Analysis and engineering of biomolecules and microorganisms: from genome-scale study of pathogens to programming of DNA and cells

Vom Promotionsausschuss der
Technischen Universität Hamburg-Harburg
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von
Lifu Song

aus
Shandong, China

2018

Dissertation Committee

Chairman: Prof. Dr.-Ing. habil. Dr. h.c. Stefan Heinrich
Supervisor & Examiner: Prof. Dr. rer. nat. habil. An-Ping Zeng
Examiner: Prof. Dr. rer. nat. habil. Christoph Wittmann
Examination date: 25-May-2018

DOI: 10.15480/882.1668

In memory of my mother

Acknowledgements

There are many people without whom this thesis would not have been possible. First, I would like to thank my supervisor, Prof. An-Ping Zeng, for providing me this opportunity to pursue my research interests. I would also like to thank Dr. Wei Wang, who basically supervised the first part of this thesis. I really appreciate her hard efforts on improving my manuscripts sentence by sentence. Thanks to both Prof. Zeng and Dr. Wei Wang for many pieces of advice about how to write scientific papers.

Next, I would like to thank my dissertation committee members. A big thank to Prof. Heinrich for agreeing to be the chairman and waiting a long time for my examination. I appreciate Prof. Wittmann for agreeing to review my thesis, the huge efforts to reach Hamburg to attend my examination and the kindness during the examination.

Then, many thanks to Dr. Sugima Rappert and Dr. Wael Sabra who have always been kind and helpful with all kinds of issues in the lab. I thank Prof. Ralf Pörtner for being always kindness and the help with contract issues. Thanks to Ms. Cornelia Hoffmann for her help with kinds of document stuff. Many thanks to Mr. Ralf Grajetzki and Mr. Olaf Schmidt who helped me a lot in setting up the PC and servers. I thank Dr. Uwe Jandt for his insightful discussions and helps with a conference presentation. Thanks to Yaeseong Hong for his help with construction of some plasmids.

After that, a big and special thank to Dr. Ke Wang. Although her major is quite different with mine, I do benefit a lot from the discussions with her. Furthermore, her kind encourages made me came through the darkest days of my life. I appreciate Dr. Chengwei Ma, as well as his wife - Ying Liu, for their hosting during the time waiting for the examination, insightful discussions, and many other bits of help. Thanks to Dr. Ying Dong for the help regarding the examination procedure and the words comforting me. I would also like to thank all my colleagues and friends, Anibal Mora, Anna Gorte, Birgit Koch, Christiane Goepfert, Christin Groeger, Feng Geng, Jan Bomnüter, Jan Sens, Jin Guo, Libang Zhou, Lin Chen, Rebekka Schmitz, Sibel Ilhan, Tyll Utesch, Yujun Zhang and Minliang Chen, for the help and all kinds of enjoyable discussions, making years of living in a foreign country an unforgettable experience.

Additionally, I would like to thank Bundesministerium für Bildung und Forschung (BMBF) for the financial support.

Last but not least, I would like to thank my families. Thank my parents for their support. Thank my wife for her love, patience, support, and understanding throughout my Ph.D. studies. Thank my little daughter who has no idea how wonderful my life becomes with her presence.

Abstract

This thesis is consisted of three major but different parts with the general aims of systems level evaluation and engineering of biomolecules and biological systems. In the first part of this thesis, comparative genomic studies of mutans streptococci strains, which are involved in the development of dental caries, were performed for better understanding their pathogenicity at the level of systems biology. A mosaic-like structure of genome arrangement was revealed by genome alignment analysis. Genes related to pathogenicity were found to have high variations among the strains, whereas genes for oxidative stress resistance are well conserved, indicating the importance of this trait in the dental biofilm community. Genome-scale metabolic network analysis revealed significant differences in 42 pathways. A striking dissimilarity is the unique presence of two lactate oxidases in *S. sobrinus* DSM 20742, probably indicating an unusual capability of this strain in producing H_2O_2 and expanding its ecological niche. In addition, lactate oxidases may form a unique energy-producing pathway with other enzymes in *S. sobrinus* DSM 20742 that can remedy its deficiency in citrate utilization pathway. An "open" pan-genome was inferred by pan-genome analysis using 67 *S. mutans* genomes currently available including the strains sequenced in this study. An online regulation database for *S. mutans*, named StrepReg, was constructed by integrating a transcription factor-based gene regulatory network, which was derived from time-series transcriptome analysis, with STRING protein-protein interaction information and KEGG pathway information (http://biosystem.bt1.tu-harburg.de:1555/homes/).

Although systems biology is a powerful tool in understanding the system level behaviors of biological systems, the establishment of predictive, multiscale models in systems biology is still a challenge due to the complexity of biological systems. For the same reason, mathematical models often fail in applications under physiological conditions, such as for identification of targets in metabolic engineering for the development of highly production strains. In the second part of this thesis, a novel multiple input-output (I/O) system was therefore proposed and verified, which allows the identification of limiting bioreactions or key enzymes in metabolic pathways and even the optimization of biomolecules in vivo. The basic idea is to design a multiple I/O system which can introduce various genetic manipulations (perturbations) into the cells and record the specific intracellular signal changes

correspondingly. This was achieved by engineering the interactions of phage with *E. coli* cells. Specifically, a multiple I/O system was implemented using M13 phage derivatives which can introduce various perturbations into *E. coli* cells after infection, such as up- or down-regulation of specific gene expressions. Using a rationally designed biological circuit, the intracellular signal changes after introduction of the perturbations by the phage infection were linked to the phage reproduction process. This means, signal changes caused by specific perturbations are linked to the specific populations of phages introducing the corresponding perturbations. In this way, the various signals are 'recorded' in forms of corresponding populations of phage derivatives. The usefulness of the multiple I/O system was demonstrated with three applications, i.e. identification of beneficial genetic manipulations, parallel evaluation of various designs of enzymes, and parallel screening of key enzymes for L-lysine biosynthesis in *E. coli*. Various gene operations related or not related to L-lysine biosynthesis in *E. coli* were used as inputs and the intracellular lysine concentration changes were used to trigger output signals. Correct predictions of beneficial genetic manipulations for enhanced lysine production in *E. coli* were achieved. New and effective variants of a key enzyme aspartate kinase III (AK-III), which is strictly inhibited by L-lysine, were obtained and evaluated in parallel. Importantly, the I/O system shows a ultra-sensitivity in capturing signal changes caused by the certain perturbations introduced. The approach developed in this work opens up new possibilities in systems metabolic engineering and synthetic biology of industrial microorganisms for practical applications.

In the third part of this thesis, a novel self-error-detecting, three-base block encoding scheme (SED3B), which takes full advantage of the inherent redundancy feature of DNA synthesis for error correction, was proposed for reliable information encoding in DNA of living cells. In addition to the high error tolerance, SED3B encoded sequences were shown to be orthogonal to natural DNA sequences, indicating for the first time a low biological relevance of the encoded sequences. Features such as effective error tolerance and low biological relevance make SED3B an appealing solution for orthogonal information encoding in living cells with low or no affections to their biological functions, e.g. as a comment language in programming cells in vivo and for biological barcode encoding. Based on error-prone PCR experiments it was estimated that more than 12,000 years of continuous replication would be required to make the SED3B encoded information in *E. coli* cells become unrecoverable. To facilitate the usage of SED3B as a comment and barcode encoding system in synthetic biology, an online encoding-decoding system was implemented and released at http://biosystem.bt1.tu-harburg.de/sed3b. In principle, SED3B is also applicable for in vitro large data storage in synthesized DNA. Although further investigation is required, preliminary analysis shows that SED3B has a great potential for increasing the storage density

to over several exabytes (EBs) per gram DNA which is theoretically much higher than that of methods reported in literature so far.

ZUSAMMENFASSUNG

Diese Doktorarbeit besteht aus drei Hauptteilen mit dem Generalziel, Biomoleküle und biologische Systeme auf Systemebene zu analysieren bzw. zu programmieren. Im ersten Teil dieser Arbeit wurden vergleichende genomische Untersuchungen von Mutans - Streptokokken - Stämmen, die an der Entstehung von Karies beteiligt sind, durchgeführt, um deren Pathogenität auf systembiologischer Ebene besser zu verstehen. Genom-Alignment ergab eine mosaikartige Struktur der Genomanordnung. Gene, die mit der Pathogenität in Zusammenhang stehen, weisen hohe Variationen unter den Stämmen auf, wohingegen Gene für die Resistenz gegen oxidativen Stress gut konserviert sind, was die Bedeutung dieses Merkmals in der dentalen Biofilm-Gemeinschaft anzeigt. Die Analyse genomweiter metabolischer Netzwerke zeigte signifikante Unterschiede in 42 Signalwegen. Eine bemerkenswerte Besonderheit ist die einzigartige Anwesenheit von zwei Lactatoxidasen in *S. sobrinus* DSM 20742, was wahrscheinlich auf eine ungewöhnliche Fähigkeit dieses Stamms hinweist, H_2O_2 zu produzieren und seine ökologische Nische zu erweitern. Zusätzlich können Lactatoxidasen einen einzigartigen energetischen Weg mit anderen Enzymen in *S. sobrinus* DSM 20742 bilden, der seinen Mangel im Citratverwertungsweg beheben kann. Unter Verwendung von derzeit verfügbaren 67 *S. mutans*-Genomen, einschließlich der in dieser Studie sequenzierten Stämme, wurde die theoretische Kerngenomgröße von *S. mutans* geschätzt und eine Modellierung von *S. mutans* pan-genom durch Anwendung verschiedener Fitting-Modelle durchgeführt. Ein "offenes" Pan-Genom wurde gezeigt. Eine Online-Regulierungsdatenbank für Streptococcus, genannt StrepReg, wurde durch Integration eines Transkriptionsfaktor-basierten Genregulationsnetzwerkes, das aus einer zeitreihen Transkriptomanalyse in Zusammenarbeit mit Projektpartnern abgeleitet wurde (http://biosystem.bt1.tu-harburg.de:1555/homes/).

Obwohl die Systembiologie ein sehr nützliches Werkzeug ist, um das Systemverhalten von biologischen Systemen zu verstehen, ist die Etablierung von prädiktiven Multiskalenmodellen aufgrund der Komplexität biologischer Systeme immer noch eine große Herausforderung. Aus dem gleichen Grund scheitern mathematische Modelle oft für Anwendungen unter physiologische Bedingungen, wie z.B. bei der Identifizierung von Targets in Metabolic Engineering für die Entwickelung von Hochleistungsproduktionsstämmen. Zur Lösung der Probleme wurde im zweiten Teil dieser Arbeit ein neuartiges Mehrfach Input-Output

(I/O) System vorgeschlagen und verifiziert, das verschiedene genetische Manipulationen in die Zellen einbringen und die entsprechenden intrazellulären Signaländerungen aufzeichnen kann, mit dem Ziel, Schlüsselreaktionen bzw. Enzyme in Stoffwechselwegen in *E. coli* zu identifizieren und Biomoleküle zu optimieren. Die Grundidee dabei war, die Interaktionen von Phagen mit *E. coli*-Zellen zu gestalten und zu nutzen. Konkret wurde ein Mehrfach-I/O-System unter Verwendung verschiedener M13-Phagenderivate implementiert, die verschiedene genetische Modifikationen (Störungen) in *E. coli*-Zellen nach einer Phageninfektion einführen können, wie etwa eine Aufwärts- oder Abwärtsregulierung spezifischer Genexpressionen. Unter Verwendung eines rational entworfenen biologischen Schaltkreises wurden die intrazellulären Signalveränderungen nach der Einführung von Störungen durch Phageninfektion mit dem Phagenreproduktionsprozess verknüpft. Dies bedeutet, dass Signaländerungen, die durch spezifische Störungen verursacht werden, mit den spezifischen Phagenpopulationen verbunden sind, die die entsprechenden Störungen einführen. Mit anderen Worten werden die verschiedenen Signale in Formen von entsprechenden Populationen von Phagenderivaten "aufgezeichnet". Die Nützlichkeit des Mehrfach-I/O-Systems wurde in drei Anwendungen gezeigt, d.h. Identifizierung von vorteilhaften genetischen Manipulationen, paralleler Bewertung verschiedener Designs von Biomolekülen und parallelem Screening von Schlüsselenzymen für die L-Lysin-Biosynthese in *E. coli*. Verschiedene Genoperationen, die mit der L-Lysinbiosynthese in *E. coli* verwandt waren oder nicht, wurden als Inputs verwendet und die intrazellulären Lysinkonzentrationsänderungen wurden verwendet, um Ausgangssignale auszulösen. Korrekte Vorhersagen von vorteilhaften genetischen Manipulationen für eine erhöhte Lysinproduktion in *E. coli* wurden erzielt. Neue und effektive Varianten eines Schlüsselenzyms Aspartatkinase III (AK-III), das durch L-Lysin streng gehemmt wird, wurden parallel erhalten und ausgewertet. Es ist anzumerken, dass das I/O-System eine besonders hohe Empfindlichkeit bei der Erfassung von Signaländerungen aufweist, die durch die eingeführten bestimmten Störungen verursacht werden. Der in dieser Arbeit entwickelte Ansatz eröffnet neue Möglichkeiten in Systems Metabolic Engineering und synthetischer Biologie industrieller Mikroorganismen für praktische Anwendungen.

Im dritten Teil dieser Arbeit wurde ein neuartiges selbstfehlererkennendes Drei-Basen-Block-Codierungsschema (SED3B) für eine zuverlässige Informationscodierung in DNA, insbesondere für Anwendungen in lebenden Zellen vorgeschlagen und verifiziert, das die inhärente Redundanz der DNA-Synthese zur Fehlerkorrektur in der DNA-Datenspeicherung voll ausnutzt. Zusätzlich zu der hohen Fehlertoleranz wurde gezeigt, dass SED3B-codierte Sequenzen sich von den natürlich gebildeten DNA-Sequenzen grundsetzlich unterscheiden, was zum ersten Mal eine geringe biologische Relevanz der zu diesem Zweck codierten Sequenzen anzeigt. Merkmale, wie die effektive Fehlertoleranz und die geringe biolo-

gische Relevanz, machen SED3B zu einer ansprechenden Lösung für die orthogonale Informationscodierung in lebenden Zellen mit geringen bzw. keinen Beeinträchtigungen ihrer biologischen Funktionen, z. als Kommentarsprache beim Programmieren von Zellen in vivo und für ein biologisches barcoding. Basierend auf einem fehleranfälligen PCR-Experiment wurde geschätzt, dass mehr als 12.000 Jahre kontinuierlicher Replikation erforderlich wären, um die SED3B-codierte Information in *E. coli*-Zellen zu verlieren. Um die Verwendung von SED3B als Kommentar- und Barcode-Kodierungssystem in der synthetischen Biologie zu erleichtern, wurde ein Online-Kodierungs-Dekodierungssystem implementiert und unter http://biosystem.bt1.tu-harburg.de/sed3b veröffentlicht. Im Prinzip ist SED3B auch für eine in vitro große Datenspeicherung in synthetisierter DNA anwendbar. Obwohl weitere Untersuchungen erforderlich sind, zeigen erste Ergebnisse, dass SED3B ein gutes Potenzial zur Erhöhung der Speicherdichte auf mehrere extaabytes (EBs) pro Gramm DNA hat, was theoretisch viel höher ist als bei den bekannten Methoden für digitale DNA-Informationskodierung.

Table of contents

List of figures **xix**

List of tables **xxi**

Nomenclature **xxiii**

1 Introduction and objectives **1**

1.1 Genome-scale comparative studies of mutants streptococci 1

1.2 A multiple input-output system for systems metabolic engineering in *E. coli* cells 2

1.3 Development of an orthogonal information encoding scheme for reliable information encoding in DNA of living cells 4

2 Materials and methods **7**

2.1 Methods for systems biology analysis 7

2.1.1 Genome sequences and strains 7

2.1.2 Genome sequencing, assembly and annotation 8

2.1.3 Genome alignment 9

2.1.4 Pan-genome and core-genome analysis 9

2.1.5 Gene content-based comparative analysis of 10 mutans streptococci strains 10

2.1.6 Identification of putative two-component signal transduction systems 10

2.1.7 Genome-scale metabolic networks construction 11

2.1.8 PCR verification of unique genes in the comparative genomics studies 12

2.1.9 Construction of lactate oxidase encoding gene knockout mutants and transformation of *S. sobrinus* DSM 20742 13

2.2 Methods for multiple input-output system 14

2.2.1 Chemicals 14

2.2.2 Bacterial strains 14

2.2.3 Phagemids, plasmids and primers 14
2.2.4 Media 15
2.2.5 Strain conservation 16
2.2.6 Molecular cloning 17
2.2.7 Preparation of infective engineered phages 18
2.2.8 Screening based on cell-phage interactions 18
2.2.9 Enzyme characterization 19
2.3 Methods for orthogonal information encoding in living cells 20
2.3.1 Detailed steps for encoding binary data into DNA string 20
2.3.2 Decoding error-containing DNA strings into binary data 20
2.3.3 Implementation of the online encoding-decoding system for SED3B 21
2.3.4 Analysis of error tolerance by *in silicon* simulation 21
2.3.5 *In vivo* verification of the error tolerance by error-prone PCR 21

3 Genome-scale comparative studies of mutans streptococci 25
3.1 Introduction 25
3.2 Genome sequencing, assembly and annotation of eight mutans streptococci strains 26
3.3 Genome rearrangement of *S. mutans* genomes 30
3.4 Core and pan-genome analysis of *S. mutans* species 31
3.4.1 Core-genome 32
3.4.2 Pan-genome 33
3.5 Gene content-based comparative analysis of mutans streptococci strains . . 34
3.5.1 Distribution of two-component signal transduction systems 37
3.5.2 High diversities of the competence development regulation module 44
3.5.3 Distribution of bacteriocin- and antibiotic resistance-related proteins 49
3.5.4 Oxidative stress defense systems in mutans streptococci 57
3.6 Metabolic network construction and analysis 61
3.6.1 Genome-scale metabolic network reconstruction 61
3.6.2 Variability and specificity in metabolic pathways and network . . . 62
3.7 Construction of StrepReg - a regulation database of *S. mutans* 64
3.8 Conclusion 65

4 Development of a multiple IO system for biological engineering in *E. coli* 71
4.1 Introduction 71
4.2 Principles of a multiple input-output system which can interact with *E. coli* cells 72

4.3 Proof of concept studies . . . 75
4.3.1 Identification of beneficial genetic manipulations . . . 76
4.3.2 Evaluation of designs . . . 77
4.3.3 Parallel and sensitive screening of biomolecules . . . 78
4.4 Conclusion and Perspective . . . 87

5 Orthogonal information encoding in living cells 89
5.1 Introduction . . . 89
5.2 Theoretical and technological backgrounds . . . 91
5.2.1 The method of Church *et al.* . . . 92
5.2.2 The method of Goldman *et al.* . . . 92
5.2.3 The method of Grass *et al.* . . . 92
5.3 Principles of a self-error-detecting, three-base block encoding scheme (SED3B) 95
5.4 High error tolerance revealed by *in silicon* simulations . . . 96
5.5 SED3B encoded DNA sequences show low biological relevance . . . 98
5.6 SED3B encoded DNA sequences show simple secondary structure . . . 100
5.7 Reliable orthogonal information encoding in living cells using SED3B . . . 102
5.8 *In vitro* data storage using SED3B . . . 103
5.9 Development of an online encoding-decoding system . . . 106
5.10 Conclusion . . . 106

6 Summary and outlook 109

References 113

Appendix A Source codes of utilized perl scripts 139
A.1 panGenomeAnalysis.pl . . . 139
A.2 shared.pl . . . 150
A.3 bin2DNA.pl . . . 152
A.4 Consensus.pl . . . 154
A.5 DNA2bin.pl . . . 155
A.6 kmerAnalysis.pl . . . 157
A.7 biologyRelevanceAnalysis.pl . . . 160
A.8 bin2DNACRCIndex.pl . . . 162

Appendix B Supplement Information 167
B.1 Sequences of mutacins used for the identification of putative mutacins in 10 mutans streptococci strains. . . . 167

List of figures

1.1 Design–Build–Test Cycle for Biomolecular and Biosystems Engineering . . 3

2.1 Detailed steps of decoding error-containing DNA strings into error free bit string. 22
2.2 The logo of our institute used as input for error tolerance simulation 23
2.3 Illustration of construction process of plasmids carrying the encoded 78bp DNA string with variant errors introduced by error-prone PCR 23

3.1 Phylogenetic analysis of 10 mutans streptococci strains compared in this study and their phylogenetic relationship to other *Streptococcus* species. . . 28
3.2 Comparison of local collinear blocks (LCBs) of chromosomal sequences of the eight *S. mutans* strains. 31
3.3 Core and pan-genome model of 67 *S. mutans* genomes. 35
3.4 Alignment of ComC and ComS amino acid sequences. 48
3.5 Cluster structure of the mutacin-K8 production system across six *S. mutans* strains. 50
3.6 Example of visualized genome-scale metabolic networks constructed based on genome annotations and KEGG pathway 67
3.7 Glycolysis/Gluconeogenesis and TCA cycle pathway in mutans streptococci 68
3.8 Screenshots of StrepReg database 69

4.1 Sample plasmid maps of inputting phages carrying out overexpression/repression operation on specific genes 73
4.2 Principle of phage based multiple IO system 74
4.3 Proof of concept application studies of the IO system 75
4.4 Illustration of the output device using the concentration of intracellular lysine as an output signal 76
4.5 Inhibition profiles of wild-type and mutants of AK-III by lysine 78
4.6 Work flow of cell robot based screening by using the IO system 82

4.7 3D structure illustration of the de-allosteric regulation mechanism of R300C mutein . . . 85
4.8 Flow Cytometry assays of cell populations harboring wild-type AK-III and AK-III mutants of R300C and V339A . . . 88

5.1 GF(47) to DNA codon wheel for mapping every element of GF(47) to three nucleotides . . . 93
5.2 Illustration of encoding binary data into DNA string using the SED3B encoding scheme. . . . 96
5.3 Error detection and repression by using the SED3B encoding scheme. . . . 97
5.4 Error correction capabilities by multiple DNA sequences encoded by SED3B encoding scheme. . . . 98
5.5 Simulation of required sequence numbers for reliable information recovery by DNA fragments with variant rates of errors. . . . 99
5.6 Comparative analysis of SED3B encoded sequences with a natural DNA sequences. . . . 100
5.7 The number of complementary matched k-mers is reduced remarkably by using the SED3B scheme. . . . 101
5.8 Correct information can be retrieved using 14 sequences with high rates of errors introduced by error-prone PCR. . . . 102
5.9 Schematic presentation of *in vitro* information storage in DNA . . . 104
5.10 Strategy of *in vitro* digital information encoding in DNA using SED3B . . . 105
5.11 Screenshots of the online encoding-decoding system . . . 107

List of tables

2.1 Eight newly sequenced and two previously sequenced mutans streptococci strains included in the analysis 8
2.2 *E. coli* strains used in the present work 14
2.3 Plasmids used in present work 16
2.4 Primers used in present work 17

3.1 Genome assembly and annotation of eight *S. mutans* strains sequenced in this study in comparison with previously sequenced *S. mutans* strains UA159 and NN2025 29
3.2 Unique protein coding sequences (CDSs) between the different strains revealed by ortholog analysis 36
3.3 Identification and classification of putative two component systems in the eightmutans streptococci strains sequenced in this study 38
3.4 Ortholog analysis and classifications of the putative TCS proteins 40
3.5 Distribution of competence development-related systems in the 10 mutans streptococci strains 46
3.6 Distribution of mutacins and mutacin immunity proteins in the 10 mutans streptococci strains 53
3.7 Distribution of antibiotic resistance-related proteins 56
3.8 Distribution of oxidative stress resistance systems 60
3.9 Compositions of the established metabolic networks of the 10 mutans streptococci strains 61

5.1 Base-3 to DNA encoding ensuring no repeated nucleotides in the Goldman's method 93
5.2 Comparison of capabilities of current available encoding schemes for digital information storage in DNA 94

Nomenclature

Roman Symbols

Ω	Theoritical core-genome size

Acronyms / Abbreviations

γ-GCS-GS	γ-Glutamylcysteine synthetase-glutathione synthetase
dpr	Dps-like Peroxide Resistance gene
k-mer	All the possible substrings of length k that are contained in a string
lysC	Gene encoding aspartokinase
AK-III	Aspartokinase III
ATP	Adenosine triphosphate
BCP	Bacterioferritin comigratory protein
Cas9	CRISPR associated protein 9
CDSs	Protein Coding Genes
CMKM	Complementarily matched *k-mer* pairs
CoA	Coenzyme A
COGs	Clusters of Orthologous Groups of proteins
CRC	Cyclical Redundancy Check
CRISPR	Clustered regularly interspaced short palindromic repeats
CSP	Competence stimulating peptide

DNA	Deoxyribonucleic Acid
EB	Exabyte
eYFP	Enhanced Yellow Fluorescent Protein
FACS	Fluorescence-activated cell sorting
G3P	Phage minor coat gene 3 protein
GF(47)	Galois Field of size 47
GS	Glutathione synthetase
GSH	L-γ-Glutamyl-L-cysteinylglycine
GSSG	Oxidized Glutathione
HGT	Horizontal gene transfer
HK	Histidine Kinase
HMM	Hidden Markov Model
HO•	Hydroxyl radical
HTS	High Throughput Screening
IO	Input-Output
IPTG	Isopropyl-β-D-Thiogalactoside
LCBs	Locally collinear blocks
LTA	Lipoteichoic acid
ML	Maximum Likelihood
multi-MUMs	Multiple Maximal Unique Matches
NAD	Nicotinamide adenine dinucleotide
NAD^+	Oxidized form of Nicotinamide adenine dinucleotide
NADH	Reduced form of Nicotinamide adenine dinucleotide
$NADP^+$	Oxidized form of Nicotinamide Adenine Dinucleotide Phosphate

NADPH	Reduced form of Nicotinamide Adenine Dinucleotide Phosphate
PB	Petabyte
PCR	Polymerase Chain Reaction
PEP	Phosphoenolpyruvate
PT	Petabytes
PTS	Phosphotransferase system
RBS	Ribosomal binding site
ROS	Reactive oxygen species
RR	Response Regulators
rRNA	Ribosomal ribonucleic acid
RS	Reed–Solomon
SED3B	Self-error-detecting, three-base block encoding scheme
SOD	Superoxide dismutases
TCA	Tricarboxylic Acid
TCS	Two-component signal transduction system
TF	Transcription factor
TG	Target gene
TM	Transmembrane helix
tRNA	Transfer Ribonucleic Acid
V-ATPases	V-type ATPases

Chapter 1

Introduction and objectives

This thesis is based on work done during my stay at the Hamburg University of Technology as a scientific coworker. It is consisted of three major but different parts with the general aims of systems level evaluation and engineering of biomolecules and biological systems. In the following, the background and objectives of each part are briefly introduced. More detailed introduction and background information are presented in the corresponding chapter for each part.

1.1 Genome-scale comparative studies of mutants streptococci

The oral microbiome is a dynamic environment inhabited by both commensals and pathogens. Among them, mutans group streptococci are considered as significant contributors to the development of dental caries [1]. This is attributed to their ability to form biofilms which are generally difficult or impossible to eradicate by antibiotic therapy because biofilm cells are resistant to antibiotics [2]. Systems biology is a holistic approach to decipher the complexity of biological systems. It is based on the understanding that live biological networks that form the whole of living organisms are more than the sum of their parts [3–5]. Systems biology studies try to design predictive, multi-scale models to discover new biomarkers for disease, drug targets, to understand pathogenicity mechamisms and to develop high performance producers in industrial biotechnology. It has been responsible for some of the most important developments in the science of biology [6–15]. In the first part of this thesis, systems biology efforts were made to understand the pathogenicity of ten mutans streptococci strains. Due to the high diversity of genetic content of different isolates, genome contents of single or just few isolates cannot represent specific species or group of strains.

Among all the species of mutans group streptococci, only the genomes of two strains of *S. mutans* were sequenced previously. In the frame of a collaboration systems biology project, six *S. mutans* strains, one *S. ratti* strain and one *S. sobrinus* strain were submitted for genome sequencing. Genome annotation, genome level comparative analysis and metabolic network analysis were performed in this work to reveal strain-specific features and potential drug targets. An online transcriptional regulatory network database of *S. mutans*, named StrepReg, was constructed by integrating time-resolved transcriptomic data from the project partners (http://biosystem.bt1.tu-harburg.de:1555/homes/). All the information and tools should be helpful for understanding the evolution and pathology of these oral pathologies.

1.2 A multiple input-output system for systems metabolic engineering in *E. coli* cells

Systems biology is a fast developing discipline making significant contributions to other disciplines. Systems biology strategy has been applied to metabolic engineering, enabling a new state-of-art technology termed 'systems metabolic engineering' [16–21]. The key challenges of metabolic engineering have been the time-, cost- and labor-intensive processes of strain development owing to the difficulties in understanding the complex interactions among the metabolic, gene regulatory and signaling networks inside the cells, which are collectively represented as overall system performance under industrial fermentation conditions. To avoid laborious try-and-error manner experiments, systems biology studies have focused on building genome-scale models of cellular functions to make predictions. However, due to the complexity of cellular functions and the technical/biological variations in omics data, establishment of predictive, multiscale models is still quite challenging. Indeed the complexity issue not only occurs on whole cell-level, it was observed even on single gene level. In consequence, effective engineering of biological parts or systems, regardless the scale of the target system, requires extensive studies and efforts in the form of design–build–test cycles as shown in Figure 1.1, in which many designs are evaluated and the process is iterated in order to improve the performance. The rate of improvements is directly related to the throughput and rounds of the design cycles.

Although cells are composed of molecules and their viability relies on extracting and using energy to maintain them, they are not 'just' matter and energy. Information processing, also called "cellular computing", is essential for cellular function. Previous studies proved that the computational abilities of biological system could be utilized in rational ways. Here, the computation abilities of cells were proposed to be utilized for systems-level prediction

and optimization of biomolecules and microorganisms. The key issue to do so is how to let cells "compute" the processes of interests and output the corresponding results to the different inputs. In other words, a multiple input-output (IO) system is required to interact with cells. In this thesis, interactions of M13 bacteriophage with *E. coli* were employed for implementation of the multiple IO system. The input system was implemented by using various M13 phage derivatives which can carry out up or down regulations targeting different genes. By a rationally designed biological circuit, the signal changes within the cells after gene operations executed by phage infection are linked to the phage reproduction process, which are in turn linked to the populations of different types of phage. In other words, the various signals are 'recorded' in forms of the populations of corresponding phage derivatives. The populations of various phages could be determined easily by sequencing. This novel IO system was utilized to aid systems metabolic engineering of L-lysine biosynthesis as a model system. For proof of concept, the IO system was demonstrated for identification of beneficial genetic manipulations, parallel evaluation of designs and parallel screening of key enzymes for effective lysine biosynthesis which represent some of the most key efforts in systems metabolic engineering.

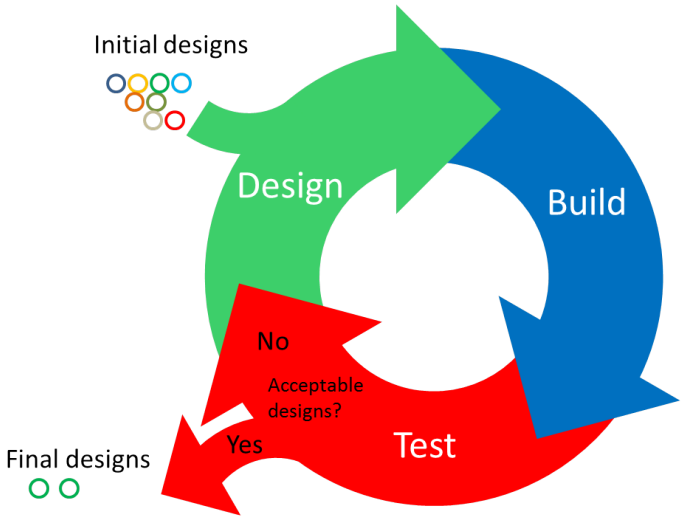


Fig. 1.1 Design–Build–Test Cycle for Biomolecular and Biosystems Engineering

The designs are initially placed within an organism that has not been optimized for specific purpose such as chemical production or logical operation. The first pass through the design step of the cycle may involve varying the levels of gene expression or exploration of mutations in enzyme activity sites. These designs are implemented through DNA synthesis and/or cloning technologies during the build step. In the test step, the newly constructed designs are evaluated for their performance. The designs with favorable performance are retained and used as starting point for the next round of design. The cycle is iterated until a design is found that meets the requirements.

1.3 Development of an orthogonal information encoding scheme for reliable information encoding in DNA of living cells

We live in the age of information explosion which imposes a big challenge to data storage technologies [22]. The presently used storage media such as magnetic tape or hard disk drivers have a decisive shortcoming of limited lifetime and density, e.g. around 50 years for hard disk drivers. The recent studies of Church *et al.* and Goldman *et al.* opened up a new and exciting possibility of storing digital information in synthetic DNA [23, 24]. Goldman *et al.* achieved an information density of 2.2 petabytes (PT)/gram DNA which is far above the current commercial technologies. Besides the advantage of high density, information storage in DNA has additional attractive features such as ultra-long lifetime and low maintenance requirements [23, 24]. However, unlike other planner storage media, relatively high rate of errors could be introduced to stored digital data by complex "writing" and "reading" processes of information storage in DNA, especially if fast and cheap synthesis and sequencing technologies are applied [25, 26]. The error rate can be even higher if the encoded DNA contains sequences with extreme GC contents, long homopolymers or complex secondary structures which are hard to be synthesized or sequenced [23, 24].

Previous studies dealt with information encoding in DNA outside living cells. It is also of interest to know if DNA data storage or information encoding in living cells are feasible and reliable. This should enable applications such as for biological barcodes of engineered biological parts (Biobricks) and as comment "language" in "programming biology" in the emerging area of synthetic biology [27]. Theoretically, the encoding schemes designed for *in vitro* data storage in DNA are also applicable for *in vivo* applications. However, to the best of our knowledge, no reported work has addressed the issue of increasing errors introduced by DNA replication. This issue is crucial for *in vivo* applications since DNA replication happens constantly under *in vivo* conditions. Furthermore, the artificial DNA fragments could interfere with the native and natural ones (being so-called biologically relevant). This is another issue which has not been studied so far. For *in vivo* applications, such as biological barcodes or comments encoding in living cells, the encoded DNA sequences should not share the same sequence space as the natural ones to avoid interference with cellular functions. In other words, they should be orthogonal to exclude biological relevance. One unique feature of information storage in DNA is that there are always many copies of DNA molecules synthesized while data writting by DNA synthesis. In other words, a high data redundancy is inherently generated during this process. In this study, we sought to design an encoding

scheme by taking advantage of the inherent redundancy feature for effective error correction with additional consideration of the biological relevance, homopolymers and extrem GC content issues.

Chapter 2

Materials and methods

2.1 Methods for systems biology analysis

2.1.1 Genome sequences and strains

Serotype c strain *S. mutans* 5DC8 was isolated from root caries by David Beighton (London, UK); serotype *c* strain *S. mutans* AC4446 was isolated from a proven case of infective endocarditis in Dillingen (Germany), serotype *c* strain *S. mutans* KK21 was isolated from enamel caries of an adult by Susanne Kneist (Jena, Germany), serotype *c* strain *S. mutans* KK23 was isolated from enamel caries of a child by Susanne Kneist (Jena, Germany), Serotype *c* strain *S. mutans* ATCC 25175 was isolated from carious dentine, serotype *f* strain *S. mutans* NCTC 11060 was isolated in Denmark from a patient's blood, serotype *b* strain *S. ratti* DSM 20564(=ATCC 19645) was isolated from caries lesion in rat, and finally, serotype non-d & non-g strain *S. sobrinus* DSM 20742 (= ATCC 33478) was isolated from human dental plaque. Serotype *c* is over-represented because 70-80% of all *S. mutans* isolates are of this serotype. However, non-*c* serotypes seem to be associated with cardiovascular diseases and this is represented in our study by the serotype *f* strain. Besides *S. mutans*, *S. sobrinus* is considered as a relevant cariogenic species in human. The genome sequences of *S. mutans* UA159 and *S. mutans* NN2025 were sequenced previously and obtained from NCBI genome database (http://www.ncbi.nlm.nih.gov/genome/). They were used in this study as reference genomes for the genome analysis. All used strains are listed in Table 2.1.

Some parts of the "Materials and Methods" presented here have been taken or modified from publications (Song *et al.* 2012, Song *et al.* 2013; Song *et al.* 2017a; Song *et al.* 2017b) with me as the first author. The experiments for Section 2.1 were carried out by project partner(s) as specified in the corresponding publications.

Table 2.1 Eight newly sequenced and two previously sequenced mutans streptococci strains included in the analysis

Species	Strain	Short description
S. mutans	UA159	The first sequenced *S. mutans* strain [28].
S. mutans	NN2025	The genome sequences of NN2025 was release in 2009 [29].
S. mutans	5DC8	Serotype *c*, isolated from root caries by David Beighton (London, UK), alterations in 16S sequence in comparison to type strain.
S. mutans	AC4446	Serotype *c*, isolated from a proven case of infective endocarditis in Dillingen (Germany).
S. mutans	KK21	Serotype *c*, isolated from enamel caries of an adult by Susanne Kneist (Jena, Germany), potent producer of bacteriocin.
S. mutans	KK23	Serotype *c*, isolated from enamel caries of a child by Susanne Kneist (Jena, Germany), potent producer of bacteriocin.
S. mutans	ATCC25175	Type strain, serotype *c*, isolated from carious dentine, quality control strain.
S. mutans	NCTC11060	Serotype *f*, isolated in Denmark from a patient's blood (bacteremia), reference strain.
S. ratti	DSM20564	Type strain (= ATCC19645), serotype *b*, isolated from caries lesion in rat, nearest neighbor to species *S. mutans* with a 94-95% similarity on 16S level.
S. sobrinus	DSM20742	Type strain (= ATCC33478), serotype non-*d* & non-*g*, isolated from human dental plaque, 93% similarity with *S. mutans* on 16S level; considered as a relevant cariogenic species in human.

2.1.2 Genome sequencing, assembly and annotation

The eight mutans streptococci strains listed above in Table 2.1 as mentioned earlier were sequenced by a shotgun strategy using the Solexa sequencing platform at the Helmholtz Center for Infection Research in Braunschweig, Germany. The "high-quality draft" [30] genome sequences of these mutans streptococci strains were assembled by a combined use of the sequence assembly tools SOAPdenovo [31], Maq [32] and Phrap [33]. All genomes were annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP, http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html) and the whole-genome shotgun sequences have been deposited at DDBJ/EMBL/GenBank under the accessions of AOBX00000000 (*S. mutans* 5DC8), AOBY00000000 (*S. mutans* KK21), AOBZ00000000 (*S. mutans* KK23), AOCA00000000 (*S. mutans* AC4446), AOCB00000000 (*S. mutans* ATCC 25175), AOCC00000000 (*S. mutans* NCTC 11060), AOCD00000000 (*S. ratti* DSM 20564) and AOCE00000000 (*S. sobrinus* DSM 20742). Manual curation based on blast searches

using known coding nucleotide sequences were performed to complement some missing coding genes.

2.1.3 Genome alignment

Multiple genome alignments have been computed using the progressive Mauve algorithm of the Mauve software [34] with default options.

2.1.4 Pan-genome and core-genome analysis

In addition to the six *S. mutans* draft genomes of this study and the previously released complete genomes of *S. mutans* UA159 and NN2025, 59 *S. mutans* genomes (2 completed and 57 drafts) available in NCBI till April 2013 were also included in the core- and pan-genome analysis of *S. mutans*. The accessions of the 59 genomes are as follows:

AGWE00000000, AHRB00000000, AHRC00000000, AHRD00000000, AHRE00000000, AHRF00000000, AHRG00000000, AHRH00000000, AHRI00000000, AHRJ00000000, AHRK00000000, AHRL00000000, AHRM00000000, AHRN00000000, AHRO00000000, AHRP00000000, AHRQ00000000, AHRR00000000, AHRS00000000, AHRT00000000, AHRU00000000, AHRV00000000, AHRW00000000, AHRX00000000, AHRY00000000, AHRZ00000000, AHSA00000000, AHSB00000000, AHSC00000000, AHSD00000000, AHSE00000000, AHSF00000000, AHSG00000000, AHSH00000000, AHSI00000000, AHSJ00000000, AHSK00000000, AHSL00000000, AHSM00000000, AHSN00000000, AHSO00000000, AHSP00000000, AHSQ00000000, AHSR00000000, AHSS00000000, AHST00000000, AHSU00000000, AHSV00000000, AHSW00000000, AHSX00000000, AHSY00000000, AHSZ00000000, AHTA00000000, AHTB00000000, AHTC00000000, AHTD00000000, AHTE00000000, CP003686, AP012336.

Data pre-processing for the core and pan-genome analysis were performed using a self-implemented perl script (the source codes are gaven in Appendix A), which is similar as described previously by Tettelin et al. [35]. Briefly, an iterative procedure was carried out to estimate total genes/core genes to be discovered per additional genome sequenced. The number of total genes/core genes provided by each added new genome depends on the selection of previously added genomes. All possible combinations of genomes from 1 to M (the maximal number of available genomes) were calculated. In the case more than 1000 combinations were possible, only 1000 random combinations were used. In order to take into consideration of core genes that are possibly missed during genome sequencing and assembly, for the calculation of core-genome size, an additional correction step was introduced, in which any one gene that is only absent in one of the 63 draft genomes was

still regarded as core gene. During the fitting step of the core genome model, the inputted genome numbers were used as fitting weight for corresponding data point.

The pre-data processing was performed using recently released pipeline PGAP [36]. The pan-genome size was calculated using a "Power law model" proposed by Tettelin previously [37, 35]. The core-genome model $F_c(n) = k_c exp[^{-n}/\tau_c] + \Omega$ (k_c, τ_c, and Ω are free parameters and Ω means the theoretical core-genome size) proposed by Tettelin et al. was also applied in this study [35].

2.1.5 Gene content-based comparative analysis of 10 mutans streptococci strains

In this work, if not otherwise specified, the uniqueness of genes is defined according to the ortholog groups constructed by using the OrthoMCL program [38]. If the ortholog of a gene from organism A is absent in "organism B", this gene was defined to be unique or specific to organism A in comparison to organism B. However, it does not imply that there is no homolog of this gene in organism B. In some cases, this gene is just an additional copy (namely paralog) of another gene whose alleles/orthologs are found in both organisms. Certainly, it does further not imply that this gene is present only in organism A. For example, the ortholog of this gene may be found in organism C from the relationship table or in other strains or species not compared in this work.

2.1.6 Identification of putative two-component signal transduction systems

The identification of histidine kinases (HKs) and response regulators (RRs) of putative two-component systems (TCSs) of the eight mutans streptococci strains (shown in Table 2.1) was carried out based on computational domain analysis of the predicted protein sequences. Two previously sequenced *S. mutans* strains, the *S. mutans* UA159 and *S. mutans* NN2025, were used as reference strains for comparison. To this end, the same identification procedure was carried out on the genomes of *S. mutans* NN2025 and UA159 to ensure that the same search criteria were applied for all the strains included in this study so that a reasonable comparison can be achieved. The genome sequences of the two reference strains were obtained from the genome database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/sites/genome). Approaches for identifying HKs and RRs were similar to those described previously [39] with slight modifications. Briefly, putative HK and RR proteins were identified by Hidden Markov Model (HMM) searches using the

related HMM profiles available in the Pfam database (http://pfam.sanger.ac.uk/) as templates [40]. The sequence homology search software HMMER3 (http://hmmer.org/) [41] was used for scanning the predicted protein sequences with the HMM profiles. All the HK related HMM profiles with the accession numbers PF00512, PF07568, PF07730, PF07536, PF06580, PF01627, PF02895, PF05384, PF10090 were used for identifying putative HKs. The HMM profile PF00072 which targets the receiver (REC) domain of RR proteins was used to recognize putative RRs. For the identification of HKs, the homology search was performed without setting E-value/score cutoffs to avoid missing any putative HKs with low scores. However, all the identified putative HKs were manually validated by judging whether at least one of the following two criteria was satisfied: (a) the presence of a cognate putative RR in the same operon as the putative HK in question; (b) the presence of both the HisKA-like and HATPase_c domains so that any HATPase_c domain possessing non-HK proteins could be excluded. For the identification of putative RRs, the E-value cut-off was set at 1e-6. Paired HK and RR present in the same operon comprise a TCS cluster. Hybrid HKs, if any, could be determined by the presence of a complete HK transmitter domain and a REC domain in a single protein. If no corresponding cognate RRs or HKs can be found in the same operon, HKs and RRs are defined as orphan HKs or RRs. The operon information used in this study was predicted by Pathway Tools [42].

2.1.7 Genome-scale metabolic networks construction

The bipartite metabolic networks were constructed based on the connection matrix of updated KEGG reactions database according to Stelzer and Zeng [43, 44] with the addition of the newly identified reactions catalyzed by lactate oxidase (Lactate + O_2 => Pyruvate + H_2O_2) with provisional R numbers of R10001 (C00186 + C00007 => C00022 + C00027) and R10002 (C00256 + C00007 => C00022 + C00027). Compared to reaction graph or metabolite graph, wherein either reactions or metabolites (called "node") are shown in an interconnected way, the bipartite network is more understandable because both the reactions and metabolites are visualized at the same time. Seventy-six non-enzymatic automatic reactions were also considered for the network construction. The construction of sub-networks was based on KEGG pathway classification (http://www.genome.jp/kegg/pathway.html) with slight modification by adding lactate oxidase to the glycolysis/gluconeogenesis pathway (MAP00010) and the pyruvate metabolism pathway (MAP00620). The software Cytoscape [45] was used for the visualization and comparative analysis of the genome-scale metabolic networks.

2.1.8 PCR verification of unique genes in the comparative genomics studies

To verify the unique presence of the lactate oxidase (consecutive) coding genes D823_06595 and D823_06598, respectively, in *S. sobrinus* DSM 20742 and to exclude the possibility of contamination with e. g. human DNA during the process of genome sequencing, PCR amplifications (using one primer pair covering both genes) with isolated DNAs from *S. sobrinus* DSM 20742 and a second *S. sobrinus* strain (AC153), as well as from *S. mutans* UA159 and *S. ratti* DSM 20564 (the latter two strains as negative controls) were performed. The primers used were: 5'- GAGCAGGATAATTGACAGTC -3' (forward primer) and 5'-ACTCAGTGACGAATCAGTT -3' (reverse primer), which were designed by using Primer Premier and Vector NTI 9.0 (InforMax), respectively. Conditions for this conventional PCR were: 94°C, 2 min; followed by 32 cycles of 94°C for 30s; annealing temperature 48°C for 30s; and 72°C for 90s; final extension at 72°C for 5 min; length of amplicon 1,175 bp.

To verify the unique presence of TCS-15 in *S. mutans* NCTC11060, PCR amplification with original DNA from this strain using two different forward primers was performed (*S. mutans* UA159 as negative control). The primers used were: 5'-TTGCTTGCTGTTGTTGTG-3' (forward primer), 5'- GGCTACCATTTAGTAGAAAAGAGG -3' (alternative forward primer) and 5'-TGTTACCATCTTCGGAAGG-3' (reverse primer), which were designed by using Primer Premier 6 and Vector NTI 9.0 (InforMax) respectively. Conditions for this conventional PCR were: 94°C, 2 min; followed by 32 cycles of 94°C for 30s; annealing temperature 49°C for 30s; and 72°C for 90s; final extension at 72°C for 5 min; length of amplicons: 1,624 bp and 504 bp, respectively.

To verify the unique presence of TCS-18 and the unique absence of TCS-13 in *S. ratti* DSM 20564, as well as the unique absence of TCS-9 and TCS-3 in *S. sobrinus* DSM 20742, PCR amplifications using original DNAs from *S. ratti* DSM20564, and *S. sobrinus* DSM 20742 was performed (*S. mutans* UA159 as negative control). The primers used, the annealing temperatures and the lengths of amplicons were as follows (all other parameters were kept the same as mentioned above): TCS-18 F 5'-CACTGTTCCTCCTGTATCC 3', TCS-18 R 5'-ATGCTGGCTATGATGTTGT-3'(Tm=50°C, length: 1,899bp covering HK and RR); TCS-13 F 5' RAKTTYATGCCYCTMACYTTYCAG 3', TCS-13 R 5' GATTCRWWRGCMGCCTC 3' (Tm = 49°C, length: 1,600 bp covering HK and RR); TCS-9 HK-F 5' ATACAGTCAATATGCYAAGC 3', TCS-9 HK-R 5' GRATAACACGGAAAA 3' (Tm = 45 C, length: 1,055 bp);

All primers in section 2.1.8 were designed by the author. The experiments in section 2.1.8 and 2.1.9 were performed by a project partner (Anke Brock, Anke.Brock@rwth-aachen.de, Division of Oral Microbiology and Immunology, Department of Operative and Preventive Dentistry Periodontology, RWTH Aachen University, Aachen, Germany).

