OLIGAMI: OLIGomer Architecture and Molecular Interface

Kazuo Fujiwara* and Masamichi Ikeguchi

Department of Bioinformatics, Soka University, 1-236 Tangi-cho, Hachioji, Tokyo 192-8577, Japan

Abstract: OLIGAMI (OLIGomer Architecture and Molecular Interface) is a database of the verified coordinates and new chain formulas for biological molecules that allows users to browse oligomers through the SCOP hierarchy and to interactively view three-dimensional structures of biological molecules for all PDB entries. OLIGAMI is publicly available at http://protein.t.soka.ac.jp/oligami/

INTRODUCTION

The nearly exponential growth in the number of structures in the Protein Data Bank (PDB) enables us to reveal the relationships between protein structures and functions. The three-dimensional structures of proteins are comprised of one or several domains, which are minimal segments of the folding and function of proteins, and then certain discrete polypeptide chains (regardless of size) interact with other polypeptides to form higher-order quaternary structures *via* non-covalent bonds. SCOP [1] and CATH [2] are both classifications of domain structures that allow us to browse domains with similar structures. There are also several databases for the quaternary structure of proteins such as PIBASE [3], SCOPPI [4], SNAPPI-DB [5], ProtBud [6] and 3D complex [7]. These databases draw information of the quaternary structures from The Protein Quaternary Structure file server (PQS) [8] and PDB.

PQS at European Bioinformatics Institute (EBI) is a widely used database that stores the coordinates of predicted quaternary structures for the PDB entries determined by X-ray crystallography. PQS is also used in various analyses of protein-protein interactions [9, 10]. The PQS web site allows us to search a database by ‘PDBidcode’, ‘KeyWords’, ‘Quaternary type’ and so on, and provides quaternary structure information, including the chain formula. Each oligomer stored in PQS has a chain formula that represents the chain composition. For example, a dimer is represented as [A2], and a trimer is represented as one of three different chain formulas, [A3], [A2B] or [ABC]. The letters represent distinct polypeptide or poly–nucleic acid subunits, and the numbers indicate the number of each subunit per trimer. Thus, the chain formula simply and specifically describes the oligomeric composition and allows us to simultaneously compare many such compositions. However, the PQS chain formulas contain small peptides or nucleic acids as well as proteins, and thus the chain formula [A2] does not necessarily represent a protein dimer. Furthermore, the PQS database yields false predictions as described in the original article [8]. The PQS database does not include structures determined by NMR spectroscopy nor information on domain structures and biological functions of proteins.

PDB is a fundamental source for all structural data, and the PDB file that is usually downloaded from the PDB web site contains the asymmetric unit, the repeating unit within a crystal. In 1999, PDB added the BIOMOLECULE section ‘REMARK 350’ describing the biological multimeric states, termed the biological unit, to the PDB files, and the coordinates of the biological unit were provided as the PDB biological unit file [11]. The quaternary structures in these files are potentially of higher quality as compared with the predicted PQS quaternary structures because the information is provided by researchers for their registered proteins. However, the REMARK 350 section also contains misinformation. For example, yeast 2,4-dienoyl-CoA reductase functions as a homodimer [12], but the biological unit of PDB entry (1GYR) for this protein has three chains. In the PDB web site, furthermore, the three-dimensional views are provided only as a JPEG image for the PDB biological unit files.

