A preliminary study on Business Process-aware Large Language Models

Mario Luca Bernardi[2,†], Angelo Casciani[1,*,†], Marta Cimitile[3,†] and Andrea Marrella[1,†]

[2]Department of Engineering, University of Sannio, Piazza Roma 21, Benevento, 82100, Italy
[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, Rome, 00185, Italy
[3]Department of Law and Digital Society, UnitelmaSapienza, Piazza Sassari, Rome, 00185, Italy

Abstract

AI-Augmented Business Process Management Systems (ABPMSs) are innovative information systems with increased flexibility, autonomy, and conversational capability. These systems can be boosted by Large Language Models (LLMs), renowned for their ability to handle natural language processing tasks. Nevertheless, no significant empirical validations exist about their usefulness in process-driven decision support. In this study, we propose a business process-oriented LLM framework, for enacting actionable conversations with workers involved in a business process, leveraging Retrieval-Augmented Generation (RAG) to enrich process-specific knowledge. The methodology has been assessed to evaluate its capacity to produce precise responses to inquiries posed by users within a public administration context. The preliminary study shows the framework's ability to identify specific activities and sequence flows within the targeted process model, thereby providing valuable insights into its potential for improving ABPMSs.

Keywords

Business Process, Decision Support Systems, Large Language Models, Retrieval-Augmented Generation

1. Introduction

AI-Augmented Business Process Management Systems (ABPMSs) embody new human-centered information systems distinguished by significant flexibility, autonomy, and extensive conversational and self-enhancement abilities. [1]. Thus, Artificial Intelligence (AI) expands conventional *process-aware Decision Support Systems* (DSS) to facilitate prompt and effective decision-making by elucidating the underlying factors influencing the decisions [2]. Integrating ABPMSs into human workflows may introduce shifts in workforce dynamics, potentially leading to a lack of trust [3]. One possible remedy for this challenge is the incorporation of Conversational Systems (CSs). The emergence of CSs presents a promising avenue for enhancing Business Process Management (BPM) initiatives, significantly empowering ABPMSs [4, 5]. The adoption of *Large Language Models* (LLMs) could push substantial advancements in these systems [5]. LLMs represent an emerging class of machine learning models showcasing great performance in accomplishing various Natural Language Processing (NLP) tasks [6]. Thanks to their huge advantages, practitioners are progressively utilizing LLMs across various domains, gaining significant benefits for industries and business operations while reshaping the dynamics of human interaction with management systems [7]. Notably, LLMs have been transforming several organizations towards the paradigm of *autonomous* enterprise and enable ABPMSs to hold a central position in assisting human activities and decisions across the system life cycle. Indeed, starting from business processes, LLMs should transcend local reasoning contexts, support the management of diverse scenarios, and enhance the business activities understanding [7]. In front of the recognized potentiality of LLMs to assist human decisions in the business landscape [1], this topic is few explored in literature [7] and, as far as we are aware, an empirical validation regarding the efficacy of LLMs for process-aware decision support is missing. In this research context, our work presents an innovative methodology for business process analysis leveraging the usage of LLMs to develop a conversational process-aware DSS. We propose to adopt a process-aware *Retrieval-Augmented Generation* (RAG) [8] framework to extend process- and domain-specific knowledge, in the direction of improving the conversational capability of a LLM to respond to business process-related inquiries. The overall system supports the user in a wide range of process comprehension and execution tasks using natural language. Our work evaluates the proficiency of the methodology in producing precise and contextually appropriate responses to process-related questions within different settings. In particular, we investigate the effi-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
bernardi@unisannio.it (M. L. Bernardi); angelo.casciani@uniroma1.it (A. Casciani); marta.cimitile@unitelasapienza.it (M. Cimitile); andrea.marrella@uniroma1.it (A. Marrella)
0000-0002-3223-7032 (M. L. Bernardi); 0009-0003-7843-8045 (A. Casciani); 0000-0003-2403-8313 (M. Cimitile); 0000-0002-1031-0374 (A. Marrella)

cacy of the approach in a real-world scenario within the realm of public administration.