TCS-9 RR-F 5' TGCTGARGACCAAGA 3', TCS-9 RR-R 5' TTAGCTGCAATTTCTT 3' (Tm = 50°C, length: 522 bp); TCS-3 HK-F 5' CAYGAYYTIMGIAAYCC 3', TCS-3 HK-R 5' GTDATIACIGTICCC 3' (Tm = 40°C, length: 505 bp).

2.1.9 Construction of lactate oxidase encoding gene knockout mutants and transformation of *S. sobrinus* DSM 20742

To clarify the functionality of the two lactate oxidases, namely D823_06598 (Llod) and D823_06595 (lod), PCR ligation mutagenesis according to the method described by Lau *et al.* [46] was used to separately replace the two genes encoding the two enzymes by an erythromycin resistance cassette via double homologous recombination. Primers P1Llod (TTACCGTTATCCGCGAATTAT) and P2Llod (GGCGCGCCAACCACCCAAGGTTGAATC), P1lod (GGCTGGTTTCCTCCATGATA) and P2lod (GGCGCGCCCCAAAACCACCTTGAGGAAT) were used to amplify the 5'flanking regions of both genes, respectively, introducing an AscI restriction site. To amplify the 3'flanking regions of both genes, the primers P3Llod (GGCCGGCCGGGAGCTCAAGGTGTTCAAA) and P4Llod (CAAATTGTTCAAAGCGGGAAC), P3lod (GGCCGGCCGGCAGCAGCCGGTAGTATT) and P4lod (GGGTGCCAACTTATGTCACGA) were used, respectively, thereby introducing restriction site for FseI. The erythromycin resistance cassette was amplified from previously constructed gene deletion mutant [47] using primers ErmFor (GGCGCGCCCCGGGCCCAAAATTTGTTTGAT) and ErmRev (GGCCGGCCAGTCGGCAGCGACTCATAGAAT), containing the restriction site for AscI and FseI, respectively. After digestion with the appropriate restriction enzymes, following purification, the three amplicons were ligated together and used for transformation.

For transformation, two natural transformation methods were first used to assay and optimize the natural transformation of the *S. sobrinus* cells. The first step was the preparation of pre-competent cells of *S. sobrinus* applying the methods according to Lefrancois *et al.* [48] and Ween *et al.* [49]. Afterwards 200 ng of the constructs prepared for mutagenesis were used for the transformation. The plasmids like pDL278 (Spr, pAT18 Emr, and suicide vector pFW5 Spr in both circular and linearized form were used as a positive control. Another transformation protocol according to Li *et al.* [50] applying pheromone CSP of *S. mutans* was additionally used to introduce genetic constructs and plasmids into *S. sobrinus* cells. In this approach two various concentrations of CSP were used: 0.2 and 1μM, respectively. Transformation of *S. mutans* was used as a parallel control. All these experiments were carried out at least three times.

All experiments in section 2.1.9 were performed by a project partner Anke Brock (Anke.Brock@rwth-aachen.de)

Later, electroporation experiment was carried out according to the procedure described by LeBlanc *et al.* [51]. Various pHs of electroporation mix (EPM) [52] as well as various pulsing conditions were tested. The electroporation was carried out by adding to the chilled electrocompetent cells 200 ng of constructs prepared for mutagenesis or plasmids. Other protocol for electroporation according to [53] was also tested.

2.2 Methods for multiple input-output system

2.2.1 Chemicals

Chemicals of analytical grade were purchased from Sigma-Aldrich Chemie GmbH (München, Germany). Other chemicals were purchased from Carl Roth GmbH (Karlsruhe, Germany). Enzymes and other reagents for molecular biology were obtained from Fermentas (St. Leon-Roth, Germany). Kits for site-directed mutagenesis were obtained from Agilent Technologies (Karlsruhe, Germany).

2.2.2 Bacterial strains

E. coli DH5α and TOP10 were used as hosts for normal vectors construction. *E. coli* BL21(DE3) was used for high level protein expression. *E. coli* XL1-Blue (Agilent Technologies) was used for M13 phage infection. The genotypes of *E. coli* strains are listed in Table 2.2.

Table 2.2 *E. coli* strains used in the present work

Strain	Genotype	Description
Top10	*F^- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(Str^R) endA1 λ^-*	Host for normal DNA cloning and transformation
DH5α	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F´ proAB $lacI^qZ\Delta M15$ Tn10 (Tet^R)]*	Host for VCSM13 phage amplification
XL1-Blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F´ proAB $lacI^qZ\Delta M15$ Tn10 (Tet^R)]*	Host for VCSM13 phage amplification
BL21 (DE3)	*B F^- ompT gal dcm lon $hsdS_B(r_B^-m_B^-)$* λ(DE3 *[lacI lacUV5-T7p07 ind1 sam7 nin5]) $[malB^+]_{K\text{-}12}(\lambda^S)$*	Host for protein overexpression

2.2.3 Phagemids, plasmids and primers

The M13 phage (VCSM13) was purchased from Agilent Technology (5301 Stevens Creek Blvd. Santa Clara, CA 95051, USA). The wild *lysC* gene encoding AK-III was amplified by

PCR from the genomic DNA of *E. coli* K12 MG1655. For over-expression and purification of the wild-type AK-III and relevant muteins, the wild-type *lysC* gene was cloned to pET-22b(+) (Novagen, Darmstadt, Germany) with the introduction of an additional His-tag at the C-terminal to generate the plasmid pET22-lysC. Site-mutagenesis was performed on pET22-lysC to generate over-expression plasmids for AK-III muteins. The *lysC* gene was also cloned to VCSM13 by replacing the original gene III to generate a phagemid M13-lysC. Similarly, site-mutagenesis was also performed on M13-lysC to generate phagemid derivations carrying different AK-III muteins.

For construction of plasmid AP-Lys-B, i.e. the device harnessed by the host cells to control the phage packaging process based on intracellular lysince concentration, we ultilized a lysine inducible promoter from *Corynebacterium glutamicum* ATCC13032 as a lysine sensor. The lysine inducible promoter, gene III from M13 phage and a GFP-encoding gene were cloned into the plasmid pZE21MCS to obtain AP-Lys-B. The transcriptional levels of gene III and GFP encoding gene are controlled by the lysine inducible promoter. The antibiotic resistance type of AP-Lys-B was changed to ampicillin resistance by replacing the kanamycin resistance gene with an ampicillin resistance gene.

The plasmids used in this study are listed in Table 2.3.

2.2.4 Media

Complex medium

LB medium

The LB (Luria-Bertani) medium was routinely used for the cultivations of *E. coli* strains. One liter LB liquid medium contained: 10 g tryptone, 5 g yeast extract and 10 g NaCl. LB solid plate was prepared by addition of 15 g/L agar. The pH was adjusted to 7.0 by 5M NaOH. Sterilization was performed at 121°C for 20 min. When necessary, appropriate antibiotics were added to the medium before usage. For *E. coli* strains, the working concentration of ampicillin and kanamycin was 100μg/mL or 50μg/ml, respectively.

SOC medium

The SOC (Super Optimal broth with Catabolite repression) medium is a nutrient-rich medium used for the regeneration of *E. coli* strains after heat shock transformation. For preparation, 20 g tryptone, 5 g yeast extract. 0.5 g NaCl and 0.186 g KCl were dissolved in 975 mL water and autoclaved at 121°C for 20 min. Subsequently, 20 mL filter-sterilized glucose (1M, 0.22 um Ultrafree-MC, Millipore) and 5 mL filter-sterilized $MgCl_2$ (2M, 0.22 um Ultrafree-MC, Millipore) were added into the cooling medium.

2XYT medium

Table 2.3 Plasmids used in present work

Phages	Description/Genotype	Source
VCSM13	*Kan*	Agilent
M13-lysC	Derived from VCSM13 by replacing gene III with wild lysC from *E.coli* K12	This study
M13-srRNA-lysC	Derived from VCSM13 by elimination of gene III and insert small RNA fragment which can inhibit lysC expression in *E. coli*	This study
M13-lysC-T253R	Derived from M13-lysC by site mutagenesis	This study
M13-lysC-R300C	Obtained by screening with a library of M13-LysC generated by in vivo random mutagenesis	This study
M13-lysC-R305A	Derived from M13-lysC by site mutagenesis	This study
M13-lysC-H320A	Derived from M13-lysC by site mutagenesis	This study
M13-lysC-I337P	Derived from M13-lysC by site mutagenesis	This study
M13-lysC-S338L	Derived from M13-lysC by site mutagenesis	This study
M13-lysC-V339A	Derived from M13-lysC by site mutagenesis	This study
pJ175e	*Amp*	Gift from David Group
pJ175e-Str	*Str*	Derived from pJ175e by changing the type of the antibiotic resistance
AP-Lys-B	*Amp*, Derived from pZE21 plasmid;	This study
pET22-lysC	*Amp*, Expression plasmid for wild-type AK-III	This study
pET22-lysC-T253R	Expression plasmid for T253R mutant of AK-III	This study
pET22-lysC-R300C	Expression plasmid for R300C mutant of AK-III	This study
pET22-lysC-R305A	Expression plasmid for R305A mutant of AK-III	This study
pET22-lysC-H320A	Expression plasmid for H320A mutant of AK-III	This study
pET22-lysC-I337P	Expression plasmid for I337P mutant of AK-III	This study
pET22-lysC-S338L	Expression plasmid for S338L mutant of AK-III	This study
pET22-lysC-V339A	Expression plasmid for V339A mutant of AK-III	This study

The 2XYT medium is a nutritionally rich medium for the propagation of M13 bacteriophage. One liter 2XYT liquid medium contained: 16g tryptone, 10g yeast extract and 10 g NaCl. Sterilization was performed at 121°C for 20 min. When necessary, appropriate antibiotics were added to the medium before usage.

2.2.5 Strain conservation

Fresh cells were grown overnight in LB medium to an OD_{600} around 2. Bacteria strains were stored either in glycerol solution or in Roti®-Store cryo-vials (Roth, Karlsruhe, Germany). For the former case, 0.5 ml cultured solution was mixed with 0.5 ml sterile 60% glycerol and stored at -80°C. For the latter case, 0.5 ml culture solution was added into the vial and mixed thoroughly. The supernatant was removed and the tube was stored at -80°C.

Table 2.4 Primers used in present work

Primers	**Description**	**Sequence**
M13Seq-G3-P1	Sequencing primer	TCTGTAGCCGTTGCTACCCTCGTT
M13Seq-G3-P2	Sequencing primer	AAGAAACAATGAAATAGCAATA
M13-ln4Genes-P1	Primer for linearization of VCSM13	CTAGTATTTCTCCTCTTTCTCTAGT ATAATTGTATCGGTT-TATCAGCTTGCT
M13-ln4Genes-P2	Primer for linearization of VCSM13	CTCCCTCAATCGGTTGAATGT
LysC-4M13-P1	For cloning of *lysC*	GAGGAGAAATACTAGATGTCTGAAA TTGTTGTCTCC
LysC-4M13-P2	For cloning of *lysC*	AACCGATTGAGGGAGTTACTCAAAC AAATTACTATG
V339A-P1	Site-directed mutagenesis of *lysC* to generate V339A mutant	GCAGACTTAATCACCACGTCAGAAG
V339A-P2	Site-directed mutagenesis of *lysC* to generate V339A mutant	CGAAATATTATGCCGCGCGAGGATG
T253R-P1	Site-directed mutagenesis of *lysC* to generate T253R mutant	CGTTTTGGTGCAAAAGTACTGC
T253R-P2	Site-directed mutagenesis of *lysC* to generate T253R mutant	TGCCATCTCTGCCGCTTCGGCA
R305A-P1	Site-directed mutagenesis of *lysC* to generate R305A mutant	TGCTCGCAATCAGACTCTGCTC
R305A-P2	Site-directed mutagenesis of *lysC* to generate R305A mutant	AGCGCCAGAGCGCGGAACAGCG
H320A-P1	Site-directed mutagenesis of *lysC* to generate H320A mutant	TTCTCGCGGTTTCCTCGCGGAA
H320A-P2	Site-directed mutagenesis of *lysC* to generate H320A mutant	GCCAGCATATTCAGGCTGTGCA
I337P-P1	Site-directed mutagenesis of *lysC* to generate I337P mutant	CTTCGGTAGACTTAATCACCAC
I337P-P2	Site-directed mutagenesis of *lysC* to generate I337P mutant	GATTATGCCGCGCGAGGATGCC
S338L-P1	Site-directed mutagenesis of *lysC* to generate S338L mutant	TGGTAGACTTAATCACCACGTC
S338L-P2	Site-directed mutagenesis of *lysC* to generate S338L mutant	AAATATTATGCCGCGCGAGGAT
R300C-P1	Site-directed mutagenesis of *lysC* to generate R300C mutant	TGCGCTCTGGCGCTTCGTCGCAATC
R300C-P2	Site-directed mutagenesis of *lysC* to generate R300C mutant	GAACAGCGGCGGATTTTCAGTTTTA

2.2.6 Molecular cloning

Genomic DNA and plasmid extraction

The extraction of genomic DNA from *E. coli* was performed using genomic DNA isolation kit NucleoSpin® Tissue (Macherey-Nagel, Düren, Germany). Fresh colony from agar plate was incubated overnight in LB medium at 37°C One milliliter culture solution was harvested

and the cell lysis was achieved by incubation of the sample in a proteinase K/SDS solution. Cell harvest and DNA purification were performed according to the manual of NucleoSpin® Tissue. Plasmid extraction was carried out by following the standard protocol of NucleoSpin® Plasmid kit (Macherey-Nagel, Düren, Germany).

Mutagenesis

Site-mutagenesis was performed using a protocol similar to the NEB Q5® Site-Directed Mutagenesis Kit. Briefly, none overlap primers were designed and synthesized which contain the desired mutations. Then PCR amplification was performed with the designed primers using the original plasmid as templates to generate linear plasmids. Template DNA was eliminated by enzymatic digestion with DpnI. Finally, phosphorylation and ligation using T4 Polynucleotide Kinase and T4 Ligase were carried out to obtain circular DNA before transformation.

Random *in vivo* mutagenesis was enabled by using the plasmid pJ184-Str harboring genes which can increase intracellular DNA replication error rates. The plasmid pJ184-Str was derived from pJ184 by replacing the chloramphenicol acetyltransferase encoding gene with a streptomycin resistance gene. The pJ184 plasmid which has been described previously was obtained from David R. Liu's group of Harvard Medical School [54].

2.2.7 Preparation of infective engineered phages

Since the engineered phages lack gene III, the helper plasmid pJ175e was harnessed by the host cells to supply gene III products intracellularly to obtain infective phages. The plasmid pJ175e was obtained from David R. Liu's group. Specifically, engineered phages were co-transformed with pJ175e into XL1-Blue cells. Overnight cultures were deposited for centrifuge and the supernatant containing the packaged infective phages was collected.

2.2.8 Screening based on cell-phage interactions

XL1-Blue/AP-Lys-B cells were incubated in LB medium to an OD_{600} value around 1.0. Roughly 200ul XL1-Blue/AP-Lys-B cells were mixed with 2ul proper diluted phages (Cells to phage number ratio above 10:1 to make sure that all phages could be captured and evaluated by host cells. Different types of phages in a total number of roughly 10,000 were used as inputs in the present study). The mixture was incubated at 37°C for 15 minutes without shaking to allow the phages to attach to the cells, following by incubation at 37°C with shaking for 1 to 2 hours. Inactivate the host cells at 65°C for 15min. The cell debris were

spinned down and the supernatant containing the "scored" phages was transferred to a fresh tube. A proper amount of "scored" phages were mixed with fresh XL1-Blue/AP-Lys-B cells and incubated at 37°C for 15 minutes without shaking to allow the host robots to absorb the highly "scored" phages. A proper amount of the culture was then sprayed on LB agar plates with kanamycin (50 mg/ml) for selection.

2.2.9 Enzyme characterization

Enzyme overexpression in *E. coli*

Enzyme overexpression was achieved with pET-22b(+) in *E. coli* BL21(DE3) cells. The recombinant cells bearing the expression vectors were firstly grown in 100 mL LB medium supplemented with appropriate antibiotics (80 μg/ml ampicillin for pET-22b(+) derivates at 37°C. When the OD_{600} of the culture reached 0.6, protein expression was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) in a final concentration of 0.1 mM, and the culture was continued for an additional 12 to 14 h at 30°C. Cells were harvested by centrifuge (10min, 5000rpm, 4°C), washed twice with 20mM Tris-HCl buffer (pH 7.5) and resuspended with 5mL lysis buffer (20 mM Tris-HCl (pH7.5), 150 mM NaCl and 500 mM $(NH_4)_2SO_4$). Cell suspensions were directly submitted for enzyme purification steps within the same day.

Enzyme purification

The supernatant was obtained by centrifugation at 4°C for 1 hour at 13,000 rpm. Targeted proteins with His-tag at C-terminal (pET-22b (+) derivates) were purified by His SpinTrapTM columns (GE Healthcare Bio-Sciences, Piscataway, USA). The protocols from the kits were followed during the purification (twice washing with washing buffer (20mM KH_2PO_4, 500mM NaCl, 20mM Imidazole, pH 7.4) and elution with elution buffer (20mM KH_2PO_4, 500mM NaCl, 500mM Imidazole, pH 7.4)). After the step of enzyme purification, PD MiniTrap G-10 columns (GE Healthcare Bio-Sciences, Piscataway, USA) were used for buffer change (20 mM Tris-HCl (pH7.5), 150 mM NaCl and 500 mM $(NH_4)_2SO_4$). Protein content was determined at 595nm by Bradford method (Bradford, 1976) with a reagent solution from Biorad (Biorad, Hercules, USA) and BSA (Bovine Serum Albumin) standard protein.

Enzyme assay

The parameters of enzyme kinetics were determined by varying the concentrations of substrates. To test the influence of allosteric inhibition, effectors with varied concentrations were additionally added into the standard reaction. The relative activities were calculated by normalizing the specific activities of enzymes under the standard conditions.

The enzyme activity of aspartokinase was detected by using the hydroxamate method [55]. The quantity of aspartate hydroxamate formed in the presence of hydroxylamine was measured at 540 nm. The standard assay reaction mixture in 1 ml contained 200 mM Tris-HCl (pH 7.5), 10 mM $MgSO_4 \cdot 6H_2O$, 10 mM aspartate, 10 mM ATP, 160 mM $NH_2OH \cdot HCl$ (neutralized with KOH), and appropriate amounts of enzyme. After incubation at 30°C for 30 min, the reaction was stopped by mixing with 1 ml of a 5% (wt/vol) $FeCl_3$ solution, and the absorbance at 540 nm was monitored.

2.3 Methods for orthogonal information encoding in living cells

2.3.1 Detailed steps for encoding binary data into DNA string

To detail the steps of encoding arbitrary digital data into DNA string, an arbitrary computer file is represented as a string (S1) of bits (often interpreted as a number between 0 and 1). The detailed steps are illustrated in Figure 5.2 and explained as follows: 1) Bit string S1 is converted to DNA string S2 of characters in A, C, G, T four bits by four bits using the scheme shown in Figure 1 (shown in rows of "Data encoding bases"). 2) One error detecting base was inserted per two bases based on assign rule I shown in Figure 5.2 to generate DNA string of S3. 3) Check presentation of "TTT" three bases by three bases and adapt the error detecting base of blocks that next to "TTT" to a new base based on rule II to generate final DNA string of S4.

2.3.2 Decoding error-containing DNA strings into binary data

The decoding process here refers to restore the original binary data from variant numbers of long error-containing DNA strings. Our encoding scheme does permit detection of deletion and insertion errors which could be achieved by detection of extensive errors emerged in continuous three-base encoding blocks in principle. For proof of concept, here we used a simple error correction decoding process without considering the indels. Additional steps for

indels treatments would enhance the error correction efficiency. The details of the decoding process were illustrated in Figure 2.1 and described as follows:

1. Generate consensus DNA string block by block as follows:

a) Read a three-base block from all DNA fragments and remove all the three-base blocks with errors detected (The rule of error detecting base was initialized with rule I); b) Make consensus block by taking the block that with largest occurrence frequency; c) Switch the error detecting rule to rule II if the consensus block is 'TTT', otherwise, switch to rule I; d) Go to next blocks and repeat a), b), c) steps until the complete consensus DNA string was generated;

2. Transfer the consensus DNA string into bit string based on the scheme shown in Figure 5.2.

2.3.3 Implementation of the online encoding-decoding system for SED3B

The online system is implemented by using CakePHP (https://cakephp.org/) web development framework. Two different applications are provided: comment encoding-decoding and biological barcode encoding-decoding. The system is available under the link: http://biosystem.bt1.tu-harburg.de/sed3b/.

2.3.4 Analysis of error tolerance by *in silicon* simulation

The 35,292 bps DNA string encoding the logo (Figure 2.2) of our institute is used as input for error tolerance simulation. The specific rate of random errors was introduced base by base by giving a specific error probability. The rates of A<->T and G<->C transition errors were doubled to that of A/T<->G/C transition errors to mimic the natural DNA replication process. Variant numbers of DNA sequences with random errors were then used for decoding to test the error tolerance.

2.3.5 *In vivo* verification of the error tolerance by error-prone PCR

To test the error tolerance capability of the SED3B encoding scheme in practical, we encoded text of "Hello, World!" into 78bp DNA string. A 168bp DNA fragment including the 78bp DNA string encoding "Hello, World!" was constructed using two primers of 5'- TCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTTAAGGATGCTCGTGCCCATGCCCATGCCGTAC -3' and 5'-GGCTCGAGCTCGAGACTAGCACCTGGTTTAGCATGGGCAAGTAAAACGGCACAAAAATATGGTTGGGGTACGGCATGGGCATGGGCACGAGCATCCTTAA -3'. We then used

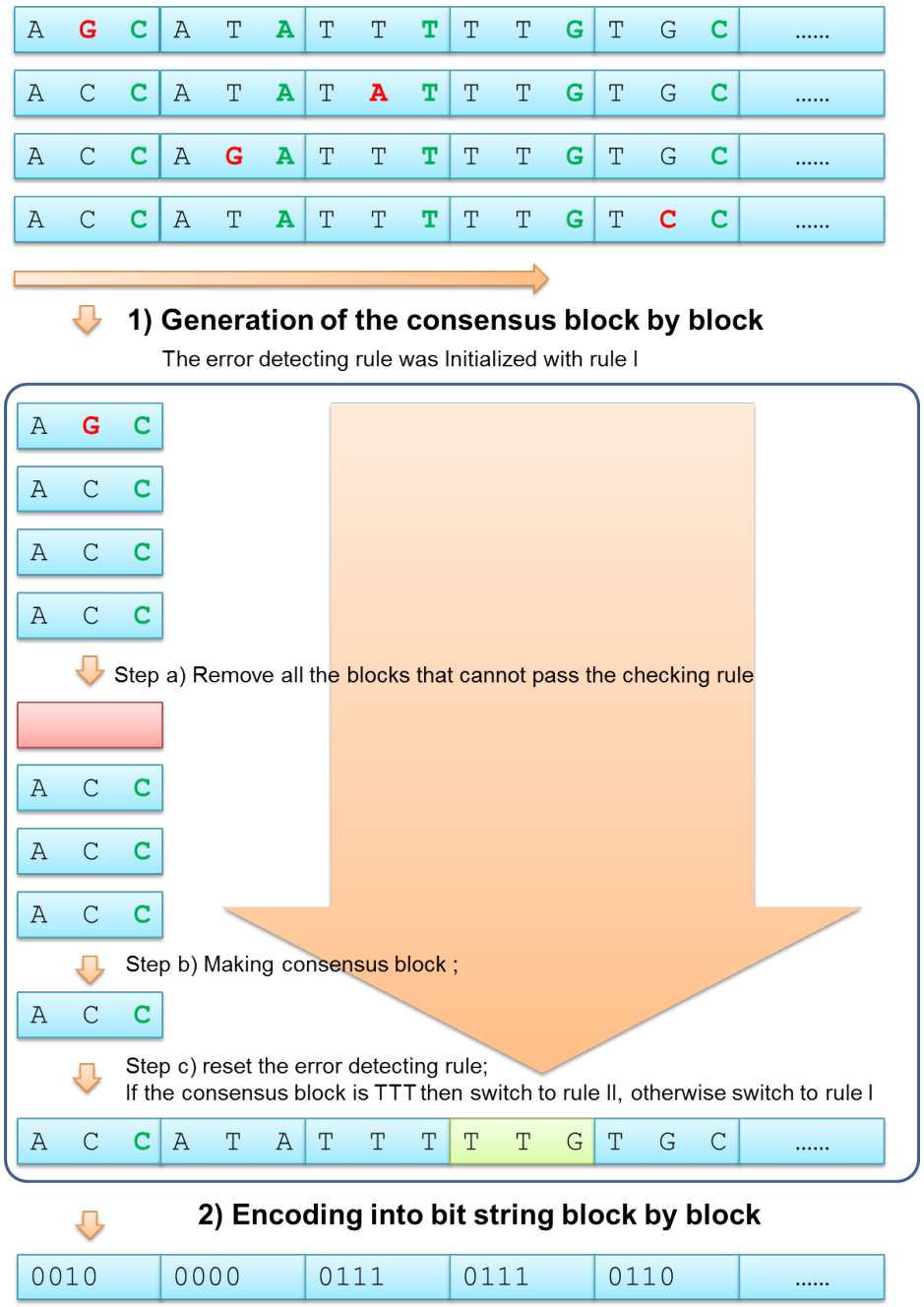


Fig. 2.1 Detailed steps of decoding error-containing DNA strings into error free bit string.

The black, green and red characters stand for the data encoding bases, error correction bases and error containing bases respectively. The encoding scheme does permit detection of insert and deletion errors by detection of continuous errors of encoding blocks. Although the decoding algorithm implemented here didn't involve a frameshift correction process which can increase the accuracy of recovered information in principle, correct information still can be recovered as proved by error-prone PCR experiments.

Fig. 2.2 The logo of our institute used as input for error tolerance simulation

error-prone PCR to introduce random errors into the 168bp DNA fragments. Error-prone PCR was performed using JBS dNTP-Mutagenesis Kit using the recommended protocol with 30 thermal cycles to introduce errors into the encoded DNA string. The amplified fragments by error-prone PCR were ligated with linearized pZE21-MCS plasmid using In-Fusion® HD Cloning Kit from Clontech© Laboratories. The ligation products were transformed into stellar *E. coli* stellar competent Cells. The plasmid map and encoded information are presented in Figure 2.3.

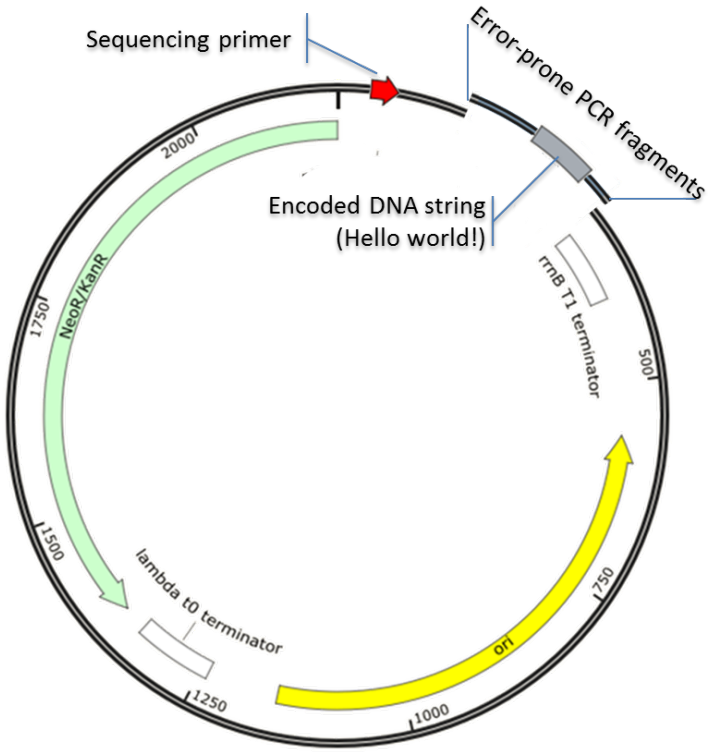


Fig. 2.3 Illustration of construction process of plasmids carrying the encoded 78bp DNA string with variant errors introduced by error-prone PCR

The plasmid abstractions of individual colonies were deposited for sequencing. Original information was recovered using the error rich DNA fragments. Primers used for error-prone amplification of the 168bp insert DNA fragment are 5'-TCTAAGAAACCATTATTATCAT-

3’ and 5’-GGCTCGAGCTCGAGACTAGCA-3’. The primers used for linearization of the plasmid are 5’- TAATGGTTTCTTAGACGTCGGAATTGCCAGCTGGG -3’ and 5’-TCTCGAGCTCGAGCCAGGCATCAAATAAAACGA AAGG-3’. The primer used for sequencing is 5’-GCGAAACGATCCTCATCCTGTCT-3’.

Chapter 3

Genome-scale comparative studies of mutans streptococci

3.1 Introduction

Traditionally and supported by 16S rRNA gene and *rnpB* gene sequence analyses, the genus *Streptococcus* is divided into several groups, with the mutans group streptococci consisting of the species *S. mutans, S. sobrinus , S. ratti , S. criceti, S. downei, S. macacae,* and – but controversially discussed – *S. ferus* [56][i]. Mutans group streptococci are considered as significant contributors to the development of dental caries [1]. By attaching to the tooth surfaces and forming biofilms, they can tolerate and adapt to the harsh and rapidly changing physiological conditions of the oral cavity such as extreme acidity, fluctuation of nutrients, reactive oxygen species, and other environmental stresses [57]. They occasionally also cause bacteremia, abscesses, and infective endocarditis [58, 59]. Many strains of mutans streptococci are genetically competent, i.e. they can take up DNA fragments from the environment and recombine them into their chromosome, an important mechanism for horizontal gene transfer (HGT). The ability of some bacteria to generate diversity through HGT provides a selective advantage to these microbes in their adaptation to host eco-niches and evasion of immune responses [60, 61]. Due to diversities in the genetic contents between

This chapter was modified based on two previous publications:
Song, Lifu; Sudhakar, Padhmanand; Wang, Wei; Conrads, Georg; Brock, Anke; Sun, Jibin *et al.* (2012): A genome-wide study of two-component signal transduction systems in eight newly sequenced mutans streptococci strains. *BMC genomics* 13, S. 128.; Song, Lifu; Wang, Wei; Conrads, Georg; Rheinberg, Anke; Sztajer, Helena; Reck, Michael *et al.* (2013): Genetic variability of mutans streptococci revealed by wide whole-genome sequencing. *BMC genomics* 14, S. 430. Some of the texts, figures, and tables may be directly used without further indication.

[i]For updates please refer to http://www.bacterio.net/s/streptococcus.html

different isolates, the genome content of a single isolate doesnot necessarily represent the genomic potential of a certain species. With the rapid development of DNA sequencing technologies, the steadily increasing genome data enable us to dig the evolutionary and genetic information of a species from a pan-genome perspective. In 2002, the release of the genome sequence of *S.mutans* UA159, the first genome sequence of mutans group streptococci, has greatly helped in understanding the robustness and complexity of *S. mutans* as an oral and odontogenic (e.g. infective endocarditis and abscesses) pathogen [28]. Later, after the genome sequence of *S. mutans* NN2025 became available, a comparative genomic analysis of *S. mutans* NN2025 and UA159 has provided insights into chromosomal shuffling and species-specific contents [29]. Recently, Cornejo *et al.* have studied the evolutionary and population genomics of *S. mutans* based on 57 *S. mutans* draft genomes and revealed a high HGT rate of *S. mutans* [62].

In this study, the whole genome of eight mutans streptococci strains, including six *S. mutans* strains (5DC8, KK21, KK23, AC4446, ATCC25175 and NCTC11060), one *S. ratti* strain (DSM20564) and one *S. sobrinus* strain (DSM20742) were sequenced. A pan-genomic model of mutans streptococci was constructed and analyzed. Cross-comparison of the genome contents of the eight mutans streptococci strains and the previously genome sequenced strains of *S. mutans* UA159 and NN2025 were carried out focusing on the genomic components that are highly related to pathogenicity. Further, by constructing and comparative analysis of genome-level metabolic networks, the diversities in sub-pathways among these strains were systematically investigated. The results are helpful for understanding the evolution and pathogenicity of these oral pathogens, which in turn will be helpful for the clinical management of diseases caused by these pathogens and for the development of diagnostics and new molecular epidemiological methods.

3.2 Genome sequencing, assembly and annotation of eight mutans streptococci strains

Shot-gun sequencing of the six *S. mutans* isolates (5DC8, KK21, KK23, AC4446, ATCC25175, and NCTC11060) as well as *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742 were carried out on a Solexa sequencing platform. An overview of the genome assemblies and annotations of the six *S. mutans* isolates, *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742 is summarized in Table 3.1 in comparison with two previously sequenced *S. mutans* strains, namely UA159 and *S. mutans* NN2025. The average GC contents are in the range of low GC organisms [28]. The genome sizes are very close to each other, with the largest one from *S. sobrinus* DSM

20742 and the smallest one from *S. mutans* KK23 showing merely 5.7% differences. The total numbers of protein-coding sequences per genome are also similar among all the strains compared.

As would be expected, the overall genomic features of the eight *S. mutans* strains (5DC8, KK21, KK23, AC4446, ATCC25175, NCTC11060, UA159 and NN2025) are more similar to each other than to *S. ratti* DSM20564 and *S. sobrinus* DSM20742. This is consistent with the results of the phylogenetic analysis, as visualized by the phylogenetic tree constructed based on 16S rRNA gene sequences and core-genome single-nucleotide polymorphisms (SNPs) shown in Figure 3.1.

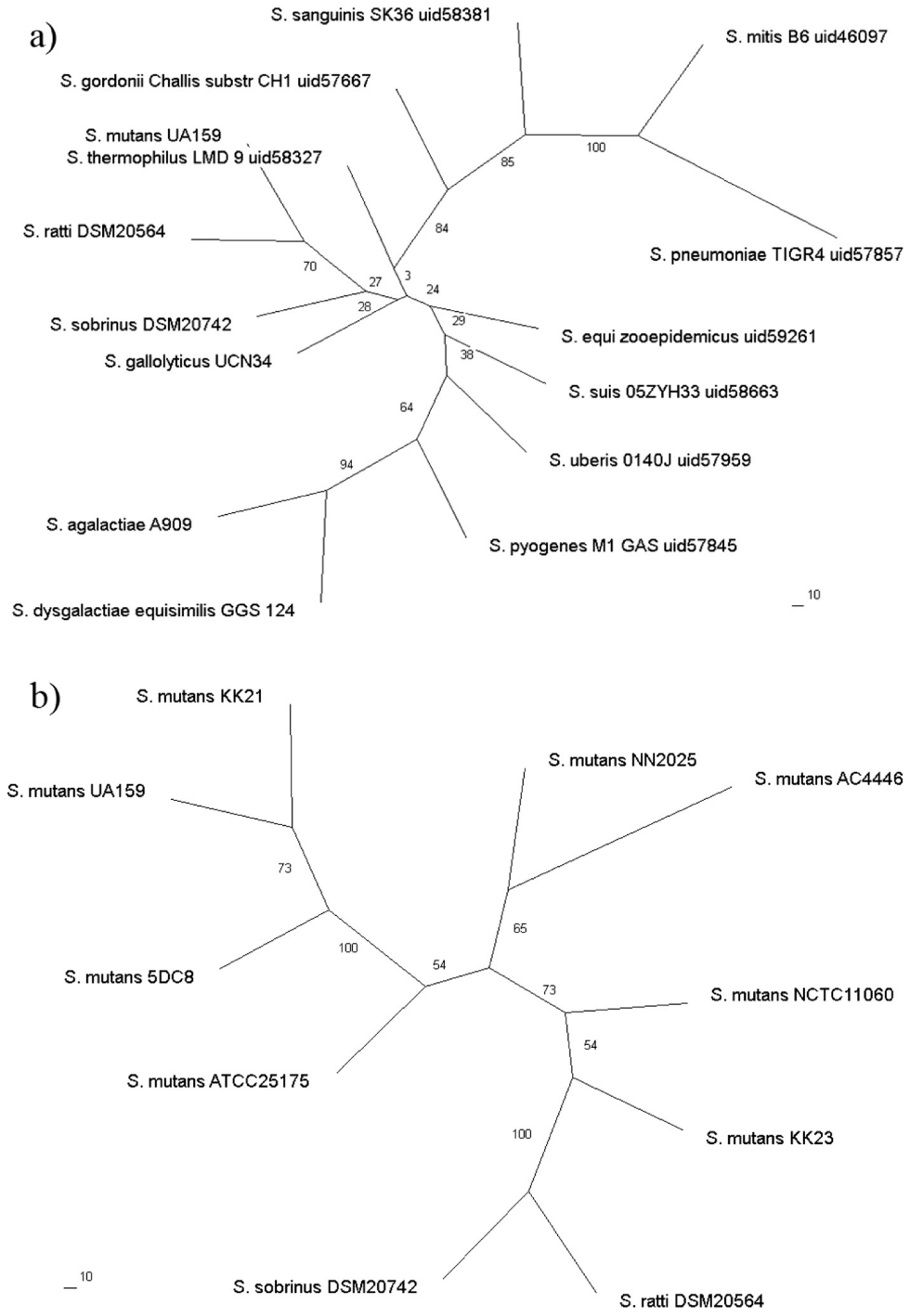


Fig. 3.1 Phylogenetic analysis of 10 mutans streptococci strains compared in this study and their phylogenetic relationship to other *Streptococcus* species.

a) 16S rRNA phylogenetic tree of *Streptococcus* species with genomes known as this study was performed (Since the 16S rRNA sequences were almost identical between the different *S. mutans* strains, only UA159 is shown as representative). b) Phylogenetic tree of the mutans streptococci compared in this study constructed with the core-genome SNPs obtained by PGAP pipeline [36]. All phylogenetic trees were constructed using ClustalX [63] and Phylip [64] by applying the maximum likelihood (ML) method with bootstrap value set to 100. Values beside branches depict ML bootstrap support values. The scale bars in the unit of "substitution/site" are shown below the trees.

Table 3.1 Genome assembly and annotation of eight *S. mutans* strains sequenced in this study in comparison with previously sequenced *S. mutans* strains UA159 and NN2025

	UA159	**NN2025**	**5DC8**	**KK21**	**KK23**	**AC4446**	**ATCC25175**	**NCTC11060**	**DSM20564**	**DSM20742**
	NC_004350.2	NC_013928.1	AOBX	AOBY	AOBZ	AOCA	AOCB	AOCC	AOCD	AOCE
Length	2,030,921	2,013,587	2,010,935	2,034,586	1,976,057	2,003,537	1,999,532	2,021,202	2,037,184	2,096,203
Contigs	Complete	Complete	9	2	38	42	10	36	182	283
N50 size	Complete	Complete	354,736	1,622,660	134,323	167,413	233,425	94,580	23,860	12,417
N90 size	Complete	Complete	325,634	411,935	38,851	26,425	107,076	43,987	6,098	3,659
G+C content	36.82%	36.85%	36.90%	36.81%	36.68%	36.90%	36.87%	36.98%	40.29%	43.46%
Total Genes	2040	1975	2,004	2,031	1,933	1,999	1,982	1,982	1,995	2,057
CDSs	1,960	1,895	1,924	1,949	1,907	1,919	1,903	1,915	1,965	2,032

3.3 Genome rearrangement of *S. mutans* genomes

Genome rearrangements have important effects on bacterial phenotypes and the evolution of bacterial genomes [65]. A comparison of the genomes of *S. mutans* NN2025 and UA159 has discovered a large genomic inversion across the replication axis and similar genomic variations were also confirmed among 95 clinical *S. mutans* isolates using long-PCR analysis [29]. In this study, genome rearrangements among the eight *S. mutans* strains were determined by genome alignment using the MAUVE software [34]. The results are presented in Figure 3.2, which shows the locally collinear blocks (LCBs) representing the landmarks, i.e. the homologous/conserved regions shared by all the input sequences in the chromosomes. A LCB is defined as a collinear (consistent) set of exactly matched subsequences (multiple maximal unique matches, namely 'multi-MUMs') which are shared by all the chromosomes considered, appear once in each chromosome and are bordered on both sides by mismatched nucleotides. The weight (the sum of the lengths of the included multi-MUMs) of a LCB serves as a measure of confidence that it is a true homologous region rather than a random match.

As shown in Figure 3.2, 16 LCBs (marked as A to P) were identified by multi-alignment of the eight *S. mutans* genome sequences. Compared to UA159 and NN2025, the chromosome segment represented by LCB E is reversely inserted between the LCB G and H in the strain AC4446, and between the LCB L and M in the strain KK23. This segment is related to the genomic island "*SMU.100-SMU.116*" of *S. mutans* UA159 which mainly contains sorbitol phosphotransferase system (PTS), transposase and hypothetical proteins [66]. LCB N is found to be reversed and relocated to the position between LCB A and B in the strain AC4446. A cluster of tRNA genes is found to be located downstream of LCB N. In KK23, LCB I and J are moved to position between LCB F and G. A tRNA-Gln and a tRNA-Tyr is found to be adjacent to the left of LCB I. LCB K in NCTC 11060, AC4446, KK23 and NN2025 are very similar to each other but much smaller than those of other strains (with sequence length reduced about two-thirds). The missing sequence corresponds to the genomic island TnSmu2 of *S. mutans* which harbors a nonribosomal peptide synthetase-polyketide synthase gene cluster responsible for the biosynthesis of pigments [67]. Using the known information about genomic islands in *S. mutans* UA159, additional genomic islands were found to be present/absent in the mutans streptococci strains of this study [68, 66][ii]. Furthermore, there are much more diversities as shown by the white areas inside the LCBs which show regions with low similarities. However, it should be noticed that there might be more genome rearrangements among the strains, because draft genome sequences were used in current

[ii]For details please refer to an online file https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S1.xlsx

analysis and all contigs in each genome were sorted according to the reference genome sequence of the strain UA159.

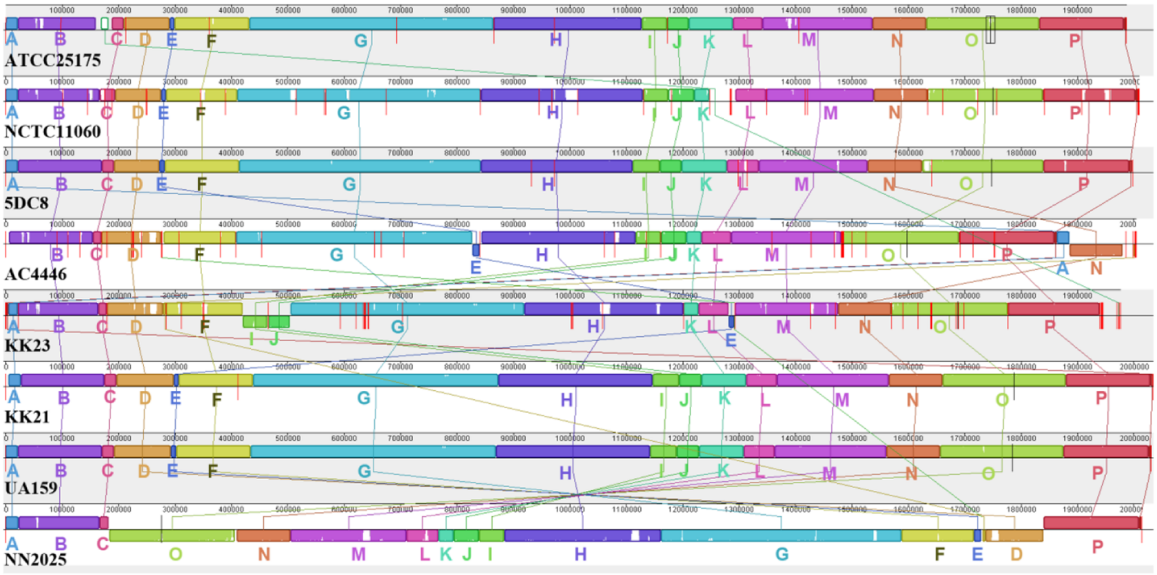


Fig. 3.2 Comparison of local collinear blocks (LCBs) of chromosomal sequences of the eight *S. mutans* strains.

In total 16 local LCBs, marked as A to P, were generated and compared by applying the MAUVE software [34, 69] with default settings and using strain UA159 as reference. The red vertical bars indicate contig ends. The white areas inside each LCB show regions with low similarities.

3.4 Core and pan-genome analysis of *S. mutans* species

The genetic variability within species in the domain *Bacteria* is much larger than that found in other domains of life. The gene content between pairs of isolates can diverge by as much as 30% in species like *Streptococcus pneumoniae* [70]. This unexpected finding led to the introduction of the pan-genome concept, which describes the sum of genes that can be found in a given bacterial species [37, 35]. The genome of any isolate is thus composed of a "core-genome" shared with all strains of this particular species, and a "dispensable genome" that accounts for the phenotypic differences between strains. The pan-genome is usually much larger than the genome of any single isolate, constituting a reservoir that could enhance the ability of many bacteria to survive in stressful environments. The pan-genome concept has important consequences for the way we understand bacterial evolution, adaptation, and population structure, as well as for more applied issues such as vaccine design or the identification of virulence genes [71]. In this study, core-genome and pan-genome

analyses of 67 *S. mutans* strains was performed, including the eight mutans streptococci strains sequenced in this study and 59 *S. mutans* strains whose genomes were available in NCBI till April 2013.

3.4.1 Core-genome

The core-genome size of the 67 *S. mutans* strains was calculated to be 1,373. For detailed information of the core genes please refer to an online file[iii]. To estimate the theoretical core-genome size achievable with an infinite number of *S. mutans* genomes, core-genome size medians corresponding to different genome numbers as shown in Figure 3.3a by the red rectangles were first calculated by random sampling 1,000 genome combinations of n genomes out of the 67 *S. mutans* genomes. Then, the exponential regression core-genome model $F_c(n) = k_c exp[^{-n}/\tau_c] + \Omega$ proposed previously by Tettelin *et al.* [37, 35] was applied to fit the median data points of the core-genome sizes, where *n* represents the number of genomes, and Ω stands for the theoretical core-genome size. To take into consideration the different deviations of the core-genome size medians, as clearly indicated by the blue error bars in Figure 3.3a, the fitting process was modified by introducing the genome number as weight to the corresponding data point. The fitting parameters thus obtained are as follows: $r^2 = 0.97403, k_c = 325.74718 \pm 10.00912, \Omega = 1,369.41225 \pm 1.986, \tau_c = 15.90248 \pm 0.66807$. Using this fitting result to describe the core-genome of *S. mutans*, the theoretical core-genome size (Ω) was estimated to be around 1,370 genes, which is comparable to the core-genome size (1,373) calculated using 67 genomes. Compared with other *streptococcus* species, the core-genome of *S. mutans* is at the same level to the core-genome of *S. pyogenes* (1,400 genes determined using 11 strains), less than that of *S. pneumoniae* (1,647 genes determined using 47 strains) and *S. agalactiae* (1,800 genes determined using eight strains) [37, 72, 73]. However, it should be cautious with such comparison. In the study of Cornejo *et al.* [62], the core genome size of *S. mutans* was determined as 1,490 by using 57 *S. mutans* genomes which is obviously different to the core genome size of *S. mutans* estimated in this study, although the 57 *S. mutans* genomes used by Cornejo *et al.* were also included in our study. The discrepancy can be caused by different reasons, such as difference in the correction step for core gene determination and, very likely, different methods and parameter settings used for determining orthologs. Apparently, a more stringent process was used in this study to determine orthologs which led to smaller core genome size of *S. mutans* estimated.

[iii] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S2.xlsx

3.4.2 Pan-genome

Three models, namely $y = a + bx^c$, $y = a - b\ln(x + c)$ and $y = a \times e^{-x/b} + c$ (where a, b and c are parameters) were applied for modeling the pan-genome of *S. mutans*, as shown in Figure 3.3b by green, blue and red curves respectively (all fitting results can be found in an online file [iv]).

