Extraction of functionally similar bioinformatics workflows

Junya Seo, Shigeto Senoo, Yoichi Takenaka, Hideo Matsuda
Graduate School of Information Science and Technology,
Osaka University, Toyonaka, Osaka, Japan
j-seo@ist.osaka-u.ac.jp

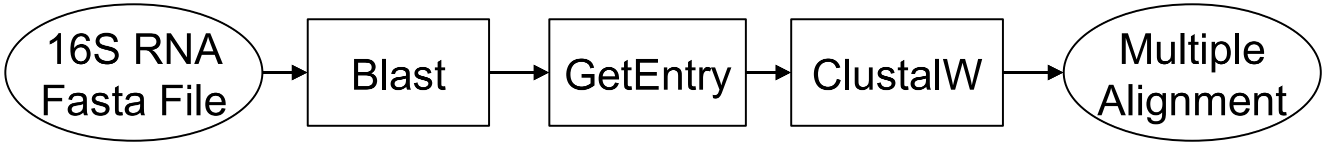
Composing workflows

- There many tools having different purpose in Bioinformatics
  - Homology search ⇒ Blast
  - Multiple sequence alignment ⇒ ClustalW

Tools are combined frequently in Bioinformatics

- There are many tools
- It is not enough to use independently

Combination of tools

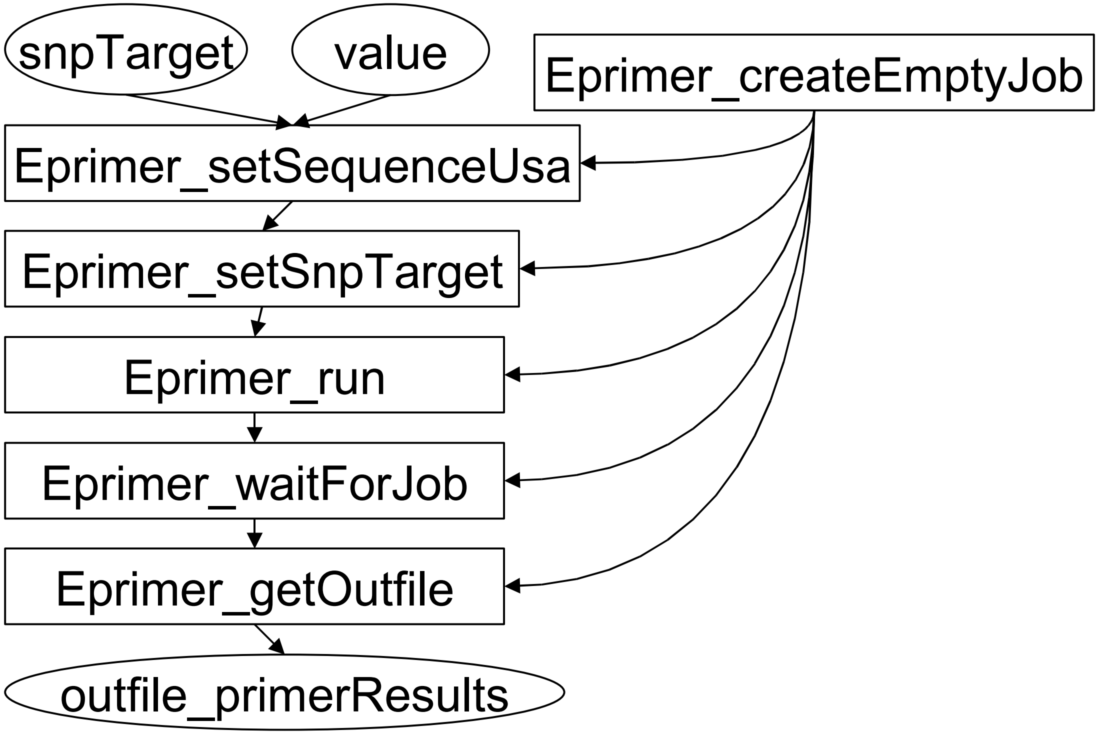


DDBJ Blast – ClustalW workflow

Workflow

Workflow is a order in which specific tool is performed

- Workflow includes…
  - Input and output data
  - Values
  - Tools
  - Data links



