When Process Mining Meets Bioinformatics

R.P. Jagadeesh Chandra Bose[1,2] and Wil M.P. van der Aalst[1]

[1] Department of Mathematics and Computer Science, University of Technology, Eindhoven, The Netherlands
[2] Philips Healthcare, Veenpluis 5–6, Best, The Netherlands
{j.c.b.rantham.prabhakara,w.m.p.v.d.aalst}@tue.nl

Abstract. Process mining techniques can be used to extract non-trivial process related knowledge and thus generate interesting insights from event logs. Similarly, bioinformatics aims at increasing the understanding of biological processes through the analysis of information associated with biological molecules. Techniques developed in both disciplines can benefit from one another, e.g., sequence analysis is a fundamental aspect in both process mining and bioinformatics. In this paper, we draw a parallel between bioinformatics and process mining. In particular, we present some initial success stories that demonstrate that the emerging process mining discipline can benefit from techniques developed for bioinformatics.

Keywords: sequence, trace, execution patterns, diagnostics, conformance, alignment, configuration

1 Introduction

Bioinformatics aims at increasing the understanding of biological processes and entails the application of computational techniques to understand and organize the information associated with biological macromolecules [1]. Sequence analysis or sequence informatics is a core aspect of bioinformatics that is concerned with the analysis of DNA/protein sequences[3] and has been an active area of research for over four decades.

Process mining is a relatively young research discipline aimed at discovering, monitoring and improving real processes by extracting knowledge from event logs readily available in today's information systems [2]. Business processes leave trails in a variety of data sources (e.g., audit trails, databases, transaction logs). Hence, every process instance can be described by a trace, i.e., a sequence of events. Process mining techniques are able to extract knowledge from such traces and provide a welcome extension to the repertoire of business process analysis techniques. The topics in process mining can be broadly classified into three

[3] DNA stores information in the form of the base nucleotide sequence, which is a string of four letters (A, T, G and C) while protein sequences are sequences defined over twenty amino acids and are the fundamental determinants of biological structure and function.

categories (i) *discovery*, (ii) *conformance*, and (iii) *enhancement*. Process discovery deals with the discovery of models from event logs. For example, there are dozens of techniques that automatically construct process models (e.g., Petri nets or BPMN models) from event logs [2]. Discovery is not restricted to control-flow; one may also discover organizational models, etc. Conformance deals with comparing an apriori model with the observed behavior as recorded in the log and aims at detecting inconsistencies/deviations between a process model and its corresponding execution log. In other words, it checks for any violation between *what was expected to happen* and *what actually happened*. Enhancement deals with extending or improving an existing model based on information about the process execution in an event log. For example, annotating a process model with performance data to show bottlenecks, throughput times etc. Some of the challenges in process mining include the discovery of process maps (navigable hierarchical process models) and the provision of process diagnostics support for auditors and analysts [3].

It is important to note that, to a large extent, sequence analysis is a fundamental aspect in almost all facets of process mining and bioinformatics. In spite of all the peculiarities specific to business processes and process mining, the relatively young field of process mining should, in our view, take account of the conceptual foundations, practical experiences, and analysis tools developed by sequence informatics researchers over the last couple of decades. In this paper, we describe some of the analogies between problems studied in both disciplines. We present some initial successes which demonstrate that process mining techniques can benefit from such a cross-fertilization.

2 Notations

We use the following notations in this paper.

- Let Σ denote the set of activities. Σ^+ is the set of all non-empty finite sequences of activities from Σ.
- A trace corresponds to a process instance expressed as a finite sequence of activities. $T \in \Sigma^+$ is a trace over Σ. $|T|$ denotes the length of the trace T.
- The ordered sequence of activities in T is denoted as $T(1)T(2)T(3)\ldots T(n)$ where $T(k)$ represents the k^{th} activity in the trace.
- An event log, $\mathcal{L}$, corresponds to a multi-set (or bag) of traces from Σ^+.

3 From Sequence to Structure

A DNA *sequence motif* is defined as a nucleic acid *sequence pattern* that has some biological significance (both structural and functional) [4]. These motifs are usually found to recur in different genes or within a single gene. For example, *tandem repeats* (tandemly repeating DNA) are associated with various regulatory mechanisms such as protein binding [5]. More often than not, sequence motifs

are also associated with *structural motifs* found in proteins thus establishing a strong correspondence between sequence and structure.

