Long-term recommendations for RIKEN's life sciences database platform (final report)

Oct. 29, 2015
The database working group
(translated into English on Apr. 07, 2017)

Table of contents

1. Database working group: background

The database working group published short-term recommendations for RIKEN's database infrastructure in August 2014. The recommendations were based on responses (approximately 200) to an internal online survey. The recommendations are summarised as follows.

i. The RIKEN Advanced Center for Computing and Communication (ACCC) should provide a method that allows databases created by RIKEN researchers to be found easily (database directory, etc.) through collaboration with the activities of all-Japan database catalogue.

ii. RIKEN ACCC should develop technologies to systematically manage metadata, i.e. data that describes data, by themselves.

iii. RIKEN ACCC should plan by themselves to develop a database infrastructure that implements a data sharing and publishing framework to facilitate data utilisation the primary principle rather than data integration.

However, the 2014 report did not consider concrete long-term recommendations based on global trends in life sciences research and databases; thus, the database working group concludes that further examination of long-term and concrete recommendations is required. In February 2015, at the first annual Bioinformatics exploratory committee meeting for FY 2014, this initiative was approved and the database working group reports this recommendation.

2. Worldwide trends in life sciences, biomedical research and databases

i. Data-oriented life sciences

In life sciences and medical sciences, the mainstream of scientific approach has been 'hypothesis-oriented' where hypotheses are formulated by scientists. However, it is expected that 'data-oriented' approaches, which discover novel life phenomena and hypotheses based on statistics drawn from large-scale comprehensive data, will be employed more frequently in future life sciences research. The various trends in life and medical sciences fields are described below.

(1) Omics analysis

A) Analysis of genomic DNA

Due to the progress of massively parallel DNA sequencing, a complete personal genome sequence and all exon sequence can be quickly determined by next generation sequencers (NGS). The 1000 Genomes Project (http://www.1000genomes.org/), a cooperative international research project, create a database of human genome sequences collected from approximately 2,500 people from 2008 to 2015. In the United Kingdom, the Genomics England programme (http://www.genomicsengland.co.uk/) was established to sequence 100,000 whole genomes in 2013. Public organisation and private venture companies in the United States and Saudi Arabia have been conducting similar projects. In Japan, the Tohoku Medical Megabank Organization (http://www.megabank.tohoku.ac.jp/), the National Center Biobank Network (NCBN; http://www.ncbiobank.org/) and other agencies has been collecting Japanese human genome sequences. The RIKEN Center for Integrative Medical Sciences (RIKEN IMS) contributes significantly to the NCBN. Some primary data are stored in the NCBI SRA (http://www.ncbi.nlm.nih.gov/sra), the DDBJ DRA (http://trace.ddbj.nig.ac.jp/dra) and the EBI ENA (http://www.ebi.ac.uk/ena), and related phenotype data are stored in the DDBJ Japanese Genotype-phenotype Archive (https://trace.ddbj.nig.ac.jp/jga/index.html). Note that use purposes are examined prior to granting permission to use these primary data and related phenotype data. Personal genome information is protected and attempts are made to track research progress. Furthermore, cancer genome sequencing is actively carried out. The International Cancer Genome Consortium (ICGC, https://icgc.org/) has lead cancer genome sequencing of more than 110,000 cases and developed a database. RIKEN IMS (formerly RIKEN CGM) significantly contributes to the ICGC.

In addition, genome sequencing projects of non-model organism are now possible because whole genome shotgun sequencing and long read sequencers are spread to various researchers. For example, the 5,000 Insect Genome Project (https://genome10k.soe.ucsc.edu/) was established to sequence the genomes of 5,000 insect and arthropod species.

In generally, primary data of NGS data is stored in FASTQ format. To reduce file size of NGS data, a new file format of genome sequence is developed and proposed. The format describes only the genetic polymorphism on a standard genome sequence using graph data structure.

A data integration technique of genome sequencing data and other omics data is also important for understanding biological phenomenon or disease. We should not only develop databases of sequence but also study a method of data visualization, statistical modelling of gene-disease relationship and, integration of various clinical information.

B) Metagenome analysis

Meta-genome analysis determines the genomic composition of biological material from environmental samples using an NGS. This enables accurate measurement of microbial ecosystems including their dynamics in soil intestinal environments even in the case of uncultured microbes that are microorganisms difficult to culture. Contributions in various fields, such as environmental preservation, soil improvement and health, are expected. RIKEN is presently conducting metagenome research, including improving biomass production (RIKEN CSRS) and intestinal immunity (RIKEN IMS).

Similar to NGS analysis, analysing large-scale data is a key for the development of metagenome analysis, and increasing data storage capacity and data analysis throughput are also required.
Increasing the speed of homology search, which is used to specify and classify species, is also an important issue. A data analysis method that infers environmental functions using biological population data using homology search is required. To address these issues, technology for discovering new knowledge by comprehensively and quantitatively analysing all related available information, including gene functions, pathways and the characteristics of each species, must be realised. RIKEN QBiC has been developing an analysis tool that effectively specifies the species from a mixture of genome sequences. Development of a database for the metadata of sampled environments is important, and semantic web technologies that enable integration of various types of information are anticipated. The goal of the National Bioscience Database Center's (NBDC) Database Integration Coordination Program (DICP) is to integrate life sciences data, and integration of microorganism metagenome information using semantic web technology has proceeded (Tokyo Institute of Technology/NBDC DICP).
Novel methodologies, such as meta-transcriptome, which captures the functional dynamics of organisms in an environment using a methodology similar to metagenomic analysis, and environmental DNA analysis that directly sequence soil and seawater organisms/microorganisms, have been developed. It is expected that mutual interaction between an environment and organisms sampled from that environment can be captured dynamically in detail. In this context, advanced methodologies will be required to analyse a large amount of data, efficiency of analysis and the development of metadata databases.