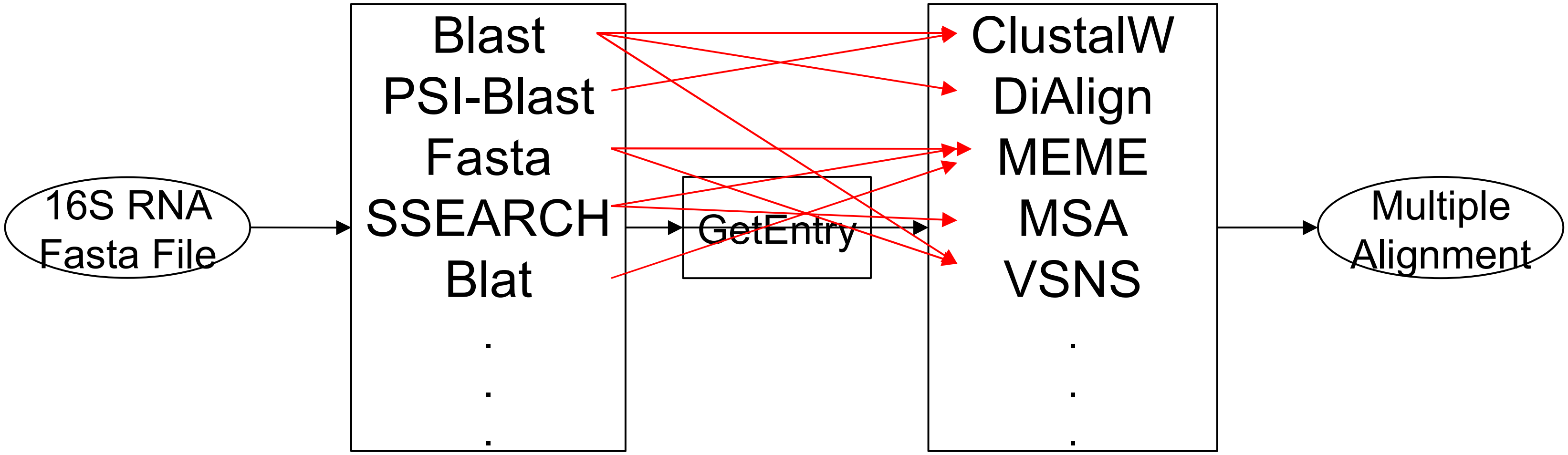
- There are some tools to make workflows
  - (e.g.) Taverna ・・・ Workflow making tool for Bioinformatics

Problems on making workflows

There are some problems on making workflows

- There are too many tools
- Which combination is better in my case?
- There are necessity of considering format between tools

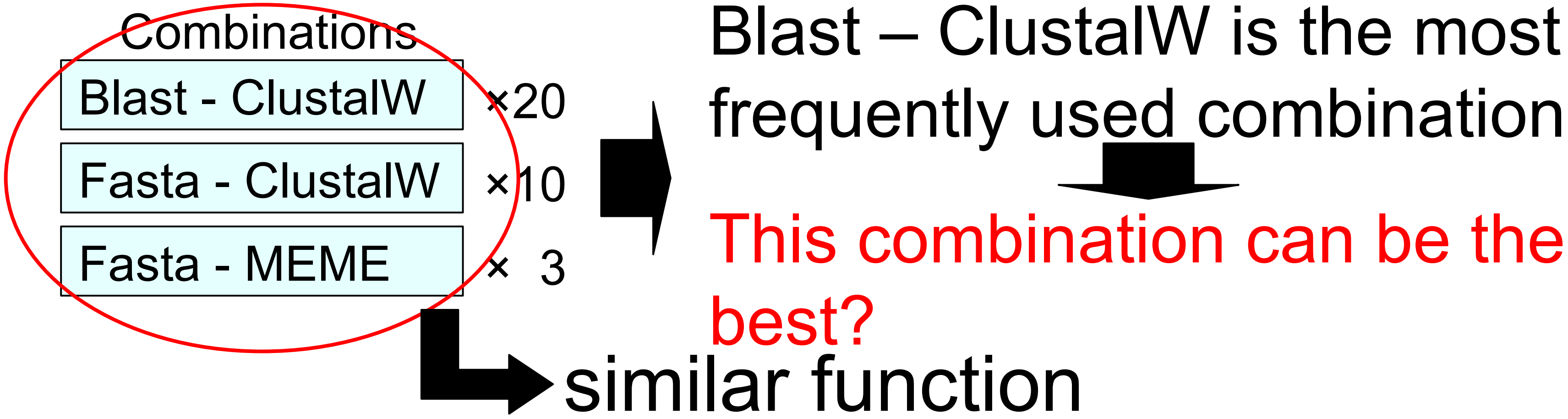
(e.g.) Homology Search and Get multiple alignment



