

Biocloud: A Systemic Review and Classification

Hongfeng Zhang^{1*}, Vincent Y. Liu², Yu Zhao²

¹Macau Polytechnic Institute, Macau.

²Macau University of Science and Technology, Taipa, Macau.

* Corresponding author. Email: ydliu@must.edu.mo.

Manuscript submitted June 7, 2014; accepted May 8, 2015.

doi: 10.17706/jsw.10.6.695-712

Abstract: Nowadays, numerous journals have published papers related to biocloud – cloud computing techniques for biological applications (e.g., genome analysis, systems biology, infection disease), indicating many researchers' interest. Therefore, the goal of this paper is to review the works that were published in journals from 2006 to 2013, suggest a classification framework of cloud computing techniques and related issues for biological applications, and explore each feature of classification framework. The literature review is conducted by using an abstract and topic search. The classification framework is developed based on cloud computing architecture and IT service management, which consists of the following six layers: concept and research layer, SaaS layer, PaaS layer, IaaS layer, DaaS layer and service management layer. Current problems and future research direction of biocloud are discussed at the end this paper.

Key words: Biocloud, cloud computing, systemic review.

1. Introduction

Currently, nearly all research has become data-centric. Novel computational approaches and enhanced capabilities are required by bio-community (e.g., larger data sequencing, research idea sharing, biodiversity research). Cloud computing should be an ideal choice as it can provide software applications and computing power (e.g., data mining, collaboration) to users as a service over the Internet [1]. Many research institutes, biotech companies, pharmaceutical companies, and computational labs start to adopt cloud computing to process and store the vast amount of biomedical data. The application of cloud computing for genomics, systems biology, biomedical data mining, health care and biodiversity research could be described as a new concept – biocloud. As a type of community cloud, biocloud could be filled with data relevant to biology, medicine and health care with the computing intensive operations, virtualization function, large number of machines and parallel computing (e.g., MapReduce, Hadoop) [2], [3]. Researches in next-generation sequencing (NGS), comparative genomics, and proteomics have already adopted biocloud to deal with data processing operation successfully [4]. For example, as one of the key cloud technologies for large-scale data analysis, MapReduce is helpful for genetic data intensive analysis [5]. The Mashup approach could be an ideal choice for bioinformatics resources delivery through Internet [5]. Amazon Elastic Compute Cloud (EC2) has been chosen as a platform for enormous number of bioinformatic data and services storage and analysis [5]. Individuals could access the data and associated service through personal devices from different locations in the world. Biocloud has revolutionized biology research, making it difficult to overstate its impact on future studies.

After first proposed the concept of cloud computing in 2006, many researchers have made efforts to

design and implement network, infrastructure, middleware, and applications which provide users with biocloud services effectively. As an exciting new paradigm, biocloud attracts many researchers' attention, and numerous journals and conferences have published papers related to this emerging topic. Therefore, we feel that this is a good time for a review analysis, since it has been five years or more since many papers were published. Currently, it is difficult to review and compare these published papers, because the available research is published in quite different journals. Prior to this study, R. L. Grossman and K. P. White presented a vision for a biomedical cloud and discussed the application for genomics, systems biology and biomedical data mining [2]. They also discussed the need of an adequate computing infrastructure for dealing with the generated data in the field of biomarkers, genetic variants, and environmental variables, the amount of which is staggering [2]. Other application fields of biocloud, including healthcare and biodiversity research, were not mentioned in the research. Therefore, a more comprehensive survey research on biocloud is needed. Other researchers, such as M. A. Titmus, *et al.*, W. R. Kubick, J. Dudley, *et al.*, R. Barga, *et al.*, and L. Stein also described a big picture of biocloud [4]-[9]. However, these researches did not provide a broad presentation of current research issues, literature review related with each biocloud layer and suggestion for further study. Accordingly, the main objectives of this paper are:

- To classify and summarize research relevant for biocloud.
- To provide a conceptual framework for the integration and classification of published papers.
- To derive suggestions and future directions for biocloud researchers based on the literature review.

The remainder of this paper is organized as follows. In Section 2, a methodology to search and extract the published papers through online database is illustrated. Section 3 describes general characteristics of the literature research. Section 4 provides the proposed classification framework. Section 5 illustrates each feature of classification framework (concept and research, SaaS, PaaS, IaaS, DaaS and service management) and contains discussions and future research directions. We conclude the paper in Section 6 with brief concluding remarks.

Table 1. The Online Databases and Subjects.

Database	Subjects	Database	Subjects
Science Direct	Human disease	ACM Digital Library	Communication, Networking
	SLA violation		Computing & Processing
	Virtual machine		Biomedical research
	Landsat datum		Benchmarking
Springer Link	Microbial sequence	Biological Sciences	Bioinformatics
	Biomedical and life Sciences		Data processing
	Computer science		Organs
	Engineering		Index medicus
IEEE Xplore	Bioengineering	Health Reference Center	Clouds
	Engineering profession		Software
	Fields, Waves		Internet
	Power, Energy		Sequence analysis, DNA
Wiley InterScience	Genome	EBSCO	Genotypes
	Heterozygote		Influenza
	Hospitals		Epidemics
	Information dissemination		Biomedical research
	Informed consent		RNA
	Information storage and retrieval		Computational biology
	Quality of health care		DBMS software
			Genetic markers
			Computerized medical records systems

2. Procedures

A total of 76 journal papers from 2006 to 2013 were obtained and reviewed. These papers were found via

computerized search of the abstract, topic, index term or keywords areas. The search was used the terms “cloud computing” and bio*. The data collection method was learned from J. Hong *et al.*'s research on context-aware systems [10]. A detail illustration of methodology for paper - collection is described in the following section.

2.1. The Selection Criteria

The selection criteria were used to search and select the biocloud papers. These downloaded papers which do not meet the selection criteria were excluded. The three criteria are described as follows:

- 1) The literature was based on a search in the descriptors for “cloud computing” and bio*. Various online journal databases were selected to search biocloud literatures. Biocloud papers were found in comprehensive subjects such as computer applications, bioinformatics, computational biology, etc. Main subjects of online database which are summarized according to fields of biocloud are shown in Table 1.
- 2) This research surveys the papers published from 2006 to 2013. The reason for selecting this time period is that researchers started to discuss cloud computing since 2006. In fact, many journals and conferences have published researches related to biocloud since 2008.
- 3) This research covers only journal papers. Other publication forms (e.g., conference proceedings, working papers, dissertations, newspapers and books) were not included. Furthermore, the poster presentation, proceedings and abstract published in journals are excluded. We did not consider journals which only publish proceeding papers either. The reason for choosing the journal papers as the searching source is that papers published in journals represent a higher level of research.