C) Transcriptome analysis

Transcriptome analysis, large-scale profiling of RNA molecules, was performed using microarrays. Its comprehensiveness, coverage of RNA structures, quantitative accuracy can be increased by using NGSs. RNA-Seq is a method widely used recently, which enables us to determine splice sites as well as quantify the levels of gene expressions, and RIKEN also has developed several technologies. Cap Analysis of Gene Expression (CAGE) is a method developed by RIKEN CLST, which enables us to determine transcription initiation sites with their frequencies, and single-cell RNA sequencing methods are also developed by RIKEN CLST, and RIKEN ACCC. A method of single-cell RNA sequencing from several hundreds to tens of thousands of cells using cell barcode technology (RIKEN IMS), mixed reactions and micro flow channel technology have been under development, which will contribute to the discovery of unknown cell subtypes in tissue samples. "Single cell project", which can be based on RNA profiling as well as other measurements, was adopted as a major theme of budget demands for life sciences research in FY 2015, and increase of the data amount is expected in future. Furthermore, RNA sequencing technologies are becoming increasingly diverse. For example, microRNA, localised RNA in nuclei and cytoplasm, transcription RNA (ribosomal profiling), and protein binding RNA have been monitored, which capture RNA metabolism and dynamics. Monitoring these aspects across and space should be effective to understand cell systems. These increase not only the data amount but also data complexity, which underline importance of computational analysis and databases.
Comparison among existing gene expression level databases is important to interpret newly acquired RNA profiles. Currently, NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) and EMBL-EBI Array Express (https://www.ebi.ac.uk/arrayexpress/) host public gene expression repositories; however, these repositories primarily store microarray data. In future, databases for RNA-Seq and CAGE data will be required. Similarity search of gene expression levels is highly relevant, which require all metadata such as sample information are organised properly. In future, annotate gene expression databases are essential, which could be implemented through semantic web technologies, text-mining and expert curation. It has to be noted that data journals, e.g. Scientific Data, are recently established to publish data and data descriptions with standards. RNA sequencing requires RNA sequence and structure databases for reference. Reference transcriptome models have been developed as Ensembl (http://ensembl.org/) and RefSeq (http://www.ncbi.nlm.nih.gov/refseq/), and recently GENCODE (http://gencodegenes.org/) has been established to achieve comprehensive and compatible models.
The RIKEN FANTOM project (http://fantom.gsc.riken.jp/jp/) publishes the promoter level expression

information of a large number of cells. However, a public expression database corresponding to NCBI GEO or EMBL-EBI Array Expression does not exist in Japan (DDBJ), and few expression databases are being developed in Japan and the rest of Asia. This issue must also be addressed.

D) Epigenome analysis

Understanding epigenetic mechanisms, where phenotypes and diseases are inherited over generations but the genome does not change, is a very important issue in current life sciences research. After completion of the Human Genome Project, Japan has focused on post-genome projects. On the other hands, in the West, ENCODE (https://www.encodeproject.org/) and the Epigenome Roadmap (http://www.roadmapepigenomics.org/) have focused on annotation of non-coding regions and developed databases that play fundamental roles in current epigenome research. Such projects are delayed in Japan, where only international databases are used. However, the Japanese team (http://crest-ihec.jp/) that works with the International Epigenome Consortium (IHEC, http://ihec-epigenomes.net/), which collects human epigenome information, contribute to provide novel sequencing techniques to IHEC members.

Another issue that should be addressed is the development of data integration technologies to understand diseases and environmental responses by comprehensively connecting relationships between epigenome (chemical modification of DNA and histone) and gene expression levels. Furthermore, one of the most important issues in life sciences is the investigation of how epigenome changes across generations affect phenotypes. RIKEN focuses on this issue, e.g. the Shinkai epigenetics project (Competitive Programme for Creative Science and Technology) and the Disease epigenome project (life sciences budget request in FY 2015-2017) are being actively promoted.

Epigenome information in the existing databases, particularly DNA and protein binding and histone modification data (ChIP-Seq), has problems with reproducibility using antibodies and reagent kits. To reuse such data, related metadata should be organised and stored in a database.

Histone modification and DNA methylation affect the higher-order structure of chromatin, and, as a result, change the gene expression level. RIKEN CDB has sequenced the higher-order structure of chromatin using HiC. In future, integration of epigenome information and image data captured using histone and chromatin live imaging technology (RIKEN QBiC) will become important.

E) Metabolome analysis

Metabolome analysis comprehensively analyses low molecular metabolites related to biological phenomena using nuclear magnetic resonance (NMR) and various mass spectrometric methods. In addition to fundamental research, such as elucidation of physiological phenomena, metabolomics is widely applied in medicine, e.g. to identify biomarkers. Due to the development of high-quality measurement technologies, Japan holds approximately 60% of all metabolome information in the world. Furthermore, in metabolome analysis, to identify compounds by comparing peaks and spectra with standard products, large-scale collection, concatenation and sharing of comprehensive quantitative metabolite detection data obtained from experiments and compound knowledge are extremely important.

RIKEN IMS has been developing technology to measure lipid metabolome, and the volume of data is expected to increase in future. In addition, RIKEN has established an integrated database that collects metabolome data (RIKEN CSRS/NBDC DICP) and is promoting the integration of compound, spectrum, activity and genome information. Kyoto University provides a metabolic pathway database, KEGG (http://www.genome.jp/kegg/).

F) Phenome analysis

A phenome represents a set of phenotypes expressed by cells, tissues, organs, organisms or multiple species. Determining the significance of a phenome involves much more than simply collecting data. Phenome is important since it is the basis of detection and visualization of biological relationships among various phenotypes (e.g. the relationships between diseases and individual pathologies, conservation of the evolutional mechanisms across species as a base concept of a model organism, and interaction among individuals in an ecological system). Phenome analysis is essential to understand all aspects of the mechanisms of life; however, there are a great many obstacles. Aiming

to overcome the obstacles, various trials to improve interoperability of experimental data have been performed using imaging and standardization of measurement protocols.

F-1) Comprehensive phenome analysis

Traditionally, phenotypic analysis has been performed as individual research based on a given hypothesis particularly in higher organisms. However, with this approach, negative data often remain unpublished because it is difficult to publish as a paper. As a result, objectivity and interoperability of experimental data produced from individual researches are relatively poor to be used in data science. These limitations would present obstacles to understanding the complete life mechanism in near future.

