Future biology is Semantic information science

Tetsuro Toyoda, Ph.D.
Director of Bioinformatics And Systems Engineering Division, RIKEN, Japan
Nature of Life = Information

Nature of Replication
（Self-replication）
Life evolving through replications and selections

Replication of genome information (DNA replication)

Various replications (Fertilized eggs)

Evolution Cycle (About 3.8 billion years)

Natural Selection

Various groups of individuals

Genome is blueprint of Life
(Genetic Information)

Genomic information is recorded by combined DNA blocks.

DNA blocks are structured by only 4 kinds of blocks: A, T, G, and C.

To map sequences of DNA blocks is Genome sequencing.

Sequenced genomic information is represented by character information as shown below.

ATGCTAGTCGCGTGCTAGTCGATCTCGTGCA・・・・・・

Human’s genome represented by about 3 billion letters.
Biological diversity resulting from diversity of Genomic information

Informatics viewpoint

Sequences of DNA blocks evolve
↓ means
Information is evolved
(Biological diversity)

Material viewpoint

DNA of human and microorganisms are consisted of the same 4 kinds of DNA blocks (G, A, C, and T).
↓
Any living organism hasn’t evolved as a material.
(Common among living organisms)
No matter the Information, Copy and Selection is the Driving Force of evolution

Phylogenetic Tree of Open-Source Software Information

Evolution of Open-source programs undergo the same Copy and Selection data evolution process as the Genome.
Previously, Digital and Genomic Information were separate

<table>
<thead>
<tr>
<th>Digital Information</th>
<th>Genomic Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recording media for information</td>
<td>Expression of information (phenotype)</td>
</tr>
<tr>
<td>CD, DVD</td>
<td>CD, DVD</td>
</tr>
<tr>
<td>Copied CD (Copied information)</td>
<td>DNA</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Digital and Genomic Information merged with introduction of machine decoding of genomic information from DNA
“Intelligent” Cloud with DNA information

Could be a New Realm of Evolution

SciNetS
Infrastructure for Collaboration

Part 2

Bioinformatics And Systems Engineering Division, RIKEN, Japan
What is SciNetS?

SciNetS: Scientist's Networking System

Infrastructure for Collaboration

Cloud Computing and Collaboration

sinets.org

As an integrated database, **SciNetS** provides a versatile database container compatible with international standards.

Using the semantic web, **SciNetS** provides a ‘total incubation infrastructure system’ for life science-related Virtual laboratories.
Single system hosts numerous databases

Number of databases (Virtual Labs) hosted by RIKEN SciNeS is larger than the number of servers in the system.

Databases > Server computers > System

Number of databases hosted by RIKEN SciNeS is larger than the number of servers in the system.

Databases > Server computers > System

Our single system can host many DBs.

Hosting databases separately has a limitation.

Number of databases to be integrated

SciNeS.org is expected to function as a new academic medium for publishing individual databases as entire sets of research results.

Available at http://database.riken.jp

Total file size: 4.5 Terabytes
Total data files: 120 Million
Public databases: 192
  incl. DBs for mammals 23
  DBs for plants 29
  DBs for protein 17
Private databases: 190
Data records (instances): 7.5 Million
Semantic relationships: 26 Million
Databases contain meta data, a Semantic Web

“Meta data” (RDF) includes semantic links among data items and resources

“Data resources” include experimental results or statistical results

Mouse Phenotype Database

Collaborating on Ontology Basis for Classification of Databases in the Cloud
A Semantic web of Mouse databases

The RIKEN integrated database of mammals project

Instances = data recodes

YAMATO-GXO Lite project

Upper Ontology

Genetic Marker

Allele

Strain

Experimental result

Phenotype

Genetic Marker in mammal

Allele in mammal

Strain resource in mammal

Mammalian phenotype

FANTOM4 project

FANTOM gene

Gdf5

MGI project

MGI gene

Gdf5

MGI allele

Gdf5

ENU-induced allele

BRC mouse strain

Result in BRC

Phenotype result for M100451

RIKEN BRC resource project

M100451

Mammalian phenotype ontology project

sinets.org/db/mammal
InterPhenome: International Data Sharing of Mouse Phenotypes

http://www.interphenome.org/
SciNetS for Life Sciences

Part 3

Bioinformatics And Systems Engineering Division, RIKEN, Japan
for Life Sciences

SciNetS: Science, Innovation, Education, Technology, Synergy

SciNetS: Scientist's Networking System
“Superbrain” solving biological problems by simulating artificial brain understanding the knowledge of biology
Limit of general search through database

