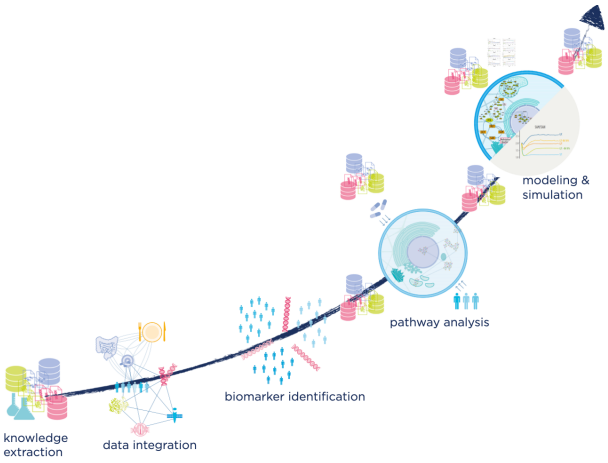
Computational approaches to understanding nutrient metabolism and metabolic disorders

Corrado Priami
Dipartimento di Informatica, Università di Pisa
Largo Pontecorvo, 56124 Pisa, Italia
corrado.priami@unipi.it

Abstract.
Computational methods are becoming more and more essential to elucidate biological systems. Many different approaches exist with pros and cons. This paper reviews the most useful technologies focusing on nutrient metabolism and metabolic disorders. Space limitation prevents from exploring the examples in details, but pointers to the relevant papers are reported.

Graphical abstract.



Highlights.

1. Identification of main computational needs for nutrient metabolism and metabolic disorders
2. Definition of an ideal pipeline to address the needs identified in 1.
3. Description of the main state of the art technologies
4. Outlook of the field

INTRODUCTION

Computational technologies experienced an exponential growth in the last decade for their application to biological data and systems. Methods can roughly be divided into two main categories depending on the outcome they are intended to elucidate: static technologies and dynamic technologies. Static technologies are analyzing snapshots of biological systems by observing a state, measuring variables of that state and trying to connect them to physiology or higher-level behavior of the system. Static technologies are not considering the variable time as a main variable of the models and analysis performed. On the contrary dynamic technologies are considering time as the driving variable of the models. These methods represent the evolution of biological systems over time and use them through simulation for predicting and controlling the behavior of biological systems [1]. Static technologies are usually collected

under the discipline of bioinformatics [2], while dynamic technologies are the core of computational biology [3, 4] (see Fig. 1).
An interesting area of biology that can gain a lot from computational models is metabolism [5, 6]. It is absolutely relevant to prevent and cure metabolic disorders like Type 2 Diabetes that is affecting a constantly growing number of individuals [7, 8]. Metabolic disorders can be controlled through drugs (with a huge interest of pharmaceutical companies), but they can also be prevented by controlling diet behavior [9]. Food is consumed three times a day, every day and with the help of microbiota is delivering a huge flow of metabolites within human bodies that interact with vital biological processes [10]. Understanding the biochemical interactions involved in food metabolism and disease onset and progression is fundamental to improve and maintain health.
An ideal pipeline to address computationally questions related to nutrient metabolism and metabolic diseases is made up of five steps:

1. data generation and data acquisition
2. data integration
3. clustering and classification
4. functional annotation and analysis
5. modeling and simulation

Computational approaches produce higher gains as we move from item 1 to item 5 above, although the technological and temporal cost of the methods is growing along the same axis.