Comprehensive phenotyping is a large-scale analysis which usually measures multiple biological properties of genetically modified organisms using standardised analytical methods. To obtain statistical results, comprehensive phenotyping is performed for multiple samples or cohorts. Through such large-scale analysis, more objective life function data can be obtained rather than small scale analysis. Recently, comprehensive phenome analysis becomes to be performed using not only microorganisms, worms and insects but also higher organisms, such as mice and Arabidopsis and human cohort analysis.

Generally, phenome analysis of small organisms contributes to the discovery of unknown life functions by obtaining highly objective data or by using advanced measurement techniques, such as image analysis, to acquire data that could otherwise not be measured by conventional methods. On the other hand, large-scale analysis is difficult and data interoperability is problematic for higher organisms. However, highly applicable data can be obtained from higher organisms.

Recently, an international collaborative project involving institutes (including RIKEN BRC) in 12 countries is performing comprehensive phenotyping as a functional analysis of mouse whole genome (http://www.mousephenotype.org). In human, multiple cohort analyses are performed for acquisition of medical data. It is expected that human cohort analysis directly contribute to medical sciences. Cohort data also play a prominent role at Genomics England and the Precision Medicine Initiative in the United States.

Linking comprehensive phenome data across species and research domains is an emerging trend. The Monarch Initiative (http://monarchinitiative.org/) is an international portal that integrates cross-species gene, genotype, variant, disease and phenotype data, including human, mouse, rat, nematode and Drosophila data. The Monarch Initiative website also shows relationships between human diseases and the phenotypes of model organisms used in experimental biology. In Japan, there is a growing need in clinical research to utilise phenotype data as the ‘components’ of a disease, which enables the identification of phenotypic similarities among diseases. Furthermore, establishing phenotype-based relationships among research resources is crucial to maximise sharing of clinical data relative to the promotion of information and communication technology (ICT) for medical treatments by the Japan Medical Research and Development Organization (http://events.biosciencedbc.jp/images/togo2015/p28-29.pdf).

At RIKEN, a phenome integration database of model organism resources (RIKEN BRC/NBDC DICP; http://jphenome.info) and a plant phenome database (RIKEN CSRS; http://metadb.riken.jp/metadb/db/SciNetS_ria61i) have been developed, and further utilisation of these data is expected.

F-2) Bioimage informatics

With the rapid development of genome science since the 1990s, measuring technology of so-called omics, such as transcriptome, proteome, metabolome, etc., also developed rapidly due to automation, speedup and quantification of such technology. Acquisition of phenotypic data depended on visual observation and recognition by researchers; thus, typically, such data had low objectivity, quantitativeness and throughput, and this has been an obstacle to realise data-oriented life sciences research. However, from the beginning of this century, a computational phenotype analysis technique that combines computer-controlled automatic capture of microscopic images and automatic recognition of image features of biological specimens in such images has been developed. Thus, the quantity and quality of phenotypic data have changed. Computational phonotype analysis can be

categorised as high content screening, quantitative analysis of biological dynamics and atlas construction.

High content screening involves automatic imaging and automatic image recognition of cells cultured on a microplate and is widely used for drug screening, etc. As a more advanced form of the system, there are cases where samples, e.g. tissue or nematodes, are cultured on a microplate. Various companies have actively engaged the development of high content screening equipment, and this equipment is commonly used in life sciences research at many universities and laboratories in Europe and the United States. However, the adoption of such equipment has been slow in Japan.

Quantitative analysis of biological dynamics uses multidimensional automatic imaging and automatic image recognition to measure and analyse cell, embryo and tissue dynamics. In this field, a microscope imaging system optimised for target biological phenomenon and an image recognition method optimised for the images captured by such a system are required. From the mid-2000s, research organisations in Western countries, such as the Janelia Farm Research Campus (USA) and the Max Planck Institute of Molecular Cell Biology and Genetics (Germany) have made significant advances toward integrating microscope imaging systems and image processing technologies. These organisations have established flagship projects, such as 'tracking whole cell nuclei in the Drosophila embryo throughout the whole embryogenesis' and are developing advanced technologies. A comparable research institution has not been established in Japan; thus, Japan is lagging behind Europe and the United States. RIKEN QBiC is the only research institution in Japan that actively promotes such activity; however, quantitative analysis of biological dynamics using bioimage informatics is not the aim of establishment of QBiC; thus, the scale of the activity is small. Quantitative analysis of biological dynamics has become indispensable for cell biology and developmental biology research in recent years, and the demand for general-purpose image analysis tools that can be used in such research has increased. Currently, ImageJ (Fiji) and CellProfiler are widely used. In addition, Janelia Farm Research Campus, Max Planck Institute of Molecular Cell Biology and Genetics, and other agencies have developed various software tools. Such generic software development in Japan has not kept pace with Europe and the United States, and urgent strategic countermeasures are required. RIKEN RAP has been developing general-purpose image analysis software for biological images.

In atlas construction, a three-dimensional sample is imaged at many different focal planes using fluorescence or electron microscopes. Then, a database is developed by computationally reconstructing the three-dimensional structure of the sample and/or a three-dimensional expression state of the genes. The Allen Institute for Brain Science (USA) was established in the early 2000s to construct a 3D atlas of mouse and human brains. Development of an image recognition tools for the 'connectome', which reconstructs the whole network of the cranial nervous system from a three-dimensional images of electron microscope is the most important task in this field, and the Allen Institute for Brain Science and the Janelia Farm Research Campus are leading such technology developments. Japan's activities and technical expertise in this field are low; however, since 2014, RIKEN has led an initiative to develop a marmoset brain connectome and related technologies (RIKEN BSI, CSLT and RAP). Note that RIKEN (including RIKEN BSI, CDB and QBiC) is a global leader of optical clearing technologies for tissues.

In addition to the three fields listed above, digital pathology, which infers pathological conditions from pathological images, has attracted significant attention. Digital pathology research will be an increasingly important field to improve diagnosis and to eliminate treatment disparities between hospitals and between regions.

Currently, most data measured using bioimage informatics technology are published using a database constructed for each project. In addition, data formats differ for each project. To promote utilisation of quantitative measurement data of biological dynamics, RIKEN has developed a unified file format, i.e. BDML and has developed the SSBD database (RIKEN QBiC/JST-NBDC). RIKEN QBiC has been working to establish an international consortium that manages the measurement data of biological dynamics on a global scale.