*Proposition:* “My job is a jail”

Boolean understands that job \( \neq \) jail
Open-endedness in the brain

Proposition: “My job is a jail”

Brain tries to understand beyond the framework of general concept
Developing a brain type database (Superbrain)

Large scale system which includes over 10 million documents and concepts for each neuron.

Research Result

Display the hit document by rounding up concept units such as ontology class, candidate gene, etc.

Out put Layer (web Service)

Expression layer (Document summary)

Inference layer (attractor)

Concept layer (statistical test)

Input layer (text search)

All calculations take within 1 second → Realization of high-speed web
New high speed technology calculates tens of thousands of spreadsheets

(Calculating signal processing among neurons on brain-type database)

The technology, creating 3-dimension spreadsheets for every gene (X=1, 2, ..., several tens of thousands of) at once.

<table>
<thead>
<tr>
<th>Include Gene X</th>
<th>Including search word</th>
<th>Excluding search word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include Gene X</td>
<td>“A “units of Documents”</td>
<td>“B “units of Documents”</td>
</tr>
<tr>
<td>Exclude Gene X</td>
<td>“C “units of Documents”</td>
<td>“D “units of Documents”</td>
</tr>
</tbody>
</table>

Gene name, concept units, etc. (~100 thousands)

Documentation files including literatures (~10 millions)

Questionnaires associated genes with documents.

All calculations within 1 second → Realization of high-speed web research
Search example: Inference of causative gene by positional cloning.

```
Pos Med
```

“Positional MEDLINE” assists your positional-cloning studies

Ranking results of candidate genes↓

Keyword: type 2 diabetes
Genome Condition: gene of no.6 chromosome

Searching for candidate genes by positional cloning.

Rough mapping based on crossing experiment
(~10Mbp wide interval)

Inference of candidate gene

Identification of causative gene variation

So far contributed to 65 achievements of searching researches on disease-causing genes.
SciNetS
for the Planet and Society

Part 4

Bioinformatics And Systems Engineering Division, RIKEN, Japan
SciNetS: Science for the Planet and Society

SciNetS: Science, Innovation, Education, Technology, Synergy

SciNetS: Scientist's Networking System
Sustainable society supported by Green Biotechnology

Fossil Resources

- Crude oil
- Petrochemistry process
  - Heavy Oil
  - Gas oil
  - Naphtha
  - Kerosene
  - Gasoline
- Energy
- Chemical Products
- Materials

↓ Search for resources
Development of oilfield (Fossil resources)

20th Century - Consumptive Society

Biomass

- Plants
- Syngas
- Sugar
- Oil, etc
- Bioprocess

→ Development of Innovation Technology

Applications of Renewable Resource

Creation of New Industry

↓ Search for resources
Development of genome (Biological Resource)

21st Century - Sustainable Society
~ Carbon-neutral ~
Toward an age of Rational Design using genomic data.
Creating bio resources from information

Omics data including genome and functional information

Bioinformatics infrastructure (SciNes)
Integrated database
Rational selection
Supercomputer

Selected elite genes

‘Informatization’ of life

Information world
Physical world

Genome design

DNA synthesis and materialization

Oomics research
Natural selection
3.8-billion-year evolution

Omics research

GenoCon Open platform for rational Genome design

GenoCon International Rational Genome Design Contest offers contestants the chance to compete in technologies for rational genome design using a browser-based programming environment provided at [http://genocon.org](http://genocon.org)

Optimize genome design for plants that absorb and resolve formaldehyde to prevent sick house syndrome.