Likewise, common subsequences of activities in an event log that are found to recur within a process instance or across process instances have some domain (functional) significance. In [6], we adopted the sequence patterns (e.g., tandem repeats, maximal repeats etc.) proposed in the bioinformatics literature, correlated them to commonly used process model constructs (e.g., tandem repeats and tandem arrays correspond to simple loop constructs) and proposed a means to form abstractions over these patterns. Using these abstractions as a basis, we proposed a *two-phase approach to process discovery* [7]. The first phase comprises of pre-processing the event log with abstractions at a desired level of granularity and the second phase deals with discovering the *process maps* with seamless zoom-in/out facility. Figure 1 summarizes the overall approach.

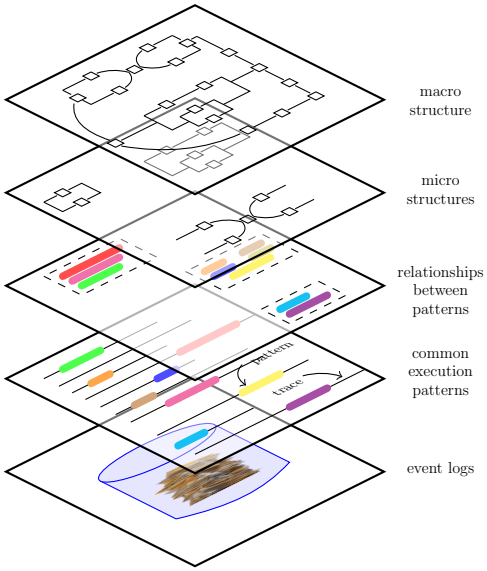

Fig. 1. Repeating subsequences of activities define the common execution patterns and carry some domain (functional) significance. Related patterns and activities pertaining to these patterns define abstractions that correspond to micro-structures (or sub-processes). The top-level process model can be viewed as a macro-structure that subsumes the micro-structures.

Figure 2 highlights the difference between the traditional approach to process discovery and the two-phase approach. Note that the process model (map) discovered using the two-phase approach is simpler. Our approach supports the abstraction of activities based on their context and type, and provides a seamless zoom-in and zoom-out functionality.

Thus the bringing together of concepts in bioinformatics to process mining has enabled the discovery of hierarchical process models and opened a new perspective in dealing with fine granular event logs.

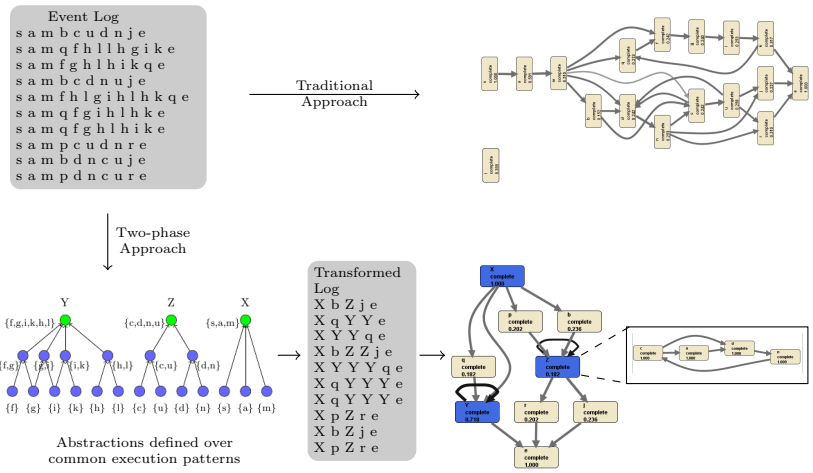

Fig. 2. Traditional approach vs. our two-phase approach to process discovery

4 Sequence Alignment and Process Diagnostics

Multiple sequence alignment has been a subject of extensive research in computational biology for over three decades. Sequence alignment is an essential tool in bioinformatics that assists in unraveling the secondary and tertiary structures of proteins and molecules, their evolution and functions, and in inferring the taxonomic, phylogenetic or cladistic relationships between organisms, diagnoses of genetic diseases etc [8, 9].