2.2. Data Sources and Procedures to Extract Downloaded Papers

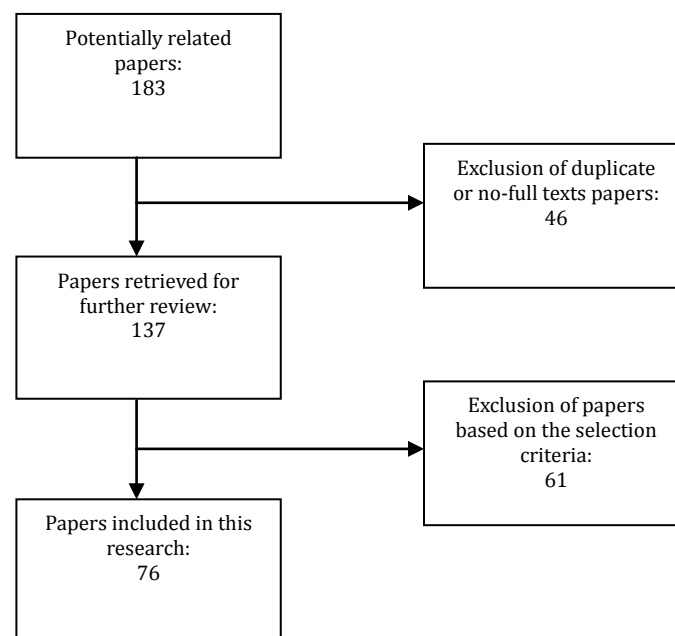


Fig. 1. The procedures to select the articles.

The papers were selected according to the procedures shown in Fig. 1. First of all, the papers were searched using 11 online databases. The total number of downloaded papers is 183. The number of papers by each online database is as follows: Science Direct (19), Springer Link Online Libraries (21), ACM Digital Library (2), IEEE Xplore (20), Wiley InterScience (21), BioMed (21), Biological Sciences (29), Health Reference Center (4), and EBSCO (46). BioOne (0), Informa Healthcare (0). Next, 46 papers were excluded

because they are duplicate papers or there is no full text provided for these papers. Next, the papers were carefully reviewed to select those that considered biocloud as the main part. Finally, 61 papers were excluded as these researches do not cover biocloud topics—they discussed biological-based systems, bio-inspired computing, self-maintenance systems or complex systems. The “section news” articles within different journals did not meet the second selection criteria either. Based on these procedures, a total of 76 articles met all the selection criteria.

3. The General Characteristics of the Papers

3.1. Classification of Papers by Publication Year

The number of papers by publication year is depicted in Fig. 2. Biocloud papers have appeared since 2008 – almost two years after the first appearance of cloud computing concept. The number of papers in 2010 have becomes 3 times more than the number of papers in 2009. The number of papers from 2009 to 2012 had been increased continuously. It is absolute that the concern about biocloud is increased and will be continued.

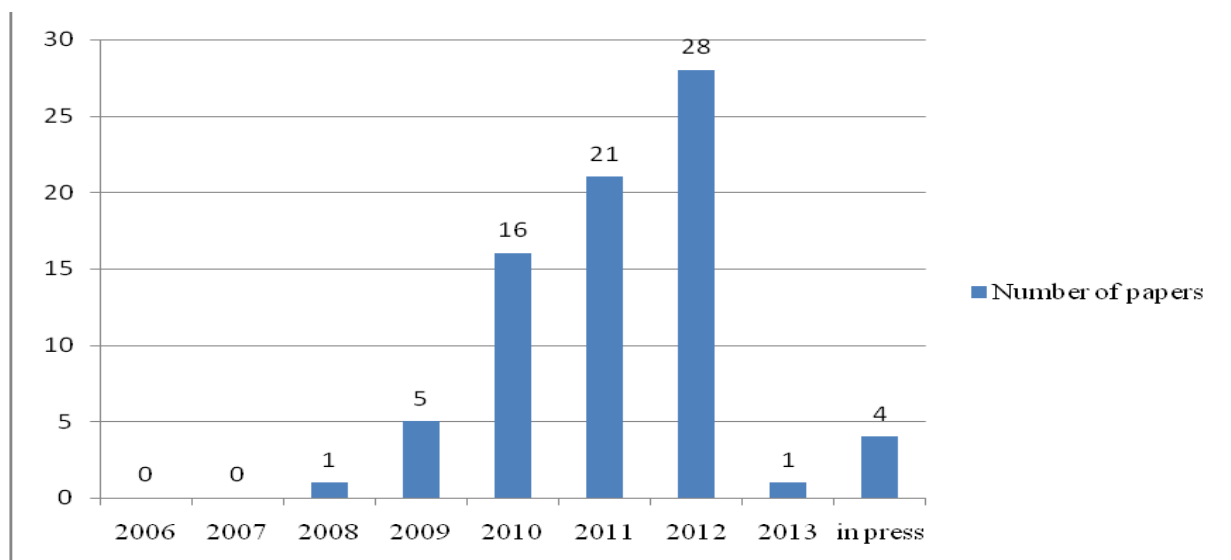


Fig. 2. Classification of articles by publication year.

3.2. Classification of Papers by Online Database

The papers by online database are categorized in Table 2. There are a total of 76 papers from 8 different databases. In Table 2, Wiley InterScience covers the highest percentage of papers (20 papers, 26.3%), because it offers papers of various study fields. BioMed stores the second highest of journal papers (14 papers, 18.4%). The results of other online databases are: Science Direct (10 papers, 13.1%), EBSCO (9 papers, 11.8% – within Academic Search Complete, Medline, Cinahl, Library Literature & Information Science Full Text and Library and Information Science & Technology Abstracts), Springer Link Online Libraries (8 papers, 10.5%), IEEE Xplore (8 papers, 10.5%), Biological Sciences (6 papers, 7.9%) and Health Reference Center (1 paper, 1.3%). As mentioned earlier, there are papers which were found repetitively among different online databases. Therefore, we add one column in Table 2 to show actual number of papers found in selected online databases.

3.3. Classification of Papers by Journal

The selected papers are categorized by related journals in Table 3. There were 44 journals that published

biocloud papers. Most of them were related to the bioinformatics, genome biology and computer science. Table 3 specifies journals that published two or more biocloud papers. Others journals are omitted. Parallel and distributed systems is closely connected with cloud computing. Therefore, journals which focus on distributed systems were also found that they have published papers relevant to biocloud. As Table 3 shows, “Concurrency and Computation: Practice and Experience” published the most papers on biocloud (10 papers, 13.2%). BMC Bioinformatics published 8 papers (10.5%). Future Generation Computer Systems and Genome Biology published 4 papers (5.3%). This result shows that various journals have published papers relevant to biocloud.

Table 2. Classification of Articles by Publication Year.

Online database	Number of papers	Actual Number of papers
Science Direct	10	10
Springer Link Online Libraries	8	8
IEEE Xplore	8	8
Wiley InterScience	20	20
BioMed	14	19
Biological Sciences	6	16
Health Reference Center	1	1
EBSCO	9	18
Total	76	100

Table 3. Classification of Papers Based on the Journal.

