SCALABILITY REDEFINED: A NEW WORKFLOW IN SPECTRONAUT TO ANALYZE 10’000+ RAW FILES ON A DESKTOP COMPUTER

Oliver M. Bernhardt[1], **Jakob Vowinckel**[1], **Tejas Gandhi**[1], **Lukas Reiter**[1]

1) Biognosys AG, Wagistrasse 21, 8952 Schlieren (Zurich), Switzerland

BIOGNOSYS
NEXT GENERATION PROTEOMICS

Oliver M. Bernhardt, MSc.
Principal Scientist, Bioinformatics

oliver.bernhardt@biognosys.com
www.biognosys.com

INTRODUCTION

In the recent years, the definition of what constitutes a large DIA experiment has shifted further and further. Today, experiments of thousands of LC-MS runs can be processed on a single powerful workstation within few days. However, studies with tens of thousands of runs are on the horizon. This poses a huge challenge for data analysis pipelines that want to ensure experiment wide quantification and identification quality. Here, we present a new workflow that combines the analysis quality of a stand-alone DIA experiment in Spectronaut, with unprecedented scalability. The SNE combine analysis pipeline allows for the combined analysis of 10’000+ raw files on cheap, consumer-grade desktop hardware.

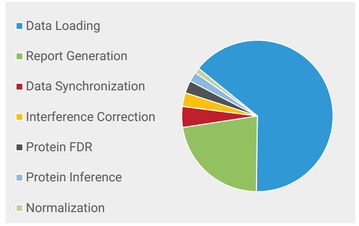


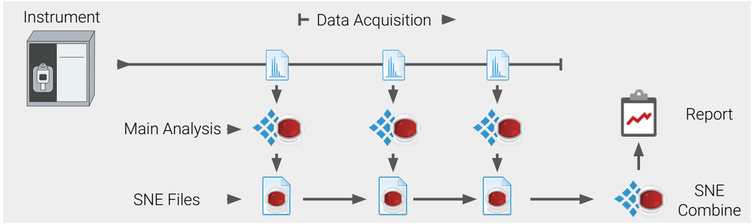
Figure 1: Time Consumption
Breakdown of the processing time for the individual tasks that make up the SNE combine pipeline. The loading of the data (SNE files) makes up over 60% of the total analysis time for this workflow.
Using faster storage media (like SSD) can reduce the time consumption of this process further.

CONCLUSIONS

- **Analyze 10’000 runs with 200’000 precursor library on a 32GB workstation**
- **Sequential analysis while acquiring data or parallel processing distributing the main workload on several computers**
- **Flexible analysis by selecting only subsets (SNE files) for intermediate results without repeating the main analysis**
- **Near endless scalability due to emphasis on minimal memory footprint**

RESULTS

Figure 2: Sequential Data Analysis
Data is analyzed in tandem with the acquisition on a single computer (e.g. LC-MS run wise). Analysis can be automated using the command line mode of Spectronaut. For the final analysis, SNE files are combined on a single, cheap desktop computer. Time to final report is significantly reduced.



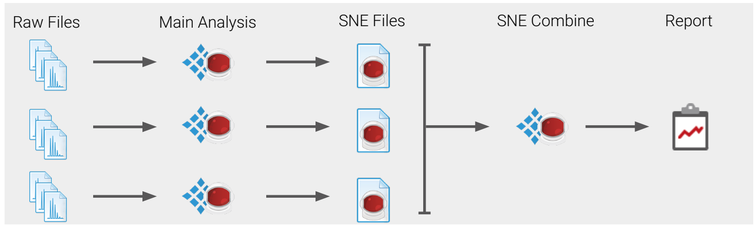


Figure 3: Parallel Data Analysis
The SNE combine pipeline can also be used to split a large experiment in smaller batches to distribute the main workload over multiple computers. For the final analysis, SNE files are combined on a single, cheap desktop computer. Time to final report is significantly reduced.

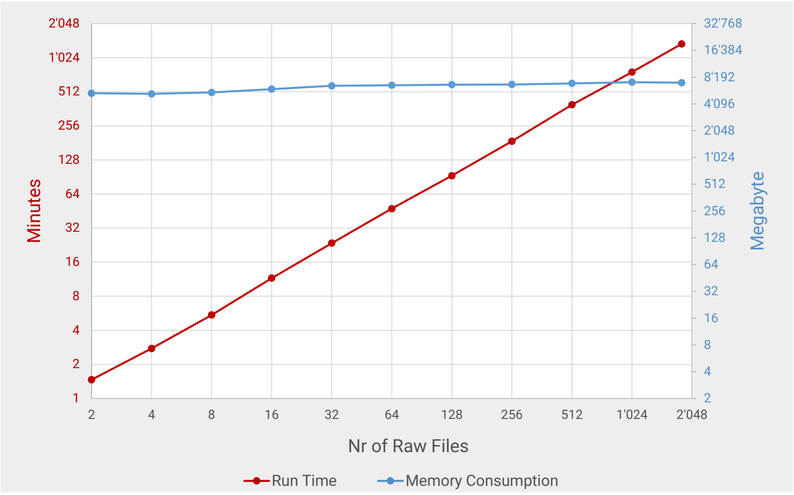


Figure 4: Scalability
Scalability of the SNE combine workflow using an experiment consisting of 1’819 raw files and a library covering ~130’000 peptides. The experiment size for the merge was consistently doubled until all 1’819 runs have been used. While run-time increases linearly with increasing the experiment size, the memory consumption shows no noticeable upwards trend even when increasing by several orders of magnitude.
The analysis took about 45 seconds per file when loading from a local hard-drive. This does not include the time for the main analysis to generate the SNE files but only the SNE combine process.