Doctoral Thesis

Tools for genome-level recoding of in vivo and in vitro production systems

Author(s):
Oesterle, Sabine

Publication Date:
2017

Permanent Link:

https://doi.org/10.3929/ethz-b-000199414 →

DISS. ETH NO. 24551

TOOLS FOR GENOME-LEVEL RECODING OF *IN VIVO* AND *IN VITRO* PRODUCTION SYSTEMS

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

SABINE ÖSTERLE

MSc ETH Interdisciplinary Sciences, ETH Zurich
born June 28th, 1988
citizen of Germany

accepted on the recommendation of

Prof. Dr. Sven Panke (ETH Zurich, Switzerland), examiner
Prof. Dr. Victor de Lorenzo (CNB, Madrid, Spain), co-examiner
Prof. Dr. Jörg Stelling (ETH Zurich, Switzerland), co-examiner

2017

ABSTRACT

Metabolic engineering is commonly used to improve biological production systems. Often, the generation and screening of huge libraries is indispensable to the identification of improved candidates. However, this screening is laborious and time consuming. Smart designs, which (i) cover a broad range of the experimental space, (ii) in which variants are uniformly distributed in this space and (iii) for which the library size is compatible with the screening throughput provide an efficient way to circumvent this issue. In this thesis tools are provided to reduce the screening effort for different applications by either reducing the bias which is created when a library is integrated into the genome of a potential production strain or by fully rationally predicting permissive sites in proteins, which can be used for small protein tag insertion. These tools were developed for facilitated genome editing, which has become a major engineering tool in biotechnology over the recent years. Briefly, with multiplex automated genome engineering (MAGE) mutations encoded on small single- or double-stranded DNA molecules are introduced as an Okazaki fragment during replication with the help of a phage protein. More recently, genome edits have become selectable by CRISPR/Cas9 counterselection. Due to their major impact on strain engineering and their critical importance for the practical aspects of this thesis, these recent developments are reviewed in chapter 1.

Metabolic engineering can be performed on *in vivo* and on *in vitro* systems, often based on the model bacterium *Escherichia coli*. While *in vivo* systems can be used to efficiently produce most metabolite-based products, *in vitro* systems can be used to couple biological and chemical synthesis in a one-pot reaction, allow for the use and formation of non-natural and/or toxic educts and products or for the application of non-cytoplasmic conditions. For both types of systems, cell free extract (CFX) or living organisms, the production pathway needs to be optimized, e.g. in terms of the stoichiometry of participating enzymes. A key element for pathway optimization are ribosome binding sites (RBSs). RBS strength regulates the protein translation level and therefore can be used to modify fluxes in a given pathway. To find the optimal RBS strength, small smart libraries of modified RBSs can be used and inserted by λ-Red and CRISPR/Cas9-facilitated genome editing, but their insertion is subject to bias due to DNA mismatch repair in response to genome editing. We develop and test a protocol called genome library optimized sequences (GLOS) that uses the known substrate profile of *E. coli*'s DNA mismatch repair to adapt input RBS libraries so that they are no longer recognized and this can be used for unbiased genome editing (chapter 2).

While RBSs can be engineered for up- and down-regulation of fluxes of interest to optimize *in vivo* and *in vitro* pathways, there is a natural limitation to the scope of this method: essential reactions cannot be downregulated extensively, as this would lead to cell death. To produce CFX, *E. coli* needs to be able to grow and therefore requires the essential gene. In an *in vitro* system, essential proteins are not required. However, they often interfere with the reaction of interest *in vitro*. For example, part of the F_1F_0- ATP-Synthase, important for energy generation in a living cell, detaches from the membrane during CFX production and continues to hydrolyze ATP in CFX applications, the regeneration of which is one of the main impediments for the application of such systems. Such proteins need to be removed at the *in vitro* stage of the experiment, for example by protein switching. A switchable protein is equipped with a tobacco etch virus (TEV) protease cleavage site at a functionally unobtrusive location in its amino acid sequence. Upon addition of a TEV protease, these proteins are cleaved and rendered nonfunctional. We inserted TEV cleavage sites into the genome of *E. coli* into different subunits of the ATP-synthase and other ATP- and ADP-degrading enzymes, evaluated protein functionality with growth assays and showed enzyme deactivation by stabilization of nucleotides in CFX, illustrating the potential of the method (chapter 3).