Determining how to share image data among research communities is a major issue in bioimage informatics. In particular, sharing images acquired using state-of-the-art microscopies, which are only deployed in a limited number of research institutions, is an important issue for cell and developmental

biology research, where the ability to acquire images directly affects the quality of the research. From this perspective, the enhancement and standardisation of image databases are required. In the United States, the ability to share high-quality images has been advanced by the development of the JCB DataViewer by the Journal of Cell Biology and the development of the Cell Image Library by the American Society of Cell Biology. However, efforts related to image databases in Japan have been largely delayed. In FY 2015, the SSBD database began collecting cutting edge cell and developmental biology images (RIKEN QBiC/JST-NBDC). The OMERO open-source image database software developed by the Open Microscopy Environment (OME), which is an international collaborative project, is widely used to share microscopy images and other images. To use images from a wide range of research communities, standardisation of metadata is required; however, adequate standards have not been developed. RIKEN QBiC is working on standardisation of image metadata in cooperation with the OME.
In addition, the amount of data has been increasing due to the enhancement of the resolution and multi-dimensionality (spatial and temporal axes) of image data. To promote bioimage informatics, it is necessary to respond to large-scale data in image format (compatibility with binary format (HDF5, etc.)) and data transfer infrastructure.

(2) Network analysis
Elucidation of the entire network structure comprised of interrelationships among molecules in the living body or intracellular relationships is central to an integrated understanding of life, while only a part of this network has been revealed to date. The acquired knowledge to date can be a powerful foundation for understanding new data. Therefore, network analysis is a major theme in life sciences, and a framework that effectively utilises analysis results is considered an important research infrastructure.
Molecular network data can be roughly classified as 1) mutual relationships extracted by experts who examine scientific literature and 2) mutual relationships obtained by comprehensive methods, such as omics analysis. For the former class, KEGG Pathway, which was established in Japan ahead of the times, is a representative example of the large-scale molecular map. However, due to changes in the usage licence etc., the utilisation of overseas databases (e.g. Reactome, Panther, WikiPathway, BioModels, etc.) has increased. In addition, the use of commercial proprietary databases is also increasing, which implies diversity in their content and use cases. In areas where no previous research has been performed or where data cannot be collected by professionals in a timely manner, the latter class, i.e. the control/similarity relationship estimated from exhaustive data, can be applied. In fact, such databases (e.g. String, CellNet, CMap, etc.) are being constructed and have been used recently.
In addition to the construction of databases, a methodology to describe a network model has been proposed for various hierarchies. The Systems Biology Markup Language (SBML; proposed by Kitano et al.), which describes the reaction process of molecules, is a pioneer methodology, and, in recent years, the COMBINE (http://co.mbine.org/) community has actively encouraged exchange among various methodologies.

(3) Computational structure biology
This is a research field that examines molecular structure and motion using structural analyses and molecular simulations to understand the functions of biomolecular complexes composed of proteins, RNA, sugar chains, etc., for application to drug discovery, etc. The Protein Data Bank Japan (PDBj) at Osaka University collaborates with RCSB and BMRB in the United States and PDBe in Europe. The PDBj publishes a database containing the three-dimensional structures of biopolymers as a globally unified PDB archive. RIKEN has made significant contributions to elucidating the basic structure of proteins and analysing functions in the protein 3000 project using NMR and large synchrotron radiation facilities, including Spring 8. Recently, research using in-cell NMR and an X-ray free electron laser for structural analysis has begun. In addition, by combining various computational methods and experimental data, a method that constructs a biomolecular structure model from experimental data using a high-performance computer, such as the K-computer, has been developed (RIKEN AICS).

(4) Artificial intelligence approach

The data handled in life sciences becomes large scale and complicated, and extraction of knowledge by computer is indispensable. Therefore, realisation of advanced analysis using artificial intelligence is extremely promising. IBM's Watson question response system (consisting of a combination of existing technologies, such as unstructured data analysis, natural language processing, fast search, etc.), which won the quiz king in 2011, is expected to be applied as a decision support system for medical diagnoses, examination of potential drug-drug interactions, etc., and collaborative research with the Institute of Medical Science at the University of Tokyo has begun (http://www-06.ibm.com/jp/press/2007/15/3001.html). Recently, expectations are growing for a deep learning method to find unknown feature quantities by multi-layered repeat learning without a teacher. In 2012, Google succeeded in making a computer learn a 'cat' (by creating a neuron that reacts to a cat) using multi-layered neural network learning without manually-designated feature points. Although technical backgrounds are diverse, in these days, basic research into practical use becomes to be established to determine the rules behind a huge amount of data using the outstanding computing power and memory capacity of modern computers. Especially in life sciences, it is expected that knowledge discovery from so-called 'big data' will be accelerated by making full use of such technologies, and this research field will be of great importance in future.

Thus, in nearly future, it becomes more important that life sciences database infrastructures are easy to be used by artificial intelligence methods. However, a term 'artificial intelligence' is so generic involves variety of technologies, such as data mining, text-mining and deep learning. Therefore, currently standard for big data cannot be generalised because clear restrictions and standards for big data depend on individual technologies. For example, in the study field of the ontology, researchers model human knowledge positively in a machine-readable way. However, in deep learning, it is considered that a 'teacher', e.g. instruction of feature points, is not required prior to modelling. Such an approach involves fewer active manual interventions, and it is possible that 'newer' knowledge that is difficult for humans to find can be obtained.

Meanwhile, modelling based on human intellectual activities and generating uniform and standard data based on such modelling is advantageous for machine processing of artificial intelligence. For example, even in case of successful deep learning, an image is highly standardised large-scale data, and processing such data requires massive computing resources. In addition, by modelling knowledge (in which bias cannot be avoided), processing according to human requirements can be performed. Furthermore, relative to correspondence with the metadata described in (5) below, in the semantic web framework, which is the global standard for metadata descriptions, it is expected that a huge knowledge base that can be accessed relatively easily by artificial intelligence can be constructed through data distribution and integration with precise and systematic metadata descriptions. Metadata and abundant knowledge resources can be used to examine the validity of artificial intelligence processing results, which is expected to facilitate the realisation of advanced knowledge processing techniques. Standardised metadata also facilitates the automatic use of data by programmes; therefore, wider realisation of a pipeline through which programmes can automatically connect and analyse data in combination with container technologies, such as docker, is expected.