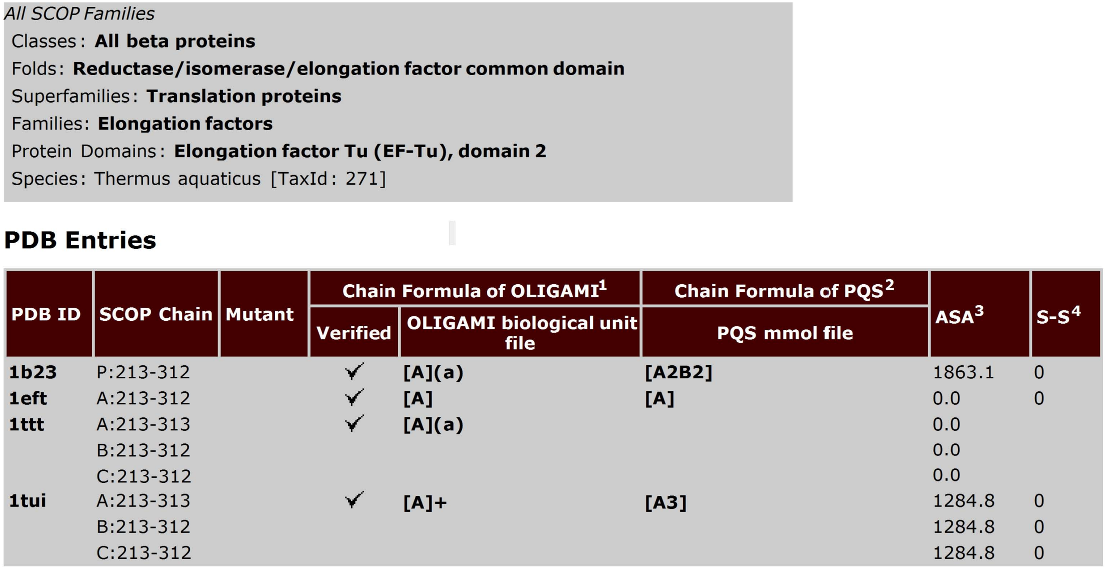
In this paper, we introduce the quaternary structural database, OLIGAMI (OLIGomer Architecture and Molecular Interface). OLIGAMI is a database of the verified coordinates for the biological molecules for all PDB entries, which are generated based on the original biological unit (described below). Furthermore, OLIGAMI provides new chain formulas, which distinguish proteins, peptides, DNAs, and RNAs for all PDB entries. The OLIGAMI web site allows users to interactively view three-dimensional structures of biological molecules for all PDB entries, to browse the molecules through SCOP hierarchy, to compare the chain formulas of OLIGAMI and PQS, and to simultaneously compare the chain formulas for a protein or for a SCOP family. OLIGAMI is publicly available at http://protein.t.soka.ac.jp/oligami/ and users will be able to find a tutorial.

OLIGAMI CHAIN FORMULA

Prior to generating a chain formula, we first determine whether the chain is a protein, peptide, DNA or RNA. For a polypeptide chain, we defined that the protein contains 30 or more residues and that the peptide has less than 30 residues. We distinguish DNA from RNA using the remediated PDB format released on August 1, 2007 (http://www.wwpdb.org/), in which the deoxyribonucleotides are identified with the residue names DA, DC, DT and DG, and the ribonucleotides are identified with the residue names A, C, T and U. Four chain types, protein, peptide, DNA and RNA are dis-

*Address correspondence to this author at the Department of Bioinformatics, Soka University, 1-236 Tangi-cho, Hachioji, Tokyo 192-8577, Japan; Tel: +81 42 691 9326; E-mail: fujiwara@soka.ac.jp

1875-0362/08 2008 Bentham Open



All SCOP Families
Classes: **All beta proteins**
Folds: **Reductase/isomerase/elongation factor common domain**
Superfamilies: **Translation proteins**
Families: **Elongation factors**
Protein Domains: **Elongation factor Tu (EF-Tu), domain 2**
Species: Thermus aquaticus [TaxId: 271]

PDB Entries

PDB ID	SCOP Chain	Mutant	Chain Formula of OLIGAMI[1]		Chain Formula of PQS[2]	ASA[3]	S-S[4]
			Verified	OLIGAMI biological unit file	PQS mmol file		
1b23	P:213-312		✓	**[A](a)**	**[A2B2]**	1863.1	0
1eft	A:213-312		✓	**[A]**	**[A]**	0.0	0
1ttt	A:213-313		✓	**[A](a)**		0.0	
	B:213-312					0.0	
	C:213-312					0.0	
1tui	A:213-313		✓	**[A]+**	**[A3]**	1284.8	0
	B:213-312					1284.8	0
	C:213-312					1284.8	0

Fig. (1). List of PDB IDs with chain formulas in OLIGAMI.

List of SCOP hierarchy and PDB IDs. For each entry, two chain formulas of OLIGAMI and PQS are shown. The plus sign (+) indicates that we modified the chain formula and coordinates.

tinguished from each other in the OLIGAMI chain formula. Polypeptide chains, both protein and peptide, are represented by square brackets [ ], whereas nucleic acid chains, both DNA and RNA, are represented by round brackets ( ). Furthermore, both protein and DNA chains are represented by uppercase letters, whereas both peptide and RNA chains are represented by lowercase letters. By definition, a homo-oligomer consists of protomers having the same amino acid sequence. We also extract the biological unit from REMARK 350 and manually verify the results as described below. The chain formulas are generated using the sequence information and the biological units.