2. Related Work

As asserted in [4], the integration of CSs holds significant potential for enhancing ABPMSs. Numerous methodologies have emerged in recent years directed at leveraging the capabilities of CSs to enhance various critical areas within BPM [5].

In the sub-field of *Descriptive Process Analytics*, describing current business processes and identifying problems and potential improvements, NLP and neural architectures, proved their effectiveness in extracting process models from natural language descriptions [9, 10, 11]. Conversely, expressing business process models in natural language aids human comprehension [12, 13]. Moreover, conversational interfaces further enhance understanding and accessibility of process mining findings [14, 15].

Predictive Process Analytics concerns building predictive models to forecast the future state and performance of business processes. Specifically, current trends in this area are centered around the development of conversational interfaces to assist the *what-if analysis* of *digital process twins* [16, 17] and *predictive process monitoring* [18, 19, 20].

Prescriptive Process Optimization primarily focuses on improving processes, often by translating insights into actionable steps aimed at enhancing process execution. CSs designed for this BPM area mainly support *automated process optimization*, suggesting adjustments to optimize process performance across various indicators [21, 22]. Additionally, these systems contribute to *prescriptive process monitoring*, providing real-time recommendations for actions to be taken, as illustrated in [23].

Augmented Process Execution embodies the concept wherein system-driven management actively oversees business process execution, with human operators providing support as needed. In this sub-field, various conversational agents have been developed to facilitate seamless interaction between systems and human users [24, 25, 26]. Furthermore, Robotic Process Automation (RPA), which involves creating software robots to automate repetitive tasks on application user interfaces, will likely benefit from the combination with CSs. Such integration enables the automation of business processes [27, 28, 29], and aids in identifying suitable routines for automation through natural language interaction [30, 31].

3. The Business Process LLM

In this study, we present a business process-oriented LLM framework, better detailed in [32]. The steps utilized for answering queries pertaining to business processes are summarized in Figure 1. The overall architecture comprises two major phases: *Knowledge Augmentation* and *Querying*.

Knowledge Augmentation The process-aware LLM pipeline starts by considering a business process model in input, resulting in the production of multiple chunks. This operation is undertaken to facilitate the LLM's understanding in generating responses. In this study, we utilized a *Directly-Follows Graph* (DFG) representation expressed in natural language.

In fact, *chunking* aims to partition broad textual content into more manageable segments, enabling the LLM to ingest only relevant context and overcoming limitations imposed by its context window. To ensure meaningful chunks and mitigate unnatural segmentation of the process model, two distinct chunking strategies were evaluated: *fixed-size* and *recursive* chunking.

Subsequently, the framework proceeds to transform the raw input chunks into model *embeddings* for storage in a vector index. These embeddings are dense, low-dimensional vectors designed to encapsulate semantic information and contextual relationships necessary for the successive retrieval and generation operations.

Afterward, the business process model embeddings are stored within a specialized vector database to enable efficient retrieval. This retrieval procedure is enacted through semantic search that, in our case, relies on *cosine similarity*.

Querying The *Querying* stage begins with the retrieval of the pertinent process model chunks needed for the crafting of precise responses to the process-related questions. In particular, this retrieval step involves fetching relevant process chunks from the vector store through semantic search utilizing cosine similarity. Following this, these segments, along with the user question, are fed into an LLM to generate an answer.

Ultimately, to offer contextually grounded answers based on the user query and the retrieved information, the proposed framework relies on two primary components: a LLM and its associated tokenizer. Initially, a prompt is formulated by merging the user query with the previously retrieved process context. Subsequently, the tokenizer converts the prompt into a format comprehensible by the model. Eventually, the prompt is fed to the LLM to generate contextually relevant answers. In particular, our process-aware approach integrates the *Llama 2 13B* [33] model as the LLM.

