Ozgur CAKICI

STRUCTURAL STUDIES OF RHIZOBIAL NODULATION PROTEINS RELATED TO NOD FACTOR BIOSYNTHESIS

Thesis presented to the Scientific Council of the Institute of Bioorganic Chemistry Polish Academy of Sciences in Poznan as a Ph.D. dissertation

Poznan 2010

The research described in this thesis has been carried out at the Institute of Bioorganic Chemistry, Polish Academy of Sciences in Poznan in the Center for Biocrystallographic Research under the supervision of Professor Dr. Michal M. Sikorski. This work was supported in part by a Marie Curie Early Stage Researcher Training (MEST-CT-2004-504066) fellowship of the 6th Framework Programme (FP6) of the European Union. PhD education was carried out under ADOPT programme (a doctoral training network in plant developmental research – from single molecule to whole plant).

I would like to address a special gratitude to
Prof. Dr. Mariusz Jaskolski
for care and understanding,
patience in discussion of significant
crystallographic details of my work
and exceptional support
during publication of this research

I would like to thank to
my thesis supervisor
Prof. Dr. Michal M. Sikorski
for support, care and understanding

Prof. Dr. Tomasz Stepkowski
for support and guidance on rhizobia

Prof. Grzegorz D. Bujacz
for advice on crystallization of proteins
and data collection

Dr. Krzysztof Brzezinski
and Dr. Robert Kolodziejczyk
for help on data collection and friendship

Dr. Mirek Gilski
for technical support on computers

Dr. Ralf Petri
Joanna Stepkowska
for help and orientation in ADOPT program

Dr. Humberto Fernandes
for being a very close friend

Wieslawa Wloszczak
Agnieszka Ciesielska
Malgorzata Konieczna
Ivona Bagdiul
Magdalena Bejger
Alina Kasperska
and other members of CBB
for help and friendship

My Didem
for patience, support and love

My family
for support and faith

List of Publications

Ozgur Cakici, Michal Sikorski, Tomasz Stepkowski, Grzegorz Bujacz, and Mariusz Jaskolski (2010). Crystal structures of NodS N-methyltransferase from *Bradyrhizobium japonicum* in ligand-free form and as SAH complex. Journal of Molecular Biology, doi:10.1016/j.jmb.2010.10.016.

Ozgur Cakici, Michal Sikorski, Tomasz Stepkowski, Grzegorz Bujacz, and Mariusz Jaskolski (2008). Cloning, expression, purification, crystallization and preliminary X-ray analysis of NodS N-methyltransferase from *Bradyrhizobium japonicum* WM9. Acta Crystallographica. Section F, Structural Biology and Crystallization Communications 64, no. Pt 12 (December 1): 1149-52. doi: 10.1107/S174430910803604X

Ozgur Cakici (2004) Biochemical and genetic characterization of *Halobacterium salinarium* strain isolated from Tuz lake in central anatolia. M.Sc. Thesis, Middle East Technical University, Ankara, Turkey.

Table of Contents

Introduction 10

Rhizobium-Legume Symbiosis 10

Nod Factor Biosynthesis 14

S-adenosyl-L-methionine Dependent Methyltransferases (MT) 17

Chemistry of SAM dependent methyltransferases 17

Structural properties of SAM-MTases 17

The SAM-binding site 20

Relationship between SAM dependent methyltransferases and Rossmann fold proteins 21

Aim of the work 22

Methods 24

Gene Amplification and Cloning 24

Primer Design 24

Polymerase Chain Reaction (PCR) 24

Agarose Gel Electrophoresis 25

Extraction and Concentration of DNA from Agarose Gel 26

Cloning and Transformation 26

Expression and Purification 26

Expression test 26

Solubility test 27

Large scale expressions for NodS and NodB proteins 27

Se-Met Derivative Expression of NodS N-methyltransferase 28

Renaturation of NodW 28

Purification of NodS 29

Purification of NodB 29

Crystallization 30

Crystallization of Ligand free NodS 30

Crystallization of NodS-SAH complex 31
Data Collection and Processing 31
Phasing 32
Multi-wavelength Anomalous Diffraction (MAD) 35
Molecular Replacement (MR) 37
Model Building 38
Refinement 38
REFMAC5 39
Validation 39
PROCHECK 39
Sequence Alignments 40
Docking of Chitooligosaccharide 40
Art Work 40
Results 42
Cloning and Expression 42
Cloning and transformation of nodulation genes 42
Expression and solubility of nodulation proteins 43
Purification 46
Purification of NodS 46
Purification of NodB 48
Purification of NodW 50
Crystallization 51
Crystallization of Ligand free NodS 51
Crystallization of Native NodS-SAH complex and SeMet Derivative NodS-SAH complex 51
Data collection and Processing 53
Ligand free NodS 53

NodS-SAH complex 54
Se-Met Derivative of NodS-SAH complex 57
Structure Solution (Phasing), Model Building and Refinement 59
NodS-SAH complex 59
Ligand free NodS 63
Model Quality 66
Overall structure of NodS 70
Crystal packing and intermolecular interactions 74
Binding of SAH molecule 74
Possible chitooligosaccharide acceptor binding site 78
Comparison of ligand-free and SAH complex structures of NodS 80
Structurally homologus proteins 83
Comparison of rhizobial NodS proteins 86
Discussion 89
Recombinant protein production and purification for crystallographic studies 89
Labeling of NodS with SeMet, crystallization and diffraction 90
NodS phasing 90
Overall structure 90
SAM/SAH binding pocket 91
Comparison of ligand-free and SAH complex structures of NodS 92
Possible reaction mechanism 92
Structurally homologus proteins 94
Comparison of rhizobial NodS proteins 95
Summary 97
Literature 100
Abbreviations 105

INTRODUCTION

Introduction

Rhizobium-Legume Symbiosis

Symbiotic relationship between rhizobium and legume is known a major contributor to the global biological nitrogen fixation. In this symbiosis, the rhizobium infects host plant and causes the formation of a novel plant organ on legume root or in some cases also on the stem. In the nodule tissues, rhizobia are transformed into bacteroids which fix atmospheric nitrogen by converting it into ammonia (Dénarié *et al.*, 1992). The origin of this complex association is unclear. It has been established that the family *Leguminosae* may have appeared approximately 70 million years ago, however, the oldest fossils have been dated for about 62 million years (Brea *et al.*, 2008). This paleobotanical evidence has recently been corroborated by phylogenetic studies that revealed rapid diversification into three subfamilies, the Caesalpinioideae, Mimosoideae and Papilionoideae. Taking into account that the ability to establish the nitrogen fixing symbiosis had evolved prior to their divergence; one can date this feature as early as 60 million years (Lavin *et al.*, 2005). In comparison to the nitrogenase, which is believed to be one of the most ancient enzymatic complexes dated for over 3 billion years, nodulation capacity has indeed evolved very recently (Raymond *et al.*, 2004).

The rhizobial species are genetically very diverse. They are divided into four different families; Rhizobiaceae, Phyllobacteriaceae, Hyphomicrobiaceae, and Bradyrhizobiaceae. Among them, only several genera have the ability to form a nitrogen-fixing symbiosis with leguminous plants. Currently these genera include *Rhizobium*, *Ensifer* (formerly known as *Sinorhizobium*), *Mesorhizobium*, *Bradyrhizobium*, *Azorhizobium*, and *Allorhizobium*. Besides their genetic diversity, rhizobia share clearly many common genetic and biochemical characteristics related to the establishment of the symbiotic relationship with legumes. One of the common features is the recognition of specific signal molecules, such as flavonoids, which are synthesized and released by host plants. This host-plant originating flavonoids induce the expression of rhizobium nodulation genes that are involved in the production of specific signal molecule, known as Nod factor. The majority of nodulation genes are confined to the root-nodule bacteria (Spaink,2000).

The bacterial symbiont forms nodules only on a restricted number of hosts, and each legume-host is only nodulated by a restricted number of rhizobium microsymbionts.

However, the degree of specificity varies greatly among rhizobia. Some isolates display a broad host range. For example, some tropical *Bradyrhizobium* strains nodulate legumes in different tribes and subfamilies of Caesalpinioideae (*Papilionoideae* and *Mimosoideae*) (Moulin *et al.*, 2004). Similarly, *Rhizobium* sp. strain NGR234 nodulates at least 35 different legume genera, belonging to 13 tribes, as well as the nonlegume *Parasponia* . In contrast, some isolates have very narrow host range. For example, *Ensifer meliloti* (formerly *Sinorhizobium meliloti*) strains induce formation of nitrogen-fixing nodules only on species of the genera *Medicago*, *Melilotus*, and *Trigonella*, whereas strains of *R. leguminosarum* bv. *trifolii* infect only species of the genus *Trifolium* (clover) (Table 1). Some strains even discriminate between genotypes within a legume species. For example, most isolates of *R. leguminosarum* nodulate European pea varieties but not certain peas from Afghanistan that require special strains with an extended host range. Rhizobial strains have been reported to form effective nodules on one plant species (or genus) and ineffective ones on another, showing that specificity is not limited to nodulation but may also affect the later stages of nodule development and the establishment of a nitrogen-fixing symbiosis (Dénarié *et al.*, 1992).

Table 1 Rhizobia-plant associations

Rhizobia	**Host plants**
Ensifer (Sinorhizobium) meliloti	Alfalfa (*Medicago*)
Rhizobium leguminosarum	
biovar *viciae*	Pea (*Pisum*), vetch (*Vicia*)
biovar *trifolii*	Clover (*Trfolium*)
biovar *phaseoli*	Bean (*Phaseolus*)
Mesorhizobium loti	Lotus
Ensifer (Sinorhizobium) fredii	Soybean (*Glycine*)
Rhizobium sp. NGR234	Tropical legumes, *Parasponia* (nonlegume)
Rhizobium tropici	Bean (*Phaseolus*), *Leucaena*
Bradyrhizobium japonicum	Soybean (*Glycine*)
Bradyrhizobium "cowpea"	Tropical legumes
Azorhizobium caulinodans	*Sesbania* (stem-nodulating)

Consequently, we can mention two different interactions for symbiotic signal exchange and host specificity. First example is successful signal exchange between alfalfa and its compatible symbiont, *Rhizobium meliloti* (Figure 1A). Alfalfa produces a spectrum of inducers (shown as yellow) that cause transcription of nod genes in *R. meliloti*. The products of the nod genes direct the synthesis of a responding signal, a lipo-oligosaccharide Nod factor that carries host-specific modifications (shown as green and blue symbols). These Nod factors successfully induce host responses. Host specificity based on the bacterial morphogen is shown on the right. The same inducers from alfalfa trigger the expression of a subset of nod genes in a bacterium, *R. leguminosarum* bv *viciae*, for which alfalfa is not a host. The resulting *R. leguminosarum* Nod factor carries different host-specific modifications (shown as red symbol), which are tailored to be active not on alfalfa but on other plants such as *Vicia*. As a result, the *R. leguminosarum* bv *viciae* signal causes no response in alfalfa. Second example is host specificity based on bacterial gene expression (Figure 1B). *R. loti*, a symbiont of *Lotus*, and *R. etli*, a symbiont of *Phaseolus*, each make a spectrum of Nod factors, of which some have the same structure. It is likely that host-specific plant inducers (as explained in the first example) restrict the host range of bacteria. However, if the expression of the nod genes in each bacterium is driven by means of an inducer-independent NodD variant protein, the bacteria extend their host range to include new plant species or genetic lines (Long,1996).

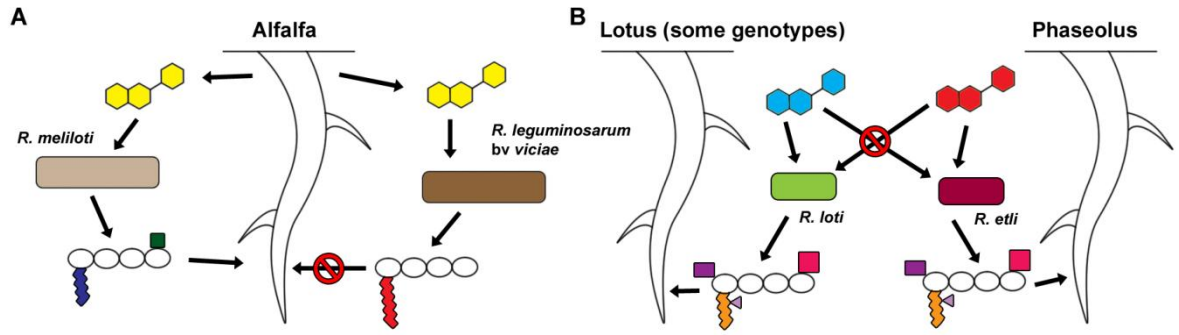


Figure 1 Illustration for symbiotic signal exchange and host specificity. (A) The Nod factor signal exchange model. (B) Host specificity based on bacterial gene expression. The arrows indicate effective signals on the target organism; the red no signs indicate compounds not effective as a signal on the target organism (adapted from **(Long,1996)**).

Genetic regulation of nodulation genes are tightly controlled in *Rhizobia*. Usually, they are expressed during a limited time in the early stages of nodulation following the induction by root exudates that contain flavonoid molecules. These inducers bind to NodD protein, a transcriptional activator, and this in turn promotes binding of NodD tetramers to specific regulatory sequence, called nod box, resulting in expression of *nod* gene operons.

While *nodD* gene is present in all rhizobia, certain regulatory genes are specific for some but not for other genera. In the genus *Bradyrhizobium*, *nod* genes are regulated by double or perhaps triple regulatory circuits. The example is a two-component regulatory system composed of *nodVW* genes, which usually form a tandem. The product of *nodV* gene is a sensor kinase that undergoes autophosphorylation after binding to a flavonoid inducer. This in turn leads to phosphorylation of NodW, which activates the NodW protein. After activation, NodW binds to a putative regulatory sequence that precedes *nod* genes. It is noteworthy that this regulatory sequence has not yet been determined. The third transcriptional activator that similarly to NodW is present only in the genus *Bradyrhizobium,* is NolA protein. The action of this regulatory protein is complex. NolA activates the expression of a second copy of *nodD* (*nodD2*) gene following the binding to its promoter region. NodD2 protein competes with NodD1 for the nod box regulatory sequence, which leads to decreased transcription of *nod* operons (Figure 2). However, *nolA* gene is also present in strains that lack *nodD*2, suggesting that NolA protein activates transcription of other genes as well (Loh *et al.*, 2003; Stepkowski *et al.*, 2003).

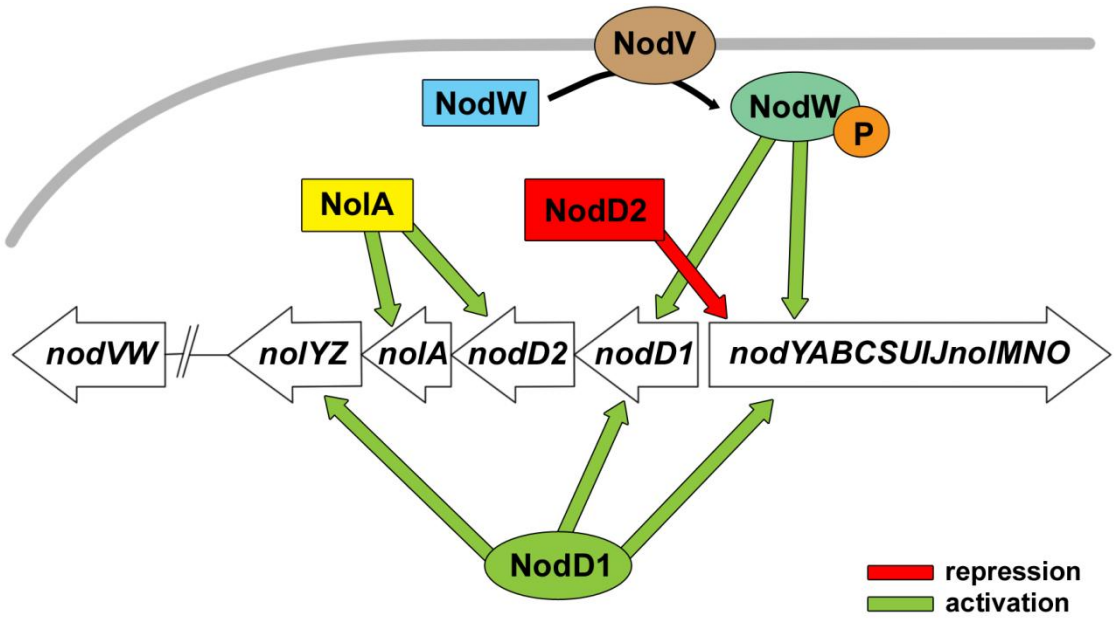


Figure 2 Regulation of the *nod* genes in *B. japonicum*. In response to genistein, NodD1 and NodVW activate the expression of the *nod* genes related to the synthesis of the Nod factor. Negative regulation of the *nod* genes is done by NolA and NodD2. NolA regulates NodD2, which then represses the *nod* genes (adapted from **(Loh *et al.*, 2003)**).

Nod Factor Biosynthesis

Nod factor (NF) is a key molecule determining the recognition of rhizobium strain by its legume host. There are around 60 genes involved in early stages of symbiosis that bear designations *nod, nol* or *noe*. Most of these genes are responsible for NF synthesis, however, only a handful are involved in the synthesis of the chitin backbone. The remaining is involved in chemical modifications that determine NF host-range specificity, in regulation of *nod* gene expression, in NF transport, or in other functions that are unrelated to NF synthesis.

The Nod factor molecule is composed of four to six (usually five) β1,4-linked N-acetyl glucosamine residues. The oligomerisation process is catalyzed by the product of the *nodC* gene, coding for a β-glycosyl transferase. The next stage in Nod factor biosynthesis is the attachment of an N-linked fatty acid to the terminal (non-reducing end) glucosamine residue. Before this happens, NodB deacetylates the glucosamine residue, leaving a free amino group, which is subsequently acylated by NodA protein (Figure 3).

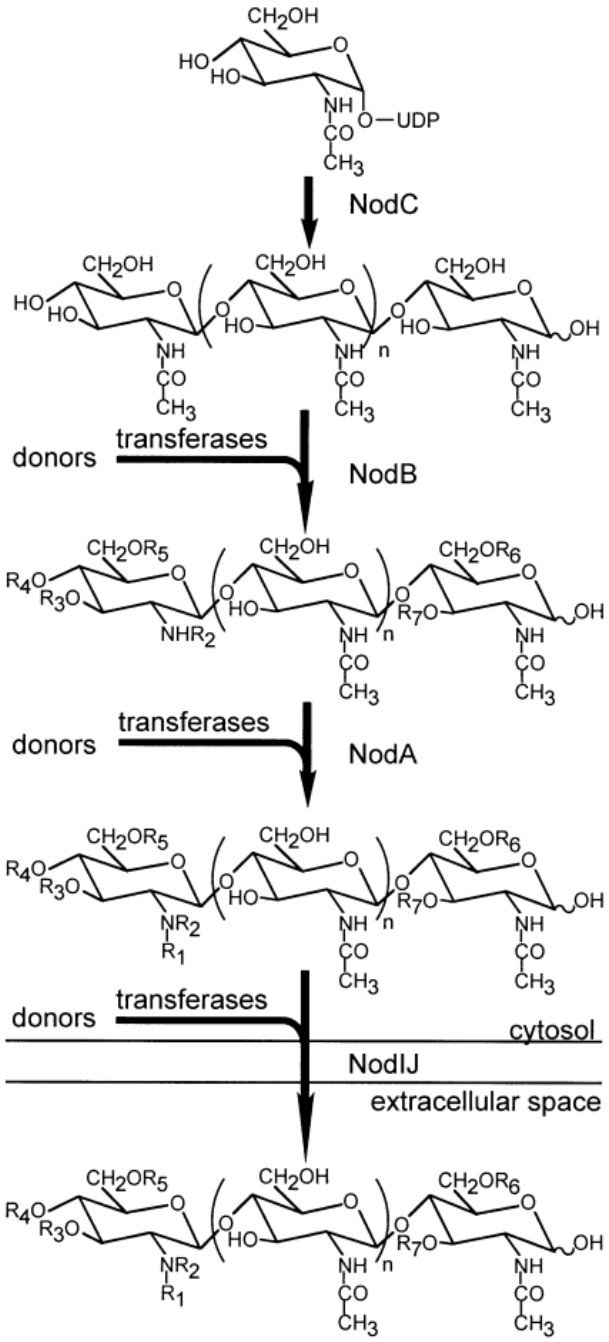


Figure 3 Biosynthesis of Nod factors. The NodC, NodB, and NodA proteins are at the heart of the Nod factor biosynthetic pathway. Depending on the *Rhizobium* strain, this pathway is extended by the introduction of strain-specific modifications catalysed by transferases. Transferases can interfere at all stages of the NodCBA pathway, for example: the fucosyltransferase NodZ (R6) acts after NodC: the methyltransferase NodS (R2) and the acetyltransferase NodL (R5) act after Nod B: and the sulphyltransferase NodH (R6) after NodA. The other transferases are not characterized biochemically and it is not known at which stages they interfere. Some of the donor substrates used by the transferases are also synthesized by Nod proteins. After the synthesis is completed, the LCOs are secreted by a protein complex involving Nodl and NodJ **(Mergaert *et al.*, 1997)**.

Rhizobia-specific modifications are usually attached to the reducing and non-reducing termini of the NF molecule (Figure 4). These modifications are carried out by the products of several genes, such as *nodS* involved in N-methylation, *nodU* and *nolO* involved in carbamoylation, *nodL* and *nolL* in acetylation and several other genes responsible for NF sulphation, fucosylation and arabinosylation (Table 2).

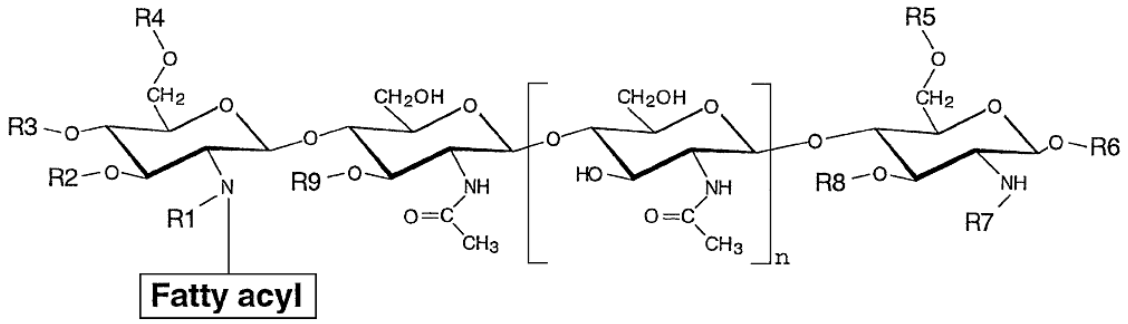


Figure 4 General structure of the Nod factors produced by rhizobia. The presence of substituents numbered R1–R9 is variable within various strains of rhizobia (for examples, Table 2). For intensive list of the identities of these substituents and references, see review **(Spaink,2000)**. In the absence of specific substituents, the R groups stand for hydrogen (R1, R2, R3, R4, R5, R6, R8, and R9), and acetyl (R7).

Table 2 Some examples of the modifications of the Nod factors and gene products responsible for substitutions.

Bacterial strain	Plant Host	GlcNAc residues (*n*)[a]	Special substituents[b]	Specialized genes[c]
S. meliloti	Galegeae	3, 4, 5	R4:Ac, R5:S, FA:C16:2, C16:3, C26(ω-1)OH	R4:NodL, R5:NodH, FA:NodAFEG
M. loti E1R	Loteae, Genisteae	4, 5	R1:Me, R3:Cb, R5:AcFuc	R1:NodS, R3:NolO, R5:NodZ & NolL
B. japonicum USDA110	Phaseoleae	5	R5:MeFuc	R5:NodZ & NoeI
A. caulinodans	Robinieae	4, 5	R1:Me, R4:Cb, R5:Fuc, R8:Ara	R1:NodS, R4:NodU, R5:NodZ, R8:NoeC

[a] The underlined numbers indicate the most abundant species. GlcNAc, *N*-acetylglucosamine
[b] For modification points, see Figure 4.
Abbreviations: Me, *N*-methyl; Cb, *O*-carbamyl; Ac, *O*-acetyl; S, *O*-sulfyl; Fuc, *α*-linked fucosyl; MeFuc, 2-*O*-methylfucosyl; FA, fatty acyl.
[c] The genes encode the enzymes which are responsible for the substitutions.

S-adenosyl-L-methionine Dependent Methyltransferases (MT)

Du Vigneaud and colleagues in 1940s found a substantial fraction of the methyl groups in cell metabolites derived from methionine. In early 1950s, the actual methyl donor was found as the ATP-activated form of methionine (Cantoni,1952). It is known as S-adenosyl-L-methionine (SAM) (Figure 5A). In those days, SAM was believed to methylate small molecules only. However, as experience with this interesting molecule accumulated, it became apparent that DNA, RNA, proteins, lipids, and polysaccharides, are subject to methylation by highly specific SAM-dependent methyltransferases after their primary synthesis. Over four decades passed between the initial characterization of SAM and the first structural characterization of a SAM dependent methyltransferase. Since 1993, a dozen of these enzymes have been structurally characterized, along with the structures of SAM and SAH (S-adenosyl-L-homocysteine) in the bound state, by X-ray crystallography and nuclear magnetic resonance (NMR).

Chemistry of SAM dependent methyltransferases

Methyl transfers are known as alkylation reactions in cellular biochemistry. There are few types of methyl donor molecules and SAM is the most common donor molecule for the methylation reactions. In the second position, various forms of folate are used as methyl donor. Around 120 types of SAM-dependent methyltransferases have been assigned E.C. numbers. Among them, 7 types methylate a sulfur atom, 14 methylate a carbon, 41 methylate a nitrogen, and 55 methylate an oxygen. These numbers do not reflect the whole extant population since some methyltransferases have not been assigned E.C. numbers yet (eg. arsenite methyltransferase). However, given the relative amounts of S, C, O, and N in biomolecules, it appears that carbon methyltransferases are relatively rare and O- and N-methyltransferases are relatively abundant.

Structural properties of SAM-MTases

Most of the methyltransferases share a common core structure referred to as an "SAM-dependent methyltransferase fold". Central to this shared core structure is a seven-stranded β-sheet. Strand 7 is antiparallel to the other six strands, and is inserted into the sheet between strands 5 and 6 (3↑ 2↑ 1↑ 4↑ 5↑ 7↓ 6↑); this feature is thus far characteristic of the SAM-dependent methyltransferases. The core structure is doubly-wound open α/β/α sandwich structure (as opposed to a "singly"-wound structure such as a β-barrel) (Figure 5B). Proteins formed by consecutive α/β motifs constitute the largest general class of enzyme structures.

The SAM-binding region of SAM dependent methyltransferases is localized to the N-terminal part of the β sheet and is formed in part by residues from loops following strands 1, 2 and 3 (Figure 5B). The substrate-binding region, which is responsible for selectively binding small molecules, proteins, lipids, polysachharides or nucleic acids in individual SAM dependent methyltransferases, is located in the C-terminal part of the β sheet (Figure 5B). Not surprisingly, given the diversity of shapes, sizes and chemistries of SAM dependent methyltransferase substrates, the substrate-binding region varies tremendously in structure and topology among individual members of the family (Figure 6).

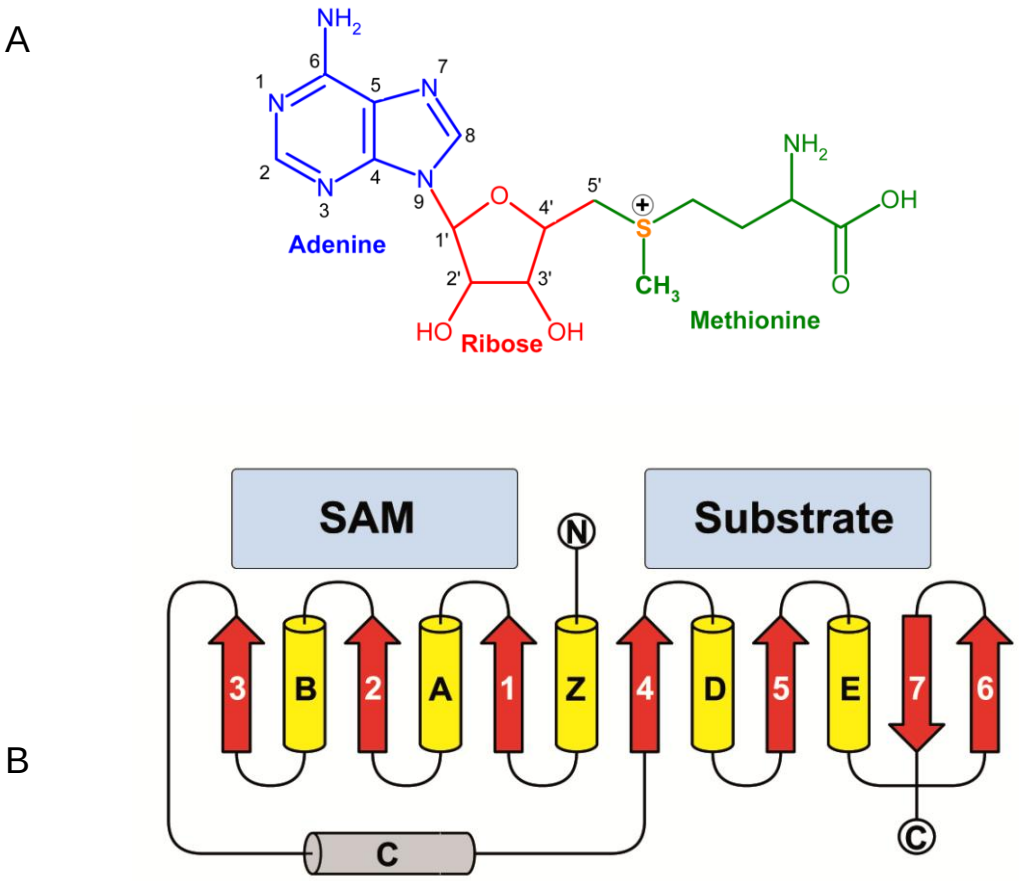


Figure 5 SAM and the SAM-MT fold. (a) Chemical structure of the methyl donor SAM. (b) Schematic showing the topology of the core fold of the SAM-MTs, indicating the SAM- and substrate-binding regions of the fold. Helices are shown as yellow cylinders, strands as red arrows. αC is shown in grey, because it is not always conserved in the core fold. The N and C termini are labeled.