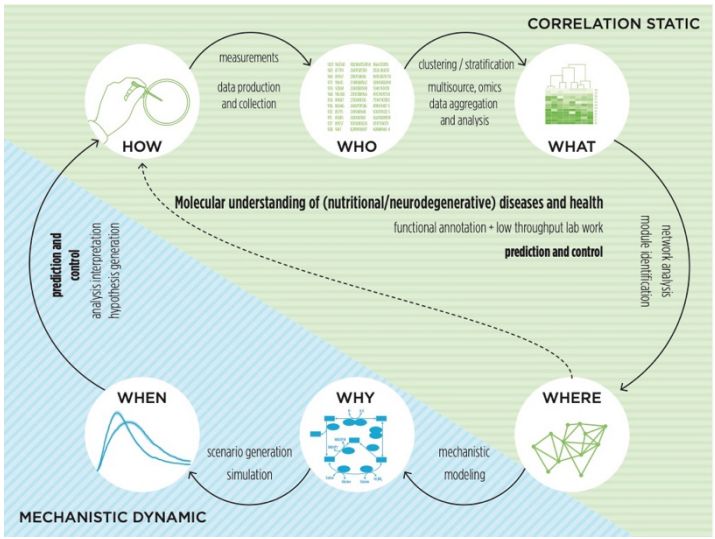


Figure 1.
The figure collects in the green part the static methods that includes the items 1 to 4 above. The blue part shows how the modeling and simulation, and hence time evolution of systems, connects with the static methods. Importantly, the figure shows how the computational support to biomedical research is an iterative approach.

The paper discusses the five steps above by using as a running example the folate metabolism.

Data and knowledge acquisition

The first step to address scientific questions is data acquisition and knowledge gathering about the problem at hand. A fundamental set of technologies in this phase is literature and data mining by exploiting both structured and unstructured sources. Artificial intelligence plays a crucial role here by exploiting natural language processing methods among which BERT [11], SciBERT [12] and BioBERT [13] approaches. This step is crucial to delineate the boundaries of the project and being sure that all the available knowledge is reviewed before embarking on a new discovery project. A preliminary step to many projects is collecting information from the literature.
A fundamental aspect to consider in this step of the ideal pipeline, and in all subsequent steps, is the quality and type of data collected. For instance, we need to fix a species, a tissue and cell

type that we are interested in and avoid mixing data collected from different systems or with different assumptions.
An example is a fully computational study to identify existing drugs that can have an impact in curing metabolic syndrome [14]. Metabolic syndrome is a pathological condition characterized by obesity, hyperglycemia, hypertension, elevated levels of triglycerides and low levels of high-density lipoprotein cholesterol that increase cardiovascular disease risk and type 2 diabetes. There are many metabolic and signaling pathways involved in the onset and progression of the disease and they can be controlled both by drugs and by adequate food habits.
The outcome of machine learning applied to literature mining to extract knowledge on a specific topic is a network connecting the relevant concepts we are diving from the literature ocean with scores that states how strong is the relationships (see Fig. 2).

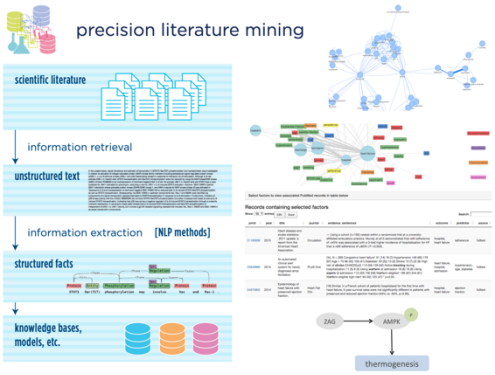


Figure 2.
The figure shows on the left the text mining pipeline that is performed to move from unstructured data to structured data that is easily searchable. On the top right there is a network representation that highlights the relationships between the concepts that have been identified has core for the scientific question. The network is linked to sentences that have been used to uncover the relationships. Eventually, in the bottom right corner, there is a representation similar to a pathway that can be inferred from semantic network just described.

Data integration

The last decade has shown an ocean of data being generated with limited orchestration. It is not easy to dive the data ocean produced by different lab technologies with different resolution and quality. Data are collected at different level of abstractions spanning from all the omics (genomics, proteomics, metabolomics, lipidomics, etc) up to physiological and clinical markers. It is not easy to compare different data sets, but the discipline of network biology is providing a useful tool for the integration [15, 16, 17, 18].
Still considering the example of the previous section, we can expand the molecular network representing the metabolic syndrome mechanisms with the network modules that represent the metabolism of available drugs or the metabolism of specific nutrients. The resulting network integrates all the available knowledge about the project under consideration and can be used to determine the topological distance between the disease module and the metabolism modules of drugs or nutrients. The closer the modules, the higher the likelihood of interaction between the disease and the food/drug.
There are many methods that are based on network biology and are used to perform many interesting analyses of metabolic processes [19, 20, 21].

