Cluster Comput
DOI 10.1007/s10586-015-0456-6

A network approach for managing and processing big cancer data in clouds

Wei Xing[1] · **Wei Jie**[2] · **Dimitrios Tsoumakos**[3] · **Moustafa Ghanem**[4]

Received: 27 February 2015 / Revised: 8 April 2015 / Accepted: 9 April 2015

Abstract Translational cancer research requires integrative analysis of multiple levels of big cancer data to identify and treat cancer. In order to address the issues that data is decentralised, growing and continually being updated, and the content living or archiving on different information sources partially overlaps creating redundancies as well as contradictions and inconsistencies, we develop a data network model and technology for constructing and managing big cancer data. To support our data network approach for data process and analysis, we employ a semantic content network approach and adopt the CELAR cloud platform. The prototype implementation shows that the CELAR cloud can satisfy the on-demanding needs of various data resources for management and process of big cancer data.

Keywords Big data · Data network · Cloud computing

✉ Wei Xing
wei.xing@cruk.manchester.ac.uk

Wei Jie
wei.jie@uwl.ac.uk

Dimitrios Tsoumakos
dtsouma@cslab.ece.ntua.gr

Moustafa Ghanem
m.ghanem@mdx.ac.uk

[1] Cancer Research UK Manchester Institute, University of Manchester, Manchester M20 4BX, UK

[2] School of Computing and Technology, University of West London, London W5 5RF, UK

[3] Computing Systems Laboratory, National Technical University of Athens, Athens 15773, Greece

[4] Department of Computer Science, University of Middlsex, London NW4 4BT, UK

1 Introduction

Translational cancer research requires to integrate big cancer data, including genomic, proteomic, and clinical information, to identify, prevent and treat cancer [1,2]. This suggests scientists to incorporate multiple levels of biological information within their studies such as phenotype, genotype, expression profiling, proteomics, protein interaction, metabolic analysis and physiological measurements, etc. [3,4].

We develop a new Cancer Data Network (CDN) model and the technology for constructing and managing content in order to support the integration of biological and clinical data with the research it is spawned from. In addition, the CDN offers the ability of track several aspects of patient care according to genetic and molecular profiles to facilitate tailoring of treatment.

In this paper, we propose the CDN architecture to stand as a novel content management model and associated system that supports end users in a distributed, dynamic and evolving information landscape. The CDN architecture shifts the view of content from being a static resource, and introduces it as a dynamic and intelligent entity that is able to perform operations such as linking itself to other relevant content. In doing so it can discover implied relationships with other content, identifying redundancies and overlap as well as updating its links with the ecosystem when new content is added or old content is removed or depreciated.

The CDN approach is thus to enable the content itself as an active object equipped with intelligence and semantic mechanisms that allow a greater degree of flexibility towards automating the procedure of content management and organization. To this end, we define active cancer data content as a logical container that contains the digital data content (i.e., patient data, clinical data, research experiment data,

Published online: 21 April 2015

publications, public gene or protein databases, etc) together with intelligent and autonomic, self-organizing mechanisms for automating content management.

Given that massive data is encoded into the CDN, it requires large amount of data and computing resources to enable the manipulation of the data network. We employ cloud computing platform to support the CDN approach. Particularly the CELAR cloud platform [5,6] is selected as the CDN cloud platform because CELAR delivers a fully automated and highly customisable cloud platform for elastic provisioning of resources.

The remainder of this paper is organised as follows. Section 2 introduces the principles and design of the CDN; Sect. 3 presents the architecture of the CDN, focusing on its software components and the main interactions between them, as well as how the components are instantiated for the implementation with the CELAR cloud platform; Sect. 4 describes the related work; and finally, Sect. 5 concludes the paper, and describes open issues and planned future work.

2 The design of the CDN

The CDN is designed to bridge the gap between translational research and targeted patient treatment. Hence, the design goals of the CDN are firstly to better analyse data obtained dynamically from various bio-instrument sources in order to answer biological question at a system level; and secondly to better translate data obtained from in vitro and in vivo discoveries into the clinic.

2.1 Problems and requirements

The key challenges that the CDN addresses is that information stored or published over the web and other specialized data sources is decentralized, growing and continually being updated. Furthermore, the contents stored or archived on different information sources may partially overlap, thus creating redundancies as well as contradictions and inconsistencies. In this section we describe the current issues in the area of personalized medicine research.

2.1.1 Redundant or irrelevant information of protein and gene sequence

Currently, over a thousand accessible data sources provide information pertaining to any gene, mRNA or protein sequence (estimated by the number of known SRS "Sequence Retrieval System") such as polymorphisms, protein interactions and expression levels. The vast majority of the data sources are specialised, maintained and updated by different organisations. In addition, data sources with the same emphasis (such as nucleotide or protein sequences) are updated and curated at different intervals and with various benchmarks and standards. As a result, many databases contain outdated, redundant or irrelevant information pertaining to the scientific questions at hand.