Journals	Number of papers
Concurrency and Computation: Practice and Experience	10
BMC Bioinformatics	8
Future Generation Computer Systems	4
Genome Biology	4
BMC Research Notes	3
Computer	3
IEEE Transactions on Parallel and Distributed Systems	3
Bioinformatics	2
PLoS computational biology	2
PLoS ONE	2
Transactions in GIS	2
The others	33
Total	76

4. Classification Framework

4.1. Abstract Architecture of Biocloud

The abstract architecture of biocloud is drawn based on the literature that explores the cloud, cloud computing, and cloud technology to offer classification criteria for dividing the literature appropriately. B. P. Rimal and E. Choi provided a taxonomical spectrum for cloud computing, including architecture, virtualization management, data governance, security, management service and fault tolerance. The taxonomical spectrum focused on the service engineering perspective of cloud [10]. L. Wang *et al.* discussed three functional aspects of cloud computing: Hardware-as-a-Service (HaaS), Software-as-a-Service (SaaS) and Data-as-a-Service (DaaS) [11]. Therefore, J. Ekanayake *et al.* differentiated the two concepts: cloud and cloud technology [12]. “Cloud” refers to a collection of infrastructure services, including Infrastructure-as-a-service (IaaS), Platform-as-a-Service (PaaS), etc [12]. “Cloud technologies” refers to various cloud runtimes such as Hadoop, Dryad, and other MapReduce frameworks [12]. As the results of literature review related with biocloud, we present general abstract layer architecture of biocloud. As Fig. 3 shows, this architecture consists of five layers: (1) Software-as-a-Service (SaaS) layer, application layer

provides users with appropriate software as a service; (2) Platform-as-a-Service (PaaS) layer, middleware layer provides developers with a platform including all the systems and environments [10]; (3) Infrastructure-as-a-service (IaaS) layer involves a collection of infrastructure services; (4) Data-as-a-Service (DaaS) layer involves data with various formats or from multiple sources which could be accessed via services by users on the network [11]; (5) Service management layer involves deployment, configuration, reporting [10]. Currently, IT service management plays an important role to guarantee the implementation quality for information technology based on different requirement.

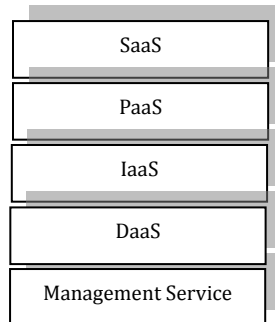


Fig. 3. Abstract layer architecture of biocloud.

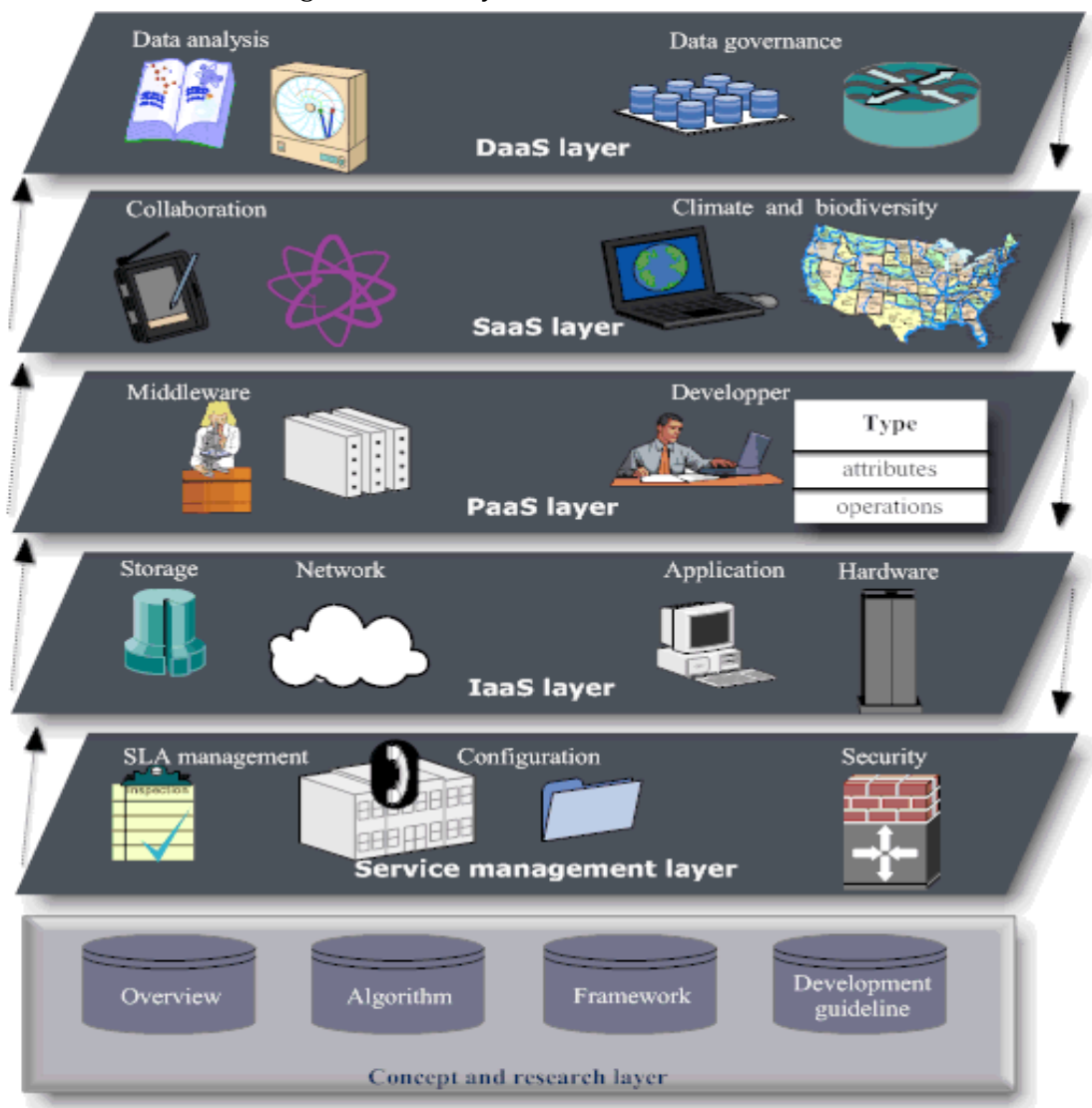


Fig. 4. Classification framework of biocloud.

4.2. Classification Framework of Biocloud

The classification framework is developed for classifying the selected biocloud literatures based on the biocloud abstract architecture in section 4.1. The classification framework developed here consists of the following six layers: concept and research layer, SaaS layer, PaaS layer, IaaS layer, DaaS layer and management service layer. Without the concept and research layer, layers of classification framework and layers of biocloud architecture are the same. As Fig. 4 shows, each layer has detailed categories for dividing the literatures in layers.

Fig. 4 depicts a graphical classification framework for biocloud. Biocloud provide applications and services based on the support of PaaS layer and IaaS layer. Moreover, services involved in IaaS layer are based on IT service management activities. Service management layer involves service level agreement (SLA) management, configuration and security management. IaaS layer involves storage, network, hardware and application [10]. PaaS layer involves developer platform and middleware. DaaS is classified as data analysis and data governance (e.g., data interoperability, data migration) [10]. Concept and research layer consists of overview research, framework, algorithms/programming model and development guideline. The SaaS layer could be more diversity – we only list climate and biodiversity application (e.g., GIS), collaboration function based on the selected publications.

5. Discuss and Suggestions

The papers by subjects are categorized in Tables 4. Table 4 shows the number of papers, percentage of subject, and percentage of all subjects, while Table 5 represent all references of the biocloud papers.

