Ph.D. dissertation

REDESIGN OF THE COLICIN E7 NUCLEASE DOMAIN INTO A CONTROLLED AND SPECIFIC ARTIFICIAL ENZYME

ESZTER NÉMETH

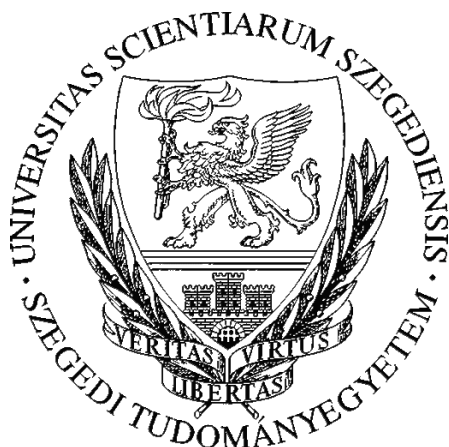**Supervisors:**

DR. BÉLA GYURCSIK
Associate professor
DR. TAMÁS KÖRTVÉLYESI
Associate professor

Doctoral School of Chemistry

Department of Inorganic and Analytical Chemistry

Department of Physical Chemistry and Material Science

Faculty of Science and Informatics University of Szeged, Hungary

2014.

TABLE OF CONTENTS

ABBREVIATIONS

Amp	ampicillin
BL21	name of an E.coli strain
bp	base pairs
CD	circular dichroism
CT-DNA	*calf thymus* DNA
CV	column volume
DH10B	name of an E.coli strain (Invitrogen TM)
DSB	double strand break
DSSP	protein secondary structure estimation program
dsDNA	double strand DNA
E.coli	Escherichia coli
EcoRI	name of a restriction endonuclease cleaving the G-AATTC sequence
EMSA	electro mobility shift assay
ESI	electrospray ionization
FLD	flow linear dichroism spectroscopy
FokI	name of a restriction enzyme
FPLC	fast protein liquid chromatography
GST	glutathione-S-transferase
HR	homologous recombination
HPLC	high performance liquid chromatography
Im7	the natural inhibitor protein of colicin E7
IPTG	isopropyl β-D-1-thiogalactopyranoside
ITC	isothermal titration calorimetry
KRNK	the 446-KRNK-449 part of the NColE7 sequence
LB	lysogeny broth or commonly Luria-Bertani medium
LD	linear dichroism
Mach1	name of an E.coli strain
MD	molecular dynamics
MM	molecular mechanics
MS	mass spectrometry
NColE7	nuclease domain of colicin E7
NColE7(m)	point mutant of NColE7
NdeI	name of a restriction endonuclease cleaving the CA-TATG sequence
NHEJ	non-homologous end joining
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEG	polyethylene-glycol
pET21a	name of a DNA plasmid
pGEX-6P-1	name of a DNA plasmid
PM6	parametric method 6, a semiempirical quantum chemical calculation method
pUC19	name of a DNA plasmid
QM/MM	hybrid quantum mechanics and molecular mechanics computations
RMSD	root mean square deviation
SCF	self consistent field
SRCD	synchrotron radiation circular dichroism
TALE	transcription activator-like effector
TOF	time of flight
XhoI	name of a restriction endonuclease cleaving the C-TCGAG sequence
ZF	zinc finger
ZFN	zinc finger nuclease

1. INTRODUCTION

The ability to cleave nucleic acids at a selected specific site is a current demand of biotechnology and gene therapy. The cleavage of chromosomal DNA at a sequence related to a disease can initiate the natural correction of DNA, and this phenomenon could be exploited in the cure of certain genetic diseases. The cleavage of nucleic acids is an important part of the cell machinery, also in normal conditions. Nature has evolved the nuclease enzymes to perform this task. However, only a few DNA sequences can be targeted by these molecules, limiting the applications for gene therapy where the sequence to target is determined by the disease. In order to cleave DNA at a desired sequence new engineered enzymes are necessary.

Zinc-finger nucleases are modular enzymes designed for the above purpose. They consist of a nonspecific catalytic center (the nuclease domain of the FokI bacterial enzyme) attached to a tunable recognition domain (three or four zinc-finger units). By the mutation of the residues that interact with the DNA bases the recognition sequence can be altered and custom-designed enzymes can be developed. Such enzymes were widely studied, but a minor cytotoxicity was found hindering the human therapeutic applications.

In this thesis the design of a novel type of zinc-finger nucleases is introduced. Instead of the commonly used catalytic center (FokI) we attempted to use the nuclease domain of colicin E7. The choice of this protein relied on previous investigations showing that the N-terminally truncated NColE7 is inactive. We hypothesized that the missing N-terminus could be used to design a novel structural control mechanism. The first two chapters describe the study of NColE7 point mutants that was important to understand the phenomenon and to get ideas on how to influence it, or how to use it for the design of a control mechanism. It was also essential to understand the Zn^{2+}-binding catalytic center. The computational design of the new enzymes is explained in the third chapter. Finally, the last section describes the expression and purification of three designed ZFN enzymes, as well as their *in vitro* nuclease activity and DNA-binding.

The results presented here may contribute to the development of a new design strategy that can be generally applied for the artificial nucleases.

2. LITERATURE SUMMARY

2.1. THE ROLE OF METAL IONS IN NUCLEASE ENZYMES

Several processes during the life of a cell require the hydrolytic cleavage of the nucleic acids, including DNA repair, DNA recombination (proofreading and primer removal), DNA topoisomerization, RNA splicing, programmed cell death, etc [1]. In bacterial cells the cleavage of foreign nucleic acids is also part of the cell defense [2]. The enzymes catalyzing this reaction are called nucleases. The hydrolysis can occur on either side of the O-P-O bridges in the nucleic acid backbone (**Fig. 2.1**).
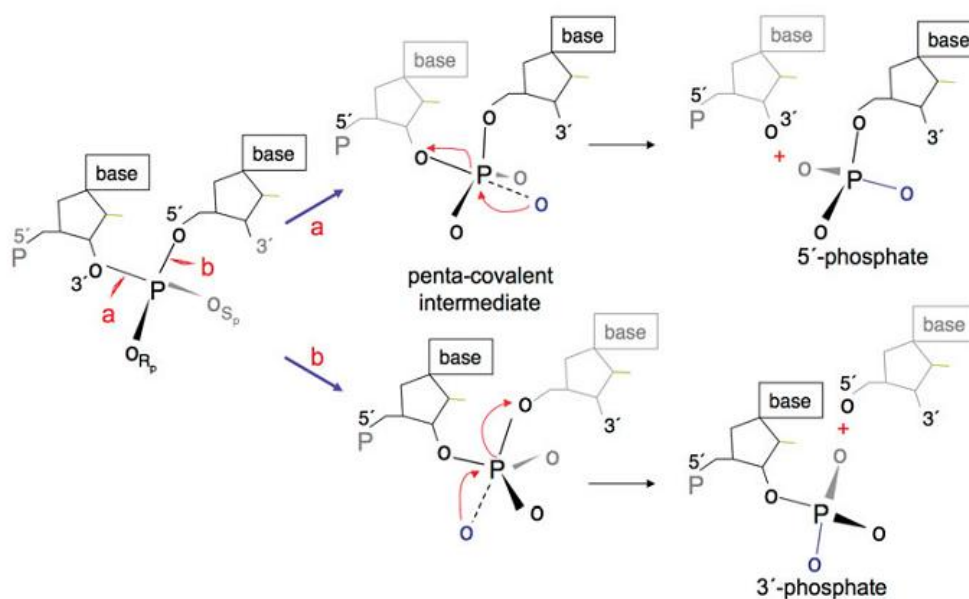
Fig. 2.1. *Reaction catalyzed by nuclease enzymes. The hydrolysis can occur on either side of the scissile phosphodiester, depending on the reaction mechanism, from* [1]

Nucleases can be divided into different groups. Specific nucleases recognize a sequence by interaction with certain DNA bases, while nonspecific nucleases interact mainly with the backbone atoms. Exonucleases perform cleavage at the termini of nucleic acid chains, and endonucleases within the sequence.

Several nuclease enzymes bind one or more metal ion cofactors fulfilling structural and/or catalytic roles [3]. The mechanism of metal-dependent nucleic acid hydrolysis varies based on the type, number and coordination sphere of the metal ions. Nuclease enzymes are exceptional in terms of diversity regarding active sites and mechanisms. Some nucleases do not need metal cofactor, others utilize one, two or three metal ions in the catalysis (**Fig. 2.2**). In contrast to site-specific nucleases that have a strong cofactor requirement the metal-binding specificity of nonspecific nucleases is debated. The promiscuity in metal-binding involving less strict metal requirement can be advantageous for nonspecific reactions [3].
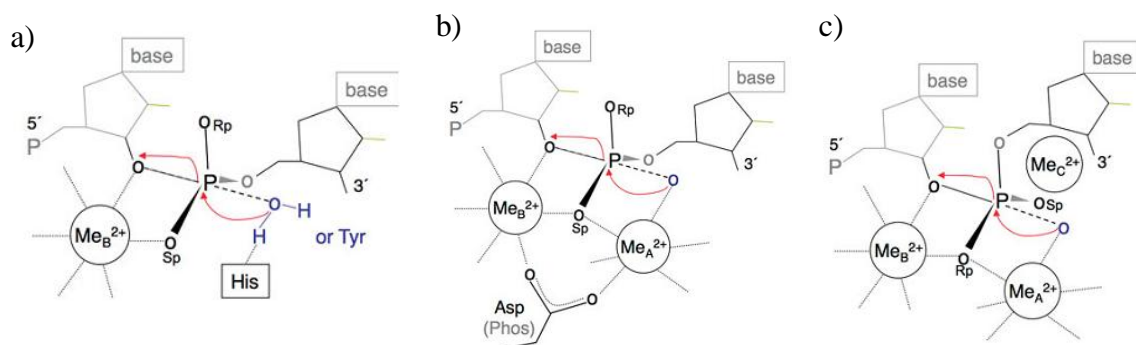
Fig. 2.2. *The general schemes for the one (a), two (b) and three (c) metal ion catalyzed phosphodiester hydrolysis reactions, from* [1].

The most common metal-ion in nucleases is Mg^{2+}, bound by hard donor atoms, such as oxygen atoms of Asp or Glu sidechains. The octahedral coordination sphere is partially filled by protein donor atoms and partially by water molecules. These water molecules are activated and take part in the reaction via the nucleophile OH^- generation for nucleophilic attack on the scissile phosphate or via protonation of the leaving group. The Mg^{2+}-bound water ligands have slow exchange rates, but can also be exchanged by the phosphate oxygen of the nucleic acid backbone, thus the metal ion is also responsible for the binding and electrostatic activation of the substrate as well as stabilization of the transition state [1]. The advantages of Mg^{2+} are the good bioavailability, stabile redox state and rigid geometry. In a few cases Ca^{2+} ions are bound and the nucleases with softer donor atoms in the metal binding site coordinate mostly Zn^{2+}-ions (e.g. colicin E7, nuclease P1, Endo VII). The proteins investigated in this project bind Zn^{2+}-ions, therefore the next chapter introduces the general role of Zn^{2+} in metalloproteins.

2.2. Zn^{2+}-BINDING SITES IN PROTEINS

Zn^{2+} was found in more than 300 enzymes [4]. The coordination sphere of Zn^{2+} can be either octahedral or tetrahedral (this is the most common) and even pentacovalent states can be formed. Since it has a filled *d* orbital, the ligand-field stabilization energy is zero and there is no preference between different coordination states. Therefore, it is able to catalyze reactions involving changes in its coordination sphere [4]. The ligand exchange reactions are rapid, thus the reaction product dissociates easily and reaction rates are often high. Zn^{2+} is a borderline metal ion between hard and soft polarization features, and can bind to several donoratoms in proteins, including most often the nitrogen donors of His, the sulfur atom of Cys, and oxygen atoms of Asp or Glu sidechains [5]. In rare cases the hydroxyl group of Tyr and the side-chain carbonyl oxygen of Asn or Gln can also serve as donor atoms for Zn^{2+} [4].

The Zn^{2+}-site can be structural, with all coordination sites occupied by protein donor atoms, or catalytic, when usually one or more coordination sites are occupied by water molecules (**Table 2.1**). Based on theoretical calculations compared with crystal structures it

was observed, that the donor atom – Zn^{2+} distances are longer in the catalytic, than in the structural Zn^{2+}-sites [6]. The longer distances may allow for the binding of an additional ligand, the change of the coordination number during reaction, and affects the Lewis acidity of Zn^{2+}. This can explain the role of protein environment in effective catalysis over metal-ligand complexes. Catalytic Zn^{2+}-sites rarely contain Cys residues, because when bound to the S-donoratoms Zn^{2+} has a higher electron density and thus it is less probable to act as a Lewis acid. As shown in **Fig. 2.3**, the water molecule in the catalytic site can be ionized to a hydroxide ion (e.g. carbonic anhydrase), polarized by a general base to generate nucleophile for catalysis (e.g. carboxypeptidase *A*) or exchanged to the phosphate (e.g. alkaline phosphatase) [4]. In the hydrolytic enzymes the Zn^{2+}-ion is responsible not only for providing an activated molecule for nucleophilic attack, but also for polarizing the scissile bond, and stabilizing the negatively charged transition state. The Zn^{2+}-polarized water molecule can also provide a proton for the leaving group.
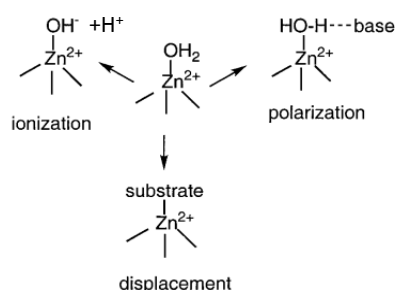
Fig. 2.3. Different functions of the Zn^{2+}-bound water molecule in the catalytic Zn^{2+}-sites [4].

***Table 2.1.** Features of different Zn^{2+}-sites [6]*

	catalytic	structural
coordination number	4,5,6	4
most common ligand	His, H_2O	Cys (2 or more)
bond lengths	longer	shorter

The coordination site of catalytic Zn^{2+}-ions frequently contains two side-chains close in sequence, separated only by one or a couple of residues, providing an anchor for the metal binding. The third ligand further in the sequence (within 200 residues) is responsible for the structural properties of the active site. This arrangement can provide a high affinity Zn^{2+}-binding site with some flexibility. It is commonly observed, that the catalytic Zn^{2+}-binding site consists of mainly hydrophilic residues, but this is surrounded by a hydrophobic core [4]. As for example, the mutation of residues F93, F95 and W97 in the carbonic anhydrase II enzyme resulted in decrease of selectivity towards Zn^{2+} over other metal ions, and decreased the stability of the structure [7].

In structural Zn^{2+}-sites the most common coordinating residue is Cys. Zn^{2+} can be stored by metallothioneins at Cys binding sites, containing several Zn^{2+}-ions in metal-thiolate clusters. Zinc-fingers are small DNA-binding proteins also containing structural Zn^{2+}-ions: in the absence of metal ions they lose their secondary structure.

2.3. ZINC-FINGER NUCLEASES

2.3.1. The need for custom designed, specific nucleases

In the last decades there is an increasing need to utilize specific nucleases to perform different kinds of gene manipulation, even for gene therapeutic purposes. In eucaryotic cells, whenever an unwanted double strand break (DSB) occurs, the cells have various mechanisms to correct this demage. Non-homologous end joining (NHEJ) means the ligation of the new termini and homologous recombination (HR) is possible in the presence of a template DNA matching the ends around the DSB (**Fig. 2.4 a**). This latter mechanism could be exploited in gene correction or addition, if the template DNA is delivered to the cell. However, the frequency of homologous recombination is very low. The reason to use nucleases is that the frequency of HR can be increased by several orders of magnitude if a DSB in the DNA chain is introduced near the targeted sequence [8] [9].
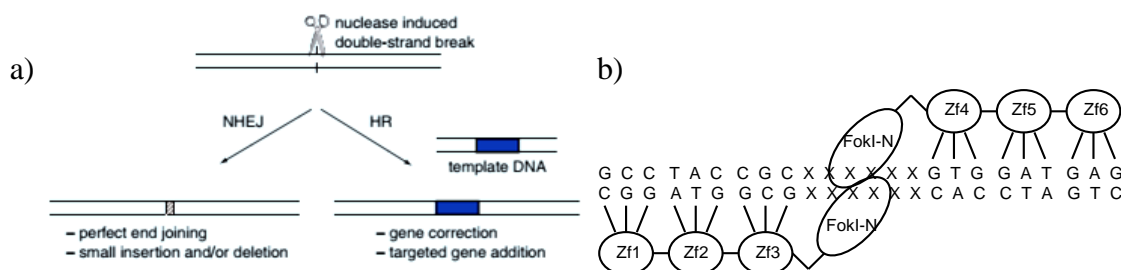
Fig. 2.4. a) *The two basic DNA repair mechanisms that can be induced by the double strand breaks* [10]: *non-homologous end joining and homologous recombination.* **b)** *Scheme of a zinc-finger nuclease. The zinc-fingers are denoted by Zf1-Zf6 and the nuclease domain of FokI by FokI-N. DNA bases recognized by ZF-s in this example are written by their one letter code, while at the X-positions any of the four bases can be present.*

The natural specific nucleases are inappropriate, since their recognition sequence is well determined by their three dimensional structure thus it is troublesome to redesign them to target an arbitrarily chosen sequence using them. Instead, several artificial nucleases have been designed with a variable recognition site [10]. The nuclease used for the above purpose has to be highly specific for a sequence that is unique in the genome. In one approach a specific enzyme can be redesigned to recognize the desired DNA sequence. However, that is a complex task and the length of the recognition sequence is limited. It is more advantageous to design modular enzymes that consist of a variable specific binding domain and a nonspecific nuclease domain. The two major groups of designed nucleases are the chemical nucleases (including metal complexes) and the chimeric enzymes. In both cases, the specificity is the most important question. The DNA-binding part can consist of triplex forming oligonucleotides, peptide nucleic acids or specific DNA binding proteins such as zinc-fingers or the transcription activator-like effector (TALE) motif. An example of such a zinc-finger nucleases (ZFN) is shown in **Fig. 2.4 b**. In this enzyme, the FokI nuclease domain is fused to a zinc-finger array responsible for specific DNA-binding. Each zinc-finger unit recognizes three subsequent bases on the DNA, thus the dimer ZFN can be designed to recognize 18-24 base pairs. The first ZFN of this kind was reported in 1996 by Chandrasegaran and coworkers [11].

2.3.2. Zinc-fingers

Zinc-fingers (ZF) are highly specific DNA-binding polypeptides, first found in transcriptional factors. They do not have a catalytic activity but bind DNA at specific sites, regulating the gene expression. Their general sequence and an example of three dimensional structure of a three ZF-protein bound to DNA are shown in **Fig. 2.5**. One ZF structure consists of two antiparalell β-strands and an α-helix. Generally two Cys residues of the β-strands and two His of the α-helix coordinate a Zn^{2+}-ion via their side chain donor atoms. The metal ion plays a structural role: it is essential for the folding of the motif.

b) **(Tyr, Phe)-X-Cys-X$_{2,4}$-Cys-X$_3$-Phe-X$_5$-Leu-X$_2$-His-X$_{3-5}$-His**
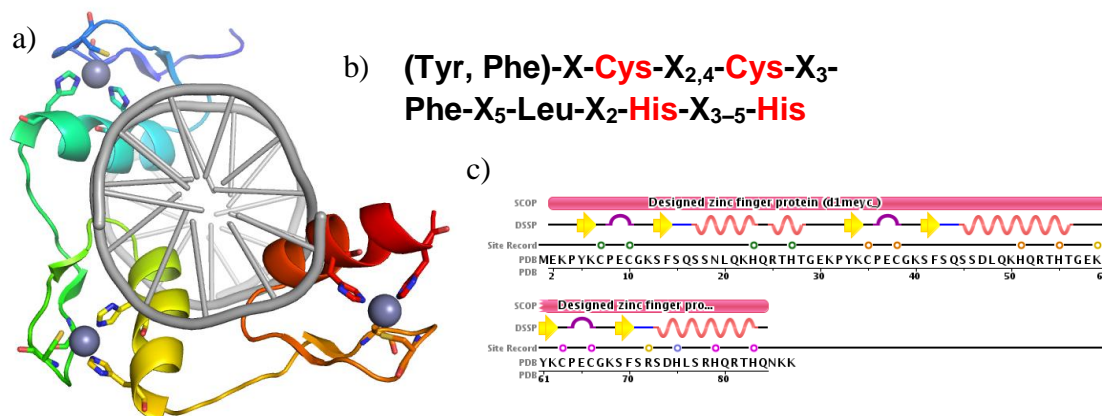
Fig. 2.5. a) The crystal structure of three zinc-finger units in complex with DNA [12]. The grey spheres represent zinc ions. b) The general sequence of zinc-fingers [13]. The metal binding residues are written in red. c) the sequence of the ZF protein (pdb ID: 1MEY) and the secondary structure elements [12].

