterns in individual mammalian genes, which result in similar patterns for different species after averaging (13). The amino acid usage dendrogram shown in Figure 3 displays, as expected, lower distances between individual sequences in comparison with the RSCU.

At present several possible applications of ISSD can be suggested. The data can be used for gene expression optimisation, including back-translation algorithms, total codon usage optimisation, optimisation of synonymous codon distribution along mRNA, and codon usage optimisation relative to protein structure. The database can be used to explore possibilities to improve protein three-dimensional structure prediction and modelling. It will also be possible to analyse the evolutionary relationship between protein structure and gene sequence.

DATABASE ACCESS

The Integrated Sequence-Structure Database can be accessed via URL http://www.protein.bio.msu.su/issd/

ACKNOWLEDGEMENTS

This work was supported by the Cancer Research Campaign (Programme Grant SP13848SN) and in part by the Russian Fund for Basic Research. IAA acknowledges support and thanks the Royal Society for a fellowship awarded under the 'Exchanges with the former FSU' scheme.

REFERENCES

1 Krasheninnikov,I.A., Komar,A.A. and Adzhubei,I.A. (1989) *Biokhimia* (*Moscow*), **54**, 187–200.
2 Krasheninnikov,I.A., Komar,A.A. and Adzhubei,I.A. (1989) *Dokl. Akad. Nauk SSSR*, **305**, 1006–1012.
3 Krasheninnikov,I.A., Komar,A.A. and Adzhubei,I.A. (1991) *J. Protein Chem.*, **10**, 445–454.
4 Thanaraj,T.A. and Argos,P. (1996) *Protein Sci.*, **5**, 1594–1612.
5 Thanaraj,T.A. and Argos,P. (1996) *Protein Sci.*, **5**, 1973–1983.
6 Brunak,S. and Engelbrecht,J. (1996) *Proteins*, **25**, 237–252.
7 Adzhubei,A.A., Adzhubei,I.A., Krasheninnikov,I.A. and Neidle,S. (1996) *FEBS Lett.*, **399**, 78–82.
8 Bernstein,F.C., Koetzle,T.F., Williams,G., Meyer,D.J., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
9 Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
10 Adzhubei,A.A. and Sternberg,M.J.E. (1993) *J. Mol. Biol.*, **229**, 472–493.
11 Sharp,P.M., Tuohy,T.M. and Mosurski,K.R. (1986) *Nucleic Acids Res.*, **14**, 5125–5143.
12 Nakamura,Y., Wada,K., Wada,Y., Doi,H., Kanaya,S., Gojobori,T. and Ikemura,T. (1996) *Nucleic Acids Res.*, **24**, 214–215. [See also this issue *Nucleic Acids Res.* (1998) **26**, 334.]
13 Wada,K., Aota,S., Tsuchiya,R., Ishibashi,F., Gojobori,T. and Ikemura,T. (1990) *Nucleic Acids Res.*, **18** (suppl.), 2367–2411.
14 Thioulouse,J. (1995) *Computat. Stat. Data Analysis*, **19**, 237–261.