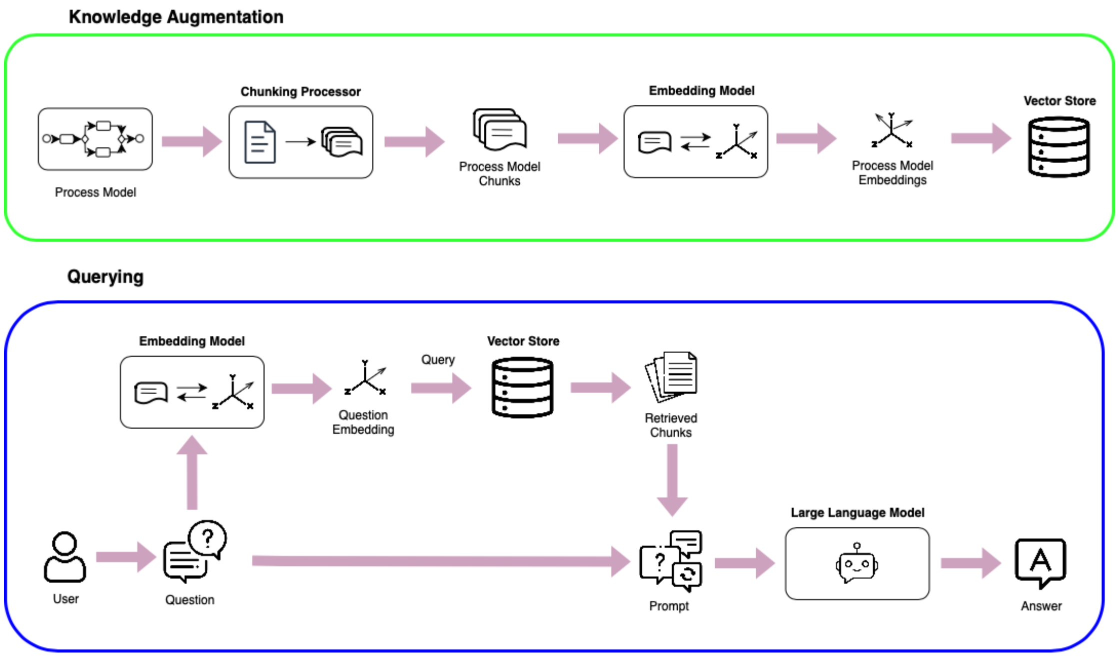

Figure 1: The business process-oriented LLM framework.

4. Evaluation

We performed a preliminary validation on the adoption of the proposed framework by applying it to a real public administration procedure. The process model, illustrated in Figure 2, involves the reimbursement of expenses for missions, a critical procedure within a university. This administrative process entails the processing of expense reports submitted by employees and the subsequent decision to either reimburse or reject these reports. In particular, the process was analyzed using textual DFG descriptions of activities and sequence flows.

The proposed framework, being rooted in generative models, provides feedback to users in natural language. To assess its effectiveness in aiding users' comprehension of business processes, the validation encompasses assessing the accuracy of the answers concerning the entities and relationships present in both the process model and the response of the LLM. The conclusion derived from this research effort centers on evaluating the approach's overall effectiveness in assisting business process users and discussing its potential applications in real-world scenarios.

4.1. Evaluation Setting

All the evaluations are performed using the reimbursement process model previously introduced, represented using the DFG expressed in natural language.

The queries adopted in this evaluation require, to be answered, to recognize both *structural* and *behavioral* information within the model. By *structural* information, it is considered the presence of activities, events, and gateways in the process model whereas *behavioral* information encompasses details concerning the sequence flows linking these entities.

Specifically, for *structural* information correctness analysis, we queried the presence of specific activities within the business process model, prompting the pipeline to answer with a simple "yes" or "no" and to provide relevant contextual references if available.

When assessing *behavioral* features, inquiries were expressed to check the presence of sequence flows between specified activities in the process representation. The LLM was prompted to state their existence in a binary manner, reporting contextual references.

Striving to obtain a thorough evaluation, we analyzed all single-pass transitions, an equivalent number of sequence flows between activities present in the model but not directly connected, and the same number of flows linking tasks that do not belong to the process.

First, we assessed the performance of the RAG-based framework in comparison to the basic version of the language model for responding to the queries within the context of the reimbursement process model.