The recognition of the DNA bases happens in the major groove. The specificity is mainly determined by 6 residues at the beginning of the α-helix, however, residues in the β-hairpin may also influence the DNA binding [14]. By mutating the mentioned residues the specificity can be changed and the desired DNA sequences can be targeted with high selectivity. However, it has to be noted that the optimization of specific binding is not a routine work and needs thorough studies. There are programs to predict the recognition sequence of zinc-fingers, such as Zinc Finger Tools [15]. To achieve a longer recognition sequence, tandem zinc-fingers can be used by the fusion of 3 or 4 domains. The fusion of more domains could destabilize the DNA-protein complex [16], but a 2×3 ZF array where the two 3 finger motifs are joined by a flexible linker sequence was reported to be effective in DNA binding [17]. The computational design of linker sequences between the individual ZF-s can improve the ZF design. Linkers can be selected to skip up to 10 bp in DNA between the individual fingers [18]. In the natural zinc-fingers the linker sequence that connects two fingers is conserved containing a TGEKP sequence. The assumed mechanism of DNA recognition is that the flexible linker allows for diffusion along the DNA chains and once the specific interactions are formed at the recognition sequence, the linker folds into an induced structure that provides a cap at the C-terminus of the preceding helix [19].

2.3.3. The FokI nuclease domain

FokI is a bacterial type II restriction endonuclease with a recognition sequence of 5'-GGATG-3'. It is under negative allosteric regulation: the catalytic reaction can occur only when the recognition site binds to the specific DNA sequence, and the conformational changes are forwarded to the cleavage domain. The specific DNA binding part is well separated from the catalytically active domain [20, 21], as shown in **Fig. 2.6**.
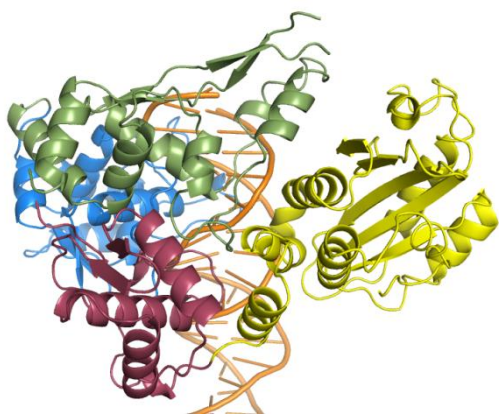
Fig. 2.6. The crystal structure of FokI in complex with DNA [20]. *DNA is shown in orange, the N-terminal, recognition domains are shown in blue, green and red, and the C-terminal catalytic domain is shown in yellow. The catalytic domain is connected to the recognition domains only via a flexible linker, that makes it possible to separate these domains.*

This feature of FokI made it possible to use its DNase domain − as a nonspecific nuclease when separated from the recognition domains − fused to zinc-fingers in a ZFN. As later it turned out, a further advantageous feature of FokI is the dimerization of the enzyme that is required for the DSB [22]. Thus, two ZFN-s have to cooperate (**Fig. 2.4 a**) forming a heterodimer and recognizing a sequence long enough to be unique even in the human genome. The studies of ZFN design revealed that a 6 bp sequence between the two binding sites is optimal for the dimerization of the FokI domain, but also longer sequence spacers allowed for a lower nuclease activity [23, 24].

2.3.4. Gene targeting with FokI-based ZFN enzymes

After the first successful *in vitro* experiments with chimeric FokI-based ZFN [11], it was not yet obvious that it is possible to introduce DSB to chromatin bound DNA *in vivo* in the same way. Soon it was shown in *Drosophila* (fruit fly) that both targeted mutagenesis [25] and gene replacement [26] could be induced by ZFN cleavage. Since then, several successful experiments have been reported in various cell types and organisms [10][27]. The experiments were extended to mammalian and human cell lines [27]. Presently, ZFN-s are utilized in clinical trials for the therapy of HIV [28].

The success of gene edition by ZFN depends on several factors including the DNA-protein interactions (specificity, affinity), interactions with DNA packaged with chromatin, DNA methylation, other proteins targeting the same sequence, etc. Furthermore, the off-target cleavages are harmful for the cell. Therefore, the design of experimental test systems is

another challenge of this research field [29]. The insertion of green fluorescent protein gene by ZFN-induced HR mechanism provides possibility for quantitative analysis of the process by measuring the fluorescence of the protein expressed from the inserted gene [30]. It has been reported, that even single strand DNA cleavage can induce HR, however, the efficiency is one quarter of that induced by DSB [31], and the process is even less toxic [32].

There are several approaches for the delivery of ZFN molecules to the target, for the application of ZFN-s to correct genes *in vivo*. These include e.g. the viral infection with engineered viruses that carry the DNA encoding the ZFN with a suitable promoter, mRNA and donor DNA injection to embryos or fusion of cell-penetrating peptides to ZFN-s. The template DNA can be delivered in the same way. To achieve therapeutic applications it is also essential to understand the DNA repair mechanisms better, and to be able to influence it, for example by favoring HR over NHEJ.

2.3.5. Specificity of ZFN enzymes

The presently available ZFN-s were shown to have a minor cytotoxicity, that can be attributed to off-target cleavages. Therefore, the specificity is a crucial point in ZFN design [33, 33]. The specificity of the enzyme depends on the DNA-binding part (ZF motifs), but not the nuclease part. However, as shown above, in order to achieve a sequence specificity of at least 18 bp, two ZFN molecules are required forming a homo- or heterodimer. To target a sequence, usually a heterodimer is required with two different recognition sites. It has been shown in *in vitro* experiments that the cleavage may also occur when only one of the two ZFN molecules is properly bound but the second is only interacting with weak nonspecific interactions at an off-target site (**Fig. 2.7 a**) [34]. This extends the recognizable sites to a large number and thus increases cytotoxicity by off-target cleavages. Furthermore, two distant sites can be looped by the ZFN dimer (**Fig. 2.7 b**).
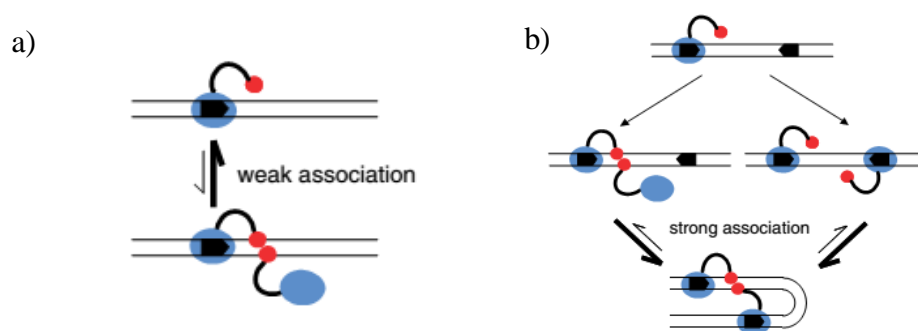
Fig. 2.7. *The specificity decreasing drawbacks of the application of ZFN dimers, taken from* [34]. ***a)*** *Cleavages at half-sites and **b)** DNA looping of two far half sites.*

The probability of half site cleavages can be diminished by careful redesign of FokI to favor heterodimers over monomers [35-38]. The recently invented transcription activator-like effector (TALE) nucleases or the clustered regularly interspaced short palindromic

repeat/CRISPR-associated (CRISPR/Cas) nuclease systems are more specific [39]. However, in spite of their advantages, none of these artificial nucleases can solve the specificity problems that are related to the degradation of the ZFN within the cell: injuries in the DNA binding domain let the nuclease operate on off-target sites.

An improvement in the prevention of cytotoxicity could be achieved by introducing control mechanisms to the nuclease. The natural FokI endonuclease works under negative control: the interaction between the recognition and nuclease domains inhibits the reaction as long as the enzyme has not found its target sequence [20, 21]. When the FokI cleavage domain is separated and fused to ZF-s, this control is lost. A higher safety process could be achieved, if the designed ZFN had a positive allosteric control. This means that it would be inactive in absence of the target sequence, but it would get reactivated via structural changes induced by the binding to to the specific substrate.

2.4. *The nuclease domain of colicin E7 (NColE7)*

There are numerous studies on FokI-zinc-finger nucleases, but the research to include other cleavage domains than FokI is limited, and the new suggested nuclease domains [40, 41] did not receive widespread interest in the construction of artificial nucleases. An attempt of exchanging two metal binding Cys residues to His in a zinc-finger domain was successful [42]. My PhD work focused on the design of a novel ZFN containing the nuclease domain of colicin E7 (NColE7). The selection of the nuclease is a former result of our research group [43], and so far there have been no examples on NColE7-fusion proteins published in the literature. Furthermore, we intend to use a redesigned form of the NColE7 protein. For these reasons it was necessary to thoroughly study its features.

2.4.1. *Biological function of NColE7*

Colicin E7 (ColE7) is a bacterial toxin produced by *E. coli* cells under stress conditions, being a member of the multistep defense mechanism of bacterial nucleases [2, 44]. It consists of three domains: receptor binding, membrane translocation and cytotoxic domains, from the N- to the C-terminus. After binding to the receptor of the attacked cell with the aid of the N-terminal domain (residues 3-308), the translocation domain (residues 288-446) helps the nuclease to get to the periplasmic space, where it goes through protease processing [2]. Only the nuclease domain (NColE7, sequence 446-576) enters the cell through the inner membrane and kills the cell by nonspecific digestion of chromosomal DNA molecules [2, 45]. RNase activity of NColE7 was also reported [46]. NColE7 is a very efficient toxin, one molecule is enough to kill a bacterial cell. The host cell protects itself by coexpression of the Im7 immunity protein forming a stable complex with NColE7. **Fig. 2.8** shows that Im7 binding does not affect the active site of NColE7, instead it prohibits the nuclease activity by hindering the DNA binding of NColE7 [46].
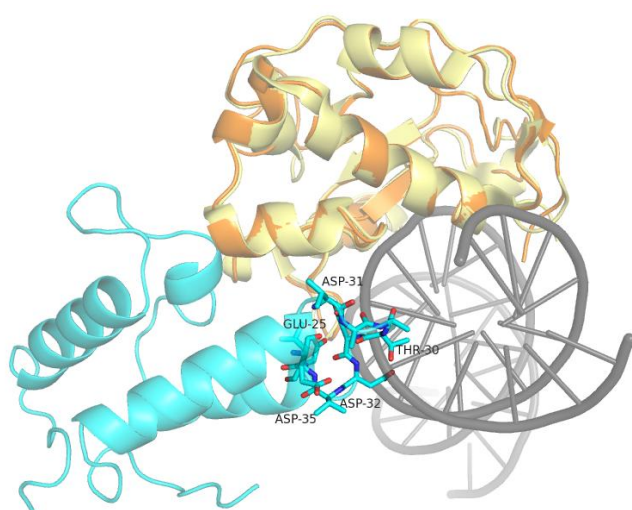
Fig. 2.8. *Superposition of the nuclease domain from the NColE7/Im7 crystal structure (7CEI, [47], Im7 in blue and NColE7 in orange) and from the NColE7/DNA crystal structure (3FBD, [48], DNA in grey and NColE7 in yellow). A negatively charged loop of the Im7 protein would overlap with the DNA, therefore the DNA binding of the nuclease domain is only possible in the absence of the Im7 protein.*

Colicin E7 is an example of colicins, for which the cell killing activity is assured by various mechanisms: pore-forming colicins depolarize the membrane and enzymatic colicins degrade RNA and/or DNA. Other nuclease colicins are E2, E8 and E9. For a detailed review on colicins see [49].

2.4.2. Structure of NColE7

NColE7 belongs to the HNH protein family. This family includes several types of proteins, among them e.g. certain homing endonucleases, restriction nucleases, inteins, introns. They were named based on three conserved residues in the sequence $EXH\underline{H}X_{14}N\underline{X}_8\underline{H}X_3H$. The HNH-proteins are DNA-binding and cleaving proteins. After the revolution of crystal structure determination, it became clear that they also share structural similarity. The HNH motif formed by ~40 residues is a His-Me finger, consisting of two antiparalell β-sheets and one α-helix, also called ββα-motif. The 3D structure of HNH motif is similar to ZF-s, however in contrast to them in the centre of this array a catalytic metal ion is coordinated. In the HNH motif of NColE7 three His sidechains coordinate a Zn^{2+}-ion.

In NColE7, the HNH-motif is the C-terminal part of the 131 residues long sequence (**Fig. 2.9**). This is the catalytic center of the protein, acting in the minor grove of DNA, while two further helices are responsible for the binding to the neighboring major grove of DNA. **Table 2.2** contains the PDB codes of the NColE7 crystal structures published so far. Several of these are point mutants of NColE7 either to inhibit the nuclease activity when cocrystallizing with DNA or to study the role of the mutated residues in metal-ion binding. It is challenging to crystallize reactants of a fast reaction, thus there is no available crystal structure of the active complex, *ie.* WT-NColE7 in complex with Zn^{2+} and DNA, which makes it difficult to understand the chemical mechanism of catalysis.
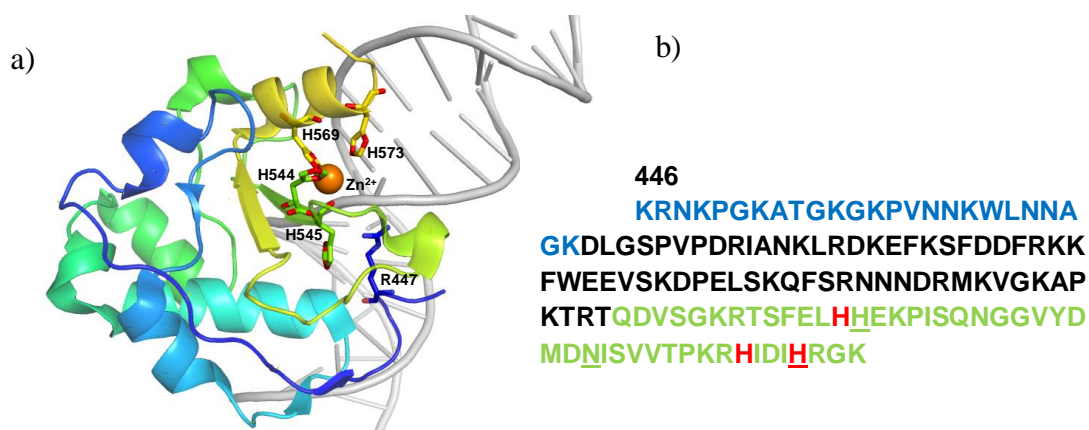
Fig. 2.9. a) *Crystal structure of NColE7 in complex with DNA (PDB: 3FBD). The orange sphere indicates the Zn^{2+}-ion, which was not included in this structure, and it was placed as a result of the alignment with the 1M08 structure (chain A). PyMol's rainbow coloring scheme was applied resulting in a color varied from blue (N-terminus) to the yellow (C-terminus).* **b)** *Sequence of NColE7, which is the C-terminal domain of the colicin E7 protein (446-576) The N-terminal loop is written in blue and the HNH motif in green with the exception of the metal binding residues in red. The underlined residues are the conserved HNH residues.*

A remarkable feature of NColE7, that the free molecule as well as the one bound to metal-ion, DNA or Im7 has the same overall structure with only minor differences in the conformation of sidechains and in the thermic factors. The commonly unresolved parts in crystal structures are the termini of the protein and the loop between the β-strands of the HNH motif. For example, the pairwise alignments for C_α atoms of 1M08, 7CEI and 1PT3 result an RMSD (root mean square deviation) value lower than 0.6 Å.

Table 2.2. *List of the published NColE7 crystal structures* [43]. *All contained a full length NColE7 and the differences in the sequences in pdb files arise from the structure determination. If the terminus is too flexible its position is not determined.*

PDB code / ref.	Mutation	Complex	Sequence in pdb file	Reason for inactivity
1M08 [50]	K446M	protein-Zn^{2+}-PO_4^{3-}	446 MRNK-HRGK 576	-
1MZ8 [51]	-	protein-Zn^{2+}-PO_4^{3-}-Im7	447 RNKP-IDIH 573	-
1PT3 [46]	-	protein-8bpDNA	449 KPGK-HRGK 576	no metal ion
1ZNS [52]	K443M/H545E	proten-Zn^{2+}-12bpDNA	450 PGKA-DIHR 574	mutation
1ZNV [52]	K443M/H545E	protein-Ni^{2+}-PO_4^{3-}-Im7	450 PGKA-HRGK 576	-
7CEI [47]	-	protein-Zn^{2+}-Im7	447 RNKP-IDIH 573	-
2IVH [53]	H545Q	protein-Zn^{2+}-18bpDNA	449 KPGK-IDIH 573	mutation
2JAZ [54]	N560D	protein-Zn^{2+}-PO_4^{3-}-Im7	450 PGKA-HRGK 576	-
2JB0 [54]	H573A	protein-Zn^{2+}-Im7	449 KPGK-HIDI 572	-
2JBG [54]	N560A	protein-Zn^{2+}-SO_4^{2-}-Im7	448 NKPG-HRGK 576	-
3GJN [55]	H545A	protein-Zn^{2+}-Im7(mut)	450 PGKA-HRGK 576	-
3GKL [55]	H545A	protein-Zn^{2+}-Im7(mut)	450 PGKA-HRGK 576	-
3FBD [48]	D493Q	protein-18bpDNA	445 SKRN-HRGK 576	no metal ion

2.4.3. DNA-binding

The shortest DNA sequence that NColE7 is able to cleave is 8 base pairs (bp) long [46]. Oppositely, several other nucleases like *Serratia* nuclease or DNaseI perform also the cleavage of dinucleotides. The reason is that in NColE7 the DNA-binding helix is distant from the catalytic centre. This feature makes it possible that the Im7 protein can inhibitt the catalysis by hindering the DNA-binding.

NColE7 has a low sequence preference [53]. The probability of cleavage at the 3'O-site of different bases is T (41,2%) > A (29,5%) > C (17,1%) > G (12,2%). The reason is probably that the A-T interaction is weaker than the G-C, thus the deformation of the DNA chain around the A-T pairs is easier. The crystal structure of NColE7 was compared with a preferred (1PT3, GCG-ATCGC or 1ZNS, CGGG-ATATCCCG) and unpreferred (2IVH, GGAATTCGAT-CGAATTCC) binding site. It was shown that the scissile phosphate is closer to the metal ion in the preferred complex: the distance is 3.18 Å and 3.72 Å in the preferred 1ZNS [52] and unpreferred 2IVH [53] complex, respectively. The mechanism of nonspecific DNA-protein interaction of NColE7 was studied in detail by molecular dynamics simulations and Brownian dynamics [56].

Binding to the major grove with the helices and in the minor grove with the HNH-motif forces DNA to bend in an angle, preparing it for the reaction. The bending angle measured depends on the length of the substrate cocrystallized with DNA: 7° (8 bp) [46], 19° (12 bp) [52], and 54° (18 bp) [53]. This is caracteristic for nonspecific nucleases, while sequence-specific restriction endonucleases induce more severe bending.

The comparison of the crystal structures of three nucleases including NColE7, VvnI and IPpoI in complex with DNA suggested that the enzymes with a ββα-type active centre cleave DNA with a similar mechanism [46]. The difference in the structure of ββα motif was only minor and they are similarly oriented in the minor grove of DNA. The roll angle between the base steps at cleavage site is around 10° in each case and the minor grove of substrate DNA has a width of 9 Å (the optimal B-DNA form has a minor grove width of 5.6 Å).