The concept and research category has the highest percentage of biocloud papers (36 papers, 47.4%). In biocloud concept and research, the category can be divided into 4 subjects. Most of the papers are related to “overview” (18 papers). Other papers are related to “framework” (7 papers), “algorithm” (8 papers) and “development guideline” (3 papers). The second highest percentage of biocloud papers is related to IaaS (20 papers, 26.3%), which has four subjects. The majority of this section is related to the “network” (15 papers). Other papers within “IaaS” category are related to “storage” (1 paper), “hardware” (2 papers) and “application” (2 papers). The amount of papers related to DaaS is 9 (11.8%). The papers of DaaS part can be also divided into two subjects. These papers are related to “data governance” (5 papers) and “data analysis” (4 papers). The SaaS category involves four papers (11.8%), which can be divided into two subjects: “climate and biodiversity” (2 papers) and “collaboration” (2 papers). The PaaS category also involves four papers (11.8%) with two subjects: “developer platform” (3 papers) and “middleware” (1 paper). The fewest of biocloud papers is related to service management (3 papers, 3.9%). This category can be divided into three parts, “service level agreement (SLA) management” (1 paper), “configuration” (1 paper) and “security management” (1 paper).

In result, the concept and research has the highest percentage of biocloud papers, in which the most explored subject is the overview research. This represents that the biocloud studies are in developing stages and much research have been conducted in discussing the concept, feature, benefit and application of biocloud. The fewest of biocloud studies is related to IT service management. This result indicates the current attention on biocloud service management is not enough. While, we need to note that IT service management research is still in its developing phase, although IT service management practices play an important role to guarantee the continuity of different operations.

Fig. 5 shows the number of distribution of papers by the year and classification framework. The total number of papers within concept and research layer and IaaS layer has been increased continuously from 2008 to 2013. Most papers which focus on framework, algorithm, hardware, application are published since 2011. It means that bio-community begins to rely on consortium grids or cloud, and the concern about

technical detail of biocloud has increased recently. This conclusion can also be supported by the distribution of PaaS papers by the publication year – researchers have discussed scriptable integration biocloud platform to develop business application and various services over the internet since 2009. The paper distribution result of DaaS and SaaS layers is interesting: there are 9 DaaS papers and 4 SaaS papers. Researchers started to discuss data analysis and data migration of biocloud since 2009, while all the SaaS papers appear since 2012, most of which are concerned with geographic information system (GIS). This phenomenon indicates that most bio-researchers choose biocloud as a data mining or scientific computing tool to deal with the enormous number of bioinformatic data resources and services at the beginning stage, instead of adopting or developing other applications or services (e.g., collaboration, mass-spectrometry, biological imaging, smarter healthcare). We also note that DaaS papers are decreased after the peak publishing point in 2010. This phenomenon was caused by recent research that covers several subject of data analysis or data governance from concept and research category (e.g., algorithm). In other words, the tendency of recent concept and research category focuses not only on basic overview but also implantation or case-study. All the service management papers appeared after 2011. Biocloud can provide better service for users from multiple institutions. However, huge amount of data stored within biocloud cause risk of security and privacy relevant to personal information. IT service management approaches or guidelines should be discussed with the advance of biocloud.

Table 4. Classification of Papers by Subject

Classification criteria	Number of papers	Percentage of subject	Percentage of all subjects
1. Concept and research	36	100.0%	47.4%
1.1 Overview	18	50.0%	23.7%
1.2 Framework	7	19.4%	9.2%
1.3 Algorithm	8	22.2%	10.5%
1.4 Development guideline	3	8.3%	3.9%
2. DaaS	9	100.0%	11.8%
2.1 Data governance	5	55.6%	6.6%
2.2 Data analysis	4	44.4%	5.3%
3. SaaS	4	100.0%	5.3%
3.1 Climate and biodiversity	2	50.0%	2.6%
3.2 Collaboration	2	50.0%	2.6%
4. PaaS	4	100.0%	5.3%
4.1 Developer platform	3	75.0%	3.9%
4.2 Middleware	1	25.0%	1.3%
5. IaaS	20	100.0%	26.3%
5.1 Storage	1	5.0%	1.3%
5.2 Network	15	75.0%	19.7%
5.3 Hardware	2	10.0%	2.6%
5.4 Application	2	10.0%	2.6%
6. Service Management	3	100.0%	3.9%
6.1 SLA management	1	33.3%	1.3%
6.2 Configuration	1	33.3%	1.3%
6.3 Security management	1	33.3%	1.3%

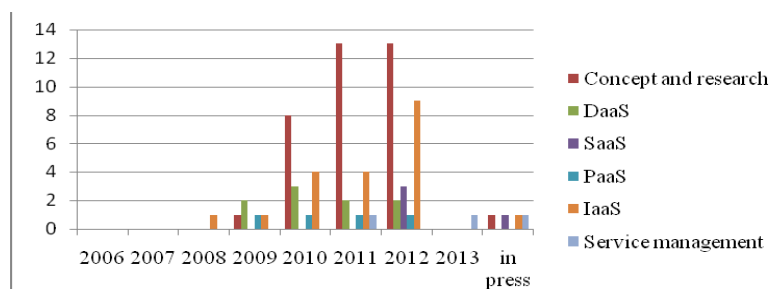


Fig. 5. Distribution of paper number by year and classification framework.

Table 5. References of Concept and Research Layer

Classification criteria			References
Concept and research	Overview		D. Reed (2011) [1], R. L. Grossman and K. P. White (2012) [2], R. Barga, D. Gannon, J. P. Hamilton and C. Robin Buell (2012) [4], M. A. Titmus, <i>et al.</i> (2012) [5], W. R. Kubick (2011) [6], J. Dudley, <i>et al.</i> (2010) [7], R. Barga, <i>et al.</i> (2011) [8], K.C. L. Stein (2010) [9], Cheng <i>et al.</i> (2012) [13], T. C. Glenn (2011) [14], S. E. Morales and W. E. Holben (2011) [15], A. N. Desai and A. Jere (2012) [16], P. Ertl (2010) [17], N. C. N. C. Roy, <i>et al.</i> (2011) [18], L. Zhang, <i>et al.</i> (2012) [19], P. Kudtarkar, <i>et al.</i> (2010) [20], T. Casci (2010) [21], L. Milanesi, <i>et al.</i> (2009) [22]
	Framework		A. Rosenthal, <i>et al.</i> (2010) [3], T. Gunarathne, <i>et al.</i> (2011) [23], E. S. Dove, <i>et al.</i> (2012) [24], J. M. Hogan, <i>et al.</i> (2011) [25], V. A. V. A. Fusaro, <i>et al.</i> (2011) [26], G. P. Krestin, <i>et al.</i> (2012) [27], K. Wilcox and T. C. Vance (2011) [28]
	Algorithm		J. Ekanayake, <i>et al.</i> (2011) [12], A. Gonzalez and R. Knight (2012) [29], H. Lee and M. C. Schatz (2012) [30], T. Dalman, <i>et al.</i> (in press) [31], J. Almeida, <i>et al.</i> (2012) [32], M. Rahman, <i>et al.</i> (2011) [33], P. Chouvarine, <i>et al.</i> (2012) [34], F. Sangermano, <i>et al.</i> (2010) [35]
	Development guideline		A. Wilke, <i>et al.</i> (2011) [36], M. Pérez-Enciso and L. Ferretti (2010) [37], I. Lanc, <i>et al.</i> (2012) [38]