Also, our continually expanding knowledge base adds new dimensions to the content. For example, the cataloging and assessment of functional impact of recently discovered mechanisms of dynamic biological regulation (including but not restricted to microRNAs and our knowledge of protein modification types and permutations) is incomplete. New categorical discoveries and their related information details need to be progressively built into any comprehensive content structures.

2.1.2 Evolving methods of data generation from multiple instrument platforms

Translational cancer research requires the integration of data from state-of-the-art technologies, for which the methods of translating and interpretation raw instrument data into relevant contextualized biological outputs are continually improving. An example of this is the interpretation of mass spectrometry peptide fragmentation data into qualitative and quantitative peptide and protein data in proteomics experiments. Different instruments produce data with different technical characteristics, including signal-to-noise ratios, raw signal intensities, and data accuracy, precision and resolution. These characteristics are continually changing for the better, but will continue to vary depending on the type and generation of instruments used, new hardware innovations, and the data acquisition and experiment style.

The bioinformatic translation of the raw fragmentation data into peptide and protein identities is also evolving. Current strategies typically employ probabilistic, stochastic or descriptive models to pattern match fragment ion profiles against theoretical profiles generated against assumed protein sequences and modification content. Personalised medicine will dictate a drift away from this data interrogation strategy since each individual labours genomic and proteomic differences that would not be represented in an assumed protein sequence database. This may involve fundamental changes to the data interrogation strategy, for example, a migration towards de novo sequencing tools, or at the very least changes to scoring of genepeptideprotein sequence assignments and the specific identification of mutations, polymorphisms or variables specific to individuals.

2.1.3 Creating genomic networks

To elucidate the wiring of cellular information processes, current research require integration of quantitative and dynamic data from several sources. Such information sources could

be genomic public database based, sequence-based or clinical information and require various algorithms and software package for data analysis. For maximal output from such data, it is important that the multidimensionality is taken into account and the data can be visualised with differential weighting of individual data sources. For example, gene mutation and gene function interactions can be measured in a static manner using techniques such as yeast 2 hybrid and complement assays as well as the dynamic and quantitative abundances can be included through platforms such as COSMIC, VerScan and Meerkat runs. While each source provides important information, the sources provide complementary aspects of information, which is important to integrate and visualize.

To address the above issues, we design the CDN system to support:

1 *Integrating heterogeneous and unstructured content* It allows scientists to incorporate multiple levels of background information within their studies, such as phenotype, genotype, expression profiling, proteomics, protein-protein interactions, biochemical metabolic studies, and physiology measurement, etc.
2 *Decentralized control and collaborative communities* The content itself either arises from biological experiments conducted by individual groups or as a result of data integration and analysis studies using data published by other groups.
3 *Muilti-displine* The information is highly relevant to researchers working on other topics and it can be shared easily between specialized data sources (including scientific literature) and databases focusing on specific topics, e.g. organisms, diseases, genes, proteins, metabolic pathways, chemical compounds or on relationships between them.

Our special focus is addressing the issues of overwhelming and continuous flood of complex information generated and published on a daily basis through the use of semantic web technology. We illustrate our approach in the next section.

2.2 Semantic approach

The CDN aims to develop novel mechanisms for constructing and generating symbiotic, semantically-described cancer data network that enables distributed heterogeneous cancer data to be linked together into data networks for integrative data analysis.

2.2.1 The cancer data networks

A key feature of cancer information is that it is continually evolving. For example, new information about a particular cancer entities (e.g. proteins, genes or diseases) is being published on a daily basis. Furthermore, the decentralized authority over the content, whereby scientists in different organizations publish and manage their own findings, means that information about the same, similar or related entities, may be stored on different sources that evolve in different ways. This inevitably results in partial overlaps in the coverage of the data sources creating redundancies as well as contradictions and inconsistencies at both the entity and the concept level.

We design the CDN to link individual elements of the digital content together. By using semantic data model and ontology, we define two type of links among the CDN nodes (i.e. content): explicit links and conceptual links.

Explicit links between different elements are typically stored with the content. At the simplest level a entry on a specific protein on a particular data source can make explicit references to other protein, gene or disease entries on other sources, or to specific supporting scientific publications. Ontologies can be used to either manually or automatically assign scientific papers, genes, proteins, to different categories.

Conceptual links between different elements are typically not stored with the content, but they can traditionally be inferred by using either statistical/probabilistic analysis techniques or domain knowledge. At the simplest level, users may wish to group proteins together based on the similarity of specific properties such as their effect on the same cellular function, or their causal implication to a similar disease phenotype.

The two types of links can represent all kinds of relationships among cancer contents (e.g., concepts and instances). And they are used to connect various cancer entities into a cancer data network.

2.2.2 Retrieval, integration and update

In [7], we developed a semantic information integration approach to integrate and update information from distributed, heterogeneous data sources dynamically. The CDN employs ActOn [7,8] as a means to retrieve, integrate, and manage the CDN Content in a intelligent and active manner.