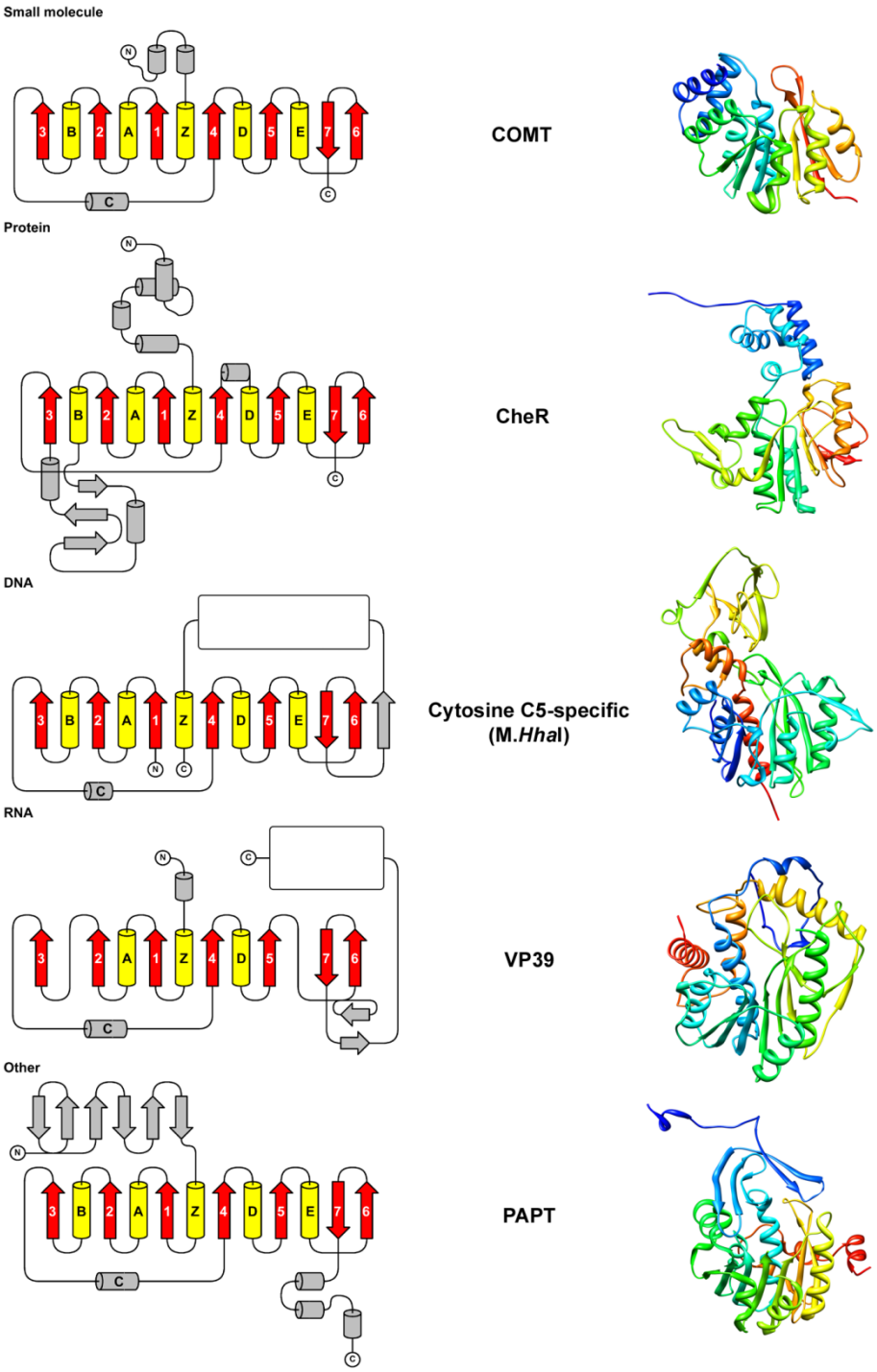


Figure 6 Variations of the SAM-MT fold. Schematic showing the variations of the core fold that have been observed in structures of small-molecule, protein, nucleic acid and other SAM-MTs. Additions to the core fold are shown in grey. Grey boxes indicate additional domains. Asterisks indicate regions that vary in specific enzymes. Dashed lines indicate variability or disorder in that part of the structure.

The SAM-binding site

SAM binding residues show poor conservation among SAM dependent methyltransferases. Generally, SAM dependent methyltransferases bind the cofactor (eg. SAM) at the same equivalent position of the fold and binding conformation of SAM is quite similar to each other. But the chemistry of the SAM-binding interaction varies tremendously. Thus, the plant OMTs have a SAM-binding site enriched in methionines, whereas that of phenylethanolamine N-MT (PNMT) has a preponderance of tyrosines and phenylalanines. The sequence alignment based on structural superimposition (Figure 7) shows that the only residues that are highly conserved in the SAM-binding N-terminal region of the core fold are the glycine-rich sequence E/DXGXGXG (often referred to as motif I) between β1 and αA, which interacts with the amino acid portion of SAM, and an acidic loop between β2 and αB (motif II), which interacts with the ribose hydroxyls. It is interesting to note that, although insertions are found at almost every other point in the fold, these two critical SAM-binding loops appear not to tolerate insertions.

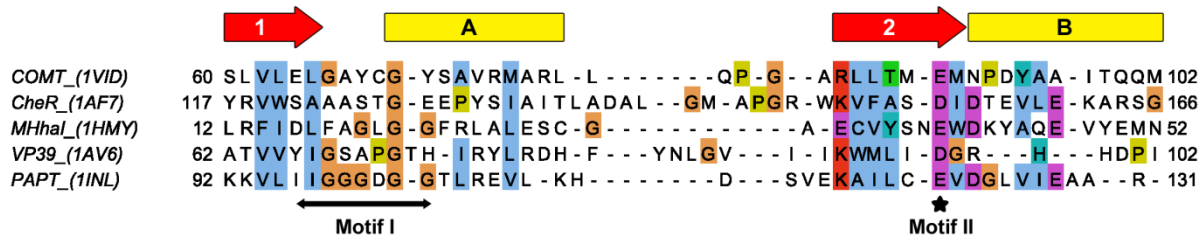


Figure 7 Structure-based sequence alignment of the core SAM-MT fold.

The sequence alignment suggests that the glycine-rich region has signature motifs that vary among the different subclasses of enzyme. For example, the DNA SAM-MTs have the motif E/DXXXGXG, whereas three of the four RNA SAM dependent methyltransferases in Figure 7 have YXGXXXG. A third region of sequence that contributes an interaction with SAM is the linker between β4 and αD. This region often incorporates a short helical insertion (Figure 6). A hydrophobic residue from the linker forms a favorable interaction with the adenine ring of SAM. This residue is not well conserved; in the structures reported to date, it has been identified variously as phenylalanine, isoleucine, valine, cysteine, tryptophan, proline or methionine. Somewhat surprising is the finding that some but not all SAM

dependent methyltransferase structures have additional SAM-binding residues contributed by insertions to the core SAM dependent methyltransferase fold. Overall, it seems that the evolutionary pressure on the SAM dependent methyltransferase fold has been to maintain the same SAM-binding region, but to allow considerable variation in the chemistry of the interaction.

Relationship between SAM dependent methyltransferases and Rossmann fold proteins

The Rossmann fold proteins are a family of proteins grouped together in the SCOP database (actual heading: "NAD(P)-binding Rossmann-fold domains") that each contain two Rossmann fold domains. The two Rossmann fold domains are linked into a continuous six-stranded parallel β-sheet. The overall architecture of these proteins is remarkably similar to that of the SAM dependent methyltransferases. The topological similarity can be seen by visual inspection in Figure 8. All proteins shown in Figure 8 have three parallel strands on each side of a central topological switchpoint, with helices sandwiching the sheet. As with the methyltransferases, the sheet has been extended in some cases. The only major and consistent difference between these two families is the insertion of the antiparallel strand 7 between strands 5 and 6 in the methyltransferases. The structural similarity of Rossmann fold and SAM dependent methyltransferases fold has been noted before, though the focus has most often been on the relationship between the SAM-binding pocket and the Rossmann single nucleotide binding fold.

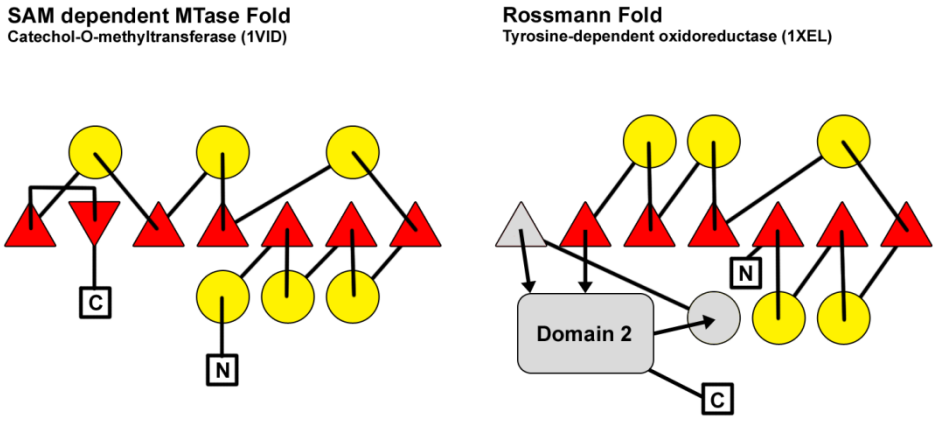


Figure 8 Comparison of SAM dependent Mtases fold and Rossmann fold.

Structural studies indicate that the SAM dependent methyltransferases represent a large structurally-conserved superfamily. General SAM dependent methyltransferases structure illustrates that the profound structural conservation is not reflected in corresponding

sequence conservation. This is consistent with the observations resulting from a more global survey in which the authors went so far as to exclude protein pairs that had more than 25% identity and still found groups of proteins exhibiting strong structural conservation. However, traditionally protein families are defined by sequence conservation (for review (Cheng,2000; Martin *et al.*, 2002)).

Aim of the work

The currently available information about the three-dimensional structure of the enzymes involved in NF synthesis is extremely limited (Brzezinski *et al.*, 2007). We presume that studies on 3D structures of nodulation proteins may contribute to better understanding of not only the biochemistry of NF synthesis, but also may help to elucidate the origin of these proteins by searching for structural similarities. The major goal of this project is to determine the 3D structures of nodulation proteins involved in both Nod factor synthesis and regulation of *nod* gene expression, such as NodA, NodB, NodD, NodS, NodW, and NolA. In particular, there are no structural data about those proteins from any organism. We have, therefore, undertaken X-ray crystallographic studies of the proteins from *Bradyrhizobium japonicum* WM9. Strain WM9 infects lupins and serradella, and its nodulation gene sequences belong to a distinct group classified as Clade II (Stepkowski *et al.*, 2003; Stepkowski *et al.*, 2007). This project is part of our wider program aimed at structural characterization of the enzymes involved in the Nod factor biosynthetic pathway.

METHODS

Methods

Gene Amplification and Cloning

Primer Design

In order to amplify nodulation genes from genomic DNA of *Bradyrhizobium* sp. WM9 strain, forward and reversed primers were designed according to gene sequences of *nodA*, *nolA*, *nodB*, *nodD*, *nodS* and *nodW*. Designed primers were 18-25 bp long and their GC percentages were higher than 50%. Primers were checked if there is any hairpin, primer pair, and mismatch. Designed primers were used for amplification of nodulation genes via polymerase chain reaction (PCR) from genomic DNA of *Bradyrhizobium* sp. WM9 (Table 3).

Table 3 Primers of target genes for PCR reactions.

Gene	Direction	Primer sequence (5' - 3')	Length (bp)	Tm (°)	%GC
nodA	Forward	CACCATGAACATGAGCGCG	19	61.3	57.9
	Reversed	TCACAGTTCTGGCCCGTTCC	20	62.9	60.0
nolA	Forward	CACCATGACAAACGCCACACC	21	64.0	57.0
	Reversed	TCAACTCTTTCCGAGCTGGTTCTCC	25	67.2	52.0
nodB	Forward	CACCATGGTGACATCCACAAACG	23	65.4	52.2
	Reversed	TCAGTGAGGTTGAGGAAGCGACC	23	65.3	56.5
nodD	Forward	CACCATGCGTTTCAAGGG	18	57.5	55.6
	Reversed	CTAGCTGGTCTCCGACGGAGAGAC	24	65.7	62.5
nodS	Forward	CACCATGGTGAGCGTAGACAACAC	24	64.2	54.2
	Reversed	TCAAGCTCGTCCGTCGGGG	19	66.3	68.4
nodW	Forward	CACCATGACCAAGCGCTCTGA	21	64.0	57.0
	Reversed	TTAGGTTTGCACCCGAACACGG	22	66.7	54.5

Polymerase Chain Reaction (PCR)

Six nodulation genes were amplified from genomic DNA of *Bradyrhizobium sp*. WM9 strain via polymerase chain reaction. To test the quality of designed primers, initial gene amplifications were done by Taq DNA polymerase (Fermentas). Amplification reactions were performed in volumes of 25 µl containing Taq DNA polymerase buffer, $MgCl_2$, dNTPs, Taq DNA polymerase, primers and template DNA. A master mix was prepared with following composition; 1x Taq DNA polymerase buffer, 1.5 mM $MgCl_2$, 0.12 mM dNTP, 0.02 U/µl Taq DNA Polymerase, 0.4 µM for each primer, 40 ng template DNA.

The reaction mixtures were mixed gently by tipping the tubes and spun briefly to remove air bubbles. The tubes were placed in the thermocycler and initial denaturation step was performed at 94°C for 5 minutes. After denaturation step, samples were subjected to 30 cycles of denaturation at 94°C for 1 minute, annealing at 55°C for 1 minute and extension at 72°C for 1 minute. Subsequently, the PCR were finalized by final extension step at 72°C for 10 minutes. Samples were stored at 4°C. PCR products were analyzed by electrophoresis in 1% (w/v) agarose gels and detected by staining with ethidium bromide.

Due to the necessity of obtaining blind end products, amplification of target gene sequences for cloning was done by using *Pwo* DNA polymerase (Roche). It is a highly processive 5'–3' DNA polymerase and possesses a 3'–5' exonuclease activity also known as proofreading activity. Two equally separated reaction mixtures were prepared for the PCR. The preparation of two separate master mixes helps to circumvent the need of hot start and in addition avoids the interaction of enzyme with primers or template without dNTPs which could lead to a partial degradation of primer and template through the 3´–5´ exonuclease activity of *Pwo* DNA Polymerase. First mixture contained dNTPs, primers, template DNA while the second mixture had *Pwo* DNA polymerase buffer and Pwo DNA polymerase. Before PCR reaction, two separate mixtures were combined and mixed gently by tipping the tubes and spun briefly to remove air bubbles. Amplification reactions were performed in total volumes of 50 µl. Final reaction mixture contained the following composition; 1x *Pwo* DNA polymerase buffer, 1.5 mM $MgSO_4$, 0.2 mM dNTP, 0.02 U/µl *Pwo* DNA Polymerase, 0.4 µM for each primer, 80 ng template DNA. The tubes were placed in the thermocycler and initial denaturation step was performed at 94°C for 5 minutes. After denaturation step, samples were subjected to 30 cycles of denaturation at 94°C for 1 minute, annealing at 55°C for 1 minute and extension at 72°C for 1 minute. Subsequently, the PCR were finalized by final extension step at 72°C for 10 minutes. Samples were stored at 4°C. PCR products were analyzed by electrophoresis in 1% (w/v) agarose gels and detected by staining with ethidium bromide.

Agarose Gel Electrophoresis

All DNA products were checked with agarose gel electrophoresis. Depending on the size of the nucleic acid, different agarose concentrations ranging from 0.6% to 1.0% (w/v) were used. Appropriate amount of agarose was added to 50 ml of 0.5 X TBE solutions and heated to melt the agarose. While cooling to room temperature, 3 µl of ethidium bromide solution was added. The gel was solidified in the tray. The electrophoresis was carried out in

0.5X TBE buffer at constant voltage 60-80 V. After visualization of DNA bands under UV illumination, photos of the gels were taken.

Extraction and Concentration of DNA from Agarose Gel

All PCR products were loaded on 0.6% of agarose gel. Amplified gene fragments were purified via agarose gel electrophoresis. Purified gene fragments were extracted and concentrated from the agarose gel by using QIAquick gel extraction kit (Qiagen).

Cloning and Transformation

Purified gene fragments were cloned into pET151/D-TOPO (Invitrogen) expression vector using Champion™ pET151 Directional TOPO Expression Kit (Invitrogen). The expression vector has an N-terminal His_6 tag, which is followed by a TEV protease cleavage site. After cloning, putatively transformed TOP10 cells known as competent *E. coli* cells were spread on plate with ampiciline (100 $mg \cdot l^{-1}$) selective LB agar medium. After one day of incubation at 37°C, 5 colonies were picked and put into liquid LB medium and grown overnight at 37°C. Plasmids were purified with QIAprep Miniprep kit. In order to check target gene insertion direction, PCR was performed by using T7 forward primers and gene specific reverse primers. After verifying that the direction of the insert was correct, DNA sequencing was performed for confirmation of the target gene sequences.

The confirmed vector was transformed into BL21-Star (DE3) competent *E. coli* cells (Invitrogen) or BL21-CodonPlus (DE3)-RIPL competent *E. coli* cells (Stratagene) for expression. Expression cells were grown in LB media with selective antibiotics (Ampicicilin for BL21-Star (DE3); Ampicicilin and Chloramphenicol for BL21-CodonPlus (DE3)-RIPL). Overnight grown cells were stocked in 15% glycerol at -80 °C.

Expression and Purification

Expression test

Ten milliliter LB medium containing 100 $mg \cdot l^{-1}$ ampicillin and 34 $mg \cdot l^{-1}$ chloramphenicol was inoculated with transformant cells and incubated at 37°C until the OD_{600} reached 0.8-1.0. Before the culture was induced by isopropyl-β-D-thiogalactopyranoside (IPTG) at a final concentration of 1 mM, 0.5 ml sample was taken and centrifuged by using minicentrifuge at top speed for one minute. Supernatant were discarded and cell pellet was stored at -20 °C. After induction, the culture was incubated for 6 hours at 37°C. During the incubation period, 0.5 ml sample was collected for each hour time points

and pelletted cells were stored at -20°C. All collected cell pellets were suspended in 100 µl of 1x sample buffer containing 63 mM Tris HCl pH 6.8, 10% glycerol, 2% SDS, 0.0025% bromophenol blue by vortexing. All samples were boiled for 5 minutes and cooled before analyzing with SDS PAGE.

Solubility test

Ten milliliter LB medium containing 100 $mg \cdot l^{-1}$ ampicillin and 34 $mg \cdot l^{-1}$ chloramphenicol was inoculated with transformant cells and incubated at 37°C until the OD_{600} reached 0.6-0.8. Before the culture was induced by isopropyl-*β*-D-thiogalactopyranoside (IPTG) at a final concentration of 1 mM, 0.5 ml sample was taken and centrifuged by using microcentrifuge at top speed for one minute. Supernatant was discarded and cell pellet was stored at -20 °C. After induction, the culture was incubated for 3 hours at 37°C. At the end of the incubation, 0.5 ml sample was collected and pelletted cells were stored at -20°C. The frozen samples were thawed and suspended in 80 µl of lysis buffer containing 50 mM potassium phosphate pH 7.8, 0.4 M NaCl, 0.1 M KCl, 10 % glycerol, 0.5% Triton X-100. The suspension was vortexed and incubated on dry ice for 3 minutes. Subsequently, the sample was thawed at 42°C for 3 minutes and vortexed. Freezing and thawing cycles were repeated 3 times. After the lysis step, the sample was centrifuged in microcentrifuge at top speed for 10 minutes at 4°C. Supernatant was collected in another eppendorf tube. The pellet was suspended in 1X sample buffer and some aliquots of the supernatant were mixed with 4X sample buffer. Then those samples were boiled for 5 minutes and 10 µl of the samples were used for SDS-PAGE analyses.

Ten ml of TB medium containing 100 $mg \cdot l^{-1}$ ampicillin and 34 $mg \cdot l^{-1}$ chloramphenicol was inoculated with the transformant cells. The cells were grown at 37°C until the OD_{600} reached 0.8-1.2. The culture was cooled to 18-20°C and induced by isopropyl-*β*-D-thiogalactopyranoside (IPTG) at a final concentration of 0.2 mM. After induction, the culture was incubated overnight at 18-20°C. Half milliliters of t each sample were collected before IPTG induction and at the end of the incubation. Those samples were pelletted and stored at -20°C. Overnight expression was analyzed with freezing and thawing method as mentioned above.

Large scale expressions for NodS and NodB proteins

40 ml of TB medium containing 100 $mg \cdot l^{-1}$ ampicillin and 34 $mg \cdot l^{-1}$ chloramphenicol were inoculated with the transformed *E. coli* cells from glycerol stock at -80°C. The cells

were grown at 37°C for 5-8 h till turbidity. The activated culture was transferred into 2 l of TB medium containing 100 $mg \cdot l^{-1}$ ampicillin and 34 $mg \cdot l^{-1}$ chloramphenicol. The cells were grown until the OD_{600} reached 0.8-1.2. The culture was cooled to 18-20°C and induced by isopropyl-*β*-D-thiogalactopyranoside (IPTG) at a final concentration of 0.2 mM. After induction, the culture was incubated overnight at 18-20°C. The cell paste was harvested and frozen on dry ice for storage at -80 °C.

Se-Met Derivative Expression of NodS N-methyltransferase

The same transformant was also used for the expression of Se-Met derivative protein. As growth media, Se-Met medium base (MDL) and nutrient mix (MDL) were used. The cultivation procedures were same as for the wild-type protein, except that the volumes were reduced by 50%. When the OD_{600} reached 0.8-1.2, the culture was cooled to 18-20°C and 10 ml of Met-Stop solution (100 $mg \cdot l^{-1}$ of Lys, Thr, Phe; 50 $mg \cdot l^{-1}$ of Ile, Val, Leu) were added to block the methionine biosynthesis pathway. The culture was supplemented with 4 ml of selenomethionine solution from MDL and after 15 min of incubation at 18-20°C, it was induced by IPTG.

Renaturation of NodW

Pelleted cells were lysed with the buffer containing 50 mM HEPES pH 7.5, 0.5 M NaCl, 5% glycerol, 1 $mg \cdot ml^{-1}$ lysozyme, 1 mM TCEP, 1 mM PMSF. After incubation on ice for 30-60 min, Triton X-100 was added to the lysate to obtain 1% final concentration and the solution was sonicated for 90 sec on ice. The extract was treated with 250 Unit Benzonase (Sigma) on ice for 15 minutes. The lysate was centrifuged at 15,000 g for 1 hour at 4°C. The supernatant was discarded. The pellet was washed twice with 1X TBS containing 1% Triton X-100 and centrifuged at 15,000 g for 1 hour at 4°C. The pellet was solubilized in 10 ml solubilizing buffer containing 50mM MES pH 6.0, 0.5 M NaCl, 5% Glycerol, 6M guanidine-HCl, 25 mM DTT. The suspension was incubated for 1 hour at 4°C. Insoluble material was removed by centrifugation at 100,000 g for 10 minutes. Protein concentration of the sample was determined and adjusted to 1 $mg \cdot ml^{-1}$ using solubilizing buffer. To renaturate the target protein, quick dilution method was used. The latter sample was quickly diluted 10 times with renaturation buffer containing 50 mM MES pH 6.0, 500 mM NaCl, 5% Glycerol, 1 mM TCEP. Diluted sample was incubated for 24 hours at 4°C. The sample was concentrated 10 times with 50 ml Amicon stirred cells (Millipore) under nitrogen pressure at 58 PSI by using ultrafiltration membranes with 10 kDa cut-off pore size (Millipore). Concentrated sample was centrifuged at 100,000g for 10 minutes to remove insoluble particles.

Purification of NodS

The cell paste was resuspended in the lysis buffer (50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP, 1 mM PMSF, 1 mg·ml^{-1} lysozyme). After incubation on ice for 30-60 min, the lysate was sonicated for 90 sec on ice. The extract was treated with 250 Unit Benzonase (Sigma) on ice for 15 minutes. The lysate was centrifuged at 15,000 g for 1 hour at 4°C. Cell debris was removed by centrifugation. The supernatant was subjected to purification using an ÄKTA Purifier system (GE Healthcare). In the first step, the proteins were applied on a HisTrap™ column equilibrated with binding buffer containing 50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. After binding, the column was washed with 30 mM imidazole in 50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. The protein was eluted from the column using a linear 30-300 mM gradient of imidazole in 50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. The eluted protein was desalted (HiPrep™ 26/10 column) against binding buffer (50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP) to remove imidazole. Imidazole-free protein solution was incubated overnight at 4°C with a His-tagged TEV protease (60 µg TEV protease / 1 mg His-tag protein) to cleave off the His-tag. Subsequently, the sample was applied on a HisTrap™ column to remove the TEV protease, the His-tag and any undigested fusion protein. The first flow-trough was applied on HiPrep™ 26/10 column to exchange the buffer to 50 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM DTT, and 1 mM EDTA before concentrating the sample up to 4 mg·ml^{-1} using Amicon Ultra-4 centrifugal devices with 10 kDa cut-off (Millipore). Concentrated sample (approximately 3-5 ml) was applied on a gel filtration column (HiLoad 16/60 Superdex 200 pg) and run at a 1 ml·min^{-1} flow rate in a buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM DTT, and 1 mM EDTA. Peak fractions were collected and analyzed by SDS-PAGE, Native PAGE and MALDI-TOF mass spectrometry. The purified protein was concentrated to 4 mg·ml^{-1} using Amicon Ultra-4 centrifugal devices with 10 kDa cut-off (Millipore) and stored in small aliquots at -80°C.

Purification of NodB

The cell paste was resuspended in a lysis buffer (50 mM Tris pH 7.2, 0.5 M NaCl, 10% glycerol, 2.5 mM TCEP, 1 mM PMSF, 1 mg·ml^{-1} lysozyme). After incubation on ice for 30-60 min, the lysate was sonicated for 90 sec on ice. The extract was treated with 250 Unit Benzonase (Sigma) on ice for 15 minutes. The lysate was centrifuged at 15,000 g for 1 hour at 4°C. Cell debris was removed by centrifugation. The supernatant was subjected to

purification using an ÄKTA Purifier system (GE Healthcare). In the first step, the proteins were applied on a HisTrap™ column which equilibrated with binding buffer containing 50 mM Tris pH 7.2, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. After binding, the column was washed with 30 mM imidazole in 50 mM Tris pH 7.2, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. The protein was eluted from the column using a linear 30-300 mM gradient of imidazole in 50 mM Tris pH 7.2, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP. The eluted protein was desalted (HiPrep™ 26/10 column) against binding buffer to remove imidazole. Imidazole-free protein solution was incubated overnight at 4°C with a His-tagged TEV protease (60 µg TEV protease / 1 mg His-tag protein) to cleave off the His-tag. Subsequently, the sample was applied on a HisTrap™ column to remove the TEV protease, the His-tag and any undigested fusion protein. The first flow-trough was collected and applied on a gel filtration column (HiLoad 16/60 Superdex 200 pg) at a 1 $ml \cdot min^{-1}$ flow rate in a buffer containing 50 mM Tris pH 7.2, 150 mM NaCl, 10% glycerol, 5 mM DTT, and 1 mM EDTA. Peak fractions were collected and analyzed by SDS-PAGE, Native PAGE and MALDI-TOF mass spectrometry. The purified protein was concentrated to 4 $mg \cdot ml^{-1}$ using Amicon Ultra-4 centrifugal devices with 10 kDa cut-off (Millipore) and stored in small aliquots at -80 °C.

Crystallization

Crystallization of Ligand free NodS

Prior to setting up the crystallization screens, the protein sample was concentrated to 10 $mg \cdot ml^{-1}$ and the buffer was changed to 10 mM Tris pH 8.0, 100 mM NaCl, 2% glycerol, 2.5 mM TCEP, using Ultrafree Centrifugal Filter Units (Millipore) with 10 kDa cut-off at 4°C, and the protein solution was passed through an Ultrafree-MC Centrifugal Filter Unit (Millipore) with 0.1 µm pore size at 4°C. Protein concentration was determined spectrophotometrically at 280 nm or by the Bradford method (Bradford,1976) with BSA as a standard. The sitting-drop vapor-diffusion screening for initial crystallization conditions was performed in the high-throughput (HT) crystallization service unit at the EMBL, Hamburg, Germany (Mueller-Dieckmann,2006). Crystal and Index Screens (Hampton) were used for the initial experiments. 200 nl protein samples were mixed with an equal amount of the reservoir solution and equilibrated against 50 µl reservoir solution, and the crystallization plates were stored at 19°C. In the next step, optimum pH, precipitant, protein and additive concentration, and the drop size were adjusted in in-house experiments according to the promising results of the HT screening. In the optimization screens, the protein sample was

mixed with reservoir solution in 1:1 ratio and equilibrated in sitting- or hanging-drops against 1 ml reservoir solution at 19°C.

Crystallization of NodS-SAH complex

For co-crystallization with S-adenosyl-L-homocysteine, a protein sample at 4 $mg \cdot ml^{-1}$ was mixed with a stoichiometric amount of 2 mM SAH and incubated overnight at 4°C. Prior to setting up the crystallization screen, the sample was concentrated to 10 $mg \cdot ml^{-1}$ and filtrated using an Ultrafree-MC Centrifugal Filter Unit with 0.1 µm pore size at 4°C.

Data Collection and Processing

The NodS crystals which are reached their final size were subjected to X-ray to obtain diffraction datasets. Mainly, a synchrotron beam line was used as X-ray source. The data from single NodS crystal were collected by using the oscillation method in a stream of cold nitrogen gas at -173 °C (i.e. 100 K). To prevent ice formation, the crystals are typically equilibrated with their mother liquor containing a cryoprotective agent. In these experiments, PEG400 were used for the NodS crystals as cryoprotectant. Each crystals equilibrated with the cryoprotectant contained solution was mounted in a small cryoloop and then the cryoloop was attached to the goniometer head which is between detector and X-ray source. The flash frozen samples were oriented by means of two perpendicular arcs of the goniometer head. During data collection, the samples were rotated through a small angle for recording of one image. After collecting of the first image, diffraction data were analyzed in order to establish parameters of the experiment such as angle of rotation, crystal-to-detector distance and exposure time. The whole oscillation range was defined based on Bravais lattice and the crystal orientation.

X-ray diffraction data processing and scaling were performed by the HKL-2000 package (Otwinowski *et al.*, 1997). The collected diffraction images were auto-indexed by using first few images. Before the indexing, peak search was performed over the selected images according to 3D window value (e.g. 3-5). After indexing, the list of Bravais lattice and unit cell values was showed. From the list, possible highest symmetry was chosen. When the Bravais lattice was defined, the diffraction geometry refinement was performed by fitting primarily the crystal rotations, unit cell parameters, and the beam positions. After the refinement was converged, other parameters were included for the final refinement. In this step, profile fitting and integration box sizes were examined. Box size and spot size were adjusted according to strongest reflection in the first image and the refinement was applied

again. When χ2 values are near 1.0 and all parameters are converged, integration of the set of frames was performed. The progress of the integration was monitored by examining the χ2, cell constants, crystal rotations, mosaicity, and distance vs. frame plots, as well as the agreement between the predicted reflections and the spots on each image. When all of these plots yield horizontal lines and χ2 values are near 1.0, scaling of data set was performed.

