Cloud-Based Data Workflow System for Food-Derived Peptide Sequences

Zhelyazko Terziyski [1], **Margarita Terziyska** [2]

[1] Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria
e-mail: terziyski@engineer.com

[2] Department of Mathematics, Physics and Information technologies, University of Food Technologies
Plovdiv, Bulgaria, e-mail: mterziyska@uft-plovdiv.bg

Abstract. **Cloud computing has established itself as an advanced and successful technology that provides services such as hardware, software, infrastructure, platform. With the help of the cloud computing paradigm, tools can be developed that can be provided as services to anyone, anywhere and through any device. On the other hand, working in a cloud environment can be facilitated by workflow management systems that support the management of resources, users, and so on. This is the reason why the field of application of cloud technologies is currently significantly expanded. This paper give basic concepts of cloud technologies, the cloud services and show how they could be useful in the field of bioinformatics. In addition, a brief overview of existing scientific cloud workflow management systems is provided. The functional structure of a cloud system for analysis of food derived peptide sequences also is presented.**

Keywords: ***cloud technologies, cloud, workflow, bioinformatics, peptides sequences***

I. Introduction

Bioinformatics is a fast-growing scientific discipline that covers the acquisition, processing, storage, dissemination, analysis and interpretation of biological and biochemical data. The scope of bioinformatics includes the so-called 'Omix' technologies [1], which aim to study certain characteristics of a large family of cellular molecules, such as genes, proteins or small metabolites. The rapid development of omix technologies is helping to integrate bioinformatics into all life sciences, including the food industry. Bioinformatics can be used to effectively access all data on genomics, proteomics and metabonomics found so far, and to provide this data to each individual company in the food industry so that it can improve the quality, taste and nutritional value of food, which will produce. Modern trends in nutrition require food to supply the body not only the necessary nutrients, but also additional substances that have a beneficial effect on health, the so-called biologically active substances (BAB). Along with the already well-known BAV, such as antioxidants, vitamins, essential fatty acids and others, the focus is increasingly shifting to proteins as carriers of health benefits. In addition to their nutritional value, proteins are also a source of biologically active peptides. They could be considered as a source of such specific substances.

Peptides are substances whose molecules are made up of a chain of amino acids linked by peptide bonds. Typically, the peptide sequence consists of 2 to 50 amino acids. Peptides are extremely important for human health. They take an active part in the growth of the body, support the work of the immune system, are natural analgesics, oppose the natural aging process, stimulate and increase the capacity of cells, regenerate tissues, etc. In other words, biologically active peptides are semi-local regulators of the activity of individual cells, tissues, organs and the organism as a whole. Peptides are known which have their own specialization, i.e. have a certain type of activity. Some lower blood pressure, others cholesterol, others have antibacterial or antioxidant properties, etc. Some of them have more than one beneficial effect on health are called multifunctional peptides. Biologically active peptides can be obtained from plant or animal sources or can be artificially synthesized. In view of all this, the object of the present project is proteomics, in particular peptides derived from dietary proteins.

As part of bioinformatics, proteomics generates large amounts of data. This is on the one hand very good for the development of science, but on the other hand it is a challenge for researchers. They need to develop and use big data methods and algorithms, as well as find suitable machines that can store them and have enough computing power to process them [2]. An appropriate solution to this problem is the use of cloud computing. At the same time, the analysis of such big data is a complex process, the modeling of which can be facilitated with the help of scientific workflow management systems. Such systems are able to interpret process descriptions, interact with participants in the flow of activities and, if necessary, call the relevant software applications and tool environments. Therefore, the purpose of this paper is to present the basic concepts of cloud computing and to indicate how they could be useful for bioinformatics researchers as a tool for big data analysis. In addition, a brief overview of scientific workflow management systems will be provided. The functional structure of a cloud system for peptide sequence analysis will also be presented.

II. Cloud computing

Cloud computing is a distributed computing resource provided to a user via a remote computer. The connection to these resources is usually via the Internet, but it is possible to use another communication line. Cloud computing includes not only software but also hardware - servers, data warehouses and more [3]. In fact, the combination of access to hardware and software is what is commonly called a cloud. The cloud is considered a metaphor for the Internet because it is usually depicted in computer network diagrams - an abstraction of the complex infrastructure that builds it.