The ActOn is an ontology-based information integration approach that is suitable for highly dynamic distributed resources. To deal with the issue that information changes frequently and information requests have to be answered quickly in order to provide up-to-date information, the ActOn employs an information cache that works with an update-on-demand policy. Due to the multitude of databases and information sources, the most appropriate sources have to be selected for each query to ensure optimal and relevant data retrieval. To deal with this issue that the most suitable information sources have to be selected from a set of different

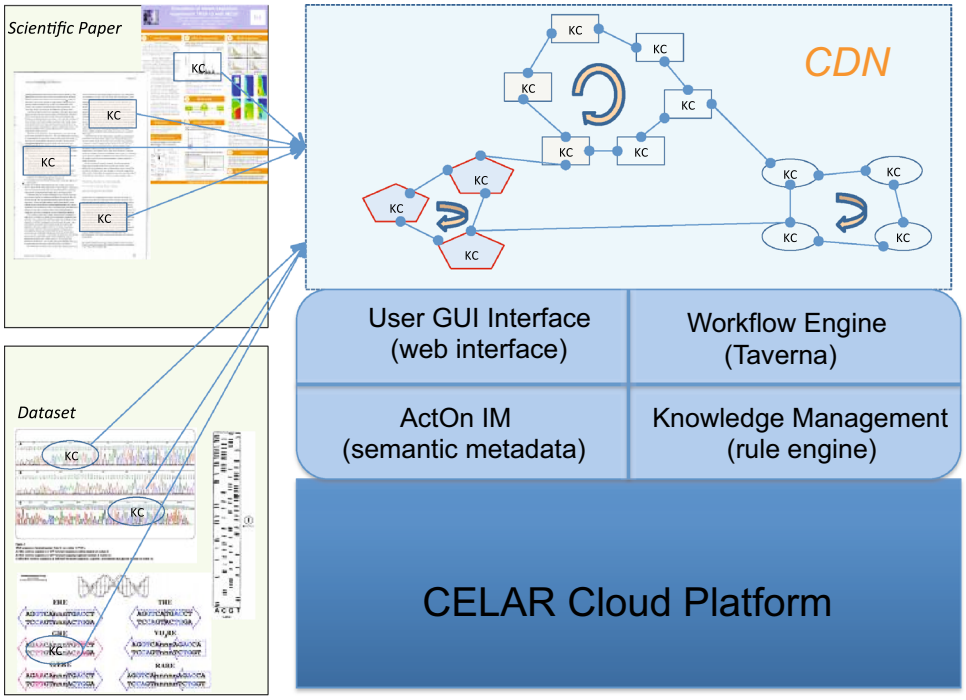

Fig. 1 Overview of the CDN architecture

distributed information sources that can provide the information needed, the ActOn adds an information source selection step to the ontology-based information integration. Thereby, the most suitable information sourcedatabase will be selected for a user query.

2.3 CDN architecture

Figure 1 shows three-tier view of the CDN architecture. At the core of the middleware lies the CDN ActOn Information Manager that represents the cancer data content and its associated information extraction tools. The CDN ActOn contains the semantic metadata and knowledge management tools that enable modelling and analyzing its life cycle and support reasoning about the content. The CDN also contains the workflow tools (Workflow engine) that enable the statistical analysis of the content enabling it to self-organise when linking with other contents. Finally, the CDN also includes semantic-aware and peer-to-peer based networking functionality that enables the content to discover other contents and communicate with them.

2.3.1 System components

We use a bottom-up description of the components shown in Fig. 1.

The CDN semantic model The bottom layer represents existing and traditional data sources that will be used within CDN. Digital content elements on the sources will be identified and extracted and represented as Knowledge Cells (KC) that represent the core of an data content object and represent nodes in abstract CDN. Semantic reasoners can be employed to infer logical consequences from a set of asserted facts of those KCs, so that KCs are able to self-manage and self-organise. Data sources can be accessed through the middleware and offered to the application platform.

The ActOn information manager The CDN middleware employs the ActOn, a semantic information integration system, to connect the data sources to the CDN system. The ActOn Information Manager can deploy the Data content and place it inside the CDN which contains extra information about the content that makes it both self-aware and context-aware together. The ActOn Information Manager can also link the data content in multiple data content networks. During system operation, the links (the edges in the network graph shown in Fig. 1) between data content entities (the nodes in the network graph in Fig. 1 can be re-organized based on statistical analysis, user preferences or other types of runtime information.

Workflow engine The Workflow engine enables data document access over preprocessing, tokenization, parsing, named entity recognition to the final consumer. It implements specialized workflows that support the different types of users of the system (publishers, curators and end users) in combining data retrieval, integration, semantic annotation and deployment of data content within data analysis tasks in end user applications. The starting point for implementing the workflow engine is to employ Tarverna workflow system (authoring tools and execution engines) for the integration and analysis of a wide variety cancer data (including genomic, proteomic data sets, as well as free text publications).