For example, the complex [A2][a](AB)(a) consists of two protein chains that have the same sequence, one peptide, two DNA chains of different sequence, and one RNA chain. We also generate the chain formulas based on the coordinates of the PDB biological unit files. Approximately, 1% of the total PDB entries with two chain formulas based on REMARK 350 and based on the coordinates of the PDB biological unit file were inconsistent. The main reason for these differences is that REMARK 350 is not properly reflected in the PDB biological unit file (for example, PDB ID 2J4H, 1I88); in addition, REMARK 350 contains some misinformation. OLIGAMI also includes chain formulas, areas of the interface, and intermolecular disulfide bonds from PQS and SCOP hierarchical information as shown in Fig. (**1**).

Consequently, the resulting numbers of monomers, homo-oligomers, and hetero-oligomers in OLIGAMI are significantly different from comparable searches obtained from PQS (Table **1**). The fraction of the monomers in OLIGAMI (50%) is greater than that in PQS (38%). On the other hand, the fraction of hetero-oligomers in OLIGAMI (12%) is less than that in PQS (20%). Some of these differences may result from the distinction between proteins and small peptides. For example, in PQS, the PDB entry for a complex between an elastase and a three-residue peptide (PDB ID: 1GVK) is assigned as a heterodimer ([AB]). In OLIGAMI, however, a small peptide is distinguished from a protein chain, and this entry is assigned as a monomeric protein ([A][a]). All of the chain formulas are listed in the OLIGAMI web site.

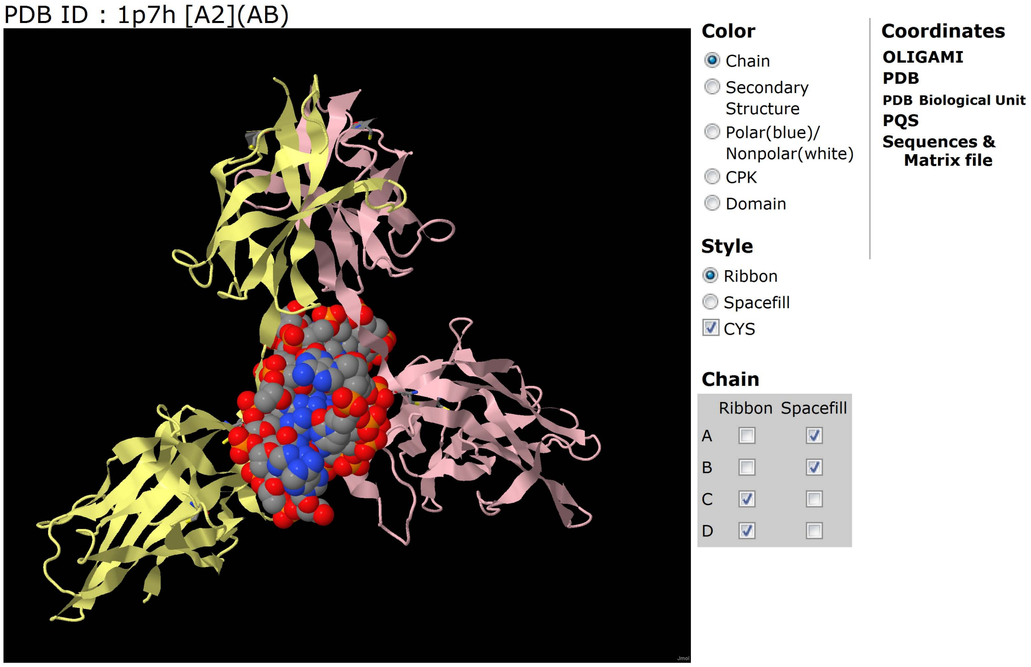
Table 1. Fractions of Each Oligomer Type in PDB

	OLIGAMI[a]	PQS[b]
Monomer	50%	38%
Homo-oligomer	38%	42%
Hetero-oligomer	12%	20%

[a]OLIGAMI percentages were computed based on the number of PDB entries with the exception of nucleic acid and nucleic acid/protein complex-related entries.
[b]PQS percentages were computed based on the number of PDB entries with the exception of nucleic acid, nucleic acid/protein complexes, and all virus coat assembly–related entries.

OLIGAMI COORDINATE AND VIEWER INFORMATION

As described above, the OLIGAMI database contains the original biological coordinate files, and the OLIGAMI web site allows users to interactively view three-dimensional structures. We selected Jmol (http://jmol.sourceforge.net/) as the molecular viewer because it is a cross-platform application that runs on Windows, Mac OS X, and Linux/Unix systems. We modified the chain names of our coordinate files to view all chains using Jmol viewer because it does not draw molecules that have the same chain names. For example, for a dimer protein with chain names A and B in a PDB file, Jmol draws two molecules. However, if chain names are defined as A and A, Jmol draws only one molecule. HETERO molecules that contact the chains included in the asymmetric unit, but not in the biological unit, are removed from the OLIGAMI coordinate file. For NMR data, we extract only the first model from PDB files. Furthermore, as


PDB ID : 1p7h [A2](AB)

Color
● Chain
○ Secondary Structure
○ Polar(blue)/ Nonpolar(white)
○ CPK
○ Domain

Style
● Ribbon
○ Spacefill
☑ CYS

Chain

	Ribbon	Spacefill
A	□	☑
B	□	☑
C	☑	□
D	☑	□

Fig. (2). Molecular viewer page in OLIGAMI.

