Alginate Epimerases: Segmental Labelling, Structures and Ligand Interactions

Buchinger, Edith

Publication date:
2011

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Buchinger, E. (2011). *Alginate Epimerases: Segmental Labelling, Structures and Ligand Interactions* (1 ed.). Institut for Kemi, Miljø og Bioteknologi, Aalborg Universitet.

Preface

This report serves as documentation for the work done under The International Doctoral School of Technology and Science at the Faculty of Engineering and Science, Aalborg University, Denmark.

This thesis is divided into a general introduction, aim of study, extended summary of the results and discussion and finally perspectives. The introduction explains protein ligation and different labelling possibilities. It describes alginates and the extracellular alginate epimerases from *Azotobacter vinelandii* and their effect on the alginates. A short introduction of NMR, ITC and SAXS is also included. This work contains five papers. Paper I and II summarize the experimental data carried out on protein *trans*-ligation with the split intein from *Nostoc punctiforme* and segmental labelling of AlgE4. Paper III and IV describe assignment of two different R-modules of AlgE6. Paper V summarized the structural characterization and substrate binding of the alginate epimerases AlgE4 and AlgE6 and some of their subunits.

This PhD project was based on collaboration between Reinhard Wimmer's group at AAU, Jan Skov Pedersen's group of Interdisciplinary Nanoscience Center, University of Aarhus, Svein Valla's and Gudmund Skjåk-Bræk's group at NTNU in Trondheim, Norway, and Dr. Hideo Iwaï's lab at Helsinki University in Finnland. The genetic work and some protein trans-ligation tests were carried out in Iwaï's lab. Segmental labelling of AlgE4 was performed in Aalborg. The assignments of the two R-modules of AlgE6 and epimerisation tests were done at NTNU. The NMR structure determinations and alginate binding tests of different R-modules were done partly at AAU and partly at NTNU. The structures obtained by SAXS measurements were performed in Aarhus.

Acknowledgement

I would like to thank my supervisor Reinhard Wimmer for excellent and dedicated supervision and for always providing invaluable help during my project.

Finn L. Aachmann thanks for the daily support during my year in Trondheim and for the many discussions and ideas.

I would also like to thank Svein Valla and Gudmund Skjåk-Bræk for providing scientific inputs and inspiration on alginate and alginate epimerases.

I am also thankful to Hideo Iwaï for introducing me to cloning and protein *trans*-ligation.

Jan Skov Pedersen and Manja Behrsen thanks for measuring and analysing SAXS data.

A. Sesilja Aranko for providing some constructs for segmental labelling.

Helga Ertesvåg and Wenche Strand thanks for showing me the alginate epimerisation tests.

Olav Aarstad thanks for kindly provided different alginate sample.

Trygve Andreassen thanks for helping with the NMR equipment at NTNU.

Daniel H. Knudsen thanks for measuring NMR and ITC titration data.

I owe Hanne Krone and Margit Paulsen thank for preparing and measuring some NMR samples.

I would like to thank colleagues and friends at NTNU, AAU and Helsinki University.

And last but not least, a great thanks to my family for their support.

For financial support:

Grant sponsor: Danish Agency for Science, Technology and Innovation;
Grant sponsors: Academy of Finland; Grant numbers: 118385 and 131413;
Grant sponsors: KMB project number 182695/I40 and the Norwegian Research Council;
The NMR laboratory at Aalborg University is supported by the Obel and SparNord foundations.

Abstract

Azotobatcer vinelandii has seven extracellular alginate epimerases which consist only of two domains, designated A- and R-module, in varying order and numbers. The A-modules (approximately 385 amino acids in size) carry the catalytic activity. They epimerise α-D-mannuronic acid moieties in alginate polymers into β-L-guluronic acid by inverting the stereochemical configuration at C5 of the sugar ring. The R-modules (approximately 155 amino acids) do not carry enzymatic activity but enhance the activity of the A-modules, if at least one R-module is bound to the A-module. It was assumed that the R-modules confer substrate specificity to the epimerases by binding alginates and thus supporting the correct orientation of the substrate for the epimerisation site. *Azotobacter vinelandii* produces 34 different R-modules and out of those, the R-modules of AlgE4 and AlgE6 were chosen be investigated in this thesis. Despite their high degree of sequence identity, AlgE4 epimerizes poly-M alginates to alternating MG-blocks while AlgE6 epimerizes poly-M alginates to GG-blocks. The structures of R-module and A-module of AlgE4 were already known and both modules are capable of binding poly-M alginates.

The NMR assignment of AlgE4R was known. In order to be able to selectively investigate the R-module within the intact AlgE4 by NMR, we produced segmentally ^{2}H, ^{15}N labelled AlgE4 isotopomeres (A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R) by protein *trans*-splicing using the naturally split intein of *Nostoc punctiforme*. An expression system was constructed for separate expression of A- and R-modules, together with the necessary parts of the split intein, in a way, that later allows the ligation of A- and R-module to full-length AlgE4. Protein expression and ligation was optimized. The ligating fragments were dimers and had to be reduced before ligation would occur. The effects of different reducing agents at various concentrations, temperature and pH values on the ligation yield were tested for AlgE4 as well as on other model proteins. The reducing agent used has a great influence on the ligation yield and in the case of A- and R-module ligation, tris(2-carboxyethyl)phosphine (TCEP) yielded by far most ligation product. The ligated AlgE4 retained its full native activity and structure. However, the overall yield of ligated AlgE4 was low.

For all three R-modules of AlgE6, near-complete resonance assignments were obtained and the three-dimensional structures were solved by NMR. The structures of all three R-modules of AlgE6 are similar to each other and to the structure of AlgE4R.

Alginate binding properties of the R-modules of AlgE4 and AlgE6 to a range of alginate oligomers of varying composition and length were investigated by NMR and ITC. Binding constants between the R-module of AlgE4 and different alginate oligomers as well as the binding sites on AlgE4R were determined.

It could be shown that the R-module of AlgE4 has a strong preference for poly-M alginates. It binds alternating MG-alginates 100 fold less efficiently than poly-M-alginates with the same degree of polymerisation. The binding constants are also affected by the length of the alginate oligomers, the highest binding constants were measured for five or more saccharide subunits. Chemical shift perturbation data yielded a good indication of the location of the binding site on the surface of the protein.

Surprisingly, none of the single R-modules of AlgE6 is capable of binding any alginate tested. This is puzzling as over 50% of the amino acids are identical and additional 20% are highly similar between the R-modules of AlgE6 and AlgE4.

Low-resolution structures of the whole AlgE4 and AlgE6 were determined from SAXS data. The results showed that the different modules adopt a well-defined orientation relative to each other with limited flexibility in between.

Resume

Azotobacter vinelandii har syv ekstracellulære alginat epimeraser. Disse består alle af kun to domæner, kaldet A- og R-modul, men domænerne optræder i forskellig rækkefølge og antal. A-modulerne (ca. 385 aminosyrer lange) er ansvarlige for den katalytiske aktivitet. De epimeriserer α-D-mannuronsyre i alginat polymerer til β-L-guluronsyre ved at invertere stereokemien på C5 i sukkerringen. R-modulerne (ca. 155 aminosyrer lange) udviser ingen katalytisk aktivitet i sig selv, men de øger aktiviteten af A-modulerne, hvis der er mindst et R-modul bundet til A-modulet. Det var formodet at R-modulerne bidrager til epimerasernes substratspecificitet ved at binde alginat og orientere det således at det aktive sæde kan gennemføre epimeriseringen. *Azotobacter vinelandii* producerer 34 forskellige R-moduler. R-modulerne fra AlgE4 og AlgE6 blev udvalgt for nærmere studier i denne afhandling. Selv om AlgE4 og AlgE6 viser en høj sekvenshomologi, epimeriserer AlgE4 poly-M alginat til MG-blokke, mens AlgE6 epimeriserer poly-M alginat til GG-blokke. Strukturerne af både A- og R-modulet af AlgE4 var allerede kendte som isolerede moduler, og begge to binder til poly-M alginat.

Ligeledes var NMR tilordningen af AlgE4R kendt. For med NMR at kunne undersøge R-modulet som en del af AlgE4, producerede vi AlgE4 isotopomerer med ^{2}H,^{15}N-mærkede moduler (A-[^{2}H-^{15}N]R og [^{2}H-^{15}N]A-R) vha. protein *trans splicing*, hvor vi brugte den naturligt delte intein af *Nostoc punctiforme*. Vi har konstrueret et vektorsystem for udtrykkelse af A- og R-modulerne, hvor de er sat sammen med de respektive dele af det opdelte intein på sådan en måde at man senere kan ligere A- og R-modulerne sammen til AlgeE4. Protein udtrykkelse og ligering blev optimeret. A- og R-modulerne foreligger som dimerer efter udtrykkelsen og skal reduceres før ligeringen kan ske. Vi har undersøgt indflydelsen af forskellige reduktionsmidler ved forskellige koncentrationer, temperaturer og pH værdier på udbytte af ligering af AlgE4 og andre modelproteiner. Reduktionsmidlet viste sig at have stor indflydelse på ligering af AlgE4, og tris(2-carboxyethyl)phosphine (TCEP) gav klart det største udbytte af ligeringen. Den ligerede AlgE4 havde samme aktivitet og struktur som den native form, men udbyttet var lavt.

NMR signalerne af alle tre R-moduler af AlgE6 blev tilordnet og deres tredimensionelle struktur i opløsning blev bestemt. Alle tre strukturer ligner meget hinanden og de ligner strukturen af R-modulet af AlgE4.

Vi har undersøgt bindingen af forskellige alginat-oligomerer med forskellig sammensætning og kædelængde til R-modulerne fra AlgE4 og AlgE6 vha. NMR og ITC, og vi har bestemt bindingskonstanter og bindings sites.

R-modulet af AlgE4 har en klar preference for poly-M alginat. MG-blok alginat bliver bundet mindst 100 gange svagere end poly-M alginat med samme kædelængde. Kædelængden havde også en indflydelse på bindingskonstanten – den stærkeste binding blev målt med kædelængder på fem eller flere sukkerenheder. Pertubationen af de kemiske skift gav en god indikering af placering af bindingsområdet på proteinoverfladen.

Overraskende nok viste det sig at ingen R-moduler fra AlgE6 binder til nogen af de undersøgte alginater, selv om R-modulerne har over 50% sekvens identitet og yderligere 20% meget lignende aminosyrer til fælles med R-modulet af AlgE4.

Strukturer af AlgE4 og AlgE6 i opløsning blev bestemt med lav opløsning ud fra SAXS data. Strukturerne viste at de forskellige moduler har veldefinerede orienteringer i forhold til hinanden og begrænset fleksibilitet.

Papers

Paper I **Aranko AS, Züger S, Buchinger E, Iwaï H (2009) *In vivo* and *in vitro* protein ligation by naturally occurring and engineered split DnaE inteins. PLoS One 4: e5185.**

Paper II **Buchinger E, Aachmann FL, Aranko AS, Valla S, Skjåk-Bræk G, Iwaï H, Wimmer R (2010) Use of protein trans-splicing to produce active and segmentally ^{2}H, ^{15}N labeled mannuronan C5-epimerase AlgE4. Protein Science 19: 1534-1543**

Paper III **Buchinger E, Skjåk-Bræk G, Valla S, Wimmer R, Aachmann FL (2011) NMR assignments of ^{1}H, ^{13}C and ^{15}N resonances of the C-terminal subunit from *Azotobacter vinelandii* mannuronan C5-epimerase 6 (AlgE6R3). Biomolecular NMR assignments 5:27-29**

Paper IV **Andreassen T, Buchinger E, Skjåk-Bræk G, Valla S, Aachmann FL (2010) ^{1}H, ^{13}C and ^{15}N resonances of the AlgE62 subunit from *Azotobacter vinelandii* mannuronan C5-epimerase. Biomolecular NMR Assignments in press**

Paper V **Buchinger E, Knudsen DH, Behrsen MA, Pedersen JS, Valla S, Skjåk-Bræk G, Wimmer R, Aachmann FL; Structural and Functional Characterization of the R-modules in Alginate C-5 Epimerase AlgE4 and AlgE6 from *Azotobacter vinelandii*; Manuscript in preparation**

Abbreviation

ABC-transporter	ADP-binding cassette
B_0	external magnetic field
BPTI	bovine pancreatic trypsin inhibitor
casAA	casamino acid
CBD	chitin binding domain
COSY	correlation spectroscopy
CPMG	Carr-Purcell-Meiboom-Gill
CPS	conditional protein splicing
CSA	chemical shift anisotropy
DD-coupling	dipole-dipole coupling
D_{max}	maximal diameter
DP_n	degree of polymerisation
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EEL	expressed enzymatic ligation
EPL	expressed protein ligation
FID	free induction decay
G	α-L-guluronic acid
GB1	B1 domain of protein G
GDL	D-glucono-δ-lactone
GDP-mannose	guanosindiphosphate-mannose
	Triphosphatase
GFP	green fluorescence protein
GG-block	regions consisting of β-L-guluronic acids
HINT motif	Hedgehog and INTein motif
HSQC	heteronuclear single quantum coherence
I(0)	forward scattering; scattering at $q = 0$
INEPT	insensitive nuclei enhanced by polarisation transfer
Int_C	C-terminal domain of a split intein
Int_N	N-terminal domain of a split intein
IPL	intein mediated protein ligation
ITC	isothermal titration calorimetry
$J(\omega)$	spectral density
K_d	dissociation constant
M	β-D-mannuronic acid
M_0	macroscopic magnetisation
MBP	maltose binding protein
MESNA	sodium salt of mercaptoethylsulfonat
MFP	membrane fusion protein
MG-block	region consisting of β-L-guluronic acid and α-D-mannuronic acid in alternating sequence
MM-block	poly- mannuronan alginate
MOPS	3(N-morpholino)propanesulfonic acid
MSG	malate synthase G
MSH	O-mesitylene-sulfonylhydroxylamine

Mtu	*Mycobacterium tuberculosis*
NCL	native chemical ligation
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
Npu	*Nostoc punctiforme*
n-SH3	N-terminal Src homology 3
OMP	outer membrane protein
p(r)	distance distribution function
RDC	residual dipolar coupling
R_g	radius of gyration
RTX-motif	repeat in toxin motif
SAIL	stereo-array isotope label
SAXS	small angle X-ray scattering
SDS-PAGE	sodium dodecyl sulfate polyacryl-gel electrophoresis
SEP	synthetic erythropoiesis protein
Ssp	*Synechocystis species*
T_1	longitudinal relaxation
T_2	transverse relaxation
TCEP	tris(2-carboxyethyl)phosphine
TISS	type I secretion system
TOCSY	total correlation spectroscopy
TRIS	tris(hydroxymethyl)-aminomethane
TROSY	transverse relaxation optimized spectroscopy
VMA-1	vacuolar membrane H^+-Adenosine
W_0, W_1 and W_2	zero- , single- and double quantum transition

Contents

1 Introduction

1.1 *Protein ligation*

Protein ligation describes every process that splices two peptides via a stable backbone bond post-translationally. The ligation partners can be either recombinant proteins or chemical synthesized peptides.

Ligation of two fully protected peptides often fails due to the growing insolubility of the extended peptide. Therefore a successful ligation mechanism should ligate unprotected peptides with high efficiency and without racemisation or other side reactions. Another request is that ligated proteins should be as similar as possible to the natural proteins and in some cases ligated proteins are indistinguishable from natural proteins.

Protein ligation can be purely chemical or a protein assisted reaction. The protein assisted reactions are subdivided into intein-mediated ligation and enzymatic ligation. Inteins are proteins that excise themselves and ligation the flanking protein sequences via peptide bonds (see Fig. 14). Enzymatic ligation is based on proteases (often mutated) that perform the protein ligation.

The ligation procedures can be divided into 3 different subgroups:

- Chemical ligations
- Intein-mediated ligation
- Expressed enzymatic ligation (EEL)

1.1.1 Application

The applications for protein ligation are diverse. One application is the ligation of synthesized peptides. Despite all optimization protein synthesis is limited to peptides below 50 amino acids as the amount of wrong synthesized peptide successes the desired one. One remarkable example is the synthesis of an all-D chiral form of the HIV-1 protease (100 residues) [1] or the preparation of the post-translationally modified variant of erythropoietin [2] (see Fig. 6).

Semisynthesis described the ligation of a recombinant and synthesized peptide. Semisynthesis is frequently used to incorporate non-natural amino acids into proteins like fluorophores.

Cyclization of proteins or peptides is used to reduce their flexibility and improve their stability *in vivo*. Protein ligation is an elegant method to obtain circular peptides (Fig. 19).

For NMR measurements protein ligation gives the opportunity to focus on one segment of a protein. This method is called segmental labelling (see Fig. 21 and also Fig. 72). *In vivo* protein-protein interaction can be also verified by protein ligation. For that purpose luciferase or green fluorescent protein (GFP) are split and only if the two protein of interest interacted with each other the fragments of the split luciferase (or GFP) are ligated (Fig. 22).

1.1.2 Chemical protein ligation

Chemical ligation is not limited to chemically synthesized proteins but also natural peptides, where at least one functional group necessary for ligation was introduced chemically, fall into this category.

1.1.2.1 Chemical protein ligation resulting in artificial backbone

Several reaction mechanisms were suggested leading to ligated peptides with artificial peptide bond analogues on the ligation site. (Tab. 1) (Fig. 1)

Tab. 1: Overview of the different ligation reaction resulting in artificial peptide analogous.

Ligation	C-terminal functional group	N-terminal functional group	Ligating bond	Ref
a) Thioester	Thioester	Bromacetyl	Thioester	[3]
b) Thioether	Thiol	Bromacetyl	Thioether	[4,5]
c) Oxime	Aldehyd	(aminooxy)acetyl	Hydrazone	[6,7]
d) Thiazolidine	Alkylaldehyd; periodate oxidation	Cys, Ser, Thr	Pseudoproline	[8,9]
e)3 + 2 cycloaddition	Alkyne	Azide	Triazole	[10,11]

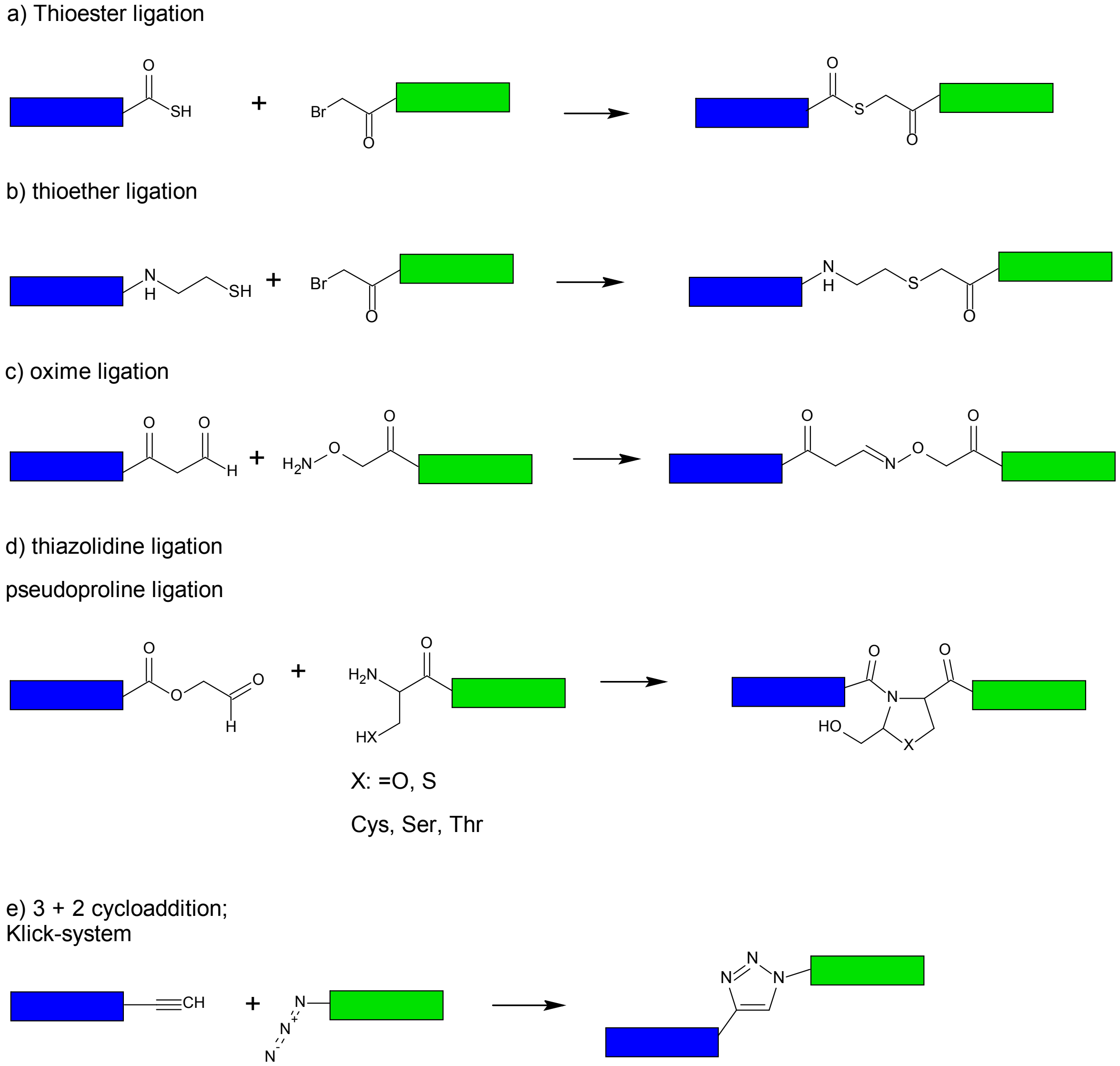


Fig. 1: Summary of different chemical ligation resulting in artificial peptide bone analogues. Those artificial proteins show similar functionality as the wild type proteins but none of the ligation method is used any more.

The maybe most prominent example of ligated peptides with artificial backbone was the 100 amino acids big, fully active HIV-protease analogue obtained by thioester ligation [3] and later by thioether ligation [4].
The artificial back bone does not automatically hinder the functionality of the proteins however ligations resulting in amide bond have outperformed all others.

1.1.2.2 Chemical protein ligation resulting in peptide backbone

1.1.2.2.1 Native chemical ligation

In 1953 Wieland et al. [12] described the reaction of S-phenyl-valine and Cys resulting in the dipeptide ValCys. (Fig. 2)

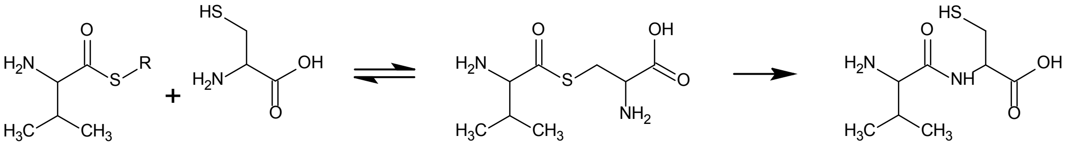


Fig. 2: The reaction of the thioester of valine with a cysteine resulting in the dipeptide ValCys. As the reaction is irreversible the ligation occurs nearly quantitative.

It took more than 4 decades to realize the potential of this reaction. Kent’s lab developed a ligation method now known as native chemical ligation (NCL) [1]. NCL is based on the reaction of a C-terminal thioester with an N-terminal cysteine resulting in a ligated protein with a peptide bond and the cysteine remaining in the sequence (Fig. 3).

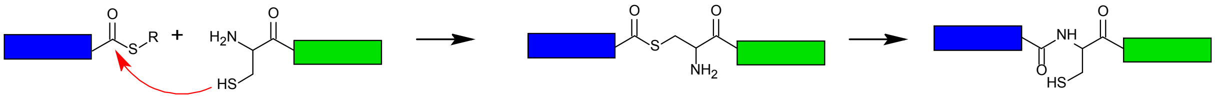


Fig. 3: Reaction mechanism of NCL. The side chain of the cysteine attacks the thioester resulting in ligation product via peptide bond.

The last reaction step is irreversible and therefore gives nearly quantitative yields of the ligated product. The great advantage of NCL in comparison to other methods is that nearly all amino acids except proline can be used as C-terminal amino acid containing the thioester [13]. The disadvantage of this reaction is considered to be the essential cysteine for ligation. Only approximately 2% of all amino acids in proteins are cysteins. Other amino acids containing hydroxyl- or amine group like serine, threonine and even histidine and tryptophan can be used instead but with lower efficiency [14]. More exotic amino acids like selenocysteine and homocysteine are even more efficient than cysteine [15]. Selenocysteine is a seldom but natural amino acid and Gieselman suggest to use NCL for obtaining selenocysteine containing peptides [16]. If selenocysteine are only used for ligation but unwanted in the product it can be used for other chemo selective reactions [15]. Homocysteine can be methylated afterwards to methionine by methyl- *p*-nitrobenzenesulfonate [17]. (Fig. 4)

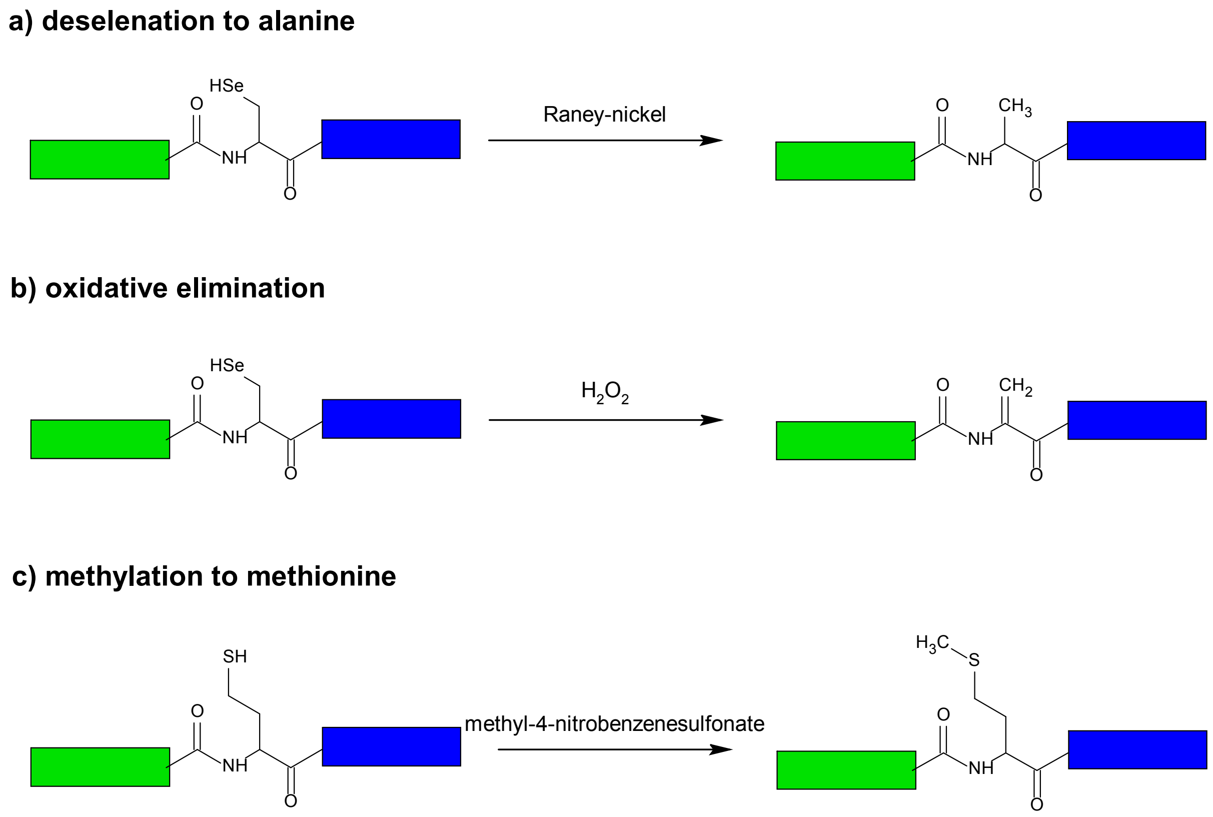

a) deselenation to alanine
HSe
O
NH
O
Raney-nickel
O
CH3
NH
O
b) oxidative elimination
HSe
O
NH
O
H2O2
O
CH2
NH
O
c) methylation to methionine
SH
O
NH
O
methyl-4-nitrobenzenesulfonate
H3C
S
O
NH
O

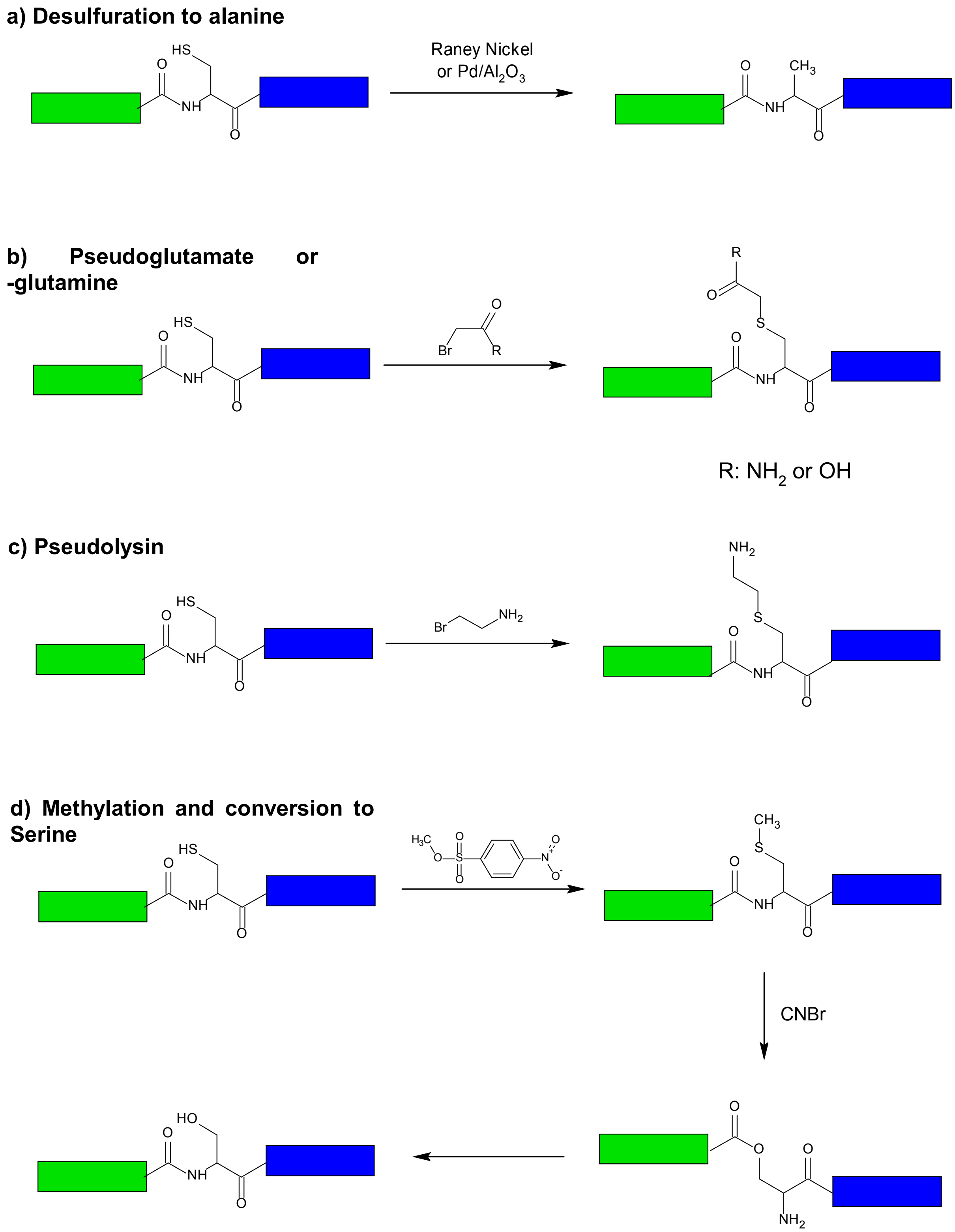
Fig. 4: Selenocysteine and homocysteine ligation yields in higher product than the N-terminal cysteine. If selenocysteine is unwanted it can be deselenized to alanine or dehydro-alanine. Dehydro-alanine can be used for further protein modifications (side chain tags). Homocysteine can be methylated to methionine.

Alkylation transforms cysteine into pseudoglutamate, pseudoglutamine [2,18] or pseudolysin [19]. These nonnatural amino acids mimic in high extend the natural amino acids. Cysteine can also be converted in serine in high yields [20]. (Fig. 5)

For any postligational modification the effect on other functional groups must be considered. Any additional cysteine must be protected (by acetamidomethyl-group). Due to the nucleophilic properties of thiol and selenol the alkylation or thiolalkylation reactions are specific and protection of other amino acids is not necessary.

Desulfuration or deselenation to alanine by Raney nickel or palladium-mediated is frequently used. Nevertheless several side reactions are known mainly the desulfuration of thioether in Met and epimerization of secondary alcohols. Additionally the harsh reaction conditions can cause aggregation.

Oxidative elimination of Cys or Sec leads to dehydro-alanine. Davis reported an oxidative elimination of a cysteine by O-mesitylene-sulfonylhydroxylamine (MSH) into dehydro-Ala [21]. The reaction was fast and tolerates various functional groups also Met. Sec can be oxidative eliminate by H_2O_2 [22,23]. The resulting dehydro-Ala derivates allows further modification [21,22].



a) Desulfuration to alanine

b) Pseudoglutamate or -glutamine

c) Pseudolysin

d) Methylation and conversion to Serine

Fig. 5: Post-ligation modification. If the cysteine is only used for protein ligation it can be modified to other amino acids like alanine and serine or to artificial amino acids like pseudoglutamine,-glutamate and-lysine.

1.1.2.2.1.1 Application of NCL

Many applications were successfully performed by NCL [1]. One of the most advanced syntheses ever performed was the production of the artificial erythropoietin protein. A four-segment ligation in combination with various additional modification processes was used to obtain an analogue of the glycoprotein hormone erythropoietin (Fig. 6).

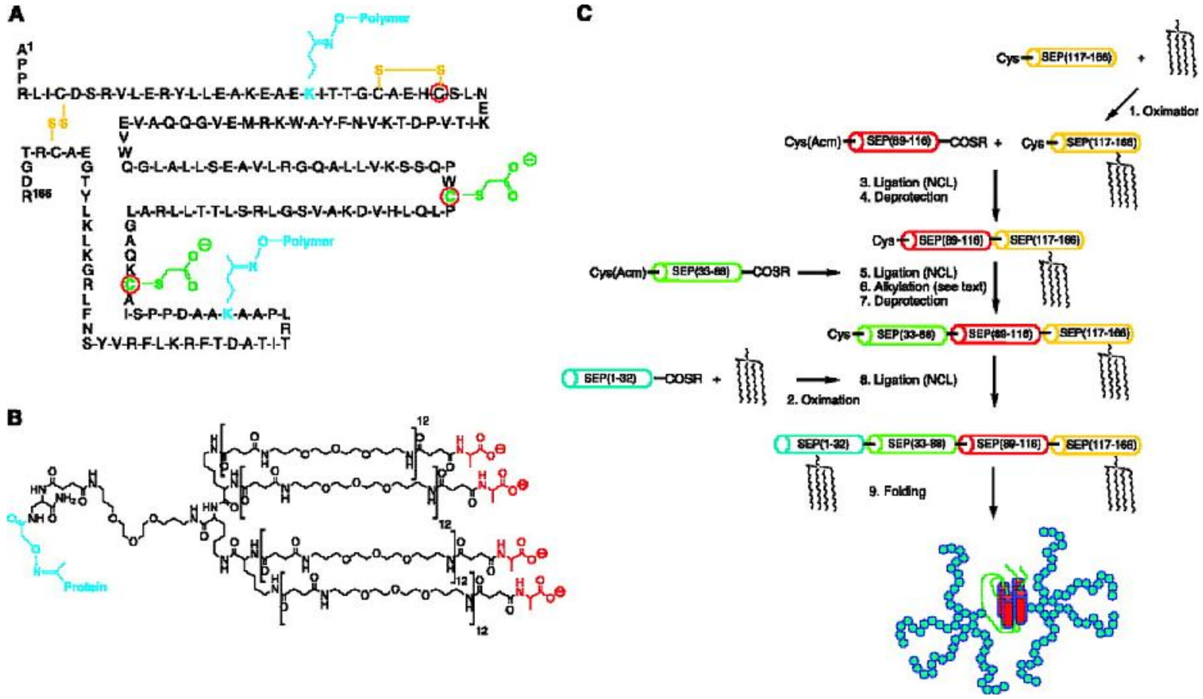

A
A1
P
P
R-L-I-C-D-S-R-V-L-E-R-Y-L-L-E-A-K-E-A-E-K-I-T-T-G-C-A-E-H-C-S-L-N
E
E-V-A-Q-Q-G-V-E-M-R-K-W-A-Y-F-N-V-K-T-D-P-V-T-I-K
V
W
Q-G-L-A-L-L-S-E-A-V-L-R-G-Q-A-L-L-V-K-S-S-Q-P
W
C
L-A-R-L-L-T-T-L-S-R-L-G-S-V-A-K-D-V-H-L-Q-L-P
G
A
Q
K
C
A
I-S-P-P-D-A-A-K-A-A-P-L
R
S-Y-V-R-F-L-K-R-F-T-D-A-T-I-T
T-R-C-A-E
G
D
R166
G T Y L K L K G R L F N
O—Polymer
O—Polymer
S
S
S S
B
Protein
12
12
12
12
C
Cys SEP(117-166) +
1. Oximation
Cys(Acm) SEP(89-116) —COSR + Cys SEP(117-166)
3. Ligation (NCL)
4. Deprotection
Cys SEP(89-116) SEP(117-166)
Cys(Acm) SEP(33-88) —COSR
5. Ligation (NCL)
6. Alkylation (see text)
7. Deprotection
Cys SEP(33-88) SEP(89-116) SEP(117-166)
SEP(1-32) —COSR +
2. Oximation
8. Ligation (NCL)
SEP(1-32) SEP(33-88) SEP(89-116) SEP(117-166)
9. Folding

Fig. 6: Molecular structure of synthetic erythropoiesis protein (SEP) [2]. One of the most advance protein ligation using NCL. (**A**) Diagram of SEP indicating its primary amino acid sequence, noncoded amino acids, and disulfide bonds. The levulinyl chemoselective sites of polymer attachment are coded in blue, and the carboxymethyl-modified cysteine residues are coded in green. The three ligation sites are circled in red. (**B**) Structure of the branched, negatively charged, precision-length polymer moiety. The aminooxyacetyl chemoselective site of protein attachment is coded in blue, and the charge control centre is coded in red. (**C**) Scheme for the synthesis of SEP by chemical ligation. The branched precision polymer was first attached to a predetermined site in each of the segments SEP (1-32) and SEP (117-166) by oxime-forming ligation at noncoded Lys (N $^{\varepsilon}$-levulinyl) residues incorporated during peptide synthesis. The 166-amino-acid residue polypeptide-polymer construct was then assembled by sequential amide-forming native chemical ligation (NCL) of the four unprotected synthetic peptide segments. The full-length polypeptide-polymer construct was folded with the concomitant formation of two intramolecular disulfide bonds to yield the SEP protein, which is represented as a cartoon.

1.1.2.2.2 Ligation by auxiliary groups

NCL has the disadvantage that a cysteine required for ligation remains in the protein sequence. N-terminal thiol-auxiliary-groups also perform protein ligations but the auxiliary moiety is removed after ligation (Fig. 7) [24,25]. The first auxiliaries were N$^{\alpha}$-ethanethiol and N$^{\alpha}$-oxyethanthiol [26,27]. After ligation N$^{\alpha}$-oxyethanthiol was removed with zinc under acid condition [28]. Since then different classes of auxiliaries for peptide ligations are reported. One of the latest inventions was photosensitive auxiliaries [29]. The auxiliary-assisted peptide ligations are very sophisticated but often the N-and C-terminal amino acids have to be glycines in order to get any ligation. Additionally the harsh condition necessary for the removal of the auxiliary after the ligation limits the range of applications. One ligation that could only be performed by auxiliary groups was the ligation and contraction to cyclic peptides [29].

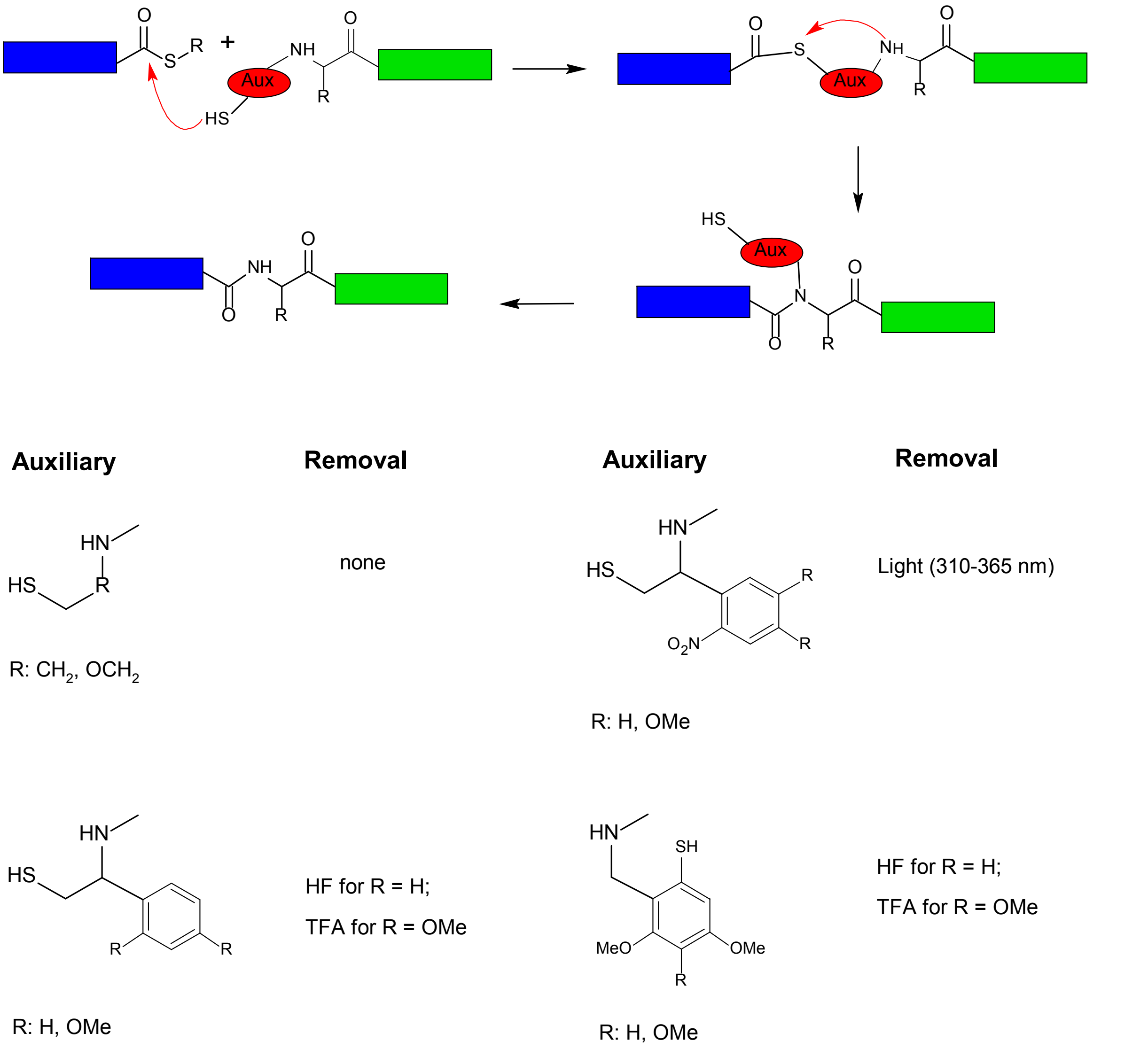


Fig. 7: Principle of ligation by auxiliary. Since the first auxiliary (N$^{\alpha}$-ethanethiol) [27] different classes of auxiliary for peptide ligations are reported [30,31,32].

1.1.2.2.3 Staudinger Ligation

1919 Staudinger and Meyer reported the reaction between azides and triphenylphosphine resulting in iminophosphoranes (Fig. 8). Iminophosphoranes have nucleophilic nitrogen which can react with many different functional groups, as ketones, aldehydes, amides and can be hydrolyzed to primary amines.

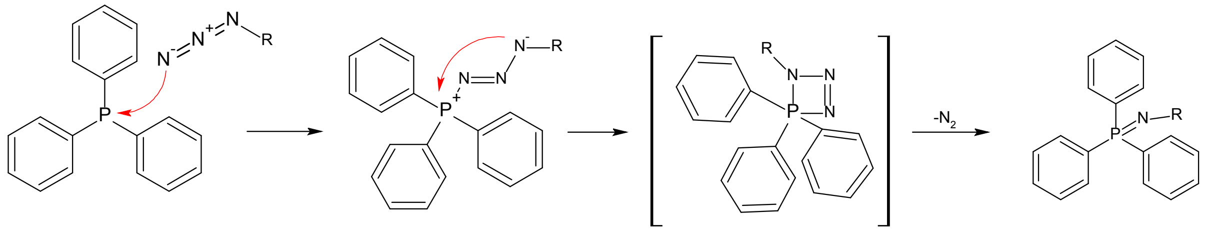


Fig. 8: Staudinger ligation; the resulting iminophosphorane is highly reactive.

The group of Bertozzi [33] reported a Staudinger reaction where the sample was linked to the detectable probe by an amide bond (Fig. 9). They were looking for abiotic reagents which react with each other but are unreactive to other biomolecules. The intramolecular cyclization of the iminophosphoranes with an ester resulted in a stabile amide bond and no hydrolysis to primary amines occurred.

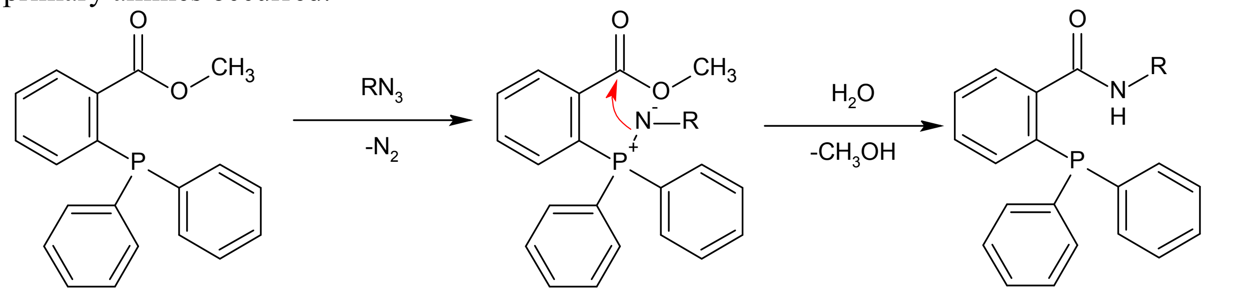


Fig. 9: Staudinger reaction performed by Bertozzi for cell surface detection. The linker between biological sample (here R) and the detectable probe is an amide bond.

3 months later the team of Raines [34] reported the ligation of a dipeptide using the Staudinger reaction. Remarkable for this reaction was that no residual atoms were left in the final product. At the same time Bertozzi et al. [35] described a similar reaction called the "Traceless" Staudinger Ligation (Fig. 10).

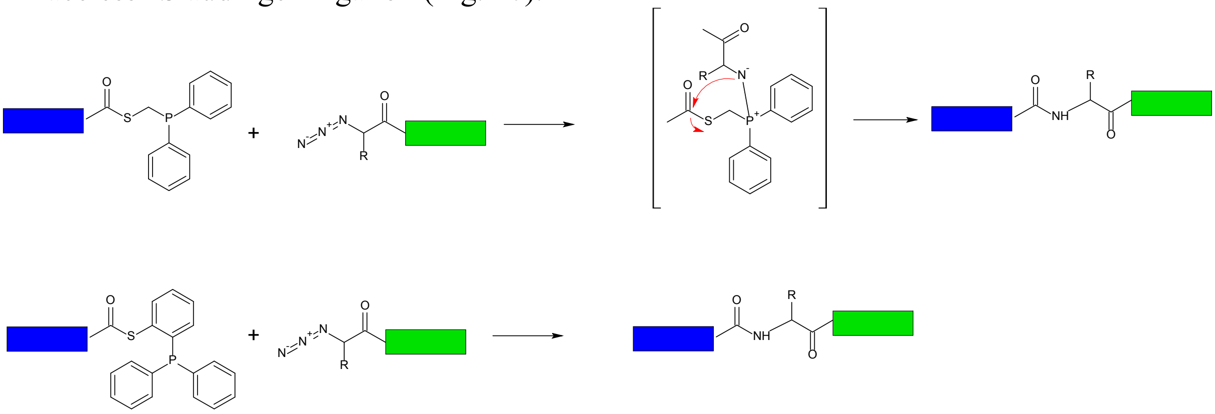


Fig. 10: Reaction mechanisms of traceless Staudinger reaction. Staudinger ligation allows having other N-terminal amino acids than cysteine. Glycines are most favourable as they don't hinder the transition state.

For high yield ligation at least one of the amino acids in the ligation junction has to be a glycine or alanine [36,37].

1.1.3 Ligation by Intein

1.1.3.1 History

20 years ago the discovery of a 120 kDa translational product that is converted posttranslationally to the 70 kDa subunit of the vacuolar membrane H^+-Adenosine Triphosphatase (VMA-1) from *Saccharomyes cerevisiae* and a so-called 50 kDa spacer protein was reported [38,39]. It was assumed and later proven [40] that the inner part is excised at the protein level. During the process the external fragments are ligated via peptide bond [41]. The intramolecular and autoprotolytic reaction needs no cofactors or other enzymes for performance [42].This spacer protein was later called intein after *int*ernal prot*ein* sequence and the flanking protein sequences were called exteins after *ext*ernal prot*ein* sequences [43].

Inteins have their name after the genus/species designation followed by the extein gene name. Inteins present in the same position in an extein homolog from different organisms are designated intein alleles [44]. Inteins from the same intein allele have a high amino acid identity (~60%), while as non-allelic inteins have normally 15-30% identity. Many inteins like the 50 kDa intein of the VMA-1 have homing endonuclease activity [45,46]. At least two different type of homing endonucleases belonging to LAGLI-DADG type and HNH type were found [47].

Sequence alignment showed that inteins contain 10 conserved motifs [44,48,49] although 4 of these motifs are now known to belong to homing endonucleases (Fig. 11)

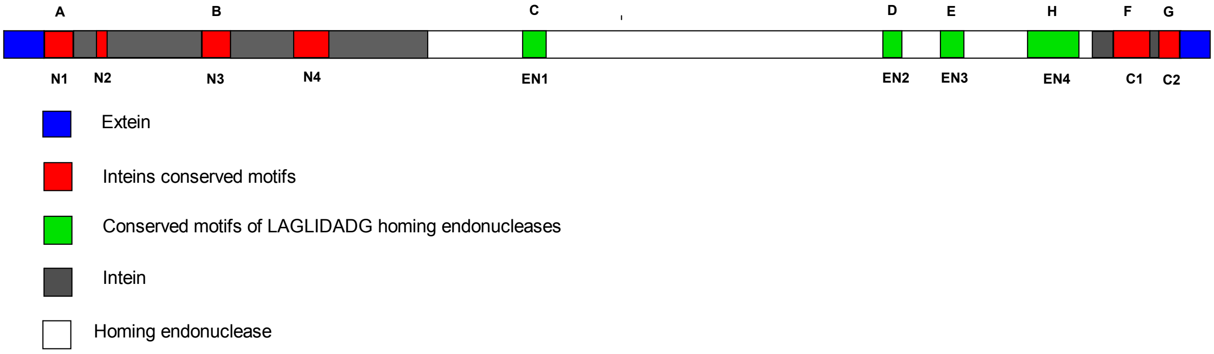


Fig. 11 Scheme of the protein sequence of inteins with the conserved blocks. Perler [44] assigned the blocks alphabetically. The block H is between block E and F because it was later discovered by Pietrokovski [48]. Pietrokoski named the blocks N for N-terminal intein, EN for endonuclease and C for C-terminal intein. The homing endonuclease (here in white and green) can be replaced either by a linker or be missed at all. The amino acid distance between the blocks A and B and between F and G is fixed but the overall length of an intein (plus endonuclease) varies. It should be noted that even in the blocks very few amino acids are really conserved. (See Fig. 15)

The splicing motif of an intein consists of the 100-180 amino acids after the N-terminal extein and approx. 40 amino acids before the C-terminal extein. (Fig. 11) [50,51,52]

Inteins can either be a two-domain protein where the second independent domain is a homing endonuclease [50,53,54] or a one-domain protein (mini-intein) [51,55] where a linker connects the N-terminal and C-terminal part of an intein. A special group of inteins are spilt intein where the N-and C-terminal fragment are separated on the DNA-level [52]. The first group of natural split inteins where found in cyanobacteria but many split inteins were engineered [55,56,57,58]. To be functional the two fragment of the intein have to join to the active protein first. The splicing of the split intein is called *trans*-splicing in comparison to *cis*-splicing by single molecule inteins.

Protein splicing is a folding dependent autocatalytic reaction [59] although there are reports where protein splicing occurs in high concentration of denaturants [60]. The first structure of an intein revealed an unusual fold for the splicing domain consisting 7 closely packed β-sheets which was later called HINT-motif (Hedgehog and INTein) [50]. Since the first description of inteins several hundreds have been found (many of them putative) in archaea, eubacteria and single cell eucarya including bacteriophages and a eukaryotic virus [61,62,63].

Few other auto-proteolysis mechanisms [64] are known (Fig. 12) but protein splicing is the most complex one as the protein back bone has to be broken at two positions and the exteins have to be ligated.

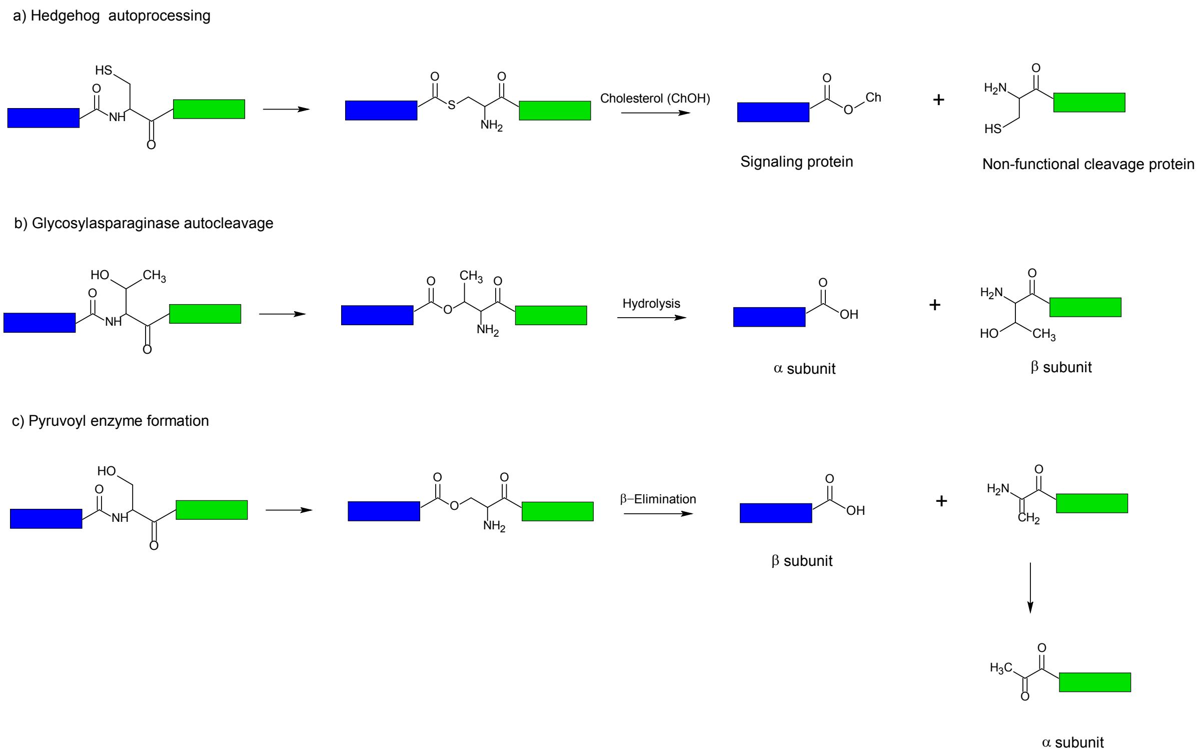


Fig. 12: Auto cleavage of protein precursor. A) Hedgehog and intein share structural similarities. The cleaved C-terminal fragment has no function whereas the N-terminal fragment esterified with Cholesterol is a signalling protein. B) The inactive precursor is cleaved between an Asp and Thr. The N-terminal threonine is essential for functional β-subunit. The mature enzyme is a heterodimer of the two subunits. C) The N-terminal pyruvoyl-moiety is necessary for enzymatic activity. Neither pyruvoyl-dependent enzymes nor glycosylasparaginase have similar sequence or structure to inteins or hedgehog proteins.

1.1.3.2 Mechanism

Intein splicing occurs very fast in fact unspliced precursors have never been identified *in vivo* in natural environment. The reaction mechanism postulated [65,66] and later proven [67,68] has similarity to the mechanism of NCL (Fig. 3 and Fig. 13). The first amino acid of the intein and C-terminal extein has to be Cys, Ser or Thr. Additional the last two amino acids of the intein are often His and Asn. One His in block B is also higher conserved (Fig. 15).

Mechanism of *cis*- or *trans*-splicing reactions:
1. Association of the two intein domains if the intein is split
2. Attack by Cys1, Ser1 or Thr1 results in a reactive (thio)-ester [67,69]
3. Attack of the N-terminal thioester by the C-terminal Cys+1, Ser+1 or Thr+1 residue yields a branched (thio)-ester [65,70]
4. Cyclization of the C-terminal asparagine residue results in spliced product [65,71]
5 S-N acyl rearrangement yields in a polypeptide chain [72]

N- and C-terminal cleavage:
a) Hydrolysis of the (thio)-ester resulting in N-terminal cleavage[40]
b) Asn cyclization causing C-terminal cleavage

For many applications cleavage is undesirable but for Express Protein Ligation the cleavage with nucleophilic agents is used advantageously (Fig. 17).

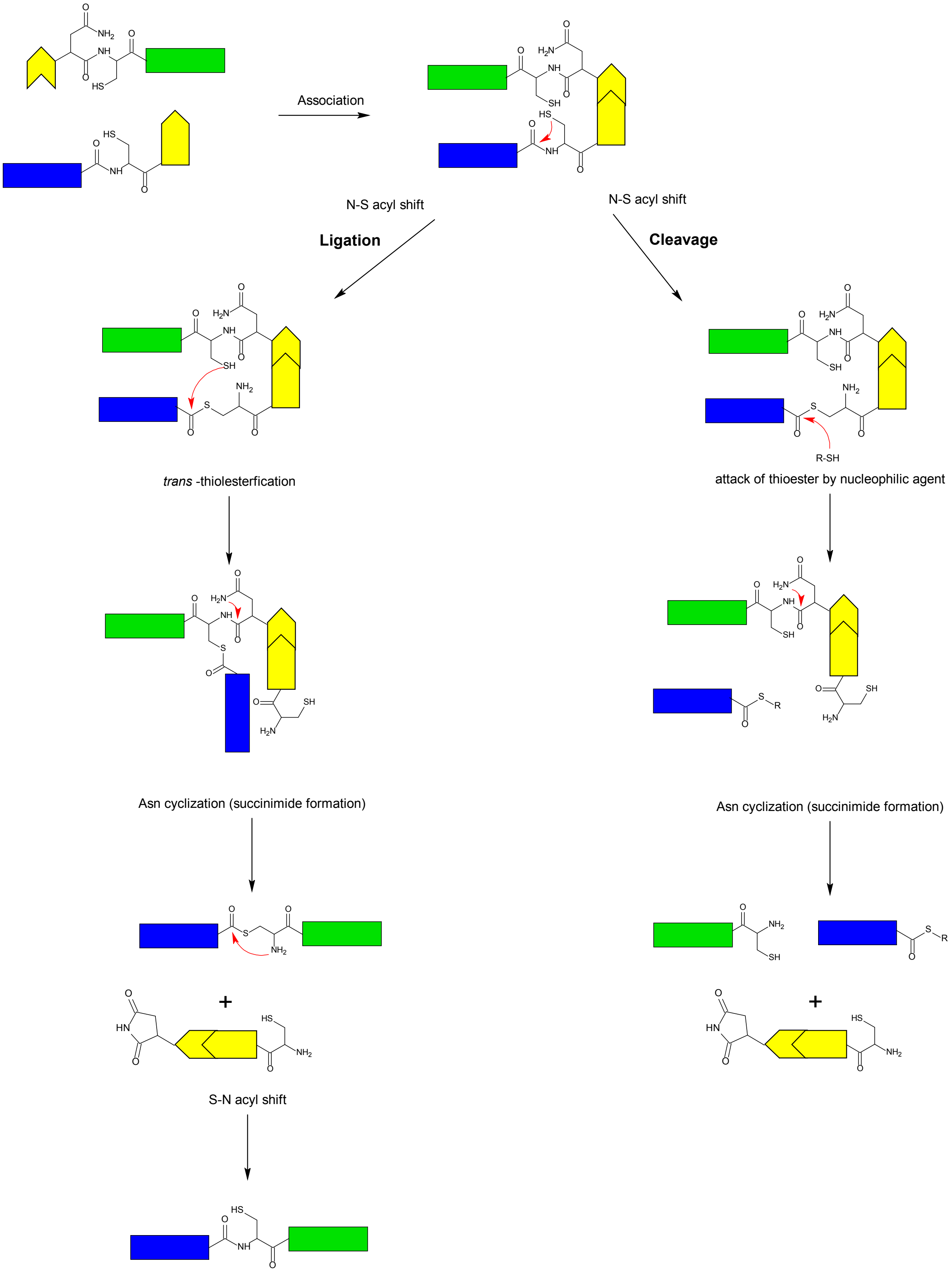

Association
N-S acyl shift
Ligation
N-S acyl shift
Cleavage
trans -thiolesterfication
attack of thioester by nucleophilic agent
R-SH
Asn cyclization (succinimide formation)
Asn cyclization (succinimide formation)
S-N acyl shift

Fig. 13: The protein splicing mechanism. If the intein is split the first step is the association of the two domains. The side chain of Cys1 or Ser1 attack the peptide bond resulting in a reactive (thio)-ester. This ester is attacked by the first amino acid of the C-terminal extein (Cys+1, Ser+1 or Thr+1) or nucleophilic detergents hydrolysis the bond. Cyclization of the asparagine (the last residue of an intein) results in spliced product and a free intein. The S-N acyl shift yields in polypeptide bind. If the reactive (thio)-ester is attack by nucleophilic detergents (like DTT) the N-and C-terminal extein are cleaved of from the intein but not ligated.

The self-association of *trans*-splicing protein is fast and so stable that "turn-over rate" (association- protein splicing-dissociation-new association) is not measureable. The two fragments of a split intein are assumedly connected by several salt bridges.[73]

In year 2000 the reaction pathway for a natural and active intein - KblA from *Methanococcus jannaschii* - was described where the first amino acid was an Ala [74]. In this case the first amino acid of the C-terminal extein attacks directly the peptide bond at the N-terminal splice junction (Fig. 14 A) This group of inteins - all belong to the same intein allele and are now called Class 2 intein- was long assumed to be inactive [48,75].

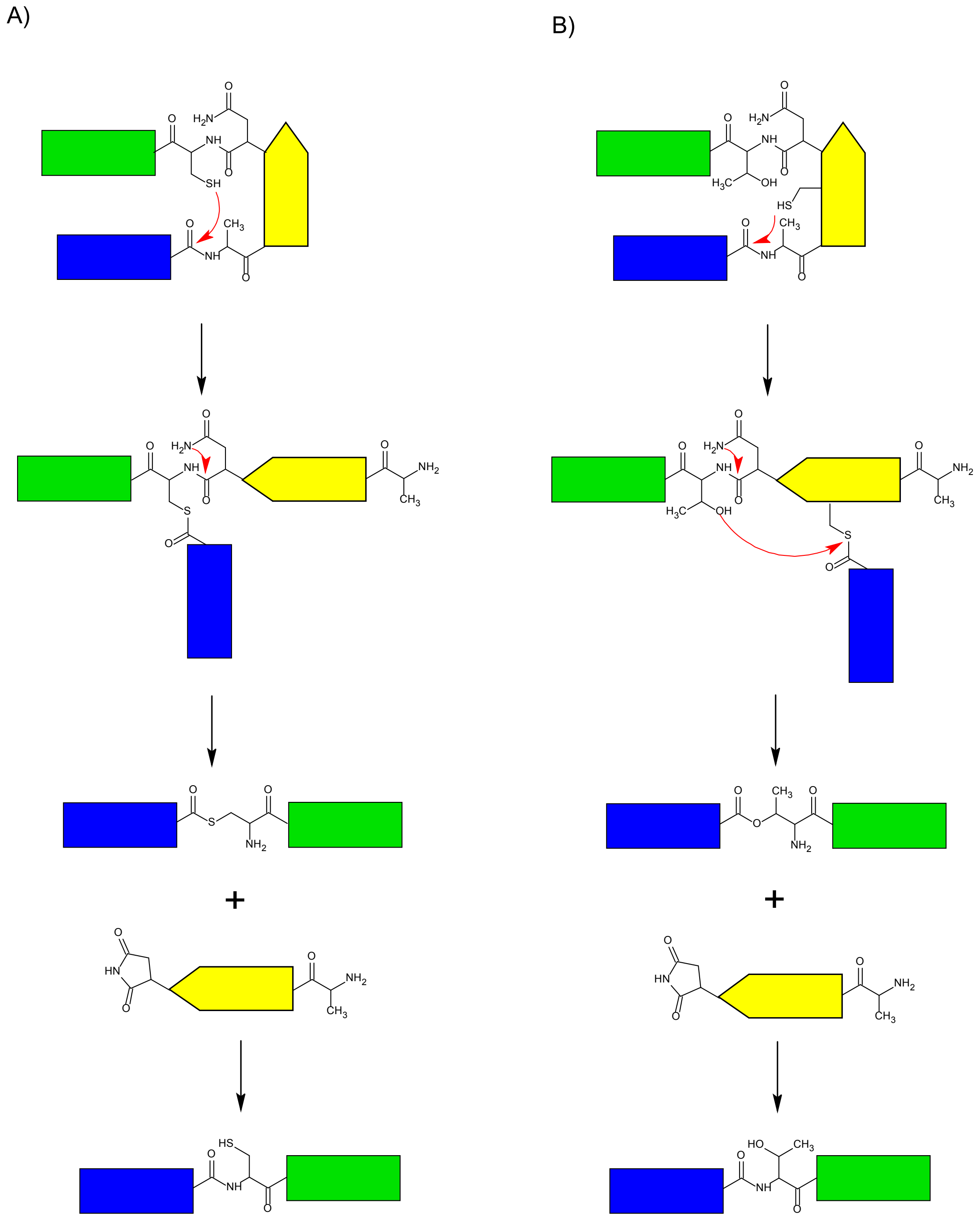


Fig. 14: A) The first residue of the C-terminal extein attacks directly the carbonyl-group of the last residue of the N-terminal extein resulting in a branched intermediate. The final steps of protein splicing are identical as described in figure 6. Inteins that follow that reaction mechanism are now called Class 2 intein. B) Reaction mechanism of Class 3 inteins. A cysteine of the intein attacks the peptide bond resulting in a thioester. That thioester is attacked by first residue of the extein.

In 2010 a third splicing pathway was described performed by the intein of Mycobacteriophage Bethlehem DnaB [76] (Class 3 intein). The type of intein forms a branched intermediate with an essential Cys in the F-block (Fig. 14). Sequence alignment revealed a conserved triplet of a Trp, Cys and Thr (WCT triplet) (Fig. 15).

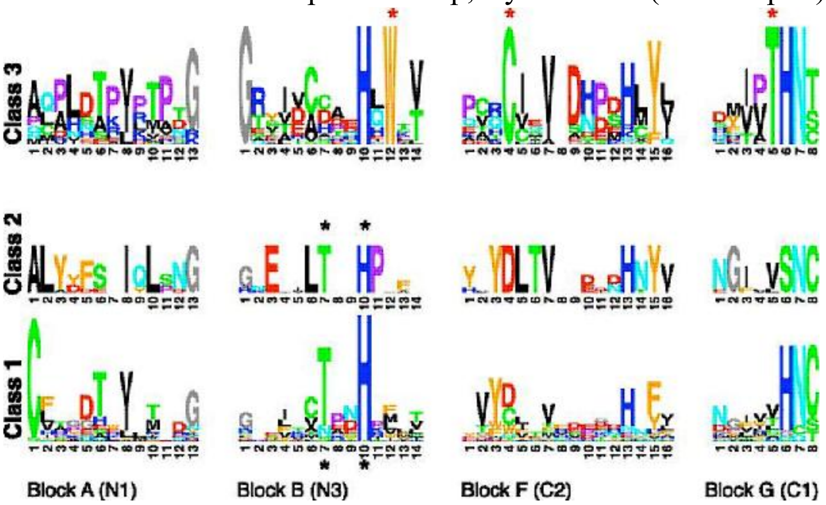


Fig. 15: Comparison of the conserved blocks [76]. Class 2 inteins have an alanine as first residue where as Class 3 inteins have different amino acid in that position. All three classes have a unique set of highly conserved amino acids. As class 2 and 3 inteins were only discovered as the first residue is not a Cys it could be possible that there are even more classes.

The only inhibitor ever found is Zn^{2+} [77]. Some other divalent ion showed with different extend inhibition (Cd^{2+}>Co^{2+}>Ni^{2+}). Mg^{2+} does not inhibit protein-splicing and copper also oxidize inteins [73]. Zn-ion inhibits *in vitro trans*- and *cis*-splicing of the intein *Mycobacterium tuberculosis* (*Mtu*) RecA and *Synechocystis species* (*Ssp*) DnaE [78]. The crystal structures of inteins Crystal structures of inteins PI-SceI [79], *Mtu* RecA [80] and *Ssp* DnaE [81] revealed that each of the three inteins has a different zinc binding site.

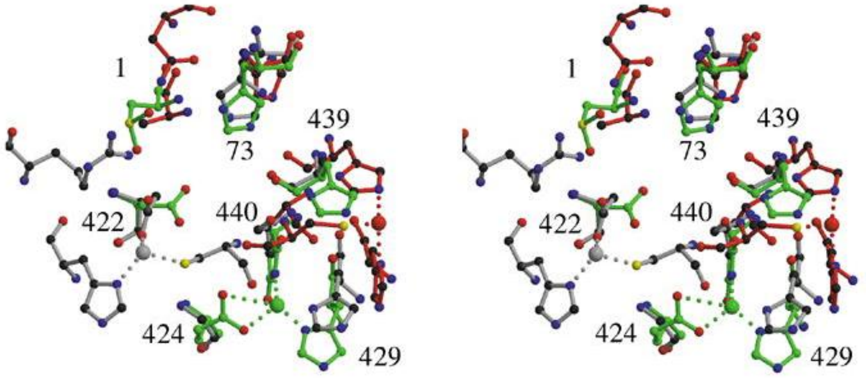


Fig. 16 Comparison of the zinc-binding sites observed in the crystal structures of three different inteins: *Mtu* RecA (green), PI-SceI (red) and *Ssp* DnaE (grey) [80]. Residue numbers shown correspond to those of *Mtu* RecA. There are at least three different zinc binding sites. In *Mtu* RecA (green) the zinc ion is coordinated by residues Glu424, His429, and the C-terminal aminosucciminide (SU440) from one molecule (green), and His439 from the second molecule in the asymmetric unit cell. The Zn-ion inhibits splicing in PI-SceI (red) by binding to the C-terminal Cys, penultimate His (here 439) and Glu (next to the conserved His in B-block) and one water molecule. In *Ssp* DnaE (grey) the Zn-binding site consist of the C-terminal Cys, an Asp (here 422) and 2 His of which one is from the second molecule in the asymmetric unit cell.

1.1.3.3 Application of Protein Ligation

For protein ligation two intein-assisted methods are used. One is Protein *trans*-splicing the other is called Expressed Protein Ligation (EPL) (Fig. 17).

EPL (or IPL after Intein-mediated Protein Ligation) use inteins to obtain a C-terminal thioester and an N-terminal cysteine for ligation [82,83]. The cleavage of the fragment from the inteins can combined with on-column purification and is commercial available (Fig. 18 see also Impact-Twin system from New England Biolabs)

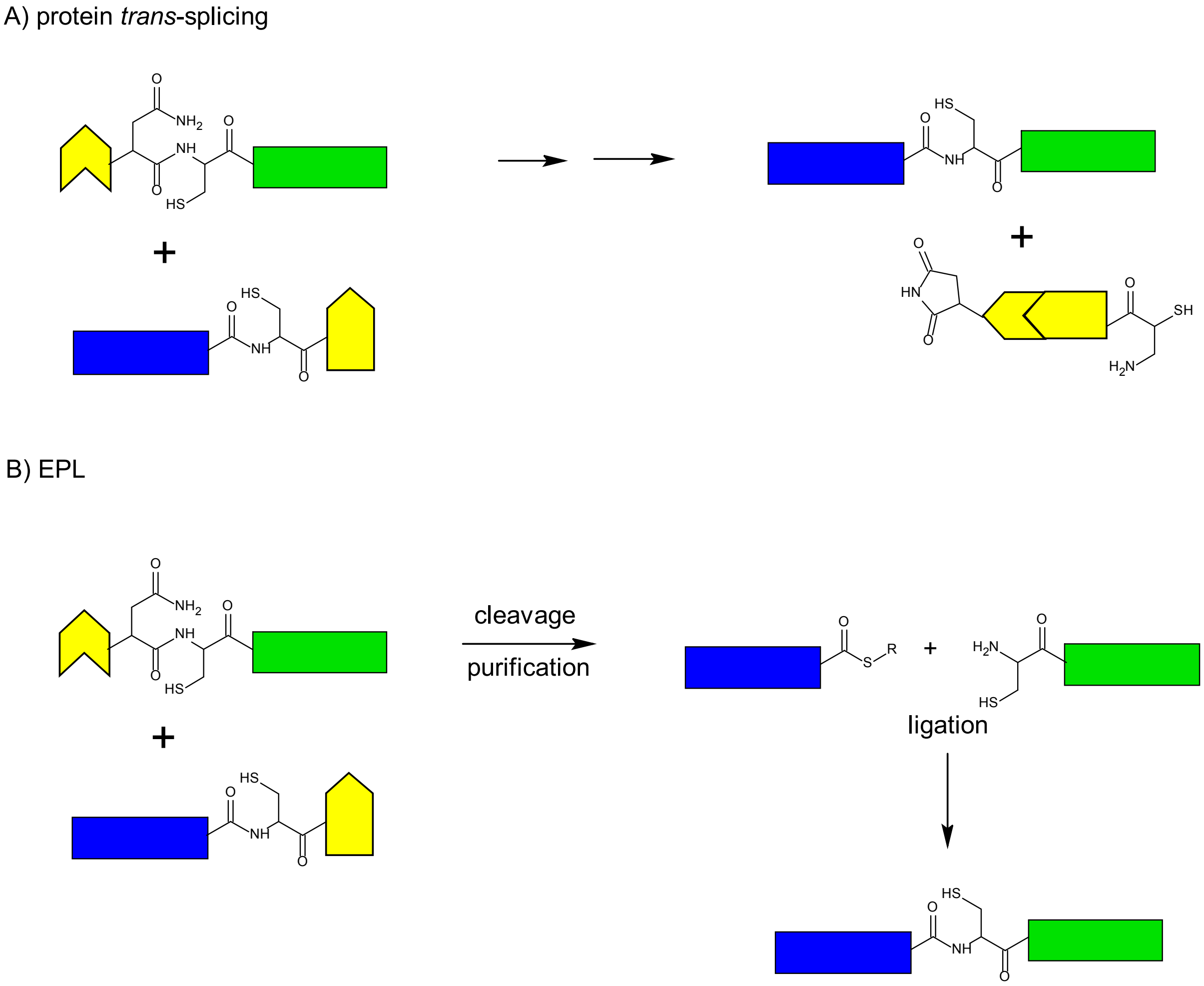


Fig. 17: Comparison of *trans*-splicing and EPL. Both ligations depend on an intein for ligation. Fragments of NCL and EPL can be combined for ligation called then semisynthesis.

Expressed protein ligation was first used for semisynthesis combining the advantages of protein splicing with chemical synthesis of protein.

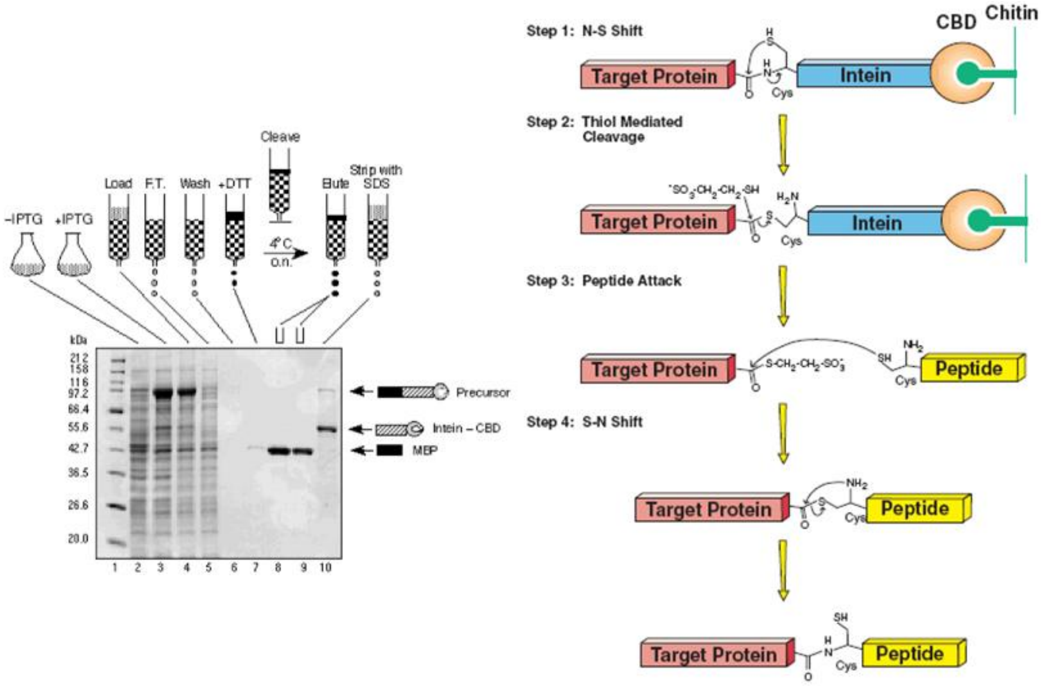


Fig. 18: The Twin IMPACT systems allows on-column purification and cleavage. The purified and cleaved product can be used for ligation. The crude cell extract containing the fusion-protein consisting of the target protein, an intein and chitin binding domain (CBD) is loaded on a chitin column. The column is washed and the target protein is cleaved from the column was a nucleophilic agent. The cleaved protein can be used further for ligation. Both fragment for the ligation can be obtained by that system. If the IMPACT system is only used for purification reason that DTT should be used as nucleophilic agent. For EPL the sodium salt of mercaptoethylsulfonate (MESNA) is used as nucleophilic agent [84,85,86].

Cytotoxic proteins [86] or cyclic peptides [87] were also produced. Cyclization of peptides were performed by the natural split intein of *Ssp*. DnaE [87] or EPL [88].

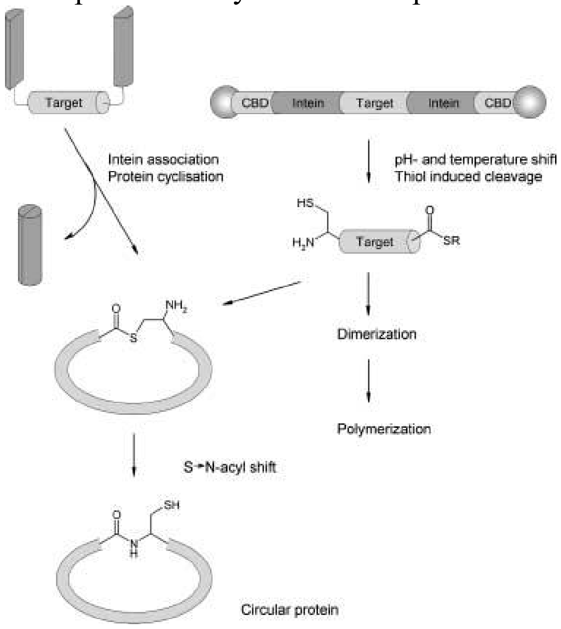


Fig. 19: Reaction mechanism of circular proteins. Cyclization can be performed by split inteins or EPL. Split inteins are now more preferred as there are less side reactions [89].

The group of Muir [90,91,92] have presented intein constructs that only perform protein splicing if a small molecule like rapamycin is present (Fig. 20). The method is called condition protein splicing (CPS).

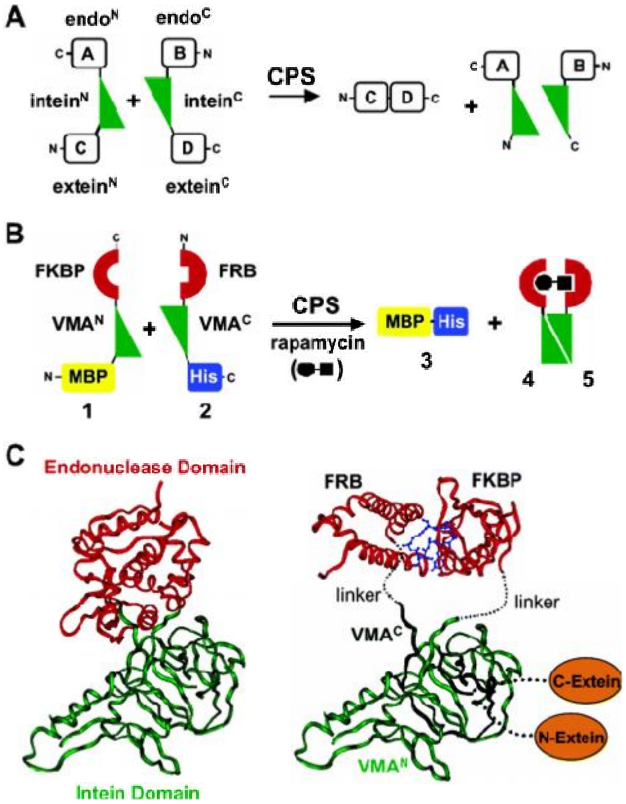


Fig. 20: Principle of conditional protein splicing (CPS) [91]**.** (A) General nomenclature for the constructs used. The extein sequences are located at the C and D positions, whereas the positions A and B on the opposite ends of the split intein halves are termed the endo sequences. (B) In the original CPS approach, the active intein is reconstituted by rapamycin, which heterodimerizes the FKBP and FRB domains located at the endo positions A and B. When using the model exteins MBP and His-tag, protein splicing results in formation of MBP−His. (C) Crystal structure of the wild-type VMA protein which is composed of the catalytic intein domain (green) and the intervening endonuclease domains (red). The modularity between the two domains should be noted (left). Rapamycin is shown in ball-and-stick representation (blue).

Other approaches of conditional splicing are temperature-sensitive [93] or light-sensitive inteins [94,95].

In 1998 segmental isotope labelling for NMR measurements was proposed (Fig. 21) [96]. The N-terminal intein of ribonucleoside-diphosphate reductase from *Pyrococcus furiosus* (*Pfu* RIR1-1) was fragmented and as the exteins the C-terminal domain of the RNA polymerase α-subunit was used.

In 1999 segmental isotopic labelling on the Abelson protein tyrosine kinase-SH(32) was prepared [97] and in the same year Otomo et al. described the first segmental labelling of a central segment [98].

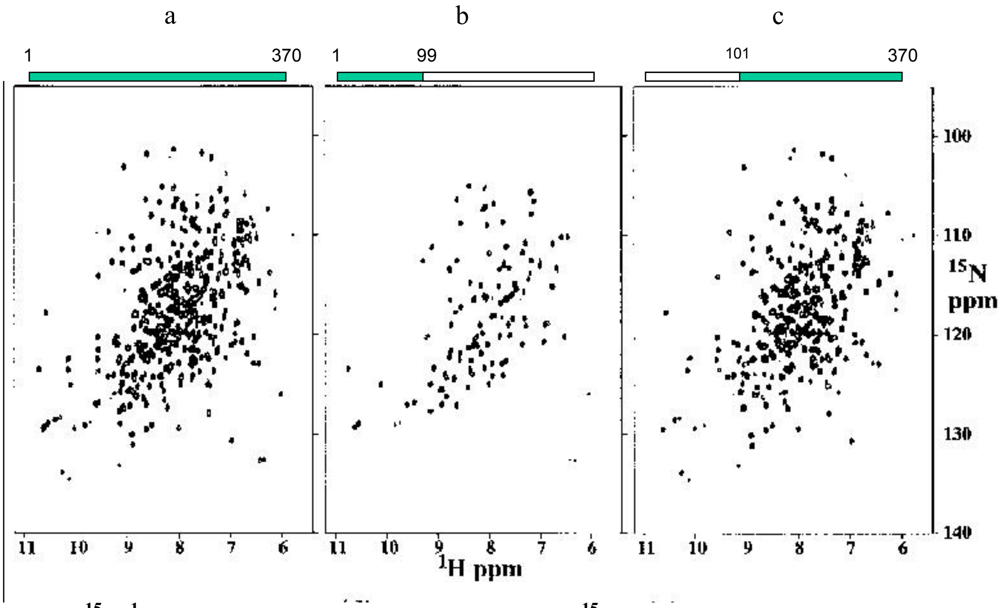


Fig. 21:**2D ^{15}N-^{1}H HSQC spectra of (a) the uniformly ^{15}N-labeled wild-type maltose binding protein, and those of MBP segmentally labelled with ^{15}N between (b) Lys1-Tyr99; (c) Gly101-Lys370 [99].**

In 2000 Ozawa et al. [100] introduced a new approach to measure protein-protein interaction *in vivo* based on protein splicing (Fig. 22). The fragment of a split intein is cloned between a split green fluorescent protein (GFP) or luciferase [101] and the two protein of interest. If the two proteins interact with each other the intein can fold correctly and ligate the fragments of GFP or luciferase.

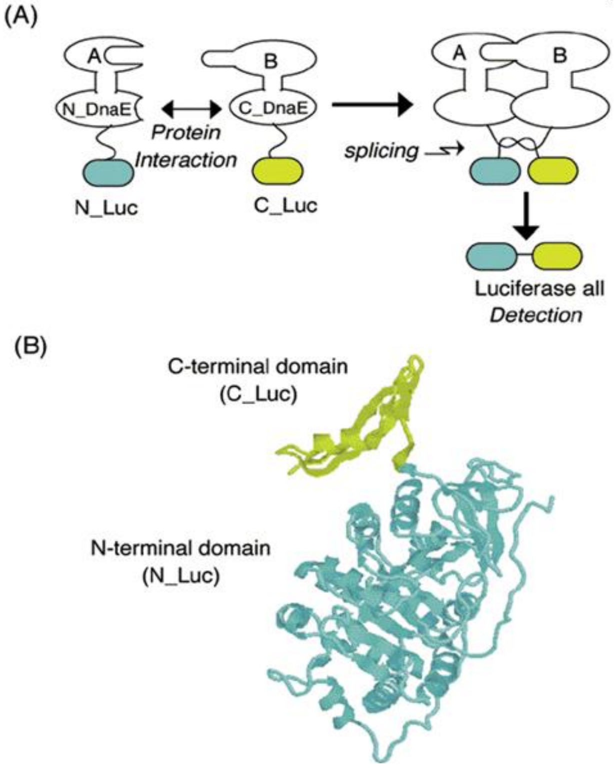


Fig. 22: A) Principle for the present protein splicing-based split luciferase enzyme system [101]**.** The N-terminal half of DnaE (N_DnaE) and C-terminal half of DnaE (C_DnaE) are connected with the N-terminal half of luciferase (1−437 amino acids; cyan) and the C-terminal half of luciferase (438−544 amino acids; yellow), respectively. Interacted protein A and protein B are linked to opposite ends of those DnaE. Interaction between protein A and protein B accelerates the folding of N_- and C_DnaE and protein splicing results. The N- and C-terminal halves of luciferase are linked together by a normal peptide bond to recover its enzymatic activity. (B) 3D structure of firefly luciferase. N- and C-terminal halves of the enzymes are shown as cyan and yellow, respectively.