There are four types of cloud models (Fig. 1) - private, public, hybrid and community. In the case of the *private* cloud, the cloud infrastructure is owned and / or leased by one organization and is used only by it, and it is even possible to hire its own IT staff to manage it. This makes the service more expensive, but also makes it very reliable and efficient. In the *public* cloud, all services are available only via the Internet for free or for a fee. Used by a huge number of users, it has a huge capacity, is easy to use and cheap, but with a low level of security. The *hybrid* cloud is a combination of the previous two. It is a modified private cloud that is used by several trusted organizations. In this way, the costs of maintaining and developing the cloud are shared. Very often, large corporations transfer much of their resources to a hybrid cloud, but keep the most confidential information in their small private cloud.

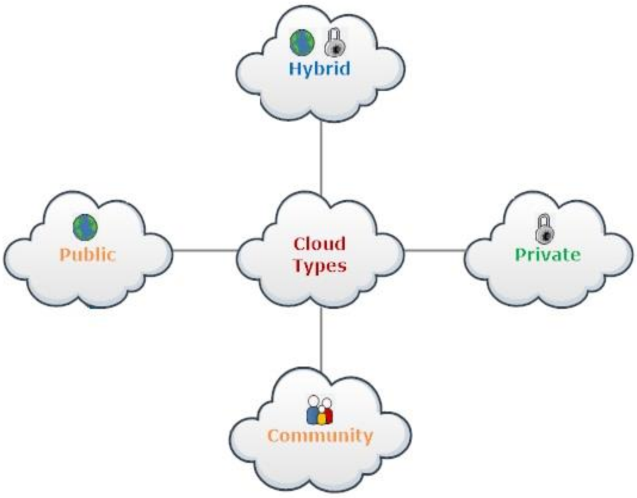

Fig. 1. Types of cloud models.

The *community* cloud is typically used by several organizations with similar interests. There are easy conditions for use, various applications are available for publishing and sharing information.

The main categories of services provided by cloud technologies (Fig. 2):

- ***Infrastructure as a Service (IaaS)***. In this service the object of rent are hardware, operating systems and system software, specialized software. At the heart of the IaaS model is virtualization, a software technology that divides a physical server into virtual resources called virtual machines. With this technology, users can emulate different environments and run multiple applications on a single physical resource. IaaS frees businesses from maintaining complex data centers and network infrastructure by reducing running and capital costs for IT infrastructure.

- ***Platform as a Service (PaaS)***. In this service, cloud tenants use both infrastructure and software applications hosted in the cloud to create web applications. In this case, it is not necessary for developers to have equipment and development environments, nor to organize their maintenance.

- ***Software as a service (SaaS)***. With this service, cloud tenants pay for the use of a software application hosted in the cloud. The main advantage of the SaaS model for the customer is the lack of costs for the installation, update and maintenance of the hardware and software running on it. The target audience here is the end users of software, and access is via an Internet browser.

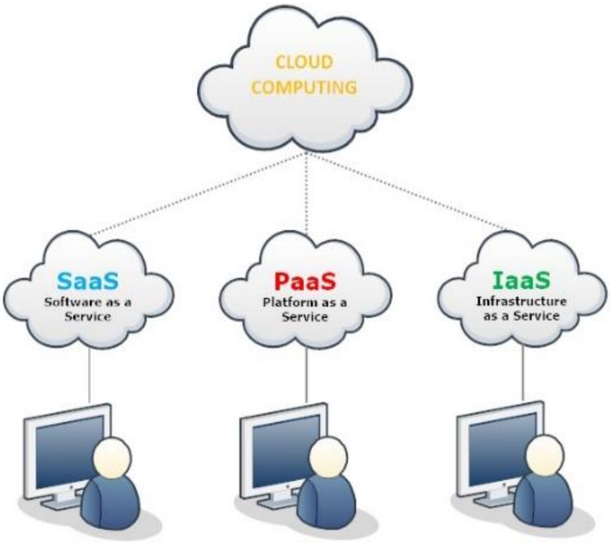

Fig. 2. The main types of cloud computing services.

In recent years, cloud technologies have been developing at a rapid pace. With the development of cloud technologies, there is a constant expansion of the capabilities of cloud services. A number of new concepts are emerging, defining services as a desktop as a service (Desktop as a Service, DaaS), a business process as a service (BPaaS) and even everything as a service (Everything as a Service, XaaS). XaaS covers any computer service (IT function) that is delivered over the Internet and paid for according to a flexible consumption model. BPaaS, on the other hand, solves typical business tasks in a cloud environment. The need for the introduction of this service arises when a user company decides to automate the same type and repetitive tasks. The same service can be used by several clients of the BPaaS provider. Payment is made at a fixed price for a period or according to the consumption of the service. Another relatively new cloud service is ***Workflow as a service (WfaaS)***. It will be discussed in more detail in the next section of the article.