2.4.4. Metal-dependent nuclease activity

NColE7 has only one metal-binding site formed by three His sidechains in the HNH motif: H544, H569 and H573. The metal ion is Zn^{2+} under physiological conditions [52], in a tetrahedral coordination sphere. Based on the quantum chemical features of the active site it was predicted already before the publication of crystal structures with DNA that the Zn^{2+}-ion in NColE7 plays a catalytic role and its coordination sphere is flexible enough to accomodate an extra ligand during the reaction [6]. The fourth coordination site can be occupied by a water molecule, as seen in the 7CEI structure [47]. Similar Zn^{2+}-binding site was found in the carbonic anhydrase, β-lactamase and carboxypeptidase enzymes [47]. The water molecule can be replaced by the phosphodiester group of DNA substrate [46, 48, 52, 53], a phosphate or sulfate ion [50, 51, 54]. These anions coordinate in the position of the scissile phosphate of

the DNA. The structure of the metal binding site is maintained even in the absence of metal ion according to crystal structures. However, the sidechain of H573 has a slightly different conformation in the presence or absence of the metal ion. It was shown, that the Zn^{2+}-ion is essential for the hydrolitic activity, but is not required for DNA-binding [57] and that it can be removed by EDTA.

NColE7 was shown to be active also in presence of other metal ions. The nuclease activity measured with a supercoiled plasmid DNA decreased in the order $Ni^{2+} > Mn^{2+} \sim Mg^{2+} > Zn^{2+}$ [52]. However, the activity with Mg^{2+} is debated, since no Mg^{2+}-binding site was found in the protein [51]. There is only one published crystal structure with other ion than Zn^{2+} (1ZNV with Ni^{2+} and Im7 [52]) in which NColE7 binds a Ni^{2+}-ion with a trigonal bipyramidal coordination sphere [52]. Three sites are occupied by the His imidazole nitrogens, one by the oxygen of a phosphate ion and one by a water molecule. The other protein in the unit cell has a tetrahedral binding site (3His and one phosphate oxygen), similarly to Zn^{2+}-bound structures.
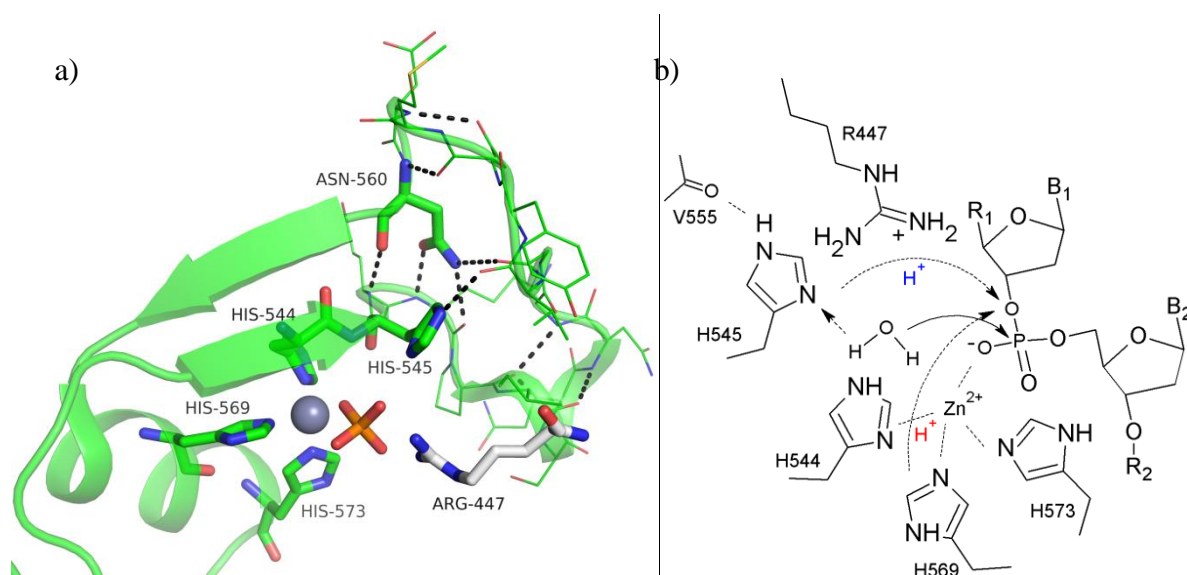
Fig. 2.10. a) *Structure of the HNH-motif in NColE7 (1M08, [50]). The conserved HNH residues (H545, N560, H573) and the metal-binding residues (H544, H569, H573) are shown in sticks. R447 is shown in white sticks.* ***b)*** *Summary of proposed mechanistic steps of the DNA hydrolysis by NColE7. The red [47] and blue [58] protonation ways are suggested approaches to explain the protonation of the leaving group.*

The mechanism of the reaction has been analyzed (**Fig. 2.10**), however it is not completely understood yet. Two of the conserved HNH residues were reported to have a direct role in catalysis. H545 is supposed to act as a general base to activate a water molecule that can perform the nucleophilic attack at the scissile phosphate group. The mutation of H545 leads to the decrease (H545Q, 11%) or cancellation (H545A, H545E) of nuclease activity [54]. The distance of H545-Nδ1 atom from the phosphate oxygen is around 2.6 Å in an appropriate geometry [51]. The second HNH residue, N560 is located in the loop between the β-sheets of the HNH motif (**Fig. 2.10 a**). This residue being far from the active site plays a structural role orienting the H545 side-chain. H545 is also in hydrogen bonding interaction

with the carbonyl group of V555 [51]. The role of N560 was studied through point mutations. The nuclease activity decreased in the following order [54]:

$$N560\,(100\%) > N560H\,(27\%) > N560A\,(24\%) > N560D\,(7\%)$$

This was in good correlation with the distance between the Nδ1-atom of H545 and the phosphate oxygen [54]:

$$N560\,(2.4\text{Å}) < N560A\,(2.9\text{Å}) < N560D\,(3.72\text{ Å})$$

This means that the decrease of the nuclease activity is related to the displacement of DNA. The leaving group of the reaction, which is the 3'-alcoholate ion has to get protonated. In other nucleases, such as I-PpoI or Serratia nuclease the proton is provided by a metal (Mg^{2+}) bound water molecule. Similarly, in NColE9 (a nuclease with 70% sequence identity to NColE7) the Ni^{2+}-bound water molecule can play the same role [59, 60]. According to crystal structures of NColE7 the tetrahedral Zn^{2+}-ion does not coordinate water molecule in the presence of the substrate, therefore the protonation of the leaving group is still a question to answer. It was suggested, that the Zn^{2+}-ion can coordinate a water molecule as the fifth ligand during reaction, or an additional Mg^{2+}-ion can promote the reaction [51], but this was not proven. R538, Q542 and H569 were also speculated to provide the proton for the leaving group in NColE7 in analogy to NColE9 [47, 61]. Recently, a shuttle mechanism was suggested in which the leaving group is protonated by the hydrogen ion originating from the same water molecule that initiated the nucleophilic attack [60].

2.4.5. The role of the N-terminus

The special structural feature of NColE7 that attracted our attention is that the N-terminal R447 and the C-terminal catalytic HNH motif are close in space (**Fig. 2.11**). The R447A mutant had only 15% of the WT nuclease activity [62], and R447 was proposed to increase the DNA binding affinity. Previously we showed that the cytotoxicity of NColE7 is completely lost upon deletion of the KRNK sequence (residues 446-449) at the N-terminus [43]. We hypothesized that the N-terminus could be used to design a control mechanism: in its absence there is no nuclease activity, but in its presence that could be restored.
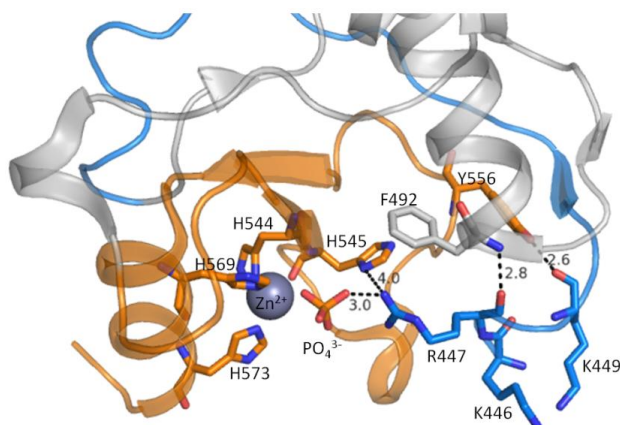
Fig. 2.11. *Crystal structure of NColE7 (PDB: 1MZ8 [51]) in complex with a Zn^{2+} and a phosphate ion. The HNH motif is in orange and the N-terminal loop is in blue. Among the N-terminal amino acids, R447 is the closest residue to the phosphate ion that forms a bridge between the Zn^{2+}-ion and R447. Hydrogen bonds of the backbone of the N-terminus are also shown* [63].

It is intriguing why R447 − lying outside the DNA-binding helices − influences the catalytic reaction. In NColE9 an arginine (corresponding to R447 in NColE7) was supposed to bind the substrate and stabilize the pentavalent transition state [60, 64]. Based on the crystal structures of Vvn − an another HNH nuclease − it was also shown that the proximal arginine side-chain binds and stabilizes the cleaved DNA to decelerate the reverse reaction [2, 65]. We have studied the published crystal structures of various HNH nucleases, and found positively charged residues in a similar orientation [43] e.g. in Vvn (PDB ID 1OUP), and Sm endonucleases (PDB ID 1G8T) and Nuclease *A* (PDB ID 1ZM8).

2.4.6. *Dimerization of NColE7*

In the FokI-based zinc-finger nucleases the dimerization of the nuclease domain is essential for the double strand break in DNA. Since we would like to substitute the FokI nuclease domain with NColE7, it is important to understand the dimerization features of NColE7.

The structures of the NColE7/DNA complexes suggest a mechanism of a single strand cut by the monomer NColE7. The one-metal mechanism is in agreement with this. However, WT NColE7 was crystallized with Zn^{2+}-ion [50] and the unit cell contained 3 molecules out of which 2 seemed to form a homodimer. The HNH motifs are facing each other, and the C-terminal K576 approaches the phosphate ion coordinated to the Zn^{2+}-ion of the other molecule by 4.6 Å. The N-termini do not interact. The distance between the two Zn^{2+}-atoms (14.7 Å) is similar to the distance between the Mg^{2+}-ions in the IPpoI-dimer (18.7 Å). The buried interface between them is 559 $Å^2$, with four hydrogen bonds and three hydrophobic residues (P548, I570, I572). The metal coordinating histidines are located in the interface. Usually the hydrophobic residues dominate the dimerization interface in solution. The ratio of apolar and polar atoms is only 48%-52%. Therefore, it is difficult to decide whether the molecules form a dimer or this is only result of crystal packing. Gel filtration experiments in solution studies gave contradictious results [50]. In experiments with low salt concentration (150 mM NaCl) and absence of transition metal ions monomers, while with high salt concentration (0.7-1.5 M ammonium-acetate) and 10 mM Zn^{2+} dimers were detected. Cross-linking experiments showed that the presence of substrate can induce the dimer formation as measured with a 27 bp DNA.

2.4.7. *NColE7 as a possible building block of a ZFN*

The choice of NColE7 for the design of a novel type ZFN is advantageous, because it meets the first criteria for building a modular ZFN, in that it is nonspecific. Furthermore, the selection of a bacterial enzyme for engineering is appropriate, since the expression is the most feasible in bacterial cells. As described in the above chapters, NColE7 has advantageous structural features in that its C-terminal metal-binding site (HNH-motif) cooperates with the N-terminus during the reaction. This could be exploited in the design of an allosteric control.

3. AIMS AND OBJECTIVES

The aim of this research is to design new ZFN molecules with a novel type of control mechanism as a proof of concept. A safe control mechanism could open a new perspective in nuclease design for gene therapy to cure monogenetic diseases. The building blocks for the novel ZFN were selected as follows: a zinc-finger protein (PDB ID: 1MEY [12], in complex with its specific DNA substrate) was intended to be used as a specific DNA-binding unit and the nonspecific NColE7 was selected to form the catalytic domain. NColE7 has not yet been applied as a part of a chimeric ZFN. Therefore, in my work first I aimed at understanding and influencing the function of the NColE7 protein, to be followed by the redesign of the protein for construction of the new ZFN molecule. The main idea is shown in **Fig. 3.1 a**. According to this scheme NColE7 would be divided into two parts, each fused to opposite termini of the ZF array. This arrangement can provide a control mechanism if none of the NColE7-parts has a nuclease activity on its own, but they function if the ZF-s orient them to get sterically close.
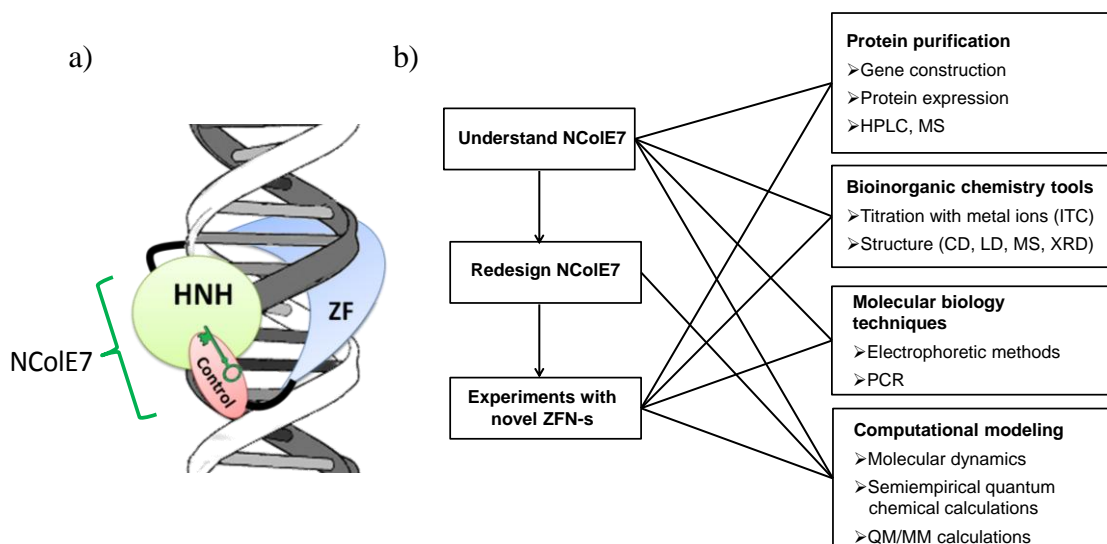
Fig. 3.1. a) *Scheme of the chimeric enzyme that we intend to design. The zinc-fingers (in blue) bind specifically to DNA, and the HNH-motif containing part of NColE7 (green) is responsible for the cleavage. However, it can only hydrolize DNA, if the N-terminal part that is fused to the other end of the ZF-s precisely fits to induce the catalytic activity (red).* ***b)*** *Workflow of this research project.*

According to the scheme shown in **Fig. 3.1 b** I applied a multidisciplinar approach for the main objectives:

1. **Understanding the properties of the NColE7 protein through point mutations with emphasis on the N-terminus and metal-binding properties**

NColE7 was extensively studied in the literature, but the potential control function of the N-terminus has not been investigated so far. Since NColE7 is a metalloprotein, it forms ternary (three component) complex with the substrate DNA. The metal-binding properties may change upon mutations altering the nuclease activity leading to a complex bioinorganic chemical problem. Since the N-terminal four residues (446-KRNK-449) proved to be

essential for the nuclease activity, we can get a deeper insight in the role of these residues through their mutations. The questions addressed in this point include:

- How much of the three positively charged residues in the N-terminal section is needed for the nuclease activity?
- Is the nuclease activity abolished when all the positive residues are missing but the backbone is present?
- Can K446 or K449 promote the reaction in the absence of R447?

2. Finding the most suitable way to control the NColE7 nuclease

To explore the use of the N-terminus to control NColE7 we need to be able to influence its function. This objective involves computational study of the contribution of each residue to the structural stabilization of the proteins and selection of the mutations to be performed experimentally. The questions to answer were:

- What is the role of the 25 residues long loop at the N-terminus? Is it responsible for keeping the KRNK motif in place?
- How does the mutation of selected residues in this seemingly not important part of the sequence influence the nuclease activity?
- How can we induce the reaction if the nuclease activity is decreased upon the mutations weakening the interactions of the loop?

3. Design of an NColE7-based ZFN by computational methods

The next goal is to design a ZF-NColE7 artificial nuclease, based on the results of the previous part of the research. Starting with published crystal structures of NColE7 with DNA and the ZF with DNA the computational design includes the following points:

- Construction of an initial model by approaching the NColE7 and the selected ZF protein simultaneously binding to a DNA substrate.
- Selection of the controlling part from the N-terminus and the catalytic part from the C-terminus.
- Design of short linker sequences appropriate to fuse the proteins.
- Study of the constructed models by MD simulations.

4. Experimental study of the designed ZFN-s

The final goal is to express and purify the computationally designed ZFN-s as recombinant proteins. The folding, metal-binding and reactivity can be significantly affected within the redesigned nuclease domain. Therefore, the new enzymes have to be studied in detail:

- Do the ZF-s maintain the sequence-specificity after the fusion of the NColE7 segments?
- Is there any observable nuclease activity?
- Is the model good enough to serve as a starting point for optimization to obtain a nuclease that is applicable in practice?

4. METHODS

Due to the complexity of the problem addressed in this work, I applied experimental techniques involving the tools of experimental and computational bioinorganic chemistry, biochemistry and molecular biology. Therefore, I will give a short introduction to these methods, focused on the specific use in this project. The detailed protocols of my experiments are listed in the appendix.

4.1. PRODUCTION OF RECOMBINANT PROTEINS

The modification of proteins, including mutation, truncation, fusion to other proteins is often required in the research. The proteins obtained from the expression of DNA (engineered by molecular cloning) in living cells are called recombinant proteins. In this work they were expressed in *Escherichia coli* (*E.coli*) cells, involving the following steps:

1) Construction of the gene of the mutant protein
2) Protein expression from the gene in cells
3) Protein purification from the cell extract
4) Validation of the obtained protein

4.1.1. Gene cloning

4.1.1.1. Introduction of mutations to the gene by PCR

If the gene of the WT protein is available, mutations can be introduced to it by polymerase chain reaction (PCR). The reaction mixture contains the following components:

- template DNA molecules
- a heat stable polymerase enzyme in a Mg^{2+}-containing buffer, e.g. *Taq* polymerase
- primers: short oligonucleotides that hybridize to the single stranded template and aid the polymerase to start the building of the complementary strand
- deoxynucleoside-triphosphates as building blocks.

The reaction is driven by heat control (**Fig. 4.1**): first it is heated to > 90 °C, where the template DNA dissociates to single strands, this is called melting. Then the temperature is lowered to the melting point of the primers (~50-65 °C) so that they can hybridize to the single stranded template (annealing). Finally the temperature is increased to the optimum for the polymerase enzyme function, which is 72 °C in case of the *Taq* polymerase, and the elongation of the primers results in the new double stranded molecule. With careful design of the primers it is possible to introduce mutations and restriction enzyme cleavage sites into the original sequence.
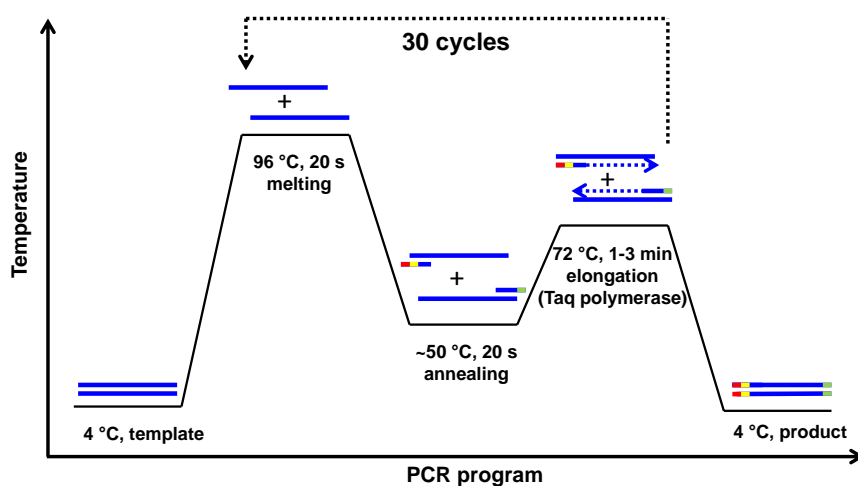
***Fig. 4.1.** The temperature program of a PCR used to introduce mutation to the end of the sequence and scheme of the reaction. Each blue lines indicate a strand of the template DNA. The red and green endings are restriction enzyme cleavage sites, carried by the primer molecules. The mutation introduced by primers is in yellow. After around 30 cycles the efficiency of the enzyme would decrease.*

The new double stranded molecules formed in the elongation step serves as templates in the next cycles. Thus the amount of DNA is exponentially increasing ($\sim 2^n$ where n is the number of cycles), and each new strand contains the modifications introduced by the primers. It is important to use a low amount of template, because that may remain in the solution as an impurity. If the required mutation site is located far from the end of a sequence, a slightly different strategy can be used. The sequence can be divided into two parts that overlap around the mutation site. They can be amplified in PCR separately, and then annealed together to give the whole sequence. The PCR product encodes the mutant protein, but it is possible to use it for in cell protein expression only after it was built into a suitable carrier DNA, so called plasmid.