1.1.4 Expressed Enzymatic Ligation (EEL)

Protein ligation by proteases in organic solvent has been known since 1970s [102,103,104] but since then few examples have been published. The double mutant of the bacterial serine protease subtilisin (S221C and P225A) designated subtiligase ligates peptides with high efficiency [105] and was used for total synthesis [106]. A heptamutant of subtiligase (M50F, N76D, N109S, L213R, N218S, S221C and P225A) named stabiligase is even active at 4M Guanidium-Chloride [105]. V8-protease [107,108] and sortase [109] mediated protein ligation were also published. Although there are successful example of peptide ligations EEL cannot compete with intein-based protein ligation.

1.2 Alginates

1.2.1 Alginate composition

Alginates are linear, anionic copolymers consisting of β-D-mannuronic acid (M) and α-L-guluronic acid (G) in (1→4) linkage [110,111,112]. Both monomers are acidic (pK_a is 3.38 for M and 3.65 for G) at neutral pH and alginates are degrading under strong acidic and alkaline condition [113] (Fig. 23).

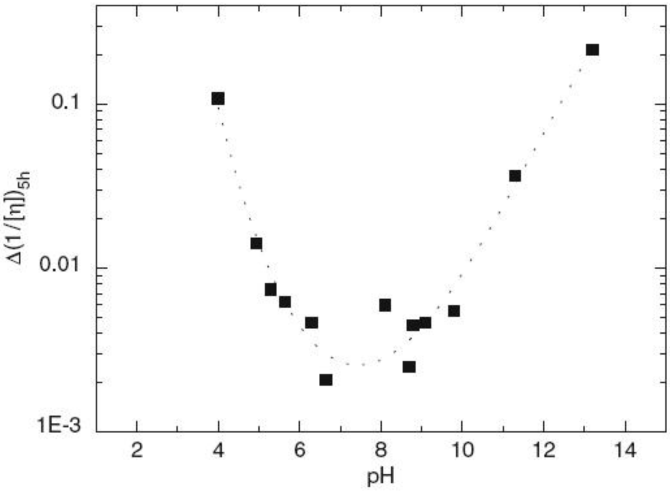


Fig. 23: Alginates are most stabile at neutral pH and degrade under acid and basic condition [114]. Degradation of alginates isolated from *Laminaria digitata* measured as the variation of the ratio 1/[η] after 5 h at different pH and at 68°C. (The *line* is drawn to guide the eye.)

In contrast to many other biopolymers, alginates do not have one defined distribution pattern of the monomers. Alginates are produced as linear polymannuronan in a $^1C_4$ conformation and alginate epimerases convert M to G [115] after polymerisation. Whenever a G is incorporated, the polymer chain bends in order to keep the acid -moiety in equatorial position (Fig. 24). Therefore, β-D-mannuronic acid is in $^1C_4$ conformation while α-L-guluronic acid adopts $^4C_1$ conformation [116,117]. Not only the ratio M to G is important to describe physical and chemical characteristics of alginates but also the short range order [118]. Alginates are therefore characterized as a block-copolymer and subdivided into three blocks named MM- and GG- and MG-blocks. MM- and GG-blocks are homopolymeric regions of mannuronic (MM-block) or guluronic acid (GG-block) respectively whereas MG-block describes the alternating sequence [119,120] (Fig. 24). The viscosity of the soluble blocks increases in the order MG<MM<GG [121]. The viscosity of the different blocks reflects the flexibility of the monomers around the glycosidic bonds and the possibility of forming hydrogen bonds (Fig. 24B). Mannuronic acids are linked diequatorially and long poly-M strains form a 3-folded left handed form with hydrogen bonds between the hydroxyl proton in position 3 and the subsequent ring oxygen [122]. Longer sequences of L-guluronic acids have a higher viscosity as the flexibility around the diaxial glucosidic bond is sterically restricted [123]. The additional weak hydrogen bonds between the carboxylic and the 2-OH group of the prior residues and the 3-OH group of the subsequent residues fixed the structure to a 2 folded helix even more. In MG-blocks the links are alternating equatorial-axial and axial-equatorial which leads to a dissimilar structure. Between the carboxylic group of M and the 2-OH and / or 3 OH group of the following G hydrogen bonds can be formed. The lack of hydrogen bonds between all the monomers results in a greater overall flexibility and therefore lowest viscosity.

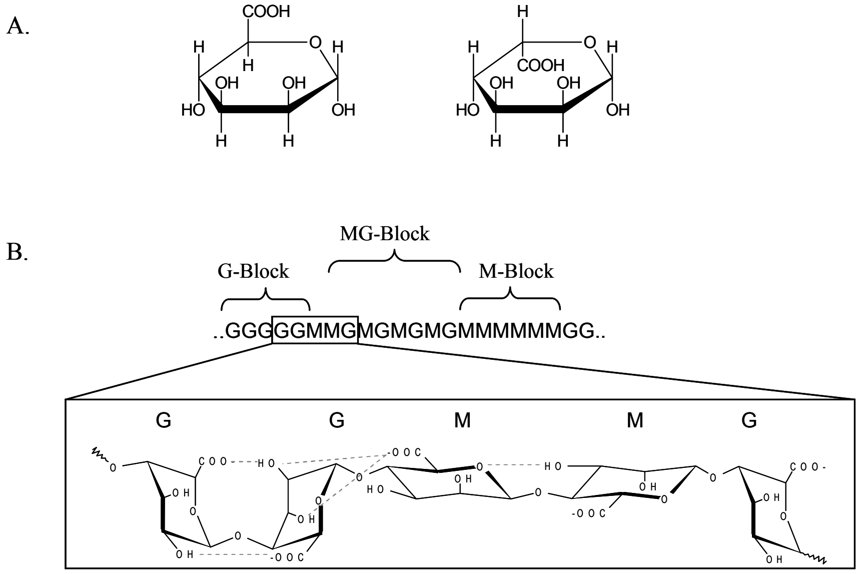


Fig. 24: A) Haworth formulae of -D-mannuronic acid (M) and α-L-guluronic acid (G). B) Different blocks of alginates. Mannuronic acid is in $^1C_4$ conformation while guluronic acid is in $^4C_1$ conformation. The hydrogen bonds are drawn as grey dashed lines [124].

All three types of blocks can form gels but under varying conditions. MM-blocks can only form acidic gels [125]. G-rich alginates, shows selective binding of certain alkaline earth metals ions and other divalent cations (e. g. $Pb^{2+}> Ba^{2+} > Ca^{2+} >> Mg^{2+}$) (Fig. 25 B) [126]. MG-blocks form soft gels induced by divalent cations such as Ca^{2+} but cannot form acidic gels [125,126,127]. The gels of the GG-blocks are very stiff, brittle and thermostable. They can be obtained either at low pH or by addition of divalent cations [125,126]. The selective binding induces a dimerization that is also called egg-box model (GG-blocks) or zipping (MG-blocks) [128,129]. The dimerization can either lead to a homer-dimer or a hetero-dimer (Fig. 25A) [127,130].

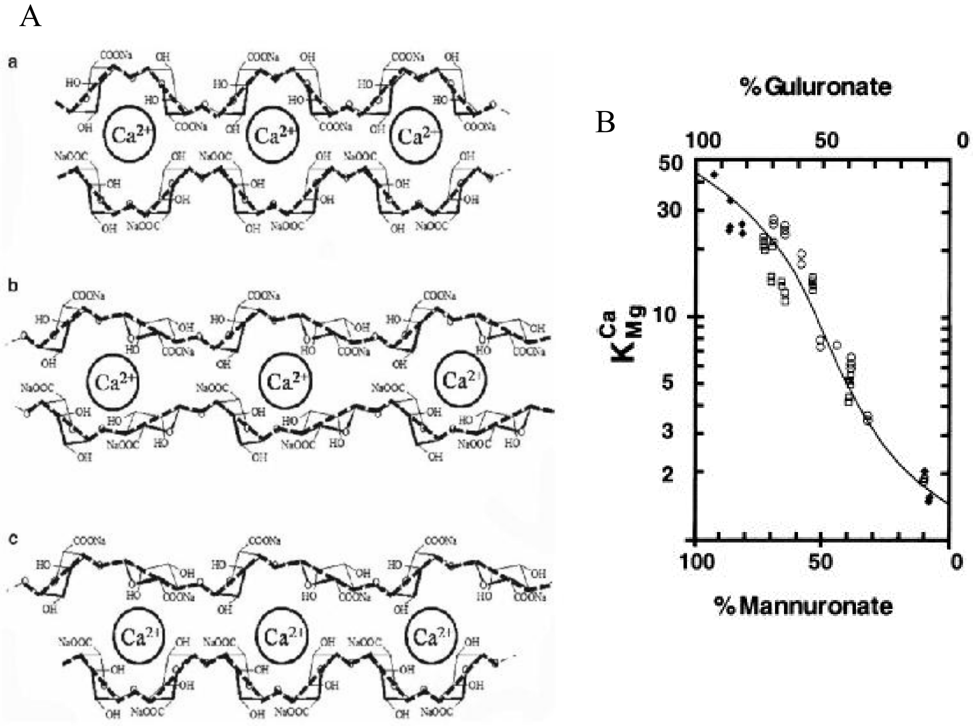


Fig. 25: A) The three possible ionic junction [114]: a) GG/GG-junction; b) MG/MG-junction; c) GG/MG-junction; **B) G-rich alginates show a preference for certain divalent cations (e.g. Ca^{2+} > Mg^{2+}) [131].** Ca^{2+}-ions are used in G-rich alginates as cross-linking ions for the junctions while in MM-blocks they just compensate the charges. The selective coefficients, K^{Ca}_{Mg} were obtained at X_{Ca}=X_{Mg}=0.5.

Bacterial alginates are often acetylated at the O2 and/or O3 position while algal alginates are never acetylated [132]. The proteins AlgI, AlgF and AlgV (AlgJ in *Pseudomonas* see Alginate Biosynthesis) are responsible for the acetylation of nascent alginate chain [133,134,135]. Only mannuronic acids are acetylated and the acetyl-moiety prevents epimerisation and lysis of the mannuronic acids and its neighbours [136]. For *Pseudomonas aeruginosa* acetylation is necessary for the formation of microcolonies during the early stage of biofilm development [137,138]. *O*-acetyl alginates protect *P. aeruginosa* from host defences [139]. For *Azotobacter vinelandii* acetylated alginates serve another purpose than for *P. aeruginosa.* It was suggested that acetylation is a control mechanism to determine the final degree of epimerisation by extracellular enzymes. Acetylated alginates contribute to the dehydration resistance of the cysts as it has a higher water binding capacity than non-acetylated alginates [140,141].

1.2.2 Alginate producing organism

Alginates are produced by brown algae and bacteria belonging to the *Pseudomonas* and *Azotobacter* genera [142,143,144]. In brown algae, alginates have structure-forming function and comprise up to 40% of the dry matter. The alginates are located in the intercellular matrix as a gel. The stiffness of the alginate gels reflect the mechanical stress the tissue is exposed to [125]. The stipe and outer cortex of *Laminaria hyberborea* have a high percentage of GG-blocks while alginates in the leaves consist of more MM- and MG- blocks [145,146].

Pseudomonas and *Azotobacter* species also produce alginates. Alginates produced by *P. aeruginosa* provide the organism an advantage during infection of patients suffering from cystic fibrosis [147]. *A. vinelandii* , a soil bacterium, produces an alginates based resting state designated cyst [148]. *A. vinelandii* epimerizes to alginates to GG-blocks by seven extracellular alginate epimerases [136].

1.2.3 Gel Formation

The gel formation properties are of prime importance for most applications. The gelformation is temperature independent. This also means that alginate gels can be heated without melting.
Alginate solutions form hydro-gels at pH < 3 or in the presence of divalent cations. The chemical composition of alginates defines the strength of the gel. The higher the ratio of GG-blocks the stronger is the gel.
Direct addition of acid or divalent cations for gel formation is not possible as the result is a dispersion of gel lumps. There are two methods for preparation of an alginate gel (Fig. 26):

- Internal setting method
- Diffusion method

Internal setting method is used for obtaining homogenous gel. This method is used both for ionic and acid gels. Formation of acid gels is controlled by D-glucono-δ-lactone (GDL). The final pH is defined by the initial concentration of the lactone. For ionic gel $CaCO_3$ (Fig. 26A) or chelated Ca-ions (Ca-EDTA, Ca-citrate). Decrease of pH (normally done by GDL) release the Ca-ions slowly [149]. If $CaCO_3$ and GDL are used in the ratio 1:2 the final pH-value does hardly differ from the beginning.

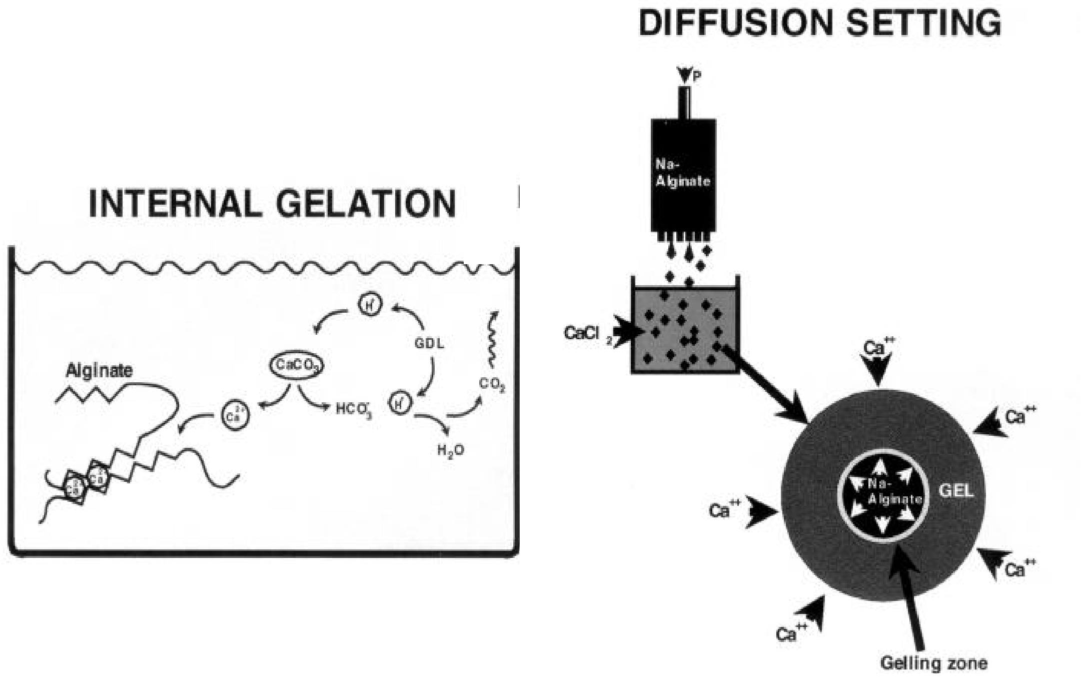


Fig. 26: Principle difference between internal gelation and diffusion method [131]. Internal gelation can be used for producing acid gel or ionic gels (as exemplified). Only ionic gels can be produced by diffusion method. The beads have a stiff shell and a soft and sometimes liquid core.

Diffusion method is characterized by allowing divalent ions to diffuse from a larger outer reservoir into the alginate solution (Fig. 26 B) [150]. The settings are often performed by dropping the alginate solution into a saturated $CaCl_2$ solution. In this case the gel formation is very rapid resulting into instantaneous formation of beads. The final gel beads often exhibit an inhomogeneous distribution of alginates where the highest concentration is on the surface and gradually decreasing towards the centre of the gel (Fig. 27) [151]. Hence, the polymer is mainly located on the external part of the bead, while a very soft and liquid core is found in its central part [150]. This method is useful for immobilization of cells and entrapment of biological compounds.

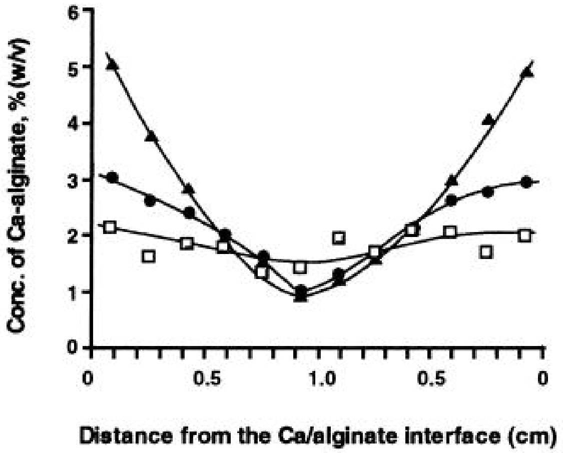


Fig. 27: Polymer concentration profiles of alginate gel cylinders formed by dialyzing a 2% (w/v) solution of Na-alginate from *Laminaria hyperborea* against 0.05 M $CaCl_2$ in the presence of different concentration of NaCl [150]. □: 0.2 M NaCl; ●: 0.05 M NaCl; ▲: no NaCl

1.2.4 Chemical and physical measurements

NMR is a good technique to gain information about M/G ratio and block compositions pioneered by Penman *et al.* [116]. ^1H and ^{13}C NMR spectroscopy provide information about the monads (F_M, F_G) and their nearest neighbours as dyads (F_{MM}, F_{MG}, $F_{GM,}$ F_{GG}) and triads (F_{MMM}, F_{MMG}, F_{GMM}, F_{GMG}, F_{MGM}, F_{MGG}, F_{GMM}, F_{GGG}).[152,153,154] Today, NMR is the dominant method for analysing the composition of alginates.

CD-measurements reveal similar information about block composition of alginates. The CD-spectrum of any alginate is a linear combination of the molar fraction of the three fundamental spectra of the pure block-alginates [155].

Mechanical properties to withstand compression or shear stress are important in any application. The strength of alginate gels depends on the length of GG-blocks [156] and MG-blocks [130] as well as the cation used for gel formation. Alginate gels containing long MG-junctions can reopen (Fig. 28). The unwinding of MG-junction under pressure gives the gel more flexibility but does not cause rupture [130].

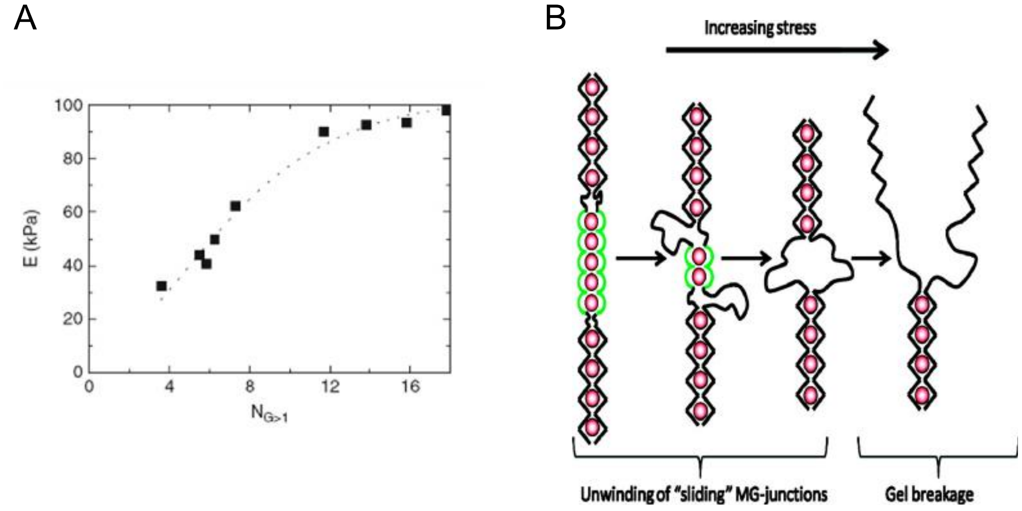


Fig. 28: A) Elastic properties of alginate gels as function of average GG-block length [156]. B) Graphical representation of the breaking of MG/MG junctions (in green) under stress [130]; this unwinding causes a plastic behaviour of the gel. GG/GG-junction (in black) breakage causes structure structural failure of the whole gel.

1.2.5 Alginate application

Alginates are used in more than 600 different applications. In the food industry, alginates are applied as additives due to their gelling and vicosifying ability. In the ice cream industry, alginates are used to prevent crystallisation and shrinkage. Alginates give ice a more homogenous and creamy texture when it is melted. Alginates are capable of forming heat stable gels therefore they are used in bakery production, jam and dairy products (Fig. 29) [157,158].

Fig. 29: An example of alginates used in food industry: the pimiento filling of olives [131]

Beside the food industry alginates are also used in textile industry for printing, in paper production as surface smoothing agent and in ceramic manufacturing for slow water evaporation.

The pharmaceutical industry uses alginates in wound dressings, where it can absorb secreted fluid as well as provide a protective environment for the healing process. Alginates are also used as dental impression material and to prevent gastric reflux. Recently oligoguluronate was suggested as modifier of cystic fibrosis mucus [159].

Encapsulation of biologically active materials can be carried out in a single step procedure under mild conditions and is suitable with most cells. In medical science encapsulation for cell transplantation holds great potential. The alginate coat protects the immobilized cells against the immune response but metabolites and small proteins are transported through the gel. The immobilization techniques can for example be used in the production of monoclonal antibodies from hybridoma [160]. Alginate beads - although easy to obtain - lead to severe complications if used in cell transplantation. Under physiological condition (0.9% NaCl) alginate beads are swelling and G-rich alginate gels tend to break. M-rich alginates cannot be used, either, as they trigger the immune response. To overcome the obstacles alginate capsules were coated with poly-L-lysine. Alginate/poly-L-lysine capsules containing pancreatic Langerhans islets have been shown to produce insulin in large animals and have also been clinically tested in humans [161,162].

1.2.6 Alginate Biosynthesis and Epimerases of *Azotobacter vinelandii*

The biosynthesis of alginate was first studied in cell free systems from *Fucus gardnerii* [163], *A. vinelandii* [115] and *P. aeruginosa* [164]. Starting with D-fructose-6-phosphate, the sugar nucleotide GDP-D-mannose is produced and oxidized to GDP-D-mannuronic acid. GDP-D-mannuronic acid is polymerized to polymannuronan. The polymer is partly acetylated and single mannuronic acids are epimerized to L-guluronic acid. The polymer chain is then transported through the periplasmic space into the medium. In *A. vinelandii* the alginate chains are further epimerized. The genes involved in the biosynthesis of alginates were first identified in *P. aeruginosa* [164]. All gene essential for alginate biosynthesis are in one gene cluster except *algC* (Fig. 30A). A similar gene cluster was found in *A. vinelandii* [165] recently.

The detailed workflow for alginate synthesis and transportation through the membrane into extracellular space contains following steps (Fig. 30B) [166]:

1) fructose -6-phosphate (F6P) to mannose-6-phosphate (m6P) by AlgA
2) mannose-6-phosphate to mannose-1-phosphate (m1P) by AlgC
3) mannose-1-phosphate to GDP-D-mannose (GDP-m) by AlgA
4) GDP-D-mannose to GDP-D-mannuronic acid (GDP-M) by AlgD
5) Polymerisation by polymannuronan by Alg8 and Alg44 (not verified)
6) Polymer enter the periplasmic space
7) Alginate becomes acetylated by AlgI, AlgF and AlgV= AlgJ in *Pseudomonas* species
8) The transport is facilitate through the periplasmic space by AlgK
9) The polymer is exposed to an alginate lyase (AlgL) and an epimerase (AlgG)
10) The acetylated and epimerized alginate is exported through the outer membrane by AlgJ= AlgE in *Pseudomonas* species
11) Further epimerisation by the AlgE-family (only in *A. vinelandii*)

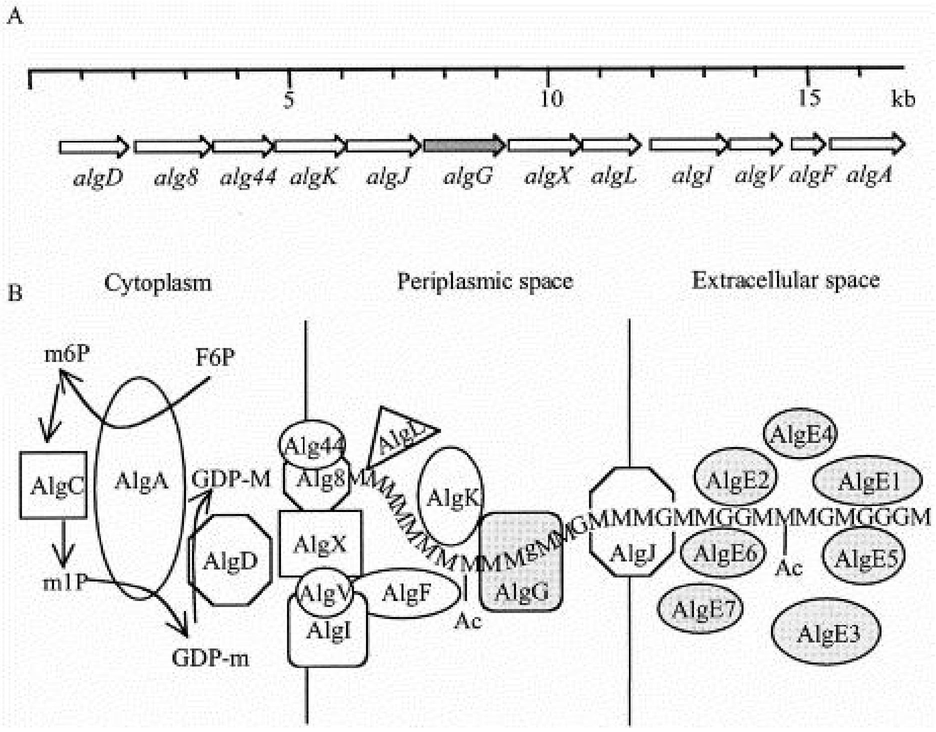


Fig. 30: A.) The alginate biosynthetic gene cluster; B) Schematic model of alginate biosynthesis in *A. vinelandii* [166]. Similar proteins for alginate biosynthesis were found in *Pseudomonas* species except the extracellular C5-epimerases. AlgG is the only found epimerase in the biosynthetic gene clusters. All the extracellular epimerases except AlgE5 are in a separate gene cluster.

1.2.7 Alginate epimerases of *Azotobacter vinelandii*

AlgG is able to epimerize M→G but two or more mannuronic acids have to be between each conversion [167].

Beside AlgG, *A. vinelandii* produces 7 extracellular C-5 alginate epimerases called AlgE1-7 (Fig. 31) [168,169]. All the members of the AlgE family are needed to form a cyst- an alginate wall surrounding the cells in a particular differentiated resting stage [170].

All extracellular alginate epimerases except AlgE5 are clustered together. In this genome cluster one additional protein is encoded, designated as ORF-9. Additionally, the protein AlgY shares sequence similarity to the AlgEs but seem to have no catalytic activity on alginates [169]. All 7 epimerases consist of two highly homologous modules designated A- (~385 amino acids) and R- (~155 amino acids) module and the last 20 amino acids of each epimerase are an unstructured secretion signal [168,171]. Both modules feature Ca^{2+} binding motifs [172,173] and calcium-ions are necessary for activity. In contrast, AlgG does not need calcium for activity [165].

All members of the AlgE family start with one A-module and AlgE1 and AlgE3 have a second one in the sequence. The number of R-modules vary from 1 (AlgE4) to 7 (AlgE3). A-modules are enzymatically active but R-modules enhance the activity of an A-module if at least one is bound after an A-module [172]. ORF9 consists of 7-R-module but lacking an A-module therefore it is presumed to be inactive.

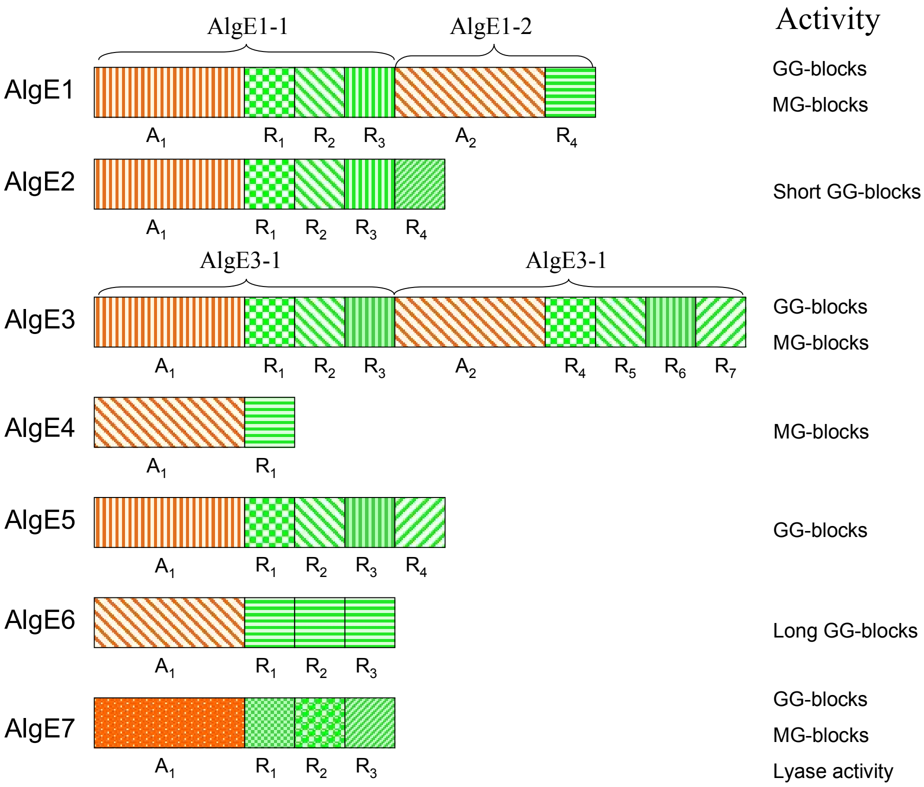


Fig. 31: ***Azotobacter vinelandii*** **expresses a family of extracellular alginate epimerases which only consists of two different modules named A- and R-module.** Only the A-modules (in orange) are catalytically active but the R-modules (in green) enhance the activity significantly if bound to an A-module. The different members of the family show different epimerisation products, given on the right side of the figure, AlgE7 acts also as a lyase.

Each epimerase converts M→G at polymer level in a unique pattern [174]. AlgE1 and AlgE3 have two A-modules and therefore they have two separate catalytic domains. One catalytic domain makes preferably MG-blocks (in AlgE1: A_2-R_4 and in AlgE3: A_2-R_4-R_5-R_6-R_7) and the other GG-blocks (A_1-R_1-R_2-R_3) [174,175]. Surprisingly the MG-blocks produced by the one catalytic site are not a substrate for the GG-block formation [176]. AlgE4 is the only epimerase that alters poly-M to MG-blocks but cannot form GG-blocks [177]. All the other extracellular epimerases make GG-blocks in different length. AlgE7 has an epimerisation and lysis activity. It is a bifunctional enzyme but with only one active site [178].

The structures of the A- and R-module of AlgE4 could be determined. The structure of the A-module of AlgE4 was determined by X-ray crystallography (Fig. 32) [179].

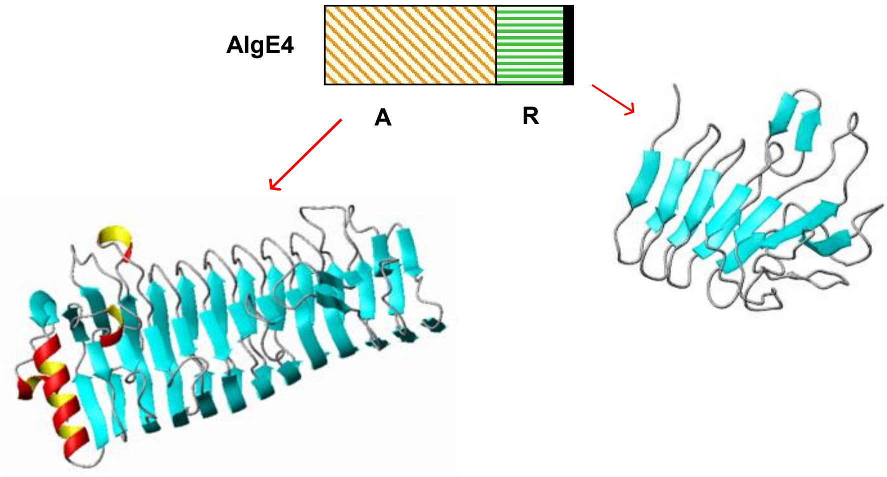


Fig. 32: Structure of A-module and R-module of AlgE4. The structure of the A-module was solved by X-ray [179] while the R-module was solved by NMR [173]. In both modules the main secondary structure elements are parallel β-strands organized as a 4-stranded β-helix or 2 stranded β-roll respectively. The black bar in the schematic draft symbolizes the unstructured tail.

The A-module is a 4-stranded β-helix. At the front side to the A-modules has a positively charged groove (Fig. 37 and Fig. 33). This groove ends in the active site close to the N-terminal end of the A-module. The catalytic site consists of the amino acids Y149, D152, H154 and D178. It was possible to co-crystallize inactive A-module with mannuronan trisaccharide (M3) and a reaction mechanism was proposed (Fig. 35) [179].

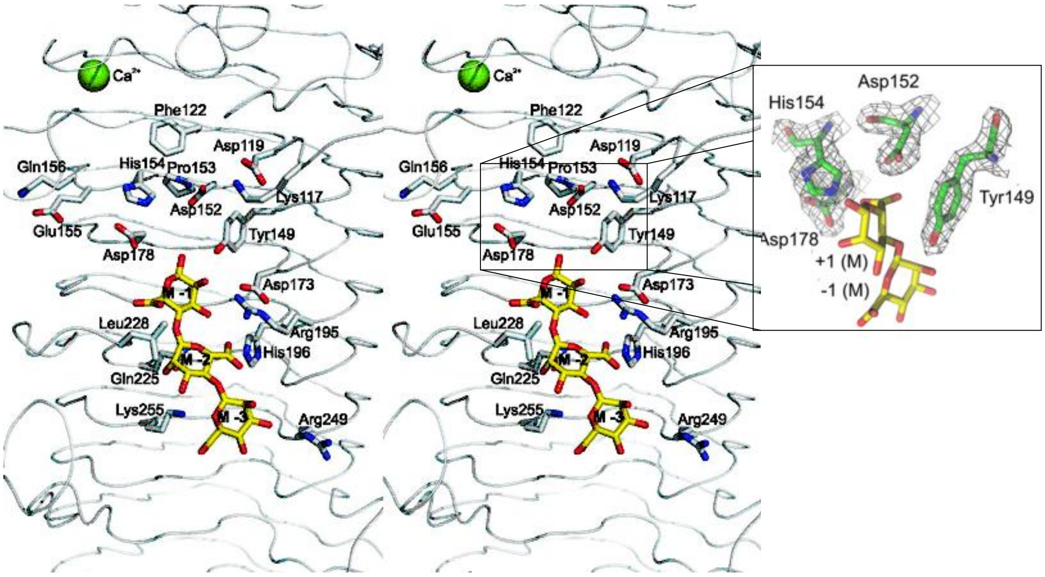


Fig. 33: Co-crystallisation of the A-module with mannuronan trisaccharide (M3) [179]. The negatively charged saccharide binds to a positively charged groove consisting of many basic amino acids like R195, H186 or K255. The positive groove ends in the active site consisting of the four amino acids Y149, D152, H154 and D178. The enlargement shows the simulation of the native active site with the substrate**.**

The reaction mechanism (Fig. 35) aligns with experimental data [172,178] and theoretical considerations [180]. In 1987 a reaction mechanism was postulated under the assumption that alginate epimerisation and lysis follow the same pathway (Fig. 34).

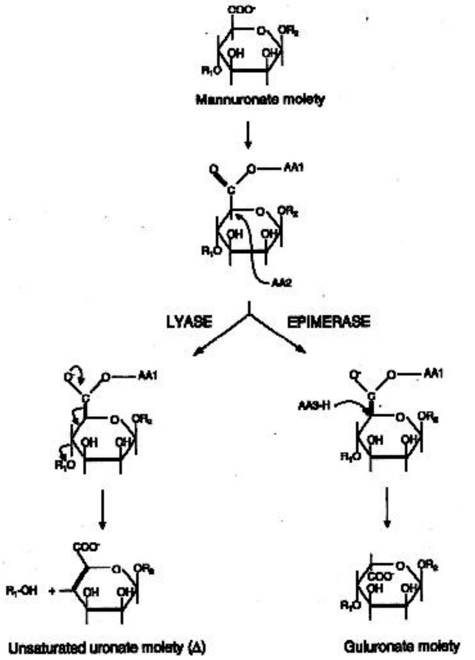


Fig. 34: The proposed reaction mechanisms for alginate lyases and epimerases [178,180] AA1-3 refers to amino acid residues on the enzyme. The negative charge on the carboxylate anion is believed to be neutralized by a positively charged amino acid (AA1) in the active site of the enzyme. Note that the abstraction of the H-5 proton is believed to occur from below the sugar plane, whereas the replacement occurs from above. For simplicity the protons are omitted from the sugar rings.

Additionally ^{3}H-5-release from a ^{3}H-5-labelled poly-mannuronan per conversion from M→G could back up the suggested epimerisation reaction [181].

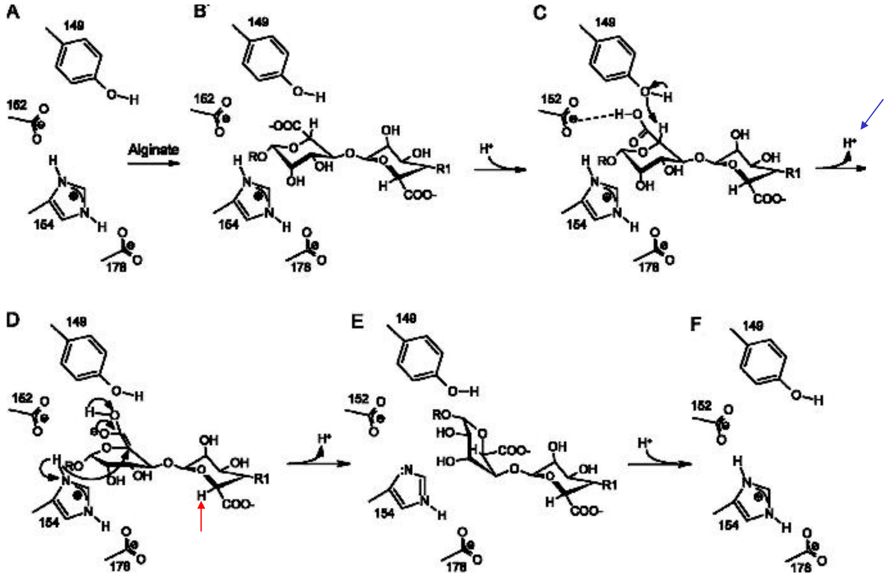


Fig. 35: The postulated reaction mechanism [179]. AlgE4*A* and *B*, the alginate polymer enters the catalytic site. *B* and *C*, the carboxylate moiety of the mannuronic acid in subsite +1 is protonated, enabling it to form a hydrogen bond with Asp^{152} (and/or 178), which stabilizes the substrate-enzyme complex. *C*, upon deprotonation of Tyr^{149} (via Arg^{195}) the alkoxide ion group extracts H-5 from the *re*-face of the mannuronic acid in subsite +1. *C* and *D*, a double bond is formed, which makes the conformation of the +1 mannuronic acid partially planar. *D*, the protonated His^{154} performs a nucleophilic attack on the C-5 atom of the +1 sugar from the *si*-face with the concomitant flip of the +1 sugar ring into the $^{1}C_{4}$ chair conformation of guluronic acid. *D* and *E*, the carboxylic

acid moiety on sugar +1 is deprotonated. *E* and *F*, the epimerized sugar leaves the active site and His[154] is protonated again. *F*, the epimerase is ready to perform a new reaction. The experimental data confirm the ^{3}H-5 release into the medium (here indicated with a blue arrow) [172]. If two or more mannuronic acids in a row are epimerised after each other, a 180° reorientation between each epimerisation step would be necessary. (red arrow indicate the second H-5)

The 3D structure of R-module of AlgE4 was solved by NMR (Fig. 32 and Fig. 36) [173]. All R-modules have 4 to 7 imperfect 9 amino acids long repeats motif called RTX-motif (Repeat in ToXin-motif) at the N-terminal end [168,182]. The sequence of the RTX-motif consists of the nona-peptide GGXGXDZUZ. The glycines and aspartic acid are highly conserved. U is mainly leucine but can also be replaced by isoleucine, valine or phenylalanine. X stands for any amino acid but mainly with short side chain and Z is for amino acid with long side chains. The first 6 amino acids form a tight loop which also binds Ca^{2+} ions whereas the last 3 amino acid form short β-strand (Fig. 36). 2 RTX-motifs make a full turn and several RTX-motifs after each other form a 2 stranded β-roll with a calcium binding motif. Ca-ions are incorporated into the structure and in absence of Calcium the proteins are intrinsically disordered [183].

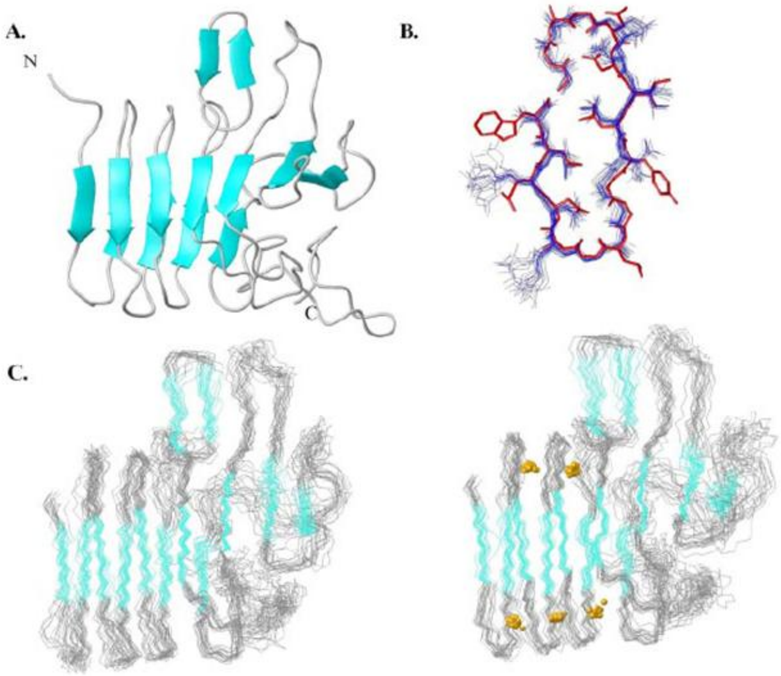


Fig. 36: A) Ribbon drawing of the energy-refined R-module structure. B) Structure of 2 RTX-motifs or a whole turn. Structure of heavy atoms in the loop region of residues 27-44 for the 20 best energy-refined conformations of the R-module (*thin lines*) and residues 351-368 of *P. aeruginosa* (*thick lines*). Both loops fixate Ca-ions above and below with the previous and following turns. C) A bundle of the 20 energy-refined conformations of the R-module in the region of amino acid residues 5-145. Shown are the R-module without calcium ions (*left*) and the R-module with calcium ions incorporated in the loops (*right*). β-sheet regions are coloured in *cyan* and calcium ions are depicted as *orange spheres* [173].

The R-module of AlgE4 has no epimerisation activity; nevertheless it binds poly-M alginates [173]. It is assumed that the R-modules help to orient the alginate chains for the active site and so increase the reaction rate. The spatial orientation of the two modules to each other is not determined but the suggestion is based on the binding site of both modules (Fig. 37). The charge distribution on the surface substantiates the suggestion. On the front across both modules the surface is positively charged while the back of both modules is negatively charged.

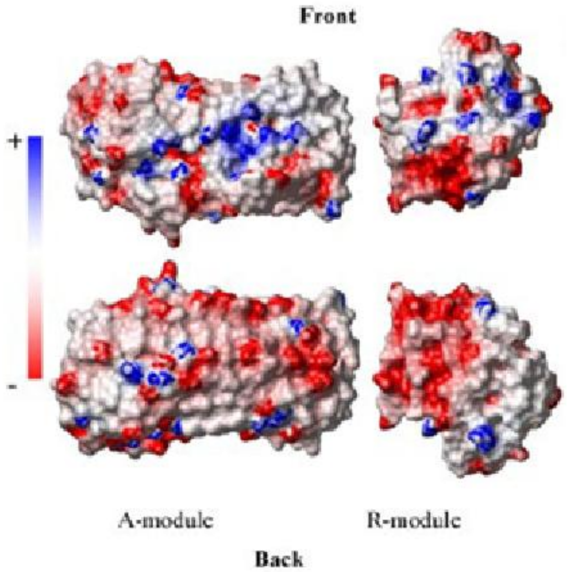


Fig. 37: Electrostatic surface potential of the AlgE4 A- (*left*) and R-modules (*right*) [173]. *Red* and *blue*, respectively, denotes regions of negative and positive potential on the protein surface.

The β-roll is an unusual structure conformation shared by some secreted proteins. These secreted proteins belong to the type I secretion system (TISS) and share beside the RTX -motifs, a C-terminal signal peptide that is not cleaved after transportation [184]. The transportation through the periplasmic space happens by the so called ABC-transporter (ADP-binding cassette) (Fig. 38) [185]. The proteins are unfolded during the secretion and refolded in the medium. The calcium concentration outside the cells is normally sufficient enough to refold into active enzymes. Recently the gene for the ABC-transporter in *A. vinelandii* could be identified [171].

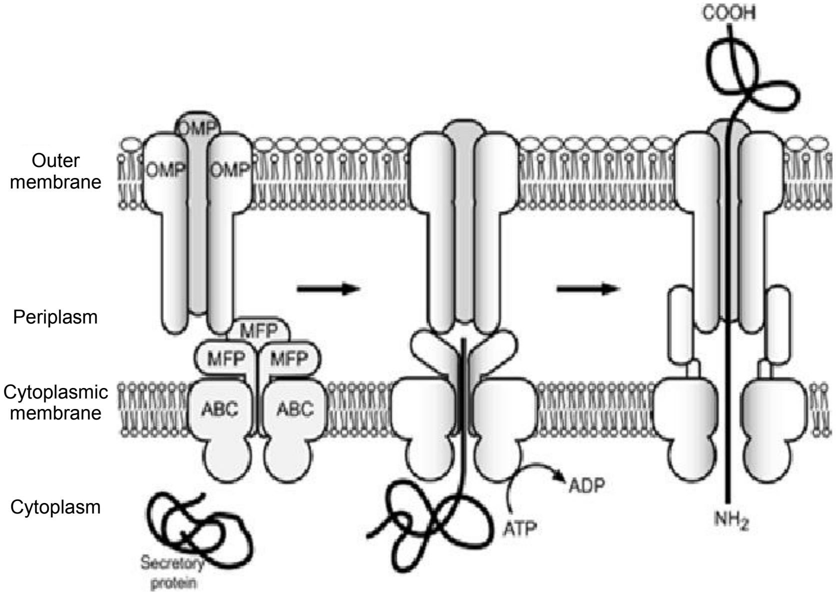


Fig. 38: Type I secretion across the cell wall depends on three specific proteins [184]: a polytopic inner membrane protein with a cytoplasmic ATPase domain operating as an ABC exporter, a membrane fusion protein (MFP) and an outer membrane protein (OMP). Upon recognition of a C-terminal secretion signal on a given RTX protein translocation substrate, the inner membrane complex formed by an energy-providing ABC transporter and a MFP contacts the trimeric OMP. A sealed channel–tunnel assembly spanning across the entire Gram-negative bacterial cell envelope is formed, through which the RTX protein is exported in a single step from the bacterial cytoplasm directly to the external bacterial surface, without transiting through the periplasmic space. The calcium concentration in the cytoplasm (<100nM) and outside the cell (~ mM) helps to unfolded and refold the secreting protein.

1.2.8 Alginate epimerisation mode of action

Epimerisation by the AlgE-family was subject to many studies. The epimerisation itself was studied but also the possibility to get well defined alginate gels and their physical properties were explored. The AlgE family can epimerise algal and bacterial alginates. Epimerisation from M→G by the AlgE family is not a random event [186] and two different modes of action were suggested to explain the experimental data. The processive mode of action can be described by binding to the alginate and converting a number of residues before dissociation. This mode of action is often visualized as protein sliding along the alginate chain. AlgE4 was suggested to epimerize by a processive mode of action [177,187]. On average 10 epimerisation events occur before the epimerase dissociates from the alginate polymer [188]. Experimental data indicate that AlgE6 introduce GG-blocks into alginate by a processive mode. But if two or more mannuronic acids in a row are to be epimerised after each other, those epimerase would have to turn 180° between each epimerisation step (Fig. 35). Processive mode is only possible if one AlgE6-protein slides along the alginate polymer incorporating MG-blocks while a second enzyme epimerizes the MG-blocks to GG-blocks [176]. The other mode of action suggested was the preferred attack where the binding to a certain block (e.g. MM-block) is preferred but between each epimerisation a dissociation and re-association occurs. AlgE2 seems to add more Gs to existing, single G by a preferred attack mode [187,189]. However it hardly converts MG-blocks to GG-blocks (GMG→GGG). AlgE5 is highly homologous to AlgE2 (98% sequence identity), nevertheless AlgE5 prefers MG-blocks as substrate to produce GG-blocks [176]. Both epimerases have lyase activity but it is very low (1-3 breaks per 1000 epimerisation) compared to AlgE7 where the average degree of polymerization (DP) of the alginate oligomers after complete lysis is 4. [176,189].
The length of poly-M alginate needed to obtain any epimerisation seems to be independent of the alginate epimerase. The length of poly-M-alginate versus the activity of the epimerases was tested for AlgE1, AlgE2, AlgE4 and AlgE6 [176,187]. No measureable activity was detected for oligomers <7. There is activity measureable for 8 and a large increase in activity was detected for alginate oligomer >10 (Fig. 39 A) [187]. In a more detailed study, different long alginate oligomers were epimerised by AlgE4 and afterwards exposed to G-lyase and the lysed fragments were analysed by MS Fig. 39 B) [188]. This lyase is specific for guluronic acid in GG and GM glucosidic linkage. The minimal length needed for epimerising one M→G is 6. Neither the first 2 nor the last three mannuronic acids from a polymer are epimerised independently of the length of the alginate chain (Fig. 39 C).

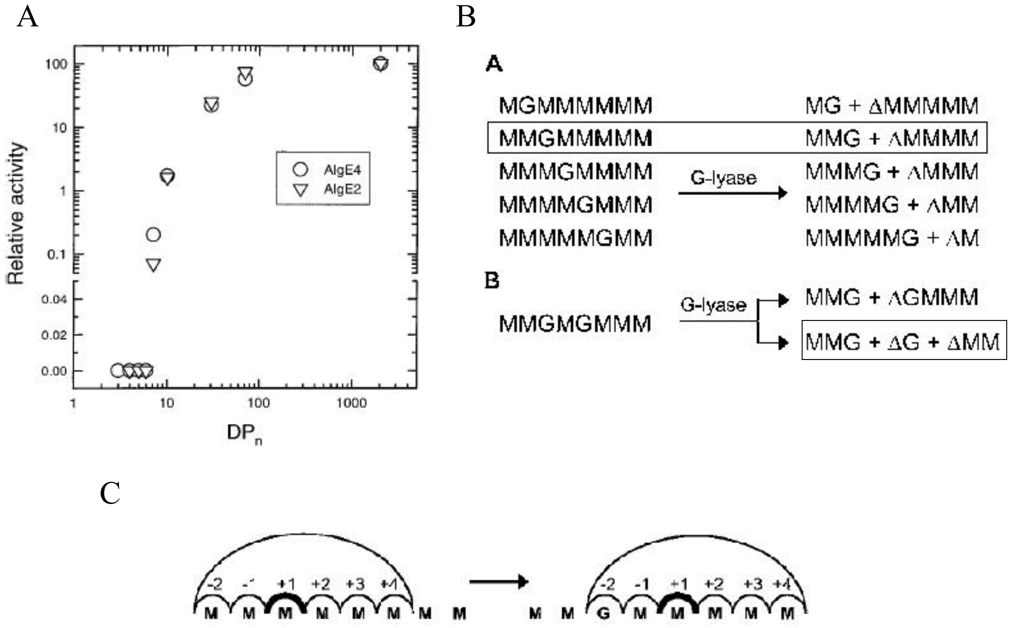


Fig. 39: A) The relative activity correlates with the length of the oligomer (DPn=degree of polymerisation) but is to be independent of alginate epimerase. B) The possible fragments of an epimerized octamer after treatment with G-lyase. The rectangle show the fragments obtained experimentally. C) Tentative subsite model for AlgE4 for two consecutive epimerisations on an octamer (the hexa- and heptamer have only one G at the position 3) [187,188].

1.2.9 Modified alginates by extracellular alginate epimerases *in vitro*

The family of extracellular alginate epimerases (AlgE1-AlgE7) from *A. vinelandii* can modify alginates *in vitro*. They are stabile in buffer solution and the only requirement is 3-5 mM calcium to perform *in vitro* epimerisation.

The epimerisation rate and activity *in vitro* of the different alginate epimerases is influenced by temperature, pH, buffer, starting material and the concentration and kind of mono- and divalent cation in solution [177,189,190]. All extracellular alginate epimerases of *A. vinelandii* have a pH optimum around 7 (6.5 -7.3) and the optimal temperature is around 37 °C (30-45 °C) (Fig. 40 A). The buffer and small additives influence the epimerisation rate greatly (Fig. 40 B). There seems to be a negative correlation between pKa-value of the buffer and the epimerisation rate. Positively charged buffer ions are attracted by the negatively charged alginate polymer (Fig. 40 B).

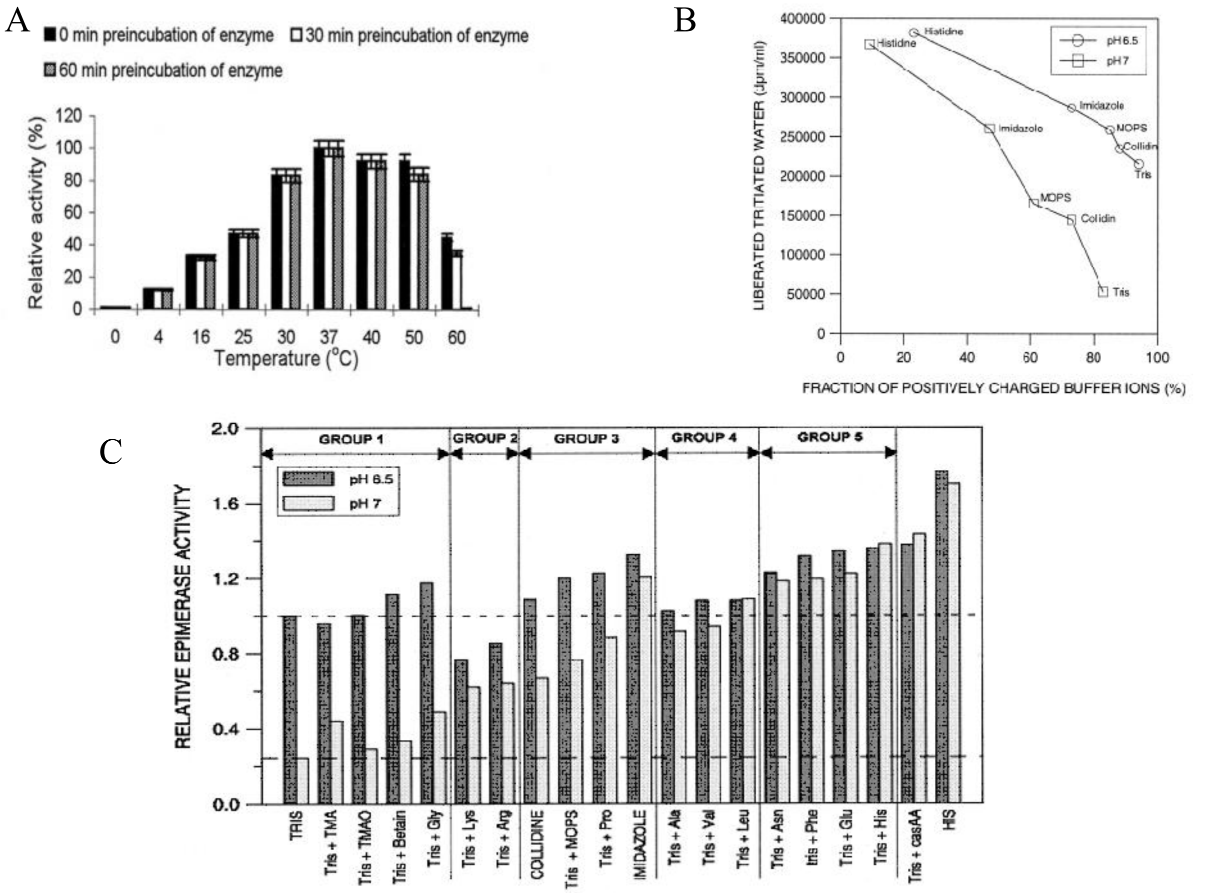


Fig. 40: A) Alginate epimerases have a relative broad temperature optimum around 37°C. The results shown here are for AlgE4 [177]. Results obtained for AlgE2 show a slightly higher optimum (45 °C). B) There is a inverse correlation between the fraction of positively charged buffer ions and epimerisation rate at a certain temperature and pH-value [190]. C) Activity of AlgE2 in Tris buffer supplemented with different compounds was measured at pH6.5 and pH7 on M-rich alginates and 3.3 mM $CaCl_2$. Nearly all additives results in a significant increase of activity at pH 7. At pH 6.5 the activity of epimerase depends less on additives. Positively charges amino acids have cause and decrease in activity while all other tested compounds have either no effect or slightly positive effect on the activity. Group 1: small positively charged or neutral compounds that resemble Tris in structure; Group 2: positively charged amino acids; Group 3: neutral or partially positive compounds with a ring structure; Group 4: uncharged amino acids; Group 5: negatively charged or aromatic amino acids; casAA: casaminoacids

Monovalent cations (expect Rb^+ and Cs^+) at low concentration increase the epimerisation rate (tested with AlgE2) probably because they can shield the negative charges of the alginate and protein and prevent so repulsion but don’t hinder binding. They have a positive influence with or without Tris in the buffer solution (Fig. 41 A + B). At higher concentration, NaCl can have negative effect on epimerisation activity (Fig. 41 C + D) but this depends also on the Ca^{2+}-concentration and AlgE epimerase. AlgE4 has the highest activity in absence of NaCl but addition of up to 100 mM does not alter the activity much (1.5 mM Ca^{2+} and MOPS buffer) [177]. In contrast, AlgE1 has an activity optimum between 100 and 200 mM NaCl (3 mM Ca^{2+}, MOPS) [175].

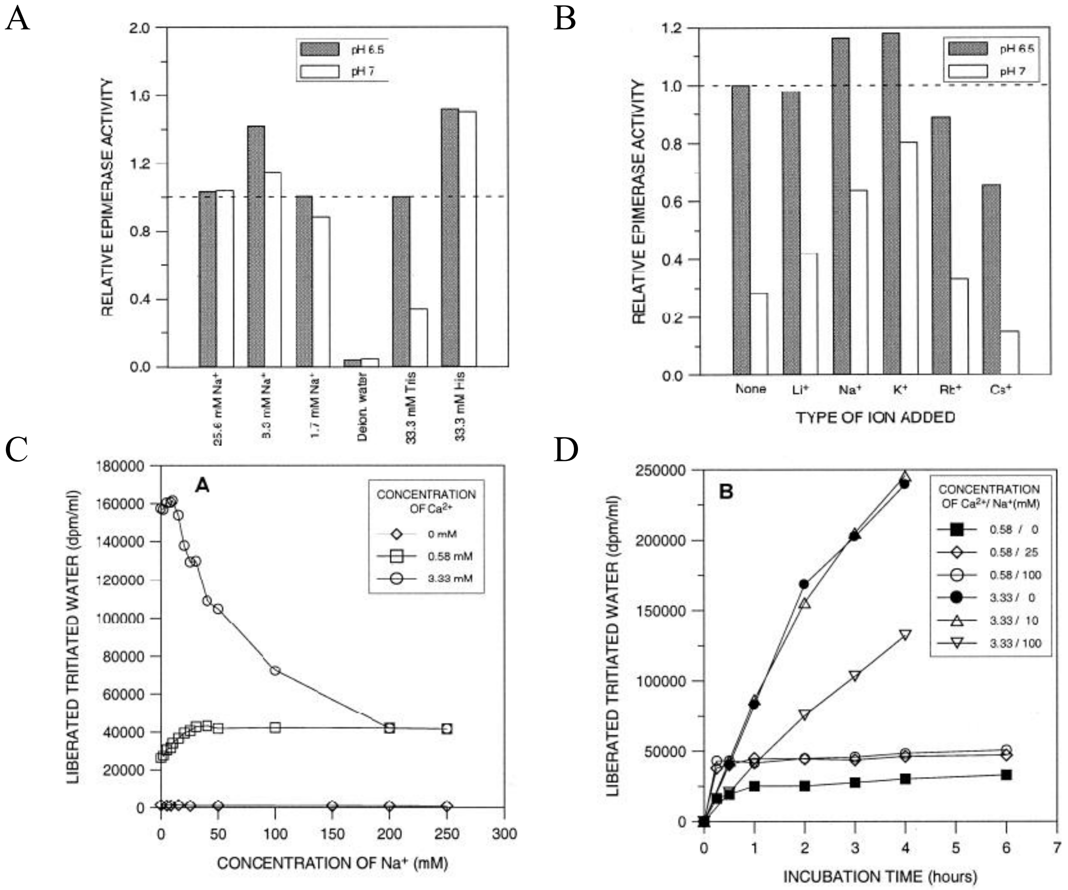


Fig. 41: A) Relative activity of AlgE2 in different reaction solution. As substrate M-rich alginate was used and [Ca^{2+}] = 3.3 mM. Na+-ion have always a positive effect at pH 7 compared to Tris. At pH 6.5 a significant positive effect can only be measured at Na^+-concentration = 8.3 mM. B) Relative activity of AlgE2 with different monovalent cations. As substrate M-rich alginate was used and the buffer contained 50 mM Tris and 3.3 mM $CaCl_2$. The monovalent cation concentration was 33.3 mM. At pH 7 only Cs^+ has a negative effect on the activity while at pH 6.5 on Na^+ and K^+-ions enhance the epimerisation activity. C and D) at low amount of Ca^{2+}, addition of Na+-ion (> 20 mM) increase the reaction rate. At higher concentration of Ca^{2+}, Na^+-ion inhibited the epimerisation [190].

The strongest effect on epimerisation has the concentration of divalent cations and especially Ca^{2+} (Fig. 42 A). It is known that A- and R-modules bind Ca^{2+}-ions and without them at least the R-module is unstructured and no epimerisation occur. At low Ca^{2+}-concentration (<1.5 mM) the epimerase activity correlates with the Ca-concentration. Above the optimal Ca^{2+} concentration epimerase activity is independent (Fig. 42 C). At low Ca^{2+}-concentration epimerisation is stopped within the first 2 hours. Interestingly the addition of Ca^{2+}-ions (final concentration is twice original concentration) does not restart epimerisation. Addition of new alginate initiates epimerisation for some hours (Fig. 42 D). For the different epimerases the optimal Ca^{2+}-concentration is between 1.5- and 3 mM. Sr^{2+} is the only cation known where the AlgE epimerases have some activity in absence or low concentration of Ca^{2+} (Fig. 42 B) [175,177,190]. AlgE epimerases have Ca^{2+}-binding motifs and it is assumed that Sr^{2+}-ions can replace some bound Ca^{2+}-ions at low Ca^{2+}-concentration and therefore keep the structure and activity intact. In presence of Ca^{2+}-ions, all other divalent cations have a negative effect on the epimerisation, probably because these cations bind to alginates and cause gel formation (Fig. 42 A).

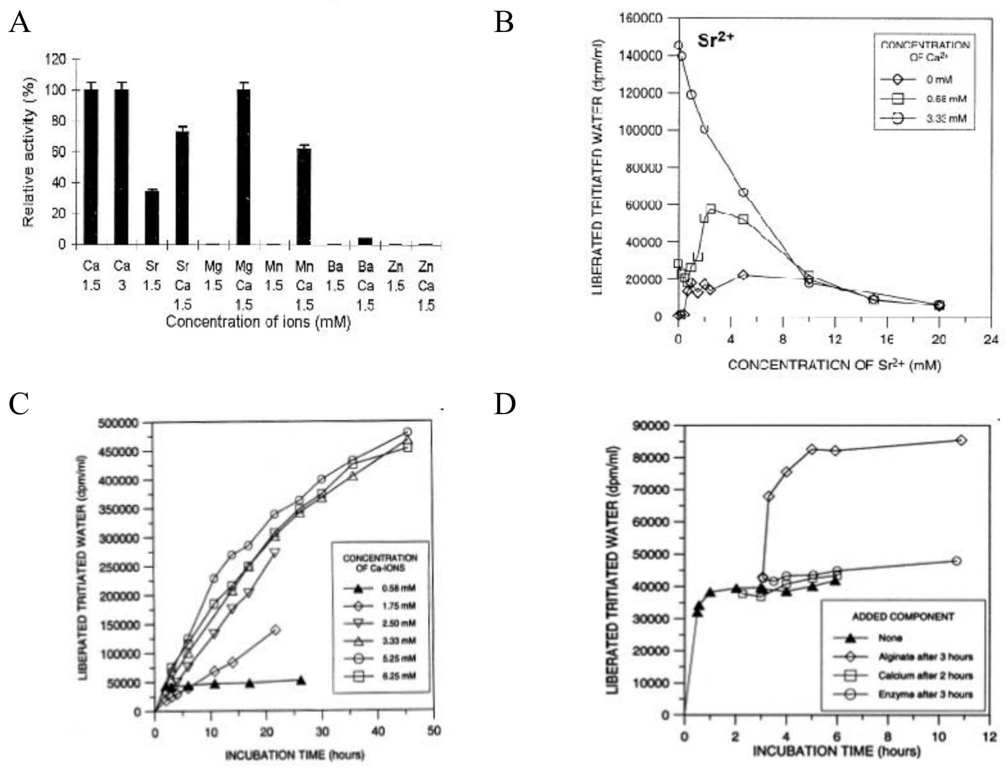


Fig. 42: A) Epimerisation activity of AlgE4 depending on different divalent cation concentration in 10 mM MOPS pH 6.8 [177]. Similar results were found also with other epimerases. When Ca^{2+}- ions are omitted only Sr^{2+} ions can support some activity. In presence of Ca^{2+}-ions all other divalent ions had a negative effect on the activity. B) Effect of Sr^{2+} concentration on epimerase activity (AlgE2) at different Ca^{2+}-concentration. At below minimal Ca^{2+} concentration (see also diagram D) Sr^{2+} concentration up to 4 mM have a positive effect on the epimerisation. Above the optimal Ca^{2+-}concentration, Sr^{2+-}ions hinder epimerisation activity. C) Above optimal Ca-level epimerisation rate is independent of Ca-concentration. At very low Ca-level (0.58 mM) the epimerisation is terminated within the first two hours. In between epimerisation activity is steady but reduced. D) At low Ca-concentration (0.58 mM) epimerisation is stopped. Surprisingly addition of Ca-ions does not restart epimerisation. Addition of new alginate initiates epimerisation some hours. All compounds were added to twice original concentration[189].

The AlgE-family epimerizes any given alginate completely if external factors like acetylation don't prevent this. In particular AlgE1, AlgE4 and AlgE6 are technically useful as the have no lyase activity. AlgE4 converts poly-M entirely to MG-blocks. AlgE1 and AlgE6 generate long GG-blocks. Alginates with long GG-blocks are difficult to produce. Nevertheless it was possible to obtain high molecular weight alginates with 97% G.

1.3 NMR -protein NMR

In the following an overview of NMR spectroscopy on proteins is given. The reader is expected to be familiar with the basics of NMR; the focus of this chapter will be on techniques and strategies for assignment and structure determination of medium to larger sized protein with the use of isotope labelling (^{2}D, ^{13}C, and ^{15}N). The literature used in this section is based on textbooks [191,192] and the references will not be repeated throughout this section.

1.3.1 Coupling of the spin

In a molecular framework, nuclei are closely associated and therefore they affect each other. That results in two types of coupling which either can be through bond (scalar coupling) or through space (dipolar coupling). In the following each type of coupling will be elaborated.

a) Scalar coupling

Scalar coupling arise due to indirect interaction between two nuclear spins that are connected by a network of covalent bonds. Magnetization transfer through bonds only occurs if the electron occupies an s-orbital and originates from Fermi contact interaction. Scalar coupling can be homonuclear or heteronuclear. The magnitude of the scalar coupling J is independent of the external magnetic field but it is affected by the magnetic field of close nuclei. The magnitude depends on the nuclei involved, bond distance and spatial orientation of the nuclei to each other. The coupling constant is denoted ^{1}J, ^{2}J, ^{3}J… where the superscript number indicated the number of bounds between the two coupling nuclei. A list for ^{1}H-^{1}H coupling constants can be found at Tab. 2.

Tab. 2: ^{1}H-^{1}H coupling constants

	Magnitude [Hz]
^{3}J- average	~ 7Hz
Vicinal J_{cis}	~ 12 Hz
Vicinal J_{trans}	~ 19 Hz
Geminal J	~ 3Hz
Aromatic J_{ortho}	~ 7 Hz
Aromatic J_{meta}	~ 3 Hz
Aromatic J_{para}	> 1Hz

Usually, coupling over more than three chemical bonds is not observed. The scalar coupling between two nuclei over 2, 3 or even 4 bounds resulting in the splitting of the signals into doublets or other multiplets. If the rotation around a bound is hindered the spatial orientation has an influence on the coupling constants (see Tab. 2 J_{trans} and J_{cis}). In proteins the rotation around a bound (especially backbone) is limited. The ^{3}J coupling constant correlates with the dihedral angle (Φ) between the coupling nuclei. The correlation between Φ and magnitude of ^{3}J is based on empiric data but can be predicted (Eq.1).The relation between Φ and ^{3}J is illustrated in the Karplus curve as in Fig. 43

$$^{3}J = A\cos^{2}\Phi + B\cos\Phi + C \qquad 1$$

Where

^{3}J	coupling constant [Hz]
Φ	dihedral angle [°]
A, B, C	constants

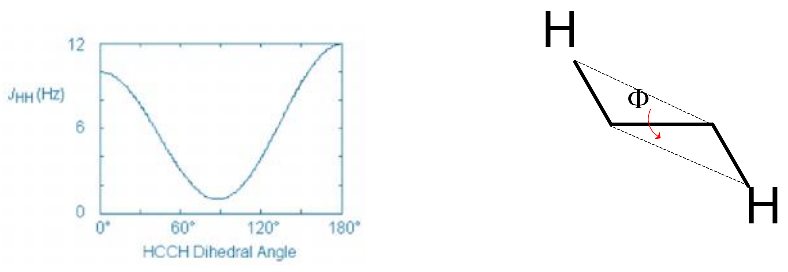


Fig. 43: The $^3J_{HH}$ coupling constant depends on the dihedral angle and can be predicted. At 90° the magnitude of ^3J is minimal.

The strong heteronuclear ^1J and ^2J coupling are utilized to create correlation spectra between nuclei found in the protein backbone such as $^1H^N$, ^{13}C (C', C^α and C^β) and ^{15}N. Magnetization transfer through J-coupling in combination with multidimensional spectra has become the standard approach today in protein assignment (see also INEPT). The typical ^1J and ^2J coupling constants found in ^{15}N and ^{13}C isotope labelled protein backbone are shown in Fig. 44.

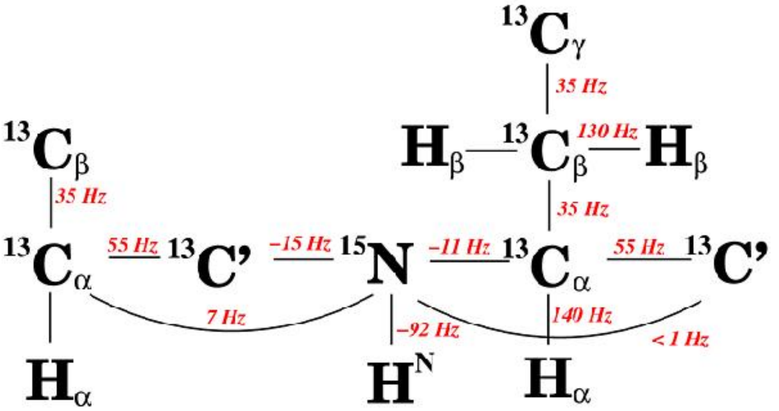


Fig. 44: ^1J and ^2J coupling constants of ^{15}N and ^{13}C isotope labelled protein.

b) Dipole-dipole coupling

If two magnetic moments are close in space they will interact with each other. The magnitude is proportional to the inverse third power of the distance and the product of the gyromagnetic ratios.

$$d = \frac{\hbar\mu_0}{4\pi}\cdot\frac{\gamma_1\gamma_2}{r^3} \tag{2}$$

Where

d	dipole-dipole coupling constant [Hz]
ħ	Planck's constant [$m^2 \cdot kg/s$]
μ_0	magnetic permeability in vacuum [$kg \cdot m/(A^2 \cdot s^2)$]
γ_1, γ_2	gyromagnetic ratio [rad/(s·T)]
r	distance between the two spins [m]

The angle θ describes the orientation between the dipole-dipole vector r and the external magnetic field (Fig. 45). In a strong magnetic field, the dipolar coupling constant depends on the orientation of the internuclear vector with the external magnetic field.

$$D \propto 3\cos^2\theta - 1 \tag{3}$$

D	dipolar coupling constant [Hz]
θ	angle between the dipole-dipole vector r and the external magnetic field B_0

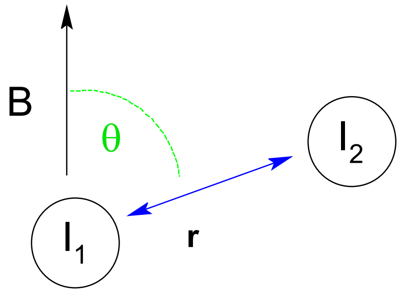


Fig. 45: Dipolar coupling depends on the angle between the external field and dipolar coupling vector. Dipolar coupling becomes 0 at an angle of θ = 54.7° also called magic angle.

The molecular reorientation in solution causes the angle θ to fluctuate and so does the dipolar coupling. Due to the fact that this constant reorientation of the molecule is very rapid on the NMR timescale and the fact that the term $3\cos^2\theta\text{-}1$ averages to zero, when all orientations are populated equally, the resulting dipolar coupling will be zero on time-average. Therefore, no line-splitting will be observed in the spectra from dipolar coupling in solution state NMR. A set of experiments known as residual dipolar coupling (RDC) measurements re-introduces dipolar coupling by partially aligning the (protein) molecules in solution. Orientation constraints obtained from the spectra can be used for protein structure calculation.

1.3.2 Relaxation

Relaxation describes every process that causes loss of signal.
Two main relaxation pathways are distinguished. One describes the loss of signal as the spins return to equilibrium (longitudinal relaxation) while the other relaxation pathway (transverse relaxation) describes the loss of phase coherence.

a) Longitudinal Relaxation (spin-lattice)
The longitudinal relaxation describes the recovery of the *z* component of the nuclear spin magnetization, M_z, towards its equilibrium value, M_0. The recovery of the magnetization can be expressed as an exponential function (Eq. 4)

$$\frac{dM_z}{dt} = \frac{M_z - M_0}{T_1}$$
$$M_0 - M_z(t) = Ae^{-\frac{t}{T_1}} \qquad 4$$

M_0	macroscopic magnetization at equilibrium
M_z	z-component of the macroscopic magnetization as a function of time
T_1	longitudinal relaxation time [s]
A	constant

T_1 relaxation involves redistributing the populations of the nuclear spin states in order to reach the thermal equilibrium distribution. Different relaxation mechanisms allow nuclear spins to exchange energy with their surroundings, the lattice, allowing the spin populations to equilibrate. Therefore longitudinal relaxation is also called spin-lattice relaxation.

T_1 depends on the magnetic field strength in bigger molecules like proteins and molecular mobility of the molecules. Paramagnetic centers have a great impact on T_1. Removing of dissolved oxygen by degassing gives longer T_1 values.

Longitudinal relaxation is measured by the inversion recovery experiment (Fig. 46). A 180° pulse inverts the net magnetisation to $-M_0$. ($M_z(t = 0) = -M_0$). The recovery follows then

$$\begin{gathered} M_z(t)=M_0(1-2e^{\frac{t}{T_1}}) \\ M_z(t)=0 \Rightarrow t=T_1\ln 2 \end{gathered} \qquad 5$$

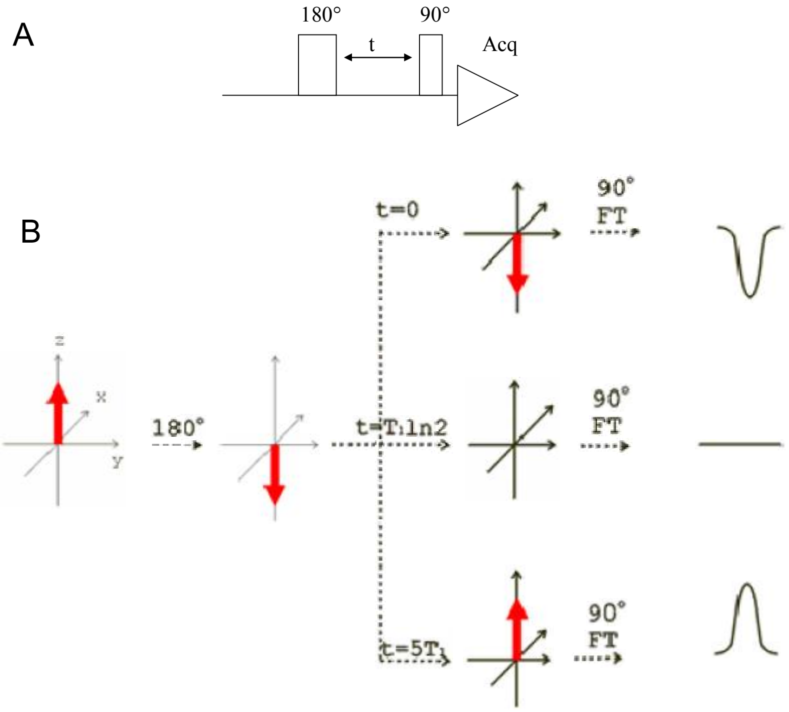


Fig. 46: Inverse Recovery; Pulse program consists of one 180° pulse and a delay time t. The 90° pulse is necessary to measure the NMR signal. The 180° pulse inverts the net magnetisation M_0. If t = 0 all signals are negative. If the delay time is increased the signals become smaller and smaller and by become zero at $t = T_1\ln2$, thereafter they become positive and increase in strength. Fitting signal intensity as function of the delay time t to Eq. **5** yields a value for T_1.

b) Transverse Relaxation (spin-spin)

Transverse relaxation describes the loss of phase coherence (Fig. 47). The main sources of transverse relaxation are dipole-dipole relaxation, scalar relaxation, spin-rotation, quadrupole relaxation (for nuclei with spin > ½ and is not relevant for protein NMR) and CSA relaxation.

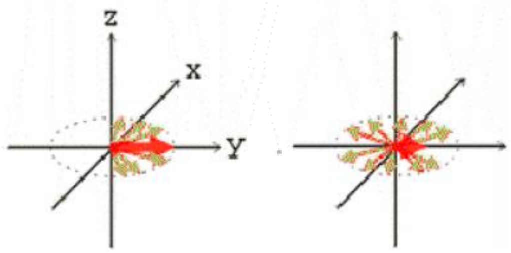


Fig. 47: Transverse relaxation describes the loss of phase coherence as the nuclei precess in the xy-plane.

$$M_{x,y} = M_{x,y}(t=0) \cdot e^{-\frac{t}{T_2}}$$

6

$$R_2 = \frac{1}{T_2} = \frac{1}{T_{DD}} + \frac{1}{T_{SR}} + \frac{1}{T_Q} + \frac{1}{T_{CSA}} + ...$$

$M_{x,y}$	transverse magnetization as a function of time
$M_{x,y}(t = 0)$	initial transverse magnetization
T_2	transverse relaxation time [s]
R_2	transverse relaxation rate [s^{-1}]
T_{DD}	dipole-dipole relaxation [s]
T_{SR}	scalar relaxation [s]
T_Q	quadrupole relaxation [s]
T_{CSA}	chemical shift anisotropy relaxation [s]

Molecular tumbling causes fluctuation in the local magnetic field experienced by the nuclei. This fluctuation is the source of transversal relaxation in proteins. Dipole-dipole relaxation is usually the dominating relaxation mechanism for nuclei with spin ½.

Transverse relaxation is measured by the CPMG pulse sequence which contains repeated spin-echo elements. In principle, T_2 can be obtained by measuring the signal width at half-height however the line width is often dominated by field inhomogeneity. Relaxation rates extracted from the line width are called $1/T_2^*$.

$$\frac{1}{T_2*} = \frac{1}{T_2} + \frac{1}{T_2^{\#}}$$

7

T_2^*	apparent T_2 based on line width of signals [s]
T_2	transverse relaxation [s]
$T_2^{\#}$	apparent relaxation due to the contribution of field inhomogeneity [s]

1.3.3 Molecular motion and Spectral density Function

The molecular reorientation in solution can be described by the correlation time (τ_c), which is the average time taken for the molecule to rotate by one radian in the solution. The frequency distribution of the fluctuating magnetic field caused by molecular tumbling is called spectral density J(ω).

$$J(\omega) = \frac{2}{5} \frac{2\tau_c}{(1+\omega^2\tau_c^2)}$$

8

Where

J(ω)	spectral density function [s/rad]
τ_c	correlation time [s/rad]
ω	angular frequency [rad/s]

The spectral density function can be seen as the probability of finding a component of the molecular motion at a given frequency. If a suitable component exists at the Lamor frequency of a spin, longitudinal relaxation will occur. According to Eq. 8 for fast molecular tumbling

(τ_c ω <<1) and for slow molecular tumbling (τ_c ω >>1) the probability of finding a frequency component at the Lamour frequency is low and hence T_1 is long. Longitudinal relaxation time is shortest when τ_c ω ~1.
Proteins are relative big and have a slow molecular tumbling therefore T_1 is long. On the other side T_2 is short for proteins as low frequency fluctuations causes dephasing of transverse coherence. The probability to find low frequency components becomes bigger for slow molecular tumbling.

1.3.4 The Nuclear-Overhauser-Effect

The Nuclear Overhauser Effect (NOE) is mediated by the dipolar interaction between the nuclei. The effect is the change of intensity of one spin I if the other spin S is perturbed from its equilibrium.

$$I = (1+\eta)\cdot I_0 \qquad 9$$

I	resulting intensity
η	NOE
I_0	original intensity

The two spins are coupled so they form a two-spin system. If the spin S is saturated the spin I is affected indirectly. The perturbation of S does not influence W^I_1 but W_0 and W_2. To regain equilibrium two spins have to flip at the same time (Fig. 48).

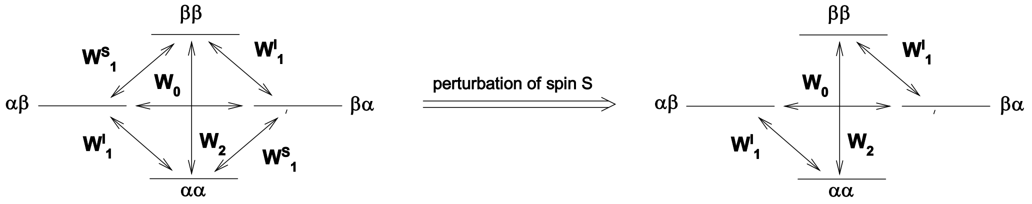


Fig. 48: The four different energy levels in a two spin system consisting of spin S and I. Saturation of spin S eliminates the W^S_1 pathway and the system can only relax through W^I_1, W_0 and W_2.

W_2 (ββ→ αα) is called double quantum transition and W_0 (βα↔βα) zero-quantum transition. Neither W_2 nor W_0 can be measured directly but indirect effect can be measured. The steady state NOE, which is the balance of the possible relaxation pathways, can be described as

$$\eta_I = \frac{W_2 - W_0}{2W_1 + W_2 + W_0}$$
$$\sigma = (W_2 - W_0) \qquad 10$$
$$\rho = (2W_1 + W_2 + W_0)$$

Where

η_I	enhancement of the spin I
σ	cross-relaxation rate
ρ	auto-relaxation rate

The cross-relaxation rate σ (see Eq. 10) determines the sign of the NOE.
W_2 relaxation will occur at a rate proportional to the spectral density at the sum of the Lamor frequencies $J(\omega_I + \omega_S)$, while W_0 relaxation will occur at a rate proportional to $J(\omega_I - \omega_S)$.

For small molecules W_2 is dominant and the peaks are enhanced by maximal $\eta_{\max} = \frac{\gamma_S}{2\gamma_I}$ that is in the homonuclear case enhancement of 50% in the heteronuclear case it can be much more (saturating ^1H can give up to200% enhancement of ^{13}C signals). For proteins W_0 is dominant, this means the signals are reduced or become zero.

1.3.5 Two-dimensional spectra

The 2D-experiment was proposed by Jean Jeener and first performed by Aue, Bartholdi and Richard Ernst [193].

All two-dimensional and multi-dimensional spectra are based on four blocks (Fig. 49). They are called preparation, evolution (t_1), mixing and acquisition (t_2). The pulse sequence starts with preparation time where the sample is excited with one or more pulses. The magnetisation is allowed to evolve for the first time in the evolution period. Next follows the mixing time in which one or more pulses transfer the magnetisation to different nuclei. In the acquisition period, the Free Induction (FID) decay is recorded.

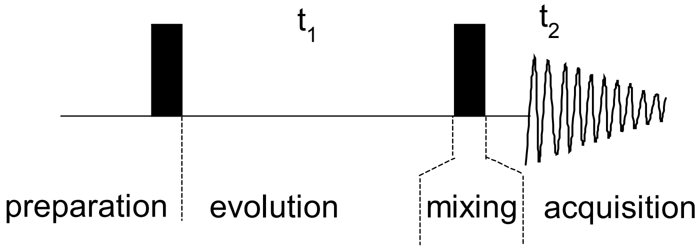


Fig. 49: Schematic of a 2D pulse sequence.

The signal is only recorded in t_2 but not in t_1. t_1 is divided into the sampling interval Δ_1. 2D spectra are a series of 1D spectra with different sampling intervals. In the first 1D-spectrum t_1 is very small. In the second t_1 is incremented by Δ_1, and in every subsequent spectrum, t_1 is incremented by an additional Δ_1. The spectra are twice Fourier transformed resulting in a 2D frequency spectrum (Fig.50).

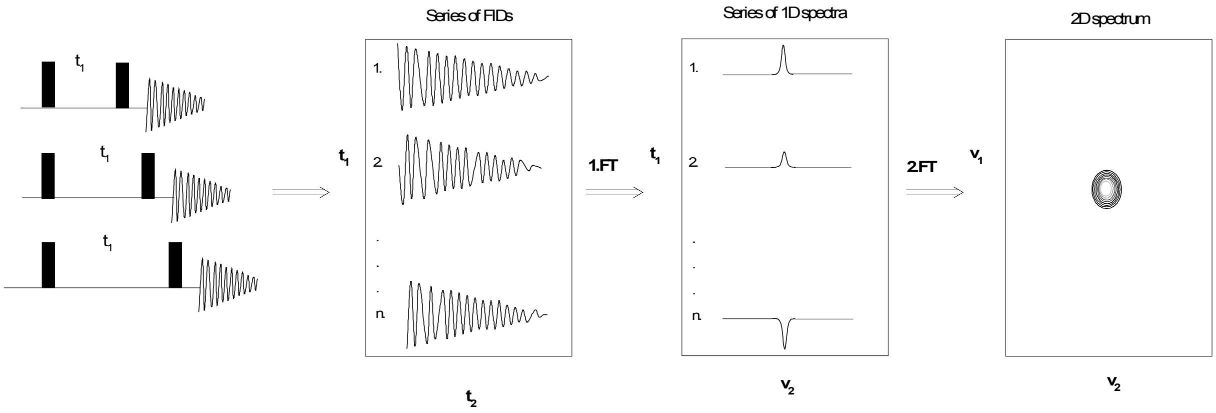


Fig. 50: A 2D -spectrum recording. Each 2D-spectrum consists of a series of 1D spectrum where the following spectrum has a longer evolution time.

A 3D spectrum has 2 evolution times and 2 mixing times. Only during the acquisition time (t_3) FID are recorded. Increments of t_1 and t_2 are changed independently.