III. WORKFLOW MANAGEMENT SYSTEMS

The term workflow is defined as a sequence of tasks performed in the processing of a set of data. Workflows are used in every type of business, industry and research. They usually transfer data between people and / or systems. Workflows are divided into workflow process and case workflow. When the set of tasks is predictable and repetitive, a workflow process is used. This means that before an item begins the workflow, you know exactly what path it should take. In the case workflow, the path required to complete the item in the beginning is unknown. The path is reveals itself as more data is gathered.

In order to automate workflows, speed up execution, visualization and monitoring of tasks related to the processing of large volumes of data, it is appropriate to use ***workflow management systems (WfMS)***. Scientists often use a programming language to describe a workflow, to create a model that determines the sequence of tasks. The refinement of complex scientific experiments with the help of work processes is becoming more and more common. As a result, a number of software tools have emerged, called ***Scientific Workflow Management Systems (SWfMS)***, that allow the specification, registration, execution, visualization, and monitoring of scientific workflows.

In fact, WfMS in general is directly related to Workflow as a Service (WFaaS). Using cloud-based workflow solutions, organizations (business and/or research) can automate and improve the efficiency of their workflows at a significantly lower cost. Cloud-based workflow platforms are already designed and hosted in the cloud by service providers. Cloud workflow technology can be used on two levels. From the point of view of cloud users, it supports definitions of cloud application processes and allows flexible configuration and automated process operation. This type of cloud workflow is considered "above the cloud". From the point of view of cloud service providers, cloud workflow offers automatic task scheduling and cloud computing resource management. This category is called "in the cloud".

Bioinformatics workflow management systems are a special form of workflow management systems designed specifically to compile and perform a series of calculations, data processing or workflow steps that are related to bioinformatics [4]. They offer many benefits to bioinformatics. SWfMS provides a common language for describing analysis workflows, contributing to reproducibility and building libraries of reusable components. They can support both incremental build and re-entry - the ability to selectively re-execute parts of a workflow with additional inputs or configuration changes and to resume execution from where the workflow stopped before. Many workflow management systems improve portability by supporting the use of containers, high-performance computing systems, and the cloud. Most importantly, workflow management systems allow bioinformatics to delegate the way their workflows run to the workflow management system and its developers. This allows scientists to focus on what these work processes need to do, on their data analysis, and on their science.

Choosing a scientific workflow management system can be a daunting task. SWfMS offers two main approaches to workflow modeling - through text and through graphical programming languages. It is generally accepted that the graphical language allows easier understanding of the program. With this type of system, the user can create and modify applications with little or no programming knowledge. However, when the workflow is more complex, then the model becomes more difficult to understand and even confusing. Therefore, many scholars prefer text-based programming. The latter is more suitable for rapid prototyping and provides a more compact presentation of complex workflows. On the other hand, writing code requires scientists to be able to program in a specific language, most often Python. Therefore, it is necessary to carefully consider which type of SWfMS to choose. Specifically in the field of bioinformatics, the most popular scientific workflow management systems are Taverna and Galaxy. Currently, the Galaxy platform is installed on more than 170 servers around the world, while the social network for scientists myExperiment of Galaxy shares almost 4,000 scientific workflows among its 11,162 members.

Taverna [5] is open-source software that integrates many different software components, including those provided by the National Center for Biotechnology Information, the European Institute of Bioinformatics, the DNA Data Bank of Japan (DDBJ), SoapLab, BioMOBY and EMBOSS. Taverna Workbench provides a desktop development environment and implementation mechanism for scientific workflows. It is available separately as a Java API, command line tool or as a server. It is used in a number of fields such as bioinformatics, chemoinformatics, medicine, astronomy and others. Workflows developed with Taverna can be shared through myExperiment, a social network for scientists.

Galaxy [6] is a collection of software packages that can be managed through a web browser on a public server. The graphical user interface it has means that its users do not need to be familiar with programming. Galaxy is open-source software that allows users to edit and improve it. The Galaxy project is based on the Python programming language. Galaxy features include next-generation sequencing tools that allow the user to convert between different file formats with sequence such as text, spreadsheet, SFF, FASTA and FASTQ, filter quality rating data, cut amino acids from sequences, search of specific character strings in data sets, running complete statistical reports on data sets and much more. Galaxy allows the use of saved (ready-made) workflows, which can be performed on different data sets in exactly the same way each time.

Along with Taverna and Galaxy, a number of other workflow management systems are used in bioinformatics. Below are some of them:

Pegasus [7] is an open-source workflow management system. It provides scientists with the abstractions they need to create scientific workflows and allows these workflows to run transparently across a variety of computing platforms, including high-performance computing clusters, clouds, and national cyber infrastructures. In Pegasus, workflows are described abstractly as acyclic control graphs (DAGs) using the provided API for Jupyter Notebooks, Python, R, or Java. During execution, Pegasus translates the constructed abstract workflow into an executable workflow that is executed and managed by HTCondor.

