DISS. ETH NO. 23931

Generalized and High-throughput ^{13}C Metabolic Flux Ratio Analysis by Machine Learning

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES OF ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

MARIA ZIMMERMANN-KOGADEEVA
Specialist in Mathematics and System Programming,
Lomonosov Moscow State University
born on 03.05.1989
citizen of Russian Federation

accepted on the recommendation of
Prof. Dr. Uwe Sauer
Dr. Nicola Zamboni
Prof. Dr. Manfred Claassen
Prof. Dr. Vassily Hatzimanikatis

2016

Abstract

Metabolism is an essential process for all living creatures. It describes the chemical conversion of consumed nutrients into biomass precursors, redox equivalents and energy, and the release of waste products. The metabolic potential of each cell is represented by a network of metabolites connected via reactions catalyzed by enzymes encoded in its genome. Living cells adjust their metabolic reaction rates, known as fluxes, in response to the external conditions, internal status and cellular requirements. Hence, metabolic fluxes reflect actual cellular behavior, and their assessment is crucial for understanding and controlling metabolic processes of biotechnological and biomedical relevance.

Metabolic fluxes cannot be measured directly, and have to be indirectly inferred from measurable quantities, e.g. gene expression, protein abundance, or temporal profiles of metabolite concentration changes. The most explicit experimental technique for flux elucidation are stable isotope tracing experiments, highly informative when alternative reactions lead to distinct labeling patterns in metabolic intermediates. These labeling patterns are detectable by nuclear magnetic resonance spectroscopy or mass spectrometry, and have to be interpreted either manually, with analytic equations, or incorporated into mathematical models in order to retrieve information on the underlying metabolic fluxes.

Existing flux analysis methods are divided into global ^{13}C metabolic flux analysis, which is based on iterative fitting of flux distributions to the labeling data; and local analysis based on calculating flux ratios from labeling data with ad hoc analytic equations. The former is generally applicable, but it requires comprehensive measurements, provides multiple equally plausible flux solutions and is computationally costly. The latter, on the contrary, is rapid and easy to apply, provides unique relative flux estimates, but is limited to few nodes and experimental conditions. Therefore, there is a demand for a fast, robust and generally applicable method that is scalable to large datasets, conceivably obtained from parallel experiments.

In this work, we present SUMOFLUX, a conceptually novel generalized method for targeted ^{13}C metabolic flux ratio analysis. This method exploits machine learning to predict flux ratios of interest from measurable data, using an *in silico*

training dataset generated with **su**rrogate **mo**delling. SUMOFLUX is applicable to virtually any type of network, substrate and measured data that can be simulated; and the actual estimation is very rapid once the flux ratio predictor is built. In **Chapter 2**, we developed the SUMOFLUX workflow, performed a proof-of-principle experiment to resolve key flux ratios in central carbon metabolism of *Escherichia coli*, and demonstrated that SUMOFLUX estimates were in good agreement with results obtained with both local and global ^{13}C flux analysis methods. Additionally, we illustrated the scalability and ease of experimental design with SUMOFLUX on a cohort of 121 *Bacillus subtilis* transcription factor mutants.

A remarkable advantage of the targeted approach is its applicability in complex systems even in case of poorly determined networks and little amount of data. This benefit became especially apparent in **Chapter 3**, where we investigated amino acid metabolism in mycobacteria in defined media and in the infection setup with macrophage-like THP-1 cells. By formulating specific flux ratios characterizing amino acid utilization, we classified amino acids by their role for central metabolism in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis* growing in composite media. Investigation of *M. tuberculosis* behavior in the complex infection setup, where inter-species and media exchange fluxes are unknown, was possible through extensive simulations of feasible flux distributions in the phagosome-bacterial network. It revealed that during infection, biosynthesis of several amino acids decreased compared to bacterial growth in rich media. This implies that in the scarce nutritional conditions inside the phagosome, the pathogen is forced to utilize any nutrient it encounters. These results underline the adaptability of mycobacterial metabolism and partially explain recurrent failures of multiple drug treatments.

We further exploited the speed and flexibility of SUMOFLUX in **Chapter 4**, coupling it with rapid labeling data acquisition by untargeted high-throughput metabolomics platform FIA-TOF (flow injection – time of flight), that enabled to perform several hundred flux analyses per day. Substantial gain in speed came with challenges of missing and overlapping isotopologue data, which we partially solved by adopting rigorous filtering and quality check procedures. We validated the high-throughput flux analysis protocol with a set of *E. coli* knockout mutants with known flux phenotypes, which SUMOFLUX succeeded to predict. The developed protocol allowed us to perform a fluxomics screen of 60 *E. coli* strains with mutations in enzyme phosphorylation sites to generate hypotheses on the functional role of this

post-translational modification, which yet has not been extensively studied in prokaryotes. Our flux screening revealed the deactivating function of isocitrate dehydrogenase phosphorylation reported earlier, and proposed several novel functionally relevant phosphorylation events. The developed high-throughput flux profiling protocol brings ^{13}C fluxomics to a new level comparable with the scale of other omics techniques.