Ideally, protein switching should be applied to large number of proteins to easily optimize CFX, which requires a reliable high throughput method to detect permissive and accessible (taggable) insertion sites for small protein tags. Since the state of the art methods are either based on random transposome mutagenesis or individual evaluation of protein structure and known functionalities, they are very time consuming. Therefore, we developed GapMiner, a tool that predicts putatively taggable internal sites based on evaluation of four criteria: sequence and length variability, relative surface accessibility and preservation of secondary structure. GapMiner gives the user a list of tagging sites sorted by predicted taggability, which reduces the experimental workload by minimizing the number of sites that need to be tested. GapMiner does not require any knowledge about the protein except for its sequence, however, if available, an expert user can use additional knowledge to reduce the number of sites to test even further. We tested GapMiner on a randomly selected set of five essential proteins of *E. coli* and found could insert two different tags into these five proteins and only one strain showed growth deficits. For four test proteins both tags worked, illustrating the usefulness of the tool (chapter 4). GLOS and GapMiner both integrate available knowledge to either produce small smart RBS libraries or predict internal protein tagging sites. Both tools will help to facilitate further engineering projects.

ZUSAMMENFASSUNG

Metabolic Engineering ist eine weitverbreitete Methode, um biologische Produktionssysteme zu verbessern. Hierzu werden oft riesige Bibliotheken an Stammvarianten generiert. Um daraus den besten Kandidaten identifizieren zu können, müssen diese mit sehr großem Arbeits- und Zeitaufwand analysiert werden. Oft reicht die Kapazität des Screeningsystems jedoch nicht aus, daher versucht man, mit gut durchdachten Designs die Größe dieser Bibliotheken zu reduzieren (Smart Designs). Um die Vielfalt in der Bibliothek nicht zu verlieren, entfernt man möglichst repetitive Designs aus der Ursprungsbibliothek. Smart Designs müssen einen großen Bereich des zu optimierenden experimentellen Parameters abdecken, die einzelnen Varianten sollten uniform über den Parameterbereich verteilt sein und die finale Bibliothekgröße sollte der Kapazität des Screeningsystems entsprechen.

Im Mittelpunkt dieser Arbeit stehen deshalb verschiedene Methoden, welche den Screeningaufwand für verschiedene Anwendungen verkleinern. Im ersten Teil reduzieren wir die Verzerrung, die entsteht, wenn Bibliotheken funktionaler DNA Sequenzen direkt in ein bakterielles Genom integriert werden. Im zweiten Teil entwickeln wir mit einem komplett rationalen Ansatz eine computergestützte Anwendung, die es erlaubt, Positionen zum Einfügen funktionaler Peptide in Proteine vorherzusagen, sodass das Protein auch nach dem Einfügen noch funktional ist. Diese Anwendung wurde für die erleichterte Sequenzeditierung von Genomen entwickelt, welche sich in den letzten Jahren zu einem wichtigen Ingenieurwerkzeug in der Biotechnologie entwickelt hat. Kurz gesagt, mit multiplexem automatisierten Genom-Engineering (MAGE) können Mutationen, die auf kleinen einzel- oder doppelsträngigen DNA-Molekülen kodiert sind, während der Zellteilung mit Hilfe eines Phagenproteins als Okazaki-Fragment eingebaut werden. Seit kurzem kann man die Effizienz des Genom-Engineering durch CRISPR / Cas9-Gegenselektion noch weiter verbessern (CRMAGE). Aufgrund ihrer großen Bedeutung für die Stammentwicklung und ihrer Wichtigkeit für die praktischen Anwendungen in dieser Arbeit, haben wir diese jüngsten Entwicklungen dieser Technologie in Kapitel 1 zusammengefasst.

Solche Ingenieursmethoden finden sowohl beim Modulieren von biotechnologischen *in vivo* als auch *in vitro* Systemen Anwendung. Während *in vivo* Systeme effizient für die Produktion von Produkten, welche *E. coli* Metaboliten ähneln, genutzt werden können, gibt es bei *in vitro* Systemen weitaus mehr Möglichkeiten. Es können zum Beispiel biologische und chemische Synthesemethoden kombiniert werden, nicht-natürliche und/oder toxische Edukte verwendet werden und/oder Produkte produziert werden oder die Reaktionsbedienungen verändert werden. Beim Arbeiten mit lebenden Zellen sind diese

Möglichkeiten sehr begrenzt, da die Bedingungen für lebende Zellen sehr schnell toxisch werden. Damit lebende Zellen oder zellfreie Systeme als Produktionssysteme verwendet werden können, müssen beide für ihr Ziel optimiert werden.