Users can change the color and style using radio buttons. Users can switch on and off the stick model of cysteine (CYS) residues, the model for each chain, and HETERO molecule. This page also contains four links to the coordinate files. Jmol viewer will not display a coordinate file that is larger than 10 MB. The link to an OLIGAMI coordinate file (XXXX_oligami.pdb) can be accessed directly by software such as MDL Chime, which allows users to view large oligomers (although you will need to install MDL Chime on your computer).

shown in Fig. (**2**), users can easily alter the color of SCOP domains, chains, secondary structures, polar/non-polar, or atom types using radio buttons; users also can select ribbon or space-filling models and switch on and off the model of each chain.

CURATION OF CHAIN FORMULA AND COORDINATES IN OLIGAMI

As of February 2008, PDB contained over 48,000 entries. It is not productive to check the references for every PDB entry to see whether the protein of the PDB entry actually exists as an oligomer as described in REMAERK 350. To select the PDB entry that should be checked, we performed the following operation.

For many proteins, there are several PDB entries. So, we compare the OLIGAMI chain formulas of PDB entries in a protein and group the proteins into four levels described below (Table **2**). We use only the protein portion of the OLIGAMI chain formula for comparison. Each protein was identified by the lowest hierarchy, species, of SCOP 1.73 release. SCOP 1.73 release includes 13,198 proteins in the species hierarchy and contains 34,494 PDB entries.

Level 1: There are four or more PDB entries for a protein and their OLIGAMI chain formulas are identical.

Level 2: There are four or more PDB entries for a protein and is a predominant OLIGAMI chain formula which accounted for greater than 75% in a protein.

Level 3: There are four or more PDB entries for a protein and is not a predominant OLIGAMI chain formula.

Level 4: Proteins contain fewer than four PDB entries.

Next step, the proteins are automatically verified by the following criterion. If the protein does not fulfill the criterion, we check the references for their PDB entries and cu-

Table 2. Four Levels for Checking the Biological Units

		Number of proteins (verified)		Number of PDB entries (verified)	
Level 1	All same chain formulas	1,576	(1,475)	9,224	(8,721)
Level 2	Predominant chain formula (greater than 75 %)	762	(369)	8,860	(4,948)
Level 3	No predominant chain formula	801	(231)	6,042	(1,409)
Level 4	Less than 4 entries	10,059	(297)	12,260	(410)
	Total	13,198	(2,372)	34,494	(14,856)

rate them if the oligomeric state of OLIGAMI is different from that in the references.

i For level 1, OLIGAMI and PQS chain formulas for each PDB entry in a protein are deemed similar if they are more than 70% consistent, and 1,119 proteins (6,152 PDB entries) were automatically identified as verified proteins. If the consistency is less than 70%, we then check their references, curate them and 1,475 proteins (8,721 PDB entries) were totally identified as verified proteins.

ii For level 2 and 3, we first select the proteins that contain chain formulas differing in only the number of different chains, such as [A] and [AB], and that do not contain chain formulas with the same number of different chains but differing in the total number of chains, such as [AB] and [A2B2]. Then, we compare the OLIGAMI and PQS chain formulas for each PDB entry in a protein and automatically identify 311 proteins as verified proteins that the consistency is greater than 70%. If the consistency is less than 70%, we then check their references, curate them and 369 proteins (4,948 PDB entries) for level 2 and 231 proteins (1,409 PDB entries) for level 3 were totally identified as verified proteins.

For example, we manually verified chain formulas and coordinates for two PDB entries of goat alpha-lactalbumin (PDB ID: 1HFY, 1HMK), which is classified as a member of the C-type lysozyme family. Alpha-lactalbumin is a monomeric protein in solution, and it is frequently used as a model protein in protein folding studies [13-16]. When comparing the two PDB entries for goat alpha-lactalbumin, one PDB entry contained the chain formula [A2] and the other entry contained the chain formula [A]. In this PDB entry, we manually corrected chain formula [A2] to chain formula [A]. To date, we have verified 2,372 proteins (14,856 PDB entries), of which we have corrected 598 PDB entries.