1.3.5.1 Homonuclear 2D-spectra

Homonuclear 2D-spectra are often ^1H-^1H spectra. The advantage of ^1H-^1H spectra is that unlabelled samples can also be used and structures of small proteins (<50 amino acids) can be determined on unlabelled samples.

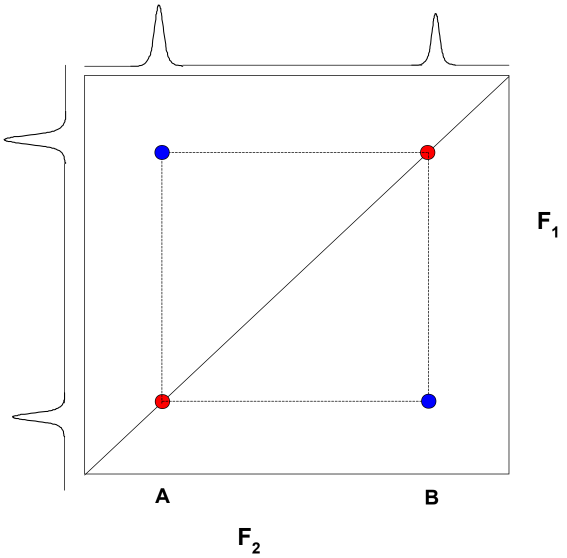


Fig. 51: A homonuclear spectrum consists of diagonal peaks (red) and cross-peaks (blue). Cross-peaks indicate interaction of two nuclei as magnetization was exchanged during mixing time.

Fig. 51 shows a cartoon of a homonuclear 2D-spectrum. Diagonal peaks have the same frequency in both dimensions and are the result of magnetization remaining on the same nucleus during both evolution and acquisition. Cross-peaks signal originate from nuclei that exchanged magnetization during the mixing time. They show an interaction between two nuclei. Cross-peaks are symmetric around the diagonal.

1.3.5.1.1 COSY

In Correlation Spectroscopy (COSY) experiment magnetization is transferred from one spin to another via scalar coupling separated by maximal three to four bonds. The pulse sequence consists of three 90° pulses and the evolution time between them (Fig. 52).

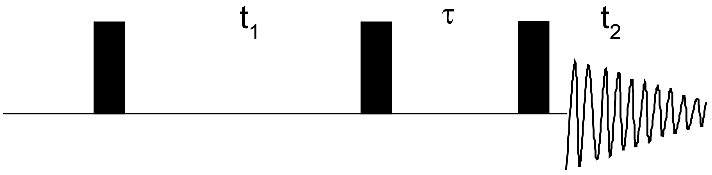


Fig. 52: Pulse sequence of a double-quantum filtered COSY.

The first 90° pulse generates a transverse-magnetization and during t_1 the spins evolve under the scalar coupling. The second 90° pulse generates different quantum coherences where the double quantum coherence is selected by pulse field gradients or a phase cycle. After the mixing pulse coupled spins show additional signal at the frequency of coupled spin (Fig. 53).

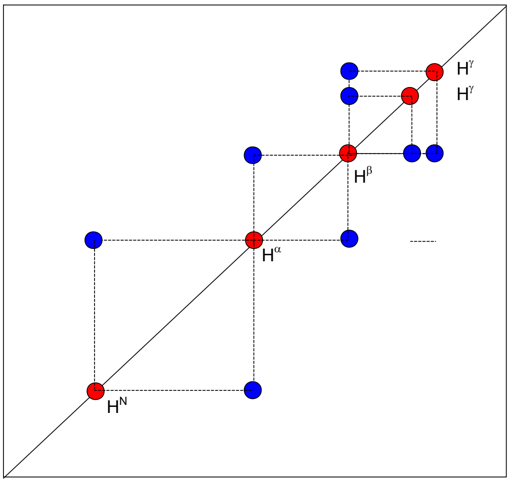


Fig. 53: Schematic COSY spectrum of a valine. Between the two H^γ there is no cross peak as a ^4J coupling constant is nearly zero.

1.3.5.1.2 NOESY

NOESY is one of the few experiments where magnetization is not transferred via bond but through space. The pulse sequence is given in Fig. 54.

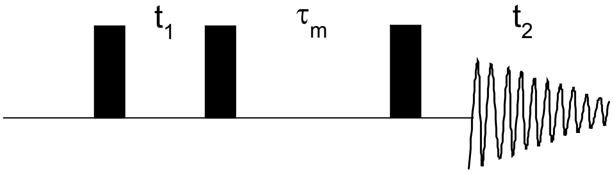


Fig. 54: Pulse sequence of 2D-NOESY.

The first 90° pulse generates a transverse-magnetization and during t_1 the spins evolve. The second 90° pulse flips the magnetization back into z-direction. During the mixing time (τ_m) NOE is building up. The cross-peaks indicate spins that are close in space and the volume of the peaks is inverse proportional to the distance to the negative sixth power. The NOE needs time to build up and at short mixing times, the signal strength is proportional to the inverse sixth power of the distance between the spins ($1/r^6$). This effect allows determining internuclear distances that forms the basis for 3D structure determination of proteins with NMR spectroscopy.

1.3.5.1.3 TOCSY

In the TOCSY experiment, magnetization is dispersed over a complete spin system (e.g. of an amino acid) by successive scalar coupling. The TOCSY experiment correlates all protons within one spin system. The pulse sequence is shown in Fig. 55.

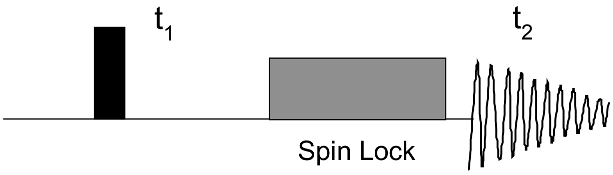


Fig. 55: TOCSY pulse sequence. Spin lock allows transferring magnetization over the whole spin system.

A series of pulses called spin lock keeps the transverse-magnetisation fixed and makes it possible to transfer magnetisation over the whole spin system (Fig. 56). For assignment of unlabelled proteins, TOCSY is an important experiment.

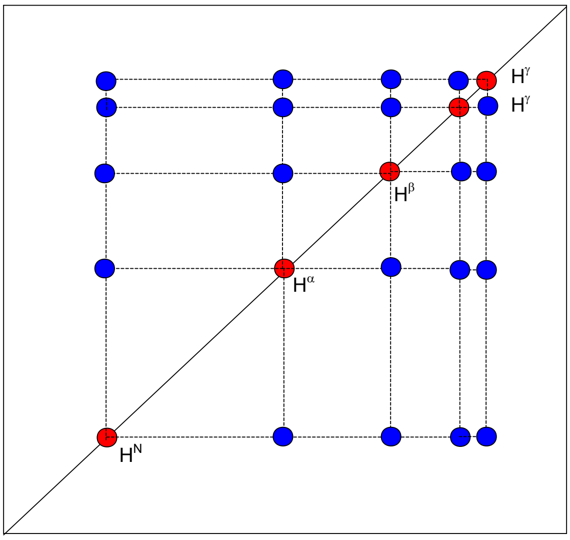


Fig. 56: The TOCSY experiment correlates all the protons of one spin system. A TOCSY spectrum has additional cross-peaks compare to a COSY spectrum.

1.3.5.2 Heteronuclear 2D-spectra

In heteronuclear 2D-spectra magnetization is transferred between different nuclei, in the case of protein NMR, usually ^{1}H and ^{13}C/^{15}N. The natural abundance of ^{13}C and ^{15}N is only 1.1% and 0.037% respectively, therefore is this kind of spectra best performed on labelled protein samples.

The intensity of an NMR-signal depends on the gyromagnetic ratio γ. Signals of ^{13}C spins are therefore 4 times less intensive than proton signals. ^{1}H is the most sensitive nucleus known (with the exception of the naturally not occurring ^{3}H). The INEPT (Insensitive Nuclei Enhance by Polarisation Transfer) pulse sequence allows magnetization transfer from a sensitive nucleus (^{1}H) to an insensitive nucleus (^{13}C or ^{15}N) by heteronuclear scalar coupling and enhances the signals of the insensitive nuclei by a factor γ_H / γ_X . In almost all heteronuclear 2D or 3D spectra the INEPT sequence is incorporated.

1.3.5.2.1 INEPT

The pulse sequence is shown in Fig. 57.

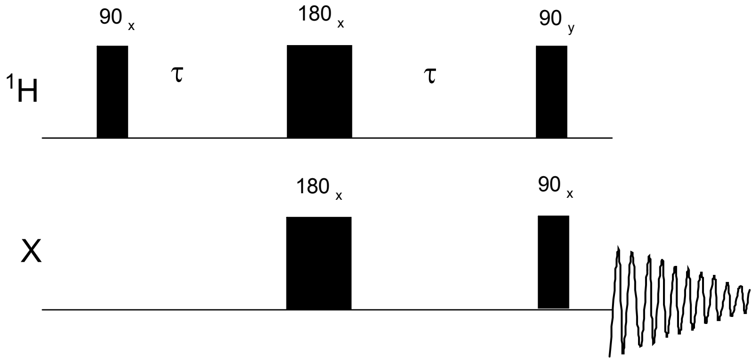


Fig. 57: INEPT pulse sequence. The signals of the insensitive nuclei X are enhanced by that pulse sequence.

INEPT transfers equilibrium population differences from a more sensitive nucleus, thus the signal of the less sensitive nucleus is enhanced. INEPT starts with a 90° pulse on the protons. They evolve under the effects of chemical shifts and heteronuclear coupling during t_1. After the period τ, simultaneous 180° pulses are applied on both protons and the heteronucleus. During the second period τ, the chemical shift evolution is refocused while the heteronuclear coupling continues to evolve. The evolution delay is adjusted such that $2\cdot\tau = 1/(2J)$ ensuring maximum magnetisation transfer. After the second period τ the chemical shift evolution for protons is refocused while heteronulcear scalar coupling between ^{1}H and the heteronucleus X has further evolved. The last 90° pulse on the proton channel transfers the proton magnetization into the z-direction creating longitudinal two-spin order I_ZS_Z between ^{1}H and the heteronucleus X. The 90° pulse on the heteronuclei creates an antiphase magnetisation of the type $N_{xy}H_z$. The heteronucleus is then detected with an enhanced intensity. The gain of sensitivity is defined by Eq.11.

$$I_{INEPT} = I_0 \frac{\gamma_H}{\gamma_X} \qquad 11$$

Where

I_0	Intensity of the less sensitive nucleus
γ_H, γ_X	gyromagnetic ratio of proton and heteronucleus [rad/(s·T)]

The polarisation transfer will result in a signal gain of a factor 4 for ^{13}C, and a factor 10 for ^{15}N.

The reverse INEPT sequence transfers from a less sensitive nucleus to a more sensitive nucleus (e. g. proton). This is done before the acquisition allowing to record the signals on the more sensitive nucleus resulting in a better signal- to- noise ratio.

1.3.5.2.2 HSQC

Heteronuclear single quantum coherence is a heteronuclear 2D experiment yielding cross peaks between atoms exhibiting heteronuclear ^{1}J-coupling. The pulse sequence for the HSQC is built up of an INEPT transfer of magnetization from ^{1}H to the insensitive nucleus followed by an evolution time t_1. Evolution of heteronuclear scalar coupling during t_1 is eliminated by a 180° pulse on the protons centred in the middle of the evolution period t_1. The magnetisation is then transferred back to the more sensitive protons by a reverse INEPT step and the FID is recorded while the heteronuclei are decoupled (Fig. 58).

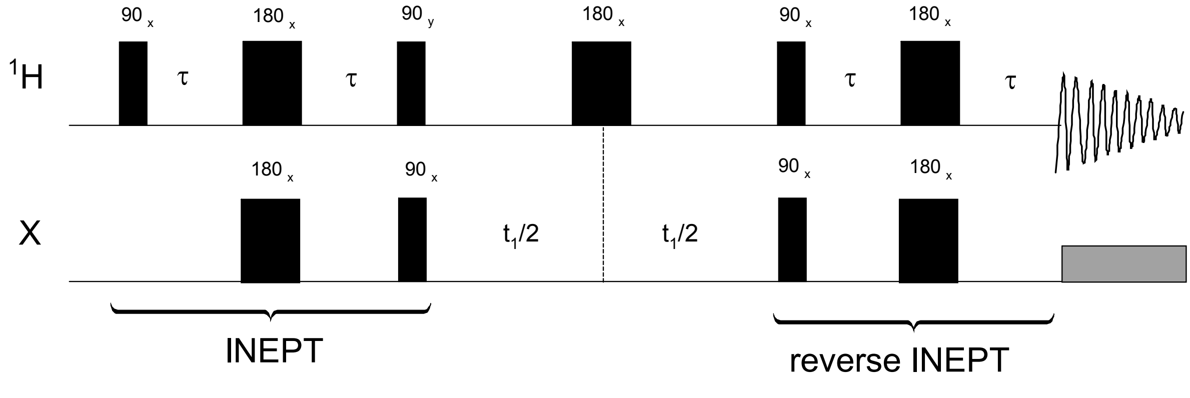


Fig. 58: Pulse sequence of HSQC.

In protein NMR, ^{15}N-HSQC is widely used to correlate H^N with the amide N on the peptide back bone. Additionally, signals of N-H moieties in the side chains of the Asn, Gln, Trp, some Arg and more seldom Lys are visible in the ^{15}N-HSQC.

1.3.6 3D experiments and assignment strategy

For the assignment of proteins bigger than 5kDa three dimensional spectra are often needed. For protein assignment at first the atoms of backbone (H^N, N, C^α, C', H^α) plus C^β and H^β are assigned. For this purpose the 3D-experiments HNCO, HN(CA)CO, HNCA, HN(CO)CA, HBHANH, HBHA(CO)NH, CBCANH and CBCA(CO)NH are used. Remaining alkyl carbons and protons of the longer amino acids can be assigned by HCCH-TOCSY (both HC(C)H-TOCSY and (H)CCH-TOCSY), and HCCH-COSY (both HC(C)H-COSY and (H)CCH-COSY) experiments. The aromatic resonances have to be assigned separately (Fig. 59).

The name of the measurements reflects the magnetization transfer pathway and the names of the atoms, whose chemical shift is recorded. Atoms taking part in the magnetization transfer but whose chemical shift is not recorded are given in brackets.

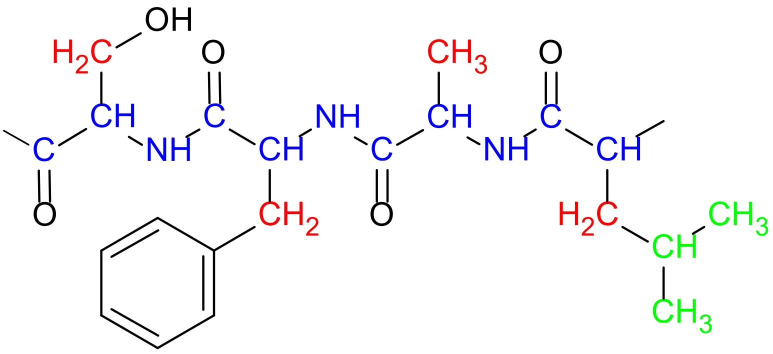


Fig. 59: Assignment plan for a protein. First the backbone atoms are assigned (in blue) and the amino acid sequence is determined. With the assignment of the C^β and H^β some amino acids are fully assigned. The remaining alkyl-carbons and protons of the longer amino acids are assigned by HCCH-TOCSY and HCCH-COSY experiments.

Nearly all aforementioned 3D experiments contain a ^{15}N-HSQC of the amide group of the backbone and the magnetization is transferred from the protons to the more insensitive heteronuclei and back to the protons for acquisition by INEPT and reverse INEPT respectively. For intensity reasons the experiments start with exciting ^{1}H and end with acquiring signal also on the proton channel. Magnetization is transferred between the nuclei by scalar coupling. As an example, the HNCA pulse sequence is shown in (Fig.60).

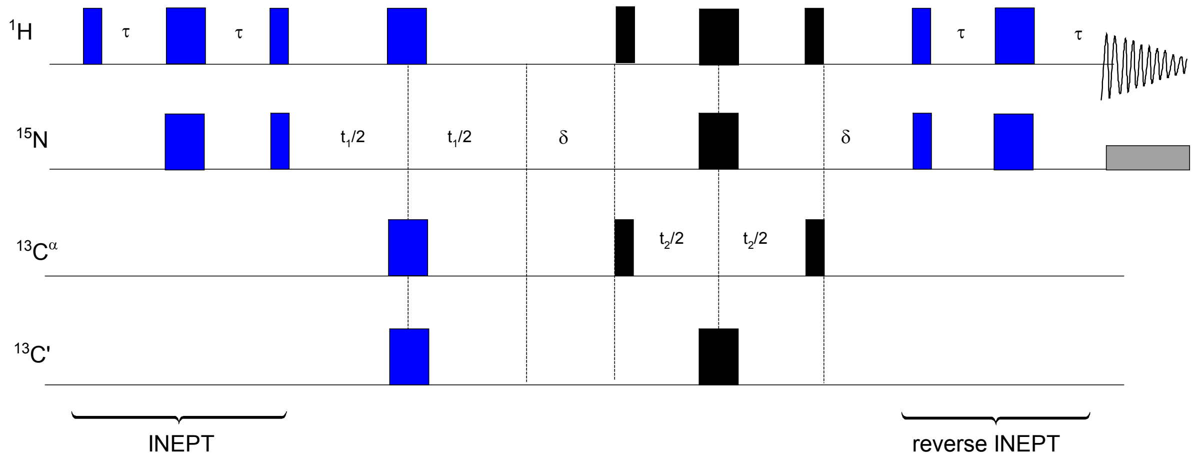


Fig. 60: Pulse sequence of HNCA. The pulses of the ^{15}N-HSQC are in blue. Magnetization is transferred from the protons to N for the first evolution and then further to C^{α} for the second evolution period. For the acquisition magnetization is transferred back to the protons via N.

The magnetisation is transferred from H^{N} to N via the INEPT pulse sequence. Magnetization evolves on N (t_1). Evolution due to scalar coupling with H^{N}, C^{α} and C' spins is eliminated by 180° pulses on each of these nuclei centred in middle of the evolution period t_1. During the delay δ, N magnetization becomes antiphase with respect to the coupled C^{α} spins. The δ delay is chosen to be $2/J_{(NH)}$ so that the N magnetization remains antiphase with respect to the coupled protons. 90° pulses on ^{1}H and C^{α} spins establish a three-spin coherence (H^{N}-N-C^{α}). The second evolution period t_2 follows and 180° pulses on H, N and C' in the middle of t_2 prevent scalar coupling. During the second δ delay N magnetization is rephased with respect to its coupled C^{α} spins but remains in antiphase with the coupled H^{N}s. A reverse INEPT element transfers the magnetization back to amide proton and the FID is recorded.

The J-constant between N and C^{α} from the preceding amino acid and within the same amino acid are have similar absolute values (11 Hz and 7 Hz respectively; see also Fig. 61). Therefore is the magnetization transferred to both C^{α}s.

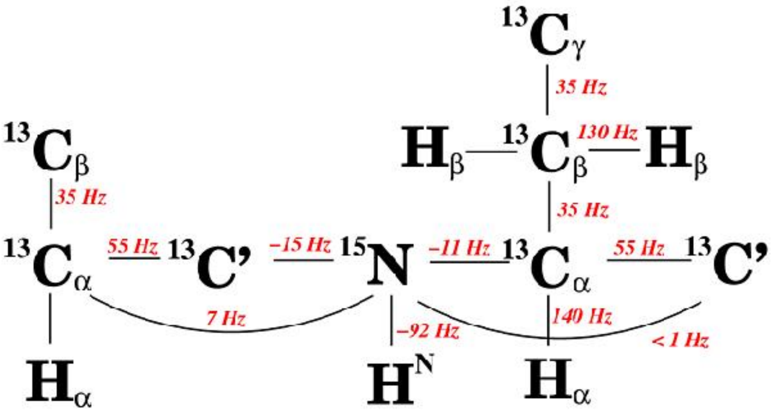


Fig. 61: ^{1}J and ^{2}J- coupling constants of proteins. $^{1}J_{NCA}$ and $^{2}J_{NCA}$ have similar values therefore the magnetization is transferred to both C^{α}. In contrast $^{1}J_{NC'}$ and $^{2}J_{NC'}$ differ enough that magnetization is only transferred to the preceding C' (C'-1).

To distinguish between preceding and intraresidual C^{α} the HN(CO)CA experiment is performed. With this pulse sequence, magnetization is only transferred to the preceding C^{α} (C^{α} -1) as the preceding and intraresidual $J_{NC'}$-constants are different enough to be selective (Fig. 61 and Fig. 62). With this pulse sequence couple it is theoretically possible to link every amino acid with its preceding one. This is also called sequential walk (Fig. 63).

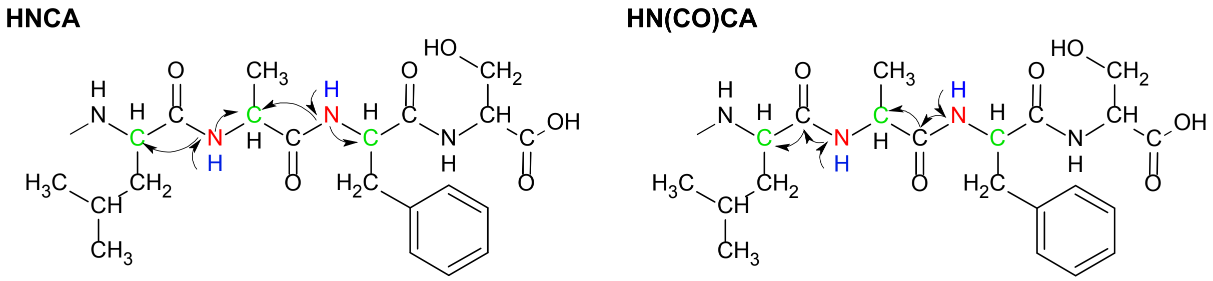


Fig. 62: Comparison of magnetization transfer in HNCA and HN(CO)CA experiments. The magnetization is transferred from H^N (blue) to N (red) and further to C^α (green). In HN(CO)CA magnetization is carried over C’ to C^α.

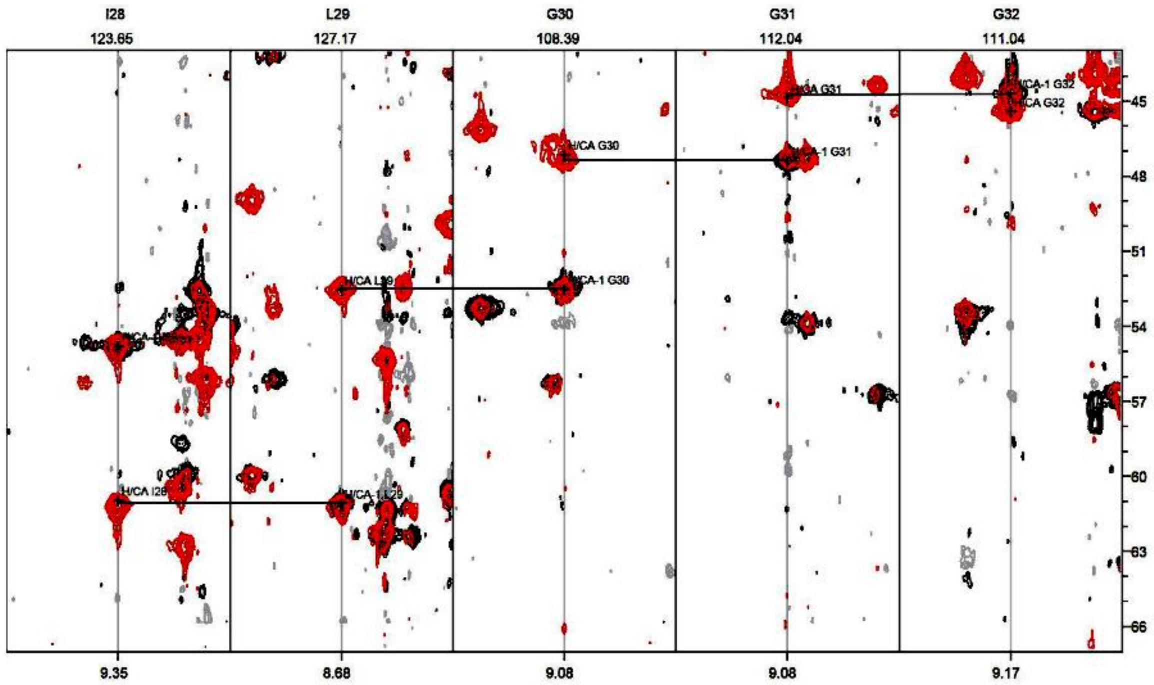
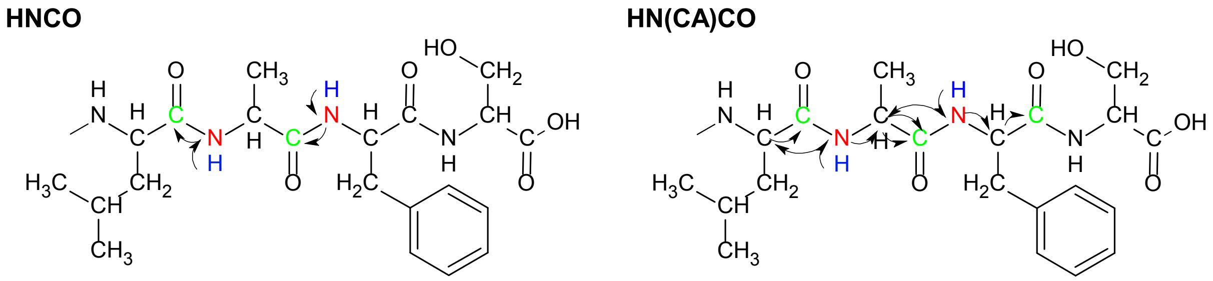


Fig. 63: Overlay of HN(CO)CA (black) and HNCA (red) in strip representation. The intraresidual and preceding C^α can be distinguished by overlaying the spectrum of HNCA with the spectrum of HN(CO)CA. It is then possible to make the attempt to link the different strips.

1.3.6.1 Backbone assignment

1.3.6.1.1 HNCO/HN(CA)CO

HNCO is the most sensitive 3D spectrum. It correlates the amide proton and nitrogen from one amino acid with the ^{13}C-carbonyl of the preceding amino acid (Fig. 64).
HN(CA)CO is less sensitive. The magnetisation is transferred from the amide nitrogen to C^α of its own amino acid but also of the preceding and then to both carbonyl. If both spectra are overlaid C’ -1 and C’ can be distinguished. Theoretically it is possible to connect all amino acid correctly.



HNCO

HN(CA)CO

Fig. 64: Comparison of magnetization transfer of HNCO and HN(CA)CO. In HNCO magnetization is only transferred to the preceding carbonyl while in HN(CA)CO magnetization is transferred first to the preceding and intraresidual C^{α}s and then to C’.

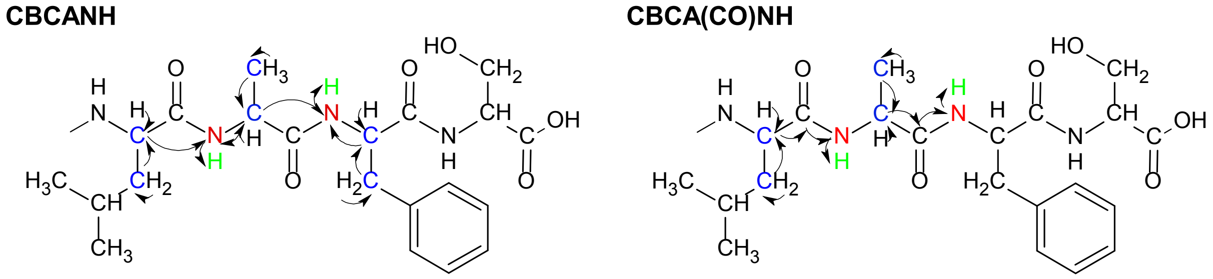
1.3.6.1.2 HNCA/HN(CO)CA

They correlate the amide proton and nitrogen from one amino acid with the C^{α} its own amino acid and the preceding one. The HNCA spectrum shows two peaks per amide group whereas the HN(CO)CA has only a correlation to C^{α}-1.

1.3.6.2 Side chain assignment

1.3.6.2.1 CBCANH/CBCA(CO)NH

The magnetisation of H^{α} and H^{β} is transferred to C^{α} and C^{β} respectively (magnetisation transfer see Fig. 65). It is further transferred to H^{N} via N (and C’). The J-constant of H^{α} and C^{α} and between H^{β} and C^{β} is so similar that both are detected in one dimension. The CBCANH experiment is only of limited value for larger proteins due to its inherently low sensitivity.



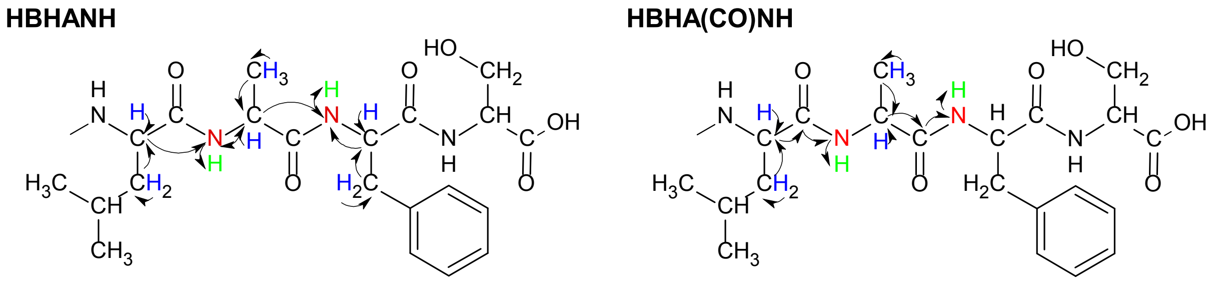
CBCANH

CBCA(CO)NH

Fig. 65: Magnetization transfer of CBCANH and CBCA(CO)NH. In CBCANH intraresidual and preceding C^{α} and C^{β} are detected while in CBCA(CO)NH magnetization is only transferred from the preceding C^{α} and C^{β} to the amide nitrogen and proton.

1.3.6.2.2 HBHANH/HBHA(CO)NH

Similar experiments as CBCANH/CBCA(CO)NH but H^{α}, H^{β}, N and H^{N} are evolved (Fig. 66).



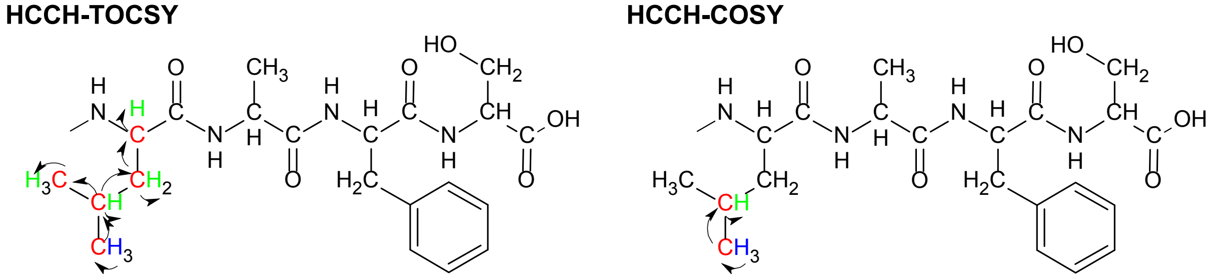
HBHANH

HBHA(CO)NH

Fig. 66: HBHANH and HBHA(CO)NH have a similar magnetization route like CBCANH and CBCA(CO)NH but H^{α} and H^{β} are evolved instead of C^{α} and C^{β}.

1.3.6.2.3 HCCH-COSY/HCCH-TOCSY

The remaining side chain alkyl-carbons and protons can be determined by these two pulse sequences. As aforementioned, HCCH-COSY is the term for (H)CCH-COSY or H(C)CH-COSY. The main difference between these two types of spectra is on which nuclei evolution occurs. H(C)CH-COSY spectrum has two proton axes and one carbon axis while (H)CCH-COSY has two evolution times on carbon spins and acquisition on the protons. In the case of (H)CCH-COSY, it would be possible to start the pulse sequence on the carbons (CCH-COSY) but the sensitivity is enhanced if the magnetisation is transferred from the protons to the carbons (see INEPT). The same applies for HCCH-TOSCY. The magnetisation of the side chain protons is transferred to the carbons. The spinlock (TOCSY) or J-coupling (COSY) is done on the carbon spins. Finally the magnetisation is transferred back to the protons (Fig. 67). The spinlock is specifically applied on the aliphatic carbons (10-70 ppm) and allows to transfer magnetisation over the whole side chain (plus C^{α} and H^{α}) of an amino acid. The ^{1}J coupling constant between aliphatic carbons is around 35 Hz (see Fig. 61). This enables to connect neighbouring carbons and their protons of one side chain (plus C^{α} and H^{α}). The spectra are normally shown like a 2D COSY or TOCSY and the third dimension is ^{13}C. The aryl-side chains cannot be detected with these spectra.



HCCH-TOCSY

HCCH-COSY

Fig. 67: Comparison of HCCH-TOCSY and HCCH-COSY. From a given proton (in this case the methyl protons of valine) magnetization is transfer via ^{1}J coupling to the binding carbon. In HCCH-TOCSY magnetization is transferred to every carbon and from there to the binding proton(s). In HCCH-COSY magnetization is transferred by ^{1}J-coupling to the next carbon and from there to the binding proton.

Assignment of aromatic carbons and protons can be done by 2D-NOESY (assignment of protons) in combination with ^{13}C-HSQC or by special pulse sequences like (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE.
With these pulse sequences, it should be possible to fully assign a protein.

1.3.7 3D-NOESY for structure calculation

Isotope labelling can also greatly aid in collection and assigning NOE constraints for structure calculation. Due to the increasing number of resonances, 2D-NOESY spectra become increasingly crowded for proteins with increasing size. If isotopically labelled protein is available, the NOESY cross peaks can be spread into a third spectroscopic dimension.

1.3.7.1 3D-^{15}N-NOESY-HSQC/ and 3D-^{13}C-NOESY-HSQC

Both spectra are essential for structure calculation. They are a combination of NOESY and HSQC. The sequence starts with a 90° on the proton channel that generates a transverse magnetisation and during t_1 the proton spins evolve. Then the magnetisation is exchanged by NOE. The magnetisation is transferred to the aliphatic (or aromatic) carbon or amide nitrogen and back to the protons for acquisition (Fig. 68). Protons that are close in space can so be detected. These spectra give the restraints necessary for structure calculation.

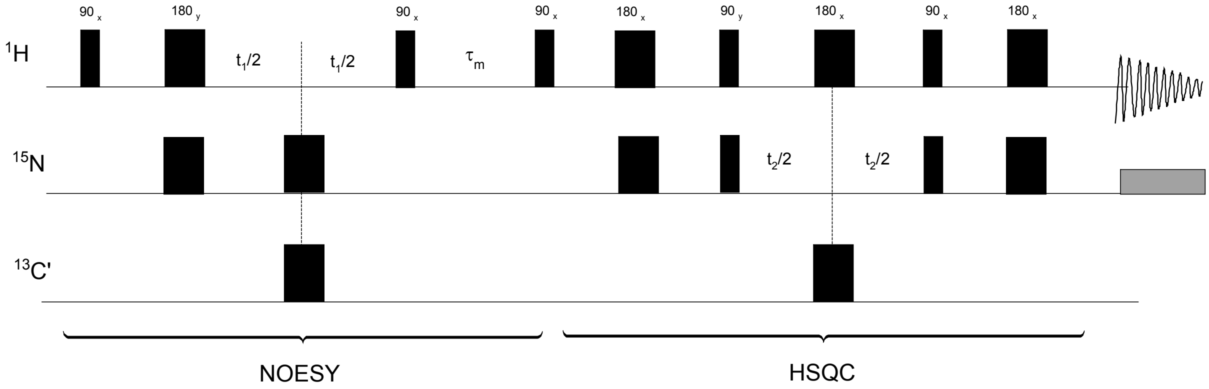


Fig. 68: Pulse sequence of 3D-^{15}N-^{1}H-NOESY-HSQC is a combination of NOESY and ^{15}N-HSQC.

NOE-derived distance constraints are the most important source of information for structure calculation. As aforementioned, the cross peak volume is proportional to the inverse sixth power of the distance between two spins. A 3D-^{15}N-^{1}H-NOESY-HSQC spectrum gives beside intraresidual and sequential NOE also secondary structure related NOEs. In a 3D-^{13}C-^{1}H-NOESY-HSQC intraresidual and long range side chain-side chain NOEs can be found. NOEs related to the secondary structure and especially long range NOEs are important for structure determination.

1.3.8 Assignment of big proteins (>30 kDa)

Enrichment of ^{15}N and ^{13}C of whole proteins is nowadays standard. But big proteins (>35 kDa) are hard to assign for two reasons:

1) Big proteins have fast T_2 relaxation. This causes broader signals as during the pulse sequence a part of the phase coherence got lost.
2) The more amino acids one protein has, the higher is the risk of severe overlapping.

If the line width is broad due to the size of the protein, methods like TROSY and/or deuterium labelling have to be considered.

Dipole-dipole relaxation can be reduced by lack of nearby nuclei to which the magnetization can be transferred (Fig. 69). Large proteins are often deuterium labelled (full or partly) to reduce relaxation.

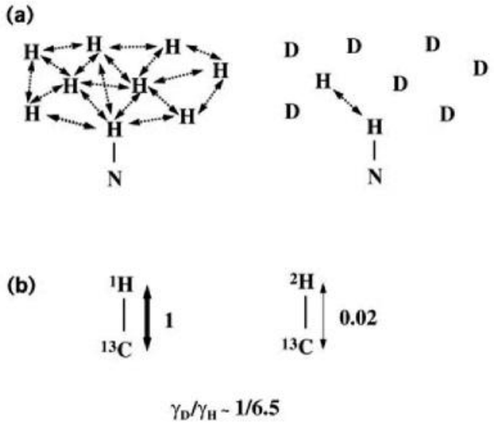


Fig. 69: Labelling with deuterium decreased relaxation rate in two ways. A) Fewer relaxation pathways as there are fewer nuclei to interact with B) relaxation rate is proportional to $\gamma^2 \rightarrow$ relaxation rate is reduced 40 times as γ_D/γ_H~1/6.5. Nevertheless many pulse sequences depend on protons**.**

Although deuterium labelling improves the line width, all pulse sequences described above depend on protons (at least H^N). The H^Ns are often exchanged back by unfolding and refolding of deuterated proteins in water. For fully deuterated proteins new 3D spectra were invented where instead of ^1H-^{15}N-HSQC a $^{13}C'$-^{15}N-HSQC is incorporated in the pulse sequences [194,195,196].

1.3.8.1.1 TROSY

Transverse Relaxation Optimized Spectroscopy [197] is specially designed for big molecules or complexes where the line widths are relatively broad.

In a normal HSQC, the signal is a doublet - in both ^1H and ^{15}N dimensions, therefore each cross peak is split into 4 signals. They are usually collapsed into one signal due to decoupling. The line widths of the 4 peaks are, however, different based on two effects: DD coupling ± CSA (Fig. 70).

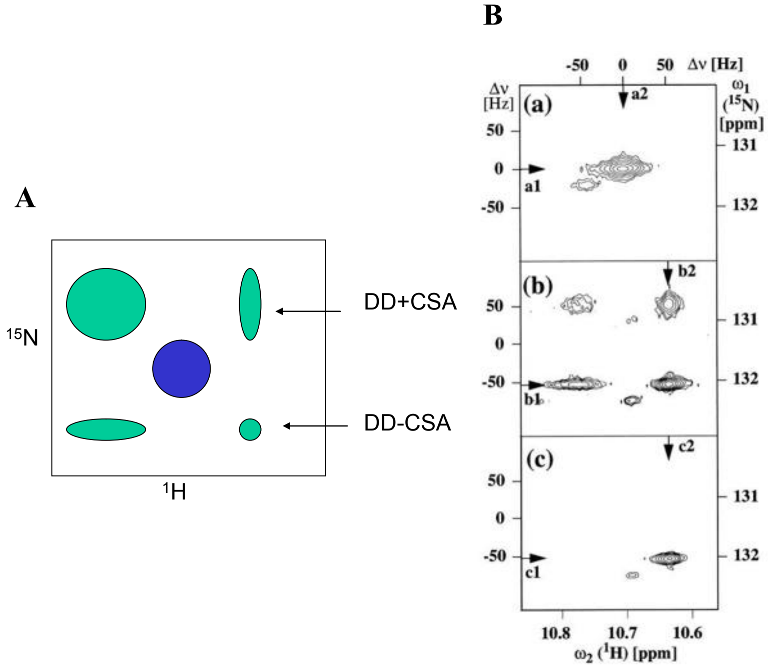


Fig. 70: TROSY effect. A) In a normal ^{15}N-HSQC the doublets in both directions are suppressed resulting in one cross peak per N-H bond (blue dot in schema or spectrum). The two doublets have different line width based on two effects: Dipole-dipole coupling and CSA. Dipole-dipole coupling is always positive while CSA can be

positive or negative. One line width of each doublet is broader (additive behaviour of DD coupling+ CSA relaxation) than the line width of the HSQC peak while the other is narrower (DD - CSA). (Green dots in the spectrum). B) Contour plots of ^{15}N,^{1}H correlation spectra showing the indole ^{15}N–^{1}H spin system of Trp-48 recorded in a 2 mM solution of uniformly ^{15}N-labeled *ftz* homeodomain complexed with an unlabeled 14-bp DNA duplex in 95% H_2O/5% ^2H_2O at 4°C, pH = 6.0, measured at the ^{1}H frequency of 750 MHz. (*a*) Conventional broad-band decoupled HSQC spectrum. The evolution caused by the $^1J(^1H,^{15}N)$ scalar coupling was refocused in the ω_1 and ω_2 dimensions by a 180° proton pulse in the middle of the ^{15}N evolution time t_1, and by waltz composite pulse decoupling of ^{15}N during data acquisition, respectively. (*b*) Conventional HSQC spectrum recorded without decoupling during t_1 and t_2. (*c*) TROSY-type ^{15}N,^{1}H correlation. Chemical shifts relative to DSS in ppm and shifts in Hz relative to the center of the multiplet are indicated in both dimensions. [197].

The TROSY experiment selects the narrowest peak. The resolution and intensity is increased especially for big proteins although the integral of the HSQC peak is bigger than of the selected peak in TROSY. The effect depends also on the magnetic field strength. It was postulated that the narrowest peaks should be obtained at ~ 1.3 GHz [198].

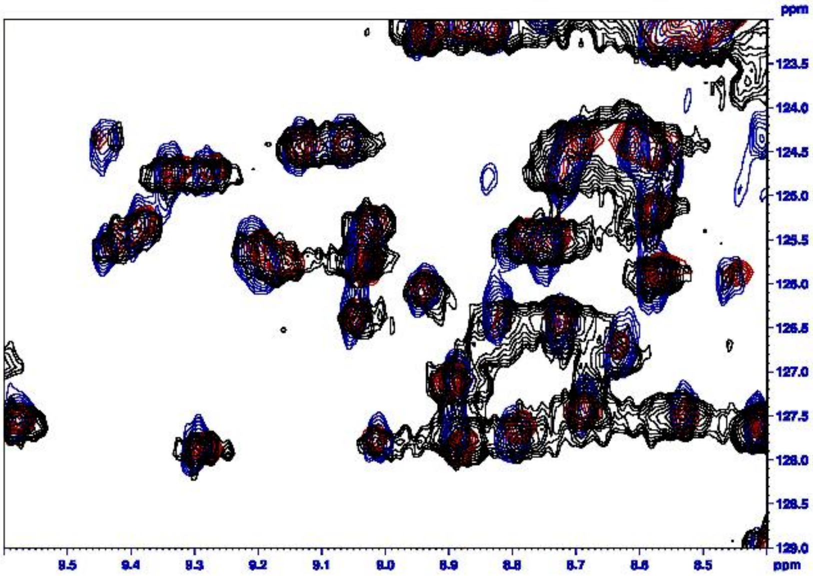


Fig. 71: Overlay of ^{15}N-HSQC of [^{15}N]-R-module of AlgE4 17 kDa (in blue) with the segmentally labelled AlgE4 A-[^{15}N]-R-module 57.6 kDa (in black). Deuterium labelling of AlgE4 A-[^{2}H, ^{15}N]-R and TROSY spectrum (red) reduces the line width significantly.

The TROSY principle can be incorporated into triple-resonance spectra (e.g. TROSY-HNCA), which allows scientists to assign resonances of large proteins [199,200]
Applying TROSY sequences on partly deuterated proteins narrows the line width even further (Fig. 71).

1.3.9 Selective Isotope labelling

The assignment of bigger and bigger proteins causing more and more overlaps. The complexity of the spectra can be reduced either by using multidimensional spectra or by labelling only specific amino acids or segments are labelled. Selective labelling does only reduce the complexity of the spectra but not improve the line width. If the line width is broad due to the size of the protein other methods like TROSY and/or deuterium labelling have to be considered.

1.3.9.1 Selective labelling of amino acid types

One or more ^{13}C and/or ^{15}N labelled amino acids are added to the medium with the other unlabelled amino acids. The resulting spectra are very simple. A full protein assignment is not possible but specific questions can be answered easily.

1.3.9.2 Segmental labelling

Segmental labelling can be done by EPL or protein *trans*-splicing (see Application of Protein Ligation). The complexity of the spectra is reduced but sequential assignment of a ^{15}N and ^{13}C labelled segment is possible (Fig. 72). The labelled segment has the correct conformation as it is incorporated in the whole native protein.

The first segmentally labelled protein was described in 1998 [96] and Otomo et al. could improve segmental labelling and introduced the first protein where a central segment was labelled [98,99] (Fig. 21). The group of Cowburn and Muir introduced EPL for labelling [97]. The advantage of EPL is that the fragments are purified on column before ligation. The N-terminal fragment was cloned to an intein with C-terminal chitin-binding-protein (CBP) and eluted from the column with ethanethiol. The C-terminal was cleaved off the column by Factor Xa to get the essential N-terminal cysteine.

Iwai's group used naturally split inteins for protein *trans*-splicing [201,202,203]. The preceding fragments can be ligated *in vivo* or *in vitro*.

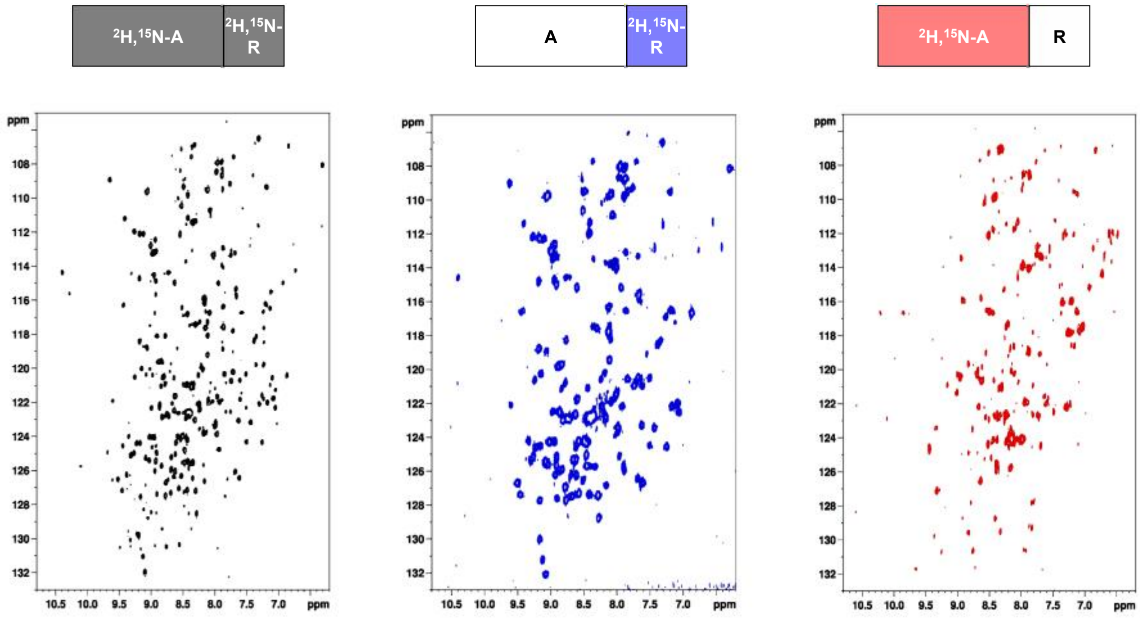


Fig. 72: TROSY NMR spectra of the fully labelled [^{2}H, ^{15}N]-AlgE4 (black) and the segmentally labelled AlgE4 where only one module was labelled. A-[^{2}H, ^{15}N]-R (blue) and [^{2}H, ^{15}N]-A-R (red) from this work [204]

1.3.9.3 Selective backbone and sidechain protonation in full ^{2}H, ^{13}C, ^{15}N labelling

Deuterium labelling narrows the line width but in completely deuterated proteins ^{1}H-^{1}H NOESY spectra are impossible. Therefore the proteins were ~ 70% randomly deuterated. This caused another problem: the methylene- and methylgroups had different isotopomers and the appeared at slightly different chemical shifts. Selective backbone and sidechain protonation prevent this.

Selective H^{α} incorporation in ^{2}H, ^{13}C, ^{15}N labelled amino acids was achieved by chemical reaction [205]. New pulse sequences based on the $^{1}H^{\alpha}$-$^{13}C^{\alpha}$ HSQC like HACACO or HACAN made it possible to assign the backbone.

Selective methyl-group protonation is done by adding a stable isotope-labelled supplement to the growth medium [206]. Methyl-groups give strong signal and methyl-methyl and methyl-H^{N} NOEs can be used for structure calculation. The assignment and structure of 723- residues malate synthase G (MSG) could so be solved (Fig. 73) [207,208].

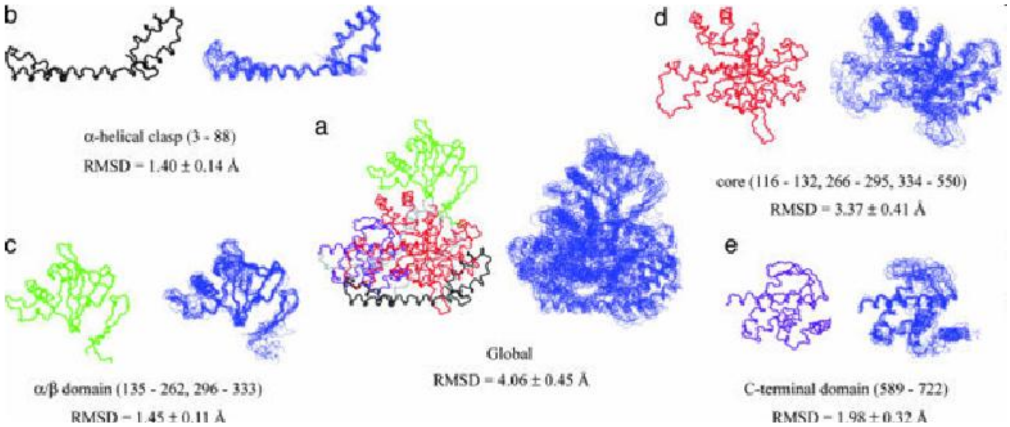


Fig. 73: Comparison of x-ray and NMR structures of malate synthase G **on a per-domain basis.** (*a*) The x-ray structure (PDB ID code 1D8C) and the 10 lowest-energy NMR structures of malate synthase G (MSG) calculated on the basis of experimental restraints. The lowest energy NMR structure (*Right*) calculated on the basis of 1,531 NOE, 1,101 dihedral angle, 415 residual dipolar couplings, and 300 carbonyl-shift restraints. Backbone traces of the x-ray structure (*Left*) and NMR structures (*Right*) are displayed and superimposed by aligning residues in elements of regular secondary structure. The α-clasp, α/β, core, and C-terminal domains are shown in black, green, red, and purple, respectively, in the x-ray structure, with the linkers shown in gray. Individual domains [α-clasp (*b*), α/β (*c*), core (*d*), and C-terminal (*e*)] are shown and superimposed by fitting over residues in regular secondary structure. The rmsd of the NMR ensemble (10 structures) is indicated for heavy backbone atoms of regular secondary structure elements for the entire molecule and individual domains [207].

The first approach to selectively deuterate only one methyl group of valine and leucine was done in the end of the 1980s [209]. The newest approach are selective protonation is called SAIL (stereo-array isotope labelling) [210]. The amino acids are synthesized chemically and enzymatically for cell free protein expression. Each methylene- and methyl-group has only one proton the rest are deuterium (Fig. 74). As this causes new chiralities to appear, each amino each was produced in enatiomeric purity.

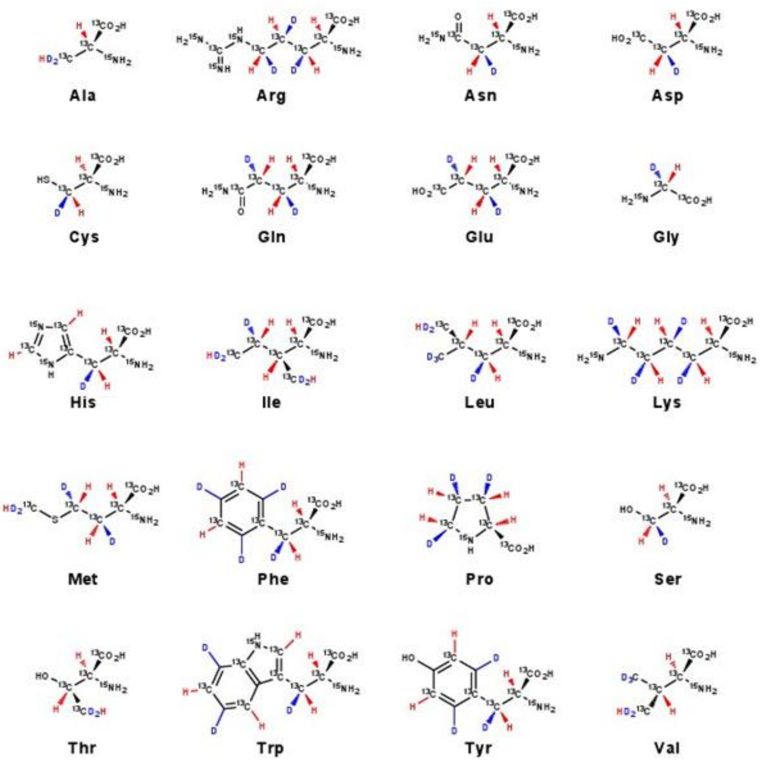


Fig. 74: The 20 partly deuterated amino acids used for SAIL.

The partly deuteration reduces the ^{1}H-^{1}H and ^{1}H-^{13}C dipolar coupling and also relaxation. Therefore is the signal-to-noise improved and also the line width is narrower than in a uniformly labelled sample. The calculated protein structures also improve as pseudo-atom correction that causes inaccuracy is not necessary. This method has great potential but until now it is extremely expensive and is mainly used by the group of Prof. Masatsune Kainosho who invented SAIL.

1.3.10 Structure calculation

Distance constraints are the most important source of information for structure calculation. As aforementioned, the cross peak volume of NOE-spectra is proportional to the inverse sixth power of the distance between two spins. A second set of constraints, the torsion angle constraints can be derived from vicinal ^{3}J coupling constants (see Fig. 43). Allowed ranges for the torsion angles are calculated from ^{3}J coupling constants and can be included into the structure calculation. The torsion angle prediction program TALOS uses the secondary shifts of C^{α}, C^{β}, CO, H^{α} and H^{N} of amino acid triplets and compares it to a database and returns the torsion angle of the ten best matching secondary chemical shifts and residue type homology [211]. TALOS+ is based on the same concept as TALOS but has some supplementary features. TALOS+ includes a neural network component whose output is used as additional term in the conventional TALOS search. This network allows to predict secondary structure in absence of NMR chemical shift data and it reports a reliable estimate of the likelihood that output values is applicable. TALOS+ also reports an estimated backbone order parameter [212]. Orientation constraints are the third group of input parameter that is used for structure calculation. Orientation constraints are obtained by new experiments called residual dipolar coupling (RDC) (see also dipole-dipole coupling)

Besides the different NMR related constraints described above, the distance geometry algorithm is also fed with the covalent structure of the protein including bond lengths, bond angles and other empirical information, such as disulfide topology, if known. There are different computer programs around for structure calculations, in this work, the structure calculations were performed by the program CYANA [213]. Normally the calculation is done about 100 times using the same experimental set of constraints on randomly generated starting structures. CYANA uses a simulated annealing procedure for the torsion angles where the temperature is set to a 10.000 K [214]. This allows the system to cross over energy barriers to find the global minimum. The system is slowly cooled down to normal temperature and finally minimized. The structures are scored by the target function. The CYANA target function is defined such that it is zero if and only if all experimental distance constraints and torsion angle constraints are fulfilled. A conformation that satisfies the constraints better than another will lead to a lower target function value. The 20 best structures selected as final result. The 20 structures are then evaluated and based on the evaluation the NMR input data is refined. Additionally, new NOE constraints can be identified based on the results from the structure calculation. The structure calculation and refinement is repeated iteratively until the results are satisfying (low target function, good superposition of the 20 structures and no violations of experimental constraints).

1.4 ITC

Isothermal Titration Calorimetry is a method to study binding of small molecules (e.g. ligands) to macromolecules (protein or DNA) [215].The heat release or uptake of the protein samples is measured after injecting a precise amount of ligand. From the measurement, thermodynamic parameters can be determined based on equation 12 using least-square methods.

$$\Delta G = \Delta H - T\Delta S = RTN \ln K_d \qquad 12$$

Where

ΔG	change in Gibb's free energy [J/mol]
ΔH	enthalpy change [J/mol]
ΔS	entropy change [J/(mol·K)]
T	temperature [K]
R	ideal gas constant (8.31 J/(mol·K))
N	number of ligand molecules binding to one protein molecule
K_d	dissociation constant [mol/L]

An isothermal titration calorimeter is composed of a pair of identical coin shaped cells enclosed in an adiabatic jacket. Before the addition of ligand, a constant heating power (< 1mW) is applied to the reference cell. This activates the heater (feedback heater) of the sample cell in order to keep the temperature in both cells equal (Fig. 75 A).

The ligand solution is titrated (injected) into the cell containing the protein solution by a computer controlled syringe. The heat released or absorbed per injection is directly proportional to the amount of binding. If an injection results into an exothermic reaction, the feedback power will decrease to keep the constant temperature between sample and reference cell. For endothermic reactions, the feedback power has to be increased. When the macromolecule in the cell becomes saturated with added ligand, the heat signal diminishes until only the background heat of dilution is observed.

From the raw ITC data, the stoichiometry, binding constant, enthalpy (ΔH) and entropy (ΔS) of the interaction is then calculated (Fig. 75 B).

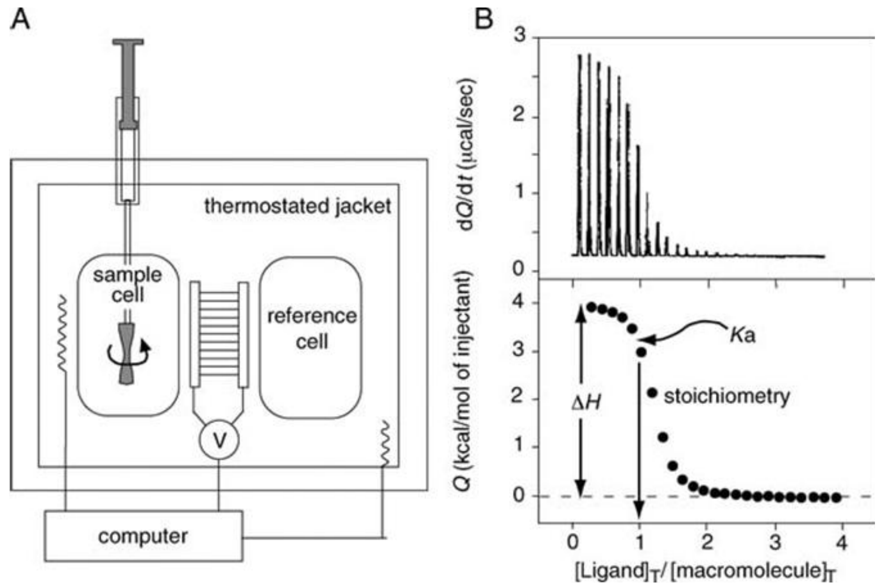


Fig. 75: A) Basic schematic illustration of the ITC instrument, showing the two cells (sample and reference) surrounded by the thermostated jacket, the injection syringe that also works as stirring device, and the computer-controlled thermostatic and feedback systems. B) Example of a typical ITC experiment. The top panel shows the sequence of peaks, each one corresponding to each injection of the solution in the syringe. This example corresponds to an endothermic binding. The bottom panel shows the integrated heat plot [216].

1.5 SAXS

1.5.1 Basics of SAXS

In the following an overview of SAXS on proteins is given. The literature used in this section is based on textbooks and articles [217,218,219] and the references will not be repeated throughout this section.

Small Angle X-ray Scattering (SAXS) is a method where radiation is elastically scattered by the sample. The scattering of the radiation in the sample leads to interference effects resulting in a scattering pattern. The range of angles used for SAXS (< 3.5°) contains information about the shape and the size of macromolecules (Fig. 76).

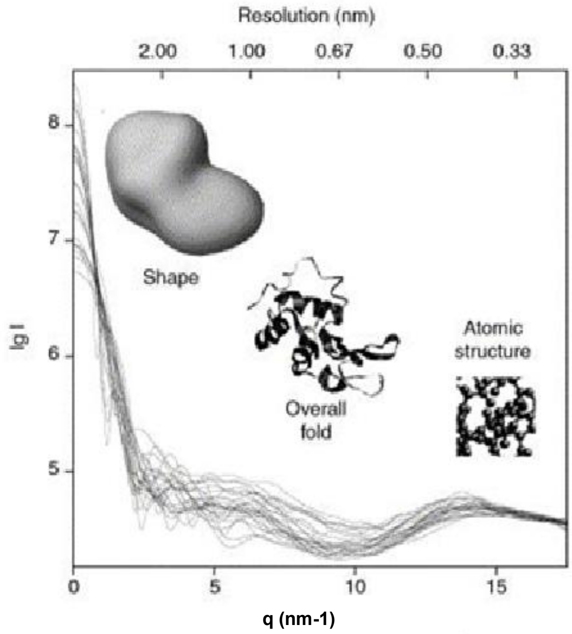


Fig. 76: X-ray solution scattering curve computed from atomic models of 25 different proteins with different molecular masses (10 -300 kDa). At low momentum transfer q the 25 curves clearly differ and the curves are determined by the overall shape of the protein at low resolution (~ 2 nm). At high resolution all curves are very similar and no information can be gained [220].

The experimental setup consists of a source that produces a monochromatic X-ray beam with a wavelength of about 0.1-0.15 nm (Fig. 77). The X-ray is scattered by electrons of the sample and the scattering pattern is detected.

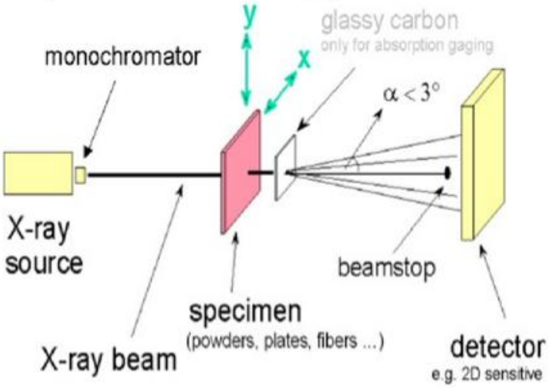


Fig. 77: SAXS setup. A monochromatic X-ray beam is scattered by the sample (e.g. protein in solution). The scattering pattern is recorded by the detector in an angle normally α<3° [221]

Theoretical background

In the solution the proteins are distributed randomly, therefore scattering from each individual particle has to be considered.

Assume a monochromatic wave is scattered elastically when a particle P is illuminated.. Under this assumption the absolute value of the monochromatic wave vector and scattered wave vector are equal (Eq.13).

$$k_0 = \left|\vec{k}_0\right| = k_s = \left|\vec{k}_s\right| = \frac{2\pi}{\lambda} \qquad 13$$

k_0 angular wavenumber of the X-ray wave [m^{-1}]

k_s angular wavenumber of the scattered wave [m^{-1}]

$\vec{k}_0$	X-ray wave vector
$\vec{k}_s$	scattered wave vector
λ	wave length [m]

The scattering vector $\vec{q}$ is described as the difference between $\vec{k}_0$ and $\vec{k}_s$

$$\vec{q}=(\vec{k}_{\boldsymbol{0}}-\vec{k}_s) \tag{14}$$

The scattering vector length is

$$|\vec{q}|=q=\frac{4\pi\sin\theta}{\lambda} \tag{15}$$

q	scattering vector length [m^{-1}]
θ	scattering angle [°]
λ	wave length [m]

The position of the particle P with respect to the origin O is described by the position vector $\vec{r}$. The scattered wave of the particle P is described as $e^{-i\vec{q}\vec{r}}$. The amplitude is the sum of all the scattering waves.

$$A(\vec{q})=\sum e^{-i\vec{q}\vec{r}} \tag{16}$$

A	amplitude of scattering pattern
$\vec{q}$	scattering vector
$\vec{r}$	position vector

The position of electrons cannot be measured precisely therefore the concept of electron density is introduced. ρ is defined as the number of electrons per unit volume and the term $\rho(\vec{r})dV$ describes that the volume element dV at the position $\vec{r}$ will then contain N electrons. The summation can be replaced by the integration over the whole volume V irradiated by the incident X-ray beam.

$$A(\vec{q})=\int\rho(\vec{r})e^{-i\vec{q}\vec{r}}dV \tag{17}$$

The intensity I is the absolute square of the scattering amplitude.

$$I(\vec{q})=|A(\vec{q})|^2 \tag{18}$$

It is the intensity at a certain position that is actually measured at the detector. To simplify the analysis two restrictions are introduced.

(1) The system is statistically isotropic. This allows to describe a scattering wave as its average taken over all directions of $\vec{r}$.

$$\left\langle e^{-i\vec{q}\vec{r}}\right\rangle=\frac{\sin(qr)}{qr} \tag{19}$$

This formula was expressed by Debye 1915 [222].

(2) There exists no long range order and it is possible to describe a uniform density function for the particle while the density function of the solvent is zero. This allows to define the autocorrelation function γ.

$$\gamma(\vec{r}) = \int \rho(\vec{r})dV = (\overline{\rho})^2 V = \rho^2 V \qquad 20$$

From the restriction (1) applies also for the correlation function i.e. also from the correlation function the spherical average is used

$$\gamma(r) = \langle \gamma(\vec{r}) \rangle \qquad 21$$

Equations 19, 20 and 21 are inserted into Eq. 18 resulting

$$I(q) = \int 4\pi \cdot r^2 \gamma(r) \frac{\sin(qr)}{qr} dr \qquad 22$$

It is the intensity that is actually measured at the detector. Until now it was assumed that the solvent does not cause any scattering. In reality the measured scattering intensity I(q) of a sample is the sum of scattering intensities of the buffer and the protein. The scattering pattern of the buffer has to be subtracted from the scattering pattern of the sample before any analysis can be done.

$$I_{sample}(q) = I_{buf}(q) + I_{prot}(q) \Rightarrow I_{prot}(q) = I_{sample}(q) - I_{buf}(q) \qquad 23$$

From the protein intensity the autocorrelation is obtained by inverse Fourier transformation of Eq.23.

$$\gamma(r) = \frac{1}{2\pi^2} \int q^2 I(q) \frac{\sin(qr)}{qr} dq \qquad 24$$

From the autocorrelation the plain shape factor $\gamma_0(r)$ can be calculated

$$\gamma_0(r) = \frac{\gamma(r)}{\gamma(0)} \qquad 25$$

$\gamma_0(r)$ describes the probability of finding a point within the particle at a distance r from a certain point. Therefore the limits are for $\gamma_0(0) = 1$ and for the maximal diameter (D_{max}) $\gamma_0(D_{max}) \rightarrow 0$.

Similar to the autocorrelation function is the distance distribution function p(r). It described the distribution of distances between all possible pairs of points within the particle as a function of the distance r. It is defined as

$$p(r) = \gamma(r) \cdot r^2 \qquad 26$$

The distance distribution can be calculated from the Intensity by inverse Fourier transformation

$$I(r) = 4\pi \int p(r) \frac{\sin(qr)}{qr} dr \xrightarrow{inverseFT} p(r) = \frac{r^2}{2\pi^2} \int q^2 I(q) \frac{\sin(qr)}{qr} dq \qquad 27$$

For first visual inspection the distance distribution versus distance r is often used. Basic information as shape and maximal diameter can be read from the plot.

The scattering patterns (intensity) and the distance distribution curve of differently shaped particles with the same maximal diameter are shown in Fig. 78. From the pattern of the curve it is possible to distinguish different geometrical bodies visually. E.g. the distance distribution plot of a spherical body has the maximum at $D_{max}/2$ while a long rod has the maximum at low r values.

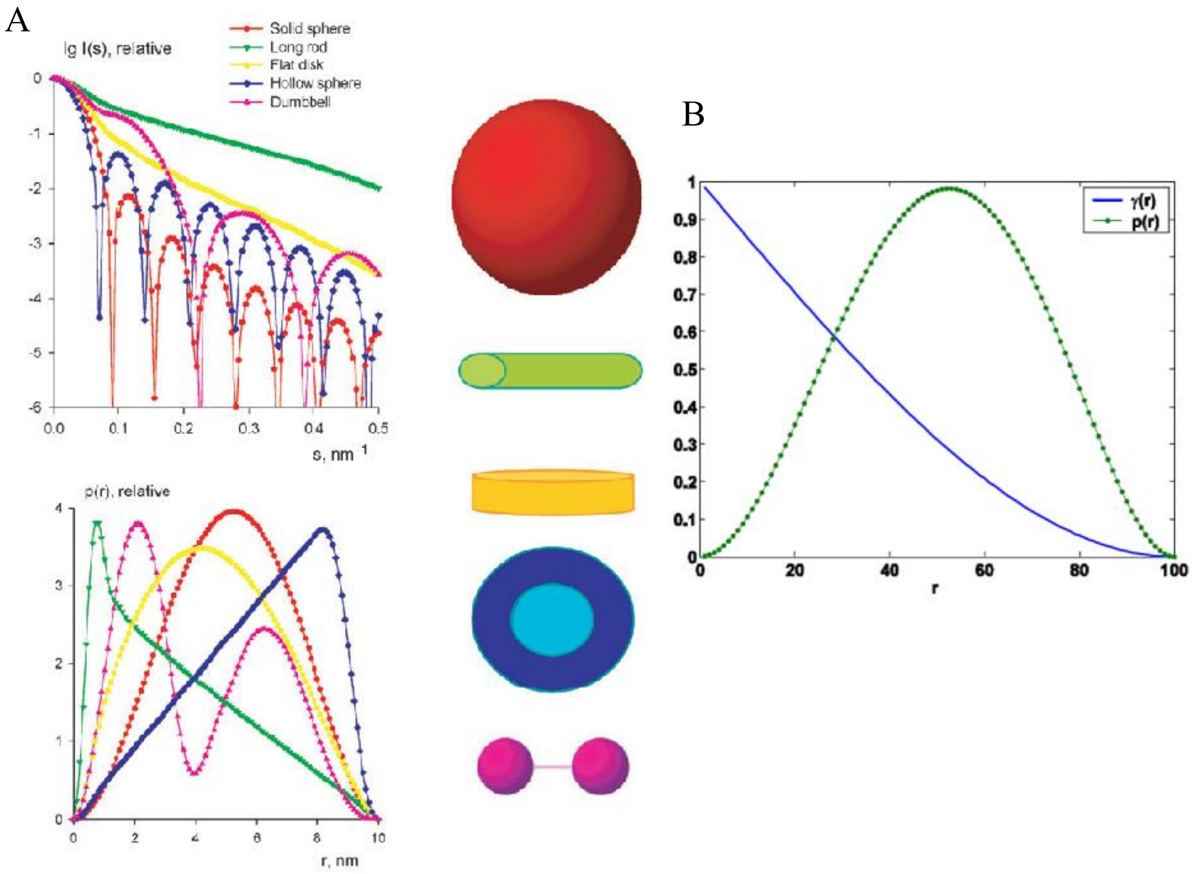


Fig. 78: A) Scattering patterns and distance distribution functions of geometrical bodies with the same maximum size. Both functions contain the same information but distance distribution functions are more straightforward to conclude but visual inspection. **B) Comparison of the plain shape factor γ(r) diagram and the distance distribution factor p(r) diagram on a spherical particle with radius R = 50 Å.** The same information can be deduced from both functions.

1.5.2 Guinier plot

At small q ($1.3 > q$r) the data reveal a linear region based

$$I(q) = I(0)e^{-\frac{R_g^2 q^2}{3}} \Rightarrow \ln(I) = \ln(I(0)) - \frac{R_g^2 q^2}{3} \qquad 28$$

Where

R_g	radius of gyration [m]
I(0)	forward scattering; scattering at $q = 0$

The equation 28 derived by Guinier has long been the most important tool of SAXS data analysis. The measured intensities I(q) plotted versus q^2 is called Guinier plot (Fig. 79). From the slope R_g can be extracted and extrapolation to $q \rightarrow 0$ reveals I(0). The Guinier plot is also a good tool for detecting aggregation (Fig. 79).

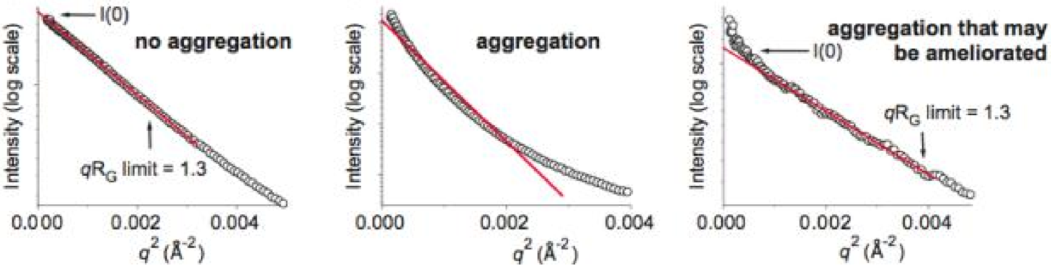


Fig. 79: Guinier plot can be used to obtain I(0) and R_g. The guinier plot also indicates aggregation. Scattering from aggregated samples strongly influences the entire data set and no further processing can be performed.

From I(0) the molecular mass of the particle can be calculated.

$$I(0) = c\Delta\rho_m^2 M_w^{protein}$$
$$\Downarrow$$
$$M_w^{protein} = \frac{I(0)}{c\Delta\rho_m^2} \qquad 29$$
$$\Downarrow in \quad kDa$$
$$M_w^{protein} = \frac{I(0)}{c\Delta\rho_m^2} \cdot N_A$$

I(0) the forward scattering; scattering at $q = 0$

c concentration of the protein [g/L]

$\Delta\rho_m$ the scattering length density difference per unit mass [electrons/gram]

N_A Avogadro number (6.022 10^{23} molecule/mol)

1.5.3 Applications of SAXS

SAXS can be used to establish the shape and size of molecules in solution. This can be used in several ways for protein studies: if there is no high-resolution structure available, *ab initio* modelling can help to obtain a low-resolution structure. Several different programs were developed for at purpose. Before computer-aided modelling was possible, the experimental data were compared with scattering patterns from different simple shapes (spheres, cylinders, ellipsoids...). SASMODEL [223] uses a similar approach, where cylinder and ellipsoids are arranged randomly to compute p(r). A newer approach is the multipole expansion method where an envelope function describes the boundary of the particle (see Fig. 80 first column). The method was implemented in the program SASHA [224].

A more detailed description can be achieved by the bead methods. A spherical volume with diameter D_{max} is filled with N densely packed spheres of a much smaller radius r_0. These spheres can either belong to the particle (index = 1) or to the solvent (index = 0). Starting from a random distribution of 0's and 1's the model is randomly modified until the shape fits the experimental data. With the beads method it is also possible to compute holes inside the particle. One of the programs performing that calculation is DAMMIN [225] (see Fig. 80

second column). SASHA and DAMMIN only use a portion of the experimental scattering pattern (see Fig. 81) which limits the resolution of their models to 2-3 nm. The program GASBOR [226] uses also the wide angle scattering data and that reduce the resolution to 0.5 nm (see Fig. 80 and Fig. 81). The C^{α} atoms of neighbouring amino acids residues are separated by approximately 0.38 nm so a 0.5 nm resolution can be regarded as an assembly of dummy residues. GASBOR starts with a known number of dummy residues (the number of amino acids of the protein) which are arranged until the model fits the scattering pattern.

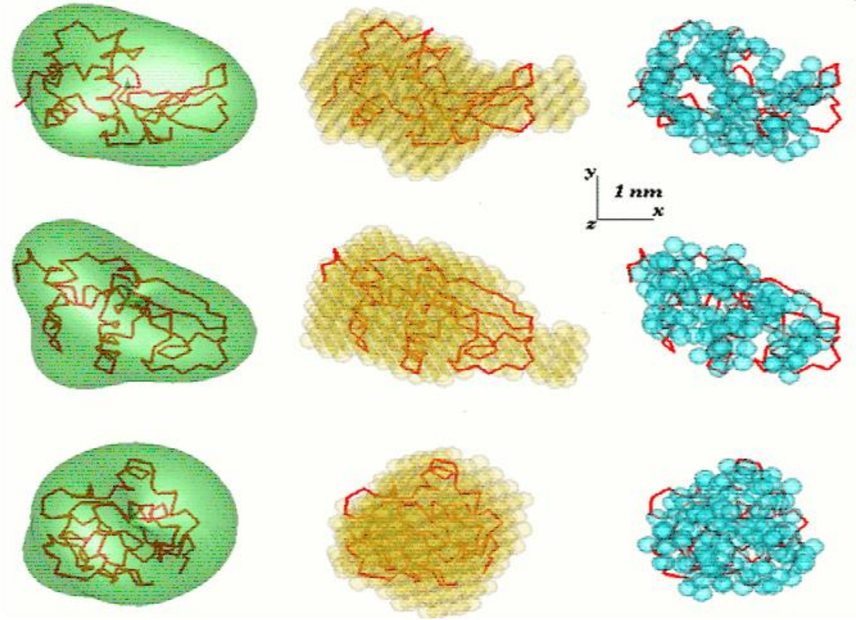


Fig. 80: Atomic model of lysozyme superimposed with *ab initio* models obtained with the program SASHA (left column, semitransparent envelope), DAMMIN (middle column, semitransparent dummy atoms) and GASBOR (right column, semitransparent dummy residues). The low-resolution models are superimposed on the atomic structure using the program SUPCOMB [227]. The middle and the bottom rows are rotated counter clockwise by 90° around x and y respectively. All three-dimensional models were displayed using the program ASSA [219].

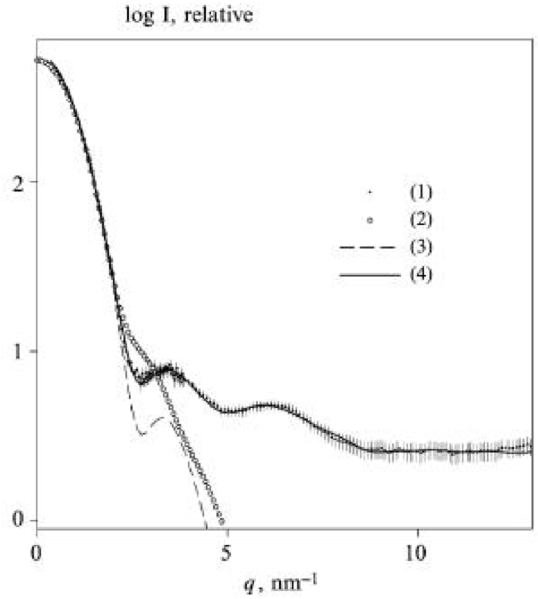


Fig. 81: X-ray scattering from the lysozyme (1) with error bars and the scattering from the *ab initio* models of Fig. 80. (2) envelope model SASHA, (3) bead model DAMMIN and (4) dummy residue model GASBOR. The scattering of the *ab initio* models fit well at low *q*. This region defines the shape of a particle (see **Fig. 76**). The scattering of the *ab initio* model calculated by GASBOR is the only one which also fit at high *q*. This allows to have a more refined model [219].

Scattering data of proteins with known structure can be modelled by the program CRYSOL [228]. It adds a 0.3 nm hydration layer with a density ρ_b (differ from the protein density and the bulk solvent ρ_S) before calculating the scattering pattern. For globular proteins or domains the experimental scattering data and calculated data based on a structure should be very similar. But if a protein is intrinsically flexible or has flexible linkers between domains, the established methods of structure determination often fail. In those cases SAXS offers the possibility of obtaining complementary and/or additional data. Different programs like CREDO, CHADD or GLOOPY [229] were developed to add missing loops or domains to structural models based on the dummy residues approach.

One example is the solution structure of the full length DNA gyrase A subunit [230]. DNA gyrase from *Escherichia coli* consists of two subunits, GyrA (97 kDa) and GyrB (90 kDa), the active enzyme is a heterotetramer A2B2. The subunit GyrA consists of two domains: an amino-terminal domain of 59 kDa (GyrA59) whose structure is known [231] and a carboxyl-terminal domain of 38 kDa (GyrA-CTD structure unknown).

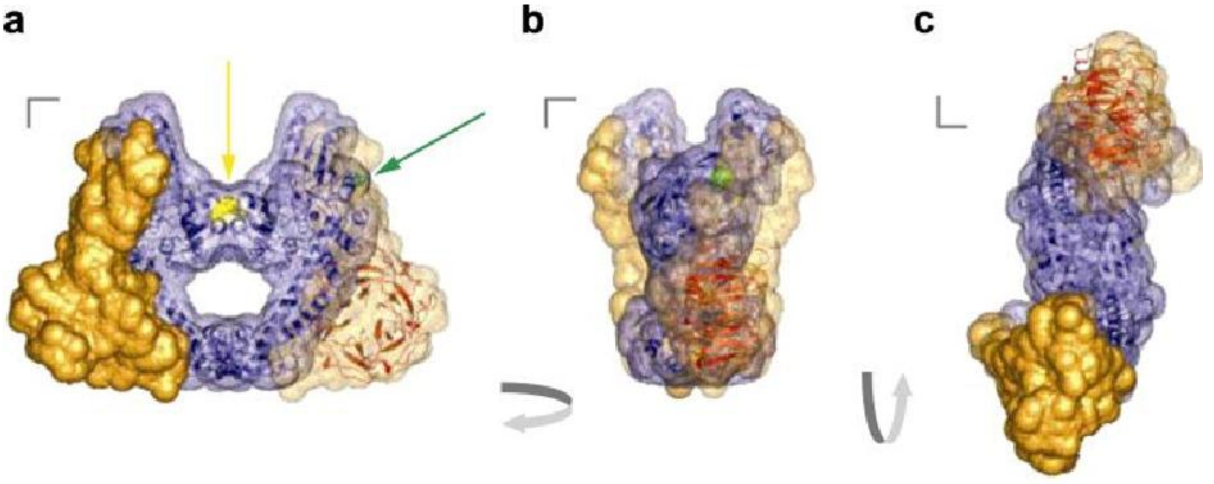


Fig. 82: GyrA solution structure. The model obtained with CREDO is represented as a surface with the fixed GyrA59 structure (*blue*) and the added densities for GyrA-CTD on both sides (*orange*). The active-site tyrosines are colored in yellow and the GyrA59 carboxyl in green, shown in space fill, and indicated by arrows in (*a*). The surface was built from a sphere radius of 5 Å for each residue. The GyrA59 crystallographic structure is shown in blue ribbons. The red ribbons represent the six-bladed β pinwheel domain of one GyrA-CTD (carboxyl-terminal domain), modelled from a homologous crystallographic structure [232] and fitted into the density added by CREDO. The views (*a–c*) are from (*a*) front, (*b*) side, and (*c*) bottom.

Rigid body refinement [233] is used to determine the quaternary structure of protein complexes or general intermolecular interactions in solution like DNA-protein complexes as well as orientation of domains in proteins. An example is αB-crystallin, a small heat shock protein (sHSP), shown in Fig. 83. The high resolution crystallographic model of a homolog of αB-crystallin from *M. janashii* was used as a template to build the model of the dimeric αB-crystallin domain. The dimeric interface of the homology model is identical to MjHSP16.5 but scattering computed from the homology model significantly deviates from the experimental scattering by the αB-crystallin (Fig. 83 b curve 1). A new dimerization interface was proposed and rigid body modelling was employed to refine the angle between the monomers (Fig. 83b curve2-6). The final model (Fig. 83 a right panel) fits the SAXS data and suggests that the αB-crystallin is composed of flexible building units with an extended surface area.

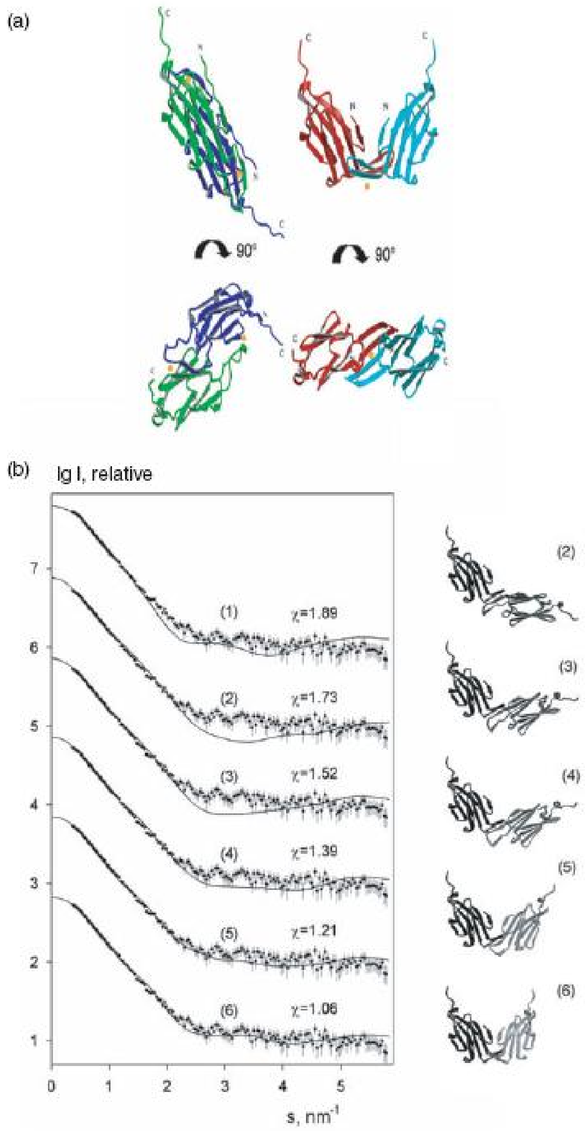


Fig. 83: (*a*) Crystallographic model of the MjHSP16.5 dimer (left panel) and the model of the dimeric *α*-crystallin domain obtained by rigid body refinement (right panel) in two perpendicular views. The monomers coloured green (left) and red (right) are in the same orientation. The residues, which should be in contact according to spin-labelling data, are indicated by orange spheres. (*b*) Experimental scattering from the *α*-crystallin domain and the fits calculated from the atomic models. (1) The crystallographic dimer of MjHSP16.5; (2)–(6) scattering from the dimeric homology models with increasing compactness in displayed on the right. The curves (1)–(6) are displaced down by one logarithmic unit for clarity and the discrepancies with the experimental data are presented [218].

SAXS measurements can help to establish the right equilibrium between monomer, dimer and other multimers. Crystal structures of complexes are likely to be affected by crystallization conditions and crystal-packing interactions. An example is bovine pancreatic trypsin inhibitor (BPTI). In different crystallization conditions either the monomer or the decamer was observed. From the crystal structure the monomer, dimer, pentamer and decamer structures were identified and their scattering patterns were computed. The experimental curves were then analyzed as linear combinations of these theoretical patterns using a non-linear curve-fitting procedure. The results confirmed two different BPTI particles in solution: a monomer and a decamer without any evidence for other intermediates.