In summary, the developed targeted ^{13}C flux ratio approach offers an unprecedented medley of advantages. First, due to the generalization power of machine learning, its application is not limited to a specific organism, experimental conditions, type of input data or flux ratios. Second, the embedded surrogate modeling allows to reduce assumptions on the metabolic network structure, and to perform analysis of poorly determined systems. Third, estimating local ratios and *in silico* testing ensures extraction of most flux information even from small amount of measurements. Finally, once the flux ratio predictors are built, their application is rapid and scalable for high-throughput analysis. Taken together, this approach is able to address local questions in complex setups, such as bacterial co-cultures, higher cells in complex media or host-microbe systems, and is best suited for targeted hypotheses validation and high-throughput flux screening. We believe that our approach, alone or in combination with global flux analysis methods, will open up new horizons in ^{13}C fluxomics, and advance future biological discoveries in cellular metabolism and its regulation.

Zusammenfassung

Der Stoffwechsel oder Metabolismus ist ein unverzichtbarer Prozess für jedes Lebewesen. Es beschreibt die chemische Umwandlung aufgenommener Nahrung in Biomassebausteine, Redox-Äquivalente, Energie und Abfallprodukte. Das metabolische Potential einer Zelle kann durch ein Netzwerk dargestellt werden, in dem die verschiedenen Metabolite durch chemische Reaktionen miteinander verbunden sind. Diese chemischen Reaktionen wiederum werden durch Enzyme katalysiert, die im Genom der Zelle codiert sind. Lebende Zellen passen ihre metabolischen Reaktionsraten, auch metabolische Flüsse genannt, an die vorgefundenen Umweltbedingungen, den intrazellulären Status und die zellulären Anforderungen an. So gesehen widerspiegeln metabolische Flüsse das eigentliche Zellverhalten und deren Bestimmung ist daher essentiell um metabolische Prozesse von biotechnologischer und biomedizinischer Wichtigkeit zu verstehen und zu kontrollieren.

Metabolische Flüsse können nicht direkt gemessen werden und müssen deshalb indirekt von messbaren Grössen, wie beispielsweise Genexpression, Proteinlevels, oder Veränderungen der Metabolitkonzentrationen abgeleitet werden. Experimentelle Ansätze, die mit stabilen Massenisotopen markierte Substrate verwenden, werden speziell für Flussbestimmunen eingesetzt. Dabei werden die Massenisotope der verschiedenen Metabolite mittels Kernspinresonanzspektroskopie oder Massenspektrometrie quantifiziert und miteinander verglichen. Die Interpretation der resultierenden Massenisotopenverteilungen erfolgt entweder manuell, mittels analytischen Formeln oder eines mathematischen Modells und gibt Aufschluss über die metabolischen Flüsse, die den gemessenen Isotopenverteilungen zu Grunde liegen.

Metabolische Flussanalysen können in globale ^{13}C Flussanalysen, welche iterativ diejenige Flussverteilung suchen, die die gemessenen Isotopenverteilungen am besten reflektiert und lokale ^{13}C Flussanalysen, die das Verhältnis mehrerer Flüsse zueinander mittels spezifisch dafür hergeleiteten analytischen Formeln bestimmen. Erstere sind allgemein anwendbar, benötigen aber umfangreiche Messdaten, resultieren in mehreren möglichen Lösungen und sind rechnerisch aufwendig. Im Gegensatz dazu sind lokale Flussanalysen rechnerisch schnell, einfach anzuwenden,

und resultieren in einer einzigen Lösung. Allerdings ist deren Anwendung auf einige wenige metabolische Knoten und experimentelle Voraussetzungen beschränkt. Daher besteht die Nachfrage für eine schnelle, robuste und allgemein anwendbare Methode, mit deren Hilfe auch die heutzutage typisch grossen Datensätze paralleler Experimente analysiert werden können.

In dieser Arbeit stellen wir SUMOFLUX, eine konzeptionell neue und allgemein anwendbare Methode für ^{13}C Flussanalyse spezifischer Flussverhältnisse vor. Diese Methode basiert auf Maschinellem Lernen um bestimmte metabolische Flussverhältnisse von gemessenen Daten vorauszusagen. Dazu wird zuerst ein *in silico* Datensatz modelliert –„**Su**rrogate **Mo**delling“. SUMOFLUX kann auf jeden Netzwerktyp, jedes Substrat und alle möglichen Messdaten angewandt werden und ist nach dem Erstellen der sogenannten „Predictors“ mittels *in silico* Modellierung rechnerisch sehr schnell. Im **2. Kapitel** führen wir SUMOFLUX ein und zeigen dass die berechneten metabolischen Flüsse in *Escherichia coli* mit den Resultaten globaler und lokaler ^{13}C Flussanalysen im Einklang sind. Weiter demonstrieren wir die Skalierbarkeit unserer Methode durch ihre Anwendung auf 121 Transkriptionsfaktorenmutanten in *Bacillus subtilis* und wir illustrieren, wie die Methode auch für die Optimierung des Experimentaldesigns verwendet werden kann.