FUTURE DEVELOPMENTS

We plan to add information for the interface residues involving protein-protein interactions. OLIGAMI database information will also be provided in a downloadable format.

ACKNOWLEDGEMENTS

We thank Taishi Yanashima of SOFTCREATE CO., LTD. (http://www.softcreate.biz/) for the database and web site design.

REFERENCES

[1] T.J. Hubbard, B. Ailey, S.E. Brenner, A.G. Murzin, and C. Chothia, "SCOP: A structural classification of proteins database", *Nucleic Acids Res.*, vol. 27, pp. 254-256, Jan 1999.

[2] C.A. Orengo, F.M. Pearl, J.E. Bray, A.E. Todd, A.C. Martin, L. Lo Conte, and J.M. Thornton, "The CATH database provides insights into protein structure/function relationships", *Nucleic Acids Res.*, vol. 27, pp. 275-279, Jan 1999.

[3] F.P. Davis, and A. Sali, "PIBASE: A comprehensive database of structurally defined protein interfaces", *Bioinformatics*, vol. 21, pp. 1901-1907, May 2005.

[4] C. Winter, A. Henschel, W.K. Kim, and M. Schroeder, "SCOPPI: A structural classification of protein-protein interfaces", *Nucleic Acids Res.*, vol. 34, pp. D310-D314, Jan 2006.

[5] E.R. Jefferson, T.P. Walsh, T.J. Roberts, and G.J. Barton, "SNAPPI-DB: A database and API of structures, interfaces and alignments for Protein-Protein interactions", *Nucleic Acids Res.*, vol. 35, pp. D580-589, Jan 2007.

[6] Q. Xu, A. Canutescu, Z. Obradovic, and R.L. Dunbrack, Jr., "ProtBuD: A database of biological unit structures of protein families and superfamilies", *Bioinformatics*, vol. 22, pp. 2876-2882, Dec 2006.

[7] E.D. Levy, J.B. Pereira-Leal, C. Chothia, and S.A. Teichmann, "3D complex: A structural classification of protein complexes", *PLoS Comput. Biol.*, vol. 2, pp. e155, Nov 2006.

[8] K. Henrick, and J.M. Thornton, "PQS: A protein quaternary structure file server", *Trends Biochem. Sci.*, vol. 23, pp. 358-361, Sep 1998.

[9] E.R. Jefferson, T.P. Walsh, and G.J. Barton, "Biological units and their effect upon the properties and prediction of protein-protein interactions", *J. Mol. Biol.*, vol. 364, pp. 1118-1129, Dec 2006.

[10] T.M. Nye, C. Berzuini, W.R. Gilks, M.M. Babu, and S.A. Teichmann, "Statistical analysis of domains in interacting protein pairs", *Bioinformatics*, vol. 21, pp. 993-1001, Apr 2005.

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissing, and I.N. Shindyalov, "The protein data bank", *Nucleic Acids Res.*, vol. 28, pp. 235-242, Jan 2000.

[12] T.T. Airenne, J.M. Torkko, S. Van den plas, R.T. Sormunen, A.J. Kastaniotis, and R.K. Wierenga, "Structure-function analysis of enoyl thioester reductase involved in mitochondrial maintenance", *J. Mol. Biol.*, vol. 327, pp. 47-59, Mar 2003.

[13] T.K. Chaudhuri, K. Horii, T. Yoda, M. Arai, S. Nagata, T.P. Terada, H. Uchiyama, T. Ikura, K. Tsumoto, H. Kataoka, M. Matsushima, K. Kuwajima, and I. Kumagai, "Effect of the extra n-terminal methionine residue on the stability and folding of recombinant alpha-lactalbumin expressed in Escherichia coli", *J. Mol. Biol.*, vol. 285, pp. 1179-1194, Jan 1999.

[14] A.C. Pike, K. Brew, and K.R. Acharya, "Crystal structures of guinea-pig, goat and bovine alpha-lactalbumin highlight the enhanced conformational flexibility of regions that are significant for its action in lactose synthase", *Structure*, vol. 4, pp. 691-703, Jun 1996.

[15] H. Van Dael, and A. Chedad, "An equilibrium and a kinetic stopped-flow fluorescence study of the binding of various metal ions to goat alpha-lactalbumin", *J. Fluoresc.*, vol. 16, pp. 361-365, Apr 2006.

[16] A. Chedad, and H. Van Dael, "Kinetics of folding and unfolding of goat alpha-lactalbumin", *Proteins*, vol. 57, pp. 345-356, Nov 2004.

Received: May 26, 2008 Revised: July 17, 2008 Accepted: July 18, 2008