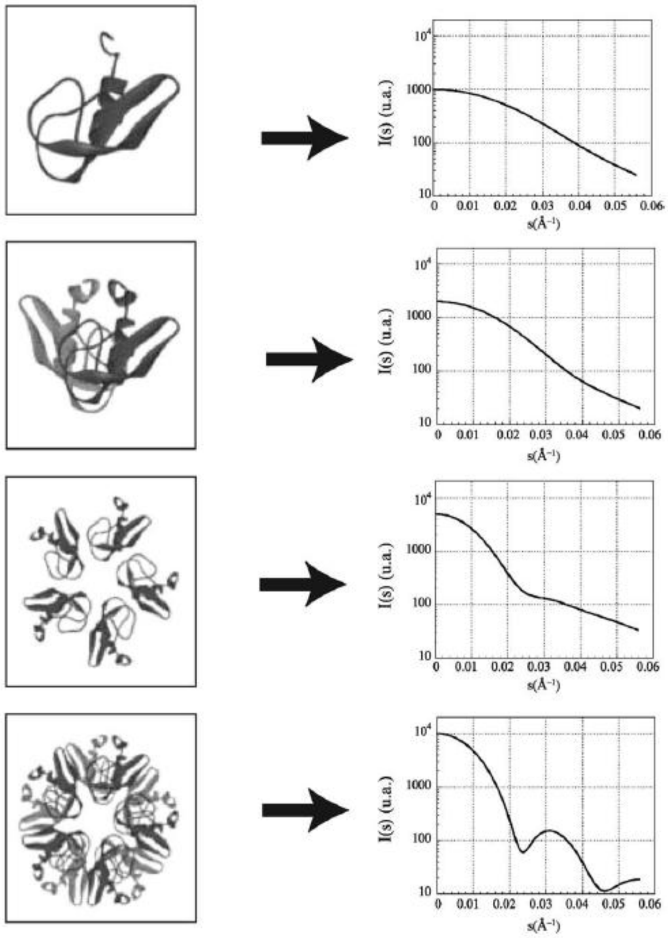


Fig. 84: Ribbon representation of the monomer, dimer, pentamer, and decamer of BPTI with their calculated scattering patterns [219].

2 Aim of the studies

This study has to be seen in the context of the ongoing work on the AlgE epimerases of *A. vinelandii*. In previous studies, the structures of the A- and R-module of AlgE4 had been determined. Both modules show a highly unusual structure consisting mainly of parallel β-sheets making up a four stranded β-helix for the A-module and a two stranded β-roll for the R-module. Both modules contain calcium ions that are essential for the stability of the fold. The substrate of the epimerases is alginates, and both the A- and the R-module of AlgE4 had been demonstrated to bind to poly-M alginates. Since the enzymatic activity is carried solely by the A-modules, whereas the R-modules only enhance this activity, the role of the R-modules on the molecular level is somewhat enigmatic.

In this study, we want to investigate the role of the R-modules from two different angles: firstly, from a structural point of view: how are the A- and R-modules oriented relative to each other in solution? Is there a fixed orientation or do the two modules tumble independently of each other? Does this change in the presence of substrate? Secondly, from a substrate specificity point of view: what is the substrate specificity of the R-module? Which types of alginate (MM, MG or GG-block type) is bound preferentially? Is there a difference in substrate preference between AlgE4, which produces MG blocks and AlgE6, which produces GG-block alginate? In an intact full-length protein, will the substrate preferentially bind to the A- or to the R-module or to both with the same affinity?

The following route of investigation was chosen in order to find answers to the questions above: it seemed desirable to obtain full-length AlgE4 protein, where only the A- or the R-modules are isotopically labelled for NMR. Full-length AlgE4 with its 553 amino acids yields very crowded NMR spectra, hence segmental isotopic labelling would greatly simplify spectral analysis. In order to achieve segmentally isotope labelled AlgE4, a technique called protein trans splicing was employed, where the A- and R-modules, respectively, are fused to a naturally split intein, and cloned into different vectors. The two modules can then be produced independently with different isotope labelling patterns. Upon mixing, the two halves of the naturally split intein would meet, form the intein protein, which excises itself and ligates the two flanking protein sequences (exteins, here the A- and R-modules) with a peptide bond.

Substrate specificity of the R-modules of AlgE4 and AlgE6 can be studied with both NMR and ITC. These two techniques yield information on both the thermodynamics and structural aspects of the protein-substrate interaction. Since only the structure of the R-module of AlgE4 was known, the structures of all three R-modules of AlgE6 should be determined, too.

SAXS would be used in order to obtain information on relative domain orientations in full-length AlgE4 (57.6 kDa) and AlgE6 (90.2 kDa). If sufficient amounts of properly isotopiclly labelled AlgE4 could be obtained, RDC measurements could in principle also be used. However, this would also need at least a partial assignment of backbone resonances of the A-module of AlgE4.

3 General discussion

3.1 Protein trans-ligation (Papers I&II)

3.1.1 In vitro

Protein *trans*-splicing is using split inteins, a rare form of intein that are split on the DNA level. Nearly all split inteins were found in cyanobacteria and belong to one intein allele (DNA polymerase III alpha subunit) [61]. More than half of the found split inteins are putative. Only the split inteins from *Nostoc punctiforme* (*Npu*) and *Synechocystis species* (*Ssp*) strain PCC6803 are used for protein *trans*-splicing repeatedly. The split intein of *Npu* was used for ligating the modules of AlgE4 for two reasons:

- The split intein of *Npu* shows a higher robustness for foreign amino acids around the splicing site
- The split intein of *Npu* is known to perform ligation in high reaction rate

It is known that the amino acids around the splicing site have a major influence on the ligation yield [201]. The first amino acid on the C-terminal extein has to be a cysteine, serine or threonine. The following amino acids on C-terminal extein are less conserved. Nevertheless, they still have a great impact on the yield of ligated protein. The last two amino acids of the N-terminal extein also affect protein ligation. From the two split intein described above *Npu* has a higher robustness for foreign amino acids [201,234]. Fig. 85 shows the influence of the second amino acid of the C-terminal extein on the ligation efficiency.

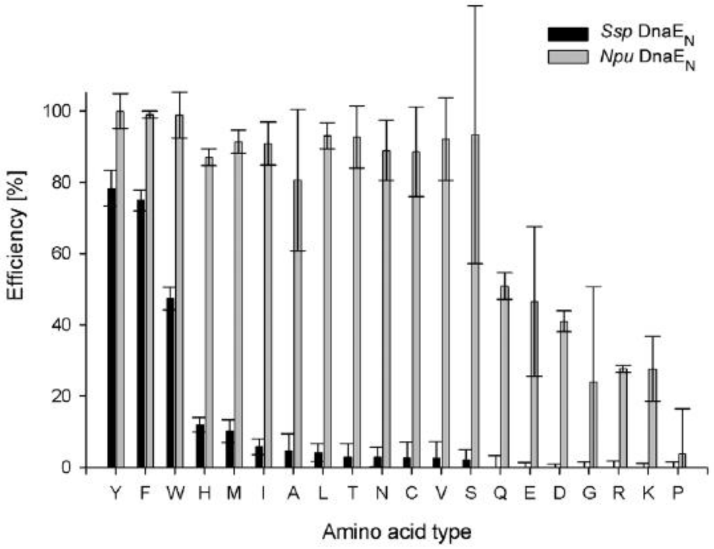


Fig. 85: Graphical representation of the ligation efficiencies. Filled and grey bars indicate the efficiency of *trans*-splicing for the variants of *Ssp* DnaE$_C$ and *Ssp* DnaE$_N$, and the variants of *Ssp* DnaE$_C$ and *Npu* DnaE$_N$, respectively. The amino acid types in the linker CXNGT are shown at the bottom, where X stands for the substituted amino acid type. The error bars indicate the error in the estimation of the protein amounts by image analysis, except for Phe (F) in which the error was estimated from the three independent experiments [201].

To optimize ligation of the modules of AlgE4, few amino acids around the splicing site were exchanged to ones known to support ligation. They are shown in Fig. 3 of Paper II.

The intein from *Nostoc punctiforme* performs robust and fast ligations. The fastest protein *trans*-splicing reaction rate ever measured was $(1.1 \pm 0.2) \cdot 10^{-2}$ s^{-1} at 37° C [234]. This means that $t_{1/2}$ is approximately 60 s.

From prior experiments the *Npu* split intein seems to be efficient for protein trans-splicing. Nevertheless the major obstacle of *in vitro* protein ligation is the dimerization of the essential cysteins.

The cysteins necessary for ligation are often dimerized and have to be reduced before ligation can occur. But most of the reducing agents (like DTT) also act as nucleophilic agents and enhance cleavage of the N-terminal intein part [235]. To test the effect of different reducing agents, the N-terminal SH3 (n-SH3) was ligated with the B1 domain of protein G (GB1) using the *Npu* split intein. The ligation occurred fast and the type of reducing agent is secondary. The yield of ligated product is independent to the reducing agent used (Fig. 86).

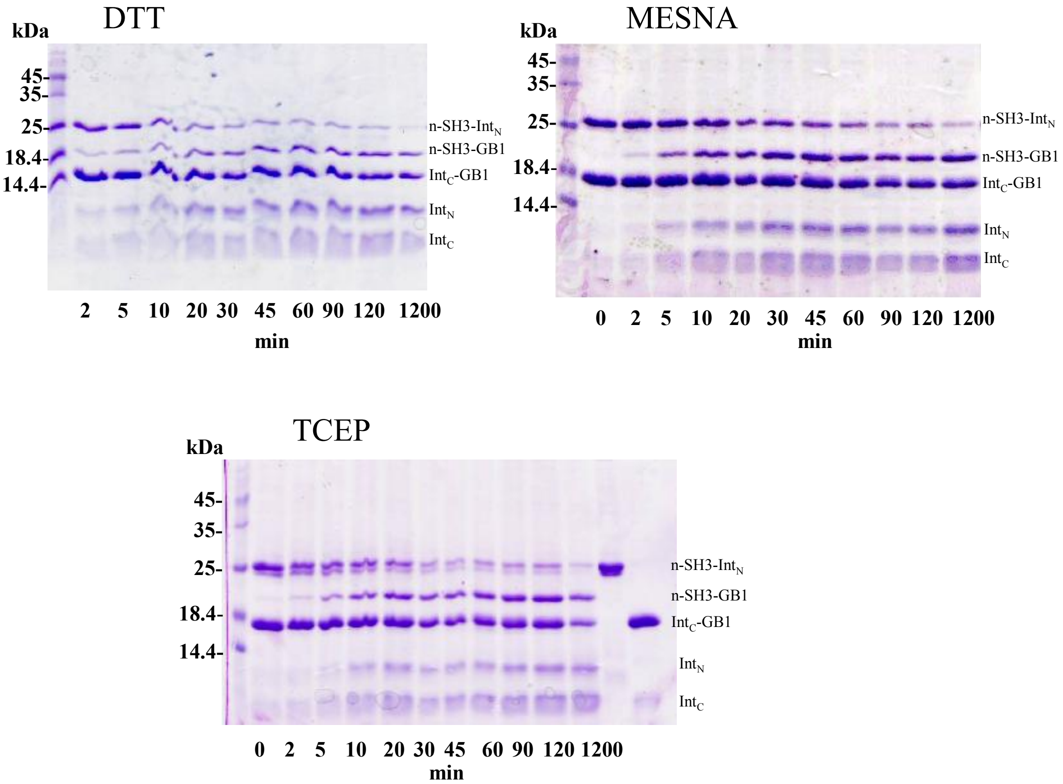


Fig. 86: The ligation yield of n-SH3 with GB1 is independent of the used reducing agent. This indicates that the rate of the cleavage reaction is at least one magnitude lower than the rate of the ligation reaction. The concentrations of the reducing agents were 50 mM DTT, 20 mM MESNA and 0.5 mM TCEP. The ligation reaction rate with 50 mM DTT as reducing agent was $2.3 \pm 0.2 \times 10^{-4}$ s^{-1}. [236].

The results of the ligation of n-SH3 and GB1 indicate that ligation using the split intein of *Npu* is fast and robust against cleavage thus the type of reducing agent is secondary.

The positive results of the aforementioned ligations corroborate our decision to use the split intein from *N. punctiforme* instead of the more commonly used *Ssp* split intein.

The results of the ligation test of AlgE4 differ from the ligation of n-SH3 and GB1. In the case of segmental labelling of AlgE4, the addition of nucleophilic reducing agent caused mainly cleavage - only the addition of TCEP - a non nucleophilic reducing agent - yielded near-complete ligation (Fig. 87).

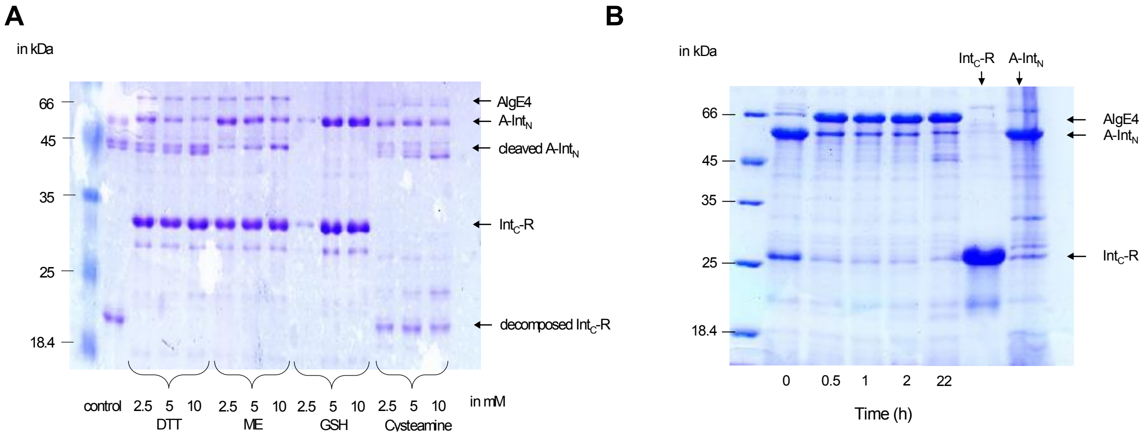


Fig. 87: Comparison of the ligation of AlgE4 in presence of different reducing agent. A) Overview of the ligation with different reducing agents after 2 days ligation at room temperature in TRIS buffer at pH 7 without protease inhibitor. In presence of nucleophilic reducing agents maximal 20% was ligated to AlgE4. (see also **Fig. 88** B) The ligation of the single modules to AlgE4 occurs fast in presence of TCEP. Within 2 hours the reaction has terminated. Data and Figure from Paper II [204].

The SDS-PAGE gels used to visualize the results of the ligation tests were scanned and analyzed by ImageJ [237]. The sum of A-module containing polypeptides (A-Int$_N$, A (cleaved) and AlgE4) is normalized to 100%. Fig. 88 shows the analysis of the gel lanes from Fig. 87 A. Glutathione is inactive at pH 7 and was therefore omitted. For comparison with the other reducing agents the results of ligation in presence of 5 mM glutathione at pH8 (Fig. 89 lane 6) and 5 mM TCEP (Fig. 87B after 22h) were added in Fig. 88. The ligation and cleavage rates were calculated from the SDS-PAGE showing the formation of AlgE4 and cleaved A-module as a function of time (see Fig. 87B and Fig. 89).

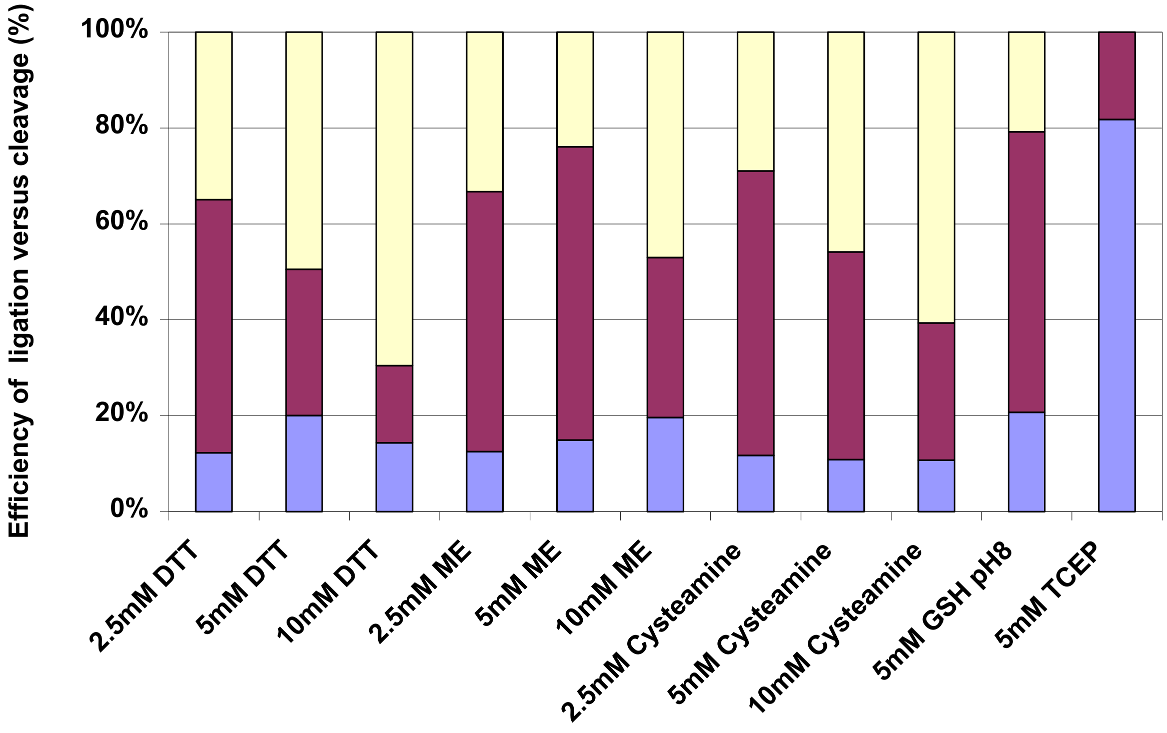


Fig. 88: **The ratio between the ligated AlgE4 (blue), unchanged A-Int$_N$ (purple) and cleaved A-module (yellow).** There was always more A-module cleaved than was ligated to AlgE4.

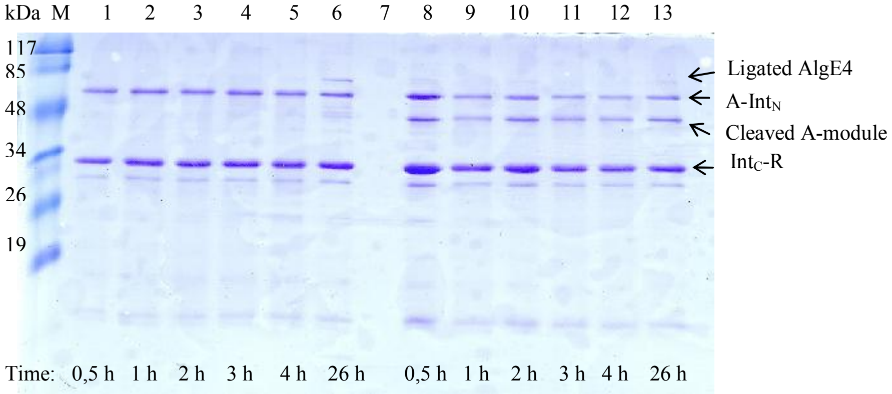
TCEP is not a nucleophilic agent. Therefore, the cleavage reaction only occurs when a water molecule attacks the thioester. From the ligation data it seems that the cleavage reaction rate with water as nucleophile is at least10 fold lower than the ligation reaction rate.

From the ligation data with other reducing agents than TCEP, the cleavage rate is similar or higher than the ligation rate. Additionally the cleavage and ligation occur at a low rate compared to TCEP (Tab. 3).

Tab. 3: Comparison of the ligation and cleavage rates using different reducing agents. All the thiol-based reducing agents show the same trend. The cleavage rate is similar to or higher than the ligation rate. Higher concentrations of reducing agent apparently enhance the cleavage rate more than the ligation rate, resulting in more cleavage product. TCEP has a 1000 times higher ligation rate than any other reducing agent tested here. The ligation rate from the experiment with TCEP was calculated as first order reaction while for the other experiments the ligation and cleavage rate were obtained by using the a first order competitive model.

	k_{lig} [s^{-1}]	k_{cleav} [s^{-1}]
2.5mM DTT	$5 \cdot 10^{-06}$	$1 \cdot 10^{-05}$
5mM DTT	$5 \cdot 10^{-06}$	$2 \cdot 10^{-05}$
10mM DTT	$6 \cdot 10^{-06}$	$3 \cdot 10^{-05}$
2.5mM ME	$1 \cdot 10^{-06}$	$1 \cdot 10^{-06}$
5mM ME	$2 \cdot 10^{-06}$	$3 \cdot 10^{-06}$
10mM ME	$3 \cdot 10^{-06}$	$5 \cdot 10^{-06}$
2.5mM Cysteamine	$7 \cdot 10^{-07}$	$1 \cdot 10^{-06}$
5mM Cysteamine	$2 \cdot 10^{-06}$	$7 \cdot 10^{-06}$
10mM Cysteamine	$2 \cdot 10^{-06}$	$2 \cdot 10^{-05}$
5mM GSH at pH 8	$3 \cdot 10^{-06}$	$3 \cdot 10^{-06}$
5 mM TCEP	$1 \cdot 10^{-03}$	---

All reducing agents except for glutathione show the highest yield of ligated protein at pH 7 and at room temperature. Glutathione was the only reducing agent active at pH 8 but cleavage was dominant (Fig. 89). Higher pH value, temperature and concentration of reducing agent resulted in complete cleavage.


kDa M 1 2 3 4 5 6 7 8 9 10 11 12 13
117
85
48
34
26
19
Ligated AlgE4
A-Int$_N$
Cleaved A-module
Int$_C$-R
Time: 0,5 h 1 h 2 h 3 h 4 h 26 h 0,5 h 1 h 2 h 3 h 4 h 26 h

Lane 1-6: Ligation progress with 5 mM glutathione at room temperature and pH 8

Lane 7: empty

Lane 8-13: Ligation progress with 5 mM cysteamine at roomtemperature and pH 8

Fig. 89: Glutathione is active at pH8 and minimal ligation occurs. Cysteamine enhances cleavage at pH8 and no ligation is detectable.

The results obtained from the ligation tests with the split intein of *N. puntiforme* described above and further ligations tests described in Paper I show the influence of the extein on the ligation yield. Although the same intein and the same amino acids around the splicing site where used in all experiments described above, the ligation rate and yield varies between different exteins. It seems that steric repulsion at the splicing site can limit or inhibit successful ligation.

3.1.2 *In vivo* ligation

Protein ligation can be performed *in vitro* and *in vivo*. For *in vivo* ligation the constructs of the exteins plus the corresponding intein fragment have to be cloned into two different vectors. These vectors must have different antibiotic resistance and inducing systems. A-Int$_N$ was cloned into a vector under control of the T7-promoter and with kanamycin resistance gene, while Int$_C$-R was cloned into a BAD vector under control of pBAD promoter with ampicillin resistance gene. For segmentally labelled proteins, *in vivo* ligation is more demanding during the expressions of the fragments as the medium has to be changed but ligation *in vivo* normally results in high yield. We also tried to ligate the modules of AlgE4 *in vivo*, however, the yield was minimal. The main problem was that at 37°C the Int$_C$-R-module fragment is produced as inclusion bodies and cannot be used for *in vivo* ligation. Lowering the temperature would reduce the formation of inclusion bodies, however, the BAD vector cannot be used for low temperature expression. The optimal temperature was found to be between 25 and 30° C but the yield was relatively low (Fig. 90). At higher concentration of L-arabinose more Int$_C$-R is produced but nevertheless, only a fraction of the modules ligate and unligated fragments are accumulating in the cell.

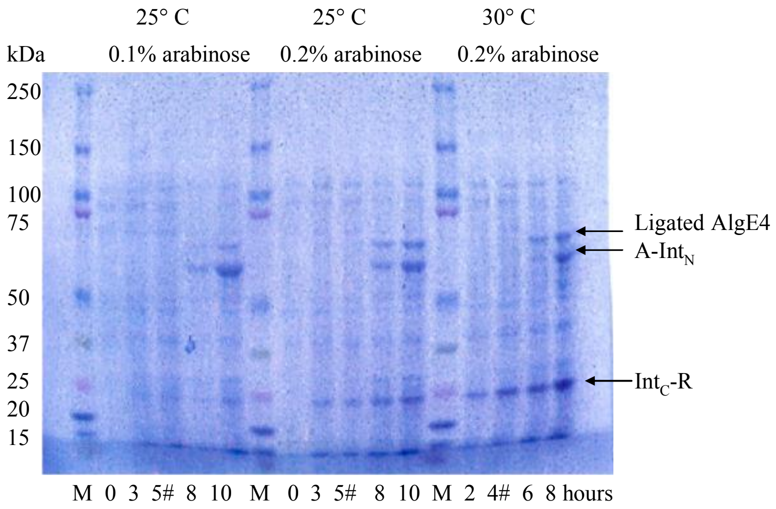
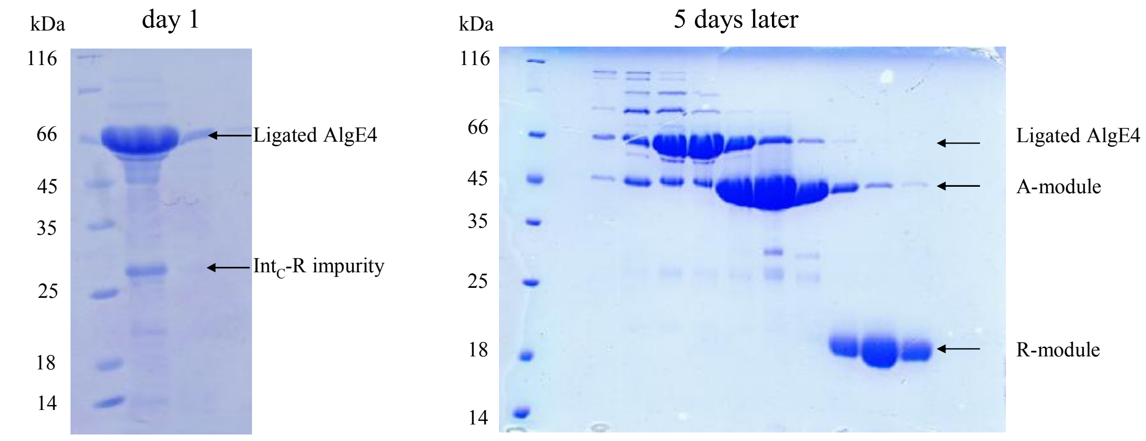


Fig. 90: Different *in vivo* ligation tests in unlabelled M9 medium. The Int$_C$-R was expressed first after induction with L-arabinose. After 5 or 4 h, respectively, the medium was exchanged, indicated here with # and A-Int$_N$ was expressed after adding IPTG. In every experiment shown here, *in vivo* ligated AlgE4 was produced but the yield was minimal.

For successful *in vivo* ligation of AlgE4 it would be best to clone the construct Int$_C$-R into a different vector system that can be used at 16-20° C.

3.1.3 Decomposition of ligated AlgE4

The first segmentally labelled AlgE4 [A-^{15}N-R] that we produced, was not completely pure after purification with His-tag (Fig. 91). Therefore, it was decided to further purify the sample by gelfiltration.



Fraction after His-tag purification

Fraction after gelfiltration

Fig. 91: SDS-**PAGE of the NMR sample.** The first NMR sample of *in vitro* ligated AlgE4 (A-^{15}N-R) that was measured had still a small amount of Int$_C$-R impurity. As this impurity was labelled, we used gelfiltration to get rid of the remaining Int$_C$-R. The results of the gelfiltration showed severe decomposition into the single modules of AlgE4. The fractions of the lowest band were collected and a ^{15}N-HSQC confirmed that it is the cleaved R-module, still correctly folded.

However, SDS-PAGE analysis of the fractions obtained from gel filtration showed that the segmentally labelled AlgE4 had undergone severe decomposition. A ^{15}N-HSQC measurement confirmed that the lowest band indeed was correctly folded R-module (Fig. 92).

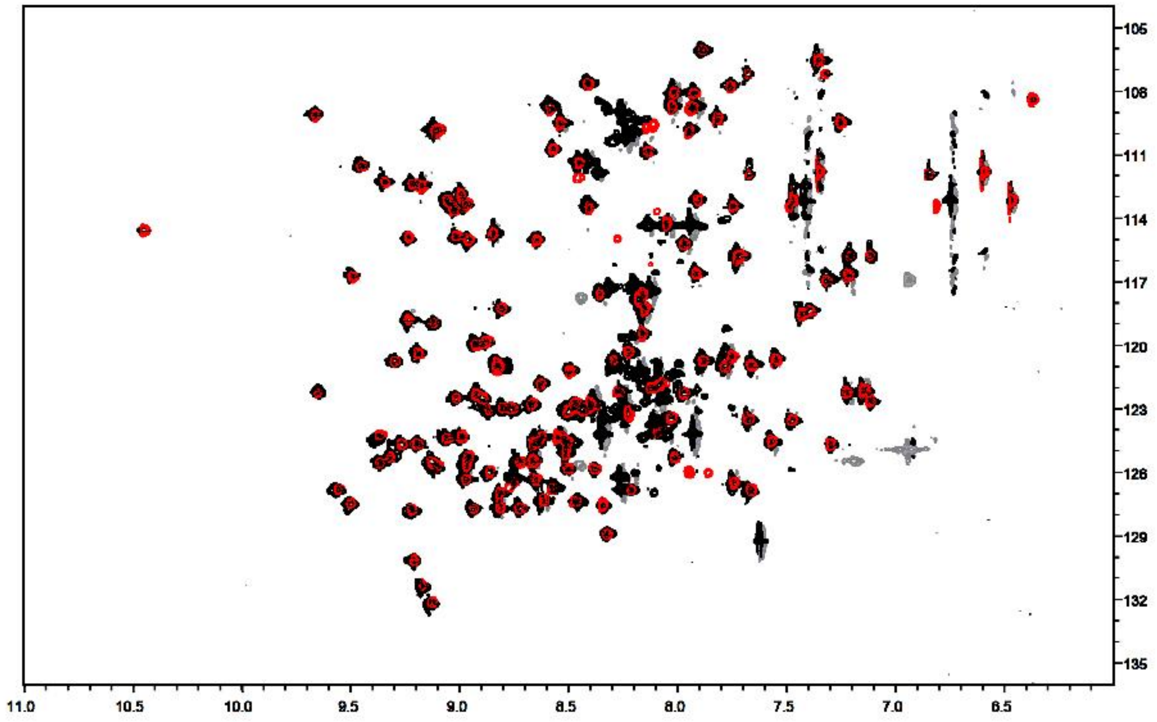


Fig. 92: Overlay of the R-module spectrum (in black) with the spectrum of the sample of the lowest band of the gelfiltration. It is obvious that the sample obtained from the gelfiltration is the R-module. It seems that the cleavage occurred at the splicing junction as the peak of the amino acid E5 could be confirmed. The R-module shows extra peaks deriving from its unstructured tail (20 amino acids) which were removed in the ligation construct.

This was surprising as AlgE4 is reported to be stable in solution. Although it was not further investigated, it seems that the exchange of the amino acids for optimal splicing resulted in lower resistance against proteases. For all further experiments, protease inhibitor was added at every step of purification and analysis to preserve the integrity of segmentally labelled AlgE4 and experiments were only performed on freshly prepared samples. Additionally, after the measurements, an SDS-PAGE was used to confirm the stability of the ligation product during the experiment.

3.1.4 Segmentally labelled AlgE4

After optimizing the ligation and adding protease inhibitor into the buffer solution, differently segmentally labelled AlgE4 samples were produced. In the first segmentally labelled AlgE4 only the R-module was ^{15}N-labelled. The ^{15}N-HSQC spectrum of this AlgE4 (A-^{15}N-R) was similar to a spectrum of an R-module alone but the line widths of the peaks of the segmentally labelled AlgE4 was - as expected - broader. This limits further investigation as severe overlapping occurs (Fig. 93). Therefore, deuterated segmentally labelled AlgE4 samples (A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R) were obtained.

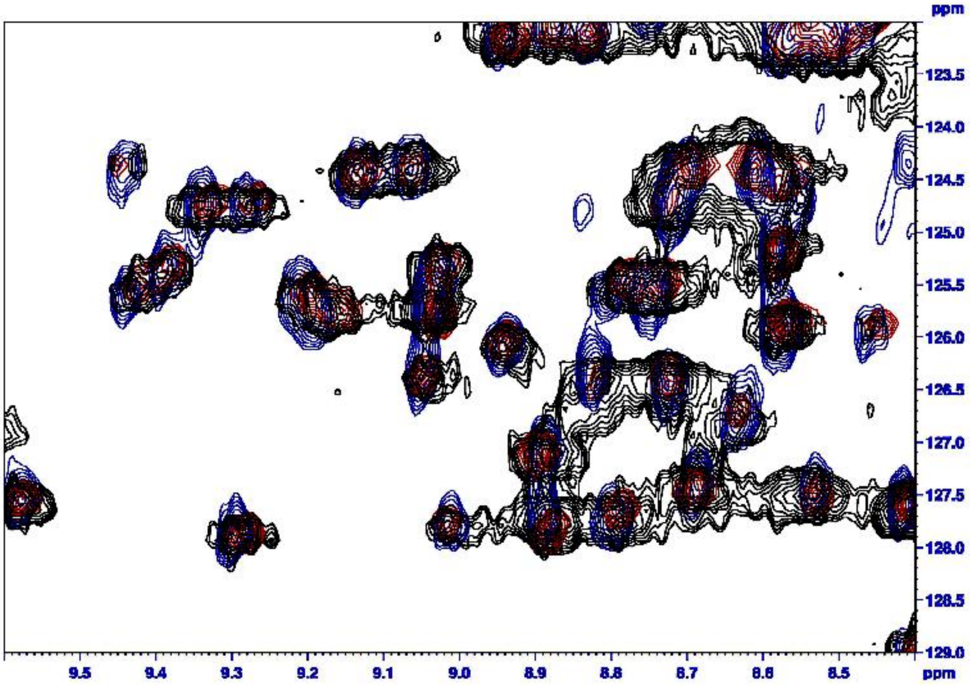


Fig. 93: Overlay and zoom in of three spectra. In red is a ^{15}N-HSQC spectrum of the R-module alone. The ^{15}N-HSQC spectrum in black is the segmentally labelled AlgE4 (A-[^{15}N]-R) and the blue labelled TROSY spectrum is the deuterated, segmentally labelled AlgE4 (A-[^{2}H, ^{15}N]-R). The line width of the segmental labelled AlgE4 is relative broad compared to the R-module alone. Deuteration and the TROSY effect reduced the line width of the segmental labelled AlgE4 from 17.35 ± 1.7 Hz to 11.2 ± 1.3 Hz. This was similar to the line width of the R-module alone 10.1 ± 0.9 Hz.

It was possible to obtain HSQC and TROSY spectra of the deuterated segmentally labelled AlgE4. However, the concentration of the pure, partially deuterated, proteins (A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R) after refolding and purification was far too low (20 µM) to allow extensive studies of structure and ligand binding. Therefore, these plans had to be abandoned.

3.2 The modules of the extracellular alginate epimerases from Azotobacter vinelandii *(Papers III, IV & V)*

3.2.1 A-module

In this thesis, the A-modules of the extracellular epimerases were hardly investigated. Nevertheless, a short summary of the A-modules is included here, as the A-modules contain the catalytic site and it is expected that the binding of alginates to the R-modules correlates with the type of epimerases. *Azotobacter vinelandii* has 7 extracellular alginate epimerases and one protein designed AlgY which has a high similarity to the alginate epimerases but does not show any alginate epimerisation activity [168,169]. Only the A-module of the extracellular alginate epimerases is active [172]. AlgE1 and AlgE3 have two A-modules of which the first A-module in both enzymes epimerize alginates to GG-blocks while the second A-module belongs to the group of MG-block epimerizing ones [175]. The function of AlgY is not known but the protein could be found in at least two *Azotobacter* genomes from different sources [169,238]. The A-modules have a high sequence identity and similarity. Sequence alignment of the A-module (plus AlgYA) shows that the A-modules can be divided into three groups (Fig. 94). One group are the MG-block epimerising A-modules plus AlgE6A (which is actually a GG-block producing A-module). The second group consists of the remaining GG-block epimerising A-modules. The third group consists of AlgE7A and AlgYA.

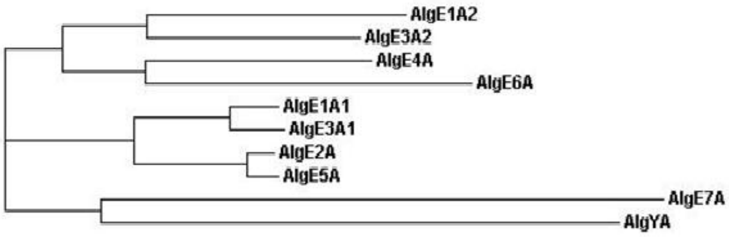


Fig. 94: Phylogram of the A-modules. The A-modules are subdivided into three groups which correlate with their functions (except for AlgE6A) [169,239].

If AlgE7 and AlgY are excluded the amino acids identity is to 73%. Additional 15% of the amino acids are highly similar. At the beginning of the A-modules nearly every amino acid is conserved while the final 80 amino acids are less conserved (Fig. 95). Amino acids that differ between MG- and GG-block epimerases were labelled. The A-modules of AlgE7 and AlgY are not included as they show significant differences to the other A-modules.

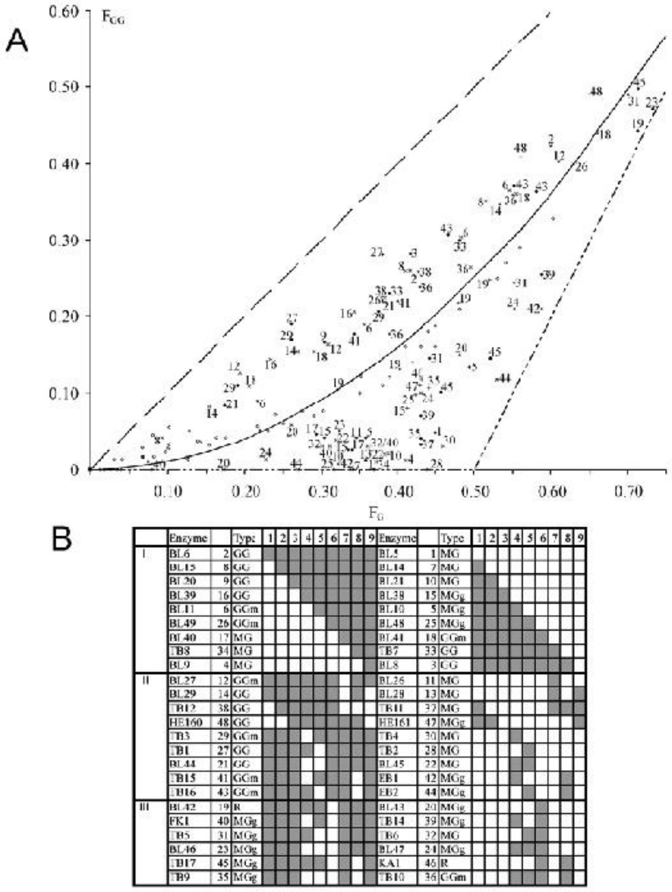
```
AlgE1A2         -VFNAKDFGALGDGASDDRPAIQAAIDAAYAAGGGTVYLPAGEYRVSPTGEPGDGCLMLK 59
AlgE3A2         -VFNAKDFGALGDGASDDRPAIQAAIDAAYAAGGGTVYLPAGEYRVSPTGDPGDGCLMLK 59
AlgE4A          MDYNVKDFGALGDGVSDDRASIQAAIDAAYAAGGGTVYLPAGEYRVSAAGEPGDGCLMLK 60
AlgE6A          MDYNVKDFGALGDGVSDDRVAIQAAIDAAHAAGGGTVYLPPGEYRVSAAGEPSDGCLTLR 60
AlgE1A1         MDYNVKDFGALGDGVSDDTAAIQAAIDAAHAAGGGTVYLPAGEYRVSGGEEPSDGCLTIK 60
AlgE3A1         MDFNVKDFGALGDGASDDTAAIQAAIDAAHAAGGGTVYLPAGEYRVSGGEEPSDGALTIK 60
AlgE2A          MDYNVKDFGALGDGVSDDTAAIQAAIDAAYAAGGGTVYLPAGEYRVSGGEEPSDGCLTIK 60
AlgE5A          MDYNVKDFGALGDGVSDDTAAIQAAIDAAYAAGGGTVYLPAGEYRVSGGEEPSDGCLTIK 60
AlgYA           MDFNVKDSGALGDGVSDDRAAIQAAIDAAHAAGGGTVYLPAGEYRVSGGERGVDGALMMK 60
AlgE7A          MEYNVKDFGAKGDGKTDDTDAIQAAIDAAHKAGGGTVYLPSGEYRVSGGDEASDGALIIK 60
                  :*.** ** *** :**  :********: *********.******      **.* ::

AlgE1A2         DGVYLAGDGIGETVIKLIDGSDQKITGMVRSAYGEETSNFGMSDLTLDGNRDN-TSGKVD 118
AlgE3A2         DGVYLVGAGMGETVIKLIDGSDQKITGMVRSAYGEETSNFGMSDLTLDGNRDN-TSGKVD 118
AlgE4A          DGVYLAGAGMGETVIKLIDGSDQKITGMVRSAYGEETSNFGMRDLTLDGNRDN-TSGKVD 119
AlgE6A          DNVYLAGAGMGQTVIKLVDGSAQKITGIVRSPFGEETSNFGMRDLTLDGNRAN-TVDKVD 119
AlgE1A1         SNVHIVGAGMGETVIKMVDGWTQNVTGMVRSAYGEETSNFGMSDLTLDGNRDN-LSAKVD 119
AlgE3A1         SNVYIVGAGMGETVIKMVDGWTQNVTGMVRSAYGEETSNFGMSDLTLDGNRDN-LSAKVD 119
AlgE2A          SNVHIVGAGMGETVIKLVDGWDQDVTGIVRSAYGEETSNFGMSDLTLDGNRDN-TSGKVD 119
AlgE5A          SNVYIVGAGMGETVIKLVDGWDQDVTGIVRSAYGEETSNFGMSDLTLDGNRDN-TSGKVD 119
AlgYA           SNVYLAGAGMGETVVKLLDGWNGHVNGMIRSSGTEETHDFGVRDLTLDGNRDNNPEGTVF 120
AlgE7A          SNVYIVGAGMGETVIKLVDGWDEKLTGIIRSANGEKTHDYGISDLTIDGNQDN-TEGEVD 119
                ..*::.* *:*:**:*::**   .:.*::**.  *:* ::*: ***:***: *     *
```

```
AlgE1A2     GWFNGYIPGQDGADRNVTIERVEIREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADY 178
AlgE3A2     GWFNGYIPGQDGADRNVTLERVEIREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADY 178
AlgE4A      GWFNGYIPGGDGADRDVTIERVEVREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADY 179
AlgE6A      GWFNGYAPGQPGADRNVTIERVEVREMSGYGFDPHEQTINLVLRDSVAHHNGLDGFVADY 179
AlgE1A1     GWFNGYIPGQDGADRDVTLERVEIREMSGYGFDPHEQTINLTIRDSVAHDNSLDGFVADY 179
AlgE3A1     GWFNGYIPGQDGADRDVTLERVEIREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADY 179
AlgE2A      GWFNGYIPGEDGADRDVTLERVEIREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADF 179
AlgE5A      GWFNGYIPGEDGADRDVTLERVEIREMSGYGFDPHEQTINLTIRDSVAHDNGLDGFVADF 179
AlgYA       GFYTGYKFG-DGADRNVIVERVEAREMSGYGFDPHARTVNLVIRDSVAHDNGFVGFVADH 179
AlgE7A      GFYTGYIPGKNGADYNVTVERVEIREVSRYAFDPHEQTINLTIRDSVAHDNGKDGFVADF 179
            *::.**  *  *** :* :**** **:* *.**** :*:**.:******.*.  *****.

AlgE1A2     LVDSVFENNVAYNNDRHGFNIVTSTYDFVMTNNVAYGNGGAGLTIQRGSEDLAQPTDILI 238
AlgE3A2     LVDSVFENNVAYNNDRHGFNVVTSTYDFTLSNNVAYGNGGAGLVIQRGAEDLAQPTDILI 238
AlgE4A      LVDSVFENNVAYANDRHGFNVVTSTHDFVMTNNVAYGNGSSGLVVQRGLEDLALPSNILI 239
AlgE6A      QIGGTFENNVAYANDRHGFNIVTSTNDFVMRNNVAYGNGGNGLVVQRGSENLAHPENILI 239
AlgE1A1     QVGGVFENNVSYNNDRHGFNIVTSTNDFVLSNNVAYGNGGAGLVVQRGSYDLPHPYDILI 239
AlgE3A1     QVGGVFENNVSYNNDRHGFNIVTSTNDFVLSNNVAYGNGGAGLVVQRGSYDLPHPYDILI 239
AlgE2A      QIGGVFENNVSYNNDRHGFNIVTSTNDFVLSNNVAYGNGGAGLVVQRGSSDVAHPYDILI 239
AlgE5A      QIGGVFENNVSYNNDRHGFNIVTSTNDFVLSNNVAYGNGGAGLVIQRGSYDVAHPYGILI 239
AlgYA       QIDGAFENNVAYNNDLHGFNVVTSSHDFTLSDNVAYGNGAAGLVVQRGSYDVPHAYNIRI 239
AlgE7A      QIGAVFENNVSYNNGRHGFNIVTSSHDIVFTNNVAYGNGANGLVVQRGSEDRDFVYNVEI 239
             :...*****:* *. ****:***: *:.: :*******. **.:***  :     .: *

AlgE1A2     DGGAYYDNALEGVLFKMTNNVTLQNAEIYGNGSSGVRLYGTEDVQILDNQIHDNSQNGTY 298
AlgE3A2     DGGAYYDNALEGVLLKMTNNITLQNAEIYGNGYSGVRLYGTEDVQILNNQIHDNAQNVAY 298
AlgE4A      DGGAYYDNAREGVLLKMTSDITLQNADIHGNGSSGVRVYGAQDVQILDNQIHDNAQAAAV 299
AlgE6A      DGGSYYDNGLEGVLVKMSNNVTVQNADIHGNGSSGVRVYGAQGVQILGNQIHDNAKTAVA 299
AlgE1A1     DGGAYYDNALEGVQLKMAHDVTLQNAEIYGNGLYGVRVYGAQDVQILDNQIHDNSQNGAY 299
AlgE3A1     DGGAYYDNALEGVQLKMTHDVTLQNAEIYGNGLYGVRVYGAQDVQLLDNQIHDNSQNGAY 299
AlgE2A      DGGAYYDNGLEGVQIKMAHDVTLQNAEIYGNGLYGVRVYGAEDVQILDNYIHDNSQNGSY 299
AlgE5A      DGGAYYDNGLEGVQIKMAHDVTLQNAEIYGNGLYGVRVYGAEDVQILDNYIHDNSQSGSY 299
AlgYA       DGGSYHDNALEGVLIKLSHDVTLQNAHIYDNGTAGVRIAGAQDVQLLDNRIHDNVQNGTY 299
AlgE7A      EGGSFHDNGQEGVLIKMSTDVTLQGAEIYGNGYAGVRVQGVEDVRILDNYIHDNAQSKAN 299
            :**:::**. *** .*:: ::*:*.*.*:.**  ***: *.:.*::*.* **** :

AlgE1A2     PEVLLQAFDDSQ-VTGELYETLNTRIEGNLIDASDNANYAVRERDDGSDYTTLVDNDISG 357
AlgE3A2     AEVLLQSFNDVG-VSGNFYATTGTWIEGNVISGSANSTYGIEERNDGTDYSSLYANTIDG 357
AlgE4A      PEVLLQSFDDTAGASGTYYTTLNTRIEGNTISGSANSTYGIQERNDGTDYSSLIDNDIAG 359
AlgE6A      PEVLLQSYDDTLGVSGNYYTTLNTRVEGNTITGSANSTYGVQERNDGTDFSSLVGNTING 359
AlgE1A1     AEVLLQSYDDTAGVSGNFYVTTGTWLEGNVISGSANSTYGIQERADGTDYSSLYANSIDG 359
AlgE3A1     AEVLLQSYDDTAGVSGNFYVTTGTWLEGNVISGSANSTFGIQERADGTDYSSLYANTIDG 359
AlgE2A      AEILLQSYDDTAGVSGNFYTTTGTWIEGNTIVGSANSTYGIQERDDGTDYSSLYANSVSN 359
AlgE5A      AEILLQSYDDTAGVSGNFYTTTGTWIEGNTIVGSANSTYGIQERADGTDYSSLYANSVSN 359
AlgYA       PEVLLQAFDDSG-ITGNVYETLNTLIEGNLITTSGDATYIVQERNDGSDYTTLRDNGISG 358
AlgE7A      AEVIVESYDDRDGPSDDYYETQNVTVKGNTIVGSANSTYGIQERADGTDYTSIGNNSVSG 359
            .*:::::::*    :.  * * .. ::** *  * ::.: :.** **:*::::  * : .

                                Linker
AlgE1A2     GQVASVQLSGAHSSLSG--GTVEVPQ-- 381
AlgE3A2     VQTGAVRLNGAHSIVSDQPGTGQQATLE 385
AlgE4A      VQQP-IQLYGPHSTVSGEPGATPQQPST 386
AlgE6A      VQEA-AHLYGPNSTVSGTVSAPPQ---- 382
AlgE1A1     VQTGAVRLYGANSTVSSQSGSGQQATLE 387
AlgE3A1     VQNGTVRLYGANSTVSEQPSSGQQATLE 387
AlgE2A      VQNGSVRLYGANSVVSDLPGTGQQATLE 387
AlgE5A      VQSGSVRLYGTNSVVSDLPGTGQQATLE 387
AlgYA       GQIASVQLSGAHSSSGPLR--------- 377
AlgE7A      TQRGIVQLSGTNSTFSGRSGDAYQFID- 386
               *     :* *.:*   .
```

Fig. 95: Alignment of the A-modules. Amino acids that maybe have an influence the epimerisation are marked. ▨: Amino acids known to be involved in alginate binding or epimerisation reaction. **Comparison of G- and MG-epimerising A-modules**; 🟩: AlgE1A2 and AlgE3A2 have different amino acids than the rest; A: AlgE1A2 and AlgE3A2 have the same residue but the rest is not identical on that position or AlgE1A2 and AlgE3A2 have different residues that also differ from the remaining A-modules; 🟦: AlgE1A2, AlgE3A2, AlgE4A and AlgE6A have one amino acid type while remaining A-modules share an other amino acid. A: Only three out of AlgE1A2, AlgE3A2, AlgE4A or AlgE6A have the same amino acid but all four have a different amino acid to the rest; 🟨: the amino acid of Alg1A2 and AlgE3A2 differ from AlgE4A and AlgE6A and both

group differ from the remaining A-modules; 🟥: The MG-epimerising A-modules share one amino acid while the rest have a different one; A: Only two out of AlgE1A2, AlgE3A2 or AlgE4A have the same amino acid but all three have a different amino acid to the rest 🟦: Alg4A and AlgE6A have the same amino acid but differ from the rest; A: Alg4A and AlgE6A have different amino acid but differ also from the remaining A-modules or AlgE4A, AlgE6A and a three A-module have the same amino acid; A: AlgE1A1, AlgE2A, AlgE3A and AlgE5 share one amino acid while at least two of the remaining have an other amino acid; * indicates residue that are identical in all A-modules, : denotes conserved amino acids and shows semi-conserved amino acids (ligation were performed with ClustalW2) [239].The fragment from residue 215-263 (AlgE4A) had significant influence on the epimerization pattern. (see also **Fig.96**)

There are only 11 positions where the MG-epimerizing A-modules differ from the GG-block producing one (red labelled). The differences between MG- and G-epimerizing A-modules were investigated [240]. For this study 46 hybrid epimerases were constructed where parts of DNA sequence encoding AlgE4A were exchanges with AlgE2A (Fig.96). Amongst others, two hybrid series were made starting from AlgE4A and exchanging one fragment after each other and vice versa. From this series it is known that the area of Y215-Q263 has a great impact on the epimerisation pattern [240]. This is very surprising as the alignment of this part does not show any position where the difference between the A-modules correlates with the epimerisation pattern. This region is downstream to the active site meaning that how the epimerase interacts with the product influences epimerisation pattern.



	Enzyme		Typ:	1	2	3	4	5	6	7	8	9	Enzyme		Type	1	2	3	4	5	6	7	8	9
I	BL6	2	GG	▨	▨	▨	▨	▨	▨	▨	▨	▨	BL5	1	MG									
	BL15	8	GG		▨	▨	▨	▨	▨	▨	▨	▨	BL14	7	MG	▨								
	BL20	9	GG			▨	▨	▨	▨	▨	▨	▨	BL21	10	MG	▨	▨							
	BL39	16	GG				▨	▨	▨	▨	▨	▨	BL38	15	MGg	▨	▨	▨						
	BL11	6	GGm					▨	▨	▨	▨	▨	BL10	5	MGg	▨	▨	▨	▨					
	BL49	26	GGm						▨	▨	▨	▨	BL48	25	MGg	▨	▨	▨	▨	▨				
	BL40	17	MG							▨	▨	▨	BL41	18	GGm	▨	▨	▨	▨	▨	▨			
	TB8	34	MG								▨	▨	TB7	33	GG	▨	▨	▨	▨	▨	▨	▨		
	BL9	4	MG									▨	BL8	3	GG	▨	▨	▨	▨	▨	▨	▨	▨	
II	BL27	12	GGm	▨	▨	▨	▨	▨	▨		▨	▨	BL26	11	MG							▨		
	BL29	14	GG	▨	▨	▨	▨	▨	▨		▨		BL28	13	MG							▨		▨
	TB12	38	GG		▨	▨	▨	▨	▨				TB11	37	MG	▨						▨	▨	▨
	HE160	48	GG			▨	▨	▨	▨	▨	▨		HE161	47	MGg	▨	▨							▨
	TB3	29	GGm	▨	▨	▨		▨	▨	▨	▨	▨	TB4	30	MG				▨					
	TB1	27	GG	▨	▨	▨	▨		▨	▨	▨	▨	TB2	28	MG					▨				
	BL44	21	GG	▨	▨	▨			▨	▨	▨	▨	BL45	22	MG				▨	▨				
	TB15	41	GGm	▨	▨	▨		▨	▨	▨		▨	EB1	42	MGg				▨				▨	
	TB16	43	GGm	▨	▨	▨	▨		▨	▨		▨	EB2	44	MGg					▨			▨	
III	BL42	19	R	▨	▨	▨	▨	▨		▨	▨	▨	BL43	20	MGg						▨			
	FK1	40	MGg	▨	▨	▨		▨		▨	▨	▨	TB14	39	MGg				▨		▨			
	TB5	31	MGg	▨	▨	▨	▨			▨	▨	▨	TB6	32	MG					▨	▨			
	BL46	23	MGg	▨	▨	▨				▨	▨	▨	BL47	24	MGg				▨	▨	▨			
	TB17	45	MGg	▨	▨	▨	▨	▨		▨		▨	KA1	46	R						▨		▨	
	TB9	35	MGg	▨	▨	▨				▨		▨	TB10	36	GGm				▨	▨	▨		▨	

Fig.96: Structure and epimerization patterns of the hybrid enzymes; A) The broken line indicates the maximal amount of GG-blocks. The dotted line represents alginates that are completely epimerized to MG-alginates before more G are inserted. The epimerisation pattern of each hybrid epimerase must lie in these boundaries. The pattern of random attack is represented by the solid line. B) The different hybrid epimerases that were made. White area means the fragment is from AlgE4A while fragments that have the sequence from AlgE2A are in grey.

The hybrid where the sequence M1-A214 was exchanged from AlgE4A to AlgE2A (BL48) has similar epimerisation pattern like AlgE4A. This is surprising as many of the amino acids that differ after epimerisation pattern and most of the amino acids that participate in epimerisation or binding have been exchanged (Fig.97). It seems that small changes in the end part of the A-modules have an influence on the epimerisation pattern. These hybrid studies could show areas which influence the epimerisation pattern but the actual amino acids that cause the difference in epimerisation could not be determined.

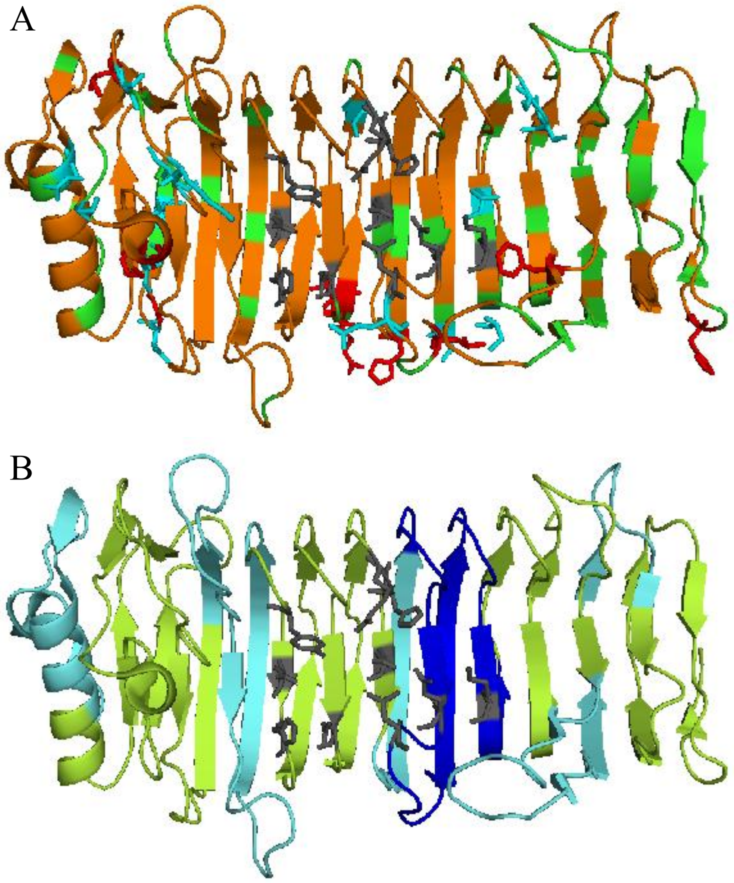


Fig.97: A) The conserved amino acids plotted on the tertiary structure of AlgE4A (in orange). The red labelled amino acids indicate positions where the MG-block epimerases differ from the GG-block epimerases. The cyan labelled amino acids indicate positions where MG-epimerases plus AlgE6 differ from the other epimerases. In gray are those amino acids that are known to participate in epimerisation reaction or in binding. B) The fragments described in Fig.96 are labelled in alternating colours on the structure of AlgE4A. The fragment 6 is labelled in dark blue as its exchange has the most influence on epimerisation. Amino acids that are known to participate in epimerisation reaction or in binding are shown in side chains in gray. The visualizing of the conserved amino acids and fragments of the A-module was done with Pymol (2006 DeLano Scientific LLC).

AlgYA is inactive, nevertheless AlgYA has the same residues at the active site as the rest of the A-modules. The only difference is that R195 involved in binding alginate is mutated to a leucine.

3.2.2 R-modules

It is known that *A. vinelandii* is expressing 34 different R-modules including the R-modules of ORF9 and AlgY. The phylogram of all 34 R-modules is shown in Fig. 98.

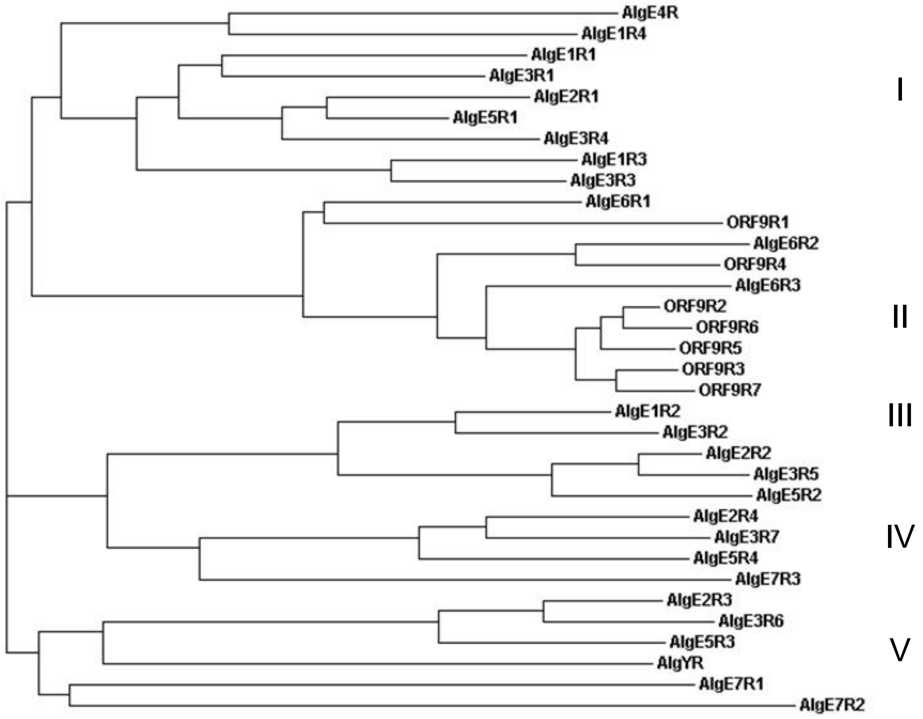


Fig. 98: Phylogram of the R-modules of *A. vinelandii*. The R-modules can be subdivided into 5 groups [239].

The R-modules can be subdivided into 5 groups. The first group consists mainly of R-modules that immediately follow an A-module like AlgE4R, AlgE1R1 or AlgE1R4. The second group consists of the R-modules of AlgE6 and ORF9. This is the only group where all the R-modules of those two proteins are clustered together. The third group consists only of R-modules that are at the second position after an A-module. The R-modules of the fourth group are the C-terminal ones of their respective epimerases. The fifth group consists on the one hand of R-modules that are the third but not the last R-module after an A-module and AlgYR, AlgE7R1 and AlgE7R2 that are relatively different to the other R-modules. It is quite evident that the single R-modules of AlgE1 and AlgE3-1 as well as the R-modules of AlgE2, AlgE5 and of AlgE3-2 always cluster together. Several amino acids are highly conserved when all the R-modules are compared (Fig. 99). For the β-roll, the conserved amino acids are mainly the Asp and Gly residues essential for calcium binding. The amino acid sequence which leads from the β-roll to the first long antiparallel β-sheet (L84 - L100 in AlgE4R) is also highly conserved. The second antiparallel β-sheet is completely conserved and also the third long antiparallel β-sheet has several conserved amino acids.

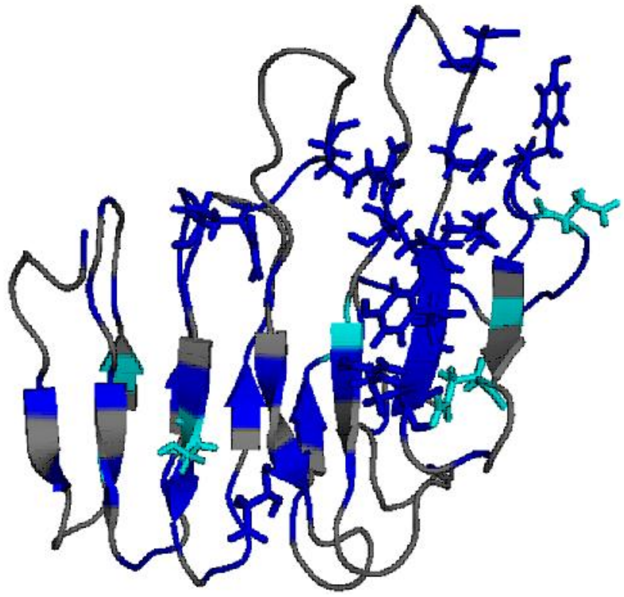


Fig. 99: The conserved and similar amino acids of the R-modules are superimposed on AlgE4R. 🟦: Identical amino acid or few R-modules (four or fewer) have a mutation on this position. ▨: All R-modules have a similar amino acid on that position. Amino acids that are conserved or similar and that are maybe involved in binding are shown here with side chain. In AlgE4R these are R40, S44, R62, D74, D94, Y96, N105, R110, Y112, K114, E117, D119 and E126. Labelling of the conserved amino acid performed in Pymol (2006 DeLano Scientific LLC).

3.2.3 Orientation studies

Low-resolution solution structures of the alginate epimerases AlgE4 and AlgE6 were determined by SAXS. Both epimerases show a defined orientation of the modules with limited flexibility in between. The angle between the A- and R-module of AlgE4 is approximately 120°. The structure of AlgE6 is more complicated; the arrangement is shown in Fig. 100.

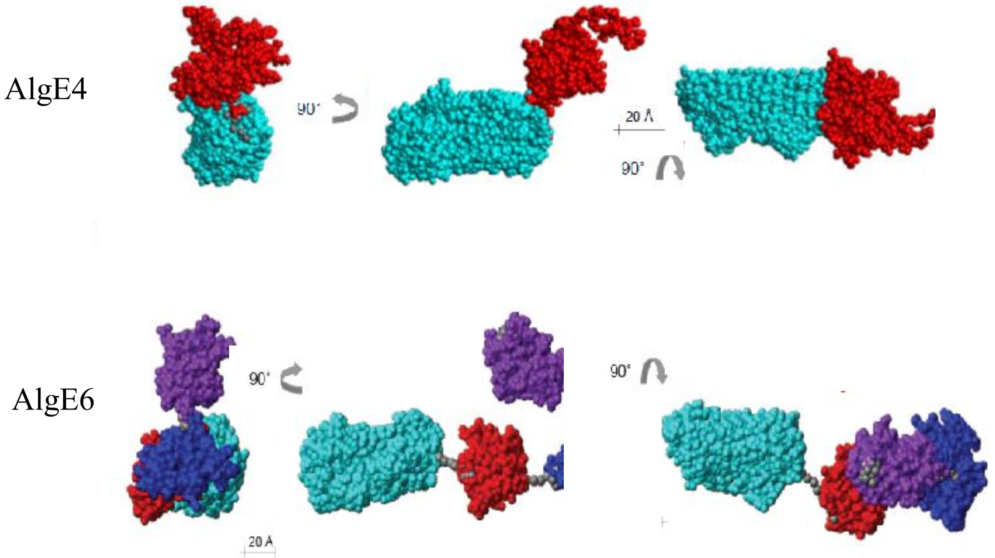


Fig. 100: The low resolution structures of AlgE4 and AlgE6 show a defined orientation between the modules. The angle between the modules of AlgE4 is around 120°. AlgE6 has a similar angle between its A-module and the first R-module moreover the arrangement of the three R-modules to each other reminds on a triangle.

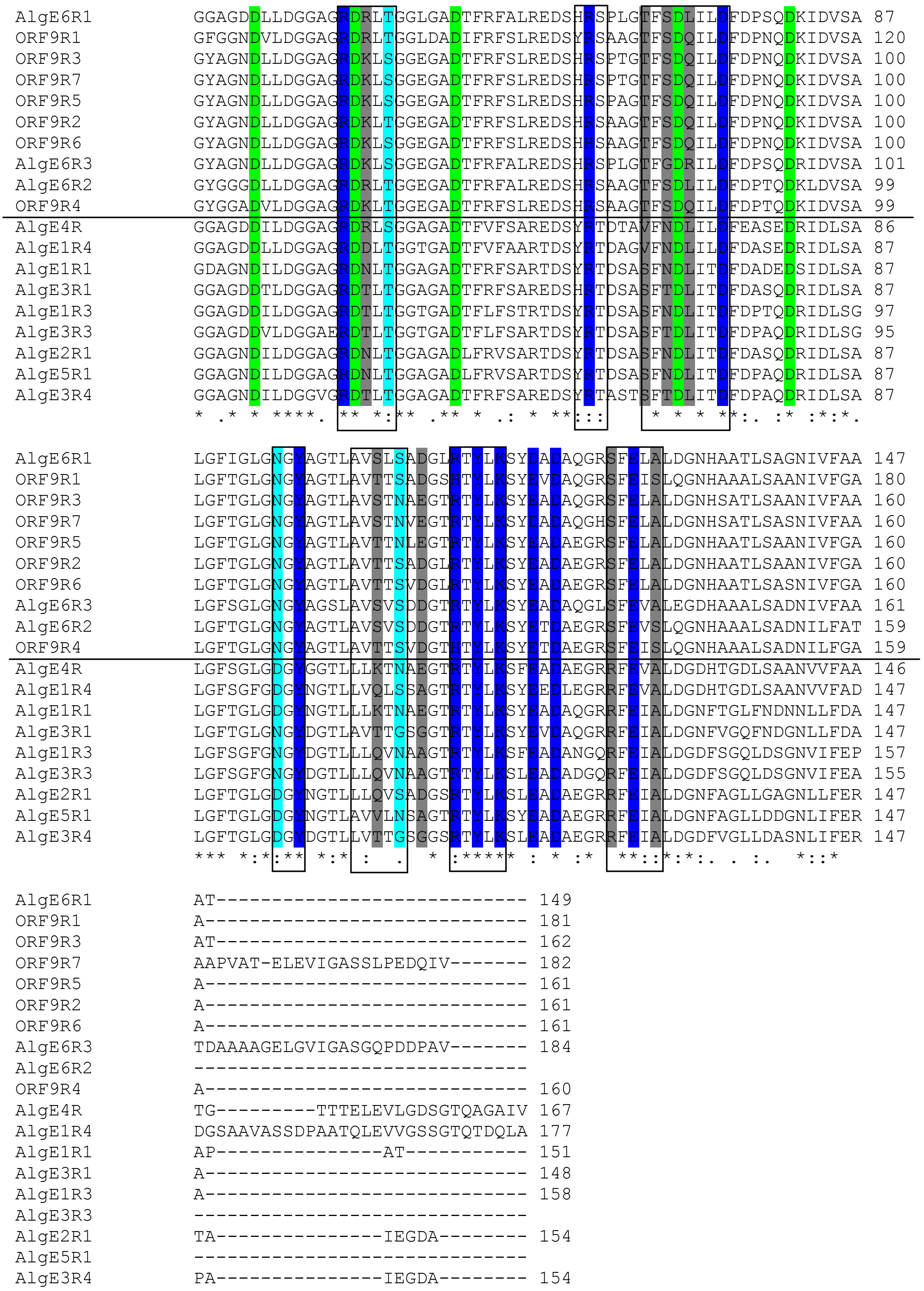
3.2.4 Binding studies

Binding of alginates was tested on the R-modules of AlgE4 and AlgE6 In a M.Sc. thesis project conducted by someone else in parallel, – alginate binding to ORF9 – a protein consisting of only nine R-modules – was investigated [241]. These R-modules belong to two different groups. AlgE4R belongs to group I and it was the first R-module whose structure was determined [173]. The epimerase AlgE6 is highly homologous to AlgE4 both in A- and R-module but AlgE4 is a MG-block forming epimerase while AlgE6 produces GG-blocks. The R-modules of AlgE6 were chosen because they are homologous to AlgE4 but are part of a GG-block forming epimerase. AlgE6 is the only epimerase where all R-modules are in one cluster. The R-modules of AlgE6 and ORF9 belong to one cluster (group II). The function of ORF9 is not known, but it is not an alginate epimerase as ORF9 does not have an A-module. It is under the same operon as most of the alginate epimerases and it could be found in at least two *Azotobacter* genomes from different sources [168,169,238].

The assignment of the R-module of AlgE4 had been determined before [242]. Therefore, binding between the AlgE4R and alginates could be studied by NMR. The single R-modules of AlgE6 were assigned in order to follow alginate binding by NMR [243,244,245]. All four single R-modules of AlgE4 and AlgE6 were also studied by ITC. ORF9 was studied as whole protein thus alginate binding was investigated only by ITC [241].

The binding results are very interesting. AlgE4R binds strongly to poly-M alginates but the binding constants to MG-alginates are 100-fold lower than to M-alginates with the same length. The shortest alginate that was tested, was M3 and elongation by one mannuronic acid resulted in 10 times stronger binding until M5 (see also Tab. 4), thereafter increase in degree of polymerisation only led to small increases in binding constant. Amino acids that are affected most by alginate binding were determined and the binding site was postulated (Fig. 101). The binding constants obtained by NMR and ITC are comparable.

```
AlgE6R1         ---------------------------------GTDGNDVLIGSDVGEQISGGAGDDRLD 27
ORF9R1          MSGQEQLVVEGTTDENGNPVVSEGPSIETTAVAGTEGNDLLYGTEVGEELVGGAGDDRLY 60
ORF9R3          --------------------VPVDPNVEGTPIVGSDLDDVLHGTLGSEQVLGGGGADQLY 40
ORF9R7          --------------------VPVDPNVEGTPVVGSDLDDELHGTLGSEQILGGGGADQLY 40
ORF9R5          --------------------VPVDPNVEGTPVIGSDLDDELHGTLGSEQILGGGGADQLY 40
ORF9R2          --------------------VPVDPNVEGTPVVGSDLDDELHGTLGSEQILGGGGADRLY 40
ORF9R6          --------------------VPVDPNVEGTPIVGSDLDDELHGTLGSEQILGGGGADQLY 40
AlgE6R3         -------------------PVPVDPGVEGTPVVGSDLDDELHGTLGSEQILGGGGADQLY 41
AlgE6R2         ---------------------PVDPSAEAQPIVGSDLDDQLHGTLLGEEISGGGGADQLY 39
ORF9R4          ---------------------PVDPSVEGTPIVGSDLDDQLHGTVLGEEISGGGGADRLY 39
AlgE4R          ---------------------------------GSDG-EPLVGGDTDDQLQGGSGADRLD 26
AlgE1R4         ---------------------------------GTDGNDVLVGSDANDQLYGGAGDDRLD 27
AlgE1R1         ---------------------------------GSAGNDALSGTEAHETLLGQAGDDRLN 27
AlgE3R1         ---------------------------------GTAGNDVLSGTGAHELILGLAGNDRLD 27
AlgE1R3         -----------------------SQGGQMTIIEGTDGNDTLQGTEANERLLGLDGRDNLN 37
AlgE3R3         -------------------------PPEQATIEGTDGNDSLQGTGADELLLGLGGRDSLN 35
AlgE2R1         ---------------------------------GTAGNDTLGGSDAHETLLGLDGNDRLN 27
AlgE5R1         ---------------------------------GTTGNDTLTGSEAHETLLGLDGNDRLN 27
AlgE3R4         ---------------------------------GTTGNDTLGGSDAHETLLGLDGDDRLD 27
                                                 *:   : * *    : : *  * * *
```



```
AlgE6R1         GGAGDDLLDGGAGRDRLTGGLGADTFRFALREDSHRSPLGTFSDLILDFDPSQDKIDVSA 87
ORF9R1          GFGGNDVLDGGAGRDRLTGGLDADIFRFSLREDSYRSAAGTFSDQILDFDPNQDKIDVSA 120
ORF9R3          GYAGNDLLDGGAGRDKLSGGEGADTFRFSLREDSHRSPTGTFSDQILDFDPNQDKIDVSA 100
ORF9R7          GYAGNDLLDGGAGRDKLSGGEGADTFRFSLREDSHRSPTGTFSDQILDFDPNQDKIDVSA 100
ORF9R5          GYAGNDLLDGGAGRDKLSGGEGADTFRFSLREDSHRSPAGTFSDQILDFDPNQDKIDVSA 100
ORF9R2          GYAGNDLLDGGAGRDKLTGGEGADTFRFSLREDSHRSAAGTFSDQILDFDPNQDKIDVSA 100
ORF9R6          GYAGNDLLDGGAGRDKLSGGEGADTFRFSLREDSHHSAAGTFSDQILDFDPNQDKIDVSA 100
AlgE6R3         GYAGNDLLDGGAGRDKLSGGEGADTFRFALREDSHRSPLGTFGDRILDFDPSQDRIDVSA 101
AlgE6R2         GYGGGDLLDGGAGRDRLTGGEGADTFRFALREDSHRSAAGTFSDLILDFDPTQDKLDVSA 99
ORF9R4          GYGGADVLDGGAGRDKLTGGEGADTFRFSLREDSHRSAAGTFSDQILDFDPTQDKIDVSA 99
AlgE4R          GGAGDDILDGGAGRDRLSGGAGADTFVFSAREDSYRTDTAVFNDLILDFEASEDRIDLSA 86
AlgE1R4         GGAGDDLLDGGAGRDDLTGGTGADTFVFAARTDSYRTDAGVFNDLILDFDASEDRIDLSA 87
AlgE1R1         GDAGNDILDGGAGRDNLTGGAGADTFRFSARTDSYRTDSASFNDLITDFDADEDSIDLSA 87
AlgE3R1         GGAGDDTLDGGAGRDTLTGGAGADTFRFSAREDSHRTDSASFTDLITDFDASQDRIDLSA 87
AlgE1R3         GGAGDDILDGGAGRDTLTGGTGADTFLFSTRTDSYRTDSASFNDLITDFDPTQDRIDLSG 97
AlgE3R3         GGAGDDVLDGGAERDTLTGGTGADTFLFSARTDSYRTDSASFTDLITDFDPAQDRIDLSG 95
AlgE2R1         GGAGNDILDGGAGRDNLTGGAGADLFRVSARTDSYRTDSASFNDLITDFDASQDRIDLSA 87
AlgE5R1         GGAGNDILDGGAGRDNLTGGAGADLFRVSARTDSYRTDSASFNDLITDFDPAQDRIDLSA 87
AlgE3R4         GGAGNDILDGGVGRDTLTGGAGADTFRFSAREDSYRTASTSFTDLITDFDPAQDRIDLSA 87
                *  .* *  ****. **  *:** .** *  .: * **:::    * * * **:. :* :*:*.

AlgE6R1         LGFIGLGNGYAGTLAVSLSADGLRTYLKSYDADAQGRSFELALDGNHAATLSAGNIVFAA 147
ORF9R1          LGFTGLGNGYAGTLAVTTSADGSHTYLKSYEVDAQGRSFEISLQGNHAAALSAANIVFGA 180
ORF9R3          LGFTGLGNGYAGTLAVSTNAEGTRTYLKSYEADAQGRSFELALDGNHSATLSAANIVFAA 160
ORF9R7          LGFTGLGNGYAGTLAVSTNVEGTRTYLKSYDADAQGHSFELALDGNHSATLSASNIVFAA 160
ORF9R5          LGFTGLGNGYAGTLAVTTNLEGTRTYLKSYEADAEGRSFELALDGNHAATLSAANIVFGA 160
ORF9R2          LGFTGLGNGYAGTLAVTTSADGLRTYLKSYEADAEGRSFELALDGNHAATLSAANIVFGA 160
ORF9R6          LGFTGLGNGYAGTLAVTTSVDGLRTYLKSYEADAEGRSFELALDGNHAATLSASNIVFGA 160
AlgE6R3         LGFSGLGNGYAGSLAVSVSDDGTRTYLKSYEADAQGLSFEVALEGDHAAALSADNIVFAA 161
AlgE6R2         LGFTGLGNGYAGTLAVSVSDDGTRTYLKSYETDAEGRSFEVSLQGNHAAALSADNILFAT 159
ORF9R4          LGFTGLGNGYAGTLAVTTSVDGTHTYLKSYETDAEGRSFEISLQGNHAAALSADNILFGA 159
AlgE4R          LGFSGLGDGYGGTLLKTNAEGTRTYLKSFEADAEGRRFEVALDGDHTGDLSAANVVFAA 146
AlgE1R4         LGFSGFGDGYNGTLLVQLSSAGTRTYLKSYEEDLEGRRFEVALDGDHTGDLSAANVVFAD 147
AlgE1R1         LGFTGLGDGYNGTLLLKTNAEGTRTYLKSYEADAQGRRFEIALDGNFTGLFNDNNLLFDA 147
AlgE3R1         LGFTGLGNGYDGTLAVTTGSGGTRTYLKSYEVDAQGRRFEIALDGNFVGQFNDGNLLFDA 147
AlgE1R3         LGFSGFGNGYDGTLLLQVNAAGTRTYLKSFEADANGQRFEIALDGDFSGQLDSGNVIFEP 157
AlgE3R3         LGFSGFGNGYDGTLLLQVNAAGTRTYLKSLEADADGQRFEIALDGDFSGQLDSGNVIFEA 155
AlgE2R1         LGFTGLGDGYNGTLLLQVSADGSRTYLKSLEADAEGRRFEIALDGNFAGLLGAGNLLFER 147
AlgE5R1         LGFTGLGDGYNGTLAVVLNSAGTRTYLKSYEADAEGRRFEIALDGNFAGLLDDGNLIFER 147
AlgE3R4         LGFTGLGDGYDGTLLVTTGSGGSRTYLKSLEADAEGRRFEIALDGDFVGLLDASNLIFER 147
                ***  *:*:**  *:*  :    .    *  :****  *  :  *  :*   **::*:*:.  .  :.   *::*

AlgE6R1         AT---------------------------- 149
ORF9R1          A----------------------------- 181
ORF9R3          AT---------------------------- 162
ORF9R7          AAPVAT-ELEVIGASSLPEDQIV------- 182
ORF9R5          A----------------------------- 161
ORF9R2          A----------------------------- 161
ORF9R6          A----------------------------- 161
AlgE6R3         TDAAAAGELGVIGASGQPDDPAV------- 184
AlgE6R2         ------------------------------
ORF9R4          A----------------------------- 160
AlgE4R          TG---------TTTELEVLGDSGTQAGAIV 167
AlgE1R4         DGSAAVASSDPAATQLEVVGSSGTQTDQLA 177
AlgE1R1         AP---------------AT----------- 151
AlgE3R1         A----------------------------- 148
AlgE1R3         A----------------------------- 158
AlgE3R3         ------------------------------
AlgE2R1         TA---------------IEGDA-------- 154
AlgE5R1         ------------------------------
AlgE3R4         PA---------------IEGDA-------- 154
```

Fig. 101: Alignment of the R-modules of group I and group II. Asp and Glu residues labelled in green point into the β-roll. The dark blue labelled amino acids are conserved in all R-modules and their side chains are on the binding surface. Amino acids that are labelled in light blue are similar in all R-modules and their side chains point into the solution. Gray labelled amino acids were affected by binding of alginate to AlgE4R. The rectangles indicate the areas that are at the "front" side of the R-modules.

In contrast, the binding constants of single R-modules of AlgE6 with any alginate were not determinable. In some cases, small energy changes were observed in the ITC thermograph. One example was the titration of AlgE6R1 with M6 or MG6 respectively. The energy release or uptake of AlgE6R1 with M5 is too low to be detected correctly, while energy release could be measured using MG5 (Fig. 102). The thermographs of the binding of M8 and MG8 alginates show a small energy uptake. However, the energy release or uptake is too small to quantify that. It seems that MG-blocks are bound better than M-alginates.

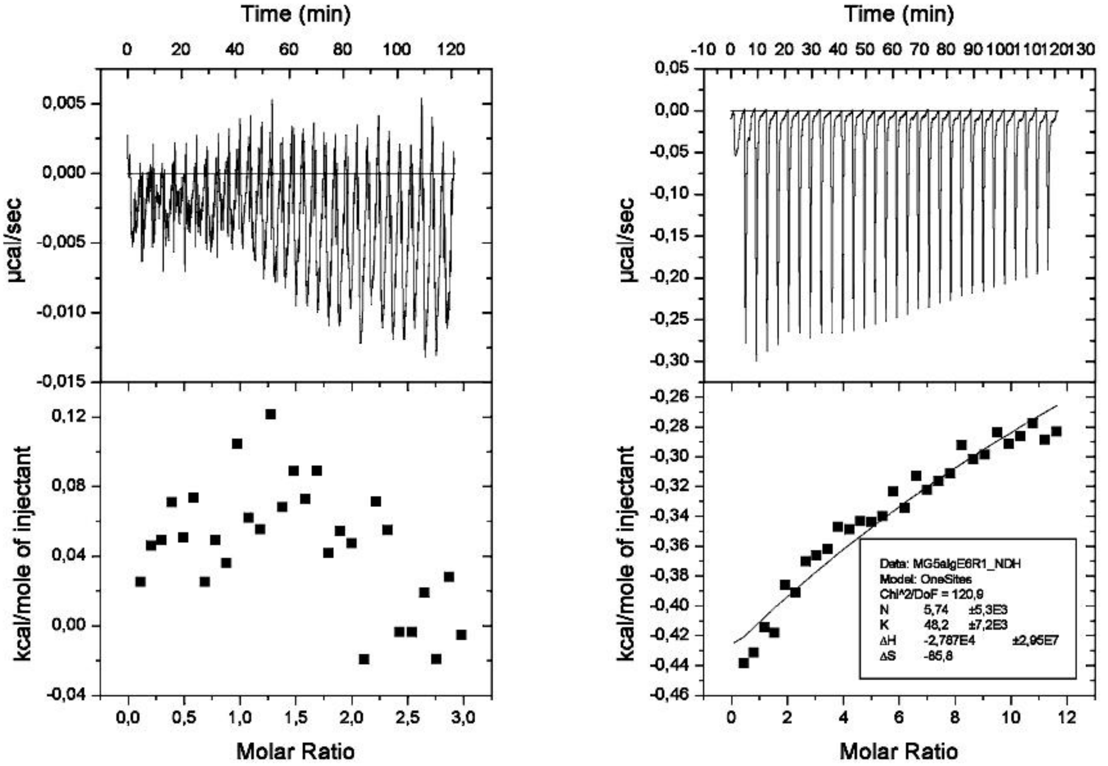


Fig. 102: ITC thermographs of AlgE6R1 with M5-alginate (left) and MG-5 alginate (right). Titration of penta mannuronic acid does not show any energy changes while titration with an alternating alginate pentamer shows a small energy release. Nevertheless, the energy release was too small to determine enthalpy or binding constant correctly. The ITC data indicate that MG-alginates are better bound than poly-M alginates.

All R-modules of AlgE6 bind short-chain alginate oligomers at least 10,000 times weaker than the R-module of AlgE4.

Given the high sequence similarity, this is interesting: first it was assumed that the one arginine (R124 in AlgE4R) is essential for binding. This arginine is present in every R-module of group I but the R-modules of AlgE6 have serines on that position (Fig. 101).

Interaction measurements between ORF9 and alginates showed that ORF9 does bind alginates [241]. The R-modules of ORF9 are most similar to the R-modules of AlgE6 and all have a serine where AlgE4R has R124. This suggests that the arginine is not essential for binding.

The results obtained from alginate titrations of ORF-9 are described in more detail in paper V. Overall, ORF-9 shows a similar trend as AlgE4R with increased binding strength with increasing degree of polymerisation (Tab. 4) and affinities strongest for poly-M alginate, weaker for MG-block alginate and none or very weak for GG-block alginate. Direct comparison of the results is, however, difficult, as only full-length ORF-9, and not single R-modules, was investigated. Further, the alginate oligomers investigated, had a higher degree of polymerisation, and some titrations had to be conducted in the absence of Ca^{2+} which also influences the results.

Tab. 4: Comparison of the binding constants of poly-M alginates with ORF9 or AlgE4R respectively. Using long alginates resulted in higher K_a-values. The binding constants of the tested M-alginates to ORF9 are similar to the binding constant of M4 to AlgE4R.

ORF9		AlgE4R	
Alginate	$K_a \cdot 10^4$ [M^{-1}]	Alginate	$K_a \cdot 10^4$ [M^{-1}]
M12	1.5 ± 0.3	M3	0.31 ± 0.003
M13-15	1.8 ± 0.1	M4	3.85 ± 0.3
M18-20	7.2 ± 0.3	M5	26.7 ± 0.45

We was not able to construct an R1-R2-R3 of AlgE6 that could be expressed yet which will give us the opportunity to investigate whether the R-modules of AlgE6, in all three together, bind alginate more strongly.

4 Future studies

4.1 Ligation studies

The ultimate goal of any ligation studies is to obtain a ligated protein that does not differ from the wild type. More than 20 split inteins were discovered but few of them were tested for their splicing ability *in vitro*. Soon more of the natural split intein should be tested.

The amino acids around the splicing site have a great impact on the ligation yield. The first amino acid of the C-terminal extein has to be a cysteine, serine or threonine. The following 3 amino acids of the C-terminal extein and the last two amino acids of the N-terminal extein influence the ligation yield. Nevertheless those amino acids are not conserved therefore it should be possible to obtain a library of split inteins with different amino acids combinations around the splicing site.

Oxidation of the essential cysteins inhibits the ligation completely and reduction often cause high amount of cleavage. It seems best to use a non nucleophilic reducing agent to avoid cleavage. Another option is to use inteins which have serines or threonines at the splicing sites.

The yield of ligated AlgE4 using TCEP was over 80% nevertheless the exchange of some amino acids around the splicing site causes severe decomposition of the ligated product. For successful use of segmentally labelled AlgE4 a combination of amino acids around the splicing site must be found that results in high amount of ligation yield but does not show post-ligational cleavage. From the linker reaction between the A- and R-module GEPGATPQQPST the red labelled amino acids were exchange to KCFNG. This exchange results in a completely different linker which can affect the orientation and flexibility between the modules and it should be tried to obtain a more similar linker by exchanging GATPQQ to GACFQQ.

