Logic programs to infer computational models of human embryonic development

co-supervised by
Jérémie Bourdon, Professeur des Universités, Université de Nantes, LS2N, UMR 6004, Nantes
Carito Guziolowski, Maître de Conférences, Centrale Nantes, LS2N, UMR 6004, Nantes
Laurent David, PhD, HDR, CRTI, MCUPH faculté de médecine, Université de Nantes

PhD topic

Background

Assisted reproductive technologies (ART), and in particular in vitro fertilization (IVF), are in direct need of novel approaches to improve the clinical outcome of infertility treatments. Current limitations of embryo culture systems and embryo quality assessments methods **limit the success rate of IVF cycles to only 25%**, leading to social, emotional and medical burden for the couple and the infertility medical team. In this context, the recent advent of novel technologies, such as transcriptomics, proteomics and imagery, represents a formidable opportunity to consider in depth each embryo individually and to understand embryo developmental steps from a genetic and metabolic point of view. The final goal of deciphering human embryo development is to refine and set up robust embryo quality assessment methods, allowing more clinically relevant patient-centred care to improve clinical outcome. However, despite those promising technologies, research on human embryos is too limited to allow systematic testing of hypothesis-driven research, therefore **a key aspect for the future of our field is to be able to generate a computational model of preimplantation development**, starting with transcription factor networks. Computational models will be invaluable to predict how perturbations impact the system.

The aim of this thesis project is to write logic programs and use state-of-the-art solvers (such as clasp, a conflict-driven Answer Set solver used in the study of NP-hard search problems [1]) to perform combinatorial searches in the massive solution space of models describing human embryonic cells. The objective of such a search logic program will be to find optimal models that best fit incomplete and noisy data of human embryos development. The logic program resolution is equivalent to a learning algorithm on an exhaustive basis (answer set programs perform global instead of local searches).

Artificial intelligence

Knowledge in Learning is a branch of Artificial Intelligence [2] dedicated to the usage of prior-knowledge to infer, deduce, and check for hypotheses. It is applied to solve problems involving combinatorial search, planning, automatic diagnosis, and decision. On this context, solving a given problem is composed of the following ingredients : (1) a set of observations or prior knowledge, (2) a logic-program of the system, defining the (logical) rules, constrains, and queries that define our system, (3) a solver, a system that can read the logical program and compute if the observations satisfy the given rules, and finally (4) a model (of new observations) that explain the rules according to the given observations. These ingredients appear in the Answer Set Programming (ASP) paradigm [3], which is used to explore massive domains of (structured) data. Through previous works (see for example [4,5]) we have observed that the nature of biological data presents the following properties: incompleteness, noise and structure. These properties allow that ASP programs can be written to model such data, under particular contexts, such as the disposal of prior-knowledge concerning the interactions among biological entities, also known as protein-networks, signalling networks or gene regulatory networks. Through previous works we have proven how ASP can be applied to concrete biological problems on Human Health such as understanding the molecular mechanisms (in the biological networks) that differentiate patients with good and bad prognosis in Acute Myeloid Leukemia [6]. Very often these multi-disciplinary efforts, also point to unexplored computer science research on the field of Knowledge in Learning and in particular, in the improvement of algorithms embedder in the solvers of logic-programs. For example, we have recently proposed an algorithm that improves the *clasp [1]* solver search of optimal solutions and proposes computational models offering diverse biological mechanisms [7]. This type of search algorithm is novel and it could be applied on different search problems, not only of biological nature.

***Preliminary knowledge/data*.**

Generation of a pseudo-time human lineage specification model (Meistermann et al, https://ssrn.com/abstract=3441907). One of the outstanding question of the field is to understand the chain of events regulating human preimplantation development leading to an implantation-competent embryo. To address this question, we generated **single cell transcriptomic data** from multiple stages of human embryos in order to precisely describe the transcriptome changes occurring in the different cell types and overtime. We then used the expression matrix to generate a pseudo-time model: a dimensionality reduction method that organizes all samples along trajectories that represent the likeliest path to go from one cell to another (Qiu et al., 2017). Of note, our dataset corresponds to embryos that have been followed by **time-lapse microscopy** prior to analysis, thus guaranteeing proper staging using the Gardner and Schoolcraft grading system (2011) (**Fig 3A**). This is crucial to link the phenotype (e.g., cavitation) to the transcriptome. Finally, we have used preimplantation genetic diagnosis laser to dissect the embryos, composed of polar TE and embryonic compartments on one side, and the mural TE compartment on the other side of the blastocyst cavity. This allowed to unequivocally assess the transcriptome differences occurring along the blastocyst axis, which is thought to be crucial for implantation (Cruz et al., 1985).