5.1. Concept and Research Layer

The published papers of concept and research layer are listed in Table 5. “Concept and research layer” involves current research, future trend and foundation of biocloud. We categorize this layer into general overview, biocloud framework, biocloud algorithm and biocloud development guidelines. With the widespread popularity of genomic applications, the “bioinformatics bottleneck” issue appears due to the uncertainty about the cost and infrastructure needed to meet increasing demands for large amount sequence analysis [39]. By comparing with traditional scientific data centers, biocloud could offer economies of scale and the ability to adjust to workload variations [3], [39]. Especially, with MapReduce or DryadLINQ based frameworks, cloud infrastructure services could provide users with scalable, highly available alternatives without the burden of managing their traditional counterparts [23]. The MapReduce framework has been adopted by many organizations as a very promising solution to map sequencing reads and reference genome sequences by utilizes hundreds or thousands of shared computers in a short time [40]. Mashup framework should be an ideal choice for delivering the vast amount of web-based bioinformatics data – the web becomes an operating system neutral and makes the Mashup be usable by anyone through web browser [3]. With the sharing and reuse feature, Mashup framework provides a real incentive environment for creating specialized bioinformatics applications and eliminating the hosting barrier [25]. A biomashup community could offer a rich library of components to support a new style of bioinformatic experiment [25]. However, large data processing and provenance record maintaining are the two challenges for biomashup implementation –appropriate and scalable algorithms are needed to achieve effective biocloud operation.

In fact, the increasing adoption of distributed cloud computing resources has positioned the MapReduce coding pattern as a foundation of scalable bioinformatics algorithm development [32]. Algorithm is an essential part in developing biocloud, and this subject involves papers related to genomic sequence analysis, SNP comparative analysis, read mapping in molecular biology and large data processing. The algorithm research within concept and research layer could be divided into three subjects: (1) designing and modeling algorithm for sequence analysis. For example, the sequence comparison algorithm by J. Almeida, *et al.* which can be thoroughly decomposed into multiple rounds of map and reduce operations [32]. From a technological point of view, next generation sequencing algorithm should deal with targeted sequencing and multiple tagging, instead of rather than total sequence throughput, (2) increasing accuracy and efficiency of algorithm for genomic sequence analysis and efficiently mapping reads to the genome, and (3)

developing algorithm or method to process terabytes of new data to biocloud (e.g., SNP frequency data parallels). After all the data are collocated in one place, algorithms could be used to integrate and process all the available data on a continuous basis. In fact, more efficient analysis algorithms are needed not only for the scalable data, but also for the research cost issues – users can not financially afford to continue performing the analysis with the current generation of algorithms [36].

5.2. DaaS layer

Genetic code is the basis of all biological life. The genome and how genes are encoded within the genome, has become a fundamental resource in biology [29]. In the post-genome era, biomedical research is intensely data-driven. The volume and complexity of biomedical data (e.g., microbial sequences result) producing by new technologies and the near-infinite number of bioinformatic data composing ways require effective data governance and data analysis. The DasS layer of biocloud could let the user access and manipulate remote multi-format data from different sources via Internet. For example, users could store and analyze data through Amazon Simple Storage Service (S3). We categorize DaaS layer into data governance and data analysis. There are different types of biocloud analytical tools, including CLoVR, Galaxy, etc. [29]. R. L. Grossman and K. P. White tried to connect phenotypic data to biomarker profiles and therapeutic treatments for creating predictive models of disease detection, progression and therapeutic response [2]. Biological data analysis includes analyzing a wide variety of biological data: (1) genomic data (DNA sequencing, RNA expression, proteomic data, metabolomic data and small molecule screening), (2) information from biological literature, medical records, etc., and (3) image (e.g., X-rays, functional MRI, new types of scanning microscopes) [2].

The multi-sources of data include experimental perturbations, genetic associations, medical text, evolutionary conservation, electronic medical records and large-scale genomic studies [2]. By moving these data in the biocloud, users may lose some capabilities (e.g., (control the action or behavior of data, data regulation, tactical and strategic decisions) [10]. Therefore, biocloud data governance needs to address data quality and consistency, including the prevention of sensitive data from discursing and losing. Successful data governance relies on data interoperability and data migration [10]. For data interoperability, we refer to the ability to communicate between multiple cloud platforms. Data interoperability could allow applications to be ported between clouds [10]. For data migration, we refer to the data migration between data centers or different cloud platforms. In the current situation, data migration needs to consider several architecture and technical issues, including legacy and target environment, data quality, data loss, availability, governance, security, scalability and cost efficiency, etc. [10]. G. Fedak, *et al.* proposed a programmable framework – BitDew. The target of BitDew is for automatic and transparent data management on computational Desktop Grids with multi-protocol file transfer and metadata abstraction [41].

5.3. SaaS layer

SaaS refers to software deployed as a hosted service and cloud be accessed over the Internet [10]. Based on SaaS, the users do not need the software license as well as they do not have to purchase and support the infrastructure that software relies [10]. Google Docs is the example of SaaS key providers. There are two types of software mentioned in the selected SaaS papers: GIS for climate and biodiversity research, and healthcare collaboration service. A biogeographic information system with the entire hardware could be built on a cloud computing environment for storing, querying, disseminating, and mapping a global biogeography data. A healthcare collaboration application could base on smart home technology. A person's vital data can be collected in a smart home environment and the bio-information can then be transferred via the Internet to medical databases and the healthcare professionals. The biocloud healthcare service

could monitor patients' complete health history to improve the quality of care. It enables on-demand, real-time access to patient data and creates collaboration among healthcare providers. Different healthcare providers can thus avoid the investment of duplicated infrastructure. A proper healthcare service should contain the following features:

Low cost and large service scale. Healthcare service is designed for the cloud service provider, healthcare providers, and patients with low operation cost.

Secure data protection. Patient data (e.g., privacy information) must be stored securely and kept for a long time. Proper healthcare service should include secure and auditable mechanisms for medical records accesses.

Patient empowerment. Healthcare service should allow patients to view their own data, regardless of which provider generated them.

SaaS for biocloud is still in its infancy. Future efforts are needed on integrating other applications (e.g., the mass-spectrometry, biological imaging) into the biocloud.

5.4. PaaS layer

PaaS could provide developers with a platform for developing, testing, deploying and hosting of sophisticated web applications as a service. PaaS waves the time-consuming work of creating and maintaining an infrastructure and provides an easier way to develop application and various services over the internet. Google App Engine and Microsoft Windows Azure are the prominent examples of PaaS [10]. We categorize PaaS papers into two subjects: developer platform and middleware. Different with SaaS, the PaaS layer of biocloud could allow experienced users to choose scripting language for quickly executing powerful commands. A PaaS should also provide open APIs that developers could use to create, integrate, and provision existing and new applications and services. Theoretically, SaaS or PaaS services can be set up on top of IaaS platforms. With the development of biocloud, there is an increasing demand for bio-informatically oriented biologists – proper programming ability is needed when available platform cannot fulfill all specific needs for computational biology. We have now arrived at an intersection point between genome technology and cloud computing.