4.1.1.2. *Processing of the gene and carrier by restriction endonucleases*

Plasmids are circular DNA molecules of bacterial origin, consisting of a few thousands of base pairs. They are widely applied in molecular biology. The gene of the recombinant protein can be inserted in the so called cloning region of the plasmid. In this region there are special sequence elements recognized and cleaved by restriction endonucleases (**Fig. 4.2 a**). The nucleases used here cut DNA leaving sticky ends, *i.e.* short single stranded leftovers at the cleavage site. The termini cut by the same enzyme match specifically. The same restriction sites can be built by the primers into the gene encoding the target proteins and thus the gene can be inserted into the plasmid via the matching ends.
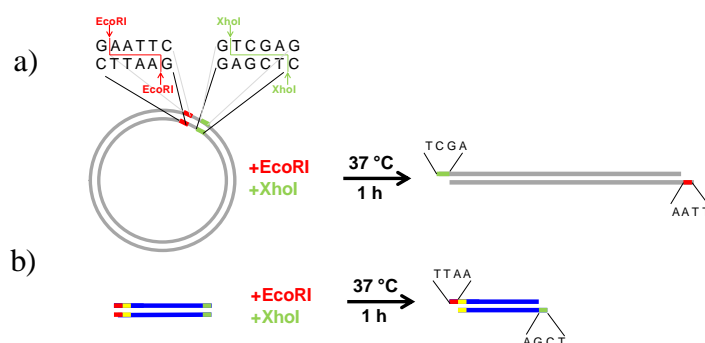
***Fig. 4.2. a)** Cleavage of the plasmid (grey) and **b)** the mutant protein gene (blue) by two restriction endonucleases, EcoRI and XhoI. These enzymes are commercially available.*

4.1.1.3. Ligation of the gene into the carrier plasmid

The enzyme catalyzing the covalent bond formation at the matching DNA termini is called ligase. The reaction results in a circular plasmid that contains the gene of the target protein inserted within the selected restriction sites (**Fig. 4.3**). There are several side reactions that can occur including linear DNA molecules that are either not ligated or that are ligated on one side only. Furthermore, two or more insert molecules can be ligated to each other and the plasmids can also be joined. Circular plasmids carrying 3 or 5 substrates are also commonly observed. To minimize these reactions, the plasmid/insert ratio in the mixture is essential, usually 1/6 is optimal. The ideal DNA/ligase ratio is also important for the effective reaction.
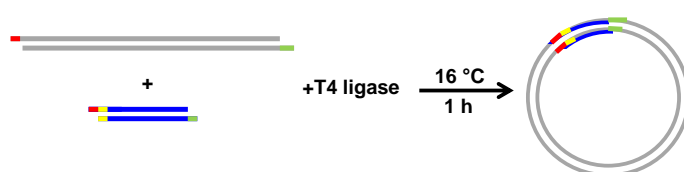
Fig. 4.3. *Scheme of the ligation reaction.*

4.1.1.4. Transformation of the DNA-expressing bacterial cells

The ligation mixture is then mixed with a suspension of competent bacterial cells suitable to copy the plasmid in high number, e.g. *E.coli* DH10B or MachI cells. The suspension is subjected to a heat-shock treatment. With an antibiotic resistance selection (the plasmid itself carries a resistance gene) it can be assured that only those bacteria are able to survive that had taken up a plasmid DNA. After cultivation of the selected colonies the success of the insertion of the target gene into the plasmid needs to be checked by agarose gel electrophoresis and DNA sequencing. The plasmid DNA can be extracted by alkalic lysis of the bacteria, followed by a column purification and ethanol precipitation.

4.1.2. Protein expression

There are several available modified plasmids that are suitable for recombinant protein expression, e.g. pGEX-6P-1 is shown in **Fig. 4.4**.
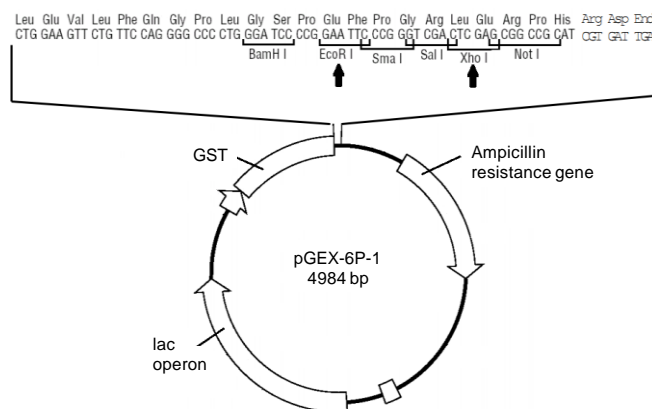
Fig. 4.4. *The scheme of pGEX-6P-1 plasmid. The cloning site is magnified. The ampicillin resistance gene aids to control experiments: if the culture media are supplied with ampicillin, only the cells that have the resistance gene, ie. contain the plasmid can survive.*

The protein expression from the gene inserted into the pGEX-6P-1 plasmid is regulated by the *lac* (lactose) operon. If the operon is active, the protein expression will occur from the inserted gene. The *lac* operon is required for lactose transportation and metabolism in *E.coli* cells. In the presence of glucose and absence of lactose a repressor protein binds and inactivates the operon. In molecular biology experiments instead of lactose an analogue, the isopropyl β-D-1-thiogalactopyranoside (IPTG) is used to induce protein expression. IPTG is usually added to the culture in the *log* phase when the OD_{600} value is ~0.6. After addition to the bacterial culture it binds to the repressor thus activating the operon, similarly to lactose. However, it is not a substrate of the β-galactosidase, so it will not be degraded and the expression will not be downregulated. This allows for the "overexpression" of the target protein. A common choice for this purpose is the *E.coli* BL21(DE3) strain.

The pGEX-6P-1 plasmid contains the gene of the *glutathion S-transferase* (GST) enzyme between the promoter of the protein expression and the cloning region, without a stop codon. Thus the result of protein expression will be a GST-fused target protein, consisting of one continuous polypeptide sequence: the N-terminal domain is the GST and the C-terminal is the target protein. There is a flexible linker between them, thus GST usually does not interfere with the folding and function of the target protein. The advantage to use the GST-tag is that it allows for affinity based purification, resulting in a high purity product. After purification the GST-protein can be cleaved with a specific protease, the *Human rhinovirus C3 protease* [66] *(PreScission protease, GE Healthcare)*. A short sequence depending on restriction site location will remain at the N-terminus of the protein. Another example of plasmids is pET21a, that fuses a C-terminal $(His)_6$ sequence to the protein, if the gene does not contain a STOP codon before the multihistidine containing part. Plasmids can also be used to express proteins without any affinity tags. However, in this case the protein purification is more challenging.

4.1.3. *Fast protein liquid chromatography*

Fast protein liquid chromatography (FPLC) is a high performance liquid chromatography (HPLC) method developed for the purification of biological samples [67]. It is characterized by high loading capacity, biocompatible aqueous buffers, fast flow rates and wide range of stationary phases, such as affinity, gel filtration, ion exchange [68]. The system can be automatized including autosampler, gradient program control, peak collection. ÄKTA FPLC explorer system (**Fig. 4.5** Amersham Pharmacia Biotech, Sweden) is recently used frequently to purify proteins and other biological samples.
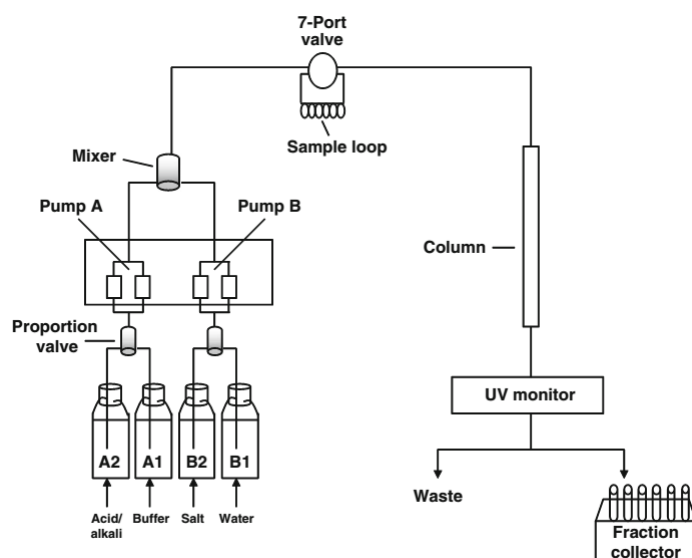
Fig. 4.5 *The scheme of the ÄKTA system used for ion exchange. Taken from* [68].

In this work affinity, cation exchange cromatography and gel filtration have been applied to purify the nucleases expressed in BL21(DE3) cells. After cultivation of the cells and protein expression, the culture was centrifuged and resuspended in an appropriate buffer. The cells were disintegrated by sonication followed by one more centrifugation step. The proteins were purified from the supernatant ("raw extract").

4.1.3.1. GST-affinity chromatography

In affinity chromatography the protein is separated based on specific, reversible interactions to a ligand that is coupled to a chromatographic matrix. The advantage of this method is the very high specificity and related to this the high sample loading capacity. After equilibration of the column, the sample is loaded and the target protein binds to the column, while all other impurities of the cell extract flow through the column, as indicated by high absorbance values on the chromatogram (**Fig. 4.6**). After washing the column for further 1-2 column volumes (CV) the sample can be eluted with an agent of higher affinity to the protein as compared to the ligand on the column.
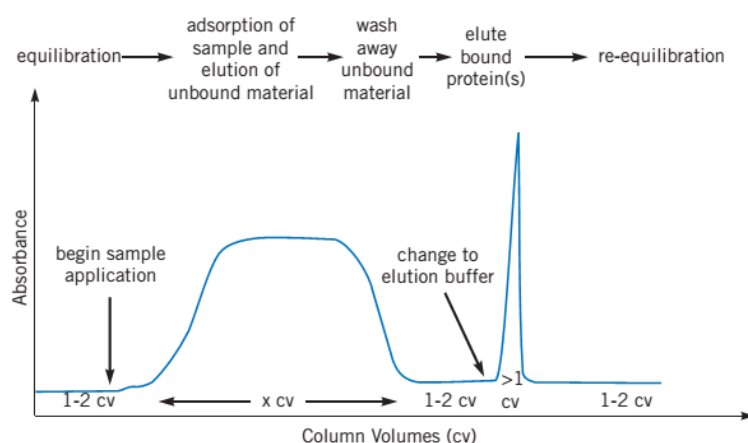
Fig. 4.6 *A typical chromatogram of affinity purification (from Handbooks of Amersham Biosciences)*

The specific interactions most frequently used in affinity chromatography include:

- enzyme – substrate analogue, e.g. glutathione-S-transferase (GST) – glutathione
- antibody – antigen
- metal ions – poly His-sequence.

The ligand is bound to a chemically inert matrix, such as sepharose (bead-form agarose), involving a linker.

If GST-affinity chromatography is intended to be used for protein purification, the GST protein is fused to the target protein: the gene of the recombinant protein is cloned to a site containing the gene of GST, and the two proteins are expressed as a continuous sequence. The column contains glutathione immobilized to a sepharose matrix (**Fig. 4.7**). The protein can be eluted with reduced glutathione solution.
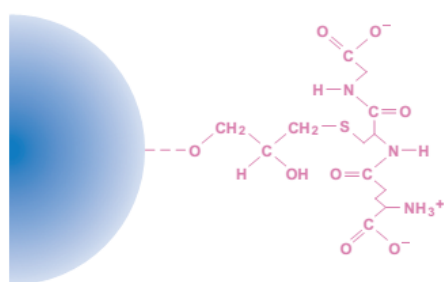
Fig. 4.7 *The glutathione linked to the sepharose matrix (from Handbooks of Amersham Biosciences).*

4.1.3.2. Cation exchange chromatography

DNA-binding proteins, such as the zinc-fingers and the nuclease enzymes are positively charged, complementing the negative charge of the nucleic acid. Therefore, they can be purified in cation exchange chromatography, in that positively charged solute molecules adsorb to the negatively charged groups immobilized to the matrix. Such cation exchangers are for example carboxymethyl groups ("*CM*", $-O-CH_2-COO^-$), sulfopropyl groups ("*SP*", $-O-CH_2-CHOH-CH_2-O-CH_2-CH_2-CH_2SO_3^-$) or methyl sulfonate groups ("*S*", $-O-CH_2-CHOH-CH_2-O-CH_2-CHOH-CH_2SO_3^-$).

In the purification process, first the system is equilibrated with the required buffer, the sample is applied to the column (adsorption) and an increasing ion gradient and/or pH-change is used for the desorption of the molecules in the order of their binding affinity to the column. It is important to optimize the gradient elution in order to get a high resolution. It has to be noted, that ion exchange chromatography is not a specific method and a well resolved peak of the chromatogram can still contain the mixture of proteins.

In this work, a Sepharose SP Fast Flow column was used. In the Fast Flow ion exchanger columns 90 μm agarose beads serve as a matrix, that is highly crosslinked (6%) providing high physical and chemical stability and can be used with high flow rates in a wide pH-range.

4.1.3.3. Gel filtration

In gel filtration molecules are separated based on different rate of movement in a gel matrix due to different size. Unlike the previously introduced techniques, in this case the molecules do not bind to the stationary phase. Thus the buffer usually has no significant effect on the resolution and a wide range of conditions can be applied. It is possible to perform the separation in the presence of cofactors or denaturing agents and at different temperatures. In order to achieve a high resolution a long column is used and the sample is injected in a high concentration in low volume. The separation is finished within one CV (**Fig 4.8**) and first the higher molecular weigth molecules are eluted, because they can not enter the small pores of the matrix and therefore go through a shorter way.
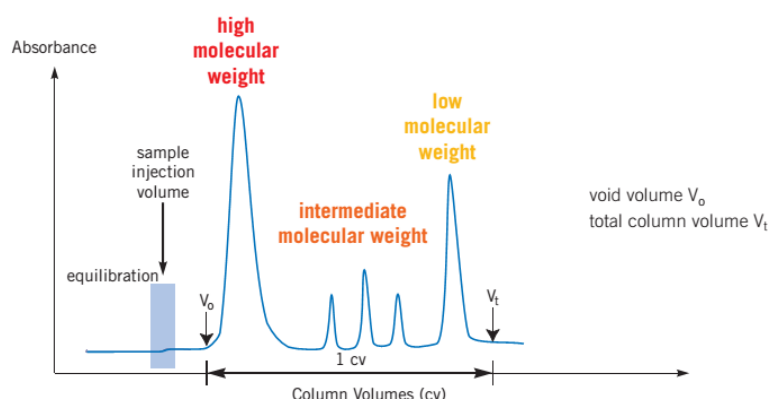
Fig 4.8 *A general gel filtration chromatogram (from Handbooks of Amersham Biosciences).*

4.2. CHARACTERIZATION OF METALLOPROTEINS

4.2.1. Detection of proteins in SDS-PAGE

The polyacrylamide gel electrophoresis with sodium dodecyl sulfate (SDS-PAGE) experiment is the easiest and fastest way to detect the protein content and purity of the sample, based on the electrophoretic mobility. The gel can be easily prepared with the radical polymerization of acrylamide, with the addition of N,N-methylenebisacrylamide to introduce crosslinks to the polymer. The reaction is initiated with ammonium-persulfate and controlled with N,N,N',N'-tetramethylethylenediamine. The electrophoresis is carried out in a vertical direction. The gel has two layers: a short stacking gel (e.g. 6% acrylamide, pH = 6.8) and a long resolving gel (e.g. 12.5% acrylamide, pH = 8.8).

The sample is prepared with the addition of a buffer (pH = 8.5), SDS and mercaptoethanol (to reduce Cys bridges). During the incubation at high temperature (95 °C)

protein is denatured. SDS binds to the hydrophobic regions, and the chain obtains a negative charge proportional to its length.

During the electrophoresis the proteins first enter the stacking gel where highly concentrated narrow bands are formed. The proteins can be stained in the gel after electrophoresis with the Comassie Brilliant Blue dye, an anionic triphenylemethane dye that nonspecifically binds proteins. Compared to sample standards the protein size (length of sequence) and concentration can be estimated.

4.2.2. Mass spectrometry of metalloproteins

Mass spectrometry is a method to determine precise molecular weights. In the mass spectrometer the sample is first ionized and transferred to gas phase. Then the charged molecules are accelerated in the mass analyzer where they fly with a velocity that depends on their m/z value, and finally they are detected.

In order to analyze biological macromolecules, the so called soft ionization techniques are used including ESI (electrospray ionization) and MALDI (matrix-assisted laser adsorption) techniques. There have been a number of technical improvements of the ESI technique providing facility for the study of sensitive macromolecular complex systems. In the conventional ESI the analyte solution is pumped at a flow rate of a few µl/min. The capillary needle is under a high voltage relative to a counter electrode. Therefore the sample forms an aerosol after getting out of the tip, in the so called Taylor-cone and forms charged droplets. These droplets get through an orifice in the counter electrode, enter the vacuum mass analyzer and undergo fissions and shrinking leading to gas phase ions. In order to decrease the necessary sample amount, nano-ESI has been evolved [69]. The recently developed nano-ESI capillary has a diameter of 1-10 µm and is coated with a conducting material, e.g. gold. 1-5 µl of sample is enough for loading and the capillary is placed 0.5-2 mm from the electrode orifice under camera control (**Fig. 4.9. a**). The flow is 20-50 nL/min. 500-1500 V is applied to induce dispersion. $10^{-5} - 10^{-7}$ M is the typical concentration range used for proteins.
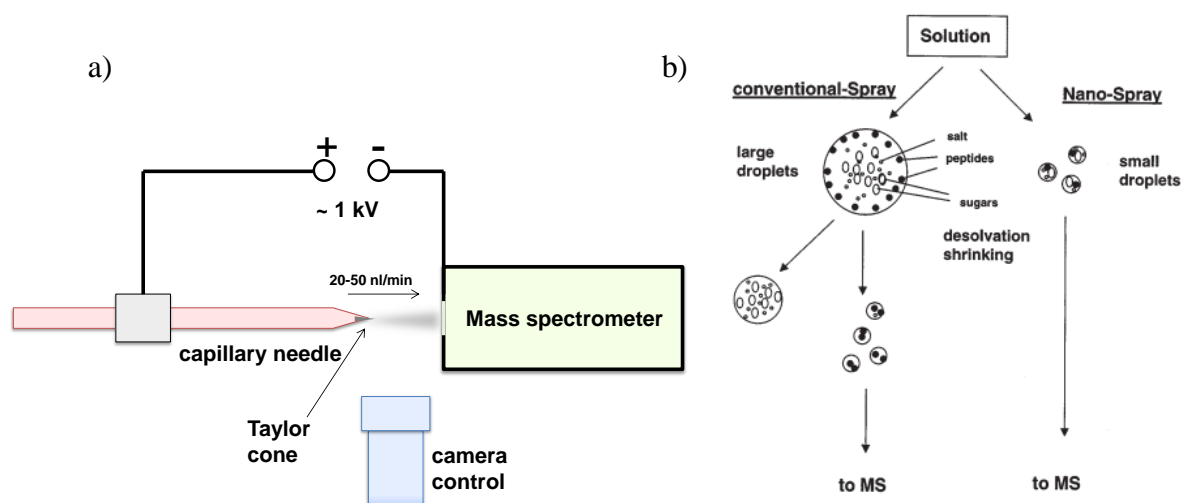
Fig. 4.9. a) *Scheme of the Nano-ESI device.* **b**) *The mechansim of ionization process in ESI and nano-ESI devices, from* [69]

A further advantage of nano-ESI is that instead of the larger droplets in ESI (starting with around 1.5 μm diameter) smaller droplets are immediately formed, with a diameter around 150 nm (**Fig. 4.9. b**). Therefore the fission and shrinkage process has less steps and the upconcetration of the sample during shrinking is less significant. Consequently, nano-ESI MS has an increased salt-tolerance, 10^{-2} M NaCl was reported not to interfere with the measurement [69]. In case of proteins the m/z spectra typically represent multiple charge states. In the most frequently used positive mode of measurements the basic sites of the proteins get protonated, and the molecular weight is $M_{prot}+z{\times}M_H$, where M_{prot} is the molecular weight of the protein at neutral pH and z is the charge of the molecule. The neighboring peaks correspond to molecules differing in one H^+.

$$m_1 = \frac{M + z_1 M_H}{z_1}$$

$$m_2 = \frac{M + z_2 M_H}{z_2}$$

$$z_2 = z_1 + 1$$
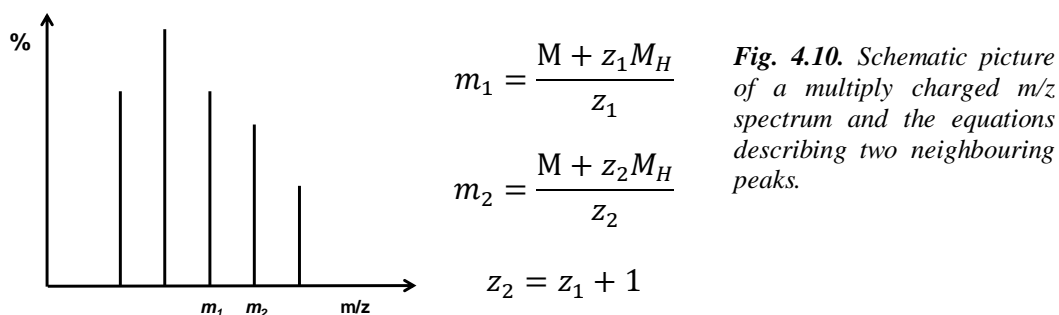
Fig. 4.10. Schematic picture of a multiply charged m/z spectrum and the equations describing two neighbouring peaks.