Ein bemerkenswerter Vorteil unseres Ansatzes ist seine mögliche Anwendung auf komplexe Systeme, auch wenn nur wenige Daten von Messungen und zum metabolischen Netzwerk verfügbar sind. Diesen Vorteil nutzten wir im **3. Kapitel** aus, in dem wir den Aminosäurestoffwechsel von Mykobakterien unter axenischen Wachstumsbedingungen und während der Infektion von Makrophagen untersuchten. Wir teilten Aminosäuren aufgrund ihrer *in vitro* Aufnahme durch *Mycobacterium tuberculosis* und *Mycobacterium smegmatis* und ihrer metabolischen Verwendung, charakterisiert durch die errechneten Flussverhältnisse, in verschiedene Klassen ein. Untersuchungen des metabolischen Verhaltens von *M. tuberculosis* während der Infektion von Makrophagen, wenn weder der Stoffaustausch mit dem Wachstumsmedium noch derjenige zwischen den beiden Organismen bekannt sind, waren mittels extensiver SUMOFLUX Simulationen aller möglicher Stoffflüsse im kombinierten metabolischen Netzwerk von Makrophagen und dem Pathogen möglich. Diese Simulationen zeigten, dass die zum Wachstum normalisierte Biosynthese bestimmter Aminosäuren während der Infektion im Vergleich zu *in vitro* Bedingungen reduziert wird. Aus diesen Daten schlossen wir, dass einerseits nur

bestimmte Aminosäuren für die im Phagosom replizierenden Tuberkelbakterien zugänglich sind und dass andererseits auch Aminosäuren zur Energiegewinnung verstoffwechselt werden. Unsere Resultate unterstreichen die Anpassungsfähigkeit des mykobakteriellen Metabolismus und erklären zum Teil die wiederholten Misserfolge vergangener Versuche chemotherapeutisch den Stoffwechsel dieses Pathogens zu inhibieren.

Im **4. Kapitel** nutzten wir die hohe Geschwindigkeit und Flexibilität von SUMOFLUX um mittels FIA-TOF (flow injection - time of flight) ^{13}C Daten mit hohen Durchsatz zu messen und zu analysieren. Dies resultierte in mehreren Hundert Flussanalysen pro Tag. Diese hohe Messgeschwindigkeit ist der Tatsache zu verdanken, dass vollends auf chromatographische Trennung verzichtet wird, da die Proben direkt ins Massenspektrometer injiziert werden, was zu fehlenden und überlappenden Isotopologen führt. Dieses Problem lösten wir durch die Entwicklung von spezifischen Datenfiltern und Qualitätskontrollen eines jeden Massensignals. Wir validierten die Kombination von SUMOFLUX mit FIA-TOF erfolgreich mittels Analyse von mehreren *E. coli* Mutanten mit bekannten Flussverteilungen. Anschliessend wandten wir das entwickelte Protokoll auf 60 *E. coli* Phosphorylierungsmutanten an um die Rolle dieser in Prokaryoten wenig untersuchten posttranslationellen Modifizierung zu untersuchen. Unsere Flussanalyse bestätigte die bereits beschriebene Inhibition der Isocitratedehydrogenase durch deren Phosphorylierung und identifizierte verschiedene neue funktionelle Proteinphosphorylierungen. Der entwickelte Ansatz zur Flussanalyse mit hohem Durchsatz stellt ^{13}C Fluxomics messtechnisch endlich auf Augenhöhe mit anderen Omics-Analysen.

Die Vorteile der entwickelten Methode lassen sich wie folgt zusammenfassen: Erstens, dank der allgemeinen Gültigkeit Maschinellen Lernens ist die Anwendung unserer Methode weder auf spezifische Organismen, experimentelle Bedingungen, Inputdaten noch bestimmte Flussverhältnisse limitiert. Zweitens, die eingebettete *in silico* Modellierung erlaubt *a priori* Annahmen zur metabolischen Netzwerkstruktur zu reduzieren und Flussanalysen auch für schlecht annotierte Netzwerke durchzuführen. Drittens, Berechnungen von lokalen Flussverhältnissen und deren extensive *in silico* Prüfung maximiert die Menge an Flussinformationen bei einem Minimum an notwendigen Messdaten. Letztlich, sobald die Modellierung abgeschlossen ist und die „Predictors“ berechnet sind, ist Integration der Messdaten

sehr rasch und kann mit dem Durchsatz modernster Messmethoden Schritt halten. Zusammenfassend lässt sich sagen, dass unser Ansatz fähig ist, lokale Flussverhältnisse in einem komplexen Setup zu beantworten, wie beispielsweise bakterielle Co-Kulturen, höhere Zellen in heterogenen Wachstumsmedien, und Wirts-Gast-Wechselwirkungen. Er eignet sich am besten zur Beantwortung gezielter Fragestellungen und Screens mit hohem Messdurchsatz. Wir sind überzeugt, dass der hier entwickelte Ansatz alleine, oder in Kombination mit globalen Flussanalysen, eine neue Ära von ^{13}C Fluxomics Analysen einläutet und künftig einen Beitrag zur Erforschung des Stoffwechsels und dessen Regulation leisten wird.