In vivo ligation to AlgE4 can only be successful after the fragment Int_C-R is transferred into a new vector. This vector must have a different induction system than used for the T7 promoter and expression at low temperature (16-20° C) is possible. Additionally it must be compatible with the vector containing A-Int_N and glucose should not affect expression. One possible candidate is the *Pm*-promoter. The amount of expressed protein can be controlled either by number of copies and/or by the concentration of induction agent which is benzoic or toluic acid [246,247].

If one succeeds in producing ligated AlgE4 in acceptable quantities for NMR, one could perform alginate binding test on the segmentally labelled AlgE4. This will show which effect the A-module has on the binding of the R-module to alginate and vice versa. The segmentally labelled AlgE4 is an active epimerase therefore measurements have to be done at 4°C to reduce the activity [177]. The orientation of the module to each other should be determined in presence of alginate by RDC. The structure of AlgE4 shows a fixed angle between the A- and R-module in absence of alginate. It is not certain it this orientation remains when alginate oligomers are binding. Moreover the complex of AlgE4 with alginate should be determined. It would be the first alginate-epimerase complex.

4.2 Complex of AlgE4R and M5-alginate

The alginate binding studies to AlgE4R are complete now but a complex structure of the R-module with poly-M alginate is still missing. Intermolecular NOE between ^{13}C-labelled R-module and unlabelled alginate pentamer should reveal the orientation of the alginate on the surface of AlgE4R and the participating amino acids. Another approach is the use of paramagnetic compounds. Addition of paramagnetic compounds to the solution should reveal the area where alginate is binding to the R-module and limits the possible interacting amino acids. The third possibility is to use selective labelled amino acids. From the NMR binding studies it is known that most of the amino acids involved in binding are charged amino acids. By incorporating labelled arginine, (lysine, aspartic and glutamic acid) into an unlabelled structure it would be possible to obtain binding constants from the side chain where binding occurs. As only few amino acids are labelled solving the complex structure could be easier as less signal overlap can occur.

4.3 AlgE6

The binding constant of the single R-modules of the AlgE6 could not be determined. As the next step, a construct consisting of all three R-modules (AlgE6R123) should be tested by NMR, ITC and SAXS. Binding should be verified by NMR and ITC. If alginates are binding to AlgE6R123 then further investigation like determining the binding site and the involved amino acids will be carried out. The structure of AlgE6R123 should be additional determined by SAXS independent of the binding ability of this construct.

4.4 ORF9

The binding studies with the R-modules of ORF9 should be finished. To have a better comparison of the binding constants between the different types of alginate all experiment should be performed without $CaCl_2$ in the buffer. The length of the alginate oligomers tested until now can maximally cover three out of the seven R-modules. It would be great to test even longer alginate oligomers, especially MG- and G-block rich alginates. But G-rich alginates tend to be highly viscous and prone to gel formation.

The overall structure of ORF9 should be also determined by SAXS. The R-modules of AlgE6 are not linearly orientated but bend nearly 135°. As the R-modules of ORF9 and AlgE6 are very similar to each other and consist of seven R-modules it could be possible that the R-modules of ORF9 have even more bends.

4.5 Hybrid epimerases

Hybrid epimerases between the AlgE4 and AlgE6 should determine which combination of amino acids change the MG-epimerising A-module into a GG-block producing one. These two A-modules are closely related and out of the 373 amino acid only 67 are different (excluding the linker). Thereby, the effect of the R-module on the A-module should be addressed. It is known that the R-modules enhance the reaction rate but it is not known if the R-modules also influence the epimerisation pattern. Hybrid epimerases consisting of the A-modules of AlgE4 with different R-modules and the A-module of AlgE6 with e.g. AlgE4R will shed some light on the influence that the R-module has on the epimerization pattern.

Another study object should be AlgE1. AlgE1 consists of two active alginate epimerases in one molecule. The first part (A_1-R_1-R_2-R_3) of AlgE1 is a GG-block epimerase while the second epimerase (A_2-R_4) converts alginate to MG-blocks. First the position of the two epimerases should be exchanged to investigate the effect on epimerisation. Further a library of hybrid epimerases between the single epimerases should be constructed. The results of this study should be compared with AlgE3. AlgE3 is the biggest epimerase has consists of 2A-modules and 7 R-modules. It contains the only MG-forming epimerase that has more than one R-module (A_2-R_4-R_5-R_6-R_7). It is also the only epimerase whose individual inactivation has an effect on the cyst formation *in vivo* [170]. The strain carrying the inactivation of *algE3* produced alginates with low G-content.

4.6 A-modules

Also the A- modules should be investigated further. The binding specificity of the A-module should also be tested. The A-modules are active and they are known to bind alginates nevertheless the binding constants and binding specificities were never measured. The epimerisation rate of the single A-module is relative low [172] and low temperature (4-15 °C) reduces the epimerisation further [177]. It should be possible to measure binding constants without significant epimerisation. First the A-module of AlgE4 and AlgE6 will be tested by ITC but finally all 10 A-modules should be tested. Additionally the A-module of AlgE4 should be assigned and the binding constant and the affect amino acids will be obtained from the segmentally labelled AlgE4.

4.7 R-modules

AlgE4R does bind alginates well while none of the single R-modules of AlgE6 can bind alginates. To test the importance of R124 (in AlgE4R) an AlgE4R mutant should be constructed where this arginine is exchanged to serine (R124S) plus the reverse AlgE6 mutants were the serine is change to arginine (S126R in AlgE6R1, S135R in AlgE6R2 and AlgE6R3). Additionally, hybrid R-modules between AlgE4R and AlgE6R1 should be investigated to identify the important amino acids for binding. Eventually, more R-modules should be tested. The next group which should be investigated are the R-modules of AlgE3 and mainly there the R-modules of the MG-epimerase. It is the only MG-epimerase which has more than one R-module and the R-modules of AlgE3-2 are similar to the R-modules of AlgE2 and AlgE5. The binding ability of single R-modules as well as the construct of the quadruple R-modules should be tested.

4.8 Orientation

The structure of AlgE6 and AlgE4 obtained by SAXS was the first structures of a full-length alginate epimerases. The SAXS data showed that both epimerases are bent and that there is little flexibility between the models. But AlgE4 and AlgE6 have proline rich linkers between the modules which are not conserved for all the epimerases. Therefore, it is not sure if all the alginate epimerases have a rigid conformation. AlgE1 and AlgE3 consist of two epimerases and the orientation and flexibility between these epimerases should also be investigated. Therefore at least AlgE4, ORF9, AlgE1 and AlgE3 should also be measured by SAXS.

4.9 Mode of action

The processive mode of action of AlgE4 should be further investigated. It could be shown that AlgE4 slides along the polyM-alginates for approximate 8 epimerisation before it dissociates [187]. The velocity of which the epimerase moves along the alginate chain should be investigated by using alginates which reducing end is labelled with a paramagnetic complex. Also the kinetic and moving direction of the single modules should be investigated.

For AlgE6 the mode of action will be determined. It is assumed to have also a processive mode of action however it was never confirmed. Epimerisation from poly-M alginates to GG-blocks is very complex. A processive mode of action would only be possible if one AlgE6 protein epimerize poly-M alginates to MG-blocks and a second AlgE6 convert then the MG-blocks to GG-blocks. This is only possible if gelling of the MG-blocks can be prevented.

5 References

1. Dawson PE, Muir TW, Clark-Lewis I, Kent SB (1994) Synthesis of proteins by native chemical ligation. Science (New York, NY 266: 776-779.
2. Kochendoerfer GG, Chen SY, Mao F, Cressman S, Traviglia S, Shao H, Hunter CL, Low DW, Cagle EN, Carnevali M, Gueriguian V, Keogh PJ, Porter H, Stratton SM, Wiedeke MC, Wilken J, Tang J, Levy JJ, Miranda LP, Crnogorac MM, Kalbag S, Botti P, Schindler-Horvat J, Savatski L, Adamson JW, Kung A, Kent SB, Bradburne JA (2003) Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. Science (New York, NY 299: 884-887.
3. Schnölzer M, Kent SB (1992) Constructing proteins by dovetailing unprotected synthetic peptides: backbone-engineered HIV protease. Science (New York, NY 256: 221-225.
4. Englebretsen DR, Garnham BG, Bergman DA, Alewood PF (1995) A Novel Thioether Linker: Chemical Synthesis of a HIV-1 Protease Analogue by Thioether Ligation. Tetrahedron Letters 36: 8871-8874.
5. Robey FA, Fields RL (1989) Automated synthesis of N-bromoacetyl-modified peptides for the preparation of synthetic peptide polymers, peptide-protein conjugates, and cyclic peptides. Analytical biochemistry 177: 373-377.
6. Gaertner HF, Offord RE, Cotton R, Timms D, Camble R, Rose K (1994) Chemo-enzymic backbone engineering of proteins. Site-specific incorporation of synthetic peptides that mimic the 64-74 disulfide loop of granulocyte colony-stimulating factor. The Journal of biological chemistry 269: 7224-7230.
7. Gaertner HF, Rose K, Cotton R, Timms D, Camble R, Offord RE (1992) Construction of protein analogues by site-specific condensation of unprotected fragments. Bioconjugate chemistry 3: 262-268.
8. Liu CF, Tam JP (1994) Peptide segment ligation strategy without use of protecting groups. Proceedings of the National Academy of Sciences of the United States of America 91: 6584-6588.
9. Tam JP, Rao C, Liu CF, Shao J (1995) Specificity and formation of unusual amino acids of an amide ligation strategy for unprotected peptides. International journal of peptide and protein research 45: 209-216.
10. Tornoe CW, Christensen C, Meldal M (2002) Peptidotriazoles on solid phase: [1,2,3]-triazoles by regiospecific copper(i)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. The Journal of organic chemistry 67: 3057-3064.
11. Agard NJ, Prescher JA, Bertozzi CR (2004) A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. Journal of the American Chemical Society 126: 15046-15047.
12. Wieland T, Bokelmann E, Bauer L, Lang HU, Lau H (1953) Über Peptidsynthesen. 8. Mitteilung Bildung von S-haltigen Peptiden durch intramolekulare Wanderung von Aminoacylresten. Libigs Ann Chem 583: 129–149.
13. Hackeng TM, Griffin JH, Dawson PE (1999) Protein synthesis by native chemical ligation: expanded scope by using straightforward methodology. Proceedings of the National Academy of Sciences of the United States of America 96: 10068-10073.
14. Tam JP, Xu J, Eom KD (2001) Methods and strategies of peptide ligation. Biopolymers 60: 194-205.
15. Gieselman MD, Zhu Y, Zhou H, Galonic D, van der Donk WA (2002) Selenocysteine derivatives for chemoselective ligations. Chembiochem 3: 709-716.
16. Gieselman MD, Xie L, van Der Donk WA (2001) Synthesis of a selenocysteine-containing peptide by native chemical ligation. Organic letters 3: 1331-1334.

17. Tam JP, Yu Q (1998) Methionine ligation strategy in the biomimetic synthesis of parathyroid hormones. Biopolymers 46: 319-327.

18. Clayton D, Shapovalov G, Maurer JA, Dougherty DA, Lester HA, Kochendoerfer GG (2004) Total chemical synthesis and electrophysiological characterization of mechanosensitive channels from *Escherichia coli* and *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences of the United States of America 101: 4764-4769.

19. Smith HB, Hartman FC (1988) Restoration of activity to catalytically deficient mutants of ribulosebisphosphate carboxylase/oxygenase by aminoethylation. The Journal of biological chemistry 263: 4921-4925.

20. Okamoto R, Kajihara Y (2008) Uncovering a latent ligation site for glycopeptide synthesis. Angewandte Chemie (International ed 47: 5402-5406.

21. Bernardes GaJL, Chalker JM, Errey JC, Davis BG (2008) Facile Conversion of Cysteine and Alkyl Cysteines to Dehydroalanine on Protein Surfaces: Versatile and Switchable Access to Functionalized Proteins. Journal of the American Chemical Society 130: 5052-5053.

22. Guo J, Wang J, Lee JS, Schultz PG (2008) Site-Specific Incorporation of Methyl- and Acetyl-Lysine Analogues into Recombinant Proteins. Angewandte Chemie 120: 6499-6501.

23. Wang J, Schiller SM, Schultz PG (2007) A Biosynthetic Route to Dehydroalanine-Containing Proteins. Angewandte Chemie 119: 6973-6975.

24. Macmillan D (2006) Strategien zur Proteinsynthese vereinigen sich mit der nativen chemischen Ligation. Angewandte Chemie 118: 7830-7834.

25. Hackenberger CPR, Schwarzer D (2008) Chemoselective Ligation and Modification Strategies for Peptides and Proteins. Angewandte Chemie International Edition 47: 10030-10074.

26. Bark SJ, Kent SBH (1999) Engineering an unnatural N[α]-anchored disulfide into BPTI by total chemical synthesis: structural and functional consequences. FEBS letters 460: 67-76.

27. Shao Y, Lu W, Kent SBH (1998) A novel method to synthesize cyclic peptides. Tetrahedron Letters 39: 3911-3914.

28. Canne LE, Bark SJ, Kent SBH (1996) Extending the Applicability of Native Chemical Ligation. Journal of the American Chemical Society 118: 5891-5896.

29. Meutermans WDF, Golding SW, Bourne GT, Miranda LP, Dooley MJ, Alewood PF, Smythe ML (1999) Synthesis of Difficult Cyclic Peptides by Inclusion of a Novel Photolabile Auxiliary in a Ring Contraction Strategy. Journal of the American Chemical Society 121: 9790-9796.

30. Botti P, Carrasco MR, Kent SBH (2001) Native chemical ligation using removable N[α]-(1-phenyl-2-mercaptoethyl) auxiliaries. Tetrahedron Letters 42: 1831-1833.

31. Marinzi C, Offer J, Longhi R, Dawson PE (2004) An o-nitrobenzyl scaffold for peptide ligation: synthesis and applications. Bioorganic & Medicinal Chemistry 12: 2749-2757.

32. Offer J, Boddy CNC, Dawson PE (2002) Extending Synthetic Access to Proteins with a Removable Acyl Transfer Auxiliary. Journal of the American Chemical Society 124: 4642-4646.

33. Saxon E, Bertozzi CR (2000) Cell surface engineering by a modified Staudinger reaction. Science (New York, NY 287: 2007-2010.

34. Nilsson BL, Kiessling LL, Raines RT (2000) Staudinger ligation: a peptide from a thioester and azide. Organic letters 2: 1939-1941.

35. Saxon E, Armstrong JI, Bertozzi CR (2000) A "traceless" Staudinger ligation for the chemoselective synthesis of amide bonds. Organic letters 2: 2141-2143.

36. Soellner MB, Tam A, Raines RT (2006) Staudinger ligation of peptides at non-glycyl residues. The Journal of organic chemistry 71: 9824-9830.
37. Nilsson BL, Kiessling LL, Raines RT (2001) High-yielding Staudinger ligation of a phosphinothioester and azide to form a peptide. Organic letters 3: 9-12.
38. Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y (1990) Molecular structure of a gene, VMA1, encoding the catalytic subunit of H^+-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. The Journal of biological chemistry 265: 6726-6733.
39. Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H^+-adenosine triphosphatase. Science (New York, NY 250: 651-657.
40. Davis EO, Jenner PJ, Brooks PC, Colston MJ, Sedgwick SG (1992) Protein splicing in the maturation of M. tuberculosis recA protein: a mechanism for tolerating a novel class of intervening sequence. Cell 71: 201-210.
41. Cooper AA, Chen YJ, Lindorfer MA, Stevens TH (1993) Protein splicing of the yeast TFP1 intervening protein sequence: a model for self-excision. The EMBO journal 12: 2575-2583.
42. Kawasaki M, Satow Y, Ohya Y, Anraku Y (1997) Protein splicing in the yeast Vma1 protozyme: evidence for an intramolecular reaction. FEBS letters 412: 518-520.
43. Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J, Belfort M (1994) Protein splicing elements: inteins and exteins--a definition of terms and recommended nomenclature. Nucleic acids research 22: 1125-1127.
44. Perler FB, Olsen GJ, Adam E (1997) Compilation and analysis of intein sequences. Nucleic acids research 25: 1087-1093.
45. Gimble FS, Thorner J (1992) Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. Nature 357: 301-306.
46. Gimble FS, Thorner J (1993) Purification and characterization of VDE, a site-specific endonuclease from the yeast *Saccharomyces cerevisiae*. The Journal of biological chemistry 268: 21844-21853.
47. Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS (1997) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. Nucleic acids research 25: 4626-4638.
48. Pietrokovski S (1998) Modular organization of inteins and C-terminal autocatalytic domains. Protein Sci 7: 64-71.
49. Pietrokovski S (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. Protein Sci 3: 2340-2350.
50. Duan X, Gimble FS, Quiocho FA (1997) Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. Cell 89: 555-564.
51. Telenti A, Southworth M, Alcaide F, Daugelat S, Jacobs WR, Jr., Perler FB (1997) The Mycobacterium xenopi GyrA protein splicing element: characterization of a minimal intein. Journal of bacteriology 179: 6378-6382.
52. Wu H, Hu Z, Liu XQ (1998) Protein *trans*-splicing by a split intein encoded in a split DnaE gene of *Synechocystis sp.* PCC6803. Proceedings of the National Academy of Sciences of the United States of America 95: 9226-9231.
53. Hodges RA, Perler FB, Noren CJ, Jack WE (1992) Protein splicing removes intervening sequences in an archaea DNA polymerase. Nucleic acids research 20: 6153-6157.
54. Chong S, Xu MQ (1997) Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. The Journal of biological chemistry 272: 15587-15590.

55. Shingledecker K, Jiang SQ, Paulus H (1998) Molecular dissection of the *Mycobacterium tuberculosis* RecA intein: design of a minimal intein and of a *trans*-splicing system involving two intein fragments. Gene 207: 187-195.
56. Lew BM, Mills KV, Paulus H (1998) Protein splicing *in vitro* with a semisynthetic two-component minimal intein. The Journal of biological chemistry 273: 15887-15890.
57. Mills KV, Lew BM, Jiang S, Paulus H (1998) Protein splicing in trans by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein. Proceedings of the National Academy of Sciences of the United States of America 95: 3543-3548.
58. Wu H, Xu MQ, Liu XQ (1998) Protein *trans*-splicing and functional mini-inteins of a cyanobacterial dnaB intein. Biochimica et biophysica acta 1387: 422-432.
59. Kawasaki M, Makino S, Matsuzawa H, Satow Y, Ohya Y, Anraku Y (1996) Folding-dependent in vitro protein splicing of the *Saccharomyces cerevisiae* VMA1 protozyme. Biochemical and biophysical research communications 222: 827-832.
60. Southworth MW, Adam E, Panne D, Byer R, Kautz R, Perler FB (1998) Control of protein splicing by intein fragment reassembly. The EMBO journal 17: 918-926.
61. Perler FB (2002) InBase: the Intein Database. Nucleic acids research 30: 383-384.
62. Ogata H, Raoult D, Claverie JM (2005) A new example of viral intein in Mimivirus. Virology journal 2: 8.
63. Pietrokovski S (1998) Identification of a virus intein and a possible variation in the protein-splicing reaction. Current Biology 8: R634-R638.
64. Perler FB, Xu MQ, Paulus H (1997) Protein splicing and autoproteolysis mechanisms. Current opinion in chemical biology 1: 292-299.
65. Xu MQ, Comb DG, Paulus H, Noren CJ, Shao Y, Perler FB (1994) Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. The EMBO journal 13: 5517-5522.
66. Wallace CJ (1993) The curious case of protein splicing: mechanistic insights suggested by protein semisynthesis. Protein Sci 2: 697-705.
67. Xu MQ, Perler FB (1996) The mechanism of protein splicing and its modulation by mutation. The EMBO journal 15: 5146-5153.
68. Chong S, Shao Y, Paulus H, Benner J, Perler FB, Xu MQ (1996) Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an *in vitro* splicing system. The Journal of biological chemistry 271: 22159-22168.
69. Shao Y, Xu MQ, Paulus H (1996) Protein splicing: evidence for an N-O acyl rearrangement as the initial step in the splicing process. Biochemistry 35: 3810-3815.
70. Xu MQ, Southworth MW, Mersha FB, Hornstra LJ, Perler FB (1993) *In vitro* protein splicing of purified precursor and the identification of a branched intermediate. Cell 75: 1371-1377.
71. Shao Y, Xu MQ, Paulus H (1995) Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. Biochemistry 34: 10844-10850.
72. Shao Y, Paulus H (1997) Protein splicing: estimation of the rate of O-N and S-N acyl rearrangements, the last step of the splicing process. J Pept Res 50: 193-198.
73. Dassa B, Amitai G, Caspi J, Schueler-Furman O, Pietrokovski S (2007) Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. Biochemistry 46: 322-330.
74. Southworth MW, Benner J, Perler FB (2000) An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. The EMBO journal 19: 5019-5026.
75. Gorbalenya AE (1998) Non-canonical inteins. Nucleic acids research 26: 1741-1748.
76. Tori K, Dassa B, Johnson MA, Southworth MW, Brace LE, Ishino Y, Pietrokovski S, Perler FB (2010) Splicing of the mycobacteriophage *Bethlehem* DnaB intein:

identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. The Journal of biological chemistry 285: 2515-2526.

77. Mills KV, Paulus H (2001) Reversible inhibition of protein splicing by zinc ion. The Journal of biological chemistry 276: 10832-10838.
78. Ghosh I, Sun L, Xu MQ (2001) Zinc inhibition of protein trans-splicing and identification of regions essential for splicing and association of a split intein. The Journal of biological chemistry 276: 24051-24058.
79. Poland BW, Xu MQ, Quiocho FA (2000) Structural insights into the protein splicing mechanism of PI-SceI. The Journal of biological chemistry 275: 16408-16413.
80. Van Roey P, Pereira B, Li Z, Hiraga K, Belfort M, Derbyshire V (2007) Crystallographic and mutational studies of *Mycobacterium tuberculosis* recA mini-inteins suggest a pivotal role for a highly conserved aspartate residue. Journal of molecular biology 367: 162-173.
81. Sun P, Ye S, Ferrandon S, Evans TC, Xu MQ, Rao Z (2005) Crystal structures of an intein from the split dnaE gene of *Synechocystis sp.* PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. Journal of molecular biology 353: 1093-1105.
82. Severinov K, Muir TW (1998) Expressed protein ligation, a novel method for studying protein-protein interactions in transcription. The Journal of biological chemistry 273: 16205-16209.
83. Muir TW, Sondhi D, Cole PA (1998) Expressed protein ligation: a general method for protein engineering. Proceedings of the National Academy of Sciences of the United States of America 95: 6705-6710.
84. Chong S, Mersha FB, Comb DG, Scott ME, Landry D, Vence LM, Perler FB, Benner J, Kucera RB, Hirvonen CA, Pelletier JJ, Paulus H, Xu M-Q (1997) Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene 192: 271-281.
85. Chong S, Montello GE, Zhang A, Cantor EJ, Liao W, Xu MQ, Benner J (1998) Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. Nucleic acids research 26: 5109-5115.
86. Evans TC, Jr., Benner J, Xu MQ (1998) Semisynthesis of cytotoxic proteins using a modified protein splicing element. Protein Sci 7: 2256-2264.
87. Scott CP, Abel-Santos E, Wall M, Wahnon DC, Benkovic SJ (1999) Production of cyclic peptides and proteins *in vivo*. Proceedings of the National Academy of Sciences of the United States of America 96: 13638-13643.
88. Iwaï H, Pluckthun A (1999) Circular beta-lactamase: stability enhancement by cyclizing the backbone. FEBS letters 459: 166-172.
89. David R, Richter MP, Beck-Sickinger AG (2004) Expressed protein ligation. Method and applications. European journal of biochemistry / FEBS 271: 663-677.
90. Mootz HD, Blum ES, Muir TW (2004) Activation of an autoregulated protein kinase by conditional protein splicing. Angewandte Chemie (International ed 43: 5189-5192.
91. Mootz HD, Blum ES, Tyszkiewicz AB, Muir TW (2003) Conditional protein splicing: a new tool to control protein structure and function *in vitro* and *in vivo*. Journal of the American Chemical Society 125: 10561-10569.
92. Mootz HD, Muir TW (2002) Protein splicing triggered by a small molecule. Journal of the American Chemical Society 124: 9044-9045.
93. Zeidler MP, Tan C, Bellaiche Y, Cherry S, Hader S, Gayko U, Perrimon N (2004) Temperature-sensitive control of protein activity by conditionally splicing inteins. Nature biotechnology 22: 871-876.

94. Vila-Perello M, Hori Y, Ribo M, Muir TW (2008) Activation of protein splicing by protease- or light-triggered O to N acyl migration. Angewandte Chemie (International ed 47: 7764-7767.
95. Tyszkiewicz AB, Muir TW (2008) Activation of protein splicing with light in yeast. Nature methods 5: 303-305.
96. Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, Ito N, Ishino Y, Nakamura H (1998) Segmental Isotope Labeling for Protein NMR Using Peptide Splicing. Journal of the American Chemical Society 120: 5591-5592.
97. Xu R, Ayers B, Cowburn D, Muir TW (1999) Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. Proceedings of the National Academy of Sciences of the United States of America 96: 388-393.
98. Otomo T, Ito N, Kyogoku Y, Yamazaki T (1999) NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation. Biochemistry 38: 16040-16044.
99. Otomo T, Teruya K, Uegaki K, Yamazaki T, Kyogoku Y (1999) Improved segmental isotope labeling of proteins and application to a larger protein. Journal of biomolecular NMR 14: 105-114.
100. Ozawa T, Takeuchi TM, Kaihara A, Sato M, Umezawa Y (2001) Protein splicing-based reconstitution of split green fluorescent protein for monitoring protein-protein interactions in bacteria: improved sensitivity and reduced screening time. Analytical chemistry 73: 5866-5874.
101. Ozawa T, Kaihara A, Sato M, Tachihara K, Umezawa Y (2001) Split luciferase as an optical probe for detecting protein-protein interactions in mammalian cells based on protein splicing. Analytical chemistry 73: 2516-2521.
102. Homandberg GA, Laskowski M, Jr. (1979) Enzymatic resynthesis of the hydrolyzed peptide bond(s) in ribonuclease S. Biochemistry 18: 586-592.
103. Homandberg GA, Mattis JA, Laskowski M, Jr. (1978) Synthesis of peptide bonds by proteinases. Addition of organic cosolvents shifts peptide bond equilibria toward synthesis. Biochemistry 17: 5220-5227.
104. Homandberg GA, Komoriya A, Chaiken IM (1982) Enzymatic condensation of nonassociated peptide fragments using a molecular trap. Biochemistry 21: 3385-3389.
105. Chang TK, Jackson DY, Burnier JP, Wells JA (1994) Subtiligase: a tool for semisynthesis of proteins. Proceedings of the National Academy of Sciences of the United States of America 91: 12544-12548.
106. Jackson DY, Burnier J, Quan C, Stanley M, Tom J, Wells JA (1994) A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. Science (New York, NY 266: 243-247.
107. Machova Z, von Eggelkraut-Gottanka R, Wehofsky N, Bordusa F, Beck-Sickinger AG (2003) Expressed enzymatic ligation for the semisynthesis of chemically modified proteins. Angewandte Chemie (International ed 42: 4916-4918.
108. Srinivasulu S, Acharya AS (2002) Product-conformation-driven ligation of peptides by V8 protease. Protein Sci 11: 1384-1392.
109. Mao H, Hart SA, Schink A, Pollok BA (2004) Sortase-mediated protein ligation: a new method for protein engineering. Journal of the American Chemical Society 126: 2670-2671.
110. Fischer FG, Dörfel H (1955) [Polyuronic acids in brown algae.]. Hoppe-Seyler's Zeitschrift fur physiologische Chemie 302: 186-203.
111. Hirst EL, Jones JKN, Jones WO (1939) Structure of Alginic Acid. J Chem Soc: 1880.
112. Nelson WL, Cretcher LH (1929) The Alginic Acid from *Macrocystis Pyrifera*. Journal of the American Chemical Society 51: 1914-1918.

113. Haug A, Larsen B, Smidsrød O (1963) The degradation of alginates at different pH values. Acta chemica Scandinavica 17: 1466 – 1468.

114. Rehm BHA, Donati I, Paoletti S (2009) Material Properties of Alginates. Alginates: Biology and Applications: Springer Berlin / Heidelberg. pp. 1-53.

115. Pindar DF, Bucke C (1975) The biosynthesis of alginic acid by *Azotobacter vinelandii*. The Biochemical journal 152: 617-622.

116. Penman A, Sanderson GR (1972) A method for the determination of uronic acid sequence in alginates. Carbohydrate research 25: 273-282.

117. Atkins EDT, Mackie W, E. SE (1970) Cystralline Structures of Alginic Acid. Nature 225: 626-628.

118. Smidsrød O (1970) Solution Properties of Alginate. Carbohydrate research 13: 359-372.

119. Haug A, Larsen B, Smidsrød O (1966) A study of the Constitution of Alginic Acid by Partial Acid Hydrolysis. Acta chemica Scandinavica 20: 183-190.

120. Haug A, Larsen B, Smidsrød O (1967) Studies on the Sequence of Uronic Acid Residues in Alginic Acid. Acta chemica Scandinavica 21: 691-704.

121. Smidsrød O (1973) The relative extension of alginates having different chemical composition. Carbohydrate research 27: 107-118.

122. Atkins EDT, Nieduszynski IA, Mackie W, Parker KD, Smolko EE (1973) Structural components of alginic acid. I. The crystalline structure of poly-β-D-mannuronic acid. Results of X-ray diffraction and polarized infrared studies. Biopolymers 12: 1865-1878.

123. Atkins EDT, Nieduszynski IA, Mackie W, Parker KD, Smolko EE (1973) Structural components of alginic acid. II. The crystalline structure of poly-α-L-guluronic acid. Results of X-ray diffraction and polarized infrared studies. Biopolymers 12: 1879-1887.

124. Aachmann FL (2005) Alginate epimerases. PhD thesis.

125. Haug A, Myklestad S, Larsen B, Smidsrød O (1967) Correlation between Chemical Structure and Physical Properties of Alginates. Acta chemica Scandinavica 21: 768-778.

126. Smidsrød O, Haug A (1972) Dependence upon the gel-sol state of the ion-exchange properties of alginates. Acta chemica Scandinavica 26: 2063-2074.

127. Donati I, Holtan S, Mørch YA, Borgogna M, Dentini M, Skjåk-Bræk G (2005) New hypothesis on the role of alternating sequences in calcium-alginate gels. Biomacromolecules 6: 1031-1040.

128. Emmerichs N, Wingender J, Flemming HC, Mayer C (2004) Interaction between alginates and manganese cations: identification of preferred cation binding sites. International journal of biological macromolecules 34: 73-79.

129. Grant GT, Morris ER, Rees DA, Smith PJC, Thom D (1973) Biological Intercations between polysaccharides and divalent Cations: The Egg-box-Model. FEBS letters 32: 195-198.

130. Donati I, Mørch YA, Strand BL, Skjåk-Bræk G, Paoletti S (2009) Effect of elongation of alternating sequences on swelling behavior and large deformation properties of natural alginate gels. The journal of physical chemistry 113: 12916-12922.

131. Draget KI, Smidsrød O, Skjåk-Bræk G (2005) Alginates from Algae: Wiley-VCH Verlag GmbH & Co. KGaA.

132. Skjåk-Bræk G, Zanetti F, Paoletti S (1989) Effect of acetylation on some solution and gelling properties of alginates. Carbohydrate research 185: 131-138.

133. Franklin MJ, Ohman DE (2002) Mutant analysis and cellular localization of the AlgI, AlgJ, and AlgF proteins required for O acetylation of alginate in *Pseudomonas aeruginosa*. Journal of bacteriology 184: 3000-3007.

134. Franklin MJ, Ohman DE (1996) Identification of algI and algJ in the *Pseudomonas aeruginosa* alginate biosynthetic gene cluster which are required for alginate O acetylation. Journal of bacteriology 178: 2186-2195.

135. Franklin MJ, Ohman DE (1993) Identification of algF in the alginate biosynthetic gene cluster of *Pseudomonas aeruginosa* which is required for alginate acetylation. Journal of bacteriology 175: 5057-5065.

136. Skjåk-Bræk G, Grasdalen H, Larsen B (1986) Monomer sequence and acetylation pattern in some bacterial alginates. Carbohydrate research 154: 239-250.

137. Tielen P, Strathmann M, Jaeger K-E, Flemming H-C, Wingender J (2005) Alginate acetylation influences initial surface colonization by mucoid *Pseudomonas aeruginosa*. Microbiological Research 160: 165-176.

138. Nivens DE, Ohman DE, Williams J, Franklin MJ (2001) Role of Alginate and Its O Acetylation in Formation of *Pseudomonas aeruginosa* Microcolonies and Biofilms. J Bacteriol 183: 1047-1057.

139. Pier GB, Coleman F, Grout M, Franklin M, Ohman DE (2001) Role of Alginate O Acetylation in Resistance of Mucoid *Pseudomonas aeruginosa* to Opsonic Phagocytosis. Infect Immun 69: 1895-1901.

140. Vazquez A, Moreno S, Guzmán J, Alvarado A, Espín G (1999) Transcriptional organization of the *Azotobacter vinelandii* algGXLVIFA genes: characterization of algF mutants. Gene 232: 217-222.

141. Skjåk-Bræk G, Larsen B, Grasdalen H (1985) The role of *O*-acetyl groups in the biosynthesis of alginate by *Azotobacter vinelandii*. Carbohydrate research 145: 169-174.

142. Linker A, Jones RS (1966) A new polysaccharide resembling alginic acid isolated from *Pseudomonas*. The Journal of biological chemistry 241: 3845-3851.

143. Stanford ECC (1881) Brit Patent 142.

144. Gorin PAJ, Spencer JFT (1966) Exocellular alginic acid From *Azotobacter vinelandii*. Can J Chem 44: 993-998.

145. Haug A, Larsen B, Smidsrod O (1974) Uronic Acid Sequence in Alginate from different Sources. Carbohydrate research 32: 217-225.

146. Smidsrod O, Draget KI (1996) Alginates: Chemistry and physical Peoperties. Carbohydr Eur 14: 6-13.

147. May TB, Shinabarger D, Maharaj R, Kato J, Chu L, DeVault JD, Roychoudhury S, Zielinski NA, Berry A, Rothmel RK, et al. (1991) Alginate synthesis by *Pseudomonas aeruginosa*: a key pathogenic factor in chronic pulmonary infections of cystic fibrosis patients. Clinical microbiology reviews 4: 191-206.

148. Sadoff HL (1975) Encystment and germination in *Azotobacter vinelandii*. Bacteriological reviews 39: 516-539.

149. Ingar Draget K, Østgaard K, Smidsrød O (1990) Homogeneous alginate gels: A technical approach. Carbohydrate Polymers 14: 159-178.

150. Skjåk-Bræk G, Grasdalen H, Smidsrød O (1989) Inhomogeneous polysaccharide ionic gels. Carbohydrate Polymers 10: 31-54.

151. Thu B, Skjåk-Bræk G, Micali F, Vittur F, Rizzo R (1997) The spatial distribution of calcium in alginate gel beads analysed by synchrotron-radiation induced X-ray emission (SRIXE). Carbohydrate research 297: 101-105.

152. Grasdalen H, Larsen B, Smidsrød O (1979) A p.m.r. study of the composition and sequence of uronate residues in alginates. Carbohydrate research 68: 23-31.

153. Grasdalen H, Larsen B, Smidsrød O (1977) ^{13}C-N.M.R. studies of alginate. Carbohydrate research 56: C11-C15.

154. Grasdalen H, Larsen B, Smidsrod O (1981) ^{13}C-N.M.R. Studies of Monomeric Composition and Sequence in Alginate. Carbohydrate research 89: 179-191.

155. Donati I, Gamini A, Skjåk-Bræk G, Vetere A, Campa C, Coslovi A, Paoletti S (2003) Determination of the diadic composition of alginate by means of circular dichroism: a fast and accurate improved method. Carbohydrate research 338: 1139-1142.
156. Skjåk-Bræk G, Smidsrød O, Larsen B (1986) Tailoring of alginates by enzymatic modification in vitro. International journal of biological macromolecules 8: 330-336.
157. Glicksman M (1987) Utilization of seaweed hydrocolloids in the food industry. Hydrobiologia 151-152: 31-47.
158. Neetoo H, Ye M, Chen H (2010) Bioactive alginate coatings to control *Listeria monocytogenes* on cold-smoked salmon slices and fillets. International Journal of Food Microbiology 136: 326-331.
159. Draget KI, Taylor C (2011) Chemical, physical and biological properties of alginates and their biomedical implications. Food Hydrocolloids 25: 251-256.
160. Skjåk-Bræk G, Espevik T (1996) Application of alginate gels in biotechnology and biomedicineA. Carbohydr Eur 14: 19-25.
161. Soon-Shiong P, Feldman E, Nelson R, Heintz R, Yao Q, Yao Z, Zheng T, Merideth N, Skjåk-Bræk G, Espevik T, et al. (1993) Long-term reversal of diabetes by the injection of immunoprotected islets. Proceedings of the National Academy of Sciences of the United States of America 90: 5843-5847.
162. Strand BL, Gåserød O, Kulseng B, Espevik T, Skjåk-Bræk G (2002) Alginate-polylysine-alginate microcapsules: effect of size reduction on capsule properties. Journal of Microencapsulation 19: 615-630.
163. Lin TY, Hassid WZ (1966) Pathway of algnic acid synthesis in the marine brown alga, *Fucus gardneri Silva*. The Journal of biological chemistry 241: 5284-5297.
164. May TB, Chakrabarty AM (1994) *Pseudomonas aeruginosa*: genes and enzymes of alginate synthesis. Trends in microbiology 2: 151-157.
165. Rehm BH, Ertesvåg H, Valla S (1996) A new *Azotobacter vinelandii* mannuronan C-5-epimerase gene (algG) is part of an alg gene cluster physically organized in a manner similar to that in *Pseudomonas aeruginosa*. Journal of bacteriology 178: 5884-5889.
166. Valla S, Li J, Ertesvåg H, Barbeyron T, Lindahl U (2001) Hexuronyl C5-epimerases in alginate and glycosaminoglycan biosynthesis. Biochimie 83: 819-830.
167. Franklin MJ, Chitnis CE, Gacesa P, Sonesson A, White DC, Ohman DE (1994) *Pseudomonas aeruginosa* AlgG is a polymer level alginate C5-mannuronan epimerase. J Bacteriol 176: 1821-1830.
168. Ertesvåg H, Høidal HK, Hals IK, Rian A, Doseth B, Valla S (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in *Azotobacter vinelandii*. Molecular microbiology 16: 719-731.
169. Svanem BI, Skjåk-Bræk G, Ertesvåg H, Valla S (1999) Cloning and expression of three new *Aazotobacter vinelandii* genes closely related to a previously described gene family encoding mannuronan C-5-epimerases. Journal of bacteriology 181: 68-77.
170. Steigedal M, Sletta H, Moreno S, Mærk M, Christensen BE, Bjerkan T, Ellingsen TE, Espin G, Ertesvåg H, Valla S (2008) The *Azotobacter vinelandii* AlgE mannuronan C-5-epimerase family is essential for the *in vivo* control of alginate monomer composition and for functional cyst formation. Environmental microbiology 10: 1760-1770.
171. Gimmestad M, Steigedal M, Ertesvåg H, Moreno S, Christensen BE, Espin G, Valla S (2006) Identification and Characterization of an *Azotobacter vinelandii* Type I Secretion System Responsible for Export of the AlgE-Type Mannuronan C-5-Epimerases. J Bacteriol 188: 5551-5560.
172. Ertesvåg H, Valla S (1999) The A modules of the *Azotobacter vinelandii* mannuronan-C-5-epimerase AlgE1 are sufficient for both epimerization and binding of Ca^{2+}. Journal of bacteriology 181: 3033-3038.

173. Aachmann FL, Svanem BI, Güntert P, Petersen SB, Valla S, Wimmer R (2006) NMR structure of the R-module: a parallel β-roll subunit from an *Azotobacter vinelandii* mannuronan C-5 epimerase. The Journal of biological chemistry 281: 7350-7356.

174. Ertesvåg H, Høidal HK, Schjerven H, Svanem BI, Valla S (1999) Mannuronan C-5-epimerases and their application for *in vitro* and *in vivo* design of new alginates useful in biotechnology. Metabolic engineering 1: 262-269.

175. Ertesvåg H, Høidal HK, Skjåk-Bræk G, Valla S (1998) The *Azotobacter vinelandii* mannuronan C-5-epimerase AlgE1 consists of two separate catalytic domains. The Journal of biological chemistry 273: 30927-30932.

176. Holtan S, Bruheim P, Skjåk-Bræk G (2006) Mode of action and subsite studies of the guluronan block-forming mannuronan C-5 epimerases AlgE1 and AlgE6. The Biochemical journal 395: 319-329.

177. Høidal HK, Ertesvåg H, Skjåk-Bræk G, Stokke BT, Valla S (1999) The recombinant *Azotobacter vinelandii* mannuronan C-5-epimerase AlgE4 epimerizes alginate by a nonrandom attack mechanism. The Journal of biological chemistry 274: 12316-12322.

178. Svanem BI, Strand WI, Ertesvåg H, Skjåk-Bræk G, Hartmann M, Barbeyron T, Valla S (2001) The catalytic activities of the bifunctional *Azotobacter vinelandii* mannuronan C-5-epimerase and alginate lyase AlgE7 probably originate from the same active site in the enzyme. The Journal of biological chemistry 276: 31542-31550.

179. Rozeboom HJ, Bjerkan TM, Kalk KH, Ertesvåg H, Holtan S, Aachmann FL, Valla S, Dijkstra BW (2008) Structural and mutational characterization of the catalytic A-module of the mannuronan C-5-epimerase AlgE4 from *Azotobacter vinelandii*. The Journal of biological chemistry 283: 23819-23828.

180. Gasesa P (1987) Alginate-modifying enzymes; A proposed unified mechanism of action for the lyases and epimerases. FEBS letters 212: 199-202.

181. Skjåk-Bræk G, Larsen B (1982) A new assay for mannuronan C-5-epimerase activity. Carbohydrate research 103: 133-136.

182. Ertesvåg H, Doseth B, Larsen B, Skjåk-Bræk G, Valla S (1994) Cloning and expression of an *Azotobacter vinelandii* mannuronan C-5-epimerase gene. Journal of bacteriology 176: 2846-2853.

183. Chenal A, Guijarro JIa, Raynal B, Delepierre M, Ladant D (2009) RTX Calcium Binding Motifs Are Intrinsically Disordered in the Absence of Calcium. Journal of Biological Chemistry 284: 1781-1789.

184. Linhartová I, Bumba L, Mašín J, Basler M, Osička R, Kamanová J, Procházková K, Adkins I, Hejnová-Holubová J, Sadílková L, Morová J, Šebo P (2010) RTX proteins: a highly diverse family secreted by a common mechanism. FEMS Microbiology Reviews 34: 1076-1112.

185. Delepelaire P (2004) Type I secretion in gram-negative bacteria. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1694: 149-161.

186. Hartmann M, Duun AS, Markussen S, Grasdalen H, Valla S, Skjåk-Bræk G (2002) Time-resolved ^{1}H and ^{13}C NMR spectroscopy for detailed analyses of the *Azotobacter vinelandii* mannuronan C-5 epimerase reaction. Biochimica et biophysica acta 1570: 104-112.

187. Hartmann M, Holm OB, Johansen GA, Skjåk-Bræk G, Stokke BT (2002) Mode of action of recombinant *Azotobacter vinelandii* mannuronan C-5 epimerases AlgE2 and AlgE4. Biopolymers 63: 77-88.

188. Campa C, Holtan S, Nilsen N, Bjerkan TM, Stokke BT, Skjåk-Bræk G (2004) Biochemical analysis of the processive mechanism for epimerization of alginate by mannuronan C-5 epimerase AlgE4. The Biochemical journal 381: 155-164.

189. Ramstad MV, Ellingsen TE, Josefsen KD, Høidal HK, Valla S, Skjåk-Bræk G, Levine DW (1999) Properties and action pattern of the recombinant mannuronan C-5-epimerase AlgE2. Enzyme and Microbial Technology 24: 636-646.

190. Ramstad MV, Markussen S, Ellingsen TE, Skjåk-Bræk G, Levine DW (2001) Influence of environmental conditions on the activity of the recombinant mannuronan C-5-epimerase AlgE2. Enzyme Microb Technol 28: 57-69.

191. Friebolin H (2004) Basic One- and Two-Dimensional NMR Spectroscopy. 406.

192. Claridge T (1999) High-Resolution NMR Techniques in Organic Chemistry. A Pergamon Title: 382.

193. Aue W, Bartholdi E, Ernst R (1976) Two-dimensional spectroscopy. Application to nuclear magnetic resonance. J Chem Phy 64: 2229-2246.

194. Bermel W, Bertini I, Duma L, Felli IC, Emsley L, Pierattelli R, Vasos PR (2005) Complete assignment of heteronuclear protein resonances by protonless NMR spectroscopy. Angewandte Chemie (International ed 44: 3089-3092.

195. Bermel W, Bertini I, Felli IC, Kümmerle R, Pierattelli R (2006) Novel ^{13}C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. J Magn Reson 178: 56-64.

196. Bermel W, Bertini I, Felli IC, Lee YM, Luchinat C, Pierattelli R (2006) Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. Journal of the American Chemical Society 128: 3918-3919.

197. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proceedings of the National Academy of Sciences of the United States of America 94: 12366-12371.

198. Pervushin K (2000) Impact of transverse relaxation optimized spectroscopy (TROSY) on NMR as a technique in structural biology. Quarterly reviews of biophysics 33: 161-197.

199. Salzmann M, Pervushin K, Wider G, Senn H, Wüthrich K (1998) TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. Proceedings of the National Academy of Sciences of the United States of America 95: 13585-13590.

200. Yang D, Kay LE (1999) TROSY Triple-Resonance Four-Dimensional NMR Spectroscopy of a 46 ns Tumbling Protein. Journal of the American Chemical Society 121: 2571-2575.

201. Iwaï H, Züger S, Jin J, Tam PH (2006) Highly efficient protein *trans*-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. FEBS letters 580: 1853-1858.

202. Züger S, Iwaï H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nature biotechnology 23: 736-740.

203. Busche AE, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dotsch V, Iwaï H (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein *trans*-splicing using only one robust DnaE intein. Angewandte Chemie (International ed 48: 6128-6131.

204. Buchinger E, Aachmann FL, Aranko AS, Valla S, Skjåk-Bræk G, Iwaï H, Wimmer R (2010) Use of protein trans-splicing to produce active and segmentally ^{2}H, ^{15}N labeled mannuronan C5-epimerase AlgE4. Protein Science 19: 1534-1543.

205. Yamazaki T, Tochio H, Furui J, Aimoto S, Kyogoku Y (1997) Assignment of Backbone Resonances for Larger Proteins Using the ^{13}C-^{1}H Coherence of a ^{1}Hα-, ^{2}H-, ^{13}C-, and ^{15}N-Labeled Sample. Journal of the American Chemical Society 119: 872-880.

206. Goto NK, Gardner KH, Mueller GA, Willis RC, Kay LE (1999) A robust and cost-effective method for the production of Val, Leu, Ile (δ 1) methyl-protonated ^{15}N-, ^{13}C-, ^{2}H-labeled proteins. Journal of biomolecular NMR 13: 369-374.
207. Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. Proceedings of the National Academy of Sciences of the United States of America 102: 622-627.
208. Tugarinov V, Kay LE (2003) Quantitative NMR studies of high molecular weight proteins: application to domain orientation and ligand binding in the 723 residue enzyme malate synthase G. Journal of molecular biology 327: 1121-1133.
209. Neri D, Szyperski T, Otting G, Senn H, Wüthrich K (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. Biochemistry 28: 7510-7516.
210. Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440: 52-57.
211. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. Journal of biomolecular NMR 13: 289-302.
212. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. Journal of biomolecular NMR 44: 213-223.
213. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of molecular biology 273: 283-298.
214. Güntert P (2004) Automated NMR Structure Calculation with CYANA. Meth Mol Biol 278: 353-378.
215. Microcal I (2001) VP-ITC Instrction Manual.
216. Velázquez-Campoy A, Ohtaka H, Nezami A, Muzammil S, Freire E (2001) Isothermal Titration Calorimetry: John Wiley & Sons, Inc.
217. Glatter O, Kratky O (1982) Small-angle X-ray Scattering. Academic Press, London.
218. Svergun DI, Koch MHJ (2003) Small-angle scattering studies of biological macromolecules in solution. Reports on Progress in Physics 66: 1735-1782.
219. Vachette P, Koch MHJ, Svergun DI, Charles W. Carter JaRMS (2003) Looking behind the Beamstop: X-Ray Solution Scattering Studies of Structure and Conformational Changes of Biological Macromolecules. Methods in Enzymology: Academic Press. pp. 584-615.
220. Svergun DI, Koch MH (2002) Advances in structure analysis using small-angle scattering in solution. Current opinion in structural biology 12: 654-660.
221. http://www.ipfdd.de/X-ray-Lab.197.0.html?&L=.
222. Debye P (1915) Zerstreuung von Röntgenstrahlen. Scattering from non-crystalline substances. Ann Phys 46: 809-823.
223. Zhao J, Hoye E, Boylan S, Walsh DA, Trewhella J (1998) Quaternary structures of a catalytic subunit-regulatory subunit dimeric complex and the holoenzyme of the cAMP-dependent protein kinase by neutron contrast variation. The Journal of biological chemistry 273: 30448-30459.
224. Svergun DI, Volkov VV, Kozin MB, Stuhrmann HB (1996) New Developments in Direct Shape Determination from Small-Angle Scattering. Acta Crystallogr 52: 419-426.
225. Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophysical journal 76: 2879-2886.

226. Svergun DI, Petoukhov MV, Koch MH (2001) Determination of domain structure of proteins from X-ray solution scattering. Biophysical journal 80: 2946-2953.

227. Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. Journal of Applied Crystallography 34: 33-41.

228. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Journal of Applied Crystallography 28: 768-773.

229. Petoukhov MV, Eady NA, Brown KA, Svergun DI (2002) Addition of missing loops and domains to protein models by x-ray solution scattering. Biophysical journal 83: 3113-3125.

230. Costenaro L, Grossmann JG, Ebel C, Maxwell A (2005) Small-Angle X-Ray Scattering Reveals the Solution Structure of the Full-Length DNA Gyrase A Subunit. Structure 13: 287-296.

231. Edwards MJ, Flatman RH, Mitchenall LA, Stevenson CEM, Le TBK, Clarke TA, McKay AR, Fiedler H-P, Buttner MJ, Lawson DM, Maxwell A (2009) A Crystal Structure of the Bifunctional Antibiotic Simocyclinone D8, Bound to DNA Gyrase. Science (New York, NY 326: 1415-1418.

232. Corbett KD, Shultzaberger RK, Berger JM (2004) The C-terminal domain of DNA gyrase A adopts a DNA-bending β-pinwheel fold. Proceedings of the National Academy of Sciences of the United States of America 101: 7293-7298.

233. Konarev PV, Petoukhov MV, Svergun DI (2001) MASSHA - a graphics system for rigid-body modelling of macromolecular complexes against solution scattering data. Journal of Applied Crystallography 34: 527-532.

234. Zettler J, Schütz V, Mootz HD (2009) The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. FEBS letters 583: 909-914.

235. Martin DD, Xu MQ, Evans TC, Jr. (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis sp.* PCC6803. Biochemistry 40: 1393-1402.

236. Aranko AS, Züger S, Buchinger E, Iwaï H (2009) *In vivo* and *in vitro* protein ligation by naturally occurring and engineered split DnaE inteins. PloS one 4: e5185.

237. Girish V, Vijayalakshmi A (2004) Affordable image analysis using NIH Image/ImageJ. 47-47 p.

238. Setubal JC, dos Santos P, Goldman BS, Ertesvåg H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D, Cromes L, Curatti L, Du Z, Godsy E, Goodner B, Hellner-Burris K, Hernandez JA, Houmiel K, Imperial J, Kennedy C, Larson TJ, Latreille P, Ligon LS, Lu J, Mærk M, Miller NM, Norton S, O'Carroll IP, Paulsen I, Raulfs EC, Roemer R, Rosser J, Segura D, Slater S, Stricklin SL, Studholme DJ, Sun J, Viana CJ, Wallin E, Wang B, Wheeler C, Zhu H, Dean DR, Dixon R, Wood D (2009) Genome Sequence of *Azotobacter vinelandii*, an Obligate Aerobe Specialized To Support Diverse Anaerobic Metabolic Processes. J Bacteriol 191: 4534-4545.

239. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics (Oxford, England) 23: 2947-2948.

240. Bjerkan TM, Lillehov BE, Strand WI, Skjåk-Bræk G, Valla S, Ertesvåg H (2004) Construction and analyses of hybrid *Azotobacter vinelandii* mannuronan C-5 epimerases with new epimerization pattern characteristics. The Biochemical journal 381: 813-821.

241. Dahlheim MØ (2010) Functional Study of its Effect(s) on Alginate Epimerisation. Masteroppgave.

242. Aachmann FL, Svanem BIG, Valla S, Petersen SB, Wimmer R (2005) NMR assignment of the R-module from the *Azotobacter vinelandii* Mannuronan C5-epimerase AlgE4. Journal of biomolecular NMR 31: 259-259.

243. Aachmann FL, Skjåk-Bræk G (2008) ^{1}H, ^{15}N, ^{13}C resonance assignment of the AlgE6R1 subunit from the *Azotobacter vinelandii* mannuronan C5-epimerase. Biomol NMR Assign 2: 123-125.

244. Andreassen T, Buchinger E, Skjåk-Bræk G, Valla S, Aachmann F (2010) ^{1}H, ^{13}C and ^{15}N resonances of the AlgE62 subunit from *Azotobacter vinelandii*; mannuronan C5-epimerase. Biomolecular NMR Assignments: 1-3.

245. Buchinger E, Skjåk-Bræk G, Valla S, Wimmer R, Aachmann F (2011) NMR assignments of ^{1}H, ^{13}C and ^{15}N resonances of the C-terminal subunit from *Azotobacter vinelandii* mannuronan C5-epimerase 6 (AlgE6R3). Biomolecular NMR Assignments 5: 27-29.

246. Winther-Larsen HC, Blatny JM, Valand B, Brautaset T, Valla S (2000) Pm Promoter Expression Mutants and Their Use in Broad-Host-Range RK2 Plasmid Vectors. Metabolic engineering 2: 92-103.

247. Blatny JM, Brautaset T, Winther-Larsen HC, Karunakaran P, Valla S (1997) Improved Broad-Host-Range RK2 Vectors Useful for High and Low Regulated Gene Expression Levels in Gram-Negative Bacteria. Plasmid 38: 35-51.

PAPER I

12.11.2010

AALBORG UNIVERSITET

Declaration for Edith Buchinger's share of work in the following articles

Aranko AS, Züger S, Buchinger E, Iwaï H (2009) In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. PLoS One 4: e5185.

Experimental design, molecular genetics, protein expression and purification for new split SspDnaE intein was performed by SZ. Construction of plasmids, protein expression and purification for new split NpuDnaE intein was performed by ASA. The *in vivo* and *in vitro* ligation tests were performed by ASA, SZ and EBU. ASA, SZ and HI wrote the paper. HI was active in the formulation of the project, fundraising and in regular discussions of progress.

S. Züger

Buchinger

A. Sesilja Aranko

Sarah Züger

Edith Buchinger

Hideo Iwaï

In Vivo and *In Vitro* Protein Ligation by Naturally Occurring and Engineered Split DnaE Inteins

A. Sesilja Aranko, Sara Züger[¤a], Edith Buchinger[¤b], Hideo Iwaï[*]

Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Abstract

Background: Protein *trans*-splicing by naturally occurring split DnaE inteins is used for protein ligation of foreign peptide fragments. In order to widen biotechnological applications of protein *trans*-splicing, it is highly desirable to have split inteins with shorter C-terminal fragments, which can be chemically synthesized.

Principal Findings: We report the identification of new functional split sites in DnaE inteins from *Synechocystis sp*. PCC6803 and from *Nostoc punctiforme*. One of the newly engineered split intein bearing C-terminal 15 residues showed more robust protein *trans*-splicing activity than naturally occurring split DnaE inteins in a foreign context. During the course of our experiments, we found that protein ligation by protein *trans*-splicing depended not only on the splicing junction sequences, but also on the foreign extein sequences. Furthermore, we could classify the protein *trans*-splicing reactions in foreign contexts with a simple kinetic model into three groups according to their kinetic parameters in the presence of various reducing agents.

Conclusion: The shorter C-intein of the newly engineered split intein could be a useful tool for biotechnological applications including protein modification, incorporation of chemical probes, and segmental isotopic labelling. Based on kinetic analysis of the protein splicing reactions, we propose a general strategy to improve ligation yields by protein *trans*-splicing, which could significantly enhance the applications of protein ligation by protein *trans*-splicing.

Citation: Aranko AS, Züger S, Buchinger E, Iwaï H (2009) *In Vivo* and *In Vitro* Protein Ligation by Naturally Occurring and Engineered Split DnaE Inteins. PLoS ONE 4(4): e5185. doi:10.1371/journal.pone.0005185

Editor: Hilal Lashuel, Swiss Federal Institute of Technology Lausanne, Switzerland

Received January 11, 2009; **Accepted** March 5, 2009; **Published** April 13, 2009

Funding: This work was supported by the grants from NSERC (298493), Sigrid Juselius Foundation, and the Academy of Finland (118385). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

[*] E-mail: hideo.iwai@helsinki.fi

[¤a] Current address: Biochemisches Institut der Universität Zürich, Zürich, Switzerland
[¤b] Current address: Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Aalborg, Denmark

Introduction

Protein splicing is a post-translational modification, in which an intervening protein splicing domain (intein) catalyzes ligation of the two flanking N- and C-terminal segments (N-extein and C-extein) by a peptide bond and concomitantly excises itself from the precursor protein [1–3]. Protein splicing can also take place in *trans* by ligating separate protein fragments containing each half of a naturally or artificially split intein (N-intein and C-intein) [4–6]. This protein *trans*-splicing (PTS) could also work in foreign contexts where the naturally occurring extein segments are replaced with other foreign protein sequences of interest. Therefore, protein *trans*-splicing can be used for ligation of polypeptide chains with a peptide bond for protein semi-synthesis, protein cyclization, segmental isotopic labelling, and site-specific protein modifications [7–12]. Protein *trans*-splicing has also been exploited to control protein functions in living organisms as a post-translational modification [13–15]. Thus, protein *trans*-splicing could be widely used in biotechnology and chemical biology [16].

Inteins usually consist of two domains, namely, a Hint domain and an endonuclease domain [3]. Since only the Hint domain is required for protein splicing, several inteins have been minimized by removing the endonuclease domain for biotechnological applications [17–19]. The Hint domain could be reduced to as small as 135 residues, which is presumably the minimal functional length [19]. Naturally occurring split inteins contain 102–111 residues for N-intein ($\mathrm{Int_N}$) and 35–36 residues for C-intein ($\mathrm{Int_C}$) [20]. Short functional fragments of inteins have been of special interest because they could be easily prepared by chemical synthesis [12] and widen applications of protein *trans*-splicing for chemical modifications and protein semi-synthesis [21]. The shortest fragment identified so far is the N-terminal 11 residues of *Synechocystis sp*. PCC6803 (*Ssp*) DnaB intein [22]. Our interest was to identify functional split DnaE inteins with a shorter C-intein. Shorter C-inteins could be used as a ligation tag that can be easily synthesized or fused with other proteins for protein ligation.

In this study, a series of split DnaE inteins with new split sites have been constructed and tested for protein ligation both *in vivo* and *in vitro* to identify a functional split DnaE intein with a minimal C-terminal fragment. The robustness of the short C-intein has been tested by ligation of two domains that could not be ligated by wild-type DnaE intein. We also investigated the effect of extein sequences on protein ligation by protein *trans*-splicing. The effect of various reducing agents on *in vitro* protein ligation was tested with several target proteins.

Results

Construction of split SspDnaE inteins with new split sites

*Ssp*DnaE intein is one of the naturally occurring split inteins widely used in biotechnological applications. Naturally occurring split inteins can spontaneously induce protein splicing in *trans* after association of the N- and C-terminal parts (Figure 1a). In contrast, artificially split inteins often require tedious denaturation and renaturation steps to restore protein splicing activity because of lower solubility of the precursor fragments [23]. Protein ligation of two flanking foreign sequences through protein *trans*-splicing by naturally split inteins usually requires no additional cofactor, but a few residues of the original extein sequences might be necessary for efficient splicing [11]. To identify new functional split inteins with a shorter C-intein, we have moved the split site in naturally split *Ssp*DnaE intein towards the C-terminus by shortening the C-terminal half (*Ssp*DnaE-Int$_C$) systematically by 6–7 residues and elongating the N-terminal half (*Ssp*DnaE-Int$_N$) by approximately the same lengths (Figure 1b, Figure S1). *Ssp*DnaE-Int$_N$s were fused with the N-terminally His-tagged B1 domain of protein G (GB1),

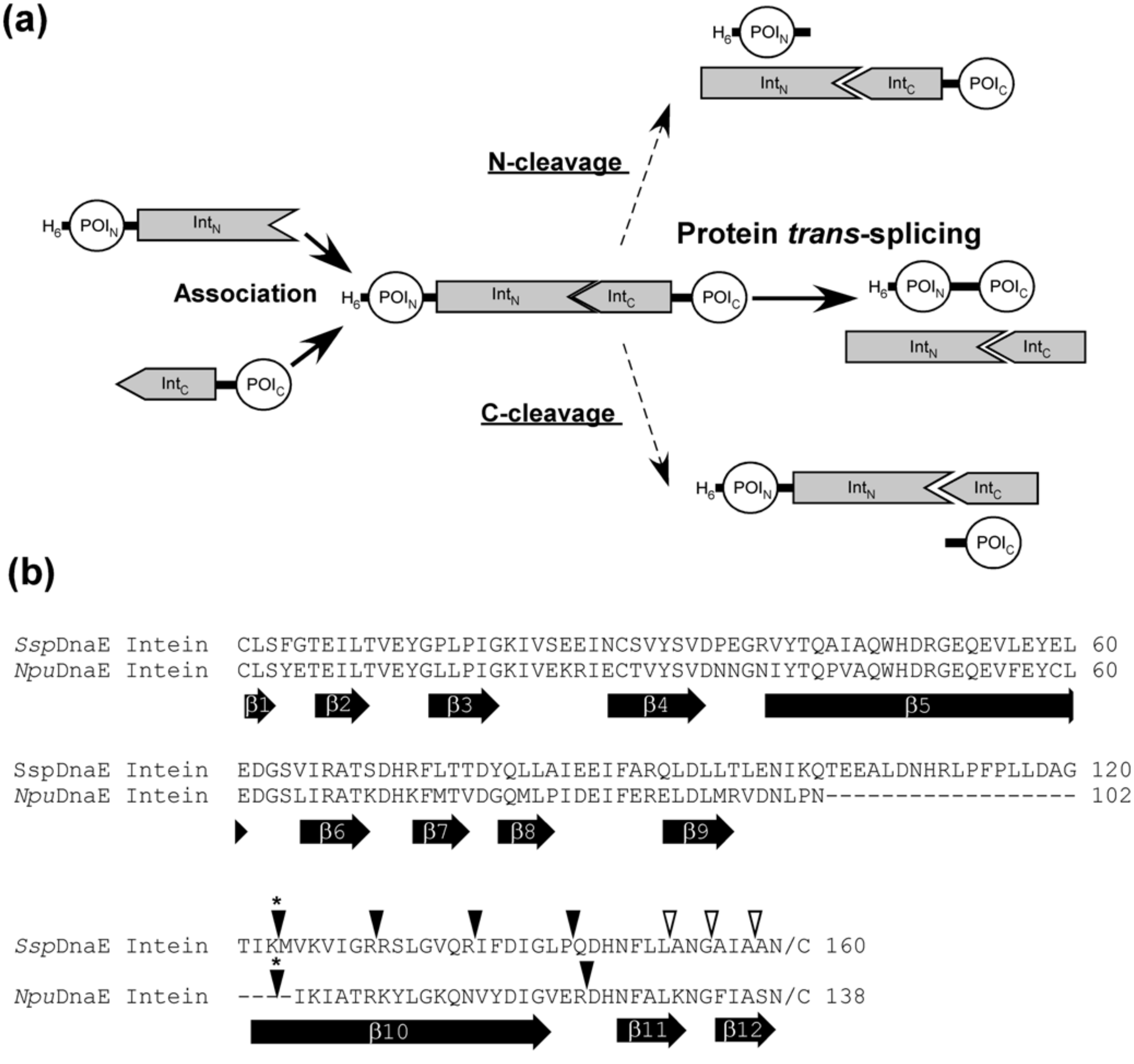

(a)
H_6 POI$_N$
Int$_N$
Int$_C$
POI$_C$
N-cleavage
Association
Protein *trans*-splicing
C-cleavage
(b)
*Ssp*DnaE Intein CLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGEQEVLEYEL 60
*Npu*DnaE Intein CLSYETEILTVEYGLLPIGKIVEKRIECTVYSVDNNGNIYTQPVAQWHDRGEQEVFEYCL 60
β1
β2
β3
β4
β5
SspDnaE Intein EDGSVIRATSDHRFLTTDYQLLAIEEIFARQLDLLTLENIKQTEEALDNHRLPFPLLDAG 120
*Npu*DnaE Intein EDGSLIRATKDHKFMTVDGQMLPIDEIFERELDLMRVDNLPN------------------ 102
β6
β7
β8
β9
*
*Ssp*DnaE Intein TIKMVKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIAAN/C 160
*
*Npu*DnaE Intein -----IKIATRKYLGKQNVYDIGVERDHNFALKNGFIASN/C 138
β10
β11
β12

Figure 1. Protein *trans*-splicing and locations of the new split sites. (a) Schematic representation of the protein *trans*-splicing process and two possible side reactions of N- and C-cleavage. Two fragments of the protein of interest (POI) can be ligated by protein *trans*-splicing reaction. (b) Sequence alignment of *Ssp*DnaE and *Npu*DnaE inteins. The locations of the experimentally tested split sites of *Ssp*DnaE and *Npu*DnaE inteins are indicated by inverse triangles on the top of the primary sequences. The asterisks above inverse triangles indicate the naturally occurring split site. Filled triangles indicate the split sites, where the split inteins retained protein *trans*-splicing activity. Open triangles indicate the split sites, where no protein *trans*-splicing activity could be detected. The location of the b-strands observed in the crystal structures of *Ssp*DnaE intein (PDB code 1ZDE) [35] and *Ssp*DnaB mini-intein (1MI8) [36] are indicated at the bottom of the sequences. The numbering for b-strands is adapted from *Ssp*DnaB mini-intein [36].

doi:10.1371/journal.pone.0005185.g001

of which expression was under the control of an inducible T7 promoter [8]. Previously, we found that the change of the N-terminal junction sequence of EY from *Ssp*DnaE to other sequences such as GS had little influence on the ligation yield [24]. Therefore, we used a linker of GS originated from the restriction site of *Bam*HI between GB1 and *Ssp*DnaE-$\mathrm{Int_N}$s. *Ssp*DnaE-$\mathrm{Int_C}$s were also fused to a chitin binding domain (CBD), of which expression was controlled by an arabinose promoter (Figure S1). We kept the sequence of CFNK from the wild-type junction sequence of *Ssp*DnaE and added GT for the cloning site of *Kpn*I as a linker between *Ssp*DnaE-$\mathrm{Int_C}$s and CBD [8]. GB1 and CBD were used here as model proteins because they are small soluble proteins. The N- and C-precursor proteins were genetically encoded into two separate plasmids that bear the compatible RSF3010 and ColE1 origins [8]. Seven plasmids for each half were constructed for testing *in vivo* and *in vitro* protein ligation (Table S1 and Figure S1).

In vivo protein *trans*-splicing by the new split *Ssp*DnaE inteins

Protein ligation by the new split inteins was tested *in vivo* using the dual vector system previously developed in our group [8]. This system allows us to conveniently check protein ligation because protein ligation could be initiated by the induction of the two precursor fragments with the two inducers, isopropyl-β-D-thiogalactoside (IPTG) and arabinose, and subsequently analyzed by SDS-PAGE [24]. Moreover, endogenous auxiliary factors such as chaperones might improve protein ligation in cells by promoting correct protein folding. The C-terminal part was always first induced for 0.5 hours ensuring an excess of the C-terminal precursor prior to the expression of the N-terminal precursor, and followed by the induction of the N-terminal precursor for another 3.5–5.5 hours. The pre-existing C-terminal precursor protein could be converted into the ligated product through protein *trans*-splicing after the association with the N-terminal part and protein splicing. The expression level of the N-terminal fragment was monitored by SDS-PAGE in order to avoid an enormous excess of the N-terminal part, which could underestimate the ligation yields. Immobilized Metal Affinity Chromatography (IMAC) was used to purify the N-terminal His-tagged precursor, the ligated product, and, if any, the cleaved N-terminal GB1 produced by the side reactions (Figure 1a). If i*n vivo* protein ligation works with 100% efficiency and if there is no excess of the N-terminal precursor, only $\mathrm{H_6}$-GB1-CBD will be purified by IMAC through the N-terminal His-tag. If the N- and C-terminal fragments associate with each other but no protein splicing is induced, both N- and C-terminal fragments ($\mathrm{H_6}$-GB1-*Ssp*DnaE-$\mathrm{Int_N}$ and *Ssp*DnaE-$\mathrm{Int_C}$-CBD) will be purified owing to the affinity between them. Furthermore, if the N- and C-inteins do not interact or if the C-terminal cleavage reaction is the dominant reaction after association of the N- and C-inteins, a single band of the N-terminal precursor is expected to be visible in the SDS gel. In some cases, during protein purification and sample preparation for SDS-PAGE, reactions such as splicing and cleavages could take place, which produced smaller bands of cleaved and spliced products. The ligated product was confirmed by mass-spectrometry (Figure S2). We could identify the ligated product $\mathrm{H_6}$-GB1-CBD in the elution fractions from IMAC only for the combinations of *Ssp*DnaE-$\mathrm{Int_{N123}}$/*Ssp*DnaE-$\mathrm{Int_{C36}}$ (wild-type), *Ssp*DnaE-$\mathrm{Int_{N130}}$/*Ssp*DnaE-$\mathrm{Int_{C30}}$, *Ssp*DnaE-$\mathrm{Int_{N137}}$/*Ssp*DnaE-$\mathrm{Int_{C23}}$, and *Ssp*DnaE-$\mathrm{Int_{N144}}$/*Ssp*DnaE-$\mathrm{Int_{C16}}$ (Figure 2a). The ligation yields were estimated from the ratios between the intensities of the ligated product and one of the most abundant residual precursor fragments in the SDS gel, which were ca. 3% for *Ssp*DnaE-$\mathrm{Int_{N144}}$/*Ssp*DnaE-$\mathrm{Int_{C16}}$, ca. 1% for *Ssp*DnaE-$\mathrm{Int_{N137}}$/*Ssp*DnaE-$\mathrm{Int_{C23}}$, and ca. 16% for *Ssp*DnaE-$\mathrm{Int_{N130}}$/*Ssp*DnaE-$\mathrm{Int_{C30}}$. These efficiencies might be underestimated if an excess of the N-terminal part was present during the expression due to the co-purification of the N-terminal precursor containing an N-terminal His-tag. The highest yield was estimated for the wild-type combination of *Ssp*DnaE-$\mathrm{Int_{N123}}$/*Ssp*DnaE-$\mathrm{Int_{C36}}$ (67%). Albeit the amounts of the ligated products produced by the newly engineered inteins were very small, the protein ligation was still detectable by SDS-PAGE. The split site of *Ssp*DnaE-$\mathrm{Int_{N144}}$/*Ssp*DnaE-$\mathrm{Int_{C16}}$ was the split site of the shortest C-intein retaining detectable splicing activity. However, the ligation efficiency was significantly lower than that of wild-type *Ssp*DnaE intein because of the low splicing activity and the side reactions. The pairs of *Ssp*DnaE-$\mathrm{Int_{N151}}$/*Ssp*DnaE-$\mathrm{Int_{C9}}$ and *Ssp*DnaE-$\mathrm{Int_{N154}}$/*Ssp*DnaE-$\mathrm{Int_{C6}}$ could not induce protein *trans*-splicing as only the N-terminal precursor was purified, indicating there was no significant interaction between them. On the other hand, the shortest C-intein construct of *Ssp*DnaE-$\mathrm{Int_{C3}}$ was purified together with the N-terminal *Ssp*DnaE-$\mathrm{Int_{N157}}$ indicating that there was sufficient interaction between them. However, we could not identify any ligated product although there was a band at 18.4 kDa in the SDS gel indicating a small amount of the N-cleavage reaction that produced $\mathrm{Int_N}$.

Split *Npu*DnaE intein with the new split site

The low ligation efficiencies of the newly functional split sites of *Ssp*DnaE intein suggest little practical use of these new split inteins. However, we have recently discovered that DnaE intein from *Nostoc punctiforme* (*Npu*) has more robust protein *trans*-splicing activity than that of *Ssp*DnaE intein and is also more tolerant of amino acid replacements at the C-terminal splicing junction [24]. Our previous study indicated that the N-terminal part (*Npu*DnaE-$\mathrm{Int_N}$) is responsible for the higher ligation efficiency [24]. Therefore, we were interested in introducing the new split site of *Ssp*DnaE intein into *Npu*DnaE intein to obtain sufficient protein *trans*-splicing activity for practical use. The new split site with the C-terminal 16 residues in *Ssp*DnaE is located between β-strands 10 and 11 (Figure 1b). We decided to shorten the C-intein by one more residue in *Npu*DnaE intein because based on the NMR structures of *Npu*DnaE intein (PDB entry, 2KEQ) the split site would be still in the loop between β-strands 10 and 11 [25,26]. Protein ligation *in vivo* by *Npu*DnaE-$\mathrm{Int_{N123}}$/*Npu*DnaE-$\mathrm{Int_{C15}}$ is demonstrated in Figure 2b. The C-terminal part (*Npu*DnaE-$\mathrm{Int_{C15}}$-GB1) was induced first by L-arabinose (lane 2, Figure 2b). After the consecutive induction of the N-terminal part ($\mathrm{H_6}$-GB1-*Npu*DnaE-$\mathrm{Int_{N123}}$), a large amount of the ligated product ($\mathrm{H_6}$-GB1-GB1) was accumulated (lane 3 and 4, Figure 2b). The fraction purified by IMAC contained almost no precursor proteins and the ligation was confirmed by mass spectrometry (lane 5, Figure 2b and Figure S3). We estimated the ligation efficiency to be ca. 96%, which is significantly better than any of the tested combinations of the newly split *Ssp*DnaE inteins. We also tested protein ligation by the combination of *Npu*DnaE-$\mathrm{Int_{N123}}$/*Ssp*DnaE-$\mathrm{Int_{C16}}$, which resulted in similar ligation efficiency (data not shown). This result emphasizes the dominant contribution of the N-intein to the ligation efficiency and suggests that the sequence variation between *Npu*DnaE-$\mathrm{Int_{C15}}$ and *Ssp*DnaE-$\mathrm{Int_{C16}}$ (the sequence identity is 66%) has little influence on protein *trans*-splicing efficiency.

Protein ligation of SH3 domains by the naturally split *Npu*DnaE intein

The robustness of naturally split *Npu*DnaE intein encouraged us to use *Npu*DnaE intein as a general tool for protein ligation and to apply it to biologically relevant proteins [24]. The Src

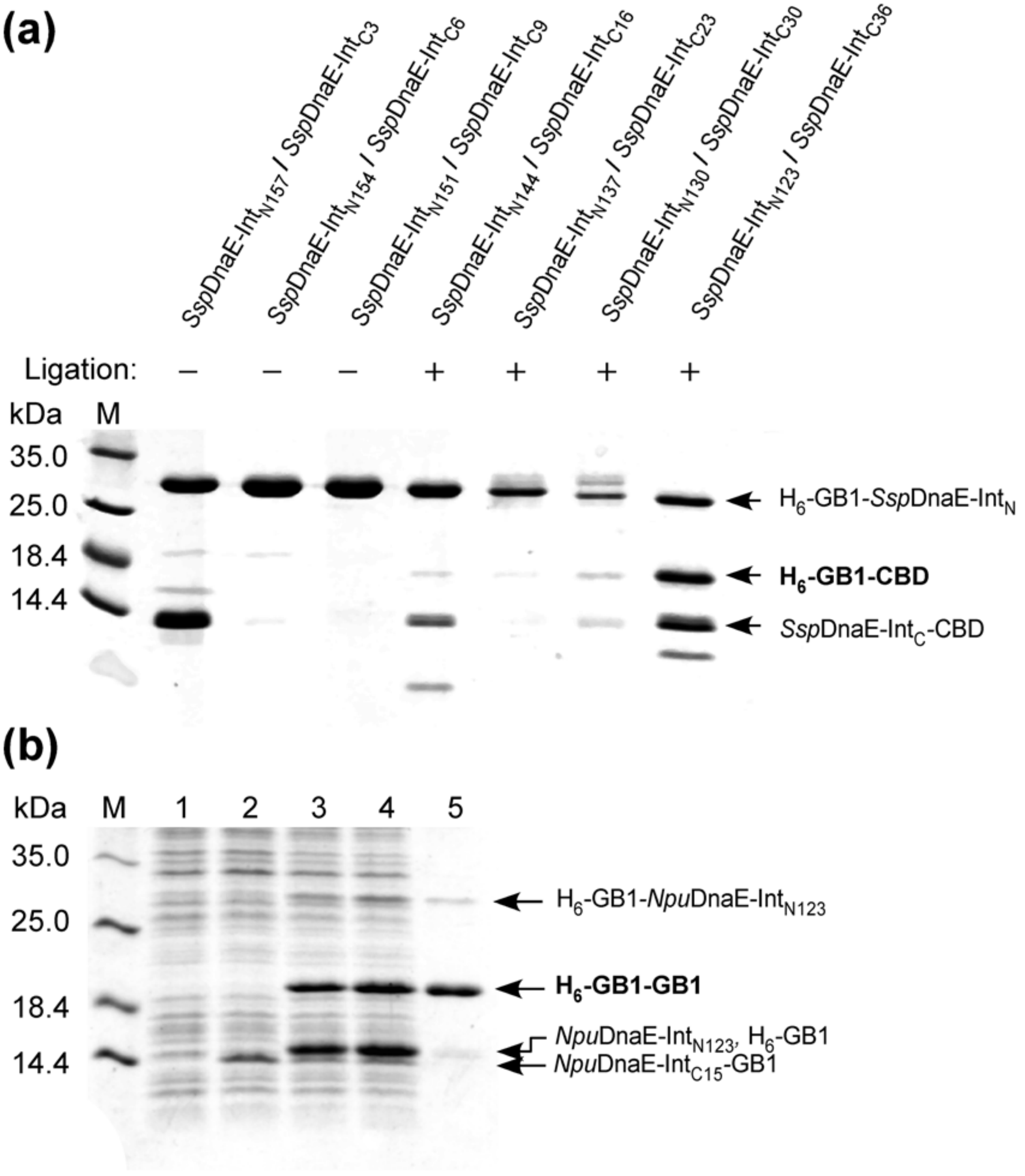

(a)
SspDnaE-IntN157 / SspDnaE-IntC3
SspDnaE-IntN154 / SspDnaE-IntC6
SspDnaE-IntN151 / SspDnaE-IntC9
SspDnaE-IntN144 / SspDnaE-IntC16
SspDnaE-IntN137 / SspDnaE-IntC23
SspDnaE-IntN130 / SspDnaE-IntC30
SspDnaE-IntN123 / SspDnaE-IntC36
Ligation: − − − + + + +
kDa M
35.0
25.0
18.4
14.4
H6-GB1-SspDnaE-IntN
H6-GB1-CBD
SspDnaE-IntC-CBD
(b)
kDa M 1 2 3 4 5
35.0
25.0
18.4
14.4
H6-GB1-NpuDnaE-IntN123
H6-GB1-GB1
NpuDnaE-IntN123, H6-GB1
NpuDnaE-IntC15-GB1

Figure 2. *In vivo* protein ligations by the newly engineered split *Ssp*DnaE and *Npu*DnaE inteins. (a) SDS-PAGE analysis of *in vivo* protein ligations by the newly engineered split *Ssp*DnaE inteins after purification with Ni-NTA. The combinations of *Ssp*DnaE-Int$_N$ and *Ssp*DnaE-Int$_C$ are indicated on the top of the lanes. (b) *In vivo* protein ligation by *Npu*DnaE intein with the newly engineered split site (*Npu*DnaE-Int$_{N123/C15}$). Lane 1, before induction; lane 2, 1.5 hours after induction only with arabinose; lane 3, 1.5 hours after additional induction with IPTG; lane 4, 3 hours after induction with IPTG and arabinose; lane 5, elution from Ni-NTA column.
doi:10.1371/journal.pone.0005185.g002

homology 3 (SH3) domain is one of the most abundant domains in multi-domain proteins. Therefore, we were interested in protein ligation of the two SH3 domains from c-Crk-II adaptor protein [27]. Despite the robustness of *Npu*DnaE intein, protein ligation of the two SH3 domains by wild-type *Npu*DnaE intein was not possible, because the side reactions were dominating the *trans*-splicing and producing mainly cleaved products (Figure 3a and 3c). When the N-terminal SH3 (nSH3) was replaced with the model protein GB1, both *in vivo* and *in vitro* ligation of the two proteins by protein *trans*-splicing was still not possible with high yields (Figure 3b and 3d, Figure S4). On the other hand, the ligation of the two proteins *in vitro* as well as *in vivo* was significantly improved after replacing the C-terminal SH3 (cSH3) with GB1 (Figure 4a, Table 1). These observations indicate that protein *trans*-splicing can be significantly influenced not only by the sequences near the splicing junctions but also by the exteins, which brings additional complexity to protein *trans*-splicing. Furthermore, the replacement of the C-terminal precursor protein suggests that the C-terminal fragment containing cSH3 negatively affects the protein ligation.

The effect of reducing agent on *trans*-splicing

In theory, protein *trans*-splicing does not require any thiol agents for the reaction [3]. However, both N- and C-inteins of *Npu*DnaE

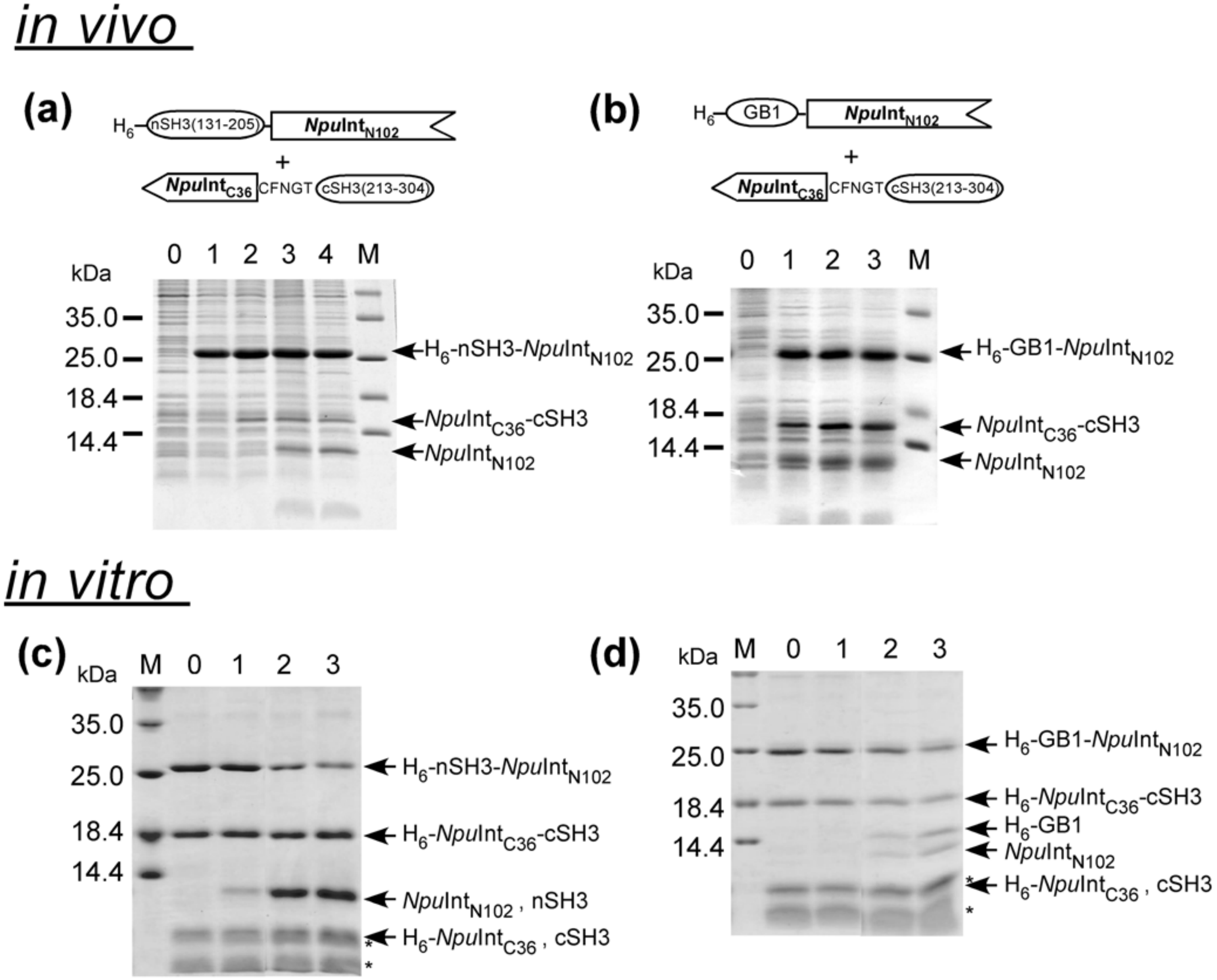


Figure 3. Protein ligation *in vivo* and *in vitro* by the naturally occurring split *Npu*DnaE intein. (a) Protein ligation of nSH3 and cSH3 *in vivo* by naturally occurring split *Npu*DnaE intein. Lane 0, before induction; lane 1, 1 hour after the induction with IPTG and arabinose; lane 2, 2 hours; lane 3, 4 hours; lane 4, 6 hours. (b) Protein ligation of GB1 and cSH3 *in vivo* by the wild-type *Npu*DnaE intein. Lane 0, before induction; lane 1, 2 hours after the induction with IPTG and arabinose; lane 2, 4 hours; lane 3, 6 hours. *In vitro* protein ligation (c) of nSH3 and cSH3 (d) of GB1 and cSH3 in the presence of 50 mM DTT. Lane 0, 0 min after the mixing; lane 1, 10 min; lane 2, 3 hours; lane 3, 24 hours for (c). Lane 0, 0 min after the mixing; lane 1, 3 min; lane 2, 3 hours; lane 3, 24 hours for (d). Asterisks indicating the bands below 14.4 kDa in (c) and (d) are impurities from the purification of H_6-*Npu*Int$_{C36}$-cSH3.

doi:10.1371/journal.pone.0005185.g003

intein contain unpaired cysteine residues that could form intermolecular disulfide bonds and they may prevent the appropriate association of the two fragments. Therefore, it is desirable to keep the reaction under reducing conditions with a sulfhydryl reductant. In a previous study on *Ssp*DnaE intein, it has been reported that the presence of 50 mM dithiothreitol (DTT) would almost totally block protein *trans*-splicing and instead shunt the reaction to *trans*-cleavage [28]. As a sulfhydryl reductant, we have tested two thiol agents (DTT and 2-mercaptoethane sulfonic acid, MESNA) and a trialkylphosphine (tris(2-carboxyethyl)phosphine, TCEP) that is unreactive with thiol groups such as cysteine (Figure 4c). In contrast to the previous report, we found that the effect of various reducing agents on protein ligation was negligible for the ligation between nSH3 and GB1 (Figure 4c) as well as for the ligation of the two SH3 domains (data not shown). In the case of nSH3 and cSH3, the reaction was always dominated by *trans*-cleavage rather than *trans*-splicing (Table 1, Figure 3). It was not possible to improve the ligation of those SH3 domains by replacing the reducing agent. For the ligation of nSH3 and GB1, *trans*-splicing was always observed regardless of the reducing agents used (Table 1, Figure 4c). To understand these puzzling effects, we analyzed the kinetics of the protein ligation. It is well accepted that protein-splicing reaction involves the four concerted steps: (1) N-S acyl shift, (2) *trans*-thioesterification, (3) Asn cyclization, and (4) S-N acyl shift, and possibly undesired side reactions of N- and C-cleavage (Figure 1a) [29]. The detailed kinetics of the individual steps has been previously characterized for *Ssp*DnaE intein [28]. We decided to approximate the reactions with a simple kinetic model as depicted in (I), in which the entire reaction was divided into the two parallel reactions: *trans*-splicing and cleavage reactions because the two reactions are both irreversible processes. In this model, we also assume that the formation of the precursor

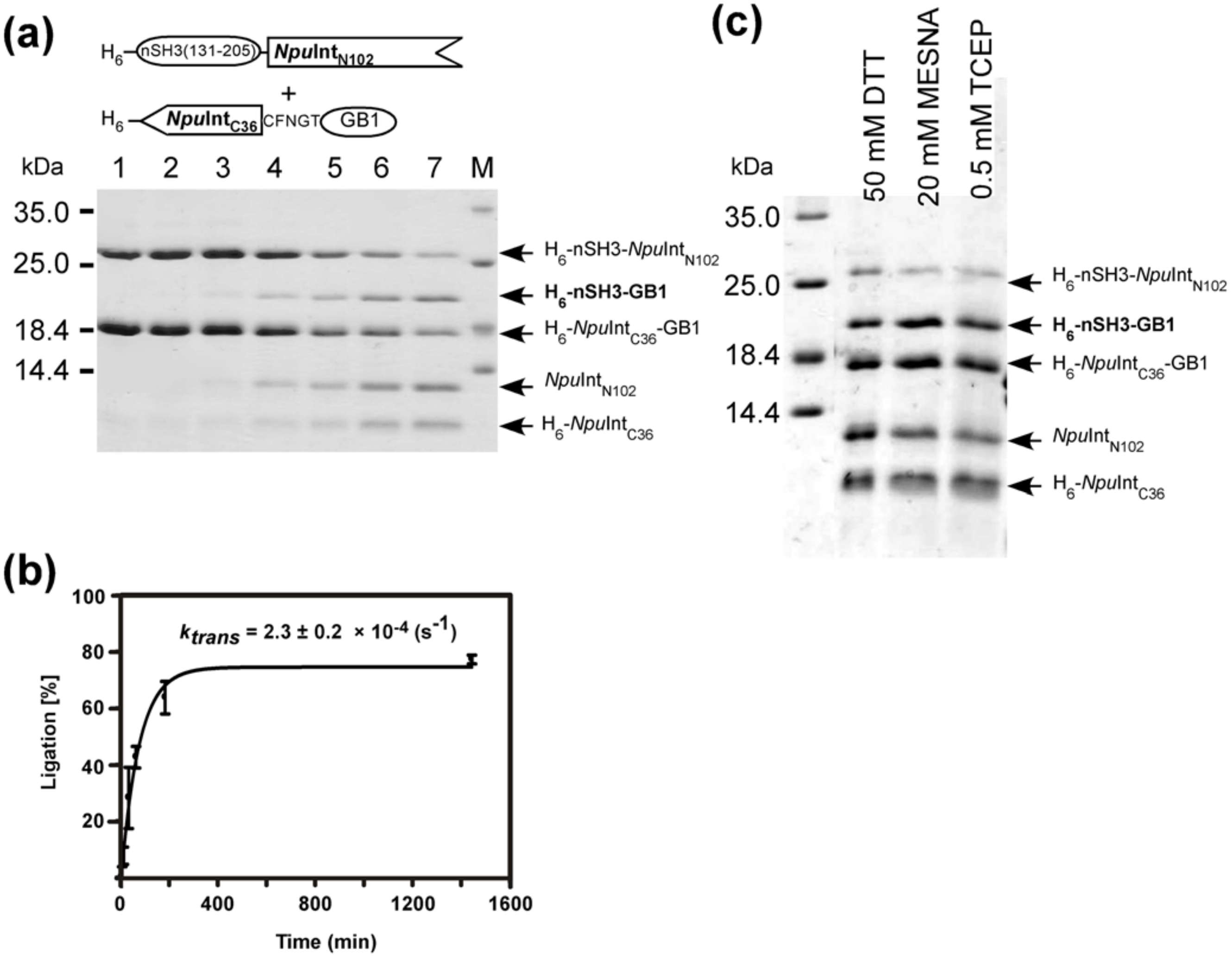


Figure 4. *In vitro* protein ligation of nSH3 and GB1 by the naturally occurring split *Npu*DnaE intein. (a) Time course of the protein ligation of nSH3 and GB1 by naturally occurring split *Npu*DnaE intein in the presence of 50 mM DTT. Lane 1, 0 min after the mixing; lane 2, 3 min; lane 3, 10 min; lane 4, 30 min; lane 5, 1 hour; lane 6, 3 hours; lane 7, 22 hours. (b) Kinetic analysis of the protein ligation from the SDS-PAGE. (c) SDS-PAGE analysis of the ligation reaction after overnight incubation in the presence of different reducing agents.
doi:10.1371/journal.pone.0005185.g004

complex is fast relative to the subsequent reaction steps and the dissociation constant is much smaller than the protein concentration used in the experiments [28].

$$C \xleftarrow{k_{unprod}} A \xrightarrow{k_{trans}} B \qquad (I)$$

A = precursor complex, B = ligated product, C = cleaved product, k_{trans} = 1st order kinetic constant for *trans*-splicing, and k_{unprod} = apparent 1st order kinetic constant for all unproductive side reactions including the N- and C-terminal cleavage reactions. Time courses of the products can be formulated by the following rate equations.