In [10], we have adapted sequence alignment to traces in an event log and showed that it carries significant promise in process diagnostics. The goal of *trace alignment* is to align traces in such a way that event logs can be easily explored. Given a set of traces $\mathbb{T} = \{T_1, T_2, \ldots, T_n\}$, trace alignment can be defined as a mapping of $\mathbb{T}$ to another set of traces $\overline{\mathbb{T}} = \{\overline{T_1}, \overline{T_2}, \ldots, \overline{T_n}\}$ where $\overline{T_i} \in (\Sigma \cup \{-\})^+$ for $1 \leq i \leq n$. In addition, the following three properties need to be satisfied with respect to $\mathbb{T}$ and $\overline{\mathbb{T}}$: (a) each trace in $\overline{\mathbb{T}}$ is of the same length i.e., there exists an $m \in \mathbb{N}$ such that $|\overline{T_1}| = |\overline{T_2}| = \cdots = |\overline{T_n}| = m$ (b) $\overline{T_i}$ is equal to T_i after removing all gap symbols '$-$' and (c) there is no $k \in \{1, \ldots, m\}$ such that $\forall_{1 \leq i \leq n}\ \overline{T_i}(k) = -$.

Trace alignment can be used to explore the process in the early stages of analysis and to answer specific questions in later stages of analysis. More specifically, trace alignment can assist in answering questions such as:

- What is the most common (likely) process behavior that is executed?
- Where do my process instances deviate and what do they have in common?
- Are there any common patterns of execution in my traces?
- What are the contexts in which an activity or a set of activities is executed in my event log?

- What are the process instances that share/capture a desired behavior either exactly or approximately?
- Are there particular patterns (e.g., milestones, concurrent activities etc.) in my process?

Figure 3 depicts the results of trace alignment for a real-life log from a rental agency. The figure shows that trace alignment can assist in answering a variety of diagnostic questions. Every row corresponds to a process instance and time increases from left to right. The horizonal position is based on *logical time* rather than real timestamps. If two rows have the same activity name in the same column, then the corresponding two events are very similar and are therefore aligned. Note that the same activity can appear in multiple columns. By reading a row from left to right, we can see the sequence of activities (i.e., the trace) that was executed for a process instance. Process instances having the same trace can be grouped into one row to simplify the diagram. The challenge is to find an alignment that is as simple and informative as possible. For example, the number of columns and gaps should be minimized while having as much consensus as possible per column.

The application of sequence alignment in bioinformatics to process mining has created an altogether new dimension to conformance checking; *deviations and violations are uncovered by analyzing just the raw event traces* (thereby avoiding the need for process models).

Finding good quality alignments is notoriously complex. The initial results of trace alignment are definitely encouraging. Nonetheless, there are various new challenges when adopting biological sequence alignment to trace alignment in the context of business processes [11]. For example, biological sequences tend to be homogenous whereas traces in semi-structured processes (e.g., care processes in hospitals) tend to be much more variable. Other differences are the fact that traces in an event log can be of very different lengths (e.g., due to loops) and may be the result of concurrency. These characteristics provide new challenges for sequence alignment.

5 Phylogeny and Process Configuration

Phylogenetics refers to the study of evolutionary relationships, and is one of the first applications in bioinformatics. A phylogeny is a tree representation of the evolutionary history of a set (family) of organisms, gene/protein sequences etc. The basic premise in phylogenetics is that genes have evolved by duplication and divergence from common ancestors [12]. The genes can therefore exist in a nested hierarchy of relatedness.

In the past couple of years, *process configuration* has gained prominence in the BPM community [13]. Process configuration is primarily concerned with managing families of business processes that are similar to one another in many ways yet differing in some other ways. For example, processes within different municipalities are very similar in many aspects and differ in some other aspects. Such discrepancies can arise due to characteristics peculiar to each municipality