The three equations shown in **Fig. 4.10** contain three unknowns (z_1, z_2, M), thus the molecular weight can be determined. This can be repeated for all peaks and the averaged results give a precise solution for the molecular weight. In practice, a computer algorithm is used to convert the *m/z* spectra into mass spectra. The obtained data are compared with expected molecular weights of the sample. The addition of the average molecular weights of each residue in the sequence is the calculated mass of the protein, and in most cases this will correspond to the expected mass. However, if there are any metal cofactors or disulfide bridges, the expected mass will differ from the calculated mass by the atom weight of the metal ion minus the number of metal charges or SH groups participating in disulfide bridges. The reason is that the algorithm computing the molecular weight will consider the additional charge of metal ions as hydrogen atoms. The expected mass is then modified as

$$M_{expected} = M_{calculated} + M_{metal} - z_{metal} + 2n_{ss\text{-}bridges}.$$

The first application of MS for metalloproteins was published in 1993 [70], focused on the metal binding of metallothioneins. In the ESI technique it is also possible to distinguish between metal cofactors and other ligands or associated metal ions, by increasing the internal energy of the complexes in the gas phase. The *m/z* spectra can add further structural details as compared to the mass spectra. On the *m/z* spectra the folded protein molecules appear in a narrow charge state range, because only certain groups on their surface can be protonated. As in opposite, unfolded proteins appear at lower m/z values due to higher charge states and in a wide range depending on the completeness of protonation. Therefore, this technique can be applied to study the metal-ion dependent structural features of the proteins [71, 72].

4.2.3. Determination of metal affinity by isothermal titration calorimetry

In the ITC experiment the protein sample is titrated with the metal solution, and the energy associated with the reaction is directly measured. The heat effect (q_i) after one injection can be described as

$$q_i = v \times \Delta H \times \Delta L_i$$

where v is the volume of the cell, ΔH the binding enthalpy and ΔL_i is the increase in the concentration of the bound ligand after the ith injection [73]. During the reaction the energy needed to maintain a constant temperature difference between the sample and reference cell is measured (µcal/s). The heat effect is then calculated as the area under the peak on the energy (µcal/s) vs. time (s) diagram (**Fig. 4.11**).
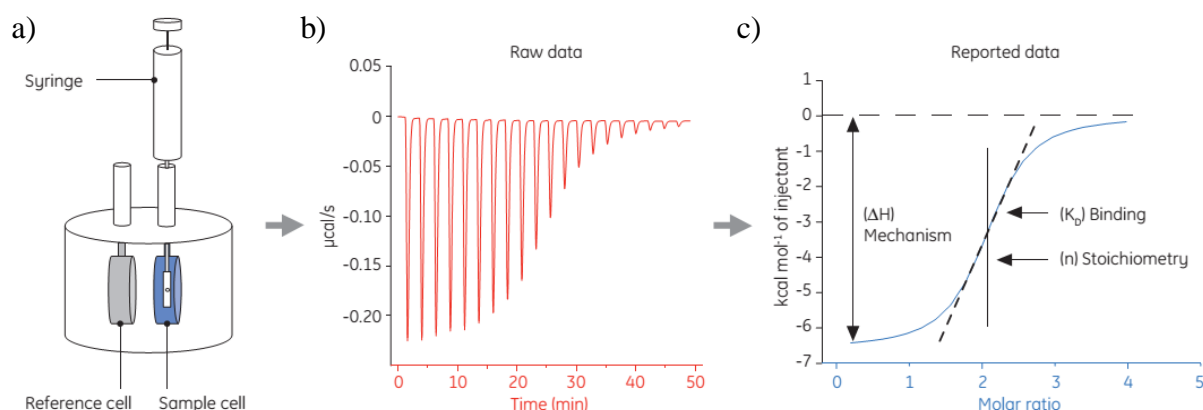
***Fig. 4.11. a)** Schematic picture of the ITC cell. **b)** Data obtained during a titration based on an exothermic reaction. **c)** Calculated titration curve and data analysis based on the raw ITC data. The figures are adopted from the official description of the GE Healthcare MicroCal iTC$_{200}$ instrument (http://biophysicscore.bsd.uchicago.edu/calorimetry.php).*

Since the free protein concentration decreases, the heat effect also decreases during the titration. The dilution heat has to be subtracted from data before analysis. It is important to avoid any other heat effects, e.g. metal ion – buffer interaction, therefore the metal salt has to be dissolved in the dialysis buffer of the protein. In the above equation ΔL_i can be expressed in terms of K_a (association equilibrium constant) and n (stoichiometry) for different binding models including one or more independent or cooperative binding sites and finally K_d can be calculated. Based on the ITC experiment several thermodynamic data can be obtained: ΔH, $\Delta G = -RTlnK_a$ and ΔS can be calculated based on $\Delta G = \Delta H\text{-}T\Delta S$ [73]. The heat capacity can also be determined by repeating the titrations at different temperatures.

High affinity complexes are a limitation for the ITC method, because the curvature of the titration is too steep **Fig. 4.12**.

It was shown, that the constant $c = K_a \times [P]$ product has to be lower than 1000 in order to the optimal interpretation of data [74]. This means that an association constant of maximum $10^8 – 10^9$ M^{-1} can be accurately determined. Complexes of higher stability can only be studied by competition assays.
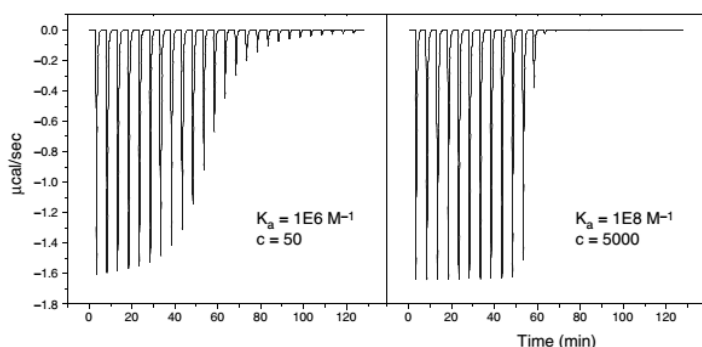
Fig. 4.12 *The effect of c = $K_a \times [P]$ on the shape of the titration curve, from [73].*

The metal binding of proteins is frequently coupled with deprotonation processes. This reaction depends on pH and the ionization enthalpy of the buffer and therefore repeating the measurement in different buffers can allow for the calculation of the number of protons participating in the coupled reaction: $\Delta H_{measured} = \Delta H_{binding} + n_H \times \Delta H_{ionozation}$ [74].

4.2.4. Study of protein solution structure by circular dichroism spectroscopy

The available methods for atomic structure determination of biomolecules (X-Ray crystallography, NMR) require high amount of sample and can provide data only in certain conditions, like single crystals or highly concentrated solutions. Circular dichroism (CD) spectroscopy is a supportive technique that is widely available and can be applied in physiologically more relevant experiments to study low resolution macromolecular structure and its changes.

The linearly polarized light can be divided into two, circularly polarized light beams. If these beams are shined on a sample that absorbs both components at the same level, the beams are recombined to the same plane. However, if the sample contains chiral chromophores that absorb the components differently, the light going through the sample will be elliptically polarized (**Fig. 4.13**).
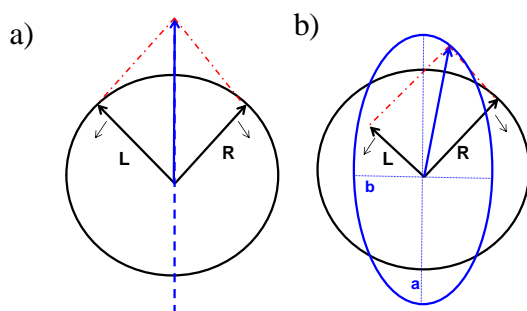
Fig. 4.13. a) *Combination of the left and right polarized light beams of the same intensity results in a linearly polarized light in the original plane.* **b)** *If the left and right polarized light beams are absorbed to different extent, the result is elliptical polarization.*

In the CD spectropolarimeter the absorbance difference $\Delta A = A_L - A_R$ is measured, where A_L and A_R is the absorbance of left and right circularly polarized light, respectively. This is usually reported in ellipticity (θ), given in degrees. By definition, $\tan\theta = b/a$, where b and a are the minor and major axes of the ellipse. The conversion can be performed as $\theta = 32.98 \times \Delta A$ [75]. It is convenient to calculate the molar ellipticity according to the formula

$$[\theta]_{\text{molear},\lambda} = 100 \times \theta_\lambda/c \times d$$

where $[\theta]_{\text{molar},\lambda}$ (deg cm^2 dmol^{-1}) is the molar ellipticity at a λ wevelength, θ_λ is the observed ellipticity, c the molar concentration of the protein and d the pathlength in cm. It has to be noted, that the differential signal is very low, usually around 10 mdeg that corresponds to around 3×10^{-4} in absorbance units. Therefore it is important to provide low absorbance backgrounds and buffers, salts and other components disturb the measurement in high concentration.

Biological macromolecules have several chiral chromophores, and thus can be studied by CD-spectroscopy [75, 76]. Not only the atomic chiral centers, but also the secondary structure elements (α-helices, β-sheets) form chiral units. In the 170 – 260 nm range the CD signal of proteins is essentially attributed to electronic transitions in the peptide bond (**Fig. 4.14 a**). The intensity and sign of the signal depends on their environment, *i.e.* on the secondary structure of the protein.
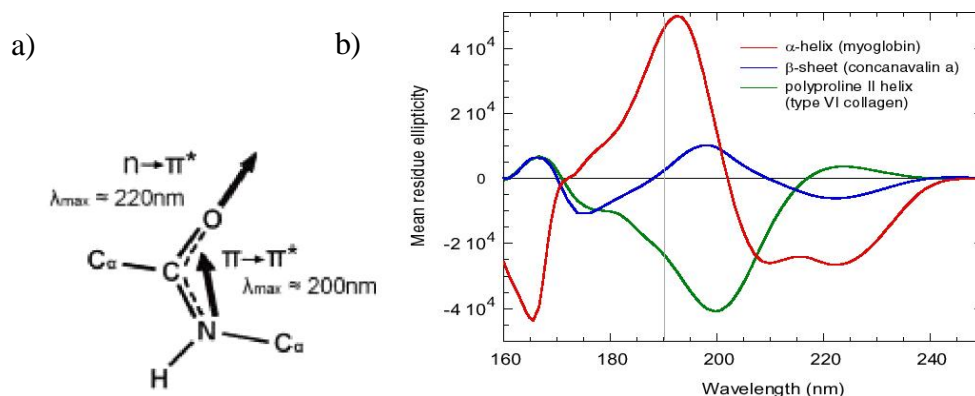
Fig. 4.14. a) *The electronic transitions of peptide bonds, taken from* [77]. **b)** *Typical CD spectra of the different protein secondary structure elements, taken from* [76]. *The grey line at 190 nm indicates the approximate lower limit for a conventional CD equipment.*

CD-spectroscopy is sensitive to conformational changes upon various reactions of the macromolecules, including the binding of metal cofactors or other macromolecules (protein, DNA). Temperature dependent CD measurements can provide information on protein stability. Computer algorithms are used to estimate the abundance of secondary structural elements in a protein by comparison to CD-datbases [77]. It has to be emphasized that for proper analysis the precise protein concentration has to be determined.

With the conventional CD instruments, e.g. the Jasco series, it is not possible to record CD spectra below ~180-190 nm for several reasons. The intensity of the Xe light

source decreases in this range. Water molecules and several buffer or salt molecules increase absorption at low wawelength. N_2 gas used to purge the optical units for protection from ozone and to exclude oxygen that would absorb in the 180-200 nm range. In the synchroton radiation CD-spectroscopy (SRCD) it is possible to measure down to ~160 nm [76] due to high intensity beam and vacuum system [78]. In our experiments at the ASTRID beamline (Aarhus STorage RIng in Denmark), with a 0.2 mm quartz cell it was possible to measure the spectra from 170 nm.

It has to be noted, that in case of metalloproteins CD-active charge transfer bands may appear, *e.g.* it was observed between 200 and 230 nm of Zn^{2+}-containing metallothioneins [79].

4.2.5. *Determination of protein concentration*

Even though concentration data is required for the interpretation of all quantitative experiments, there is no straightforward method available to determine it accurately [80]. One way is to calculate protein concentration from the UV-VIS absorbance spectrum of the sample. The spectrum has two peaks: ~280 nm (aromatic residues) and ~205 nm (peptide bonds) peaks. The extinction coefficient of the protein can be estimated based on the number of Trp, Tyr residues and disulfide bridges [76] at 280 nm:

$$\varepsilon_{protein} = n_{Trp} \times \varepsilon_{Trp} + n_{Tyr} \times \varepsilon_{Tyr} + n_{Cys} \times \varepsilon_{Cys}$$

or alternatively [75],

$$A_{280}(1mg/ml; 1\ cm) = (5690 \times n_{Trp} + 1280 \times n_{Tyr} + 60 \times n_{Cys})/M_{protein}$$

Based on the Lambert-Beer law the concentration can be calculated. It has to be noted, that both formulae refer to samples denatured in 6M guanidium hydrochloride. The determined value is reliable, if [75]:

- the protein sample is homogenous,
- there is no contribution from light scattering (indicated by the negative slope of the spectrum at 310-400 nm),
- there are no absorbing contaminants (the nucleic acid impuirity is *e.g.* indicated by the A_{280}/A_{260} ratio)
- a correction is performed for the difference between unfolded and folded protein – this can be performed by the comparison of the absorbance with and without guanidium hydrochloride. Usually this ratio is between 0.9 and 1.1.

The extinction coefficient of proteins in native state can be also estimated using experimental parameters based on a database of Trp and Tyr containing proteins. Assuming that all cysteins form disulfide bridges (otherwise the Cys term is neglected) [81]:

$$A_{280}(1mg/ml; 1\ cm) = (5500 \times n_{Trp} + 1490 \times n_{Tyr} + 62.5 \times n_{Cys})/M_{protein}$$

In this case there is no need to correct for unfolded/folded state, but the other previously listed criteria have to be fulfilled. The method was reported to be precise for Trp containing proteins but to have more than 10% error for the proteins without Trp. There is an online tool to estimate the extinction coefficient in aqueous solution as well as other parameters of the protein based on its sequence (http://web.expasy.org/protparam/). This uses the above equation and computes for both reduced and oxidized cysteines.

There are also several methods that use a reference sample to determine the sample concentration, including Comassie blue staining, biuret method, Lowry method, Bradford method. However, all these methods may give results influenced by contaminations of the sample and are also invasive. The most precise method is the quantitative amino acid analysis, which is based on hydrolysis in 6M HCl or other special buffer followed by reverse phase HPLC and amino acid quantification. The method applies an internal standard, such as norleucine in a known amount. It gives multiple concentration data based on the amount of each amino acid (except for the one that may be destroyed by the method, like Ser and Trp) to give the final, averaged result. This method also indicates the purity of the sample, based on the ratios of different amino acids.

4.2.6. X-Ray crystallography

X-Ray crystallography provides detailed information of the atomic structure of proteins in crystalline state. It relies on the measurement of diffraction of X-Ray radiation on single crystals and the computation of the structural parameters based on the observations.

4.2.6.1. Protein crystallization

For the crystallization experiments a highly concentrated protein sample (>3mg/ml) is required. In the vapor diffusion methods the protein sample is mixed with a precipitant matrix in a small volume (0.2-2 µl), and the drop is placed into a special dish, either sitting on a plate or hanging on a silica surface (**Fig. 4.15 a** and **b**, respectively). Near the drop a higher volume of the precipitant (reservoir) is placed. After closing the compartment containing the drop and the reservoir the solutions start to equilibrate: water evaporates from the drop, because it is less concentrated than the high volume reservoir. The aim is to lead the solution on the phase diagram (**Fig. 4.15 c**) from the soluble phase to the labile zone, where the nucleation can occur and continue in the metastable zone where crystals grow.
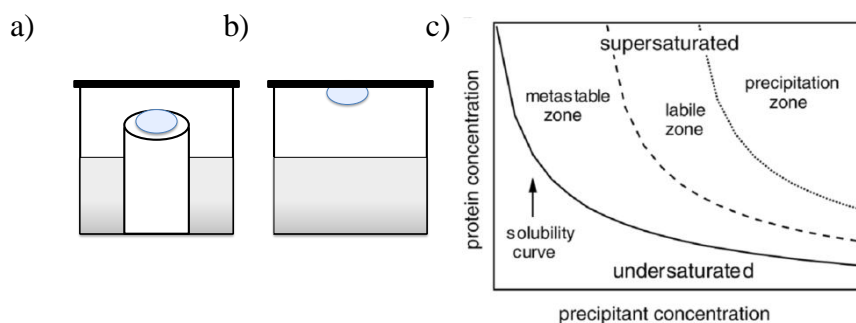
Fig. 4.15. a) *Sitting drop crystallization method.* **b)** *Hanging drop crystallization method.* **c)** *General phase diagram for protein solubility taken from* [82]. *The nucleation occurs in the labile zone and the metastable zone is favorable for crystal growing.*

The difficulty is that the phase diagram is unknown prior crystallization. Different proteins have different solubility and the understanding of the phenomenon is still limited. The most common strategy is to try several conditions in so called screens (24 or 96 different conditions in one experiment).