At RIKEN, discussion about bioinformatics analysis infrastructures using artificial intelligence technology has begun, and use of the Garuda platform (http://www.garuda-alliance.org) has been proposed as a data infrastructure for statistical data mining, such as data classification, association analysis and correlation analysis (RIKEN IMS). The development of an automatic evolution experiment system that combines comprehensive analysis methods, evolution experiments and mathematical model analyses is also underway (RIKEN QBiC; http://www.qbic.riken.jp/mbd/index.html). However, few research papers on this subject have been published. Therefore, it is supposed that collaboration between artificial intelligence and metadata, researchers with a clear vision of the collaboration are few both inside and outside RIKEN

(5) Utilisation of metadata in life sciences

In life sciences, ongoing efforts are being made toward the standardisation of a vocabulary that describes metadata as ontology, e.g. the Gene Ontology Consortium (http://geneontology.org/), which has attempted to formulate a common vocabulary for functional descriptions of genes since 1998. The

Open Biological and Biomedical Ontologies (OBO) Consortium (http://www.obofoundry.org/), which aims at developing common controlled vocabularies (ontologies) across various biomedical fields, was established in 2001. Furthermore, the National Center for Biomedical Ontology (http://www.bioontology.org/), which aims at research and development of an innovative technology and methods on ontologies was launched in 2005. The use of ontologies developed rapidly after the Web Ontology Language (OWL), a semantic web technology used to exchange ontology data on the web, became popular. Currently, a lot of ontologies have been proposed and published from various research fields in OBO or OWL formats. The semantic web of life data has been developed in response to the development of such ontologies.
In 2006, a research paper from Bio2RDF (http://bio2rdf.org/), a website that converts and publishes existing well-known databases in a semantic web format, was published, and, in 2013, EBI also published six databases, including UniProt, which corresponds to the semantic web (https://www.ebi.ac.uk/rdf/platform). A semantic web research community for life science databases has been developed, including medical and drug discovery fields. Life sciences is the prominent topic at the International Semantic Web Conference (http://iswc 2015.semanticweb.org/), and Semantic Web Applications and Tools for Life Sciences (http://www.swat4ls.org/) is an international conference specialised in the implementation of life sciences applications. At these international conferences, the results of research into ontologies that integrate electronic medical records and disease names primarily in Europe and the United States have been presented in recent years.
In Japan, at the Database Center for Life Science (DBCLS) launched in 2007, the semantic web began to be utilised for organic use of content around 2010 and National Bioscience Database Center (NBDC), which launched in 2011, has been developing programmes that support database publication based on the semantic web.
Prior to these activities, RIKEN has been promoting database construction research based on semantic web technology since 2005. In 2009, the semantic web technology was adopted to integrate the database published by RIKEN (press release: The disclosure standard of RIKEN's database construction base is unified into semantic web, http://www.riken.jp/pr/press/2009/20090331_2/). In 2015, based on semantic web technology, the RIKEN MetaDatabase (http://metadb.riken.jp), which was designed to realise data distribution requirements proposed by the database working group, was published. The RIKEN MetaDatabase supports data format of IntegBio (http://integbio.jp/en/), which is an integrated database catalogue from four ministries in Japan and the World Wide Web Consortium's (W3C) Dataset Descriptions: Health Care and Life Sciences (HCLS) Community Profile (http://www.w3.org/TR/hcls-dataset/), which contains biomedical database metadata.
In a keynote speech given at Bit-IT World (http://www.bio-itworld.com/) in 2005, Tim Berners-Lee, the inventor of the World Wide Web, stated that the life sciences field was expected to demonstrate leadership relative to semantic web technology. As expected, information infrastructure of life sciences has been formed and developed as a global data distribution framework.

ii. Database technologies

(1) Conventional relational databases

A database is an information system that collects and manages data according to a predetermined format. In information science, research into data models that abstract data structures, management methods and practical search algorithms has been promoted to make it possible to search and extract data in a database quickly and accurately. Currently, the most popular type of database is the relational database, and the relational model was devised based on set theory with a foundation in predicate logic.

(2) Transition of database technology relative to diversification and increasing volumes of life sciences data

Data represented by a relational model can be described as a table. Furthermore, tables can be connected via data constituting rows and columns using a set of data called relations. The table format is commonly used by experimental researchers in biology, and free practical implementations have

become common, such as MySQL (https://www-jp.mysql.com/) and PostgreSQL (https://www.postgresql.jp/). Therefore, life science databases have been constructed in the relational model. However, with the diversification of life sciences data, the demand for collecting data from multiple databases and conducting integrated analyses has increased. A relational database constitutes a closed system in a single database. Cooperation among multiple databases is difficult due to differences among table data formats and relations. In addition, data semantics can differ even when the data formats are the same among multiple databases. Efforts, such as myGrid (http://www.mygrid.org.uk/), have been made to deal with the XML format and utilise the WebService API (http://www.mygrid.org.uk/); however, due to high implementation cost, etc., such efforts have not yielded widespread data integration.

Processing of big data, which is a huge amount of data that cannot be processed using conventional technology, including relational databases, can be considered a framework for general-purpose data processing with large-scale data. Thus, a new mechanism is required to process accumulated and enormous data in a short period to obtain processing results in real time, such as trend analyses. The NoSQL database technology is expected to be utilised in future because it satisfies this requirement. Relational databases implement functions, such as data processing exclusion control and consistency guarantees that can handle data correctly. However, NoSQL does not guarantee data processing consistency and prioritises improving data processing speed by simplifying the data structure and using distributed parallel processing. Life sciences data processing often deals with collections of huge volumes of similar data, i.e. the problem (data to be processed) can be divided and solved using parallel processing performed by mutually-independent processes. Apache HBase (https://hbase.apache.org/) of the open-source Apache Hadoop (http://hadoop.apache.org/) project is a prominent concrete database system inspired by a system that realises Google's large data as distributed storage. Hadoop is suitable for genome data analysis for life sciences. For instance, the large-scale Contrail genome assembly processing system and a prototype system for analysis of NGS data, which are performed in the cloud, are good examples.

