

Léa Chuzel

**Application of functional
metagenomics to the field
of glycobiology**



Application of functional metagenomics to the field of glycobiology

Dissertation
zur Erlangung des akademischen Grades

**Doctor rerum naturalium
(Dr. rer. nat.)**

von Dipl. Biotech. Léa Chuzel
geboren am 15. November 1993 in Lyon (Frankreich)

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik der Otto-
von-Guericke-Universität Magdeburg

Promotionskommission:

Prof. Dr. rer. nat. habil. Weiß (Vorsitz)
Prof. Dr.-Ing. Udo Reichl (Gutachter)
Dr. Erdmann Rapp (Gutachter)
Dr. Christopher H. Taron (Gutachter)

eingereicht am: 30. Juni 2021

Promotionskolloquium am: 1. November 2021

Forschungsberichte aus dem Max-Planck-Institut
für Dynamik komplexer technischer Systeme

Band 55

Léa Chuzel

**Application of functional metagenomics
to the field of glycobiology**

Shaker Verlag
Düren 2022

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Magdeburg, Univ., Diss., 2021

Copyright Shaker Verlag 2022

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-8437-5

ISSN 1439-4804

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren

Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9

Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

Bacteria account for ~15% of the earth's biomass and are the second largest biomass contributor after plants. Bacteria are ubiquitous and have adapted to live in all habitable ecosystems on our planet resulting in a vast diversity estimated to be 1 trillion (10^{12}) different species. The genomes of these microbes represent an extraordinary resource for novel enzyme discovery. The field of metagenomics strives to access these genomes, in particular, those from uncultivable species that have been otherwise out of reach.

Glycobiology is a field of research that addresses the structure, function, and biology of glycans and glycoconjugates. Glycans have wide-ranging importance in both basic biology and pharmaceutical science. Aberrant glycosylation has been implicated in many diseases. Additionally, glycosylation of therapeutic proteins affects their safety and efficacy, and is monitored during drug development and manufacturing. Enzymes acting on glycans are important and can represent therapeutic targets (*e.g.*, influenza A neuraminidase), therapeutic agents (*e.g.*, use of sialidases in cancer immunotherapy), or essential glycoanalytical tools (*e.g.*, PNGase F) that help deconvolute glycan structure. In this thesis, a functional metagenomic workflow was created for discovery of new enzymes that act upon glycans. This was used to answer fundamental and application-driven questions in the field of glycobiology.

The functional metagenomic workflow established in this thesis relies on large-insert metagenomic libraries created in *Escherichia coli*. A collection of almost 100,000 clones was created from diverse ecosystems including extreme environments, and is estimated to contain 3-4 million of environmental genes. Three activity-based screens were executed using libraries in this collection. The first led to identification of a novel exosialidase having a unique catalytic mechanism and new protein structure that defined a new glycoside hydrolase family (GH156). The second screen isolated two sialidases with a preference for a non-human form of sialic acid, a specificity not previously described. Finally, a third screen identified

two enzymatic activities: a sugar-specific sulfatase and a sulfate-dependent hexosaminidase that can act on sulfated glycans, an important chemical modification of *N*-glycans for which well-defined analytical tools (*e.g.*, enzymes acting specifically on sulfated sugars) have yet to be established.

To summarize, this work demonstrates the benefit of using functional metagenomics to identify precise enzyme specificities that can answer questions in the field of glycobiology and address technical challenges in glycoanalytics. It highlights that this method of discovery can yield novel protein families that act on glycans, unusual enzymatic specificities, and needed analytical tools. The workflow employed in this thesis is versatile and can easily be adapted to other enzyme discovery projects.

Zusammenfassung

Mit einer Biomasse von etwa 15% sind Bakterien, nach den Pflanzen, der zweitgrößte Produzent von Biomasse auf der Erde. Bakterien sind ubiquitär und haben sich an das Leben in allen Habitaten auf unserem Planeten angepasst, was zu einer enormen Vielfalt von geschätzten 1 Billion (10^{12}) verschiedenen Arten führt. Die Genome dieser Mikroben stellen eine außergewöhnliche Ressource für die Entdeckung (und Nutzung) neuer bakterieller Enzyme dar. Der Bereich der Metagenomik strebt nach der Erforschung dieser Genome insbesondere von Arten, die nicht kultiviert werden können und sonst unzugänglich wären.

Die Glykobiologie ist ein Forschungsgebiet, das sich mit der Struktur, Funktion und Biologie von Glykanen und Glykokonjugaten befasst. Glykane haben eine weitreichende Bedeutung sowohl in der Grundlagenbiologie als auch in der (bio)pharmazeutischen Forschung und Produktion.

