

Studies in Computational Intelligence

Data, Semantics and Cloud Computing

Volume 806

Series editor

Amandeep S. Sidhu, Biological Mapping Research Institute, Perth, WA, Australia
e-mail: dscc@biomap.org

More information about this series at <http://www.springer.com/series/11756>

Chung Yik Cho · Rong Kun Jason Tan
John A. Leong · Amandeep S. Sidhu

Large Scale Data Analytics

Chung Yik Cho
Curtin Malaysia Research Institute
Curtin University
Miri, Sarawak, Malaysia

John A. Leong
Curtin Malaysia Research Institute
Curtin University
Miri, Sarawak, Malaysia

Rong Kun Jason Tan
Curtin Malaysia Research Institute
Curtin University
Miri, Sarawak, Malaysia

Amandeep S. Sidhu
Biological Mapping Research Institute
Perth, WA, Australia

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISSN 2524-6593 ISSN 2524-6607 (electronic)
Data, Semantics and Cloud Computing
ISBN 978-3-030-03891-5 ISBN 978-3-030-03892-2 (eBook)
<https://doi.org/10.1007/978-3-030-03892-2>

Library of Congress Control Number: 2018960754

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1	Introduction	1
1.1	Project Overview	1
1.2	Research Background	2
1.3	Problem Statement	3
1.4	Objective	4
1.5	Outline	4
2	Background	5
2.1	Literature Reviews	5
2.1.1	Process of Life Science Discovery	5
2.1.2	The Biological Data Nature	6
2.1.3	Constant Evolution of a Domain	7
2.1.4	Data Integration Challenges	8
2.1.5	Semantic Integration Challenges	10
2.1.6	Biomedical Ontologies	11
2.1.7	Creation of Ontology Methodologies	12
2.1.8	Ontology-Based Approach for Semantic Integration	16
3	Large Scale Data Analytics	19
3.1	Language Integrated Query	19
3.2	Cloud Computing as a Platform	20
3.3	Algebraic Operators for Biomedical Ontologies	21
3.3.1	Select Operator	21
3.3.2	Union Operator	22
3.3.3	Intersection Operator	23
3.3.4	Except Operator	24
4	Query Framework	27
4.1	Functions for Querying RCSB Protein Data Bank (PDB)	27
4.1.1	Make Query Function	27
4.1.2	Do Search Function	28

4.1.3	Do Protsym Search Function	29
4.1.4	Get All Function	30
4.2	Functions for Looking up Information Given PDB ID	30
4.2.1	Get Info Function	30
4.2.2	Get PDB File Function	31
4.2.3	Get All Info Function	31
4.2.4	Get Raw Blast Function	32
4.2.5	Parse Blast Function	32
4.2.6	Get Blast Wrapper Function	33
4.2.7	Describe PDB Function	33
4.2.8	Get Entity Info Function	34
4.2.9	Describe Chemical Function	35
4.2.10	Get Ligands Function	35
4.2.11	Get Gene Ontology Function	36
4.2.12	Get Sequence Cluster Function	37
4.2.13	Get Blast Function	38
4.2.14	Get PFAM Function	38
4.2.15	Get Clusters Function	39
4.2.16	Find Results Generator Function	39
4.2.17	Parse Results Generator Function	39
4.2.18	Find Papers Function	40
4.2.19	Find Authors Function	41
4.2.20	Find Dates Function	42
4.2.21	List Taxonomy Function	42
4.2.22	List Types Function	43
4.3	Functions for Querying Information with PDB ID	44
4.3.1	To Dictionary Function	44
4.3.2	Remove at Sign Function	44
4.3.3	Remove Duplicates Function	44
4.3.4	Walk Nested Dictionary Function	45
5	Results and Discussion	47
5.1	Query Web Portal	47
5.2	Summary	50
6	Conclusion and Future Works	51
6.1	Conclusion	51
6.2	Limitations	52
6.3	Future Works	52
	Appendix	53
	Bibliography	87

List of Figures

Fig. 2.1	Process of life science discovery.....	6
Fig. 2.2	Process of On-To-Knowledge.....	14
Fig. 2.3	Ontology development with On-To-Knowledge.....	15
Fig. 2.4	OPSDS architecture.....	17
Fig. 2.5	Process of global ontology.....	17
Fig. 3.1	Usage of select operator in instances of family concept.....	22
Fig. 3.2	Usage of union operator.....	23
Fig. 3.3	Usage of intersection operator.....	25
Fig. 4.1	Make query function.....	28
Fig. 4.2	Do search function.....	29
Fig. 4.3	Do protsym search function.....	29
Fig. 4.4	Get all function.....	30
Fig. 4.5	Get info function.....	31
Fig. 4.6	Get PDB file function.....	31
Fig. 4.7	Get all info function.....	32
Fig. 4.8	Get raw blast function.....	32
Fig. 4.9	Parse blast function.....	33
Fig. 4.10	Get blast wrapper function.....	33
Fig. 4.11	Describe PDB function.....	34
Fig. 4.12	Sample output for describe PDB function.....	34
Fig. 4.13	Get entity info function.....	34
Fig. 4.14	Sample output for get entity info function.....	35
Fig. 4.15	Describe chemical function.....	35
Fig. 4.16	Sample output for chemical function.....	35
Fig. 4.17	Get ligands function.....	36
Fig. 4.18	Sample output for get ligands function.....	36
Fig. 4.19	Get gene ontology function.....	36
Fig. 4.20	Sample output for get gene ontology function.....	37
Fig. 4.21	Get sequence cluster function.....	37
Fig. 4.22	Sample output for get sequence cluster function.....	37
Fig. 4.23	Get blast function.....	38

Fig. 4.24	Sample output for get blast function	38
Fig. 4.25	Get PFAM function	38
Fig. 4.26	Sample output for get PFAM function	39
Fig. 4.27	Get clusters function	39
Fig. 4.28	Sample output for get clusters function	39
Fig. 4.29	Find results generator function	40
Fig. 4.30	Sample output for find results generator function	40
Fig. 4.31	Parse results generator function	40
Fig. 4.32	Find papers function	41
Fig. 4.33	Sample output for find papers function	41
Fig. 4.34	Find authors function	41
Fig. 4.35	Sample output for find authors function	42
Fig. 4.36	Find dates function	42
Fig. 4.37	List taxonomy function	42
Fig. 4.38	Sample output for list taxonomy function	43
Fig. 4.39	List types function	43
Fig. 4.40	To dictionary function	44
Fig. 4.41	Remove at sign function	44
Fig. 4.42	Remove duplicates function	45
Fig. 4.43	Walk nested dictionary function	45
Fig. 5.1	Homepage of query web portal	48
Fig. 5.2	Search page of query web portal	48
Fig. 5.3	Search result for keyword 'crispr'	48
Fig. 5.4	Information related to protein ID '1WJ9'	49
Fig. 5.5	Detailed information of protein ID '1WJ9'	49
Fig. 5.6	Contact page of query web portal	50

List of Tables

Table 1.1	Outline.....	4
-----------	--------------	---