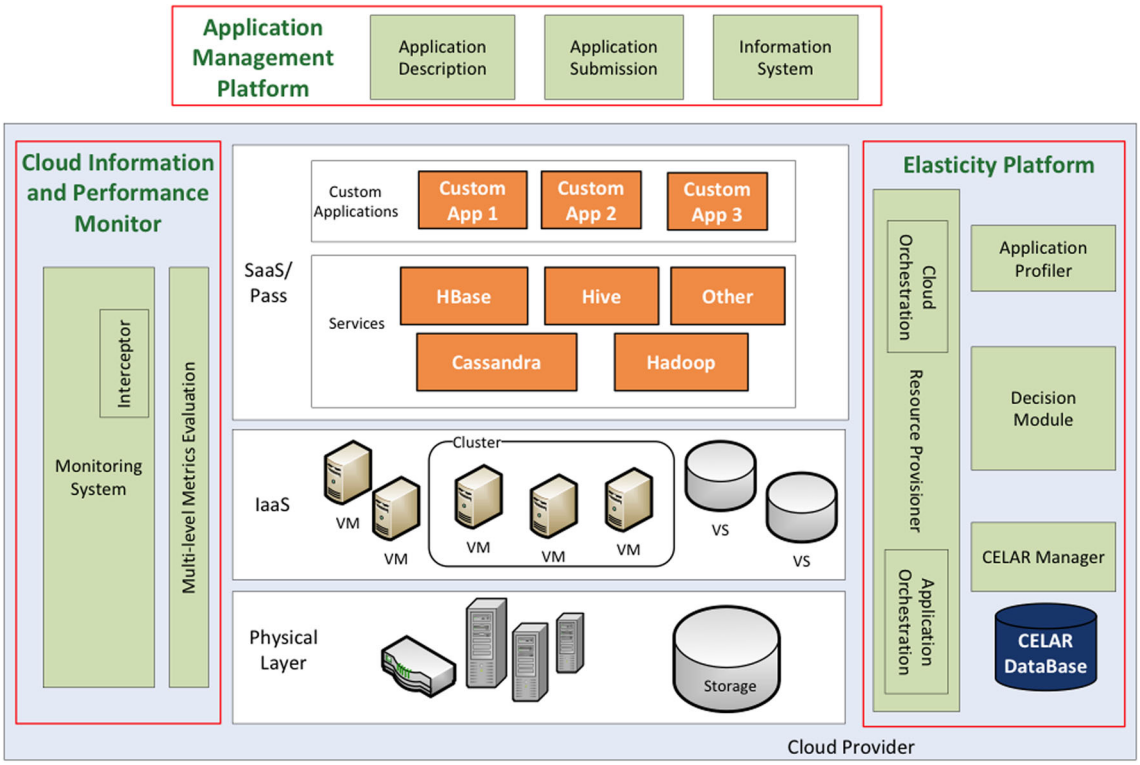

Fig. 2 The CELAR cloud platform

Semantic web user interface At the top-level, a user interface includes functions that can be used to connect to the CDN middleware so that contents can be retrieved from the distributed data sources (shown as different databases in the bottom of Fig. 1) to create the relevant data content. When a data content is created, it is passed to the CDN middleware and then added in the CDNs. The user interface can also interact with the user, issue user-queries and get back results which the CDN present to the users in an advanced way, showing the KCs inside the retrieved content. As such, the user can use the KC to issue/refine further queries or even browse based on a given KC in order to find similar KCs and iteratively refine his queries within the CDN to get satisfactory results.

3 A cloud platform for CDN

In order to assess whether the genome can tell us more about the undue burden, the CDN needs to manage and process large amount of genomic and proteomic data to identify the driven mutation of tumor samples, and then to associate the identified mutation with protein functions within a cell signal network. Given 3 billion DNA base pairs of human genome and 25,000 human protein-code genes, the CDN actually requires massive, various data and computing resources as well as associated software environments. This implies that the elasticity of cloud computing [9–11] can play a key role for the CDN approach. Taking advantages of cloud computing, the CDN can be supported in a way that continually involved massive data will always get enough data resources dynamically and seamlessly for its needs.

3.1 The selection of CELAR cloud platform

The EU CELAR platform is a fully automated and highly customisable system for elastic provisioning of resources in cloud computing platforms (Fig. 2). The CELAR aims at providing an elasticity layer for applications that need to take advantage of the elastic, pay-as-you-go resource provisioning nature of cloud infrastructures in a transparent and customizable manner. Therefore it is a suitable cloud platform for managing and manipulating big omic data of the CDN. More precisely, the CELAR is able to allocate the data and computing resources to the CDN according to the size of its genome data and the data processes needed. In this section, we introduce the CELAR platform and describe how the CELAR platform manages data elasticity in the CDN level to analyze large scale cancer data efficiently and economically.