Both the fitting results of using $y = a + bx^c$ and $y = a - b\ln(x + c)$ indicated an infinite pan-genome, while the fitting result of using $y = a \times e^{-x/b} + c$ resulted in a negative value of the parameter a, suggesting a finite pan-genome However, the last fitting shows obvious deviations to many of the data points. Especially, the deviations even become larger with increased genome numbers, indicating that this model is not suitable. The best fitting result obtained with the model $y = a + bx^c$ shows fittings to all the data points with very high confidence. According to this model, the pan-genome of *S. mutans* is still "open" although 67 genomes were included for the estimation, and the expected average new gene number with the addition of a new genome is estimated to be 15. The infinite pan-genome was first proposed by Tettelin *et al.* for *S. agalactiae* based on the use of 9 *S. agalactiae* genomes. The three regression models used in this study are all based on the assumption that contingency genes are independently sampled from the pan-genome with equal probability, except in the case of "specific/unique genes", which are modeled as unique events that appear only once in the entire global population. Hogg *et al.* [74] proposed a finite supragenome model for pan-genome based on a different supposition that contingency genes are sampled from the pan-genome with unequal probability. By applying this finite supragenome model to 44 *S. pneumoniae* genomes, the predicted number of new genes drops sharply to zero when the number of genomes exceeds 50. However, in the case of *S. mutans* such sharp decrease of new gene number could not be observed even after 67 genomes were included. In the study of Cornejo *et al.* [62], they proposed a finite pan-genome for *S. mutans*, after they used a special "pseudogene cluster" identification process to exclude about 30% of the rare genes that are considered to be pseudogenes. However, they didn't provide detailed parameters they obtained from fitting. Our modeling using the 67 *S. mutans* genomes by applying the model described above without any restrictions pointed to an infinite pan-genome of *S. mutans*. However, this predicted "infinite" pan-genome should be understood as follows: 1) a "pan-genome" should be considered as "dynamic" rather than "static", which means the pan-genome content is changing during the evolution, it does not matter if its size is infinite or finite; 2) The change of a pan-genome content can be caused either by the acquirement of new genes or by the loss of existing genes; 3) The actual pan-genome size can be more stable than the content of the pan-genome but can also change during evolution coupled

[iv] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S3.docx

with the change of the environment. Thus, without considering "gene loss events", it's quite understandable to have a "growing" or "infinite" pan-genome as gene acquirement occurs no matter how slow it might be. Interestingly, Cornejo *et al.* found a high rate of HGT in *S. mutans*, where many genes were acquired from related streptococci and bacterial strains predominantly residing not only in the oral cavity, but also in the respiratory tract, the digestive tract, genitalia, in insect pathogens and in the environment in general [62]. Such high rate of HGT might also lead to a continuously growing (infinite) pan-genome.

3.5 Gene content-based comparative analysis of mutans streptococci strains

The annotated protein sequences of the ten mutans streptococci genomes studied were cross-compared based on alleles/ortholog groups established by the program OrthoMCL [38]. In total, 2,211 putative alleles/ortholog groups were established[v]. A pair-wise comparison of the protein coding sequences between each two strains is shown in Table 3.2. It is clear to see that remarkable differences in protein coding sequences (CDSs) exist between the strains, even inside the same species of *S. mutans* . In the following sections, systems that are highly related to stress resistance and pathogenicity are presented and discussed. As all the following results are based on putative alleles/ortholog groups established by OrthoMCL, if not otherwise specified, the word "putative allele(s)/ortholog(s)" is omitted in the following text.

[v] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S4.xlsx

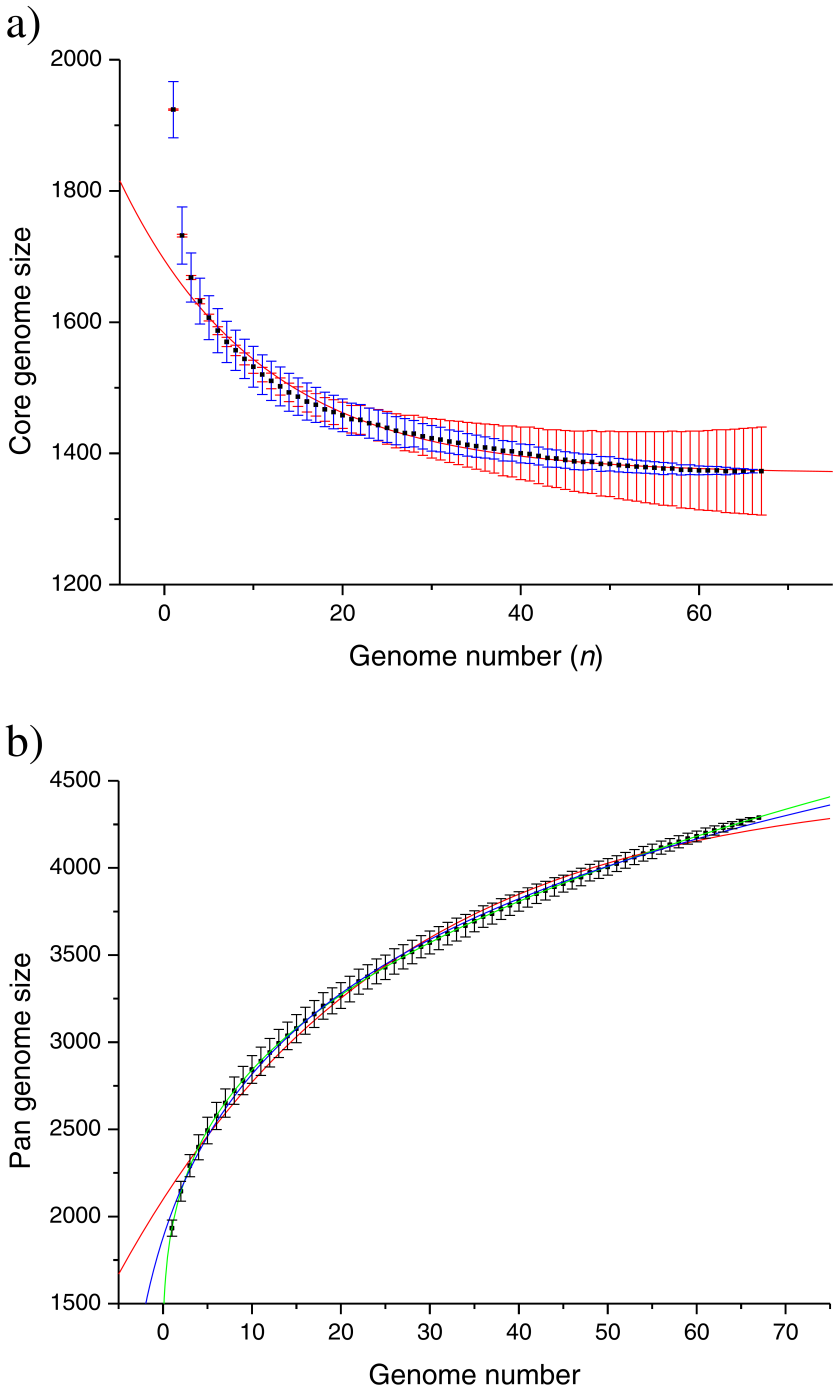


Fig. 3.3 Core and pan-genome model of 67 *S. mutans* genomes.

a) Core-genome model of *S. mutans*. The core-genome size (number of common genes) is plotted as a function of the number (n) of genomes according to a previously proposed model $F_c(n) = k_c exp[^{-n}/\tau_c] + \Omega$, where k_c, τ_c, and Ω are model parameters. Red rectangles are the medians of the core-genome sizes calculated by random sampling 1,000 different genome combinations of n genomes out of 67 genomes. Blue bars are the standard deviations of the medians. The red bars are weights used for model fitting and the red curve is the fitting result. b) Pan genome modeling of *S. mutans* genomes using three models, $y = a + bx^c$, $y = a - b\ln(x+c)$ and $y = a \times e^{-x/b} + c$ (where a, b and c are parameters), represented by green, blue and red curves respectively. Black rectangles are the medians of the pan-genome sizes calculated by random sampling 1,000 different genomes combination of n genomes out of 67 genomes, and black bars are the standard deviations of the medians.

Table 3.2 Unique protein coding sequences (CDSs) between the different strains revealed by ortholog analysis

	Unique CDSs in comparision to										
	UA159	**NN2025**	**5DC8**	**KK21**	**KK23**	**AC4446**	**ATCC25175**	**NCTC11060**	**DSM20564**	**DSM20742**	**All others**
UA159		216	125	63	230	221	166	212	427	566	42
NN2025	150		150	150	133	102	182	167	358	510	24
5DC8	85	176		52	164	161	132	153	379	522	31
KK21	47	200	76		190	184	127	175	402	544	3
KK23	183	152	157	159		146	173	175	387	525	56
AC4446	145	92	125	124	117		159	146	364	502	31
ATCC25175	117	199	123	94	171	186		146	373	525	33
NCTC11060	126	147	107	105	136	136	109		334	488	34
DSM20564	432	429	424	423	439	445	427	425		564	289
DSM20742	616	626	612	610	622	628	624	624	609		492

3.5.1 Distribution of two-component signal transduction systems

Bacterial two-component signal transduction systems (TCS) play important roles for many bacteria by enabling them to detect and respond to diverse changes/stresses in the environment. The conspicuous absence of TCS proteins in mammalian genomes makes them interesting potential targets for the development of novel antibacterial drugs. A bacterial two-component system comprises, in general, a transmembrane sensor histidine kinase (HK) and a corresponding cytoplasmic response regulator (RR) encoded by genes located adjacently within the same operon, although stand-alone genes coding for HKs or RRs (without a corresponding cognate HK or RR in the same operon) have also been reported. In some cases, a HK and a RR were found to be merged in the same polypeptide, giving rise to a so called 'hybrid' HK protein. A HK protein is autophosphorylated at its conserved histidine (His) residue upon the recognition of a specific environmental stimulus. The phosphoryl group is then transferred to the aspartate (Asp) residue of the corresponding response regulator [75]. While HKs, in general, serve to detect signals, the most common function of the RRs is to bind directly to DNA and thereby modulating the expression of a certain set of genes which are necessary for mounting a physiological response to the perceived signals [75]. HK and RR proteins are composed of domains which are structurally and functionally conserved and can be used for their classification. 14 TCS clusters have previously been identified in *S. mutans* UA159 [28, 76] and many of them have been reported to be involved in its virulence, adaptation and survival [76–80].

Table 3.3 Identification and classification of putative two component systems in the eightmutans streptococci strains sequenced in this study

Strain	*S.mutans* UA159	*S.mutans* NN2025	*S.mutans* 5DC8	*S.mutans* KK21	*S.mutans* KK23	*S.mutans* AC4446	*S.mutans* ATCC25175	*S.mutans* NCTC11060	*S.ratti* DSM20564	*S.sobrinus* DSM20742
Identification										
Total TCS proteins	29	29	29	29	29	27	25	27	28	21
Total paired HKs/RRs	14	14	14	14	14	13	12	13	13	9
Orphan HKs	0	0	0	0	0	0	0	0	1	1
Orphan RRs	1	1	1	1	1	1	1	1	1	2
Classification										
HK type										
HPK1	8	8	8	8	8	8	7	7	9	5
HPK7	3	3	3	3	3	3	3	4	3	2
HPK8	1	1	1	1	1	1	1	1	1	1
HPK10	1	1	1	1	1	1	1	1	0	1
HPK11	0	0	0	0	0	0	0	0	1	0
unclassified	1	1	1	1	1	0	0	0	0	1
RR type										
NarL	3	3	3	3	3	3	3	4	3	2
LytTR	2	2	2	2	2	2	2	2	1	3
AmiR	0	0	0	0	0	0	0	0	1	0
OmpR	9	9	9	9	9	9	8	8	9	6
unclassified	1	1	1	1	1	0	0	0	0	0

By combining the HMM profiling results and the information on putative operon organization, repertoires of potential TCS proteins (HKs and RRs) in the eight mutans streptococci strains sequenced in this study were obtained, as shown in Table 3.3 in comparison to *S. mutans* NN2025 and UA159. The total numbers of TCS proteins identified are comparable among the 10 mutans streptococci strains.

By analyzing the putative operon organizations of the identified TCS proteins, 98.5% of the total putative HKs and 92.2% of the total putative RR were found to be constituting HK-RR pairs. Ortholog analysis of the paired or non-paired TCS proteins among the 10 mutans streptococci strains revealed a total of 18 different TCS clusters, 2 orphan HKs and 2 orphan RRs (Table 3.4). The numbering of the TCS clusters was based on an existing numbering system used by Levesque [81], and extended to the new TCS clusters identified in this study. Co-evolution of TCS proteins could be clearly observed. This means HKs and RRs, which belong to a particular TCS cluster, are usually co-present or co-absent in a specific strain.

In addition, putative alleles/orthologs of the corresponding HKs and RRs were found to be highly conserved (similarity $\geq 95\%$) among the *S. mutans* strains. The conservation was clearly lower across the species. Furthermore, it is obvious that, in most cases, putative HK alleles/orthologs within one TCS cluster exhibited a higher degree of diversity than the corresponding putative RR alleles/orthologs across the species. This could be attributed to the high variability of sensing/input domains harbored by the individual HKs, as will be discussed later in the classification of HKs and RRs.

Table 3.4 Ortholog analysis and classifications of the putative TCS proteins

	TCS Protein	RR family	C1	C2	UA159	NN2025	5DC8	KK21	KK23	AC4446	ATCC25175	NCTC11060	DSM20564	DSM20742
TCS-1	**HK (VicK)**	-	HPK1	C	SMU.1516	GI|290580114	smc|01510	smd|01560	sme|01514	smf|01497	smg|01519	smh|01557	sral1000270(90)	sobl6900029(74)
	RR (VicR)	OmpR	-		SMU.1517	GI|290580113	smc|01511	smd|01561	sme|01515	smf|01498	smg|01520	smh|01558	sral1000269(96)	sobl6900028(87)
TCS-2	**HK (CiaH)**	-	HPK1	E	SMU.1128	GI|290580439	smc|01121	smd|01170	sme|01262	smf|01137	smg|01147	smh|01166	sral3300059(83)	sobl10000005(53)
	RR (CiaR)	OmpR	-		SMU.1129	GI|290580438	smc|01122	smd|01171	sme|01263	smf|01138	smg|01148	smh|01167	sral3300058(94)	sobl10000006(86)
TCS-3	**HK (CovS)**	-	HPK1	E	SMU.1145c	GI|290580424	smc|01138	smd|01187	sme|01279	smf|01154	smg|01163	smh|01183	sral3300038(48)	Absent
	RR (CovR)	OmpR	-		SMU.1146c	GI|290580423	smc|01139	smd|01188	sme|01280	smf|01155	smg|01164	smh|01184	sral3300037(76)	Absent
TCS-4	**HK (KinF)**	-	HPK1	E	SMU.928	GI|290580625	smc|00919	smd|00964	sme|01045	smf|00943	smg|00948	smh|00938	sral1200002(75)	sobl15200007(50)
	RR (LlrF)	OmpR	-		SMU.927	GI|290580626	smc|00918	smd|00963	sme|01044	smf|00942	smg|00947	smh|00937	sral1200003(89)	sobl15200008(70)
TCS-5	**HK (ScnK)**	-	HPK1	E	SMU.1814	GI|290579846	smc|01808	smd|01870	sme|01810	smf|01679	Absent	Absent	Absent	Absent
	RR (ScnR)	OmpR	-		SMU.1815	GI|290579845	smc|01809	smd|01871	sme|01811	smf|01680	Absent	Absent	Absent	Absent
TCS-6	**HK (SpaK)**	-	HPK1	E	SMU.660	GI|290580857	smc|00645	smd|00684	sme|00777	smf|00640	smg|00672	smh|00668	sral1000132(71)	sobl3600002(47)
	RR (SpaR)	OmpR	-		SMU.659	GI|290580858	smc|00643	smd|00682	sme|00775	smf|00639	smg|00670	smh|00666	sral1000130(80)	sobl3600003(67)
TCS-7	**HK (PhoR)**	-	HPK1	E	SMU.1037c	GI|290580522	smc|01032	smd|01081	sme|01153	smf|01049	smg|01055	smh|01052	Absent	Absent
	RR (YcbL)	OmpR	-		SMU.1038c	GI|290580521	smc|01033	smd|01082	sme|01154	smf|01050	smg|01056	smh|01053	Absent	Absent
TCS-8	**HK (KinG)**	-	HPK1	M	SMU.1009	GI|290580539	smc|01005	smd|01053	sme|01135	smf|01030	smg|01036	smh|01024	sral900051(80)	sobl6900007(55)
	RR (LlrG)	OmpR	-		SMU.1008	GI|290580540	smc|01004	smd|01052	sme|01134	smf|01029	smg|01035	smh|01023	sral900049(89)	sobl6900006(64)
TCS-9	**HK (LevS)**	-	HPK7	M	SMU.1965c	GI|290579718	smc|01957	smd|02016	sme|01953	smf|01824	smg|01946	smh|01994	sral3500035(88)	Absent
	RR (LevR)	NarL	-		SMU.1964c	GI|290579719	smc|01956	smd|02015	sme|01952	smf|01823	smg|01945	smh|01993	sral3500036(91)	Absent
TCS-10	**HK (LytS)**	-	HPK8	M	SMU.577	GI|290580924	smc|00560	smd|00598	sme|00685	smf|00568	smg|00588	smh|00583	sral1000030(95)	sobl800016(80)
	RR (LytT)	LytTR	-		SMU.576	GI|290580925	smc|00559	smd|00597	sme|00684	smf|00567	smg|00587	smh|00582	sral1000029(92)	sobl800017(70)
TCS-11	**HK (LiaS)**	-	HPK7	M	SMU.486	GI|290581011	smc|00465	smd|00505	sme|00595	smf|00473	smg|00493	smh|00475	sral2900015(89)	sobl4200072(64)
	RR (LiaR)	NarL	-		SMU.487	GI|290581010	smc|00466	smd|00506	sme|00596	smf|00474	smg|00494	smh|00476	sral2900013(92)	sobl4200071(82)
TCS-12	**HK (HK11)**	-	HPK7	M	SMU.1548c	GI|290580085	smc|01542	smd|01591	sme|01547	smf|01527	smg|01550	smh|01590	sral1000231(53)	sobl11300005(50)
	RR (RR11)	NarL	-		SMU.1547c	GI|290580086	smc|01541	smd|01590	sme|01545	smf|01526	smg|01548	smh|01588	sral1000232(82)	sobl11300004(72)
TCS-13	**HK(ComD)**	-	HPK10	M	SMU.1916	GI|290579761	smc|01910	smd|01968	sme|01906	smf|01777	smg|01900	smh|01947	Absent	sobl8500005(37)
	RR (ComE)	LytTR	-		SMU.1917	GI|290579760	smc|01911	smd|01969	sme|01907	smf|01778	smg|01901	smh|01948	Absent	sobl8500004(43)
TCS-14	**HK**	-	?	C	SMU.45	GI|290579565	smc|00046	smd|00053	sme|00049	Absent	Absent	Absent	Absent	Absent
	RR	?	-		SMU.46	GI|290579566	smc|00047	smd|00054	sme|00050	Absent	Absent	Absent	Absent	Absent
TCS-15	**HK(ComP)**	-	HPK7	M	Absent	Absent	Absent	Absent	Absent	Absent	Absent	smh|00177	Absent	Absent
	RR(CmpR)	NarL	-		Absent	Absent	Absent	Absent	Absent	Absent	Absent	smh|00178	Absent	Absent
TCS-16	**HK**		HPK1	E	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral800020	Absent
	RR	OmpR			Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral800019	Absent
TCS-17	**HK**		HPK1	E	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral3500015	Absent
	RR	OmpR			Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral3500014	Absent
TCS-18	**HK**		HPK11	C	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral1400052	Absent
	RR	AmiR			Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral1400053	Absent
Orphan HK1	**HK**		HPK1	N	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sral3800008	Absent
Orphan HK2	**HK**		?	M	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sobl13200014
Orphan RR1	**RR (GcrR)**	OmpR	-		SMU.1924	GI|290579753	smc|01920	smd|01978	sme|01916	smf|01787	smg|01910	smh|01957	sral200007(87)	sobl100018(55)
Orphan RR2		LytTR			Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	sobl14100005

TCS proteins common to all the 10 mutans streptococci strains

Proteins of the TCS clusters 1, 2, 4, 6, 8, 10, 11, 12 and the orphan RR1 are common to all the 10 mutans streptococci strains compared here, indicating probably the functional importance of these TCS clusters for the adaptation and survival of these mutans streptococci. For instance, Orphan RR1 is highly conserved across the 10 mutans streptococci strains. In *S. mutans* UA159, this orphan RR is encoded by *gcrR* (*SMU.1924c*) and has been found to play a vital role in sucrose-dependent adherence and cariogenesis [82]. Therefore, it is conceivable that conservation of this gene across the mutans streptococci strains is essential for their primary pathogenicity.

TCS proteins uniquely present/absent in one or several strains

The TCS-3 (CovSR) cluster was predicted to be absent in *S. sobrinus* DSM20742. CovSR is involved in the acid tolerance response of *S. mutans* UA159 [81], and has also been reported to be involved in counteracting oxidative stress and reducing susceptibility to phagocytic killing [83]. TCS-9 (LevRS), which affects the acid tolerance response as well [81], was also absent in *S. sobrinus* DSM20742. The absence of the *covS* and *levS* genes was experimentally supported by the PCR results. In *S. mutans* UA159, the *levRS* gene cluster is flanked by *levQ* and *levT*, which code for two carbohydrate-binding proteins. These four genes together constitute a four-component signal transduction system levQRST controlling the transcription of the fructan hydrolase gene (*fruA*) and a four-gene cluster *levDEFG*, which encode a fructose/mannose sugar:phosphotransferase system located immediately downstream of *levQRST* [84]. *S. sobrinus* was also found to lack the *levQ*, *levT* and *levDEFG* genes. Taking together, these findings indicate dramatic differences in the regulation of fructan catabolism and the acid tolerance response of *S. sobrinus* DSM20742 in comparison to the *S. mutans* strains.

It should be pointed out that the TCS nomenclature is unfortunately inconsistent among the published articles on TCSs of *S. mutans* strains. Many publications [85–89, 82] on CovR actually address the orphan response regulator RR1 which is known as GcrR (*SMU.1924*) in *S. mutans* UA159 [82]. On the other hand, CovSR of *S. mutans* has also been confusingly named as ScnKR [81, 83]. In this study, the nomenclature of TCS genes/proteins was based primarily on the Oralgen Pathogen Sequence Database (http://www.oralgen.lanl.gov). In cases where several names were given for a gene in Oralgen, gene names used by Biswas *et al.* [76] were preferably used.

TCS-5 (ScnKR) could be neither found in the two *S. mutans* strains ATCC25175 and NCTC11060 nor in *S. ratti* DSM20564 and *S. sobrinus* DSM20742. The asymmetric

distribution of TCS-5 was also observed in a previous study, in which TCS-5 was found to be present only in two of the 10 *S. mutans* strains compared [76]. In *S. mutans* UA159, an insertion mutant of *scnK* gene displayed no significant difference to the wild type with respect to growth under various stress conditions [76]. In *Streptococcus pyogenes*, *scnKR* was found to be essential for the production of a bacteriocin (SAFF22) [90]. By a closer look at the genes in the neighborhood of *scnKR* in the 10 strains studied, it was found that *S. mutans* KK23 and *S. mutans* NN2025 carried two and three genes, respectively, which encode putative bacteriocin precursor peptides sharing more than 60% similarity with SAFF22. In addition, a putative bacteriocin biosynthesis protein coding gene was also found downstream of the *scnKR* operon in the two *S. mutans* strains. It was therefore inferred that TCS-5 (ScnKR) might be involved in the regulation of mutacin production at least in *S. mutans* KK23 and *S. mutans* NN2025.

TCS-7 (PhoR/YcbL) was only shared by the eight *S. mutans* strains. In *S. mutans* UA159, an insertion mutant of the gene encoding PhoR displayed no significant difference to the wild type with respect to growth under various stress conditions [76]. The clear function of TCS-7 is still unknown.

As mentioned before in the identification and classification of TCS proteins, TCS-13 (ComDE) was absent in *S. ratti* DSM20564. This finding was also supported by the PCR experiment. In *S. mutans*, ComDE is the most extensively studied two-component signaling system involved in quorum sensing and competence development. Mediated by the so-called competence stimulating peptide (CSP), it is involved in multiple stress responses and has been implicated in competence development, bacteriocin production, virulence, biofilm formation, and cariogenicity [79, 91–95]. Further analysis showed that the ortholog of the *comC* gene, whose product is the precursor of the signal peptide CSP sensed by ComDE in *S. mutans*, was absent in *S. ratti*. Putative orthologs of *comD* and *comE* were found in *S. sobrinus* DSM20742. However, with a similarity of merely 37% and 43%, respectively, to the *comD* and *comE* of *S. mutans* UA159, it's highly possible that the actual function of TCS-13 in *S. sobrinus* DSM20742 might be quite different from that known for *S. mutans*. It is also worthy to mention that no *comC*-like gene was found in *S. sobrinus* DSM20742.

Putative alleles of the HK and RR proteins of TCS-14 are present in five of the *S. mutans* strains, namely 5DC8, KK23, KK21, UA159 and NN2025. This cluster was first identified in *S. mutans* UA159 by Biswas *et al.* and the corresponding HK and RR are encoded by *SMU.45* and *SMU.46*, respectively [76]. They also found that TCS-14 was present only in two of the 13 *S. mutans* strains compared in their study. HKs of this TCS cluster contain only one recognizable HATPase_c domain. In addition, no known output domain was identified in the cognate RRs. Thus, neither HKs nor RRs of TCS-14 could be classified into any known

HK and RR families. In addition, by the multi-sequence alignment of the putative TCS-14 HK alleles, the open reading frame predictions carried out in this study revealed that while *SMU.45* and its upstream gene *SMU.44* clearly constitute two separate genes in *S. mutans* UA159, they are merged to constitute the parts of a single gene coding for the HKs in the *S. mutans* strains 5 DC8, KK23 and NN2025. The corresponding HK of KK21 was also split into two proteins by the lacking of a single glutamine (Q) residue.

TCS-15 was found exclusively in the genome of the serotype *f* blood isolate *S. mutans* NCTC11060. Genes of a TCS located on the genomic island TnSmu2 of *S. mutans* UA140, as recently reported by the research group Qi [67], could be possible alleles of TCS-15 genes. This was based on the fact that the predicted HKs and RRs of both TCS 15 in this study and the TCS found by Qi's group have the HK (YP_002747386.1) and RR (YP_002124238.1) of *Streptococcus equi*, respectively, as the best matched homologs. The HK (smh|00177) of TCS-15 is also the only histidine kinase found in this study that contains a PDZ domain. PDZ domain was first reported to be present in animals. In 1997, Ponting *et al.* claimed that PDZ domains exist also in diverse signaling proteins of bacteria, yeasts and plants. Experimental evidence was first provided by Liao *et al.* through the determination of the first crystal structure of a bacterial PDZ domain [96]. The most remarkable feature of PDZ domains is their ability to specifically recognize and bind to short C-terminal peptide motifs. This allows them especially to bind membrane proteins such as ion channels, which have very small free C-termini. To exclude the possibility of contamination with *e. g.* human DNA during the process of genome sequencing, the existence of the gene coding for this unusual PDZ domain-containing HK protein in the DNA of the NCTC11060 strain was experimentally verified by using two different forward primers in the PCR experiment.

The PDZ domain of smh|00177 is flanked by one transmembrane helix (TM) at its N-terminal side and 9 TMs at its C-terminal side, which is characteristic of a ComP-like HK. ComP-like HKs are a group of sensors of another peptide-dependent quorum sensing system related to cell density-responsive regulation other than ComDE in gram-positive bacteria. In *B. subtilis*, ComP is the sensor histidine kinase of the four-component *comQXPA* quorum sensing system, where ComA stands for the corresponding response regulator, ComX is the pheromone precursor and ComQ the protein required for the proteolytic cleavage and modification of the pheromone precursor molecule [97] . The most similar homolog of smh|00177 found from a BLASTp search in the NCBI database (http://blast.ncbi.nlm.nih.gov/) with a sequence identity of around 65% is the histidine kinase (YP_003353659.1) of a TCS from *Lactococcus lactis subsp. lactis* KF147. Furthermore, the cognate response regulator of TCS-15 showed a sequence identity of around 78% with the corresponding response regulator (YP_003353660.1) from the same *L. lactis subsp. lactis* strain. In this study, the RR

(smh|00177) of TCS-15 was termed as CmpR. It should be mentioned that neither *L. lactis subsp. lactis* KF147 nor *S. mutans* NCTC11060 possesses homologs of the *B. subtilis* ComX and ComQ. Thus, the signal peptide sensed by ComP/CmpR in *S. mutans* NCTC11060 remains unknown.

TCS-16, 17 and 18 are uniquely present in *S. ratti* DSM20564. According to BLASTP searches against the NCBI database, the top matches to the TCS-16 HK protein (sra|800020) or the RR protein (sra|800019) are all from *Streptococcus* species such as *Streptococcus infantarius* and *Streptococcus agalactiae*. In addition, the HK and RR homologs in the different *Streptococcus* species are all encoded by two adjacent genes and annotated in some *S. agalactiae* strains as sensor histidine kinase DltS and DNA-binding response regulator DltR. The DltSR has been reported to be involved in the regulation of D-alanyl-lipoteichoic acid biosynthesis in *S. agalactiae* [98]. Lipoteichoic acid (LTA) is a major cell wall constituent of Gram-positive bacteria which is phosphoglycerol substituted with a D-Ala ester or a glycosyl residue and anchored in the membrane by its glycolipid moiety. D-alanylation of lipoteichoic acid has been proven to contribute to the virulence of *Streptococcus suis* [99], as well as to the biofilm formation and resistance to antimicrobial peptides in *enterococci* [100]. Thus, the TCS-16 cluster might also be an important virulence factor in *S. ratti*.

The TCS-17 is composed of a HAMP-containing HK and an OmpR-type RR. The best homologs of the HK protein (sra|3500015) and the RR protein (sra|3500014) are from *S. agalactiae* strains. But the functions of these homologous proteins remain unknown.

TCS-18 is the only cluster that comprises a HPK 11 family HK and an AmiR family RR, which possesses a novel RNA-binding type output domain. The top ten best hits of the BLASTp search in the NCBI database showed that the closest homologs for the HK and RR of TCS-18 are all from *Listeria* species. Since the HK of TCS-18 possesses a PAS domain that is commonly involved in sensing intracellular signals such as redox potential, similar to the sensing mechanism described for the HK (VicK) of TCS-1, the function of TCS-18 might also be related to the sensing and response to signal(s) originated in the cytoplasm. The unique presence of TCS-18 in *S. ratti* DSM20564 was also confirmed by the PCR experiment.

3.5.2 High diversities of the competence development regulation module

In the previous section, the two-component signal transduction systems (TCS) in the 10 mutans streptococci strains were systematically discussed. ComDE, one of the TCS is directly related to competence development. Competence development is a complex process

involving sophisticated regulatory networks that trigger the capacity of bacterial cells to take up exogenous DNA from the environment. This phenomenon is frequently encountered in bacteria of the oral cavity, e.g., *S. mutans* [101]. In *S. mutans* , ComX, an alternative sigma factor, drives the transcription of the so called 'late-competence genes' required for genetic transformation. ComX activity is modulated by the inputs from two types of signal pathways, namely the competence-stimulating peptide (CSP) dependent competence regulation system and CSP-independent competence regulation system. ComX and the 'late-competence genes' regulated by ComX as labeled in boldface in Table 3.5, are highly conserved even between the species, indicating that all mutans streptococci studied here might have the ability of transforming to genetic competence state. On the other hand, the upstream signal pathways regulating the activity of ComX show high variety as discussed in details below.

Table 3.5 Distribution of competence development-related systems in the 10 mutans streptococci strains

Name	UA159	NN2025	5DC8	KK21	KK23	AC4446	ATCC 25175	NCTC 11060	DSM 20564	DSM 20742
ComA	SMU.286	GI|290581206	D816_01150	D817_01300	D818_01134	D819_01163	D820_01336	D821_01208	D822_01584	D823_05343
	SMU.1881c	GI|290579788	D816_08453	D817_08643	-	D819_07724		D821_08449	D822_08325	D823_01400
ComB	SMU.287	GI|290581205	D816_01155	D817_01305	D818_01139	D819_01168	D820_01341	D821_01213	D822_01589	D823_05923
ComC	SMU.1915	GI|290579762	D816_08588	D817_08778	D818_08368	D819_07839	D820_08520	D821_08549	-	-
SepM	SMU.518	GI|290580977	D816_02205	D817_02448	D818_02735	D819_02254	D820_02420	D821_02274	D822_04126	D823_08607
ComD	SMU.1916	GI|290579761	D816_08593	D817_08783	D818_08373	D819_07844	D820_08525	D821_08554		D823_05333
ComE	SMU.1917	GI|290579760	D816_08598	D817_08788	D818_08378	D819_07849	D820_08530	D821_08559		D823_05328[a]
HtrA	SMU.2164	GI|290581420	D816_09733	D817_00015[b]	D818_00020	D819_09056	D820_09650	D821_09748	D822_05851	D823_03191
HdrM	SMU.1855	GI|290579809	D816_08353	D817_08543	D818_08143	D819_07614	D820_08345	D821_08319	D822_08240	D823_08222
HdrR	SMU.1854	GI|290579810	D816_08348	D817_08538	D818_08138	D819_07609	D820_08340	D821_08314	-	-
BrsM	SMU.2081	GI|290581347	D816_09358	D817_09538	D818_09198	D819_08671	D820_09275	D821_09348	-	-
BrsR	SMU.2080	GI|290581346	D816_09353	D817_09533	D818_09193	D819_08666	D820_09270	D821_09343	D822_05085	-
OppD	SMU.258	GI|290581226	D816_01025	D817_01175	D818_01039	D819_01063	D820_01211	D821_01051	D822_05611	D823_04322
ComS	NC_004350.2	NC_013928.1	D816_00277	D817_00297	D818_00297	D819_00203	D820_00247	D821_00253	D822_01077	-
ComR	SMU.61	GI|290579576	D816_00275	D817_00295	D818_00294	D819_00200	D820_00245	D821_00250	D822_01080	
ComX	SMU.1997	GI|290579687	D816_08973	D817_09163	D818_08748	D819_08219	D820_08900	D821_08929	D822_07328	D823_08887
ComEA	**SMU.625**	**GI|290580890**	**D816_02675**	**D817_02923**	**D818_03217**	**D819_02694**	**D820_02880**	**D821_02784**	**D822_02674**	**D823_08107**
ComEC	**SMU.626**	**GI|290580889**	**D816_02680**	**D817_02928**	**D818_03222**	**D819_02699**	**D820_02885**	**D821_02789**	**D822_02679**	**D823_08117**
CoiA	**SMU.644**	**GI|290580870**	**D816_02775**	**D817_03018**	**D818_03322**	**D819_02786**	**D820_02970**	**D821_02879**	**D822_02739**	**D823_01025**
EndA	**SMU.1523**	**GI|290580108**	**D816_06842**	**D817_07008**	**D818_06659**	**D819_06647**	**D820_06860**	**D821_06857**	**D822_03254**	**D823_09687**
ComG	**SMU.1981c**	**GI|290579702**	**D816_08898**	**D817_09088**	**D818_08673**	**D819_08144**	**D820_08825**	**D821_08854**	**D822_07418**	**D823_01170**
ComYD	**SMU.1983**	**GI|290579700**	**D816_08908**	**D817_09098**	**D818_08683**	**D819_08154**	**D820_08835**	**D821_08864**	**D822_07408**	**D823_01160**
ComYC	**SMU.1984**	**GI|290579699**	**D816_08913**	**D817_09103**	**D818_08688**	**D819_08159**	**D820_08840**	**D821_08869**	**D822_07403**	**D823_01155**
	SMU.2075c	**GI|290581342**	**D816_09328**	**D817_09508**	**D818_09168**	**D819_08641**	**D820_09245**	**D821_09318**	**D822_05110**	**D823_03558**
CinA	**SMU.2086**	**GI|290581351**	**D816_09383**	**D817_09563**	**D818_09218**	**D819_08691**	**D820_09295**	**D821_09368**	**D822_05060**	**D823_03593**
ComYB	**SMU.1985**	**GI|290579698**	**D816_08918**	**D817_09108**	**D818_08693**	**D819_08164**	**D820_08845**	**D821_08874**	**D822_07398**	**D823_01150**
ComYA	**SMU.1987**	**GI|290579697**	**D816_08923**	**D817_09113**	**D818_08698**	**D819_08169**	**D820_08850**	**D821_08879**	**D822_07393**	**D823_01145**
ComFC	**SMU.499**	**GI|290580999**	**D816_02100**	**D817_02348**	**D818_02650**	**D819_02154**	**D820_02290**	**D821_02159**	**D822_06218**	**D823_02981**
ComFA	**SMU.498**	**GI|290581000**	**D816_02095**	**D817_02343**	**D818_02645**	**D819_02149**	**D820_02285**	**D821_02154**	**D822_06223**	**D823_02986**
CinA	**SMU.2086**	**290581351**	**D816_09383**	**D817_09563**	**D818_09218**	**D819_08691**	**D820_09295**	**D821_09368**	**D822_05060**	**D823_03593**

[a]Additional ComE like protein identified D823_7992 [b]Additional HtrA like protein identified D817_9913. The rows related to highly conserved 'late-competence genes' were shown in boldface. The missing of ComDE in *S. ratti* DSM 20564 has been discussed in previous section about two compenent systems

CSP-dependent competence regulation system

It has been reported that the ComABCDE system in *S. mutans* combines the action of the two ortholog systems which are present as ComABCDE and BlpABCRH in *S. pneumoniae* and involved in competence regulation and bacteriocins regulation, respectively. It should be noticed that, ComAB have been primarily considered to be the transporter of ComC, the precursor of CSP. Later, ComAB have been renamed as NlmTE as they were found to function together as transporter of nonlantibiotic bacteriocins, while another gene pair CslAB was supposed to be the transporter of ComC [102]. However, a recent study confirms that ComAB is indeed a transporter both for nonlantibiotic bacteriocin and the peptide pheromone CSP [103].

In *S. mutans* , the *comC*-encoded prepeptide of CSP has a leader sequence containing a conserved double glycine (GG), at which the leader sequence is cleaved during transporting by ComAB to generate the mature signal peptide (CSP-21) containing 21 amino acid residues [102, 104, 105]. Recent studies show that an extracellular protease, SepM (*SMU.518*), is involved in the further processing of CSP-21 by removing the "LGK" residues in the C-terminal to generate a 18-residue peptide (CSP-18), which can work at a concentration much lower than that of CSP-21 [103, 106]. SepM is identified in all the 10 strains compared in this study, although putative comC alleles are present only in the eight *S. mutans* strains, not in the *S. sobrinus* DSM 20742 and *S. ratti* DSM 20564. Multi-alignment of the ComC sequences shows clear variations among different *S. mutans* strains (Figure 3.4a). Genetic variation of ComC in *S. mutans* has been reported previously [107]. Interestingly, the C-terminal amino acid sequence "LGK" of ComC is absent in the ComC prepetides of *S. mutans* KK23 and AC4446, which have also been observed previously in other *S. mutans* strains by Allan *et al.* [107]. ATCC 25175 possesses a unique ComC sequence ended with "LGKIR" at its C-terminal. In addition to the variations at the carboxyl end, substitutions of single amino acid residues at different positions were also found.

All the variants of *comC* revealed in this study have been verified by PCR experiments. Although Allan *et al.* pointed out that different *comC* alleles in some clinical strains of *S. mutans* exist but their products are functionally equivalent and there is no evidence of phenotype specificity [107], considering the complexity of phenotype evaluation, whether and how the variations found in this study may affect the natural genetic competence of these *S. mutans* strains requires further investigation.

The CSP-initiated activation of the response regulator ComE, through its cognate receptor kinase ComD, leads to the induction of competence through the alternative sigma factor ComX, and at the same time ComE directly induces a set of bacteriocin-related genes [79, 91, 95, 102, 104, 108, 109]. The comparison of the two-component signal transduction

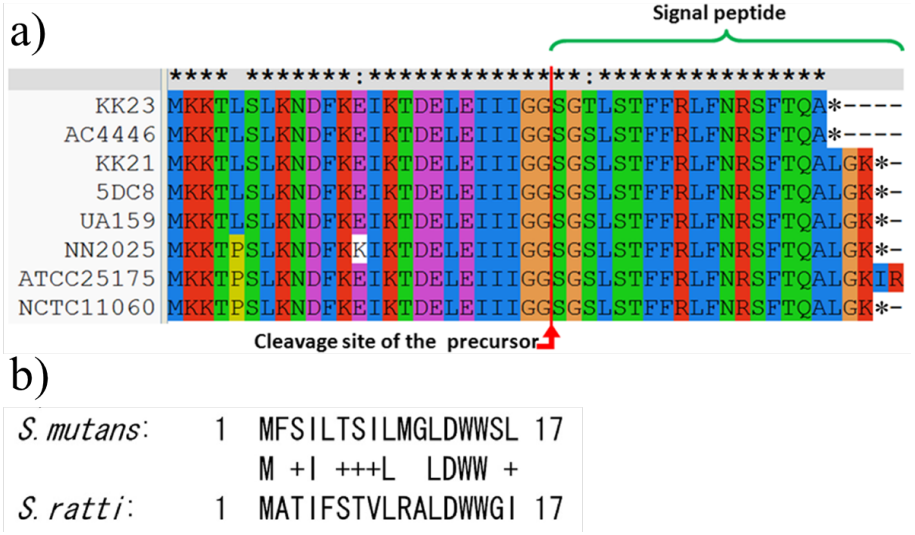


Fig. 3.4 Alignment of ComC and ComS amino acid sequences.

a) Alignment of ComC amino acid sequences identified in *S. mutans* species using CLUSTALX. Conserved residues are marked with "*" above the figure. The diversity in the ComC sequences have been verified by PCR experiments (data not shown). b) BlastP alignment of the ComS sequence of *S. mutans* (identical among the eight *S. mutans* strains) with that of *S. ratti* DSM 20564 (No ComS was identified in *S. sobrinus*). "+" stands for similar amino acid residues.

systems of these mutans streptococci strains has revealed the complete missing of ComDE in *S. ratti* DSM 20564 and the low similarities of putative ComDE in *S. sobrinus* DSM 20742 to the ComDE of *S. mutans* strains. Accordingly, no comC-like genes could be identified in *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742. Thus, it can be inferred that *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742 are quit different to the *S. mutans* strains regarding cellular functions including genetic competence associated with the ComABCDE system.

In *S. mutans* , no binding motif for ComE is present in the promoter region of ComX, suggesting that ComE is not a direct regulator of ComX, whereas a new peptide regulator system (ComSR) downstream of ComE that directly activates ComX has been identified by Mashburn-Warren *et al.* ComR activates the expression of ComS, which is secreted, processed, and internalized through the peptide transporter OppD. The processed peptide, designated XIP (for sigma X-inducing peptide), modulates the activity of ComR, which in turn activates the expression of ComX. Deletion of *comR* or *comS* gene completely abolished the competence in *S. mutans* [110]. In this study, the ComSR regulating system is identified in most of the strains, except for *S. sobrinus* DSM 20742 which lacks the ComSR-coding genes. This well explains the fact that despite the presence of comX and the 'late-competence genes' it was not able to obtain the genetic competence state of *S. sobrinus* DSM 20742 (see discussion later in the "Variability and specificity in metabolic pathways and network" part).

It is also worth to mention that the putative ComS ortholog found in *S. ratti* DSM 20564 is quite different to those of *S. mutans* strains, as shown in Figure 3.4b.

CSP-independent competence regulation system

It has been reported that a basal level of competence remains (referred as CSP-independent competence) after the deletion of comE from *S. mutans* , suggesting that the CSP-dependent regulation system is only one of the signaling pathways involved in ComX activation [91]. Indeed, under conditions of biofilm growth the HdrMR system, a novel two-gene regulatory system, has been shown to contribute to competence development through the activation of ComX by a yet unknown signal [111]. Moreover, microarray analysis revealed that both regulators, ComE and HdrR, activate a large set of overlapping genes [111, 112]. Recently, Xie *et al.* identified in *S. mutans* another regulatory system, designated BsrRM, that primarily regulates bacteriocin-related genes but also affects the HdrMR system and thus indirectly contributes to competence development [113]. In this study, HdrR, the response regulator of the HdrMR system, is found neither present in *S. ratti* DSM 20564 nor in *S. sobrinus* DSM 20742. Furthermore, the response regulator BrsR of the BsrRM system is also absent in *S. ratti* DSM 20564, and *S. sobrinus* DSM 20742 lacks the complete BsrRM system. However, a competence damage-inducible protein CinA, which is regulated via ComX and has been proven to be related to DNA damage, genetic transformation and cell survival [114], is present in all strains.

Taking together, both the CSP-dependent and CSP-independent competence regulation systems in *S. ratti* DSM 20564 and especially in *S. sobrinus* DSM 20742 are very different to those of the *S. mutans* strains.

3.5.3 Distribution of bacteriocin- and antibiotic resistance-related proteins

Bacteriocin-related proteins

Bacteriocins are proteinaceous toxins produced by bacteria to kill or inhibit the growth of similar or closely related bacterial strain(s). Bacteriocins produced by mutans streptococci were named “mutacins”. As dental plaque, the dominating niche of mutans streptococci, is a multispecies biofilm community that harbors many microorganism species, mutans group strains have developed a variety of mutacins to inhibit the growth of competitors, such as mitis group streptococci [103, 115, 116]. In this study, information about known mutacins as well as mutacin-immunity proteins was collected from the NCBI (http://www.ncbi.nlm.nih.gov)

and Oralgen (http://www.oralgen.lanl.gov/) databases, as well as by searching for related publications. The collected protein sequences, as detailed in Appendix B, were used to blast against the proteomes of the 10 strains to see whether or not these known mutacins and mutacin-immunity proteins do exist in the mutans streptococci strains of this study. Distributions of identified mutacins and mutacin-immunity proteins are summarized in Table 3.6. Using this approach it is, however, not possible to identify any new types of mutacins.

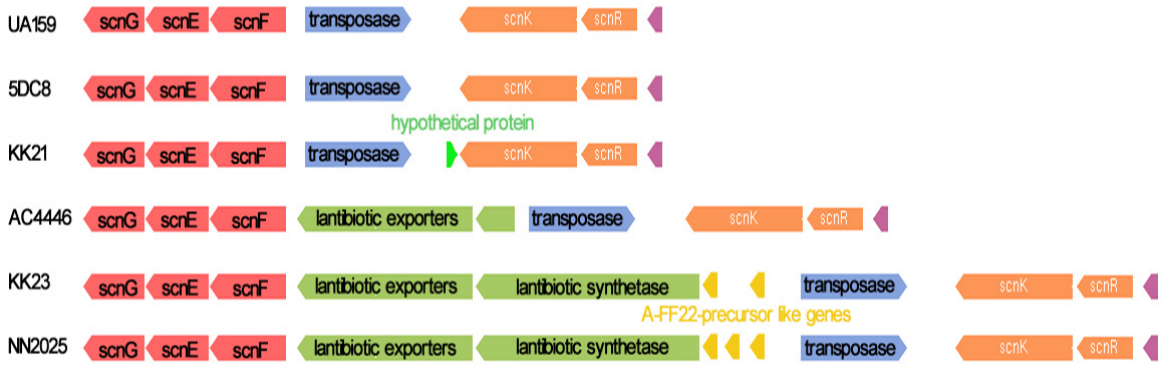


Fig. 3.5 Cluster structure of the mutacin-K8 production system across six *S. mutans* strains.

The ORFs colored in yellow are the possible mutacin-K8 precursor genes. scnGEF: bacteriocin related ABC element; possible immunity system; *scnK*: histidine kinase of two component system; *scnR*: response regulator of two component system (*Note: mutacin-K8 production system was failed to be identified in S. mutans NCTC 11060, S. mutans ATCC 25175, S. ratti DSM 20564 and S. sobrinus DSM 20742).*

Diversity of *Streptococcus* bacteriocins has been reported previously [117, 118]. The mutacin assortments of the 10 strains in this study also demonstrate certain variations. An interesting result is that in contrast to *S. mutans* strains and *S. ratti* DSM 20564, *S. sobrinus* DSM 20742 does not possess any genes coding for mutacin-like proteins. Mutacin-SMB has been identified in *S. mutans* and *S. ratti* previously [119, 120]. In our study, mutacin-SMB cluster was only identified in *S. ratti* DSM 20564 comprising 7 genes, including the mutacin-coding genes smbA and smbB, as well as 5 mutacin-related genes (*smbG* -> D822_07603, *smbT* -> D822_07593, *smbM* -> D822_07578, *smbF* -> D822_07588, and *smbM2* -> D822_07598). Lantibiotic mutacins, namely mutacin-I [121], mutacin-II [122] and mutacin-III [123], are completely absent in the 10 mutans streptococci strains. However, three gene copies possibly encoding the precursor of the lantibiotic mutancin mutacin-K8 were identified in the *S. mutans* strains KK23 and NN2025. Mutacin-K8 is an ortholog of the bacteriocin Streptococcin A-FF22 identified in group-A streptococci [124], and its production system has previously also been identified in the *S. mutans* strain K8 [125]. By carefully examining the genes surrounding mutacin-K8 precursor genes the gene cluster coding for a complete mutacin-K8 production system was also revealed in the strains KK23 and NN2025

(Figure 3.5). A partial ortholog of the mutacin-K8 production system was found in *S. mutans* UA159, 5DC8 and KK21, with only genes responsible for the immunity (*scnFEG*) left behind. Orthologous genes coding for a part of the mutacin-K8 production system were also found in *S. mutans* AC4446, consisting of only *scnFEG*, *scnT* (coding a lantibiotic exporter) and a part of *scnM* (coding the lantibiotic synthetase). Since a gene encoding ISSmu2-type transposase was found to be located upstream of mutacin-K8 precursor genes, the variety of mutacin-K8 production system in *S. mutans* strains studied here is highly possible to be caused by transposase actions.