[illegible] “GetEntry”?

To solve problems

- Refer example of past workflows
  - Combination of tools
    - Count frequency of combination



Combinations

Blast - ClustalW	×20
Fasta - ClustalW	×10
Fasta - MEME	× 3

Blast – ClustalW is the most frequently used combination

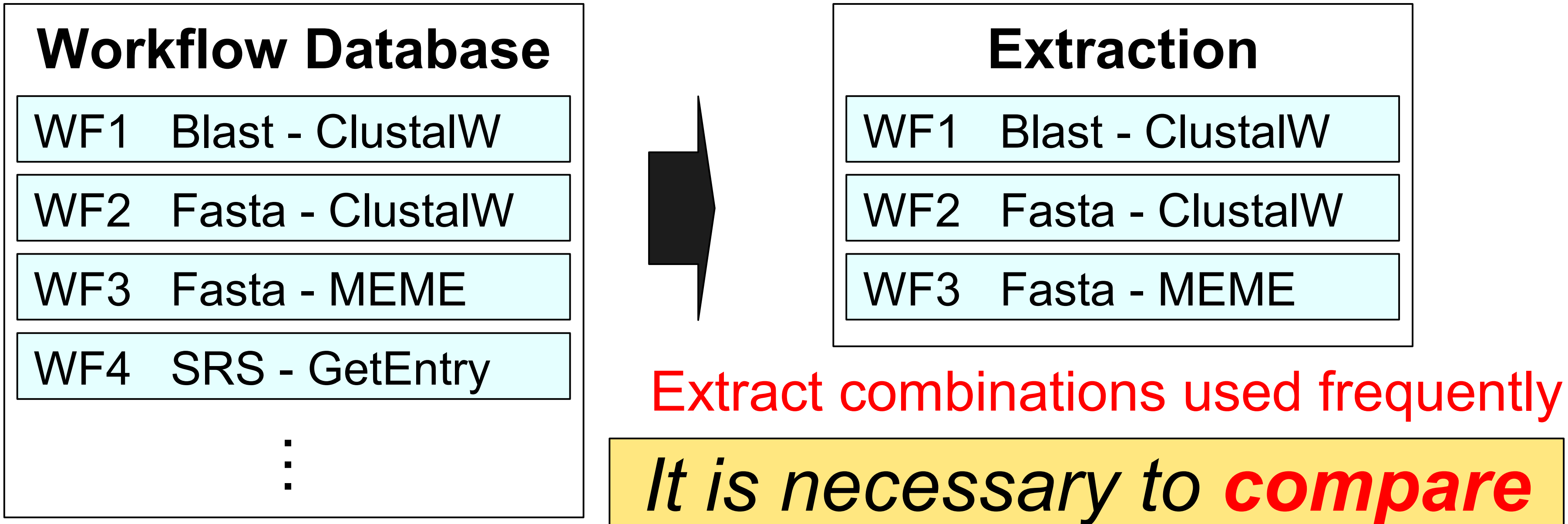
This combination can be the best?