Veränderungen der Glykosylierung werden mit einer Vielzahl an Erkrankungen in Verbindung gebracht. Darüber hinaus beeinflusst die Glykosylierung therapeutischer Proteine hinsichtlich deren Sicherheit und Wirksamkeit, und wird daher während der Arzneimittelentwicklung und -herstellung überwacht. Enzyme, die auf Glykane wirken, sog. Glyko-Enzyme können therapeutische Ziele (z. B. Influenza-A-Neuraminidase) oder therapeutische Wirkstoffe (z. B. Verwendung von Sialidasen in Krebsimmuntherapie) darstellen oder als essenzielle glykoanalytische Werkzeuge (z.B. PNGase F) fungieren, die zur Aufklärung der Glykanstruktur beitragen. In dieser Dissertation wurde ein funktionaler metagenomischer Workflow zur Entdeckung neuer Glyko-Enzyme entwickelt. Dieser Workflow wurde zur Beantwortung grundlegender und anwendungsorientierter Fragen im Bereich der Glykobiologie eingesetzt.

Der in dieser Arbeit etablierte funktionelle Metagenomik-Workflow basiert auf metagenomischen Bibliotheken mit großen Inserts, die in *Escherichia coli* erstellt wurden.

Eine Sammlung von fast 100.000 Klonen wurde aus verschiedenen Ökosystemen einschließlich extremer Umgebungen erstellt und enthält schätzungsweise 3-4 Millionen Umweltgene. Drei aktivitätsbasierte Screenings wurden unter Verwendung von Bibliotheken in dieser Sammlung ausgeführt. Das erste Screening führte zur Identifizierung einer neuen Exosialidase mit einem einzigartigen katalytischen Mechanismus und einer neuen Proteinstruktur, die eine neue Glykosid-Hydrolase Familie (GH156) definierte. Das zweite Screening isolierte zwei Sialidasen mit einer Präferenz für eine nicht-humane Form von Sialinsäure, eine Spezifität, die zuvor nicht beschrieben wurde. Schließlich identifizierte ein drittes Screening zwei enzymatische Aktivitäten: Eine zuckerabhängige Sulfatase und eine sulfatabhängige Hexosaminidase, die auf sulfatierte Glykane einwirken können, eine wichtige chemische Modifikation von N-Glykanen, für die noch genau definierte analytische Werkzeuge (z.B. Enzyme die spezifisch auf sulfatierte Zucker wirken) entwickelt werden müssen.

Zusammenfassend demonstriert diese Arbeit den Nutzen der funktionellen Metagenomik zur Identifizierung präziser Enzymspezifitäten, die Fragestellungen im Bereichen der Glykobiologie beantworten können und technische Herausforderungen der Glykoanalytik angehen. Es unterstreicht, dass diese Entdeckungsmethode neue, auf Glykane wirkende Proteinfamilien, ungewöhnliche enzymatische Spezifitäten und benötigte analytische Werkzeuge hervorbringen kann. Der in dieser Arbeit verwendete Workflow ist vielseitig und kann problemlos für Projekte zur Entdeckung anderer Enzyme angepasst werden.

List of abbreviations	IX
1 Introduction	1
2 Theoretical background	6
2.1 Metagenomics.....	8
2.1.1 Microorganisms: a tremendous diversity previously inaccessible	8
2.1.2 Sequence-based metagenomics	10
2.1.3 Functional metagenomics.....	13
2.1.3.1 Choosing a vector-host system	14
2.1.3.2 Screening strategies	16
2.1.3.3 Application of functional metagenomics.....	18
2.1.4 Contrasting approaches to enzyme discovery	20
2.2 Glycobiology	23
2.2.1 A few definitions	23
2.2.2 Glycan synthesis	23
2.2.3 Tremendous glycan diversity	24
2.2.4 Glycan functions	26
2.2.5 Importance of protein glycosylation in the pharmaceutical industry.....	28
2.2.6 Glycoanalytics	29
3 Development of a functional metagenomic workflow for enzyme discovery	36
3.1 Introduction	38
3.2 Material and methods	39
3.2.1 Environmental DNA extraction.....	39
3.2.1.1 eDNA isolation from terrestrial samples.....	39
3.2.1.2 eDNA isolation from aquatic environments.....	41
3.2.1.3 eDNA isolation from human feces.....	42
3.2.2 Fosmid eDNA library construction in <i>E. coli</i>	43
3.2.2.1 Library principle	43
3.2.2.2 Library assembly procedure.....	46
3.2.3 Library quality assessments	47
3.2.4 High-throughput functional screening	48
3.2.4.1 Agar plate-based enzyme screening	48
3.2.4.2 Lysate-based enzyme screening	48
3.2.5 Fosmid sequencing and bioinformatics	49
3.2.5.1 Single clone sequencing.....	49
3.2.5.2 Multiplexed-sequencing	50
3.2.5.3 Blue Pippin size-selection	51
3.2.5.4 RSII sequencing and <i>de novo</i> assembly.....	51
3.2.5.5 ORF prediction and ORF map drawings.....	52
3.3 Results and discussion	53
3.3.1 Metagenomics libraries and collection	53
3.3.1.1 Environmental DNA	53
3.3.1.2 The NEB Collection in November 2019	55