Tavaxy [8] is a system for creating and executing workflows based on the use of an extensible set of ready-made workflow templates that allow reuse. Tavaxy offers a range of new features that simplify and improve the development of sequence analysis applications. Allows the integration of existing Taverna and Galaxy workflows into a single environment and supports the use of cloud computing capabilities. The integration of existing Taverna and Galaxy workflows is seamlessly maintained at both runtime and design time levels, based on hierarchical workflow concepts and workflow models. The use of cloud computing in Tavaxy is flexible, whereby users can either instantiate the entire system in the cloud or delegate the execution of certain sub-streams to the cloud infrastructure. Tavaxy reduces the workflow development cycle by introducing the use of workflow templates. It allows the reuse and integration of existing (sub-) workflows from Taverna and Galaxy and allows the creation of hybrid workflows. Its additional features use the latest advances in high-performance cloud computing to cope with the growing size of data and the complexity of analysis. The system can be accessed either through a cloud web interface or downloaded and installed to run in the user's local environment.

Kepler [9] is a workflow management system developed by a collaboration of universities, including UC Davis, UC Santa Barbara, and UC San Diego, United

States. Kepler has been adopted in various scientific projects including the fluid dynamics and computational biology. This SWfMS provides compatibility to run on different platforms, including Windows, OSX, and Unix systems.

HyperFlow [10] is a computational model, programming approach, and also a workflow engine for scientific workflows. It provides a simple declarative description based on JavaScript. HyperFlow supports the workflow deployment in container-based infrastructures such as docker and Kubernetes clusters. HyperFlow is also able to utilize the serverless architecture for deploying Montage workflow in AWS Lambda and Google Function.

IV. Functional structure of a cloud-based system for peptide sequence analysis

Peptides are organic compounds whose molecules consist of two or more amino acids linked by a peptide bond. They are extremely important for the human body and its health. They take an active part in the growth of the body, maintain the good functioning of the immune system. Peptides can be used successfully in therapies against various diseases, but they also have a number of disadvantages - low systemic stability, poor membrane permeability, low solubility, rapid clearance, poor oral bioavailability and high production costs. Identification of suitable peptides by traditional experimental approaches involves screening peptide libraries or examining whole proteins using overlapping windows in the areas of the peptide chains, and then evaluating each part of the activity. This process is time consuming, expensive and time consuming and depends on various factors. Overcoming these limitations requires proteomics to make full use of modern information and communication technologies. On the one hand, new and efficient computational approaches should be sought to be used to detect candidate peptides. On the other hand, it is imperative to find a way to store and process the vast amount of data generated by proteomics. Therefore, the authors aim to develop a cloud-based system for the analysis of peptide sequences derived from food. As the first stage of this project, the functional structure of this system is presented in Fig. 3. Each customer accesses it via a smart device (computer, tablet, telephone) connected to the Internet (1). The peptide sequences to be analyzed can be entered directly by the users of the system or can be extracted from private and/or public data warehouses (5-6). Some of the public databases with food-derived peptides are summarized in Table. 1.

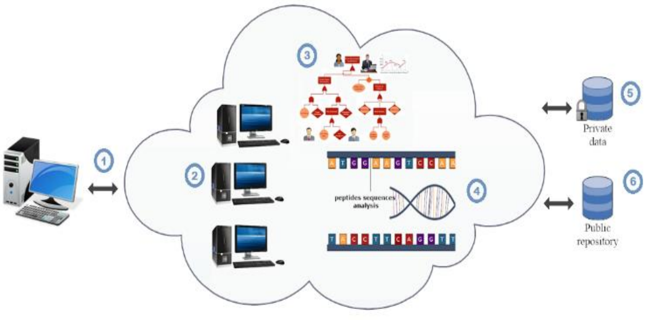

Fig. 3. The main types of cloud computing services.

The functional system must include certain computing resources (2) necessary for the analysis, which is carried out with the help of workflow systems (3) and specific software applications (4). It is envisaged that after pre-processing the data, they can be directed to a module for feature extracting and encoding of peptide sequences, or to a module for predicting their biological activity using machine learning methods.