similar function

It is necessary to extract similar workflows

To extract workflows

- Get example of combinations from workflow database



Workflow Database

WF1	Blast - ClustalW
WF2	Fasta - ClustalW
WF3	Fasta - MEME
WF4	SRS - GetEntry

⋮

Extraction

WF1	Blast - ClustalW
WF2	Fasta - ClustalW
WF3	Fasta - MEME

Extract combinations used frequently

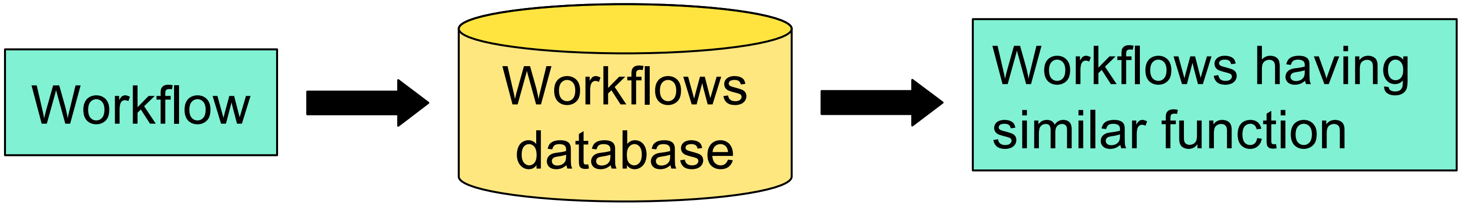
It is necessary to ***compare*** *workflows focused on functional similarity*

Extraction of functionally similar workflows

- We focused on functional similarity of workflow
- Extract functionally similar workflows
  - Extract by biological purpose (Ex. Get multiple alignment)
    - But it was difficult to associate biological purpose and workflows
  - Extract workflows functionally similar to the input
    - (Blast – ClustalW) → (Fasta – MEME),(Blat – MEME)
    - These workflows have similar function

Our Method

- Input ･･･ Workflow
  - This workflow has target function
- Output ･･･ Workflows
  - These workflows have similar function to the target
- By using input workflow, we extract workflows from database.