Specifically, we estimated the capability of the LLaMA

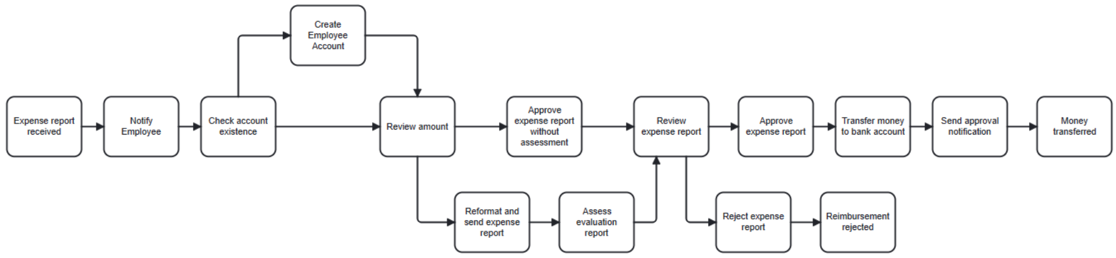

Figure 2: The DFG model of the reimbursement process in a university.

2 13B model and the RAG-based pipeline in addressing related to business processes, employing *accuracy* as the measure.

For this reason, we designed an evaluation approach for assessing the performance of the framework relying on binary response questions (expecting either a "yes" or a "no" as allowed answers) to allow a rigorous assessment of the provided answers. The *accuracy* quantifies the proportion of exact predictions generated by the LLM in answering the user's questions out of the total responses provided. We classify predictions given by the framework as *true positives* (TP) when they correspond to positive expected outcomes and as *true negatives* (TN) when they match negative expected outcomes. Vice versa, *false positives* (FP) arise when the approach produces positive answers opposite to negative expectations, whereas *false negatives* (FN) derive from negative answers generated by the framework despite positive expected ones.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Subsequently, we estimated the effects of employing various chunking techniques within the process-aware LLM pipeline, alongside investigating how prompt engineering can further augment the framework's performance. Fixed-size and recursive chunking with different sizes are tested.

In both cases, the accuracy (reported in Formula 1) of the framework in answering the queries is evaluated.

We carried out this evaluation employing an *oracle* that considers both the query and the corresponding binary response as input. Such oracle compares the answers of the pipeline with the expected ones and computes the accuracy as the ratio of correct predictions to the total number of tests conducted in that particular assessment.

In our experimentation, we found that by retrieving the top 20 chunks, we were always able to capture a comprehensive overview of the process model, enabling the language model to generate grounded responses.

Table 1
Accuracy obtained for basic LLM and RAG framework.

Methodology	Representation	Accuracy
Basic LLM	None	40.18%
RAG-based framework	Natural Language DFG	**72.37%**

The experiments were conducted on a workstation running the Linux/Ubuntu 22.04.3 LTS operating system and equipped with an NVIDIA A100 GPU.

4.2. Evaluation results

We proceed to analyze the results obtained during the evaluation phase under various experimental conditions.

The results in terms of accuracy for the basic LLM and the RAG-based pipeline on the reimbursement DFG model described in natural language are presented in Table 1.

The table demonstrates a notable improvement in accuracy upon utilizing the RAG-based LLM, which is consistent with our expectations for the test. This enhancement exhibits an acceptable performance level (*72.37 percent*) for the framework, relying on the natural language representation to drive more informed and accurate decision-making.

Our observations revealed instances of hallucination, wherein the pure LLM would provide responses despite lacking pertinent information about the process model, occasionally asserting familiarity with certain activities even when such knowledge was absent.

Table 2 illustrates the accuracy computed using various chunking methods, including no chunking, fixed-size chunking, and recursive chunking.

Comparable outcomes are achieved through the usage of a fixed-size strategy and a recursive technique

Table 2
Accuracy obtained using different chunking strategies.

Chunking	Accuracy
No Chunking	79.52%
Fixed	81.58%
Recursive	**82.89%**

for chunking leveraging the natural language representation. In both cases, the ideal size for the chunks is identified as 128 tokens with a 10-token overlap. We can attribute this observation to the relatively modest scale of the process model, which causes its content to be nearly encapsulated within a single chunk. Additionally, the above consideration clarifies why the absence of chunking yields analogous results.