TABLE I. CURRENTLY AVAILABLE DATABASES OF FOOD-DERIVED BIOACTIVE PEPTIDES

Database	[a]Website
FeptideDB [11] Database of food-derived bioactive peptides	http://www4g.biotec.or.th/FeptideDB/
BioPepDB [12] Database of bioactive peptides of food origin	http://bis.zju.edu.cn/biopepdbr/
AHTPDB [13] Database of food-derived AHTPs	http://crdd.osdd.net/raghava/ahtpdb/info2.php
BIOPEP-UWM DB [14] Database of bioactive peptides of food origin	http://www.uwm.edu.pl/biochemia/index.php/en/biopep
MBPDB [15] Milk Bioactive Peptide database	http://mbpdb.nws.oregonstate.edu

[a] *All websites indicated in the table were accessed in November, 2021.*

V. Conclusions

Many large companies have already moved their business to the cloud. The benefits for them are reduced running costs, faster upgrades and increased security. In this respect, research teams are lagging behind. In this article, an attempt is made to show that cloud technologies and the services they provide also have a place in the scientific community. Building virtual labs will help bridge the divide between poorly funded and well-funded real-life labs, help fill research teams with "remote" staff and interns, and help research continue despite the blockage of life due to another Covid-19 lockdown. In addition, cloud operation allows for potentially easier setup, faster processing, and lower construction and management costs compared to on-premises workstations. The article also demonstrates that a number of cloud applications have already been developed, specifically in bioinformatics. An overview of the more well-known scientific workflow management systems has been reviewed, the main purpose of which is to help automate workflows, speed up execution, visualization and monitoring of tasks related to the processing of large volumes of data. The functional structure of a cloud-based system for the analysis of peptide sequences derived from food is also presented.

Acknowledgment

The authors would like to acknowledge the support of the Research and International Sector at the University of Food Technologies (UFT), Plovdiv, Bulgaria under research project № 08/21-H funded through the UFT's Science Fund.

References

[1] Blazhko, Natalia, Sultan Vyshegurov, and Kirill Shatokhin. "Application of omix technologies in studying of BLV biological diversity by gag gene." International Scientific and Practical Conference "Digitization of Agriculture-Development Strategy"(ISPC 2019), Advances in Intelligent Systems Research. Vol. 167. 2019.

[2] Shakil, Kashish Ara, and Mansaf Alam. "Cloud computing in bioinformatics and big data analytics: Current status and future research." Big Data Analytics. Springer, Singapore, 629-640, 2018.

[3] Srivastava, Priyanshu, and Rizwan Khan. "A review paper on cloud computing." International Journal of Advanced Research in Computer Science and Software Engineering 8.6 (2018): 17-20.

[4] Larsonneur, Elise, et al. Evaluating workflow management systems: a bioinformatics use case." IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018.

[5] Oinn, Tom, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20.17, 3045-3054, 2004.

[6] Jalili, Vahid, et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update." Nucleic acids research 48.W1, W395-W402, 2020.

[7] Deelman, Ewa, et al. "Pegasus, a workflow management system for science automation." Future Generation Computer Systems 46, 17-35, 2015.

[8] Abouelhoda, Mohamed, Shadi Alaa Issa, and Moustafa Ghanem. "Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support." BMC bioinformatics 13.1, 1-19, 2012.

[9] Altintas, Ilkay, et al. "Kepler: an extensible system for design and execution of scientific workflows." Proceedings. 16th IEEE International Conference on Scientific and Statistical Database Management, 2004.

[10] Balis, Bartosz. "HyperFlow: A model of computation, programming approach and enactment engine for complex distributed workflows." Future Generation Computer Systems 55, 147-162, 2016.

[11] Panyayai, T., Ngamphiw, C., Tongsima, S., Mhuantong, W., Limsripraphan, W., Choowongkomon, K., Sawatdichaikul, O.: FeptideDB: A web application for new bioactive peptides from food protein. Heliyon, 5(7), e02076 2019.

[12] Li, Q., Zhang, C., Chen, H., Xue, J., Guo, X., Liang, M., Chen, M.: BioPepDB: An integrated data platform for food-derived bioactive peptides. International Journal of Food Sciences and Nutrition, 69(8), 963-968, 2018.

[13] Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J., Singh, S., Gautam, A., Raghava, G.: AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. Nucleic acids research, 43(D1), D956-D962, 2015.

[14] Minkiewicz, P., Iwaniak, A., Darewicz, M.: BIOPEP-UWM database of bioactive peptides: Current opportunities. International journal of molecular sciences, 20(23), 5978, 2019.

[15] Nielsen, S. D., Beverly, R. L., Qu, Y., & Dallas, D. C.: Milk bioactive peptide database: A comprehensive database of milk protein-derived bioactive peptides and novel visualization. Food Chemistry, 232, 673-682, 2017.