During scaling process, the data was converted to a common scale and symmetry related reflections were merged. The scaling operation was started with lowest symmetry. Error model was adjusted to bring χ2 values of the resolution shells close to 1. The characteristic of the dataset was determined up on statistical evidences: <I/σ(I)>, completeness and Rmerge. In order to define final resolution of the dataset, the data was discarded where final resolution shell <I/σ(I)> was below 2. To decide space group of the dataset, different Laue groups were checked starting from the lowest symmetry. When χ2 went to unreasonable values, the previous choice was the correct one. Screw axes were assigned based on systematic absences. In case of expected anomalous signal in the dataset, the anomalous flag was set during scaling. Thereby, I+ and I- reflections are treated as two separate measurements within the data set. When the dataset was scaled well, the presence of an anomalous signal was detected by examining the graph, χ2 and Rmerge versus resolution. If there is no useful anomalous signal in the data, the curves showing the χ2 resolution dependence should be flat and about 1 for scaling with merged and un-merged Friedel pairs. On the other hand, if χ2 is greater than 1 and the clear resolution dependence of the χ2 for scaling with merged Friedel pairs, there is a strong indication of the presence of an anomalous signal (For complete methodology of data processing, please check HKL-2000 manual).

Phasing

Achieving a structure in a crystal, two terms are important. These are amplitude and phase. Crystallographic diffraction experiment supplies only the amplitude but not the phase (so-called phase problem). The intensity of diffracted beam $I(hkl)$ is proportional to the square of the amplitude of the structure factor $F(hkl)$:

$$I(hkl) = |F(hkl)|^2$$

The structure factor is a function of the electron density distribution in the unit cell:

$$F(hkl) = \sum_{j=1}^{N} f_j(hkl) e^{2\pi i(hx_j + ky_j + lz_j)}$$

The structure factor *F*(*hkl*), which can be represented as a complex vector in the Argand diagram (Figure 9) with an amplitude $|F(hkl)|$ and a phase angle $\varphi(hkl)$, is a complete description of a diffracted ray recorded as reflection hkl.

The aim of the phasing is to calculated electron density ρ at every position x, y, z in the unit cell. Electron density distribution *ρ(x,y,z)* is Fourier transform of *F*(*hkl*) and, therefore, *ρ(x,y,z)* can be written as a function of all *F*(*hkl*):

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F(hkl) e^{-2\pi i(hx + ky + lz)}$$

Because $F(hkl) = |F(hkl)| e^{i\varphi(hkl)}$, we can also write

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F(hkl)| e^{-2\pi i(hx + ky + lz) + i\varphi(hkl)}$$

Although, the structure factor amplitude$|F(hkl)|$ can be derived from the intensities $I(hkl)$, the phase angles $\varphi(hkl)$ cannot be derived straightforwardly from the diffraction pattern. Fortunately, several methods have been developed to solve this problem. These indirect methods are molecular replacement, isomorphous replacement, and single- or multi-wavelength anomalous diffraction.

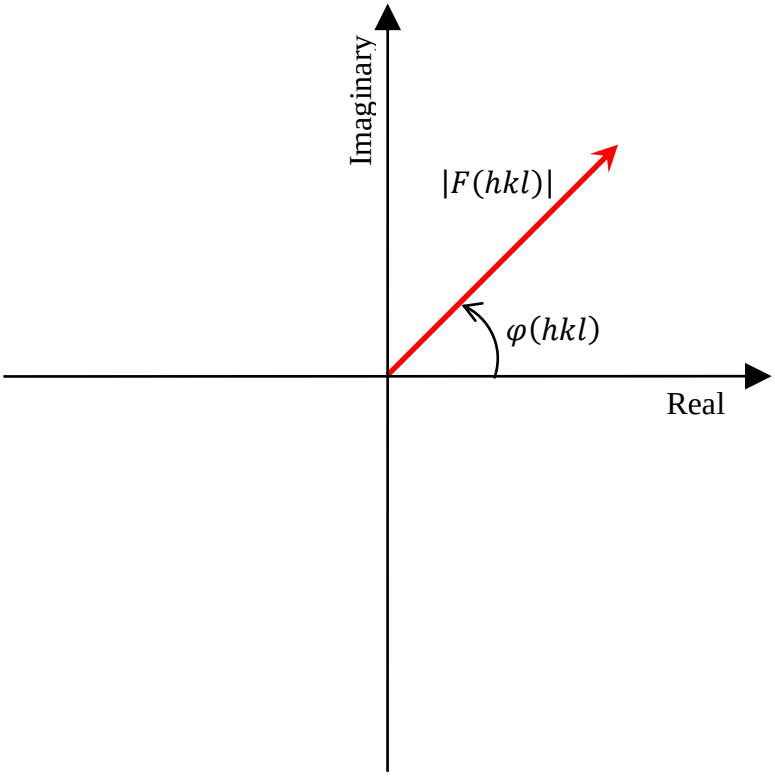


Figure 9 The structure factor F represented as a vector on the plane of complex numbers. $|\boldsymbol{F(hkl)}|$ is a length of the vector, which is proportional to the squared root of measured intensity. $\boldsymbol{\varphi(hkl)}$ is an angle between the vector and the positive real axis.

Multi-wavelength Anomalous Diffraction (MAD)

The possibility to obtain phase information from anomalous scattering has been known since late 1940s (Bijvoet,1949). Anomalous scattering by an atom is due to the fact that its electrons cannot be regarded as completely free electrons. This effect depends on the wavelength. But it is generally stronger for heavier atoms than for the light atoms in the periodic system. So, presence of heavy atoms in a protein structure are resulted in anomalous scattering and the intensities of a reflection hkl and its Bijvoet mate $\bar{h}\bar{k}\bar{l}$ are no longer equal. If there is anomalous scatterer in the crystal, the atomic scattering factor $f(hkl)$ for the heavy atom is described as a complex number $f = f° + f' + if''$ (Figure 10), where $f°$ is the "normal" atomic scattering factor and $(f' + if'')$ is the anomalous scattering correction to take account of the fact that the atomic scattering factor depends on the wavelength near the absorption edge of the atom. The real component f' (the dispersive correction) scatters in phase with the primary wave while the imaginary component f'' (absorptive or Bijvoet correction) is out of phase by $\pi/2$. As a result of the absorption, Friedel's law is broken and the reflections hkl and $\bar{h}\bar{k}\bar{l}$ are not equal in intensity (Figure 10).

There are two general approaches to solve structure by MAD:

a) Algebraic formalism based on direct solving of MAD observational equation (Hendrickson *et al.*, 1988):

$$\left|{}^{\lambda}F_{obs}(\pm hkl)\right|^2 = |F_T|^2 + a_\lambda|F_A|^2 + b_\lambda|F_T||F_A|\cos(\varphi_T - \varphi_A) \pm c_\lambda|F_T||F_A|\sin(\varphi_T - \varphi_A)$$

Where $a_\lambda = [(f'')^2 + (f')^2]/(f°)^2$, $b_\lambda = 2f'/f°$ and $c_\lambda = 2f''/f°$

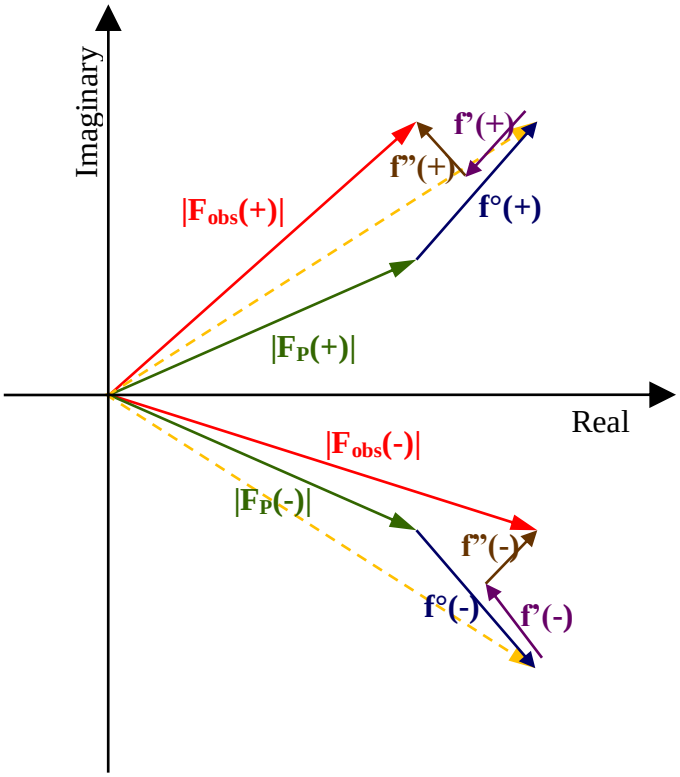


Figure 10 Argand diagram illustrating anomalous scattering. $|F_{obs}(+)|$ and $|F_{obs}(-)|$: observed structure factor amplitudes for the reflections *hkl* and *–h-k-l*, respectively. $|F_P|$: amplitude for normal scattering of the protein atoms, *f*°: amplitude for normal scattering of the anomalous scatterers, *f*': dispersive component of anomalous scattering factor, *f*": Bijvoet component of anomalous scattering factor. In yellow, $|F_T|$: amplitude for normal scattering of all the atoms.

The symbols used here are equivalent to that in Figure 10. $|F_A|$ is structure factor amplitude for normal scattering of the anomalous scatterers. The wavelength-dependent components (corresponding to a, b and c) can be estimated directly and those that are wavelength independent namely $|F_T|$, $|F_A|$ and the phase difference ($\varphi_T - \varphi_A$) can be obtained by least-square fitting to the observational equation. From the derived $|F_A|$, the anomalous scatterer substructure is determined by Patterson or direct methods and φ_A can be calculated. To calculate an electron density map based on normal scattering by all atoms, φ_T is derived from the phase difference ($\varphi_T - \varphi_A$).

b) Another approach to solve the structure by MAD is pseudo-MIR method where data at one wavelength are considered as native while the other wavelengths as derivative data (Burling *et al.*, 1996).

Typically, for MAD experiment three wavelengths are used to collect dataset, a dataset corresponding to the minimum value of f' (inflection point), a dataset close to the maximum f'' (peak), and a remote dataset which is usually chosen on the high-energy side of the absorption edge due to the larger Bijvoet signal. Since the chemical environment of the anomalous scatterer shifts the edge it is necessary to find the peak wavelength by measuring the X-ray absorption spectrum as function of the energy, and then to calculate f' from f'' by the Kramers-Kronig transformation (James,1982).

MAD dataset for SeMet derivative of NodS-SAH complex were initially analyzed by SHELX C, D and E (Sheldrick,2008). For MAD phasing and initial model building, AutoSol wizard in PHENIX package was used (Adams *et al.*, 2002). Scaled dataset, sequence of the protein, number of expected heavy atom sites, type of the anomalously-scattering atom and f' and f'' for each wavelength were supplied to run the wizard. The AutoSol Wizard was defined to use HYSS, SOLVE, RESOLVE, xtriage and phenix.refine to solve a structure and generate experimental phases with the MAD methods. The Wizard begins with data files containing amplitudes of structure factors, identifies heavy-atom sites, calculates phases, carries out density modification and NCS identification, and builds and refines a preliminary model.

Molecular Replacement (MR)

Obtaining a structure by molecular replacement method was first introduced in beginning of 1960s (Rossmann *et al.*, 1962). The idea behind the method is that identical or similar structures found in different crystallographic environments are expected to have some similarities between their diffraction patterns. Thereby, a known homologous model helps to determine orientation (rotation function) and position (translation function) of the searching model within the crystal cell. The molecular replacement can be based on the comparison of the observed Patterson map and that which is calculated for the model. As a model structure is available it is also possible to directly compare correlation coefficient (or Rf) between the observed and calculated intensities or structure factor amplitudes. The success of this method depends strongly on the similarity between the model and searched protein. Generally, it is assumed that the amino acid sequence identity between the two proteins should be higher than 20% and the root mean square deviation for the superpositions of their Cα atoms should not exceed 2.0 Å. Even if it is possible to obtain the solution using low-homology model, the phase angles are poor estimates of the true phase angles and there is a high bias towards the

model making it difficult to refine the structure. The molecular replacement was performed using automated molecular replacement with AutoMR in PHENIX package (Adams *et al.*, 2002). Search models for molecular replacement were prepared from the structure of the NodS-SAH complex.

Model Building

Preliminary model building for ligand free NodS and SeMet derivative NodS-SAH complex was performed automatically by using AutoBuild wizard in PHENIX package (Adams *et al.*, 2002). Model building for native NodS-SAH complex was preliminary performed by using and ARP/wARP (Perrakis *et al.*, 1999). Starting models were manually edited to obtain complete model by using COOT (Emsley *et al.*, 2004). Calculated 2Fo-Fc and Fo-Fc maps were used to build missing parts in models. Building of solvent structure was performed via ARP/wARP by using its automated solvent building option. Water molecules located by ARP/wARP were visually revised and validate by using COOT.

Refinement

The refinement techniques in protein X-ray crystallography are based on the principal of least squares or maximum likelihood (Tronrud,2004). In the least-squares refinement, the function $\sum_{hkl}(|F_{obs}| - |F_{calc}|)^2$ is minimized. The least-squares method assumes that the errors in the observations obey a normal distribution. Therefore, incomplete models may lead to some refinement problems. This is not a case in maximum-likelihood refinement where the model is adjusted to maximize the probability of given observations.

In macromolecular refinement, the ratio between experimental observations to refinable parameters (x, y, z, B factor) in most cases is insufficient (below 1). Increase in data-to-parameters ratio can be achieved by using constraints or restraints. Constraints are fixed values for given parameters. For example in early stages, rigid body refinement can be conducted where only orientation and position of the molecule is refined and other parameters, such as bond lengths and angles are fixed. Alternatively, the number of data can be increased by adding information from small molecule crystal structures in the form of restraints on bond lengths, bond angles, dihedral angles, chiral volume and planarity.

The number of data available influence the way of temperature factors refinement. In the most desirable situation, individual atomic displacement parameters are described as ellipsoids represented by six parameters per atom. To reduce the number of parameters at

medium resolution, isotropic B factors are refined where it is assumed that the motion is the same in each direction. To introduce anisotropy in the description of disorder it is possible to refine anisotropic rigid-body displacement parameters in the form of TLS refinement. Here, the correlated motion of rigid groups is modeled by a single tensor that describes the translation, libration and screw-rotation. Mainly REFMAC5 was used for structural refinement of the models.

REFMAC5

In Refmac5 (Murshudov *et al.*, 1997) the unrestrained refinement of atomic positions and B factors by maximum likelihood method was carried out. The TLS parameters (Winn *et al.*, 2001), defined as separate TLS group for each molecule in the asymmetric unit, were refined for 10 cycles prior to individual atomic refinement. In subsequent refinement cycles the file containing refined TLS parameters was used. The refinement worked best with the default settings with exception that isotropic B factors were not pre-set to a constant value as suggested by authors.

Validation

Finalized structures were validated by using PROCHECK (Laskowski *et al.*, 1993).

PROCHECK

The PROCHECK suite of programs was used throughout the refinement to control the stereochemical correctness of the structure as well as for final validation. The program performs the following checks:

1. Ramachandran plot of residues φ-ψ torsion angles (Ramachandran *et al.*, 1963).
2. Ramachandran plots for all residue types
3. Plots of side chains torsion angle values.
4. General main chain parameters.
5. General side chains parameters.
6. List of properties of individual residues.
7. Plots of main chain bonds length and angle distribution.
8. RMS deviations from planarity in side chains of Arg, Asn, Asp, Gln, Glu, His, Phe, Trp and Tyr.
9. The list of residues with distorted geometry (main chain bond lengths or angles and planarity).

Sequence Alignments

Multiple sequence alignments were performed in ClustalW program (Thompson *et al.*, 1994). First, it does a pairwise comparison of every sequence. The multiple alignment is performed starting with the pair of sequences that is most similar. Then, the sequences are added one by one to the alignment based on their similarities to the starting pair. The alignments were visualized in Jalview (Clamp *et al.*, 2004).

Docking of Chitooligosaccharide

The docking simulation of a chitooligosaccharide substrate in the NodS-SAH complex structure was carried out using the AutoDock4 software trough the AutoDockTools GUI (Morris *et al.*, 1998). Prior to the simulation, the SAH molecule in the complex structure was converted to SAM, and a disordered N-terminal tail consisting of twelve residues (including the cloning heptapeptide) was removed. Hydrogen atoms were added to all molecules at predicted positions and the Gasteiger charges were calculated by the AutoDock ADT tool. During the docking calculations, the complex structure was treated as a rigid molecule. As a ligand for the docking procedure, a chitotriose (COS) molecule was retrieved from the GLYCAM (Kirschner *et al.*, 2008) oligosaccharide library. It was modified by removal (done in PyMOL) of the acetyl group from the non-reducing end of the molecule. Three torsion angles of the ligand were set as active rotatable bonds (C2-N2 bond of $GlcNH_2$, C1-O4 bonds between the monomers). As the docking site, the most pronounced groove on the surface of the NodS-SAH structure was chosen, within which a 15 x 15 x 22.5 Å grid box was built with default grid spacing (0.375Å). Lamarckian Genetic Algorithm was executed for the docking simulation with the following parameters: 10 runs, 150 population size, 2,500,000 evaluations, and 27,000 generations. Mostly default values were used for all other parameters. The calculated binding orientations of the COS molecule were ranked according to the interaction energy. Among the solutions, the final conformation of the docked ligand was chosen based on the orientation of the ligand, the calculated interaction energy and the RMSD from the reference (starting) conformation.

Art Work

Molecular and electron-density illustrations were prepared using the programs PyMOL (Delano,2008).

RESULTS

Results

Cloning and Expression

Cloning and transformation of nodulation genes

Genomic DNA of *Bradyrhizobium japonicum* WM9 was used as template DNA to isolate (clone) target nodulation genes which are *nolA, nodA, nodB, nodD, nodS* and *nodW*. The designed primers for target genes were tested by using Taq DNA polymerase in polymerase chain reaction (PCR). All target genes were successfully amplified as single products (Figure 11A). To obtain blunt end PCR products, *Pwo* DNA polymerase was used and only five of the target genes were amplified as a single fragment (Figure 11B). Three fragments were obtained from amplification of *nodA*. One of those bands was referred to exact size of the amplified *nodA*. All PCR products were loaded on 0.6% of agarose gel. After agarose gel electrophoresis, target bands were extracted and purified from the agarose gel and they were used for TOPO cloning.

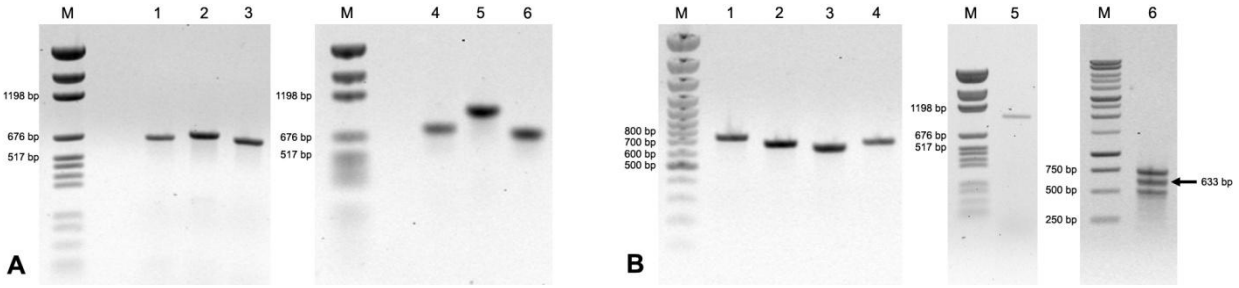


Figure 11 Nodulation genes amplified from genomic DNA of *Bradyrhizobium sp*. WM9. PCR products were analyzed with 1% agarose gel. A) Amplification with Taq DNA polymerase, Lane1: *nodA* (633bp), Lane2: *nodB* (663bp), Lane3: *nodS* (633bp), Lane4: *nolA* (720bp), Lane5: *nodD* (924bp), Lane6: *nodW* (684bp), B) Amplification with Pwo DNA Polymerase, Lane 1: *nolA* (720 bp), Lane 2: *nodB* (663 bp), Lane 3: *nodS* (633 bp), Lane 4: *nodW* (684 bp), Lane 5: *nodD* (924 bp), Lane 6: *nodA* (633 bp).

Amplified gene fragments were cloned into pET151/D-TOPO expression vector containing N-terminal His-tag. The cloning was carried out using the TOPO system. TOP10 cells were transformed with the recombinant vectors for selection. The cells were streaked on selective solid growth media containing selective antibiotics. After overnight growth at 37 °C, five colonies were picked for plasmid isolation. Isolated plasmids were analyzed via PCR

for correct oriented insertion of the target gene fragment. Proved plasmid vectors were used for sequencing analysis. Positive constructs were successfully obtained for *nolA*, *nodB*, *nodD*, *nodS* and *nodW* genes. These were used for successive protein expression.

Expression and solubility of nodulation proteins

Competent *E.coli* BL21 cells were transformed with the positive constructs of the five nodulation genes. In order to test overexpression of the target genes, the transformants were grown in broth media containing appropriate antibiotic at 37°C. Expression of the nodulation protein was induced by isopropyl-β-D-thiogalactopyranoside (IPTG) when the growth was between the mid-log phase and stationary phase. After three hours of induction, overexpression of five nodulation proteins indicated as NolA, NodB, NodD, NodS and NodW was observed (Figure 12).

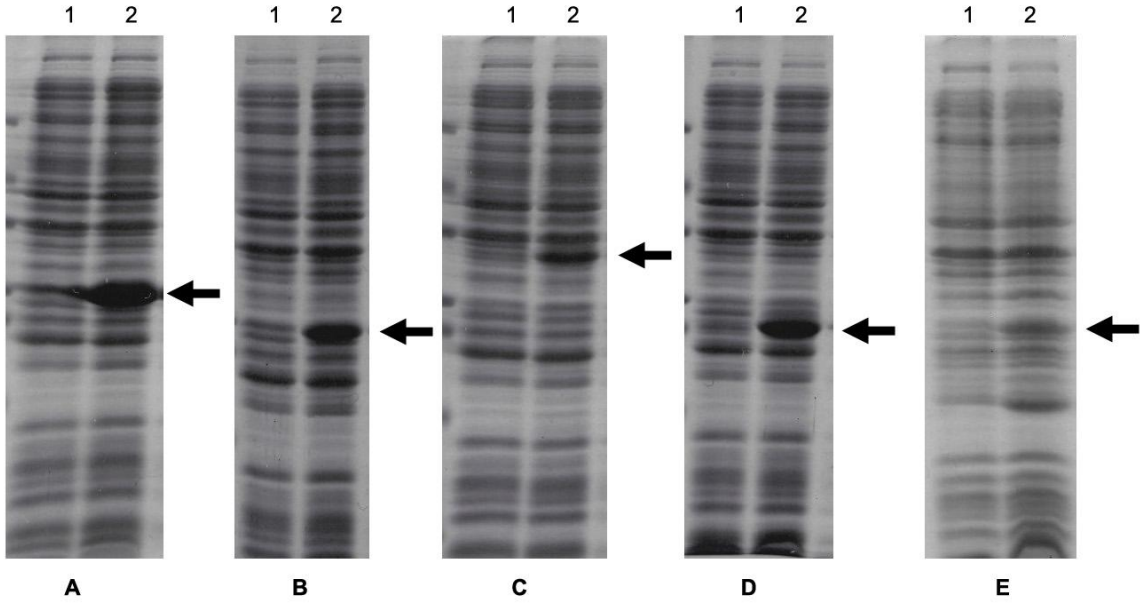


Figure 12 Expression of recombinant nodulation proteins in *E. coli* BL21 competent cells. Lane 1: before induction Lane 2: after induction with IPTG. A) NolA, B) NodB, C) NodD, D) NodS, E) NodW.

Overexpressed proteins were analyzed for their solubility in small scale. Two different conditions were tested for the protein solubility. These were three hours growth in LB media at 37 °C induced with 1mM IPTG and overnight growth in TB media at 18-20 °C induced with 0.2 mM IPTG. After overexpression, pelletted cells were lysed with a buffer containing 50 mM Potassium phosphate pH 7.8, 0.4 M NaCl, 0.1 M KCl, 10% Glycerol,

0.5% Triton X-100, 10 mM Imidazole. Soluble and insoluble fractions were analyzed by SDS-PAGE. Only NodS and NodB, which were expressed overnight at low temperature, showed significant solubility in the buffer (Figure 13).

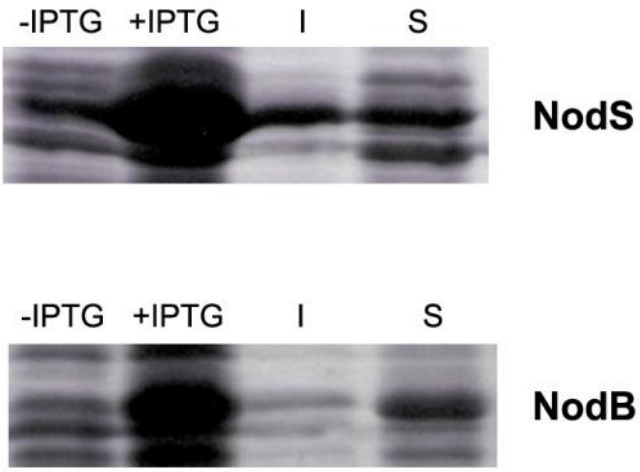


Figure 13 Solubility analysis of over-expressed recombinant proteins. (I: insoluble fraction, S: soluble fraction)

Insoluble proteins were subjected to basic renaturation procedure. Pelletted cells were lysed by a buffer containing 50 mM HEPES pH 7.5, 0.5 M NaCl, 5% glycerol, 1 $mg \cdot ml^{-1}$ lysozyme, 1 mM TCEP, 1 mM PMSF. After sonication and centrifugation, obtained inclusion bodies were resolubilized with a denaturizing agent containing 50mM Buffer, 0.5 M NaCl, 5% Glycerol, 6M guanidine-HCl, 25 mM DTT. The buffer and its pH were chosen according to pI values of the proteins. After resolubilization, denatured proteins were tried to refold via diluting the denaturizing agents either quick dilution or overnight dialysis against 50 mM Buffer, 500 mM NaCl, 5% Glycerol, 1 mM TCEP at 4 °C. Renaturation trials for NolA and NodD were resulted in aggregation. Only NodW was obtained as refolded protein by using quick dilution method and 1 day incubation at 4 °C (Figure 14).

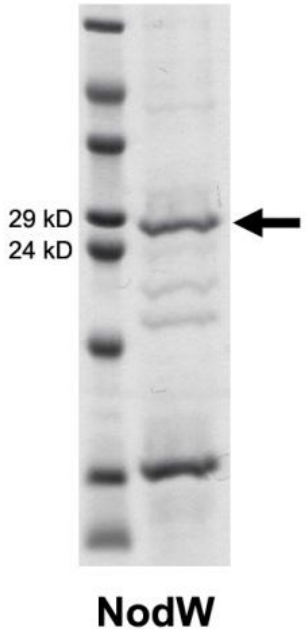


Figure 14 NodW after renaturation.

Purification

Soluble recombinant NodS and NodB protein from *Bradyrhizobium* sp. WM9 was obtained by overnight cultivation of transformed *E. coli* cells at 18-20 °C. NodS and NodB were expressed in large scale for further purification method.

Purification of NodS

The cell paste obtained from NodS expression was lysed with a buffer containing 50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP, 1 mM PMSF, 1 $mg \cdot ml^{-1}$ lysozyme. After sonication and centrifugation, the supernatant was used for purification of the protein by using FPLC. In the first step, His-tagged NodS protein was purified by using HisTrap™ column containing charged Ni sepharose. The purified protein was desalted against binding buffer (50 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 2.5 mM TCEP) to remove imidazole and incubated overnight at 4°C with His-tagged TEV protease to remove His-tag fusion from the protein. Final suspension was applied to HisTrap™ column again to remove the TEV protease, the His-tag and any undigested fusion protein. First flow trough from the latter purification step were desalted against 50 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM DTT, and 1 mM EDTA and concentrated up to 4 $mg \cdot ml^{-1}$. Concentrated sample was applied on a gel filtration column (GFC) (HiLoad 16/60 Superdex 200 pg). After 80 ml elution from GFC, one big narrow peak was observed (Figure 15). Peak fractions were collected and analyzed with SDS-PAGE and Native PAGE. The peak position from the GFC was referred to monomeric homogenous NodS protein (Figure 15). The sample was also applied to mass spectrometry by using MALDI-TOF. Expected molecular weight for NodS is 24,123 Da. Observed data was for m/z =1, 24,071 Da and for m/z = 1/2, 12,060 Da. Because highest molecular marker was about 20 kDa, the value for m/z =1 is not exact. So estimated molecular weight value was calculated from m/z = 1/2 and it was 24,120 Da for the NodS (Figure 16). NodS was also produced and purified in selenomethionyl form using non-auxotrophic *E. coli* cells and a cultivation protocol that blocks the methionine biosynthetic pathway. Successful Met→Se-Met substitution of the six methionine sites was confirmed by mass spectrometry. Expected molecular weight for SeMet derivative NodS is 24,404.7 Da. Observed data was for m/z =1, 24,327.9 Da and for m/z = 1/2, 12,192.4 Da. Because highest molecular marker was about 20 kDa value for m/z =1 is not exact. So estimated molecular weight value was calculated from m/z = 1/2 and it was 24,384.8 Da for the NodS Se-Met (Figure 16).