Functional similarity on workflow

Functional similarity is included in input data and output on workflow

Because …

A workflow is composed by some tools

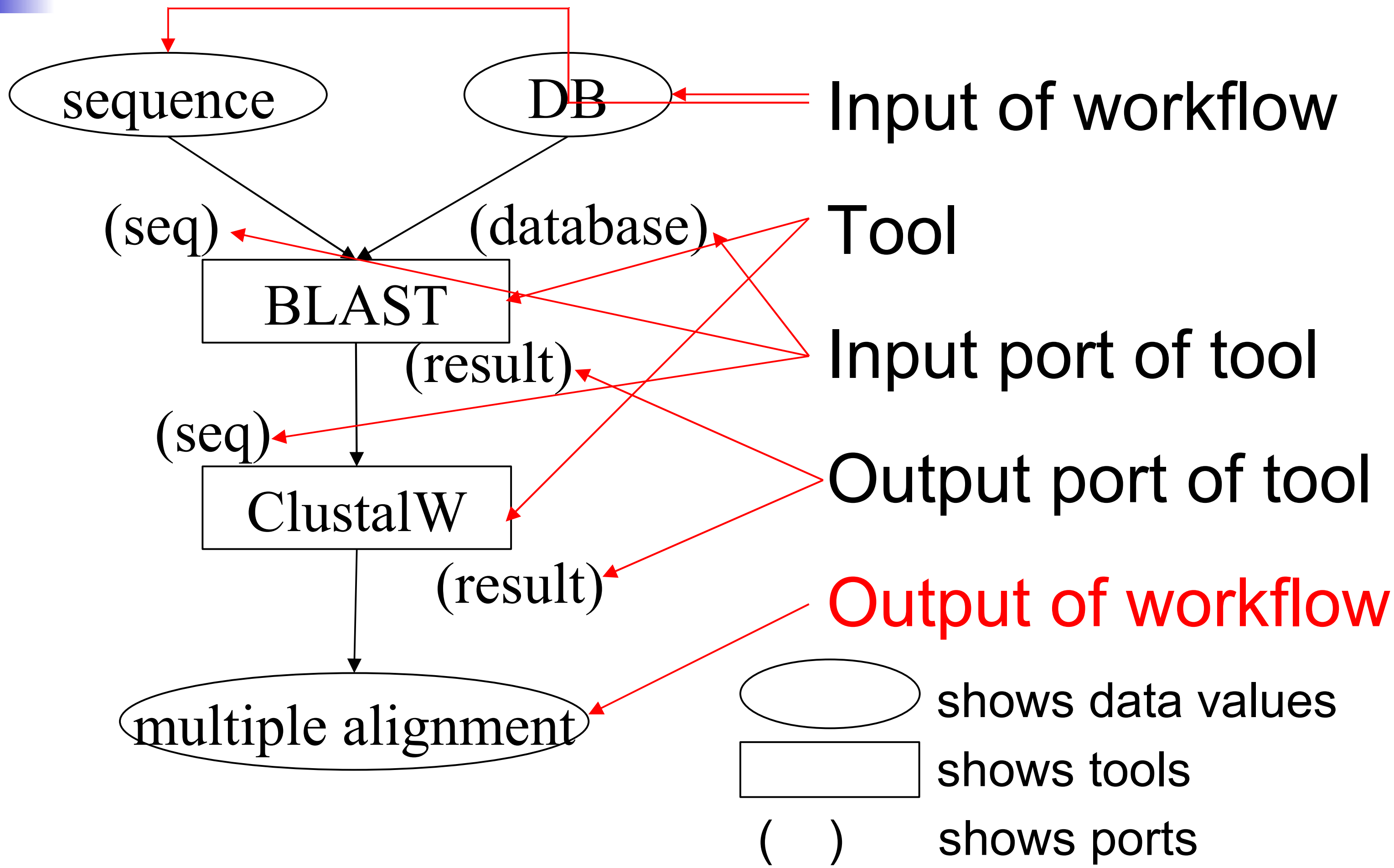
and

Each tools have input data and output data on workflow

↓

Similar workflows have similar data (input and output) each other

Details of workflow



Input of workflow

Tool

Input port of tool

Output port of tool

Output of workflow

shows data values

shows tools

( ) shows ports

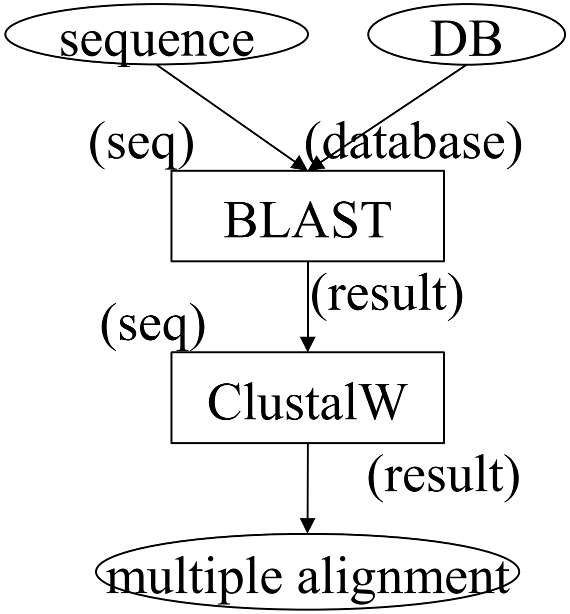
Comparing workflows

- To compare workflows
  - We used some names on workflows
    - Names of inputs and outputs
    - Names of tools
    - Port names of tools (input and output)
  - We calculate matching ratio of string