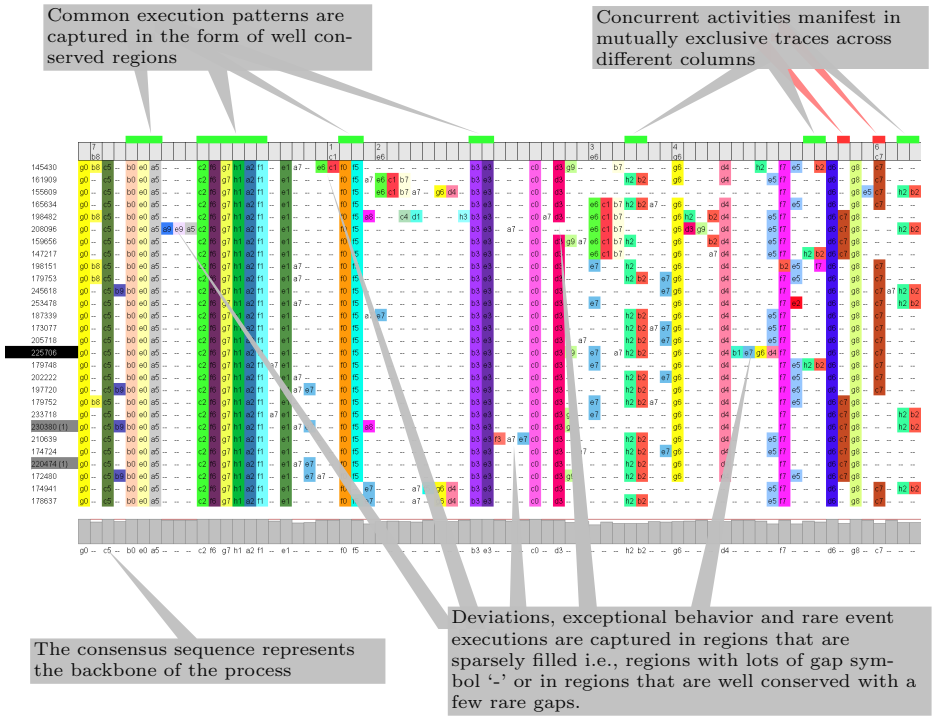

Fig. 3. An example of trace alignment for a real-life log from a rental agency. Each row refers to a process instance. Columns describe positions in traces. Consider now the cell in row y and column x. If the cell contains an activity name a, then a occurred for case y at position x. If the cell contains no activity name (i.e., a gap "–"), then nothing happened for y at position x.

(e.g., differences in size, demographics, problems, and policies) that need to be maintained. Furthermore, operational processes need to change to adapt to changing circumstances, e.g., new legislation, extreme variations in supply and demand, seasonal effects, etc. A configurable process model describes a family of similar process models in a given domain [13], and can be thought of as the genesis (root) of the family. All variants in the family can be derived from the configurable model through a series of change patterns [14]. One of the core research problems in *process configuration* is to automatically derive configurable process models from specific models and event logs.

One can find stark similarity between phylogenetics and process configuration. Techniques have been proposed in the bioinformatics literature to discover phylogenies both from (protein) structure as well as from sequences. This can be compared to deriving configurable process models from specific models and from event logs respectively. The adaptability of phylogeny construction techniques to process configuration needs to be explored.

Techniques from bioinformatics have also been adopted to trace clustering in process mining [15, 16]. Sequence clustering techniques have been applied to deal with unlabeled event logs[4] in process mining [17]. Experiences from bioinformatics can also contribute to tooling and infrastructure efforts in process mining. For example, visualization is one of the challenging problems in process mining tooling[5]. A lot of current visualization means in process mining become unmanageable when dealing with large event logs thereby compromising the comprehensibility. *Visualization* is used in many areas within bioinformatics (e.g., sequence matching, genome browsing, multiple sequence alignment etc.), with varying success, and good tools already exist. As another example, to cater to the rapidly increasing accumulation of biological data, lots of efforts had been initiated in bioinformatics to create advanced databases with analysis capabilities devoted to particular categories e.g., Genbank (cataloguing DNA data), SWISS-PROT/TrEMBL (repository of protein sequences) etc. Recently, similar efforts had been initiated in the process modeling and process mining community to create repositories with advanced support for dealing with process model collections e.g., APROMORE [18]. Such an overlap between the goals combined with the promising initial results calls for a more rigorous attempt at understanding and exploiting the synergy between these two disciplines.

6 Conclusions

Bioinformatics and process mining share some common goals. In this paper, we presented the commonalities between the problems and techniques studied in bioinformatics and process mining. Exploiting these commonalities, we demonstrated that process mining can benefit from the plethora of techniques developed in bioinformatics. Initial attempts at such a crossover have enabled the discovery of hierarchical process models and helped extending the scope of conformance checking to also cover the direct inspection of traces. Although this is just a first step towards an interaction between the two disciplines, the results are very promising and the relationship will be explored further in our future work.