Mutacin-IV, nonlantibiotic bacteriocins coded by *nlmA/B* (SMU.150/151, Note: hereinafter whenever needed/possible the locus_tag of the reference strain *S. mutans* UA159 is given for convenience) was discovered first in *S. mutans* UA140 to be active against the mitis group streptococci [126]. In this study, nlmA/B were found to be present in six of the *S. mutans* strains, including UA159, 5DC8, KK21, KK23, ATCC 25175 and NCTC 11060, but not in *S. mutans* NN2025 and AC4446, nor in *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742. On the other hand, the immunity protein for mutacin-IV (*SMU.152*), was identified in all strains, consistent with the fact that no inhibition phenomenon has been observed yet among different mutans streptococci strains. A mutacin-IV like protein found before in the strain UA159 (*SMU.283*) was also identified in all strains except for *S. sobrinus* DSM 20742.

Mutacin-V, another nonlantibiotic peptide coded by *cipB* (*SMU.1914*) present in all strains studied here, except for*S. sobrinus* DSM 20742 and *S. mutans* strains ATCC 15175 and NCTC 11060. There are two homologs of mutacin-V immunity protein in *S. mutans* UA159, namely CipI (*SMU.925*) and *SMU.1913* [28, 127]. These two immunity proteins share a sequence identity of 82%. However, it has been reported that though very likely co-transcribed with *cipB*, *SMU.1913* cannot prevent CipB-caused cell lysis in *S. mutans* UA159, and the key immunity factor of mutacin-V has been supposed to be CipI (*SMU.925*) rather than *SMU.1913* [127]. All the 10 strains including *S. sobrinus* DSM 20742 possess at least one orthologous gene encoding one of the two mutacin-V immunity proteins. Based on the similarity scores *S. mutans* NN2025 does not have an ortholog of CipI, but it possesses an ortholog (GI|290579764) of *SMU.1913*, which is possibly co-transcribed with GI|290579764, the *cipB* ortholog in *S. mutans* NN2025. Furthermore, the only putative immunity protein D822_3349 in *S. ratti* DSM 20564 shows very close similarities to *SMU.925* (61%) and SMU.1913 (56%) and is possibly co-transcribed with D822_03354, the CipB ortholog in *S. ratti* DSM 20564. From these results, it is tempting to suppose that *SMU.1913*, which is co-transcribed with *cipB* (*SMU.1914*), might be the ancestor gene coding for the mutacin-V immunity factor. The additional copy, like *SMU.925* in *S. mutans* UA159, might be generated

by duplication action and evolved as the dominant immunity factor in some of the mutans streptococci strains.

Furthermore, a possible nonlantibiotic bacteriocin peptide (*SMU.423*) is found to be present in all strains except for *S. ratti* DSM 20564. Putative ComAB, which has been proved to be the transporter complex of mutacin IV in *S. mutans* [102], were identified in all strains, supporting the suggestion that ComAB might function as a common transporter for multi-type nonlantibiotic bacteriocins rather than merely for mutacin IV. Moreover, an additional paralog of ComA is present in most of the strains except for *S. mutans* KK23 and *S. mutans* ATCC 25175.

To summarize, a differed distribution of mutacin/bacteriocin encoding genes accompanied with a high conservation of genes coding for mutacin-immunity proteins were revealed for the 10 mutans streptococci strains/species. The conservation of mutacin immunity proteins apparently plays an important role in the survival of mutans streptococci strains under a bacteriocin-rich environment.

Table 3.6 Distribution of mutacins and mutacin immunity proteins in the 10 mutans streptococci strains

	UA159	NN2025	5DC8	KK21	KK23	AC4446	ATCC 25175	NCTC 11060	DSM 20564	DSM 20742
Mutacin-SMB (lantibiotic mutacin)	-	-	-	-	-	-	-	-	D822_07608 D822_07613	-
Mutacin-I (lantibiotic mutacin)	-	-	-	-	-	-	-	-	-	-
Mutacin-II (lantibiotic mutacin)	-	-	-	-	-	-	-	-	-	-
Mutacin-III (lantibiotic mutacin)	-	-	-	-	-	-	-	-	-	-
Mutacin-K8 (lantibiotic mutacin)	-	GI|290579849 GI|290579848 GI|290579850	-	-	D818_07928 D818_07933 D818_07938	-	-	-	-	-
Mutacin-IV (NlmA)	SMU.150	-	D816_00655	D817_00675	D818_00659	-	D820_00642	D821_00661	-	-
Mutacin-IV (NlmB)	SMU.151	-	D816_00660	D817_00680	D818_00664	-	D820_00647	D821_00666	-	-
Mutacin-IV like	SMU.283	GI|290581209	D816_01135	D817_01285	D818_01099	D819_01148	D820_01321	D821_01193	D822_03404	-
Mutacin-IV*	SMU.152	GI|290580110	D816_06832	D817_06998	D818_06649	D819_06637	D820_06850	D821_06847	D822_03264	D823_04636
Mutacin-V (CipB)	SMU.1914c	GI|290579763	D816_08583	D817_08773	D818_08363	D819_07834	-	-	D822_03354	-
CipI (Mutacin-V) *	SMU.925		D816_04020	D817_04283	D818_04522	D819_04119	D820_04232	D821_04089	D822_03349	
Homolog of CipI	SMU.1913c	GI|290579764	D816_08578	D817_08768	D818_08358	D819_07829				D823_03992
	SMU.423	GI|290581063	D816_01775	D817_01930	D818_01847	D819_01823	D820_01975	D821_01862		D823_05348
NlmT/ComA	SMU.286	GI|290581206	D816_01150	D817_01300	D818_01134	D819_01163	D820_01336	D821_01208	D822_01584	D823_05343
	SMU.1881c	GI|290579788	D816_08453	D817_08643	-	D819_07724	-	D821_08449	D822_08325	D823_01400
NlmE/ComB	SMU.287	GI|290581205	D816_01155	D817_01305	D818_01139	D819_01168	D820_01341	D821_01213	D822_01589	D823_05923

Note: as a multi-function exporter, the entries of NlmTE(ComAB) have been shown in Table 3.5 and here again. * *stands for a immunity protein*

Antibiotic resistance-related proteins

Bacteria and other microorganisms that cause infections are remarkably resilient and can develop ways to survive drugs meant to kill or weaken them. Antibiotic resistance can be a result of horizontal gene transfer [128], and also of unlinked point mutations in the pathogen genome at a rate of about 1 in 108 per chromosomal replication [129]. The antibiotic action against the pathogen can be seen as an environmental selective pressure and bacteria which have developed mutations allowing them to survive will live on to reproduce. They will then pass this trait to their offsprings, which will result in the evolution of fully resistant colonies. In this study, putative antibiotic resistance-related genes were identified in the genomes of the ten mutans streptococci strains/species and are listed in Table 3.7.

The *S. mutans* species is known to be intrinsically resistant to bacitracins produced by Bacillus subtilis. This was also confirmed in this study by testing all the 10 strains with a bacitracin-E-test (data not shown). All strains including *S. ratti* DSM 20564 and *S. sobrinus* DSM 20742 showed a minimum inhibitory concentration between 128 and 256 μg/l. In fact, bacitracins have been used to isolate mutans-streptococci from highly heterogeneous oral microflora. It has been reported that *bceABRS* (also named as *mbrABCD*) system, encoding a two component signal transduction system and an ABC-transporter, is required for bacitracin resistance in *S. mutans* [130, 131]. As expected, ortholog of *bceABRS* system was found to be present in all strains of this study. Furthermore, an ortholog of a putative bacitracin resistant protein UppP (*SMU.244*, undecaprenyl-diphosphatase) is also present in all strains. It has been proved that overexpression of UppP in Escherichia coli and Bacillus subtilis results in bacitracin resistance [132, 133]. However, the function of UppP in bacitracin resistance in mutans streptococci has not yet been investigated. Based on its conservation in all strains studied here, it is reasonable to suppose that UppP might play an important role in bacitracin resistance of mutans streptococci species as well.

Two penicillin-binding proteins (*SMU.75* and *SMU.455*) were identified in all the ten strains, indicating that they are potentially all susceptible to penicillin. Phenotypically all strains were tested to be susceptible to penicillin (data not shown). On the other hand, all the strains possess orthologs of *SMU.368c*, *SMU.400*, *SMU.1444c* and *SMU.1515*, which are homologs to beta-lactamases (EC 3.5.2.6), as well as orthologs of two so called beta-lactam resistance factors (SMU.716, SMU.717). Thus, all the strains are potentially capable of resistance against beta-lactam antibiotics. Orthologs of macrolide-efflux transporter proteins, as coded by GI|290581182 and GI|290581181 in *S. mutans* NN2025, were found to be also present in *S. mutans* 5DC8 and *S. mutans* KK21. A vancomycin b-type resistance-associated protein (D822_01634) is uniquely present in *S. ratti* DSM 20564, although a phenotypic testing showed as expected that *S. ratti* DSM 20564 was still susceptible

to vancomycin. Furthermore, several putative multidrug resistance-associated proteins (*SMU.745*, *SMU.1611c* and *SMU.905* except for *SMU.1286c*) were found to be present in all strains.

Table 3.7 Distribution of antibiotic resistance-related proteins

Name	Resistance	UA159	NN2025	5DC8	KK21	KK23	AC4446	ATCC 25175	NCTC 11060	DSM 20564	DSM 20742
UppP	Bacitracin	SMU.244	GI|290581239	D816_00960	D817_01110	D818_00974	D819_00998	D820_01146	D821_00986	D822_05517	D823_09307
BceA	Bacitracin	SMU.1006	GI|290580542	D816_04484	D817_04663	D818_04902	D819_04489	D820_04607	D821_04449	D822_02154	D823_04551
BceB	Bacitracin	SMU.1007	GI|290580541	D816_04489	D817_04668	D818_04907	D819_04494	D820_04612	D821_04454	D822_02159	D823_04556
DacF	Penicillin	SMU.75	GI|290579588	D816_00335	D817_00355	D818_00354	D819_00260	D820_00330	D821_00310	D822_07803	D823_05036
	Penicillin										
Pbp2X	Penicillin	SMU.455	GI|290581039	D816_01905	D817_02153	D818_01967	D819_01954	D820_02095	D821_01964	D822_00802	D823_06528
	Penicillin										
	beta-Lactam	SMU.368c	GI|290581108	D816_01525	D817_01680	D818_01583	D819_01608	D820_01711	D821_01583	D822_04346	D823_00655
	beta-Lactam										
	beta-Lactam	SMU.400	GI|290581086	D816_01660	D817_01815	D818_01732	D819_01708	D820_01860	D821_01747	D822_05706	D823_03675
	beta-Lactam										
YqgA		SMU.1444c	GI|290580186	D816_06482	D817_06653	D818_06314	D819_06285	D820_06483	D821_06502	D822_08877	D823_08387
	beta-Lactam										
	beta-Lactam	SMU.1515	GI|290580115	D816_06807	D817_06973	D818_06624	D819_06612	D820_06825	D821_06822	D822_03289	D823_04661
MurN	beta-Lactam	SMU.716	GI|290580807	D816_03100	D817_03358	D818_03627	D819_03104	D820_03315	D821_03199	D822_00265	D823_09452
MurM	beta-Lactam	SMU.717	GI|290580806	D816_03105	D817_03363	D818_03632	D819_03109	D820_03320	D821_03204	D822_00260	D823_09457
	Macrolide		GI|290581182			D818_01269	D819_01313				
	multidrug		GI|290581181			D818_01274	D819_01318				
VanW	Vancomycin									D822_01634	
	Multidrug	SMU.745	GI|290580783	D816_03220	D817_03478	D818_03732	D819_03234[a]	D820_03442	D821_03314	D822_00530	D823_08347
PmrA	Multidrug	SMU.1611c	GI|290580030	D816_07242	D817_07403	D818_07009	D819_07037	D820_07260	D821_07242	D822_07918	D823_02317
YitG	Multidrug	SMU.1286c	GI|290580299	D816_05764	D817_05958	D818_02360	D819_05785	D820_05818	D821_05850	D822_01559	
	Multidrug	SMU.905	GI|290580642	D816_03940	D817_04208	D818_04447	D819_03949	D820_04157	D821_04009	D822_09885	D823_08492

[a] Additional homolog was identified D819_09750

3.5.4 Oxidative stress defense systems in mutans streptococci

For protection against reactive oxygen species (such as O_2^-, H_2O_2, HO·) or adaptation to oxidative stresses aerobes and facultative anaerobes have evolved efficient defense systems, comprising an array of antioxidant enzymes such as catalase, superoxide dismutase (SOD), Dps-like peroxide resistance protein, alkylhydroperoxide reductase (AhpCF), glutathione reductase, and thiol reductase, which have been identified in many bacterial species.

Although the first genome sequence of *S. mutans* UA159 has already been published in 2002, the oxidative stress defense systems in the group of mutans streptococci have not yet been systematically discussed. By searching for known antioxidant systems in the genomes of the sequenced mutans streptococci strains of this study, an overview of putative oxidative defense systems in these mutans streptococci strains/species was obtained, which are composed of superoxide dismutase (SOD), AhpF/AhpC system, Dpr, thioredoxin system and glutaredoxin system, as shown in Table 3.8.

SOD, which catalyzes the dismutation of superoxide into oxygen and hydrogen peroxide, is an important antioxidant defense in nearly all cells exposed to oxygen [134]. SOD was found in all strains of this study. Catalase, which catalyzes the decomposition of hydrogen peroxide, was not found in any of the mutans streptococci strains of this study. It is known that although most streptococci can grow in the presence of air, they do not possess a catalase, implying that hydrogen peroxide defense mechanism, by which lactic acid bacteria established their growth in air, are very different to those of aerobes. It has been reported that both the bi-component peroxidase system AhpF/AhpC and Dps-like peroxide resistance protein confer tolerance to oxidative stress in *S. mutans* [135].

The AhpF/AhpC system catalyzes the NADH-dependent reduction of organic hydroperoxides and/or H_2O_2 to their respective alcohol and/or H_2O. Both AhpF and AhpC are present in all *S. mutans* strains of this study and in *S. ratti* DSM 20564, but are absent in *S. sobrinus* DSM 20742. The natural missing of AhpF and AhpC in *S. sobrinus* indicates that AhpF/AhpC system is not an essential peroxide tolerance system for some mutans streptococci species. While studying a ahpF and ahpC double deletion mutant of *S. mutans* , Higuchi *et al.* [136] found that the mutant still showed the same level of peroxide tolerance as did the wild-type strain that led them to the finding of the *dpr* gene, which encodes a ferritin-like iron-binding protein involved in oxygen tolerance by limiting the nonenzymatic hydroxyl radical synthesis via iron-catalyzed 'Fenton reaction' in *S. mutans* . Their further studies on the biological function of dpr found that dpr gene from *S. mutans* chromosome was capable of complementing an alkyl hydroperoxide reductase-deficient mutant of *E. coli*, as well as complementing the defect in peroxidase activity caused by the deletion of *ahpF/ahpC* in *S. mutans* , indicating that dpr plays an indispensable role in oxygen tolerance of *S. mutans*

[135, 137]. Dpr homologs were found in all strains of this study as expected by the supposed essential function of *dpr* gene in oxygen tolerance.

Thioredoxins are a class of small redox mediator proteins known to be present in all organisms. They are involved in many important biological processes, including redox signaling. Thioredoxins are kept in the reduced state by the flavor enzyme thioredoxin reductase in a NADPH-dependent reaction [138]. They act as electron donors to many proteins including thiol peroxidases [139]. Thioredoxin, thioredoxin reductase and thiol peroxidase, the components of thioredoxin system, were identified in all the strains of this study. Two putative thioredoxin reductases (*SMU.463* and *SMU.869*) were found in all strains/species. It has been reported that in some species thioredoxin reductases have been evolved to be activated by both NADPH and NADH [140]. Since *SMU.463* and *SMU.869* shares less than 20% similarities, it is reasonable to speculate that *SMU.463* and *SMU.869* might have been evolved to have different preferences to NADPH and NADH. If it holds true, this could be advantageous for these mutans streptococci, as the extra amount of NADH produced from glycolysis/gluconeogenesis pathway under anaerobic conditions could be directly used for oxidative stress resistance. Thioredoxin (*SMU.1869*) and two thioredoxin family proteins (*SMU.1971c* and *SMU.1169c*) were found to be present in nearly all strains, except for *S. sobrinus* DSM 20742, which lacks any ortholog of SMU1169c. An ortholog of a thiol peroxidase-coding gene (*tpx*) was identified in all strains.

Glutaredoxins share many functions of thioredoxins but are reduced by glutathione (L-γ-glutamyl-L-cysteinylglycine, GSH) rather than by a specific reductase. This means that glutaredoxins are oxidized by their corresponding substrates, and reduced non-enzymatically by GSH [141]. Oxidized glutathione (GSSG) is then regenerated by glutathione reductase. Together, these components comprise the glutathione system [142]. GSH is a well-characterized antioxidant in eukaryotes and Gram-negative bacteria, where it is synthesized by the sequential action of two enzymes, γ-glutamylcysteine synthetase (γ-GCS) and glutathione synthetase (GS). Among Gram-positive bacteria only a few species contain GSH. It has been reported that streptococci lack the moderate-to-high levels of intracellular glutathione normally found in Gram-negative bacteria [143]. Using *Streptococcus agalactiae* as a model, it has been discovered that in GSH-containing Gram-positive bacteria GSH synthesis is catalyzed by one bifunctional protein, γ-glutamylcysteine synthetase-glutathione synthetase (γ-GCS-GS), encoded by one gene, *gshAB*. Homologs of γ-GCS-GS have been identified in the genomes of 19 mostly studied Gram-positive bacteria, including *S. mutans* [144]. All components of the glutathione system were identified in all the 10 strains of this study. Several *S. mutans* strains, namely UA159, 5DC8, KK21, KK23, ATCC 25175, and NCTC 11060, as well as *S. ratti* DSM 20564, possess two glutathione reductase orthologs (*SMU.140* and

SMU.838). This could possibly convey these strains certain advantages in the re-generation of GSH from GSSG, which in turn would be helpful for oxidative resistance.

In addition, 3'-phosphoadenosine-5'-phosphate phosphatase activity has recently been reported to be required for superoxide stress tolerance in *S. mutans* [145]. Putative 3'-phosphoadenosine-5'-phosphate phosphatase coding genes were identified in all strains of this study as well (Table 3.8).

Table 3.8 Distribution of oxidative stress resistance systems

Class	Name	UA159	NN2025	5DC8	KK21	KK23	AC4446	ATCC25175	NCTC11060	DSM20564	DSM20742
SOD	Sod[a]	SMU.629	GI|290580884	D816_02695	D817_02943	D818_03247	D819_02714	D820_02900	D821_02804	D822_02694	D823_08152
	None[b]	SMU.1297	GI|290580288	D816_05819	D817_06013	D818_02305	D819_05840	D820_05873	D821_05905	D822_08440	D823_09052
AhpF/AhpC	AhpC[c]	SMU.764	GI|290580768	D816_03290	D817_03548	D818_03807	D819_03314	D820_03512	D821_03389	D822_08028	-
	AhpF[d]	SMU.765	GI|290580767	D816_03295	D817_03553	D818_03812	D819_03319	D820_03517	D821_03394	D822_08023	-
Dpr	Dpr[e]	SMU.540	GI|290580957	D816_02305	D817_02548	D818_02835	D819_02354	D820_02520	D821_02374	D822_04226	D823_02352
	TrxB[f]	SMU.463	GI|290581031	D816_01940	D817_02188	D818_02007	D819_01989	D820_02130	D821_01999	D822_06878	D823_01947
Thioredoxin	TrxB[g]	SMU.869	GI|290580673	D816_03785	D817_04038	D818_04292	D819_03804	D820_04002	D821_03854	D822_03499	D823_01550
	TrxA[h]	SMU.1869	GI|290579800	D816_08398	D817_08588	D818_08193	D819_07664	D820_08390	D821_08394	D822_08270	D823_06913
	TrxH[i]	SMU.1971c	GI|290579712	D816_08848	D817_09038	D818_08623	D819_08094	D820_08775	D821_08804	D822_07458	D823_08552
	None[j]	SMU.1169c	GI|290580401	D816_05229	D817_05413	D818_05692	D819_05219[α]	D820_05307	D821_05309	D822_06958	-
	Tpx[k]	SMU.924	GI|290580628	D816_04015	D817_04278	D818_04517	D819_04114	D820_04227	D821_04084	D822_03359	D823_07595
Glutaredoxin	GshAB[l]	SMU.267c	GI|290581223	D816_01065	D817_01215	D818_01054	D819_01078	D820_01251	D821_01091	D822_01287	D823_06703
	GshR[m]	SMU.838	GI|290580702	D816_03640	D817_03893	D818_04147	D819_03659	D820_03857	D821_03709	D822_01904	D823_04976
	GshR[n]	SMU.140	-	D816_00620	D817_00640	D818_00624	-	D820_00607	D821_00626	D822_06143	-
	NrdH[o]	SMU.669c	GI|290580848	D816_02885	D817_03143	D818_03447	D819_02894	D820_03090	D821_03009	D822_02899	D823_05398

[a] Superoxide dismutase; [b] 3'-Phosphoadenosine-5'-phosphate phosphatase; [c] Alkyl hydroper oxide reductase, subunit C; [d] Alkyl hydroperoxide reductase, subunit F; [e] Peroxide resistance protein / iron binding protein; [f] Thioredoxin reductase (NADPH); [g] Thioredoxin reductase; [h] Thioredoxin; [i] Thioredoxin family protein ; [j] Thioredoxin family protein; [k] Thiol peroxidase ; [l] Glutathione biosynthesis bifunctional protein; [m] Glutathione reductase; [n]Glutathione reductase ;[o] Glutaredoxin; [α]Additional homolog was identified D819_05259;

3.6 Metabolic network construction and analysis

3.6.1 Genome-scale metabolic network reconstruction

In order to systematically reveal the metabolic variability of the mutans streptococci in this study, the genome-scale metabolic networks of all the strains sequenced were reconstructed and analyzed according to the method proposed by Ma and Zeng [43, 44]. All annotated protein sequences having EC numbers were considered for the network reconstruction. From the functional annotation discussed in chapter 3.2, total EC numbers identified in the 10 strains are very close to each other. A summary of the total numbers of reactions and metabolites in each of the reconstructed metabolic networks is shown in Table 3.9, and all the constructed metabolic networks can be found in an online file[vi] in *.cys format which can be opened with Cytoscape [45], a software for visualization and analysis of biological networks. The sizes of the constructed metabolic networks of the eight *S. mutans* strains are very close to each other, with UA159, NN2025, AC4446, 5DC8 and KK21 having almost exactly the same size, and the networks of KK23, ATCC 25175 and NCTC 11060 being merely about 2% larger. While the size of the metabolic network of *S. ratti* DSM 20564 is comparable to those of the *S. mutans* strains, the metabolic network of *S. sobrinus* with 833 reactions and 853 metabolites is the smallest one, which have 62 less reactions and 60 less metabolites compared to the largest one of *S. mutans* NCTC 11060 (895 reactions and 913 metabolites).

Table 3.9 Compositions of the established metabolic networks of the 10 mutans streptococci strains

Strain	EC Numbers	Reactions	Metabolites
S. mutans UA159	454	875	893
S. mutans NN2025	450	874	892
S. mutans 5DC8	453	875	893
S. mutans KK21	453	875	893
S. mutans KK23	452	893	911
S. mutans AC4446	449	874	893
S. mutans ATCC 25175	453	891	911
S. mutans NCTC 11060	456	895	913
S. ratti DSM20564	435	888	893
S. sobrinus DSM20742	434	833	853

[vi] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S6.cys

3.6.2 Variability and specificity in metabolic pathways and network

Despite the comparable network sizes, however, all the strains possess or lack certain reactions/metabolites, as revealed by detailed comparative analyses. Using the metabolic network of *S. mutans* UA159 as reference, the presence and absence of reactions in each of the strains/species compared were discovered and mapped into sub-pathways based on the KEGG pathway classification (http://www.genome.jp/kegg/pathway.html). As the result, among the 416 sub-pathways defined in the KEGG pathway database certain variations between the strains/species were found in 46 sub-pathways(Please refer to an online file for details[vii]).

A key feature of the oral environment is that the nutrients available to the oral bacteria are always fluctuating between abundance and famine associated with human diet. Thus, the ability to quickly acquire and metabolize carbohydrates to produce energy and precursors for biosynthesis is essential for the survival of all oral bacteria. Due to their key roles in carbohydrates metabolism and energy production, glycolysis/gluconeogenesis, TCA cycle and pyruvate metabolism pathways are generally considered to be highly conserved among these oral bacteria. Although mutans streptococci strains/species are closely related species as revealed by phylogenetic tree analysis in this study (Figure 3.1), differences in these central carbon metabolic pathways were found as shown in Figure 3.6.

Facultative anaerobes such as lactic acid bacteria including *Streptococcus* lack cytochrome oxidases required for energy-linked oxygen metabolism. Energy (in the form of ATP) required for survival and growth are generated by substrate level phosphorylation in the glycolysis pathway [136]. L-lactate oxidase (D823_06598) with a similarity of 73% to YP_003064450.1 (accession number) of *Lactobacillus plantarum* JDM1 and lactate oxidase (D823_06595) with a similarity of 65% to ZP_09448656.1 (accession number) of *Lactobacillus mali* KCTC 3596, were found to be uniquely present in *S. sobrinus* DSM 20742. These two enzymes catalyze the reaction of L-Lactate + O_2 => Pyruvate + H_2O_2 and/or D-Lactate + O_2= > Pyruvate + H_2O_2. It has been reported that in *S. pneumonia* concerted action of lactate oxidase and pyruvate oxidase forms a novel energy-generation pathway by converting lactate acid to acetic acid under aerobic growth conditions [146]. Because no pyruvate oxidase could be identified in *S. sobrinus* DSM 20742, the function of the lactate oxidases in *S. sobrinus* DSM 20742 should be different to that of *S. pneumoniae*. By a close examination, it is reasonable to hypothesize that lactate oxidase, together with pyruvate dehydrogenase, phosphate acetyl transferase and acetate kinase, could form a novel energy production pathway to convert lactate acid to acetate and simultaneously produce one additional ATP, as depicted in Figure 3.7. By doing so, the lactate oxidases of *S. sobrinus*

[vii] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S7.docx

DSM 20742 could also play a role in consuming lactate to regulate pH, which would be an advantage for *S. sobrinus* DSM 20742 in resistance to acid stress. In addition, this pathway could replenish acetyl-CoA, an important intermediate for the biosynthesis of fatty acids and amino acids. This is for the first time that such an energy production pathway is proposed in *Streptococcus* species. Furthermore, lactate oxidase and lactate dehydrogenase could form a local NAD^+ regeneration system, which would be certainly advantageous to *S. sobrinus* DSM 20742 under aerobic growth conditions. Moreover, it is known that mutans group streptococci and the mitis group streptococci are competitors, with *S. mutans* producing mutacins to kill the mitis group streptococci and the mitis group streptococci in turn produce H_2O_2 to kill mutans group streptococci [67, 147]. Favored by possessing the lactate oxidases, *S. sobrinus* DSM 20742 has the potential ability of producing H_2O_2 to kill not only competitors (oxygen sensitive *S. mutans*, oral anaerobes) but also macrophages [148], and thereby defend its ecological niche. The unique presence of lactate oxidases in *S. sobrinus* DSM 20742 was verified by PCR experiments (Please refer to an online file for details [viii]). Later, another *S. sobrinus* strain AC153 was also found to harbor homologous genes of lactate oxidase, suggesting that lactate oxidase may be conserved and play an important role in *S. sobrinus*. In the effort to clarify the functionality of lactate oxidase it was tried to knock out the two genes encoding the two enzymes by PCR ligation mutagenesis according to the method of Lau PC *et al.* (2002). Although different transformation methods (two natural transformation methods and two electroporation methods) were applied, but it was failed to obtain the desired recombinants. Then, to find out if *S. sobrinus* DSM 20742 is in the least able to enter genetic competence state, transforming *S. sobrinus* with plasmids which are replicative in other *Streptococcus spp.* like pDL278 (Sp^r, pAT18 Em^r, with suicide vector pFW5 Sp^r) in both circular and linearized forms were tested but no transformants could be obtained. Therefore, it is clear that the genetic competence behavior of *S. sobrinus* DSM 20742 is very different to that of *S. mutans*, attributing very likely to the lacking of the genes *comSR* and *comC*.

In contrast to the unique harboring of lactate oxidases in *S. sobrinus* DSM 20742, citrate lyase (EC 4.1.3.6), which catalyzes the cleavage of citrate into oxaloacetate and acetate, and oxaloacetate decarboxylase (EC 4.1.1.3), catalyzing the irreversible decarboxylation of oxaloacetate to pyruvate and CO_2, are not present in *S. sobrinus* DSM 20742, as shown in Figure 3.7 by the blue dotted lines. It has been reported that citrate lyase functions as a key enzyme in initiating the anaerobic utilization of citrate by a number of bacteria, and the further catabolism of oxaloacetate formed takes place either by decarboxylation or by reduction. In some organisms, oxaloacetate is decarboxylated to pyruvate by oxaloacetate

[viii] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751929/bin/1471-2164-14-430-S8.docx

decarboxylase, which is also induced in the presence of citrate. The two enzymatic reactions, which occur sequentially, constitute the 'citrate fermentation pathway' [149]. The absence of citrate lyase and oxaloacetate decarboxylase implies that *S. sobrinus* DSM 20742 might lacks the ability in anaerobic utilization of citrate as a substrate. However, the disadvantages of *S. sobrinus* DSM 20742 in citrate utilization could be offset by the novel energy production pathway from lactate to acetate proposed above.

A putative pyruvate-phosphate dikinase (EC 2.7.9.1), which catalyzes the interconversion between PEP and pyruvate, was found to be uniquely present in *S. ratti* DSM 20564. Pyruvate-phosphate dikinase has been found in propionic acid bacteria [150] . The large difference in the standard free energy of hydrolysis for ATP to AMP and pyrophosphate (-7.6 kcal/mole) and for PEP to pyruvate (-13.6 kcal/mole) at pH 7.0 indicates that the equilibrium for the reaction it catalyzes would strongly favor pyruvate formation. But studies in *Acetobacter xylinum* clearly indicate that the function of this enzyme under physiological conditions favors the process of gluconeogenesis [151]. Metabolite interconversion at the PEP-pyruvate-oxaloacetate node involves a structurally entangled set of reactions that interconnect the major pathways of carbon metabolism and thus, is responsible for the distribution of the carbon flux among catabolism, anabolism and energy supply of the cell [152]. Under glycolytic conditions oxaloacetate is generated by carboxylation of PEP and/or pyruvate catalyzed by PEP carboxylase (PEPCx) and/or pyruvate carboxylase (PCx). In this study PCx was not found in any of the mutans streptococci strains.

All the 10 strains of this study possess similarly an incomplete TCA cycle and the primary role of the existing TCA enzymes is most likely the synthesis of amino acid precursors as has been reported previously [28, 153].

3.7 Construction of StrepReg - a regulation database of *S. mutans*

Detailed knowledge of all interactions between proteins/genes in a given cell would represent an important milestone towards a comprehensive description of cellular mechanisms and functions. To enable a global view of all interactions in *S. mutans*, three sources of functional information of genes/proteins were integrated into one database named StrepReg: 1) the regulatory network constructed based on assigning predicted relationships between transcription factors (TFs) and target genes (TGs). This regulatory network is a "static" connection network consisting 1,785 TF-TG relationships corresponding to 32 regulons [154].

2) protein-protein association network released by STRING database (https://string-db.org/). 3) KEGG pathway information (http://www.genome.jp/kegg/).

The platform was constructed using the open source web framework CakePHP (https://cakephp.org/) and the open-source relational database management system MySQL (https://www.mysql.com/). The visualization function of the regulatory networks was implemented using D3.js (https://d3js.org/). The database was provided at: http://biosystem.bt1.tu-harburg.de:1555/homes/. Figure 3.8 shows two screen shots of the StrepReg database.

3.8 Conclusion

The genomes of 8 mutans streptococci strains, including six *S. mutans* strains, one *S. ratti* strain and one *S. sobrinus* strain were sequenced, annotated and compared together with *S. mutans* UA159 and NN2025. Multiple genome alignment showed extensive genome rearrangement among the eight strains of *S. mutans* . The core-genome size of *S. mutans* was determined to be around 1,370 genes by including 67 *S. mutans* genomes available in the NCBI database. A possibly open pan-genome of *S. mutans* was inferred.

Systematic comparative analyses were focused on competence regulation, bacteriocin (mutacin) production, antibiotic resistance, oxidative stress resistance, as well as central carbon metabolism and energy production pathways. Most of these cellular functional systems show remarkable differences between the strains, especially between the species with the mutans group streptococci, except for oxidative stress resistance systems which are well conserved. For example, CSP-dependent and independent competence regulation systems are highly diverse in mutans streptococci while no comC-like genes could be identified in *S. ratti* and *S. sobrinus*; putative ComC amino acid sequences of *S. mutans* strains show clear variations; ComS and ComR are also absent in *S. sobrinus* which well explains the fact that it was not able to obtain genetic competence state of *S. sobrinus* by experiment, even though the ComX and the downstream competence development genes are well reserved. Furthermore, the response regulators of the HdrMR and BsrRM systems, which are known to be also involved in competence development, are missing in both *S. ratti* and *S. sobrinus*.

Variation in the presence/absence of mutacin-encoding genes is accompanied with the conservation of mutacin immunity proteins, which indicates apparently important roles of the mutacin immunity proteins for the survival of these mutans streptococci in a bacteriocin rich environment. The presence of various antibiotic resistance factors, together with the open pan-genome inferred, implies that attention should be paid to the potential of mutans group streptococci in the development of antibiotic resistance.

The sizes of the genome-scale metabolic networks of the 10 strains are very close to each other. Comparative analysis of sub-pathways using *S. mutans* UA159 as reference reveals that 46 sub-pathways of all 416 sub-pathways as defined in KEGG pathway database show variations between the strains. By identifying lactate oxidases to be uniquely present in *S. sobrinus* DSM 20742, for the first time a novel energy production pathway in *S. sobrinus* is proposed. Additional functions of the lactate oxidases in connection with the proposed energy production pathway are also discussed.

An online regulation database for *S. mutans*, named StrepReg, was constructed by integrating transcription factor-based gene regulatory network, which was derived from time-series transcriptome analysis, with information from STRING interaction database and KEGG pathway database (http://biosystem.bt1.tu-harburg.de:1555/homes/).

In conclusion, the genomes of mutans group streptococci display remarkable differences, especially between different species. The strain-specific information provided in this study can be helpful in understanding the evolution and adaptive mechanisms of those oral pathogens.

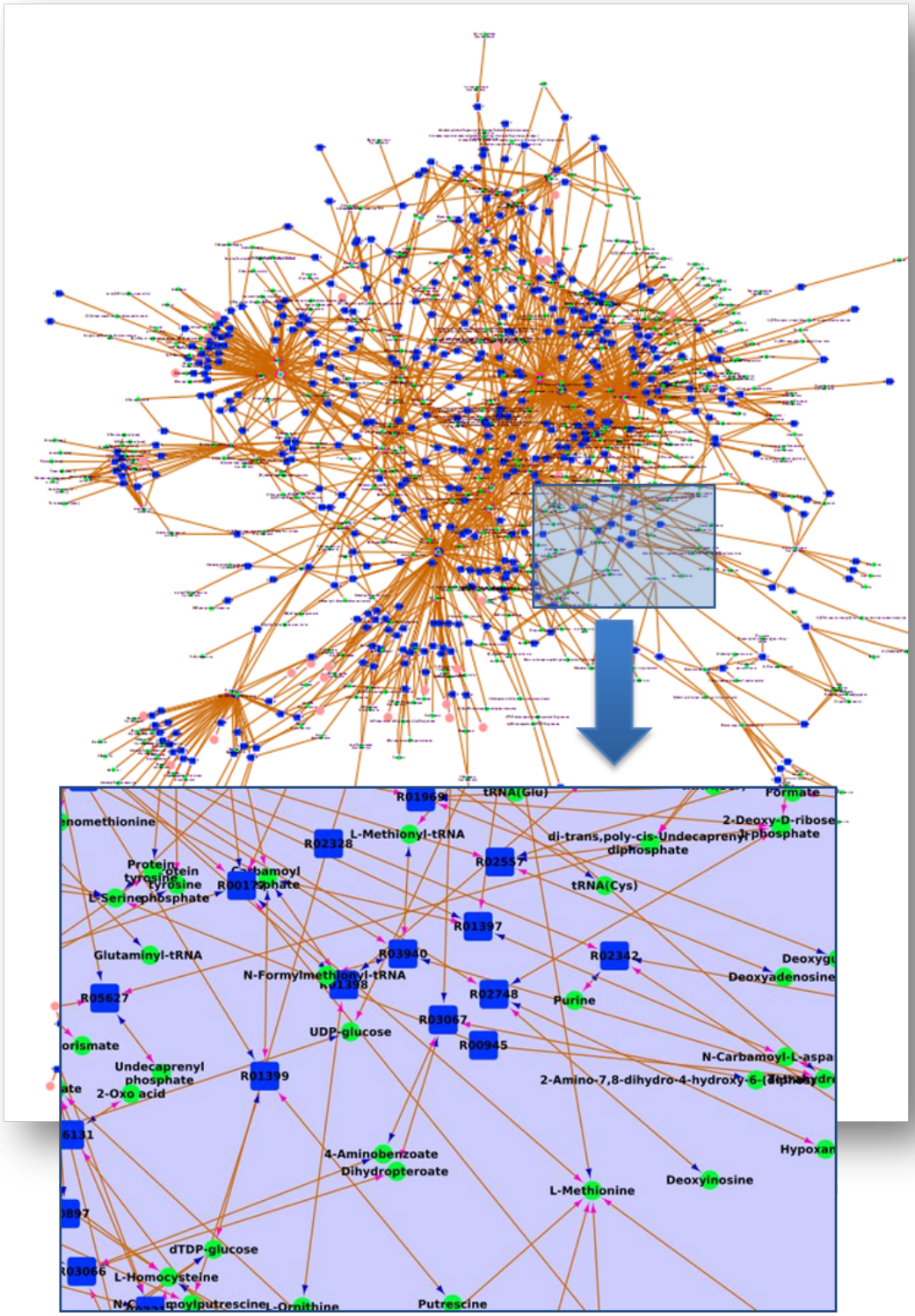


Fig. 3.6 Example of visualized genome-scale metabolic networks constructed based on genome annotations and KEGG pathway

The blue rectangle nodes represent the reactions and the circle green nodes represent the metabolites.

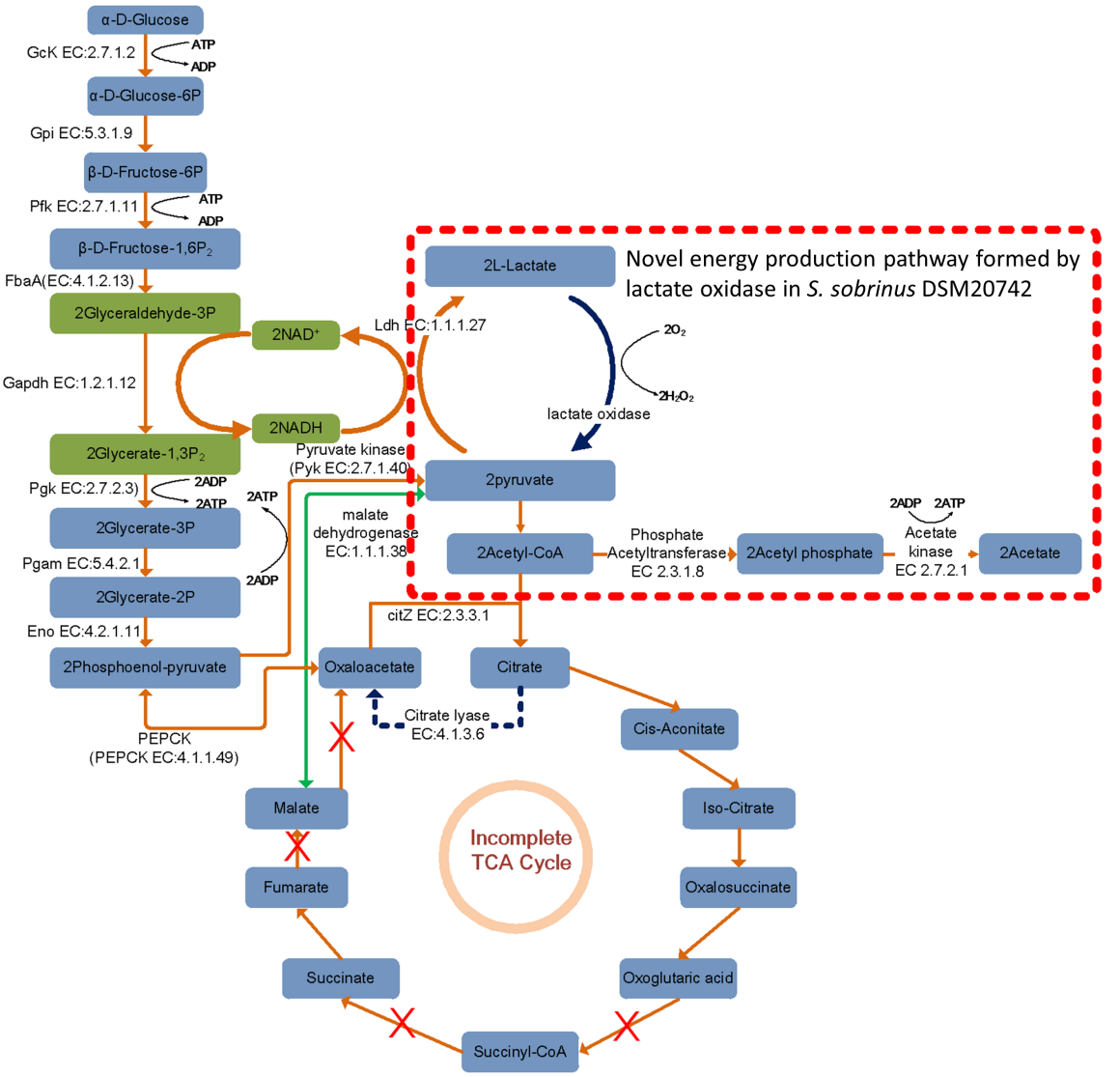


Fig. 3.7 Glycolysis/Gluconeogenesis and TCA cycle pathway in mutans streptococci

The rectangle nodes represent the metabolites. The yellow lines represent enzymes and the blue line represent enzymes with diversities across mutans streptococci strains studied here. The yellow line with cross means this enzyme is not present in all strains. Dotted blue line means this enzyme is absent in *S. sobrinus* DSM20742 and solid blue line means this enzyme is uniquely present in *S. sobrinus* DSM20742. Malate dehydrogenase represented in green line was absent in *S. mutans* NN2025 and *S. mutans* AC4446.

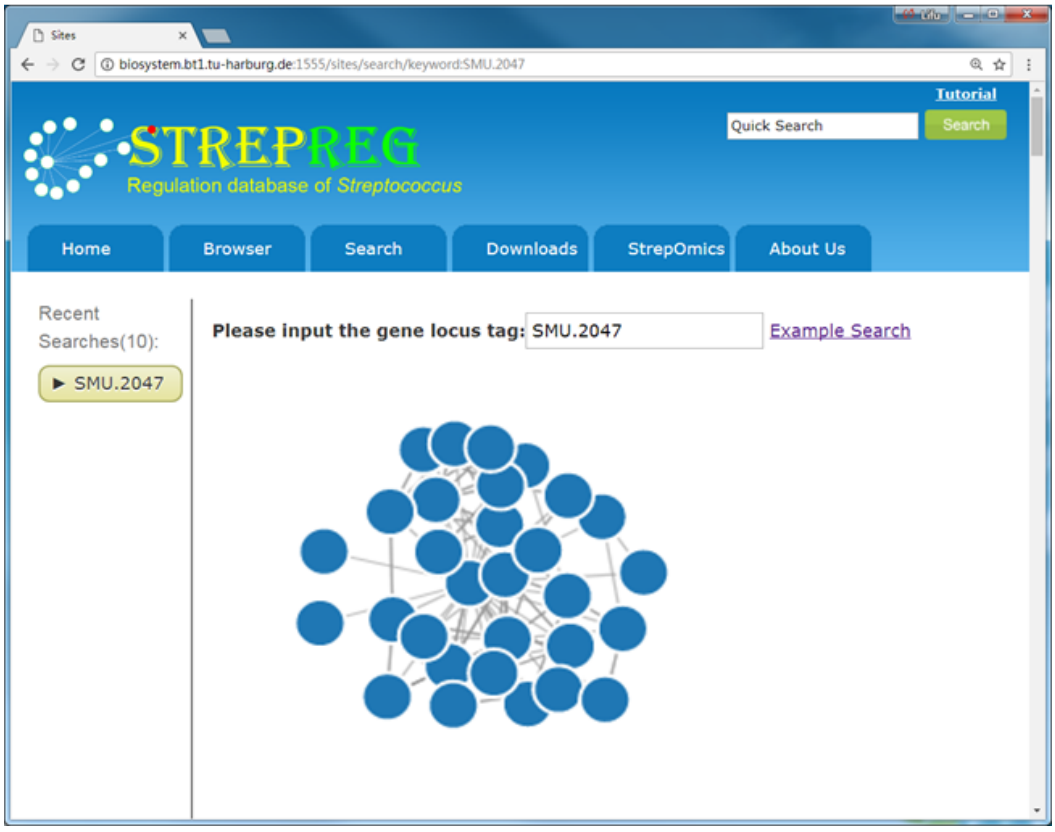


Fig. 3.8 Screenshots of StrepReg database

Chapter 4

Development of a multiple IO system for biological engineering in *E. coli*

4.1 Introduction

Although cells are composed of molecules and their viability relies on extracting and using energy to maintain them, they are not 'just' matter and energy [155]. Cells can respond to their environment, make decisions, build structures, and coordinate tasks based on computational operations performed by networks of regulatory proteins that integrate signals and control the timing of gene expression [155]. It has been shown that cells can be programmed using synthetic genetic circuits composed of regulators organized to generate desired operations [155–159]. Stimulated by the great potential of engineering biological systems to achieve novel tasks, an emerging discipline termed synthetic biology is drawing more and more attentions [160–184]. It focuses on designing and building novel biological functions and systems by combining science and engineering principles, including the design and construction of new biological parts, devices, and systems, as well as the re-design of existing, natural biological systems for useful purposes. In general, the overall process of biological engineering is similar to programming in computer science. However, unlike programming on a computer, "programming" a biological system is much more time- and labor-intensive. One reason is that changing the "biological codes" is much more difficult than changing digital codes on a computer. It always takes days or even weeks to enable the editing of the "genetic codes". Recently, this process has been greatly simplified by the recently emerged CRISPR/Cas9 based genetic editing tools [185–188]. Another more crucial

This chapter was a modified and extended version of a recent publication: Song, Lifu; Zeng, An-Ping (2017): Engineering 'cell robots' for parallel and highly sensitive screening of biomolecules under *in vivo* conditions. Scientific Reports 7 (1), p. 15145.

reason is the inherent complexity and uncertainty of the genotype-phenotype relationships of biological systems. Despite the complicate interactions among the metabolic, gene regulatory and signaling networks at the cellular level, it is not possible to precisely predict the consequences of even a single base change at the single gene level. Hence, the biological engineering process is held by the time- and labor- intensive design–build–test cycles as shown in Figure 1.1, in which many designs have to be evaluated and iterated on in order to improve the performance of target system. The rate of improvements is directly related to the throughput and rounds of the design cycles, with higher throughputs and more rounds resulting in reduced development period. Although recent advances have enabled the design and construction of billions of genetic variants per day, but evaluation capacity is still limited to thousands of variants per day.

