CA 21169, Information, Coding, and Biological Function: the Dynamics of Life, Dynalife

Deliverable D2A: Report on a universal platform for validation and testing, including recommended tools and procedures (month 24).

Background: This deliverable focuses on the synergy between WG1 and WG2 by creating a report of a universal platform that can be used for testing and validation. The objective of WG1 is to develop a theoretical framework that unifies the interpretation of patterns observed in the genetic code and coding regions of DNA by comparing and integrating existing models and theories. These theoretical approaches will not only address the genetic code and coding regions but will also extend to noncoding regions of the genome and potentially to other biological codes. WG2, on the other hand, provides expertise in statistics, bioinformatics, and computer science for biological data analysis, that can support the theoretical insights from WG1. This connection between WG1 and WG2 aims to bridge theoretical models and real datasets, fostering collaboration and innovation. It can be expected that the theoretical problems that are addressed by WG1 also pose new challenges that require modern technical expertise in the fields of bioinformatics, machine learning, and statistics.

On the interface between the theoretical models and the technological means, Dynalife seeks to identify the requirements and specifications of a platform that brings together both existing and new tools for validating and testing theoretical models using biological data. The main objective of the platform will be to serve as a robust and comprehensive resource, enabling researchers to apply and refine models through a structured framework of validation.

To achieve this, identifying relevant data sources is the first challenge. Initially, freely available datasets may be utilized, but future efforts may involve specifically designed experiments that cater to unique requirements. This will involve meticulous experimental design and appropriate protocols. Given the inherent complexity of biological data and the dynamic nature of molecular information, the platform will help researchers to select and configure methodologies that provide accurate parameter values for theoretical models. Hypothesis testing and goodness-of-fit measures will further refine these models, enabling necessary adjustments to model assumptions for optimal accuracy.

In addition, the platform will allow to compare model predictions against data generated through numerical simulations and machine learning (ML) predictors. This integration will allow for a dual approach: enhancing the extraction of knowledge from ML methods and improving ML methodologies and outcomes. Dynalife aims to connect these efforts with the broader theory of Dynamical Systems, leveraging elements such as genomic sequence patterns and reservoir computing approaches to build a comprehensive, flexible, and impactful validation framework. A dedicated workshop, "Statistical and machine learning methods for biological data, Nijmegen", October 2024, has been organized to deepen the connection between ML and statistical models, recognizing that while ML has greatly advanced the field, statistical models remain crucial for addressing uncertainties and achieving robust results.

Progress so far: The organization of a WG2 Meeting to focus on WG2 tasks and cooperation research between WG1 and WG2 was one of the recommendations from the WG1 DYNALIFE Meeting "Modelling Information Flow in Biological Coding Systems" in Belgrade, in June 2023. This meeting was organized at Thessaloniki in June 2024: "DYNALIFE WG1-WG2, Interaction Meeting, Data driven evidence: theoretical models and complex biological data". Discussions about the interaction between WG1 and WG2 were further pursued in this meeting, also at the level of the Core Group and MC of Dynalife. In particular, one of the conclusions is that there remains a gap between theoretical models and the specific datasets and instruments needed to validate and test these models. Moreover, in developing a cohesive framework for quantitative theoretical biology, two challenges have emerged:

1. Theoretical models, though powerful, sometimes overlook the complexities of biological data, which can make them difficult for theoretical scientists to interpret and apply.
2. Conversely, advancements in biological sciences often follow an empirical approach, and some biologists may hesitate to engage with theoretical models due to their mathematical nature.

Approach: A first discussion was put forward during the first meeting in Venice in May 2022, where it was noted that the problem seems to be a huge one to the extent that there exist COST actions exclusively dedicated to biological databases. This was also discussed during the WG1 meeting “Modelling Information Flow in Biological Coding Systems” in June 2023, and it was decided to begin facilitating collaborations between WG1 and WG2 members and in particular deliverable D2A was discussed. Following discussions, a call for proposals was issued to facilitate joint projects between WG1 and WG2 and for identifying testing tools and data sets. The call, prepared by the leaders of WG1 and WG2 and supported by the DYNALIFE Core Group and the Chair of the Action, received five proposals focusing on areas such as the genetic code, information flow, and agent-based modeling (see annex 1).