Table 1. The final yields of the protein ligation by protein *trans*-splicing.

Intein	N-extein	C-extein	Yield (%) with 50 mM DTT	Yield (%) with 0.5 mM TCEP
*Npu*DnaE-Int$_{N102/C36}$ (wild-type)	nSH3	cSH3	n.d.	n.d.
	GB1	cSH3	n.d.	n.d.
	nSH3	GB1	77±10	61±8
*Npu*DnaE-Int$_{N123/C15}$	nSH3	cSH3	27±10	50±7
	Smt3	GB1	9±3	65±7

n.d., not detectable.
doi:10.1371/journal.pone.0005185.t001

$$\frac{d[A]}{dt} = -\left(k_{unprod} + k_{trans}\right)[A],$$
$$\frac{d[B]}{dt} = k_{trans}[A],\ \frac{d[C]}{dt} = k_{unprod}[A] \tag{II}$$

These equations can be easily solved [30]. The yield of the ligation at an infinite time can be derived from the two kinetic constants for *trans*-splicing and cleavage according to Eq. (III).

$$E(\%) = 100 \times \frac{k_{trans}}{k_{trans} + k_{unprod}} \tag{III}$$

With this model, we should be able to estimate ligation yields from the rate constants of *trans*-splicing and side reactions, and *vice versa*. For the ligation between nSH3 and GB1, $2.3 \pm 0.2 \times 10^{-4}$ (s^{-1}) was estimated for k_{trans} in the presence of 50 mM DTT (Figure 4b). According to Eq. (III) using the obtained kinetic constants and the reported DTT induced cleavage rate constant for *Ssp*DnaE intein ($1.0 \pm 0.5 \times 10^{-3}$ (s^{-1})) [28], the ligation yield for nSH3 and GB1 is expected to be 12–33%. However, the obtained final yield of close to 80% might suggest that the rate constant of DTT induced cleavage is about 1×10^{-4} with this system (Table 1). *Trans*-splicing was not detectable for nSH3 and cSH3, but the side reactions were dominant with the kinetic constant $k_{unprod} = 5.4 \pm 0.4 \times 10^{-4}$ (s^{-1}) in the presence of 50 mM DTT (data not shown). Although the replacement of DTT with TCEP as a reducing agent slowed the unproductive cleavage reactions, *trans*-splicing was not detectable. This suggests that *trans*-splicing reaction occurs at a significantly slower rate than the cleavage reaction. This model assumes that the association rates are fast and that the dissociation rates are similarly low for different exteins compared with the experimental concentration. Therefore, when the estimation of the yield is largely discrepant with the kinetic constants, the limiting factor is likely to be imposed by the association rate. Thus, this simple model and the kinetic analysis might provide a useful tool to predict final yields as well as to identify the rate-limiting step in protein *trans*-splicing reaction.

Protein ligation by the newly engineered split *Npu*DnaE

From the aforementioned results with the SH3 domains, we assumed that the C-intein fused with cSH3 is the limiting factor for the protein ligation of two SH3 domains, inducing fast cleavage reactions. We believe that cSH3 probably interferes with association of N- and C-inteins of wild-type *Npu*DnaE intein and that the shorter C-intein might not interfere the ligation of the two SH3 domains. Therefore, we decided to replace the intein with the newly engineered *Npu*DnaE intein (*Npu*DnaE-Int$_{N123}$/*Npu*DnaE-Int$_{C15}$)

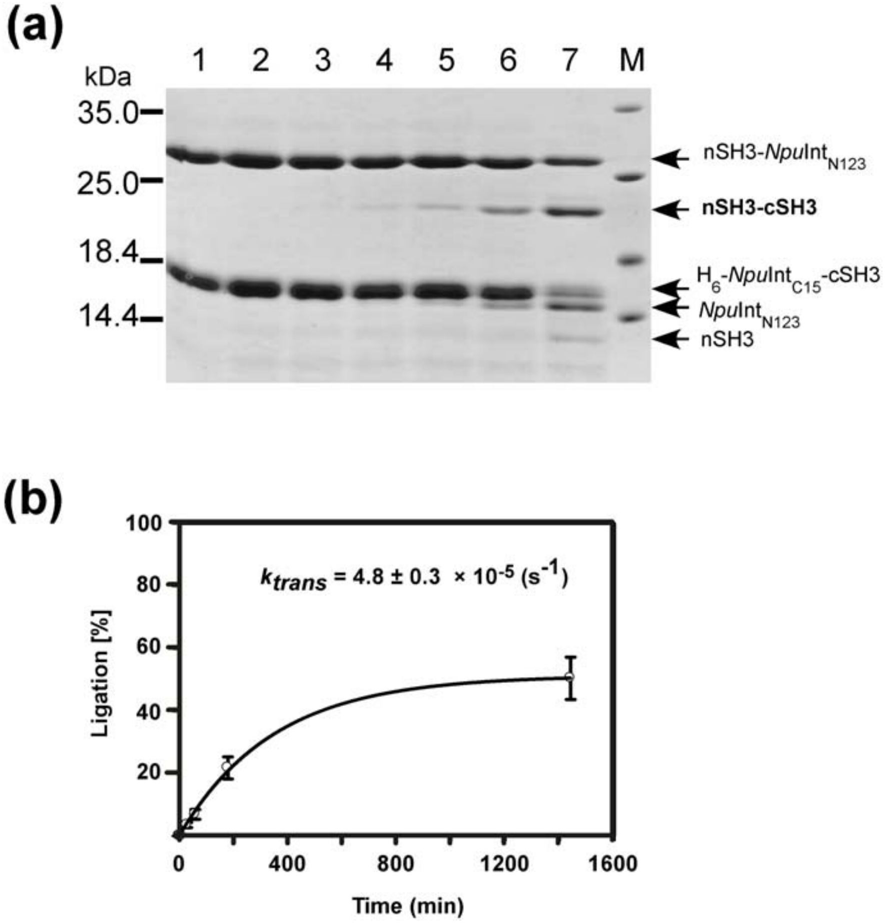


Figure 5. Protein ligation of two SH3 domains by the newly engineered split *Npu*DnaE intein. (a) SDS-PAGE analysis of the time course from the protein ligation reaction of nSH3 and cSH3 in the presence of 0.5 mM TCEP. Lane 1, 0 min after the mixing; lane 2, 3 min; lane 3, 10 min; lane 4, 30 min; lane 5, 1 hour; lane 6, 3 hours; lane 7, 22 hours. (b) Kinetic analysis of the protein ligation from the SDS-PAGE.
doi:10.1371/journal.pone.0005185.g005

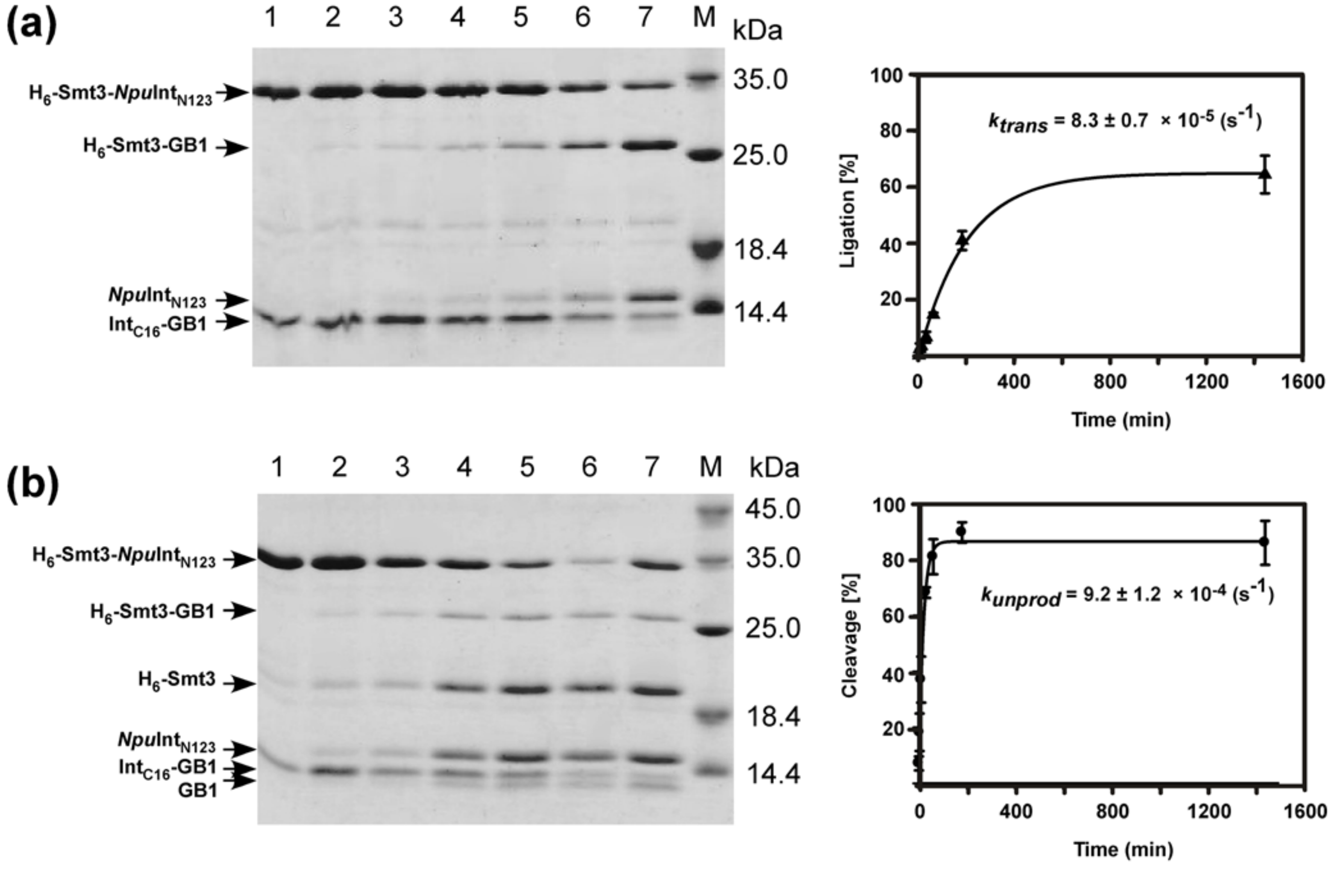


Figure 6. Protein ligation of Smt3 and GB1 by the newly engineered split *Npu*DnaE intein. Time courses and kinetic analysis of protein ligation in the presence of (a) 0.5 mM TCEP or (b) 50 mM DTT. SDS-PAGE: lane 1, 0 min after the mixing; lane 2, 3 min; lane 3, 10 min; lane 4, 30 min; lane 5, 1 hour; lane 6, 3 hours; lane 7, 24 hours.
doi:10.1371/journal.pone.0005185.g006

for the ligation. As demonstrated in Figure 5, the new split *Npu*DnaE intein could indeed ligate nSH3 and cSH3 that were not possible to be ligated by the naturally occurring split *Npu*DnaE intein. It demonstrates the effectiveness of the shorter C-intein in the case of difficult ligations such as the one between the two SH3 domains. The kinetic constants for *trans*-splicing were estimated to be $4.8 \pm 0.3 \times 10^{-5}$ (s^{-1}) in the presence of 0.5 mM TCEP (Figure 5b). Thus, the engineered split *Npu*DnaE intein can significantly improve the protein ligation by accelerating the *trans*-splicing reaction.

Protein ligation of Smt3 and GB1

Because of the strong influence of the extein sequences on protein *trans*-splicing, we wanted to test another small protein with a similar size, the yeast ubiquitin-like protein Smt3, for protein ligation [31]. Protein ligation of His-tagged Smt3 and GB1 by *Npu*DnaE-Int$_{N123}$ was tested in the presence of either 50 mM DTT or 0.5 mM TCEP. The protein ligation of Smt3 and GB1 responded differently to the two different reducing agents. When 0.5 mM TCEP was used, the yield was more than 60–70%. On the contrary, only about 10% of the protein ligation was achieved in the presence of 50 mM DTT, where the cleavage reaction dominated. In this case the kinetic constant for *trans*-splicing in presence of 0.5 mM TCEP was estimated to be $8.3 \pm 0.7 \times 10^{-5}$ (s^{-1}). DTT induced the dominant cleavage reaction with a kinetic constant of $9.2 \pm 1.2 \times 10^{-4}$ (s^{-1}) (Figure 6). The protein ligation yield in the presence of 50 mM DTT is expected to be around 10% as it can be derived from Eq. (III) with an assumption that *trans*-splicing rates are similar for both DTT and TCEP. This is in good agreement with the yield obtained experimentally, suggesting that the simple model is appropriate for roughly estimating the yield without any intricate methods.

Discussion

In this article, we demonstrated that C-intein from *Ssp*DnaE and *Npu*DnaE inteins could be shortened to C-terminal 16 or 15 residues without abolishing protein *trans*-splicing activity. The newly engineered split *Npu*DnaE intein bearing the C-terminal 15 residues as C-intein retained robust protein *trans*-splicing activity. The use of the shorter C-intein was even more effective for the ligation of the two SH3 domains that could not be ligated by the wild-type split DnaE inteins. The shorter length of C-intein of the engineered split *Npu*DnaE intein could be attractive for chemical synthesis and suitable for incorporation of chemically modified peptides by protein *trans*-splicing [32]. Moreover, the kinetic analysis of the ligation reaction could be important because the kinetic parameters are the key factor determining the ligation yields. The analysis using a simple parallel model to approximate the reaction could be a convenient tool to investigate the rate-limiting steps in the reaction and to estimate the ligation yields based on the kinetic parameters. Protein *trans*-splicing reaction in foreign contexts can be categorized into three groups. In the first group only side reactions of cleavages can be observed. Various reducing agents such as TCEP or DTT have little effect on improving protein ligation in this group. In this case, the cleavage reaction has a typical kinetic constant of $>1 \times 10^{-4}$ (s^{-1}) and the *trans*-splicing rate is much slower than the cleavage rate. In the second group, regardless of the used reducing agents, protein ligation by protein *trans*-splicing can be observed. Here, the *trans*-splicing reaction is faster ($>1\times10^{-4}$ (s^{-1})) than the unproductive cleavage reactions induced by various reducing agents. In the third group, *trans*-splicing reaction is slower than the side reactions induced by DTT, but faster than the side reactions in the presence of TCEP. Therefore, the reducing agent could greatly influence the final yield. This is presumably because the thiol group of DTT is a nucleophile competing with the thiol of the first cysteine of C-intein and induces dominant cleavage reactions. However, the side reactions in the presence of TCEP are usually slower because it has no thiol group that functions as nucleophile competing with *trans*-splicing reaction.

In summary, we could create new functional split inteins with shorter C-inteins, which retained *trans*-splicing activity. Protein *trans*-splicing was found to be dependent on the protein sequences of the exteins even if the sequence around the splicing junctions were identical. How the exteins influence protein *trans*-splicing remains unclear. However, monitoring the kinetics of the protein *trans*-splicing reaction could be a useful tool to identify the rate-limiting steps in protein ligation reaction. To achieve a higher yield of protein ligation by protein *trans*-splicing, it is of importance to keep the competing side reactions slower than the *trans*-splicing reaction by replacing the reducing agent with non-thiol reducing agents such as TCEP or by accelerating the *trans*-splicing reaction using a more efficient split intein. The ligation between self-contained domains by protein *trans*-splicing was investigated in this article. However, the model describing the relation between *trans*-splicing and side reactions should be generally applicable even for the ligation within a single domain although such ligation may require refolding of the precursors which could be a more dominant factor affecting the yield. Further understanding of the factors influencing protein *trans*-splicing reaction rates such as folding processes of split inteins and engineering of split inteins could make protein *trans*-splicing a more versatile tool for protein modification, protein semi-synthesis, and segmental isotopic labelling.

Materials and Methods

Construction of plasmids

The N-terminal fragments of *Ssp*DnaE intein (*Ssp*DnaE-Int$_N$) of various lengths were previously constructed [24]. The coding sequences of *Ssp*DnaE-Int$_N$s were subcloned into pJJDuet30 between *Bam*HI and *Hin*dIII sites [8], resulting in the sequences coding for H$_6$-GB1-*Ssp*DnaE-Int$_N$s (Figure. S1). *Ssp*DnaE-Int$_C$s of various lengths were constructed from the plasmid pSZRS1 containing the gene of the full-length *Ssp*DnaE-Int$_C$ and the chitin binding domain (CBD) using synthetic oligonucleotides (Table S1) and cloned into pBAD vector (Figure S1).

The plasmid for H$_6$-GB1-*Npu*DnaE-Int$_{N123}$ was constructed by replacing the codon of residue 124 of the full-length *Npu*DnaE intein with a stop codon in the plasmid of pSKDuet16, resulting in pHYDuet36 [26]. The plasmid pSKDuet16 contains an additional two mutations of HM to LG at the front of GB1 due to the replacement of *Nde*I site by *Avr*II site, compared with the plasmids derived from pJJDuet30 [8]. The gene of *Npu*DnaE-Int$_{C15}$-GB1 was amplified from pSKDuet16 and cloned into a pBAD vector (pHYBAD44) [24]. *Npu*DnaE-Int$_{N123}$ was subcloned from pHYDuet36 into pHYRSF53LA using *Bam*HI and *Hin*dIII sites, which resulted in pHYRSF53-36 coding for H$_6$-Smt3-*Npu*DnaE-Int$_{N123}$ [26].

The gene of the N-terminal SH3 domain was amplified from pAT044 [33] with the two oligonucleotides (#HK009 and #SK202) and cloned into pHYRSF1-12 using *Nde*I and *Ahd*I sites, which resulted in pTMRSF07 (Table S1). The plasmid pHYRSF1-12 was previously constructed by transferring the gene of GB1 and the N-terminal *Npu*DnaE from pSKDuet1 into pRSF-1b using *Nco*I and *Hin*dIII sites. The plasmid of pHYRSF1-12 contains additional mutations of GS to TK to introduce *Ahd*I site at the front of *Npu*DnaE intein. The gene of the C-terminal SH3 domain was amplified from pAT044 with the two oligonucleotides (#SK199 and #SK200) and cloned into pSKBAD2 using *Kpn*I and *Hin*dIII sites (pHYBAD2-03) (Table S1). The plasmid pSARSF03, which codes for H$_6$-*Npu*DnaE-Int$_{C36}$-cSH3, was constructed by subcloning the genes of Int$_{C36}$ and cSH3 into pHYRSF1-2 by *Nde*I and *Hin*dIII sites. The plasmids (pMMRSF17 for nSH3 and pMMRSF1-16 for cSH3) coding for the SH3 domains fused to the newly designed split *Npu*DnaE intein were previously described [34].

Expression and purification of new split inteins

His-tagged DnaE-Int$_N$s fused with GB1, Smt3, or SH3 domains were purified using His-Trap columns (GE Healthcare) under native condition. The DnaE-Int$_C$s without a His-tag fused with GB1 were purified with IgG sepharose (GE Healthcare) according to the manufacturer's protocol. The eluted fractions of DnaE-Int$_C$s were dialyzed against 10 mM Tris, 500 mM NaCl, 1 mM EDTA, pH 7.0 prior to protein ligation.

In vivo protein ligation

Each pair of the two plasmids encoding N- or C-terminal precursor proteins was transformed into *E.coli* ER2566 (New England Biolabs) for protein expression. The cells bearing these two plasmids were grown in 25 ml LB medium supplemented with 100 μg/ml ampicillin and 25 μg/ml kanamycin. The plasmid containing DnaE-Int$_C$ was first induced for 0.5 hours at a final concentration of 0.04% arabinose when the cell density reached OD$_{600}$ = 0.5–0.8, followed by an additional induction of the N-terminal part with addition of a final concentration of 1 mM isopropyl-β-D-thiogalactoside (IPTG). Expression was carried out for another 4–5.5 hours. The cells were spun down at 4,500×*g* for 10 min and stored at −20°C for further purification. The harvested cells were lysed by ultrasonication in lysis buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, pH 8.0). The cell debris was removed from the protein solution by centrifugation for 15 min at 18,000×*g*. The entire amount of the supernatant was loaded on a Ni-NTA spin column (Qiagen) equilibrated with lysis buffer and centrifuged for 2 min at 700×*g*. The column was washed twice with 600 μl washing buffer (50 mM sodium phosphate, 300 mM NaCl, 30 mM imidazole, pH 8.0). The bound protein was eluted from the spin column by washing twice with 200 μl elution buffer (50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 8.0).

In vitro protein ligation

Equal amounts of the two precursor fragments (final concentrations of 15 μM) were mixed in the presence of final concentrations of 1 mM EDTA and either 50 mM DTT (dithiothreitol), 20 mM MESNA (2-mercaptoethane sulfonic acid), or 0.5 mM TCEP (tris(2-carboxyethyl)phosphine). The reactions were incubated at 25°C with shaking. The samples for SDS-PAGE analysis were typically taken at 0 min, 3 min, 10 min, 30 min, 1 hour, 3 hours, and 24 hours after mixing. The reactions were stopped by adding an equal amount of 1× SDS sample buffer containing 2-mercaptoethanol and stored at −20°C for over night. The samples were loaded on 18% SDS polyacrylamide gels after the incubation at 95°C for 5 min. The ligation yields were estimated from the intensities of the bands in the SDS-gels colored with Coomassie brilliant blue (PhastGel Blue R, GE Healthcare) by quantifying the scanned gels with ImageJ (NIH). The amounts of proteins were calculated with the assumption that the staining dye binds to the proteins equally. The errors were estimated by at least three independent reactions.

Supporting Information

Table S1 List of the used oligonucleotides
Found at: doi:10.1371/journal.pone.0005185.s001 (0.06 MB DOC)

Figure S1 The summary of the constructs for the newly engineered split SspDnaE inteins
Found at: doi:10.1371/journal.pone.0005185.s002 (0.01 MB PDF)

Figure S2 The mass spectrum of the elution fraction from In vivo ligation of GB1 and CBD by SspDnaE intein.
Found at: doi:10.1371/journal.pone.0005185.s003 (0.17 MB PDF)

Figure S3 The mass spectrum of the ligated product, H6-GB1-GB1 by the newly engineered NpuDnaE intein.
Found at: doi:10.1371/journal.pone.0005185.s004 (0.05 MB PDF)

Figure S4 The mass spectra of the ligated and cleaved products from the ligation of nSH3 and GB1 by NpuDnaE intein.
Found at: doi:10.1371/journal.pone.0005185.s005 (0.34 MB PDF)

Acknowledgments

The authors thank Pui-Hang Tam and Tanel Mahlakõiv for their contribution at the early stage of the project and Markku Hakala for his help in the protein production.

Author Contributions

Conceived and designed the experiments: ASA HI. Performed the experiments: ASA SZ EB HI. Analyzed the data: ASA HI. Wrote the paper: ASA SZ HI.

References

1. Hirata R, Ohsumi Y, Nakano A, Kawasaki H, Suzuki K, et al. (1990) Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. J Biol Chem 265: 6726–6733.
2. Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, et al. (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science 250: 651–657.
3. Paulus H (2000) Protein splicing and related forms of protein autoprocessing. Ann Rev Biochem 69: 447–496.
4. Mills KV, Lew BM, Jiang S, Paulus H (1998) Protein splicing in trans by purified N- and C-terminal fragments of the Mycobacterium tuberculosis RecA intein. Proc Natl Acad Sci U.S.A. 95: 3543–3548.
5. Southworth MW, Adam E, Panne D, Byer R, Kautz R, et al. (1998) Control of protein splicing by intein fragment reassembly. EMBO J 17: 918–926.
6. Wu H, Hu Z, Liu XQ (1998) Protein *trans*-splicing by a split intein encoded in a split DnaE gene of *Synechocystis sp.* PCC6803. Proc Natl Acad Sci U.S.A. 95: 9226–9231.
7. Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, et al. (1998) Segmental isotope labeling for protein NMR using peptide splicing. J Am Chem Soc 120: 5591–5592.
8. Züger S, Iwai H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat Biotechnol 23: 736–740.
9. Iwai H, Lingel A, Plückthun A (2001) Cyclic green fluorescent protein produced *in vivo* using an artificially split PI-*Pfu*I intein from *Pyrococcus furiosus*. J Biol Chem 276: 16548–16554.
10. Williams NK, Prosselkov P, Liepinsh E, Line I, Sharipo A, et al. (2002) *In vivo* protein cyclization promoted by a circularly permuted *Synechocystis* sp. PCC6803 DnaB mini-intein. J Biol Chem 277: 7790–7798.
11. Evans TC, Martin D, Kolly R, Panne D, Sun L, et al. (2000) Protein *trans*-splicing and cyclization by a naturally split intein from the *dnaE* gene of *Synechocystis* species PCC6803. J Biol Chem 275: 9091–9094.
12. Ludwig C, Pfeiff M, Linne U, Mootz HD (2006) Ligation of a synthetic peptide to the N terminus of a recombinant protein using semisynthetic protein *trans*-splicing. Angew Chem Int Ed 45: 5218–5221.
13. Mootz HD, Blum ES, Tyszkiewicz AB, Muir TW (2003) Conditional protein splicing: a new tool to control protein structure and function *in vitro* and *in vivo*. J Am Chem Soc 125: 10561–10569.
14. Schwartz EC, Saez L, Young MW, Muir TW (2007) Post-translational enzyme activation in an animal via optimized conditional protein splicing. Nat Chem Biol 3: 50–54.
15. Buskirk AR, Ong YC, Gartner ZJ, Liu DR (2004) Directed evolution of ligand dependence: small-molecule-activated protein splicing. Proc Natl Acad Sci U.S.A. 101: 10505–10510.
16. Xu MQ, Evans TC Jr (2004) Recent advances in protein splicing: manipulating proteins *in vitro* and *in vivo*. Curr Opin Biotechnol 16: 440–446.
17. Derbyshire V, Wood DW, Wu W, Dansereau JT, Dalgaard JZ, et al. (1997) Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution. Proc Natl Acad Sci U.S.A. 94: 11466–11471.
18. Wu H, Xu MQ, Liu XQ (1998) Protein trans-splicing and functional mini-inteins of a cyanobacterial dnaB intein. Biochim Biophys Acta 1387: 422–432.
19. Hiraga K, Derbyshire V, Dansereau JT, Van Roey P, Belfort M (2005) Minimization and stabilization of the *Mycobacterium tuberculosis* recA intein. J Mol Biol 354: 916–926.
20. Dassa B, Amitai G, Caspi J, Schueler-Furman O, Pietrokovski S (2007) Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. Biochemistry 46: 322–330.
21. Kurpiers T, Mootz HD (2007) Regioselective cysteine bioconjugation by appending a labeled cystein tag to a protein by using protein splicing in trans. Angew Chem Int Ed Engl 46: 5234–5237.
22. Sun W, Yang J, Liu XQ (2004) Synthetic two-piece and three-piece split inteins for protein trans-splicing. J Biol Chem 279: 35281–35286.
23. Otomo T, Teruya K, Uegaki K, Yamazaki T, Kyogoku Y (1999) Improved segmental isotope labeling of proteins and application to a larger protein. J Biomol NMR 14: 105–114.
24. Iwai H, Züger S, Jin J, Tam PH (2006) Highly efficient protein *trans*-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. FEBS Lett 580: 1853–1858.
25. Oeemig JS, Aranko AS, Djupsjöbacka J, Heinämäki K, Iwaï H (2009) Solution structure of DnaE intein from *Nostoc punctiforme*: Structural basis for the design of a new split intein suitable for site-specific chemical modification. 10.1016/j.febslet.2009.03.058.
26. Heinämäki K, Oeemig JS, Pääkkonen K, Djupsjöbacka J, Iwaï H (2008) NMR resonance assignment of DnaE intein from *Nostoc punctiforme*. Biomol NMR assign; in press. DOI: 10.1007/s12104-008-9137-1.
27. Reichman CT, Mayer BJ, Khawer S, Hanafusa H (1992) The product of the cellular crk gene consists primarily of SH2 and SH3 regions. Cell Growth Differ 3: 451–460.
28. Martin DD, Xu MQ, Evans TC Jr (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis sp.* PCC6803. Biochemistry 40: 1393–1402.
29. Perler FB, Xu MQ, Paulus H (1997) Protein splicing and autoproteolysis mechanisms. Curr Opin Chem Biol 1: 292–299.
30. Fersht A (1999) Structure and mechanism in protein science, 2nd Ed. New York: W. H. Freeman and Company.
31. Mossessova E, Lima CD (2000) Ulp1-SUMO structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. Mol Cell 5: 865–876.
32. Iwai H, Aranko AS, Djupsjöbacka J (2008) Protein ligation using protein trans-splicing. J Pept Sci 14: Suppl. 183.
33. Forrer P, Jaussi R (1998) High-level expression of soluble heterologous proteins in the cytoplasm of Escherichia coli by fusion to the bacteriophage lambda head protein D. Gene 224: 45–52.
34. Muona M, Aranko AS, Iwai H (2008) Segmental isotopic labelling of a multi-domain protein by protein ligation using protein trans-splicing. ChemBioChem 9: 2958–2961.
35. Sun P, Ye S, Ferrandon S, Evans TC, Xu MQ, et al. (2005) Crystal structures of an intein from the split *dnaE* gene of *Synechocystis* sp PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. J Mol Biol 353: 1093–1105.
36. Ding Y, Xu MQ, Ghosh I, Chen X, Ferrandon S, et al. (2003) Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. J Biol Chem 278: 39133–39142.

Supplementary Table: List of the used oligonucleotides

Plasmid	Gene	oligonucleotides
pSZBAD09PG	*Ssp*DnaE-Int$_{C3}$	#SZ007:TGAATTTCATATGGCCAACTGTTTTAACAAAGG #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD10PG	*Ssp*DnaE-Int$_{C6}$	#SK038: AATCATATGGCTATCGCCGCCAACTG #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD08PG	*Ssp*DnaE-Int$_{C9}$	#SZ006: TGGATTTCATATGGCTAATGGTGCTATCG #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD07PG,	*Ssp*DnaE-Int$_{C16}$	#SZ005: TGACTTTCATATGCAAGACCATAATTTTCTGC #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD06PG	*Ssp*DnaE-Int$_{C23}$	#SZ004: TGAATTGCATATGATCTTTGATATCGGTCTGC #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD05PG	*Ssp*DnaE-Int$_{C30}$	#SZ003:TGAATTTCATATGCGATCCCTGGGTGTGC #T7_term: CGTTTAGAGGCCCCAAGGGG
pSZBAD01PG	*Ssp*DnaE-Int$_{C36}$	#SZ001:TGAATTTCATATGGTTAAAGTTATCG #SZ002:TTGGGTACCTTTGTTAAAACAGTTGGC
pHYBAD44	*Npu*DnaE-Int$_{C15}$	#HK146: TACATATGGACCATAATTTTGCACTC #T7_term: CGTTTAGAGGCCCCAAGGGG
pTMRSF07	nSH3	#HK009: 5’CTTCCTGGTTACCTCCAATC #SK202: 5’TCATATGCAGGAGGAGGCAGAGTATGTG
pHYBAD2-03	cSH3	#SK199: TTGGTACCCTGGGTGGGCCGGAGCCTG #SK200: CGCAAGCTTAGCTGAAGTCCTCATCGGGATTC

Supplementary Figure 1

The summary of the constructs for the newly engineered split *Ssp*DnaE inteins

(a) $\mathrm{Int_N}$ constructs

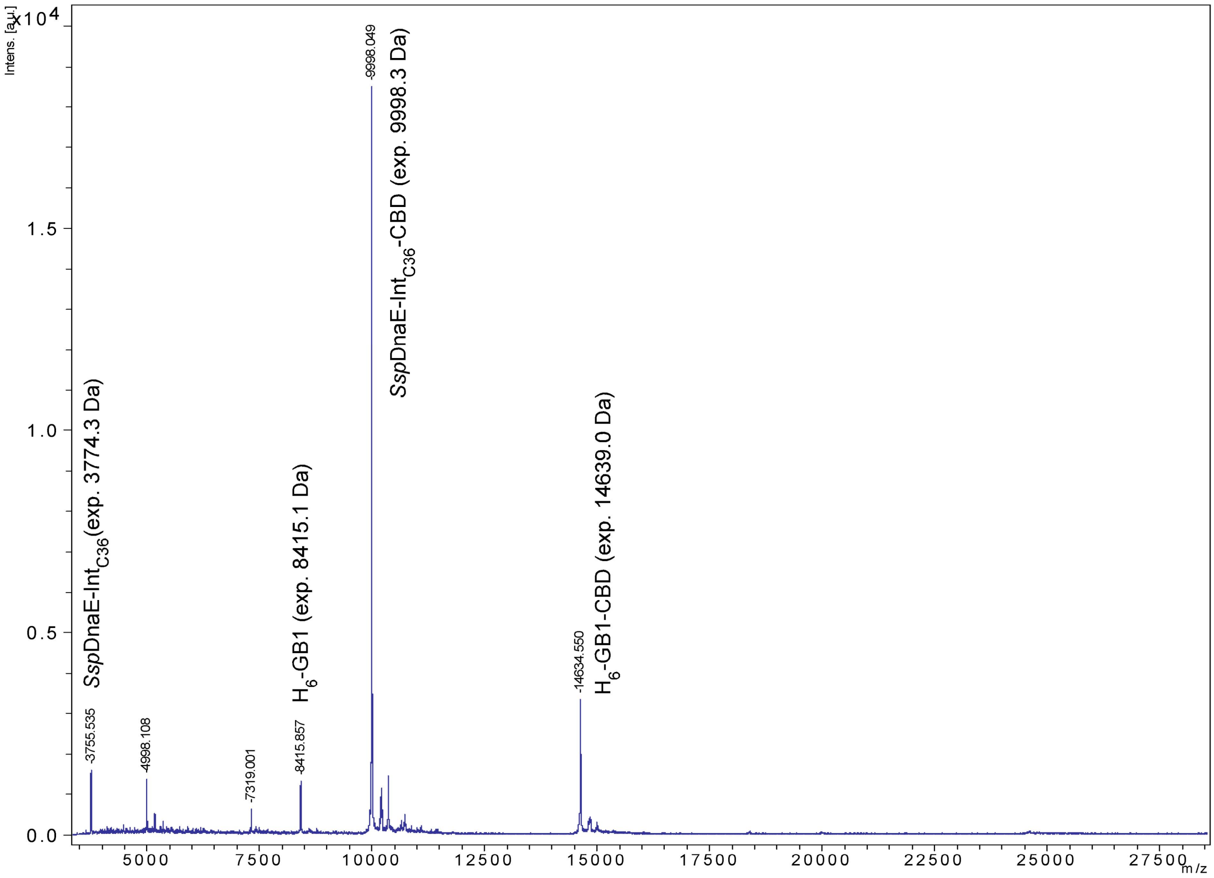
*Ssp*DnaE-$\mathrm{Int_{N157}}$	(pTTDuet19)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIA
*Ssp*DnaE-$\mathrm{Int_{N154}}$	(pSZDuet03)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGRRSLGVQRIFDIGLPQDHNFLLANG
*Ssp*DnaE-$\mathrm{Int_{N151}}$	(pTTDuet18)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGRRSLGVQRIFDIGLPQDHNFLL
*Ssp*DnaE-$\mathrm{Int_{N144}}$	(pTTDuet17)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGRRSLGVQRIFDIGLP
*Ssp*DnaE-$\mathrm{Int_{N137}}$	(pTTDuet15)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGRRSLGVQR
*Ssp*DnaE-$\mathrm{Int_{N130}}$	(pTTDuet05)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	-MVKVIGR
*Ssp*DnaE-$\mathrm{Int_{N123}}$	(pTTDuet02)	$\mathrm{H_6}$-GB1 or $\mathrm{H_6}$	-GS	*Ssp*$\mathrm{DnaE_N}$(1-123)	

(b) $\mathrm{Int_C}$ constructs

*Ssp*DnaE-$\mathrm{Int_{C3}}$	(pSZBAD09)	MANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C6}}$	(pSZBAD10)	MAIAANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C9}}$	(pSZBAD08)	MANGAIAANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C16}}$	(pSZBAD07)	MQDHNFLLANGAIAANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C23}}$	(pSZBAD06)	MIFDIGLPQDHNFLLANGAIAANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C30}}$	(pSZBAD05)	MRSLGVQRIFDIGLPQDHNFLLANGAIAANCFNKGT-	CBD
*Ssp*DnaE-$\mathrm{Int_{C36}}$	(pSZBAD01)	MVKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIAANCFNKGT-	CBD

Supplementary Figure 2

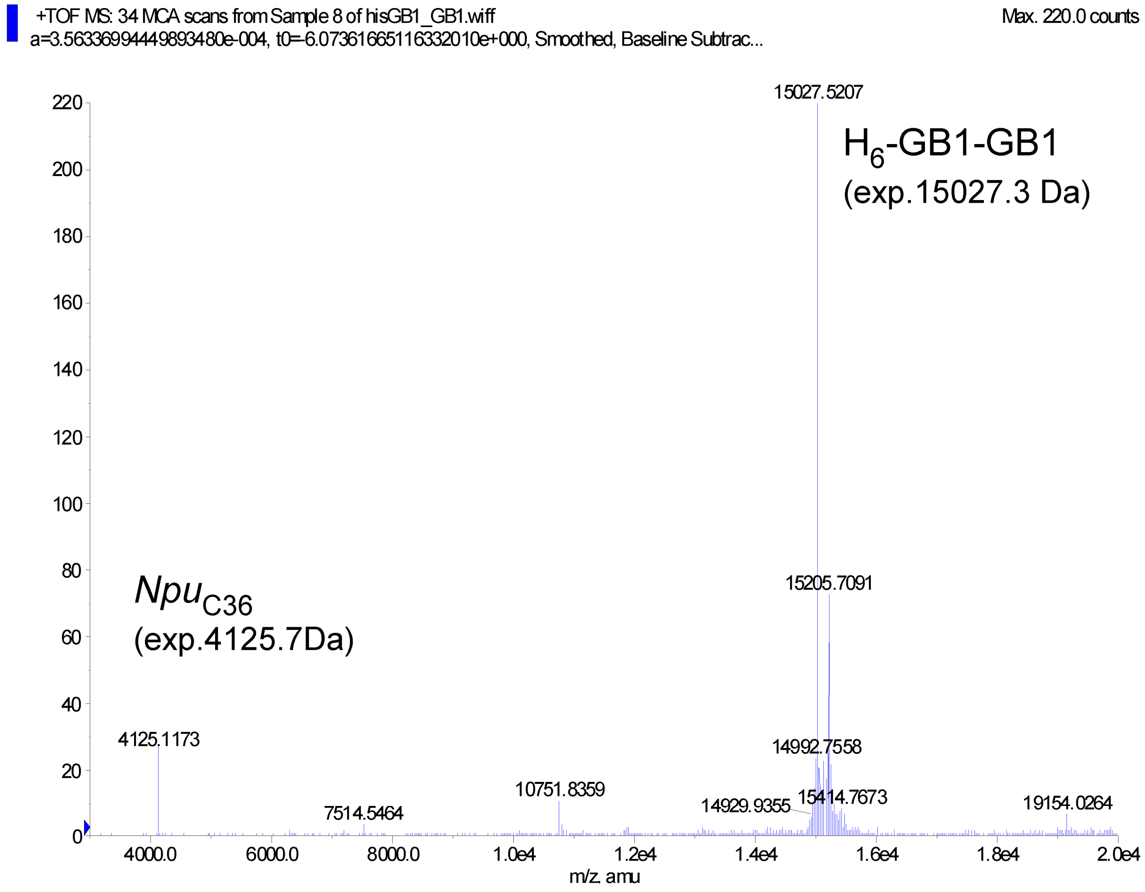
The mass spectrum of the elution fraction from *In vivo* ligation of GB1 and CBD by *Ssp*DnaE intein.



Supplemental Fig.2

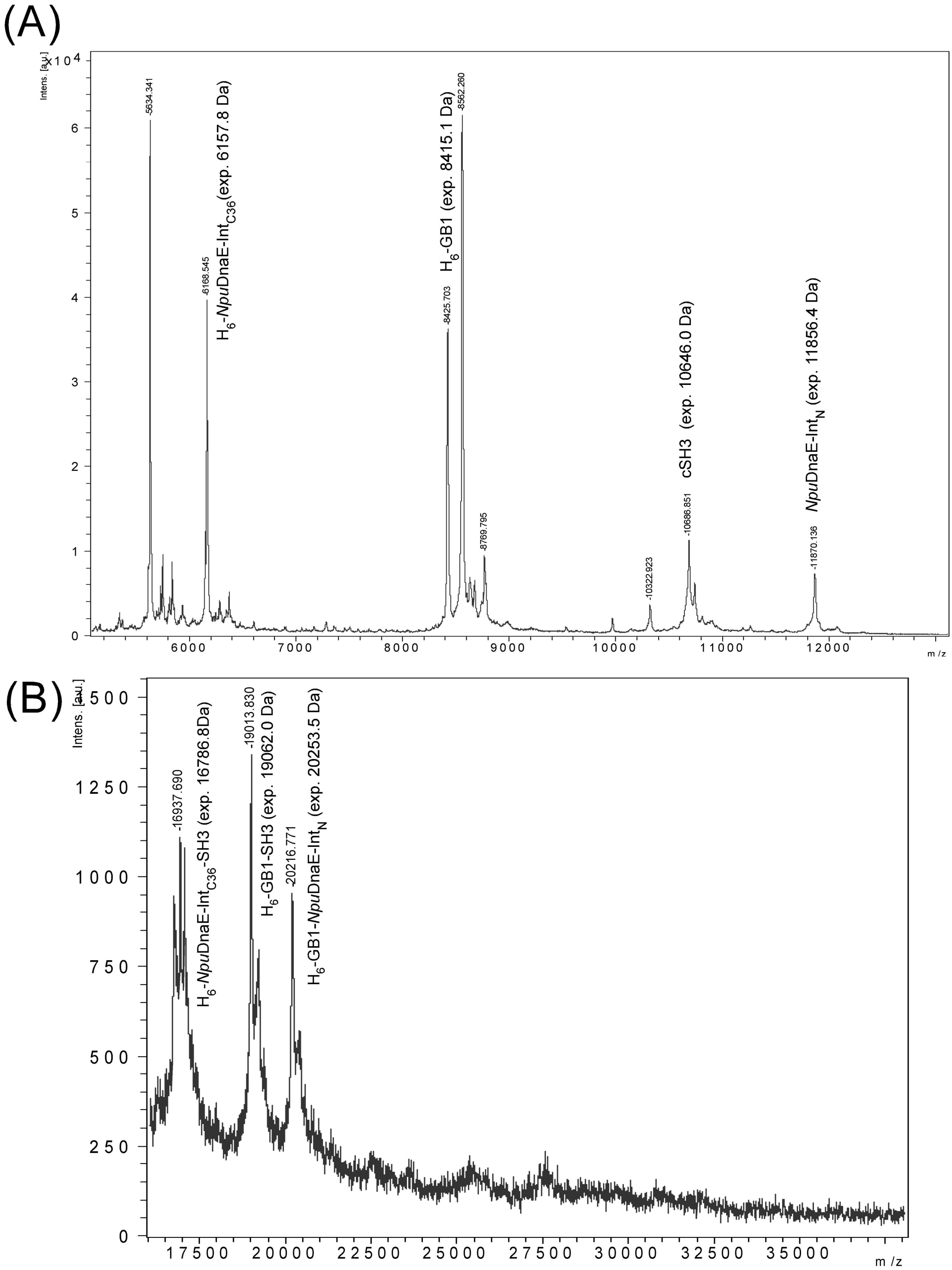
Supplementary Figure 3

The mass spectrum of the ligated product, H_6-GB1-GB1 by the newly engineered *Npu*DnaE intein.



Supplementary Figure 4

The mass spectra of the ligated and cleaved products from the ligation of nSH3 and GB1 by *Npu*DnaE intein.



Supplemental Fig.3

PAPER II

12.11.2010

AALBORG UNIVERSITET

Declaration for Edith Buchinger's share of work in the following articles

Buchinger E, Aachmann FL, Aranko AS, Valla S, Skjåk-Bræk G, Iwaï H, Wimmer R (2010) Use of protein trans-splicing to produce active and segmentally ^{2}H, ^{15}N labeled mannuronan C5-epimerase AlgE4. Protein Sci 19: 1534-1543

The plasmid constructs were done by ASA and EBU. Expression, purification, *in vivo* and *in vitro* ligation test and segmental labelling were performed by EBU. NMR data were recorded and treated by RW, FLA and EBU. Alginate activity studies were carried out by FLA and EBU. EBU and RW wrote the paper with input from the other authors. SV, GSB, HI and RW were active in the formulation of the project, fundraising and in regular discussion of the progress and further strategy for the project.

Buchinger

Edith Buchinger

Finn L. Aachmann

A. Sesilja Aranko

Svein Valla

Svein Valla

Gudmund Skjåk-Bræk

Hideo Iwaï

Reinhard Wimmer

Use of protein trans-splicing to produce active and segmentally ^{2}H, ^{15}N labeled mannuronan C5-epimerase AlgE4

Edith Buchinger,[1,2,3] Finn L. Aachmann,[2] A. Sesilja Aranko,[3] Svein Valla,[2] Gudmund Skjåk-BræK,[2] Hideo Iwaï,[3] and Reinhard Wimmer[1]*

[1]Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Aalborg DK-9000, Denmark

[2]Department of Biotechnology, NOBIPOL, Norwegian University of Science and Technology, Trondheim 7491, Norway

[3]Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki,
Helsinki FIN-00014, Finland

Received 12 April 2010; Revised 29 May 2010; Accepted 2 June 2010
DOI: 10.1002/pro.432
Published online 15 June 2010 proteinscience.org

Abstract: Alginate epimerases are large multidomain proteins capable of epimerising C5 on β-D-mannuronic acid (M) turning it into α-L-guluronic acid (G) in a polymeric alginate. *Azotobacter vinelandii* secretes a family of seven epimerases, each of which is capable of producing alginates with characteristic G distribution patterns. All seven epimerases consist of two types of modules, denoted A and R, in varying numbers. Attempts to study these enzymes with solution-state NMR are hampered by their size—the smallest epimerase, AlgE4, consisting of one A- and one R-module, is 58 kDa, resulting in heavy signal overlap impairing the interpretation of NMR spectra. Thus we obtained segmentally ^{2}H, ^{15}N labeled AlgE4 isotopomeres (A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R) by protein *trans*-splicing using the naturally split intein of *Nostoc punctiforme*. The NMR spectra of native AlgE4 and the ligated versions coincide well proving the conservation of protein structure. The activity of the ligated AlgE4 was verified by two different enzyme activity assays, demonstrating that ligated AlgE4 displays the same catalytic activity as wild-type AlgE4.

Keywords: trans-splicing; inteins; protein ligation; alginate epimerases

Introduction

Production of isotopically enriched proteins that are suitable for structural and functional studies by nuclear magnetic resonance (NMR) is still a major limiting step. Structural studies of biomolecules by NMR have made tremendous progress mainly due to improved recombinant protein expression and ^{13}C, ^{15}N labeling, and/or deuteration. New ways of producing labeled proteins and nucleotides have stimulated the development

Abbreviations: DTT, Dithiothreitol; EDTA, Ethylenediaminetetraacetic acid; EPL, Expressed Protein Ligation; FPLC, Fast Protein Liquid Chromatography; G, α-L- guluronic acid; GdmCl, Guanidium chloride; GSH, reduced Glutathione; HEPES, N-2-Hydroxyethylpiperazine-N′-2-ethanesulfonic acid; HSQC Heteronuclear Single Quantum Coherence; IMAC, Immobilized Metal Affinity Chromatography; Int_C, C-terminal part of intein; Int_N, N-terminal part of intein; IPTG, Isopropyl β-D-1-thiogalactopyranoside; M, β-D-mannuronic acid; ME, Mercaptoethanol; NCL, Native Chemical Ligation; NMR, Nuclear Magnetic Resonance; *Npu*, *Nostoc punctiforme;* PTS, Protein *Trans*-Splicing; SAIL, Stereo-Array Isotope Labelling; SDS-PAGE, Sodium Dodecylsulfate Polyacrylamide Gel Electrophoresis; TCEP, *Tris*(2-carboxyethyl)phosphine; TRIS, *Tris*-(hydroxymethyl)aminomethane; TROSY, Transverse Relaxation Optimized Spectroscopy; *Amp*®, Ampilicin resistance gene; *Kan*®, Kanamycin resistance gene.

Additional Supporting Information may be found in the online version of this article

Grant sponsors: Academy of Finland; Grant numbers: 118385 and 131413; Grant sponsor: Danish Agency for Science, Technology and Innovation, KMB project number 182695/I40 and the Norwegian Research Council; The NMR laboratory at Aalborg University is supported by the Obel and SparNord foundations.

**Correspondence to:* Reinhard Wimmer, Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Aalborg DK-9000, Denmark. E-mail: rw@bio.aau.dk

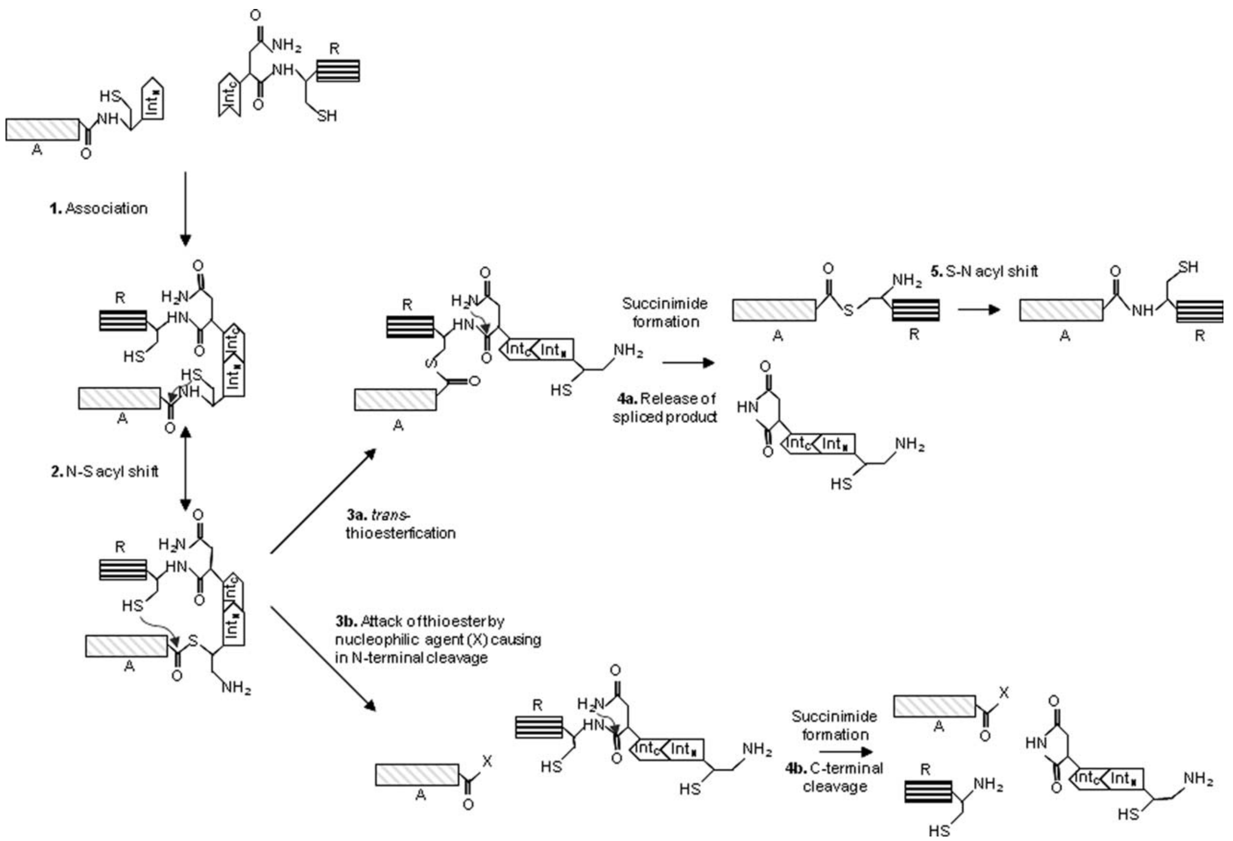


Figure 1. Mechanism of the trans-splicing and trans-cleavage reactions. The reaction steps are: 1. Association of the two intein domains, 2. Attack by Cys1 of intein results in a reactive thioester, 3. either (a) attack of the N-terminal thioester by the first cysteine residue in the C-terminal extein to the intein yields a branched thioester or (b) attack of the thioester by nucleophilic reagents—also water (X) yielding in N-terminal cleavage, 4. Cyclization of the C-terminal asparagine residue results in (a) spliced product or (b) C-terminal cleavage, 5. *S-N* acyl rearrangement restores a native peptide bond.

of novel NMR experiments.[1,2] It has also expanded the scope of NMR with biological macromolecules such as larger proteins by new labeling technology. Whereas NMR structure determination of globular domains of below 20 kDa has increasingly become a routine procedure, resonance assignment of larger proteins can be very time-consuming. A variety of labeling methods such as methyl labeling,[3,4] selective isotopic labelling[5] and stereo-array isotope labeling (SAIL)[6] have been developed and exploited for NMR studies of larger proteins. One of the potentially powerful labeling methods is segmental isotopic labeling, where a segment or a domain in a protein are selectively labeled with stable isotopes.[7,8] Isotopic labeling of a segment or domain in a large protein not only simplifies the spectral complexity but also allows investigation of a region of interest in an intact protein with conventional triple-resonance NMR experiments.[9] Segmental isotopic labeling has been achieved by native chemical ligation (NCL),[10] expressed protein ligation (EPL),[11,12] protein *Trans*-splicing (PTS)[7,9] and enzymatic ligation by sortase and V8 protease.[13,14] One frequently used reaction is called native chemical ligation where an N-terminal cysteine reacts with C-terminal thioester via *trans*-thioesterification and *S-N* acyl shift to a peptide bond. A family of proteins called inteins perform a similar reaction. Inteins are intervening peptide sequences, which excise themselves post-translationally and ligate two flanking N-and C-terminal segments (exteins) via a peptide bond.[15–20] In EPL the intein variant is cleaved by thiol reagents resulting in an extein with a C-terminal thioester. Together with the other extein with an N-terminal cysteine, the two fragments then ligate like in NCL. PTS is an intein mediated method, where the two parts of the split intein associate and perform the peptide ligation (Fig. 1). PTS is a very robust method for ligation and in the last few years *in vivo* and *in vitro* ligation of two segments has been established[21] and recently also three fragments could be ligated.[22] Here, we describe an application of segmental isotopic labeling by protein *trans*-splicing to an enzyme, alginate C5-epimerase AlgE4. This enzyme belongs to a family of seven secreted, structurally related, Ca^{2+} dependent mannuronan C5-epimerases (AlgE1–7) produced by *Azotobacter vinelandii*.[23] These epimerases catalyse epimerisation around C-5 of β-D-mannuronic acid (M) to α-L-guluronic acid (G) at the polymer level in the alginate

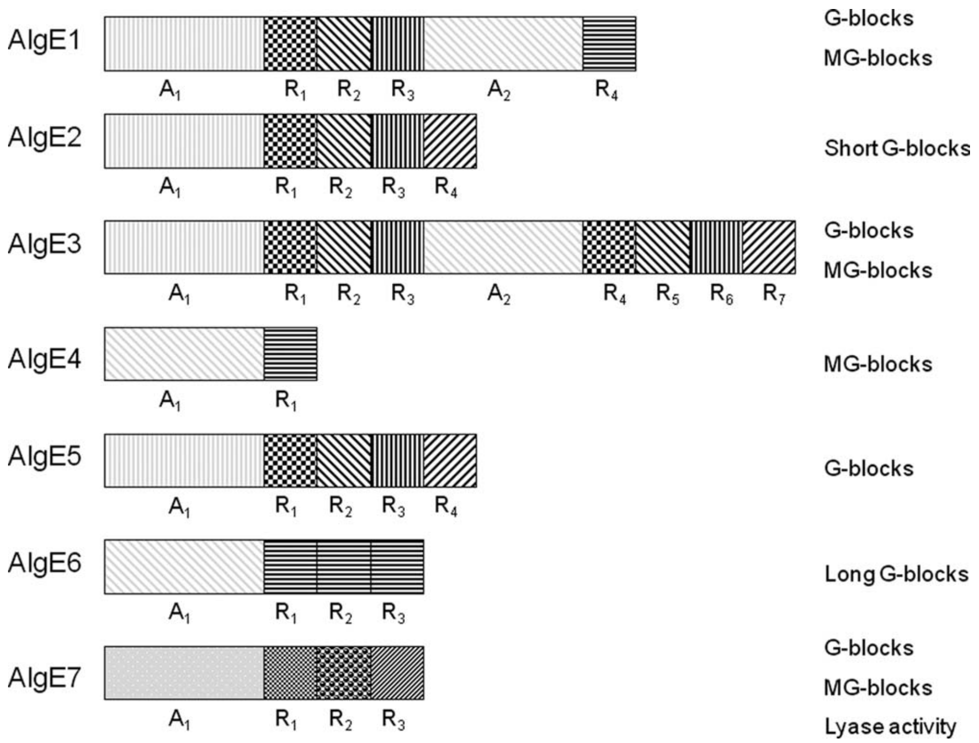


Figure 2. *Azotobacter vinelandii* expresses a family of extracellular alginate epimerases which only consists of two different modules named A- and R-module. Only the A-modules are catalytically active but the R-modules enhance the activity significantly if bound to an A-module. The different members of the family show different epimerisation products, given on the right side of the figure, AlgE7 acts also as a lyase.[23]

polysaccharide. Each AlgE epimerase produces a unique epimerisation pattern of M and G subunits, and AlgE4 produces an alternating structure (MG-blocks).[24]

AlgE1–7 consist of a unique combination of two different modules, designated A and R, and a C-terminal peptide presumably involved in translocation of the protein. The A- and R-modules consist of ~385 and 155 amino acids, respectively. In all the AlgE enzymes there is an A-module located N-terminally, and AlgE1 and AlgE3 have one additional A-module in their sequences. The number of R-modules varies from 1 to 7[25] (Fig. 2). All A-modules and all R-modules share extensive sequence similarities, indicating that the *algE* gene family was generated by a series of gene duplication events.[26,27]

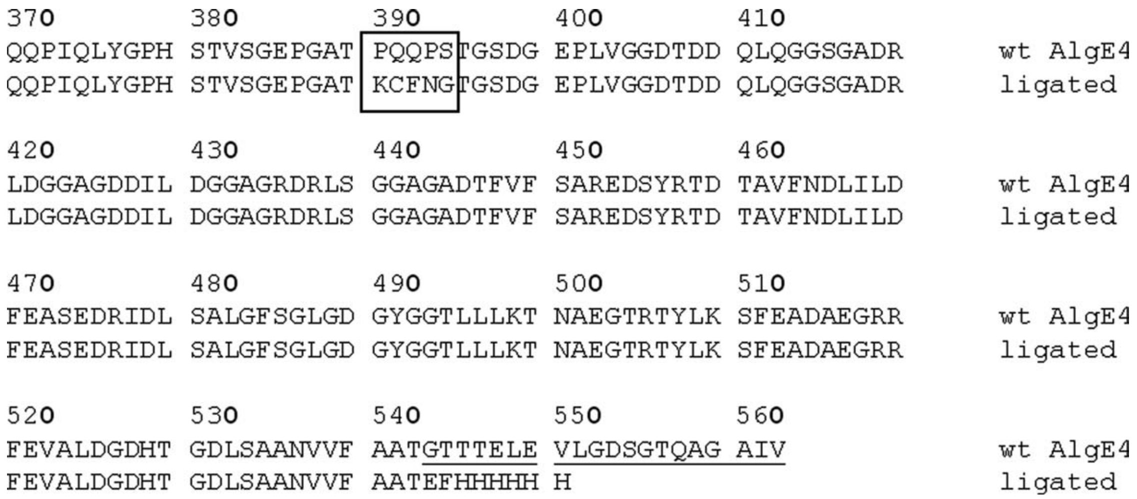
AlgE4 is the smallest of the alginate epimerases and has the composition A-R. The structures of the AlgE4 A- and R-module have been solved separately by X-ray crystallography and NMR, respectively.[28,29] Both proteins showed a highly unusual structure consisting mainly of parallel β-sheets making up a four stranded β-helix and a two stranded β-roll, respectively. The A-modules are catalytically active on their own.[30] The R-modules do not posses any catalytic activity but strongly enhance the reaction rate if at least one R-module is linked to an A-module. Furthermore, the R-module was also shown to interact with alginate oligomers.[29] The strength of the interaction depends on the relative content of M and G.

The aim of this study is to simplify further investigations into the overall structure and substrate binding of active AlgE4 by using segmentally labeled AlgE4, both A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R. The size of AlgE4 (57.7 kDa) makes it necessary to fully deuterate the NMR observable domain. We have chosen to use the naturally split intein of *Nostoc punctiforme* (*N. punctiforme*) for PTS as it shows high tolerance of sequence variations at the splicing junctions, high splicing activity with foreign exteins, and a high solubility.[31,32]

Results

Cloning and production

To obtain a segmentally labeled AlgE4, the epimerase was divided into two parts within the intermediate region connecting the A- and the R-module. The A-module was cloned upstream of the N-terminal part of the naturally split intein *Npu*DnaE intein (Int$_N$), while the R-module was cloned between the C-terminal *Npu*DnaE intein (Int$_C$) and a His-tag for purification. The short amino acid sequence KCFNG around the splicing site was used to obtain optimal splicing. This gives rise to a slightly different amino acid sequence of the produced AlgE4. In addition, the C-terminal 20 residues were omitted, as they are



```
370        380        390        400        410
QQPIQLYGPH STVSGEPGAT PQQPSTGSDG EPLVGGDTDD QLQGGSGADR      wt AlgE4
QQPIQLYGPH STVSGEPGAT KCFNGTGSDG EPLVGGDTDD QLQGGSGADR      ligated

420        430        440        450        460
LDGGAGDDIL DGGAGRDRLS GGAGADTFVF SAREDSYRTD TAVFNDLILD      wt AlgE4
LDGGAGDDIL DGGAGRDRLS GGAGADTFVF SAREDSYRTD TAVFNDLILD      ligated

470        480        490        500        510
FEASEDRIDL SALGFSGLGD GYGGTLLLKT NAEGTRTYLK SFEADAEGRR      wt AlgE4
FEASEDRIDL SALGFSGLGD GYGGTLLLKT NAEGTRTYLK SFEADAEGRR      ligated

520        530        540        550        560
FEVALDGDHT GDLSAANVVF AATGTTTELE VLGDSGTQAG AIV             wt AlgE4
FEVALDGDHT GDLSAANVVF AATEFHHHHH H                          ligated
```

Figure 3. Part of the sequence alignment of native AlgE4 and the segmentally labeled AlgE4 construct. The last 20 amino acids (underlined) of the wild-type AlgE4 are known to be unstructured and were exchanged to a His-Tag for purification. The rectangle shows the splicing site, where few amino acids were exchange to obtain optimal ligation.

known to be unstructured, and a C-terminal His-tag was included. Figure 3 illustrates the differences between the sequences of wild-type and ligated AlgE4.

Expression studies showed that the A-Int$_N$ is soluble if the *Escherichia coli* cells were induced at low temperature (15°C, over night). However, the expression of Int$_C$-R always resulted in insoluble fractions that had to be refolded from 6 *M* Guanidium chloride (GdmCl) before protein ligation. On average, 24 mg (0.45 μmol) of purified A-Int$_N$ was obtained from 1 L growth medium. The yield of the purified Int$_C$-R was 7 mg (0.30 μmol).

Reducing agent/protein ligation

The refolded R-module turned out to be a dimer, where dimerisation occurred by the formation of a disulfide bridge between the single cysteine residues from two Int$_C$-R molecules. This could be seen from comparing SDS-PAGE gels run under reducing and nonreducing conditions (data not shown). As the presence of the free cysteine is required for the ligation to proceed (see Fig. 1), the dimeric Int$_C$-R had to be reduced before PTS. As previously reported, the choice of reducing agents can have a significant effect on the *trans*-splicing efficiency.[33] For the ligation tests, purified A-Int$_N$ and Int$_C$-R were used. To facilitate the purification, a construct in which an N-terminal His-tag was attached to A-Int$_N$ was used, while Int$_C$-R had a C-terminal His-tag. Both proteins were purified by Immobilized Metal Affinity Chromatography (IMAC). For the ligation, concentrations of both parts were adjusted to 0.01 m*M*. Four different thiol-agents were tested (DTT, mercaptoethanol, reduced Glutathione (GSH) and cysteamine), at different temperatures (room temperature, 30 and 37°C), at different final concentrations (2.5, 5, and 10 m*M*) and at two different pH values (7 and 8). For one set of experiments the buffer was exchanged to HEPES buffer. None of the experiments gave a satisfying result; in all cases ligation was very slow. The conditions yielding the highest levels of ligated product were: room temperature, pH 7 and 5 m*M* DTT, mercaptoethanol or cysteamine. For Glutathione as reducing agent, the optimal condition for ligation was room temperature, pH 8 and 5 m*M* concentration. Even in the best cases, the amount of cleaved product exceeded the amount of ligated AlgE4. (Fig. 4A,C). Higher temperature and pH 8 resulted mainly in cleavage of A-Int$_N$. Changing to HEPES buffer had no effect on the ligation, the addition of Ca^{2+}-EDTA completely blocked ligation. The ligation tests also showed that neither A-Int$_N$ nor Int$_C$-R were stable in solution at room temperature. Cleavage of the N-terminal intein part has been described before,[34–36] the instability of Int$_C$-R was probably caused by traces of proteolytic enzymes, as the degradation of Int$_C$-R could be prevented by adding protease inhibitors.

While none of the thiol-based reducing agents yielded a satisfying level of ligation, the *Tris*(2-carboxyethyl)phosphine (TCEP) reducing agent gave high yields of *trans*-splicing without cleavage [Fig. 4(B,D)]. Thus, we continued the work using TCEP as reducing agent to initiate ligation.

In vitro *ligation of A-[^{15}N]-R*

Protein ligation was initiated by adding TCEP to a final concentration of 5 m*M*. *Trans*-splicing was very robust—the ligation could be performed fast in crude cell extract without preliminary purification [Fig. 4(B,D)]. Within one hour, ligation yields up to 90% were reached.

After 2 hours of ligation, the ligated product was purified by IMAC. The eluted fractions contain both the segmentally labeled AlgE4 (A-[^{15}N]-R) and the precursor [^{15}N]-Int$_C$-R, as both compounds carry

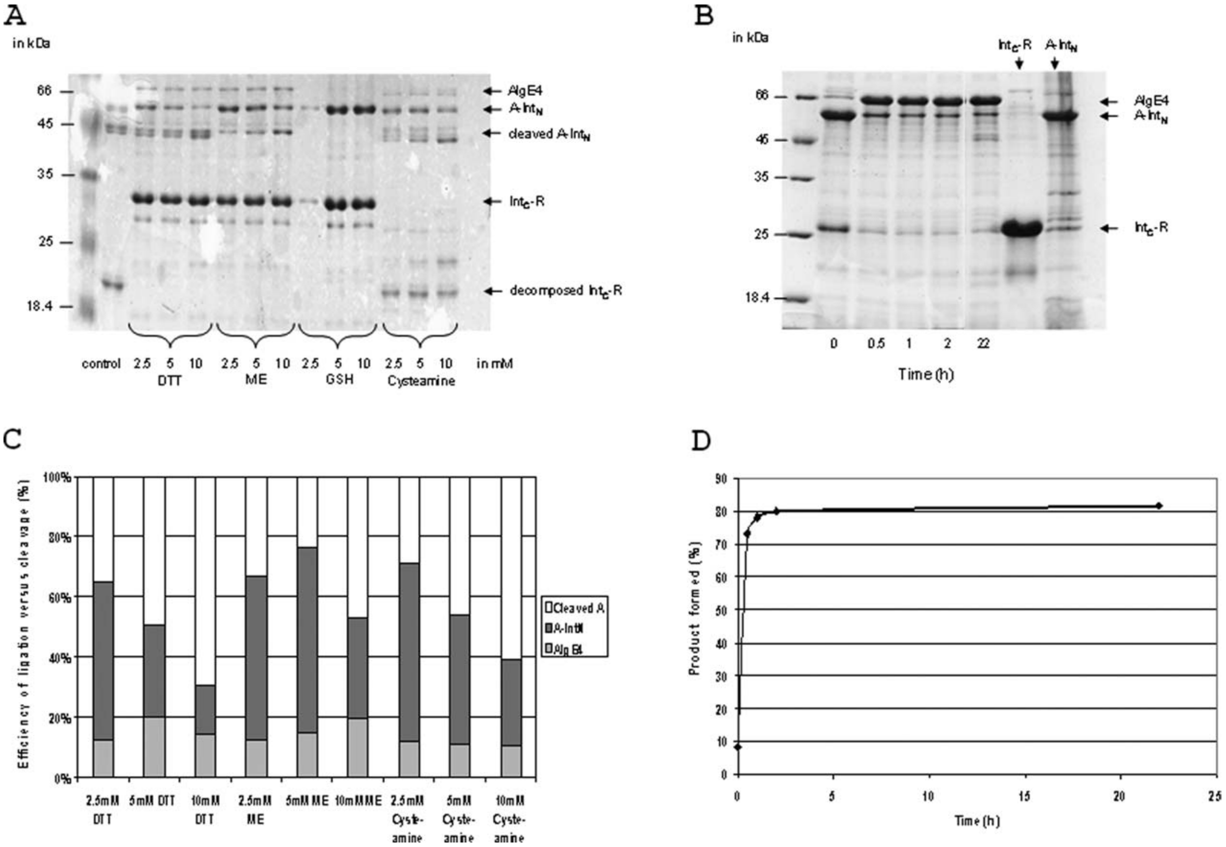


Figure 4. (A) Overview of the ligation with different reducing agents after 2 days ligation at room temperature in Tris buffer at pH 7 without protease inhibitor. A-Int$_N$ or Int$_C$-R was purified before ligation tests. The sample without any reducing agent showed that the A-Int$_N$ or Int$_C$-R is not stable without protease inhibitor. The reducing agents were Dithiothreitol (DTT), Mercaptoethanol (ME), Glutathione reduced (GSH) and Cysteamine. (B) SDS-PAGE of the ligation mixture containing A-Int$_N$ and Int$_C$-R. Both proteins were not purified before ligation although Int$_C$-R had to be refolded. To avoid degradation of Int$_C$-R, protease inhibitor was added and the reaction started after the addition of TCEP. Samples for SDS-PAGE were taken at the given time point. (C) Analysis of the gel lanes from panel A by ImageJ. The sum of A-module containing polypeptides (A-IntN, A (cleaved) and AlgE4) is normalized to 100%. Glutathione is inactive at pH 7 and was therefore omitted. (D) Formation (by ligation) of AlgE4 as a function of time. Analysis of the gel lanes from panel B by ImageJ.

the C-terminal His-tag. Therefore, the eluted fractions had to be further purified by gel-filtration. The amount of purified segmentally labeled A-[^{15}N]-R was ~15 mg (0.25 μmol) from A-Int$_N$ and Int$_C$-R produced in 1 L medium each.

In vitro ligation of A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R

The line width of the A-[^{15}N]-R was broader than that of the R-module alone. For NMR studies, segmentally labeled A-[^{15}N]-R is therefore only of limited use.

To obtain sharper signal of the R-module and due to the size of the A-module two different segmentally deuterated and [^{15}N]-labeled AlgE4 (A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R) were obtained. The protein ligation seemed to proceed slower with deuterated material, independent of which domain was labeled. Therefore the ligation reaction was allowed to continue for 16 h. The yield of the purified A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-A-R was 3 mg (0.05 μmol) from A-Int$_N$ and Int$_C$-R produced in 1 L growth medium each. TROSY spectra of [^{2}H, ^{15}N]-A-R, A-[^{2}H, ^{15}N]-R and [^{2}H, ^{15}N]-AlgE4 were overlaid to confirm the correct fold of segmentally labeled AlgE4. Figure 5 shows the TROSY-NMR spectra of wild-type AlgE4 and the two segmentally labeled variants.

Line width

For NMR studies, it is very important to have narrow lines, as broad lines have a deleterious effect on both resolution and signal-to-noise ratio. The peaks of the segmentally labeled AlgE4 (A-[^{15}N]-R) are expected to be broader than the peaks of the [^{15}N]-R-module alone due to the size difference. This was confirmed by comparing the line widths of 14 peaks from both spectra. The average line width in the HSQC-spectrum of segmentally labeled AlgE4 (A-[^{15}N]-R) was 17.35 ± 1.7 Hz which is significantly broader than the average line width of 10.1 Hz ± 0.9 Hz in the HSQC of the R-module alone. Deuteration and the use of TROSY reduced the line width of

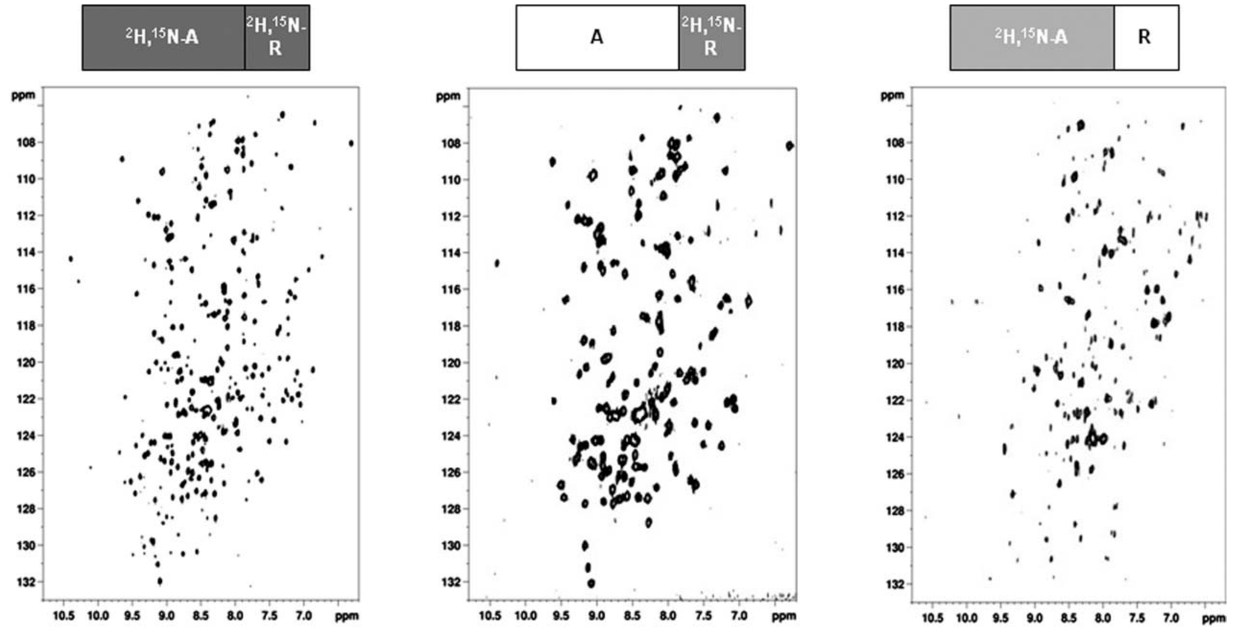


Figure 5. TROSY NMR spectra of [2H, 15N]-AlgE4, A-[2H, 15N]-R and [2H, 15N]-A-R.

the segmentally, labeled AlgE4 (A-[^{2}H, ^{15}N]-R) significantly to 11.2 Hz ± 1.3 Hz.

Activity studies

Two different epimerisation tests verified the activity of the segmentally labeled AlgE4. The NMR-based assay,[23] whose result is shown in Figure 6, certified that A-[^{15}N]-R epimerizes poly-M-alginate to the typical MG-pattern characteristic for AlgE4.[24] As the A-module alone is also active,[30] this is not necessarily a proof for the presence of active AlgE4. However, the epimerisation rate of the A-module alone is significantly lower than that of full-size AlgE4. Hence, a determination of the specific activity of the ligated product is needed to prove that the ligation product is as active as the native protein. By help of a tritium-release assay,[26] it was proven that ligated AlgE4 was able to epimerise poly-M into the expected MG-blocks with comparable specific activity to that of native AlgE4. Wild-type AlgE4 showed a specific activity of 887 ± 140 counts μg^{-1} $hrs^{-1,}$ while *in vitro* ligated AlgE4 showed an activity of 1300 ± 88 counts μg^{-1} hrs^{-1}.

Discussion

Previously, structural studies of AlgE4[28,29] were conducted either on the A- or R-module alone. These gave useful information on the reaction mechanism of the A-module as well as indications of the importance of the R-module for an active epimerase. But no information about the interactions between the two domains, functional aspects and their orientation with respect to each other in the presence and absence of alginate oligomers were gained from these studies. Yet, this information could explain the puzzling effect that the R-module connected to an A-module enhances the reaction rate although it has no activity on its own. Analysis of domain-domain interactions by NMR is difficult with conventional uniform labeling due to the size of AlgE4. Thus segmental isotopic labeling of individual domains in the full-length protein is a very attractive possibility to

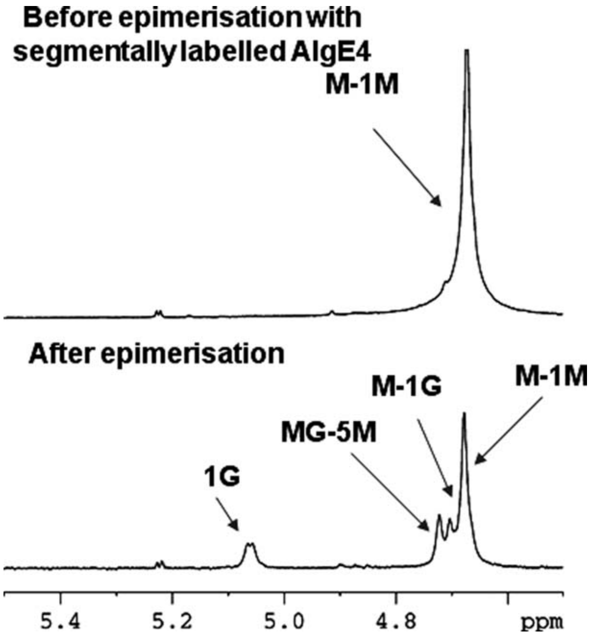


Figure 6. 1D NMR spectra of an alginate sample before (upper) and after (lower) epimerisation with ligated AlgE4. Poly-M-alginate was epimerized by AlgE4 over night, partially hydrolysed, freeze-dried and redissolved in D_2O. The NMR spectrum confirms the typical MG-pattern produced by AlgE4.

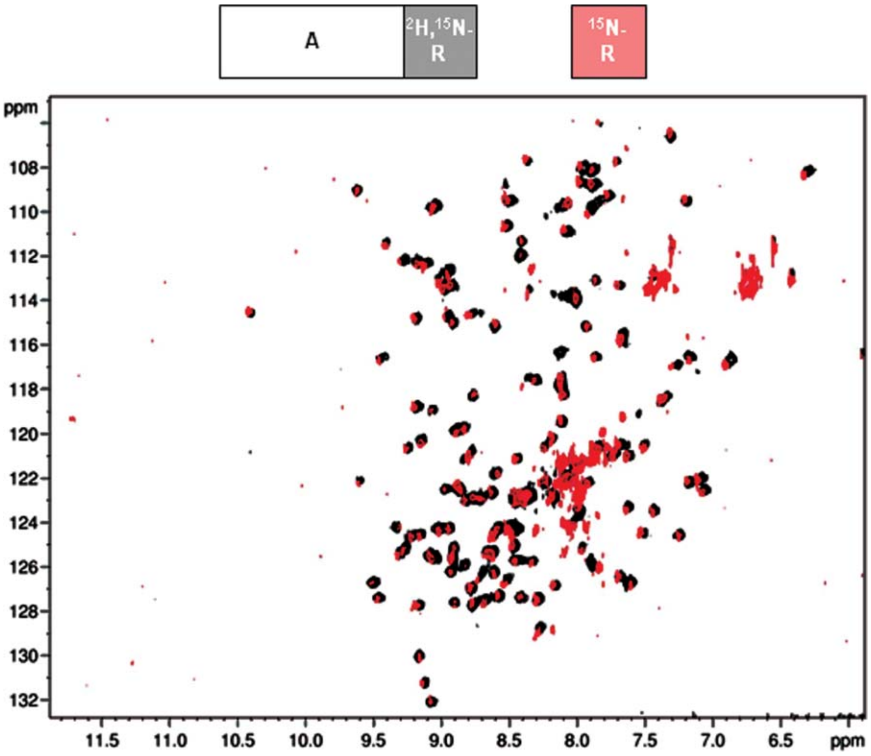


Figure 7. TROSY-NMR of A-[^{2}H, ^{15}N]-R (black) and [^{1}H,^{15}N]-HSQC of [^{15}N]-R-module (red). Extra peaks occurring in the spectrum of the R-module around 8 ppm/122 ppm stem from the unstructured, C-terminal signal sequence that is present in the R-module, but not in the Int$_C$-R construct (Fig. 3).

analyze the structure-function relationship of the individual AlgE4 modules. Figure 7 shows an overlay of NMR spectra of the R-module alone and segmentally labeled AlgE4 (A-[^{2}H, ^{15}N]-R). It demonstrates clearly that the overall structure of the R-module is essentially the same, as most chemical shifts are identical. Only a few residues in the N-terminus of the R-module show changes in chemical shifts. This result suggests that the two domains of AlgE4 do not interact with each other in solution, at least in the absence of substrate. For *trans*-splicing the natural split intein *Npu*DnaE was used as it shows high robustness in *trans*-splicing with non-native exteins, as long as few, here only three amino acids around the splicing site are kept from the native exteins.

In this study, the choice of reducing agent had a significant effect on protein ligation. Without any reducing agent the reaction was blocked as the single cysteins of two Int$_C$-R molecules formed a disulfide bond. Most of the reducing agents tested here were thiol agents. While they reduce disulfide bridges, they also perform a nucleophilic attack on the high energetic thioester occurring after the *N-S* acyl shift as an intermediate of the ligation reaction and thus cause cleavage of the A-module—especially, at higher concentrations of reducing agent (Fig. 1). TCEP is a good alternative, as it has the necessary reduction power to reduce the dimer but has no possibility for a nucleophilic attack. The ligation was performed in the crude cell extract without initial purification. TCEP is probably most often the best choice as reducing agent for *trans*-splicing but there are cases where the reducing agent used has no effect on the amount of ligated product.[33]

We also attempted an *in vivo* ligation following a protocol from the literature.[21] However, all attempts at *in vivo* ligation yielded only minute or undetectable amounts of ligated AlgE4 (data not shown). We ascribe this to the lower solubility of Int$_C$-R, its tendency to misfold when produced at higher temperatures and its susceptibility to proteolytic cleavage. The amounts produced would by no means suffice for NMR spectroscopy, however, we could determine the specific activity of the *in vivo* ligated AlgE4 and found it to be 975 ± 126 counts μg^{-1} hrs^{-1}, that is quite close to the specific activity of the native AlgE4.

To get the narrowest line width possible, a segmentally labeled, deuterated AlgE4 (A-[^{2}H, ^{15}N]-R) was produced. It narrowed the line width of full-length AlgE4 signals in a TROSY experiment to a value only slightly bigger than the line width obtained from the R-module alone in a standard HSQC. Additionally, also the complementary segmentally labeled deuterated AlgE4 ([^{2}H, ^{15}N]-A-R) was produced. Alginate epimerisation tests confirm the activity of the ligated AlgE4. Altogether, segmentally labeled AlgE4 is a strong tool for a better understanding of the structure-function relationship of the epimerases.