We design the CDN as a data network module that can run on top of the CELAR cloud platform. In particular, we allow the CDN can support computational and data elasticity so that the CELAR can intelligently orchestrate and adjust the

computing resource allocation according to needs of cancer diagnose and the nature of cancer data of individual patients.

3.2 CELAR middleware for CDN

The CELAR enhances the functionality provided by current cloud infrastructures and provide automated, multi-grained, elastic resource provisioning for cloud-based applications, such as the CDN. In this section, we explain how the CDN can co-operate with the CELAR middleware components.

As shown in Fig. 2, the application management modules will be developed and provided under the c-Eclipse framework and exposed via meaningful, user-friendly UIs to the end-users and application experts. The CDN will interact with the application management modules to control the CDN data network accordingly. The modules enable intelligent, application- and user-aware description and the deployment of the CDN. It can also monitor the changes of the CDN, exposing an overview of the current and past status of the CDN processes as well as the available resources (software and hardware) from the underlying Infrastructure as a Service (IaaS).

The CDN Data can be stored in plain files and accessed through data wrappers or via database systems that range from typical relational databases to NoSQL stores. The provisioning layer consists of well-known database systems that can be described and profiled by the CELAR platform. Some of these systems, such as the distributed NoSQL stores, exhibit horizontal elastic behavior that can be exploited by the CELAR platform; others, such as centralized RDBMS, exhibit vertical resizing functionality based on the resources dedicated on a single virtual machine.

Similar to the storage resource, the CDN uses the CELAR Provisioner to prepare largest amount of computing resources used for the CDN and needs to be elastically scaled. To do so, the CDN will provide application-lever information that can be used to predict the computing resources required in the CDN operation. Apart from dynamic provisioning of computing resources, the CELAR can also be used to provide online resizing of the resources allocated to the CDN when the involved data is removed or added.

3.3 CDN data processed by CELAR SCAN

SCAN [12] as a CELAR application platform can be applied to support data analyses on top of the CDN. The key objective of SCAN application platform is to match the resource demand required by a variety of bio-applications or by different volume of cancer data. SCAN is comprised of a number of genomic and/or proteomic applications, which may incorporate multiple levels of biological information of a CDN within studies such as phenotype, genotype, expression profiling, proteomics, protein interaction, metabolic analysis and physiological measurements, etc.

The SCAN processes CDN data with cloud resources. With different requests and stages of process, the SCAN can talk to the CELAR middleware to obtain substantially different levels and types of resource ideally suited. For example, mapping of deep sequencing data to genome annotation via a relational database such as ENS-EMBL [13] relies on the ability to perform frequent joins across multiple tables containing millions of rows, while computation of downstream statistics is often dependent on repeated numerical calculations over permuted data in order to provide a null distribution. Also SCAN may help CDN to have different resource needs due to the size and complexity of the data of CDN. For example, SCAN mutation detection process can take 4 CPU/hours for Whole Exome Sequencing data in the CDN network or 10 CPU/hours for whole genome sequencing (WGS) data of CDN network. In general, SCAN can help to process more than thirty kinds of genome data of CDN, which can be used for cancer research, such as Whole Exome Sequencing data, Whole Genome Sequencing data, total RNA data, miRNA sequence data etc.

3.4 Implementation

We implement the CDN system using Java Spring Framework and RDF Jena API. Spring is a software toolkit that can be used to program web-based application and data management system. Jena is a Java API that can be used to create and manipulate RDF models. Using Spring framework, we are enable to code the system in Java following the WSRF specification. We use Jena OWL toolkit for creating, manipulating and querying the semantic metadata of data content.

In view of the large volume of the cancer data, we use the CELAR cloud platform, a elastic cloud computing platform, to process and build the CDN system. The EU CELAR cloud can deliver a fully automated and highly customisable system for elastic provisioning of resources within cloud computing infrastructures. It therefore can provide large scale computation resources required by the CDN. In addition the CELAR platform can also provision particular types of computing resources required by the CDN dynamically, such as windows system with large memory or large amount CPU resources of linux systems, etc. Currently our prototype implementation is mainly for creating and managing gene mutation data and the next generation sequencing variation detection data. We have processed about 10TB whole genome sequence data and linked them with Uniprot protein database to generate the sample-protein-gene network in Fig. 3. The initial results shows that data within CDN can be linked based on domain ontology. As shown in Fig. 3, the big cancer data processing can be executed efficiently without delay since relevant data are already be retrieved into