In this work the JCSG+ [83, 84], a commercially available reagent set with 96 solutions was used as a starting point for crystallization experiments. This set is a general sparse matrix or incomplete factorial screen which means that instead of systematically changing the conditions it is taking diverse points of the multidimensional parameter space to cover as many different conditions as possible. This is a good starting point for crystallization of a new protein. It contains different precipitants (polymers 59%, salts 22%, organics 19%) that are used to decrease the solubility of the protein by competing for the water molecules, making the solution oversaturated. There are other sets available that offer a systematic screen ("grid search") based on the systematic variation of conditions such as pH, PEG, cations and anions (PACT screen, INDEX screen). The additive screen (*e.g.* HR2-138 of Hampton Research) can be used to influence the morphology of crystals found in the initial hits. Additives are usually small molecules that can interact with the macromolecules and/or with the solvent molecules thus altering the protein-protein and protein-water interactions. This can influence not only the solubility of the protein but also the quality of crystals.

After setting up the experiment the changes are observed with a microscope. For the record and analysis of the results a scoring system is used (**Table 4.1**).

Table 4.1. *Scoring system of Hampton Research*

1	Clear drop
2	Phase separation
3	Regular granular precipitate
4	Birefringent precipitate or microcrystals
5	Posettes or spherulites
6	Needles (1D growth)
7	Plates (2D growth)
8	Single crystals (3D growth, < 0.2 mm)
9	Single crystals (3D growth, > 0.2 mm)

The crystallization of protein/DNA complexes is usually more challenging, because one more component means a new parameter that influences the experiment. The length of DNA chain, its stability (especially if it is a short hybridized oligonucleotide), the K_d of protein-DNA complex, the change in intermolecular interactions all influence the crystallization process.

To study metal binding the desired metal ion can be added to the sample in an appropriate concentration prior crystallization. Alternatively, the crystal can be soaked into metal-ion containing solution. The shorter the soaking time is, the lower the chance for the disruption of the crystal.

4.2.6.2. Crystal mounting and data collection

The single crystals are birefringent and can be distinguished from other crystals with a polarizer on the microscope. However, the crystals obtained in the screens may not only consist of macromolecules, but commonly salt crystals or small organic molecule crystals are obtained instead. The protein crystals can be selected by dyes that can diffuse in the water content of the crystal and readily binds to protein molecules turning the crystal colored. However, the test crystal can not be used for data collection.

For data collection the crystal has to be mounted from the drop. This can be done under microscope with a special nylon loop. It has to be quickly transferred to liquid nitrogen. The low temperature increases the lifetime of the crystal and also decreases the damage caused by X-Ray radiation during data collection. However, in most cases this method would lead to the crystallization of the water content of the drop disrupting the protein crystals. Therefore, the crystal is transferred for a short time to a drop of cryoprotectant solution (e.g. 20% glycerol) prior freezing.
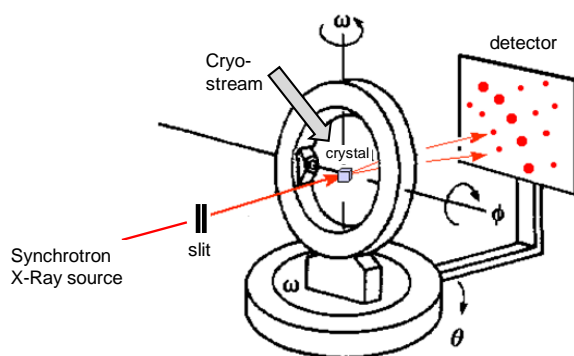
Fig. 4.16. Scheme of the experimental arrangement for data collection. The position of the crystal can be rotated in all directions and the detector can be moved as well. The crystal is cooled during data collection in the fume of liquid nitrogen.

Not only the features of the crystal but also the type of X-Ray source determines the quality of the obtained data. High intensity and microfocus X-Ray source is available at synchrotron beamlines allowing for data collections from crystals as small as 5 μm. The data collection is done following a strategy, because the high intensity X-Ray radiation quickly destroys the crystal by starting free radical reactions. The crystal can be moved in all directions in the beam (**Fig. 4.16**). The first step is to record two frames at a 90° difference. Based on the diffraction pattern it can be seen whether it is a protein crystal, if yes what is the unit cell and resolution of the crystal. After that several (500-5000) frames are recorded in different orientations of the crystal, usually rotating it by 180-360°. Thus a redundant data set is obtained decreasing the experimental error.

4.2.6.3. Solving a protein crystal structure

Interpreting crystallographic data of proteins into structural information is a complex task [85]. The molecules are regularly organized in the single crystal, and behave as a scattering grid. In most directions the scattered beams extinguish each other, but in certain directions, following Bragg's low the intensity is enhanced:

$$2dsin\theta = n\lambda$$

where *d* is the distance difference in the way of light beam of two scattering centers (*e.g.* atoms), θ is the angle between the light beam and scattering plane, λ is the wavelength of the light and *n* is an integer. This condition of constructive interference determines the directions where the scattered light beams ("reflections" in the traditional terminology) can be observed as spots on the detector. *d* depends on the structure of the molecule raising the opportunity to calculate structural data from the diffraction pattern.

In the first step of solving a structure the unit cell and space group is determined. Then the data of different frames collected at different angles as compared to the detector are merged and the Miller indices are assigned to the reflections, and the software integrates the intensities of each spots.

The main difficulty is to solve the "phase problem": the intensity of the diffracted beams is observed, but there is no tool to detect the phases of the beams. In case of small molecules direct mathematical methods can be applied, but that is not the case with macromolecules. One method is to add high electron density atoms, like metal ions or halide ions to the crystal by soaking and use their signal to get the initial phases. However, if there is a related structure available, a much easier way is to perform molecular replacement. The model of the template structure is fit with a subsequent rotational and translational functions and then used to calculate the initial phases that can be used as a starting point for model building and refinement.

The final step is to refine the structure. This includes automated computer algorithms as well as manual modifications of the model. The new model is used then to calculate an electron density. The atomic positions are adjusted so that they fit the measured data better. The success of refinement is characterized by the R_{work} factor, a statistical parameter showing how far the calculated and observed amplitudes are. It is defined as $R_{work} = \sum|F_{obs}-F_{calc}|/\sum F_{obs}$ where F_{obs} is the observed and F_{calc} the calculated amplitude. The lower the R_{work} factor, the better the fit. Similarly to the fitting of polynomial expressions to a data set, this can result in overinterpretation of the data. A related parameter, R_{free} is calculated in a similar way but on a set of data (usually 5-10%) excluded from the refinement process. In a good refinement R_{free} does not exceed R_{work} by more than 7% [85]. In the opposite case the changes induced to the model would interpret the experimental noise as signal. The refinement is continued until the R-values can not be further minimized significantly and the structure is chemically reasonable. To check this the molecules are analyzed with the aid of empirically derived geometrical rules, like usual bond length, angle, common rotamers of residues etc. The generation of Ramachandran diagram of protein residues reveal outliers that may be misinterpreted. At low resolution it is possible to observe alternative conformations of the proteins.

4.2.6.4. *Deposition of crystal structure*

The experimentally determined coordinates of biological macromolecules are written in the .pdb file format, and are collected in the Protein Data Bank (PDB Database, www.pdb.org) [86]. The pdb file contains information of the publication related to the

structure, details of experimental and computational methods of structure determination, statistics, secondary structure elements, missing residues etc. In the second part the Descartes-coordinates of each atoms are listed line by line. Each line contains the occupancy and B-factor of the atoms as well. The occupancy is lower than 1.0 if only a fraction of the molecules contains the given atom at the given position in the crystal. The B-factor or temperature factor shows how strict the position of the atom is in the crystal. The higher the temperature factor the more flexible the atom and the less precise its position is in the electron density map.

4.3. FURTHER METHODS TO STUDY DNA-BINDING AND NUCLEASE ACTIVITY

4.3.1. Electrophoretic mobility shift assays

Electrophoretic mobility shift assays (EMSA) rely on the observation that molecules with different charge, shape and size have different electrophoretic mobility in a gel matrix in an electric field. For DNA molecules longer than ~100 bp, agarose gels are suitable. The experiment is easy to perform and robust. For low amount of DNA or short sequences low % (5-10%) native PAGE is more appropriate.

The cleavage of DNA by nucleases can be followed using the EMSA technique, in time course experiments. Bacterial plasmids, common choice for such experiments are naturally in supercoiled form, and once a single strand cut (nick) occurs, the DNA turns to a relaxed opencircular form. Double strand cleavage or two single strand cleavages at the opposite strand, close to each other lead to a linear fragment. These forms of DNA have different mobility properties in the agarose gel, thus even a single strand cleavage can be observed (**Fig. 4.17 a** and **b**).
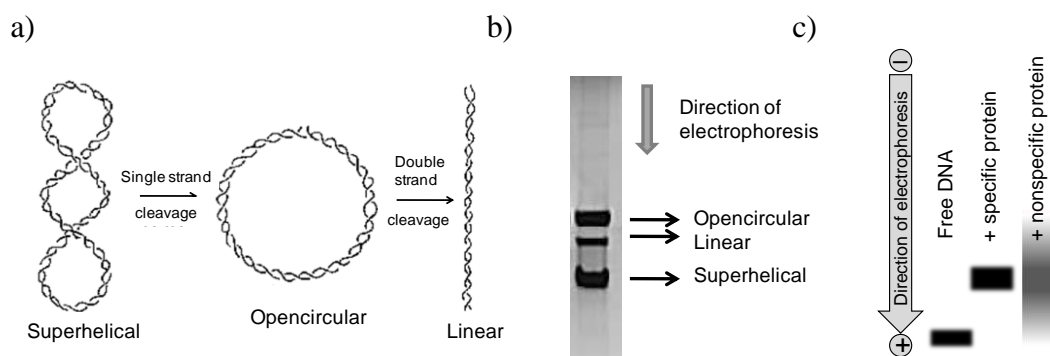
Fig. 4.17. a) *Different forms of bacterial plasmid DNA* **b)** *Electrophoresis of a plasmid DNA sample* **c)** *Schematic representation of protein binding in EMSA*

Protein binding to DNA can also be observed in EMSA experiments. In most of the cases the protein/DNA complex moves slower in the gel as compared to free DNA molecules, due to altered shape and charge. Applying a range of protein concentrations, the DNA-binding of the protein can be quantified. It is a semiequilibrium technique and the DNA-

binding affinity can be either slightly increased or decreased in the gel cages. It is also possible to distinguish between specific and nonspecific binding (**Fig. 4.17 c**). In case of sequence specific binding the composition of the complex is determined and the shifted band is sharp and intense. If the binding is nonspecific, several protein molecules may bind to the DNA, and a smear is observed consisting of a mixture of complexes with different composition and stability. It is also important to optimize the buffers used for electrophoresis and sample loading. A number of parameters needs to be optimized: ionic compounds, such as NaCl can be added to decrease the ionic interactions thus increasing specificity, and to stabilize the hybridized form of short DNA sequences; EDTA or metal cofactors can be used if the protein function requires that; nonionic detergents can be used to improve protein solubility and decrease aggregation.

The detection of DNA can be performed by several methods. In the noncovalent detection methods fluorescent molecules are used, that significantly change fluorescence intensity upon DNA binding. Such probee is e.g. the ethidium-bromide, increasing fluorescence after intercalating between neighboring base pairs of the DNA molecule. However, the detection level of these methods is low causing severe limitations of applicability. Furthermore, they compete with other ligands binding to the DNA molecule. Another option is to use covalent labeling of DNA-molecules including radiolabeled groups, *e.g.* γ^{32}P-ATP. Covalent fluorescent labeling molecules and biotin can also be used to detect DNA molecules. In case of short DNA sequences that have a low melting temperature, the whole procedure has to be carried out at 4 °C.

4.3.2. *Linear dichroism spectroscopy*

Linear dichroism (LD) spectroscopy is a technique related to circular dichroism spectroscopy, but measuring the difference in the absorption of two perpendicular linearly polarized beams of light. In a homogenous sample there is no such absorption difference. An LD signal can only be observed if the sample has an anisotropic orientation, and thus the transition dipole moments of molecules are also oriented to a specific direction. Consequently, the LD signal depends not only on the concentration, but also on the degree of orientation in the sample. UV/VIS measurements can provide information about the total concentration of the sample, and LD about the degree of orientation. It is defined as $LD = A_{\parallel} - A_{\perp}$, where $A_{\parallel}$ and $A_{\perp}$ are the absorbances of the paralell and perpendicularly linearized light, respectively, compared to the orientation axis of the molecule. The orientation can be achieved *e.g.* in stretched polymer films or in the so called micro-volume Couette cell that was designed for the study of macromolecules [87]. It consists of a quartz cylinder and a quartz rod that fits into the cylinder. The sample is filled into the narrow gap between them, and one of them (usually the outer cylinder) is rotated at a high speed, *e.g.* 3000 rpm. As a result, fiber-shaped molecules are oriented in the flow by mechanical shear forces (**Fig. 4.18**).
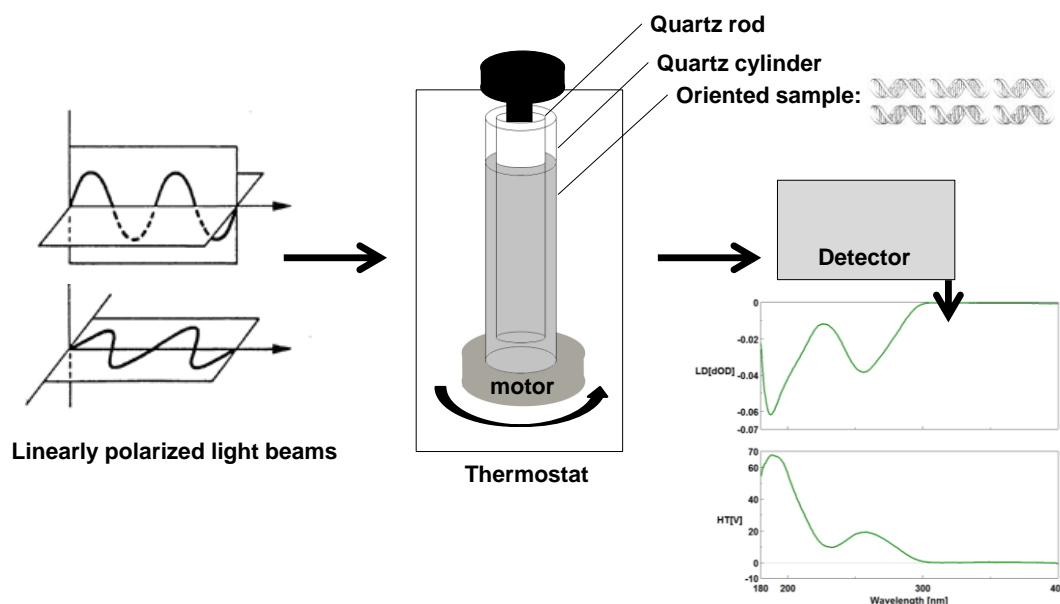
Fig. 4.18. *Scheme of flow LD-spectroscopy. The sample is oriented in the Couette cell and thus absorbs the components of the two linearly plarized lights differently. The LD-spectrum of a long (thousends of bp-s) DNA chain is shown as an example, which has the same peaks but opposite sign as compared to the absorption spectrum. The high tension (HT) value is proportional to the UV-VIS absorption and simultaneously measured by the instrument.*

The bases in DNA molecules are stacked and twisted around the axis of the helix. If the molecules are oriented in the cell, their π-π* transitions below 300 nm give rise to LD-signal [88]. Since the bases and their in-plane transition moments are perpendicular to the orientation axis, the LD-spectrum of DNA is negative. It has peaks at the same wavelength as the UV-VIS spectrum, but the intensity depends on the concentration and orientation efficacy that is influenced by the length, flexibility and ligand binding of the DNA. If the ligands are oriented by DNA-binding, they can have LD-signal themselves. Ligand binding also influences the dynamics of DNA chain affecting the intensity of the spectrum.

Enzymatic reactions resulting in the modification of DNA can be followed by LD spectroscopy. Since the orientation ability decreases with the decreasing length of DNA chains, DNA cleavage can be also studied. Kinetic parameters of restriction endonucleases digesting plasmid DNA have been determined with LD-spectroscopy, and the results were in agreement with those obtained by other techniques [89-91]. However, with a protein having more than one cleavage site in the DNA chain, the phenomenon becomes complex to describe mathematically. Nevertheless, LD-spectroscopy provides a possibility to study nuclease activity in a non-invasive manner.

4.3.3. *Atomic force microscopy*

Atomic force microscopy (AFM) is a technique that can be used to study surfaces with molecular detail. The sample is fixed to the piezoelectric scanner positioned with a subnanometer accuracy (**Fig. 4.19**). The end of the tip has a radius of curvature in the order of nanometers. Its vertical movement is detected by a laser beam, reflected to the detector. The interatomic interactions between the sample and the tip are directly measured, and the

topography of molecules on the surface can be analyzed. In contrast, STM (scanning tunneling microscope) would require the electric conductivity of the sample.
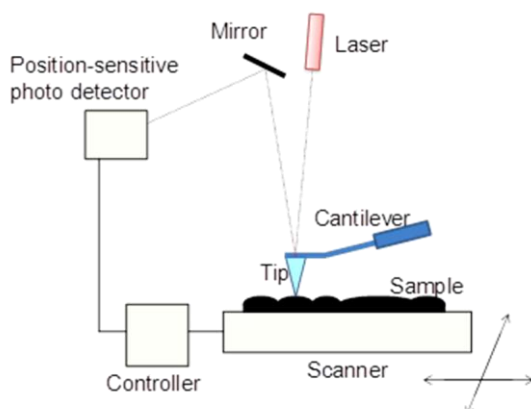
Fig. 4.19 *Scheme of AFM measuring atomic interactions between the sample and the tip* [92].

The surface can be scanned in various modes. In the contact mode the repulsion while in the non-contact mode the attraction forces between atoms are used. However, the strong contact may damage the biological sample, and therefore these molecules are studied in the so called tapping mode (or intermittent contact mode). In this mode the cantilever is vibrated at a frequency close to its resonance frequency, with a small amplitude. As the tip approaches the sample surface, the van der Waals attractive forces change the amplitude and phase of the vibration.

A further difficulty in case of biomolecules is that they are usually hydrophilic in nature. The mica surface, that is commonly used to bind the biomolecules is also hydrophilic, likewise the tip. They can bind water molecules that form a bridge when the tip approaches the sample. The oscillating tip in the tapping mode can break this force, and thus the effect can be eliminated.

DNA can be deposited onto a mica surface in presence of Mg^{2+}- or other divalent metal ions [93]. The role of cations is to bind the negatively charged DNA molecules to the negative mica surface by electrostatic interactions. During this the DNA-chain behaves as a flexible polymer and equilibrates on the surface [94]. Functionalization of the mica surface can allow for different binding mode of DNA molecules [92]. Protein binding to DNA, protein induced DNA-bending [95-97], and cooperative DNA-binding of DNA methyltransferase 3*a* [98] were visualized by AFM. In the latter case the WT protein binds to DNA cooperatively and protein molecules cover DNA in all length, thus a high profile is observed, while in case of a mutant cooperativity is lost and discrete protein molecules are observed on the DNA-surface.

4.4. COMPUTATIONAL METHODS

In contrast to experiments that study macroscopic features of materials, computational modeling aids to describe features of individual molecules. The more atoms are included in the system, the higher is the computational cost of the modeling. This results in limitations and the method has to be chosen according to the question asked. Quantum chemical methods augmented with approximations are appropriate to study macromolecular systems, but such calculations have high computational requirements. Newtonian mechanics can be routinely used for macromolecules, but due to its classical nature it can not be applied to describe electronic states. The combination of different methods according to the question to answer can provide a valuable tool augmenting the experimental studies.

4.4.1. The visualization of macromolecular structures

The structure of macromolecules can be determined on atomic detail *e.g.* by X-Ray crystallography or NMR-spectroscopy. Both the experimental and modeled structures can be understood via visualizing them in three dimensions by using the Cartesian coordinates of the atoms. Numerous softwares are able to perform this based on the list of coordinates of the atoms in the molecule. The figures in this work are made in PyMOL 1.3 (*The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.*), a free software for academic use. It has a graphical interface and command line as well. Apart from the publication quality representation of structures in different modes (line, stick, cartoon, surface etc.) it can perform the alignment of different structures based on structural similarity. The root mean square deviation (RMSD) of the superposed structures is also computed. The different operations in PyMOL can be controlled and automated by *python* scripts.