(e.g.) sequence ↔ DNAsequence

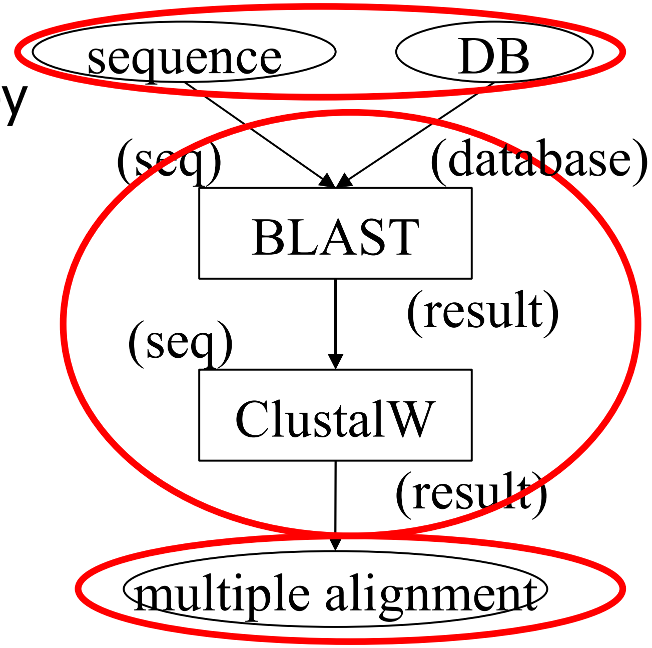
77%

    - We use this rate to narrow down candidates



Steps of extraction

- Step1:
  - Search for candidates of similar workflows by names of output or port names of output from the database
- Step2:
  - Rank the candidates of similar workflows by names of input or names of input port from Step1 result respectively
- Step3:
  - Examine these results and determine result workflows



Experiment

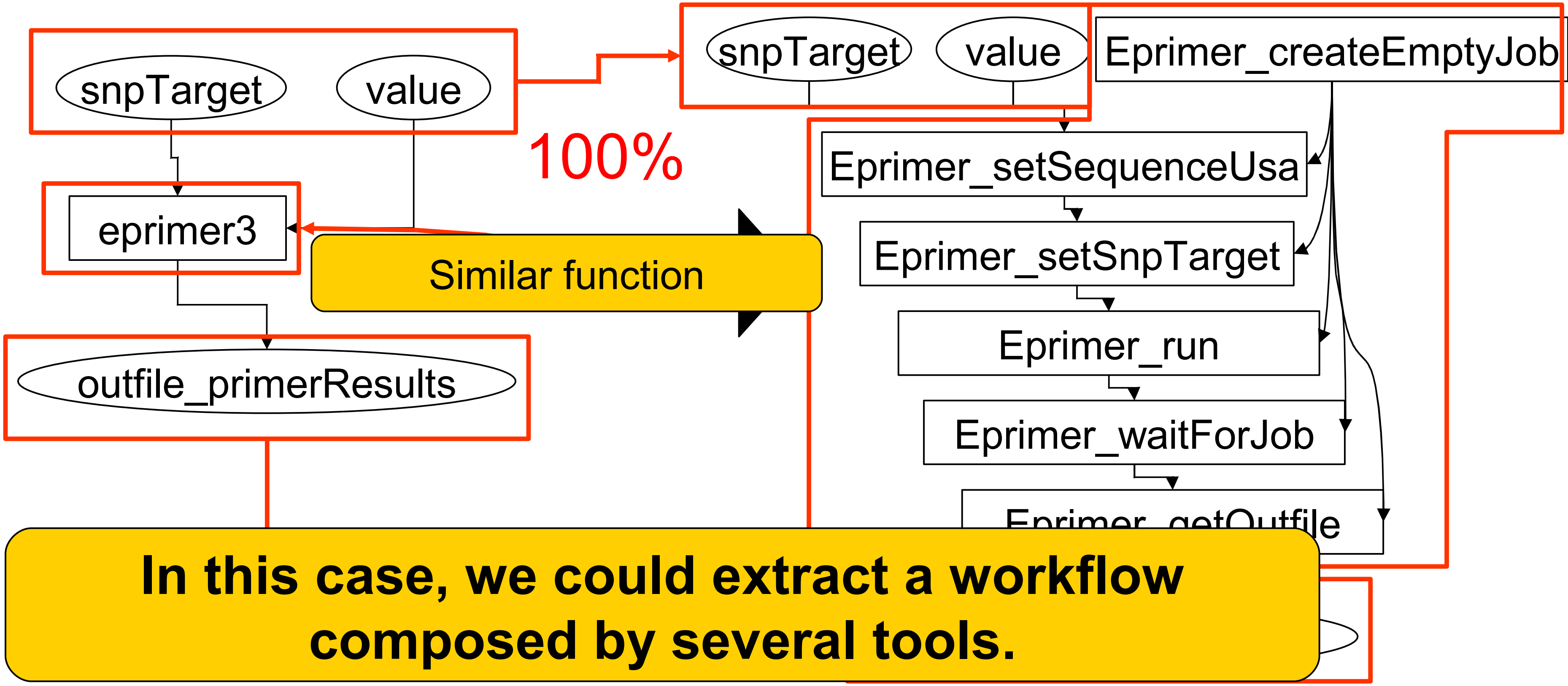
- Workflows data we used
  - 57 workflows (Taverna[1])
  - 398 tools
    - We used each workflow as the input and searched for the similar ones from the rest
  - We used Taverna workflow[1]
- Machine
  - Pentium3 700MHz
  - 256MB Main Memory
- Execution time was a few minutes.