Eine wichtige Möglichkeit, den Fluss durch einen Stoffwechselweg zu verändern, ist die Modifizierung ribosomaler Bindungsstellen (RBS) auf mRNA Molekülen. So kann die Affinität zwischen RBS und Ribosomen durch die Veränderung einzelner Basen beeinflusst werden, wobei eine höhere Affinität die Proteinbiosynthese positiv beeinflusst. Da die optimale RBS-Stärke vorab unbekannt ist, versucht man diese mit Hilfe von smartem Design zu optimieren. Um die resultierenden Bibliotheken direkt auf dem Genom von *E. coli* zu integrieren, verwenden wir CRMAGE. Zu Problemen führt dabei das unterschiedliche Erkennen und Reparieren der modifizierten Basen durch das DNA-Reparatursystem von *E. coli*, da sich diese neuen DNA-Abschnitte in ein bis sechs Basen zur ursprünglichen DNA unterscheiden. Dies führt zu einer niedrigeren Effizienz der Integration und einer unerwünschten ungleichen Häufigkeitsverteilung der einzelnen Sequenzen nach der Integration. Unser neu entwickeltes Protokoll, welches wir „Genome Library Optimized Sequences“ (GLOS) nennen, verhindert diese ungleiche Verteilung, da keine der eingefügten Sequenzen vom DNA-Reparatursystem erkannt wird, sodass die Vielfalt der Bibliothek gewährleistet ist (Kapitel 2).

Während wir RBSs dazu verwenden können, um die Translation vieler Proteinen zu regulieren, ist diese Möglichkeit für essentielle Proteine limitiert, da hier die Translationsrate nur bedingt verringert werden kann, denn die Zelle ist auf die korrekte Produktion dieser Protein angewiesen. Für die Anwendung von *in vitro* Systeme sind wir nicht auf diese Proteine angewiesen, zur Produktion der zellfreien Extrakte, auf denen die *in vitro* Systeme aufbauen, jedoch schon. Daher können wir auf die Gene für essentielle Proteine auch für *in vitro* Produktionssysteme nicht verzichten. Oft sind es aber genau diese Enzyme, die uns in einem Produktionssystem stören. Ein Beispiel hierfür ist die F_1F_0- ATP-Synthase, die in lebenden Zellen wichtig für die Energiegewinnung ist. Sie wird im zellfreien Extrakt nicht mehr in die Membran integriert und hydrolysiert ATP unspezifisch. Nucleotide, insbesondere ATP und ADP, werden in fast allen *in vitro* Produktionssystemen gebraucht und müssen deshalb zugegeben und mit komplexen Systemen regeneriert werden. Mit „schaltbaren Enzymen“ gelang es uns, solche essentiellen Enzyme erst in der *in vitro* Phase zu entfernen, indem wir in die Aminosäuresequenz des Enzyms ein kurzes Peptidstück („Tag“) eingefügt hatten, ohne dass dadurch die Funktionalität des Enzyms verloren ging. Dieser Peptidtag ist in unserem Fall die Erkennungssequenz für eine Protease des Tobacco Etch Virus (TEV). Durch Expression oder Zugabe der TEV Protease

werden Proteine mit Erkennungssequenz geschnitten und damit deaktiviert. Da ADP und ATP im zellfreien Extrakt schnell abgebaut werden, aber wichtig für fast alle Produktionssysteme sind, versuchten wir die Nucleotide durch Inaktivierung der Nucleotid-abbauenden Enzyme zu stabilisieren. Durch das Einfügen des Peptid-Tags und nachfolgendes Zerschneiden einzelner ATP- und ADP-abbauender Enzyme konnte die Stabilität des Nukleotidpools verbessert werden (Kapitel 3). Diese Methode eröffnet uns aber auch die Möglichkeit, andere Nebenreaktionen sowie Kofaktor- und Energieträger-abbauende Reaktionen auszuschalten. Diese Anwendung setzt jedoch eine Methode voraus, die es erlaubt, schnell und möglichst ohne detailliertes Wissen in vielen verschiedenen Zielenzymen eine geeignete Position zum Einsetzen eines Tags zu identifizieren. Alle derzeitigen Methoden, die dafür in Frage kämen, sind entweder sehr arbeitsintensiv, wie zum Beispiel Transposon-Mutagenese, oder individuell auf das Zielprotein zugeschnitten. Um ein schnelles und universelles Beschicken von Enzymen mit Peptiden („Taggen“) von Enzymen zu ermöglichen, haben wir GapMiner entwickelt, ein selbstlernenden Algorithmus, welcher auf die vier Eigenschaften Sequenz- und Längenvariabilität, relative Oberflächenzugänglichkeit und Erhaltung der Sekundärstruktur trainiert ist. GapMiner ermöglicht es, potentielle Tag-Positionen vorherzusagen, sodass Tag und Enzym nach Einfügen noch funktional sind. Hierfür braucht das Programm nur die Aminosäuresequenz des Proteins (Kapitel 4).

Diese beiden Methoden, GLOS und GapMiner, werden es ermöglichen, in der Zukunft Produktionssysteme effizienter zu entwickeln.