5. Conclusion

In conclusion, this work introduced a *business process-aware LLM*, an innovative framework designed to facilitate actionable conversations and support process-aware DSSs, thereby laying the ground for intelligent interaction with ABPMSs. The proposed methodology, tailored for aiding business process analysis, aims to enhance the conversational skills of LLMs in the business process context. This objective is realized through the development of a RAG-based architecture, which extends its knowledge of the structural and behavioral aspects of process models by ingesting contextual information concerning specific inquiries. Consequently, the process-aware framework is equipped to assist users in understanding and executing business processes through a natural language interface. Additionally, we assessed the performance of the process-aware LLM in providing precise and pertinent answers to the queries posed by the users across diverse evaluation scenarios.

In future research within the domain of process discovery [34], we intend to delve into the analysis of the business process execution information and explore the impact of different embedding models on the developed technique. Furthermore, investigating the integration of the framework with symbolic AI solvers to embed reasoning capabilities could present another intriguing avenue for future work.

Acknowledgments

The work of Angelo Casciani has been carried out in the range of the Italian National Doctorate on AI run by Sapienza. The work of Andrea Marrella is partially supported by the Sapienza project FOND-AIBPM and the PNRR MUR project PE0000013-FAIR.

References

[1] M. Dumas, F. Fournier, L. Limonad, A. Marrella, M. Montali, et al., AI-augmented Business Process Management Systems: A Research Manifesto, ACM Trans. Manage. Inf. Syst. 14 (2023). doi:10.1145/3576047.

[2] P. Agarwal, B. Gao, S. Huo, P. Reddy, et al., A Process-Aware Decision Support System for Business Processes, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, 2022, p. 2673–2681. doi:10.1145/3534678.3539088.

[3] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (2004).

[4] D. Chapela-Campa, M. Dumas, From process mining to augmented process execution, Software and Systems Modeling (2023) 1–10.

[5] A. Casciani, M. L. Bernardi, M. Cimitile, A. Marrella, Conversational Systems for AI-Augmented Business Process Management, in: Proceedings of the 18th Research Challenges in Information Science (RCIS 2024), 2024, pp. 1–16.

[6] I. Ozkaya, Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications, IEEE Software 40 (2023) 4–8. doi:10.1109/MS.2023.3248401.

[7] D. Fahland, F. Fournier, L. Limonad, I. Skarbovsky, et al., How well can large language models explain business processes?, 2024. arXiv:2401.12846.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, 2020, pp. 1–16.

[9] K. Sintoris, K. Vergidis, Extracting business process models using natural language processing (NLP) techniques, in: Proceedings - 2017 IEEE 19th Conference on Business Informatics, CBI 2017, volume 1, Institute of Electrical and Electronics Engineers Inc., 2017, p. 135 – 139.

[10] H. van der Aa, K. J. Balder, F. M. Maggi, A. Nolte, Say it in your own words: Defining declarative process models using speech recognition, BPM Forum (2020).

[11] C. Qian, L. Wen, A. Kumar, BEPT: A behavior-based process translator for interpreting and understanding process models, Int. Conf. on Information and Knowledge Management, Proceedings (2019).

[12] L. Ackermann, S. SchöNig, et al., Natural language generation for declarative process models, CAiSE Workshops LNBIP 231 (2015) 3 – 19.

[13] Y. Fontenla-Seco, M. Lama, A. Bugarín, Process-To-Text: A Framework for the Quantitative Description of Processes in Natural Language, Trustworthy