Acknowledgments The authors are grateful to Philips Healthcare for funding the research in process mining.

References

1. Luscombe, N., Greenbaum, D., Gerstein, M.: What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine **40**(4) (2001) 346–358

[4] In an unlabeled event log, the case to which an event belongs to is unknown.

[5] ProM is an extensible framework that provides a comprehensive set of tools/plugins for the discovery and analysis of process models from event logs. See http://www.processmining.org for more information and to download ProM.

2. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
3. van der Aalst, W.M.P.: Challenges in Business Process Mining. Technical Report BPM-10-01, Business Process Management (BPM) Center (2010)
4. Das, M.K., Dai, H.K.: A Survey of DNA Motif Finding Algorithms. BMC Bioinformatics **8**(Suppl 7) (2007) S21
5. Kolpakov, R., Bana, G., Kucherov, G.: mreps: Efficient and Flexible Detection of Tandem Repeats in DNA. Nucleic Acids Research **31**(13) (2003) 3672–3678
6. Bose, R.P.J.C., van der Aalst, W.M.P.: Abstractions in Process Mining: A Taxonomy of Patterns. In Dayal, U., Eder, J., Koehler, J., Reijers, H., eds.: Business Process Management. Volume 5701 of LNCS., Springer-Verlag (2009) 159–175
7. Li, J., Bose, R.P.J.C., van der Aalst, W.M.P.: Mining Context-Dependent and Interactive Business Process Maps using Execution Patterns. In zur Muehlen, M., Su, J., eds.: BPM 2010 Workshops. Volume 66 of LNBIP., Springer-Verlag (2011) 109–121
8. Chan, S., Wong, A.K.C., Chiu, D.: A Survey of Multiple Sequence Comparison Methods. Bulletin of Mathematical Biology **54**(4) (1992) 563–598
9. Gotoh, O.: Multiple Sequence Alignment: Algorithms and Applications. Advanced Biophysics **36** (1999) 159–206
10. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace Alignment in Process Mining: Opportunities for Process Diagnostics. In Hull, R., Mendling, J., Tai, S., eds.: Proceedings of the 8th International Conference on Business Process Management (BPM). Volume 6336 of LNCS., Springer-Verlag (2010) 227–242
11. Notredame, C.: Recent Progress in Multiple Sequence Alignment: A Survey. Pharmacogenomics **3** (2002) 131–144
12. Thornton, J.W., DeSalle, R.: Gene Family Evolution and Homology: Genomics Meets Phylogenetics. Annual Review of Genomics and Human Genetics **1**(1) (2000) 41–73
13. van der Aalst, W.M.P., Lohmann, N., Rosa, M.L., Xu, J.: Correctness Ensuring Process Configuration: An Approach Based on Partner Synthesis. In Hull, R., Mendling, J., Tai, S., eds.: Proceedings of the 8th International Conference on Business Process Management (BPM). Volume 6336 of LNCS., Springer-Verlag (2010) 95–111
14. Weber, B., Rinderle, S., Reichert, M.: Change Patterns and Change Support Features in Process-Aware Information Systems. In: Proceedings of the 19th International Conference on Advanced Information Systems Engineering (CAiSE), Springer-Verlag (2007) 574–588
15. Bose, R.P.J.C., van der Aalst, W.M.P.: Context Aware Trace Clustering: Towards Improving Process Mining Results. In: Proceedings of the SIAM International Conference on Data Mining (SDM). (2009) 401–412
16. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models. In: Business Process Management Workshops. Volume 43 of LNBIP., Springer (2010) 170–181
17. Ferreira, D., Zacarias, M., Malheiros, M., Ferreira, P.: Approaching Process Mining with Sequence Clustering: Experiments and Findings. In: Proceedings of the 5th International Conference on Business Process Management (BPM). Volume 4714 of LNCS., Springer (2007) 360–374
18. Rosa, M.L., Reijers, H.A., van der Aalst, W.M.P., Dijkman, R.M., Mendling, J., Dumas, M., Garcia-Banuelos, L.: APROMORE: An Advanced Process Model Repository. Expert Systems with Applications **38**(6) (2011) 7029–7040