5.5. IaaS layer

We categorize IaaS papers into four subjects: storage, network, hardware and application. As we mentioned, the explosion of bio-data has cause a large and ever-growing need for computational infrastructure to assemble, manage and mine the enormous and rapidly growing data. For example, the bioinformatics community needs to consider how to make use of distributed resources effectively to satisfy the demand for data analysis. An interconnected network of data centre scale facilities with the appropriate security architecture is a proper foundation for the bio-community. IaaS is the delivery of computer infrastructure as a service. Due to the higher flexibility, IaaS could provide the latest technology to users with a usage-based payment scheme. There are many features for IaaS, including dynamic scaling, usage-based pricing, reduced capital and personnel costs and access to superior IT resources, etc. [10]. Typical examples are GoGrid, GNi, Flexiscale, Layered Technologies, etc. [10]. IaaS is also called Hardware-as-a-Service (HaaS). Amazon is an example of HaaS (Amazon Elastic Compute Cloud (EC2)) [10]. Cloud computing infrastructure could allow the user to create a virtual compute cluster through virtual machines on host servers by using the virtualization technology. Virtualization can be found in the form of network, hardware and application [10]. In the field of genome informatics, large datasets typically are stored by taking the form of virtual disk images which can be attached to virtual machines or shared as networked volumes (e.g., GenBank) [9]. Web services that run as virtual machines based on these datasets. Computational biologists could move the compute cluster to the data – they may configure a virtual

machine image with the software they plan to run, instead of moving the data to the compute cluster [9]. The benefits of virtualization include maximization of resources and reduce operation costs. Computational biologists could rent these virtual machines, instead of purchasing server, storage and networking hardware. The network of biocloud requires automated analysis pipelines to guarantee transparency, reproducibility and portability across different operating systems, including the commercial Amazon Elastic Compute Cloud (EC2) [39]. These pipelines could be implemented in virtual machine [39]. In the field of genome informatics, pipelines or workflows could chain a number of discrete compute and data intensive analysis tasks into sophisticated analysis procedures [42].

5.6. Service Management Layer

We categorize service management papers into three subjects: SLA management, configuration and security. Besides the powerful computing and networking power, cloud computing also creates a requirement of configuration options of the application and infrastructure. Currently, IT service management plays an important role to guarantee the service quality of information technology. The service management of cloud computing should consist of deployment, configuration, reporting, SLA etc. and the typical applications include VMware Capacity Planner and CIRBA [10]. Biocloud relies on secure and reliable services for the biomedical-infomics synthesis. The basic best practices for data security includes use public/private key pair, create restricted user accounts and control access using firewalls [3]. However, the IT service management for biocloud maybe a little bit different – biocloud typically bases on large data centers which are acquired and managed by one organization with fairly uniform in configuration. In fact, data center also requires automated configuration and maintenance of individual or networked computers from the policy specification to handle the vast services within cloud environment [10]. SLA is a powerful component to guarantee the vision of biocloud QoS goals and configure resource agreements between service users and service providers (e.g., guarantees on execution time). It involves little human-based interaction in order to guarantee the scalability and efficient biocloud resource utilization [43]. Biocloud users may also make decisions on capacity planning based on SLA due to the biocloud infrastructure is considered to be highly available (e.g., 24/7 running).

G. Future research direction

Biocloud is in its infancy and is still developing. Many researchers have been concerned about applications, services and infrastructures of biocloud, and as shown in Fig. 2, the concern has increased from 2008 to 2013. Moreover, reports or white paper of many biocloud projects have been published increasingly. However, biocloud is not fully implemented. As shown in Table 4, much research has focused on the concept and research layer. The scope of applications or services in most articles is limited to data sequencing and GIS. In addition, strategic alternatives or operation models for gaining the benefit of biocloud are very few. Furthermore, technologies related to biocloud are merely standardized. The architecture, method of biocloud algorithm and network implementation in these published papers is different.

We have now arrived at an intersection point between biology and cloud computing. With the advance of biocloud, there is an increasing demand for bio-informatically oriented biologists to deal with the volume and complexity of biomedical data producing by new technologies. Biocloud also relies on secure and reliable data governance service for the biomedical-infomics synthesis. In the current scenario, biocloud lacks data governance. An effective biocloud should support the scalable ingestion of biological, medical and healthcare data from genomics databases or other clouds. It should also manage and archive all the bioinformatics data by providing a scalable storage based on users' demand as well as a mechanism for exporting data to another cloud or facility [2]. A biocloud could contain a mixture of public data and restricted data (e.g., human genome data). Therefore, documentations of quality, SLA management and

resource monitoring are needed. Specialized and secure private clouds could be needed for the most restricted data.

Besides, the immediate future of biocloud application and service may increasingly lie in everyday life (e.g., health care, smart home) – as researchers are beginning to realize the impossibility of developing systems and application only in laboratory environments. Therefore, possible future research agenda may include answers to the following questions.

What are design patterns of biocloud? Design patterns are defined as recurring solutions to design problems by focusing on the reuse of recurring architectural design themes. Design patterns have been proposed in many domains as a solution to capture and share design knowledge, examples and their solutions to recurring problems between designers or developers. Therefore, design patterns could assist designers in developing biocloud by avoiding the potential design risk. However, very little research has been conducted on design patterns for biocloud and the suggested design patterns are not specific. Therefore, future effort is need on this area. Besides, different APIs are needed by developers when available biocloud platform cannot fulfill all the specific needs for computational biology. How to make these APIs be interoperable or portable? Development guidelines are needed for biocloud to solve the issue.

How to evaluate performance of biocloud? It is essential to improve managing and planning biocloud based on performance evaluation, Recently, various models introduced by IT service management, such as Balanced Score Card (BSC), IS Success Model, Control Objectives for Information and Related Technology (COBIT), Information Economics, and so on, have been applied to evaluate the performance of information systems from different perspectives. Due to the unique characteristics of biocloud such as user-centric interfaces, on-demand service provisioning, autonomous system, scalability and flexibility, the current technology for evaluating information systems (IS) does not provide techniques to formally define, verify, implement, and analyze these biocloud features. Therefore, research on the model and performance measure reflecting the characteristics of biocloud should be conducted.

How to save data safely in biocloud? Users must look towards the safety and segregation of their data in the multi-tenancy architecture of biocloud. Service providers need to assure that data/information of different user can not be accessed without permission. Successful data governance relies on policy and rules. In fact, sharing an individual's genetic information with prospective employers or insurance companies might result in genetic information-based discrimination [16]. However, there is no conceptual framework to integrate and no way to compute over all the data within biocloud [2]. Suitable protocols and secure communications between different biocloud are also needed as bioinformatics data may be located at medical research centers and complex analyses can be carried out between multiple clouds.

6. Conclusions

The review was organized according to the framework developed with the purpose of providing a comprehensive overview of research on biocloud. The review introduced the biocloud concept, infrastructure, platform, application, service management and presented an exhaustive list of each layer of biocloud. In this paper, we reviewed the literatures for the concept and applications, and examined them using dimensions related to ongoing and emerging issues in biocloud. This paper is based on a literature review of biocloud from 2006 to 2013 using a keyword index and abstract search. Overall, we found that the research related with biocloud seems to increase dramatically and can be classified into SaaS layer, PaaS layer, IaaS layer, DaaS layer, service management layer and concept and research layer which provides the base of biocloud infrastructure, application and service. This paper provides sufficient literature for the researchers on the use of biocloud, and we hope that it will motivate the researchers and practitioners to examine biocloud issues and its applications. The framework of biocloud offers a general development guideline, fundamental component and relationships within component of biocloud. Furthermore, on the

basis of this framework and the literature review, selected opportunities for future research were discussed.