Inspired by the debugging system in computer science, a versatile diagnosis system was proposed to reduce the development burden for biological engineering. This was achieved by a novel multiple input-output (IO) system which can interact with the cells and output multiple signals corresponding to various perturbations (inputs). Despite impressive progress in systems metabolic engineering and synthetic biology, there are still unsolved major problems in their practical applications for developing effective microorganisms for biosynthesis, such as identification of relevant targets for pathway engineering, designed elements or devices from synthetic biology often not working well inside cells under industrially relevant conditions. For proof of concept, the IO system used for target identification, evaluation of designs, evolution and selection of key enzymes for bioproduction.

4.2 Principles of a multiple input-output system which can interact with *E. coli* cells

The inputs here refer to operations/perturbations that can alter intracellular conditions. The concentrations of chemicals, e.g. IPTG, arabinose, have been widely used as inputs in previous studies. However, the availability of well-characterized chemicals is limited and intracellular genetic parts have to be built in order to sense the desired signal and change intracellular gene expression pattern accordingly. Here various DNA fragments were proposed to be utilized as inputs directly. The DNA fragments with different functions can be easily designed and they are different from each other inherently in their sequences. The possible inputs are unlimited in principle. The challenge is how to introduce these DNA fragments into the cells.

M13 filamentous phage can infect F^+ *E. coli* cells without lysing the cells, making it an ideal raw material for designing the input system. There are two kinds of disturbances to the gene expressions, over-expression and repression. For the over-expression of target genes, the specific genes cloned into M13 phages could be easily over-expressed by using a strong RBS because the copy number of the phages is up to 200-300 per cell. For the repression of target genes, the small regulatory RNA mediated repression system is applied here [189, 190]. The target gene expression could be repressed by a designed small RNA which can reversely bind to the mRNAs of the target gene. As shown in Figure 4.1, for over-expression of a specific gene, the phage structure of a) was used and for repression of a specific gene, structure b) was used.

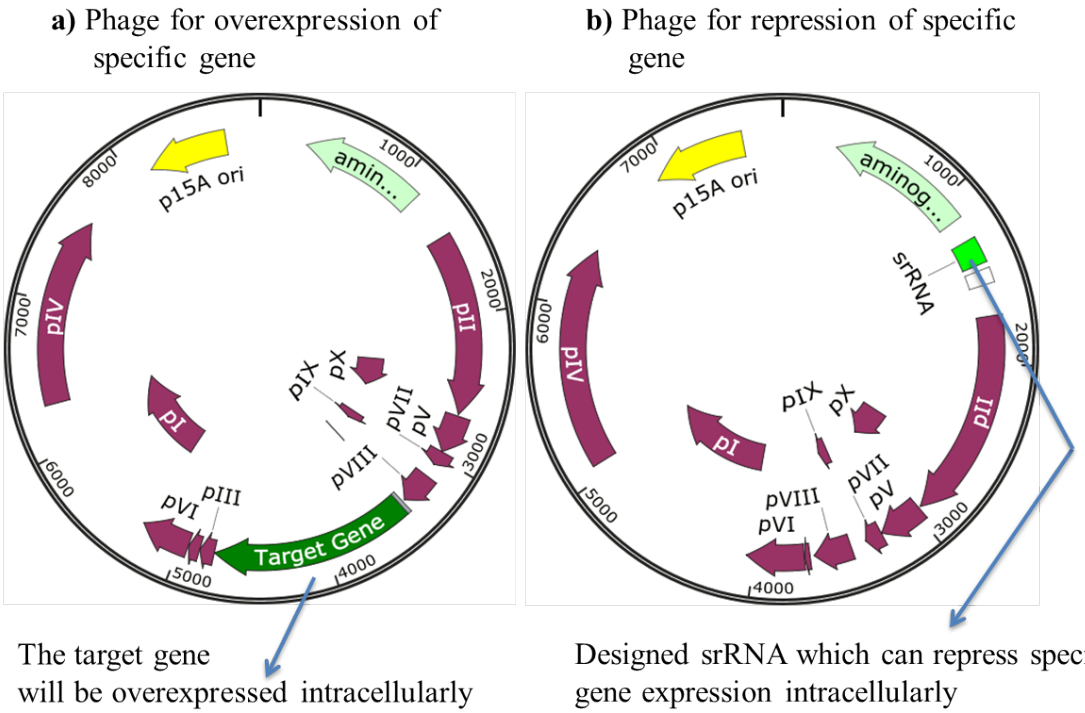


Fig. 4.1 Sample plasmid maps of inputting phages carrying out overexpression/repression operation on specific genes

Multiple inputs require various M13 phages carrying different DNA fragments to conduct the input operations. The populations of various types of phages are ideal 'variants' for recording the intracellular signal changes caused by the corresponding phages. To enable this an intracellular sensor was used to control the phage populations to record the specific intracellular signal changes in the form of corresponding phage populations. The principle of phage based multiple IO system is shown in Figure 4.2. Specifically, *E. coli* cells were first enabled to control the infectivity of the packaged phages by transferring an essential

gene for phage infectivity from the M13 phage into *E. coli* cells. The essential gene applied in this study is Gene III encoding the attachment protein G3P which mediates adsorption of the phage to its primary receptor, the tip of *E. coli* F-pilus [191]. Next, an intracellular biological circuit was designed and implemented to control the infectivity of the packaged phages by controlling the expression level of Gene III based on a specific intracellular signal, such as the concentration of a product or an upstream metabolite. Thus, the expression of protein III (G3P) varies depending on the signal strength and in turn will affect the input phage population after several rounds of phage replications, which means that the signal is 'recorded' in the form of corresponding phage populations. As illustrated in Figure 4.2, phages a, b, c carry out different operations A, B, C in the host cells. After infecting the host cells, phage b could produce more new generations of phages since the intracellular signal is increased by the operation B, which leads to a high expression level of G3P while phage A can only produce a limited amount of phages and phage C cannot produce infective phages. The resulted phage populations are related to the intracellular signals caused by the operations introduced by the corresponding phages. The outputs can be easily read out by colony sequencing or various high-throughput sequencing technologies.

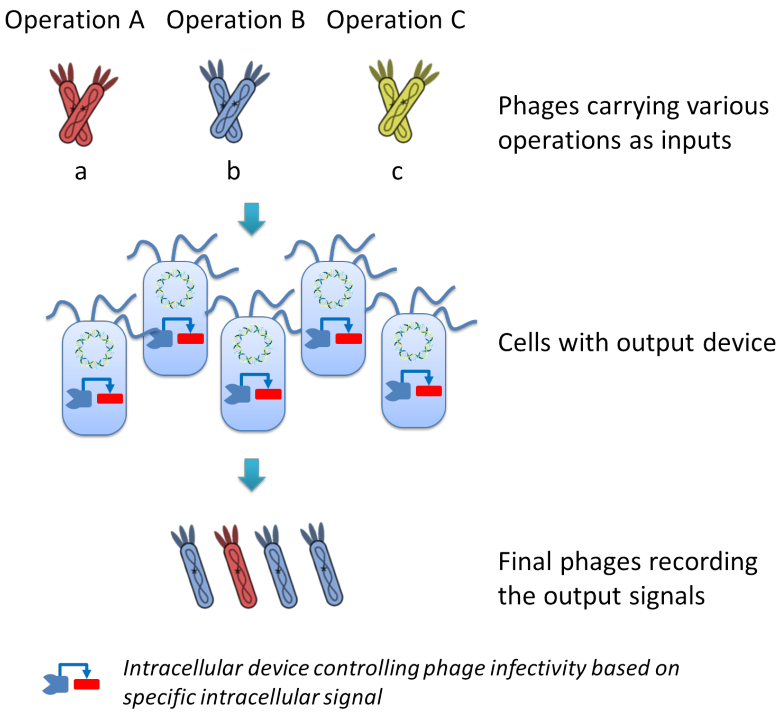


Fig. 4.2 Principle of phage based multiple IO system

4.3 Proof of concept studies

Three different proof of concept studies, e.g., identification of beneficial genetic manipulations, parallel evaluation of designs, evolution and selection of key enzymes in the lysine biosynthesis by *E. coli* as shown in Figure 4.3, have been performed.

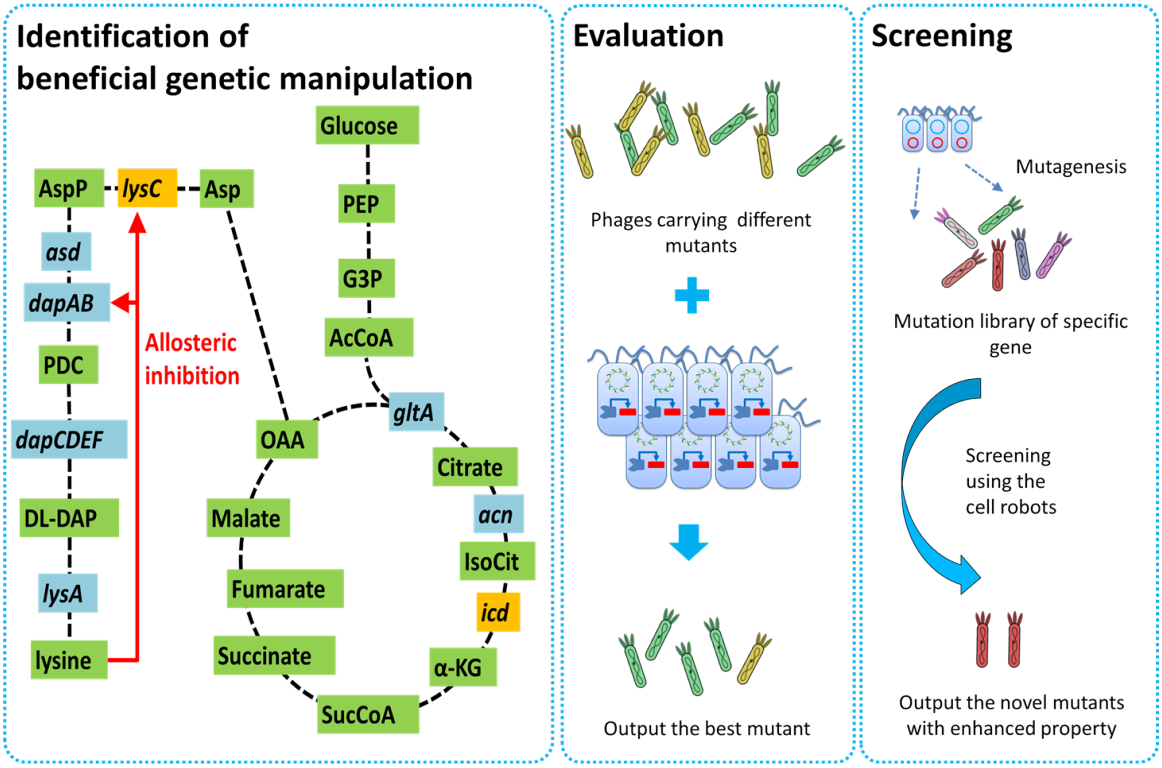


Fig. 4.3 Proof of concept application studies of the IO system

Various genes/operations related or not related to the lysine biosynthesis in *E. coli* were used as inputs and the changes of intracellular lysine concentration were used to trigger output signals. To use the intracellular lysine concentration as a signal for outputs, a lysine inducible promoter was cloned from *Corynebacterium glutamine* ATCC13032 as the lysine sensor [192]. The Gene III from M13 phage was cloned to a plasmid under the control of the lysine inducible promoter (Figure 4.4. The obtained plasmid namely AP-Lys-B is the output device using lysine as a signal. The host cells used in the following studies, namely *E. coli* XL-Blue-AP-Lys-B, were generated by transforming *E. coli* XL-Blue cells with the plasmid AP-Lys-B. If a phage carrying a gene/operation that can increase the intracellular lysine concentration is absorbed by the *E. coli* XL-Blue-AP-Lys-B cells, it will increase the Gene III expression level and this will in turn result in the production of infectious phages. Thus, the final populations of different types of phages will reflect the corresponding changes

of intracellular lysine concentration caused by the corresponding types of phages. In other words, the signals are recorded in forms of various phage populations.

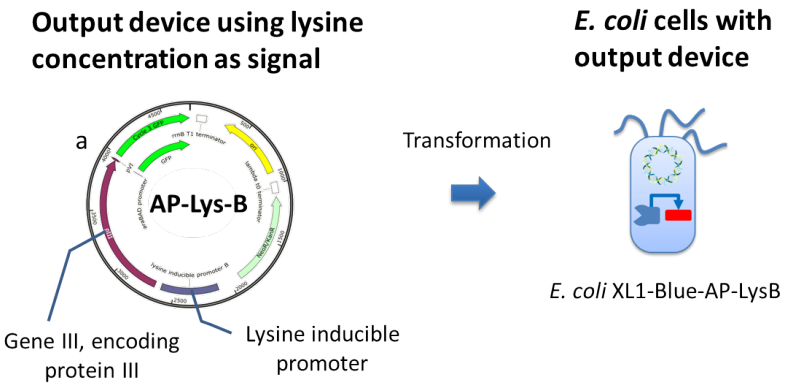


Fig. 4.4 Illustration of the output device using the concentration of intracellular lysine as an output signal

a – Plasmid map of the biological output device using intracellular lysine concentration as signal. The gene III transcription level is designed to be controlled by a lysine inducible promoter cloned from *Corynebacterium glutamicum* ATCC13032. A green fluorescence protein encoding gene is placed downstream of gene III under the control of the same promoter. The green fluorescence protein encoding gene is not required for the IO system. It was used for comparing the sensitivity of IO system with flow cytometry-based methods.

4.3.1 Identification of beneficial genetic manipulations

Due to the complexity of the cellular functions, identification of the beneficial genetic manipulations is one of the key challenges in metabolic engineering. To avoid laborious try-and-error experiments, many studies have been focused on building genome-scale models of cellular functions to make predictions [193]. One of the most useful systems-based tools for metabolic engineering is the *in silico* genome-scale metabolic reconstruction and flux analysis [193]. The recent studies in kinetic modeling show many difficulties and are limited to small scale networks [194]. Metabolic and kinetic models lack the crucial information on regulation and interactions which are, in principle, essential for prediction and reprogramming of cellular functions. We are still far away from having kinetic and regulatory models that are good enough for the design of industrially competitive cell factories [194]. Although cells are composed of molecules and their viability relies on extracting and using energy to maintain them, they are not 'just' matter and energy. Information processing, also called "cellular computing", is essential for cellular function [155] . Several previous studies have proved that the computational abilities of biological system could be used in rational ways [155–159]. It is interesting to ask the question "Can we rein the computation abilities

of cells for systems-level prediction and optimization of microorganisms?" If so, we then do not need to build laborious mathematic models of the whole cell since the perfect model, the cells, is already there. The key issue is how to let the cells "compute" the processes we are interested in and output the results corresponding to the different inputs? In the current study, efforts on utilizing the computation abilities of the cells to make "predictions" by the phage-based multiple IO systems have been made.

For a proof of the principle, the efforts were focused on *lysC* gene encoding aspartate kinase III (AK-III) which catalyzes the phosphorylation of aspartate and controls the biosynthesis of several industrially important amino acids such as lysine, threonine, and methionine in *E. coli* [195]. For the construction of the inputting phage to over-express *lysC*, the wild-type *lysC* gene was cloned from *E. coli* MG1655 to VCSM13 phage replacing the Gene III. The obtained plasmid is named as M13-lysC. For the construction of M13-srRNA-lysC which can repress *lysC* expression, a srRNA fragment which targeting *lysC* was synthesized by Invitrogen and cloned into the intergenic region (upstream gene II) of M13-rmGIII. Roughly the same amount of two phages were put together with *E. coli* XL1 blue F^+ host cells carrying the designed output system. After 4 hours of co-cultivation, the different phage populations were determined by colony PCR and sequencing. Only a large amount of M13-lysC phages was detected. According to the principle of the output system, higher phage population indicates a higher intracellular lysine concentration caused by the corresponding operation. Thus, the correct 'prediction', e.g. up-regulation of *lysC* expression can enhance the lysine productivity, is obtained.

4.3.2 Evaluation of designs

As mentioned above, due to the inherent complexity of biological systems, biological engineering efforts always have to evaluate many designs/variants to obtain optimized biological parts/devices with desired properties for pathway engineering or biological circuits as shown in Figure 1.1. Here some efforts on utilizing the proposed IO system for parallel evaluation of multiple designs/variants were made. For proof of concept, the evaluation of different mutants of *lysC* gene was investigated.

Several variations of the plasmid were constructed by introducing site mutations to the *lysC* gene of M13-lysC individually. The included mutations are T253R, R305A, H320A, I337P, S338L and V339A, which have been previously proved to be resistant to allosteric inhibition by lysine [196] at various levels. Another mutant R300C obtained from the screening studies was also included as shown in Figure 4.5. Roughly equal amounts of different phages were mixed and screened using the designed host cells. In the first run of screening, 12 colonies were submitted for sequencing which gave the following score of

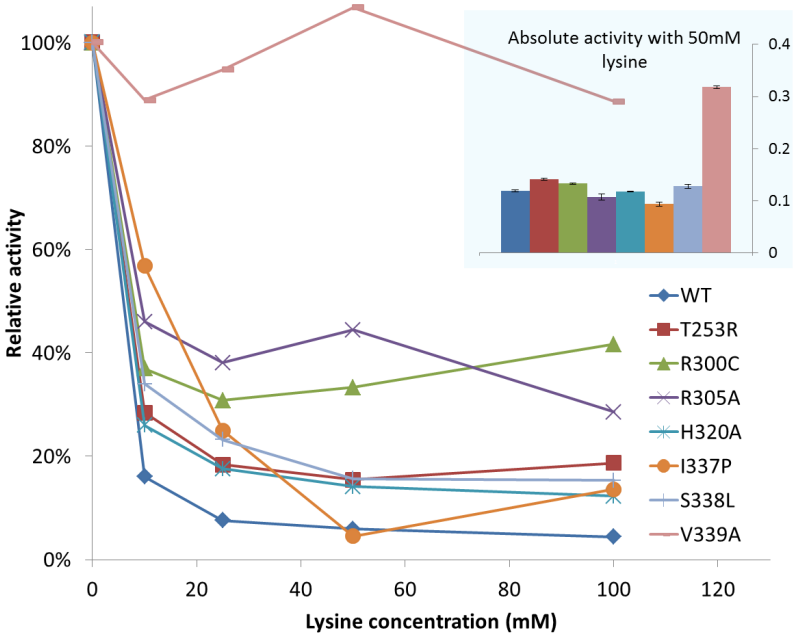


Fig. 4.5 Inhibition profiles of wild-type and mutants of AK-III by lysine

The activities were displayed as relative activities normalized by the specific activities without lysine inhibition. The specific activities with 50mM lysine are presented by measured absorbance normalized by protein concentration show by the small histogram top-right. Data represent mean values and standard deviation from three assays. V339A mutant is shown to be the best one.

the variants: 5 for the variant V339A, 3 for the variant I337P, 2 for the variant S338L, 1 for each of the variant T253R and H320A. In the second run using the phage mixture from the first round of screening, only one colony of V339A was obtained. As shown in Figure 4.5 the variant V339A has the highest activity and resistance against lysine, confirming the effectiveness of the IO system in parallel evaluation of designs.

4.3.3 Parallel and sensitive screening of biomolecules

High-throughput screening (HTS) technologies are powerful tools with many successful applications, especially in the directed evolution of biomolecules such as enzymes. They are primarily based on chemical or physical readouts such as fluorescence and assisted with miniaturized and/or parallel devices such as microfluidics and microchip, increasingly in an automated manner with the help of robotics [197–200]. These systems require expensive infrastructure and special expertise. The major focus was put on speeding up the screening process. For example, the state-of-the-art HTS technology based on fluorescence activated cell sorting (FACS) can reach 18,000-20,000 events per second [201]. However, signal

detection with fast moving cells is challenging resulting with noisy signals as shown in previous studies [202–205]. Furthermore, single cell variations are another source of signal noise which cannot be avoided by FACS based methods [206]. These represent some of the shortcomings of presently used HTS technologies when the molecules to be evolved and optimized are to be used for regulation and improvement of metabolic pathways in the context of metabolic engineering or for creation of new synthetic pathways and regulation tools.

Similar to the electric robots, microbial cells can be considered as a kind of "biological robots" that can sense the information of fast changing environment, compute and make decisions for survival. Cells are highly programmable as proved by recent developments in the field of synthetic biology. Programming cells to perform specific tasks have been successfully achieved in many cases. For example, cells have been programmed to produce pharmaceuticals, fuels, amino acids, fine and bulk chemicals and even metal nanoparticles [207–214]. Cells also have been programmed to sense toxic compounds in environments [215], to record the environment signal in human gut [216] and to eradicate human pathogen[217]. Although the capability of a single cell is limited, cells can reproduce themselves exponentially and work simultaneously to solve complicated tasks or accomplish sophisticated tasks in principle. However, these capabilities of cells have not yet been well exploited, especially for the purpose of HTS.

Recently, concentrations of intracellular molecules have been used as a signal for overexpression of fluorescence for screening purposes in the context of strain improvement [203]. For example, Binder et al. successfully used the intracellular concentration of lysine, a natural lysine-responsible transcriptional activator LysG and fused expression of a yellow fluorescent protein (eYFP) to screen high lysine producer from *Corynebacterium glutamicum* [203]. Later, by using the same sensor for in vivo detection of the desired end-product in single cells, they established a screening method with FACS to screen enzymes without allosteric inhibition. However, due to the complexity of the metabolic pathways, one enhanced enzyme usually has limited effects on productivity of the end-products. Genetic modifications are required to enhance the signals of the end-product in their studies [204].

Esvelt *et al.* (2011) presented an interesting phage-assisted method for continuous evolution of a specific gene-coded biomolecule that is linked to the infectivity of the phage mediated by the expression of a specific protein in host cells [218]. Specifically, M13 filamentous bacteriophage carrying the molecule of interest was used to infect *E. coli* cells in a lagoon with continuous inflow and outflow of the host cells, where the evolving gene is transferred from host cell to host cell in a manner that is dependent on the activity of the molecule of interest. The method was demonstrated with the evolution of a T7 RNA

polymerase with new binding properties. It was later on used to successfully evolve proteases with significantly increased drug resistance to protease inhibitor [219, 220, 220].

The cells can be considered as a kind of “biological robots”. Compared to physical robots the biological robots have the decisive advantage of fast replication, resulting in a large pool for simultaneously screening under *in vivo* conditions. Thus, the screening throughput can be expanded simply by using a larger population of cells, indicating a massively parallel screening manner potentially far beyond the current HTS technologies. It is also worth to mention that the cost for such an approach is almost zero compared to methods based on expensive FACS or microcapillary arrays, making it applicable in almost all biological labs.

Here, the novel IO system designed in this part of study was proposed to programm cells as "screening robots" for parallel and highly sensitive screening of biomolecules for metabolic pathway optimization under *in vivo* conditions. The basic idea is to input may phages carrying various mutants and outputting the phages carrying mutants with desired properties. The concept was demonstrated by screening mutants of a protein with reduced allosteric inhibition. Allosteric regulation is one of the fundamental mechanisms that control almost all cellular metabolisms and gene regulation [221]. Deregulation of allosteric inhibition is essential in designing and optimizing metabolic pathways for the production of target metabolites such as amino acids [196]. AK-III is allosterically inhibited by L-lysine strictly. AK-III was chosen in this work as a model enzyme because of our extensive previous work on the rational design of this enzyme[196, 222]. The new approach is shown to be more sensitive than the widely used flow cytometry method by novel way of signal capturing.

Principle and work-flow of cell robot based screening utilizing the IO system

The workflow of programming cells as robots for the screening of molecule of interest (target) is shown in Figure 4.6. Briefly, instead of placing the screening targets inside of the host cells as in most of the traditional screening methods, the targets to be screened were placed on M13 phages. The host cells were then engineered so that they can screen for phages carrying the targets with desired properties. Specifically, *E. coli* XL1-Blue cells were used as the host cells for this purpose. To enable the host cells to control the infectivity of packaged phages, an essential gene for phage infectivity was transfered from the M13 phage to the host cells. The essential gene applied in this study is gene III encoding the attachment protein pIII which mediates adsorption of the phage to its primary receptor, the tip of *E. coli* F-pilus[191]. An intracellular biological circuit was implemented to control the infectivity of packaged phages by controlling the expression level of gene III based on a specific intracellular signal that is related to the performance of the biomolecules to be screened, such as the concentration of an end product or an intermediate metabolite of a metabolic pathway. The targets are then

cloned into VCSM13 by replacing the original gene III. A helper plasmid pJ175-Str which can supply the gene III product is used for preparing infective phage library at the first step of screening (see below). Elimination of gene III does not affect the phage secretion. However, the infectivity of the produced phages is very low. Thus, to enable an effective screening, a two-step strategy was utilized as illustrated in Figure 4.6.

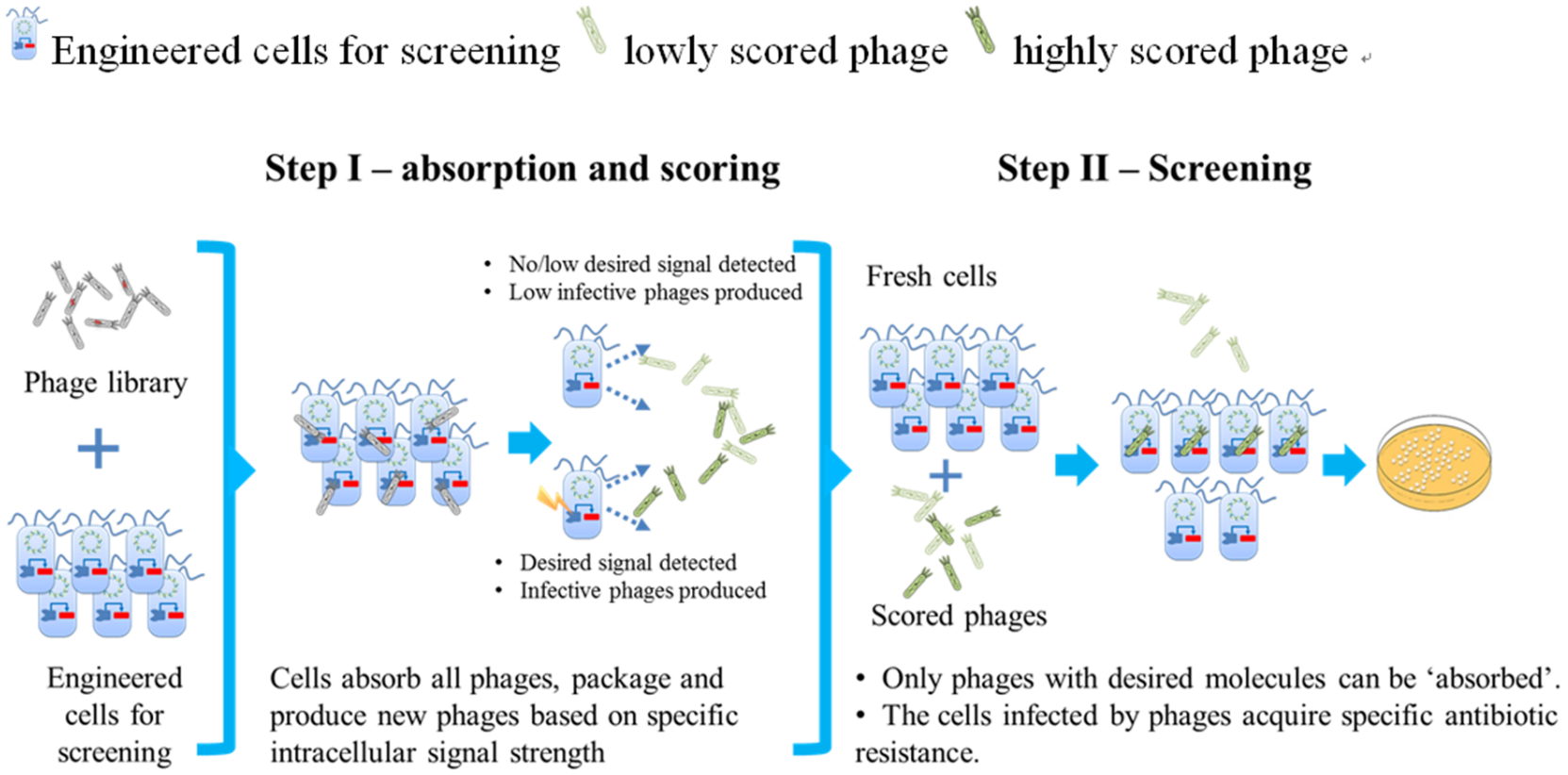


Fig. 4.6 Work flow of cell robot based screening by using the IO system

A two-step screening strategy is suggested. First, phages are absorbed by engineered cells and packaged (scored) based on the performance of the molecules carried by the phages. Only the phages carrying molecules with desired properties are packaged in an effective way. Second, the 'scored' phages are absorbed by fresh host cells. In this step, only the infective phages, i.e. phages carrying molecules with the desired properties, are 'absorbed' by the host cells. The cells infected by the phages with desired molecules/targets acquire kanamycin resistance and can be easily identified by cultivation under kanamycin stress.

In the first step, the phage library with the variants is 'absorbed' and 'scored' by the host cells based on the strength of the specific signal representing the performance of the target molecule. High-performance targets will produce more infectious phages than the low-performance ones. In the second step, the 'scored' phages are collected and screened in another round of cell–phage interactions. In this step, only infectious phages carrying the molecule with desired property can be 'absorbed'. Since a kanamycin resistance gene (*aph*) is placed on the M13 phage, the cells capturing phages with desired properties can be easily selected by incubation under the antibiotic pressure. In such a way, the target with the best performance under *in vivo* conditions can be effectively identified.

Experimental verification of the method

To experimentally demonstrate the functioning of the method, roughly equal amounts of M13-lysC and M13-lysC-V339A phages were mixed and screened using the designed host cells. If the screening robots function as expected, the phages of M13-lysC-V339A should be screened out. The experiments were repeated for three times with the designed host cells using lysine as signals for screening. Once used cells cultivated with LB medium and twice used cells cultivated with M9 medium , concerning the differential expression levels of lysine synthesis pathway genes under various conditions. To verify the genotypes of the resulting phages, plasmids extracted from six individual colonies were sequenced for each experiment. All colonies were verified to be M13-lysC-V339A in all three experiments, confirming a robust screening function of the designed "cell robots".

Generation and cell robot based screening of an AK-III mutation library

In-*vivo* mutagenesis enabled by a mutation plasmid reported previously [54, 223] was applied to introduce random mutations to the *lysC* gene. A library in a size of around 10^4 was created and screened by using roughly 10^5 cellular robots cultivated with LB medium. Many colonies appeared after screening. 16 individual colonies were submitted for sequencing. Two colonies show the same mutation on *lysC* gene resulting in an arginine to cysteine replacement at residue 300 (The mutein formed is named as R300C). All the rest colonies are verified to be the wild-type. The R300C mutein shows significant resistance to lysine as proven by *in vitro* enzyme activity assay as shown in Figure 4.5.

Novel de-regulation mechanism of R300C mutation

AK-III comprises a regulatory domain and a catalytic domain linked by a flexible loop. By comparing the structures of active and inactive state AK-III, it has been indicated that binding

of lysine to the regulatory domain of the active state AK-III triggers a series of changes that release a "latch", the loop (residues 355-365) colored in yellow color as shown in Figure 4.7, from the catalytic domain, which in turn undergoes large rotational rearrangements, promoting tetramer formation and completion of the transition to the inactive state [224]. Rearrangement of the catalytic domain blocks the ATP-binding site, which is the structural basis for allosteric inhibition of AK-III by lysine [224]. All previous studies of removing the allosteric inhibition focused on mutagenesis of the regulatory domains, especially on mutagenesis of the binding sites [225, 55, 224, 226]. Surprisingly, the R300C mutation discovered here located in the flexible loop A (residues 291-300) which link the regulatory and catalytic domains. Besides as a linker of the regulatory domain and the catalytic domain, this flexible loop has not been reported to be related to the allosteric transition process so far. After comparing the structures of the active and inactive states of the wild type AK-III, it was found that the residue 300 arginine forms a hydrogen bond with the residue 293 threonine in the inactive state of AK-III after binding of lysine. The appeal between these two residues forms a 'dragging' force which can deform the flexible loop and form a hairpin structure after the binding of lysine as shown in Figure 4.7a-b. The deformation of the flexible loop might play key roles in the initial steps of the rotational rearrangement of active state AK-III to the inactive state AK-III. The replacement of arginine 300 by cysteine will abolish this 'dragging' force and block the formation of the hydrogen bond. Thus, by blocking the transition from active state to inactive state and unstable the inactive state, the R300C show significant resistance to lysine inhibition.

Screening with a higher sensitivity than fluorescence-based method

To compare the sensitivity of the current screening system with methods based on fluorescence and flow cytometry, the state-of-the-art screening technology, we transformed *E. coli* XL1-Blue/AP-Lys-B cells with M13-lysC, M13-lysC-V339A, and M13-lysC-R300C individually to obtain cells of XL1-Blue/AP-Lys-B/M13-lysC-WT, XL1-Blue/AP-Lys-B/M13-lysC-V339A, and XL1-Blue/AP-Lys-B/M13-lysC-R300C. A GFP-encoding gene was placed under the control of the same lysine sensor in *E. coli* XL1-Blue/AP-Lys-B. Over-night cultivated cells of XL1-Blue/AP-Lys-B, XL1-Blue/AP-Lys-B/M13-lysC-WT, XL1-Blue/AP-Lys-B/M13-lysC-V339A and XL1-Blue/AP-Lys-B/M13-lysC-R300C were harvested and washed twice with 50mM, ice-cooled PBS buffer. The fluorescence activities of the four different cell populations were measured using flow cytometry. As shown in 4.8 a and b, although slight differences could be observed, it is not possible to set up a gain setting to select the mutants. In other words, the different cell populations cannot be distinguished by the flow cytometry method. On the other hand, our method based on cell-phage interaction can

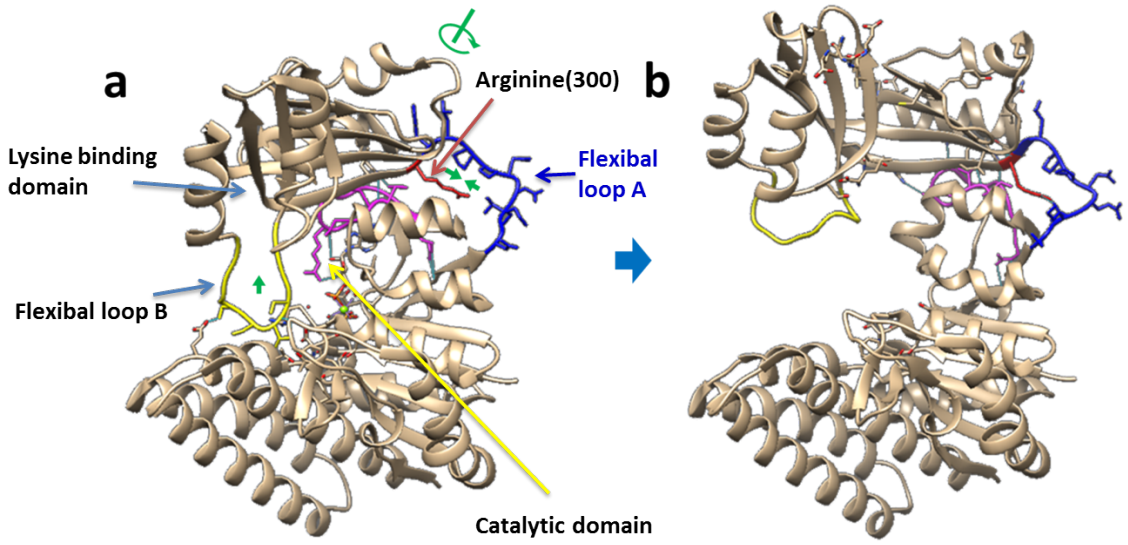


Fig. 4.7 3D structure illustration of the de-allosteric regulation mechanism of R300C mutein

a - active state of wild AK-III without lysine binding. *b* - inactive state of wild AK-III with lysine bound to the regulatory domain. The visualizations were generated using the UCSF Chimera software [227].

successfully screen out V339A as the best mutant of AK-III, confirming a higher sensitivity of the cell robot based screening method.

Biological systems are complex and highly adaptive, meaning that the cells always try to reduce the perturbations introduced. Introducing molecular variants into cells can be regarded as perturbations to the cells. As shown in 4.8c, after the introduction of molecular variants, the cells may undergo three stages of signal change: enhancement, attenuation and stabilization. The signal enhancement stage is the direct consequence of the perturbations induced by the introduced molecular variants. The signal attenuation stage is caused by the adaptive response of cells to the perturbation. Finally, the signal reaches a stable state which might be slightly different from the state before the perturbation. The time interval for these changes may be relatively short. The curves in 4.8c are theoretical response patterns of cells to the disturbance by over-expression of the different AK-III variants respectively. For the FACS-based method, the cells to be measured may have already reached the stable stage where the signal strength may not be significantly different in the cell populations with different variants. However, our method captures signals during the whole response and adaptation processes which correspond to the area below the curves and can be therefore more sensitive.

Furthermore, the “cell robots” based screening works in principle like an autocatalytic process of signal amplification: the target molecule with desired performance will increase the intracellular concentration of the signal molecule (lysine in this case) in the cell, the

increased concentration of the signal molecule will amplify the population of phage carrying the target molecule. The amplified phages can infect other cells to further enhance the signal. In such a way the screening process is highly effective and sensitive compared to the other presently used methods, such as those based on single cells using fluorescence as the readout signal [204].

Summary of cell robot based screening

For conclusion, it was demonstrated that the biological 'robots', i.e. the cells, can be engineered to perform screening tasks in protein engineering. By capturing the signals during the whole response and adaptation processes, which cannot be achieved by screening based on electric machines, the cell-phage based screening system has an inherent higher sensitivity. The current proof of concept study shows that cell-phage interaction system does not require any genetic modifications of the host cells to enhance the signal for screening. In a recent similar work which used FACS as the screening method, genetic modifications are required to enhance the signal [204]. Furthermore, FACS-based screening often suffers signal noise caused by single cell variations and signal detection under conditions of fast moving cells [206, 204, 203]. By equally accessing all cells, the cell-phage interaction system can avoid the problem of single cell variation in principle. Cells as biological 'robots' have a unique feature of reproducing themselves to generate a vast population exponentially and cheaply. Thus, the screening throughput can be simply expanded by using a larger population of cells with minimal additional costs, indicating a massively parallel screening manner beyond the current electric machines. As proved by the power of parallel computing in computational science [228], parallelization is a great solution for speeding up the process of parallel tasks. The sensitivity and throughput are key factors determining the success of a screening experiment. The cell-phage screening system shows clear advantages in both sensitivity and throughput. Furthermore, the cost of cell robots is almost zero compared to that of expensive electric machines/robots. It should be mentioned that, while electric machines can utilize various types of signals for screening, screening based on the cell robots uses a "biological signal and sensor", which might represent a limitation in some cases. However, many natural or purposefully designed biological elements or sensors such as promoters and riboswitches can be used for this purpose [229, 230] and the signal molecules can be intermediates of metabolic pathways.

4.4 Conclusion and Perspective

Engineering biological system is complex and kind of unpredictable. In this study, a multiple IO system was proposed to simplify the development process of biological engineering. The IO system was implemented based on M13 phages. For proof of concept, the method was demonstrated for target identification, evaluation of designs, evolution and selection of key enzymes for the lysine biosynthesis in *E. coli*. Correct prediction of beneficial genetic manipulation for enhanced lysine production in *E. coli* was achieved. New and effective variants of AK-III which is strictly inhibited by L-lysine, were obtained. Importantly, the cellular IO system showed an ultra-sensitivity in capturing the signal changes caused by the perturbations introduced. The author believes that the approach developed in this work opens up new possibilities for systems metabolic engineering and synthetic biology of industrial microorganisms in practical applications.

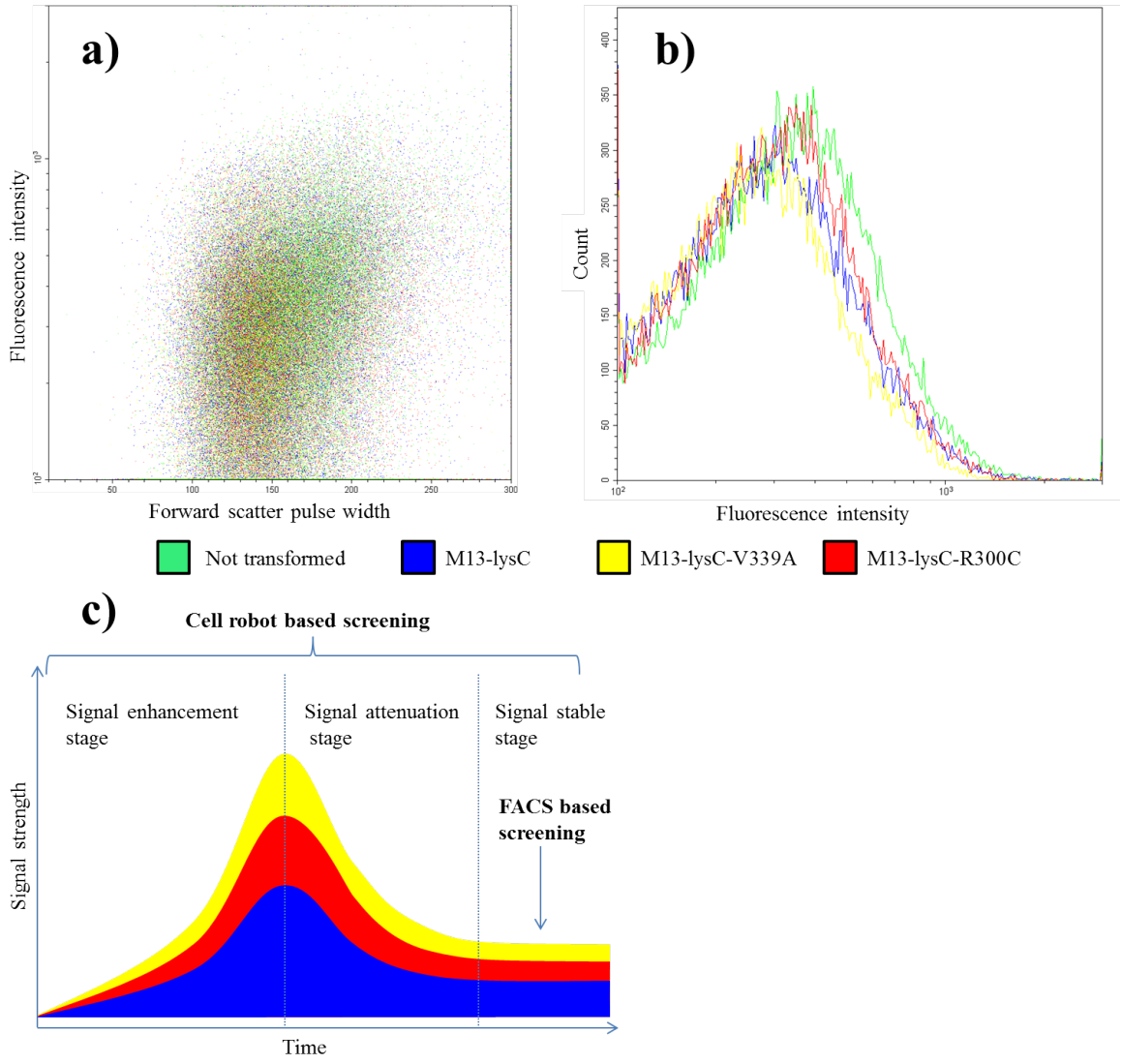


Fig. 4.8 Flow Cytometry assays of cell populations harboring wild-type AK-III and AK-III mutants of R300C and V339A

a) Flow Cytometry assays of different cell populations harboring the wild-type AK-III, the mutant R300C or the mutant V339A. By using the same lysine responsive promoter to control a GFP encoding gene intracellularly, the different cell populations cannot be distinguished by measuring green fluorescence intensity using flow cytometry. b) Illustration of differences in signal capture based on the cell-robots and that of the conventional FACS-based screening method. Introduction of molecular variants into cells can be regarded as perturbations to the cells. After introduction of variants, the cells may undergo three stages of signal change: enhancement, attenuation and stabilization. The signal enhancement stage is caused by the direct consequence of perturbations. The signal attenuation stage is caused by adaptive responds of cells to perturbations. Finally, the signal reaches a stable state which is slightly different from the state before the perturbation. The curves in figure b) show theoretical response patterns of cells to the disturbances by over-expression of the different AK-III variants independently. For the FACS based method, the cells to be measured already reach the stable stage while introducing the variants into cells by transformation resulting in slight differences of responding signals which is hard to be distinguished. The present method captures signals during the whole adaptation process, which correspond to the area below the curves and can be therefore more sensitive.

Chapter 5

Orthogonal information encoding in living cells

5.1 Introduction

Deoxyribonucleic acid (DNA) is the natural information carrier utilized in all living organisms on earth [231]. The first report about artificial information in DNA was published in 1996 by Davis *et al.* They encoded a binary graphic data into a synthetic DNA molecule using a simple bit-mapping mechanism. Later in 1999, Clelland *et al.* proposed a method for hiding messages in DNA microdots for data encryption purpose [232]. Bancroft *et al.* first proposed used DNA as a long-term information storage media [233]. In 2007, Yachie *et al.* proposed an alignment-based approach for durable data storage into living organisms [234]. In 2009, Gustafsson *et al.* encoded a poem into DNA [235]. Ailenberg *et al.* proposed an improved Huffman coding method for archiving text, images, and music characters in DNA [236]. In 2010, Gibson *et al.* wrote a watermark message into a chemically synthesized genome[237]. In addition to these applications, artificial information encoding in DNA has more attractive potential applications such as barcoding and comments encoding for programming cells in synthetic biology [27], and even for large and long-term data storage [232, 27, 22, 238–241]. Information stored in DNA can be distributed in a three dimensional space while the traditional planner media can only store information on a two dimensional surface. The extra dimension remarkably enhances the information density as recently demonstrated by Church *et al.*, Goldman *et al.* and Erlich *et al.* [23, 24, 242]. Furthermore, unlike the presently used storage media such as magnetic tape or hard drives which have a

This chapter was modified based on a recent publication: Song, Lifu; Zeng, An-Ping (2017): Orthogonal information encoding in living cells with high error-tolerance, safety, and fidelity. ACS Synthetic Biology 7 (3), pp. 866–874.

decisive shortcoming of a limited lifetime, e.g. around 50 years for hard drivers, information storage in DNA can have a lifetime of many thousands of years and is of low maintenance costs [23, 24, 243, 244, 233]. By using silica glass spheres for DNA encapsulation, Grass *et al.* predicted an error-free retrieval of information from the DNA after more than 1 million years if stored at -18°C and 2000 years at 10°C [245]. However, relatively high error rates could be introduced in the information "writing", "reading" and "copying" processes (i.e. DNA synthesis, sequencing and polymerase chain reaction (PCR)) [25, 26]. The error rate is even higher if the encoded DNA sequences contain extreme GC contents, long homopolymers or complex secondary structures which are hard to be synthesized, sequenced and amplified. Errors were detected in early studies lacking an effective error correction mechanism [23, 24]. In the later studies of Yim *et al.* and Grass *et al.*, error correction codes such as "Reed–Solomon" or "low-density parity-check" were introduced and information could be retrieved successfully [245, 246]. Yazdi *et al.* implemented a DNA-based storage architecture that enables random access to data blocks and rewriting of information stored at arbitrary locations within the blocks [247]. Later on, additional error correction codes were introduced and the scale of encoded data volume increased remarkably [244, 248, 249]. Interestingly, a recent study has successfully retrieved information stored in synthesized long DNA fragments using a portable sequencer - MinION [25, 250, 251]. More recently, by introducing Fountain codes, Erlich *et al.* established a robust and efficient storage strategy called DNA Fountain with a data density of 215 petabytes per gram of DNA [242].

All the studies mentioned above dealt with information encoding in DNA outside living cells. It is also of great interest to know if DNA data storage or information encoding in living cells is feasible and reliable. This should enable applications such as biological barcodes of engineered biological parts (Biobricks) and comment "language" in "programming biology" in the emerging area of synthetic biology [27]. Theoretically, the encoding schemes designed for *in vitro* data storage in DNA are also applicable for in vivo applications. However, to the best of our knowledge, no reported work has addressed the issue of error enrichment introduced by DNA replication which is crucial for *in vivo* applications since DNA replication happens constantly under *in vivo* conditions. In the study of Erlich *et al.*, they retrieved the original data from a deep copy of the data by PCR amplifying the oligo pool in a total of nine successive reactions [242]. Theoretically, the successive PCR reactions could generate $218{\times}10^{12}$ copies of the original data, showing a great advantage of the exponential "copying" process of DNA storage by PCR to generate vast data copies quickly and cheaply. However, the number of rounds of the "copying" process is only demonstrated within limited cycles which are insufficient for *in vivo* applications [242]. Furthermore, the artificial DNA fragments could interfere with the native and natural ones (being so-called biologically relevant).

