Extraction and Representation of *in silico* Biological Methods from the Literature

Geraint Duck Email: duckg@cs.man.ac.uk Web: http://www.cs.man.ac.uk/~duckg/

Supervisors: Robert Stevens, Goran Nenadic and David Robertson

Research Group: Text Mining (TM)

The current shift towards *in silico* biological experimentation is making deciding on the "best" method for an experiment challenging for current researchers. We hypothesise that each field can be characterised by the methods employed providing a way of exploring potential methods for a given task. For *in silico* experiments, we define databases and tools as the key elements of a method.

Introduction

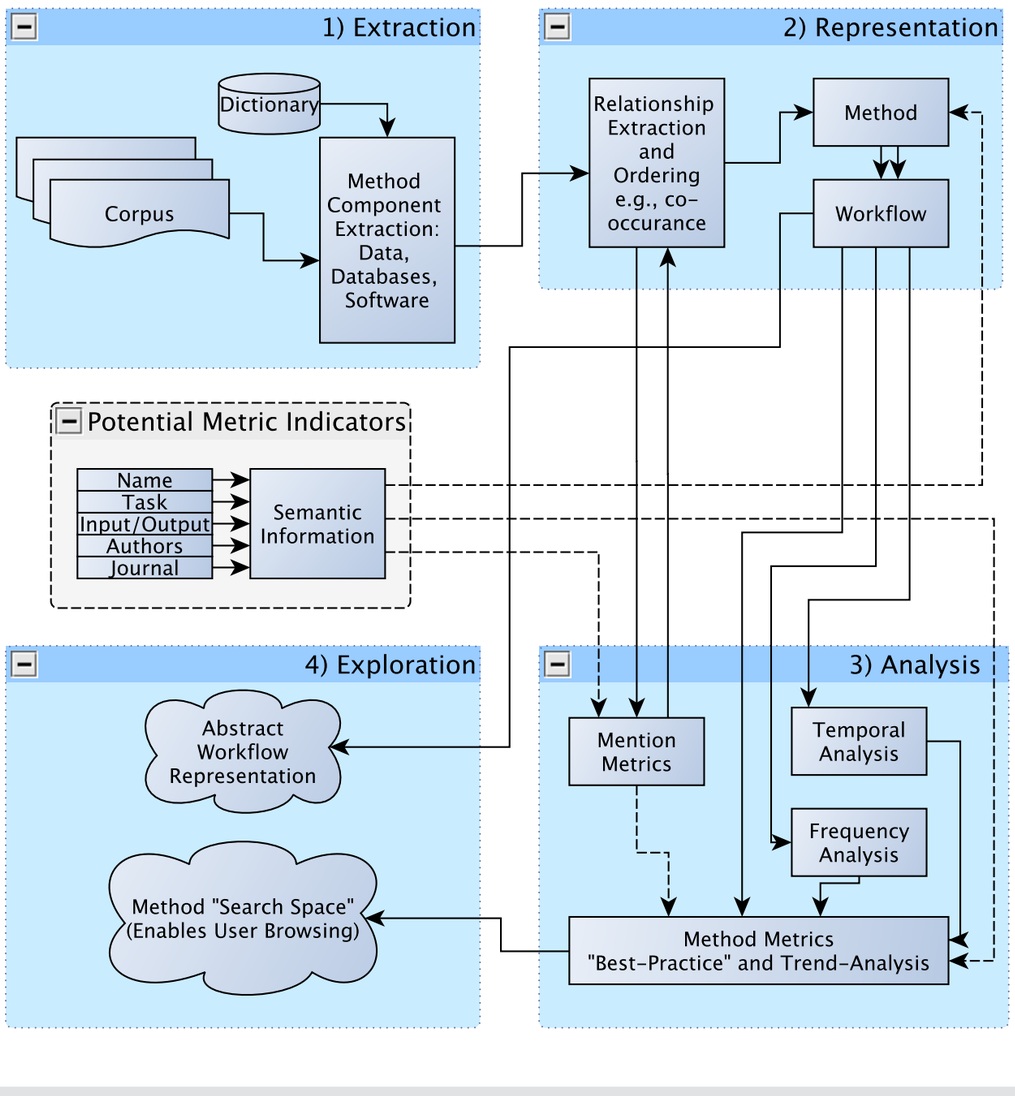
- Methods enable research and new knowledge
- They answer: What? Why? Where? How? When?
- Define "Current Approach" and "Best Practice"
- Issues of Reproducibility and Replication
- Our Focus: Bioinformatics and Computation Biology

Aim and Hypothesis

- Aim: To design, develop and evaluate a framework for literature extraction and representation of method
- Hypothesis: We can extract *in silico* biological methods from literature assisting method selection

Method

Figure 1: Method Overview



Discussion

- Preliminary analysis has highlighted some of the challenges of this task (e.g., see Figure 2)
- These challenges arise due to the volatile nature of the fields we are investigating making a solely dictionary based approach insufficient
- Once resolved, we will analyse database and tool usage throughout the literature

Acknowledgements

Acknowledgements go to my supervisors for their continued help and support, as well as to the Biotechnology and Biological Sciences Research Council (BBSRC) for providing the funding for this research.