Our preliminary pseudo-time model shows that the main progression is the development time of the cells: it starts with a single branch that bifurcates into 3 major branches, corresponding to the EPI, PrE and TE lineages (**Fig 3B**). This **continuous model allowed us to hierarchize expression of genes along preimplantation development**. In particular, we validated two transcription factors (TFs) associated with TE lineage specification: GATA3, which is expressed from morula stage, and NR2F2, which is expressed in late blastocysts (**Fig 3D**). Immunofluorescence showed that GATA3 is among the first markers associated with TE specification (**Fig 3E**) and that NR2F2 marks the mature polar TE cells, the cells that are thought to mediate the initial adhesion to the uterus (**Fig 3F**). This highlights the quality and biological relevance of our dataset and of our transcriptomic model. This preliminary work generated the currently **most comprehensive analysis describing the progression of the transcriptome of human preimplantation embryos**. This data serves as an invaluable basis to study the molecular pathways that are active during developmental progressions. This work is currently under revision and available online. Our transcriptomic analysis represents a perfect basis to generate a computational models allowing to challenge hypothesis *in silico*.

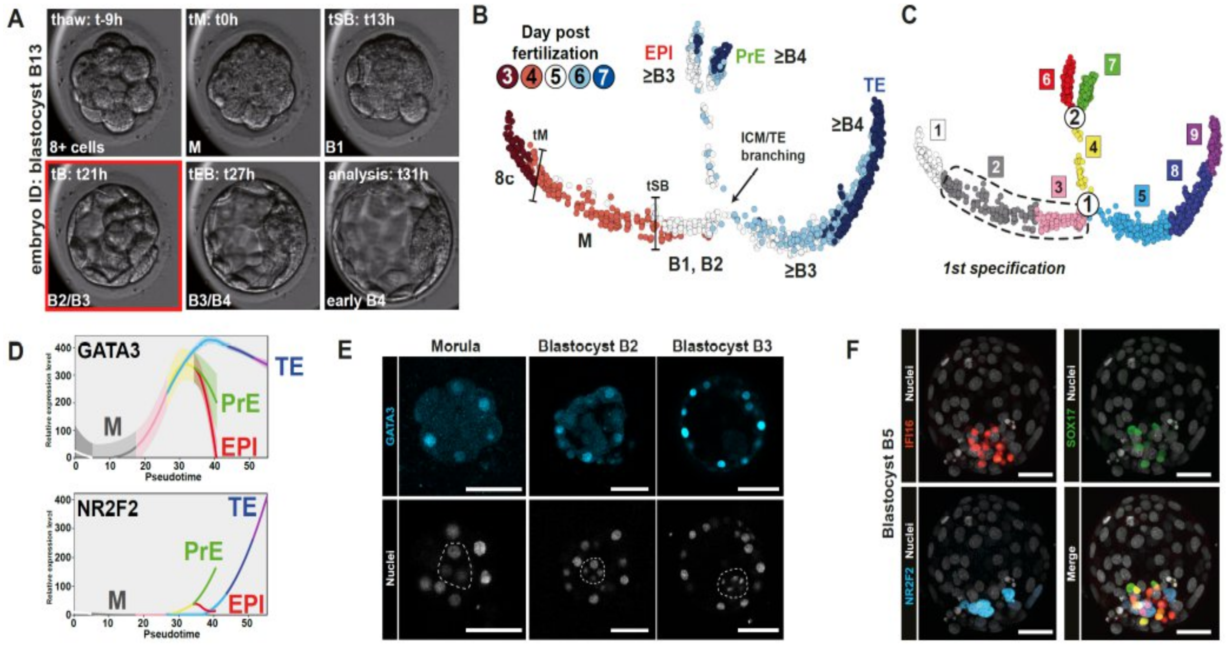

Figure 3. Automated image analysis coupled to scRNAseq pseudotime model.

A. Frames from time-lapse microscopy for embryo "B13". For each embryo sequenced in this study, their morphokinetics were acquired by time-lapse microscopy. Developmental events include morula compaction (tM), blastulation (tSB) leading to B1 stage, full blastocyst (tB) at B3 stage and expanded blastocyst once the zona pelucida thickness is halved (tEB) at B4 stage. tM is used as t0 to compare thawed embryos. **B.** Projection of developmental day (3 dpf to 7 dpf) for all samples combined for this study, and the result of our refined staging. TE and ICM distinct transcriptomes are detectable at the B2/B3 stage transition. **C.** Subdivision of branches of the pseudotime model according to gene module expression: Pre-specification branch is subdivided into 3 states (1.Pre−morula, 2.Morula, 3.Unspecified B1/B2) and the TE branch is subdivided into 3 states (5.Early-TE, 8.TE.NR2F2-, 9.TE.NR2F2+). This yields a total of 9 states, numbered by their order of apparition in the pseudotime. GATA3 and NR2F2 expression are represented using the colored segments. **D-E.** IF staining of indicated proteins at specific developmental stages. Dashed line: ICM. Scale = 47µm.