This is another issue which has not been studied so far. For *in vivo* applications, such as biological barcodes or comments encoding in living cells, the encoded DNA sequences should not share the same sequence space as the natural ones to avoid interference with cellular functions. In other words, they should be orthogonal to exclude biological relevance.

One unique feature of information storage in DNA is that there are always many copies of DNA molecules synthesized which represent the same data. In other words, there is a high inherent data redundancy. In this study, using a novel way of adding error detection codes block by block, an efficient self-error-detecting, three-base block encoding scheme (SED3B) which can take full advantage of the inherent redundancy feature for error correction was established. SED3B can effectively repress error enrichment emerging from DNA replication as proved by *in silicon* and experimental verifications. With merely 30 sequences for error correction, the SED3B scheme can tolerate a high error rate of 19.1%. Errors in a rate of 40% still can be corrected with 180 DNA sequences as proved by in silicon simulations. Over 12,100 years of continuous replication are estimated to be required to make the SED3B encoded information in growing *E. coli* cells unrecoverable as proved by *in vivo*, error-prone PCR experiments. In addition to limited extreme GC contents, homopolymers, and simple secondary structure, SED3B encoded sequences also show very low biological relevance as proved by comparative studies with naturally formed sequences. Features of high error tolerance and low biological relevance make SED3B promising for orthogonal information encoding in living cells, *e.g.* as comment language in programming cell or for biological barcode encoding. To facilitate the usage of SED3B as a univeral information encoding scheme in living cells, an online encoding-decoding system with cases of comment and barcode encoding is implemented and released in http://biosystem.bt1.tu-harburg.de/sed3b/.

5.2 Theoretical and technological backgrounds

The focus of this part of work is to design an encoding scheme for reliable digital data encoding in DNA with regarding to the unique features of DNA as data storage media. There are many methods available for data encoding in DNA. In the previous section, the available methods for data encoding in DNA have been briefly introduced. In this section, four representative state-of-the-art methods released in recent years are detailed, their merits and limitations or disadvantages are mentioned. Other efforts on associative memory and DNA computing were not included because they were designed for different purposes for DNA information storage needs. In a recent study, Erlich *et al.* reported a storage strategy, called DNA Fountain. They proved that the 2.14×10^6 bytes encoded data could be retrieved

by 2.18×10^{15} times, indicating a highly robust system. However, this method was not detailed here since this strategy is not applicable for *in vivo* applications to the author.

5.2.1 The method of Church *et al.*

In the study of Church *et al.* in 2012, they used a "one bit per base" coding system with the base "A/C" for zero and "G/T" for one [23]. To avoid the formation of extreme GC, homopolymers and secondary structures in the encoded DNA sequences, they applied random disruption mechanism. They encoded an html version draft of a book that included 53,426 words, 11 JPG images, and one JavaScript program into a 5.27-megabit bitstream and all data blocks were recovered with merely 10 bit errors emerged. The errors identified after sequencing are mainly due to the lack of an error correction mechanism. Furthermore, this method sacrifices half of the storage capacity which in turn would double the costs.

5.2.2 The method of Goldman *et al.*

In the study of Goldman *et al.*, a base-3 encoding scheme was applied. Digital information was first converted to base-3 using a Huffman code that replaces each byte with five or six base-3 digits (trits) [24].This in turn was converted *in silico* to DNA code by replacement of each trit with one of the three nucleotides different from the previous one used as shown in Table 5.1. DNA homopolymers are abolished while sacrificing one fourth of the encoding capacity. However, this method cannot avoid extreme GC and complex secondary structure contents effectively. Furthermore, to make sure a full coverage of every fragment during sequencing, a fourfold redundancy was created by fragment overlapping which resulted with an efficiency of (3/4)/4=18.75% without considering the index and compress issues. Together a simple parity-check for single base error-detection with 1.2×10^5 copies of each DNA string, the information could be recovered without any errors (1.2×10^7 copies of each DNA string were actually used in Goldman's experiments and they supposed that 1% of them are enough for reliable information storage). However, such high coverage reduces the data density and raises the cost for information storage in DNA.

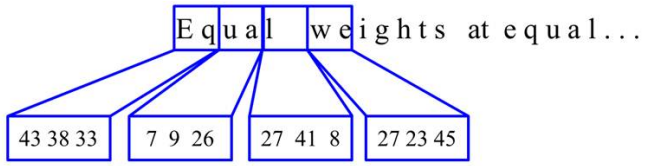
5.2.3 The method of Grass *et al.*

In 2015, Grass *et al.* reported an encoding strategy applying the Reed–Solomon (RS) coding to data storage in DNA [245]. First, two bytes of a digital file are mapped to three elements of the Galois Field of size 47 (GF(47)) by base conversion (256^2 to 47^3). Second, RS codes are employed to add redundancy A to the individual blocks. Finally, the data blocks were

Table 5.1 Base-3 to DNA encoding ensuring no repeated nucleotides in the Goldman's method

Previous base written	Next base to encode		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

converted into DNA by mapping every element of GF(47) to three nucleotides by utilizing the GF(47) to DNA codon wheel as shown in Figure 5.1, thereby guaranteeing that no base is repeated more than three times. By encapsulating the DNA in an inorganic matrix, they estimated a reliable information storage in DNA for 2000 years, which is far beyond the capabilities of transitional digital information storage media (<50 years).



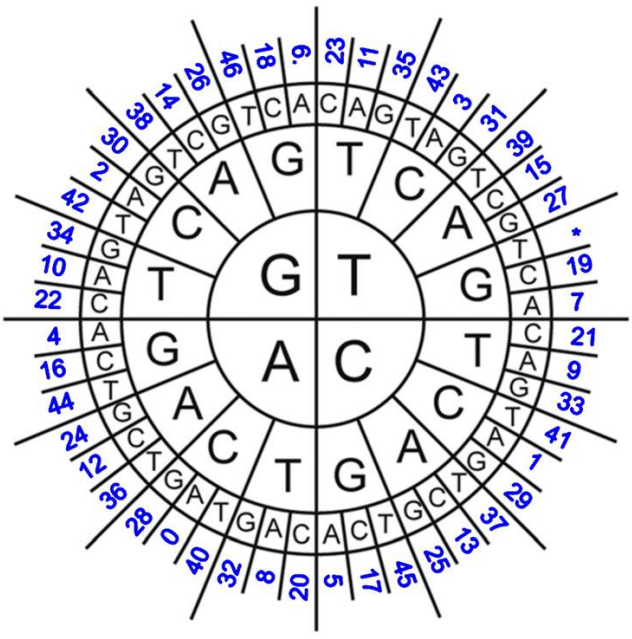


Fig. 5.1 GF(47) to DNA codon wheel for mapping every element of GF(47) to three nucleotides

Table 5.2 Comparison of capabilities of current available encoding schemes for digital information storage in DNA

	Church *et al.* 2012	Goldman *et al.* 2013	Grass *et al.* 2014	This study
Extreme GC	Yes	No	No	GC% <66.7%
Length of homopolymers	Up to 1	Up to 1	Up to 3	Up to 3G, 5A, 7T, 5C
Secondary structures	Yes	No	No	Yes
Error correction codes	No	Parity checking	RS codes	SEDTB
Encoding Efficiency [i]	50%	18.75%	63.37%	66.70%
Low Biological relevance	No	No	No	Yes

[i]The encoding efficiencies were calculated without consideration of index and compress issues.

5.3 Principles of a self-error-detecting, three-base block encoding scheme (SED3B)

To fully utilize the redundancy feature of DNA molecules for error correction, a novel self-error-detecting, three-base block (SED3B) encoding scheme was proposed for effective and flexible error correction. In details, binary bits are first transformed into data encoding DNA bases four bits by four bits using the scheme shown in Figure 5.2. Then one error checking base was inserted per two data encoding bases to form a three-base block encoding manner. The third base is designed to detect whether there are errors emerged in the two encoding bases. A simple way of error checking by the third base is the checksum principle [252]. However, the checksum method has no optimization option for homopolymers and extreme GC contents. Instead, a novel strategy was utilized to enable error checking by the third base. At first, all possible 16 two-base combinations were divided into four groups based on the principle that all the four two-base combinations in the same group do not share any identical base in neither the first nor the second base, and then every group was assigned with an error detecting base as shown in Figure 5.2. Thus, the data encoding two-base and the error detecting base won't match to each other anymore if error emerges in any of the three bases. In other words, a single base error on any of the three bases can be detected. To avoid extreme GC, long homopolymers and complex secondary structures generated in the encoding DNA strings, three additional principles are followed while assigning error checking bases to the four groups of two-base combinations: 1) no more than 3 G/C present in the three base block to avoid extreme GC contents; 2) no identical bases present in all the three bases to avoid long homopolymers; 3) no complementarily matched three base blocks present. However, principals 1) and 2) cannot be satisfied simultaneously. To address this issue, two different rules were introduced for assigning error detecting bases as shown in Figure 5.2, rows of "error detecting base rules". Rule I satisfies the principle that there are no more than 2 G/C present in all the three bases while a "TTT" homopolymer does present there. Rule II abolishes any three base homopolymer while enabling G/C presents in all three bases. During the encoding process, Rule I is used in general and only if "TTT" is present, the rule for assigning the error detecting bases is switched to Rule II temporarily and then switched back to Rule I after having encoded once. Thus, continuous "T" homopolymers can be avoided as the error detecting base for "TT" is switched to "G", not "T" in Rule II and extreme GC content can also be avoided as three G/C combinations in Rule II are only present if the previous encoding three-base block is "TTT". Finally, no more than seven continuous "T", five continuous "A/C" and three continuous "G" are possible to exist in the encoded DNA strings which have been proved to be acceptable by current DNA synthesis

and sequencing technologies [26]. The GC content can be controlled below 67.7%. Since two-thirds of the total bases are used for data encoding, the SED3B scheme has a theoretic encoding efficiency of 66.7% regardless the addressing and compress problems.

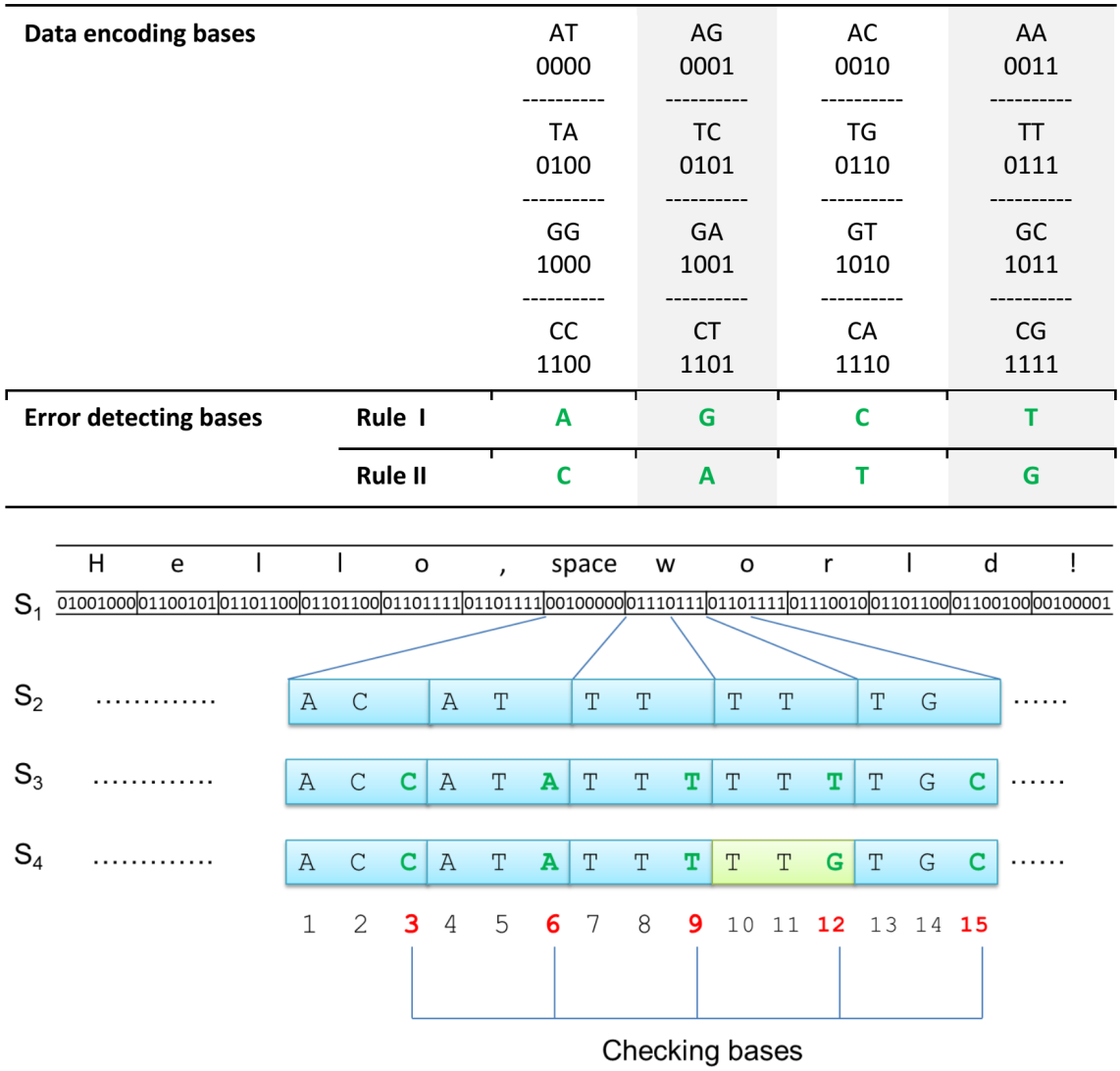


Fig. 5.2 Illustration of encoding binary data into DNA string using the SED3B encoding scheme.

5.4 High error tolerance revealed by *in silicon* simulations

To test the error detection capability, different rates of random errors were introduced into the SED3B encoded DNA fragments, and calculated the percentage of errors that could be detected by SED3B. As shown by the green triangles in Figure 5.3, more than 90% errors can

be detected while an error rate less than 10% and 78% errors still can be detected even when the error rate is as high as 30%. The error rates after error repression shown in red crisscross are more than one magnitude lower than the untreated ones. Next, the error correction capability was tested using variant numbers of DNA sequences. Simulations with 10 and 100 DNA sequences were performed individually at first. As shown in Figure 5.4, the error tolerance ability by using 100 DNA sequences is higher than that using 10 DNA fragments as expected. The error tolerance is up to 5% using 10 DNA fragments and up to 33% using 100 DNA fragments. To estimate the number of sequences required for reliable correction of a specific rate of errors, series of simulations with error rates ranging from 1% to 40% were performed, with a step increment of 1%. At each simulated error rate, the simulation started with a small number of sequences to retrial the data for 500 iterations. If errors have emerged in any of the 500 iterations, the sequence number was increased by one and the process was iterated until there are no errors emerged all 500 iterations. As shown in Figure 5.5, although the required sequence number increased exponentially with the increase of the error rate, 200 sequences were enough to correct a very high error rate of 40%.

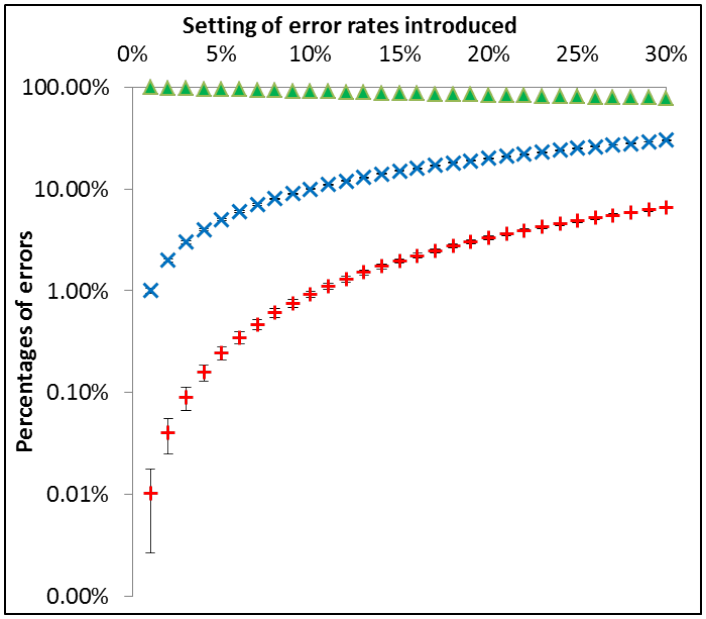


Fig. 5.3 Error detection and repression by using the SED3B encoding scheme.

▲Percentages of errors detected by SED3B method. + Remained percentages of errors in DNA fragments after removing the errors detected. × Percentages of random errors introduced during simulations. Errors were introduced into DNA fragments randomly base by base. A range of error rates from 1 to 30% was simulated with a stepping increment of 1%. Random errors were introduced in each step with a specific error rate setting, and each step was iterated for 500 times. More than 90% errors could be detected while the error rate less than 10%. More than 78% errors have been detected even the error rate is as high as 30%. The error rates after error repression shown in red crisscross are more than one magnitude lower than the untreated ones.

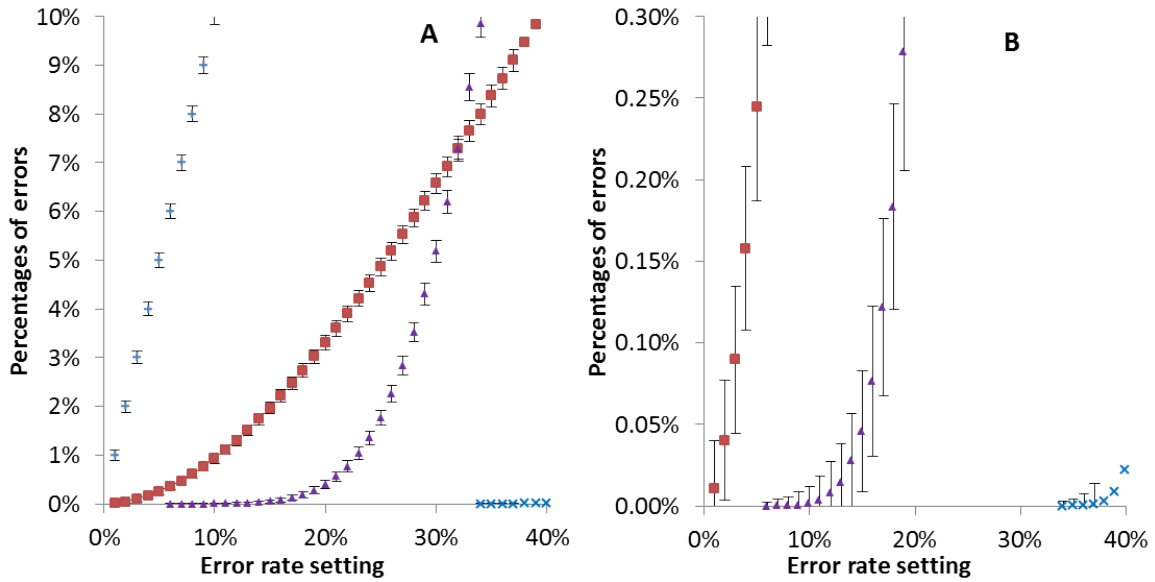


Fig. 5.4 Error correction capabilities by multiple DNA sequences encoded by SED3B encoding scheme.

+ Percentages of errors introduced during simulations. ■ Remained percentages of errors in DNA strings after removing the detected errors. ▲ The emerged percentages of errors in final recovered information using 10 error-containing DNA strings for retrieval of the information. ×The emerged percentages of errors in final recovered information using 100 DNA strings for information retrieval. Errors were introduced into DNA fragments randomly base by base. A range of error rates from 1% to 40% was simulated with a stepping increment of 1%. Random errors were introduced in each step with a specific error rate setting, and each step was iterated for 500 times.

5.5 SED3B encoded DNA sequences show low biological relevance

The biological safety issue has been widely discussed in synthetic biology [253, 254]. However, this problem didn't draw enough attention in previous studies of information encoding in DNA. Large data storage in DNA will produce a huge amount of novel DNA fragments. The encoding scheme should provide mechanisms to avoid that the encoded DNA fragments could be utilized by microbes in nature, especially for large data storage application.

Similar to the life coding system in nature, SED3B also uses a three-base block encoding manner. However, only one-fourth of the 64 possible three-base combinations are used in SED3B in general and another one-fourth is used only in cases that "TTT" is present in the previous encoding block. Such an encoding scheme imposes strong limitations on the encoded DNA string, making it hard to form "biologically meaningful" sequences. To prove this, a Perl script was implemented to search for sub-sequences that satisfy our encoding rules

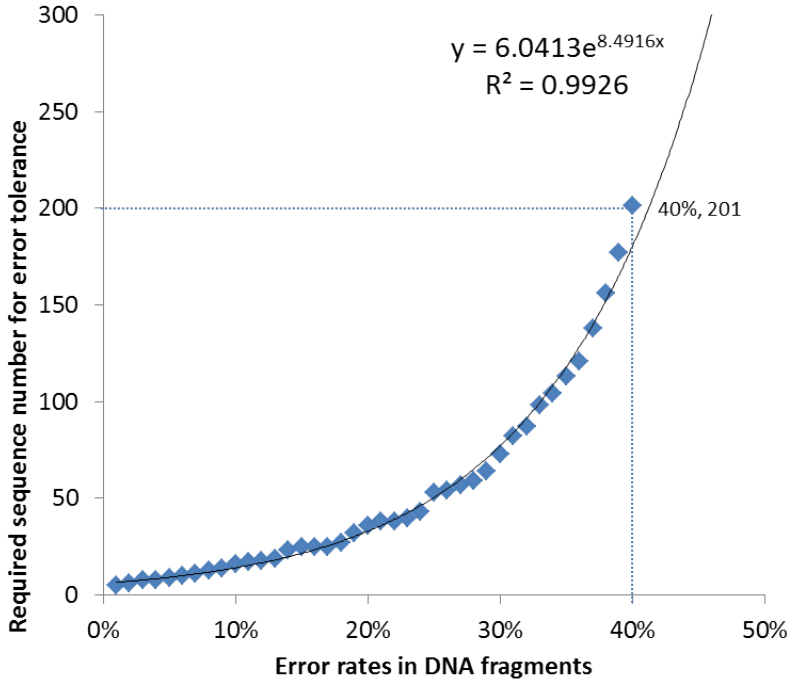


Fig. 5.5 Simulation of required sequence numbers for reliable information recovery by DNA fragments with variant rates of errors.

To estimate the number of sequences required for reliable correction of a specific rate of errors, series of simulations were performed with a range of error rates from 1% to 40%, with a step increment of 1%. At each simulated error rate, simulation started with a small number of sequences to retrial the data for 500 iterations. If errors emerged in any of the 500 iterations, the sequence number increaseed by one and the process was repeated until there is no errors emerged all 500 iterations.

in natural DNA sequences (the Perl script is detailed in Appendix A). All the 30,151,123 nucleotide sequences available in the NCBI nucleotide collection (nt) database (collected on May 28, 2015) were analyzed considering all three frames. The results showed that none of the entire coding sequences can fit our encoding rules and the number of matched partial sequences decreases rapidly along with the increase of the cut-off length as shown by the blue dots in Figure 5.6. Furthermore, large amounts of partial sequences are found to be tandem repeat structure containing sequences which have low biological meanings as shown by the red dots in Figure 5.6. Indeed, all partial sequences with a length longer than 65bp are found to be tandem repeats. These results imply that the SED3B encoded DNA sequences and naturally formed DNA sequences are located in different sequence spaces with slight space overlaps of tandem repeat sequences. In other words, SED3B encoded DNA strings show very low biological relevance.

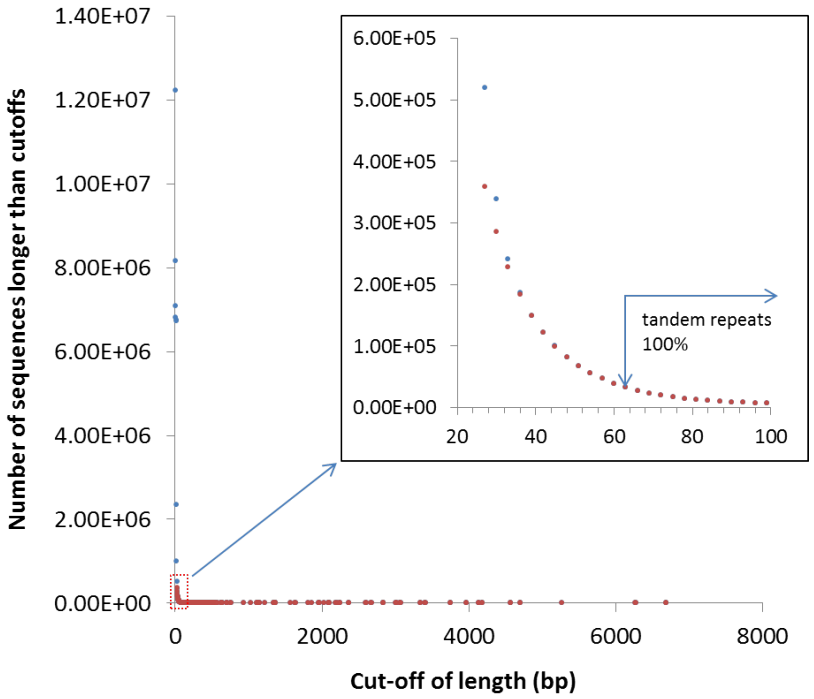


Fig. 5.6 Comparative analysis of SED3B encoded sequences with a natural DNA sequences.

All partial sequences longer than 66bp that satisfy the SED3B scheme are all tandem repeats. The 30,151,123 sequences in NCBI Nucleotide database collected on May 28, 2015 were used as inputs. All three frames were analyzed. The horizontal axis stands for length cut-off of partial sequences. The vertical axis stands for the total number of partial sequences that have a length equal or longer than the cut-offs. The small chart in the top-right is a zoom-in of the large chart. Blue dots stand for the total numbers of matched partial sequences equal or longer than a specific length. Red dots stand for the total numbers of matched partial sequences which are found to be tandem repeats.

5.6 SED3B encoded DNA sequences show simple secondary structure

Synthesis and sequencing of DNA fragments with complex secondary structures is a not-well-solved problem yet [255]. Retrieving information stored in DNA with complex secondary structures is a challenge which has been shown in previous studies [23, 24].

Complex secondary structures are formed by complementary subsequences. In SED3B, all the reverse complementary three bases combinations are abolished, which strongly prohibits the encoded DNA strings to form complex secondary structures in principle. To verify this, three files in different sizes were encoded into DNA strings by using the SED3B encoding scheme and without using the third base optimization individually. Since it's difficult to predict and compare the secondary structure complexity of DNA sequences

directly, the total number and percentage of complementarily matched k-mer pairs (CMKM) among all *k-mers* were used as an indicator of the complexity of secondary structures. Although the SED3B encoded DNA strings are 1.5 folds longer in length compared to the ones without the third base optimization, the total CMKMs are reduced by more than 80% using SED3B as shown in Figure 5.7. Furthermore, the percentage of reduced CMKMs increases while enlarging the data volume. In the case of File C with a size of 9,797 Kilobyte, CMKMs are reduced by 94%. This implies that the DNA strings encoded by SED3B show much simple secondary structures.

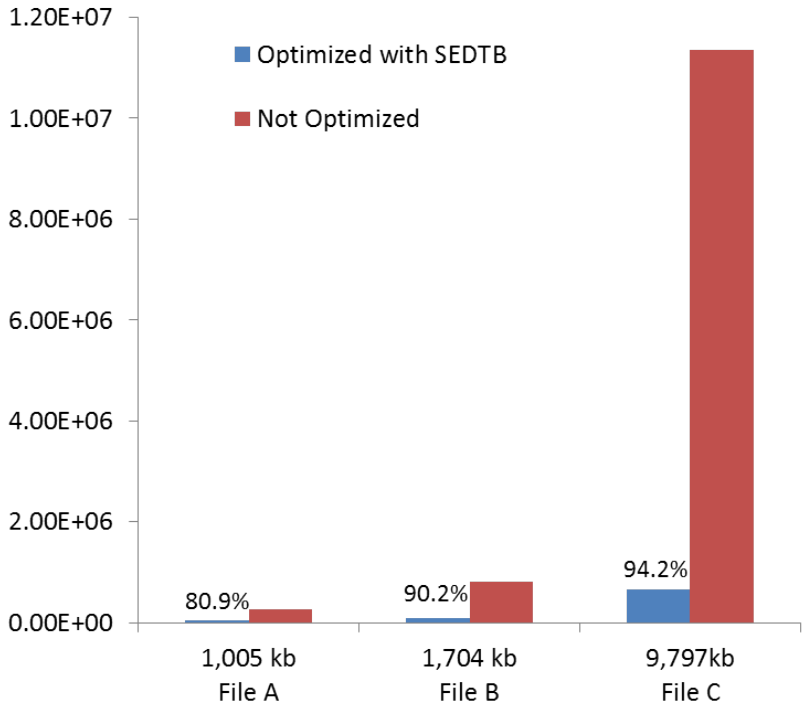


Fig. 5.7 The number of complementary matched k-mers is reduced remarkably by using the SED3B scheme.

The numbers above the blue bars stand for the percentages of reduced complementary matched *k*-mers by applying SED3B for secondary structure optimization corresponding to different input files.

5.7 Reliable orthogonal information encoding in living cells using SED3B

The features of effective error correction and low biological relevance make SED3B very promising for orthogonal information encoding in living cells. To test the reliability of the information written with SED3B and stored in living cells in practical, the digital information "Hello, World!" was first encoded in a plasmid. Since the replication error rate of *E. coli* cells is very low, error-prone PCR was employed to speed up the error enrichment process. The JBS dNTP-Mutagenesis Kit with a very high mutation rate of up to 20% was used to perform the error-prone PCR experiment. The error-prone PCR products were transformed into *E. coli* stellar competent cells. 14 individual colonies were picked for plasmid abstraction and sequencing. The sequencing results revealed that variant error rates ranging from 11% to 30% were introduced and the average error rate was 19.1%. The original information can be retrieved correctly from all the 14 sequences.

Fig. 5.8 Correct information can be retrieved using 14 sequences with high rates of errors introduced by error-prone PCR.

The three-base blocks with errors detected are replaced with "—". The average error rate is 19.1%.

Random errors could emerge and be enriched exponentially during replications of DNA. The final rate of errors is related to the fidelity of DNA replication and the number of replications. To destruct the stored information by DNA replication, the enriched errors

should be higher than the error rate that can be tolerated by the encoding scheme. Thus, we get the following unequal constraint:

$$E = 1 - (1 - P)^n > E_r \qquad \text{Eq. 5.1}$$

where E denotes the final rate of errors after n times of replication with a replication error probability of P per base and E_r is the rate of errors that can be tolerated. It has been reported that the DNA replication error rate of *E. coli* cells is as low as 10^{-9} to 10^{-11} per base pair [256]. Here, the highest error rate, i.e. 10^{-9}, was used to make a confident estimation. Although the simulation results show that SED3B can tolerance as high as 40% rate of errors, an error rate of 19.1% which has been proved in the error-prone PCR experiment in practical was used for calculation. Using these numbers, we obtain:

$$n > log(1 - E_r) \div log(1 - P) = log(1 - 0.191) \div log(1 - 10^{-9}) \approx 2.12E8$$

The doubling time of *E. coli* is around 0.5 to 1 hour. We use a doubling time of 0.5 hour for the following calculation. Thus, the minimal replication time $T_{\min}$ required to destruct the information is obtained from the following equation:

$$T_{min} = n \times T_d = 2.12E8 \times 0.5 hours \approx 12,100 \text{ years}$$

Thus, it would take more than 12,000 years' of replication time to make the information distorted, indicating a reliable information encoding in living cells.

To fascinate the utilization of SED3B as an information encoding system in living cells, an online encoding-decoding system for comment and barcode encoding-decoding has been implemented and released in http://biosystem.bt1.tu-harburg.de/sed3b/.

5.8 *In vitro* data storage using SED3B

A typical process of *in vitro* data storage in DNA is shown in Figure 5.9. For the information writing process, the digital information represented by a string of zero and one is first encoded into a DNA string. Due to the limitation in DNA synthesis, especially in high throughput DNA synthesis, the encoded DNA string should be fragmented and indexed in this step. The second step is to generate the realistic DNA fragments by high throughput DNA synthesis.

SED3B is also applicable for *in vitro* data storage in DNA in principle. Indeed, SED3B has some advantages in large data storage theoretically. We notice that by using merely five DNA sequences SED3B can correct an error rate of 5%. It has been reported that the error rate of high throughput DNA synthesis technology is around 0.5% currently [26]. Thus, five

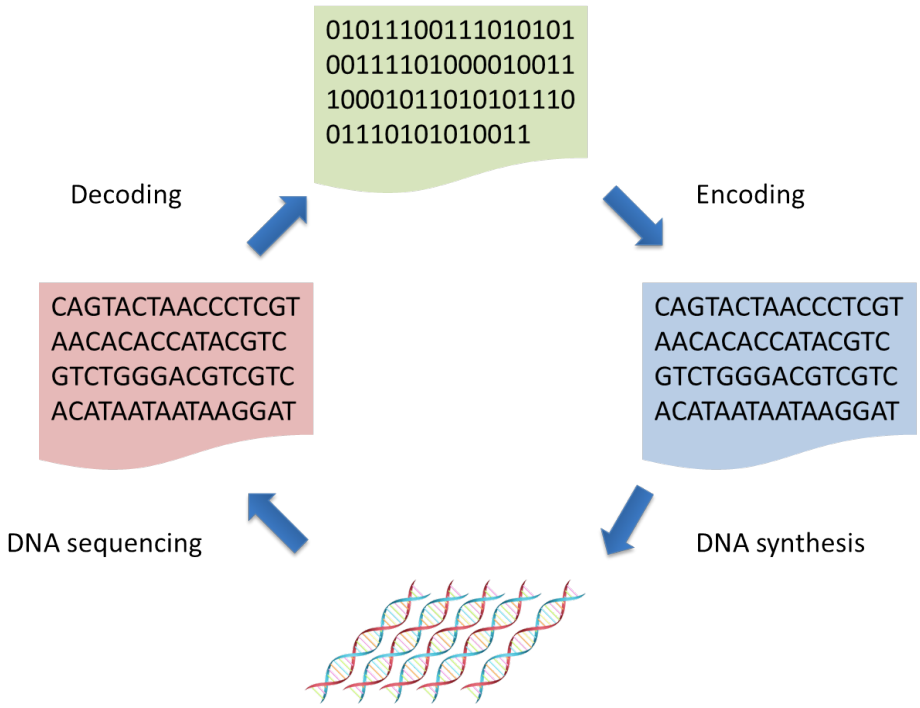


Fig. 5.9 Schematic presentation of *in vitro* information storage in DNA

sequences are enough for reliable information encoding with SED3B using the state of the art DNA synthesis technology theoretically. In Goldman's study, 1.2×10^5 copies of each DNA fragments were proposed for reliable data storage in DNA. Our simulation results show that this copy number could be reduced remarkably which in turn will greatly enhance the storage density using the SED3B encoding scheme. Even take one hundred copies instead of five for a conservative estimation, the storage density can still be increased by 2,400 times compared to the Goldman's method, resulting in a storage density around 4.7EB ($2.2\text{PB}\times2400\times8/9$) per gram DNA. Additionally, releasing huge amount of artificial DNA fragments into the environment might cause potential biological safety issues especially for large data storage. For example, the microbes in nature may employ the novel DNA fragments to generate diversity. It in turn may accelerate antibiotic resistance development of microbes, which is one of the most critical problems to human health at present [257]. Thus, the encoding scheme should provide mechanisms to avoid or reduce the formation of biologically relevant DNA sequences. With a unique feature of low biological relevance, SED3B shows potential in solving the biological safety issue of large data storage for the first time.

To investigate the potential of SEB3B for *in vitro* data storage applications, we encoded a famous picture showing the first release of the IMB new Ramac 305 super computer with the first disk drive storage into DNA fragments as shown in Figure 5.10a. The picture in size of 75kb was first translated into a DNA string in length of 459,630bp by SED3B. We then

fragmented it into 5,892 fragments in length of 78bp. For every 10 fragments, we inserted a Cyclical Redundancy Check (CRC) fragment which enable recovering the full information in case any one of the ten fragments is missing. 12bp SED3B encoded index and two 15bp PCR adaptors were also inserted as shown in Figure 5.10c. Finally, we obtained 6,483 fragments each in length of 120bp. We synthesised the DNA fragments using the service provided by Synbio Technologies LLC (Suite 101,Building C20 Biobay, 218 Xinghu Street, SIP, Suzhou, 215123 China).

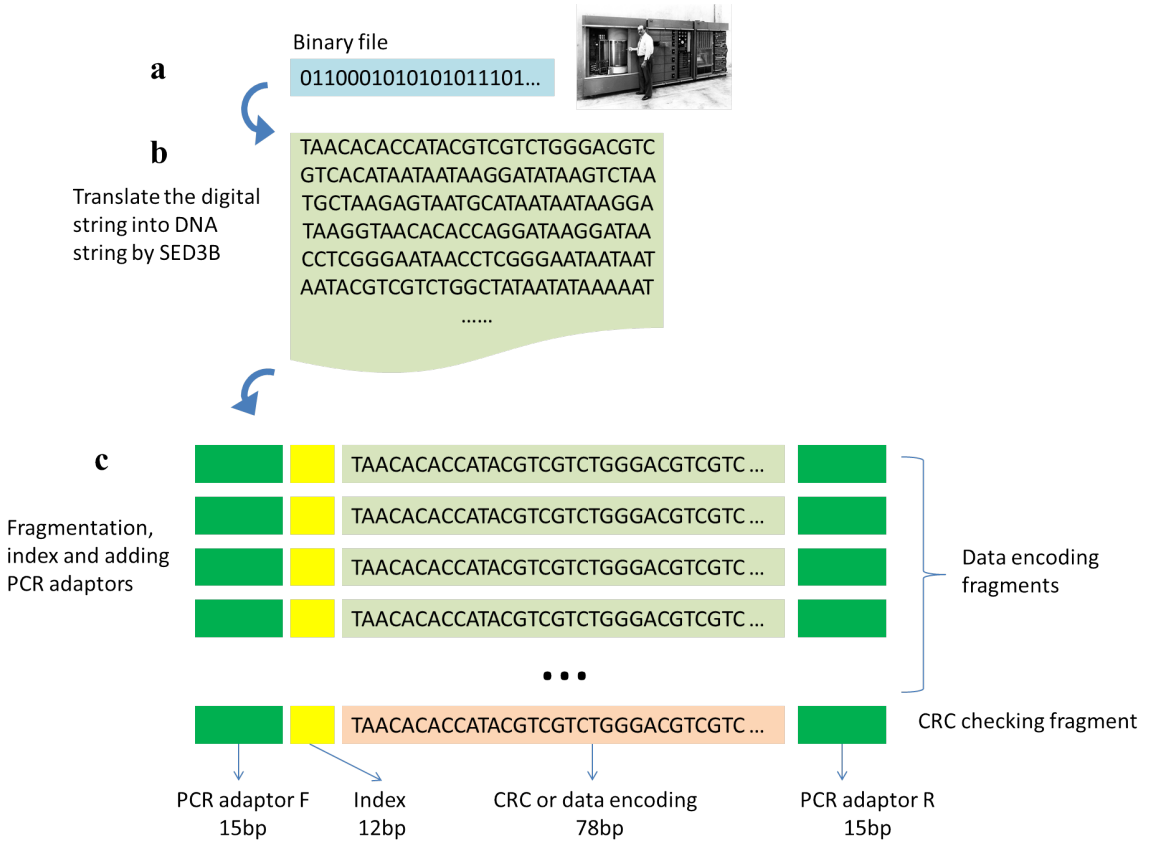


Fig. 5.10 Strategy of *in vitro* digital information encoding in DNA using SED3B

a, Digital data represented by a string of zero and one. b, SED3B encoded DNA string. c, Fragmented DNA string with CRC, index and PCR adaptors inserted. The final DNA sequences can be submitted for DNA synthesis to complete the data writing process

Although most of the previous studies using Illumina technologies as high throughput DNA sequencing resolution for information reading, we focus on a recently released portable sequencing device - MinION Nanopore sequencer [251]. After setting up the device (MinION), 1,191 pores were detected which are excellent according to the documentation provided by the company. We followed the standard library preparation and sequencing protocol. However, the quality of the sequencing results was found to be not good enough to recover

the information (data not shown). A possible reason is the nanopore sequencing protocol is designed for sequencing of long fragments (>500bp). In our study, 120bp fragments were used which may lead to a low quality of the sequencing results. In a more recent study by Yazdi *et al.*, 1kb DNA fragments were used to encode data and the data can be read correctly using MinION [250]. Since high throughput DNA synthesis is limited to short DNA oligos currently, our results indicate that further efforts are required to apply nanopore sequencing technology to fetch the data encoded in short DNA oligos.

5.9 Development of an online encoding-decoding system

To facilitate the utilization of SED3B as an information encoding system in living cells, an online encoding-decoding system for comment and barcode encoding-decoding has been implemented and released (http://biosystem.bt1.tu-harburg.de/sed3b/) as shown in Figure 5.11.

5.10 Conclusion

Reliable information storage *in vivo* in error rich DNA molecules is still represents a challenge since more and more errors could be introduced and enriched exponentially by rounds of replications. In this study, we presented a novel encoding scheme named SED3B, which can take full advantage of the inherent redundancy of DNA molecules for error correction. By using a small number of DNA molecules for error correction, SED3B can effectively correct the exponentially enriching errors during DNA replications as proved by *in silicon* simulation and error-prone PCR experiments for the first time. Based on error-prone PCR*in vivo* experiments with *E. coli* cells, more than 12,000 years of continuous replications are estimated to be required to make the SED3B encoded information unrecoverable in growing *E. coli* cells. Furthermore, for the first time we showed that SED3B encoded DNA sequences have little biological relevance to known natural DNA sequences, indicating its excellent orthogonality. Synthetic biologists are trying to design biological devices and algorithms to programm cells for various functions. Similar to the situation in programming of computers and machines, we need to write information such as comments or barcodes in the synthetic molecular programs. In these cases, SED3B is well suitable for reliable information encoding with no or low affections to the biological functions.

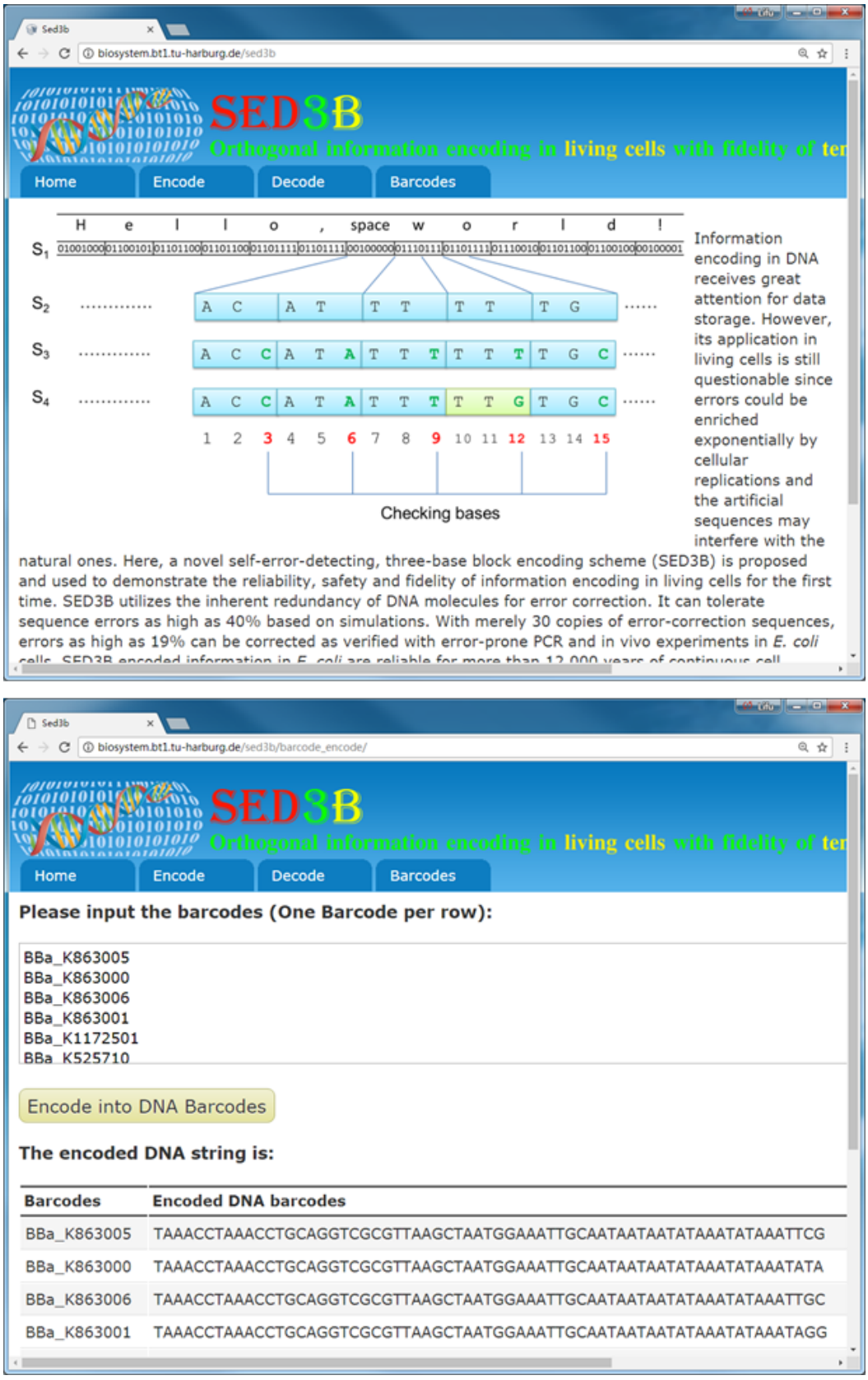


Fig. 5.11 Screenshots of the online encoding-decoding system

Chapter 6

Summary and outlook

Systems biology studies have been performed on ten strains of mutans streptococci for the sake of identification of the targets of carolacton and better understanding of the pathogenicity of mutants streptococci strains. Eight *S. mutans* strains, one *S. ratti* strains and one *S. sobrinus* strain were genome sequenced. Mathematic modeling of the *S.mutans* pan-genome displayed a possible opened pan-genome with high lateral gene transfer rate than other established pan-genome models of *Streptococcus* species. Distribution of factors which are thought to be highly related to pathology was systematically studied. Most of the studied systems show high diversities across different species except oxidative resistance system. The whole genome level metabolic networks were constructed and cross-compared. Diversities in central metabolic pathways were observed and a novel energy production pathway uniquely presented in *S. sobrinus* DSM20742 was revealed for the first time with the involvement of two novel lactate oxidases. Possible opened pan-genome, high lateral transferred genes rate, high diversities in resistance related systems and metabolic pathways – the pathogenicity of mutans streptococci should be re-evaluated. The strain-specific information provided in this study is helpful to understand the evolution and adaptive mechanisms of mutans streptococci and in turn will be very helpful for researchers to better understand those oral pathogens. An online regulation database for *S. mutans*, named StrepReg, was constructed by integrating time-resolved transcription factor based regulation network, STRING interaction database and KEGG pathway database (http://biosystem.bt1.tu-harburg.de:1555/homes/).

Due to the inherent complexity of the biological systems, biological engineering is unpredictable and cannot be performed in a rational way like other engineering disciplines such as electronic engineering. The biological engineering process have to go through many design-build-test cycles, within which many designs/variations have to be evaluated to generate some targets with desired properties. In other words, the biological engineering is suffered from the time- and labor- intensive 'debugging' process. Inspired by the debugging

process of programming on a computer, a debugging system is proposed to speed up the debugging process of biological engineering. To this end, we implemented a multiple IO system based on M13 phages which can be used as a debugging system for biological engineering in *E. coli*. Our proof of concept studies show that the current IO system can be utilized for applications of parallel evaluation of designs, beneficial genetic operation prediction and screening. The IO system shows higher sensitivity compared to the state of the art method of flow cytometry. Although further experiments are required, the IO system does show great potential in scaling up the input-output numbers. Scaling up the inputs to genome-level and combining the electric machine based robots may open up a new way to study the complex interactions of various intracellular components. Consequently, the massive information obtained could be helpful for mathematic modeling studies in systems biology.