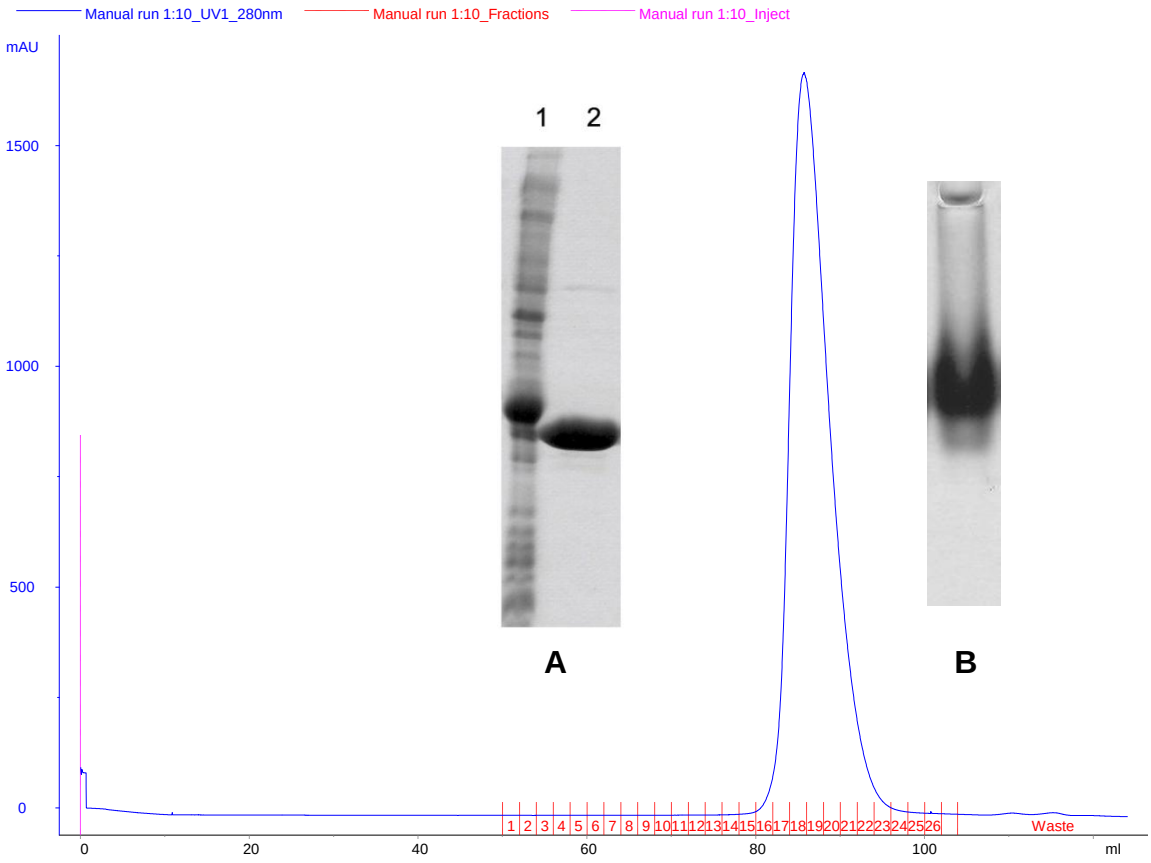


Figure 15 Gel filtration chromatography result for NodS. A) SDS PAGE analysis of purified NodS; Lane 1: Over-expressed protein with 6xHis-Tag, Lane 2: Pure Protein without 6xHis-Tag. B) Native PAGE (10%) analysis of purified NodS.

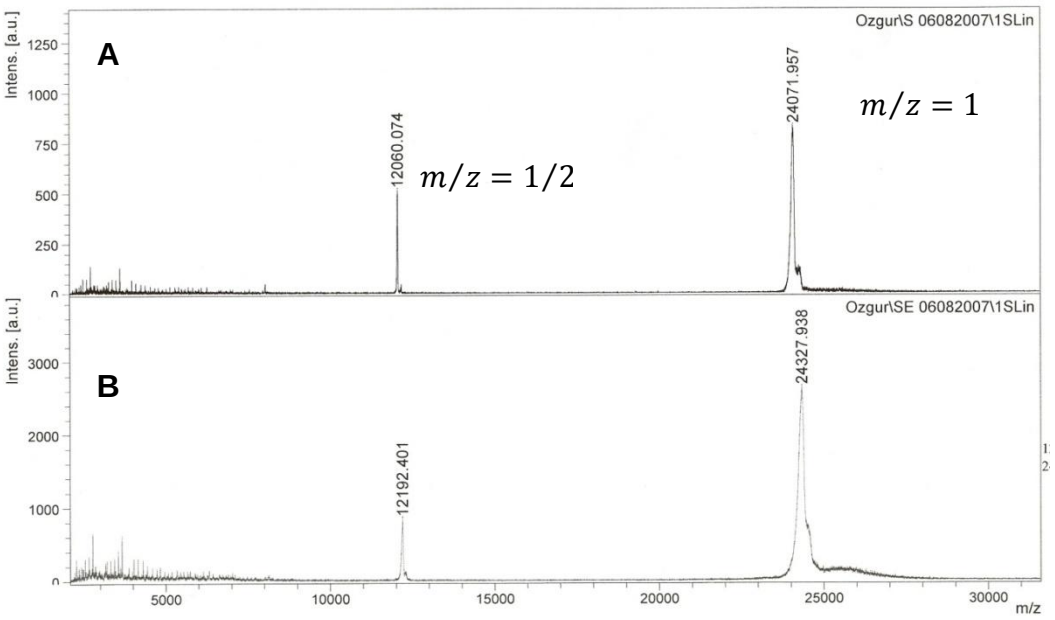


Figure 16 Mass spectrometry analysis of purified NodS and SeMet derivative NodS. A) NodS, B) SeMet derivative NodS.

Purification of NodB

The cell paste obtained from NodB expression was lysed with a buffer containing 50 mM Tris pH 7.2, 0.5 M NaCl, 10% glycerol, 2.5 mM TCEP, 1 mM PMSF, 1 $mg \cdot ml^{-1}$ lysozyme. After sonication and centrifugation, the supernatant was used for purification of the protein by using FPLC. In the first step, His-tagged NodB protein was purified by using HisTrap™ column containing charged Ni sepharose. The purified protein was desalted against binding buffer (50 mM Tris pH 7.2, 0.5 M NaCl, 10% glycerol, 2.5 mM TCEP) to remove imidazole and incubated overnight at 4°C with His-tagged TEV protease to remove His-tag fusion from the protein. Final suspension was applied to HisTrap™ column again to remove the TEV protease, the His-tag and any undigested fusion protein. First flow trough from the latter purification step was desalted against 50 mM Tris pH 7.2, 150 mM NaCl, 10% glycerol, 5 mM DTT, and 1 mM EDTA and concentrated up to 4 $mg \cdot ml^{-1}$. Concentrated sample was applied on a gel filtration column (GFC) (HiLoad 16/60 Superdex 200 pg). After 80 ml elution from GFC, one small peak was observed. Peak fractions were collected and analyzed with SDS-PAGE and Native PAGE. The peak position from the GFC was referred to monomeric NodB protein. Homogeneity of NodB was analyzed by using Native PAGE. Several bands were observed on the electrophoregram (Figure 17).

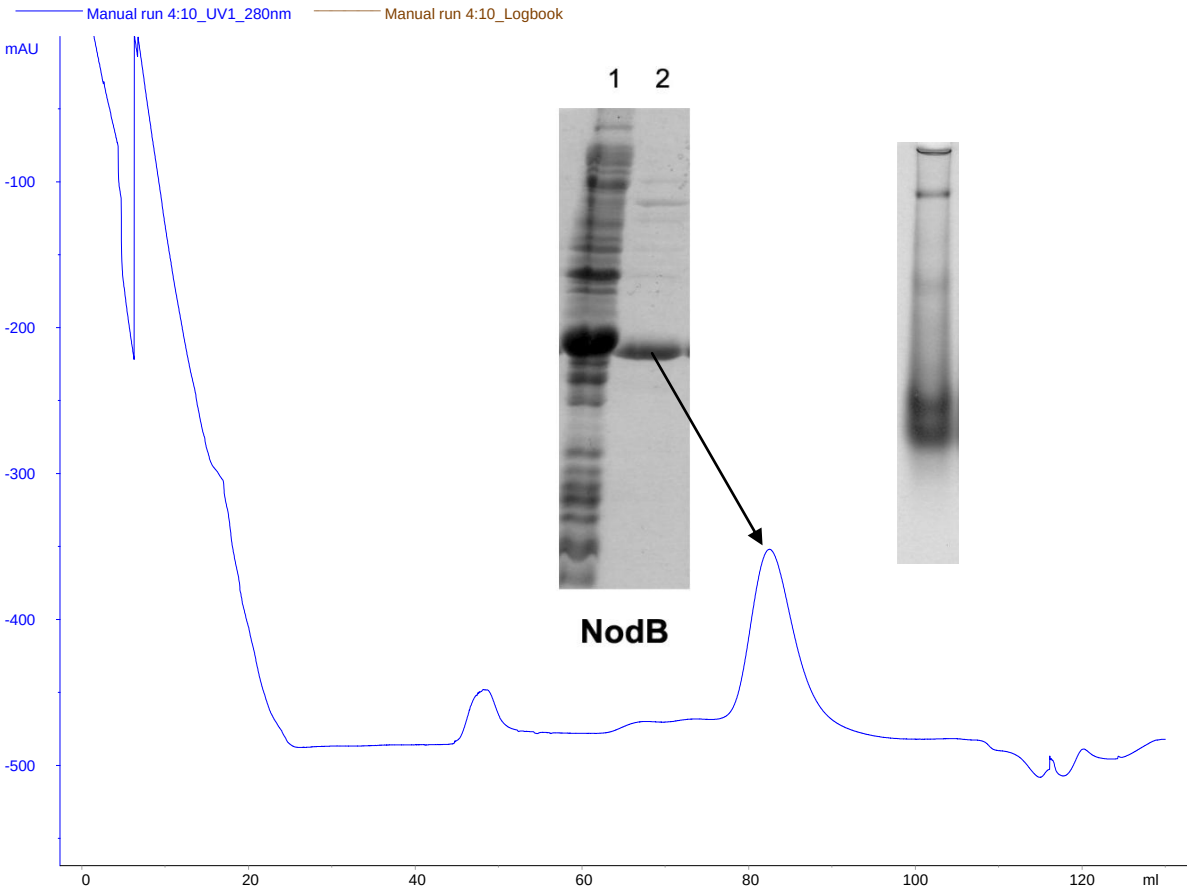


Figure 17 Gel filtration chromatography result for NodB. A) SDS PAGE analysis of purified NodB; Lane 1: Over-expressed protein with 6xHis-Tag, Lane 2: Pure Protein without 6xHis-Tag. B) Native PAGE (10%) analysis of purified NodB.

The sample was also applied to mass spectroscopy by using MALDI-TOF. Expected molecular weight for NodB is 24,228 Da. Observed data was for m/z =1, 24,167.7 Da and for m/z = 1/2, 12,106.8 Da. Because highest molecular marker was about 20 kDa value for m/z =1 is not exact. So estimated molecular weight value was calculated from m/z = 1/2 and it was 24,213.6 Da for the NodB (Figure 18).

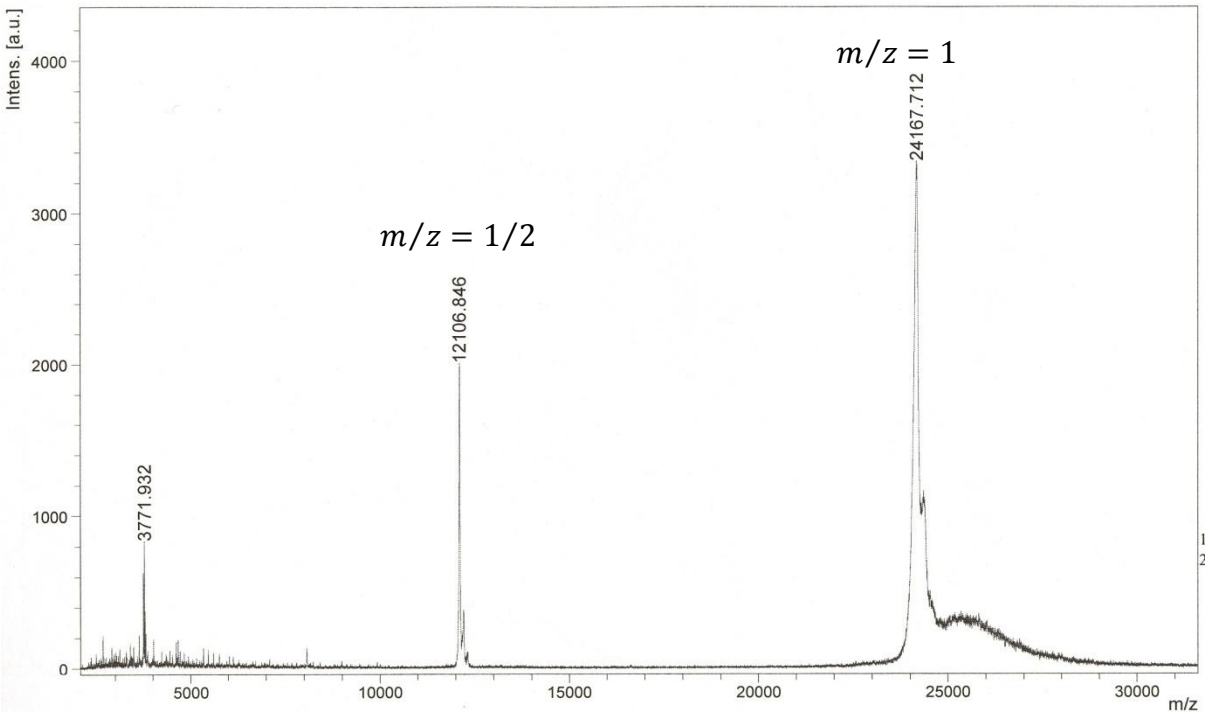


Figure 18 Mass spectrometry analysis of purified NodB.

Purification of NodW

Refolded NodW was applied to Ni-NTA gravity column equilibrated with binding buffer containing 50 mM MES pH 6.0, 500 mM NaCl, 5% Glycerol, 1 mM TCEP. The column was washed with 5, 10, 20, 50 mM imidazole in binding buffer. The protein was eluted from the column with 250 mM imidazole in binding buffer (Figure 19). Buffer of the eluted sample was exchanged with binding buffer by using Ultrafree Centrifugal Filter Units (Millipore) with 10 kDa cut-off at 4°C and stored at -20 °C.

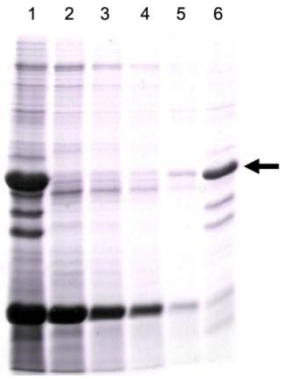
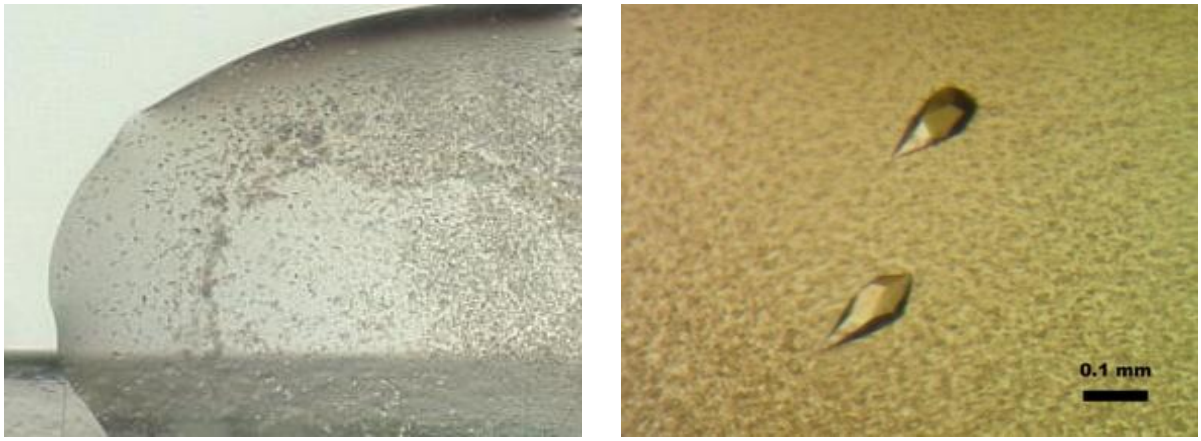


Figure 19 Purification of NodW with Ni-NTA column. Lane 1 Applied sample, Lane 2 -6 Elution 5, 10, 20, 50 and 250 mM Imidazole, respectively.

Crystallization

Crystallization of Ligand free NodS

Approximately 10 mg·ml^{-1} of homogenous NodS and Se-Met derivative NodS proteins were used for high-throughput crystallization screening. Crystal and Index Screens (Hampton Research) were used for the initial high-throughput experiments. Initial hit for ligand-free NodS were found in index screen (Figure 20A). Microcrystals were observed in 30 minutes after setting the drop. The condition was optimized to obtain large crystals. During optimization screening, effect of variation at protein concentration, precipitant concentration, salt concentration, pH and temperature were searched. After optimization trials, the best diffracting crystals of ligand-free NodS were obtained at 19°C in 28% PEG3350, 0.1 M $MgCl_2$, pH 7.9 (Figure 20B). The crystals appeared after one day and grew to their final dimensions in five days. Similar crystallization conditions were not applicable to obtain diffracting crystal from Se-Met derivative NodS.



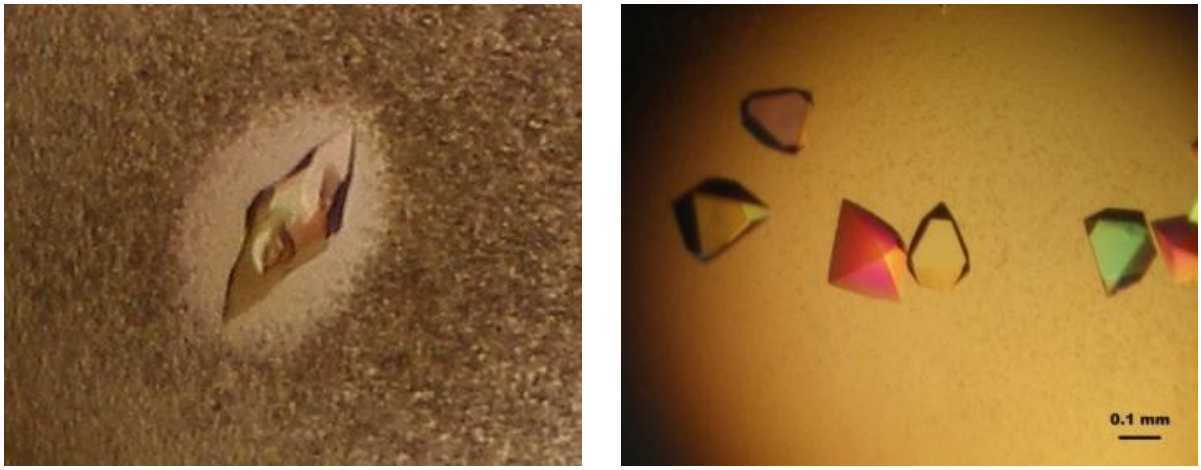
A **B**

Figure 20 Crystallization results for ligand free NodS. A) Initial hit after HT screening. B) Diffractable crystals.

Crystallization of Native NodS-SAH complex and SeMet Derivative NodS-SAH complex

The crystal condition for ligand free NodS was also tested for co-crystallization of NodS with S-adenosyl-L-homocysteine (SAH). Effect of different PEG precipitants and their concentrations (eg. PEG4K, PEG6K and PEG8K) were searched for SeMet derivative NodS-SAH complex. Initial hit for SeMet derivative NodS-SAH complex were found in 25% PEG 8000, 50 mM $MgCl_2$, pH 7.5 (Figure 21A). After optimization trials, the best crystals of SeMet derivative NodS-SAH complex were obtained at 19 °C in 14% PEG8000, 40 mM $MgCl_2$, pH 8.5, using the streak seeding method and protein concentration of 4 mg·ml^{-1}. The crystals appeared after one day and grew to their final dimensions in seven days at 19°C

(Figure 21B). The native NodS-SAH complex was crystallized in similar conditions (10 $mg \cdot ml^{-1}$ protein concentration, in 16% PEG8000, 5 mM $MgCl_2$, pH 8.5, at 19°C), also with the aid of streak seeding (Figure 22).



A B

Figure 21 Crystallization results for SeMet derivative NodS-SAH complex. A) Initial hit B) Diffractable crystals obtained by using streak seeding method.

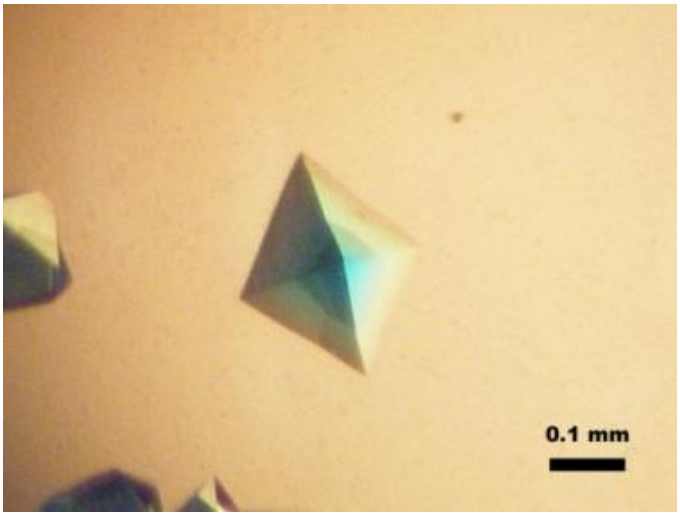


Figure 22 Crystallization result for native NodS-SAH complex. The crystal was obtained by using streak seeding method.

Data collection and Processing

Ligand free NodS

X-ray diffraction data of the ligand free NodS crystal were collected at the EMBL beamline X13 of the DESY synchrotron in Hamburg, using a 165 mm MAR CCD detector (Figure 23). During data collection, cryoprotection of the crystal was supplied with 25% (v/v) PEG400. Indexing, integration and scaling of all diffraction images were performed in HKL-2000 (Otwinowski *et al.*, 1997). The crystals of ligand free NodS are tetragonal, space group $P4_322$, with one protein molecules in the asymmetric unit estimated at a Matthews coefficient (Matthews,1968) of 1.74 $Å^3.Da^{-1}$ (solvent content 29.2%). The related information and statistics were given in the Table 4.

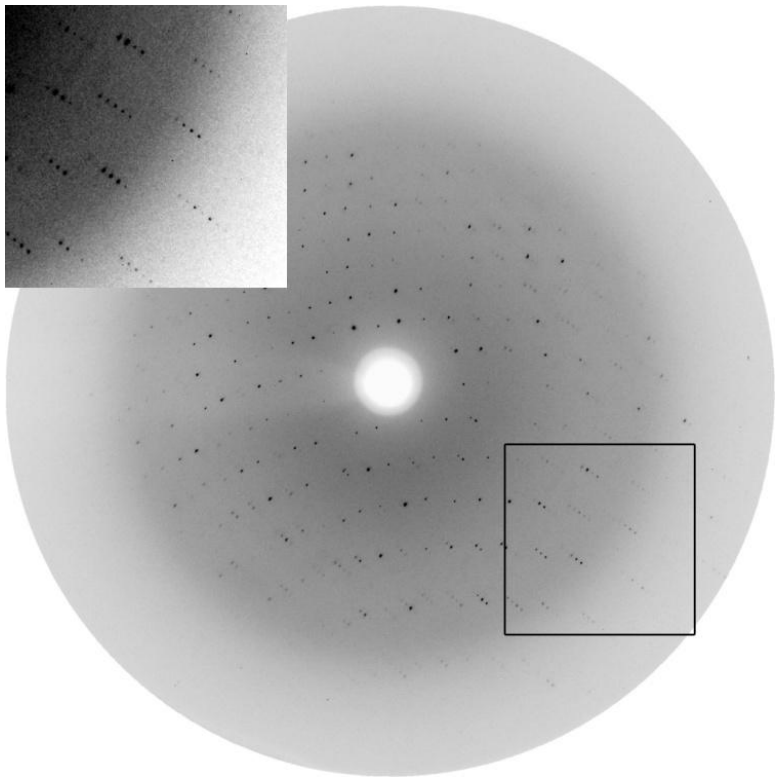


Figure 23 X-ray diffraction patterns recorded for crystal of ligand free NodS (0.75° oscillation) The edge of the detector (framed, inset) corresponds to a resolution of 2.3.

Table 4 Data collection and processing statistics of ligand free NodS protein (values in the parenthesis indicate last resolution shell).

	Native ligand free NodS
Space group	$P4_322$
Unit-cell parameters (Å)	
a	48.68
b	48.68
c	141.46
Radiation source	DESY (Hamburg, Germany)
Beamline	EMBL X13
Wavelength (Å)	0.8086
Temperature (K)	100
No. of molecules in ASU	1
Resolution (Å)	20-2.42 (2.51-2.42)
Mosaicity	0.69
Crystal-to-detector distance (mm)	210
Oscillation (º) /no. of images	0.75/120
No. of observations	46874
No. of unique reflections	6954
Completeness (%)	99.6 (97.3)
Redundancy	6.7
R_{merge} [a]	0.087 (0.440)
$R_{p.i.m.}$ [b]	0.034 (0.176)
<I/σ(I)>	16.7 (2.1)

[a] $R_{merge} = \Sigma_{hkl}\Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th observation of reflection *hkl* and $\langle I(hkl)\rangle$ is the weighted average intensity for all observations *i* of reflection *hkl*.
[b] $R_{p.i.m.} = \Sigma_{hkl} [1/(N-1)]^{1/2} \Sigma_i |I_{i(hkl)} - \langle I_{(hkl)}\rangle| / \Sigma_{hkl}\Sigma_i I_{i(hkl)}$, where *N* is redundancy (Weiss,2001), calculated in *SCALA* (Evans,2006) using data processed with *DENZO* (Otwinowski *et al.*, 1997).

NodS-SAH complex

X-ray diffraction data of the NodS-SAH complex crystal were collected in two runs, corresponding to low (2.3 Å) and high (1.71 Å) resolution at beamline BL 14.2 of the BESSY synchrotron in Berlin using a 165 mm MAR CCD detector (Figure 24). During data collection, cryoprotection of the crystal was supplied with 14% (v/v) PEG400. Indexing, integration and scaling of all diffraction images were performed in HKL-2000 (Otwinowski *et al.*, 1997). The NodS-SAH complex crystallizes in the orthorhombic space group $P2_12_12$ with four molecules in the asymmetric unit estimated at a Matthews coefficient (Matthews,1968) of 2.28 $Å^3.Da^{-1}$ (solvent content 46.6%). The related information and statistics were given in the Table 5.

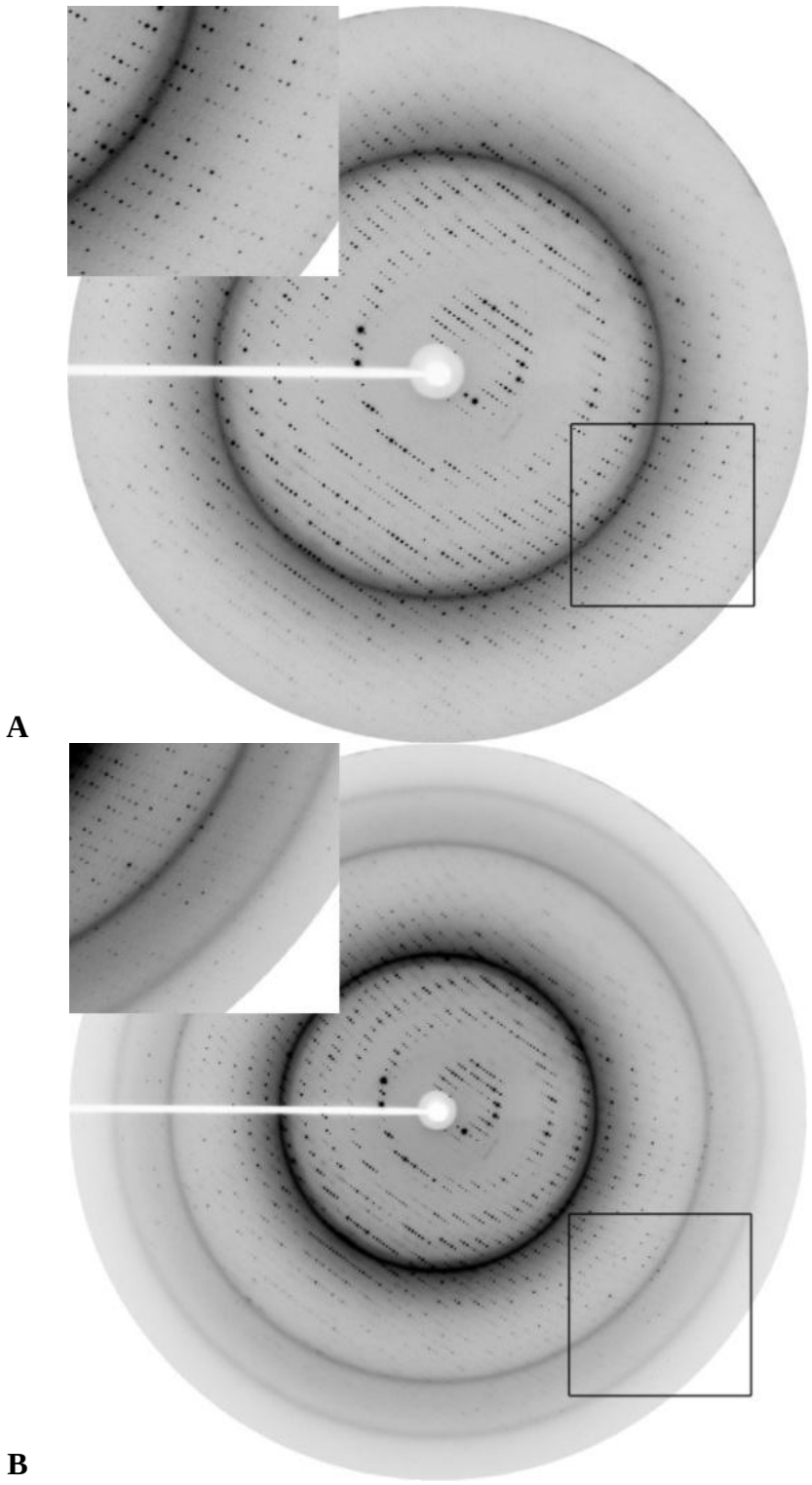


Figure 24 X-ray diffraction patterns recorded for crystal of native NodS-SAH complex. A: diffraction for low resolution, the edge of the detector (framed, inset) corresponds to a resolution of 2.3 Å. B: diffraction for high resolution, the edge of the detector (framed, inset) corresponds to a resolution of 1.71Å.

Table 5 Data collection and processing statistics of native NodS-SAH complex (values in the parenthesis indicate last resolution shell).

	Native NodS-SAH complex
Space group	$P2_12_12$
Unit-cell parameters (Å)	
a	81.01
b	143.30
c	75.85
Radiation source	BESSY (Berlin,Germany)
Beamline	BL 14.2
Wavelength (Å)	0.9184
Temperature (K)	100
No. of molecules in ASU	4
Resolution (Å)	50-1.85 (1.92-1.85)
Mosaicity	0.66
Crystal-to-detector distance (mm)	190 low res.; 135 high res.
Oscillation (º) /no. of images	1/100 low res.; 0.5/200 high res.
No. of observations	366793
No. of unique reflections	72827
Completeness (%)	96.0 (97.8)
Redundancy	5.0
R_{merge} [a]	0.052 (0.347)
$R_{p.i.m.}$ [b]	0.024 (0.146)
<I/σ(I)>	24.6 (2.4)

[a] $R_{merge} = \Sigma_{hkl}\Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th observation of reflection *hkl* and $\langle I(hkl)\rangle$ is the weighted average intensity for all observations *i* of reflection *hkl*.
[b] $R_{p.i.m.} = \Sigma_{hkl} [1/(N-1)]^{1/2} \Sigma_i |I_{i(hkl)} - \langle I_{(hkl)}\rangle| / \Sigma_{hkl}\Sigma_i I_{i(hkl)}$, where *N* is redundancy (Weiss,2001), calculated in *SCALA* (Evans,2006) using data processed with *DENZO* (Otwinowski *et al.*, 1997).