Although considerable attention was given to the biocloud framework design and classification of the literature review, some limitations still exist. First, some relevant papers might have been overlooked. Much the literature has been found by the keyword or abstract only. The keyword or abstract describes the content quite well in most cases, but this is not always so – that is why we need to review the downloaded paper carefully and select the qualified ones based on the criteria in Section 2. In order to conduct a comprehensive literature review, the topical focus was kept relatively narrow on “cloud computing”. Moreover, this paper makes a brief literature review of biocloud from 2006 to 2013 in order to explore how biocloud have developed recently. White paper, practical reporter and many articles related with biocloud in the selected databases were not reviewed. Period and searching limitation may not satisfy all the need of readers looking for a review on biocloud. Finally, the correlation between the framework and real projects of biocloud are not illustrated.

Acknowledgement

The review was supported by Macau Foundation, Faculty Research Grant of Macau University of Science and Technology and Research Grant of Macau Polytechnic Institute (RP/OTHER-01/2014). This review learned the idea from J. Hong *et al.*'s research on context-aware systems [10].

Reference

- [1] Barga, R., Gannon, D., & Reed, D. (2012). The client and the cloud democratizing research computing. *IEEE Internet Computing*, 15(1), 72-75.
- [2] Grossman, R. L., & White, K. P. (2012). A vision for a biomedical cloud. *Journal of Internal Medicine*, 271(2), 122-130.
- [3] Rosenthal, A., *et al.* (2010). Cloud computing: A new business paradigm for biomedical information sharing. *Journal of Biomedical Informatics*, 43(2), 342-353.
- [4] Hamilton, J. P., & Robin, B. C. (2012). Advances in plant genome sequencing. *The Plant Journal*, 70(1), 177-190.
- [5] Titmus, M. A., *et al.* (2012). Answering the demands of digital genomics. *Concurrency and Computation: Practice and Experience*, 26(4), 917-928.
- [6] Kubick, W. R. (2011). Are we ready to fly into the cloud? *Applied Clinical Trials*, 20(2), 28-30.
- [7] Dudley, J., *et al.* (2010). Translational bioinformatics in the cloud: An affordable alternative. *Genome Medicine*, 2(8), 51.
- [8] Barga, R., *et al.* (2011). Bioinformatics and data-intensive scientific discovery in the beginning of the 21st century. *A Journal of Integrative Biology*, 15(4), 199.
- [9] Stein, L. (2010). The case for cloud computing in genome informatics. *Genome Biology*, 11(5), 207.
- [10] Hong, J., *et al.* (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications*, 36, 8509-8522.
- [11] Rimal, B. P., & Choi, E. (2012). A service-oriented taxonomical spectrum, cloudy challenges and opportunities of cloud computing. *International Journal of Communication Systems*, 25(6), 796-819.
- [12] Wang, L., *et al.* (2010). Cloud computing: A perspective study. *New Generation Computing*, 28(2), 137-146.
- [13] Ekanayake, J., *et al.* (2011). Cloud technologies for bioinformatics applications. *IEEE Transactions on Parallel and Distributed Systems*, 22(6), 998-1011.
- [14] Cheng, K. C., *et al.* (2012). Aquatic models, genomics and chemical risk management. *Comparative Biochemistry and Physiology Part C: Toxicol Pharmacol*, 155(1), 169-73.

- [15] Glenn, T. C. (2011). Field guide to next-generation DNA sequences. *Molecular Ecology Resources*, 11(5), 759-769.
- [16] Morales, S. E., & Holben, W. E. (2011). Linking bacterial identities and ecosystem processes: can 'omic' analyses be more than the sum of their parts. *FEMS Microbiology Ecology*, 75(1), 2-16.
- [17] Desai, A. N., & Jere, A. (2012). Next-generation sequencing: ready for the clinics. *Clinical Genetics*, 81(6), 503-510.
- [18] Ertl, P. (2010). Molecular structure input on the web. *Journal of Cheminformatics*, 2(1), 1.
- [19] Roy, N. C. N. C., et al. (2011). A comparison of analog and next-generation transcriptomic tools for mammalian studies. *Briefings in functional genomics*, 10(3), 135-150.
- [20] Zhang, L., et al. (2012). Gene set analysis in the cloud. *Bioinformatics (Oxford, England)*, 28(2), 294-295.
- [21] Kudtarkar, P., et al. (2010). Cost-effective cloud computing: A case study using the comparative genomics tool, roundup. *Evolutionary Bioinformatics*, 6, 197-203.
- [22] Casci, T. (2010). Bioinformatics: SNPs while you wait. *Nature Reviews Genetics*, 11(1), 4.
- [23] Milanese, L., et al. (2009). Trends in modeling biomedical complex systems. *BMC Bioinformatics*, 10.
- [24] Gunarathne, T., et al. (2011). Cloud computing paradigms for pleasingly parallel biomedical applications. *Concurrency and Computation: Practice and Experience*, 23(17), 2338-2354.
- [25] Dove, E. S., et al. (2012). Harnessing omics sciences, population databases, and open innovation models for theranostics-guided drug discovery and development. *Drug Development Research*, 73(7), 439-446.
- [26] Hogan, J. M., et al. (2011). Biomashups: the new world of exploratory bioinformatics. *Concurrency and Computation: Practice and Experience*, 23(11), 1169-1178.
- [27] Fusaro, V. A. V. A., et al. (2011). Biomedical cloud computing with Amazon web services. *PLoS Computational Biology*, 7(8), 1-e1002147.
- [28] Krestin, G. P., et al. (2012). Integrated diagnostics. *Proceedings of the 9th Biennial Symposium of the International Society for Strategic Studies in Radiology. European Radiology* (pp. 2283-2294).
- [29] Wilcox, K., & Vance, T. C. (2011). Ride the wave. *GEO World*, 24(9), 18-21.
- [30] Gonzalez, A., & Knight, R. (2012). Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology*, 23(1), 64-71.
- [31] Lee, H., & Schatz, M. C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics (Oxford, England)*, 28(16), 2097-2105.
- [32] Dalman, T., et al. Cloud Mapreduce for Monte Carlo bootstrap applied to metabolic flux analysis. *Future Generation Computer Systems*.
- [33] Almeida, J., et al. (2012). Fractal mapreduce decomposition of sequence alignment. *Algorithms for Molecular Biology*, 7(1), 12.
- [34] Rahman, M., et al. (2011). A taxonomy and survey on autonomic management of applications in grid computing environments. *Concurrency and Computation: Practice and Experience*, 23(16), 1990-2019.
- [35] Chouvarine, P., et al. (2012). Transcriptome-based differentiation of closely-related miscanthus lines. *PLoS ONE*, 7(1).
- [36] Sangermano, F., et al. (2010). Similarity weighted instance-based learning for the generation of transition potentials in land use change modeling. *Transactions in GIS*, 14(5), 569-580.
- [37] Wilke, A., et al. (2011). An experience report: porting the MG-RAST rapid metagenomics analysis pipeline to the cloud. *Concurrency and Computation: Practice and Experience*, 23(17), 2250-2257.
- [38] Pérez, E. M., & Ferretti, L. (2010). Massive parallel sequencing in animal genetics: Wherefroms and wheretos. *Animal Genetics*, 41(6), 561-569.
- [39] Bullock, A. W., et al. (2010). Dockomatic - automated ligand creation and docking. *BMC Research Notes*, 3, 289-296.