Extraction Example

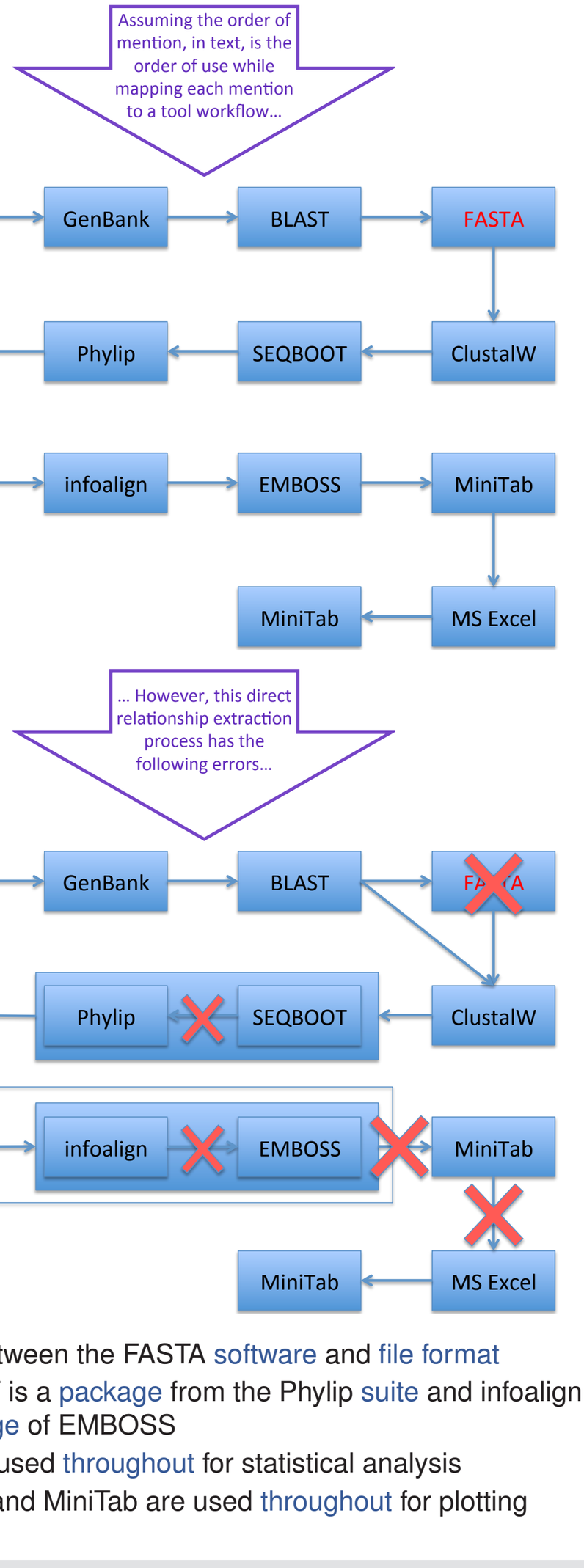
Figure 2: Example Extraction Workflow Process; Lagerström et al. (2006)

... all sequences were aligned ... using ... ***BLAT*** 3.0 ... in which case the ***GenBank*** sequence was used...

... divided ... by ***BLAST*** searches ... were combined into a *FASTA* file and aligned using ... ***ClustalW*** 1.82 ... The alignment was bootstrapped ... using ***SEQBOOT*** from the ... ***Phylip*** 3.6 package ... *[excerpt removed]*

... branch lengths were estimated in ***TreePuzzle*** using the following parameters ...

... constructed and scored automatically using a *bash-script* that utilized ***ClustalW*** as alignment engine and ***infoalign*** from the ***EMBOSS*** 2.8.0 package for scoring, ... All statistical analysis was performed using ***MiniTab***. Graphs were plotted using ***Microsoft Excel*** and ***MiniTab***.



Issues:

- Conflict between the FASTA software and file format
- SEQBOOT is a package from the Phylip suite and infoalign is a package of EMBOSS
- MiniTab is used throughout for statistical analysis
- MS Excel and MiniTab are used throughout for plotting graphs