Formaldehyde tolerance designed by Prof. Izui is the inaugural challenge in GenoCon

- Introduce two enzymes, extracted from bacteria, into Calvin cycle of plants

- Enhancement of formaldehyde-resistance by shepherd's-purse and Tobacco.

\[ \text{CH}_2\text{OH} \xrightarrow{\text{RuSP}} \text{CH}_2\text{OH} \xrightarrow{3\text{-Hu6P}} \text{HCHO} \xrightarrow{\text{HPS}} \text{CH}_2\text{O-P} \]

\[ \text{CH}_2\text{OH} \xrightarrow{\text{RuBP}} \text{CO} \xrightarrow{3\text{-PGA}} \text{HOCH} \xrightarrow{\text{HCOH}} \text{HCOH} \xrightarrow{\text{HCOH}} \text{CH}_2\text{O-P} \]

\[ \text{CH}_2\text{OH} \xrightarrow{\text{Hexulose 6-phosphate synthase}} \text{HOCH} \xrightarrow{\text{Hexulose 6-phosphate isomerase}} \text{HCOH} \xrightarrow{\text{HCOH}} \text{CH}_2\text{O-P} \]

AB: Gene introduction strain
CK: Control
shepherd's-purse
Tobacco

Patent applied by Prof. Izui, Kyoto/Kinki University

genocon.org
SciNetS provides a web-based programming interface for a user to process each data item in our database cloud, where all the programs and files the user has created are stored.
Creating Bio Resources from Information Resources

Info-Synthetic Biology

Open Laboratory

Closed Laboratory

Rational Genome Design

Database (RIKEN SciNeS)

Experiments in Laboratory Safety Control

Creation of Biological Resources

Short URL services such as http://bit.ly are often used in social media such as twitter. These offer similar domain mapping services by mapping external URLs under various domain names to internal URLs of a uniform domain with a short internal ID that corresponds to the external one in the repository.

Semantic-JSON upgrades the mapping service to have semantic relationships among the internal IDs that can be used to accelerate semantic communications in social media of the future.
Semantic-JSON supports most languages used by Bioinformaticians

Sample Operation in Mathematica

Load the Semantic-JSON package

Set the server’s URL

Obtain labels of two individuals

Obtain the shortest paths between two individuals

Similarly, Semantic-JSON is available in popular languages including Java, Ruby, Perl, Python.
Web-based programming interface provided by Semantic-JSON for user to process each data item in our database cloud where the user stored all the programs and files they created.

// 1. Obtain protein sequences of HPS.
var hpsSeq =SciNeS.DB.get("/sw/protein/cria214s2r214u053481")
SciNeS.write("HPS Protein sequence has been obtained from the server<br/>

// 2. Reverse translate the obtained protein sequences of HPS to DNA sequences.
var hpsDNASeq = reverseTranslation(hpsSeq);
SciNeS.write("HPS DNA Sequence has been generated<br/>

// 3. Obtain protein sequences of PHI.
var phiSeq =SciNeS.DB.get("/sw/protein/cria214s2r214u032097")
SciNeS.write("PHI Protein sequence has been obtained from the server<br/>

// 4. Reverse-translate the obtained protein sequences of PHI to DNA sequences.
var phiDNASeq = reverseTranslation(phiSeq);
SciNeS.write("PHI DNA Sequence has been generated<br/>

For more information, visit semantic-json.org.
Semantic JSON Application

Snapshot of the Plant Ontology tree viewer embedded in SciNetS.org.

When the user opens the web page the Tree structure is programmably drawn using Semantic-JSON and published into human readable content.

Ontology Browser

semantic-json.org

Semantic-JSON: API to access SemanticWeb data with JSON

Please Visit http://semantic-json.org

Current Cloud Table Applications don’t support Semantic Web, so can not be used for Life-Sciences

- Current Table applications only use simple Semantic Data, so do not harness the power of the Semantic web + Ontology

http://dabbledb.com
http://factual.com
http://socrata.com
http://google.com/fusiontables
>> Biological data formats used in life science research

- Biologists like to use the table format for their research results.
- Since most biological research results are essentially structured kinds of data, however, it is appropriate to represent these research results as Semantic Web graphs consisting of subject-property-object triples.