These proposals highlighted the potential for cross-disciplinary progress but also underscored the need for specific datasets that directly support the theoretical models. However, the rapid growth in both the quantity and quality of biological data introduces challenges in data selection and the definition of tools for a platform is therefore a difficult problem. Recognizing this, Dynalife anticipates that tool selection may take time and probably cannot be completed within the duration of the Action; however, Dynalife will continue to work diligently toward this goal. Progress will be achieved through the development of specific models that already have identified datasets, allowing for the incremental building of tools and testing strategies in the Action’s repository. This repository will be equipped with metadata that guide new models through established procedures and strategies, creating a valuable resource that will remain accessible and developable beyond the duration of the Action (see the description of the repository/dissemination platform, in Deliverable D4B).

Looking forward, the repository will serve as a foundation for a comprehensive testing platform. This effort will be supported by further exploration of theoretical models in conjunction with ML, leveraging rigorous statistical approaches. This final endeavour will pave the way for a synergistic integration of ML and theoretical models, creating a dynamic, adaptive platform that evolves with future advancements in the field.

Annex 1

Call for theoretical models to be used as case studies for the testing and validation platform:

Dear WG1 Members,

The organization of a **WG2 Meeting to focus on WG2 tasks and cooperation research between WG1 and WG2** is one of the recommendations from the WG1 meeting in Belgrade, which was also attended by WG2 Leader Jeanine Houwing-Duistermaat. On that occasion, an online mini-meeting was held. Some of the conclusions are that such an **extended WG2 Meeting should be organized in the spring of 2024 and that WG1 should more specifically define its research proposal for WG2 as the first step in creating cooperation between the two groups**.

We invite WG1 members to submit their Research Proposals designed for data science with the goal of establishing WG1 and WG2 cooperation and validating your theoretical models. The proposal should introduce the problem you are investigating, why this problem is important, the hypothesis to be tested, a brief summary of the methodology and a conclusion. In more detail, the following information should be included in the research proposal:

1. **Title** – the short title of your research proposal.
2. **Background/Introduction** – a description of background facts and the importance of the research area you are investigating (your research area, research motivation, the implications for knowledge and application).
3. **Problem Statement & Hypothesis** – a summary of the particular issues to be addressed (such as a lack of knowledge of...) and formulating the hypothesis to be tested.
4. **Specific Aims/Objectives** – a list of objectives to be achieved for testing your hypothesis with a description of the benefits/impacts that will be generated if the research problem (such as some mechanisms of biology) is answered.
5. **Methodology** – a definition of research methods and logical steps to solve the problem and achieve the proposed objectives, and in particular specifying the data sets to be studied (define input parameters, features, methods, data sets...).
6. **Literature Review** – a short summary of previous related research on the research problem and your main references related to the research proposal.

Note 1: Include figures, tables, and datasets, if necessary.

Note 2: The research proposal should contain up to three A4 pages (without datasets).

Note 3: Your research proposal should include datasets, but generally you have 3 options..

Option 1 (well-defined datasets and provided in advance) – The best option is that you/your_team can well-defined and provide in advance the datasets. Below are descriptions of the datasets that are required:

1. **Types of biological datasets** (genomes, proteomes, transcriptomes, epigenomes, metabolomes, microbiome…).
2. **Estimated dataset sizes and availability of metadata**.
3. **Bioinformatics platforms used for data production**.
4. **Data sources** (e.g., blood, biopsies, stool, environmental…).
5. **Shotgun or targeted sequencing data** (i.e., define whether you are working with the whole or partial genome).

Option 2 (partially defined datasets or provided with the help of WG2 experts) – It is not necessary for a research proposal to immediately provide datasets, but you do need to describe them as best you can. If you need some additional guidelines, you can count on the help of WG2 members.

Option 3 (without datasets) – The least preferred option is to submit a research proposal without defining any datasets. In that case, it is preferable to present the research idea as clearly as possible.

For a brief introduction to bioinformatics databases, software and tools, the appropriate text is attached.

For any questions regarding this Call, please contact the email addresses of WG1&2 Leaders at the same time - J.Duistermaat@math.ru.nl & nmisic@rcub.bg.ac.rs.