4.4.2. Molecular dynamics

Molecular dynamics (MD) simulations provide details of atomic structure and movement. MD can be used to model conformational changes and flexibility of molecules on a time dependent manner. Classical MD is a molecular mechanics based method, which means that the atoms are considered as objects moving according to classical mechanical laws. The required input data includes the initial coordinates of the biomolecule, this can be obtained from crystal structures or from other modeling tools. Then the initial velocities are generated based on *e.g.* Maxwell-Boltzmann velocity distribution. The potentials and forces acting on each atoms are computed. Based on the results the geometry is updated, the new physical properties (energies) are computed and written to the output file. The process is repeated millions of times and the changes of the computed parameters as a function of time steps compose the trajectories. After the MD algorithm finished, the required information can be extracted or computed from the trajectories. The calculation follows the scheme shown in **Fig. 4.20**.
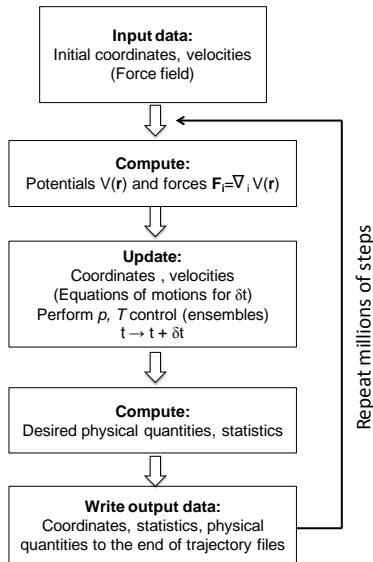
Fig. 4.20 *Scheme of an MD simulation. In the update step the Newtonian equation of motion is solved numerically:*

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i}$$

There are numerous methods for the integration of the equations of motion, e.g. Verlet-algorithm, Leap-frog algorithm and velocity-Verlet algorithm. They differ in the order of computation of forces, velocities and positions.

4.4.2.1. Computation of the MM energy

In order to determine the forces acting on each atoms of the system it is necessary to compute the interaction energies of each atom with the surrounding atoms. The energy calculations in basic molecular mechanics involve separated terms that contribute to the total energy (*V*) of the system: energy related to bond stretching, bond angle bending and torsion are the bonded terms and the nonbonded interactions involve electrostatic (Coulomb's law) and van der Waals terms (Lennard-Jones potential).

$$V = V_{stretch} + V_{bend} + V_{torsion} + V_{nonbonded}$$
$$= \frac{1}{2}\sum_i k_{i,stretch}\,(r_i - r_{i,0})^2 + \frac{1}{2}\sum_i k_{i,bend}\,(\theta_i - \theta_{i,0})^2$$
$$+ \sum_{n=0}^{N} \frac{V_n}{2}\,(1 + \cos(n\omega - \gamma)) + \sum_{i<j}^{N}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + \varepsilon_{ij}\left[\left(\frac{R_{0,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{0,ij}}{r_{ij}}\right)^{6}\right]\right)$$

The position of the atoms (r_i), the angle of the bonds (θ_i) and the distances between atoms (r_{ij}) are determined by the geometry of the system. $k_{i,stretch}$, $k_{i,bend}$, are constants Hooke's law, n is the number of different torsion positions, $R_{0,ij}$ is obtained from the van der Waals parameters of i and j atoms. These latter constants have to be known for each type of atom in the molecule, and therefore they are stored in force fields. They are empirically derived data of interaction energies, geometry constraints etc. Different force fields are optimized for different group of molecules (protein, DNA, alkanes etc.). The bonded terms are easily obtained, but in case of nonbonded interactions it is more problematic to determine which interactions have to be computed.

The simulations can represent NVT (canonical), NVE (micorcanonical) or NPT (isothermal-isobaric) ensembles, where N means the constant number of atoms, V, E and T the constant volume, energy and temperature of the system, respectively. These are achieved

by restraints of features related to the constant, such as kinetic energy in case of the temperature.

4.4.2.2. Set-up of the simulation

Several preparation steps are required prior to the simulation, even when the starting structure is already known. In the present work the following strategy was applied:

1) Construction of topology files containing sequence data of protein, DNA or combined and references to the corresponding force field files
2) Setting up distance restraints: if a restraint is needed, it has to be specified according to the atom numbering used by the MD software
3) Convert the initial geometry to a special file format according to the software used
4) Generate coordinates for hydrogen atoms if the initial file does not contain them
5) Energy minimization to relax hydrogen atoms
6) Generate solvent box around the macromolecule with the selected solvent model
7) Energy minimization with solvent
8) Substitute the required number of water molecules by ions to provide ionic strength resembling natural solution
9) Equilibrate system
10) Run productive MD simulation

The water molecules can be explicitely involved in the simulations with different water models, such as SPC (regarding a water molecule as point charges at the position of the atoms) or the more complex TIP3P, TIP4P. The solvent box may have a cubic or other shape. In the calculation the edge of this box would extend to the vacuum. To avoid this phenomenon the so called periodic boundary conditions are used. In this algorithm if an atom moves out from the box at the edge, it will move in at the opposite side of the box. This means that the box appears to be repeated infinite times, each containing the same macromolecule. The solutes molecules in the neighboring boxes must not interact, because that would mean an artifact: the molecule woould interact with itself (with its image in the neighboring box). Therefore the minimal size of the box is $a = 2\, r_{cutoff} + d_{molecule}$, where r_{cutoff} is the limit of the van der Waals interactions and $d_{molecule}$ is the diameter of the macromolecule.

It is important to mention the limitations of MD simulations. The simulations are classical, thus excited states and chemical reactions can not be modeled. The force fields are approximate and long range interactions are cut off at a certain value. The boundary conditions do not represent a real ensemble but an unnatural state. Nevertheless, MD is a highly supportive technique for the experimental tools both in predicting and understanding dynamic features of macromolecules.

4.4.3. Semiempirical quantum-chemical calculations

Although molecular mechanics provides a possibility to calculate structure and energetics of macromolecular systems, it is not applicable for problems involving quantum features, such as electronic states. The quantum chemical equations can describe these phenomena, including the formation or disruption of chemical bonds and precise atomic interactions. However, these calculations require a high computational cost limiting their use for small molecules. They rely on the Hartree-Fock methods (self-consistent field calculations), solving the nonrelativistic Schrödinger equation for molecules within the Born-Oppenheimer approximation. Semiempirical quantum chemical methods provide a compromise between these two extremes introducing further approximations, at the same time using empirical parameters to augment the calculations. The parameters are derived to best agree with experimental or *ab initio* (without parameters) calculations. The two electron integrals are approximated or omitted.

Parametric method 6 (PM6) is based on the neglect of the diatomic differential overlap (NDDO) [99]. It has several improvements as compared to earlier NDDO-based methods (AM1, PM3) including parametrization based on a larger data set with an emphasis on biological molecules, parametrization of several metal ions and improved empirical description of hydrogen bonds. It contains a core-core diatomic interaction term and an approximation for d-orbitals. This method can be used effectively to describe macromolecules. However, the weak interactions are not adequately described. The DH approximation was introduced to include an empirical dispersion energy and an electrostatic term to improve the description of hydrogen-bonded complexes [100]. The hydrogen bonding scheme was further improved resulting in the DH2 augmentation [101]. The PM6-DH2 method was proved to model proteins with a good accuracy.

In order to use semiempirical methods involving all atoms of the macromolecules, an effective method was needed for the solution of the self-consistent field equations. In the method usually used for semiempirical calculations the SCF equations are solved with matrix algebra methods, resulting in a computational time proportional to the 3rd power of the number of atoms in the system. This is a severe limitation for the biological systems. It could be decreased to a quadratic function by dividing the molecule into groups. In the extreme case one atom is considered to be one group. This is the base of localized molecular orbitals (MOZYME method, [102]), where the molecular orbitals are obtained from the Lewis-structure of the molecule instead of orbitals involving all atoms in the molecule. This results in the linear scaling of the method, and thus it can be effectively used for macromolecular systems [103].

4.4.4. QM/MM calculations

As a result of the great improvement of computational facilities in the last years it became possible to compute as large systems as 100-200 atoms in a quantum chemical way. However, even these systems are too small for the modeling of biomolecules containing thousands of atoms. Such large systems can be described by classical molecular mechanic (MM) models, but the accuracy of those methods is often not enough. The hybrid QM/MM methods were developed to unite the advantages of the molecular mechanics (fast calculation on large systems) and that of quantum chemical calculations (accurate description of the structure involving electrons). While the global protein structures and folds are undoubtedly best predicted by X-ray crystallography, the QM/MM calculations seem highly competitive and complementary method for the local structural details [104, 105].

4.4.4.1. Definition of the QM/MM energy

In the QM/MM model the system is devided into layers (**Fig 4.21**). The region of interest is calculated on quantum chemical (QM) level, while its surroundings are treated by MM methods. This approach proved to be successful in modelling enzymatic reactions. However, several problems arising from the interface of the two regions have to be considered. Usually, there are 3 layers in the model (**Table 4.2**).

Table 4.2 The layers of a typical QM/MM calculation

Name	Level of computation	Fixed atom positions?
System 1	QM	No
System 2	MM	No
System 3	MM	Yes

System 1 is the part of the system calculated on QM level (usually DFT methods). System 2 and 3 are the parts calculated by MM methods. The difference between these two is that in system 2 the atoms are allowed to move, while in system 3 they are in fixed positions. The inclusion of system 3 does not fasten the calculation, the reason for using it is a theoretical problem for locating energy minima and transition states by QM/MM methods. The problem is arising from the large number of local minima found for the system, making it difficult to identify the desired points of the energy surface. This effect can be decreased by fixing the positions of the outer atoms. In this way also the hydrogen bond network reorganization at the vacuum-water interface can be avoided.
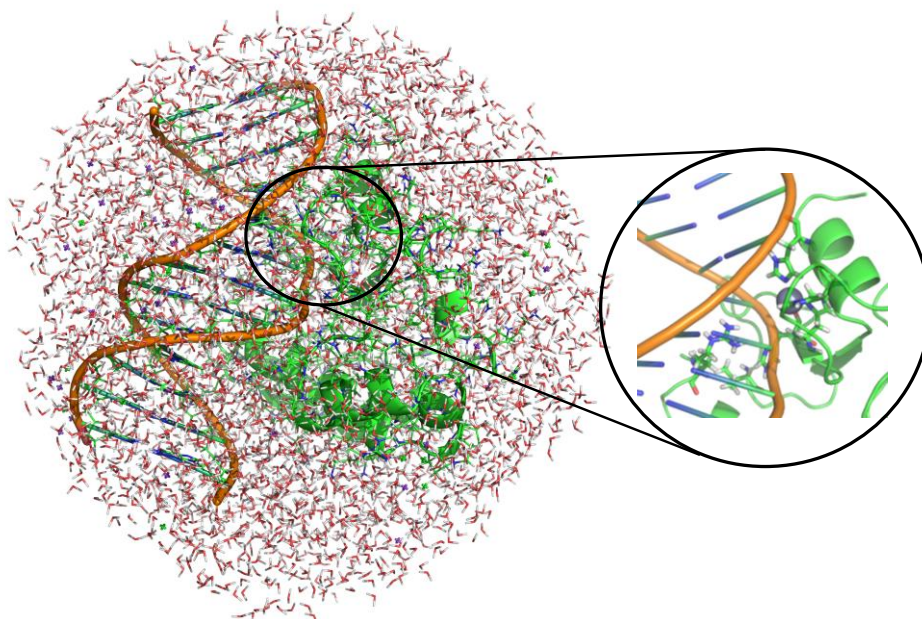
Fig 4.21 *The QM/MM model of the NColE7-DNA complex. The magnified atoms in stick representation constitute the QM region.*

The energy of the system is not a simple sum of the QM and MM parts, because these two regions, being a part of the same molecule, strongly interact with each other. For this reason the energy of the system is defined as

$$E_{QM/MM} = E_{MM123} - E_{MM1} + E_{QM1}$$

where E is the total energy of the system, E_{MM123} is the energy of the whole system computed on MM level, E_{MM1} is the MM energy of the quantum system and E_{QM1} is the energy of the quantum system.

4.4.4.2. The QM/MM interface

During the selection of quantum region covalent bonds have to be "separated". In the resulted quantum system there will be missing valances of these atoms, which would cause serious artifacts in the calculation. To avoid this, so called link atoms are used that are normally hydrogens added in position of the neighbouring atom in system 2. It is best to cut nonpolar C-C bonds, because the C-H(link) bond also has a low polarity. In the MM calculations parameter sets are used for the different residues. Near the QM/MM interface several residues are included only partially in system 2, which means that the parameter sets for such truncated residues have to be constructed.

4.4.4.3. The ComQum program

ComQum is a program for combined QM/MM optimisations [106, 107]. It performs quantum chemical calculations with the program Turbomole and molecular mechanics calculations with the program package AMBER. The scheme of its computation is shown in **Fig. 4.22**.
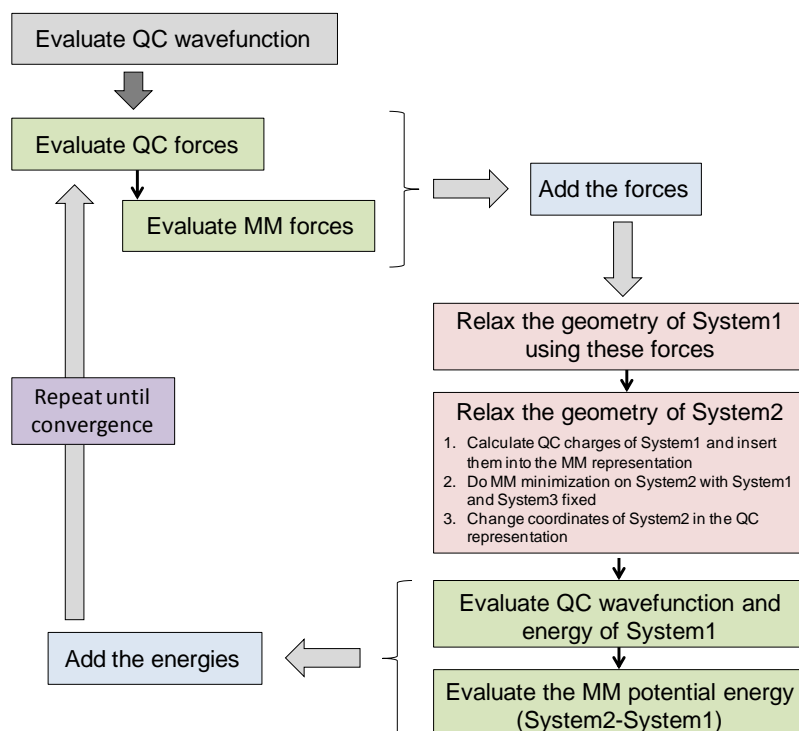
Fig. 4.22 *Flow scheme of the original ComQum program based on the description found at http://www.teokem.lu.se/~ulf/ comqum.html.*

5. RESULTS

5.1. NColE7 AS A POSSIBLE BUILDING BLOCK FOR A CONTROLLED ZINC-FINGER NUCLEASE

As it turned out from the results on the N-terminally truncated mutants [43] and described in the previous chapters, NColE7 has an interesting feature in that its termini are cooperating during the catalytic reaction. The C-terminal HNH motif is considered to be the active centre, but the protein can not function if four residues (446-KRNK-449 referred as "KRNK" in the followings) are deleted from N-terminus – even though the protein structure is not significantly altered. The understanding of this phenomenon is essential for the design of a controlled nuclease based on the function of N-terminus. Therefore, I addressed the question, how these positive residues influence the activity of the protein. The four point mutants included KGNK, KGNG, GGNK and GGNG named after the sequence substituting the KRNK part of the protein. The mutation sites are shown in **Fig. 5.1**.


```
NColE7  438-GPLGSPEFKRNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
KGNK    438-GPLGSPEFKGNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
KGNG    438-GPLGSPEFKGNGPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
GGNK    438-GPLGSPEFGGNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
GGNG    438-GPLGSPEFGGNGPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL

NColE7  SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
KGNK    SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
KGNG    SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
GGNK    SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
GGNG    SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
```
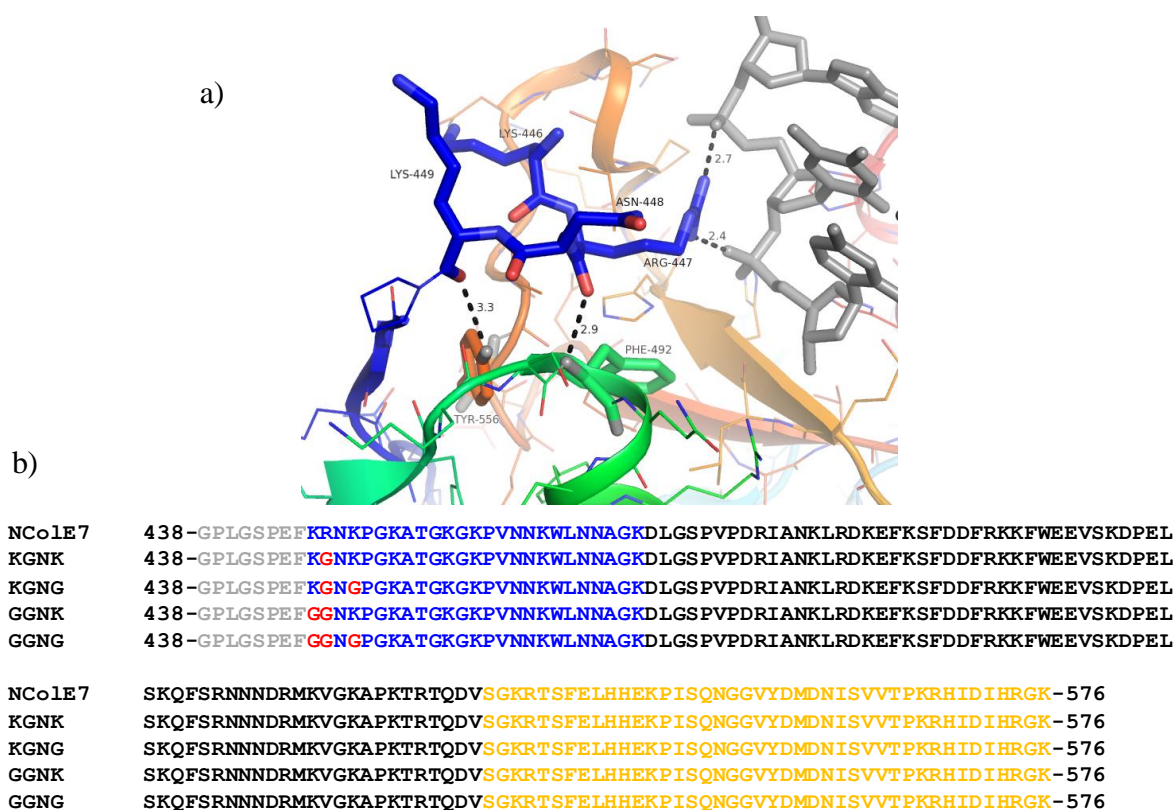
Fig. 5.1. a) *Location of the selected mutation sites shown in the 3FBD crystal structure [48]. They constitute the N-terminus (blue) but get sterically close to the C-terminal HNH-motif (yellow) and other parts of the protein (green). Apart from the intramolecular H-bonds of the N-terminal backbone the R447 sidechain interacts with two phosphate groups of DNA (grey) including the scissile phosphate.* **b)** *Sequence of the WT NColE7 and the N-terminal point mutants of this study. The first 8 residues (grey) are the result of the applied purification procedure (remaining from the GST-tag). The N-terminal loop is written in blue with the mutations in red. The HNH-motif is written in yellow.*

The sequence shown in **Fig. 5.1 b** contains additional residues as compared to the natural enzyme on the N-terminus. This is a result of the applied purification procedure, remaining from the GST-tag after protease cleavage. Since this short sequence does not contain any positively charged sidechains, it is considered not to interfere with our experiments.

The detailed experimental conditions and computational methods applied related to the results of this chapter can be found in the Appendix chapter 10.1.