3.3.1.3 Library quality	62
3.3.2 Plate-based and lysate-based screenings	64
3.3.2.1 Plate-based screening	65
3.3.2.2 Lysate-based screening	67
3.3.2.3 Hit definition	70
3.3.3 Sequencing the hits and generation of fosmid maps	72
3.4 Chapter 3 conclusion	77
4 Screening metagenomic libraries for sialidases	79
4.1 Introduction to sialic acid biology	80
4.2 Discovery of the novel GH156 sialidase family	86
4.2.1 Material and methods	86
4.2.1.1 Screening for sialidases	86
4.2.1.2 Tn5 mutagenesis	86
4.2.1.3 <i>In vitro</i> and <i>in vivo</i> sialidase expression	87
4.2.1.4 Sialidase biochemical characterization	88
4.2.1.5 NMR spectroscopy	90
4.2.1.6 <i>Armatimonadetes</i> homolog expression	91
4.2.2 Results	91
4.2.2.1 Functional screening	91
4.2.2.2 Sialidase gene identification	93
4.2.2.3 Sialidase biochemical characterization	97
4.2.2.4 ORF12p sialidase reaction mechanism	101
4.2.2.5 ORF12p sialidase family phylogeny	104
4.2.3 Discussion	106
4.3 Biostructural characterization of GH156	108
4.3.1 Material and methods	108
4.3.1.1 Selenomethionine protein labeling	108
4.3.1.2 SEC-MALLS	110
4.3.1.3 Crystallization of EnvSia156 and EnvSia156 substrate and inhibitor complexes	110
4.3.1.4 3D structure solution	111
4.3.1.5 Site directed mutagenesis and activity assay of generated mutants	112
4.3.2 Results	113
4.3.2.1 Expression and purification of EnvSia156 SeMet mutant for MAD	113
4.3.2.2 Structure of EnvSia156 defines an unusual sialidase fold ...	116
4.3.2.3 Mechanism of EnvSia156 and definition of its active center	121
4.3.3 Discussion	128
4.4 Identification of unconventional sialidases	129
4.4.1 Material and methods	130
4.4.1.1 Screening for Neu5Gc specific sialidases	130
4.4.1.2 PacBio sequencing and enzyme identification	130
4.4.1.3 Expression and purification of C19 and C22 sialidases	130
4.4.1.4 Determination of sialidase substrate preference	131

4.4.2 Results.....	131
4.4.2.1 Functional metagenomic screening	131
4.4.2.2 Identification of C22 and C19 sialidases	134
4.4.2.3 Neu5Gc preferring sialidases activity.....	135
4.4.3 Discussion	136
4.5 Chapter 4 conclusion	140
5 Applying functional metagenomics to post-glycosylation modifications.....	141
5.1 Introduction to post glycosylation modifications with a focus on sulfation	142
5.2 Material and methods	148
5.2.1 Screening for sulfated glycan using a coupled assay	148
5.2.2 F1-ORF13 and F10-ORF19 hexosaminidase <i>in vivo</i> expression and purification.....	148
5.2.3 F1-ORF13 activity on sulfated monosaccharides	149
5.2.4 Enzyme activities on <i>N</i> -glycans	150
5.2.4.1 <i>N</i> -glycan release and APTS-labeling	150
5.2.4.2 Substrate preparation	150
5.2.4.3 F1-ORF13 activity on <i>N</i> -glycans released from hlgA and human urokinase.....	152
5.2.4.4 <i>N</i> -Glycan enrichment using F1-ORF13.....	153
5.2.4.5 F10-ORF19 activity on <i>N</i> -glycans released from hlgA and human urokinase.....	153
5.2.4.6 xCGE-LIF analysis.....	154
5.3 Results	154
5.3.1 Functional metagenomic screening for sulfatases	154
5.3.2 Analysis of fosmid DNA sequences.....	157
5.3.3 Identifying genes encoding active sulfatases using <i>in vitro</i> protein expression.....	161
5.3.4 Protein sequence analysis of active sulfatases	162
5.3.5 Characterization of F1-ORF13 sulfatase	163
5.3.5.1 Determination of F1-ORF13 sulfatase specificity using sulfated monosaccharides	163
5.3.5.2 F1-ORF13 sulfatase activity on GlcNAc-6-SO ₄ in intact <i>N</i> -glycans.....	164
5.3.5.3 F1-ORF13 binds GlcNAc-6-SO ₄ -containing <i>N</i> -glycans in absence of calcium	167
5.3.6 Identifying genes encoding active hexosaminidases using <i>in vitro</i> protein expression.....	171
5.3.7 Protein sequence analysis of active hexosaminidases	172
5.3.8 F10-ORF19 hexosaminidase activity upon GlcNAc-6-SO ₄ in intact <i>N</i> -glycans.....	174
5.4 Chapter 5 conclusion	177
6 Conclusion and outlook	180
Bibliography.....	186
List of tables and figures	211