Se-Met Derivative of NodS-SAH complex

The Se-Met derivative NodS-SAH complex crystal were used to collect Multiwavelength Anomalous Diffraction (MAD) data at three different wavelengths. Wavelengths for peak, edge and remote were indentified according to fluorescence scanning of the crystal (Figure 25).

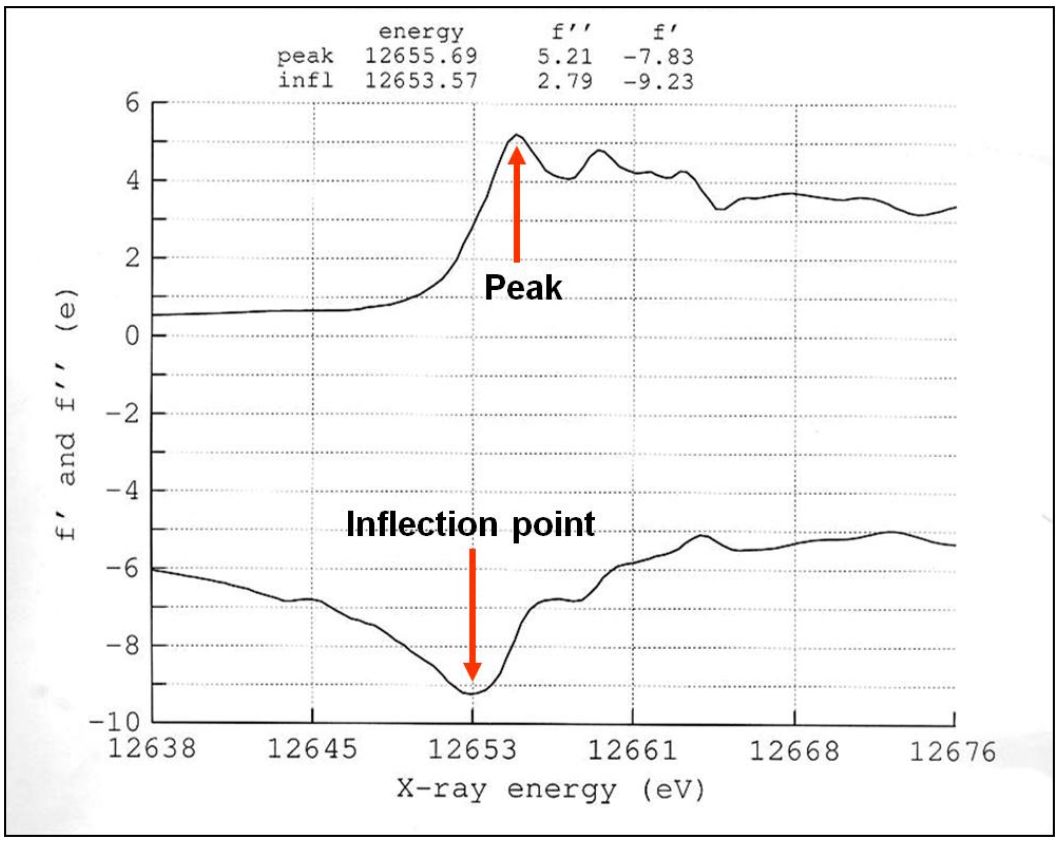


Figure 25 Fluorescence scanning for SeMet derivative NodS-SAH complex crystal.

Data collection was performed at beamline BL 14.2 of the BESSY synchrotron in Berlin, equipped with a 165 mm MAR CCD detector. During data collection, cryoprotection of the crystal was supplied with 12% (v/v) PEG400. Indexing, integration and scaling of all diffraction images were performed in HKL-2000 (Otwinowski *et al.*, 1997). The related information and statistics were given in the Table 6.

Table 6 Data collection and processing statistics of SeMet derivative NodS-SAH complex (values in the parenthesis indicate last resolution shell).

	Se-Met derivative NodS-SAH complex		
	Peak	**Edge**	**Remote**
Space group		$P2_12_12$	
Unit-cell parameters (Å)			
a		81.34	
b		143.52	
c		75.98	
Radiation source		BESSY (Berlin, Germany)	
Beamline		BL 14.2	
Wavelength (Å)	0.97968	0.97984	0.95000
Temperature (K)		100	
No. of molecules in ASU		4	
Resolution (Å)		50-2.00 (2.07-2.00)	
Mosaicity		0.45	
Crystal-to-detector distance (mm)		150	
Oscillation (º) /no. of images	0.75/135	0.75/135	0.75/135
No. of observations	246696	247187	250132
No. of unique reflections	59766	59858	59874
Completeness (%)	98.2 (83.1)	98.5 (85.5)	99.9 (100)
Redundancy	4.1	4.1	4.2
R_{merge} [a]	0.067 (0.634)	0.054 (0.727)	0.055 (0.423)
$R_{p.i.m.}$ [b]	0.037 (0.267)	0.030 (0.306)	0.031 (0.229)
$\langle I/\sigma(I)\rangle$	15.7 (2.4)	21.0 (2.3)	21 (2.9)

[a] $R_{merge} = \Sigma_{hkl}\Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th observation of reflection *hkl* and $\langle I(hkl)\rangle$ is the weighted average intensity for all observations *i* of reflection *hkl*.
[b] $R_{p.i.m.} = \Sigma_{hkl} [1/(N-1)]^{1/2} \Sigma_i |I_{i(hkl)} - \langle I_{(hkl)}\rangle| / \Sigma_{hkl}\Sigma_i I_{i(hkl)}$, where *N* is redundancy (Weiss,2001), calculated in *SCALA* (Evans,2006) using data processed with *DENZO* (Otwinowski *et al.*, 1997).

Structure Solution (Phasing), Model Building and Refinement

NodS-SAH complex

The phase problem for NodS-SAH complex crystal was solved by using Se MAD dataset. As experimental phasing, three wavelength dataset (eg. peak, edge and remote) were analyzed by using SHELX C, D and E (Sheldrick,2008). Hkl2map graphical user interface (GUI) was used to run SHELX C/ D/ E (Pape *et al.*, 2004). Scaled and merged Se MAD dataset was used as input for SHELXC. SHELXC is a program which prepares the three files necessary for running SHELXD (heavy atom location) and SHELXE (phasing and density modification). These contain estimated heavy-atom structure factors $|F_A|$ from Se MAD experiment and the phase shifts that should be added to the heavy-atom reference phases to obtain starting values for the native phases. However, SHELXC outputs also give some useful statistics about the data. For example, anomalous correlation coefficient (CC) versus resolution graph helps to define high resolution cut-off for SHELXD (Figure 26).

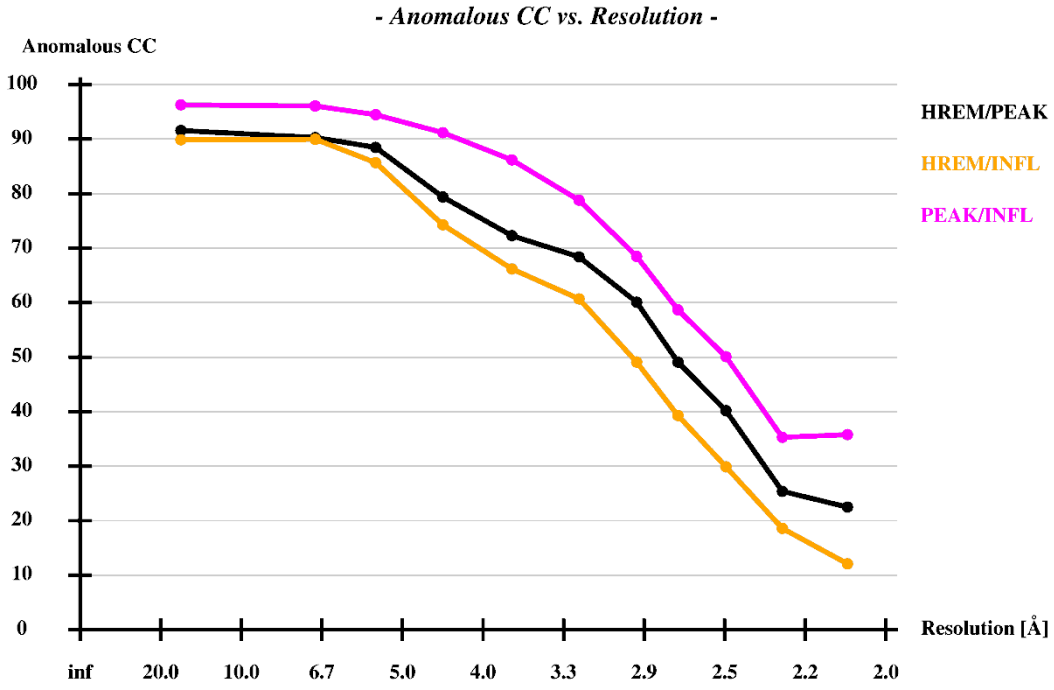


Figure 26 Anomalous CC versus Resolution

Substructure structures factors calculated by SHELXC was used to find heavy atom sites. In this analysis, high resolution cut-off was set to 2.5Å resolution and 24 of Se atoms were expected to find in the asymmetric unit. 76 of 100 trials were resulted in 45% CCall. Site occupancy for 21 of peaks was found higher than 0.3. Peaks less than 0.2 were accepted as noise (Figure 27).

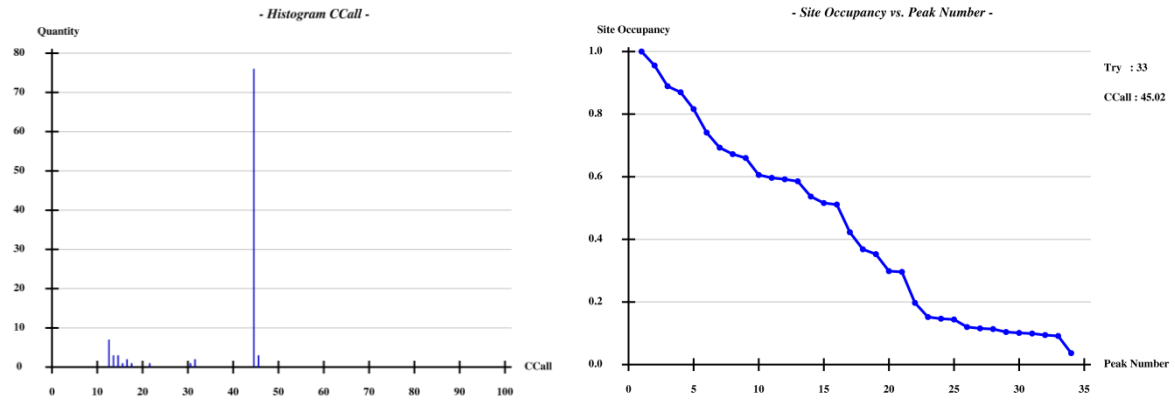


Figure 27 SHELXD outputs

Native phases and the corresponding weights (figure of merit) were estimated by SHELXE. After phasing, electron density map and density modification was calculated (Figure 28).

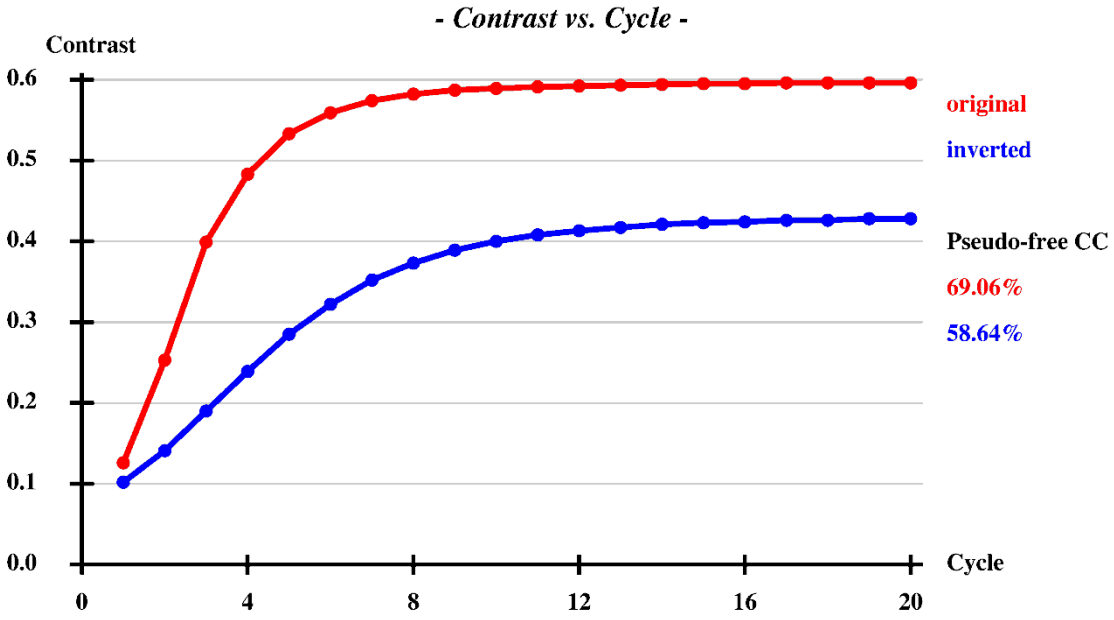


Figure 28 Variation of the contrast during SHELXE density modification starting from substructure.

Electron density map were inspected in COOT for significant density of any secondary structural element (eg. α-helices and β-strand). The calculated phases were used in ARP/wARP for automatic model building (Figure 29).

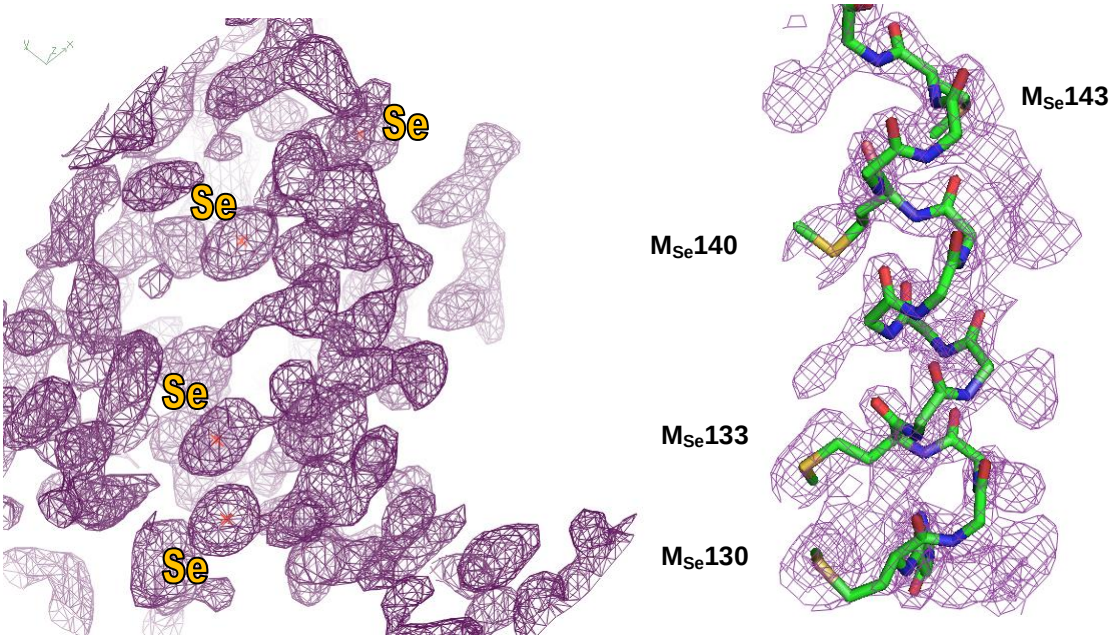


Figure 29 2Fo-Fc map (Contour level 1.5 σ) from SHELXE heavy atom sites from SHELXD. Electron density for a helix compared to model built map after ARP/wARP.

Best result for phasing and model building of Se-Met derivative structure was obtained by using automated structure solution (AutoSol) in the PHENIX software suite was used (Adams *et al.*, 2002).

The AutoSol Wizard uses several programs for complete solutions. These are HYSS to find heavy-atom sites, SOLVE to calculate phases from those sites, RESOLVE for density modification and initial model building, phenix.refine for refinement. 21 heavy-atom sites were found by HYSS (Table 7).

Table 7 Table Statistical results for best solution which is found by HYSS.

Score type	CC	RFACTOR	SKEW	FOM	NCS_OVERLAP
Raw scores	0.58	0.41	0.27	0.55	0.79
Z-scores	29.55	71.81	34.07	11.00	39.34

Best solution from HYSS was used to build preliminary model. R and R-free of the model were calculated as 0.29 and 0.32, respectively. Map model CC was calculated as 0.72.

To obtain native structure of NodS-SAH complex, the derivative model structure was used as a probe for model building. Automatic model building via ARP/wARP (Perrakis *et al.*, 1999) was performed by using the native dataset and phases from derivative model. The

last model (Figure 30) was used as a starting model to complete the structure manually by using COOT (Emsley *et al.*, 2004).

Refinement of the model was carried out with REFMAC5 (Murshudov *et al.*, 1997). The complete model was used to calculate TLS Motion Determination. For that purpose, TLSMD server was used to define number of TLS segments (Painter *et al.*, 2006). The obtained data was used in REFMAC5 for TLS and restrained refinement (Table 8). Water molecules were located by using ARP/wARP Solvent within CCP4i GUI. The located water molecules were manually reviewed and edited by using COOT.

Table 8 Refinement statistics for native NodS-SAH complex

	NodS+SAH
Program used	REFMAC5
Resolution limits (Å)	29.70 - 1.85
Number of reflections	70528
Number of reflections in test set	1165
Number of atoms	7103
Water molecules	847
R/R_{free}	0.189/ 0.237
Overall B value (Å^2)	39.8
RMSD bond length (Å)	0.018
RMSD bond angle (°)	1.71
Ramachandran φ/ψ angles (%)	
Most favored	92.9
Additionally allowed	6.5
Generously allowed	0.6
PDB Code	3OFK

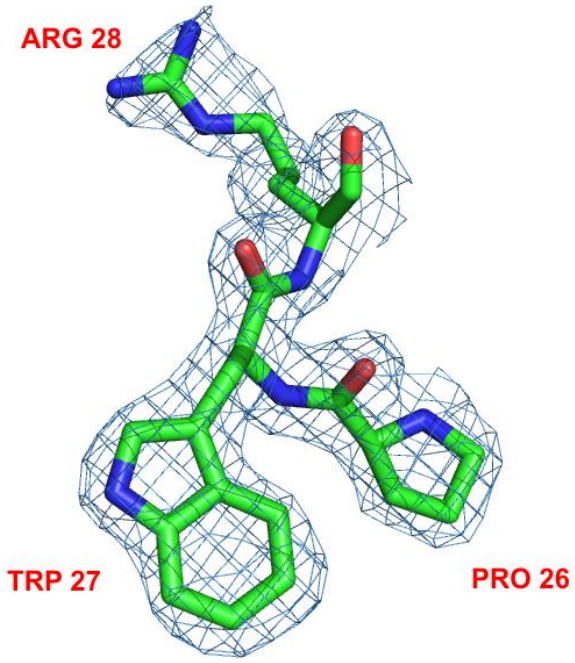


Figure 30 Part of NodS-SAH complex - 2Fo-Fc Map, Contour level 1.5 σ

Ligand free NodS

Because any derivative of ligand free NodS crystals was not able to obtain, phasing problem for the structure was solved by molecular replacement method. For that purpose, SAH complex structure of NodS was used as a search model. Structure of the model was edited to remove high B factor parts such as loop, N terminal, C terminal and unstable first helix in the structure. The edited model was used as an ensemble in AutoMR wizard in PHENIX suite. AutoMR was performed automated maximum likelihood molecular replacement with PHASER (Table 9).

Table 9 Molecular replacement solution by AutoMR

Space group of the solution	P $4_3$22
Log likelihood gain for the solution	685
Solution annotation (history)	
The rotation function Z-score (RFZ)	12.7
The translation function Z-score (TFZ)	28.4

When the solution was found, the automatic model building was carried out by AutoBuild in PHENIX suite (Table 10). Final model was used as a starting model to complete the structure by using COOT (Figure 31).

Table 10 Best model statistics for model building

R / Rfree	0.24 / 0.31
Residues rebuilt	148
Model-map CC	0.81

Refinement of the model was carried out with REFMAC5 (Murshudov *et al.*, 1997). The complete model was used to calculate TLS Motion Determination. For that purpose, TLSMD server was used to define number of TLS segments (Painter *et al.*, 2006). The obtained data was used in REFMAC5 for TLS and restrained refinement (Table 11). Water molecules were located by using ARP/wARP Solvent within CCP4i GUI. The located water molecules were manually reviewed and edited by using COOT.

Table 11 Refinement statistics for ligand free NodS

	NodS
Program used	REFMAC5
Resolution limits (Å)	20.06- 2.43
Number of reflections	5300
Number of reflections in test set	768
Number of atoms	1344
Water molecules	18
R/R_{free}	0.203/ 0.276
Overall B value ($Å^2$)	52.1
RMSD bond length (Å)	0.018
RMSD bond angle (°)	1.84
Ramachandran φ/ψ angles (%)	
Most favored	88.8
Additionally allowed	10.6
Generously allowed	0.6
PDB codes	3OFJ

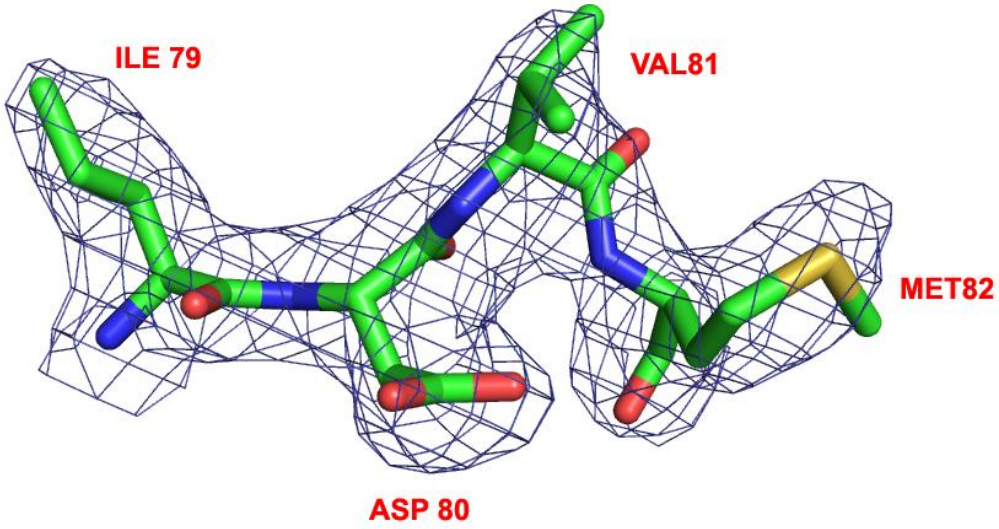


Figure 31 Part of ligand free NodS - 2Fo-Fc Map, Contour level 1.0 σ

Model Quality

Although one protein molecule is found in the asymmetric unit of ligand-free NodS, there are four molecules in the asymmetric unit of the NodS-SAH complex and each protein molecule has a bound ligand at full occupancy. However, no meaningful non-crystallographic symmetry (NCS) can be found in the asymmetric unit of the NodS-SAH complex.

Some N- and C-terminal residues do not have visible electron density in both structures. In addition, the first helix (αA) and the loop linking αA and αB of the ligand-free structure could not be modeled due to their placement in a highly disordered area within the crystal. A similar problem was encountered for one of the four molecules (chain D) in the asymmetric unit of the SAH-complex structure. In general, however, the NodS polypeptide model in the SAH-complex structure is much more complete (93.1% SAH-complex form *vs* 85.2% ligand-free form) (Table 12).

Table 12 Number of the amino acids in the crystal structures of NodS.

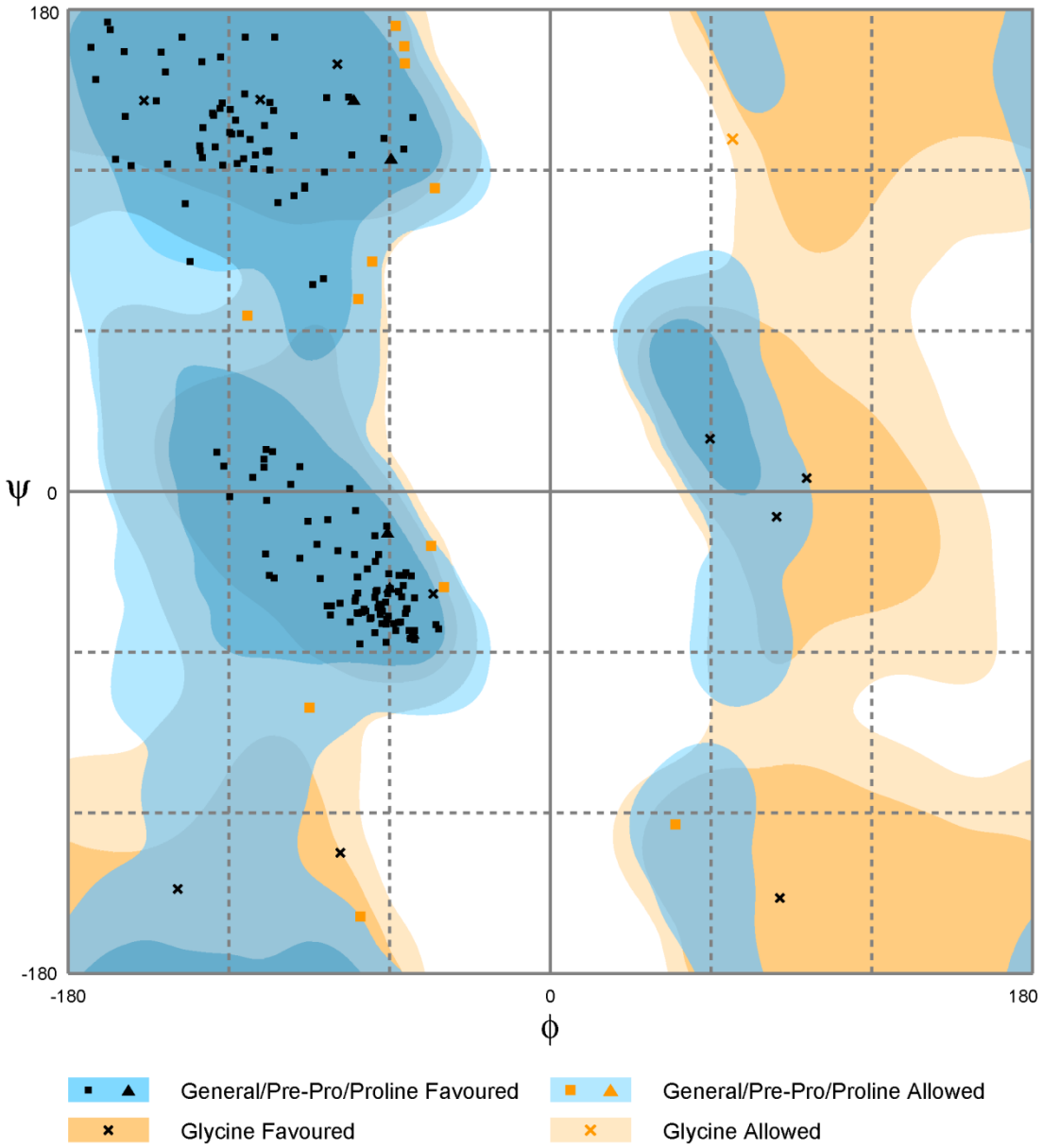
NodS+SAH		
Chain	Range	Number of amino acids
A	-4 - 199	204
B	4 - 199	196
C	4 - 205	202
D	23 - 203	181
NodS		
Chain	Range	Number of amino acids
A	22 - 199	178

The ligand-free and SAH-complex structures of NodS were validated using PROCHECK (Laskowski *et al.*, 1993). Both structures have good stereochemical parameters and low RMS (root-mean-square) deviation from ideal bond lengths and angles (Table 13). Most of the non-glycine residues are located within the most favored or additionally allowed

regions of the Ramachandran plot. The plot for ligand free structure shows 92.6% of residues in favored region, 7.4% of residues in allowed region, and there is no residue in outlier regions (Figure 32). For SAH-complex structure, it shows 97.2% of residues in favored region, 2.8% of residues in allowed region, and there is no residue in outlier region (Figure 33).