Relative to technologies to address large-scale data, technological proposals specialised in life sciences have been performed in addition to the above-mentioned general-purpose big data processing infrastructure. In particular, to address large-scale data generated by an NGS, high-speed processing of genomics-specific queries has been developed. Specifically, several file formats have been proposed, such as BAM (https://samtools.github.io/hts-specs/SAMv1.pdf), VCF (http://samtools.github.io/hts-specs/VCFv4.2.pdf) and bigWig/bigBed (Kent et al. PMID:20639541). These file formats are defined by storage and index methods according to data type, and tools that support these data formats are used extensively.

Regarding biological image data, OMERO (https://www.openmicroscopy.org/site) was developed as a software platform for image management, display and analysis, and is employed by the JCB DataViewer (http://jcb-dataviewer.rupress.org/). OMERO is becoming a de facto standard. Since biological image data are very large, corresponding network bandwidth is required for both intra- and inter-institution data transfer.

(3) Metadata technology to promote data distribution on a global scale

Although the progress of the standardisation of data formats specialised for various data types, data distribution technology for comprehensive data analysis has been difficult using the conventional frameworks. Relative to data integration and distribution, it has become desirable to realise database integration at the data layer and the database implementation layer. As a result, metadata technology, especially utilisation of semantic web technology (http://www.w3.org/standards/semanticweb/), to describe and distribute metadata in a form that corresponds to a global network is attracting significant attention.

Tim Berners-Lee of the W3C has advocated the use of semantic web technology for development of the World Wide Web. Documents on the web are written in HTML, which describes the structure of documents and links between documents. However, HTML does not describe the content of a written document or its semantics. On the other hand, the semantic web attempts to handle the semantics of documents and data by standardising data formats, interfaces and tools. The core of the semantic web is a mechanism that exposes data and implements a data network by establishing links with

meaningful relationships among datasets. It is expected that sharing and reusing data on the Web, including documents, will be promoted not only by people but also in information processing systems.

iii. Data visualisation

When extracting meanings and rules from enormous volumes of data, visualisation plays a very important role because, through visualisation, humans can intuitively recognise the phenomena and structures contained in data. In addition, in big data analysis in many scientific fields, statistical analysis is an important technique for intellectual discovery. However, as demonstrated with Anscombe's quarted, different data can have the same statistics; therefore, data visualisation is essential for accurate analysis of big data. Visualisation techniques include interactive three-dimensional data display, coarse visualisation and latent variable discovery, and further include the acceleration of computation to enable the interactive processing of big data. Generally, Gnuplot (http://www.gnuplot.info/), MATLAB (https://www.mathworks.com/products/matlab.html), GrADS (http://cola.gmu.edu/grads/), ParaView (http://www.paraview.org/), etc. are widely used visualisation and analysis tools. In addition, many other applications specialised for scientific visualisation of three-dimensional data and software specialised for specific data have also been developed. However, the development of a visualisation tool to accelerate knowledge extraction from large-scale data has only just begun. This field has been studied actively, and major international conferences, such as IEEE VIS (http://ieeevis.org/), exist overseas. However, recognition in Japan remains low. At RIKEN, the development of visualisation technology for simulation results obtained using the K-computer (RIKEN AICS) and analysis data of biolobical dynamics (RIKEN QBiC) is underway.

3. Long-term recommendations for RIKEN's life sciences database platform

i. Summary of issues and problems facing life sciences in Japan

From the above, the issues facing RIKEN's life sciences database are summarised as follows.

A. Improving information infrastructure to utilise big data

a. Large-capacity storage improvement, speedup of calculation processing and acceleration of data transfer due to increased data volumes

Increased data volumes (such as NGS and image data) are expected in nearly all research fields. Thus, a high-performance infrastructure to store and utilise such large-scale data without delay is required.

b. Promotion of innovation by open data

Use of data or knowledge from other research fields is indispensable. Thus, the promotion of innovation by opening RIKEN's data is required. It is also necessary to widely disseminate the types of data that exist in RIKEN's information infrastructures.

c. Description standardisation for integration and concatenation of data or metadata

To mutually utilise a wide variety of data and knowledge, systematically interconnecting metadata and making such metadata widely available in a standardised manner are necessary.

B. Knowledge extraction from big data

d. Artificial intelligence approach

It is expected that life sciences data will become increasingly large and complex; thus, an artificial intelligence approach will become essential for data analysis. RIKEN should proactively develop such technology for that purpose. In the development of a data infrastructure, the development and standardisation of metadata to facilitate ease of use with artificial intelligence is necessary.

e. Development of statistical analysis methodology for large-scale data analysis

Conventional statistical analysis methods have problems when applied to big data. For example, too many false negatives can be generated with large-scale, multi-item data. Thus, it is necessary to generalise such methodologies such that they can be used for a wide variety of analyses. To efficiently extract the meaning of data, such methodologies should be applied to both domestic and RIKEN's big data analysis research.

f. Data visualisation and coarse visualisation

The importance of visualisation technology for big data analysis is expected to increase rapidly as technology to solve the theoretical problems with artificial intelligence and statistical analytic approaches in future. RIKEN should proactively promote the development of an environment that enables efficient development and sharing of visualisation software, such as a collection of visualisation software and unification of the data formats such software handles.

g. Information extraction from images and knowledge extraction

Extraction of information from images yields quantitative and reproducible data is expected to play a significant role in future life sciences big data and sensor data and is also expected to be applied to many other research fields. In the life sciences field, both developing advanced algorithms specialised for flagship projects and developing highly versatile tools applicable to individual research projects are required. The current situation in Japan is lagging far behind the West from both perspectives. Images in the life sciences field are diverse, and it is necessary to select and develop appropriate image analysis methods for individual research projects; however, strategies that provide solutions for a wide range of requirements remain missing. Therefore, for Japan and RIKEN to lead the world in this field, selecting strategies that satisfy a broad range of requirements is important.