Clustering and classification

When studying a biological problem at pre-clinical or clinical level very often imposes to find markers to separate or cluster animals or subjects in classes with similar characteristics (examples are responders vs non-responders, patients vs controls, different sub-types of a disease etc.). The variables that properly discriminates the objects of the study are called biomarkers and usually have a relevant role in the biological processes under study. Artificial intelligence is playing a fundamental role in biomarker identification, complementing classical

statistical approaches like PCA. Supervised and unsupervised methods are used to extract from large datasets features that distinguishes the population from which the data has been recorded into distinct groups. Particularly promising are the so-called rank-based methods that do not consider the actual values of the measured variables, but only their rank in an ordered list [22, 23]. Once the list of variables is ordered, the top ranked and bottom ranked are considered as a signature of a subject. Defining a distance on the subject signatures allows to group subjects with similar molecular behavior. These approaches allow to limit the bench-effect and inter-center variability of lab activity (see Fig. 3).

Figure 3.
Assuming to have transcriptomic data, the second panel shows the ranked gene expression for each subject. The third panel selects the most and least expressed genes as individual signatures. The fourth panel defines a distance between each pair of signatures. A threshold is selected to represent the distance matrix as a graph where subjects with close distances cluster together and can be visually explored. The solution in the figure can be improved by equipping the methods with genetic algorithms that can tune parameters to optimize the size or number of expected clusters, the length of the biomarker, the size of the signature.

Many methods have been proposed recently to integrate different data types especially in the field of multi-source biomarker identification. These methods overlap with the previous section of data integration, but we mention them here because have been mainly applied to identify by markers. Among the most relevant, we mention Multiple Canonical Correlation Analysis, Multiple Co-Inertia Analysis, Multiple Factor Analysis, Joint and Individual Variation Explained and Similarity Network Fusion.
Examples of metabolic biomarker identification are in [24, 25, 26].

Functional annotation and analysis

The biomarker identification technologies tell us what to measure and compare to classify and cluster objects according to a specific end-point of interest. Unfortunately, there is no clue of why those markers are biologically important. We then need to bind the markers to biological processes so that the why question can be answered. Connecting the marker and the molecular interactions produces a network of limited size and focused on the scientific question that we are addressing. There are many tools for computational analysis of pathways [27, 28, 29, 30, 31]. A promising approach is considering topological information, experimental data and a priori knowledge at the same time. The intuition is to rely on a human, tissue-specific interactome (that collectively represent the available knowledge and the topological structure of the interactions). The interactome is then tagged with differentially expressed genes obtained directly from experimental data. Eventually, the researcher can tag the interactome with molecules that he thinks are particularly relevant for the scientific question under investigation. The resulting labelled interaction network is used as a backbone for running an information flow algorithm that ranks the level of activity of different paths. The modules identified can be used to determine significative canonical pathways with an enrichment analysis.
For instance, we might want to study the effect of fructose on maturation of adipocytes (see Fig. 4) [32]. Additional studies that shows applications of network biology to metabolism are in [33].