In addition, a novel encoding scheme SED3B which can take full advantage of the inherent redundancy of DNA molecules for error correction was implemented. SED3B can effectively correct the exponentially enriching errors during DNA replications by using a small population of DNA molecules as proved by *in silicon* simulation and error-prone PCR experiments. Based on error-prone PCR *in vivo* experiments with *E. coli* cells, more than 12,000 years of continuous replications are estimated to be required to make the SED3B encoded information unrecoverable in growing *E. coli* cells. Furthermore, for the first time we showed that SED3B encoded DNA sequences have little biological relevance to known natural DNA sequences. Although further investigations are definitely required, the SED3B encoding scheme is also applicable for *in vitro* data storage in DNA in principle. Indeed, SED3B does show advantages in large data storage theoretically. We noticed that by using merely five DNA sequences SED3B can correct an error rate of 5%. It has been reported that the error rate of high throughput DNA synthesis technology is around 0.5% currently [26]. Thus, five sequences are enough for reliable information encoding by the state of the art DNA synthesis technology using SED3B theoretically. In Goldman's study, 1.2×10^5 copies of each DNA fragments were proposed for reliable data storage in DNA. Our results show that this copy number could be reduced remarkably which in turn will greatly enhance the storage density. Even take one hundred instead of five for reliable estimation, the storage density still can be increased by 2,400 times compared to the Goldman's method, resulting in a storage density around 4.7EB (2.2PB$\times$2400$\times$8/9) per gram DNA considering the encoding efficiency. Additionally, releasing huge amounts of artificial DNA fragments into the environment might cause potential biological safety issues especially for large data storage. For example, the microbes in nature may employ the novel DNA fragments to generate diversity. It in turn may accelerate antibiotic resistance development of microbes, which is one of the most critical

problems to human health at present [257]. Thus, the encoding scheme should provide mechanisms to avoid or reduce the formation of biologically relevant DNA sequences. With a unique feature of low biological relevance, SED3B shows potential in large data storage concerning about the biological safety issue for the first time.

References

[1] W. J. Loesche, "Role of *Streptococcus mutans* in human dental decay," *Microbiological Reviews*, vol. 50, no. 4, pp. 353–380, 1986.

[2] P. Stoodley, K. Sauer, D. G. Davies, and J. W. Costerton, "Biofilms as complex differentiated communities," *Annual Review of Microbiology*, vol. 56, pp. 187–209, 2002.

[3] F. J. Bruggeman and H. V. Westerhoff, "The nature of systems biology," *Trends in Microbiology*, vol. 15, no. 1, pp. 45–50, 2007.

[4] C. J. Cain, D. A. Conte, M. E. Garcia-Ojeda, L. G. Daglio, L. Johnson, E. H. Lau, J. O. Manilay, J. B. Phillips, N. S. Rogers, N. S. Stolberg, H. F. Swift, and M. N. Dawson, "Integrative biology: What systems biology is (not, yet)," *Science*, vol. 320, no. 5879, pp. 1013a–1014a, 2008.

[5] H.-Y. Chuang, M. Hofree, and T. Ideker, "A decade of systems biology," *Annual Review of Cell and Developmental Biology*, vol. 26, pp. 721–744, 2010.

[6] A. Chiappino-Pepe, V. Pandey, M. Ataman, and V. Hatzimanikatis, "Integration of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses," *Current Opinion in Systems Biology*, vol. 2, pp. 58–65, 2017.

[7] K.-H. Cho, S. Lee, D. Kim, D. Shin, J. I. Joo, and S.-M. Park, "Cancer reversion, a renewed challenge in systems biology," *Current Opinion in Systems Biology*, vol. 2, pp. 48–57, 2017.

[8] A. Dominguez, E. Munoz, M. C. Lopez, M. Cordero, J. P. Martinez, and M. Vinas, "Transcriptomics as a tool to discover new antibacterial targets," *Biotechnology Letters*, 2017.

[9] S. Durmus, T. Cakir, A. Ozgur, and R. Guthke, "A review on computational systems biology of pathogen-host interactions," *Frontiers in Microbiology*, vol. 6, p. 235, 2015.

[10] J. Geng and J. Nielsen, "*In silico* analysis of human metabolism: Reconstruction, contextualization and application of genome-scale models," *Current Opinion in Systems Biology*, vol. 2, pp. 28–37, 2017.

[11] W. J. Kim, H. U. Kim, and S. Y. Lee, "Current state and applications of microbial genome-scale metabolic models," *Current Opinion in Systems Biology*, vol. 2, pp. 9–17, 2017.

[12] S. Magi, K. Iwamoto, and M. Okada-Hatakeyama, "Current status of mathematical modeling of cancer – from the viewpoint of cancer hallmarks," *Current Opinion in Systems Biology*, vol. 2, pp. 38–47, 2017.

[13] J. Nielsen, "Systems biology of metabolism: A driver for developing personalized and precision medicine," *Cell Metabolism*, vol. 25, no. 3, pp. 572–579, 2017.

[14] E. J. Strobel, K. E. Watters, D. Loughrey, and J. B. Lucks, "Rna systems biology: uniting functional discoveries and structural tools to understand global roles of rnas," *Current Opinion in Biotechnology*, vol. 39, pp. 182–191, 2016.

[15] L. S. Yilmaz and A. J. Walhout, "Metabolic network modeling with model organisms," *Current Opinion in Chemical Biology*, vol. 36, pp. 32–39, 2017.

[16] J. W. Lee, T. Y. Kim, Y.-S. Jang, S. Choi, and S. Y. Lee, "Systems metabolic engineering for chemicals and materials," *Trends in Biotechnology*, vol. 29, no. 8, pp. 370–378, 2011.

[17] S. Y. Lee and H. U. Kim, "Systems strategies for developing industrial microbial strains," *Nature Biotechnology*, vol. 33, no. 10, pp. 1061–1072, 2015.

[18] M. Gustavsson and S. Y. Lee, "Prospects of microbial cell factories developed through systems metabolic engineering," *Microbial Biotechnology*, vol. 9, no. 5, pp. 610–617, 2016.

[19] A. S. L. Hansen, R. M. Lennen, N. Sonnenschein, and M. J. Herrgard, "Systems biology solutions for biochemical production challenges," *Current Opinion in Biotechnology*, vol. 45, pp. 85–91, 2017.

[20] H. M. Purdy and J. L. Reed, "Evaluating the capabilities of microbial chemical production using genome-scale metabolic models," *Current Opinion in Systems Biology*, vol. 2, pp. 90–96, 2017.

[21] Y. Vervoort, A. G. Linares, M. Roncoroni, C. Liu, J. Steensels, and K. J. Verstrepen, "High-throughput system-wide engineering and screening for microbial biotechnology," *Current Opinion in Biotechnology*, vol. 46, pp. 120–125, 2017.

[22] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, and W. L. Hughes, "Nucleic acid memory," *Nature Materials*, vol. 15, no. 4, pp. 366–370, 2016.

[23] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in dna," *Science (New York, N.Y.)*, vol. 337, no. 6102, p. 1628, 2012.

[24] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized dna," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[25] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the minion nanopore sequencer," *Nature Methods*, vol. 12, no. 4, pp. 351–356, 2015.

[26] S. Kosuri and G. M. Church, "Large-scale de novo dna synthesis: technologies and applications," *Nature Methods*, vol. 11, no. 5, pp. 499–507, 2014.

[27] "http://syntheticbiology.org/vectors/barcode.html."

[28] D. Ajdic, W. M. McShan, R. E. McLaughlin, G. Savic, J. Chang, M. B. Carson, C. Primeaux, R. Tian, S. Kenton, H. Jia, S. Lin, Y. Qian, S. Li, H. Zhu, F. Najar, H. Lai, J. White, B. A. Roe, and J. J. Ferretti, "Genome sequence of *Streptococcus mutans* ua159, a cariogenic dental pathogen," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14434–14439, 2002.

[29] F. Maruyama, M. Kobata, K. Kurokawa, K. Nishida, A. Sakurai, K. Nakano, R. Nomura, S. Kawabata, T. Ooshima, K. Nakai, M. Hattori, S. Hamada, and I. Nakagawa, "Comparative genomic analyses of *Streptococcus mutans* provide insights into chromosomal shuffling and species-specific content," *BMC Genomics*, vol. 10, p. 358, 2009.

[30] P. S. G. Chain, D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D.

Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter, "Genomics. genome project standards in a new era of sequencing," *Science (New York, N.Y.)*, vol. 326, no. 5950, pp. 236–237, 2009.

[31] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang, "Soap2: an improved ultrafast tool for short read alignment," *Bioinformatics (Oxford, England)*, vol. 25, no. 15, pp. 1966–1967, 2009.

[32] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform," *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754–1760, 2009.

[33] M. de La Bastide and W. R. McCombie, "Assembling genomic dna sequences with phrap," *Current Protocols in Bioinformatics*, vol. Chapter 11, p. Unit11.4, 2007.

[34] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: multiple alignment of conserved genomic sequence with rearrangements," *Genome Research*, vol. 14, no. 7, pp. 1394–1403, 2004.

[35] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Current Opinion in Microbiology*, vol. 11, no. 5, pp. 472–477, 2008.

[36] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu, "Pgap: pan-genomes analysis pipeline," *Bioinformatics (Oxford, England)*, vol. 28, no. 3, pp. 416–418, 2012.

[37] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[38] L. Li, C. J. Stoeckert, JR, and D. S. Roos, "Orthomcl: identification of ortholog groups for eukaryotic genomes," *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[39] J. L. Lavin, K. Kiil, O. Resano, D. W. Ussery, and J. A. Oguiza, "Comparative genomic analysis of two-component regulatory proteins in *Pseudomonas syringae*," *BMC Genomics*, vol. 8, p. 397, 2007.

[40] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The pfam protein families database," *Nucleic Acids Research*, vol. 38, no. Database issue, pp. D211–22, 2010.

[41] S. R. Eddy, "Profile hidden markov models," *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998.

[42] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi, "Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 40–79, 2010.

[43] H. Ma and A.-P. Zeng, "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms," *Bioinformatics (Oxford, England)*, vol. 19, no. 2, pp. 270–277, 2003.

[44] M. Stelzer, J. Sun, T. Kamphans, S. P. Fekete, and A.-P. Zeng, "An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks," *Integrative Biology : quantitative biosciences from nano to macro*, vol. 3, no. 11, pp. 1071–1086, 2011.

[45] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: software for visualization and analysis of biological networks," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 696, pp. 291–303, 2011.

[46] P. C. Y. Lau, C. K. Sung, J. H. Lee, D. A. Morrison, and D. G. Cvitkovitch, "Pcr ligation mutagenesis in transformable streptococci: application and efficiency," *Journal of microbiological methods*, vol. 49, no. 2, pp. 193–205, 2002.

[47] M. Reck, K. Rutz, B. Kunze, J. Tomasch, S. K. Surapaneni, S. Schulz, and I. Wagner-Döbler, "The biofilm inhibitor carolacton disturbs membrane integrity and cell division of streptococcus mutans through the serine/threonine protein kinase pknb," *Journal of bacteriology*, vol. 193, no. 20, pp. 5692–5706, 2011.

[48] J. Lefrançois, M. M. Samrakandi, and A. M. Sicard, "Electrotransformation and natural transformation of streptococcus pneumoniae: requirement of dna processing for recombination," *Microbiology (Reading, England)*, vol. 144 (Pt 11), pp. 3061–3068, 1998.

[49] O. Ween, S. Teigen, P. Gaustad, M. Kilian, and L. S. Havarstein, "Competence without a competence pheromone in a natural isolate of streptococcus infantis," *Journal of bacteriology*, vol. 184, no. 13, pp. 3426–3432, 2002.

[50] Y. H. Li, P. C. Lau, J. H. Lee, R. P. Ellen, and D. G. Cvitkovitch, "Natural genetic transformation of streptococcus mutans growing in biofilms," *Journal of bacteriology*, vol. 183, no. 3, pp. 897–908, 2001.

[51] D. J. LeBlanc, Y.-Y. Chen, N. D. Buckley, and L. N. Lee, "Genetic transfer methods for streptococcus sobrinus and other oral streptococci," *Methods in Cell Science*, vol. 20, no. 1/4, pp. 85–93, 1998.

[52] M. G. Caparon and J. R. Scott, "Genetic manipulation of pathogenic streptococci," *Methods in enzymology*, vol. 204, pp. 556–586, 1991.

[53] R. E. McLaughlin and J. J. Ferretti, "Electrotransformation of streptococci," *Methods in molecular biology (Clifton, N.J.)*, vol. 47, pp. 185–193, 1995.

[54] A. H. Badran and D. R. Liu, "Development of potent in vivo mutagenesis plasmids with broad mutational spectra," *Nature Communications*, vol. 6, p. 8425, 2015.

[55] S. BLACK and N. G. WRIGHT, "beta-aspartokinase and beta-aspartyl phosphate," *The Journal of Biological Chemistry*, vol. 213, no. 1, pp. 27–38, 1955.

[56] J. Täpp, M. Thollesson, and B. Herrmann, "Phylogenetic relationships and genotyping of the genus streptococcus by sequence determination of the rnase p rna gene, rnpb," *International journal of systematic and evolutionary microbiology*, vol. 53, no. Pt 6, pp. 1861–1871, 2003.

[57] J. A. Lemos and R. A. Burne, "A model of efficiency: stress tolerance by *Streptococcus mutans*," *Microbiology (Reading, England)*, vol. 154, no. Pt 11, pp. 3247–3255, 2008.

[58] K. Nakano, R. Nomura, M. Matsumoto, and T. Ooshima, "Roles of oral bacteria in cardiovascular diseases–from molecular mechanisms to clinical cases: Cell-surface structures of novel serotype k *Streptococcus mutans* strains and their correlation to virulence," *Journal of Pharmacological Sciences*, vol. 113, no. 2, pp. 120–125, 2010.

[59] R. Nomura, K. Nakano, N. Taniguchi, J. Lapirattanakul, H. Nemoto, L. Gronroos, S. Alaluusua, and T. Ooshima, "Molecular and clinical analyses of the gene encoding the collagen-binding adhesin of *Streptococcus mutans*," *Journal of Medical Microbiology*, vol. 58, no. Pt 4, pp. 469–475, 2009.

[60] R. J. Redfield, W. A. Findlay, J. Bosse, J. S. Kroll, A. D. S. Cameron, and J. H. Nash, "Evolution of competence and dna uptake specificity in the *Pasteurellaceae*," *BMC Evolutionary Biology*, vol. 6, p. 82, 2006.

[61] G. D. Ehrlich, F. Z. Hu, K. Shen, P. Stoodley, and J. C. Post, "Bacterial plurality as a general mechanism driving persistence in chronic infections," *Clinical Orthopaedics and Related Research*, no. 437, pp. 20–24, 2005.

[62] O. E. Cornejo, T. Lefébure, P. D. P. Bitar, P. Lang, V. P. Richards, K. Eilertson, T. Do, D. Beighton, L. Zeng, S.-J. Ahn, R. A. Burne, A. Siepel, C. D. Bustamante, and M. J. Stanhope, "Evolutionary and population genomics of the cavity causing bacteria streptococcus mutans," *Molecular biology and evolution*, vol. 30, no. 4, pp. 881–893, 2013.

[63] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal w and clustal x version 2.0," *Bioinformatics (Oxford, England)*, vol. 23, no. 21, pp. 2947–2948, 2007.

[64] J. D. Retief, "Phylogenetic analysis using phylip," *Methods in molecular biology (Clifton, N.J.)*, vol. 132, pp. 243–258, 2000.

[65] A. E. Darling, I. Miklós, and M. A. Ragan, "Dynamics of genome rearrangement in bacterial populations," *PLoS genetics*, vol. 4, no. 7, p. e1000128, 2008.

[66] J. C. Waterhouse, D. C. Swan, and R. R. B. Russell, "Comparative genome hybridization of streptococcus mutans strains," *Oral microbiology and immunology*, vol. 22, no. 2, pp. 103–110, 2007.

[67] C. Wu, R. Cichewicz, Y. Li, J. Liu, B. Roe, J. Ferretti, J. Merritt, and F. Qi, "Genomic island tnsmu2 of *Streptococcus mutans* harbors a nonribosomal peptide synthetase-polyketide synthase gene cluster responsible for the biosynthesis of pigments involved in oxygen and h2o2 tolerance," *Applied and Environmental Microbiology*, vol. 76, no. 17, pp. 5815–5826, 2010.

[68] J. C. Waterhouse and R. R. B. Russell, "Dispensable genes and foreign dna in streptococcus mutans," *Microbiology (Reading, England)*, vol. 152, no. Pt 6, pp. 1777–1788, 2006.

[69] A. E. Darling, B. Mau, and N. T. Perna, "progressivemauve: multiple genome alignment with gene gain, loss and rearrangement," *PloS one*, vol. 5, no. 6, p. e11147, 2010.

[70] A. Muzzi and C. Donati, "Population genetics and evolution of the pan-genome of streptococcus pneumoniae," *International journal of medical microbiology : IJMM*, vol. 301, no. 8, pp. 619–622, 2011.

[71] A. Mira, A. B. Martín-Cuadrado, G. D'Auria, and F. Rodríguez-Valera, "The bacterial pan-genome:a new paradigm in microbiology," *International microbiology : the official journal of the Spanish Society for Microbiology*, vol. 13, no. 2, pp. 45–57, 2010.

[72] C. Donati, N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V. Angiuoli, M. Oggioni, J. C. Dunning Hotopp, F. Z. Hu, D. R. Riley, A. Covacci, T. J. Mitchell, S. D. Bentley, M. Kilian, G. D. Ehrlich, R. Rappuoli, E. R. Moxon, and V. Masignani, "Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species," *Genome Biology*, vol. 11, no. 10, p. R107, 2010.

[73] T. Lefebure and M. J. Stanhope, "Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition," *Genome Biology*, vol. 8, no. 5, p. R71, 2007.

[74] J. S. Hogg, F. Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J. C. Post, and G. D. Ehrlich, "Characterization and modeling of the haemophilus influenzae core and supragenomes based on the complete genomic sequences of rd and 12 clinical nontypeable strains," *Genome Biology*, vol. 8, no. 6, p. R103, 2007.

[75] A. M. Stock, V. L. Robinson, and P. N. Goudreau, "Two-component signal transduction," *Annual Review of Biochemistry*, vol. 69, pp. 183–215, 2000.

[76] I. Biswas, L. Drake, D. Erkina, and S. Biswas, "Involvement of sensor kinases in the stress tolerance response of *Streptococcus mutans*," *Journal of Bacteriology*, vol. 190, no. 1, pp. 68–77, 2008.

[77] Y. Liu and R. A. Burne, "Multiple two-component systems of *Streptococcus mutans* regulate agmatine deiminase gene expression and stress tolerance," *Journal of Bacteriology*, vol. 191, no. 23, pp. 7363–7366, 2009.

[78] S. F. Lee, G. D. Delaney, and M. Elkhateeb, "A two-component covrs regulatory system regulates expression of fructosyltransferase and a novel extracellular carbohydrate in *Streptococcus mutans*," *Infection and Immunity*, vol. 72, no. 7, pp. 3968–3973, 2004.

[79] van der Ploeg, Jan R, "Regulation of bacteriocin production in *Streptococcus mutans* by the quorum-sensing system required for development of genetic competence," *Journal of Bacteriology*, vol. 187, no. 12, pp. 3980–3989, 2005.

[80] M. D. Senadheera, B. Guggenheim, G. A. Spatafora, Y.-C. C. Huang, J. Choi, D. C. I. Hung, J. S. Treglown, S. D. Goodman, R. P. Ellen, and D. G. Cvitkovitch, "A vicrk signal transduction system in *Streptococcus mutans* affects gtfbcd, gbpb, and ftf expression, biofilm formation, and genetic competence development," *Journal of Bacteriology*, vol. 187, no. 12, pp. 4064–4076, 2005.

[81] C. M. Levesque, R. W. Mair, J. A. Perry, P. C. Y. Lau, Y.-H. Li, and D. G. Cvitkovitch, "Systemic inactivation and phenotypic characterization of two-component systems in expression of *Streptococcus mutans* virulence properties," *Letters in Applied Microbiology*, vol. 45, no. 4, pp. 398–404, 2007.

[82] V. Idone, S. Brendtro, R. Gillespie, S. Kocaj, E. Peterson, M. Rendi, W. Warren, S. Michalek, K. Krastel, D. Cvitkovitch, and G. Spatafora, "Effect of an orphan response regulator on *Streptococcus mutans* sucrose-dependent adherence and cariogenesis," *Infection and Immunity*, vol. 71, no. 8, pp. 4351–4360, 2003.

[83] P.-M. Chen, H.-C. Chen, C.-T. Ho, C.-J. Jung, H.-T. Lien, J.-Y. Chen, and J.-S. Chia, "The two-component system scnrk of *Streptococcus mutans* affects hydrogen peroxide resistance and murine macrophage killing," *Microbes and Infection*, vol. 10, no. 3, pp. 293–301, 2008.

[84] L. Zeng, S. Das, and R. A. Burne, "Genetic analysis of the functions and interactions of components of the levqrst signal transduction complex of *Streptococcus mutans*," *PloS One*, vol. 6, no. 2, p. e17335, 2011.

[85] P. Chong, L. Drake, and I. Biswas, "Modulation of covr expression in *Streptococcus mutans* ua159," *Journal of Bacteriology*, vol. 190, no. 13, pp. 4478–4488, 2008.

[86] P. Chong, P. Chattoraj, and I. Biswas, "Activation of the smu.1882 transcription by covr in *Streptococcus mutans*," *PloS One*, vol. 5, no. 11, p. e15528, 2010.

[87] A. Dmitriev, S. S. Mohapatra, P. Chong, M. Neely, S. Biswas, and I. Biswas, "Covr-controlled global regulation of gene expression in *Streptococcus mutans*," *PloS One*, vol. 6, no. 5, p. e20127, 2011.

[88] I. Biswas, L. Drake, and S. Biswas, "Regulation of gbpc expression in *Streptococcus mutans*," *Journal of Bacteriology*, vol. 189, no. 18, pp. 6521–6531, 2007.

[89] I. Biswas and J. R. Scott, "Identification of roca, a positive regulator of covr expression in the group a *Streptococcus*," *Journal of Bacteriology*, vol. 185, no. 10, pp. 3081–3090, 2003.

[90] R. E. McLaughlin, J. J. Ferretti, and W. L. Hynes, "Nucleotide sequence of the streptococcin a-ff22 lantibiotic regulon: model for production of the lantibiotic sa-ff22 by strains of *Streptococcus pyogenes*," *FEMS Microbiology Letters*, vol. 175, no. 2, pp. 171–177, 1999.

[91] S.-J. Ahn, Z. T. Wen, and R. A. Burne, "Multilevel control of competence development and stress tolerance in *Streptococcus mutans* UA159," *Infection and Immunity*, vol. 74, no. 3, pp. 1631–1642, 2006.

[92] J. A. Perry, C. M. Levesque, P. Suntharaligam, R. W. Mair, M. Bu, R. T. Cline, S. N. Peterson, and D. G. Cvitkovitch, "Involvement of *Streptococcus mutans* regulator rr11 in oxidative stress response during biofilm growth and in the development of genetic competence," *Letters in Applied Microbiology*, vol. 47, no. 5, pp. 439–444, 2008.

[93] Y.-H. Li, N. Tang, M. B. Aspiras, P. C. Y. Lau, J. H. Lee, R. P. Ellen, and D. G. Cvitkovitch, "A quorum-sensing signaling system essential for genetic competence in *Streptococcus mutans* is involved in biofilm formation," *Journal of Bacteriology*, vol. 184, no. 10, pp. 2699–2708, 2002.

[94] Y.-H. Li, X.-L. Tian, G. Layton, C. Norgaard, and G. Sisson, "Additive attenuation of virulence and cariogenic potential of *Streptococcus mutans* by simultaneous inactivation of the comcde quorum-sensing system and hk/rr11 two-component regulatory system," *Microbiology (Reading, England)*, vol. 154, no. Pt 11, pp. 3256–3265, 2008.

[95] J. Kreth, D. C. I. Hung, J. Merritt, J. Perry, L. Zhu, S. D. Goodman, D. G. Cvitkovitch, W. Shi, and F. Qi, "The response regulator come in *Streptococcus mutans* functions

both as a transcription activator of mutacin production and repressor of csp biosynthesis," *Microbiology (Reading, England)*, vol. 153, no. Pt 6, pp. 1799–1807, 2007.

[96] D. I. Liao, J. Qian, D. A. Chisholm, D. B. Jordan, and B. A. Diner, "Crystal structures of the photosystem ii d1 c-terminal processing protease," *Nature structural biology*, vol. 7, no. 9, pp. 749–753, 2000.

[97] M. Ansaldi and D. Dubnau, "Diversifying selection at the bacillus quorum-sensing locus and determinants of modification specificity during synthesis of the comx pheromone," *Journal of Bacteriology*, vol. 186, no. 1, pp. 15–21, 2004.

[98] C. Poyart, M. C. Lamy, C. Boumaila, F. Fiedler, and P. Trieu-Cuot, "Regulation of d-alanyl-lipoteichoic acid biosynthesis in *Streptococcus agalactiae* involves a novel two-component regulatory system," *Journal of Bacteriology*, vol. 183, no. 21, pp. 6324–6334, 2001.

[99] N. Fittipaldi, T. Sekizaki, D. Takamatsu, J. Harel, Dominguez-Punaro, Maria de la Cruz, S. von Aulock, C. Draing, C. Marois, M. Kobisch, and M. Gottschalk, "D-alanylation of lipoteichoic acid contributes to the virulence of *Streptococcus suis*," *Infection and Immunity*, vol. 76, no. 8, pp. 3587–3594, 2008.

[100] F. Fabretti, C. Theilacker, L. Baldassarri, Z. Kaczynski, A. Kropec, O. Holst, and J. Huebner, "Alanine esters of enterococcal lipoteichoic acid play a role in biofilm formation and resistance to antimicrobial peptides," *Infection and Immunity*, vol. 74, no. 7, pp. 4164–4171, 2006.

[101] J. M. Tanzer, J. Livingston, and A. M. Thompson, "The microbiology of primary dental caries in humans," *Journal of Dental Education*, vol. 65, no. 10, pp. 1028–1037, 2001.

[102] J. D. F. Hale, N. C. K. Heng, R. W. Jack, and J. R. Tagg, "Identification of nlmte, the locus encoding the abc transport system required for export of nonlantibiotic mutacins in *Streptococcus mutans*," *Journal of Bacteriology*, vol. 187, no. 14, pp. 5036–5039, 2005.

[103] M. S. Hossain and I. Biswas, "Mutacins from *Streptococcus mutans* ua159 are active against multiple streptococcal species," *Applied and Environmental Microbiology*, vol. 77, no. 7, pp. 2428–2434, 2011.

[104] Y. Liu and H. Lu, "Microfluidics in systems biology-hype or truly useful?," *Current Opinion in Biotechnology*, vol. 39, pp. 215–220, 2016.

[105] F. C. Petersen and A. A. Scheie, "Genetic transformation in streptococcus mutans requires a peptide secretion-like apparatus," *Oral Microbiology and Immunology*, vol. 15, no. 5, pp. 329–334, 2000.

[106] F. C. Petersen, G. Fimland, and A. A. Scheie, "Purification and functional studies of a potent modified quorum-sensing peptide and a two-peptide bacteriocin in streptococcus mutans," *Molecular microbiology*, vol. 61, no. 5, pp. 1322–1334, 2006.

[107] E. Allan, H. A. Hussain, K. R. Crawford, S. Miah, Z. K. Ascott, M. H. Khwaja, and A. H. F. Hosie, "Genetic variation in comc, the gene encoding competence-stimulating peptide (csp) in *Streptococcus mutans*," *FEMS Microbiology Letters*, vol. 268, no. 1, pp. 47–51, 2007.

[108] J. Kreth, J. Merritt, W. Shi, and F. Qi, "Co-ordinated bacteriocin production and competence development: a possible mechanism for taking up dna from neighbouring species," *Molecular Microbiology*, vol. 57, no. 2, pp. 392–404, 2005.

[109] J. Kreth, J. Merritt, L. Zhu, W. Shi, and F. Qi, "Cell density- and come-dependent expression of a group of mutacin and mutacin-like genes in *Streptococcus mutans*," *FEMS Microbiology Letters*, vol. 265, no. 1, pp. 11–17, 2006.

[110] L. Mashburn-Warren, D. A. Morrison, and M. J. Federle, "A novel double-tryptophan peptide pheromone controls competence in *Streptococcus spp.* via an rgg regulator," *Molecular Microbiology*, vol. 78, no. 3, pp. 589–606, 2010.

[111] T. Okinaga, G. Niu, Z. Xie, F. Qi, and J. Merritt, "The hdrrm operon of *Streptococcus mutans* encodes a novel regulatory system for coordinated competence development and bacteriocin production," *Journal of Bacteriology*, vol. 192, no. 7, pp. 1844–1852, 2010.

[112] T. Okinaga, Z. Xie, G. Niu, F. Qi, and J. Merritt, "Examination of the hdrrm regulon yields insight into the competence system of *Streptococcus mutans*," *Molecular Oral Microbiology*, vol. 25, no. 3, pp. 165–177, 2010.

[113] Z. Xie, T. Okinaga, G. Niu, F. Qi, and J. Merritt, "Identification of a novel bacteriocin regulatory system in *Streptococcus mutans*," *Molecular Microbiology*, vol. 78, no. 6, pp. 1431–1447, 2010.

[114] R. W. Mair, D. B. Senadheera, and D. G. Cvitkovitch, "Cina is regulated via comx to modulate genetic transformation and cell viability in streptococcus mutans," *FEMS microbiology letters*, vol. 331, no. 1, pp. 44–52, 2012.

[115] S. Alaluusua, T. Takei, T. Ooshima, and S. Hamada, "Mutacin activity of strains isolated from children with varying levels of mutants streptococci and caries," *Archives of Oral Biology*, vol. 36, no. 4, pp. 251–255, 1991.

[116] T. Baba and O. Schneewind, "Instruments of microbial warfare: bacteriocin synthesis, toxicity and immunity," *Trends in Microbiology*, vol. 6, no. 2, pp. 66–71, 1998.

[117] I. F. Nes, D. B. Diep, and H. Holo, "Bacteriocin diversity in *Streptococcus* and *Enterococcus*," *Journal of Bacteriology*, vol. 189, no. 4, pp. 1189–1198, 2007.

[118] S. Bekal-Si Ali, Y. Hurtubise, M. C. Lavoie, and G. LaPointe, "Diversity of *Streptococcus mutans* bacteriocins as confirmed by dna analysis using specific molecular probes," *Gene*, vol. 283, no. 1-2, pp. 125–131, 2002.

[119] H. Yonezawa and H. K. Kuramitsu, "Genetic analysis of a unique bacteriocin, smb, produced by *Streptococcus mutans* gs5," *Antimicrobial Agents and Chemotherapy*, vol. 49, no. 2, pp. 541–548, 2005.

[120] O. Hyink, M. Balakrishnan, and J. R. Tagg, "Streptococcus rattus strain bht produces both a class i two-component lantibiotic and a class ii bacteriocin," *FEMS microbiology letters*, vol. 252, no. 2, pp. 235–241, 2005.

[121] T. Nguyen, Z. Zhang, I.-H. Huang, C. Wu, J. Merritt, W. Shi, and F. Qi, "Genes involved in the repression of mutacin i production in *Streptococcus mutans*," *Microbiology (Reading, England)*, vol. 155, no. Pt 2, pp. 551–556, 2009.

[122] P. Chen, F. Qi, J. Novak, and P. W. Caufield, "The specific genes for lantibiotic mutacin ii biosynthesis in *Streptococcus mutans* t8 are clustered and can be transferred en bloc," *Applied and Environmental Microbiology*, vol. 65, no. 3, pp. 1356–1360, 1999.

[123] F. Qi, P. Chen, and P. W. Caufield, "Purification of mutacin iii from group iii *Streptococcus mutans* ua787 and genetic analyses of mutacin iii biosynthesis genes," *Applied and Environmental Microbiology*, vol. 65, no. 9, pp. 3880–3887, 1999.

[124] D. W. Johnson, J. R. Tagg, and L. W. Wannamaker, "Production of a bacteriocine-like substance by group-a streptococci of m-type 4 and t-pattern 4," *Journal of Medical Microbiology*, vol. 12, no. 4, pp. 413–427, 1979.

[125] C. L. Robson, P. A. Wescombe, N. A. Klesse, and J. R. Tagg, "Isolation and partial characterization of the *Streptococcus mutans* type aii lantibiotic mutacin k8," *Microbiology (Reading, England)*, vol. 153, no. Pt 5, pp. 1631–1641, 2007.

[126] F. Qi, P. Chen, and P. W. Caufield, "The group i strain of *Streptococcus mutans*, ua140, produces both the lantibiotic mutacin i and a nonlantibiotic bacteriocin, mutacin iv," *Applied and Environmental Microbiology*, vol. 67, no. 1, pp. 15–21, 2001.

[127] J. A. Perry, M. B. Jones, S. N. Peterson, D. G. Cvitkovitch, and C. M. Levesque, "Peptide alarmone signalling triggers an auto-active bacteriocin necessary for genetic competence," *Molecular Microbiology*, vol. 72, no. 4, pp. 905–917, 2009.

[128] A. K. Chatterjee and M. P. Starr, "Transfer among erwinia spp. and other enterobacteria of antibiotic resistance carried on r factors," *Journal of Bacteriology*, vol. 112, no. 1, pp. 576–584, 1972.

[129] H. Yano, A. Kuga, R. Okamoto, H. Kitasato, T. Kobayashi, and M. Inoue, "Plasmid-encoded metallo-beta-lactamase (imp-6) conferring resistance to carbapenems, especially meropenem," *Antimicrobial agents and chemotherapy*, vol. 45, no. 5, pp. 1343–1348, 2001.

[130] J. Ouyang, X.-L. Tian, J. Versey, A. Wishart, and Y.-H. Li, "The bceabrs four-component system regulates the bacitracin-induced cell envelope stress response in streptococcus mutans," *Antimicrobial agents and chemotherapy*, vol. 54, no. 9, pp. 3895–3906, 2010.

[131] H. Tsuda, Y. Yamashita, Y. Shibata, Y. Nakano, and T. Koga, "Genes involved in bacitracin resistance in streptococcus mutans," *Antimicrobial agents and chemotherapy*, vol. 46, no. 12, pp. 3756–3764, 2002.

[132] M. El Ghachi, A. Bouhss, D. Blanot, and D. Mengin-Lecreulx, "The baca gene of escherichia coli encodes an undecaprenyl pyrophosphate phosphatase activity," *The Journal of biological chemistry*, vol. 279, no. 29, pp. 30106–30113, 2004.

[133] R. Bernard, M. El Ghachi, D. Mengin-Lecreulx, M. Chippaux, and F. Denizot, "Bcrc from bacillus subtilis acts as an undecaprenyl pyrophosphate phosphatase in bacitracin resistance," *The Journal of biological chemistry*, vol. 280, no. 32, pp. 28852–28857, 2005.

[134] J. M. McCord and I. Fridovich, "Superoxide dismutase: The first twenty years (1968–1988)," *Free Radical Biology and Medicine*, vol. 5, no. 5-6, pp. 363–369, 1988.

[135] Y. Yamamoto, M. Higuchi, L. B. Poole, and Y. Kamio, "Role of the dpr product in oxygen tolerance in *Streptococcus mutans*," *Journal of Bacteriology*, vol. 182, no. 13, pp. 3740–3747, 2000.

[136] M. Higuchi, Y. Yamamoto, and Y. Kamio, "Molecular biology of oxygen tolerance in lactic acid bacteria: Functions of nadh oxidases and dpr in oxidative stress," *Journal of Bioscience and Bioengineering*, vol. 90, no. 5, pp. 484–493, 2000.

[137] Y. Yamamoto, L. B. Poole, R. R. Hantgan, and Y. Kamio, "An iron-binding protein, dpr, from *Streptococcus mutans* prevents iron-dependent hydroxyl radical formation in vitro," *Journal of Bacteriology*, vol. 184, no. 11, pp. 2931–2939, 2002.

[138] D. Mustacich and G. Powis, "Thioredoxin reductase," *The Biochemical journal*, vol. 346 Pt 1, pp. 1–8, 2000.

[139] E. S. Arner and A. Holmgren, "Physiological functions of thioredoxin and thioredoxin reductase," *European Journal of Biochemistry*, vol. 267, no. 20, pp. 6102–6109, 2000.

[140] H.-J. Seo and Y. N. Lee, "Characterization of deinococcus radiophilus thioredoxin reductase active with both nadh and nadph," *Journal of microbiology (Seoul, Korea)*, vol. 48, no. 5, pp. 637–643, 2010.

[141] A. P. Fernandes and A. Holmgren, "Glutaredoxins: glutathione-dependent redox enzymes with functions far beyond a simple thioredoxin backup system," *Antioxidants & Redox Signaling*, vol. 6, no. 1, pp. 63–74, 2004.

[142] A. Holmgren, "Thioredoxin and glutaredoxin systems," *The Journal of Biological Chemistry*, vol. 264, no. 24, pp. 13963–13966, 1989.

[143] R. C. Fahey, W. C. Brown, W. B. Adams, and M. B. Worsham, "Occurrence of glutathione in bacteria," *Journal of Bacteriology*, vol. 133, no. 3, pp. 1126–1129, 1978.

[144] B. E. Janowiak and O. W. Griffith, "Glutathione synthesis in streptococcus agalactiae. one protein accounts for gamma-glutamylcysteine synthetase and glutathione synthetase activities," *The Journal of Biological Chemistry*, vol. 280, no. 12, pp. 11829–11839, 2005.

[145] J. Zhang and I. Biswas, "3'-phosphoadenosine-5'-phosphate phosphatase activity is required for superoxide stress tolerance in streptococcus mutans," *Journal of bacteriology*, vol. 191, no. 13, pp. 4330–4340, 2009.

[146] H. Taniai, K.-i. Iida, M. Seki, M. Saito, S. Shiota, H. Nakayama, and S.-i. Yoshida, "Concerted action of lactate oxidase and pyruvate oxidase in aerobic growth of *Streptococcus pneumoniae*: role of lactate as an energy source," *Journal of Bacteriology*, vol. 190, no. 10, pp. 3572–3579, 2008.

[147] J. Kreth, J. Merritt, W. Shi, and F. Qi, "Competition and coexistence between *Streptococcus mutans* and *Streptococcus sanguinis* in the dental biofilm," *Journal of Bacteriology*, vol. 187, no. 21, pp. 7193–7203, 2005.

[148] N. Okahashi, M. Nakata, T. Sumitomo, Y. Terao, and S. Kawabata, "Hydrogen peroxide produced by oral streptococci induces macrophage cell death," *PloS one*, vol. 8, no. 5, p. e62563, 2013.

[149] S. Subramanian and C. Sivaraman, "Bacterial citrate lyase," *Journal of Biosciences*, vol. 6, no. 4, pp. 379–401, 1984.

[150] H. J. Evans and H. G. Wood, "The mechanism of the pyruvate, phosphate dikinase reaction," *Proceedings of the National Academy of Sciences*, vol. 61, no. 4, pp. 1448–1453, 1968.

[151] M. Benziman, N. Eisen, and A. Palgi, "Properties and physiological role of the pep-synthase of a. xylinum," *FEBS letters*, vol. 3, no. 2, pp. 156–159, 1969.

[152] U. Sauer and B. J. Eikmanns, "The pep-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria," *FEMS microbiology reviews*, vol. 29, no. 4, pp. 765–794, 2005.

[153] D. G. Cvitkovitch, J. A. Gutierrez, and A. S. Bleiweis, "Role of the citrate pathway in glutamate biosynthesis by *Streptococcus mutans*," *Journal of Bacteriology*, vol. 179, no. 3, pp. 650–655, 1997.

[154] P. Sudhakar, M. Reck, W. Wang, F. Q. He, I. Wagner-Döbler, I. W. Dobler, and A.-P. Zeng, "Construction and verification of the transcriptional regulatory response network of *Streptococcus mutans* upon treatment with the biofilm inhibitor carolacton," *BMC Genomics*, vol. 15, p. 362, 2014.

[155] J. Macia, F. Posas, and R. V. Sole, "Distributed computation: the new wave of synthetic biology devices," *Trends in Biotechnology*, vol. 30, no. 6, pp. 342–349, 2012.

[156] Y. Benenson, "Biocomputers: from test tubes to live cells," *Molecular bioSystems*, vol. 5, no. 7, pp. 675–685, 2009.

[157] S. Regot, J. Macia, N. Conde, K. Furukawa, J. Kjellen, T. Peeters, S. Hohmann, E. de Nadal, F. Posas, and R. Sole, "Distributed biological computation with multicellular engineered networks," *Nature*, vol. 469, no. 7329, pp. 207–211, 2011.

[158] J. Bonnet, P. Yin, M. E. Ortiz, P. Subsoontorn, and D. Endy, "Amplifying genetic logic gates," *Science (New York, N.Y.)*, vol. 340, no. 6132, pp. 599–603, 2013.

[159] J. Macia and R. Sole, "How to make a synthetic multicellular computer," *PloS One*, vol. 9, no. 2, p. e81248, 2014.

[160] N. V. Hud and D. G. Lynn, "From life's origins to a synthetic biology," *Current Opinion in Chemical Biology*, vol. 8, no. 6, pp. 627–628, 2004.

[161] S. A. Benner and A. M. Sismour, "Synthetic biology," *Nature reviews. Genetics*, vol. 6, no. 7, pp. 533–543, 2005.

[162] J. D. Keasling, "Synthetic biology for synthetic chemistry," *ACS Chemical Biology*, vol. 3, no. 1, pp. 64–76, 2008.

[163] P. E. M. Purnick and R. Weiss, "The second wave of synthetic biology: from modules to systems," *Nature reviews. Molecular Cell Biology*, vol. 10, no. 6, pp. 410–422, 2009.

[164] A. S. Khalil and J. J. Collins, "Synthetic biology: applications come of age," *Nature reviews. Genetics*, vol. 11, no. 5, pp. 367–379, 2010.

[165] J. Liang, Y. Luo, and H. Zhao, "Synthetic biology: putting synthesis into biology," *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, vol. 3, no. 1, pp. 7–20, 2011.

[166] N. Nandagopal and M. B. Elowitz, "Synthetic biology: integrated gene circuits," *Science (New York, N.Y.)*, vol. 333, no. 6047, pp. 1244–1248, 2011.

[167] A. A. Cheng and T. K. Lu, "Synthetic biology: an emerging engineering discipline," *Annual Review of Biomedical Engineering*, vol. 14, pp. 155–178, 2012.

[168] R. Kitney and P. Freemont, "Synthetic biology - the state of play," *FEBS Letters*, vol. 586, no. 15, pp. 2029–2036, 2012.

[169] S. A. Lynch and R. T. Gill, "Synthetic biology: new strategies for directing design," *Metabolic Engineering*, vol. 14, no. 3, pp. 205–211, 2012.

[170] K. A. Riccione, R. P. Smith, A. J. Lee, and L. You, "A synthetic biology approach to understanding cellular information processing," *ACS Synthetic Biology*, vol. 1, no. 9, pp. 389–402, 2012.

[171] G. Stephanopoulos, "Synthetic biology and metabolic engineering," *ACS Synthetic Biology*, vol. 1, no. 11, pp. 514–525, 2012.

[172] Y.-H. Wang, K. Y. Wei, and C. D. Smolke, "Synthetic biology: advancing the design of diverse genetic systems," *Annual Review of Chemical and Biomolecular Engineering*, vol. 4, pp. 69–102, 2013.

[173] B. Zakeri and T. K. Lu, "Synthetic biology of antimicrobial discovery," *ACS Synthetic Biology*, vol. 2, no. 7, pp. 358–372, 2013.

[174] D. E. Cameron, C. J. Bashor, and J. J. Collins, "A brief history of synthetic biology," *Nature reviews. Microbiology*, vol. 12, no. 5, pp. 381–390, 2014.

[175] G. M. Church, M. B. Elowitz, C. D. Smolke, C. A. Voigt, and R. Weiss, "Realizing the potential of synthetic biology," *Nature reviews. Molecular Cell Biology*, vol. 15, no. 4, pp. 289–294, 2014.

[176] J. C. Way, J. J. Collins, J. D. Keasling, and P. A. Silver, "Integrating biological redesign: where synthetic biology came from and where it needs to go," *Cell*, vol. 157, no. 1, pp. 151–161, 2014.

[177] R. Chao, Y. Yuan, and H. Zhao, "Building biological foundries for next-generation synthetic biology," *Science China. Life sciences*, vol. 58, no. 7, pp. 658–665, 2015.

[178] K. A. Markham and H. S. Alper, "Synthetic biology for specialty chemicals," *Annual Review of Chemical and Biomolecular Engineering*, vol. 6, pp. 35–52, 2015.

[179] S. Slomovic, K. Pardee, and J. J. Collins, "Synthetic biology devices for *in vitro* and *in vivo* diagnostics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 47, pp. 14429–14435, 2015.

[180] S. Auslander, D. Auslander, and M. Fussenegger, "Synthetic biology-the synthesis of biology," *Angewandte Chemie (International ed. in English)*, 2016.

[181] M. C. Bassalo, R. Liu, and R. T. Gill, "Directed evolution and synthetic biology applications to microbial systems," *Current Opinion in Biotechnology*, vol. 39, pp. 126–133, 2016.

[182] R. Breitling and E. Takano, "Synthetic biology of natural products," *Cold Spring Harbor Perspectives in Biology*, vol. 8, no. 10, 2016.

[183] M. J. Smanski, H. Zhou, J. Claesen, B. Shen, M. A. Fischbach, and C. A. Voigt, "Synthetic biology to access and expand nature's chemical diversity," *Nature reviews. Microbiology*, vol. 14, no. 3, pp. 135–149, 2016.

[184] A. A. K. Nielsen, B. S. Der, J. Shin, P. Vaidyanathan, V. Paralanov, E. A. Strychalski, D. Ross, D. Densmore, and C. A. Voigt, "Genetic circuit design automation," *Science (New York, N.Y.)*, vol. 352, no. 6281, p. aac7341, 2016.

[185] P. Mali, K. M. Esvelt, and G. M. Church, "Cas9 as a versatile tool for engineering biology," *Nature Methods*, vol. 10, no. 10, pp. 957–963, 2013.

[186] J. A. Doudna and E. Charpentier, "Genome editing. the new frontier of genome engineering with crispr-cas9," *Science (New York, N.Y.)*, vol. 346, no. 6213, p. 1258096, 2014.

[187] P. D. Hsu, E. S. Lander, and F. Zhang, "Development and applications of crispr-cas9 for genome engineering," *Cell*, vol. 157, no. 6, pp. 1262–1278, 2014.

[188] J. D. Sander and J. K. Joung, "Crispr-cas systems for editing, regulating and targeting genomes," *Nature Biotechnology*, vol. 32, no. 4, pp. 347–355, 2014.

[189] D. Na, S. M. Yoo, H. Chung, H. Park, J. H. Park, and S. Y. Lee, "Metabolic engineering of *Escherichia coli* using synthetic small regulatory rnas," *Nature Biotechnology*, vol. 31, no. 2, pp. 170–174, 2013.

[190] S. M. Yoo, D. Na, and S. Y. Lee, "Design and use of synthetic regulatory small rnas to control gene expression in *Escherichia coli*," *Nature Protocols*, vol. 8, no. 9, pp. 1694–1707, 2013.

[191] V. Gailus, U. Ramsperger, C. Johner, H. Kramer, and I. Rasched, "The role of the adsorption complex in the termination of filamentous phage assembly," *Research in Microbiology*, vol. 145, no. 9, pp. 699–709, 1994.

[192] Y. H. Park, H. M. Koo, J. O. Moon, S. J. Kim, H. J. Kim, and J. K. Lee, "Novel l-lysine-inducible promoter," 2006.

[193] J. Blazeck and H. Alper, "Systems metabolic engineering: genome-scale models and beyond," *Biotechnology Journal*, vol. 5, no. 7, pp. 647–659, 2010.

[194] J. Almquist, M. Cvijovic, V. Hatzimanikatis, J. Nielsen, and M. Jirstrand, "Kinetic models in industrial biotechnology - improving cell factory performance," *Metabolic Engineering*, vol. 24, pp. 38–60, 2014.

[195] J. Theze, D. Margarita, G. N. Cohen, F. Borne, and J. C. Patte, "Mapping of the structural genes of the three aspartokinases and of the two homoserine dehydrogenases of *Escherichia coli* k-12," *Journal of Bacteriology*, vol. 117, no. 1, pp. 133–143, 1974.

[196] Z. Chen, W. Meyer, S. Rappert, J. Sun, and A.-P. Zeng, "Coevolutionary analysis enabled rational deregulation of allosteric enzyme inhibition in *Corynebacterium glutamicum* for lysine production," *Applied and Environmental Microbiology*, vol. 77, no. 13, pp. 4352–4360, 2011.

[197] U. T. Bornscheuer, "Protein engineering: Beating the odds," *Nature Chemical Biology*, vol. 12, no. 2, pp. 54–55, 2016.