C. Innovation through data utilisation

h. Direct contribution to disease, drug discovery and health promotion through integrated understanding of omics information

Due to the integrated understanding of life, it is considered that disease, drug discovery and health promotion technologies will be further developed in future. RIKEN is required to contribute directly to such technologies. Based on the above circumstances, it is desirable that the database infrastructure covered in this report be constructed to support more direct contributions and broader information dissemination among RIKEN's life sciences research departments.

ii. Recommendation for database infrastructure

RIKEN's life sciences research activities are unconventionally extensive and large-scale, and the comprehensive scientific capabilities and large facilities at RIKEN are uncommon elsewhere in the world. RIKEN can be considered a microcosm of global life sciences activities. Therefore, solving the life sciences data infrastructure issues described above will help ensure that RIKEN remains a global leader.

The data infrastructure can be divided roughly into a data analysis infrastructure that promotes advanced research and a database infrastructure that promotes data sharing and utilisation. The database infrastructure is intended to promote the outcomes of RIKEN externally and should work closely and efficiently with various research and analysis infrastructures. The database infrastructure is further divided into (1) the operational infrastructure of raw data (experimental resultant data without metadata) produced by each research project (direct support of each project) and (2) metadata operational infrastructure that maximises the classification of the integrated analysis and utilisation of data produced by RIKEN.

(1) Infrastructure to support for raw data operation

The problem with infrastructure for operation of raw data corresponds to 'a. Large-capacity storage improvement, speedup of calculation processing and acceleration of data transfer due to increased data volumes' described above. The RIKEN Cloud Experimental service started in April 2015 provides virtual machines with bioinformatics tools and multipurpose virtual machines via a network using a computer managed and operated by RIKEN ACCC as the foundation for the raw data infrastructure. In addition, the HOKUSAI-GreatWave, which began service in Wako in April 2015, connects three subsystems to different functions using a high-speed network to fuse experimental/simulation/data analyses, where a single computer system environment is assumed. The first subsystem is an online storage system and hierarchical storage system for large data, the second subsystem is the massively parallel computer system, and the third subsystem is the application computing server system with GPU and large memory. The online storage system, which is the core of the HOKUSAI system, has a total effective capacity of 2.2 PB, a total theoretical bandwidth of 190 GB/s, and a 7.9PB hierarchical storage management system. The massively parallel computer system has 1,080 nodes with 34,560 cores with 1PFLOPS theoretical peak performance. To realise further computational speedup, the application computing server includes 30 server nodes with four Tesla K20Xs per node and two server nodes with 1TB of memory capacity (http://www.riken.jp/pr/press/2015/20150403_1/).

The hierarchical storage system is used by HOKUSAI-GreatWave users and as a backup system for all RIKEN laboratories (HOKUSAI Data Depository Service; http://accc.riken.jp/en/data-depo/; previous D2S system). In addition, the hierarchical storage management system employs a mechanism to transfer and backup large volumes of research data quickly and securely. Furthermore, by introducing SHOBE, which has achieved the world's top greenest supercomputer as an advanced HPC system, the latest environment can be prepared promptly and is expected to further accelerate data processing.

Proper provision of such services according to the needs of RIKEN researchers is considered important direct support for handling raw data. In particular, the amount of output data already exceeds the capacity of conventional hard disk drives; thus, continuous measures are necessary. In addition, problems have already occurred when backing up large-scale data, e.g. images do not work properly when passing through the network between RIKEN branches. Furthermore, RIKEN

experienced great loss after losing all data stored at a particular branch that suffered significant damage due to a large-scale disaster. To prevent such situations, comprehensive measures, including data backup and networking, are required.

(2) Metadata infrastructure

In addition to the above raw data infrastructure, the working group believes that the importance of the metadata infrastructure will increase over time. As mentioned previously, integrated data analysis is becoming an extremely crucial component of life sciences. However, large and heterogeneous data are obstacles to data analysis. Solving this problem and realising an information infrastructure that integrates RIKEN data and openly provides an interface to data analysis programmes will be a great contribution to life sciences. RIKEN has already conducted database research relative to a semantic web technology infrastructure. Therefore, RIKEN is expected to produce well advanced research results over the next five to ten years by extending advanced data analysis techniques that make full use of the advantages of metadata in order to describe knowledge in graph structure distributed relative to RIKEN's information strategy, including big data analysis, artificial intelligence and the K-computer, which received again the world first in Graph 500. As for the infrastructure, it is necessary to develop metadata from raw life sciences data and corresponding databases, which are connected to global data and accumulate knowledge referenced by artificial intelligence systems.

A meta-database provides multifaceted solutions regarding the issues and problems mentioned in 'i. Summary of issues and problems facing life sciences in Japan'.

With respect to 'b. Promotion of innovation by open data', each laboratory in RIKEN must assess data publication itself; however, it is indispensable to provide appropriate and standardised metadata for raw data relative to innovative utilisation of open data. Appropriate operation of a single-source meta-database for RIKEN's raw data is expected to contribute efficiently to the problem.

It is expected that a meta-database will contribute directly to 'c. Description standardisation for integration and concatenation of data or metadata'. However, practical measures for appropriate and continuous standardisation are required.

With respect to 'd. Artificial intelligence approach', metadata is expected to contributes to improving the availability of RIKEN's data from artificial intelligence systems because modern metadata standardisations are performed using a machine-readable knowledge format (i.e. an ontology), and the semantic web, which is artificial intelligence technology, realises machine readability that can be used as an artificial intelligence knowledge source. Artificial intelligence technology is developing; however, adopting globally standardised technology is relatively stable and can interface with advanced research and general social contributions.

'e. Development of statistical analysis methodology for large-scale data analysis', 'f. Data visualisation and coarse visualisation' and 'g. Information extraction from images and knowledge extraction' are advanced research topics; thus, the database infrastructure does not directly contribute to these issues. Instead, it contributes to determining what metadata should be used to increase the degree of utilisation of such research results.

In addition, 'h. Direct contribution to disease, drug discovery and health promotion through integrated understanding of omics information' is an issue for each individual research field. However, a metadata infrastructure provides RIKEN's data as data sources for research fields with an interface that is comprehensive, easy to understand and accessible by artificial intelligence systems.

In addition, metadata technology, such as the semantic web, has achieved practical use. However, many types of data are considered incompatible with current semantic web technology, such as DNA sequences and highly coded data for high-speed processing. Such data will continue to be generated by life sciences technology in future. Even so, to handle data in a comprehensive way, adding metadata in some form to raw data is necessary (e.g. adding metadata to each dataset). In other words, discovering the optimum contact point between metadata conversion technology and raw data that can be used at the time is necessary and this is very important to promote data integration across the globe.