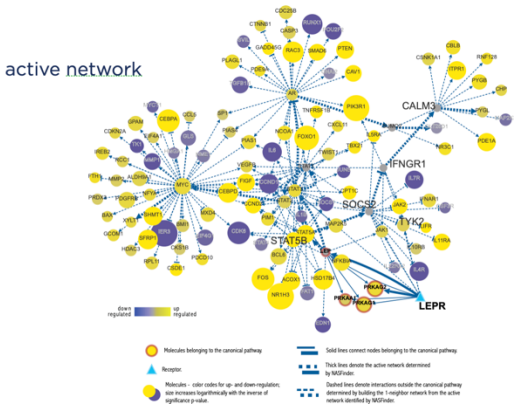


Figure 4.
The effect of fructose on maturation of adipocytes can be observed by mapping the genes differentially expressed on the human interactome restricted to adipose tissue. Researcher can select cellular receptors as molecule of interest for the problem at hand. The result of the functional analysis is a set of active networks like the one in the figure (here referring to the leptin receptor). Eventually it is possible to determine the canonical pathways that intersect the active network through enrichment analysis.

Modeling and simulation

The network identified in the previous step is the basis to build a mathematical model that can be simulated by a computer to describe the temporal behavior of the biological system [34, 35]. This step is the most expensive in terms of resources, but it produces the highest reward because it allows to predict the behavior of the system under investigation and eventually devise strategies to control it.

There are three main simulation approaches: deterministic, stochastic and hybrid [36]. Deterministic simulation provides the same result for each run and shows an average behavior that is good enough when the concentrations of the components of the system are large. The execution time of deterministic simulations is short. Stochastic simulation has a longer execution time, but reproduces an exact behavior of the system taking the noise fluctuation typical of biological systems into account. To cover all the space of possible behaviors it is necessary to run many times the same simulation (that does not produce the same result at each run). Averaging on the obtained results we approach the result of the deterministic simulation when the components are numerous enough. Hybrid is a mixed approach in which the numerous components are simulated deterministically and the components with limited amount are simulated stochastically.

A great example for modeling and simulation is elucidating folate metabolism (see Fig. 5) [37, 38]. Additional examples of dynamic modeling methods are in [39].

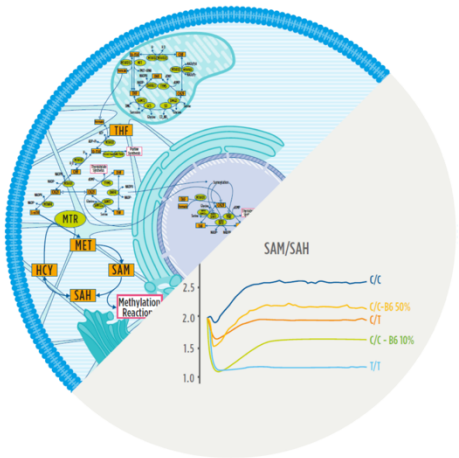


Figure 5.
The light blue part of the figure shows the one-carbon mediated folate pathway that is simulated under different MTHFR genotypes. The bottom part of the figure shows how the value of the fraction SAM/SAH is changing according to the genotype and also how the level of vitamin B9/folate is affecting that value. If we consider SAM/SAH as an indirect measure of the quality of the metabolism, we have direct link to vitamin B9/folate supplementation to control the system [37].

CONCLUSIONS

The paper presented a discussion of the computational methods available for addressing biological questions and exemplified the technologies on food metabolism and metabolic diseases. The technologies are organized into an ideal pipeline that cover data acquisition and integration, classification and functional analysis of markers, modeling and simulation. There has been an exponential growth in the application of computational technologies in biological studies over the last decade that makes computational approaches a new tool for biologist. Among these, artificial intelligence and especially machine learning is playing a pivotal role [40, 41, 42, 43].

An effort is needed to make the technologies accessible to the non-experts of the field in order to complete the last step of acceptance of these methods and making them a new microscope for biomedical scientists [44, 45, 46].

REFERENCES

[1] Morine M., Priami C. Analysis of biological systems. Imperial College Press, 2015.

[2] Altman R. Current Progress in Bioinformatics 2016. Brief Bioinform. 2016;17(1):1.

[3] Chen L. Computational systems biology for omics data analysis. J Mol Cell Biol. 2019, 19;11(8):631-632.

[4] Brown JB, Bajorath J. Computational chemical biology on the rise. Future Med Chem. 2019;11(1):1-3.