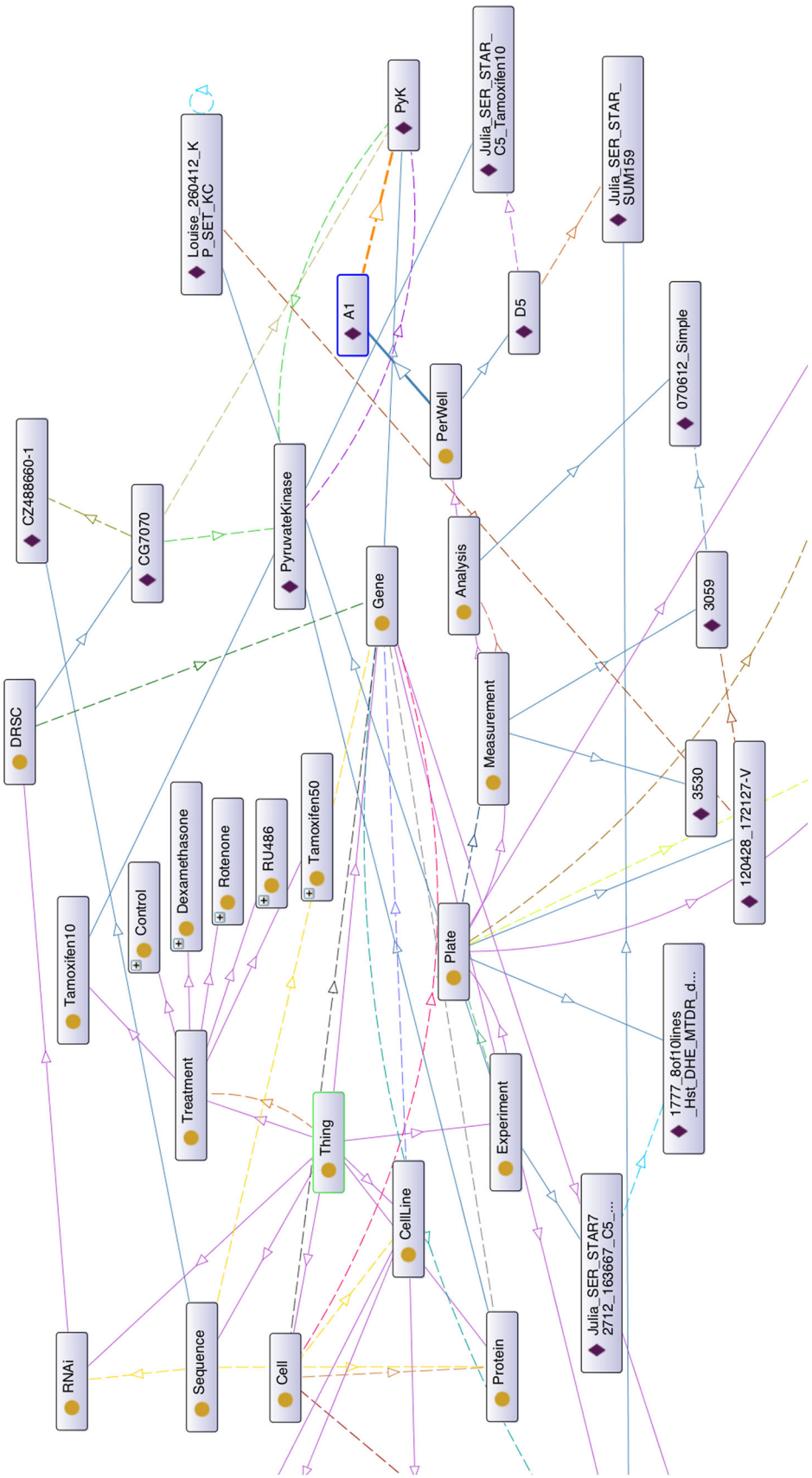

Fig. 3 An example of CDN network

the CDN. For example, the retrieval of information about DRSC gene to PyK protein in the CDN. Also the link between Tamoxifen10 to PyruvateKinase is identified automatically by CDN without manually search.

4 Related work

In the life sciences area there are several systems available that add semantic annotations, primarily these are done through Medline or similar literature databases. Some examples include iHop2, WhatIzIt3 and EBIMed4, and BioAlma5 [14–16]. Entities (such as gene names, protein names, drug names) are recognised and links are added, however, a disadvantage is that the recognition is not active, it is done once, off-line, and is not active in the CDN sense. Since the semantic framework to recognize entity identity across different services is currently missing, these services all point to a small subset of the data that is available for these entities, i.e. these systems are like isolated silos, compared to the CDN's model of self-organizing, distributed structure. Adding semantic markups to more structured data is a relatively new area that has not been systematically addressed in the biosciences. Such a system would take protein sequence files, or entire EMBL databases [17] to search against. It then would add database cross-reference information, and also add semantic annotations statically [18,19], similar as the iHop [16].