We could extract some pairs of workflow. From the following slide, we show you two of results.

[1]Oinn, T., et al.: Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics **20**(17) (2004) 3045–3054

Result: similar workflows(1)

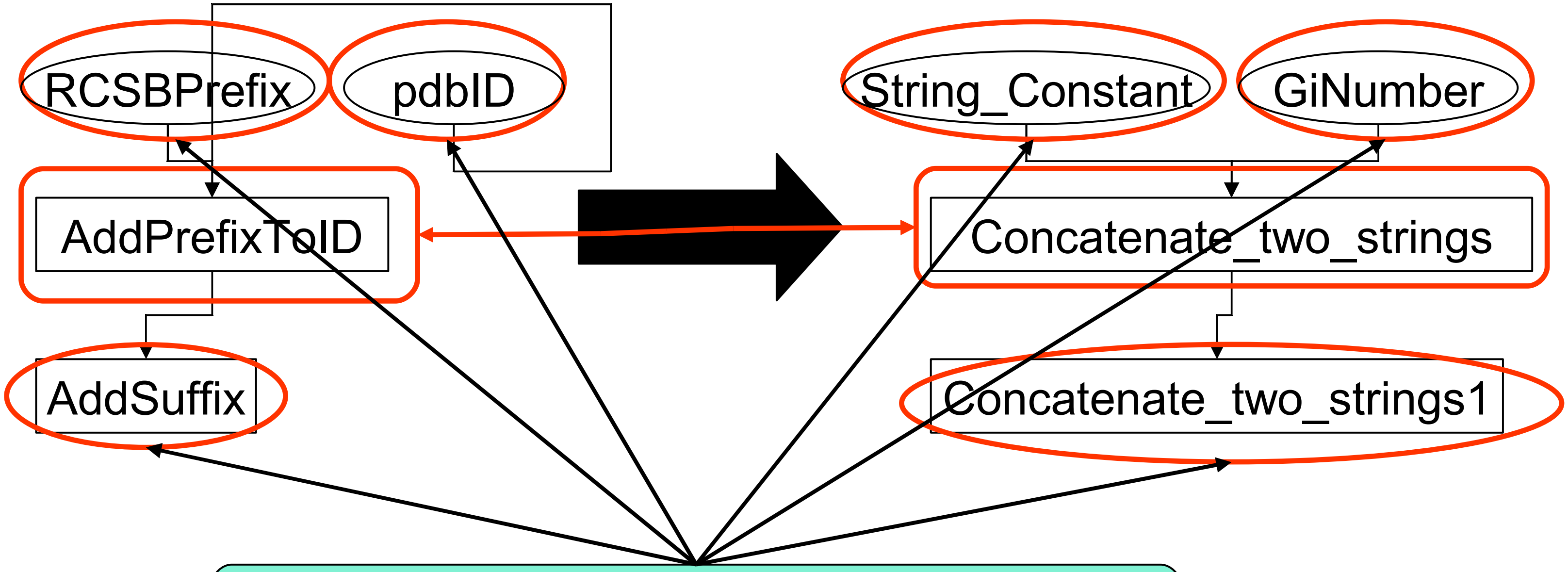
Function “pick primers and hybridization oligos for PCR reactions”



In this case, we could extract a workflow composed by several tools.