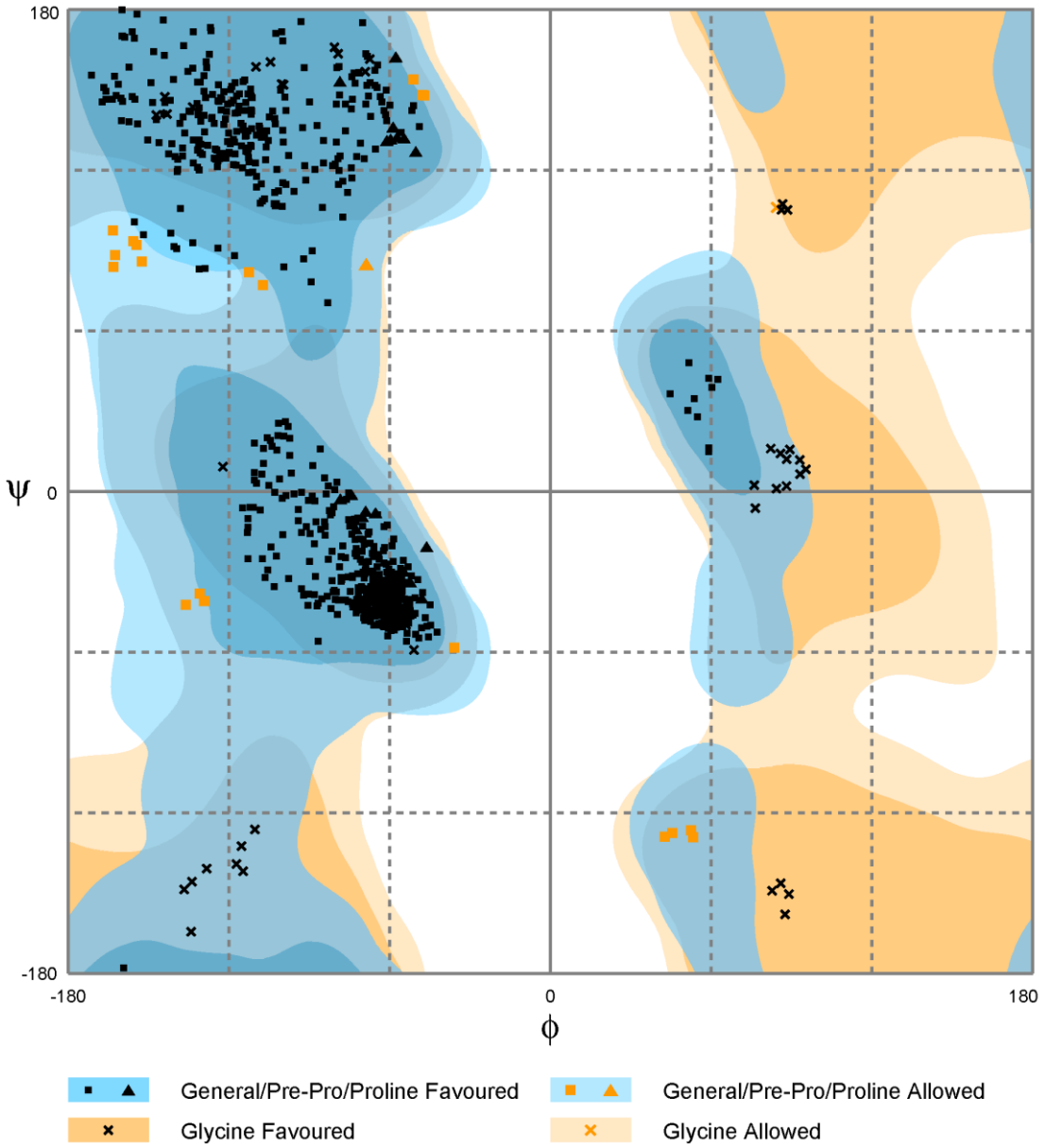
Table 13 RMS deviations of ligand free NodS and NodS-SAH complex

RMS deviation from ideal values	**NodS**	**NodS+SAH**
Bond lengths [Å]	0.018	0.018
Bond angles [°]	1.843	1.706
Torsion angles [°]	6.905	6.185
Chiral-center restraints [$Å^3$]	0.114	0.108



Number of residues in favoured region (~98.0% expected) : 163 (92.6%)
Number of residues in allowed region (~2.0% expected) : 13 (7.4%)
Number of residues in outlier region : 0 (0.0%)

Figure 32 Ramachandran plot for ligand free NodS



Number of residues in favoured region (~98.0% expected)	:	753 (97.2%)
Number of residues in allowed region (~2.0% expected)	:	22 (2.8%)
Number of residues in outlier region	:	0 (0.0%)

Figure 33 Ramachandran plot for NodS-SAH complex

Overall structure of NodS

The native sequence of NodS from *Bradyrhizobium japonicum* WM9 consists of 209 amino acid residues with a molecular weight of 23.4 kDa. However, recombinant NodS includes an additional heptapeptide (GIDPFTM-) at the N-terminus, introduced as a cloning artifact. During protein biosynthesis in *B. japonicum* WM9, the first codon (gtg) of *nodS* is translated as methionine rather than valine. Since the present protein has Val in the first position, it is considered as a single-site variant of the natural polypeptide.

The topology of the NodS molecule is the same in both structures and can be described as in Figure 34, where the secondary-structure elements are defined by DSSP (Kabsch *et al.*, 1983). Starting from the N-terminus, the consecutive β-chains are numbered by Arabic numerals and the consecutive α-helices by capital letters. There are also three short (single turn) 3_{10} helices, left without separate annotation (Figure 34). The overall fold of NodS consists of a seven-stranded open β-sheet, flanked by three α-helices on each side. The seventh α-helix, αA, which is the first secondary structure element in the NodS sequence and which gets ordered only in the SAH complex, is effectively located above the β-sheet structure rather than on its side. A C-terminal β-hairpin is formed between strands β6 and β7, instead of the more typical β-α-β crossover that generally characterizes the remaining β-β connections (Figure 35). According to its folding pattern, NodS is, therefore, classified as Class I methyltransferase (Schubert *et al.*, 2003).

Both structures have irregular regions called β-bulges which give the β-sheet a visible curvature. Three types of β-bulges are found in both structures. The first one, parallel bent type, in which the normal pattern of hydrogen bonding is disrupted with one extra residue on both strands, is found at S46 on β1 and K67 on β2. The second β-bulge, parallel special type, is found at D109 on β4, which is opposed by A138, P139, G140, and G141 on β5. The third β-bulge has antiparallel classic type and is seen between R194 of β7 and the opposing V176 and E177 on β6 (Chan *et al.*, 1993).

In the tertiary structure of NodS, the first six strands of the β-sheet are parallel to each other. Strand β7 is antiparallel to the remaining strands, and is inserted into the sheet between strands β5 and β6. The topology of the β-sheet (Richardson,1977) is therefore +1x,+1x,-3x,-1x,-2x,+1 (Figure 34, 36). The β-sheet contains mostly hydrophobic residues which interact with the hydrophobic faces of the amphipathic helices. Although the protein has an open-sheet α/β/α folding pattern resembling the Rossmann motif, it cannot be classified as a classic

Rossmann fold (Rao *et al.*, 1973) because of the presence of the antiparallel β-strand inserted into the β-sheet structure.

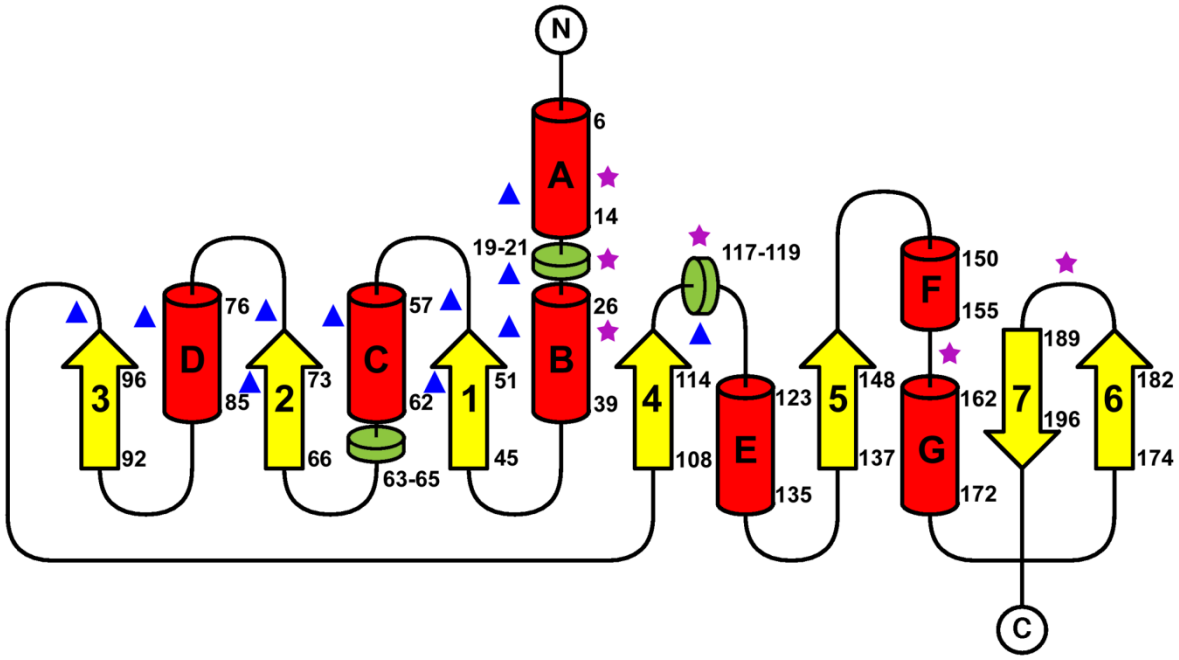


Figure 34 Topology diagram of NodS. The α-helices (capital letters, red) and β-strands (Arabic numerals, yellow) are numbered consecutively. The inclusive residue numbers correspond to DSSP **(Kabsch *et al.*, 1983)** secondary structure assignments. Green disks represent 3_{10}-helices. The blue triangles mark the positions which form the SAH binding cavity. The purple stars mark the positions in the groove on the protein surface occupied by the non-reducing end of the docked COS molecule.

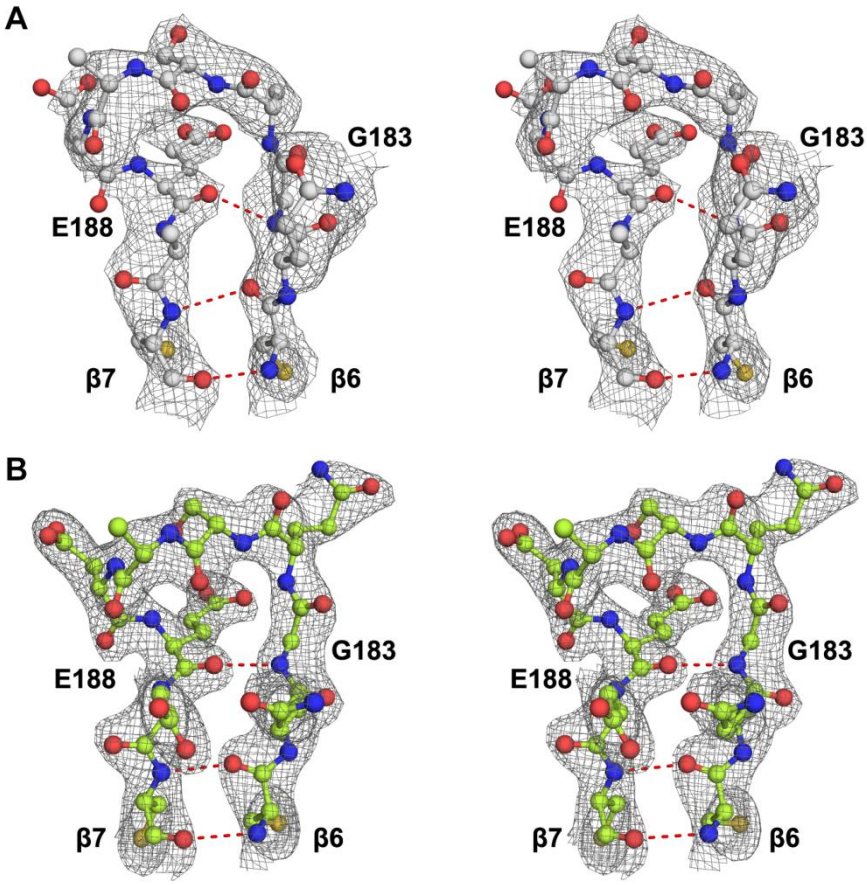


Figure 35 Stereoview of the β6-β7 β-hairpin motif in the structure of NodS; (A) ligand-free form, (B) SAH complex. The 2Fo-Fc electron density maps are shown at the 1.0σ contour level. In this and all other figures, unless stated otherwise, the ligand free NodS model is indicated by light gray carbon atoms while that from the SAH complex by light green carbon atoms.

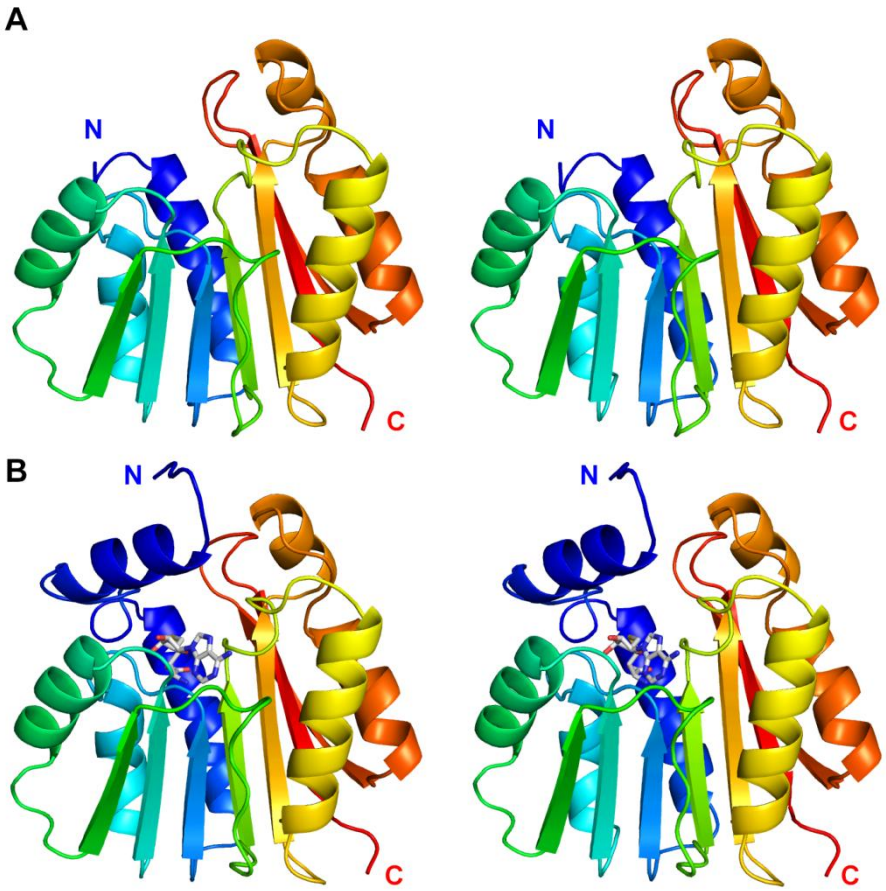


Figure 36 Stereoview of the structure of NodS in rainbow color representation from blue (N terminus) to red (C terminus); (A) ligand-free form, (B) SAH complex, with the SAH molecule in stick representation.

Crystal packing and intermolecular interactions

The four identical molecules in the asymmetric unit of the NodS-SAH complex (labeled A-D) show polar and non-polar intermolecular interactions (Table 14). Despite all of these interactions, NodS does not form oligomers in solution, as gel filtration during the purification protocol clearly indicated only monomeric protein (Cakici *et al.*, 2008). It is interesting to note, however, that very similar packing interactions are found in the crystal lattice of the ligand-free NodS structure, even though the symmetry and unit cell contents are completely different.

Table 14 Statistics of intermolecular interactions in the asymmetric unit of the NodS-SAH complex.

Chains	No. of interface residues	Interface area ($Å^2$)	No. of hydrogen bonds	No. of non-bonded contacts
A-B	11:11	656:649	2	33
A-D	4:4	152:155	4	28
B-C	19:19	919:910	7	83
C-D	10:9	471:464	1	43

Binding of SAH molecule

The SAH molecule consists of the adenine, D-ribose and L-homocysteine moieties (Figure 37A). During the enzymatic reaction, NodS converts S-adenosyl-L-methionine (SAM) to SAH, which is a byproduct of the catalysis. In this study, SAH was used as a ligand during the crystallization experiments. The interactions of the adenine nucleobase, the sugar ring, and the main chain of the amino acid residue should be identical for both the substrate and product complexes owing to the high degree of their chemical similarity.

The conformation of the four SAH molecules in the crystal structure of the complex is very similar. The ribose ring has an envelope conformation with C2'-*endo* pucker, and the orientation of the base ring system around the glycosidic bond is *anti*, as defined by the O4'-C1'-N9-C8 torsion angle (~80°). The Sδ-C5'-C4'-O4' torsion angle (~180°), which defines whether the nucleotide is extended or folded, shows that the molecules are extended, with the Sδ substituent *trans* relative to the ribose ring oxygen atom. Selected torsion angles of the SAH molecules are listed in Table 15.

Table 15 Selected torsion angles (°) of SAH molecules in the NodS-SAH complex.

	A	**B**	**C**	**D**
OXT-C-Cα-N	166	176	173	176
N-Cα-Cβ-Cγ	66	72	64	51
Cα-Cβ-Cγ-Sδ	172	179	-177	179
Cβ-Cγ-Sδ-C5'	105	101	104	108
Cγ-Sδ-C5'-C4'	-92	-98	-98	-95
Sδ-C5'-C4'-O4'	169	166	167	173
C4'-C3'-C2'-C1'	-33	-36	-28	-30
C3'-C2'-C1'-O4'	40	42	36	35
C2'-C1'-O4'-C4'	-31	-30	-30	-28
C1'-O4'-C4'-C3'	9	6	10	9
O4'-C4'-C3'-C2'	16	20	13	14
O4'-C1'-N9-C8	79	78	79	84

The SAH ligand is located in the N-terminal half of the NodS molecule. Its binding site is formed in a depression on the surface of the NodS molecule, and the bound ligand is covered from the top by fragments of the αA plus αA-αB element. In consequence, the SAH (or, by inference, SAM) ligand is enclosed in a cavity, where it is precisely oriented for the catalytic reaction. There is only a small opening leading to the cavity, through which the buried ligand is presenting its sulfonium group to the catalytic center (Figure 38A). It is also obvious that docking of the SAM/SAH ligand in its binding site is accompanied by a significant structural response of the NodS protein, namely by ordering of the αA plus αA-αB N-terminal fragment, which becomes part of the SAM binding site.

Although the SAM/SAH ligand molecule is docked in a predominantly hydrophobic cavity, hydrogen bonds are mainly responsible for its orientation in the binding site (Figure 37B). In general, the interior of the cavity is lined with side chains contributed by the following secondary-structure elements: αA (L10, L14), αA-αB loop (W20, L22), αB (R31, H32, L35), β1 (E50, I51), β1-αC loop (G52, C53, A54), αC (F58), β2 (I72, D73), β2-αD loop (V74, M75), αD (A78), β3-β4 loop (T97, D98, I99), β4 (A114), and β4-αE loop (E115, V116, Y119, L120).

The α-carboxylic group of the amino acid moiety of the SAH molecule is coordinated by the side chains of the basic amino acid residues R31 (O···Nη1, OXT···Nη2) and H32 (OXT···Nε2) in helix αB. The amino group of the homocysteine fragment of the SAH molecule is anchored by the main-chain carbonyl groups of G52 and A114 from loop β1-αC and strand β4, respectively. The hydroxyl groups of the ribose fragment of the ligand form a pair of forked hydrogen bonds with both oxygen atoms of the side chain carboxylic group of D73 (O2'···Oδ1, O3'···Oδ2) at the tip of the β2 strand, plus there is a hydrogen bond donated

by the main-chain nitrogen atom of A54 (N···O3') in loop β1-αC. Finally, the adenine ring of the ligand is oriented by the main-chain nitrogen of I99 (N···N1) and the acidic side chain of D98 (N6···Oδ1), both from the β3-β4 loop (Table 16). The ligand molecule also forms hydrogen bonds with water molecules in the binding pocket. The α-amino group of the amino acid part of the SAH molecule coordinates two water molecules while the α-carboxylate group coordinates one water molecule. One water molecule is coordinated by a pair of hydrogen bonds with two nitrogen atoms (N6 and N7) of the adenine ring.

Table 16 Polar interactions of SAH molecules in their binding cavities, with corresponding donor···acceptor distances (Å) in parentheses.

Atom	Chain A	Chain B	Chain C	Chain D
O	Nη1 R31 (2.80)	Nη1 R31 (2.80)	Nη1 R31 (2.89)	Nη1 R31 (2.91)
OXT	Nη2 R31 (3.26)	Nη2 R31 (3.17)	Nη2 R31 (3.26)	Nη2 R31 (3.10)
	Nε2 H32 (2.82)	Nε2 H32 (2.87)	Nε2 H32 (2.81)	Nε2 H32 (2.92)
	O Wat39 (2.66)	O Wat5 (2.92)	O Wat9 (2.84)	O Wat112 (2.68)
N	O A114 (2.94)	O A114 (2.80)	O A114 (2.74)	O A114 (2.86)
	O G52 (2.74)	O G52 (2.68)	O G52 (2.94)	O G52 (2.90)
	O Wat44 (2.73)	O Wat14 (2.76)	O Wat28 (3.00)	O Wat17 (2.85)
	O Wat120 (2.78)	O Wat2 (2.65)	O Wat3 (2.89)	O Wat70 (2.79)
O3'	Oδ2 D73 (2.50)	Oδ2 D73 (2.57)	Oδ2 D73 (2.61)	Oδ2 D73 (2.58)
	N A54 (2.89)	N A54 (2.95)	N A54 (2.98)	N A54 (3.08)
				O Wat298 (2.82)
O2'	Oδ1 D73 (2.73)	Oδ1 D73 (2.62)	Oδ1 D73 (2.54)	Oδ1 D73 (2.74)
				O Wat298 (3.40)
N1	N I99 (3.03)	N I99 (3.09)	N I99 (3.20)	N I99 (3.00)
N3	N V74 (3.28)	N V74 (3.21)	N V74 (3.31)	N V74 (3.14)
N6	Oδ1 D98 (2.96)	Oδ1 D98 (2.83)	Oδ1 D98 (2.95)	Oδ1 D98 (2.85)
	O Wat268 (2.87)	O Wat255 (3.41)	O Wat335 (3.39)	O Wat770 (3.04)
		O Wat220 (3.24)	O Wat132 (3.03)	
N7	O Wat268 (2.94)	O Wat220 (3.08)	O Wat132 (2.91)	O Wat770 (3.54)

There are also CH···π interactions between two aliphatic residues (V74 in loop β2-αD and V116 in loop β4-αE) and the adenine ring of the SAH molecule (V74, Cβ to pyrimidine ring of adenine ~3.7 Å; Cγ1 to imidazole ring of adenine ~3.6 Å; V116, Cγ2 to imidazole ring of adenine ~3.5 Å). Although SAH is not capable of forming cation···π interactions, the six-membered ring of W20 is facing the sulfonium group at ~3.8 Å distance. This contact could become a plausible cation···π interaction when the SAM donor (with the $–(CH_3)S^+–$ group) binds in this site (Figure 37C).

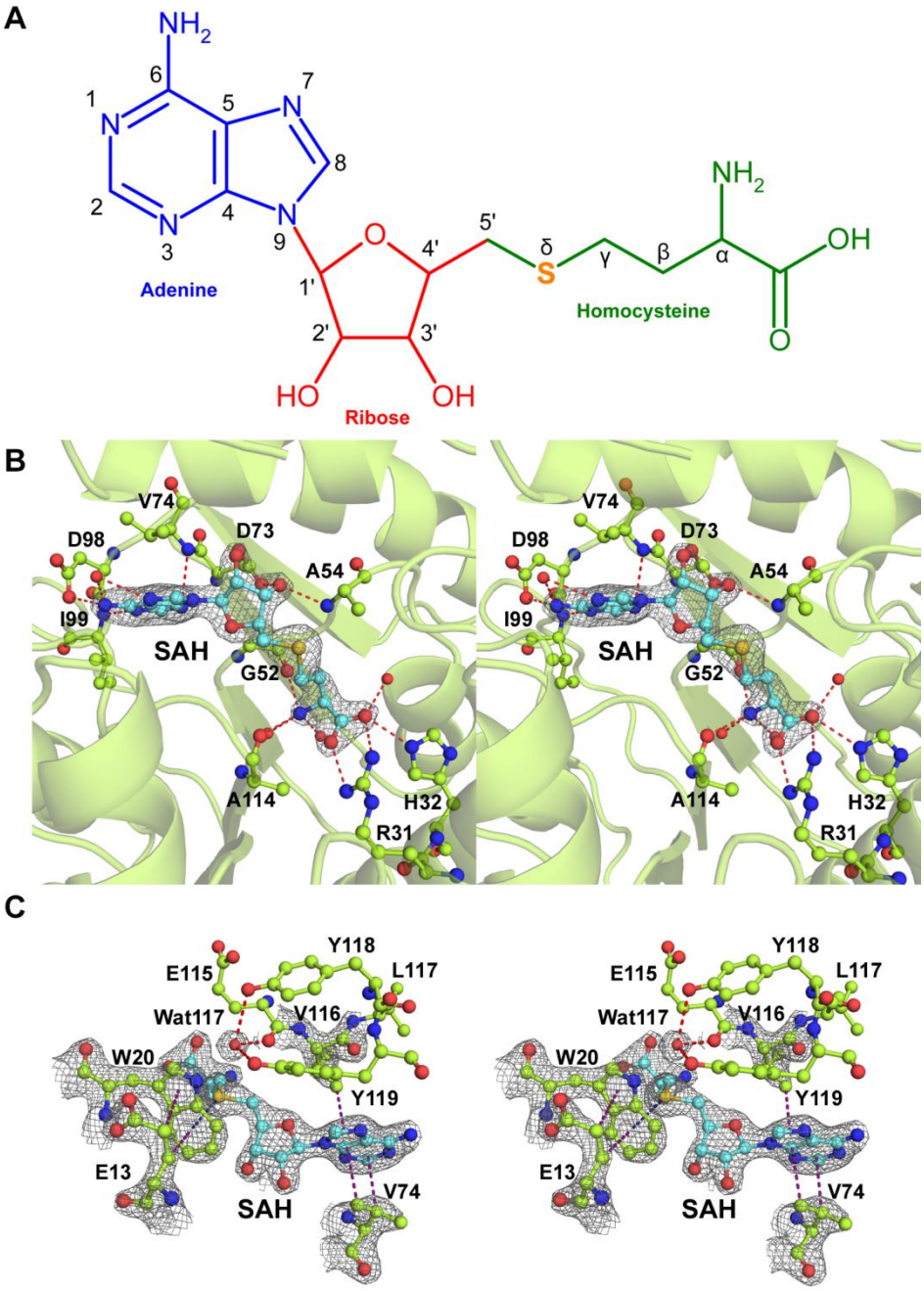


Figure 37 Binding of SAH molecule in NodS structure. (A) A schematic diagram of the SAH molecule with atom numbering. (B) Stereoview of the binding site of the SAH molecule, shown in 2Fo-Fc electron density map contoured at the 1.5σ level. (C) Stereoview of a possible cation···π interaction (blue dashed line) between the positively charged sulfur atom of SAM and the centroid of the six-membered ring of W20. Such an interaction could be formed if the SAH ligand in the NodS-SAH complex were replaced by a SAM molecule. A well ordered water molecule (with hydrogen bonds marked by red dashed lines) is found close to the SAH sulfur atom at a distance of about 3.4 Å. Parts of the structure are shown in 2Fo-Fc electron density contoured at 1.5σ. CH···π interactions are shown in purple dashed lines. In this and all following figures, the SAH (or SAM) molecule in depicted with cyan carbon atoms.

Possible chitooligosaccharide acceptor binding site

Inspection of the van der Waals model of the NodS-SAH complex reveals a distinct, deep canyon running across the molecular surface in the direction demarcated by the N-terminal helix αA, which forms one of its walls. The canyon is ~22 Å long, ~10.5 Å wide and ~10.5 Å deep, and terminates at the orifice that leads to the internal cavity accommodating the SAM methyl donor (Figure 38A). As defined in Figure 34, the canyon is located near the N-terminal end of helix αB and the C-terminal end of strand β4. The walls of the groove are lined by residues from αA (S9, E13), loop αA-αB (W20), αB (F27, E28, R31), loop β4-αE (E115, Y118, Y119), loop αF-αG (W156, H158), and loop β6-β7 (S185, D187, E188) (Figure 34, 38B).

To identify a possible binding mode of the chitooligosaccharide substrate in the NodS-SAH complex, the AutoDock4 (Morris *et al.*, 1998) software was used to dock a deacetylated chitotriose (COS) molecule in the selected canyon region. Prior to the simulation, the SAH molecule was switched to SAM. The result shows that the chitotriose molecule can be docked with a very good fit and in an energetically favorable conformation. The successful docking shows that the COS substrate is locked in the binding groove mainly by polar interactions (Table 17). In particular, the acceptor amino group of the non-reducing end of the COS molecule is anchored by the Oε1 atom of E115 and the Oη atom of Y118 (Figure 38C) in a position where it has an excellent fit near the entrance to the SAM cavity. Therefore, according to this model, the methyl donor and acceptor would be juxtaposed in a nearly ideal arrangement for the catalytic reaction. The position and orientation of the docked COS molecule also suggests the possibility of CH···π interactions between the C-H donors of the ligand and the aromatic side chain of W156.

Table 17 Possible polar interactions of the COS molecule in the binding groove of NodS, predicted by AutoDock4 software **(Morris *et al.*, 1998)**. The calculated hydrogen bond distances are in Å.

Atom	$GlcNH_2$	GlcNAc(1)	GlcNAc(2)
N2	Oη Y118 (2.79) Oε1 E115 (3.53)	Oε2 E13 (2.46)	
O2			O W156 (3.43)
O3	Oε1 E28 (3.36) Oε2 E28 (2.90)		
O4	O D187 (2.73)		
O5			Oγ S9 (2.73)
O6	Oδ1 D187 (3.31)	Nε2 H158 (3.46) Oη Y119 (2.95)	

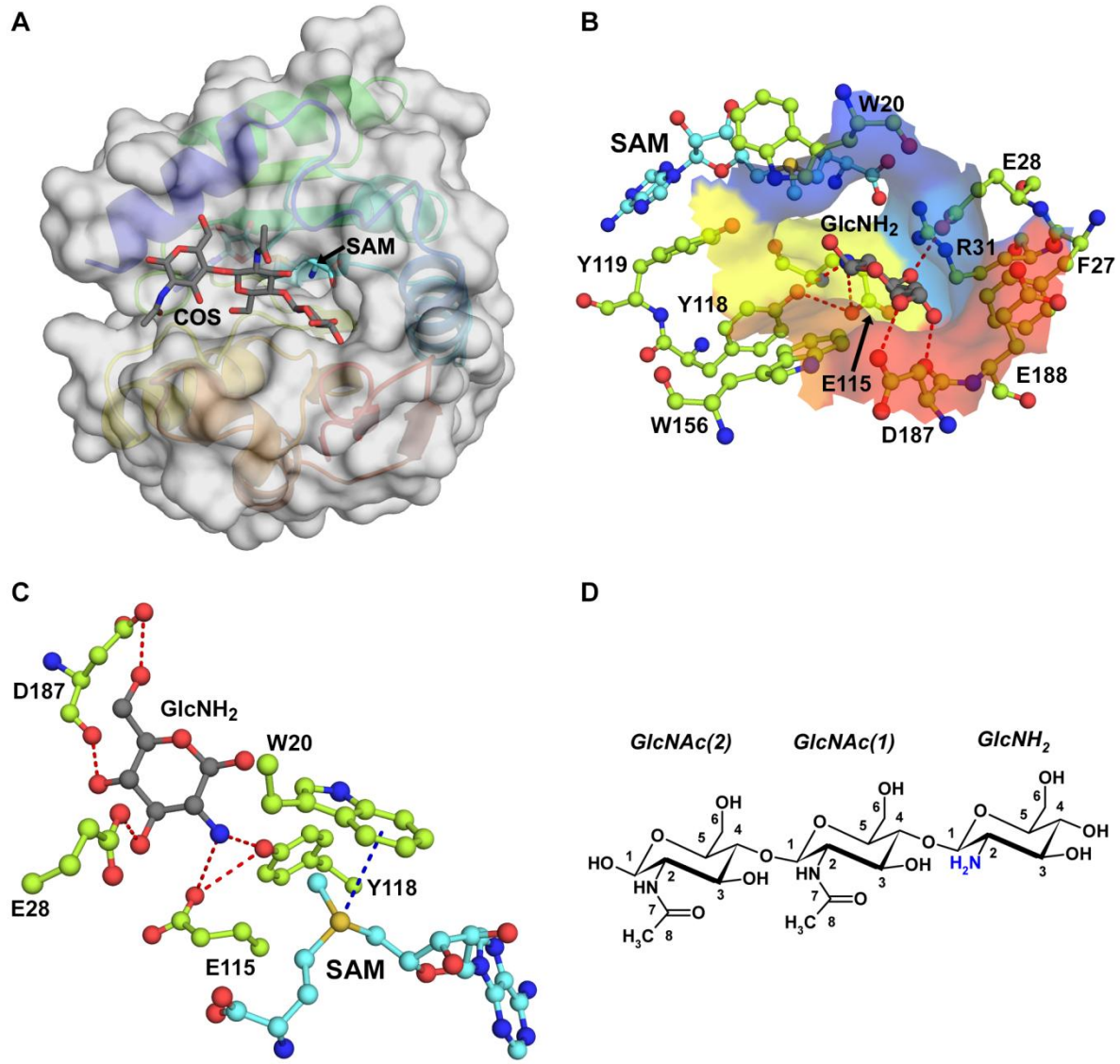


Figure 38 Putative binding site for the chitooligosaccharide substrate (dark gray) in a groove on the surface of NodS. (A) A trisaccharide molecule (COS) docked in the binding groove of NodS by the AutoDock4 software **(Morris *et al.*, 1998)**. The non-reducing end of the substrate is located over the entrance to the SAM binding cavity, here marked in cyan by the SAM molecule, which has been modeled by attaching a methyl group to the sulfur atom of the SAH template. Under the transparent surface, the NodS structure is shown in rainbow representation from blue (N terminus) to red (C terminus) (as in Fig. 36). (B) The amino acid residues that surround the non-reducing end of the COS substrate form the lining of the orifice that leads to the SAM binding cavity. The amino acid residues and their corresponding surfaces are shown in rainbow colors corresponding to their position in the NodS sequence (as in Fig. 36). (C) Polar interactions of the $GlcNH_2$ unit of the COS substrate, shown to emphasize the juxtaposition of the S^+–CH_3 donor of SAM and the NH_2 acceptor of the $GlcNH_2$ unit. (D) A schematic diagram of the trisaccharide COS molecule with atom numbering.