Materials and Methods

Vectors for in vivo *and* in vitro *protein ligation*

The vectors used for expression of the segmentally labeled proteins have been deposited to the GenBank with accession numbers HM070247 (pSABAD92A), HM070248 (pEBDuet23A) and HM070249 (pEBDuet28A).

pSABAD92A constructs

The plasmid pSABAD92A encodes a fusion protein consisting of the C-terminal fragment of DnaE intein from *Nostoc punctiforme* ($\mathrm{Int_C}$) and the R-module of the alginate epimerase AlgE4 (residue 385–533). The last 20 residue (534–553) of AlgE4 were exchange to a C-terminal His-tag for purification (EFHHHHHH).

The construct pSABAD92A has a ColE1 origin for replication and the arabinose promoter. pSABAD92A has also the ampilicin resistance gene (Amp^{R}). (For detailed cloning steps and the vector maps see Supporting Information).

pEBDuet23A constructs

The plasmid pEBDuet23A encodes a fusion protein consisting of the A-module of AlgE4 (residues 1–379) and the N-terminal part of the DnaE intein from *Nostoc punctiforme* ($\mathrm{Int_N}$). The vector has also the kanamycin resistance gene (Kan^{R}), RSF origin and the expression of the fusion gene is tightly controlled by *T7/lac* promoter.

pEBDuet28A constructs

pEBDuet28A has also the kanamycin resistance gene (Kan^{R}), RSF origin and the *T7/lac* promoter. The fusion protein, consisting of an N-terminal His-Tag for purification, the A-module of AlgE4 and the N-terminal fragment of the DnaE intein from *Nostoc puntiforme,* is expressed after adding IPTG.

Buffers and expression media

1 L LB-medium contained 10 g Tryptone, 5 g Yeast and 5 g NaCl. The pH was adjusted to 7.2 with 4 *M* NaOH and the medium was sterilized by autoclaving.

For 1 L M9-medium 7.2 g $Na_2HPO_4{\cdot}2H_2O$, 3 g KH_2PO_4, 0.5 g NaCl and 1 g NH_4Cl were dissolved in 1 L H_2O. The pH was adjusted to 7.4 and autoclaved. Before the expression 2 mL of 1 *M* $MgSO_4$, 20 mL of trace metal, 5 mL MEM Vitamins 100x (Invitrogen) and 2 g glucose dissolved in 10 mL H_2O was added.

Trace metal for M9-medium. 0.1 g/L $ZnSO_4$, 0.8 g/L $MnSO_4$, 0.5 g/L $FeSO_4$, 0.1 g/L $CuSO_4$, and 1 g/L $CaCl_2$ unlabeled proteins were produced in LB-medium. For production of the ^{15}N-labeled proteins M9-medium was supplemented with 1 g $(^{15}NH_4)_2SO_4$. ^{2}H, ^{15}N labeled protein were expressed in M9-medium prepared with D^2O (99% D) and supplemented with 1 g $(^{15}NH_4)_2SO_4$ and 2 g U-^{2}H-D -glucose.

Lysis buffer contained 20 m*M* HEPES pH 6.9, 800 m*M* NaCl, 10 m*M* $CaCl_2$, 0.1% Triton X. Folding buffer contained 20 m*M* HEPES pH 6.9, 800 m*M* NaCl, 5 m*M* $CaCl_2$. The elution buffer consists of 20 m*M* HEPES pH 6.9, 800 m*M* NaCl, 250 m*M* imidazole, 5 m*M* $CaCl_2$. One-hundred milliliter 2 × SDS buffer contained 10 mL of 1.5 *M* TRIS (pH 6.8), 6 mL 20% SDS, 30 mL glycerol and 1.8 mg bromophenol blue. To 2 mL aliquots 100 μL 1 *M* DTT was added.

Expression of $\mathrm{Int_C}$–R, [^{15}N]-$\mathrm{Int_C}$–R and [^{2}H, ^{15}N]-$\mathrm{Int_C}$–R

The *E.coli* cells with the plasmid pSABAD92A were grown in 1 L LB-medium, 1 L M9-medium supplemented with 1 g $(^{15}NH_4)_2SO_4$ or M9-medium prepared with 99% D_2O and supplemented with 1 g $(^{15}NH_4)_2SO_4$ and 2 g U-^{2}H-D-glucose. 100 μg/mL ampicillin was added. The cells were grown at 37 °C to an OD_{600} of 0.5–0.7 and induced with 0.2% (w/v) arabinose and further incubated for 3 h. The cells were harvested, resuspended in lysis buffer and stored at −20 °C for further purification.

Expression of A–$\mathrm{Int_N}$ and [^{2}H, ^{15}N]-A–$\mathrm{Int_N}$

Cells harboring the plasmid (pEBDuet23A or pEBDuet28A used only for ligation test) for A–$\mathrm{Int_N}$ were grown in 1 L LB-medium or 1 L M9-medium in D_2O supplemented with 1 g $(^{15}NH_4)_2SO_4$ and 2 g U-^{2}H-D-glucose, respectively, containing 50 μg/mL kanamycin at 37°C to an OD_{600} of 0.5–0.7. The cell culture was incubated on ice for 5 min. The protein was induced with 1 m*M* IPTG at 15 °C and incubated over night. The cells were harvested and resuspended in lysis buffer.

Refolding of [^{15}N]-$\mathrm{Int_C}$–R/[^{2}H,^{15}N]-$\mathrm{int_C}$–R and [^{2}H, ^{15}N]-A–$\mathrm{Int_N}$

The cells containing the expressed protein were thawed and lysed by sonication. The pellets were solubilised in 3 mL 20 m*M* HEPES pH 6.9, 6 *M* GdmCl, 800 m*M* NaCl, 5 m*M* $CaCl_2$ at 4 °C. The solution was diluted 10 times in folding buffer and was dialyzed against the same buffer at 4 °C.

In vitro *ligation tests*

For one reaction 100 μL of A-$\mathrm{Int_N}$ with a concentration of 0.1 m*M* were mixed with 100 μL of $\mathrm{Int_C}$-R. The reactions were started by adding reducing agent. The effect of temperature (room temperature, 30 and 37°C), pH (7 and 8), concentration of reducing agents (2.5, 5, and 10 m*M*) as well as different reducing agents (DTT, mercaptoethanol, GSH and cysteamine) were tested. At selected time points 20 μL samples for SDS-PAGE analysis were collected. The ligation was stopped by adding aliquots of 2 × SDS-buffer to the samples, followed by heating at 95°C for 5 min. The samples were loaded on 12% SDS polyacrylamide gels.

In vitro ligation of segmentally labeled AlgE4

Cells containing A-Int$_N$ were sonicated. The crude cell extract was mixed with the refolded R-Int$_C$, where one tablet of protease inhibitor ("complete EDTA-free" from Roche Diagnostics) and TCEP to a final concentration of 5 m*M* had been added to initiate the reaction. After incubation for 2 hours at room temperature, the segmentally labeled AlgE4 was purified from the reaction mixture.

Purification of ligated AlgE4

The ligation solution was loaded onto a 1 mL Ni^{2+} His-Trap^{TM} FF crude (GE Healthcare) column equilibrated in folding buffer. The column was washed with 10 column volumes of this buffer. Ligated AlgE4 and Int$_C$-R were eluted by a linear gradient to the elution buffer. Ligated AlgE4 was then separated from Int$_C$-R by size-exclusion chromatography (Superdex 75 HR 10/30).

NMR measurements

Protein NMR spectra were recorded at 298 K on a BRUKER DRX 600 spectrometer equipped with a 5 mm xyz-gradient TXI (H/C/N) probe using TopSpin 1.3 and a BRUKER Avance 800 MHz spectrometer equipped with 4 channels operating at a field strength of 18.8 T, equipped with a 5 mm TCI Cryoprobe, using TopSpin 2.0. The sample buffer was 10 m*M* HEPES pH 6.9 and 25 m*M* $CaCl_2$.

NMR spectra of alginate samples were recorded at 363 K on a BRUKER DRX 400 spectrometer equipped with a 5 mm z-gradient DUL (H/C) probe.

To reduce the viscosity of the alginate samples for obtaining higher resolution of the NMR analyses, the alginate samples were degraded by mild acid hydrolysis[37] and dialyzed, freeze-dried and redissolved in D_2O (99.9% D) before the NMR measurement.

Measurement of epimerase activity by radioisotope assay and NMR

The two different activity tests were performed as reported.[30] The relative amount of MG after the epimerisation was calculated as described.[38]

Line width

A Lorentzian line shape function was fitted to slices taken through 14 different peaks with the line width as variable parameter.

$$I = \sum_{x} \frac{w^2}{(w^2 + (x-t)^2)}$$

where
w line width ppm
x point in ppm
t middle of the Lorentz curve in ppm.

Acknowledgments

The expert technical assistance by Hanne Krone Nielsen is gratefully acknowledged. The 800 MHz spectra were obtained using the Bruker Avance spectrometer of the Danish Instrument Center for NMR Spectroscopy of Biological Macromolecules. Furthermore, we thank Wenche I. Strand and Helga Ertesvåg for kind assistance in the epimerzation assays.

References

1. Bermel W, Bertini I, Felli IC, Kümmerle R, Pierattelli R (2006) Novel 13C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. J Magn Reson 178:56–64.
2. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94:12366–12371.
3. Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE (1996) Selective methyl group protonation of perdeuterated proteins. J Mol Biol 263:627–636.
4. Neri D, Szyperski T, Otting G, Senn H, Wüthrich K (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. Biochemistry 28:7510–7516.
5. Kainosho M, Tsuji T (1982) Assignment of the three methionyl carbonyl carbon resonances in *Streptomyces subtilisin* inhibitor by a carbon-13 and nitrogen-15 double-labeling technique. A new strategy for structural studies of proteins in solution. Biochemistry 21: 6273–6279.
6. Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440: 52–57.
7. Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, Ito N, Ishino Y, Nakamura H (1998) Segmental isotope labeling for protein NMR using peptide splicing. J Am Chem Soc 120:5591–5592.
8. Xu R, Ayers B, Cowburn D, Muir TW (1999) Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. Proc Natl Acad Sci USA 96:388–393.
9. Otomo T, Teruya K, Uegaki K, Yamazaki T, Kyogoku Y (1999) Improved segmental isotope labeling of proteins and application to a larger protein. J Biomol NMR 14: 105–114.
10. Dawson PE, Muir TW, Clark-Lewis I, Kent SB (1994) Synthesis of proteins by native chemical ligation. Science 266:776–779.
11. Severinov K, Muir TW (1998) Expressed Protein Ligation, a novel method for studying protein-protein interactions in transcription. J Biol Chem 273:16205–16209.
12. Zhao W, Zhang Y, Cui C, Li Q, Wang J (2008) An efficient on-column expressed protein ligation strategy: application to segmental triple labeling of human apolipoprotein E3. Protein Sci 17:736–747.
13. Mao H, Hart SA, Schink A, Pollok BA (2004) Sortase-mediated protein ligation: A new method for protein engineering. J Am Chem Soc 126:2670–2671.
14. Machova Z, Eggelkraut-Gottanka Rv, Wehofsky N, Bordusa F, Beck-Sickinger AG (2003) Expressed enzymatic ligation for the semisynthesis of chemically modified proteins. Angew Chem Int Edit 42:4916–4918.
15. Mootz HD (2009) Split inteins as versatile tools for protein semisynthesis. Chembiochem 10:2579–2589.

16. David R, Richter MP, Beck-Sickinger AG (2004) Expressed protein ligation. Method and applications. Eur J Biochem 271:663–677.
17. Muralidharan V, Muir TW (2006) Protein ligation: an enabling technology for the biophysical analysis of proteins. Nat Methods 3:429–438.
18. Iwaï H, Züger S (2007) Protein ligation: applications in NMR studies of proteins. Biotechnol Genet Eng Rev 24: 129–145.
19. Paulus H (2000) Protein splicing and related forms of protein autoprocessing. Annu Rev Biochem 69:447–496.
20. Perler FB (2002) InBase: the intein database. Nucl Acids Res 30:383–384.
21. Züger S, Iwaï H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat Biotechnol 23: 736–740.
22. Busche AE, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dötsch V, Iwaï H (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. Angew Chem Int Ed Engl 48:6128–6131.
23. Ertesvåg H, Høidal HK, Schjerven H, Svanem BI, Valla S (1999) Mannuronan C-5-epimerases and their application for in vitro and in vivo design of new alginates useful in biotechnology. Metab Eng 1:262–269.
24. Høidal HK, Ertesvåg H, Skjåk-Bræk G, Stokke BT, Valla S (1999) The recombinant *Azotobacter vinelandii* mannuronan C-5-epimerase AlgE4 epimerizes alginate by a nonrandom attack mechanism. J Biol Chem 274: 12316–12322.
25. Ertesvåg H, Doseth B, Larsen B, Skjåk-Bræk G, Valla S (1994) Cloning and expression of an *Azotobacter vinelandii* mannuronan C-5-epimerase gene. J Bacteriol 176:2846–2853.
26. Svanem BI, Skjåk-Bræk G, Ertesvåg H, Valla S (1999) Cloning and expression of three new *Azotobacter vinelandii* genes closely related to a previously described gene family encoding mannuronan C-5-epimerases. J Bacteriol 181:68–77.
27. Ertesvåg H, Høidal HK, Hals IK, Rian A, Doseth B, Valla S (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in *Azotobacter vinelandii*. Mol Microbiol 16:719–731.
28. Rozeboom HJ, Bjerkan TM, Kalk KH, Ertesvåg H, Holtan S, Aachmann FL, Valla S, Dijkstra BW (2008) Structural and mutational characterization of the catalytic A-module of the mannuronan C-5-epimerase AlgE4 from *Azotobacter vinelandii*. J Biol Chem 283: 23819–23828.
29. Aachmann FL, Svanem BI, Güntert P, Petersen SB, Valla S, Wimmer R (2006) NMR structure of the R-module: a parallel beta-roll subunit from an *Azotobacter vinelandii* mannuronan C-5 epimerase. J Biol Chem 281:7350–7356.
30. Ertesvåg H, Valla S (1999)The A modules of the *Azotobacter vinelandii* mannuronan-C-5-epimerase AlgE1 are sufficient for both epimerization and binding of Ca2+. J Bacteriol 181:3033–3038.
31. Iwaï H, Züger S, Jin J, Tam PH (2006) Highly efficient protein trans-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. FEBS Lett 580:1853–1858.
32. Zettler J, Schutz V, Mootz HD (2009)The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. FEBS Lett 583: 909–914.
33. Aranko AS, Züger S, Buchinger E, Iwaï H (2009) In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. PLoS One 4: e5185.
34. Martin DD, Xu MQ, Evans TC (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. Biochemistry 40:1393–1402.
35. Nichols NM, Benner JS, Martin DD, Evans TC (2003) Zinc ion effects on individual Ssp DnaE intein splicing steps: regulating pathway progression. Biochemistry 42:5301–5311.
36. Nichols NM, Evans TC (2004) Mutational analysis of protein splicing, cleavage, and self-association reactions mediated by the naturally split Ssp DnaE intein. Biochemistry 43:10265–10276.
37. Ertesvåg H, Skjåk-Bræk G, (1999) Modification of alginate using mannuronan C-5-epimerases. In: Bucke C, Ed. Carbohydrate biotechnology protocols. Humana Press: Totowa, NJ, 71–78.
38. Grasdalen H, Larsen B, Smidsrød O (1979) A p.m.r. study of the composition and sequence of uronate residues in alginates. Carbohyd Res 68:23–31.

pSABAD92A

pSKDuet12 was constructed by inversion PCR of pSKDuet1A [1] using oligonucleotides #SK165 (CACCGTAACGGAGACAAAATGTCTAAGCTATG) and #SK160 (TGTCTCCGTTACGGTGTAGGTTTTGG) and carries a fusion protein with an N-terminal His-Tag, the, B1-domain of G-protein (GB1) and the N-terminal fragment (Int$_N$) of the native split *Npu*DnaE intein[2]. The primers #DuetMCS-fw (GGATCTCGACGCTCTCCCT) and #HK010 (CGCATCTCGAGTTCGGCAAATTATCAAC) were used to amplify part of pSKDuet12 with PCR. The PCR product was then ligated into pCDFDuet-1 (plasmid 15917 from Addgene) using the *Nco*I and *Xho*I sites resulting in pTMCDF01. An His-Tag was inserted by ligating the oligonucleotide #SZ008 (TCGACTCATCATCATCATCATCATTAA) and #SZ009 (TCGACTTAATGATGATGATGATGATGA) into pTMCDF01 using newly created *Xho*I site to result in plasmid pTMCDF02. Furthermore, an *Eco*RI site and an EF-linker (Glu-Phe) was created by inversion PCR using oligonucleotides #HK038 (GATAATTTGCCGAACGAATTCCATCATCATCATC) and #HK039 (GATGATGATGATGGAATTCGTT) to end up into pHKCDF22-3. Finally the primers #SK012 (TCCTTACATATGCAGTACAAACTTATC) and #HK122 (CTAAAGCTTAATGATGATGATGATGATG) were used to amplify a part of pHKCDF22-3 and the PCR product was ligated between the *Nde*I and *Hin*dIII sites of pSKBAD2A [1] (plasmid 15335 from Addgene) vector to yield pMHBAD10. To gain the fusion-protein Int$_C$- B1-domain of G-protein (GB1)-His-Tag, a part of pSKBAD2A was amplified with oligonucleotides #SK094 (TAACATATGATCAAAATAGCCACACG) and #HK158 (AGAATTCCGTTACGGTGTAGGTTTTG), and ligated into pMHBAD10 between the *Nde*I and *Eco*RI sites resulting in pMHBAD14. Then GB1 was replaced by the R-module of AlgE4.

To obtain the R-module template, pSKDuet1A was digested with *Nco*I and *Hin*dIII and ligated into pRSF-1b (Novagen) resulting in pHYRSF1. pSABAD28 was constructed by PCR using pSKBAD2A as a template and using oligonucleotides #SK094 (TAACATATGATCAAAATAGCCACACG), #HK036 (CCGCGGGCGTTCGTGCAATTAGAAGCTATGAAGCC), and #HK037 (CAGGTACCGCCAGCCCCGCGGGCGTTCGTGC) and ligating the product into pSKBAD2A between the *Nde*I and *Kpn*I sites. pSABAD28 was digested with *Nde*I and *Hin*dIII and the digestion product was ligated into pHYRSF1 obtaining pSARSF1-28. pSARSF1-66 was constructed by inversion PCR of pSARSF1-28 using oligonucleotides #HK152 (GACGCTGCTACCGCCGAAAAAGTTTTCAAAC) and #HK153 (GTTTGAAAACTTTTTCGGCGGTAGCAGCGTC). Finally pSARSF1-LICI-1 was constructed by ligation independent cloning (LIC) by amplifying the part of pFA1 [3], that contains the gene of the R-module of AlgE4, with oligonucleotides #HK137 (GCACGAACGCCCAAGGAAGCGACGGCGAGCCAC) and #HK138 (ACCGCCAGCCCCTTAGACGATCGCCCCGGCCTG) and inserting the ligation product into pSARSF1-66 using the *Sac*II site. The gene of the R-module of AlgE4 was amplified from pEBBAD30A with oligonucleotides #HK057 (AAGGTACCGGAAGCGACGGCGAGCCAC) and#HK197 (GTCTTCGCCGCGACCGAATTCA) and inserted it into pMHBAD14 between the *Kpn*I and *Eco*RI sites. pEBBAD30A was obtained by ligating the product from *Nco*I and *Hin*dIII digestion of pSARSF1LICI-1f into pSKBAD2A. The resulting plasmid pSABAD92 encodes a fusion protein consisting of the C-terminal 36 residues (Int$_C$) of the *Npu*DnaE intein from *Nostoc punctiforme*¸the AlgE4 R-module and the C-terminal His-Tag with an EF linker (Glu-Phe).

pEBDuet23A

The DNA fragment encoding the A-module (residues 1-379) of AlgE4 was amplified from pBS32 by the two oligonucleotides # HK54 (CCTACCTGAAAAGTTTCGAGGCGGATGC) and #HK141 (CCGGCGAACCCGGCGCGACAA) and cloned into pSKDuet12 using the *Nco*I and *Ahd*I sites, resulting in the plasmid pEBDuet23A.
pBS32 is a derivative of pTYB4 (NEB) in which a 1.65 kb *Nco*I-*Xma*I DNA fragment corresponding to the full-length AlgE4 from pHH4 [4] was subcloned.

pEBDuet28A:

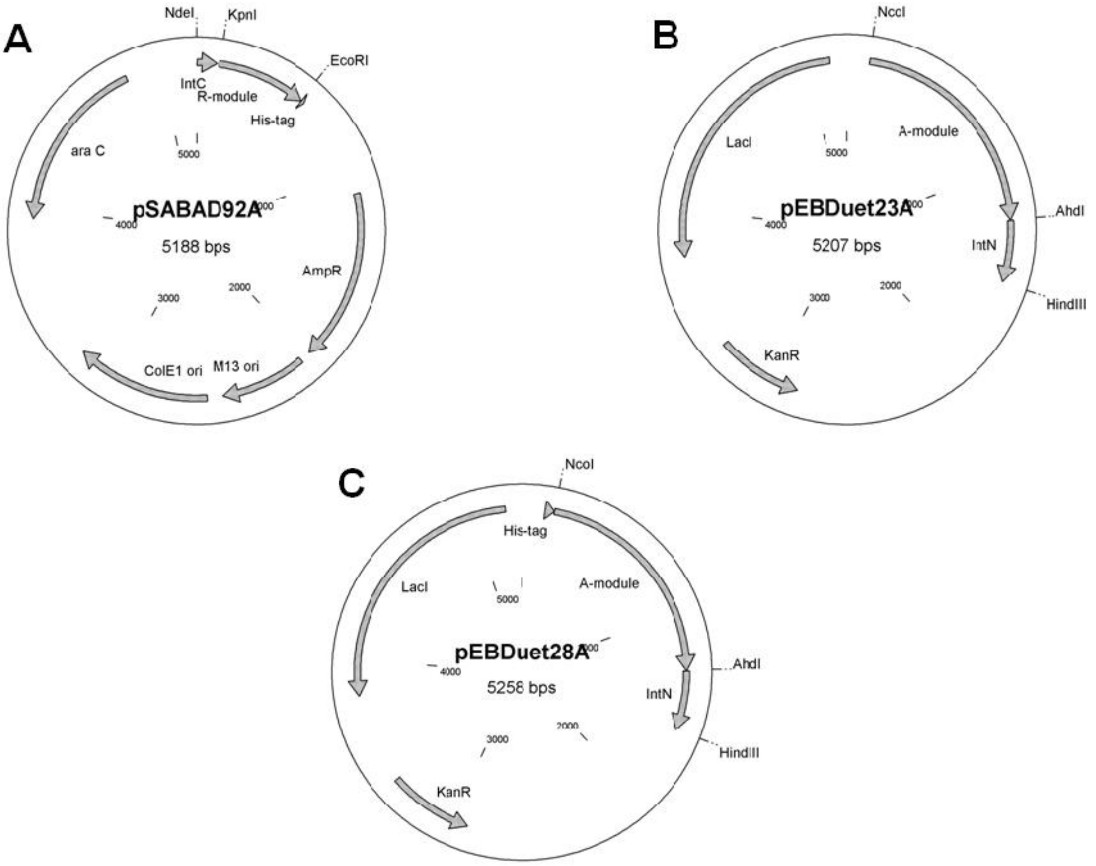
pTMDuet03 was constructed by cloning the N-terminal domain of the ClpX zinc finger from the chromosomal DNA of *E. coli* DH5α using oligonucleotides #HK001 (TAGACCATGGCAGATAAACGCAAAGATG) and #HK002 (TTTCACGATGCGGTGCAAC), ligating the PCR product into pSKDuet12 using the *Nco*I and *Ahd*I sites. The N-terminal His-Tag and TEV-digestion site was created by inversion PCR using oligonucleotides #HK014 (CATGCGGGGTTCTCATCATCATCATCATCATGAGAATTTGTATTTTCAGTCCATG) and #HK015 (ATGGACTGAAAATACAAATTCTCATGATGATGATGATGATGAGAACCCCG).
pEBDuet23A was digested with *Hin*dIII and *Nco*I. The gene encoding the N-terminal precursor A-*Int$_N$* was inserted into pTMDuet03 to yield pEBDuet28A.

Figure S1:
A)
The plasmid pSABAD92A encodes a fusion protein consisting of the C-terminal 36 residues (Int$_C$) of the DnaE intein from *Nostoc punctiforme* and the AlgE4 R-module (residue 385-533) of *Azotobacter vinelandii*. The last 20 residues of the wild type R-module were exchanged to a C-terminal His-tag for purification. The construct pSABAD92A has a ColE1 origin for replication, the arabinose promoter and has also the ampilicin resistance gene (*amp*^{R})
B)
The plasmid pEBDuet23A encodes a fusion protein consisting of the A-module (residues 1-379) of the alginate epimerase AlgE4 and the N-terminal part (Int$_N$) of the *Npu*DnaE intein .
The vector has also the kanamycin resistance gene (*kan*^{R}), RSF origin and the expression of the fusion gene is tightly controlled by *T7/lac* promoter.
C)
pEBDuet28A has also the kanamycin resistance gene (*kan*^{R}), RSF origin and the *T7/lac* promoter. A fusion protein consisting of a N-terminal His-Tag, the A-module of AlgE4 and the N-terminal part (Int$_N$) of the *Npu*DnaE intein.

References

1. Iwaï H, Züger S, Jin J, Tam PH (2006) Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. FEBS Lett 580: 1853-1858.

2. Caspi J, Amitai G, Belenkiy O, Pietrokovski S (2003) Distribution of split DnaE inteins in cyanobacteria. Mol Microbiol 50: 1569-1577.
3. Aachmann FL, Svanem BG, Valla S, Petersen SB, Wimmer R (2005) NMR assignment of the R-module from the Azotobacter vinelandii Mannuronan C5-epimerase AlgE4. J Biomol NMR 31: 259.
4. Ertesvåg H, Høidal HK, Hals IK, Rian A, Doseth B, et al. (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in Azotobacter vinelandii. Mol Microbiol 16: 719-731.


A
NdeI
KpnI
EcoRI
IntC
R-module
His-tag
ara C
5000
pSABAD92A
1000
4000
5188 bps
AmpR
3000
2000
ColE1 ori
M13 ori
B
NcoI
LacI
5000
A-module
pEBDuet23A
1000
4000
AhdI
5207 bps
IntN
3000
2000
HindIII
KanR
C
NcoI
His-tag
LacI
5000
A-module
pEBDuet28A
1000
4000
AhdI
5258 bps
IntN
3000
2000
HindIII
KanR

PAPER III

10.12.2010

AALBORG UNIVERSITET

Declaration for Edith Buchinger's share of work in the following articles

Buchinger E, Skjåk-Bræk G, Valla S, Wimmer R, Aachmann FL (2010) NMR assignments of ^{1}H, ^{13}C and ^{15}N resonances of the C-terminal subunit from *Azotobacter vinelandii* mannuronan C5-epimerase 6 (AlgE6R3). Biomolecular NMR assignments

The plasmid was constructed by FLA. Expression and purification was effected by EBU and FLA. NMR data were recorded by RW, FLA and EBU. The assignment was done by EBU and FLA. EBU wrote this article with help of FLA and the input of the other authors. SV, GSB and RW conceptualized the project and participated in regular discussion of the progress and further strategy for the project.

Buchinger
Edith Buchinger

Gudmund Skjåk-Bræk

Svein Valla
Svein Valla

Finn L. Aachmann

Reinhard Wimmer

Biomol NMR Assign (2011) 5:27–29
DOI 10.1007/s12104-010-9259-0

ARTICLE

NMR assignments of ^{1}H, ^{13}C and ^{15}N resonances of the C-terminal subunit from *Azotobacter vinelandii* mannuronan C5-epimerase 6 (AlgE6R3)

Edith Buchinger · Gudmund Skjåk-Bræk · Svein Valla · Reinhard Wimmer · Finn L. Aachmann

Received: 8 June 2010 / Accepted: 2 August 2010 / Published online: 14 August 2010

Abstract The 19.9 kDa C-terminal module (R3) from *Azotobacter vinelandii* mannronan C5-epimerase AlgE6 has been ^{13}C, ^{15}N isotopically labelled and recombinantly expressed. We report here the ^{1}H, ^{13}C, ^{15}N resonance assignment of AlgE6R3.

Keywords Alginate · Mannuronan C5-epimerases · A and R-module

Abbreviations

DSS	4,4-Dimethyl-4-silapentane-1-sulfonic acid
DTT	Dithiothreitol
G	α-L-Guluronic acid
HEPES	*N*-2-Hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid
IPTG	Isopropyl *β*-D-1-thiogalactopyranoside
M	*β*-D-Mannuronic acid

Biological context

AlgE6 belongs to a family of seven structurally related alginate epimerases called AlgE1-7 produced by *Azotobacter vinelandii* (Ertesvåg et al. 1994; Ertesvåg et al. 1995). Alginate is initially produced as poly-*β*-D-mannuronic acid (M) and alginate epimerases introduce α-L-guluronic acid (G) into the polysaccharide (Hartmann et al. 2002; Campa et al. 2004). Each member of the AlgE-family produces a unique sequence of M and G subunits in the alginate polymer (Ertesvåg and Skjåk-Bræk 1999). All these epimerases consist of two types of structural modules, designated A (~385 amino acids) and R (~150 amino acids). All epimerases contain one N-terminal A-module, and AlgE1 and AlgE3 in addition contain a second such module internally in their sequences (Ertesvåg et al. 1998). The A-modules are the catalytically active parts of the epimerases, but the R-modules strongly enhance this activity although they don't possess any catalytic activity themselves (Ertesvåg and Valla 1999). The number of R-modules vary from 1 (AlgE4) to 7 (AlgE3), and AlgE6 contains three such modules. Compared to the other two R-modules of AlgE6, R3 has 69 and 64% sequence identity to R1 and R2, respectively. In addition, AlgE6R3 (located C-terminally) contains a predicted C-terminal signal peptide for secretion of the epimerase. The core structures of the R-modules are similar which is also reflected in some conserved chemical shift patterns found in ^{15}N HSQC fingerprint spectra. However, the individual R-modules show quite different affinity for different specifically tailored alginate polymers. Therefore structures of the three R-modules and their affinities to different alginates will allow us to gain a deeper insight into the role of the R-modules in epimerase functionality.

Methods and experiments

The gene coding for the AlgE6R3-module (residues 694–874) was synthesized *de novo* (GenScript, Piscataway, USA). The sequence was extended by one amino acid (Ala1) for optimal cleavage from the intein tag during purification.

E. Buchinger · R. Wimmer
Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, 9000 Aalborg, Denmark

E. Buchinger · G. Skjåk-Bræk · S. Valla · F. L. Aachmann (✉)
Department of Biotechnology, Norwegian University of Science and Technology, Sem Sælands vei 6/8, 7491 Trondheim, Norway
e-mail: finn.aachmann@biotech.ntnu.no

The DNA sequence corresponding to AlgE6R3 was cloned into pTYB12 (IMPACT-CN system, New England Biolabs.) using BsmI and XmaI sites, generating pFA13 which codes for a fusion protein consisting of AlgE6R3 and a chitin-binding domain. Uniform labelling of AlgE6R3 (181 amino acids) was achieved by overexpressing the protein in *Escherichia coli* ER2566 containing the plasmid pFA13. The cells were grown at 37°C to OD_{600} ~0.8 in M9-medium supplemented with $(^{15}NH_4)_2SO_4$ (1 g/L), $^{13}C_6$-D-glucose (2 g/L) (Sigma-Aldrich), 0.2 mM $CaCl_2$ and 200 µg/L ampicillin. Expression was induced by 1 mM ITPG at 15°C and allowed to continue over night. The cells were harvested and resuspended in 20 mM HEPES pH 6.9, 800 mM NaCl, 10 mM $CaCl_2$ and 0.1% Triton X-100 (Sigma-Aldrich). They were then lysed by sonification and the supernatant was loaded on a column with chitin beads (New England BioLabs). The column was washed with 20 mM HEPES pH 6.9, 800 mM NaCl and 5 mM $CaCl_2$. AlgE6R3 was cleaved from the chitin binding tag by incubating the column with the bound fusion protein with 50 mM DTT in 20 mM HEPES pH 6.9, 800 mM NaCl and 5 mM $CaCl_2$ at room temperature for ~16 h, whereafter it could be eluted from the column. The eluted AlgE6R3 was dialysed against 20 mM HEPES, pH 6.9, 25 mM $CaCl_2$.

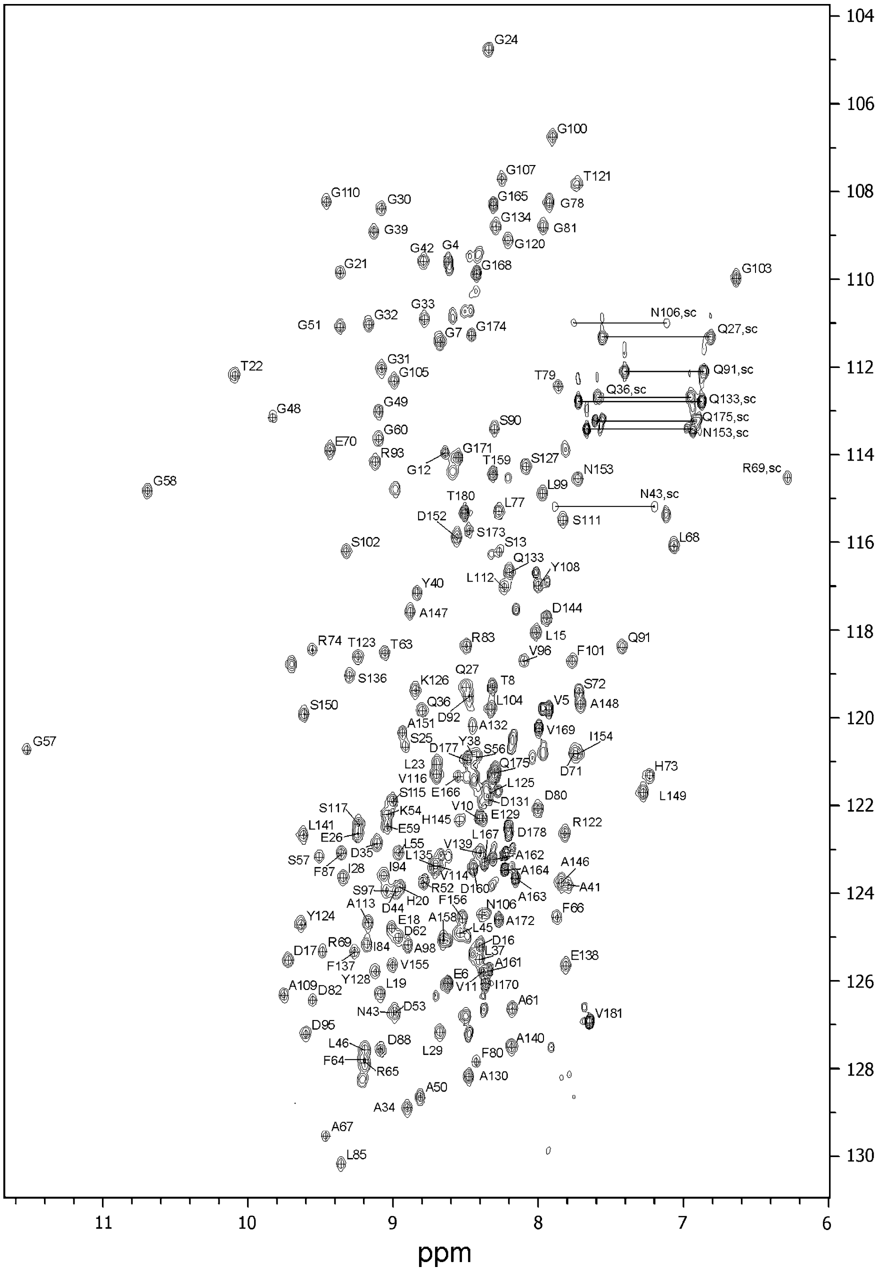


Fig. 1 ^{1}H, ^{15}N HSQC spectrum of the ^{13}C, ^{15}N-labelled AlgE6R3 subunit from *Azotobacter vinelandii* in 90:10 $H_2O:D_2O$ at pH 6.9, 298 K. Residue numbers are indicated. Side-chain resonances of Asn and Gln residues are connected by *lines*. Other side-chain amine resonances are indicated with amino acid number and sc

Samples for NMR studies contained 1.0–1.4 mM AlgE6R3 in 20 mM HEPES buffer, pH 6.9 with 25 mM $CaCl_2$ dissolved in either 90% H_2O/10% D_2O or 99.9% D_2O.

The NMR measurements were performed on a Bruker Avance 600 spectrometer equipped with 5 mm z-gradient TXI (H/C/N) cryogenic probe and on a Bruker DRX 600 spectrometer equipped with 5 mm xyz-gradient TXI (H/C/N) probe. All experiments were performed at 298 K. Proton and carbon chemical shifts were referenced relative to internal 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS); ^{15}N chemical shifts were referenced indirectly to DSS, based on the absolute frequency ratios (Zhang et al. 2003). For the sequence specific backbone- and side-chain assignment, the following experiments were used: ^{15}N-HSQC, ^{13}C-HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH, CBCA(CO)NH, HBHANH, HBHA(CO)NH, HCCH-TOCSY and HCCH-COSY. The assignment of aromatic side chains were based on 2D ^{13}C-HSQC and 2D-NOESY, 2D-COSY and 2D-TOCSY recorded on samples dissolved in D_2O. The NMR data were recorded and processed with Bruker XWinNMR version 3.5 or Bruker TopSpin 1.3 software and spectral analysis was performed using CARA version 1.4.1/1.8.4 (Keller 2004).

Assignment and data composition

We report here the backbone resonance assignments of the third R-module of AlgE6R3. The ^{15}N-HSQC spectrum of AlgE6R3, together with the assignments of the resonances, is shown in Fig. 1. The backbone and side chain assignment were essentially complete: 96.9% of the backbone H^{N}, H^{α}, C′, C^{α} and N atoms, and 95.7% of the side-chain atoms has been assigned. A1 and D2 were not assigned. The amide groups (H^{N}, N) of D14, D47, D118, D119, E142, G143 and A157 could not be found, although other nuclei of these residues were assigned. Except for H^{ε} of R69, none of exchangeable side-chain protons of Arg and Lys residues were assigned. Side-chain amide protons of all Asn and Gln residues were assigned. All aromatic protons were assigned except H^{ζ} of F101. Protonated carbon atoms of aromatic residues were assigned to a large extent.The chemical shift data have been deposited in the BioMagResBank data-base under the accession number 16956.

Acknowledgments This research was supported by the Norwegian Research Council (NFR) via KMB project (182695/I40). EB ad RW acknowledge financial support from the Danish Agency for Science, Technology and Innovation. The NMR laboratory at Aalborg University is supported by the Obel and SparNord foundations.

References

Campa C, Holtan S et al (2004) Biochemical analysis of the processive mechanism for epimerization of alginate by mannuronan C-5 epimerase AlgE4. Biochem J 381(Pt 1):155–164

Ertesvåg H, Skjåk-Bræk G (1999) Modification of alginate using mannuronan C-5-epimerases. In: Bucke C (ed) Carbohydrate biotechnology protocols, vol 10. Humana Press, Totowa, pp 71–78

Ertesvåg H, Valla S (1999) The A modules of the *Azotobacter vinelandii* mannuronan-C-5-epimerase AlgE1 are sufficient for both epimerization and binding of Ca2+. J Bacteriol 181(10): 3033–3038

Ertesvåg H, Doseth B et al (1994) Cloning and expression of an *Azotobacter vinelandii* mannuronan C-5-epimerase gene. J Bacteriol 176(10):2846–2853

Ertesvåg H, Høidal HK et al (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in *Azotobacter vinelandii*. Mol Microbiol 16(4):719–731

Ertesvåg H, Høidal HK et al (1998) The *Azotobacter vinelandii* mannuronan C-5-epimerase AlgE1 consists of two separate catalytic domains. J Biol Chem 273(47):30927–30932

Hartmann M, Holm OB et al (2002) Mode of action of recombinant *Azotobacter vinelandii* mannuronan C-5 epimerases AlgE2 and AlgE4. Biopolymers 63(2):77–88

Keller RLJ (2004) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment. Cantina Verlag, Goldau

Zhang H, Neal S et al (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25(3): 173–195

PAPER IV

10.12.2010

AALBORG UNIVERSITET

Declaration for Edith Buchinger's share of work in the following articles

Andreassen T, Buchinger E, Skjåk-Bræk G, Valla S, Aachmann FL (2010) ^{1}H, ^{13}C and ^{15}N resonances of the AlgE62 subunit from *Azotobacter vinelandii* mannuronan C5-epimerase. Biomolecular NMR Assignments

The plasmid was constructed by FLA. Expression and purification of the protein was performed by FLA and EBU. The assignment was performed by TA and FLA. NMR spectra were recorded by TA, EBU and FLA. FLA and TA wrote this article with help of the other authors. SV, GSB and FLA designed the project and discussed further strategy for the project regularly.

Buchinger

Trygve Andreassen

Edith Buchinger

Gudmund Skjåk-Bræk

Svein Valla

Svein Valla

Finn L. Aachmann

Biomol NMR Assign (2011) 5:147–149
DOI 10.1007/s12104-010-9288-8

ARTICLE

^{1}H, ^{13}C and ^{15}N resonances of the AlgE62 subunit from *Azotobacter vinelandii* mannuronan C5-epimerase

Trygve Andreassen · Edith Buchinger ·
Gudmund Skjåk-Bræk · Svein Valla ·
Finn L. Aachmann

Received: 21 October 2010 / Accepted: 13 December 2010 / Published online: 28 December 2010

Abstract The 17.7 kDa R2 module from *Azotobacter vinelandii* mannronan C5-epimerase AlgE6 has been isotopically labeled (^{13}C,^{15}N) and recombinantly expressed. Here we report the ^{1}H, ^{13}C, ^{15}N resonance assignment of AlgE6R2.

Keywords Alginate · Mannuronan C5-Epimerases · A and R-module

Biological context

In *Azotobacter vinelandii* a family of seven secreted and calcium-dependent, mannuronan C5–epimerases (AlgE1–7) has been identified (Ertesvåg et al. 1999). These enzymes catalyse the epimersation of β-D-mannuronic acid (M) to α-L-guluronic acid (G) in the polysaccharide alginate. The epimerases are composed of two different structural modules, designated A (~385 amino acids each, with 1 or 2 copies per enzyme) and R (~155 amino acids each, with one to seven copies per enzyme) and a C-terminal putative signal peptide (Ertesvåg et al. 1994, 1995). The A-modules alone are catalytically active, but their reaction rates are increased when covalently bound to at least one R-module (Ertesvåg and Valla 1999). The epimerases generate different monomer sequence distributions in their reaction products, and these patterns appear to a large extent to be controlled by the A-modules alone. The exact functional roles of the R-modules are not yet understood.

Previously the R-module from the smallest epimerase, AlgE4 (A–R), was found to fold as an all parallel β-roll protein similar to the repeats in toxin (RTX) proteins, but with a positively charged shallow grove on the front side with the ability to bind the polyanionic alginate (Aachmann et al. 2006, 2005). AlgE4 exhibit by a processive mode of action, while AlgE6 (A-R1-R2-R3) preferentially introduces GG-blocks in the alginate (Campa et al. 2004, Hartmann et al. 2002). Both the first and last R-module from AlgE6 has been studied by NMR (Buchinger et al. 2010, Aachmann and Skjåk-Bræk 2008) and the 3D structures are currently being determined. The sequence identity of AlgE6, R2 is 69% and 69% compared to R1 and R3, respectively. Some conserved chemical shift patterns seem to be shared in ^{15}N HSQC fingerprint spectrum. Preliminary results from the structure determination confirm this observation, but it has also been observed additional structure elements, which are not found in R1 and R3. Therefore, it is interesting to study the structure of R2-module as well as R1 and R3 in detail in order to gain insight into its role in AlgE6 function. Altogether, access to the structures of the three AlgE6 R-modules and their affinities to different tailored made alginates will hopefully lead to new insights that may be used to deduce their functionality in all the seven epimerases. Here, we report the complete sequence-specific assignments of the AlgE6 R2 module.

Methods and experiments

The gene coding for the AlgE6R2-module (residues 534–693) has been codon optimized for *Escherichia coli* protein expression and synthesized de novo (GenScript, Piscataway, USA). The synthetic gene for AlgE6R2 was

T. Andreassen · E. Buchinger · G. Skjåk-Bræk · S. Valla ·
F. L. Aachmann (✉)
Department of Biotechnology, Norwegian University of Science
and Techology, Sem Sælands vei 6/8, 7491 Trondheim, Norway
e-mail: finn.aachmann@biotech.ntnu.no

E. Buchinger
Department of Biotechnology, Chemistry and Environmental
Engineering, Aalborg University, 9000 Aalborg, Denmark

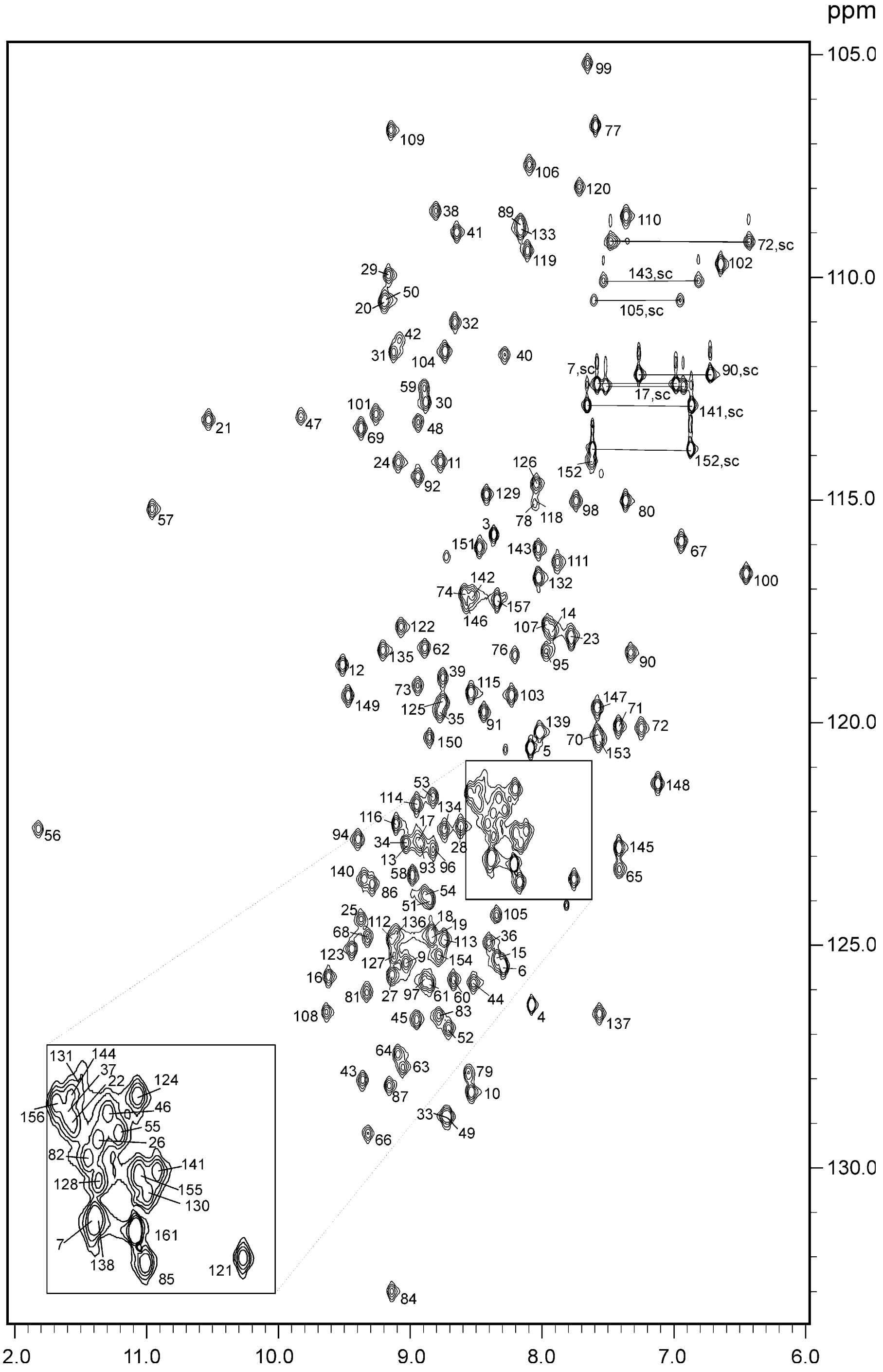


Fig. 1 ^1H, ^{15}N HSQC spectrum of the ^{13}C, ^{15}N-labelled AlgE6R2 subunit from *Azotobacter vinelandii* in 90:10 H_2O:D_2O at pH 6.9, 298 K. Residue numbers are indicated. Side-chain resonances of Asn and Gln residues are connected by *lines*. Other side-chain amine resonances are indicated with amino acid number and sc

inserted into pTYB12 vector (IMPACT-CN system, New England Biolabs) using BsmI and PstI sites, hereafter purified via gel electrophoresis and ligated with T4 ligase at 289 K 2 h generating pFA12 plasmid. The pFA12 plasmid was confirmed by restriction mapping. This plasmid was transformed into the production host *E. coli* ER2566. For uniform labelling of AlgE6R2 (161 amino acids) the cells were grown in M9-medium supplemented with $(^{15}NH_4)_2SO_4$ (1 g/L), $^{13}C_6$-D-glucose (2 g/L) (Sigma–Aldrich) with additional 0.2 mM $CaCl_2$. The fusion protein was overexpressed by growing the cells at 310 K until an OD_{600} ~0.8 was reached and the expression was started by addition 1 mM IPTG (*in toto*) and subsequently incubated at 289 K for 16 h. The cells were harvested by centrifugation and resuspended in 20 mM HEPES pH 6.9, 500 mM NaCl, 5 mM $CaCl_2$ and 0.1% Triton X-100 (Sigma–Aldrich). The cells were disrupted by sonication and centrifugated. The supernatant was applied onto chitin bead column (New England BioLabs). The column was washed with 20 mM HEPES pH 6.9, 500 mM NaCl and

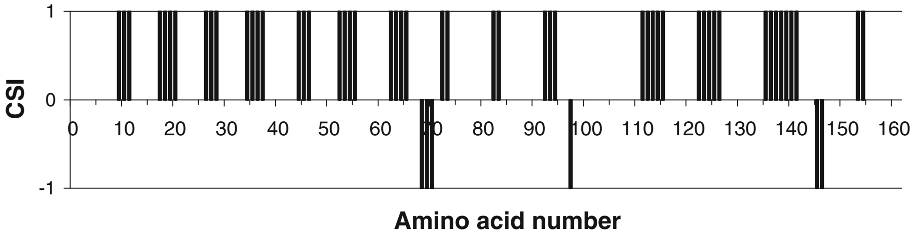


Fig. 2 The chemical shift index (CSI) for AlgE6R2 subunit from *Azotobacter vinelandii*

5 mM $CaCl_2$, whereafter followed by cleavage with 50 mM DTT at room temperature over ~16 h resulting in the release of the AlgE6R2 from the chitin bound intein tag. The eluted AlgE6R2 was dialysed against 10 mM HEPES, pH 6.9, 10 mM $CaCl_2$ in order to remove a 1.6 kDa peptide that occurred as a by-product from the cleavage reaction.

The NMR spectra were recorded at 298 K on a Bruker Avance 600 spectrometer equipped with 5 mm Z-gradient TCI(H/C/N) cryogenic probe. ^{1}H and ^{13}C chemical shifts were referenced internal to the sodium salt of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), while ^{15}N chemical shifts were referenced indirectly to DSS, based on the absolute frequency ratios (Zhang et al. 2003). Sequence-specific backbone and side-chain assignments of AlgE6R2 were accomplished using ^{15}N HSQC, ^{13}C HSQC, HNCA, HN(CO)CA, HNCO, HN(CA)CO, CBCANH, CBCA(-CO)NH, HBHANH, HBHA(CO)NH, HCCH-TOCSY and HCCH-COSY spectra. The assignments of the aromatic side chains were obtained from 2D IP-COSY, TOCSY, NOESY and ^{13}C HSQC experiments. The NMR data were recorded and processed with BRUKER XWinNMR version 3.5 or TopSpin 3.0 software and spectral analysis was performed using CARA version 1.4.1/1.5.1/1.8.4 (Keller 2004).

Assignment and data composition

Here we report the resonance assignments of the R2-module of AlgE6R2. The ^{15}N-HSQC spectrum of AlgE6R2, together with the assignments of the resonances, is shown in Fig. 1. The backbone and the side-chain assignments are essentially complete (H^N,H^α, N, C^α, C' > 97%; H and C side-chains > 94%). The amide groups (H^N, N) of Asp1, Gln35, Ala75, and, Asp117 could not be found, although other nuclei of these residues were assigned. None of the exchangeable side-chain protons of Arg and Lys residues were assigned. Side-chain amide groups of all Asn and Gln were assigned. Most of the protons and the carbon atoms of aromatic side-chains were assigned. Results obtained for the chemical shift index (CSI) point toward mainly β-strands for the proteins secondary structure (See Fig. 2) that fit also well to an overall 3D β-roll structure common for the R-modules. The chemical shift data have been deposited in the BioMagResBank under the accession number 17249.

Acknowledgments FLA acknowledges financial support from the Norwegian Research Council (NFR) via KMB project (182695/I40). EB acknowledges financial support from the Danish Agency for Science, Technology and Innovation.

References

Aachmann FL, Skjåk-Bræk G (2008) 1H, 15 N, 13C resonance assignment of the AlgE6R1 subunit from the *Azotobacter vinelandii* mannuronan C5-epimerase. Biomol NMR Assigm 2: 123–125

Aachmann FL, Svanem BIG, Valla S et al (2005) NMR assignment of the R-module from the *Azotobacter vinelandii* Mannuronan C5-epimerase AlgE4. J Biomol NMR 31:259

Aachmann FL, Svanem BIG, Guntert P et al (2006) NMR structure of the R-module - A parallel beta-roll subunit from an *Azotobacter vinelandii* mannuronan C-5 epimerase. J Biol Chem 281:7350–7356

Buchinger E, Skjak-Braek G, Valla S et al. (2010) NMR assignments of ^{1}H, ^{13}C and ^{15}N resonances of the C-terminal subunit from *Azotobacter vinelandii* mannuronan C5-epimerase 6 (AlgE6R3). Biomol NMR Assign

Campa C, Holtan S, Nilsen N et al (2004) Biochemical analysis of the processive mechanism for epimerization of alginate by mannuronan C-5 epimerase AlgE4. Biochem J 381:155–164

Ertesvåg H, Valla S (1999) The A modules of the *Azotobacter vinelandii* mannuronan-C-5-epimerase AlgE1 are sufficient for both epimerization and binding of Ca2+. J Bacteriol 181: 3033–3038

Ertesvåg H, Doseth B, Larsen B et al (1994) Cloning and expression of an *Azotobacter vinelandii* mannuronan C-5-epimerase gene. J Bacteriol 176:2846–2853

Ertesvåg H, Høidal HK, Hals IK et al (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in *Azotobacter vinelandii*. Mol Microbiol 16:719–731

Ertesvåg H, Høidal HK, Schjerven H et al (1999) Mannuronan C-5-epimerases and their application for in vitro and in vivo design of new alginates useful in biotechnology. Metab Eng 1:262–269

Hartmann M, Duun AS, Markussen S et al (2002) Time-resolved 1H and 13C NMR spectroscopy for detailed analyses of the *Azotobacter vinelandii* mannuronan C-5 epimerase reaction. Biochim Biophys Acta 1570:104–112

Keller RLJ (2004) Optimizing the Process of Nuclear Magnetic Resonance Spectrum Analysis and Computer Aided Resonance Assignment. CANTINA Verlag, Zürich

Zhang HY, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195

PAPER V

10.12.2010

AALBORG UNIVERSITET

Declaration for Edith Buchinger's share of work in the following articles

Buchinger E, Knudsen DH, Behrsen MA, Pedersen JS, Valla S, Skjåk-Bræk G, Wimmer R, Aachmann FL; Structural and Functional Characterization of the R-modules in Alginate C-5 Epimerase AlgE4 and AlgE6 from *Azotobacter vinelandii*; The manuscript is in preparation

The plasmid constructs were carried out by FLA. The NMR spectra were recorded and treated by RW, FLA and EBU. The structure calculations of the R-modules were done by FLA and EBU. Alginate titration studies were performed by DHK and FLA. SAXS data were obtained and treated by MAB and EBU. The paper was written by EBU and RW with guidance of the other authors. SV, GSB, RW and JSB designed the project. All authors discuss the development of the project regularly.

Buchinger
Edith Buchinger

Daniel
Daniel H. Knudsen

Manja A. Behrsen

Jan Skov Pedersen

Svein Valla
Svein Valla

Gudmund Skjåk-Bræk

Reinhard Wimmer

Finn L. Aachmann

Structural and Functional Characterization of the R-modules in Alginate C-5 Epimerase AlgE4 and AlgE6 from *Azotobacter vinelandii*

Edith Buchinger[a,b], Daniel H. Knudsen[a], Manja A. Behrens[c], Jan Skov Pedersen[c], Svein Valla [b], Gudmund Skjåk-Bræk [b], Reinhard Wimmer[a] and Finn Lillelund Aachmann [b*]

Adresses

[a]Department of Biotechnology, Chemistry and Environmental Engineering,
Sohngaardsholmsvej 49, DK-9000 Aalborg , Denmark.

[b]NOBIPOL, Department of Biotechnology, Norwegian University of Science and Technology, Sem Sælands vei 6/8, 7491 Trondheim, Norway.

[c]Department of Chemistry and iNANO Interdisciplinary Nanoscience Center, Aarhus University, Langelandsgade 140, DK-8000, Denmark

[*]To whom correspondence should be addressed: finn.aachmann@biotech.ntnu.no, Tel. +4773593317

Introduction

Alginates are unbranched biopolymers consisting of β-D-mannuronic (M) and its epimer α-L-guluronic (G) acid [1,2]. The physical and chemical properties of alginate are influenced by relative amount of M to G as well as the distribution in the polymer chain [3]. Therefore alginate can be sub-divided in three different blocks called M-, G- and MG-block [3,4,5]. M-blocks form acidic gels while the gels of G- and MG-blocks are induced by Ca^{2+}-ions or other divalent ions [3,6]. The stiffness of the gels decrease in this order: G- > M- >MG-blocks [7]. Alginate is produced as poly-M and epimerases on the polymer level. One alginate producing bacteria is *Azotobacter vinelandii,* which incorporate single G residues into the alginate polymer during secretion of the polymer [8]. In addition, *A. vinelandii* produces seven extracellular C-5 alginate epimerases called AlgE1-7 [9,10]. Each of the epimerases converts mannuronic acid to guluronic acid in different pattern and AlgE7 shows also lyase activity [11,12]. These seven epimerases consist of two structurally modules designed A- and R-module (Fig. 1).

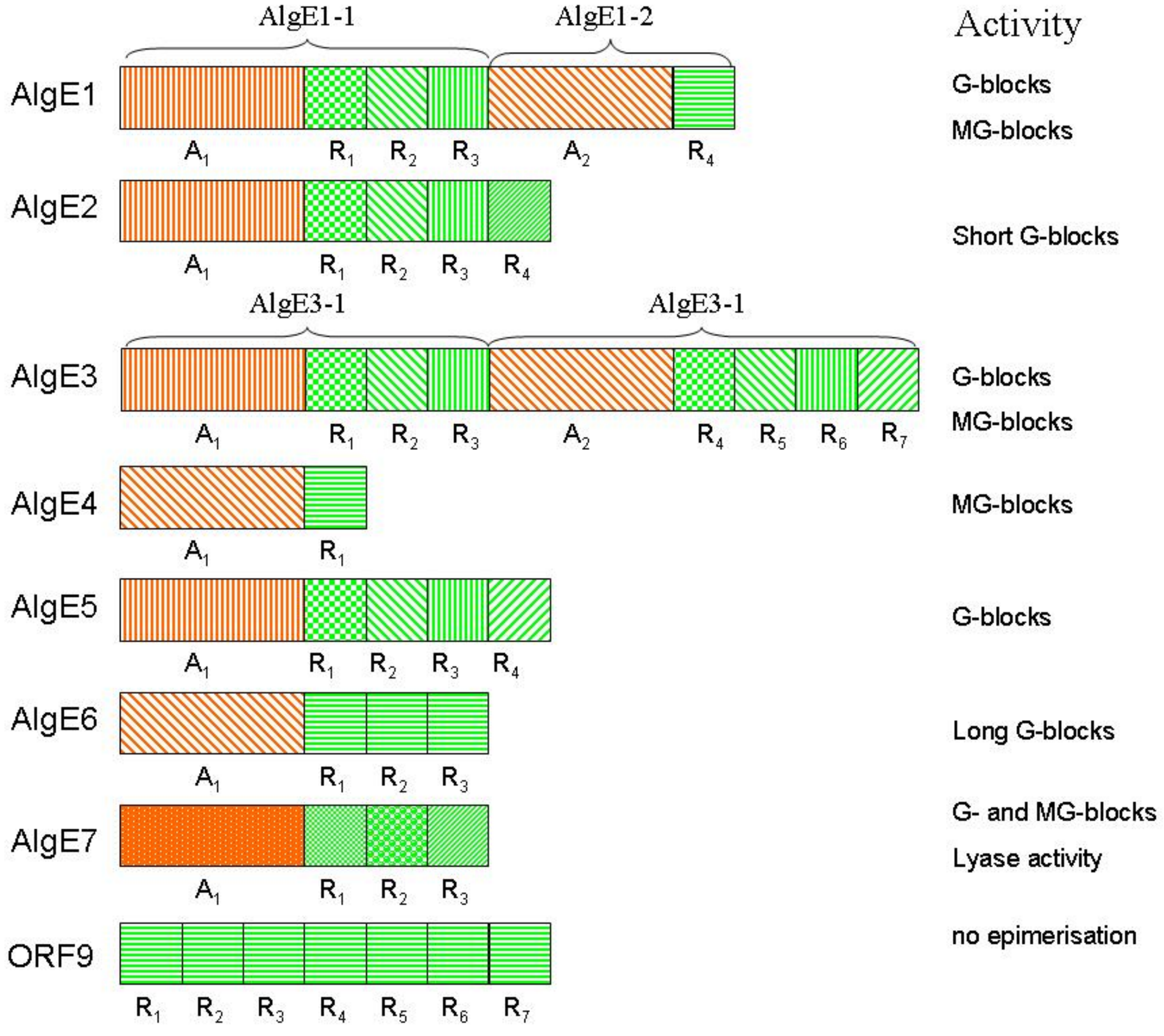


Fig. 1:Alginate epimerases. *Azotobacter vinelandii* expresses a family of extracellular alginate epimerases, which only consists of two different modules named A- and R-module. All extracellular epimerases are under the same operon except AlgE5. Only the A-modules (in orange) are catalytically active but the R-modules (in green) enhance the activity significantly if bound to an A-module. The different members of the family show different epimerisation products, given on the right side of the figure, AlgE7 acts also as a lyase. ORF9 is not an alginate epimerase and has no A-module but it is in the same operon as the other epimerases (except AlgE5). Its function is unknown.

The smallest of these epimerases is AlgE4 consisting of one A- and R-module. AlgE4 incorporates MG-blocks into the alginate sequence. The atomic-resolution structures of A- [13] and R-module [14] of AlgE4 were determined recently. It could also be shown that both modules bind to alginate [14,15].

The second smallest alginate epimerase is AlgE6, which consists of one A-module followed by three R-modules (A-R_1-R_2-R_3). The A-modules and R-modules of AlgE4 and AlgE6 share a high sequence identity and similarity [10] but produce alginates with quite different G content. While AlgE4 incorporates MG-blocks into the alginate sequence, AlgE6 produces G-blocks as end products. Only the A-modules are catalytically active [15]. Co-crystallisation of A-module with alginate revealed the catalytic site and a reaction mechanism was postulated [13].

The role of the R-module is more enigmatic. R-modules show no catalytic activity on alginates but if one R-module is bound after an A-module the epimerisation rate is ten-fold increased [15]. As mention earlier the R-module of AlgE4 is able to bind alginate [14] and epimerisation of alginate by the whole AlgE4 occurs by processive mode meaning the protein is sliding along the alginate polymer [16,17]. Therefore, it was suggested that the R-module helps to orientate the alginate for the active site and to keep the AlgE4 tightly to the polymer during epimerisation.

The role of the R-modules in G-block producing epimerases is not clear. The R-modules enhance the reaction rate [15] but if two or more mannuronic acids in row are epimerised after each other, those epimerase would have to turn 180° between each epimerisation step. Processive mode is only possible if one AlgE6 slides along the alginate polymer incorporating MG-block while a second enzyme epimerizes the MG-block to G-block. To understand the role of multiple R-modules in G-block producing epimerases we determined the atomic-resolution structures of the three R-modules of AlgE6 by NMR. The results are presented here. Additionally the binding of these R-modules and AlgE4R with alginate of different composition was investigated using NMR and isothermical titration calorimetry (ITC). It is known that AlgE4R binds to alginate but more detailed binding studies should reveal possible preferences of these R-modules to certain types of alginate. The A- and R-modules have defined elongated structures but initial tests showed flexibility between the modules [18]. The orientation of the modules towards each other was further investigated. For that, overall structures of AlgE4, AlgE6 and different modules were also determined in the present work by small angle X-ray scattering (SAXS).

Materials&Methods

Plasmids

Plasmids and bacterial strains used in this study are summarized in Tab. 1.

Tab. 1: Plasmids used in this study.

Plasmid	Description	Reference
pTYB12	IMPACT-CN fusion vector	obtained from NEB
pTYB4	IMPACT-CN fusion vector containing a C-terminal chitin-binding-tag	obtained from NEB
pFA8	Derivate of pTYB12 in which the DNA sequence of algE6R1 was cloned using BsmI and XmaI sites	[19]
pFA12	Derivate of pTYB12 in which a BsmI-XmaI fragment containing algE6R2 was inserted	[20]
pFA13	algE6R3 was cloned into pTYB12 between BsmI and XmaI sites	[21]
pBS32	Derivative of pTYB4 in which the NcoI-XmaI fragment containing *algE4*	unpublished results
pFA1	Derivate of pTYB11 opened with SapI and treated with PeaR7I to insert the sequence of *algE4R*	[22]
pTB26	Derivative of pTYB4 in which the NcoI-XmaI fragment containing *algE4A* was inserted	[13]

Bacterial strains		
DH5α	*Escherichia coli* DH1 derivate with deoR, nupG, Φ80d*lac*ZΔM15 and Δ(*lac*ZY-*arg*F)U169	[23]
ER2566	$F^-\lambda^-$fhuA2 [Ion] ompT lac::T7gene1 gal SulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10-TetS)2R(zgb-210::Tn10) (TetS) endA1 [dcm]	Obtained from NEB

All plasmids have a T7/lac promoter and ampicillin resistance gene and were transformed into ER2566 strain for protein expression.

Buffers and expression media

1 L LB medium contained 10 g Tryptone, 5 g Yeast extract and 5 g NaCl. The pH was adjusted to 7.2 with 4 M NaOH and the medium was sterilized by autoclaving.

For 1 L M9 medium 7.2 g $Na_2HPO_4x2H_2O$, 3 g KH_2PO_4, 0.5 g NaCl and 1 g NH_4Cl were dissolved in 1 L H_2O. The pH was adjusted to 7.4 and the medium was autoclaved. Before the expression 2 mL of 1 M $MgSO_4$, 20 mL of trace metal and 2 g glucose dissolved in 10 mL H_2O were sterilized by filtration and added to the medium.

Trace metal for M9-medium:

0.1 g/L $ZnSO_4$, 0.8 g/L $MnSO_4$, 0.5 g/L $FeSO_4$, 0.1 g/L $CuSO_4$ and 1 g/L $CaCl_2$

For production of ^{15}N- and $^{15}N,^{13}C$-labelled proteins M9-medium was prepared with 1 g/L $(^{15}NH_4)_2SO_4$ and 2 g/L U-^{13}C-D-glucose. $^2H,^{15}N$ labelled proteins were expressed in 1 L M9-medium prepared with 99% D_2O and supplemented with 1 g/L $(^{15}NH_4)_2SO_4$ and 2 g/L U-^2H-D-glucose.

Expression

Independent of which medium was used all cells were grown at 37° C to OD_{600nm}~ 0.8. The cell culture was incubated on ice for 5 min. The expression was induced by IPTG (final concentration 1 mM) and then culture was incubated at 15° C over night. The cells were harvested, resuspended in 20 mM HEPES pH 6.9, 800 mM NaCl and 5 mM $CaCl_2$ and 0.1% Triton X and the cell can be stored at −20° C for further purification if no purified immediately. The expression of AlgE6R1, AlgE6R2 and AlgE6R3 for structure determination was described recently [19,21]. AlgE4 was expressed as uniformly deuterated and ^{15}N-labelled proteins. For the SAXS measurement no isotope labelling was required and protein expression was thus conducted in LB-medium.

Refolding of ^{2}H,^{15}N-AlgE4 and ^{2}H,^{15}N-AlgE4A

The cells containing the expressed protein were thawed and lysed by sonication. The pellets were solubilised in 3 mL 20 mM HEPES pH 6.9, 6 M GdmCl, 800 mM NaCl, 5 mM $CaCl_2$ at 4 °C. The solution was diluted 10-times in 20 mM HEPES pH 6.9, 800 mM NaCl, 5 mM $CaCl_2$ and was dialysed against the same buffer at 4 °C.

Purification

The cells were sonicated and the crude cell extracts were loaded on a chitin column. The column was washed with 20 mM HEPES pH 6.9, 800 mM NaCl and 5 mM $CaCl_2$ with at least 10 column volumes. The proteins were cleaved from the chitin-binding-tag by loading 50 mM DTT dissolved in the washing buffer on the column and letting it react for 40 h at room temperature. The purified proteins were eluted from the column and dialyzed against 20 mM HEPES pH 6.9 and 25 mM $CaCl_2$. If the proteins were not used immediately they were freeze-dried and stored at −20° C. Protein used for SAXS was further purified by gel filtration in order to remove any aggregates.

Gel filtration

The samples for SAXS were concentrated to less than 0.5 ml or if they had been freeze dried they were dissolved in less than 0.5 ml MilliQ water. The protein solutions were loaded on Superdex 200 10/300 GL (AlgE6 and AlgE4) or superdex 75 HT 10/30 (AlgE6R1, AlgE6R2, AlgE6R3, AlgE4R and AlgE4A), respectively. The samples were eluted from the column with 5 mM HEPES pH 6.9, 50 mM NaCl, 10 mM $CaCl_2$ and 0.5 M glycine. The monomeric fractions were concentrated to less than 0.5 ml and dialyzed against 20 mM HEPES pH 6.9, 125 mM NaCl, 25 mM $CaCl_2$ and 0.5 M glycine.

Alginate binding by NMR

Binding of alginate oligomers to AlgE4R, AlgE6R1, AlgE6R2 and AlgE6R3 was investigated by NMR at 25 °C. The titration of alginate to the protein was done as described [14]. For NMR measurements, ^{15}N-labelled material was produced as described above. The protein concentrations were between 0.2 and 0.5 mM. The chemical shift changes of N and H^N atoms from the back bone of the R-module upon titration are given as an absolute change in chemical shift by the following formula:

$$\Delta\delta_{abs} = \sqrt{(\Delta\delta_H)^2 + x(\Delta\delta_N)^2}$$

$\Delta\delta_{abs}$	absolute change in chemical shift [Hz]
$\Delta\delta_H$	chemical shift change of the amide proton [Hz]
$\Delta\delta_N$	change in chemical shift of the amide nitrogen atom [Hz]
x	constant to achieve equal contribution from changes in N and H^N- shifts. The constant was set to 5

The absolute change in chemical shift was plotted versus residue. Residues that experience a $\Delta\delta_{abs} >$ 100 Hz were considered to be affected by binding. Approximately seven of the strongest shifting peaks were used to calculate dissociation constants by fitting the experimental data to the following equation:

$$\Delta\delta = \sqrt{(\Delta\delta_H)^2 + 5(\Delta\delta_N)^2}$$

$$\Downarrow$$

$$\Delta\delta_{obs} = \delta_{obs} - \delta_{free}$$

$$\Delta\delta_{bound} = \delta_{bound} - \delta_{free}$$

$$\Downarrow$$

$$\frac{\Delta\delta_{obs}}{\Delta\delta_{bound}} = \frac{[PL]}{[P]_0}$$

$$\Downarrow$$

$$K_a = \frac{[PL]}{[L]\cdot[P]} = \frac{[PL]}{([L]_0 - [PL])\cdot([P]_0 - [PL])} = \frac{1}{K_d}$$

$$\Downarrow$$

$$\frac{\Delta\delta_{obs}}{\Delta\delta_{bound}} = \frac{[PL]}{[P]_0} = \frac{1}{2[P]_0}\left[([P]_0 + [L]_0 + K_d) \pm \sqrt{([P]_0 + [L]_0 + K_d)^2 - 4[P]_0[L]_0}\right]$$

where

$\Delta\delta_N$, $\Delta\delta_H$ chemical shift change in the nitrogen or proton dimension [Hz]

δ_{free} chemical shift of the atom in the free protein without ligand in solution [Hz]

δ_{bound} chemical shift of the atom in the protein-ligand-complex [Hz]

δ_{obs} chemical shift of the atom in the protein at a certain ligand concentration [Hz]

$\Delta\delta_{obs}$ observed chemical shift changes at a certain point in titration [Hz]

$\Delta\delta_{bound}$ chemical shift changes of the atom in the protein due to ligand binding[Hz]

$[PL]$ protein-ligand complex concentration [mol/L]

$[P]$ free protein concentration [mol/L]

$[L]$ free ligand concentration [mol/L]

$[P]_0$ total protein concentration (free and bound) [mol/L]

$[L]_0$ total ligand concentration (free and bound) [mol/L]

K_a	association constant [L/mol]
K_d	dissociation constant [mol/L]

Dissociation constant determination by ITC

Calorimetric measurements were carried out with a MircoCal VP-ITC microcalorimeter at 25 °C. For the R-module from AlgE4 the alginate oligomers dissolved in 20 mM HEPES pH 6.9 and 50 mM $CaCl_2$ were added to a 0.04 mM R-module solution in 20 mM HEPES pH 6.9 with 50 mM $CaCl_2$ and the energy release/consumption was measured. For the R-modules from AlgE6 the alginate oligomers dissolved in 20 mM HEPES pH 6.9, 25 mM $CaCl_2$, and, 40 mM NaCl were added to a 0.04 mM R-module solution in 20 mM HEPES pH 6.9 with 25 mM $CaCl_2$, and, 40 mM NaCl. The energy release/consumption was measured. The calorimeter was controlled by VPViewer 2000 v.1.4.8® (MicroCal) and the parameters for the experiments are summarized below.

Total injections	30
Reference power [µcal/s]	10
Initial delay [s]	60
String speed [rpm]	290
Volume of the first injection [µL]	2
Volume of the 2.-30. injection [µL]	10
Spacing time [s]	240
Filter period [s]	2

Data analysis was performed in Origin 7.0 ® with a Raw-ITC analysis template. Thermodynamic parameters were determined based on following equation using least-square methods.

$$\Delta G = \Delta H - T\Delta S = RTN \ln K_d$$

Where

ΔG	Gibb's free energy [J/mol]
ΔH	enthalpy changes [J/mol]
ΔS	entropy changes [J/(mol·K)]
T	temperature [K]
R	ideal gas constant (8.31 J/(mol·K))
N	number of ligand molecules binding to one protein molecule
K_d	dissociation constant [mol/L]

SAXS measurement

The small-angle X-ray scattering measurements were preformed on a laboratory-based instrument at Aarhus University, Denmark [24]. The data was collected in a reusable quartz capillary and all measurements were carried out at 4° C. The scattering patterns were collected for between 3 and 4 hours, dependent on protein concentration. The protein concentrations were between 1 and 2 mg/mL. Additionally buffers was collected for all samples and background subtraction and conversion of the data to absolute scale by use of water as a primary standard was preformed using the SUPERSAXS program package (Oliveira, C. L. P. and Pedersen, J. S., unpublished). The final intensity is displayed as a function of the scattering vector $q = 4\pi \sin\theta/\lambda$, where λ is the X-ray wavelength and 2θ is the angle between the incident and scattered X-rays.

SAXS analysis and modelling

The first step in the data analysis was to perform an indirect Fourier transformation (ITF) to obtain the pair distance distribution function, $p(r)$ function. This was done by use of the program WIFT ([25] and Oliveira, C. L. P and Pedersen, J. S., unpublished). From the IFT several parameters are obtained: the maximum diameter, D_{max}, the radius of gyration, R_{g}, and the forward scattering, $I(q = 0)$. Using the forward scattering the molecular mass of the molecule in solution can be calculated as $M_w^{protein} = I(0)/[c\Delta\rho_m^2]$, where c is the protein concentration and $\Delta\rho_m$ is the scattering length density difference per unit mass for which a standard value of 2.0 x 10^{10} cm/g was used.

The scattering data of the single modules were compared with atomic resolution structures, either by use of the known pdb files (AlgE4A, AlgE4R, AlgE6R1 and AlgE6R3) or if not available by pdb-files and homology models (AlgE6A and AlgE6R2) generated by SwissModel [26]. The theoretical scattering for the models is computed for the structures in solution taking a hydration layer on the molecules into account. Finally the discrepancy between the experimental scattering data and the computed scattering pattern were calculated. This procedure is implemented in the program CRYSOL [27].

Flexible structures are analysed using the Ensemble Optimization Method (EOM) [28], where the flexibility is modelled by representing the scattering data as an ensemble of protein conformations. The structures constituting the ensemble are selected by a genetic algorithm from a large pool (10,000 conformations) of randomly generated structures. The structures are selected to minimize the discrepancy between the average scattering profile of the ensemble and the experimental scattering data.

Scattering from the full length proteins were modelled using rigid-body optimization of models for the modules. The relative position of the individual modules is optimized to ensure the best agreement with the experimental scattering data. The optimization is performed using a simulated annealing protocol, where interconnection between the modules are imposed and steric clashes are avoided. The interconnection, i.e. the linker region, is determined from the amino acid sequence and is defined as the residues not found in the subunits with atomic resolution. If necessary, like between the R-modules of AlgE6, residues were deleted from the atomic resolution structures. The procedure is implemented in the program BUNCH [29]. The models obtained are not unique, due to the randomness involved in the search. Therefore a minimum of 10 runs are preformed and the individual models are aligned, compared and filtered using the programs SUPCOMP and DAMAVER [30]. In comparing the models the most representative model from the set of the models is provided.

NMR spectra

All NMR spectra were measured on Bruker Avance 600 spectrometer equipped with 5 mm Z-gradient TCI (H/C/N) cryogenic probe and on a Bruker DRX 600 spectrometer equipped with 5 mm xyz-gradient TXI (H/C/N) probe. The measurements were performed at 25 °C.

For the NMR structure calculation uniformly ^{15}N, ^{13}C labelled proteins were produced. Cells containing the plasmid pFA8, pFA12 and pFA13 were grown in double labelled M9-media as described recently [19,21]. For the structure calculations, three different types of NOESY spectra were used: a 3D-^{15}N-edited NOESY (mixing time 80ms) recorded on a sample dissolved in 95% H_2O/5% D_2O, a 3D-^{13}C-edited NOESY (mixing time 80 ms) on a sample dissolved in 99% D_2O and a 2D-NOESY (mixing time 70 ms) spectrum recorded on an unlabelled sample dissolved in 99% D_2O.

Structure calculation

NOE cross peaks were assigned and integrated manually using the program NEASY [31]. The structure calculations were performed by the program CYANA [32]. For the first structure calculations, torsion angle constraints obtained from the program TALOS [33] were used. After obtaining more refined structures of AlgE6R1, AlgE6R2 and AlgE6R3 the TALOS constraints were excluded from the calculations. Structure calculations started from 100 random conformations, the 20 final conformations with the lowest target function were selected. Ca^{2+}-ions were incorporated in the structure calculation as described in [14].

Line width

A Lorentzian line shape function was fitted to slices taken through 10 different peaks of each module of AlgE4 with the line width as variable parameter. Peaks used for the R-module were assigned as N/H^N D14, L16, G21, G46, T51, F52, E76, D83, A118 and A 131. For the A-module there is not any assignment available. Therefore, the ^{15}N-HSQC spectrum of AlgE4 was overlaid with a spectrum of the A-module. 10 peaks definitively belonging to the A-module were picked from both spectra.

$$I = \sum_x \frac{w^2}{\left(w^2 + (x-t)^2\right)}$$

Where

w line width in ppm

x point in ppm

t center of the Lorentz curve in ppm

Protein alignment and surface calculation

The amino acid sequences of AlgE4R, AlgE6R1, AlgE6R2 and AlgE6R3 were aligned by ClustalW2 [34] at the European Bioinformatics Institute (EBI). The pair wise structure alignment of the β-rolls or core proteins was performed in PyMol (2006 DeLano Scientific LLC). For the alignment the β-roll is defined from the first identical amino acid (G1 in AlgE4R, G2 in AlgE6R1, G11 in AlgE6R2 and AlgE6R3 see also Fig. 3) until the phenylalanine 54 amino acids later (F54 in AlgE4, F55 in AlgE6R1, F65 in AlgE6R2 and AlgE6R3). The core-R-modules are from the first identical amino acid until the C-terminal end of AlgE6R1 and AlgE6R2 or the amino acid that is in the same position when the four structures are aligned (T150 in AlgE4R and A161 in AlgE6R3). The structures were pair wise aligned and the Root Mean Square Deviation (RMSD) of the backbone atoms (N, C^α, C') was calculated.

The electrostatic on the surface of the different R-modules was visualized by the APBS Plugin written by Michael Lerner [35]. Ca^{2+}-ions could not be incorporated and therefore aspartic acid and glutamic acid residues pointing into the β-roll (see Fig. 3) were artificially protonated. Pdb files were transformed to pqr files by pdb2pqr [36] and these pqr files were used as input files for the APBS-calculation.

For better comparison all the R-modules in the different figure have the same orientation. The front and back side of the proteins were arbitrarily defined based on AlgE4R. On the front side of AlgE4R the alginate is binding.

Results

Sequence Alignment

The sequences of the three R-modules of AlgE6 were aligned with the R-module of AlgE4 for determining differences in the primary sequence. For purification reason the sequence of AlgE6R1 and AlgE6R3 was extended by an additional alanine at the N-terminal. The last R-module of each extracellular alginate epimerase from *A. vinelandii* has an approx. 20 amino acids long and unstructured signal peptide essential for the secretion via an ATP-binding cassette (ABC)-transporter to the extracellular environment [37]. For the following comparison neither the tail nor the N-terminal alanine were considered. The amino acid identity between the four R-modules is relatively high (75 out of 150 amino acids) (Fig. 2 and Fig. 3).

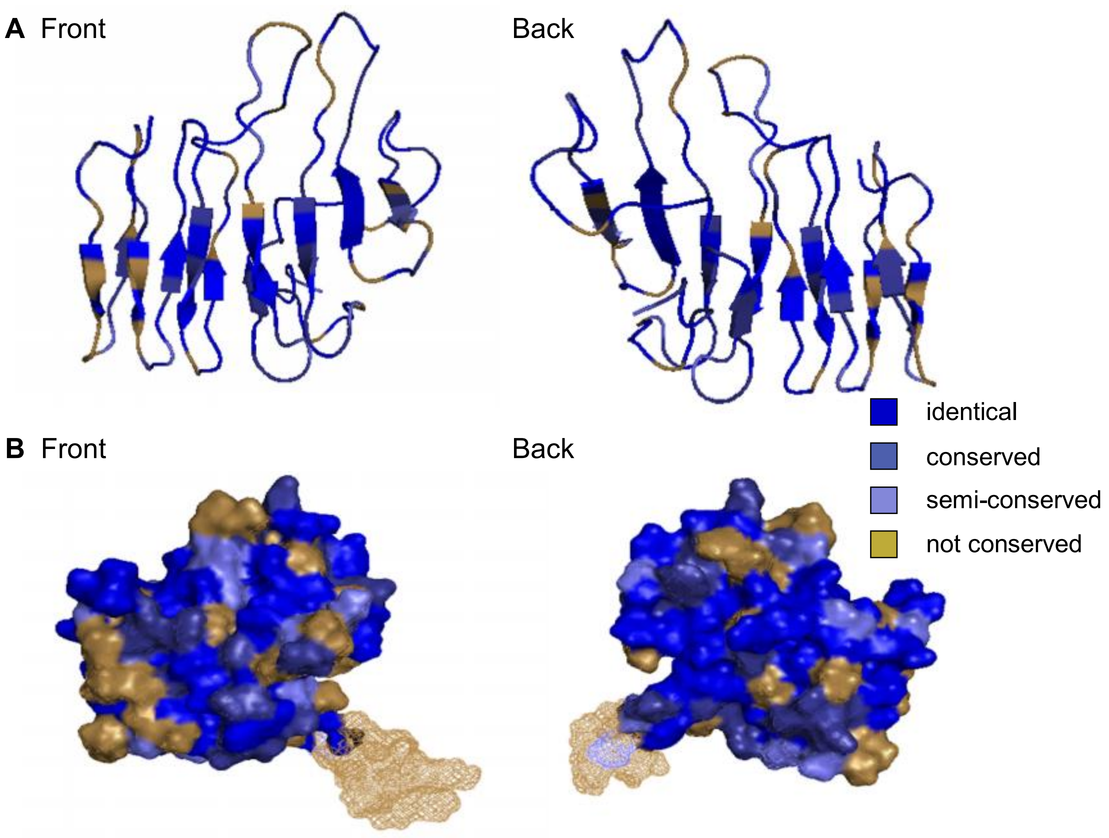


Fig. 2: Distribution of the conserved amino acids superimposed on AlgE4R. A) The structure is shown with the secondary structure elements the tail is not shown. **B)** Distribution of the conserved amino acids on the surface. The unstructured tail is indicated by a mesh-surface.

AlgE6R2 and AlgE6R3 have an N-terminal nona-sequence which is the proline rich linker between the R-modules of AlgE6.

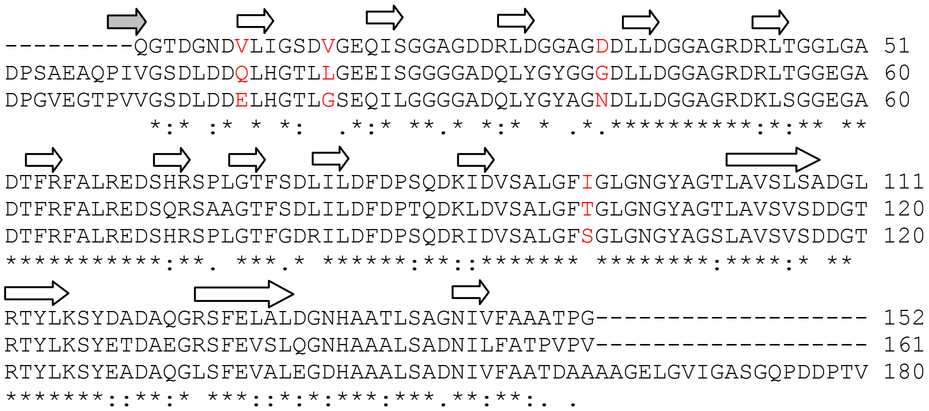
The four R-modules share 50% identity but the conserved amino acids are not equally distributed (Fig. 3). Out of the first 18 amino acids (where all four structures are aligned) only four are identical. For the next 18 amino acids the identity increases to 10. For the rest of the alignment nearly all the amino acids are conserved except two regions of five amino acids. Those regions are the first antiparallel β-hairpin and an antiparallel β-strand.

```
AlgE4R          ----------GSD-GEPLVGGDTDDQLQGGSGADRLDGGAGDDILDGGAGRDRLSGGAGA 49
AlgE6R1         ---------QGTDGNDVLIGSDVGEQISGGAGDDRLDGGAGDDLLDGGAGRDRLTGGLGA 51
AlgE6R2         DPSAEAQPIVGSDLDDQLHGTLLGEEISGGGGADQLYGYGGGDLLDGGAGRDRLTGGEGA 60
AlgE6R3         DPGVEGTPVVGSDLDDELHGTLGSEQILGGGGADQLYGYAGNDLLDGGAGRDKLSGGEGA 60
                          *:* .: * *   .::: **.* *:* * .*.*:********:*:** **

AlgE4R          DTFVFSAREDSYRTDTAVFNDLILDFEASEDRIDLSALGFSGLGDGYGGTLLLKTNAEGT 109
AlgE6R1         DTFRFALREDSHRSPLGTFSDLILDFDPSQDKIDVSALGFIGLGNGYAGTLAVSLSADGL 111
AlgE6R2         DTFRFALREDSQRSAAGTFSDLILDFDPTQDKLDVSALGFTGLGNGYAGTLAVSVSDDGT 120
AlgE6R3         DTFRFALREDSHRSPLGTFGDRILDFDPSQDRIDVSALGFSGLGNGYAGSLAVSVSDDGT 120
                *** *: **** *:  ..*.* ****:.::*::*:***** ***:**.*:* :.  . :*
```

```
AlgE4R          RTYLKSFEADAEGRRFEVALDGDHTGDLSAANVVFAATGTTT--ELEVLGDSGTQAGAIV 167
AlgE6R1         RTYLKSYDADAQGRSFELALDGNHAATLSAGNIVFAAATPG------------------- 152
AlgE6R2         RTYLKSYETDAEGRSFEVSLQGNHAAALSADNILFATPVPV------------------- 161
AlgE6R3         RTYLKSYEADAQGLSFEVALEGDHAAALSADNIVFAATDAAAAGELGVIGASGQPDDPTV 180
                ******:::**:*  **::*:*:*:. *** *::**:.  .
```

Fig. 3: Alignment of the R-modules of AlgE6 with AlgE4R. Positively charged residues are coloured in blue and negatively charged amino acids are red. HIS are labelled in cyan. ASP and GLU that point into the β-roll are marked in green. For the surface calculation these amino acids were artificially protonated. * indicates residue identical in all 4 R-modules, : denotes conserved amino acids and · shows semi-conserved amino acids (explanation for the symbols can be found atClustalW2) [34].