Result: similar workflows(2)

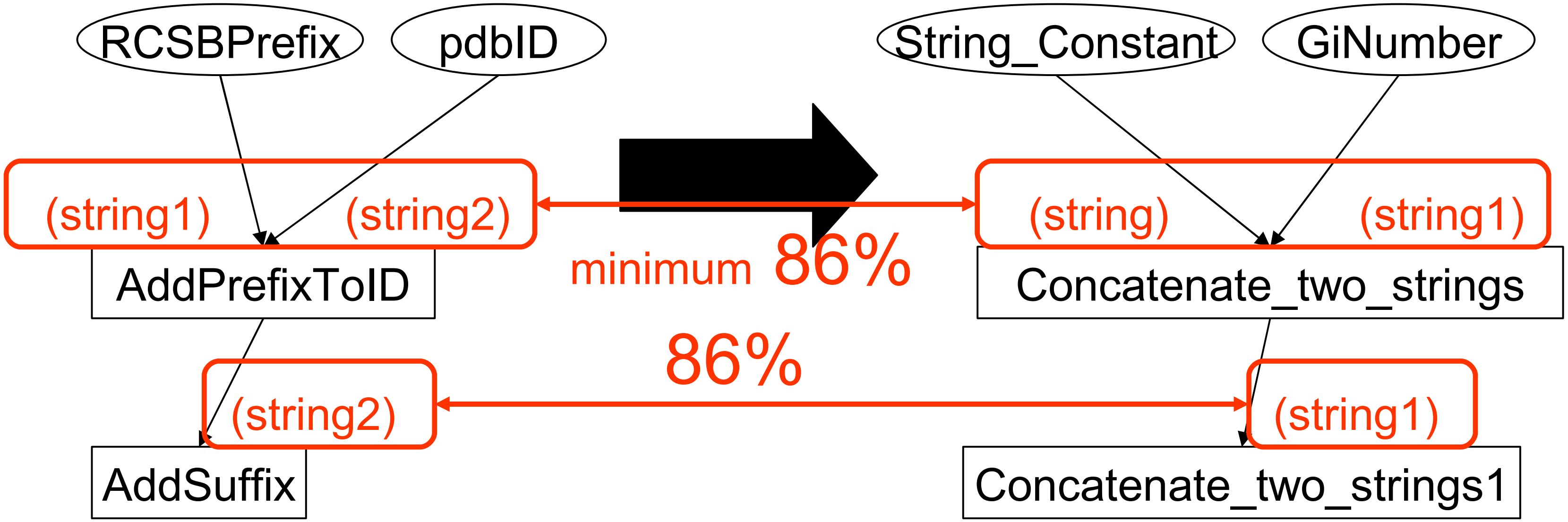
Function “Concatenate strings”



These names are not similar

Result: similar workflows(2)

Function “Concatenate strings”



In this case, we could extract workflow by the names of input port and output port

Discussion

- We showed two results
  - Our method could extract workflows having similar function
  - These were similar to the input workflow (having target function)
- We have to think about…
  - Association between biological purpose and workflows
  - Calculation of frequency (analyzing our result)

Conclusion

- We proposed a method for extraction of functionally similar bioinformatics workflow
    - Comparing and extracting workflows from the database that is similar to the query workflow.
- Future works
    - Improvement of accuracy
        - There are some pairs of workflows we couldn't extract in spite of their similarity
    - Association and calculation with workflows (described in discussion)