Comparison of ligand-free and SAH complex structures of NodS

The N-terminal α-helix (αA) and the following αA-αB loop of the NodS molecule cannot be compared in detail since they could not be modeled in the ligand-free structure because of disorder. Obviously, ordering of the αA plus αA-αB element is connected with docking and burial of the SAM/SAH ligand, as residues from this element (L10, L14, W20, L22) participate in the formation of the binding cavity in the NodS-SAH complex structure.

Table 18 lists the RMSD values for the superpositions of the Cα atoms of the four monomers of the NodS-SAH complex and of the ligand-free protein. The deviations between the four monomers in the asymmetric unit of the NodS-SAH complex are rather small, not exceeding 0.5 Å. However, the deviations of the ligand-free model of NodS are about twice as big, between 0.8-0.9 Å. Those elevated overall discrepancies are due to pronounced changes visible in several regions. The most conspicuous areas include the loop between strand β1 and helix αC (A54, A55), helix αD (R77-I79), strand β3 (S93, W94), loop β3-β4 (D98-L100, S103-E106), loop β4-αE (E115-L120), helix αF (T152-V159), and loop β6-β7 (Q182-D189) (Figure 39). Most of the residues in these areas form the walls of the SAM/SAH binding cavity and of the predicted binding groove for the chitooligosaccharide substrate.

Table 18 Comparisons of NodS monomers, characterized by RMSD (Å) between corresponding Cα atoms.

Protein	NodS-SAH (B)	NodS-SAH (C)	NodS-SAH (D)	NodS
NodS-SAH (A)	0.314	0.418	0.459	0.813
NodS-SAH (B)	-	0.476	0.459	0.923
NodS-SAH (C)	-	-	0.502	0.930
NodS-SAH (D)	-	-	-	0.782

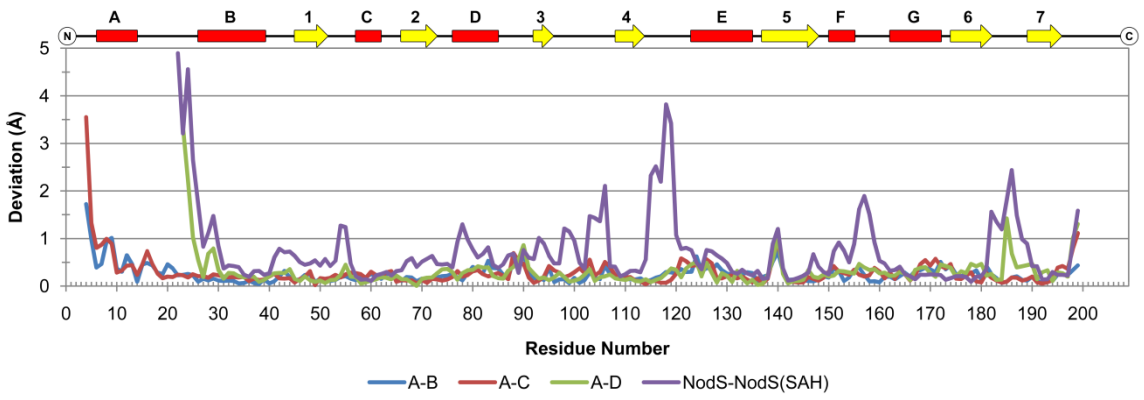


Figure 39 Comparison of Cα pairs of the NodS structures. Chain A in the asymmetric unit of the NodS-SAH complex was superposed on the other monomers (B, C, D) in the asymmetric unit and on ligand-free NodS. The pictogram above the plot shows the secondary structure elements of NodS, according to Figure 34.

Besides the major conformational rearrangements of the αA plus αA-αB element, also the loop between strand β4 and helix αE shows a significant change in its conformation upon SAH binding (Figure 40A). In particular, the Cα atoms of two consecutive tyrosine residues (Y118, Y119) in the loop change their positions by about 3.8 and 3.4 Å, respectively. The movement of the corresponding Oη atoms of Y118 and Y119 is about 10.3 and 10.7 Å, respectively (Figure 40B, C). In their complex-specific conformation, the tandem tyrosine residues Y118-Y119 have the hydroxyl groups of their side chains by hydrogen bonds mediated through a very well ordered water molecule recruited to the active site, which is also in a hydrogen-bonding distance of the sulfur atom of the SAH molecule (Figure 37C). The conformational changes in this region lead also to hydrogen bonding between the Nε1 atom of W20 in the αA-αB loop and the Oη atom of Y119, and between the Oε1 atom of E115 and Oη of Y118 in the β4-αE loop. In addition, a positional change of V116 leads to a CH···π interaction between the Cγ2 atom of V116 and the adenine imidazole ring of the SAH molecule. Although the ligand-free structure of NodS has a lower resolution, a 2Fo-Fc electron density map shows a clear picture of all the side chains involved in these interactions (Figure 40C).

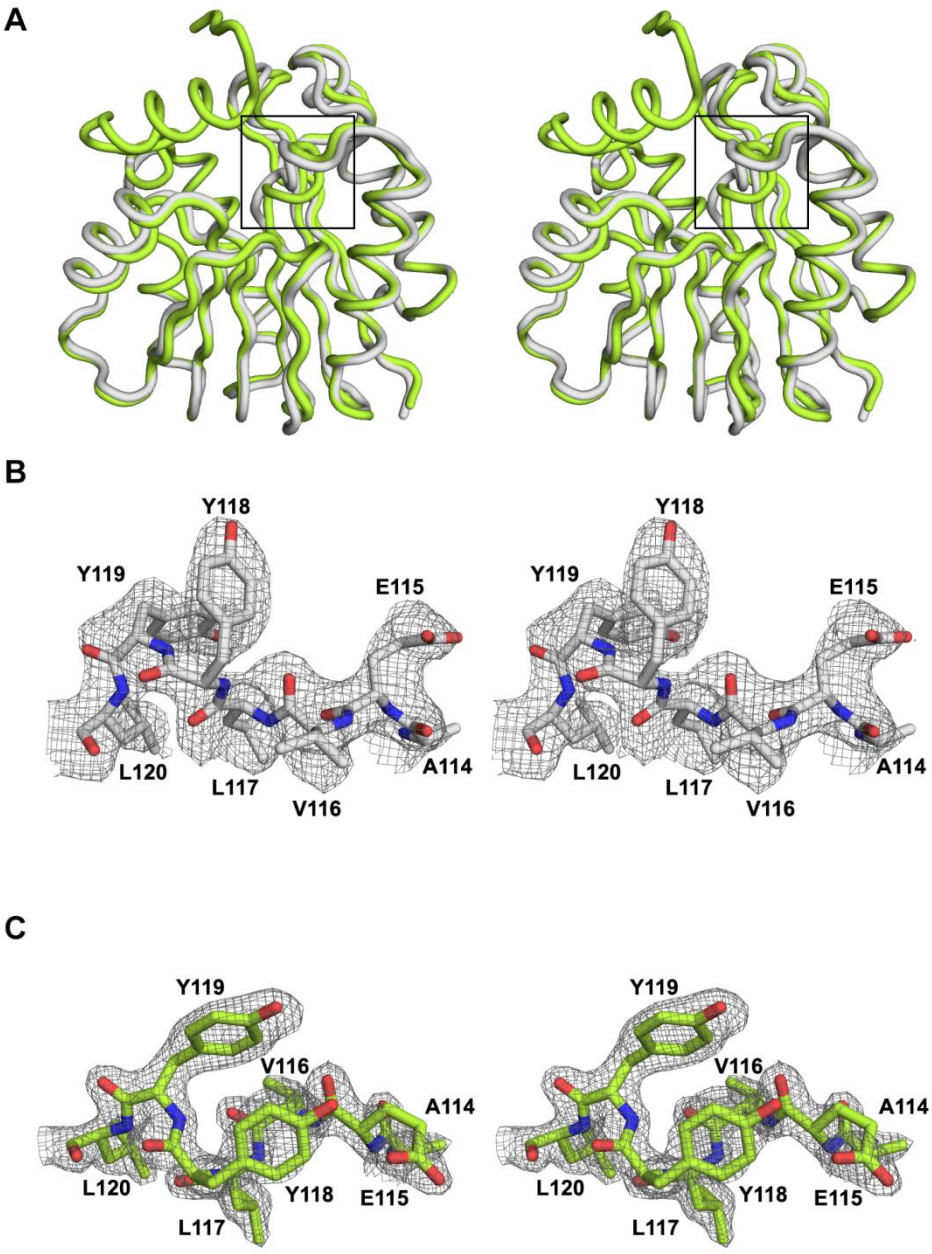


Figure 40 Stereoviews illustrating the conformational changes between ligand-free NodS (light gray) and NodS-SAH complex (light green). (A) Superposed main-chain traces of NodS in ligand-free form and in SAH complex (chain A). The framed area corresponds to loop β4-αE, shown in atomic detail in 2Fo-Fc electron density maps in panels B and C. (B) Ligand-free NodS, the map contour level is 1.0σ. (C) NodS-SAH complex, the map contour level is 1.5σ.

The ordered helix αA and loop αA-αB in the NodS-SAH complex show a number of non-polar and polar interactions. The aliphatic residues in αA (L10 and L14) and in loop αA-αB (L22) face the adenosine ring and the homocysteine moiety of the SAH molecule, respectively. There is a possible CH···π interaction between L10 and Y119 at 3.7 Å distance (Cδ1 of L10 to Y119 phenyl ring centroid). Another CH···π interaction is between M75 and Y7 at 3.4 Å distance (Cε of M75 to Y7 phenyl ring centroid). Polar interactions are formed between atoms in the αA-αB loop and atoms in segments of αB, αD, and loops β4-αE and β6-β7 (Table 19).

Table 19 Polar interactions of helix αA and loop αA-αB ("Atom") in the NodS-SAH complex, with corresponding donor···acceptor distances (Å) in parentheses.

Atom	Chain A	Chain B	Chain C	Position
O L14	Nη1 R81 (2.83)	Nη1 R81 (2.88)	Nη1 R81 (2.90)	αD
O N16	Nη1 R81 (2.54)	Nη1 R81 (2.69)	Nη1 R81 (2.60)	αD
O D17	Nη2 R81 (2.81)	Nη2 R81 (3.16)	Nη2 R81 (3.07)	αD
Nε1 W20	Oη Y119 (3.04)	Oη Y119 (3.22)	Oη Y119 (3.20)	β4-αE loop
Oδ2 D23	Nη1 R85 (2.74)	Nη1 R85 (2.89)	Nη1 R85 (2.98)	αD
Oδ1 D23	Nη2 R85 (2.86)	Nη2 R85 (2.97)	Nη2 R85 (3.01)	αD
O D23	Nη1 R29 (3.06)			αB
O D24	Nη1 R29 (3.16)			αB
Nδ2 N25	Oε2 D188 (2.93)	Oε2 D188 (2.72)	Oε2 D188 (2.66)	β6-β7 loop

Structurally homologus proteins

Two 3D alignment methods were used to search for protein models structurally homologous to NodS, using chain A of the NodS-SAH complex structure as the query. First, pairwise alignment was performed using all PDB entries and the Secondary Structure Matching (SSM) server (Krissinel *et al.*, 2004). Selected proteins from the result (Table 20), which are known to be SAM-dependent methyltransferases, were then used to carry out a multiple 3D alignment with the Chimera software (Pettersen *et al.*, 2004).

Table 20 Selected hits from SSM search of PDB entries. Q score represents the quality function of Cα alignment, the maximum being 1.00 for 100% identical structures; RMSD, Cα root-mean-square deviation (Å) between NodS-SAH and the target molecule; N_{algn}, number of matched residues or length of alignment; $\%_{seq}$, percent of sequence identity. AD-003, adrenal gland protein (unpublished); TPMT, thiopurine S-MTase **(Wu *et al.*, 2007)**; RebM, rebeccamycin sugar O-MTase **(Singh *et al.*, 2008)**; MtfA, glycopeptide N-MTase **(Shi *et al.*, 2009)**; COMT, catechol O-MTase **(Rutherford *et al.*, 2008)**; GAMT, guanidinoacetate MTase **(Komoto *et al.*, 2003)**; DesVI, N,N-dimethyltransferase **(Burgie *et al.*, 2008)**; SynOMT, CCoAOMT-like O-MTase **(Kopycki *et al.*, 2008)**; GNMT, glycine N-MTase **(Huang *et al.*, 2000)**.

Protein	Source	PDB ID	Chain	Q score	RMSD	N_{algn}	$\%_{seq}$
AD-003	*Homo sapiens*	2EX4	B	0.48	2.21	181	13
TPMT	*Homo sapiens*	2BZG	A	0.39	2.67	178	11
RebM	*Lechevalieria aerocolonigenes*	3BUS	B	0.37	2.09	162	17
MtfA	*Amycolatopsis orientalis*	3G2Q	B	0.36	2.11	158	15
COMT	*Homo sapiens*	3BWY	A	0.36	2.29	156	12
GAMT	*Rattus norvegicus*	1P1C	A	0.34	2.64	152	16
DesVI	*Streptomyces venezuelae*	3BXO	B	0.33	2.38	160	16
SynOMT	*Synechocystis sp.* PCC 6803	3CBG	A	0.32	2.58	155	16
GNMT	*Rattus norvegicus*	1D2H	D	0.30	2.38	157	18

The pairwise alignment returned homologous structures that mostly belong to the methyltranferase superfamily, but there is no match that would be really close. The highest Q score of 0.48 (with max. at 1.00) was obtained for a human adrenal gland protein in complex with SAH (2EX4, chain B, unpublished results). The structure has a β-sheet insertion between β5 and αG, instead of the α helix (αF) seen in the NodS structure. The RMSD value for this protein is 2.21 Å for the superposition of 181 Cα atoms.

Although the proteins in Table 20 have a low level of sequence identity (11-18%), the multiple 3D alignment of the selected structures shows a conserved fold of the structural core, especially with regard to the seven-stranded β-sheet and five helices, αB, αC, αD, αE and αG. Furthermore, despite the lack of overall sequence similarity, the 3D alignment revealed some highly conserved residues among the selected structures. The acidic residue D73 in the β2-αD loop, which coordinates the two hydroxyl groups of the ribose moiety of SAH/SAM is highly conserved (Figure 41). A region in the β1-αC loop, which has contacts with the homocysteine and ribose moieties of the SAH molecule, has a conserved sequence signature (E/DXGXGXG) defined as motif I among SAM-dependent methyltransferases (Cheng,2000; Martin *et al.*, 2002).

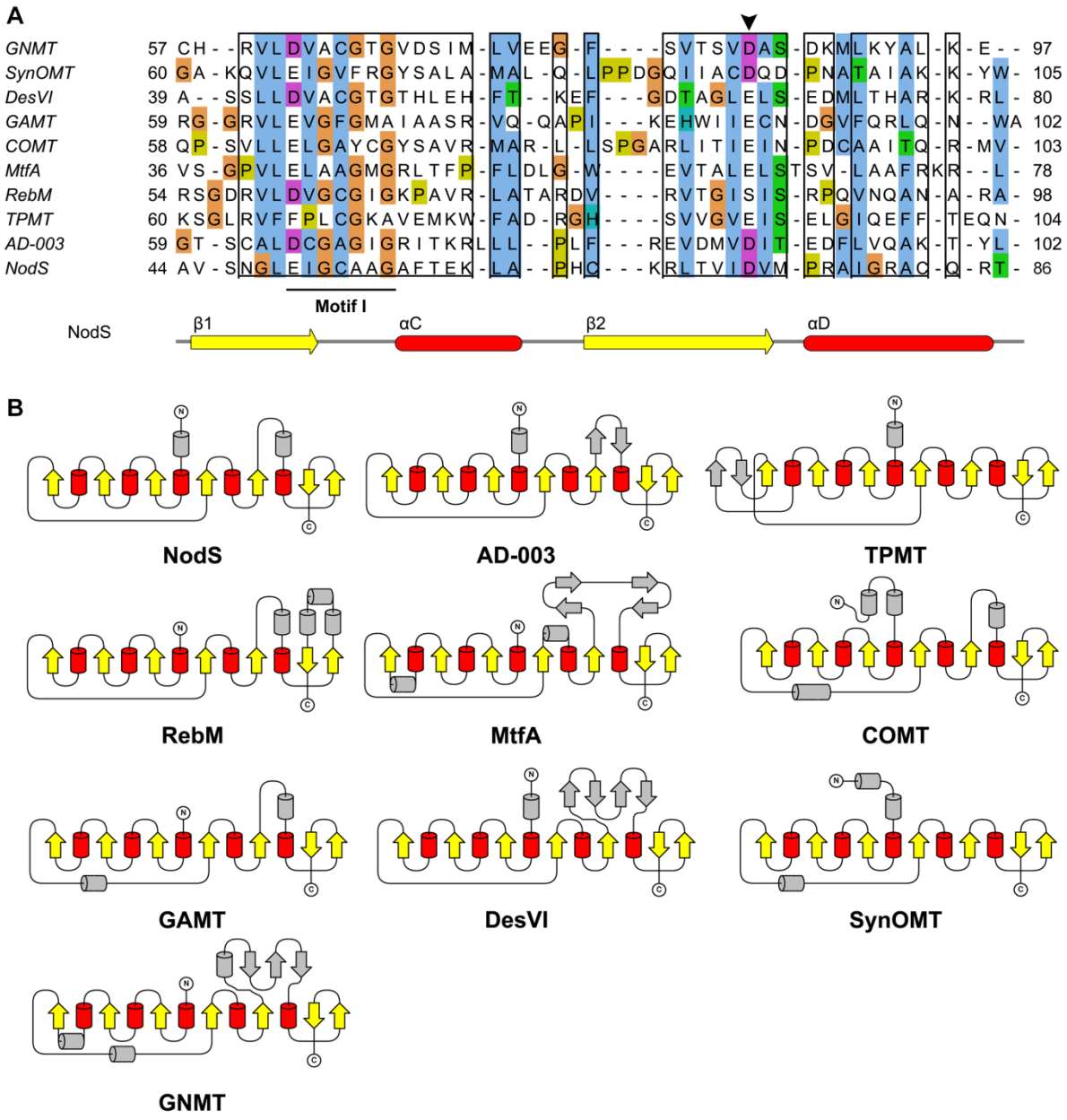


Figure 41 (A) Multiple 3D alignment of selected structures homologous to NodS. The alignment was calculated using the Chimera software **(Pettersen *et al.*, 2004)**. The figure was prepared using Jalview **(Clamp *et al.*, 2004)** and the Clustal X color scheme was used. The secondary structure elements are indicated as present in the NodS structure from *Bradyrhizobium japonicum* WM9, according to the scheme defined in Fig. 34. The arrowhead shows the highly conserved acidic residue that coordinates the ribose moiety of the SAH or SAM molecule. The black boxes indicate the superposed regions without gaps among the compared structures. Motif I is the consensus signature E/DXGXGXG, defined as the glycine-rich region between β1 and the following α-helix **(Cheng,2000; Martin *et al.*, 2002)**. (B) Topology diagrams of the selected homologous structures. Cylinders represent α-helices and arrows represent β-strands. Gray colored secondary structure elements indicate differences among the homologous structures. AD-003, adrenal gland protein (unpublished); TPMT, thiopurine S-MTase **(Wu *et al.*, 2007)**; RebM, rebeccamycin sugar O-MTase **(Singh *et al.*, 2008)**; MtfA, glycopeptide N-MTase **(Shi *et al.*, 2009)**; COMT, catechol O-MTase **(Rutherford *et al.*, 2008)**; GAMT, guanidinoacetate MTase **(Komoto *et al.*, 2003)**; DesVI, N,N-dimethyltransferase **(Burgie *et al.*, 2008)**; SynOMT, CCoAOMT-like O-MTase **(Kopycki *et al.*, 2008)**; GNMT, glycine N-MTase **(Huang *et al.*, 2000)**.

There are also some conserved residues that are not directly related to SAH or SAM binding. For instance, L49 (β1) and F108 (β4) form a CH···π interaction within the molecular β-sheet. The residues L137 (β5) and I111 (β4) have non-polar interactions while D109 (β4) forms hydrogen bonds with main-chain N atoms of S46 and N47 (β1). G140 is one of the residues in β5 which form the β-bulge in opposition to D109 in strand β4. It is difficult to attribute precisely defined roles to these residues as they are located far from the reaction center at the SAH ligand or from the putative acceptor binding groove. It is clear, however, that they are all important for the integrity of the central β1-β4-β5 part of the β-sheet, which forms the very core of the NodS structure.

Comparison of rhizobial NodS proteins

The sequence of NodS from *Bradyrhizobium japonicum* WM9 (GenBank accession no. AAK00158) was aligned with selected NodS sequences from other species that belong to the order Rhizobiales using ClustalW2 (Larkin *et al.*, 2007). The selected species are *Rhizobium* sp. NGR234 (GenBank accession no. NP_443986), *Rhizobium tropici* CIAT899 (Q53514), *Rhizobium etli* CFN 42 (NP_659762), *Sinorhizobium* sp. BR816 (CAH04370), *Sinorhizobium fredii* USDA257 (Q02007), *Mesorhizobium loti* E1R (AAG14460), *Mesorhizobium loti* MAFF303099 (NP_106712), *Bradyrhizobium japonicum* USDA 110 (P26026) and *Azorhizobium caulinodans* ORS 571 (YP_001526731) (Figure 42).

The alignment revealed a deletion in the N-terminal part of NodS from *Sinorhizobium fredii* strain USDA257, which is known to produce unmethylated NF (Bec-Ferté *et al.*, 1994). The deletion corresponds to αA, αB, β1, and the connecting loops of the present NodS structure and will certainly impair the binding of the SAM donor, thus explaining the lack of N-methyl group at the non-reducing end of the Nod factor from this bacterial strain.

The active-site residues responsible for polar interactions with the SAM ligand (R31, H32, G52, A54, D73, I99, D98 and A114) are quite conserved. Non-polar interactions involving L10, W20, L22, C53, V74, M75, A78, T97, E115, V116, Y119 and L120 (residue numbering corresponding to the present NodS protein) are also conserved. With reference to the predicted binding of the oligosaccharide substrate, the residues E13, E28, Y118, W156 and H158 implicated in interactions with the COS ligand are also highly conserved among the selected rhizobia.

The overall scheme of the sequence alignment identifies a particularly highly conserved region. It comprises the β4 strand and the loop linking β4 and αE, and is located in

the middle of the structure. In this region, there is a strictly conserved motif, EVLYY. If we exclude the sequence of *Azorhizobium caulinodans* ORS 571 from the alignment, this motif can be extended to FDLIVVAEVLYY.

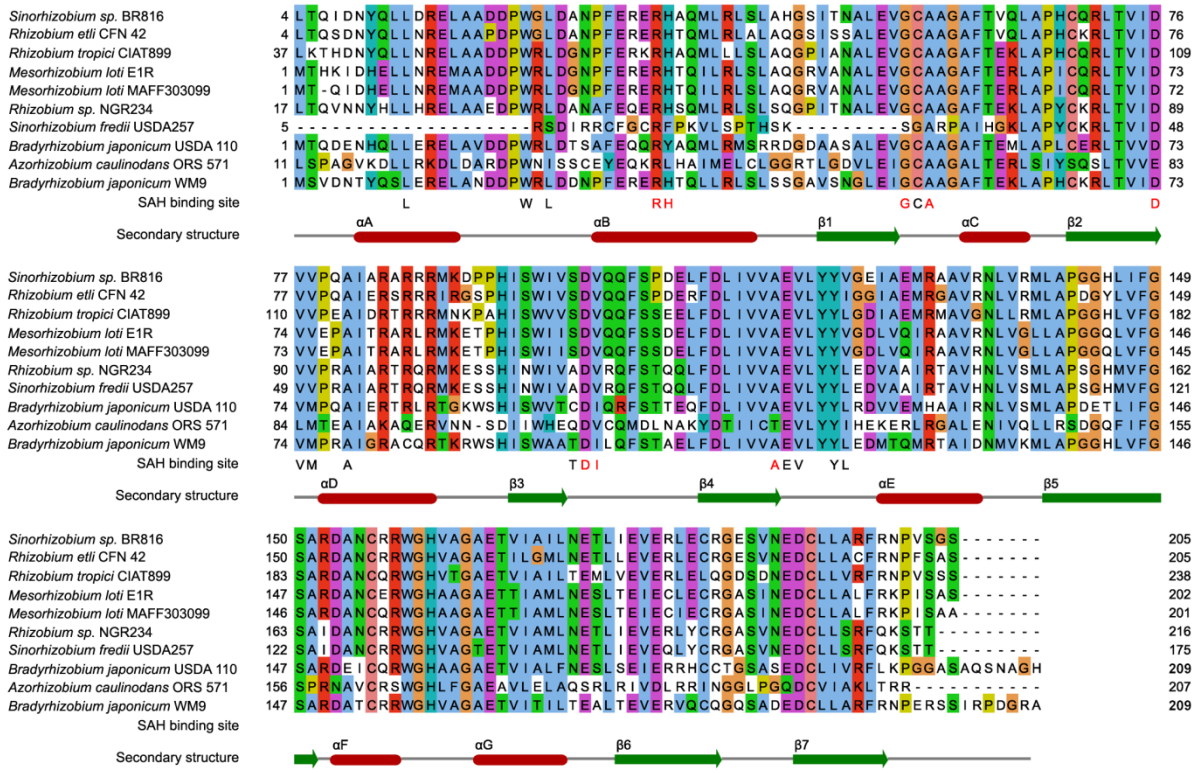


Figure 42 Multiple sequence alignment of selected NodS proteins from Rhizobia, calculated using ClustalW2 **(Larkin *et al.*, 2007)**. The figure was prepared in Jalview **(Clamp *et al.*, 2004)** and the Clustal X color scheme was used. Secondary structure elements are indicated as present in the NodS structure from *Bradyrhizobium japonicum* WM9, according to Fig. 34. Below the alignment, residues forming polar interactions with the SAH molecule are indicated by red lettering, while residues forming non-polar contacts are shown in black.

DISCUSSION

Discussion

Recombinant protein production and purification for crystallographic studies

To obtain big amount and highly purified protein for crystallization, the target genes (*nolA*, *nodB*, *nodD*, *nodS* and *nodW*) from *Bradyrhizobium japonicum* WM9 were successfully cloned and overexpressed in *E. coli* using as host organism. Although most of them were found insoluble, NodS and NodB proteins were successfully obtained as soluble after changing the growth conditions. The solubility of the both protein were improved by overnight expression at lower temperature.

Beside their solubility, high cysteine contents of the proteins lead to aggregation during initial purification of the proteins. To overcome this problem, several reducing agents were tested such as beta-mercaptoethanol, DDT and TCEP. Because of its stability and compatibility with nickel columns, TCEP mainly was used during purifications. Both proteins were purified via FPLC system for crystallization screening. Protein yield for NodS after the purification was about 10 mg per liter culture. As to NodB, it was about 5 mg per liter culture. Homogenity analysis of the pure proteins showed that NodS protein was monomeric and homogenous as it was shown by gel filtration chromatography and Native PAGE. On the other side, similar results were not obtained for purified NodB protein sample. Although NodB protein was obtained in the monomeric form after gel filtration chromatography, Native PAGE analysis showed that the protein was not homogenous. It indicates that NodB is not stable as much as NodS. NodB seems to need more effort to be obtained as homogenous and stable protein which is suitable for crystallization screening.

Insoluble proteins other than NodS and NodB were tested for renaturation from their inclusion bodies. NodW was only protein which is renatured from the inclusion bodies as a soluble protein. Preliminary purification showed that it can bind to nickel affinity column. After purification, the protein was stored at -20ºC. Unfortunately, aggregation was observed after one week. Besides its aggregation problem, NodW needs activity test to prove that the refold is correct. From the literature knowledge, it is known that NodW should be phosphorylated before its binding to the target DNA sequence, see review (Loh *et al.*, 2003).

Labeling of NodS with SeMet, crystallization and diffraction

In this study, NodS was the only protein which was homogeneous and stable. So it was the promising protein for crystallization screens. In order to have a solution for phasing, SeMet derivative NodS was also expressed and purified beside native NodS. Crystallization screens were carried out for ligand free NodS and NodS in complex with S-adenosyl-L-homocysteine (SAH). Although SeMet derivative ligand free NodS crystals were not obtained, SeMet derivative NodS-SAH crystals were grown spontaneously. Because its diffraction was not so good (data not shown), those initial crystals were used to apply streak seeding method for native and SeMet derivative NodS-SAH complex.

Crystals of native ligand free NodS, SeMet derivative NodS-SAH complex and native NodS-SAH complex diffracted at 2.42 Å, 2.0 Å and 1.85 Å, respectively. The crystals of ligand free NodS are tetragonal, space group $P4_322$, with one protein molecule in the asymmetric unit estimated at a Matthews coefficient (Matthews 1968) of 1.74 $Å^3.Da^{-1}$ (solvent content 29.2%). The NodS-SAH complex crystallizes in the orthorhombic space group $P2_12_12$ with four molecules in the asymmetric unit estimated at a Matthews coefficient (Matthews,1968) of 2.28 $Å^3.Da^{-1}$ (solvent content 46.6%).