AI - Integrating Learning, Optimization and Reasoning Workshop (2020).
[14] L. Barbieri, E. Madeira, K. Stroeh, W. van der Aalst, A natural language querying interface for process mining, Journal of Intelligent Information Systems 61 (2023) 113 – 142.
[15] H. Yeo, E. Khorasani, V. Sheinin, I. Manotas, N. P. A. Vo, O. Popescu, P. Zerfos, Natural Language Interface for Process Mining Queries in Healthcare, in: Proceedings - 2022 IEEE Int. Conf. on Big Data, Big Data 2022, Institute of Electrical and Electronics Engineers Inc., 2022, p. 4443 – 4452.
[16] D. Barón-Espitia, M. Dumas, O. González-Rojas, Coral: Conversational What-If Process Analysis, in: ICPM Demo, 2022, pp. 118–122.
[17] M. Li, R. Wang, X. Zhou, Z. Zhu, Y. Wen, R. Tan, ChatTwin: Toward Automated Digital Twin Generation for Data Center via Large Language Models, in: Proceedings of the 10th ACM Int. Conf. on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '23, Association for Computing Machinery, 2023, p. 208–211.
[18] K. Brennig, K. Benkert, B. Löhr, O. Müller, Text-Aware Predictive Process Monitoring of Knowledge-Intensive Processes: Does Control Flow Matter?, in: Int. Conf. on Business Process Management, Springer, 2023, pp. 440–452.
[19] L. Cabrera, S. Weinzierl, S. Zilker, M. Matzner, Text-Aware Predictive Process Monitoring with Contextualized Word Embeddings, in: BPM Workshops, volume 460 LNBIP, 2022, p. 303 – 314.
[20] C. Warmuth, H. Leopold, On the Potential of Textual Data for Explainable Predictive Process Monitoring, ICPM Workshops 468 LNBIP (2023) 190 – 202.
[21] S. Badini, S. Regondi, E. Frontoni, R. Pugliese, Assessing the capabilities of ChatGPT to improve additive manufacturing troubleshooting, Advanced Industrial and Engineering Polymer Research (2023).
[22] A. Mustansir, K. Shahzad, M. K. Malik, Towards automatic business process redesign: an NLP based approach to extract redesign suggestions, Automated Software Engineering 29 (2022).
[23] S. Zeltyn, S. Shlomov, et al., Prescriptive Process Monitoring in Intelligent Process Automation with Chatbot Orchestration, in: PMAI@IJCAI, 2022, pp. 49–60.
[24] T. Chakraborti, S. Agarwal, Y. Khazaeni, et al., D3BA: A Tool for Optimizing Business Processes Using Non-deterministic Planning, in: BPM Workshops, 2020, pp. 181–193.
[25] L. F. Lins, G. Melo, T. Oliveira, et al., PACAs: Process-Aware Conversational Agents, in: BPM Workshops, 2021, pp. 312–318.
[26] J. Bandlamudi, K. Mukherjee, P. Agarwal, et al., Towards Hybrid Automation by Bootstrapping Conversational Interfaces for IT Operation Tasks, in: AAAI, 2023, pp. 15654–15660.
[27] P. D. Hung, D. T. Trang, T. Khai, Integrating Chatbot and RPA into Enterprise Applications Based on Open, Flexible and Extensible Platforms, in: Int. Conf. on Cooperative Design, Visualization and Engineering, 2021, pp. 183–194.
[28] G. Dan, D. Claudiu, F. Alexandra, et al., Multi-Channel Chatbot and Robotic Process Automation, in: IEEE Int. Conf. on Automation, Quality and Testing, Robotics, 2022, pp. 1–6.
[29] Y. Rizk, V. Isahagian, S. Boag, et al., A Conversational Digital Assistant for Intelligent Process Automation, BPM Forum (2020).
[30] H. van der Aa, H. Leopold, Automatically identifying process automation candidates using natural language processing, Blockchain and Robotic Process Automation (2022) 77 – 86.
[31] Z. Zeng, W. Watson, N. Cho, et al., FlowMind: Automatic Workflow Generation with LLMs, in: ACM Int. Conf. on AI in Finance, 2023, pp. 73–81.
[32] A. Casciani, M. L. Bernardi, M. Cimitile, A. Marrella, Conversational systems for ai-augmented business process management, PREPRINT (Version 1) available at Research Square (2024). doi:https://doi.org/10.21203/rs.3.rs-4125790/v1.
[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, et al., LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971.
[34] M. L. Bernardi, M. Cimitile, C. Di Francescomarino, F. M. Maggi, Using discriminative rule mining to discover declarative process models with non-atomic activities, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8620 LNCS (2014) 281 – 295. doi:10.1007/978-3-319-09870-8_21.