[198] B. Chen, S. Lim, A. Kannan, S. C. Alford, F. Sunden, D. Herschlag, I. K. Dimov, T. M. Baer, and J. R. Cochran, "High-throughput analysis and protein engineering using microcapillary arrays," *Nature Chemical Biology*, vol. 12, no. 2, pp. 76–81, 2016.

[199] P.-Y. Colin, B. Kintses, F. Gielen, C. M. Miton, G. Fischer, M. F. Mohamed, M. Hyvonen, D. P. Morgavi, D. B. Janssen, and F. Hollfelder, "Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics," *Nature Communications*, vol. 6, p. 10008, 2015.

[200] P.-Y. Colin, A. Zinchenko, and F. Hollfelder, "Enzyme engineering in biomimetic compartments," *Current Opinion in Structural Biology*, vol. 33, pp. 42–51, 2015.

[201] N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, and R. H. Scheuermann, "Critical assessment of automated flow cytometry data analysis techniques," *Nature Methods*, vol. 10, no. 3, pp. 228–238, 2013.

[202] L. W. Arnold and J. Lannigan, "Practical issues in high-speed cell sorting," *Current Protocols in Cytometry*, vol. Chapter 1, pp. Unit 1.24.1–30, 2010.

[203] S. Binder, G. Schendzielorz, N. Stabler, K. Krumbach, K. Hoffmann, M. Bott, and L. Eggeling, "A high-throughput approach to identify genomic variants of bacterial metabolite producers at the single-cell level," *Genome Biology*, vol. 13, no. 5, p. R40, 2012.

[204] G. Schendzielorz, M. Dippong, A. Grunberger, D. Kohlheyer, A. Yoshida, S. Binder, C. Nishiyama, M. Nishiyama, M. Bott, and L. Eggeling, "Taking control over control: use of product sensing in single cells to remove flux control at key enzymes in biosynthesis pathways," *ACS Synthetic Biology*, vol. 3, no. 1, pp. 21–29, 2014.

[205] Y. Wang, Q. Li, P. Zheng, Y. Guo, L. Wang, T. Zhang, J. Sun, and Y. Ma, "Evolving the l-lysine high-producing strain of escherichia coli using a newly developed high-throughput screening method," *Journal of Industrial Microbiology & Biotechnology*, 2016.

[206] J. W. Young, J. C. W. Locke, A. Altinok, N. Rosenfeld, T. Bacarian, P. S. Swain, E. Mjolsness, and M. B. Elowitz, "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy," *Nature Protocols*, vol. 7, no. 1, pp. 80–88, 2011.

[207] D.-K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, M. C. Y. Chang, S. T. Withers, Y. Shiba, R. Sarpong, and J. D. Keasling, "Production of the antimalarial drug precursor artemisinic acid in engineered yeast," *Nature*, vol. 440, no. 7086, pp. 940–943, 2006.

[208] P. P. Peralta-Yahya, F. Zhang, S. B. del Cardayre, and J. D. Keasling, "Microbial engineering for the production of advanced biofuels," *Nature*, vol. 488, no. 7411, pp. 320–328, 2012.

[209] J. H. Ahn, Y.-S. Jang, and S. Y. Lee, "Production of succinic acid by metabolically engineered microorganisms," *Current Opinion in Biotechnology*, vol. 42, pp. 54–66, 2016.

[210] S. Y. Choi, S. J. Park, W. J. Kim, J. E. Yang, H. Lee, J. Shin, and S. Y. Lee, "One-step fermentative production of poly(lactate-co-glycolate) from carbohydrates in *Escherichia coli*," *Nature Biotechnology*, vol. 34, no. 4, pp. 435–440, 2016.

[211] T. J. Park, K. G. Lee, and S. Y. Lee, "Advances in microbial biosynthesis of metal nanoparticles," *Applied Microbiology and Biotechnology*, vol. 100, no. 2, pp. 521–534, 2016.

[212] S. H. Park, H. U. Kim, T. Y. Kim, J. S. Park, S.-S. Kim, and S. Y. Lee, "Metabolic engineering of *Corynebacterium glutamicum* for l-arginine production," *Nature Communications*, vol. 5, p. 4618, 2014.

[213] J. C. Liao, L. Mi, S. Pontrelli, and S. Luo, "Fuelling the future: microbial engineering for the production of sustainable biofuels," *Nature reviews. Microbiology*, vol. 14, no. 5, pp. 288–304, 2016.

[214] R. R. Bommareddy, Z. Chen, S. Rappert, and A.-P. Zeng, "A de novo nadph generation pathway for improving lysine production of *Corynebacterium glutamicum* by rational

design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase," *Metabolic Engineering*, vol. 25, pp. 30–37, 2014.

[215] J. H. Lee, R. J. Mitchell, B. C. Kim, D. C. Cullen, and M. B. Gu, "A cell array biosensor for environmental toxicity analysis," *Biosensors & Bioelectronics*, vol. 21, no. 3, pp. 500–507, 2005.

[216] J. W. Kotula, S. J. Kerns, L. A. Shaket, L. Siraj, J. J. Collins, J. C. Way, and P. A. Silver, "Programmable bacteria detect and record an environmental signal in the mammalian gut," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 13, pp. 4838–4843, 2014.

[217] N. Saeidi, C. K. Wong, T.-M. Lo, H. X. Nguyen, H. Ling, S. S. J. Leong, C. L. Poh, and M. W. Chang, "Engineering microbes to sense and eradicate *Pseudomonas aeruginosa*, a human pathogen," *Molecular Systems Biology*, vol. 7, p. 521, 2011.

[218] K. M. Esvelt, J. C. Carlson, and D. R. Liu, "A system for the continuous directed evolution of biomolecules," *Nature*, vol. 472, no. 7344, pp. 499–503, 2011.

[219] B. C. Dickinson, M. S. Packer, A. H. Badran, and D. R. Liu, "A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations," *Nature Communications*, vol. 5, p. 5352, 2014.

[220] J. C. Carlson, A. H. Badran, D. A. Guggiana-Nilo, and D. R. Liu, "Negative selection and stringency modulation in phage-assisted continuous evolution," *Nature Chemical Biology*, vol. 10, no. 3, pp. 216–222, 2014.

[221] C.-J. Tsai, A. Del Sol, and R. Nussinov, "Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms," *Molecular bioSystems*, vol. 5, no. 3, pp. 207–216, 2009.

[222] C.-W. Ma, Z.-L. Xiu, and A.-P. Zeng, "A new concept to reveal protein dynamics based on energy dissipation," *PloS One*, vol. 6, no. 10, p. e26453, 2011.

[223] A. H. Badran and D. R. Liu, "*In vivo* continuous directed evolution," *Current Opinion in Chemical Biology*, vol. 24, pp. 1–10, 2015.

[224] M. Kotaka, J. Ren, M. Lockyer, A. R. Hawkins, and D. K. Stammers, "Structures of r- and t-state escherichia coli aspartokinase iii. mechanisms of the allosteric transition and inhibition by lysine," *The Journal of Biological Chemistry*, vol. 281, no. 42, pp. 31544–31552, 2006.

[225] Z. Chen, S. Rappert, J. Sun, and A.-P. Zeng, "Integrating molecular dynamics and co-evolutionary analysis for reliable target prediction and deregulation of the allosteric inhibition of aspartokinase for amino acid production," *Journal of Biotechnology*, vol. 154, no. 4, pp. 248–254, 2011.

[226] E. R. STADTMAN, G. N. Cohen, and G. LEBRAS, "Feedback inhibition and repression of aspartokinase activity in *Escherichia coli*," *Annals of the New York Academy of Sciences*, vol. 94, pp. 952–959, 1961.

[227] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "Ucsf chimera–a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[228] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, "Gpu computing," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 879–899, 2008.

[229] M. Deaner and H. S. Alper, "Promoter and terminator discovery and engineering," *Advances in Biochemical Engineering/Biotechnology*, 2016.

[230] C.-W. Ma, L.-B. Zhou, and A.-P. Zeng, "Engineering biomolecular switches for dynamic metabolic control," *Advances in Biochemical Engineering/Biotechnology*, 2016.

[231] J. D. WATSON and F. H. C. CRICK, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[232] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in dna microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, 1999.

[233] C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland, "Long-term storage of information in dna," *Science (New York, N.Y.)*, vol. 293, no. 5536, pp. 1763–1765, 2001.

[234] N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, and M. Tomita, "Alignment-based approach for durable data storage into living organisms," *Biotechnology Progress*, vol. 23, no. 2, pp. 501–505, 2007.

[235] C. Gustafsson, "For anyone who ever said there's no such thing as a poetic gene," *Nature*, vol. 458, no. 7239, p. 703, 2009.

[236] M. Ailenberg and O. Rotstein, "An improved huffman coding method for archiving text, images, and music characters in dna," *BioTechniques*, vol. 47, no. 3, pp. 747–754, 2009.

[237] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith, and J. C. Venter, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science (New York, N.Y.)*, vol. 329, no. 5987, pp. 52–56, 2010.

[238] A. Baccaro, A.-L. Steck, and A. Marx, "Barcoded nucleotides," *Angewandte Chemie (International ed. in English)*, vol. 51, no. 1, pp. 254–257, 2012.

[239] N. Yachie, Y. Ohashi, and M. Tomita, "Stabilizing synthetic data in the dna of living organisms," *Systems and Synthetic Biology*, vol. 2, no. 1-2, pp. 19–25, 2008.

[240] S.-H. Jiao and R. Goutte, "Hiding data in dna of living organisms," *Natural Science*, vol. 01, no. 03, pp. 181–184, 2009.

[241] A. Akhmetov, A. Ellington, and E. Marcotte, *A highly parallel strategy for storage of digital information in living cells*. Cold Spring Harbor Labs Journals, 2016.

[242] Y. Erlich and D. Zielinski, "Dna fountain enables a robust and efficient storage architecture," *Science (New York, N.Y.)*, vol. 355, no. 6328, pp. 950–954, 2017.

[243] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prufer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-Garcia, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B. Viola, and S. Paabo, "Genome sequence of a 45,000-year-old modern human from western siberia," *Nature*, vol. 514, no. 7523, pp. 445–449, 2014.

[244] A. Extance, "How dna could store all the world's data," *Nature*, vol. 537, no. 7618, pp. 22–24, 2016.

[245] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on dna in silica with error-correcting codes," *Angewandte Chemie (International ed. in English)*, vol. 54, no. 8, pp. 2552–2555, 2015.

[246] A. K.-Y. Yim, A. C.-S. Yu, J.-W. Li, A. I.-C. Wong, J. F. C. Loo, K. M. Chan, S. K. Kong, K. Y. Yip, and T.-F. Chan, "The essential component in dna-based information storage system: Robust error-tolerating module," *Frontiers in Bioengineering and Biotechnology*, vol. 2, p. 49, 2014.

[247] Yazdi, S M Hossein Tabatabaei, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access dna-based storage system," *Scientific Reports*, vol. 5, p. 14138, 2015.

[248] M. Blawat, K. Gaedke, I. Hütter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church, "Forward error correction for dna data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[249] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A dna-based archival storage system," in *the Twenty-First International Conference* (T. Conte and Y. Zhou, eds.), pp. 637–649, 2016.

[250] S. M. H. Tabatabaei Yazdi, R. Gabrys, and O. Milenkovic, *Portable and Error-Free DNA-Based Data Storage*. Cold Spring Harbor Labs Journals, 2016.

[251] "https://www.nanoporetech.com/."

[252] B. Wei and T. Chen, *Verifying Data Migration Correctness: The Checksum Principle*. RTI Press, 2014.

[253] D. J. Mandell, M. J. Lajoie, M. T. Mee, R. Takeuchi, G. Kuznetsov, J. E. Norville, C. J. Gregg, B. L. Stoddard, and G. M. Church, "Biocontainment of genetically modified organisms by synthetic protein design," *Nature*, vol. 518, no. 7537, pp. 55–60, 2015.

[254] A. J. Rovner, A. D. Haimovich, S. R. Katz, Z. Li, M. W. Grome, B. M. Gassaway, M. Amiram, J. R. Patel, R. R. Gallagher, J. Rinehart, and F. J. Isaacs, "Recoded organisms engineered to depend on synthetic amino acids," *Nature*, vol. 518, no. 7537, pp. 89–93, 2015.

[255] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature reviews. Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[256] I. J. Fijalkowska, R. M. Schaaper, and P. Jonczyk, "Dna replication fidelity in *Escherichia coli*: a multi-dna polymerase affair," *FEMS Microbiology Reviews*, vol. 36, no. 6, pp. 1105–1121, 2012.

[257] J. M. A. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. V. Piddock, "Molecular mechanisms of antibiotic resistance," *Nature reviews. Microbiology*, vol. 13, no. 1, pp. 42–51, 2015.

Appendix A

Source codes of utilized perl scripts

A.1 panGenomeAnalysis.pl

```
#!/usr/bin/perl

=head1 Name

panGenomeAnalysis.pl

=head1 Description

run pan- and core-genome analysis from ortholog groups output file.

=head1 Version

Author: Lifu Song   (lifu.song@tuhh.de)
Version: 1.1

=head1 Usage

perl panGenomeAnalysis.pl [options] input_file
-run      run core/pan (runcp) or run new genes (default)
-gnumer      Specific the total genome number of input file
-clunum      Specific the total cluster number of the input file
-samples     How many random samples will be outputed.

=head1 Exmple
```

```
1. Run core and pan analysis:
perl panGenomeAnalysis.pl  -run runcp -gnumber 67 -clunum 3000
     -samples 1000 input > output

2. Run new genes analysis:
perl panGenomeAnalysis.pl  -run newgene -gnumber 67 -clunum 3000
     -samples 1000 input > output

=cut

use strict;
use Getopt::Long;
use Data::Dumper;
use Math::Combinatorics;
use List::Util qw/sum/;


my $run;
my $g_number;
my $clu_num;
my $total_genes_file;
my $samples;
my $help;


GetOptions(
"run:s"=>\$run,
"gnumber:i"=>\$g_number,
"clunum:i"=>\$clu_num,
"totalgenes:s"=>\$total_genes_file,
"samples:i"=>\$samples,
"help"=>\$help
);
die `pod2text $0` if (@ARGV==0 || $help);

my $in=shift;

my @data;
```

```
open(IN,$in)||die("can not open the input file!\n");
while(<IN>){
    chomp;
    s/\a//;
    my @a;
    my $i;
    @a=split(/\t/,$_);
    for($i=1;$i<$g_number+1; $i++){
        if(!$a[$i]){$a[$i]=0;}
        push(@{$data[$i-1]},$a[$i]);
    }
}


my @arr_spec_genes_num;
open(TO,"$total_genes_file")||die("cannot open total genes number file\n!");
my $ss=0;
while(<TO>){
    chomp;
    if($_){
        my @arr=split(/\t/);
        push(@arr_spec_genes_num, $arr[1] - sum(@{$data[$ss]}) );
        $ss++;
    }
}

warn "Data loaded!\n";

if($run eq "runcp"){&runCP;}
else{&runNewGene;}




sub runCP(){
    warn "Starting to run the core & pan analysis.....\n";
    my $i;
    for($i=1;$i<=$g_number; $i++){
```

```
if(&combCount($i,$g_number) <= $samples*2){

    my $combinat=Math::Combinatorics->new(count => $i,
       data => [@data]);
    my @combo=();
    my %spec;
    my $sx;
    my $ii;

    for($ii=0;$ii<@data;$ii++){
        $spec{$data[$ii]}= $arr_spec_genes_num[$ii];
    }
    while(@combo=$combinat->next_combination){
        print "$i\t";
        print coreFix(\@combo);
        print "\t";
        print core(\@combo);
        print "\t";
        my $ttt=(int(pan(\@combo))+ int(&allSpec(\@combo,\%spec)));
        print "$ttt";
        print "\n";

    }
}else{
    my @arr_samples=randConm($samples,$i,$g_number);
    my $sX;

    foreach $sX (@arr_samples){

        my $sY;
        my $spec_genes=0;
        my @combo=();
        foreach $sY (@{$sX}){
            push(@combo,$data[$sY]);
            $spec_genes+=$arr_spec_genes_num[$sY];
        }


        print "$i\t";
```

```
                print coreFix(\@combo);

                print "\t";
                print core(\@combo);

                print "\t";

                print pan(\@combo)+$spec_genes;

                print "\n";

            }
        }
    }
}
sub runNewGene(){
    warn "Starting to run the new genes analysis.....\n";
    my $i;
    for($i=2;$i<=$g_number; $i++){
        if(&newCount($i) <= $samples ){
            my $combinat=Math::Combinatorics->new(count => $i-1,
                        data => [@data]);
            my @combo=();
            while(@combo=$combinat->next_combination){
                my $j=0;
                for($j=0;$j<$g_number;$j++){
                    if(!isInArr(\@combo,$data[$j])){
                        print "$i\t";
                        print newgenes(\@combo,$data[$j])
                        +$arr_spec_genes_num[$j];
                        print "\n";}else{}
                }
            }
        }else{
            my @arr_samples=randConmN($samples,$i,$g_number);
            my $sX;
            foreach $sX (@arr_samples){
                my $sY;
                my @combo=();
```

```
                my $newGn=shift(@{$sX});
                my $newGenome=$data[$newGn];


                foreach $sY (@{$sX}){
                    push(@combo,$data[$sY]);
                }
                print "$i\t";
                print newgenes(\@combo,$newGenome)
                +$arr_spec_genes_num[$newGn];
                print "\t";
                print "\n";
            }
        }
    }
}
sub isInArr(){
    my $arr=shift;
    my $ele=shift;
    my $sItem;
    foreach $sItem (@{$arr}){

        if($sItem eq $ele){
            return 1;
        }
    }

    return 0;

}

sub combCount(){
    my $a=shift;
    my $b=shift;

    return nn($b)/(nn($a)*nn($b-$a));
}
sub newCount(){
    my $a=shift;
```

```
    my $b=shift;
    return nn($b)/(nn($a)*nn($b-$a-1));

}

sub nn(){
    my $n=int(shift);
    if($n>1){
        return $n* &nn($n-1);
    }else{return 1}
}
sub allSpec(){
    my $combo=shift;
    my $spec=shift;
    my $sx;
    my $total=0;
    foreach $sx(@{$combo}){
        $total += $spec->{$sx};
    }
    return $total;
}

sub newgenes(){
    my $data=shift;
    my $newgenome=shift;
    my $i;
    my $newGnum=0;
    for($i=0;$i<$clu_num;$i++){

        my $tmp=0;
        my $j;
        for($j=0;$j<= $#{$data};$j++){
            $tmp+=$data->[$j]->[$i];
        }

        if($newgenome->[$i] > 0  and $tmp <=0){

            $newGnum+=$newgenome->[$i];
        }
```

```
    }
    return $newGnum;

}

sub core(){
    my $d=shift;
    my $c_size=0;
    my $i;
    for($i=0;$i<$clu_num;$i++){

        my @tmp=();
        my $j;
        for($j=0;$j<=$#{$d};$j++){
            push(@tmp,$d->[$j]->[$i]);

        }
        $c_size+=c(@tmp);
    }
    return $c_size;
}

sub coreFix(){
    my $d=shift;
    my $c_size=0;
    my $i;


    for($i=0;$i<$clu_num;$i++){
        my @tmp=();
        my $j=0;
        for($j=0;$j<= $#{$d};$j++){
            push(@tmp,$d->[$j]->[$i]);

        }
        my @tmp2=();
        for($j=0;$j<$g_number;$j++){
            push(@tmp2,$data[$j]->[$i]);
```

```
        }

        $c_size+=cFix(\@tmp,\@tmp2);


    }
    return $c_size;
}


sub pan(){
    my $d=shift;
    my $p_size=0;
    my $i;
    for($i=0;$i<$clu_num;$i++){
        my  @tmp=();
        my $j;
        for($j=0;$j<@{$d};$j++){
            push(@tmp,$d->[$j]->[$i]);
        }
        $p_size+=p(@tmp);
    }
    return $p_size;
}

sub cFix(){
    my $a=shift;
    my $b=shift;
    my $c=$a->[0];
    my $sx;
    foreach $sx (@{$a}){
        if($sx<$c){$c=$sx;}
    }

    if($c > 0){return $c;}else{
        if(&zeroNum($b) > 1){return 0;}else{

            if($b->[0]*$b->[1]*$b->[2]*$b->[3] >0){
```

```
                return 1; #For Complete genomes.
            }else{return 0;}
        }

    }
}

sub zeroNum(){
    my $arr=shift;
    my $a;
    my $n;
    foreach $a (@{$arr}){
        if($a <= 0){
            $n++;
        }
    }
    return $n;
}

sub c(){
    my $a=shift;
    my $b=$a;
    while($b ne ""){
        if($b<$a){
            $a=$b;
        }
        $b=shift;
    }
    if(defined($a)){
        return $a;}else{
        return 0;
    }
}

sub p(){
    my $a=shift;
    my $b=shift;
    while($b ne ""){
        if($b>$a){$a=$b}
```

```
        $b=shift;
    }
    return $a;
}
sub randCnm(){
    my $n=shift;
    my $m=shift;
    my @arr1=(0..($m-1));
    my $i;
    my $rn;
    my %hs=();
    my @arr2=();
    for($i=0;$i<$n;$i++){
        $rn=int(rand($m-$i));
        @arr1=(sort{$a<=>$b}(@arr1));
        $hs{$arr1[$rn]}=1;
        push(@arr2,($arr1[$rn]));
        $arr1[$rn]=$m+1;
    }
    return  @arr2;
}

sub randConm(){
    my $o=shift;
    my $n=shift;
    my $m=shift;
    my @data;
    my %hs;
    my $i;
    for($i=0;$i<$o;$i++){
        my @arr = &randCnm($n,$m);
        while(defined($hs{join(" ", sort {$a<=>$b} @arr)}))
        { @arr = &randCnm($n,$m); }
        $hs{join(" ",sort {$a<=>$b} @arr)}=1;
        push(@{$data[$i]},@arr);
    }
    return @data;
}
```

```
sub randConmN(){
    my $o=shift;
    my $n=shift;
    my $m=shift;
    my @data;
    my %hs;
    my $i;
    for($i=0;$i<$o;$i++){
        my @arr = &randCnm($n,$m);
        while(defined($hs{join(" ", ($arr[0],sort {$a<=>$b} @arr))}))
        { @arr = &randCnm($n,$m); }
        $hs{join(" ",($arr[0],sort {$a<=>$b} @arr))}=1;
        push(@{$data[$i]},@arr);
    }
    return @data;
}
```

A.2 shared.pl

Script function: a shared script which is required by some of the following scripts.

```
#!/usr/bin/perl
my %codonLmA;
#//Group A
$codonLmA{"AT"} ="A";    $codonLmA{"TA"} ="A";
$codonLmA{"GG"} ="A";    $codonLmA{"CC"} ="A";

#Group G
$codonLmA{"AG"} ="G";    $codonLmA{"TC"} ="G";
$codonLmA{"GA"} ="G";    $codonLmA{"CT"} ="G";

#Group T
$codonLmA{"AC"} ="C";    $codonLmA{"TG"} ="C";
$codonLmA{"GT"} ="C";    $codonLmA{"CA"} ="C";

#Group C
$codonLmA{"AA"} ="T";    $codonLmA{"TT"} ="T";
```

```
$codonLmA{"GC"} ="T";    $codonLmA{"CG"} ="T";

my %codonLmB;
#//Group A
$codonLmB{"AT"} ="C";    $codonLmB{"TA"} ="C";
$codonLmB{"GG"} ="C";    $codonLmB{"CC"} ="C";

#Group G
$codonLmB{"AG"} ="A";    $codonLmB{"TC"} ="A";
$codonLmB{"GA"} ="A";    $codonLmB{"CT"} ="A";

#Group T
$codonLmB{"AC"} ="T";    $codonLmB{"TG"} ="T";
$codonLmB{"GT"} ="T";    $codonLmB{"CA"} ="T";

#Group C
$codonLmB{"AA"} ="G";    $codonLmB{"TT"} ="G";
$codonLmB{"GC"} ="G";    $codonLmB{"CG"} ="G";

my @cm;
push(@cm,\%codonLmA);
push(@cm,\%codonLmB);

my @bin2AT;
$bin2AT[0]  = "AT"; $bin2AT[1]  = "AG";
$bin2AT[2]  = "AC"; $bin2AT[3]  = "AA";

$bin2AT[4]  = "TA"; $bin2AT[5]  = "TC";
$bin2AT[6]  = "TG"; $bin2AT[7]  = "TT";

$bin2AT[8]  = "GG"; $bin2AT[9]  = "GA";
$bin2AT[10] = "GT"; $bin2AT[11] = "GC";

$bin2AT[12] = "CC"; $bin2AT[13] = "CT";
$bin2AT[14] = "CA"; $bin2AT[15] = "CG";

my %AT2bin;

$AT2bin{"AT"} =  0; $AT2bin{"AG"} =  1;
```

```
$AT2bin{"AC"} =  2; $AT2bin{"AA"} =  3;

$AT2bin{"TA"} =  4; $AT2bin{"TC"} =  5;
$AT2bin{"TG"} =  6; $AT2bin{"TT"} =  7;

$AT2bin{"GG"} =  8; $AT2bin{"GA"} =  9;
$AT2bin{"GT"} = 10; $AT2bin{"GC"} = 11;

$AT2bin{"CC"} = 12; $AT2bin{"CT"} = 13;
$AT2bin{"CA"} = 14; $AT2bin{"CG"} = 15;
```

A.3 bin2DNA.pl

Script function: transformation of any binary file into a DNA string in a text file.
Usage: bin2DNA.pl a-binary-file > DNA-String-file

```
#!/usr/bin/perl
require ("shared.pl");

open(IN,shift) or die "cannot open input file!\n";

binmode IN;
my $buf;
my $DNAstr="";
my $codType=0;
my $mode=0; #encoding mode, default 0;
while (read(IN, $buf, 1)) {
    my @x = unpack('B8', $buf);
    foreach $sx (@x){
        my $rr=oct("0b".$sx);
        my $rrA=int($rr/16);
        my $rrB=$rr%16;
        $DNAstr=$DNAstr . $bin2AT[$rrA];
        $DNAstr= $DNAstr . $cm[$mode]->{$bin2AT[$rrA]};
        if($rrA == 7 && $mode == 0){$mode=1;}else{$mode=0;}
        $DNAstr=$DNAstr . $bin2AT[$rrB];
        $DNAstr= $DNAstr . $cm[$mode]->{$bin2AT[$rrB]};
```

```
            if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}
        }
        if(length($DNAstr)>= 90){
            print $DNAstr. "\n";
            $DNAstr="";
        }
}

print $DNAstr. "\n" if($DNAstr);
$DNAstr="";

sub formatStr()
{
    my $str=shift;
    my $char=shift or $char="\n";
    my $num=shift or $num=100;
    $char or $char="\n";
    $num or $num=100;
    my $i=0;
    my $tmpStr="";
    while(($i+1)*$num<length($str))
    {
        $tmpStr.=substr($str,$i*$num,$num);
        $tmpStr.="$char";
    }
    return $tmpStr;
}

sub optSecondStructure(){
    my $str=shift;
    my $kmer_len=15;
    my $str_len=length($str);
}

sub complementSeq(){
    my $str=shift;
    $str=~tr/ATGC/TACG/;
    $str=reverse($str);
    return $str;
```

```
}
```

A.4 Consensus.pl

Script function: generation of the consensus DNA string from various numbers of DNA strings with base errors.

Usage: Consensus.pl input-sequences-file > consensus.seq

```
#!/usr/bin/perl
require ("shared.pl");

my @seqs;
my $seqLength=0;
$/=">";
$file=shift or die("consensus.pl\tcannot open the input file!\n");

while($file){
    open(IN,$file) or die("cannot open the input file $file!\n");
    while(<IN>){
        chomp;
        if($_){
            if(/[^ATGCatgc\n]/){
                my @arr1=split(/\n/,$_,2);
                $seq=$arr1[1];
                $seq=~s/\n//g;
            }else{
                $seq=$_;
            }
            $seq=~s/[\n\s\a]//g;
            $seq=uc($seq);
            push(@seqs,$seq);
            if(length($seq) > $seqLength){$seqLength = length($seq);}
        }
    }
    $file=shift;
}

my $pos=0;
```

```
my $consensus="";
my $encodeRule=0;
while($pos<$seqLength){
    my %blocks=();
    foreach $sItem (@seqs){
        my $aBlock=substr($sItem,$pos,3);
        if(cMatch($aBlock,$encodeRule)){$blocks{$aBlock}++;}
    }
    my $fBlock = cBlock(\%blocks);
    $consensus = $consensus . $fBlock;
    if($fBlock eq "TTT"){$encodeRule =1}else{$encodeRule=0;}
    $pos=$pos+3;
}

print ">consensus\n$consensus\n";

sub cBase{
    my $sta=shift;
    my $base="-";
    foreach $ss (keys %{$sta} ){
        $sta->{$ss}++;
        if($ss ne "N"){
            if($sta->{$ss} > $sta->{$base}){
                $base = $ss;
            }
        }
    }
    return $base;
}
```

A.5 DNA2bin.pl

Script function: transformation of a DNA consensus string to a binary file.
Usage: DNA2bin.pl consensus.seq binary-file

```
open(IN,shift) or die "cannot open input file!\n";
```

```
open(OUT,">".shift)||die("cannot open output!\n");

binmode OUT;

while (<IN>) {
    chomp;
    if($_){
        my $DNAstr=$_;
        $DNAstr=~s/[\n\s\a]//g;
        my $pos=0;
        while($pos<length($DNAstr)){
            my $twoBases=substr($DNAstr,$pos,6);
            my $twoBases1=substr($DNAstr,$pos,2);
            my $twoBases2=substr($DNAstr,$pos+3,2);

            my $octNum1=$AT2bin{$twoBases1};
            my $octNum2=$AT2bin{$twoBases2};


            my $octNum=$octNum1*16+$octNum2;


            my $binNum=sprintf( "%b", $octNum);
            my $binNum= "0" x (8- length($binNum)) . $binNum;
            my $binNumPack=pack("C",$octNum);
            my $binNumUPack=unpack('C', $binNumPack);

            syswrite(OUT,$binNumPack,1);
            $pos+=6;
        }

    }
}

sub dec2bin()#10to2
{
    my $dec = shift;
    my $bin = unpack("b4", pack("i",$dec));
    return $bin;
```

```
}
```

A.6 kmerAnalysis.pl

Script function: Analysis of complementary matched k-mers in DNA strings.
Usage: kmerAnalysis.pl dna-sequences.fasta > output

```
#!/usr/bin/perl

open(IN,shift) or die("cannot open the input file!\n");

my $kmLength=shift or die("Please specific the ");
my $seqLength=0;
my $lastSeq="";
my %kmers;

warn "Hashing the DNA Kmers\n";
while(<IN>){
    chomp;
    if($_){
        my $seq=$_;
        $seq=~s/[\n\s\a]//g;
        $seq=uc($seq);
        $seq=$lastSeq . $seq;
        my $seqLength =length($seq);
        my $pos=0;
        while($pos<$seqLength-$kmLength){
            my $kmer=substr($seq,$pos,$kmLength);
            my $octKmer=DNA2Oct($kmer);
            $kmers{$octKmer}++;
            $pos++;
        }
        $lastSeq= substr($seq,length($seq)-$kmLength,$kmLength-1);
    }
}

warn "Hash finished, working on the statics!\n";
```

```
my $km;
foreach $km (keys %kmers){
    if(exists($kmers{$km})){
        my $cpSeqNum=complementNum($km);
        if(exists($kmers{$cpSeqNum})){
            print oct2DNA($km) . "\t";
            print $kmers{$km};
            print "\t" . oct2DNA($cpSeqNum) . "\t";
            print $kmers{$cpSeqNum};
            print "\n";
            delete($kmers{$cpSeqNum});
        }
    }
    delete($kmers{$km});
}

sub complementSeq{
    my $str=shift;
    $str=~tr/ATGC/TACG/;
    $str=reverse($str);
    return $str;
}

sub DNA2Oct{
    my $DNAstr=shift;
    my %AT2bin;

    $AT2bin{"AT"} =  0; $AT2bin{"AG"} = 1;
    $AT2bin{"AC"} =  2; $AT2bin{"AA"} = 3;

    $AT2bin{"TA"} =  4; $AT2bin{"TC"} = 5;
    $AT2bin{"TG"} =  6; $AT2bin{"TT"} = 7;

    $AT2bin{"GG"} =  8; $AT2bin{"GA"} = 9;
    $AT2bin{"GT"} = 10; $AT2bin{"GC"} = 11;

    $AT2bin{"CC"} = 12; $AT2bin{"CT"} = 13;
    $AT2bin{"CA"} = 14; $AT2bin{"CG"} = 15;
```

```
    my $pos=0;
    my $fOct=0;
    while($pos<length($DNAstr)-1){
        my $twoBases=substr($DNAstr,$pos,2);

        my $octNum=$AT2bin{$twoBases};
        $fOct=$fOct*16+$octNum;
        $pos+=2;
    }
    return $fOct;
}


sub complementNum{
    my $num=shift;
    my $seq=oct2DNA($num);
    my $cSeq=complementSeq($seq);
    return DNA2Oct($cSeq);
}

sub oct2DNA{
    my $num=shift;
    my $DNA="";
    my @bin2AT;

    $bin2AT[0]  = "AT"; $bin2AT[1]  = "AG";
    $bin2AT[2]  = "AC"; $bin2AT[3]  = "AA";

    $bin2AT[4]  = "TA"; $bin2AT[5]  = "TC";
    $bin2AT[6]  = "TG"; $bin2AT[7]  = "TT";

    $bin2AT[8]  = "GG"; $bin2AT[9]  = "GA";
    $bin2AT[10] = "GT"; $bin2AT[11] = "GC";

    $bin2AT[12] = "CC"; $bin2AT[13] = "CT";
    $bin2AT[14] = "CA"; $bin2AT[15] = "CG";

    while($num > 15){
        $a=$num%16;
```

```
        $num=int($num/16);
        $DNA= $bin2AT[$a] . $DNA ;
    }
    $DNA= $bin2AT[$num] . $DNA ;
    $DNA= ("AT" x int(($kmLength - length($DNA))/2) ) . $DNA;
    return $DNA;

}
```

A.7 biologyRelevanceAnalysis.pl

Script function: Searching and analysis of sub-sequences that matching SED3B encoding rules in natural generated DNA sequences. The results are used to evaluate the biological relevance of SED3B encoded sequences.

Usage: biologyRelevanceAnalysis.pl dna-sequences.fasta > output

```
#!/usr/bin/perl
require ("shared.pl");

open(IN,shift)||die("not opened!\n");
$/="\n>";

$mSeqLength = shift;

while(<IN>){
    chomp;
    my @arr=split(/\n/,$_,2);
    my @arr2=split(/\s+/,$arr[0],2);
    my $seq=$arr[1];
    $seq=~s/[\n\s\a]//g;
    $seqf1=uc($seq);
    $seqf2=substr($seqf1,1,length($seqf1)-1);
    $seqf3=substr($seqf1,2,length($seqf1)-2);
    bSafeSeq($seqf1,"$arr2[0]_Frame1_");
    bSafeSeq($seqf2,"$arr2[0]_Frame2_");
    bSafeSeq($seqf3,"$arr2[0]_Frame3_");
}
```

```
sub cMatch{
    my $codon=shift;
    my $mode=shift;
    if($mode == 0){
        if( $codonLmA{substr($codon,0,2)} eq substr($codon,2,1))
        {return 1;}else{return 0;}
    }
    if($mode == 1){
        if( $codonLmB{substr($codon,0,2)} eq substr($codon,2,1))
        {return 1;}else{return 0;}
    }
}

sub bSafeSeq{
    my $seq=uc(shift);
    my $seqName=shift;
    my $pos=0;
    my $safeSeq;
    my $seqNum=0;
    my $encodeMode=0;
    my $mSeqLength=21;

    while($pos<length($seq)){
        my $threeBase=substr($seq,$pos,3);
        if(length($threeBase) == 3){
            if(cMatch($threeBase,$encodeMode)){
                $safeSeq = $safeSeq . $threeBase;
                if($encodeMode == 0 and ($threeBase eq "TTT") )
                {$encodeMode=1;}else{$encodeMode=0;}
            }else{
                if(length($safeSeq)>=$mSeqLength){
                    my $len=length($safeSeq);
                    print ">$seqName$seqNum $len\n$safeSeq\n";
                    $safeSeq ="";
                    $seqNum++;
                    $encodeMode=0;
                }
                $safeSeq ="";
            }
```

```
        }
        $pos+=3;
    }
    my $len=length($safeSeq);
    print ">$seqName$seqNum $len\n$safeSeq\n";
    $safeSeq ="";
}
```

A.8 bin2DNACRCIndex.pl

Script function: encode a digital file into indexed DNA sequences with CRC checking.
Usage: bin2DNACRCIndex.pl input-digital-file [start of index] > output.dna

```
#!/usr/bin/perl
require ("shared.pl");

open(IN,shift) or die "cannot open input file!\n";

binmode IN;
my $buf;
my $DNAstr="";
my $codType=0;
my $mode=0; #encoding mode, default 0;
my $rowLenth=78; #fragment length;
my $pcrF="TGCCTCTTTATCTGT";      #"CAGTACTAACCCTCG";
my $pcrR="CATTTCCGATACACC";      #"CCGCAAGAGTGTCGA";
my $indexStart=shift or $indexStart=0;
my $crcSize=2;
my $lcrSize=10;
my @crc=();

while (read(IN, $buf, 1)) {
    my @x = unpack('B8', $buf);
    foreach $sx (@x){
        my $rr=oct("0b".$sx);
        my $rrA=int($rr/16);
        my $rrB=$rr%16;
```

```
            #Encoding data
            $DNAstr=$DNAstr . $bin2AT[$rrA];
            #Encoding check base
            $DNAstr= $DNAstr . $cm[$mode]->{$bin2AT[$rrA]};

            if($rrA == 7 && $mode == 0){$mode=1;}else{$mode=0;}
            #Encoding data
            $DNAstr=$DNAstr . $bin2AT[$rrB];
            #Encoding check base
            $DNAstr= $DNAstr . $cm[$mode]->{$bin2AT[$rrB]};
            if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}
        }

        if(length($DNAstr)>= $rowLenth){
            push(@crc, $DNAstr);
            print "Encoding\t$indexStart\t$pcrF";
            print num2dna($indexStart);
            print "". $DNAstr;
            print "$pcrR\n";
            $DNAstr="";

            $indexStart++;
            #output crc information
            if(scalar(@crc) >= $crcSize){
                print "     CRC\t$indexStart\t$pcrF";
                print num2dna($indexStart);
                print "". crcString(\@crc);
                print "$pcrR\n";
                @crc=();
                $indexStart++;
            }
        }
    }

    if($DNAstr){
        push(@crc, $DNAstr);
        print "Encoding\t$indexStart\t$pcrF";
        print num2dna($indexStart);
        print "". $DNAstr;
```

```
    print "$pcrR\n";
    $indexStart++;

    print "     CRC\t$indexStart\t$pcrF";
    print num2dna($indexStart);
    print "". crcString(\@crc);
    print "$pcrR\n";
    $DNAstr="";
}

$DNAstr="";

sub crcString()
{
    my $strs=shift;
    my $sNum=scalar(@{$strs});
    my $sLength=length($strs->[0]);
    my $x=0;
    my $y=0;
    my $crcStr="";
    while($x < $sLength){
        my $totalNum=0;
        foreach my $sItem (@{$strs}){
            my $char=substr($sItem, $x, 1 );
            if($char eq "A" or $char eq "a") { $totalNum+=0;}
            if($char eq "T" or $char eq "t") { $totalNum+=1;}
            if($char eq "G" or $char eq "g") { $totalNum+=2;}
            if($char eq "C" or $char eq "c") { $totalNum+=3;}
        }

        $totalNum=   $totalNum%4;
        if($totalNum == 0 ) {$crcStr = $crcStr . "A" }
        if($totalNum == 1 ) {$crcStr = $crcStr . "T" }
        if($totalNum == 2 ) {$crcStr = $crcStr . "G" }
        if($totalNum == 3 ) {$crcStr = $crcStr . "C" }
        $x++;
    }
    return $crcStr;
```

```
}

sub formatStr()
{
    my $str=shift;
    my $char=shift or $char="\n";
    my $num=shift or $num=100;
    $char or $char="\n";
    $num or $num=100;
    my $i=0;
    my $tmpStr="";
    while(($i+1)*$num<length($str))
    {
        $tmpStr.=substr($str,$i*$num,$num);
        $tmpStr.="$char";
    }
    return $tmpStr;

}

sub optSecondStructure(){
    my $str=shift;
    my $kmer_len=15;
    my $str_len=length($str);
}

sub complementSeq(){
    my $str=shift;
    $str=~tr/ATGC/TACG/;
    $str=reverse($str);
    return $str;
}

sub num2dna(){
    my $num=shift;
    my $mode=0;
    my $DNAstr="";
    my $rrA=$num;
    my $rrB=$rrA%16;
```

```
    #Encoding data
    $DNAstr= $bin2AT[$rrB] . $cm[$mode]->{$bin2AT[$rrB]} . $DNAstr;
    #Encoding check base

    if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}

    my $rrA=int($rrA/16);
    my $rrB=$rrA%16;
    #Encoding data
    $DNAstr= $bin2AT[$rrB] . $cm[$mode]->{$bin2AT[$rrB]} . $DNAstr;
    #Encoding check base
    if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}

    my $rrA=int($rrA/16);
    my $rrB=$rrA%16;
    #Encoding data
    $DNAstr= $bin2AT[$rrB] . $cm[$mode]->{$bin2AT[$rrB]} . $DNAstr;
    #Encoding check base
    if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}

    my $rrA=int($rrA/16);
    my $rrB=$rrA%16;
    #Encoding data
    $DNAstr= $bin2AT[$rrB] . $cm[$mode]->{$bin2AT[$rrB]} . $DNAstr;
    #Encoding check base
    if($rrB == 7 && $mode == 0){$mode=1;}else{$mode=0;}
    return $DNAstr;
}
```

Appendix B

Supplement Information

B.1 Sequences of mutacins used for the identification of putative mutacins in 10 mutans streptococci strains.

Sequences of known mutacins as well as mutacin-immunity proteins collected from the NCBI (http://www.ncbi.nlm.nih.gov) and Oralgen (http://www.oralgen.lanl.gov/) databases, as well as by searching for related publications.

>SmbA
MKSNLLKINNVTEMEKNMVTLIKDEDMLAGGSTPACAIGVVGITVAVTGIST
ACTSRCINK
>SmbB
MKEIQKAGLQEELSILMDDANNLEQLTAGIGTTVVNSTFSIVLGNKGYICTV
TVECMRNCSK
>Mutacin-I (Isolated from strain UA140 AND CH43)
MSNTQLLEVLGTETFDVQEDLFAFDTTDTTIVASNDDPDTRFSSLSLCSLGC
TGVKNPSFNSYCC
>Mutacin-II AAC38144.1
MNKLNSNAVVSLNEVSDSELDTILGGNRWWQGVVPTVSYECRMN
>Mut-III (Isolated from strain UA787 1140)
MSNTQLLEVLGTETFDVQEDLFAFDTTDTTIVASNDDPDTRFKSWSLCTPG
CARTGSFNSYCC
>Mutacin-IV SMU.150 nlmA non-lantibiotic mutacin IV A
MDTQAFEQFDVMDSQTLSTVEGGKVSGGEAVAAIGICATASAAIGGLAGA
TLVTPYCVGTWGLIRSH
>Mutacin-IV SMU.151 nlmB non-lantibiotic mutacin IV B

MEWRINTMELNVNNYKSLTNDELSEVFGGDKQAADTFLSAVGGAASGFTY
CASNGVWHPYILAGCAGVGAVGSVVFPH
>Mutacin-V SMU.1914c CipB
MNTQAFEQFNVMDNEALSAVEGGGRGWNCAAGIALGAGQGYMATAGGTAF
LGPYAIGTGAFGAIAGGIGGALNSCG
>SMU.423 possible bacteriocin Kreth et al., 2005, state
that this mutacin-like gene is regulated by the competence
system.
MNTQAFEQFNVMDNEALSTVEGGGMIRCALGTAGSAGLGFVGGMGAGTVT
LPVVGTVSGAALGGWSGAAVGAATF
>Mutacin-AII (S. pyo FF22, homolog found in S.mutans)
MEKNNEVINSIQEVSLEELDQIIGAGKNGVFKTISHECHLNTWA

Lebenslauf

Name	Song
Vorname	Lifu
Geburtsdatum	12.Feb.1982
Geburtsort, -land	Shandong, China
08.1989 - 07.1994	Grundschule in Weifang, Shandong/China
08.1994 - 07.1997	Mittelschule in Weifang, Shandong/China
08.1997 - 07.2000	Oberschule in Weifang, Shandong/China
08.2000 - 07.2004	Studium Pharmaceutical Engineering an der Shandong University, Shandong/China Abschluss: Bachelor
08.2004 - 07.2007	Studium Fermentation Engineering an der Shandong University, Shandong/China Abschluss: Master of Engineering
08.2007 - 08.2008	Mitarbeiter in der Arbeitung Bioinformatik am Beijing Genomics Institute (BGI), Beijing/China
09.2008 - 05.2010	Wissenschaftlicher Mitarbeiter am Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin/China
06.2010 - 07.2011	Wissenschaftlicher Mitarbeiter Im Institut für Bioprozess– und Biosystemtechnik, Technische Universität Hamburg–Harburg, Hamburg, Deutschland
08.2011 - 09.2012	Doktorand an der Tianjin University, Tianjin/China
10.2012 - 12.2016	Promotion Institut für Bioprozess– und Biosystemtechnik, Technische Universität Hamburg–Harburg, Hamburg, Deutschland
01.2017 – jetzt	Zusammenschreiben der Doktorarbeit und Arbeitssuchend

Publications

Song, Lifu; Zeng, An-Ping (2017): Engineering "cell robots" for parallel and highly sensitive screening of biomolecules under *in vivo* conditions. In: *Scientific reports* 7 (1), p. 15145. DOI: 10.1038/s41598-017-15621-0.

Song, Lifu; Zeng, An-Ping (2017): Orthogonal information encoding in living cells with high error-tolerance, safety, and fidelity. In: *ACS synthetic biology* 7 (3), pp. 866–874. DOI: 10.1021/acssynbio.7b00382.

Song, Lifu; Sudhakar, Padhmanand; Wang, Wei; Conrads, Georg; Brock, Anke; Sun, Jibin et al. (2012): A genome-wide study of two-component signal transduction systems in eight newly sequenced mutans streptococci strains. In: *BMC Genomics* 13, S. 128. DOI: 10.1186/1471-2164-13-128.

Song, Lifu; Wang, Wei; Conrads, Georg; Rheinberg, Anke; Sztajer, Helena; Reck, Michael et al. (2013): Genetic variability of mutans streptococci revealed by wide whole-genome sequencing. In: *BMC Genomics* 14, S. 430. DOI: 10.1186/1471-2164-14-430.

Conrads, Georg; Soet, Johannes J. de; Song, Lifu; Henne, Karsten; Sztajer, Helena; Wagner- Dobler, Irene; Zeng, An-Ping (2014): Comparing the cariogenic species *Streptococcus sobrinus* and *S. mutans* on whole genome level. In: *Journal of Oral Microbiology* 6, S. 26189. DOI: 10.3402/jom.v6.26189. **Walkhoff prize of the 'Deutsche Gesellschaft für Zahnerhaltung' (DGZ)**.

Rappert, Sugima; Song, Lifu; Sabra, Wael; Wang, Wei; Zeng, An-Ping (2013): Draft Genome Sequence of Type Strain *Clostridium pasteurianum* DSM 525 (ATCC 6013), a Promising Producer of Chemicals and Fuels. In: *Genome Announcements* 1 (1). DOI: 10.1128/genomeA.00232-12.

Poster & Presentations

Song, Lifu and Zeng, An-Ping: Rein the computational abilities of cells to make predictions by letting them 'listen' and 'talk' to us, Scale-up and scale-down of

bioprocesses, 12.May.2015, Ramada Hotel Hamburg-Bergedorf **(Best Poster Award 2015)**

Song, Lifu and Zeng, An-Ping: A Novel, Ultra-Sensitive Multiple Input-Output System for Target Identification in Systems Metabolic Engineering of *E. coli*, Metabolic Engineering 11, 28 June 2016, Japan **(Selected as Rapid Fire Poster)**

Song, Lifu and Zeng, An-Ping: Digital information storage in DNA (In Chinese), The 8th China Summit Forum on Industrial Biotechnology Development, 11-12 Dec 2015, Tianjin China

Patent applications

Song, Lifu and Zeng, An-Ping (2017): Methods for encoding and decoding a binary string and system, PCT/EP2016/078122