Currently, the RIKEN MetaDatabase is being operated experimentally by RIKEN ACCC. The RIKEN MetaDatabase provides practical functions and processing speed as a backend database for the Resource Description Framework (RDF), which is the basis of the semantic web. Furthermore, the RIKEN MetaDatabase has become an advanced worldwide meta-database because it displays data

intuitively using a simple tabular form, which is familiar to experimental researchers (biologists). Thus, it is easy for general life sciences researchers to publish data associated with metadata. This is why RIKEN ACCC has faithfully implemented the short-term recommendations published by the database working group.

There is concern that limits of capacity and speed will be reached within several years when sufficient metadata is added to the precise data produced by RIKEN. The RIKEN MetaDatabase is operated as the world's top knowledge extraction infrastructure because the currently distributed huge data must be introduced from both inside and outside RIKEN, such as all genome data and gene function annotations. Using semantic web technology, data can be distributed over the Internet and federated query searches are available for distributed data as a single data source. However, primarily due to performance problems, it has not achieved wide use. It is also necessary to improve search performance, realise true data distribution and reduce waste data management loads.

Thus, it is expected that the methodology and technology of metadata description (data representation) will continue to improve. In addition, it is expected that the RIKEN MetaDatabase will not become obsolete through technology development, and the development of data management technologies and efforts toward continuous operation will be necessary to ensure continued use.

4. Results

Based on the above discussion, the database working group (WG) proposes the following for the database infrastructure from a long-term perspective (five to ten years) relative to the development of information technologies, such as artificial intelligence.

1. Raw data infrastructure

1.1. For a large-capacity data infrastructure, the WG proposes that RIKEN ACCC should provide infrastructure services, such as large capacity, high-speed storage, calculation processing, data transfer processing, etc. according to the needs of RIKEN researchers. To support the explosive increase in the volumes of data, even if it is impossible to store all produced data, sufficient storage should be provided to store at least part of the data. Furthermore, to avoid the loss of data due to disasters, we recommend reinforcement of various infrastructures, such as storage and network systems, such that data can be backed up among various remote sites. In addition, a comprehensive response that allows real data and meta-databases to cooperate more closely is desired.

2. Metadata infrastructure

2.1. The working group proposes that RIKEN ACCC operate a database (meta-database) of meta-information for data produced by RIKEN based on life sciences metadata technology in order to manage metadata, which is expected to become increasingly important in future. Specifically, the working group concludes that continued operation of the RIKEN MetaDatabase is appropriate. In addition, relative to the above raw data, the working group should continue to improve capacity and speed in preparation for RIKEN's future big data and metadata requirements.

2.2. To prevent meta-database obsolescence, the WG proposes that RIKEN ACCC should promote technology development through initiatives using advanced world-class information technology. For example, the WG requests that RIKEN ACCC promotes solutions to technical research problems. From a system perspective, RIKEN ACCC should develop an advanced world-class metadata distribution system that implements the above-mentioned highly-efficient federated query to minimise metadata management costs. In addition, from a data perspective, RIKEN should address technological issues such as development of ontologies that describes RIKEN's data and a function to convert existing data to globally standardised metadata.

2.3. Regarding vocabularies to describe metadata stored in the meta-database (2.1), the WG proposes that RIKEN should establish a system to enhance research trends, including bioinformatics, data analysis and collaboration with artificial intelligence research. We urgently request correspondence regarding the following issues.

2.3.1. Toward development of mechanism with which information science and biology researchers can collaborate and promote database activities across RIKEN, the WG proposes to organise a committee or working group to discuss utilities of the database infrastructure.
In addition, it may be necessary to consider whether new subcommittees of the Bioinformatics exploratory committee or the database working group (including subcommittee organisations) should be organised.

2.3.2. Regarding realisation of the database, the WG proposes that RIKEN ACCC, the department in charge of database maintenance, should actively engage research and development to maximise data distribution by applying appropriate technologies by investigating bioinformatics research and the development trends of web technologies.

2.3.3. To improve the utility of the meta-database infrastructure, the WG proposes to establish a

system to receive advice and evaluations from database experts in Japan and overseas. For example, evaluation committees composed of researchers inside and outside RIKEN should be organised so that periodic evaluations can be performed.

2.3.4. Metadata development and collaboration among researchers both inside and outside RIKEN should be promoted to expand the use of data by applying knowledge discovery (automatic extraction, artificial intelligence, utilisation of mathematical models, deep learning, etc.) from large-scale graph data, including various content based on systematised metadata (i.e. ontologies) using research data from RIKEN and outside.

3. Other issues

The following issues, which are related to future database development at RIKEN, should be considered carefully:

d. Artificial intelligence approaches,
e. Development of statistical analysis methodology for large-scale data analysis,
f. Data visualisation and coarse visualisation, and
g. Information extraction from images and knowledge extraction.

The WG proposes to organize other working groups that consider each of the above issues separately.

A. Abbreviations

RIKEN ACCC:	Advanced Center for Computing and Communication, RIKEN
RIKEN AICS:	RIKEN Advanced Institute for Computational Science
RIKEN BRC:	RIKEN BioResource Center
RIKEN CDB:	RIKEN Center for Developmental Biology
RIKEN CGM:	RIKEN Center for Genomic Medicine
RIKEN CLST:	RIKEN Center for Life Science Technologies
RIKEN CSRS:	RIKEN Center for Sustainable Resource Science
RIKEN IMS:	RIKEN Center for Integrative Medical Sciences
RIKEN QBiC:	RIKEN Quantitative Biology Center
RIKEN RAP:	RIKEN Center for Advanced Photonics

B. Database working group

Chairperson	Dr. Hiroshi Masuya	RIKEN BRC
Members	Dr. Shuichi Onami	RIKEN QBiC
	Dr. Hideya Kawaji	RIKEN ACCC
	Dr. Itoshi Nikaido	RIKEN ACCC
	Mr. Shigeho Noda	RIKEN ACCC
Member and Secretariat	Dr. Norio Kobayashi	RIKEN ACCC