Most cloud platforms focus on on-demand (elastic) provisioning that allows for better performance for the customers [20–22]. However, it is difficult for a user to figure out the proper scaling conditions based on input data of an application, especially when the CDN is executed on a third-party virtualized cloud computing infrastructure. Furthermore, client needs change dynamically, requiring different optimizations relative to the amount of reserved resources. Most cloud platforms are proprietary services that run on dedicated servers, translating to lack of elasticity due to vendor lock-in and questionable performance. The works in [23] solve the problem of optimizing the resources of each virtual machine (CPU, memory, etc) to achieve maximum performance, while the work in [24] mainly target at energy efficient cloud solution. We need a fully automated and highly customizable cloud platform that performs elastic resource provisioning to various data networks [6,25].

5 Conclusion

In this paper we present the CDN, an active content management system for personalized medicine research. The CDN is based on a semantic content network approach which overcomes some of the limitations of current content management approaches when dealing with dynamic, distributed and redundant bio-data sources.

Our main contribution over the state of the art in content management systems is that we propose the CDN architecture supporting deployment of the CDN, defining the containers and the networking capabilities that allow remote interactions between data content entities. We also develop a CDN prototype system as a cloud-based, networking middleware for cancer data content discovery and communication.

The initial results show the CDN can facilitate both the cataloguing of samples collected during routine research and the management of datasets generated by numerous multi-step experiments carried out from a single sample. For example, the CDN can provide a platform whereby all tissue samples, experimental step samples, datasets and analysis can be compiled and linked allowing ease of access to every stage in a open manner in order to streamline research, immortalise and protect scientific data and increase productivity.

In the future, we intend to apply real patient data and evaluate performance of the CDN. We also plan to adopt new network algorithms to guide the data integration, and enhance the interaction between CDN and Cloud middleware.

Acknowledgments We thank the Scientific Computing team and RNA Biology Group at CRUK MI for their helpful comments. We would like to thank EU CELAR project partners, in particular, the Laboratory for Internet Computing (LINC), University of Cyprus.

References

1. Lawrence, M., Stojanov, P., Mermel, C., Robinson, J., Garraway, L., Golub, T., Meyerson, M., Gabriel, S., Lander, E., Getz, G.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **505**(5), 495–501 (2014)
2. Chen, R., Mias, G., Li-Pook-Than, J., Jiang, L., Lam, H., Chen, R., Miriami, E., Karczewski, K., Hariharan, M., Dewey, F., Cheng, Y., Clark, M., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M., Greenberg, P., Snyder, P., Klein, T., Altman, R., Butte, A.J., Ashley, E., Gerstein, M., Nadeau, K., Tang, H., Snyder, M.: Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell **148**(6), 1293–1307 (2012)
3. Hanahan, D., Weinberg, R.: Hallmarks of cancer: the next generation. Cell **144**(5), 646–674 (2011)
4. Weinberg, R.A.: Coming full circlefrom endless complexity to simplicity and back again. Cell **157**(1), 267–271 (2014)
5. Giannakopoulos, I., Papailiou, N., Mantas, C., Konstantinou, I., Tsoumakos, D., Koziris, N.: CELAR: Automated Application Elasticity Platform. IEEE International Conference on Big Data (2014)
6. Copil, G., Moldovan, D., Le, D.-H., Truong, H.-L., Dustdar, S., Sofokleous, C., Loulloudes, N., Trihinas, D., Pallis, G., Dikaiakos, M.D., Sheridan, C., Floros, E., Loverdos, C.K., Star, K., Xing, W.: On controlling elasticity of cloud applications in celar. In: Emerging Research in Cloud Distributed Computing Systems. Software Engineering, and High Performance Computing Book Series, Advances in Systems Analysis (2015)