If only the three R-modules of AlgE6 are compared, 2/3 of all amino acids are identical. Only on four positions all three R-modules have different amino acids (Fig. 4). After the first 36 amino acids (in AlgE6R1) or 45 amino acids (in AlgE6R2 and 3), respectively, nearly every amino acid is conserved.



```
AlgE6R1         ---------QGTDGNDVLIGSDVGEQISGGAGDDRLDGGAGDDLLDGGAGRDRLTGGLGA 51
AlgE6R2         DPSAEAQPIVGSDLDDQLHGTLLGEEISGGGGADQLYGYGGGDLLDGGAGRDRLTGGEGA 60
AlgE6R3         DPGVEGTPVVGSDLDDELHGTLGSEQILGGGGADQLYGYAGNDLLDGGAGRDKLSGGEGA 60
                          *:* :* * *:  .*:* **.* *:* * .*.**********:*:** **

AlgE6R1         DTFRFALREDSHRSPLGTFSDLILDFDPSQDKIDVSALGFIGLGNGYAGTLAVSLSADGL 111
AlgE6R2         DTFRFALREDSQRSAAGTFSDLILDFDPTQDKLDVSALGFTGLGNGYAGTLAVSVSDDGT 120
AlgE6R3         DTFRFALREDSHRSPLGTFGDRILDFDPSQDRIDVSALGFSGLGNGYAGSLAVSVSDDGT 120
                ***********:**. ***.* ******:**::******* ********:****:* **

AlgE6R1         RTYLKSYDADAQGRSFELALDGNHAATLSAGNIVFAAATPG------------------- 152
AlgE6R2         RTYLKSYETDAEGRSFEVSLQGNHAAALSADNILFATPVPV------------------- 161
AlgE6R3         RTYLKSYEADAQGLSFEVALEGDHAAALSADNIVFAATDAAAAGELGVIGASGQPDDPTV 180
                *******::**:* ***::*:*:***:***.**:**:.  .
```

Fig. 4: Alignment of the 3 R-modules of AlgE6. The arrows indicate β-strands. The grey arrow shows the first β-strand of AlgE6R2 and AlgE6R3. The amino acids marked in red differ in all 3 R-modules. * indicates residue identity, : means conserved amino acids and · shows semi-conserved amino (definition for conserved and semi-conserved can be found in ClustalW2) [34].

AlgE6R2 and AlgE6R3 are most similar to each other and share 77% amino acid identity.

Structure determination

The structures of 3 R-modules of AlgE6 were determined by NMR on the basis of NOE upper distance limits. The experimental data for the three structures are summarized in Supp. Tab. 1. The structures of the three R-modules of AlgE6 are very similar to each other and to AlgE4R (Fig. 5).

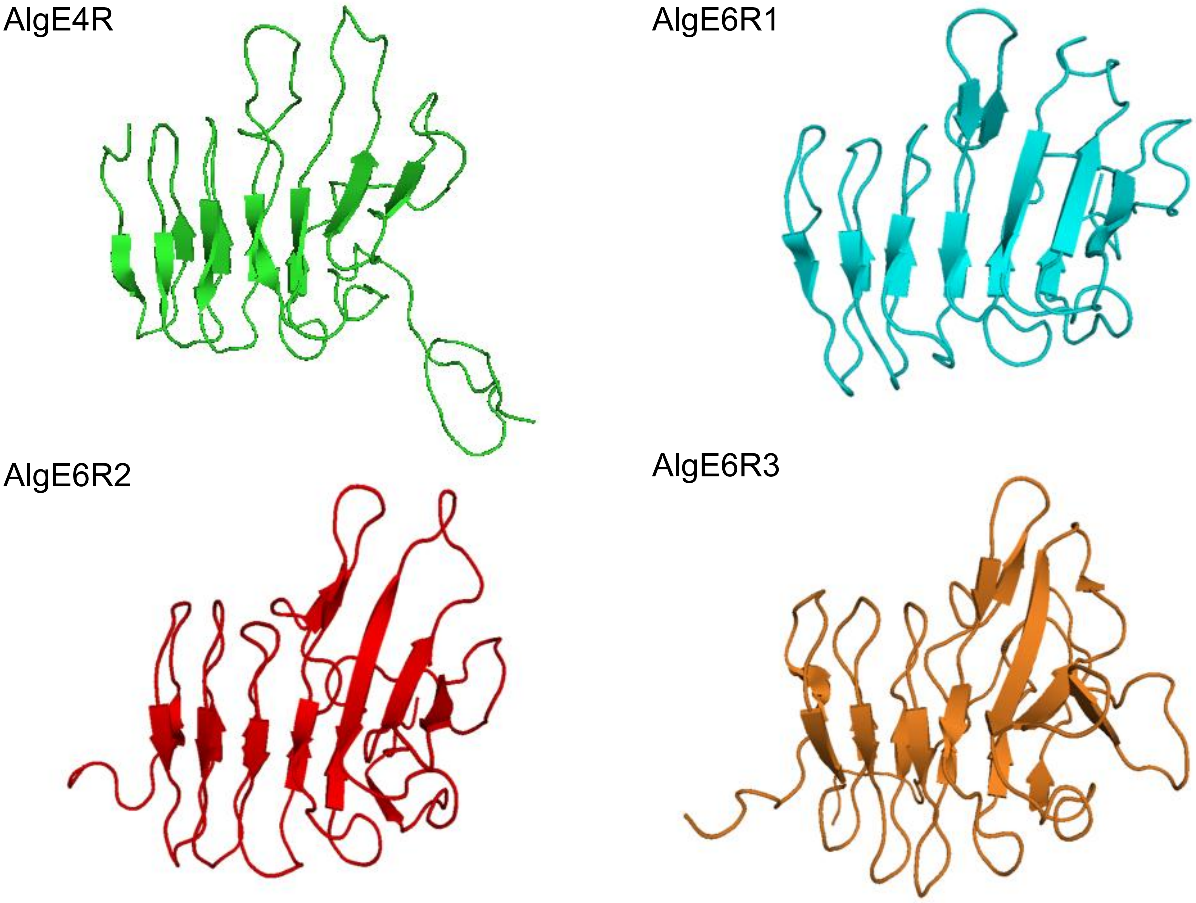


Fig. 5: Structure comparison of AlgE4R and the R-modules of AlgE6. The β-strands are indicated by arrows (see also Fig. 4). The structures are very similar to each other.

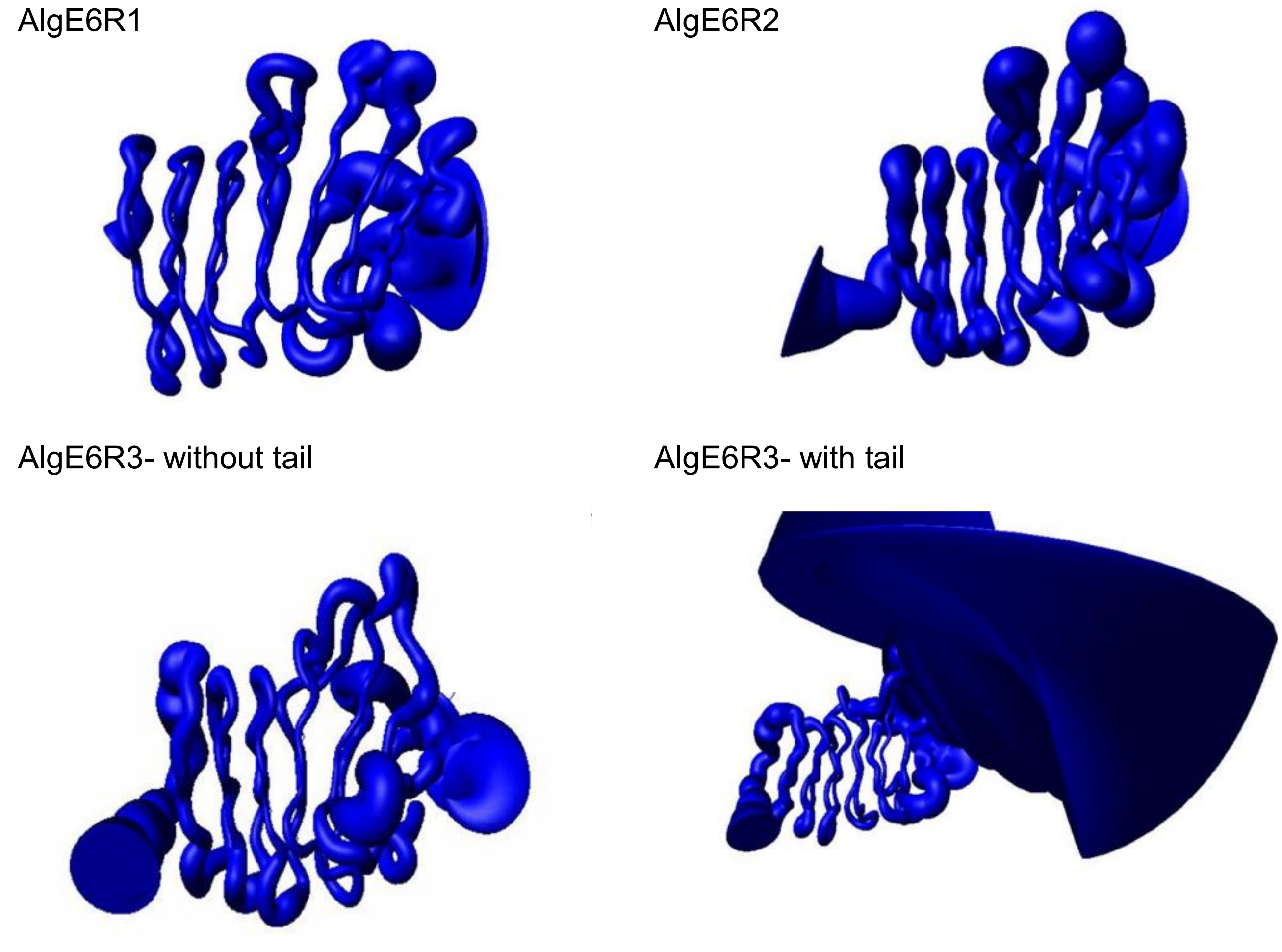


Fig. 6: Average structures of the AlgE6 R-modules plus the degree of flexibility based on the 20 structures with the lowest target functions. Although the β-roll has only short β-strands and long loops (three amino acids in β-strands and six amino acids in loops) the structure is relatively stiff. The most flexible areas can be found in the loops and between 91-100 amino acids for AlgE6R1 (100-109 amino acids for AlgE6R2 and AlgE6R3 respectively). It is a crossover between the β-roll and the first antiparallel β-strand. The flexible tail of AlgE6R3 overshadows the whole structure therefore the structure of AlgE6R3 is shown twice with and without the tail.

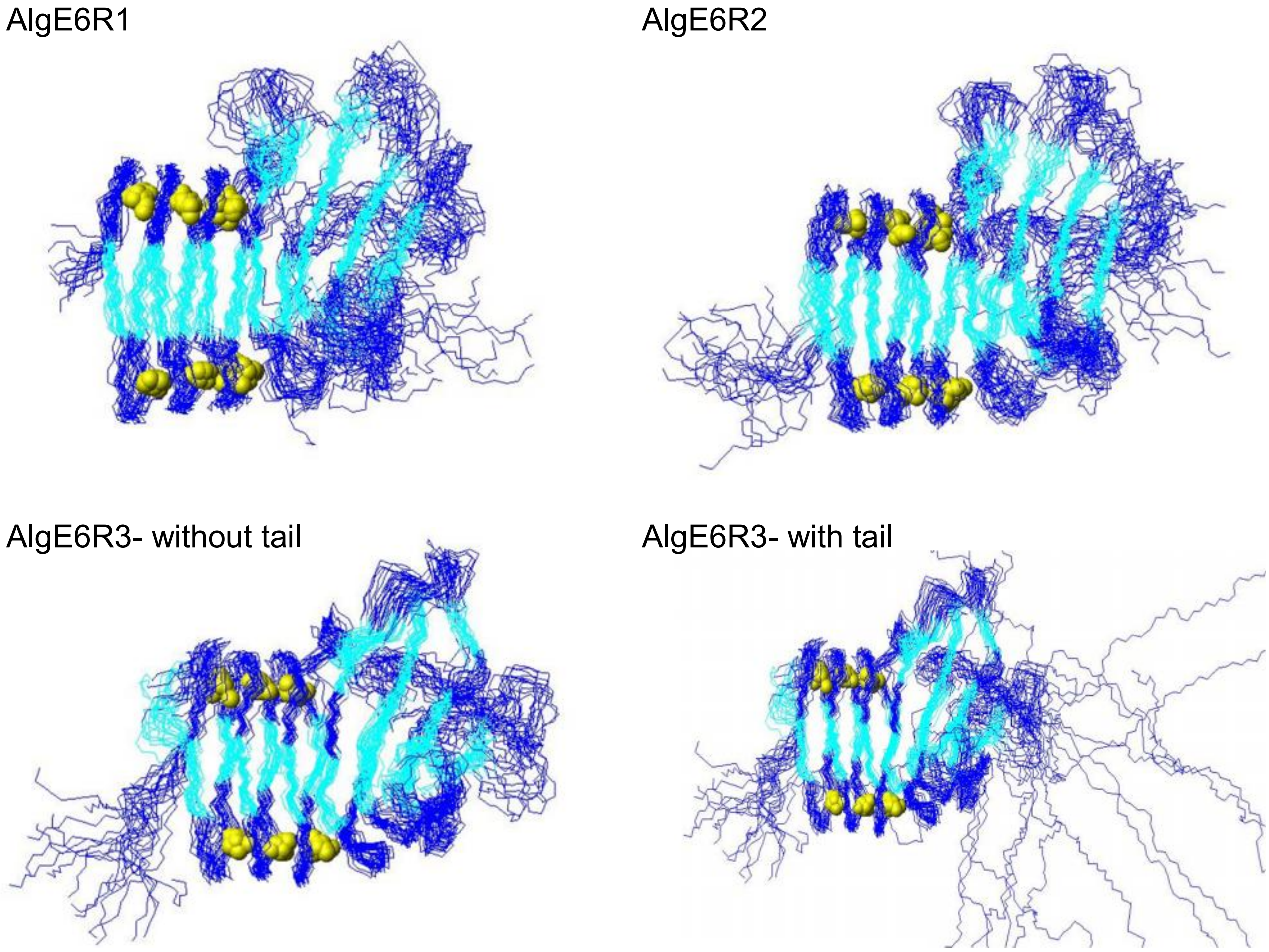


Fig. 7: The overlay of the best 20 structures with the lowest target function and Ca^{2+}-ion incorporation in the loops of the β-roll. The β-strands are shown here in cyan and the Ca^{2+}-ions are yellow spheres. AlgE6R3 is shown with and without tail.

AlgE6R2 and AlgE6R3 have nine amino acids in the beginning which is the linker between the R-modules. The first 54 amino acids (of AlgE6R1) or 63 (AlgE6R2&3), respectively, form a right handed parallel β-roll where each complete turn consists of 18 amino acids or two RTX-motifs (Repeat in Toxin-motif) [38]. The sequence of the RTX-motif consists of the nonapeptide GGXGXDZUZ. The glycine and aspartic acid residues are highly conserved; the U is a large hydrophobic amino acid, which normally is leucine and sometimes replaced by isoleucine, valine or phenylalanine. X stands for any amino acid but mainly with short side chain and Z is for an amino acid with long side chain. The first six amino acids form a tight loop which also binds a Ca^{2+} ion whereas the last three amino acids form short β-strand (Fig. 7). The compact structure of the β-roll is opened by the sequence FRF (53-55 in AlgE6R1, 63-65 in AlgE6R2 and AlgE6R3) followed by an anti-parallel hairpin structure. The chain folds back and make a new but less defined turn extending the β-roll. The next 10 amino acids are less well defined followed by three long antiparallel β-strands (Fig. 6). The last β-strand is also parallel to the penultimate β-strand of the β-roll. There is an additional long loop and the last β-strand completes the β-roll. The last 21 amino acids of AlgE6R3 are unstructured and it is known that this is a signal peptide essential for secretion [37].

The R-modules have the same secondary structure elements. The major differences are in orientation of the hairpin structure and the orientation of the antiparallel β-strands relative to the β-roll (Fig. 5).

If only the β-rolls (see Material and Methods) were compared the Root Mean Square Deviation (RMSD) is between 1.42 and 2.18. For the pair wise alignment of the core-R-modules the RMSD is between 2.07 until 3.76. The compact structure of the β-roll results in low RMSDs but also comparison of the structures of the R-modules (core-R-modules) results in relative low the RMSD (also compare to the RMSD of the single R-module Supp. Tab. 1).

The electrostatic surfaces of the R-modules also show some differences (Fig. 8). The electrostatic of the surface of the R-modules is important as alginate is a polyanionic at neutral pH. Negative charges on the surface of the proteins would reject the alginate. In general the front side is positively charged while the electrostatic surface on the back side is rather negatively charged. The electrostatic surface of AlgE6R1 is similar to the surface of AlgE4R. AlgE6R2 and AlgE6R3 do not show as clear charge separation as AlgE6R1. The charges on the surface of AlgE6R2 are homogeneously distributed. AlgE6R3 shows a groove on the front side that shows a positive electrostatic surface. This groove is made up by Arg51, Lys53, Arg82 and Arg121, whereof all except Arg82 are conserved throughout the R-modules. The other R-modules have a leucine on this position.

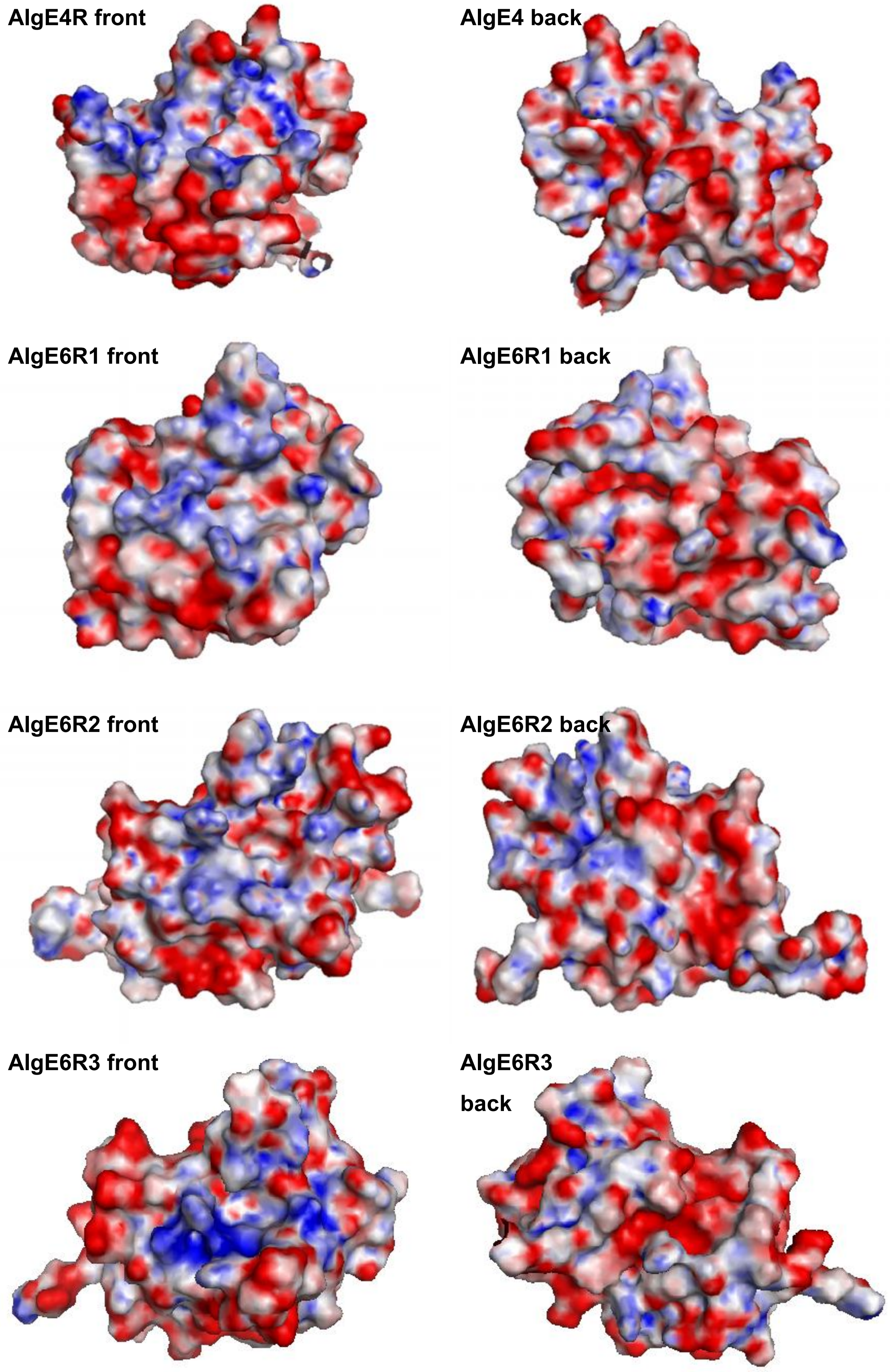


Fig. 8: Charge distribution on the surface. ASP and GLU marked in green in Fig. 1 were artificial protonated before surface electrostatics calculation. The same colour code was used in all four structures and positively charged surfaces are in blue and negatively charged in red. From the charge distribution AlgE4R and AlgE6R1

are most similar. AlgE6R2 is slightly positively charged on both sides. AlgE6R3 has a strongly positively charged groove on the front side.

Line width

The line width of 10 peaks belonging to the A-module or to the R-module, respectively, in a U-^{2}H, ^{15}N-AlgE4 were measured. For comparison the line width of the same 10 peak of the A-module were measured in a U-^{2}H, ^{15}N-AlgE4A. The average line width of the peaks from the R- module is 6.03 ± 0.72 Hz and is significant narrower than from the A-module (7.89 ± 1.37 Hz) supporting the hypothesis that the modules have a flexible linker between them. Nevertheless the average line width of the A-module in the whole AlgE4 is relative narrow considering the molecular size of 35 kDa. Another indication of a flexible linker between the A- and the R-module in the whole AlgE4 is the intensity of the peaks. The area and height of a peak from the R-module is approximately five times bigger than from the A-module also indicating that the A- and R-module relax with different rates (**Fig. 9**).

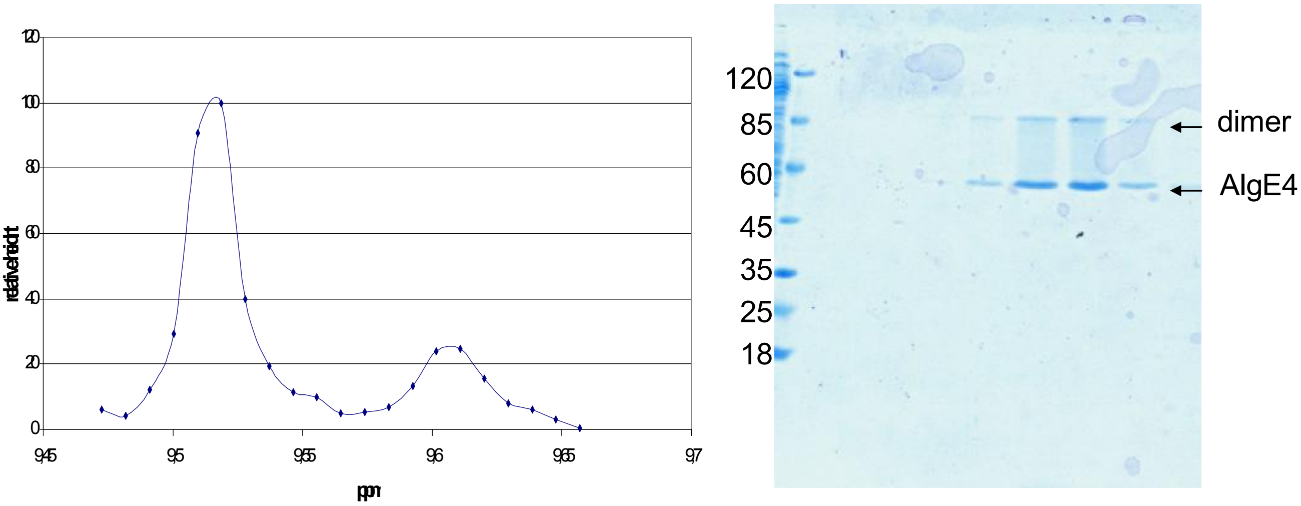


Fig. 9: Comparison of the peak intensity from the R- and A-module of ^{2}H,^{15}N-AlgE4. The peak belonging to the R-module is approximately five times higher than the peak from the A-module. Neither SDS gel nor MS-data (not shown) show any degradation of AlgE4 into the A- and R-module.

In silica analysis of the linker region between A- and R-module of AlgE4 predicted no secondary structure elements [39,40] and the linker region is relative proline rich (GEPGATPQQPST).

Module orientation

Small-angle X-ray scattering data (Fig. 10) was collected for the R- and A-modules separately, and for the full length AlgE4. The scattering intensity, $I(q)$, is displayed as a function of the modules of the scattering vector, q. The pair distance distribution function $p(r)$ was obtained by indirect Fourier transformation of the scattering data, and the following parameters were determined: The maximum diameter, D_{max}, the radius of gyration, R_{g}, and the forward scattering, $I(q = 0)$. The forward scattering allows the determination of the molecular mass of the macromolecules in solution. The results for the full length AlgE4 and the R- and A-modules (Tab. 2) show that these are present in there monomeric state in solution. To further investigate the solution structure of the R- and A-modules of AlgE4 *a priori* information was utilised. Atomic resolution structures are know for the two modules and these were used to calculated the theoretical scattering pattern and compare these to the experimentally obtained scattering data using the program CRYSOL [27]. The fits agree nicely with the scattering data, with reduced χ^2 values of 1.91 for the R-module and 1.94 for the A-module. Thus the two protein modules have a structure in solution in agreement with the atomic resolution models.

The solution structure of full length AlgE4 was determined from the SAXS data by rigid body modelling using the modules structures with atomic resolution. These were connected by dummy residues to mimic the linkers between the different modules. The optimization of the modules relative to each other was preformed using the program BUNCH [29] that uses a simulated annealing procedure where steric clashes are avoided and the modules are interconnected. Multiple BUNCH runs were performed and only one population of structures was found. The models were compared and averaged using the program packaged DAMAVER [30]. Here the most representative model is also determined, defined as the model having the highest degree of similarities to all the other models. This model yields a good fit with a reduced χ^2 of 1.23 (Fig. 10).

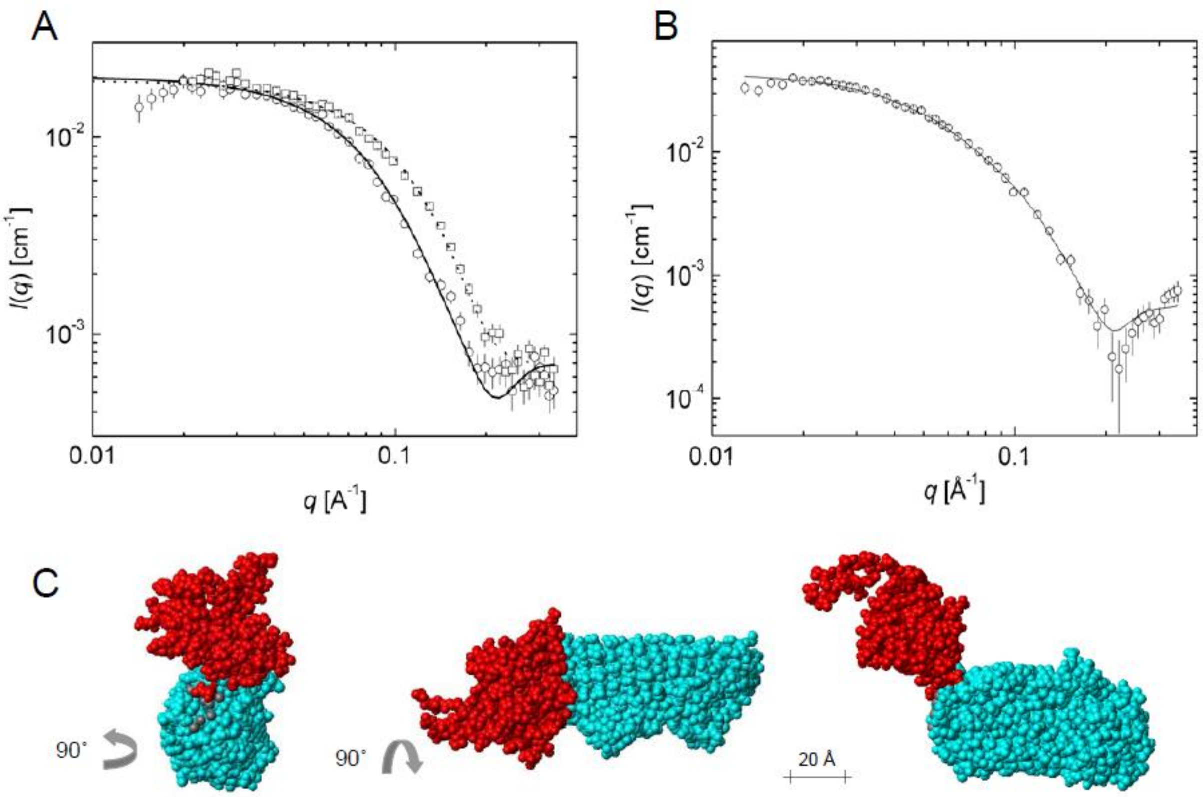


Fig. 10: A) Scattering data obtained for the A (circle) and an R (square) module of AlgE4 with their respective CRYSOL fits (solid and dashed line). B) Scattering data for AlgE4 (circle) with the corresponding fit obtained through rigid body modelling (solid line). C) Most representative model obtained through rigid body modelling, with the A module (cyan), the R module (red). The connecting residues are displayed as dummy residues (gray beads).

Tab. 2: Results obtained from the SAXS data by performing indirect Fourier transformation. Molecular mass was calculated from *I*(0) as described in the method section.The calculated molecular mass was determined by ProtParam [41].

Protein	R_g [Å]	D_{max}[Å]	$M_w^{I(0)}$ [kDa]	M_w^{Cal} [kDa]
AlgE4	31.3 ± 0.5	100 ± 10	53 ± 6	57.7
AlgE4A	22 ± 0.3	65 ± 10	44 ± 5	39.9
AlgE4R	18.7 ± 0.2	55 ± 5	17 ± 2	17.0
AlgE6	55.4 ± 1.1	180 ± 10	100 ± 10	90.2
AlgE6R1	15.6 ± 0.2	50 ± 5	16 ± 1.5	15.5
AlgE6R2	16.0 ± 0.8	50 ± 5	20 ± 2	16.5
AlgE6R3	16.6 ± 0.2	50 ± 5	21 ± 3	18.2

SAXS data was also collected for full length AlgE6 and its individual R-modules. From the initial data analysis the molecular weights and the $D_{\max}$ values suggests presence of monomeric proteins in all solutions (Tab. 2). However for the R modules data below q = 0.02 Å^{-1} was discarded due to the upturn of the data by the presence of large aggregated in the solution. To further investigate the solution structures of the individual modules the theoretical scattering patters from the available models with atomic resolution, for the A- and R-modules, is calculated using the program CRYSOL [27]. From the fits (Fig. 11) it is evident that the R1- and R2-modules have a shape in solution similar to that of their models with atomic resolution. The respective obtained reduced χ^2 values are 1.73 and 4.60. However, for the R3-module a clear difference is found between the theoretical scattering pattern and the experimental data with a reduced χ^2 value of 37.1. The poor quality of the fit originates from the large discrepancy between data and fit at high q above 0.15 Å^{-1}, where the model underestimates the data. This additional intensity could possibly originate from fluctuation scattering, which is observed for flexible protein structures. To investigate this possible explanation the additional intensity contribution was described by the scattering from a Gaussian chain [42]. The addition of scattering from a Gaussian chain increased the fit quality substantially from χ^2 37.1 to 2.05 (Fig. 11). Thus some flexibility should be present in structure which agrees well with the fact that the R3-module has a 20 residue long flexible tail.

To investigate the flexibility of the R3-module further the ensemble optimization method (EOM) [28] was applied. In EOM a large ensemble of structures are generated and a subset of these are selected by a generic algorithm to best fit the scattering data. With this method a good agreement between the model and the scattering data was obtained with a reduced χ^2 of 0.95 (Fig. 11).

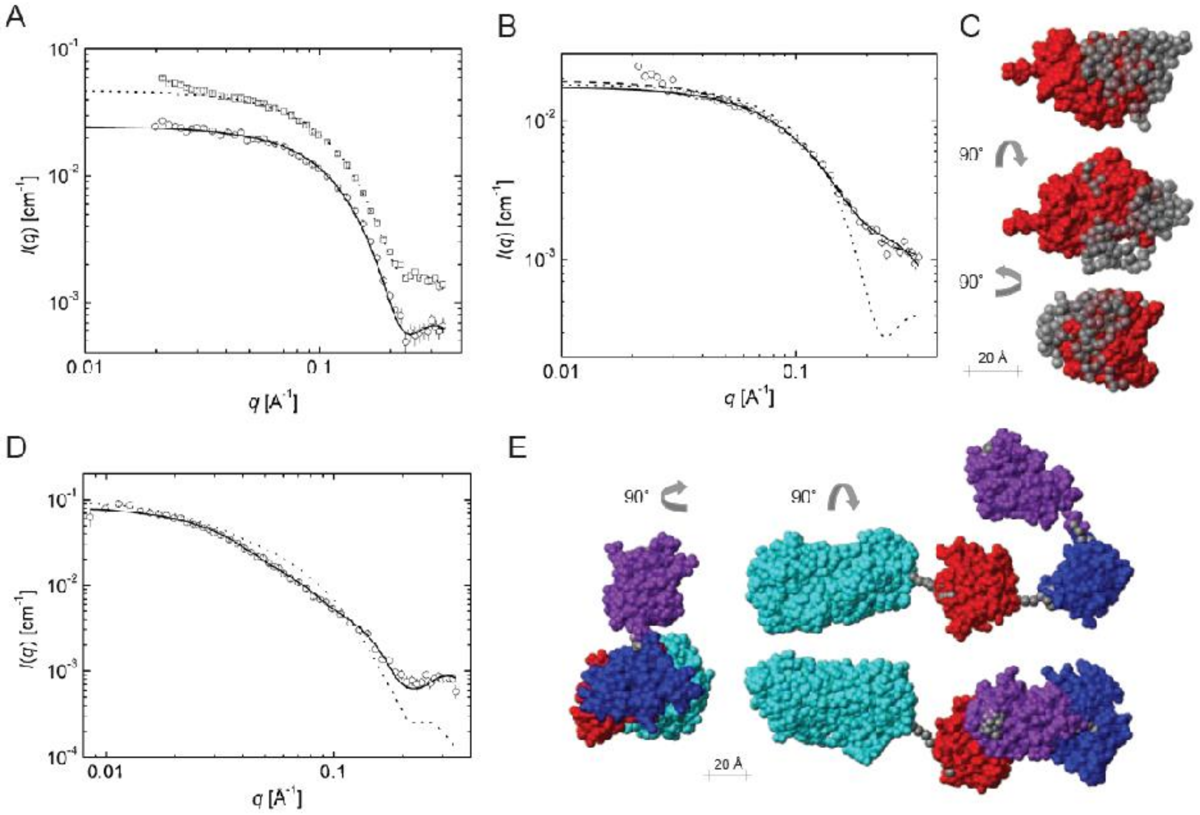


Fig. 11: A) Scattering data obtained for the R1 (circle) and R2 module (square) with their CRYSOL fits, solid and dashed line respectively. B) Scattering data obtained for the R3 module with the CRYSOL fit (dashed line), CRYSOL fit added scattering for a Gaussian chain (gray line) and EOM model fit (black line). C) EOM models where the R3 module is displayed in red and the possible placement of the 20 residue tail is displayed in semitransparent gray beads. D) Scattering data obtained for AlgE6 and the corresponding fit from the rigid body modelling (solid line) in addition to theoretical scattering data AlgE6 in its fully stretched structure (dashed line) . E) Most representative model obtained through rigid body modelling, with the A module (cyan), the R1 (red), R2 (blue) and R3 module (purple). The connecting residues and the 20 residue tail of the R3 module is displayed as dummy residues (gray besds).

Investigation of the full length AlgE6 in solution was preformed by rigid body modelling of the SAXS data, as it was done for full length AlgE4. In addition to the exploration of the module structures with atomic resolution, the 20 residue flexible tail of the R3-moduel was included as dummy residues. The optimization of the modules relative to each other and comparison of multiple runs were performed in the same manner as for AlgE4. Only one population of structures was found and the most representative model yields a good fit with a reduced χ^2 value of 1.58 (Fig. 11). A comparison of the SAXS data with a model for AlgE6 in a fully stretched state was also done by calculating the theoretical scattering by CRYSOL of such a model (Fig. 11D). From the fit quality it is evident that this structure can not describe the experimental scattering data satisfactorily, yielding a reduced χ^2 value of 20.6. The result of the stretched AlgE6 was compared to the χ^2 obtained in the rigid body modelling which is close to perfect agreement with the experimental data.

Alginate binding studies:

The A- and R-module of AlgE4 binds alginate [13,14]. Binding studies between AlgE4R and alginates that varied in length and composition were performed by NMR and ITC. From the NMR data the chemical shift change of each single amino acid was recorded as described above. Fig. 12 shows the affected amino acids in the protein sequence versus the used alginate oligomer. Most of these amino acids are located in three clusters. The first cluster reaches from residue 38 to 43. The second cluster reaches from residue 61 to 71. In this cluster three amino acids disappear from the spectrum. The last cluster spreads between amino acids 101 and 136 but not every amino acid is affected. The length and type of alginate oligomer has little effect on the amino acid distribution. Fig. 16 shows the position of those amino acids plotted on the tertiary structure of AlgE4R. Most of the amino acids experience major shift changes are clustered on the front side of AlgE4R at the antiparallel β-hairpin and the long antiparallel β-strands.

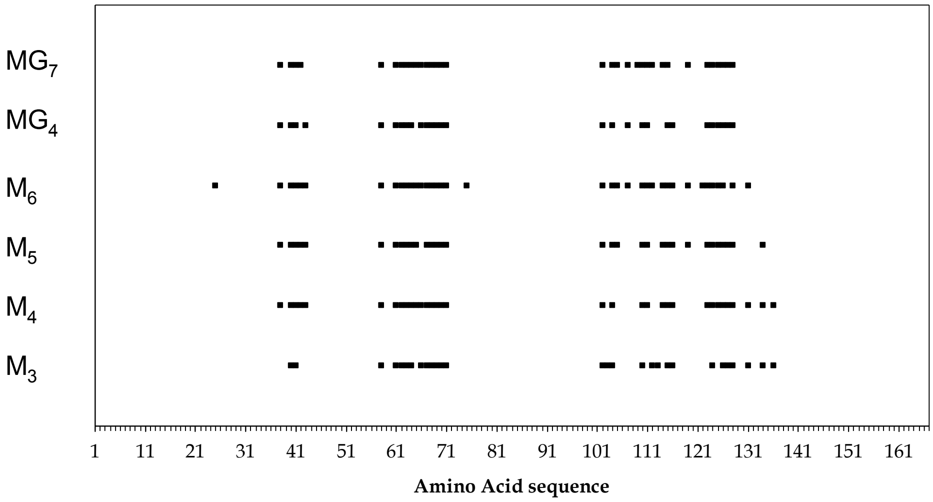


Fig. 12: Interaction patterns for the alginate on the R-module measured by NMR. Amino acids that experience a chemical shift change $|\Delta\delta_{bound}| > 100$ Hz are plotted for the alginate oligomers that were used in the experiments. The alginate oligomer type and length has hardly any influence on the location of the binding site.

From the NMR and ITC data, dissociation constants for different oligomers of alginate were calculated. The dissociation constants are summarized in Supp. Tab. 2 and Fig. 14. Dissociation constants calculated from the NMR data were for a 1:1 complex. The K_d obtained from the ITC data were calculated twice. For one calculation the binding ratio was an additional parameter to be optimized whereas for the second time the binding ratio was fixed for a 1:1 complex. Nearly all the thermograph could be fitted with the fixed binding ratio. The dissociation constants obtained from NMR data and ITC are similar (Fig. 14). For oligo-mannuronan alginate the highest binding energies were found for M5 and longer. For M8 and higher the data could not be fitted to one binding event and titration with M8 and M9 showed two individual binding events. (Fig. 13)

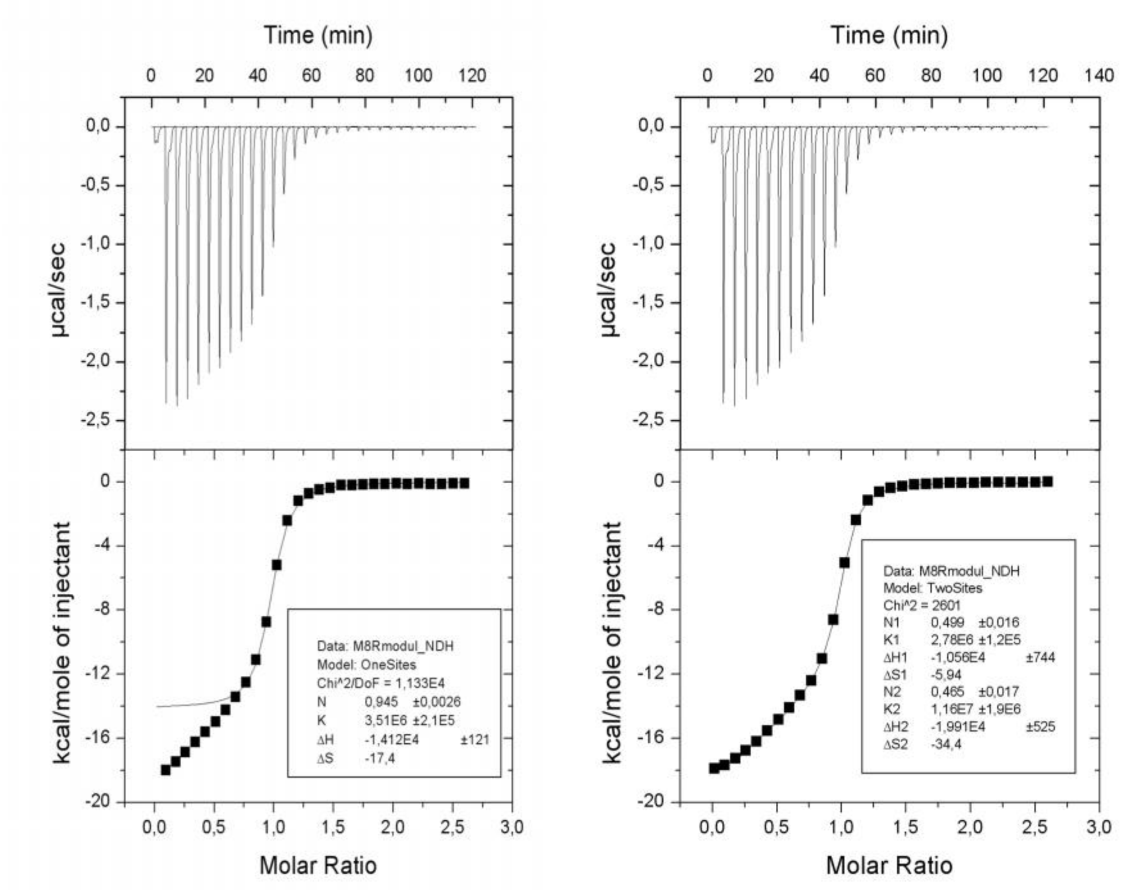


Fig. 13: ITC thermograph. The first two sheets show the experimental output of M8 and below the fitting. For the first fit the data points in the beginning were omitted to get one binding event. Therefore the fitted curve and the experimental data do not coincide over the whole experiment. The second fit shows the result of two individual binding events. The experimental and calculated data fit well.

The R-module of AlgE4 binds better to poly-M than to MG-blocks with the same size. In the case of MG-oligomers most of the ITC performance conditions were suboptimal therefore several binding ratios could be fitted to the ITC thermograph. However, setting the binding ratio to unity did not improve the results for MG5 and MG8. The enthalpy and entropy are unusually strongly negative. (Supp. Tab. 2) These measurements should be performed with 4-5 mM R-module in order to obtain reasonable result with ITC, which is a quite high amount of protein.

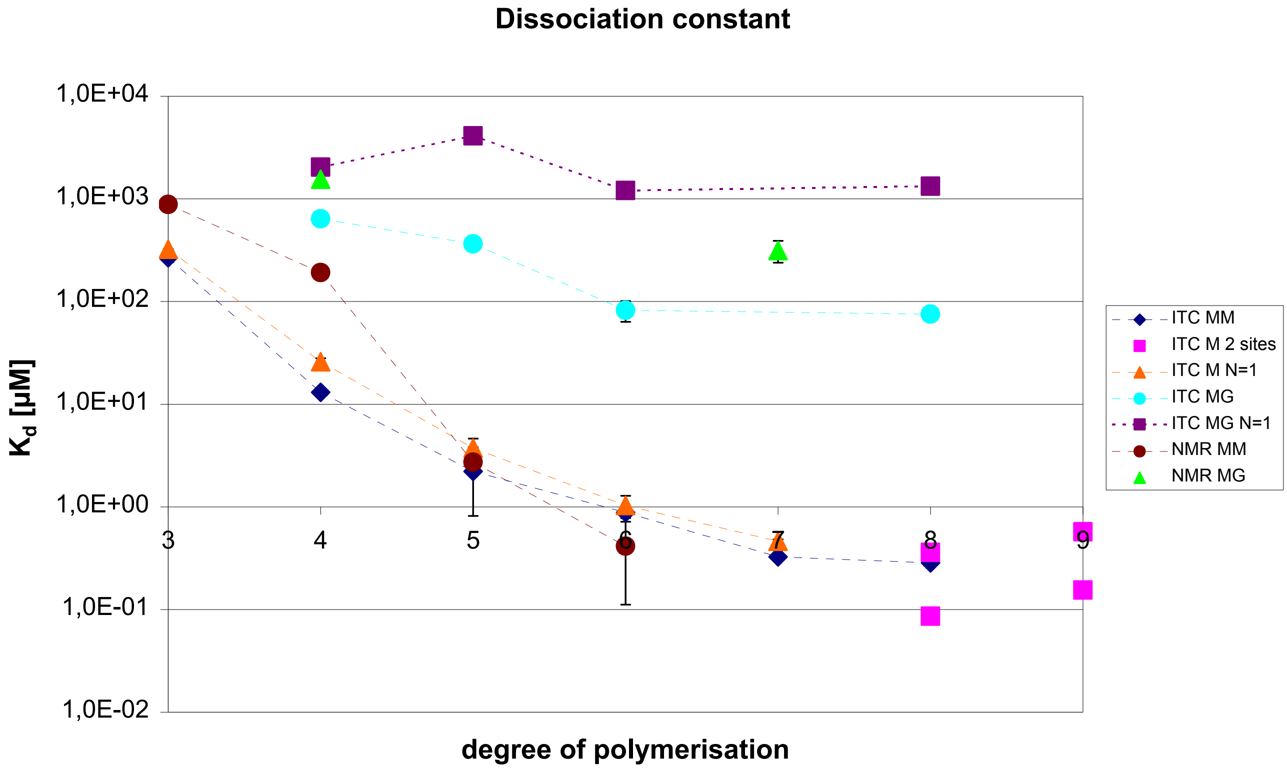


Fig. 14: **Results for interaction between R-module and alginate obtained with ITC and NMR.** The graph shows K_d obtained from the ITC and NMR measurements. In general MG oligomers have higher dissociation constants than poly-M at the same degree of polymerisation. NMR data fit well to the data obtained by ITC. The K_d depends also on the degree of polymerisation. Until M5 the increase of alginate oligomer by one sugar unit causes the decrease of K_d by 10 fold. M5 and higher have more similar dissociation constants.

Oligo-M, MG and G- alginates were tested with each single R-module of AlgE6 by either NMR or ITC. These data revealed that none of the single R-modules of AlgE6 binds to any alginate (Fig. 15). This was very surprising as the sequence and structural identity AlgE4R and the R-modules of AlgE6 is very high (Fig. 2 and Fig. 3).

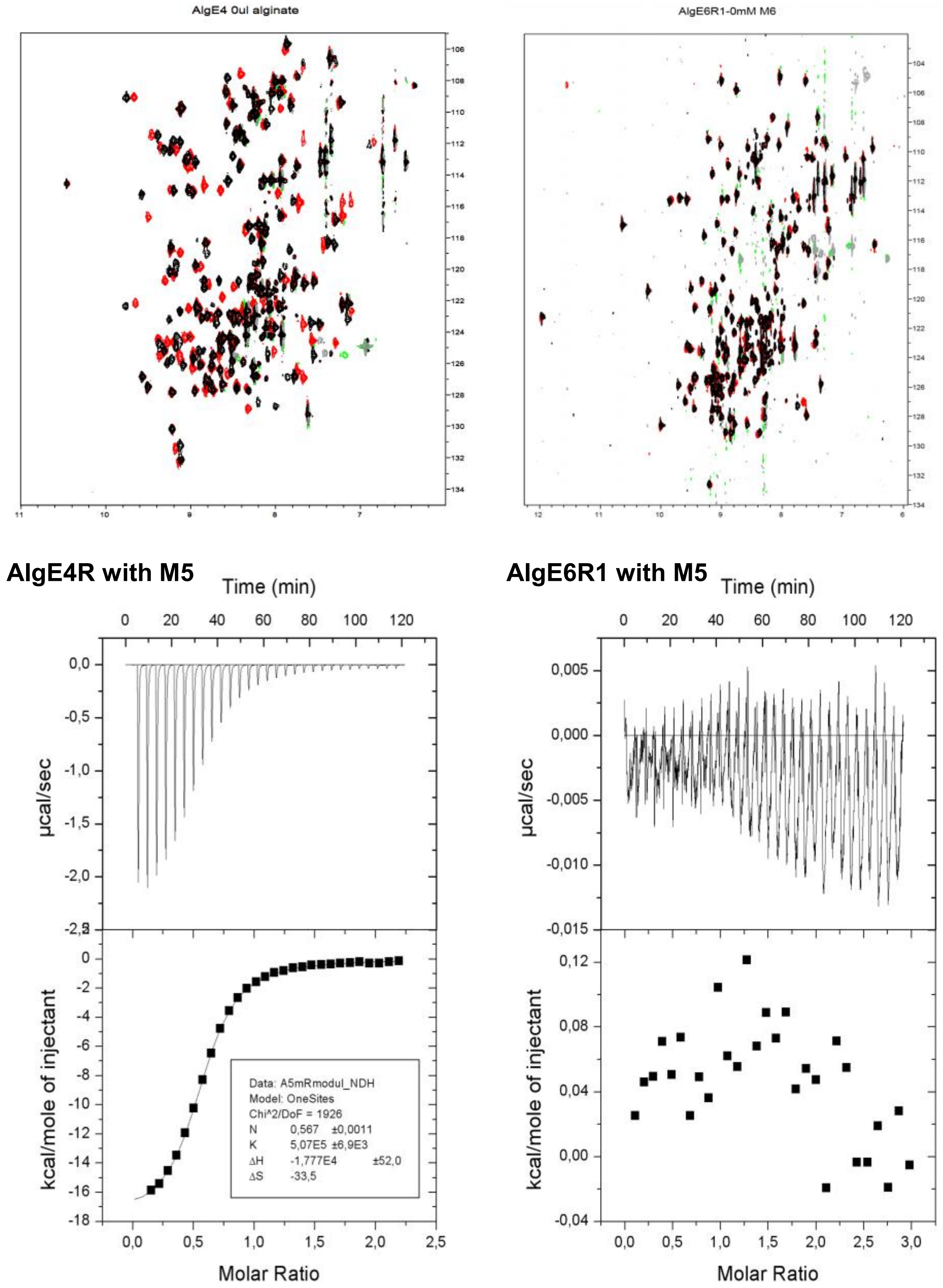
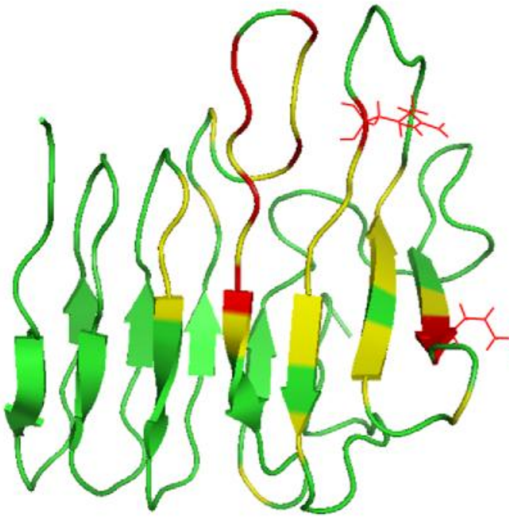


Fig. 15: Comparison of the alginate titration with AlgE4R and AlgE6R1. The HSQC spectra of the R-modules without alginate are in red while the black peaks represent the spectra of the R-modules with saturated amount of alginate. In the case of AlgE4R large chemical shift changes are observed whereas the changes in the spectra of AlgE6R1 are minimal. ITC data showed the same results. AlgE4R binds strongly to alginate while AlgE6R1 does not bind at all.

Of the amino acids of AlgE4R that experience the greatest chemical shift changes, most of them are in the front side where the binding of the alginate takes place and are conserved in all four structures. (Fig. 16).



B

```
GSDGEPLVGG DTDDQLQGGS GADRLDGGAG   30
DDILDGGAGR DRLSGGAGAD TFVFSAREDS   60
YRTDTAVFND LILDFEASED RIDLSALGFS   90
GLGDGYGGTL LLKTNAEGTR TYLKSFEADA  120
EGRRFEVALD GDHTGDLSAA NVVFAATGTT  150
TELEVLGDSG TQAGAIV                167
```

Legend:

🟩 Conserved but the side chain is not pointing to the binding site

🟨 Conserved and alginate interaction possible

🟦 Not conserved and the side chain is not pointing to the binding site

🟥 Not conserved but alginate interaction possible

Fig. 16: Amino acids that experience |Δδ_{bound}|>100 Hz are highlighted here. Yellow labelled amino acids are conversed in all four R-modules while red ones are not. Most of these conserved amino acids are clustered on the front side of AlgE4R. The red labelled amino acids are in more flexible areas or next to conserved amino acids. AlgE4R has two additional basic amino acids that are affected by binding. These Lys103 and Arg124 are shown as side chain. All R-modules of AlgE6 have serines at those positions. **B)** Approximately 40 amino acids are affected by alginate binding. Half of them can only be influenced indirectly because the side chain is pointing away from the possible binding site (green for conserved and blue for non- conserved amino acids). Amino acids side chains that are on the surface of AlgE4 are label yellow (for conserved) and red (not conserved).

Approximately half of the amino acids that are experiencing large chemical shift changes can not be involved in alginate binding directly (Fig. 16 B blue and green) as their side chains are pointing away from the possible binding site. From the remaining 16 amino acids 11 are conserved (Fig. 16 B yellow) of which five have basic and two have polar side chains. From the five amino acids, that are not conserved but probably involved in alginate binding, two are basic (Lys 103 and Arg 124) in AlgE4. The R-modules of AlgE6 have serines at those positions. At the position of Val 67 all R-modules of AlgE6 have a threonine. Leu 71 is conserved in AlgE6R1 and AlgE6R2 but AlgE6R3 has an arginine on this position. At the position of Asn 69 AlgE6R1 and AlgE6R2 have a serine and AlgE6R3 has a glycine.

The binding between alginate and the R-module of AlgE4 seems to be electrostatic as from the 16 residues that maybe interact with alginate are seven basic and three polar. The R-modules of AlgE6 have only five to six basic but five to six polar amino acids at the same positions nevertheless alginate binding is not measurable from the single R-modules.

On the possible binding site of AlgE4R are eight basic (R24, R40, R42, R62, K103, R110, K114 and R124), eight acidic (D26, D74, D94, E107, E117, D119, E126 and D130) and six polar amino acids (S44, N69, Y96, N105, T109 and Y112) located. AlgE6R1 has six positively charged (R26, R42, R44, R64, R112 and K116), seven negatively charged (D24, D28, D76, D109, D119, E128 and D132) and nine polar (T46, T69, S71, Y98, S105, S107, Y114, Y118 and S126) amino acids that can be involve in alginate binding. AlgE6R2 has the most polar residues (Q17, Q35, Y37, T55, T78, S80, N108, Y107, S114, S116, T120, Y123, Y127, S135, S139 and Q141), the least basic (R51, R53, R73, R121 and K125) and acids amino acids (D15, D85, D118, E128, D130 and E137) on the front side. AlgE6R3 has eight acidic (D15, E17, D85, D118, E128; D130, E137 and E141), six basic (R51, K53, R73, R82, R121 and K125) and nine polar (Q35, Y37, Y39, S55, T78, S114, S116, T120, Y123, Y127 and S135) residues on the supposable binding site (Fig. 17).

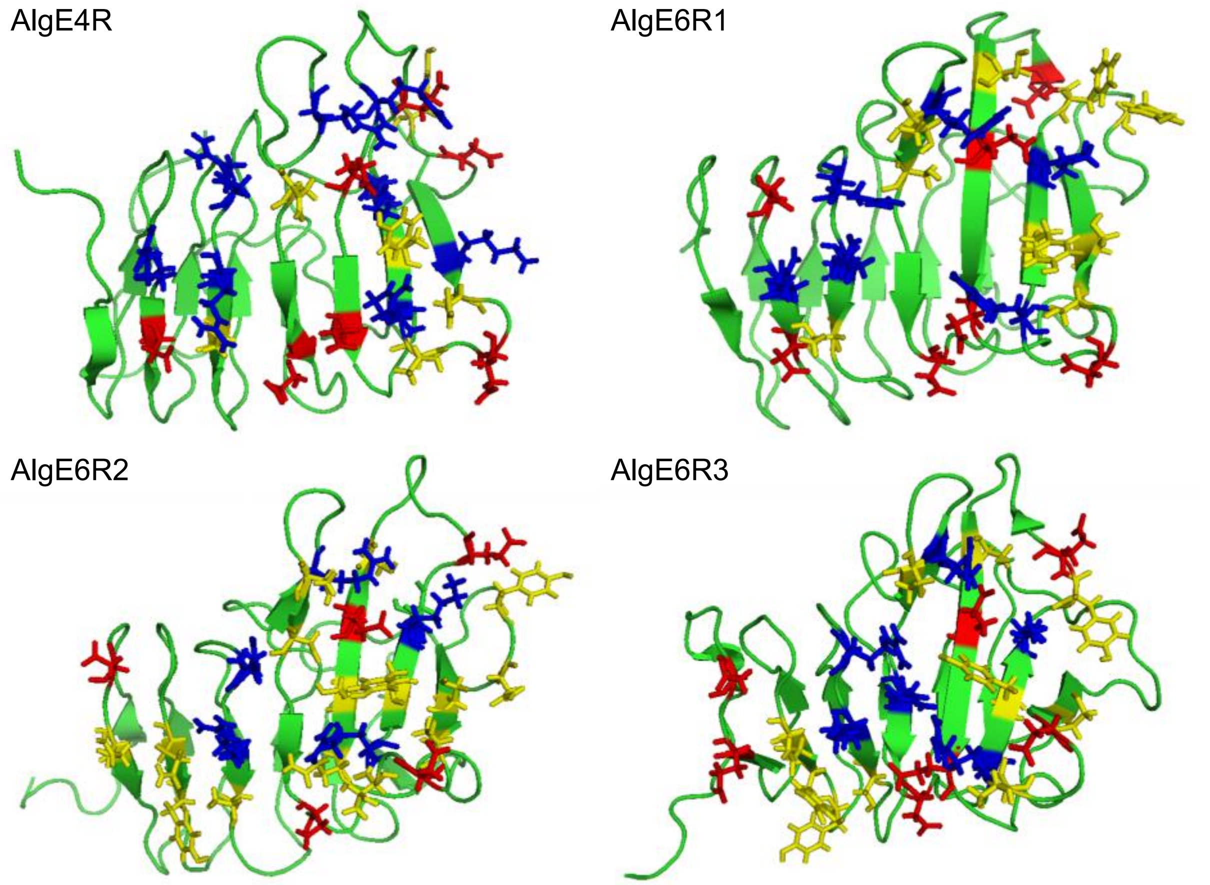


Fig. 17: Charged and polar amino acids on the assumed alginate binding site. Each R-module has 5-7 basic amino acids (in blue) and 5-7 acidic amino acids (in red) those side chains are on the surface. Polar amino acids are in yellow. AlgE4R has the more basic amino acids but the R-modules of AlgE6 have more polar amino acids.

Neither the amino acid distribution nor electronegative surface of the different R-modules can explain why alginate binding could not be detected on the individual R-modules of AlgE6.

Discussion

Structure

The overall structures for the R-modules of AlgE6 and AlgE4R are nearly identical, which is the result of the high primary sequence identity and similarity for the four R-modules. It has been shown by radioactive assay with ^{45}Ca [15,43] that R-modules bind calcium. There is a high sequence- and structure similarity between the initial β-roll of the R-modules and the RTX-domain of the metalloproteases from *Pseudomonas aeruginosa* (PDB code: 1KAP) and

Serrantia marcescens (PDB: 1SAT), which coordinate one Ca^{2+}-ion between two neighbouring loops in the β-roll [44,45]. We incorporated the geometry from these Ca^{2+}-binding motifs into the structure calculations as described for AlgE4R [46]. Introducing the Ca^{2+}-ions into our structures assuming they are bound in the same way as in AlgE4R (Fig. 7) did not change the structures or introduce additional violations (distance constraints and van der Waals repulsions) for the calculated structures. However, this is not evidence that Ca^{2+}-ions indeed are bound exactly there or exactly in that way, but it demonstrates that Ca^{2+}-ion binding in that way in these sites is possible.

The overall shape, orientation and interaction between the A- and R-modules and between multiple R-modules can provide us with a deeper understanding on how the epimerases bind alginate and what their mode of action might be.

The line widths of the peaks measured by NMR for the full length AlgE4 shows that the average line width of 10 peaks belonging to the A-module is 7.89 ± 1.37 Hz while the average line width of 10 peaks from the R-module 6.03 ± 0.72 Hz. This suggests that the A- and R-module of the full length AlgE4 relax with different rates. However, from the scattering data the orientation between A- and R-module in the full length AlgE4 was modelled and there is a defined angle between both modules. These results seem to contradict themselves but rotation and molecular motion of the AlgE4 on the modules could also explain the different relaxation rate.

Additionally, it is possible to obtain a model for the full length AlgE6 from the scattering data. Rigid body modelling of the multi-module protein AlgE6 showed that the modules have a defined orientation relative each other. Interestingly, the model indicated that the three R-modules do not have a linear orientation but form a bend (Fig. 11). The low resolution structures of AlgE4 and AlgE6 are the first structural models of a full length extracellular alginate epimerases.

In total, the results from NMR and SAXS data suggest that the epimerases adopt an overall elongated shape and with only limited flexibility between the individual modules.

Alginate Binding Site

As mentioned before, the structures of both A- and R-modules have a rather positively charged surface potential along this groove. This can possible support the binding of the strongly negatively charged alginates. Therefore, the electrostatic potential and the amount of charged and polar amino acids at the binding site of the R-modules could play a crucial role for their interaction with alginates. Alginates are both the substrate and the product of the enzymatic conversion, and substrate or product binding could be the role of the R-modules. In order to explain different alginate binding behaviour of the four R-modules despite of their structural similarity and sequence homology, we discuss similarities and dissimilarities of the charge distribution on the electrostatic surface for the four R-modules. Especially, the R-modules of AlgE6 are very similar to each other in the amino acid sequence but vary significantly in their charge distribution (Fig. 8). On the possible binding site of AlgE4R are eight basic amino acids located. From these eight amino acids three are not conserved in all four R-modules. K103 - one of the non-conserved basic amino acids in AlgE4R- is relatively far away from any other positively charged amino acid (approximately 9Å from R110 and K114). However, R124 and R24 -the other two not conversed basic amino acids- are close to other basic amino acids. R24 is close to R40 and R42 while R124 is surrounded by R62 and K114. R24 introduces a positive surface potential on a spot, where AlgE6R2 and AlgE6R3 have a negative surface potential. R124 forms a kind of bridge between R62 and K114 but the positive charge is shielded by negatively charged side chains of surrounding amino acids (E64, D94, E117, D119, E121, E126). The front site of AlgE4R has also many acidic amino acids. Many of them are on the bottom loop resulting and a strong negative area is formed by amino acids D26, D74, E76, E79, E107, D130, D132 and E136, of which E79 and D132 are non-conserved amino acids.

All R-modules of AlgE6 have one additional arginine located between two phenylalanines that open the β-roll (FRF 53-55 in AlgE6R1, 63-65 in AlgE6R2 & AlgE6R3). This arginine is on the back side of the proteins. Beside this one arginine, AlgE6R1 has an arginine extra (R26) that is at the same position as R24 in AlgE4R (see Fig. 17). R26 is flanked by two aspartic acids (D24 and D28). The charge distribution is very similar to AlgE4R although AlgE6R1 does not have that strong a negative groove on the bottom as on both non-conserved positions there are amidic instead of acidic amino acids. In the case of AlgE6R2 the charges are more equally distributed and there are no areas on the electrostatic surface that have a strong positive or strong negative potential. Although AlgE6R2 and AlgE6R3 have such similar sequence, the electrostatic surfaces are very different. Most pronounced is the electropositive area on the front side of AlgE6R3. It consists of the four basic amino acids Arg51, Lys53, Arg82 and Arg121 (Fig. 8 and Fig. 17). Arg82 is the only one that is not conserved in all four R-modules. In the other modules there is a leucine at this position. The distance between Arg82 and Arg51 is about 4.5 Å and Arg82 and Lys53 are about 8 Å away from each other. Particularly the distance between Arg51 and Arg82 is extremely close considering that both are positively charged. Additionally, Arg121 is also relatively close to Arg82 with 8 Å. In the other R-modules, Arg121 has a far distance to any other basic amino acid. On the other hand AlgE6R3 is missing one arginine which is conserved in all other R-modules. AlgE6R3 has on the position 134 a leucine while all other R-modules have an arginine at that position. This arginine does not participate in the binding.

Beside the basic residues also the acidic and polar amino acids have to be considered. Many of the acidic and polar amino acids on the front side are also conserved in all four R-modules (in AlgE4 they are D74, E107, D119, E126 and S44, Y96, N105, Y112). Both groups of amino acids can bind alginate by hydrogen bonds and probably stabilizing the proteins through avoiding electrostatic repulsion of the basic residues that might otherwise cause labile protein structure. Moreover if the binding groove only had basic residues the alginate polymer might not be able to dissociate from the enzyme surface or/and the epimerase can not move along the alginate polymer chain.

If all the R-modules are considered AlgE4R and the R-modules of AlgE6 are in two different clusters. The R-modules of AlgE6 and ORF9 form one cluster while AlgE4R with other R-modules that are the first R-module after an A-module (AlgE2R1, AlgE5R1, AlgE3R4, AlgE1R4,…). All the R-modules AlgE6 and ORF9 have a proline rich linker between them (the most common sequence is VPVDPNVEGTPVV) which doesn't exist in any other cluster. All the R-modules of this cluster have a serine or threonine on the position where AlgE4R has K103 and R124 and half of the R-modules have an arginine on the same position as R24 in AlgE4R. On the other hand R82 in AlgE6R3 seems to be a mutation as all the other R-modules have a glutamine on that position. All R-module of this cluster have an arginine that the same position of R124 in AlgE4R. R24 inAlgE4R is shared by many R-modules in that cluster except two which have an arginine two position before. R40 in AlgE4R is conserved in all R-modules of this cluster but AlgE4R is the only has a R42. This is in contrast to the cluster of AlgE6 and ORF9 where both arginine are conserved in all R-modules. K103 of AlgE4R seems to be a mutation as only one other R-module as a lysine on the same position.

Alginate Binding

The R-module of AlgE4 shows a clear preference for poly-M alginate over MG alginate (Fig. 14). Poly-G alginate was not tested, as it could not be dissolved in the buffer conditions used for AlgE4R. The protein needs Ca^{2+} in order to retain its fold, but poly-G alginate binds Ca^{2+} and forms gels. An attempt to perform an ITC measurement in calcium free buffer failed - probably due to the structural instability of the R-module from AlgE4 in the absence of calcium [46]. The highest binding energies were measured with the M5 and higher (Fig. 14 and Supp. Tab. 2). M5 has approximately the same length as the maximal distance between the basic amino acids of AlgE4 R (R24 and K103 Fig. 17). Most of the amino acids that are affected by the alginate binding are in the groove with an electropositive surface potential (see Fig. 8 and Fig. 16) and it seems that alginate binds over the whole length of AlgE4R. The fact that M3-M5 lead to chemical shift perturbations over nearly the same surface area of the R-module, while the dissociation constant of the binding decreases by 10 fold with increasing number of M subunits, suggests that there is a multiple number of binding sites available, and that short-chained alginate oligomers that cannot fulfil all binding sites simultaneously, but can freely move between different binding subsites in the alginate binding groove. The longer M8 and M9 show two distinguished binding events (Fig. 13). It is assumed that at low alginate concentration, two R-modules bind to one alginate chain. At higher alginate concentration each R-module binds to one alginate chain. The ITC experimental data, NMR titration and structural data clearly show a single binding site in the alginate binding groove. The ITC thermograph for poly-MG shows binding of alginate, however, the software was not able to fit the obtained data due to the low amounts of heat generated by this weak binding. In order to get better ITC results, the protein concentration should be 50-200 fold higher, which is not possible. The NMR data also show a weak binding for poly-MG and the obtained binding constant also indicates that ITC titrations are at its limits for poly-MG.

In contrast to the results obtained for AlgE4R none of the individual R-modules of AlgE6 showed measurable interaction to any alginate oligomers (See Fig. 15; Tested were M5 - AlgE6R1, M8 - AlgE6R1, MG8 - AlgE6R1, GG6 - AlgE6R1, GG7 -AlgE6R1, M6 - AlgE6R2, M6 - AlgE6R3, MG6 - AlgE6R2 and MG6 - AlgE6R3). This is extremely surprising as the amino acid sequence of R-modules of AlgE6 and AlgE4R and also the amount and distribution of charged and polar amino acids are alike at the assumed binding site (Fig. 3 and Fig. 17).

Recent results from study of the protein ORF9 [47], which consist of only seven R-modules from *A.vinelandii*, have showed that it binds alginate oligomers (Tested were M12, M13-15, M18-20 and MG18-22). The R-modules in ORF9 are more similar to the R-modules of AlgE6 than to AlgE4. A possible explanation for the lack of alginate binding by the individual R-modules from AlgE6 is that the number of interaction sites on the alginate binding groove might be too low to bind the alginate oligomer.

Mode of action

Although the R-modules show no epimerisation activity, they enhance the activity of the A-module by ten-fold if an R-module is bound to an A-module [15]. In the case of AlgE4 the R-module can bind to alginate oligomers with different affinity. It can be assumed that the charged and polar amino acids influence the orientation of the alginate on the epimerase surface before and after each epimerisation reaction. The charge-charge attraction and repulsion can help to move the whole epimerase on the alginate polymer forward in processive mode. The R-module is following after the A-module meaning that in the beginning the R-module may bind to poly-M alginate but after some epimerisation steps the R-module comes into contact with the MG-blocks produced. The R-module does not bind well to the alginate anymore and the whole AlgE4 detaches. This model fits well experimental data obtained by Campa *et al.* [48], which observed that AlgE4 makes ~10 epimerisation steps for each binding on the alginate polymer.

In the case of the G-block forming epimerases, the R-modules have maybe two different functions. The one function is to enhance the reaction rate by binding to M- and MG-blocks while the other function is maybe to prevent the alginate from gelling prematurely (which would disrupt epimerisation) by limiting the water-soluble surface of the alginate and hereby lowering the accessibility of the divalent cations to the alginate.

Concluding Remarks

Here we have determined the 3D solution structure of the three individual R-modules from AlgE6 with NMR spectroscopy. In general, they are all β-sheet folded into an elongated roll with a positively charged groove along the long axis of the protein on one side. Calcium ions can be incorporated into the loops of the β-roll without an increase in the target function for the structure calculation. The line width measurement for AlgE4 shows that the A- and R-module have different relaxation rates. SAXS analyses of AlgE6 and AlgE4 display a defined overall orientation for both epimerases. AlgE4R binds alginate, while the individual R-modules from AlgE6 are not able to interact with alginate. Furthermore, AlgE4R has a higher affinity for poly-M then poly-MG oligomeres which correlates well with AlgE4s mode of action and degree of processivity.

Acknowledgements

EB, RW, MAB and JSP acknowledge financial support from the Danish Agency for Science, Technology and Innovation. The 800 MHz spectra were obtained using the Bruker Avance spectrometer of the Danish Instrument Center for NMR Spectroscopy of Biological Macromolecules. The NMR laboratory at Aalborg University is supported by the Obel and SparNord foundations. FLA thanks the KMB project 182695/I40 and the Norwegian Research Council for financial support.

References

1. Fischer FG, H D (1955) [Polyuronic acids in brown algae.]. Hoppe Seylers Z Physiol Chem 302: 186-203.
2. Hirst EL, Jones JKN, Jones WO (1939) Structure of Alginic Acid. J Chem Soc: 1880.
3. Haug A, Myklestad S, Larsen B, Smidsrod O (1967) Correlation between Chemical Structure and Physical Properties of Alginates. Acta Chem Scand 21: 768-778.
4. Smidsrod O (1970) Solution Properties of Alginate. Carbohydr Res 13: 359-372.
5. Haug A, Larsen B, Smidsrod O (1967) Studies on the Sequence of Uronic Acid Residues in Alginic Acid. Acta Chem Scand 21: 691-704.
6. Smidsrod O, Haug A (1972) Dependence upon the gel-sol state of the ion-exchange properties of alginates. Acta Chem Scand 26: 2063-2074.
7. Smidsrod O (1973) The relative extension of alginates having different chemical composition. Carbohydr Res 27: 107-118.

8. Pindar DF, Bucke C (1975) The biosynthesis of alginic acid by Azotobacter vinelandii. The Biochemical journal 152: 617-622.

9. Ertesvåg H, Høidal HK, Hals IK, Rian A, Doseth B, Valla S (1995) A family of modular type mannuronan C-5-epimerase genes controls alginate structure in Azotobacter vinelandii. Mol Microbiol 16: 719-731.

10. Svanem BI, Skjak-Braek G, Ertesvag H, Valla S (1999) Cloning and expression of three new Aazotobacter vinelandii genes closely related to a previously described gene family encoding mannuronan C-5-epimerases. J Bacteriol 181: 68-77.

11. Ertesvag H, Hoidal HK, Schjerven H, Svanem BI, Valla S (1999) Mannuronan C-5-epimerases and their application for in vitro and in vivo design of new alginates useful in biotechnology. Metabolic engineering 1: 262-269.

12. Svanem BI, Strand WI, Ertesvag H, Skjak-Braek G, Hartmann M, Barbeyron T, Valla S (2001) The catalytic activities of the bifunctional Azotobacter vinelandii mannuronan C-5-epimerase and alginate lyase AlgE7 probably originate from the same active site in the enzyme. The Journal of biological chemistry 276: 31542-31550.

13. Rozeboom HJ, Bjerkan TM, Kalk KH, Ertesvag H, Holtan S, Aachmann FL, Valla S, Dijkstra BW (2008) Structural and mutational characterization of the catalytic A-module of the mannuronan C-5-epimerase AlgE4 from Azotobacter vinelandii. The Journal of biological chemistry 283: 23819-23828.

14. Aachmann FL, Svanem BI, Güntert P, Petersen SB, Valla S, Wimmer R (2006) NMR structure of the R-module: a parallel beta-roll subunit from an Azotobacter vinelandii mannuronan C-5 epimerase. The Journal of biological chemistry 281: 7350-7356.

15. Ertesvåg H, Valla S (1999) The A modules of the Azotobacter vinelandii mannuronan-C-5-epimerase AlgE1 are sufficient for both epimerization and binding of Ca2+. J Bacteriol 181: 3033-3038.

16. Hoidal HK, Ertesvag H, Skjak-Braek G, Stokke BT, Valla S (1999) The recombinant Azotobacter vinelandii mannuronan C-5-epimerase AlgE4 epimerizes alginate by a nonrandom attack mechanism. The Journal of biological chemistry 274: 12316-12322.

17. Hartmann M, Holm OB, Johansen GA, Skjak-Braek G, Stokke BT (2002) Mode of action of recombinant Azotobacter vinelandii mannuronan C-5 epimerases AlgE2 and AlgE4. Biopolymers 63: 77-88.

18. Buchinger E, Aachmann FL, Aranko AS, Valla S, Skjak-Braek G, Iwai H, Wimmer R Use of protein trans-splicing to produce active and segmentally (2)H, (15)N labeled mannuronan C5-epimerase AlgE4. Protein Sci 19: 1534-1543.

19. Aachmann FL, Skjåk-Bræk G (2008) 1H, 15N, 13C resonance assignment of the AlgE6R1 subunit from the Azotobacter vinelandii mannuronan C5-epimerase. Biomolecular NMR assignments 2: 123-125.

20. Andreassen T, Buchinger E, Skjåk-Bræk G, Valla S, Aachmann F 1H, 13C and 15N resonances of the AlgE6R2 subunit from Azotobacter vinelandiimannuronan C5-epimerase. Biomolecular NMR assignments: 1-3.

21. Buchinger E, Skjak-Braek G, Valla S, Wimmer R, Aachmann FL NMR assignments of (1)H, (13)C and (15)N resonances of the C-terminal subunit from Azotobacter vinelandii mannuronan C5-epimerase 6 (AlgE6R3). Biomolecular NMR assignments.

22. Aachmann FL, Svanem BG, Valla S, Petersen SB, Wimmer R (2005) NMR assignment of the R-module from the Azotobacter vinelandii Mannuronan C5-epimerase AlgE4. Journal of biomolecular NMR 31: 259.

23. Hanahan D, Meselson M (1983) Plasmid screening at high colony density. Methods in enzymology 100: 333-342.

24. Pedersen J (2004) A flux- and background-optimized version of the NanoSTAR small-angle X-ray scattering camera for solution scattering. Journal of Applied Crystallography 37: 369-380.

25. Pedersen JS, Hansen S, Bauer R (1994) The aggregation behavior of zinc-free insulin studied by small-angle neutron scattering. Eur Biophys J 22: 379-389.
26. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics (Oxford, England) 22: 195-201.
27. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Journal of Applied Crystallography 28: 768-773.
28. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. Journal of the American Chemical Society 129: 5656-5664.
29. Petoukhov MV, Svergun DI (2005) Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. Biophysical Journal 89: 1237-1250.
30. Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. Journal of Applied Crystallography 34: 33-41.
31. Keller RLJ (2004) Optimizing the Process of Nuclear Magnetic Resonance Spectrum Analysis and Computer Aided Resonance Assignment: Cantina Verlag, Goldau.
32. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of molecular biology 273: 283-298.
33. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. Journal of biomolecular NMR 13: 289-302.
34. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics (Oxford, England) 23: 2947-2948.
35. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proceedings of the National Academy of Sciences of the United States of America 98: 10037-10041.
36. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Research 35: W522-W525.
37. Gimmestad M, Steigedal M, Ertesvag H, Moreno S, Christensen BE, Espin G, Valla S (2006) Identification and characterization of an Azotobacter vinelandii type I secretion system responsible for export of the AlgE-type mannuronan C-5-epimerases. J Bacteriol 188: 5551-5560.
38. Delepelaire P (2004) Type I secretion in gram-negative bacteria. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1694: 149-161.
39. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33: W36-38.
40. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 292: 195-202.
41. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A, Walker JM (2005) Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook: Humana Press. pp. 571-607.
42. Debye P (1947) Molecular-weight determination by light scattering. The Journal of physical and colloid chemistry 51: 18-32.

43. Ertesvag H, Valla S (1999) The A Modules of the Azotobacter vinelandii Mannuronan-C-5-Epimerase AlgE1 Are Sufficient for both Epimerization and Binding of Ca2+. J Bacteriol 181: 3033-3038.
44. Baumann U, Wu S, Flaherty KM, McKay DB (1993) Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif. The EMBO journal 12: 3357-3364.
45. Baumann U (1994) Crystal structure of the 50 kDa metallo protease from Serratia marcescens. Journal of molecular biology 242: 244-251.
46. Aachmann FL (2005) The NMR Structure of the R-module - A Parallel beta-roll Subunit from Azotobacter vinelandii Alginate C-5 Epimerase. PhD thesis.
47. Dahlheim MØ (2010) Functional Study of its Effect(s) on Alginate Epimerisation. Masteroppgave.
48. Hartmann M, Holm OB, Johansen GAB, Skjåk-Bræk G, Stokke BT (2002) Mode of action of recombinant Azotobacter vinelandii mannuronan C-5 epimerases AlgE2 and AlgE4. Biopolymers 63: 77-88.
49. Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. Journal of molecular graphics 14: 51-55, 29-32.

	AlgE6R1		AlgE6R2		AlgE6R3	
	Without Ca^{2+}-ions	With Ca^{2+}-ions	Without Ca^{2+}-ions	With Ca^{2+}-ions	Without Ca^{2+}-ions	With Ca^{2+}-ions
Total number of NOE constraints	3510	3554	3535	3579	2057	2146
Intra	2181	2181	2316	2316	865	889
Short	782	782	760	760	589	606
Medium	112	112	106	106	118	98
Long	435	479	353	397	485	553
Cyana Target function value (Å)	2.81 ± 1.18	2.55 ± 0.92	1.17 ± 0.30	2.65 ± 0.4	2.70 ± 0.86	1.17 ± 0.42
Distance constraint violation (> 10 structures)	< 0.2	< 0.2	< 0.2	< 0.2	< 0.1	< 0.1
RMSD						
N, C$^{\alpha}$, C' (core-R-module)	1.651	2.397	2.256	1.967	1.769	1.864
N, C$^{\alpha}$, C' (secondary structure element)	0.671	1.178	1.191	1.038	0.954	0.941
Heavy Atoms (core R-module)	1.945	2.840	2.732	2.455	2.015	2.077
Heavy Atoms (secondary structure)	1.163	1.933	1.649	1.521	1.379	1.354

Supp. Tab. 1: Summary of the results of the structure calculations of the three R-module structures. The structure of each module was calculated with and without Ca^{2+}-ion incorporation. All RMSDs calculations were calculated in MolMol [49]. Core-R-modules are from the first identical amino (G1 in AlgE4R, G3 in AlgE6R1, G11 in AlgE6R2 and G11 in AlgE6R3 see also Fig. 3) acid until the C-terminal end of AlgE6R1 and AlgE6R2 or the amino acid that is in the same position when the 4 structures are aligned (T150 in AlgE4R and A162 in AlgE6R3). The secondary structure elements are only the beta-strands (see Fig. 4 and Fig. 7)

	NMR	ITC					ITC N = 1			
Ligand	K_d[μM]	N	K_d[μM]	ΔH [kJ/mol]	ΔS [J/(mol·K)]	ΔG [kJ/mol]	K_d[μM]	ΔH [kJ/mol]	ΔS [J/(mol·K)]	ΔG [kJ/mol]
M3	883 ± 98.1	1.58 ± 0.22	267.1 ± 12.3	-20.2 ± 3.1	-1.1	-19.9 ± 3.1	323 ± 3.8	-33.8 ± 0.2	-46.4	-19.9 ± 1.4
M4	191 ± 27.2	1.49 ± 0.02	13.06 ± 0.32	-29.4 ± 0.4	-5.09	-27.9 ± 0.4	26 ± 2.0	-45.4 ± 1.4	-64.5	-26.1 ± 2.4
M5	2.72 ± 1.91	0.6 ± 0.007	2.21 ± 0.04	-71.6 ± 0.3	-132.23	-32.2 ± 4.0	3.75 ± 0.06	-41.9 ± 0.2	-36.78	-31.0 ± 1.1
M6	0.41 ± 0.30	0.93 ± 0.003	0.88 ± 0.03	-47.5 ± 0.2	-43.46	-34.5 ± 1.3	1.03 ± 0.25	-46.8 ± 1.7	-42.4	-34.2 ± 2.1
M7	---	0.702 ± 0.002	0.33 ± 0.01	-62.3 ± 0.3	-84.99	-37.0 ± 2.5	0.46 ±0.02	-45.0 ± 0.2	-29.55	-36.2 ± 0.9
M8 *	---	0.945 ± 0.003	0.28 ± 0.02	-59.0 ± 0.5	-72.73	-37.3 ± 2.2	---	---	---	---
M8 -1 #	---	0.499 ± 0.016	0.36 ± 0.02	-44.1 ± 3.1	-24.83	-36.7 ± 3.2	---	---	---	---
M8 - 2 #	---	0.465 ± 0.017	0.09 ± 0.01	-83.2 ± 2.2	-143.79	-40.4 ± 4.8	---	---	---	---
M9 - 1 #	---	0.418 ± 0.018	0.57 ± 0.03	-11.6 ± 6.3	80.67	-35.6 ± 6.7	---	---	---	---
M9 - 2 #	---	0.601 ± 0.02	0.15 ± 0.02	-91.8 ± 1.8	-177.65	-38.9 ± 5.6	---	---	---	---
MG4	156 ± 23.4	30.3 ± 0.83	641 ± 36.8	-1.16 ± 0.04	57.27	-18.2 ± 1.7	2040 ± 110	-47.8 ± 1.55	-108.7	-15.4 ± 3.6
MG5	---	17.3 ± 0.25	364 ± 44	-3.84 ±0.18	52.88	19.6 ± 1.6	4115 ± 239	-230 ± 12.8	-727.3	-13.9 ± 25.2
MG6	---	19.15 ± 0.93	82.2 ± 18.8	2.16 ± 0.15	70.64	-23.2 ± 2.1	1204 ± 67	-98.5 ± 2.99	-274.63	-16.6 ± 8.7
MG7	314 ± 75.2	---	---	---	---	---	---	---	---	---
MG8		5.78 ± 0.10	75.5 ± 9.25	-9.7 ± 0.32	46.13	23.5 ± 1.4	1330 ± 132	210 ± 20.1	-652.1	-15.9 ± 27.9

Supp. Tab. 2: Dissociation constants and other thermodymanic data. NMR data were calculated for a 1:1 (N = 1) complex. ITC data were calculated once as 1:1 complex and once the binding ratio was an additional parameter. Nearly all the ITC data could be fitted to N = 1. ITC data that gave an huge error are labelled in blue. M8 and M9 could not be fitted with one binding event. * indicate that the first data points were omitted to fit the data to one binding event. Data obtained by calculating two binding events are indicated by #.