>> Characteristics of table data and semantic data

<table>
<thead>
<tr>
<th>Data container format</th>
<th>Table format</th>
<th>Semantic web format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td>Easy for people to browse and understand content.</td>
<td>Most research results can be represented as structured data.</td>
</tr>
<tr>
<td></td>
<td>Easy for biologists to process and edit each row of data.</td>
<td>Compatible with internet technology.</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Difficult to explicitly express semantics and metadata representing the conceptual relationships between data contained in the table.</td>
<td>Difficult for life science researchers to process data without assistance.</td>
</tr>
<tr>
<td></td>
<td>Difficult to represent tree-structured data.</td>
<td>Difficult for people to browse and understand content.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Example</strong></th>
<th>Property A</th>
<th>Property B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>Object 1-1</td>
<td>Object 1-2</td>
</tr>
<tr>
<td>Subject 2</td>
<td>Object 2-1</td>
<td>Object 2-2</td>
</tr>
</tbody>
</table>

![Semantic Web graph example](image-url)
What is the SciNetS Semantic Table?

- SciNetS SemanticTable is a data container with advantages of both tables and Semantic Web graphs:

1. It provides a table view for biological data convertible to TSV format (tab-separated-value) text table data popular among biologists for computer processing.

2. It is convertible to explicit Semantic Web graph such as RDF that a computer can read and process intelligently.

---

Table format

Semantic Web graph
This table presents implicitly the following semantic meanings, which may be understood by people, but not by computers.

<table>
<thead>
<tr>
<th>Yellow row - Table Header</th>
<th>Blue Row - Table Data</th>
<th>Green Column - Subjects</th>
<th>Red column - Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>age</td>
<td>sex</td>
<td></td>
</tr>
<tr>
<td>Michael</td>
<td>23</td>
<td>Male</td>
<td></td>
</tr>
<tr>
<td>Jennifer</td>
<td>27</td>
<td>Female</td>
<td></td>
</tr>
</tbody>
</table>

- The first row means: “The Person named Michael is 23 years of age and his gender is Male”
- The second row means: “The Person named Jennifer is 27 years of age and her gender is Female”

For Intelligent processing by a computer, the table data needs to be presented with explicit semantic data as follows in computer-readable Semantic Web graph format.
Semantic Table uses Ontology to represent Semantic Web data about the considered subject.
SemanticTable for Life Sciences

By adding Ontology on a huge scale, we can apply SemanticTable to the life sciences.

1. SciNetS page of Class RCSE Protein Data Bank
2. SciNetS page of Instance 1BCP
3. SciNetS page of Instance Homo sapiens
4. Link to the detailed page of property "organism"
5. Link to the detailed page of object class "species"
A "Table" is the most common type of data format used in biological research. The Semantic Web, on the other hand, where knowledge is represented by statements comprising a subject, predicate and object, is also becoming a major container for data accessible by automatic computational reasoners; i.e. computers.

Here we present "SemanticTable.org", a data-mart server providing a huge amount of biological semantic-web data as simple table-style data files that are reversely convertible to the semantic-web format.

The SemanticTable.org server contains **644 tables** and **7.5 million rows** of SemanticTable data as of December 2010.

Conversion from table data files to semantic-web data files and vice versa are supported by our web site (http://semantictable.org) and the World Wide Web Consortium (W3C) web site RDFa Distiller and Parser page at (http://www.w3.org/2007/08/pyRdfa/).

We propose SemanticTable, based on the standard RDFa specification, as the most understandable way for biologists not familiar with semantic-web specifications to utilize semantic-web data.
Using Semantic data could allow automatic drag and drop field rearrangement by a program or browser running on a PC.

**Semantic Table Merge Program**

**Features:**
- Merging Multiple Class Semantic Tables
- Subject Column Swapping
- Dumping to Text
Join the **Semantic** Table Consortium

- Call for developers to join the consortium to create and author such programs; the fundamental semantic groundwork is ready to use.
- Open collaboration for the standard will jumpstart the new platform.
- Open source programs for **Semantic** Table run on cloud-computing technology, as do the programs for spreadsheet tables.
- We hope the **Semantic** Table format will be distributed broadly and freely for general use, as well as for scientific use.

**CONTACT US**

Bioinformatics And Systems Engineering (BASE) division, RIKEN Yokohama Institute of Chemical and Physical Science

[SciNetS.org](http://SciNetS.org)  [SemanticTable.org](http://SemanticTable.org)

semantictable@base.riken.jp

End

Thank you very much!

Bioinformatics And Systems Engineering Division, RIKEN, Japan

www.base.riken.jp