NodS phasing

Because there is no identical or highly similar structure, multiwavelength anomalous diffraction (MAD) method was mainly used for NodS-SAH complex to solve the phase problem. Anomalous signals obtained via Se MAD experiment were used to calculate the position of the heavy atom sites. Although the expected number of Se atoms in the asymmetric unit is 24, it was found 21 Se atoms using the programs of SHELXD and HYSS. The final model of NodS-SAH complex clearly shows the reason. The methionine amino acid at the N terminal of the Chains B, C and D of the NodS-SAH complex is referred to the disordered region of the asymmetric unit.

Obtained phase information was used for model building of the NodS-SAH complex structure. The latter model was also used to solve ligand free NodS structure via Molecular Replacement method.

Overall structure

The structure of the NodS N-methyltransferase, which is involved in the biosynthesis of the Nod factor in Bradyrhizobia, has been elucidated in this study. As an enzyme that

methylates a chitooligosaccharide substrate, NodS is the first example among all SAM-dependent methyltranferases to have its three-dimensional structure determined. Although, the function is known (Geelen *et al.*, 1995; López-Lara *et al.*, 2001) , there is no E.C. number for chitooligosaccharide N-methyltransferases within the system of international Enzyme Classification.

The NodS N-methyltransferase from *Bradyrhizobium japonicum* WM9 shows an overall structural similarity to other Class I SAM-dependent methyltransferases, all of which have a common fold composed of six parallel and one antiparallel β-strands, surrounded by two layers of α-helices. In terms of acceptor molecules, those methyltransferases have a broad range of substrates. They can methylate DNA, proteins, lipids, carbohydrates, and a number of small molecules (Cheng,2000; Martin *et al.*, 2002).

SAM/SAH binding pocket

The methyl donor SAM molecule has different conformation in solution and in complex with a protein (Markham *et al.*, 2002). The SAH molecules in the crystal structure of the NodS-SAH complex have the less favored extended conformation (defined by the Sδ-C5'-C4'-O4' torsion angle) which is usually observed when nucleotides are bound to proteins. The SAH conformation is in agreement with the three-dimensional structures of other methyltransferase complexes (Ishida *et al.*, 1982; Saenger,1983; Markham *et al.*, 2002). The ribose ring of the SAH molecules in the NodS-SAH complex has C2'-*endo* pucker, with the adenine base in *anti* orientation around the glycosidic bond (defined by the O4'-C1'-N9-C8 torsion angle). The same conformation of the nucleoside fragment of the ligand molecule has been found in methyltransferases defined as Class I (Schubert *et al.*, 2003). The SAH molecule has excellent definition in electron density (Figure 37B) and its placement within the NodS fold is similar to the location of SAH/SAM in the core of other Class I methyltransferases. For example, structural superposition of the Cα atoms of the NodS-SAH complex and the SAM complex of MtfA (Shi *et al.*, 2009) (glycopeptide N-methyltransferase) illustrates a good overlay of the cofactor molecules (not shown), despite the fact that MtfA is a dimeric enzyme.

Comparison of ligand-free and SAH complex structures of NodS

A comparison of the ligand-free NodS and NodS-SAH complex structures reveals some important conformational differences. The changes in the N-terminal fragment of the protein, such as ordering of the αA helix and the αA-αB loop, are obviously related to the burying of the SAH molecule in the methyl donor cavity. At the same time, they lead to the formation of one wall of the binding groove for receiving the acceptor COS molecule of the methylation reaction. In addition to the rearrangement of the N-terminal fragment of the protein, a significant conformational change of the β4-αE loop leads to stabilization of the donor molecule in its binding cavity and to shaping of the remaining elements of the acceptor groove. Other changes, seen in the C-terminal part of the structure, especially in the loop αF-αG and loop β6-β7, might be related to interactions with the non-reducing end of the oligosaccharide substrate in the groove. In summary, as a result of SAH/SAM binding, the protein molecule becomes more compact and ready to receive the oligosaccharide acceptor molecule into its docking groove.

Possible reaction mechanism

The catalytic mechanism of SAM-dependent methyltransferases is thought to be a classic S_N2 reaction (Coward *et al.*, 1971). This mechanism requires linear arrangement of the nucleophile, the methyl carbon atom, and the thioester leaving group in the transition state of the reaction. The predicted binding orientation of the non-reducing end of the COS molecule is capable of reproducing the linear arrangement of the acceptor nitrogen, $C(H_3)$, and sulfonium atoms for nucleophilic attack. The catalytic mechanism of NodS is highlighted in Figure 43 by the putative interactions of the SAM and COS substrates in the enzyme's active site.

In preparation for the methylation reaction, the SAM substrate, which is bound first in the active site, is stabilized in the binding cavity by polar and non-polar interactions as mentioned earlier. Subsequently, two successive sugar units at the non-reducing end of the COS molecule could be locked by the side chains of E13, E28, E115, Y118, W156, H158 and D187 in the putative COS binding groove, bringing the acceptor NH_2 group into close proximity of the sulfonium group of the SAM molecule (Figure 38).

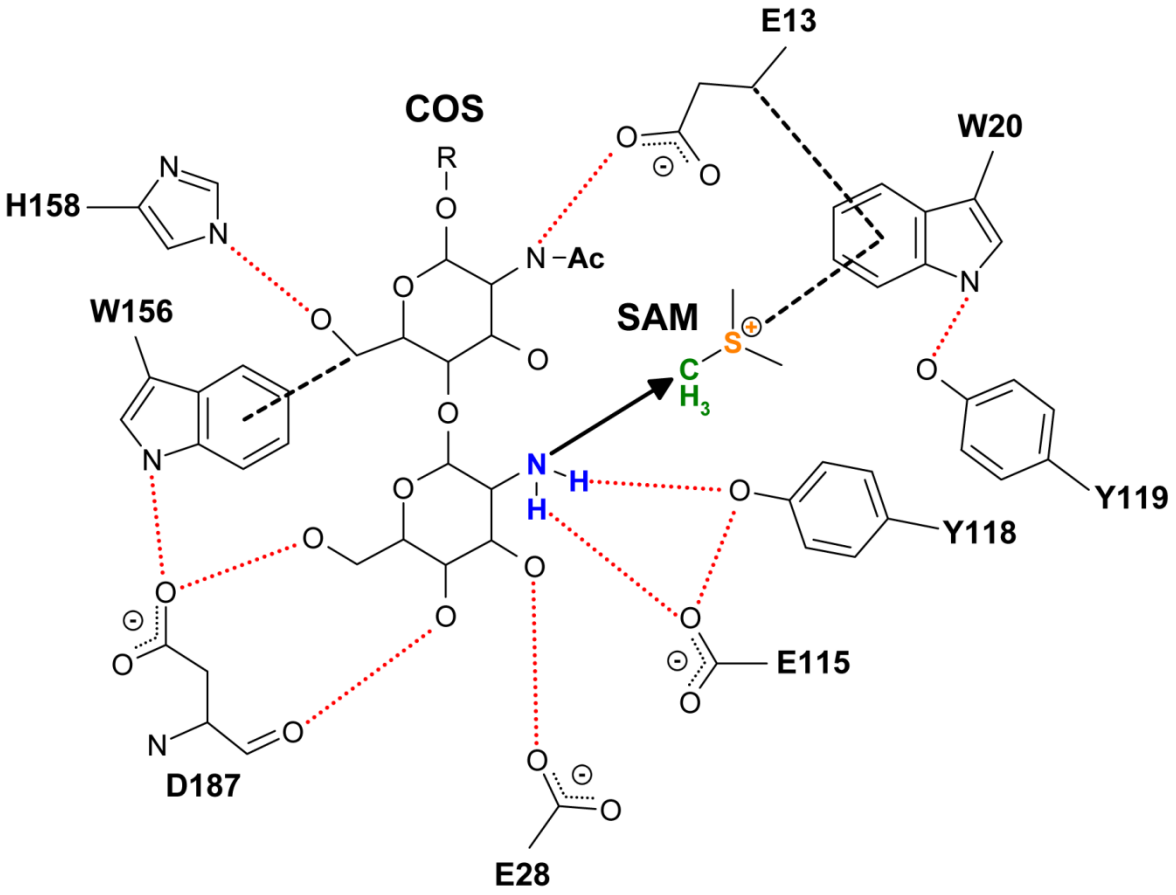


Figure 43 Scheme of the proposed speculative catalytic mechanism of the NodS protein. The methyl transfer reaction is thought to be a classic S_N2 reaction **(Coward *et al.*, 1971)**. The nucleophilic nitrogen atom (blue) of the non-reducing end of the COS molecule has a roughly co-linear orientation with the methyl donor $C(H_3)–S^+$ group of the SAM molecule in the transition state of the reaction. Red dotted lines indicate hydrogen bonds, black dashed lines show CH···π or cation···π interactions. The black arrow shows the direction of the nucleophilic attack of the $–NH_2$ group of the oligosaccharide substrate on the methyl C atom of the sulfonium group of the SAM molecule.

Through the hydrogen bonding network involving the NH_2 group of the non-reducing end of the COS molecule, the Oη atom of Y118 and Oε1 atom of E115 would increase the electronegativity of the substrate $–NH_2$ group and orient the lone electron pair on the nitrogen nucleophile to point towards the incoming methyl group. In particular, the carboxylate group of E115 would be an excellent candidate for proton abstraction from the substrate amino group and, possibly, for protonation of the leaving methyl-amino product. Proton abstraction would be especially important if, in the chemical environment of the NodS reaction, the amino substrate exists in protonated $–NH_3^+$ form. In addition, a cation···π interaction between the positively charged sulfur atom of SAM and the six-membered ring of W20 would ease the departure of the methyl carbon from the sulfonium group (Figure 43).

After the methyl-transfer reaction, the presence of the bulky $–CH_3$ group would destabilize the COS amino group in the active site by weakening its hydrogen bonds with E115 and Y118, thus promoting the departure of the methylated COS molecule from the enzyme. At the donor side of the reaction center, the transfer of the methyl group neutralizes the Sδ atom of SAH, thus removing the stabilizing cation···π interaction with the aromatic side chain of W20. This would promote the departure of the SAH product from the active site, especially if there would be competition for the binding cavity from another SAM donor molecule. During the release of the SAH molecule, the αA plus αA-αB element would become transiently disordered, as seen in the ligand-free structure. It appears form the NodS-SAH structure that the αA plus αA-αB element does not form a very strong lock, especially in molecule D, where it is close to a solvent channel in the crystal packing. A competition between SAM and SAH binding is, therefore, not impossible. If this three-component equilibrium were true, then the methylation reaction would be driven by both an excess of SAM or by removal of the SAH molecules in a secondary process (for instance, enzymatic hydrolysis).

Structurally homologus proteins

A search for structures homologous to the NodS-SAH complex resulted mostly in SAM-dependent methyltransferases, with no homologs that would be really close to NodS. An adrenal gland protein AD-003 (PDB ID 2EX4) from *H. sapiens* was the closest structural homolog, with Cα superposition characterized by an RMSD value of 2.21 Å. However, the C-terminal part of this protein has a β-sheet insertion between β5 and αG, instead of the α helix present in the NodS structure. This significant difference might indicate that the two enzymes utilize quite different acceptor molecules. Although the structure of the adrenal gland protein AD-003 has been solved in complex with SAH, the exact function of the protein is unknown. A multiple 3D alignment of NodS with selected homologous protein structures supports our current view about SAM-dependent methyltransferases. However, it is not possible to make a precise assignment of catalytic residues based on these alignments, as the acceptor substrates are very different. For instance, the above-mentioned superposition with MtfA, which methylates the N-terminus of a glycopeptide, highlights residues in an MtfA-specific appending structure (Shi *et al.*, 2009), absent in NodS.

Although SAM-dependent methyltransferases show a rather poor overall conservation of their SAM-binding residues, the SAH binding region of the NodS-SAH complex is consistent with a pattern noted in other SAM-dependent methyltransferases. With reference

to the NodS structure, this pattern can be described as consisting of three motifs (Cheng,2000; Martin *et al.*, 2002): (i) a glycine-rich sequence in the β1-αC loop, (ii) an acidic amino acid residue in the β2-αD loop, and (iii) a highly conserved aliphatic residue in strand β4. In the case of NodS, the glycine-rich region between β1 and αC has the sequence EXGXAXG instead of the consensus signature E/DXGXGXG, often referred to as motif I, but the Ala replacement of the second Gly position is fairly conservative. The highly conserved D73 residue at the tip of the β2 strand is involved in hydrogen bonding with both hydroxyl groups of the ribose moiety of the SAH molecule. The highly conserved aliphatic position in strand β4 corresponds to I111 in NodS.

The alignment also reveals important structural differences in the C-terminal part of the proteins, in which their acceptor substrates are bound. In particular, the type of insertion (i.e. a helix or a β-sheet) found between β5 and αG, and the size of the β6-β7 loop, differ among the selected structures.

Comparison of rhizobial NodS proteins

A multiple alignment of the NodS sequence from *Bradyrhizobium japonicum* WM9 with other rhizobial NodS sequences has revealed important conserved motifs related to the three-dimensional structure of NodS. All residues that are responsible for the binding of the SAH molecule are highly conserved among the NodS proteins from different species. Within the overall scheme of the alignment, strand β4 and the loop β4-αE elements, which belong to the C-terminal part of the molecule, contain a strictly conserved motif EVLYY (residues 115-119). If the sequence of *Azorhizobium caulinodans* ORS 571 is excluded from the alignment, the motif can be extended to the FDLIVVAEVLYY (residues 108-119) dodecapeptide. This highly conserved sequence motif corresponds to a site within loop β4-αE, where significant conformational changes are observed upon binding of the SAH molecule. The residues of this motif form part of the wall of the SAH binding cavity and of the COS binding groove. V116 and Y119 act to stabilize the SAH molecule as well as the αA-αB loop, which itself changes conformation upon SAH binding. The positions of E115 and Y118 in the SAH complex structure might indicate their possible role in the catalytic mechanism of NodS, as explained above. Consequently, this highly conserved sequence signature, located at the very heart of the structure, could be considered as a fingerprint of NodS proteins.

It should be mentioned here that *Azorhizobium caulinodans* produces N-methylated Nod factor (Mergaert *et al.*, 1997). It is also noteworthy that the nodulation genes of

Azorhizobium caulinodans and their predicted amino acid sequences are the most dissimilar ones with respect to nodulation genes harbored by other rhizobia. Thus, the NodS sequence of this rhizobium species could be regarded as a gauge showing how diverging amino acid sequences could still retain their specific enzymatic activity.

Another interesting result of the sequence alignments is the detection of the loss of a functional *nodS* gene in the soybean-nodulating strain *Sinorhizobium fredii* USDA257. This strain is closely related to the broad-host-range strain NGR234, with which it shares 98% sequence similarity in its nodulation loci. Unlike strain USDA257, strain NGR234 does not form productive symbiosis with soybean, despite the enormous host range of this strain (Pueppke *et al.*, 1999). The *nodS* gene in *S. fredii* USDA257 does not have a start codon and carries two single-nucleotide deletions in its 5' half, becoming in effect a pseudogene (data not shown). It is conceivable that the loss of *nodS* in this strain was not accidental, and might reflect a specific adaptation enabling effective nodulation of soybean plants.

In some soybean-nodulating strains e.g. in *Bradyrhizobium japonicum* USDA110, acetylation of carbon C-6 of the non-reducing end N-glucosamine unit by the product of the *nodL* gene (which codes for carbon C-6 specific acetyltransferase) interferes with N-methylation dependent on NodS (López-Lara *et al.*, 2001). In strain USDA110, the *nodS* gene appears to be functional. Therefore, loss, and not only gain, of a particular nodulation gene in the course of adaptation for nodulation of certain legumes, cold be a rather common phenomenon as revealed by the identification of pseudogenes of fucose-specific methyl- and sulfur-transferases in soybean and lupine nodulating rhizobia (Stepkowski *et al.*, 2007).

Summary

In this project, six target proteins (NodA, NodB, NodD, NodS, NodW, and NolA), involved in Nod factor biosynthesis in rhizobia were chosen to elucidiate their three dimensional structure by x-ray crystallography. For that purposes, the genes from *Bradyrhizobium japonicum* WM9 which is a nitrogen assimilating symbiont of lupine and serradella were cloned and expressed in *E. coli* cells. Two of them (NodB and NodS) were obtained soluble and purified for crystal screening. The screening studies were only resulted in NodS crystals. After the optimization of the crystal condition, two different forms of NodS (ligand free and SAH complex) crystals were successfully obtained. To solve phase problem, the crystals of SeMet derivatives of NodS in complex with SAH were used for multiwavelength anomalous diffraction (MAD). At the end, the structures of ligand-free NodS and of NodS-SAH complex were refined to 2.43 and 1.85 Å resolution, respectively.

NodS is a SAM-dependent methyltranferase that modifies a chitooligosaccharide substrate during the synthesis of the Nod factor. Although the overall fold of this NodS N-methyltransferase is similar to other SAM-dependent methyltrasnferases of Class I, and consists of a seven-stranded open β-sheet flanked on each side by a layer of α-helices, the organization of the β-sheet in its C-terminal part is different, reflecting the unique character of the chitooligosaccharide substrate as acceptor during the methylation reaction. It was possible to crystallize a complex between the NodS protein and S-adenosyl-L-homocysteine, which is a byproduct of the methylation reaction and a very close analog of the SAM methyl donor substrate. A comparison of the ligand-free NodS structure and the NodS-SAH complex reveals that functionally important conformational changes take place upon binding by the enzyme of the methyl donor molecule. In particular, ordering of the N-terminal helix αA and loop αA-αB, which cover the SAM/SAH molecule from the top, creates in effect a donor substrate binding cavity, with only one small opening, through which the sulfonium (i.e. methyl-donating) group is presented. At the same time, the now-ordered N-terminal elements of the NodS fold form one wall of a clearly discernible canyon on the protein surface. Structural elements from the C-terminal fragment of the protein form the opposite wall of the groove. Part of the canyon wall is formed by residues from a loop connecting strand β4 and helix αE, which contains a conserved sequence motif, quite exceptional within the poorly conserved family of SAM-dependent methyltransferases. The β4-αE loop also changes its conformation upon SAH binding. The canyon formed on the surface of the NodS-SAH

complex could be unambiguously recognized by automatic molecular docking as the binding site for the chitooligosaccharide (COS) substrate. The docking routine confidently placed the acceptor NH_2 group at the non-reducing end of the COS molecule opposite to the "sulfonium entrance" of the SAM binding cavity. The model of the ternary NodS-SAM-COS complex thus created is, therefore, quite credible and can be used for detailed studies of the mechanism of NodS. In particular, it is quite obvious from the present structural data, that the first substrate bound by NodS is the SAM methyl donor, whose binding is required for the creation of the docking site for the methyl acceptor substrate. After completion of the methyl transfer reaction, the products would be released in the reverse order, with methylated chitooligosaccharide leaving the complex first, and the SAH molecule leaving later, perhaps through exchange with a new SAM molecule. The crystal structure of NodS in combination with the molecular docking experiments provides important clues about the binding mode and protein interactions of the oligosaccharide substrate of the N-methylation reaction. Finally, the present structures highlight the amino acid residues that are crucial for the binding of both substrate molecules, as well as for the catalytic mechanism, and in consequence advance our understanding of the results of amino-acid sequence alignments of NodS proteins from different rhizobia in the context of their function and evolution.

LITERATURE

Literature

Adams P.,D., Grosse-Kunstleve R.,W., Hung L.W., Ioerger T.,R., McCoy A.,J., Moriarty N.,W., *et al.* (2002) PHENIX: Building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58:** 1948-1954.

Bec-Ferté M.,P., Krishnan H.,B., Promé D., Savagnac A., Pueppke S.,G., and Promé J.,C. (1994) Structures of nodulation factors from the nitrogen-fixing soybean symbiont rhizobium fredii USDA257. *Biochemistry* **33:** 11782-11788.

Bijvoet J.M. (1949) Phase determination in direct fourier-synthesis of crystal structures. *Proc Acad Sci Amst* **B52:** 313-314.

Bradford M.,M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72:** 248-54.

Brea M., Zamuner A., Matheos S., Iglesias A., and Zucol A. (2008) Fossil wood of the mimosoideae from the early paleocene of patagonia, argentina. *Alcheringa* **32:** 427-441.

Brzezinski K., Stepkowski T., Panjikar S., Bujacz G., and Jaskolski M. (2007) High-resolution structure of NodZ fucosyltransferase involved in the biosynthesis of the nodulation factor. *Acta Biochim Pol* **54:** 537-549.

Burgie E.S., and Holden H.M. (2008) Three-dimensional structure of DesVI from streptomyces venezuelae: A sugar N,N-dimethyltransferase required for dTDP-desosamine biosynthesis *Biochemistry* **47:** 3982-3988.

Burling F.,T., Weis W.,I., Flaherty K.,M., and Brünger A.,T. (1996) Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* **271:** 72-7.

Cakici O., Sikorski M., Stepkowski T., Bujacz G., and Jaskolski M. (2008) Cloning, expression, purification, crystallization and preliminary X-ray analysis of NodS N-methyltransferase from bradyrhizobium japonicum WM9. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **64:** 1149-1152.

Cantoni G.,L. (1952) The nature of the active methyl donor formed enzymatically from L-methionine and adenosine triphosphate. *J Am Chem Soc* **74:** 2942-2943.

Chan A.W., Hutchinson E.G., Harris D., and Thornton J.M. (1993) Identification, classification, and analysis of beta-bulges in proteins *Protein Sci* **2:** 1574-1590.

Cheng X. (2000) S-adenosylmethionine-dependent methyltransferases: Structures and functions. World Scientific Publishing Company.

Clamp M., Cuff J., Searle S.,M., and Barton G.,J. (2004) The jalview java alignment editor. *Bioinformatics* **20:** 426-427.

Coward J.K., and Sweet W.D. (1971) Kinetics and mechanism of methyl transfer from sulfonium compounds to various nucleophiles. *J Org Chem* **36:** 2337-2346.

Delano W. (2008) The PyMOL molecular graphics system. .

Dénarié J., Debellé F., and Rosenberg C. (1992) Signaling and host range variation in nodulation. *Annu Rev Microbiol* **46:** 497-531.

Emsley P., and Cowtan K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60:** 2126-2132.

Evans P. (2006) Scaling and assessment of data quality. *Acta Crystallographica Section D-Biological Crystallography* **62:** 72-82.

Geelen D., Leyman B., Mergaert P., Klarskov K., Van Montagu M., Geremia R., and Holsters M. (1995) NodS is an S-adenosyl-L-methionine-dependent methyltransferase that methylates chitooligosaccharides deacetylated at the non-reducing end. *Mol Microbiol* **17:** 387-397.

Hendrickson W.,A., Smith J.,L., Phizackerley R.,P., and Merritt E.,A. (1988) Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins* **4:** 77-88.

Huang Y., Komoto J., Konishi K., Takata Y., Ogawa H., Gomi T., *et al.* (2000) Mechanisms for auto-inhibition and forced product release in glycine N-methyltransferase: Crystal structures of wild-type, mutant R175K and S-adenosylhomocysteine-bound R175K enzymes *J Mol Biol* **298:** 149-162.

Ishida T., Tanaka A., Inoue M., Fujiwara T., and Tomita K. (1982) Conformational studies of S-adenosyl-L-homocysteine, a potential inhibitor of S-adenosyl-L-methionine-dependent methyltransferases. *J Am Chem Soc* **104:** 7239-7248.

James R. (1982) Optical principles of the diffraction of X-rays. Ox Bow Press.

Kabsch W., and Sander C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features *Biopolymers* **22:** 2577-2637.

Kirschner K.N., Yongye A.B., Tschampel S.M., Gonzalez-Outeirino J., Daniels C.R., Foley B.L., and Woods R.J. (2008) GLYCAM06: A generalizable biomolecular force field. carbohydrates *J Comput Chem* **29:** 622-655.

Komoto J., Takata Y., Yamada T., Konishi K., Ogawa H., Gomi T., *et al.* (2003) Monoclinic guanidinoacetate methyltransferase and gadolinium ion-binding characteristics *Acta Crystallogr D Biol Crystallogr* **59:** 1589-1596.

Kopycki J.G., Stubbs M.T., Brandt W., Hagemann M., Porzel A., Schmidt J., *et al.* (2008) Functional and structural characterization of a cation-dependent O-methyltransferase from the cyanobacterium synechocystis sp. strain PCC 6803 *J Biol Chem* **283:** 20888-20896.

Krissinel E., and Henrick K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60:** 2256-2268.

Larkin M.,A., Blackshields G., Brown N.,P., Chenna R., McGettigan P.,A., McWilliam H., *et al.* (2007) Clustal W and clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Laskowski R.,A., MacArthur M.,W., Moss D.,S., and Thornton J.,M. (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26:** 283-291.

Lavin M., Herendeen P.S., and Wojciechowski M.F. (2005) Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst Biol* **54:** 575-594.

Loh J., and Stacey G. (2003) Nodulation gene regulation in bradyrhizobium japonicum: A unique integration of global regulatory circuits. *Appl Environ Microbiol* **69:** 10-7.

Long S.,R. (1996) Rhizobium symbiosis: Nod factors in perspective. *Plant Cell* **8:** 1885-98.

López-Lara I.,M., Kafetzopoulos D., Spaink H.,P., and Thomas-Oates J.,E. (2001) Rhizobial NodL O-acetyl transferase and NodS N-methyl transferase functionally interfere in production of modified nod factors. *J Bacteriol* **183:** 3408-3416.

Markham G.D., Norrby P.O., and Bock C.W. (2002) S-adenosylmethionine conformations in solution and in protein complexes: Conformational influences of the sulfonium group *Biochemistry* **41:** 7636-7646.

Martin J.L., and McMillan F.M. (2002) SAM (dependent) I AM: The S-adenosylmethionine-dependent methyltransferase fold. *Curr Opin Struct Biol* **12:** 783-793.

Matthews B.,W. (1968) Solvent content of protein crystals. *J Mol Biol* **33:** 491-7.

Mergaert P., Van Montagu M., and Holsters M. (1997) Molecular mechanisms of nod factor diversity. *Mol Microbiol* **25:** 811-817.

Morris G.M., Goodsell D.S., Halliday R.S., Huey R., Hart W.E., Belew R.K., and Olson A.J. (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19:** 1639-1662.

Moulin L., Béna G., Boivin-Masson C., and Stepkowski T. (2004) Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the bradyrhizobium genus. *Mol Phylogenet Evol* **30:** 720-32.

Mueller-Dieckmann J. (2006) The open-access high-throughput crystallization facility at EMBL hamburg. *Acta Crystallogr D Biol Crystallogr* **62:** 1446-52.

Murshudov G.,N., Vagin A.,A., and Dodson E.,J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53:** 240-255.

Otwinowski Z., and Minor W. (1997) Processing of X-ray diffraction data collected in oscillation mode. Processing of X-ray diffraction data collected in oscillation mode. New York: Academic Press, pp. 307-326.

Painter J., and Merritt E.,A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr* **62:** 439-450.

Pape T., and Schneider T.R. (2004) HKL2MAP: A graphical user interface for macromolecular phasing with SHELX programs. *J Appl Cryst* **37:** 843-844.

Perrakis A., Morris R., and Lamzin V.,S. (1999) Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* **6:** 458-463.

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., and Ferrin T.E. (2004) UCSF chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25:** 1605-1612.

Pueppke S.G., and Broughton W.J. (1999) Rhizobium sp. strain NGR234 and R-fredii USDA257 share exceptionally broad, nested host ranges. *Mol Plant-Microbe Interact* **12:** 293-318.

Ramachandran G.,N., Ramakrishnan C., and Sasisekharan V. (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7:** 95-9.

Rao S.,T., and Rossmann M.,G. (1973) Comparison of super-secondary structures in proteins. *J Mol Biol* **76:** 241-256.

Raymond J., Siefert J.,L., Staples C.,R., and Blankenship R.,E. (2004) The natural history of nitrogen fixation. *Mol Biol Evol* **21:** 541-54.

Richardson J.S. (1977) Beta-sheet topology and the relatedness of proteins *Nature* **268:** 495-500.

Rossmann M.,G., and Blow D.,M. (1962) The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr* **15:** 24–31-24–31.

Rutherford K., Le Trong I., Stenkamp R.E., and Parson W.W. (2008) Crystal structures of human 108V and 108M catechol O-methyltransferase *J Mol Biol* **380:** 120-130.

Saenger W. (1983) Principles of nucleic acid structure. Springer-Verlag New York, LLC.

Schubert H.L., Blumenthal R.M., and Cheng X. (2003) Many paths to methyltransfer: A chronicle of convergence *Trends Biochem Sci* **28:** 329-335.

Sheldrick G.,M. (2008) A short history of SHELX. *Acta Crystallogr A* **64:** 112-22.

Shi R., Lamb S.S., Zakeri B., Proteau A., Cui Q., Sulea T., *et al.* (2009) Structure and function of the glycopeptide N-methyltransferase MtfA, a tool for the biosynthesis of modified glycopeptide antibiotics *Chem Biol* **16:** 401-410.

Singh S., McCoy J.G., Zhang C., Bingman C.A., Phillips G.N.,Jr, and Thorson J.S. (2008) Structure and mechanism of the rebeccamycin sugar 4'-O-methyltransferase RebM *J Biol Chem* **283:** 22628-22636.

Spaink H.,P. (2000) Root nodulation and infection factors produced by rhizobial bacteria. *Annu Rev Microbiol* **54:** 257-88.

Stepkowski T., Hughes C.,E., Law I.,J., Markiewicz Ł., Gurda D., Chlebicka A., and Moulin L. (2007) Diversification of lupine bradyrhizobium strains: Evidence from nodulation gene trees. *Appl Environ Microbiol* **73:** 3254-3264.

Stepkowski T., Swiderska A., Miedzinska K., Czaplińska M., Swiderski M., Biesiadka J., and Legocki A.,B. (2003) Low sequence similarity and gene content of symbiotic clusters of bradyrhizobium sp. WM9 (lupinus) indicate early divergence of "lupin" lineage in the genus bradyrhizobium. *Antonie Van Leeuwenhoek* **84:** 115-124.

Thompson J.,D., Higgins D.,G., and Gibson T.,J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

Tronrud D.,E. (2004) Introduction to macromolecular refinement. *Acta Crystallogr D Biol Crystallogr* **60:** 2156-68.

Weiss M.S. (2001) Global indicators of X-ray data quality. *Journal of Applied Crystallography* **34:** 130-135.

Winn M.,D., Isupov M.,N., and Murshudov G.,N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr D Biol Crystallogr* **57:** 122-33.

Wu H., Horton J.R., Battaile K., Allali-Hassani A., Martin F., Zeng H., *et al.* (2007) Structural basis of allele variation of human thiopurine-S-methyltransferase *Proteins* **67:** 198-208.

Abbreviations

COS	chitooligosaccharide
NCS	non-crystallographic symmetry
NF	Nod factor
RMSD	root-mean-square deviation
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl-L-methionine