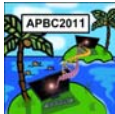


Semantically enabled and statistically supported biological hypothesis testing with tissue microarray databases

Young Soo Song^{1,2}, Chan Hee Park², Hee-Joon Chung², Hyunjung Shin¹, Jihun Kim¹ and Ju Han Kim^{2,3*}

¹Department of Industrial & Information Systems Engineering, Ajou University, Korea
²Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Korea



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Biological Database

- Collections of scientific experiments, published literatures, computational analyses
- Essential resources to biologists in their daily researches
- **Not directly answer** for the questions biologists really want to ask
 - Is my **hypothesis** true based on stored experimental data?
- Only storing of a lot of experimental data is not useful anymore.
 - **Design a system** performing hypothesis testing based on stored experimental data



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Simplified workflow of biological research

1. Hypothesis generation
2. Design of experiment
3. Performing an experiment
4. Data collection from experiment results or other sources
5. Analysis and interpretation of data
6. Evaluation of the hypothesis

If all possible experimental results were stored in DB...

1. Hypothesis generation
2. Design of experiment
3. Performing an experiment
4. Data collection from experiment results or the other sources → Data collection from DB
5. Analysis and interpretation of data
6. Evaluation of the hypothesis



APBC2011
The 9th
Asia Pacific Bioinformatics Conference



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Main processes to design a system evaluating biological hypotheses

1. Machine-readable formal representation of hypothesis and experimental data
2. Query generation
 - Based on the hypothesis and stored data
3. Statistical tests for query results
 - Based on the type of hypothesis



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

An example of hypothesis testing

- **Reduced expression of Apaf-1 in colorectal cancer** correlates with **high-grade phenotype** (Paik *et al.*, 2007).

- Search for the sample of colorectal cancer
- Make a contingency table according to Apaf-1 intensity and grade

	Strong Apaf-1	Weak Apaf-1
High grade	a	b
Low grade	c	d

- Fisher's exact test(a, b, c, d)



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Candidates for a prototype system

- The complexity of the semantics of biological hypothesis is comparable to that of natural languages.
- To reduce the complexity,
 - Experimental data should be **high throughput** one.
 - **Simply formalized** hypotheses should be tested.
 - Existence of **standard** data structure for experimental results

→ **Tissue microarray data**



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Tissue Microarray (TMA)

- **Array-based, high-throughput technology**
- Examine molecular alterations in a number of tissues on a single slide in parallel.
- **High-throughput validation tool** of the marker genes identified from DNA microarray experiments



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

TMA experimental database : Xperanto-TMA

- A web-based application for TMA experiments
- Based on an object model, **TMA-OM** (Lee *et al.*, 2006) and an exchange format, **TMA-TAB** (Song *et al.*, 2010)
- More than 100 TMA experiments are stored.



Syntax of hypothesis for TMA experiment

- **correlate(con, ind, dep)**
 - *con* : context
 - *ind* : independent entity
 - *dep* : dependent entity
- An example
 - In colon cancer, high Apaf-1 intensity is correlated with low histologic grade.
 - *correlate(colon cancer, high Apaf-1 intensity, low histologic grade)*



Syntax of hypothesis determines statistical test.

- Statistical tests for a hypothesis that can be represented as **correlate(con, ind, dep)**
 - Fisher's exact test, χ^2 test
- An example
 - In colon cancer, high Apaf-1 intensity is correlated with low histologic grade
 - Contingency table for colon cancer

	Strong Apaf-1	Weak Apaf-1
High grade	a	b
Low grade	c	d

- Fisher's exact test(a, b, c, d)



Construction of a system for hypothesis testing using TMA data (Xperanto-RDF)

- Construction of database
 - **RDF-represented TMA database**
 - Data source: Xperanto-TMA
- Hypothesis processing
 - Hypothesis editing
 - Generation of hypothesis model
 - Query generation
 - Statistical test



RDF representation of TMA data

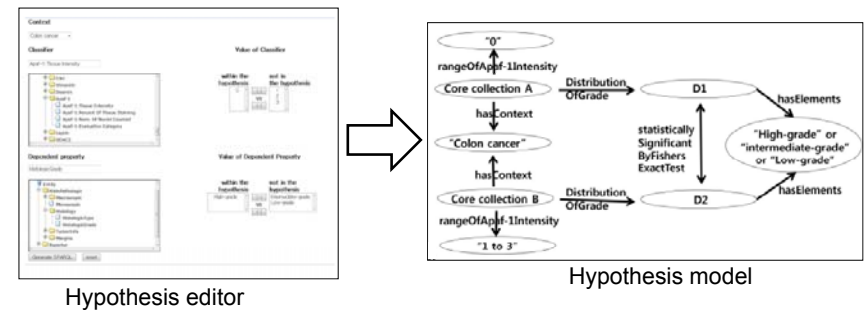
- Reasons
 - RDF-based models support **richer semantics** than RDB-based ones.
 - An entity and the relationships between entities can be **explicitly represented**.
 - Schema is implemented in the system together with data.
 - **SPARQL** as a query language is more intuitively understandable to human-beings.
 - **Integration** with the other TMA or different types of omics data was considered for the future works.
- Data from Xperanto-TMA was represented as RDF.



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Hypothesis editing and hypothesis model

- *correlate(con, ind, dep)*
 - *con* : context
 - *ind* : independent entity
 - *dep* : dependent entity



Query generation according to hypothesis model

$SPARQL_i := \text{"SELECT count(distinct ?cr) WHERE \{'' + \sum phrase_{ij} +$

$\text{"?sl xpc:Slide Block ?b. ?sl xpc:Slide_Reporter ?r.\}"}'$

$Phrase_{ij} := Factor2SPARQL_j(Factor_j, Valueset_{ij})$

$Valueset_{ij}$: if $j = 1$, the shared properties among the samples if $((i = 1, 3) \text{ and } j = 2) \text{ or}$

$((i = 1, 2) \text{ and } j = 3)$, hypothesis-describing value set for $Factor_j$,

otherwise its complementary value set

$(i = 1, 2, 3, 4, j = 1, 2, 3)$.



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Statistical test module

- Type of statistical test is determined by syntax of the input hypothesis.
- Query results are delivered to statistical test module as arguments.



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Testing of the reliability of Xperanto-RDF by experimentally proved hypotheses

Hypothesis	# of experiments	# of slides	# of samples	P value (Fisher's exact test)
Reduced expression of Apaf-1 in colorectal carcinoma correlates with high-grade phenotype. (Paik <i>et al.</i> , 2007)	3	5	55	< 0.0001
In gastric cancer, HDAC2 expression is associated with negative lymph node metastasis. (Weichert <i>et al.</i> , 2008)	3	4	52	< 0.0001
In colon cancer, the expression of Leptin is associated with negative lymph node metastasis. (Paik <i>et al.</i> , 2009)	2	3	50	< 0.0001

Interpretation of results of hypothesis testing by Xperanto-RDF

- The meaning of positive results
 - According to the TMA experiments stored in Xperanto-RDF, your hypothesis is likely to be true.
 - But we cannot decide whether your hypothesis would be still true in the real world.
 - Highly controlled experiment should be designed and performed to prove the hypothesis.
 - But as data is more **accumulated**, the difference between real world and Xperanto-RDF will be reduced.



APBC2011
The 9th
Asia Pacific Bioinformatics Conference

Availability

- Xperanto-RDF
<http://clara.snubi.org/Xperanto-RDF>
- Xperanto-TMA
<http://xperanto.snubi.org/tma>



APBC2011
The 9th
Asia Pacific Bioinformatics Conference