[5] Notebaart RA, Kintses B, Feist AM, Papp B. Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol. 2018;49:108-114.

[6] Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020;16(1):60-68.

[7] Roden M, Shulman GI. The integrative biology of type 2 diabetes. Nature. 2019;576(7785):51-60.

** This paper is a good example of how network biology can help elucidate mechnisms by relying on model diagrams. The paper shows a holistic view of Type 2 diabetes at molecula level.

[8] Magliano DJ, Islam RM, Barr ELM, et al. Trends in incidence of total or type 2 diabetes: systematic review. BMJ. 2019;366:l5003.

[9] Uusitupa M, Khan TA, Viguiliouk E, et al. Prevention of Type 2 Diabetes by Lifestyle Changes: A Systematic Review and Meta-Analysis. Nutrients. 2019;11(11):2611.

[10] Salgaço MK, Oliveira LGS, Costa GN, Bianchi F, Sivieri K. Relationship between gut microbiota, probiotics, and type 2 diabetes mellitus. Appl Microbiol Biotechnol. 2019;103(23-24):9229-9238.

* This paper shows how integration of different data types is fundamental to better understanding mechanisms of action and progression of complex diseases. The paper shows the relationships between gut microbiota and type 2 diabetes.

[11] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018; arXiv 1810.04805.

[12] Beltagy I, Kyle L, Cohan A. SciBERT: Pretrained Language Model for Scientific Text. In Proceedings of EMNLP. 2019, arXiv:1903.10676.

[13] Jinhyuk L, Wonjin Y, Sungdong K, Donghyeon K, Sunkyu K, Chan Ho So, Jaewoo K. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36(4):1234–1240.

[14] Misselbeck K, Parolo S, et al. Metabolic syndrome: identification of deregulated pathways and drug effects by network analysis. Nat. Comm., 10:5215, 2019.

[15] Tini G et al. Multi-omics integration a comparison of unsupervised clustering methodologies. Briefings in Bioinformatics, doi: 10.1093/bib/bbx167, 2017.
[16] Granata I, Troiano E, Sangiovanni M, Guarracino MR. Integration of transcriptomic data in a genome-scale metabolic model to investigate the link between obesity and breast cancer. BMC Bioinformatics. 2019;20(Suppl 4):162.
[17] Visconti A, Le Roy CI, Rosa F, et al. Interplay between the human gut microbiome and host metabolism. Nat Commun. 2019;10(1):4505.
[18] Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Curr Protoc Bioinformatics. 2019;68(1):e86.
[19] Singh D, Lercher MJ. Network reduction methods for genome-scale metabolic models. Cell Mol Life Sci. 2020;77(3):481-488.
[20] Li GH, Dai S, Han F, Li W, Huang J, Xiao W. FastMM: an efficient toolbox for personalized constraint-based metabolic modeling. BMC Bioinformatics. 2020;21(1):67.
[21] Bruins MJ, Van Dael P, Eggersdorfer M. The Role of Nutrients in Reducing the Risk for Noncommunicable Diseases during Aging. Nutrients. 2019;11(1):85.
[22] Lauria M, Moyseos P, Priami C. SCUDO: a tool for signature-based clustering of expression profiles. Nucleic Acid Research, 2015, doi: 10.1093/nar/gkv449.
[23] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-477.
[24] Collins C, McNamara AE, Brennan L. Role of metabolomics in identification of biomarkers related to food intake. Proc Nutr Soc. 2019;78(2):189-196.
[25] Malik VS, Guasch-Ferre M, Hu FB, et al. Identification of Plasma Lipid Metabolites Associated with Nut Consumption in US Men and Women. J Nutr. 2019;149(7):1215-1221.
[26] Ulaszewska MM, Weinert CH, Trimigno A, et al. Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. Mol Nutr Food Res. 2019;63(1):e1800384.
[27] Xi Y, Wang F. Extreme pathway analysis reveals the organizing rules of metabolic regulation. PLoS One. 2019;14(2):e0210539.
[28] Vieira V, Rocha M. CoBAMP: a Python framework for metabolic pathway analysis in constraint-based models. Bioinformatics. 2019;35(24):5361-5362.
[29] Mehra R, Jasrotia RS, Mahajan A, et al. Transcriptome analysis of Snow Mountain Garlic for unraveling the organosulfur metabolic pathway. Genomics. 2020;112(1):99-107.
[30] Wu C, Jiang H, Kalra I, et al. A generalized computational framework to streamline thermodynamics and kinetics analysis of metabolic pathways. *Metab Eng*. 2020;57:140-150.
[31] Toubiana D, Puzis R, Wen L, et al. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. Commun Biol. 2019;2:214.
[32] Nassiri I et al. Systems view of adipogenesis via novel omics-driven and tissue-specific activity scoring of network functional modules. Nat. Sci. Rep.,6:28851, 2016.
[33] Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17(2):159-162.
[34] Dupont D, Alric M, Blanquet-Diot S, et al. Can dynamic in vitro digestion systems mimic the physiological reality?. Crit Rev Food Sci Nutr. 2019;59(10):1546-1562.
[35] Blanco-Morales V, Garcia-Llatas G, Yebra MJ, Sentandreu V, Lagarda MJ, Alegría A. Impact of a Plant Sterol- and Galactooligosaccharide-Enriched Beverage on