7. Xing, W., Corcho, O., Goble, C., Dikaiakos, M.D.: An ActOn-based semantic information service for Grids. J. Future Gener. Comput. Syst. 26(3), March (2010)
8. Xing, W., Corcho, O., Goble, C., Dikaiakos, M.: Active ontology: an information integration approach for highly dynamic information sources. In: European Semantic Web Conference. Innsbruck, Austria (2007)
9. Wang, L., Khan, S.U., Chen, D., Kolodziej, J., Ranjan, R., Xu, C., Zomaya, A.Y.: Energy-aware parallel task scheduling in a cluster. Future Gener. Comput. Syst. **29**(7), 1661–1670 (2013)
10. Wang, L., Kunze, M., Tao, J., von Laszewski, G.: Towards building a cloud for scientific applications. Adv. Eng. Softw. 42(9), September (2011)
11. Wang, L., Chen, D., Hu, Y., Ma, Y., Wang, J.: Towards enabling cyberinfrastructure as a service in clouds. Comput. Electr. Eng. **39**(1), 3–14 (2013)
12. Xing, W., Liabotis, I., Tsoumakos, D., Sofokleous, S., Floros, V., Loverdos, C.: Translational cancer detection pipeline design (v1.0). Tech. Rep, EU CELAR Project (March 2013)
13. EMBL-EBI Services. http://www.ebi.ac.uk/services
14. Rebholz-Schuhmann, D., Kirsch, H., Gaudan, S., Arregui, M., Nenadic, G.: Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In: Proceedings of the EACL Workshop on Multi-Dimensional Markup in NLP, Trente, Italy (2006)
15. del Castillo, J.C.: Bioalma's text mining solutions for biomedical research. ALMA Bioinformatics, S.L. (2002)
16. Fernandez, J., Hoffmann, R., Valencia, A.: iHOP web services family. In: Freitas, A., Navarro, A. (eds.) Bioinformatics for Personalized Medicine, ser. Lecture Notes in Computer Science, vol. 6620, pp. 102–107 (2012)
17. EMBL-EBI Databases. http://www.ebi.ac.uk/services/dna-rna
18. Zdobnov, E.M., Lopez, R., Apweiler, R., Etzold, T.: The EBI SRS server recent developments. Bioinformatics **18**(2), 368–373 (2002)
19. Hekkelman, H.L., Vriend, G.: MRS: a fast and compact retrieval system for biological data. Nucl. Acids Res. **33**(Web-Server-Issue), 766–769, 2005
20. Xiong, P., Chi, Y., Zhu, S., Moon, H.J., Pu, C., Hacigumus, H.: Intelligent management of virtualized resources for database systems in cloud environment. In: IEEE 27th International Conference on Data Engineering, pp. 87–98, April (2011)
21. Wang, L., von Laszewski, G., Dayal, J., He, X., Younge, A.J., Furlani, T.R.: Towards thermal aware workload scheduling in a data center. In: Proceedings of the 10th International Symposium on Pervasive Systems, Algorithms, and Networks, pp. 116–122, December (2009)
22. Wang, L., von Laszewski, G., Younge, A.J., He, X., Kunze, M., Tao, J., Fu, C.: Cloud computing: a perspective study. New Gener. Comput. **28**(2), 137–146 (2010)
23. Rao, J., Bu, X., Xu, C.-Z., Wang, K.: A distributed self-learning approach for elastic provisioning of virtualized cloud resources. In: 2011 IEEE 19th International Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems, pp. 45–54, July (2011)
24. Sharma, U., Shenoy, P., Sahu, S., Shaikh, A.: A cost-aware elasticity provisioning system for the cloud. In: 31st International Conference on Distributed Computing Systems, pp. 559–570, June (2011)
25. Giannakopoulos, I., Papailiou, N., Mantas, C., Konstantinou, I., Tsoumakos, D., Koziris, N.: CELAR: automated application elasticity platform. In: 2014 IEEE International Conference on Big Data, Big Data 2014, pp. 23–25 (2014)

Wei Xing is the Head of Scientific Computing and Principle Investigator at the Cancer Research UK Manchester Institute (CRUK MI), University of Manchester. Before he joined CRUK MI, Dr Xing was a senior HPC engineer at the Institute of Cancer Research, University of London. Prior he was the head of QA team and a EU research project manager at InforSense Ltd., London, UK. Dr Xing has participated in a large number of European and international projects in the areas of translational cancer research, large-scale data management, high performance computing, and intelligent workflow platform. His current research interests focus on big omic data integrative analysis, translational cancer research, and advanced biocomputing infrastructure.

Wei Jie has been actively involved in the area of parallel and distributed computing for many years, and published over forty papers in international journals and conferences. His current research interests include cloud computing, big data processing and analytics, computing security technologies, and multidisciplinary research. Dr Wei Jie is currently a senior lecturer at school of computing, University of West London, UK. Prior to this, he was a research fellow at the University of Manchester, and a senior research engineer at the Institute of High Performance Computing in Singapore. He was awarded PhD in Computer Engineering from Nanyang Technological University in Singapore.

Dimitrios Tsoumakos is an Assistant Professor in the Department of Informatics of the Ionian University. He is also a senior researcher at the Computing Systems Laboratory of the National Technical University of Athens (NTUA). He received his Diploma in Electrical and Computer Engineering from NTUA in 1999, joined the graduate program in Computer Sciences at the University of Maryland in 2000, where he received his M.Sc. (2002) and Ph.D. (2006).

Moustafa Ghanem is a professor of Software Development and Programming at Middlesex University, London. His research interests are in large scale informatics applications, including large scale data and text mining applications and infrastructures, Grid and Cloud computing and workflow systems for e-Science applications. Before joined Middlesex University, he was a Research Fellow at the Department of Computing, Imperial College London. At Imperial College London, he had also been involved in teaching a number of courses in data mining and bioinformatics. He has been the Research Director of the spinout company InforSense Ltd since its inception in 2000, where he has led the design and development of their TextSense product and also led its participation in a number of EU-funded Research Projects with applications in drug discovery, healthcare and collaborative R&D infrastructures. Over the past few years, he has helped establish the Centre of Informatics Science at Nile University in Egypt focusing on the use of modern informatics methods for addressing problems of national importance in healthcare, agriculture, environment and cultural heritage as well as local IT industry competitiveness.