Establishing a Gene Regulatory Network and modelling

So far we have compiled a list of ASP systems to interrogate and analyse biological networks: Iggy [4], for checking the consistency between a biological network and experimental observations, and learn new observations; Caspo [5], Caspo-ts [8], for discovering boolean models from static, and dynamic data respectively; and a system to discover the dynamic bifurcation mechanisms on biological regulatory networks [9]. We have handled so far transcriptomic and phospho-proteomics data related in particular to Human cell-lines and cancer patients cohorts.

Once the experimental observations are given, we usually infer a network, which will be the basis to write the rules of the logical programs. The semantics of this network is simple, but it needs to respect a given logic. For this, we have developed a system to retrieve automatically this type of information, named Bravo [10], from public knowledge databases.

For the current PhD thesis project, a preliminary plan could be stated as:

- First retrieving the network from databases using Bravo and a list of relevant genes from the previous section study.
- Once the network will be built we will decide which ASP system can be useful for the problematic. We believe that a combination of Iggy and the bifurcation analysis could fit well. However this deserves a

rigorous study, since possibly a new ASP system has to be conceived. The data from Laurent David's project are particular and new regarding our previously developed systems. They are different since they do not use a time unit as classic time-series data points. It can be interesting to set the logic rules to describe pseudo-time.

- Check if the computational model reproduces the known facts.
- Predict new hypothesis, or run tests of the behaviour of the biological system if certain genes are altered

Pre-requisite The PhD candidate will have a Computer Science or Bioinformatic profile (Master degree or equivalent) with knowledge on logic programming or artificial intelligence. Previous experience of analysing (or computationally modelling) massive datasets of biological nature will be helpful.

Funding This PhD topic is eligible to the AIby4 call (https://aiby4.ls2n.fr/), which funds multidisciplinary Phd thesis propositions on Artificial Intelligence.

Application

- Please feel free to contact us if you have any question concerning the project, your eligibility, or the application procedure to:
  - carito [dot] guziolowski [at] ec-nantes [dot] fr
  - Jeremie.Bourdon@univ-nantes.fr
  - laurent.david@univ-nantes.fr
- Your application must contain : (1) your CV, (2) your Cover Letter stating your professional project, (3) your transcript from Bac +3 to Bac +5 or equivalent (for the results of Master or equivalent, attach the documents in your possession), and (4) contact information for 2 referees.
- Please send us your application by email before the **30/04/2021**

References

[1] Martin Gebser, Benjamin Kaufmann, André Neumann, et al. clasp: A conflict-driven an-swer set solver. InLogic Programming and Nonmonotonic Reasoning, pages 260–265, Berlin,Heidelberg, 2007. Springer Berlin Heidelberg.

[2] MLA. Russell, Stuart J. (Stuart Jonathan). Artificial Intelligence : a Modern Approach. Upper Saddle River, N.J. :Prentice Hall, 2010.

[3] Baral, C.: Knowledge Representation, Reasoning, and DeclarativeProblem Solving. Cambridge University Press, New York, NY, USA (2003)

[4] Thiele, S., et al.: Extended notions of sign consistency to relate experimental data to signaling and regulatorynetwork topologies. BMC Bioinformatics16(1) (2015). doi:10.1186/s12859-015-0733-7

[5] Videla, S., Saez-Rodriguez, J., Guziolowski, C., Siegel, A.: caspo: atoolbox for automated reasoning on the response of logical signalingnetworks families. Bioinformatics33(6), 947–950 (2017). doi:10.1093/bioinformatics/btw738

[6] L Chebouba, B Miannay, D Boughaci and C Guziolowski. Discriminate the response of Acute Myeloid Leukemia patients to treatment by using proteomics data and Answer Set Programming. *BMC Bioinformatics* 2018 19(Suppl 2):59

[7] Misbah Razzaq, Roland Kaminiski, Javier Romero, Torsten Schaub, Jeremie Bourdon and Carito Guziolowski: Computing Diverse Boolean Networks from Phosphoproteomic Time Series Data. In *CMSB 2018*, *Brno*, *Czech Republic*, *September* **2018**. LNCS, vol 11095, pp 59-74

[8] Razzaq M, Paulevé L, Siegel A, Saez-Rodriguez J, Bourdon J, Guziolowski C. Computational discovery of dynamic cell line specific Boolean networks from multiplex time-course data. PLoS Comput Biol. **2018** Oct 29;14(10):e1006538.

[9] L. Fippo Fitime, O. Roux, C. Guziolowski†, L. Paulevé†. Identification of Bifurcation Transitions in Biological Regulatory Networks using Answer-Set Programming, *Algorithms Mol Biol.* **2017** Jul 20;12:19. doi: 10.1186/s13015-017-0110-3

[10] M Lefebvre, A Gaignard, M Folschette, J Bourdon, C Guziolowski, Large-scale regulatory and signaling network assembly through linked open data, *Database*, Volume 2021, 2021, baaa113, https://doi.org/10.1093/database/baaa113