- [40] Evani, U., *et al.* (2012). Atlas2 Cloud: a framework for personal genome analysis in the cloud. *BMC Genomics*, 13.
- [41] Schuster, E. W., *et al.* (2011). Machine-to-machine communication for agricultural systems: An XML-based auxiliary language to enhance semantic interoperability. *Computers and Electronics in Agriculture*, 78(2), 150-161.
- [42] Nguyen, T., *et al.* (2011). CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. *BMC Research Notes*, 4(1), 171.
- [43] Angiuoli, S., *et al.* (2011). CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*, 12(1), 356.
- [44] Lanc, *et al.* (2012). Adapting bioinformatics applications for heterogeneous systems: A case study. *Concurrency and Computation: Practice and Experience*.
- [45] Feng, X., *et al.* (2011). PeakRanger: A cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, 12(1), 139.
- [46] Reese, M., *et al.* (2010). A standard variation file format for human genome sequences. *Genome Biology*, 11(8), R88.
- [47] Fedak, G., *et al.* (2009). BitDew: A data management and distribution service with multi-protocol file transfer and metadata abstraction. *Journal of Network and Computer Applications*, 32(5), 961-975.
- [48] Moretti, C., *et al.* (2010). All-Pairs: An abstraction for data-intensive computing on campus grids. *IEEE Transactions on Parallel and Distributed Systems*, 21(1), 33-46.
- [49] Langmead, B., *et al.* (2009). Searching for SNPs with cloud computing. *Genome Biology*, 10(11), R134.
- [50] Hosny, A. M., *et al.* (2012). Cloud-based parallel solution for estimating statistical significance of megabyte-scale DNA sequences. *Concurrency and Computation: Practice and Experience*.
- [51] Wal, I D., *et al.* (2010). Cloud computing for comparative genomics. *BMC Bioinformatics*, 11(1), 259.
- [52] Luo, D., *et al.* (2012). A cloud computing system to quickly implement new microarray data pre-processing methods. *Open Bioinformatics Journal*, 6, 37-42.
- [53] Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment*, 122, 66-74.
- [54] Fujioka, E., *et al.* (2012). Advancing global marine biogeography research with open-source gis software and cloud computing. *Transactions in GIS*, 16(2), 143-160.
- [55] Ekins, S., *et al.* (2012). Open drug discovery teams: A chemistry mobile app for collaboration. *Molecular Informatics*, 31(8), 585-597.
- [56] Li, K. F. Smart home technology for telemedicine and emergency management. *Journal of Ambient Intelligence and Humanized Computing*.
- [57] Spjuth, *et al.* (2009). Bioclipse 2: A scriptable integration platform for the life sciences. *BMC Bioinformatics*, 10.
- [58] Karp, A. H., *et al.* (2012). Fusion managing healthcare records at cloud scale. *Computer*, 45(11), 42-49.
- [59] Humphrey, M., *et al.* (2010). Publication and consumption of caBIG data services using. *Concurrency and Computation: Practice and Experience*, 22(17), 2313-2322.
- [60] Coutinho, F., *et al.* (2011). Many task computing for orthologous genes identification in protozoan genomes using hydra. *Concurrency and Computation: Practice and Experience*, 23(17), 2326-2337.
- [61] Willighagen, E. L., *et al.* (2011). Computational toxicology using the OpenTox application programming interface and Bioclipse. *BMC Research Notes*, 4.
- [62] Taura, K., *et al.* Design and implementation of GXP make-A workflow system based on make. *Future Generation Computer Systems*.
- [63] Langmead, B., *et al.* (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna.

Genome Biology, 11(8).

- [64] Velikhov, V. E. (2010). Information communication platform for the NBIC center at the Russian Research Center Kurchatov Institute. *Crystallography Reports*, 55(7), 1245-1248.
- [65] Wagener, J., et al. (2009). XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services. *Bioinformatics*, 10.
- [66] Lee, H., et al. (2012). BioVLAB-MMIA: A cloud environment for microRNA and mRNA Integrated Analysis (MMIA) on Amazon EC2. *IEEE Transactions on NanoBioscience*, 11(3), 266-272.
- [67] Moretti, C., et al. (2012). A framework for scalable genome assembly on clusters, clouds, and grids. *IEEE Transactions on Parallel and Distributed Systems*, 23(12), 2189-2197.
- [68] Gannon, D., et al. (2008). TeraGrid science gateways and their impact on science. *Computer*, 41(11), 32-41.
- [69] Abdel, B. M., et al. (2012). Enabling high-performance computing as a service. *Computer*, 45(10), 72-80.
- [70] Hsieh, J. C., & Hsu, M. W. (2012). A cloud computing based 12-lead ECG telemedicine service. *BMC Medical Informatics and Decision Making*, 12(1), 77.
- [71] Ropella, G., & Hunt, C. A. (2010). Cloud computing and validation of expandable in silico livers. *BMC Systems Biology*, 4(1), 168.
- [72] Abouelhoda, M., et al. (2012). Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. *BMC Bioinformatics*, 13(1), 77.
- [73] Hunter, A., et al. (2012). Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine*, 7(1), 1.
- [74] Angiuoli, S. V., et al. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS ONE*, 6(10), 1-10.
- [75] Krampis, K., et al. (2012). Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*, 13, 42-49.
- [76] Simão, J., et al. (2012). A checkpointing-enabled and resource-aware Java Virtual Machine for efficient and robust e-Science applications in grid environments. *Concurrency and Computation: Practice and Experience*, 24(13), 1421-1442.
- [77] Kertesz, A., et al. An interoperable and self-adaptive approach for SLA-based service virtualization in heterogeneous cloud environments. *Future Generation Computer Systems*.
- [78] Maurer, M., et al. (2013). Adaptive resource configuration for Cloud infrastructure management. *Future Generation Computer Systems*, 29(2), 472-487.
- [79] Greenbaum, D., et al. (2011). Genomics and Privacy: Implications of the new reality of closed data for the field. *PLoS Computational Biology*, 7(12), 1-6.



Hongfeng Zhang is an associate professor of Macao Polytechnic Institute. Before coming to Macau he was a pedagogic teacher, researcher of Jiangsu Normal University.

He has been awarded “Excellent Doctoral Dissertation of Higher Education Study” by Chinese Association of Higher Education in 2011, etc. Now he continues to explore the dimensions of General Education, Creativity in Higher Education and Higher Education Change from the perspective of Game theory.



Vincent Y. Liu is an assistant professor at Macau University of Science and Technology. Before he joined in MUST, he had done advanced studies at Carleton University (Canada) as a visiting research fellow from 2008 to 2010. He got his Ph.D. degree from Beihang University in 2010, in China. He has consulted for the Gezhou Dam Water Conservancy and Hydropower Project Group in China on projects related to the digital divide, ICT capacity building and the development of Business Intelligence System. In 2006, he served as a program manager for the National Development and Reform Commission (NDRC) China-sponsored “The Digital Canal for the Middle Route of South-to-North Water Transfer Project”. He also assisted in electronic authentication and credit systems development project (NDRC China sponsored) in 2007.



Zhao Yu got his master degree in management studies from Macau University of Science and Technology in 2013. He focuses on online marketing research.