Colonic Metabolism and Gut Microbiota Composition Using an In Vitro Dynamic Model. J Agric Food Chem. 2020;68(7):1884-1895.

[36] Marchetti L, Thanh VH, Priami C. Simulation algorithms for computational systems biology. In Texts in Theoretical Computer Science. An EATCS Series, Springer, ISBN: 978-3-319-63111-0, 2017.

[37] Misselbeck K et al. A hybrid stochastic model of folate-mediated one-carbon metabolism: Effect of the common C677T MTHFR variant on de novo thymidylate biosynthesis. Nat. Sci. Rep., 7:797, 2017.

[38] Berndt N, Egners A, Mastrobuoni G, et al. Kinetic modelling of quantitative proteome data predicts metabolic reprogramming of liver cancer. Br J Cancer. 2020;122(2):233-244.

[39] Misselbeck K et al. The 5-formyltetrahydrofolate futile cycle reduces pathway stochasticity in an extended hybrid-stochastic model of folate-mediated one-carbon metabolism Nat. Sci. Rep., 9(1):4322, 2019.

* This paper shows how dynamic modeling can help predicting system behavior and hence understanding how to control the system. The paper deals with folate metabolism.

[40] Alag A. Machine learning approach yields epigenetic biomarkers of food allergy: A novel 13-gene signature to diagnose clinical reactivity [published correction appears in PLoS One. 2019 Jul 24;14(7):e0220470]. PLoS One. 2019;14(6):e0218253.

* This paper shows how machine learning can be used to obtain a complete classification with very high accuracy of people who have food allergies and those who are only sensitized to food.

[41] Erban A, Fehrle I, Martinez-Seidel F, et al. Discovery of food identity markers by metabolomics and machine learning technology. Sci Rep. 2019;9(1):9697.

[42] De Clercq D, Wen Z, Fei F, Caicedo L, Yuan K, Shang R. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Sci Total Environ. 2020;712:134574.

[43] Berry SE, Valdes AM, Drew DA, et al. Human postprandial responses to food and potential for precision nutrition. Nat Med. 2020;26(6):964-973.

* This paper shows presents a machine-learning model that predicted both triglyceride and glycemic responses to food intake with the aim of developing personalized diets.

[44] Guerriero ML et al. Narrative-based computational modelling of the Gp130/JAK/STAT signalling pathway. BMC Systems Biology, 3:40, 2009.

[45] Gostner R et al. Graphical Modelling Tools for Systems Biology. ACM Computing Surveys, 47(2), 2014.

[46] Rougny A, Touré V, Moodie S, et al. Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. J Integr Bioinform. 2019;16(2):20190022.