5.1.1. Cloning and toxicity of NColE7 and its mutants

The genes of the mutant proteins were attempted to be constructed with the method described in chapter 4.1.1, including the PCR, restriction enzyme digestion and ligation. The mutations were introduced with the primers. The resulted PCR fragment was inserted into a pGEX-6P-1 plasmid between the restriction sites EcoRI and XhoI. The new plasmid was then transformed to *E.coli* DH10B cells. However, DNA sequencing showed unwanted base mutations that resulted also in amino acid mutations. After several trials the statistics showed that the unwanted mutations occur mostly within or close to the active site. By investigation of the process we found that these mutations arose from the PCR as a minuscule side reaction (**Fig. 5.2**). In DH10B cells there is a basic level of protein expression. The target proteins were toxic for the cells, while the side products containing active site mutations (**Fig. 5.3**) were nontoxic and thus the cells could survive with them. Thus, the bacteria that contained an erroneous gene after transformation were selected in the procedure.
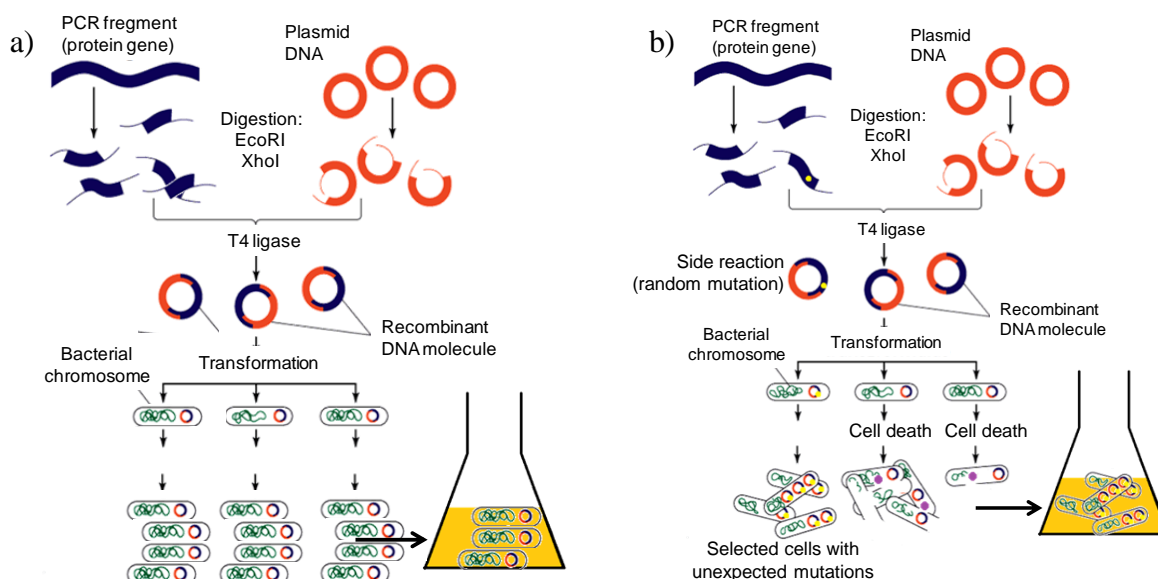
Fig. 5.2. a) *A scheme of recombinant gene production. A template is modified by PCR, and then built into a carrier plasmid providing antibiotic resistance. After transformation to bacteria and their cultivation the plasmid containing the gene can be extracted.* **b)** *The scheme of the process leading to genes with random mutations. The unexpected mutations are formed in a side reaction of PCR. If they abolish the toxic nuclease activity of the NColE7 mutants, the cells can survive.*
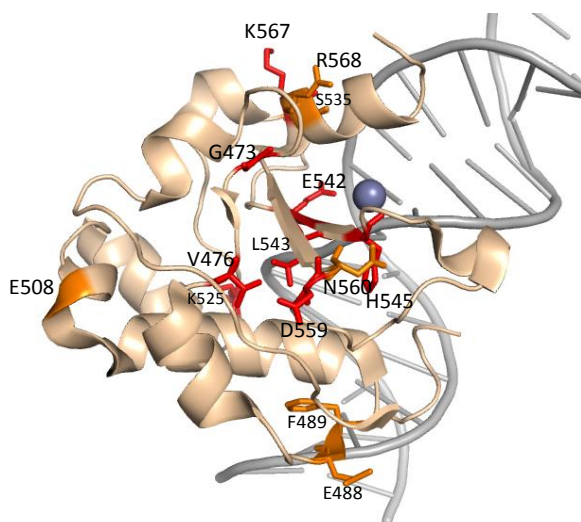
Fig. 5.3. *Location of the random mutations, as shown on the crystal structure 3FBD (D493Q-NColE7 in complex with a 18 bp DNA). The residues in red occur as single site mutations, while the orange ones are double site mutations.*

This implies that the designed N-terminal point mutants (including KGNK, KGNG, GGNK and GGNG) did not lose the cytotoxic effect. The several obtained erroneous plasmids could not be used for protein production, but these experiments provided useful information regarding which residues might be important in the function of NColE7. Later, in the process of NColE7 redesign this information was supporting the computational design [108].

In the following, we used a different strategy for cloning of the toxic proteins. The genes of the mutants were amplified together with the Im7 protein and cloned into the pGEX-6P-1 vector [47, 63]. As described in the previous sections, NColE7 and Im7 form a stable protein complex that prevents the host cell from the toxic effect of the nuclease (**Fig. 5.4**).
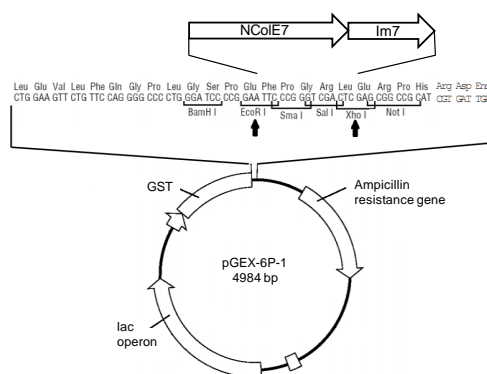
Fig. 5.4. *Scheme of the plasmid that contains the NColE7 gene as well as the Im7 gene.*

5.1.2. *Protein purification*

The NColE7 mutant proteins with the Im7 protein could be overexpressed in BL21 (DE3) cells [63]. The drawback of this procedure is that the nuclease and the immunity protein have to be separated during purification. To achieve this, two sequential HPLC purification steps were applied following a modified protocol of [50]. First the GST-NColE7/Im7 complex was separated from the other proteins of the cells by GST-affinity chromatography. After that the pH was decreased from 7.9 to 3.0. The protonation of several sidechains at low pH induces the repelling forces between NColE7 (pI = 10.1) and the Im7 protein (pI = 5.2) which became also protonated and positively charged. Thus the complex

dissociated and a cation exchange chromatography at pH = 3.0 was performed to separate the proteins. NColE7 and its mutants were eluted at ~1M NaCl concentration, while the immunity protein was collected after the gradient elution, at pH 7.9. **Fig. 5.5** shows a representative example of the sequential purification.
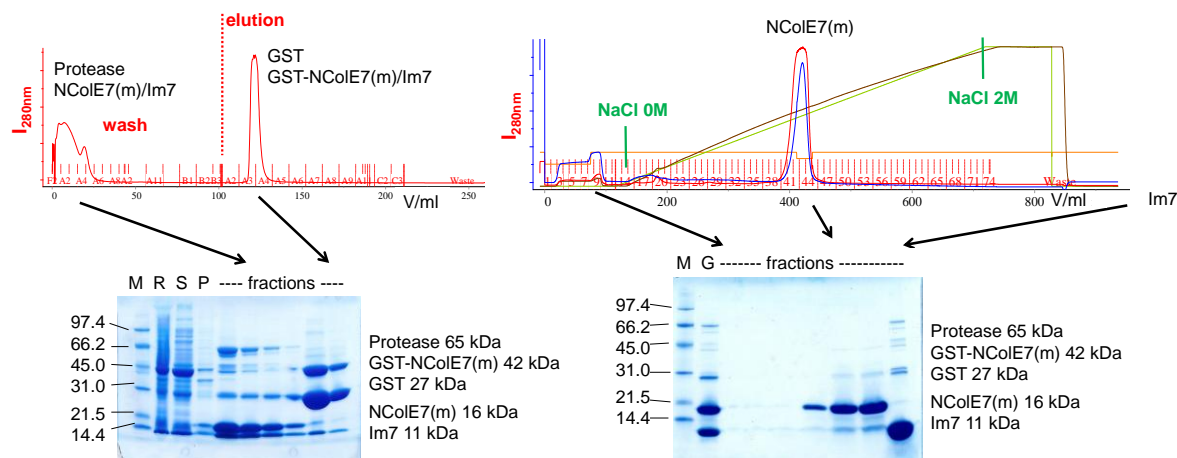
Fig. 5.5. *Example of the two-step HPLC purification of the mutant proteins: the chromatograms and SDS-PAGE analysis. After the GST purification step (left side) the GST-NColE7(m) was obtained (27 kDa protein). This was followed by the protease cleavage and a cation exchange chromatography (right side) that lead to the 16 kDa target protein. The Im7 protein fraction was also collected during the regeneration of the column. NColE7(m) – mutant NColE7, M – protein standard marker, R – raw extract from sonicated cells, S – supernatant of R, P – pellet, G – collected fractions from GST affinity chromatography.*

The fractions containing the target protein were collected, and the elution buffer was exchanged to 20 mM HEPES (pH = 7.7) by ultrafiltration. The molecular weight was measured by nano-ESI MS technique, after the addition of one equivalent Zn^{2+}-ion. The summary of protein production can be found in **Table 5.1**. The measured and expected molecular weights suggest the presence of one Zn^{2+}-ion per molecule.

	$MW_{expected}$ / Da	$MW_{measured}$ / Da	c / µM	m /mg
NColE7	15875.2	15874.9	316	~25
KGNK	15776.1	15776.2	377	~30
KGNG	15705.0	15704.5	415	~30
GGNK	15705.0	15704.8	273	~20
GGNG	15633.9	15632.8	118	~15
Im7	10961.1	10961.5 / 10412.8	293	~30

Table 5.1. *Expected and measured (ESI-MS) molecular weight of the proteins and their amount produced. The molecular weights refer to the holo form of the WT NColE7 and its mutants containing one Zn^{2+}-ion. The spectra were recorded after short incubation of the proteins with one equivalent of Zn^{2+}-acetate.*

5.1.3. N-terminal mutations decrease nuclease activity of NColE7

The gene cloning experiments revealed that the N-terminal point mutants are cytotoxic, which is most probably caused by their nuclease activity. Using the purified proteins we performed further experiments to describe the nuclease activity in detail. The most convenient tool to follow DNA cleavage is to check the DNA-content of the reaction

mixture by agarose gel electrophoresis. Bacterial plasmid DNA is frequently applied in this assay, because it is sensitive for even a single strand cleavage, as shown in chapter 4.3.1. The natural superhelical form is circular and compact, thus it moves easily in the pores of the gel as compared to the opencircular or linear forms that arise as a result of single or double strand cleavage, respectively.
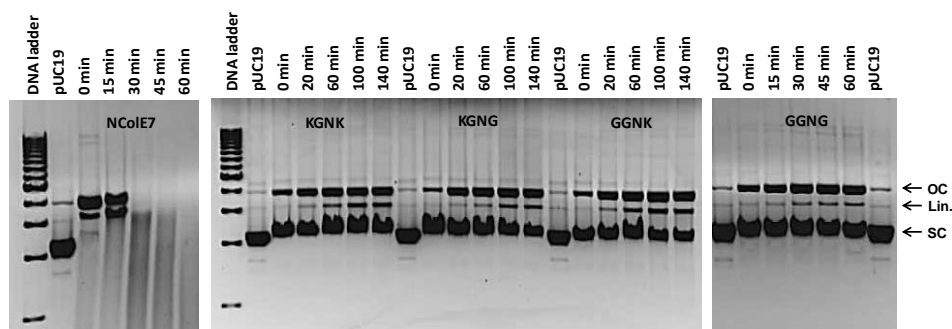
Fig. 5.6. *Digestion of 28 nM pUC19 (that is 74 μM calculated for base pairs) by 2.8 μM NColE7 mutants, incubated with one equivalent zinc(II)-acetate before mixing with DNA* [63]*. The supercoiled (SC) DNA is first cleaved on one strand resulting the open circular (OC) form. Cleavage on the opposite site of the single strand cleavage results in the linear (Lin.) form. The samples were kept at 37 °C and run subsequently on 1% agarose gel. 1 kb Molecular Ruler (BioRad) served as the reference.*

As **Fig. 5.6.** shows, the supercoiled form of the plasmid already disappeared at the first measurement point (approximately 2 minutes after mixing the solutions) in the presence of NColE7, while it was still detectable after 140 minutes in the DNA solutions containing the NColE7 mutants. The increasing amount of linearized plasmid could be observed in case of the mutants indicating the slow reaction.

The DNA-cleavage was further investigated by FLD-spectroscopy with a different type of DNA molecule. C*alf thymus* (CT)-DNA consists of extended linear and hence easily orientable chains. As the nuclease cleaves the CT-DNA the chains shorten and thus aquire a lower degree of orientation in the Couette flow cell yielding an LD intensity decrease in absolute value (**Fig. 5.7**). The absorbance was constant during the reaction, indicating that no DNA was lost e.g. by precipitation during the experiment. The LD intensities at 260 nm were evaluated as a function of time **Fig. 5.7 b**.
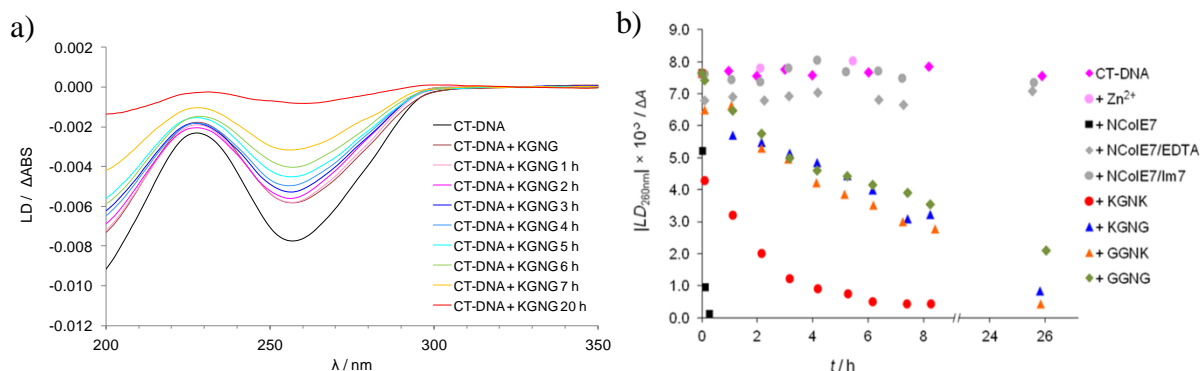
Fig. 5.7. a) *Cleavage of 130 μM (base pairs) CT-DNA by the KGNK mutant (0.5 μM) followed by flow linear dichroism spectroscopy.* **b)** *The LD signal intensity at 260 nm is taken from baseline corrected spectra of aliquots of the stock solutions, incubated at 37 °C [63]. Control experiments included CT-DNA incubated at 37 °C without or in the presence of 0.5 μM ZnCl₂, 0.5 μM NColE7 and 60 μM EDTA or 0.5 μM NColE7 and 1 μM Im7. The proteins were pre-incubated for 30 min in the presence of one equivalent ZnCl₂.*

In agreement with the plasmid cleavage experiments the reaction of the WT NColE7 is very fast, and all the mutants cleave DNA significantly slower. The exact evaluation of the curves is difficult, because the effect of single- or double-strand cleavages of DNA have a complex effect on the orientability of the DNA chains and thus on the LD-spectra. From these experiments we concluded that KGNK retained higher activity than the mutants lacking two or more positive charges from the N-terminus. This implies, that the presence of both lysines facilitates the cleavage to a low extent. At the same time, there is only minor difference between KGNG or GGNK and GGNG. It is worth to note, that even GGNG had a low nuclease activity, which is in agreement with the observation that during the DNA cloning procedure it was cytotoxic. Nevertheless, all the mutants showed a dramatic decrease in nuclease activity confirming our previous idea, that the N-terminus of the protein is essential for the reaction. In the following we studied the features of the mutants to get a deeper insight into the role of the N-terminus.

5.1.4. Investigation of reasons that lead to decreased nuclease activity

The nuclease activity of the N-terminal point mutants decreased in the order KRNK >> KGNK > KGNG ~ GGNK > GGNG. This is in agreement with the observation that the R447A mutant (KANK) has the 15% of the WT activity [62], however, the structural reasons of this phenomenon have not been studied.

5.1.4.1. Solution structure and metal binding

The SRCD spectra of each mutants and WT NColE7 are shown in **Fig. 5.8 a**. The shape of the CD-spectra is the same and the minor differences in the intensity might be caused by differences in concentration [63]. The effect of Zn^{2+}-addition on NColE7 was also studied. It caused only small shift towards high wavelength in the 180-210 nm region, and intensity

decrease in the shoulder between 210 and 240 nm, that may be related to a charge transfer. The same phenomenon was observed for the mutants. This suggests that the binding of the metal ion does not significantly influence the solution structure of NColE7 and its N-terminal point mutants.
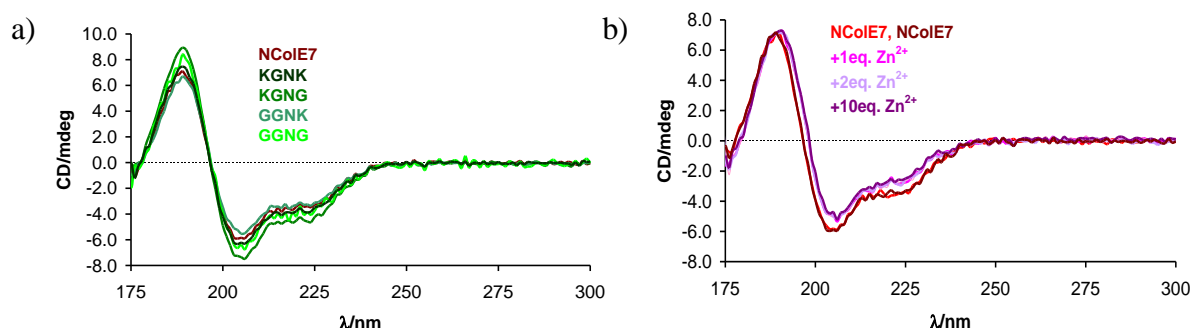
Fig. 5.8 . a) *SRCD spectra of the NColE7 variant proteins* [63]. **b)** *SRCD spectra of WT NColE7 in the absence (two independent measurements) and upon the addition of Zn^{2+}-ions* [63]. *All spectra were normalized to the same 31.6 μM protein concentration.*

The metal binding affinity was further quantified by isothermal microcalorimetric titrations with $ZnCl_2$ (**Fig. 5.9**). Based on the integrated titration curves the ΔH of binding in the selected buffer, as well as K_d were determined. As expected, the stoichiometry of binding is 1:1. The data obtained and shown in **Table 5.2** suggest, that the metal-binding affinity of the mutants is not significantly affected by the change of the N-terminus.
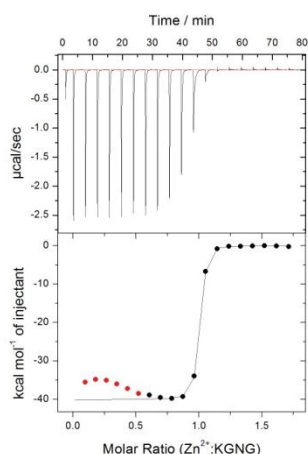
Table 5.2. *Apparent dissociation constants determined by the titration of the WT and mutant proteins with Zn^{2+}* [63].

	K_d / nM
NColE7	16.6 ± 3.5
KGNK	13.2 ± 1.5
KGNG	13.4 ± 1.4
GGNK	11.9 ± 2.6
GGNG	12.8 ± 1.9

Fig. 5.9. *Titration of the KGNG mutant with $ZnCl_2$. The part of the titration curve below the molar ratio of 0.5 could not be fitted with the applied model.*

5.1.4.2. Structure and dynamics of the proteins as studied by computational modeling

Since no obvious reason for the decreased activity could be identified, we performed calculations to get more insight also into the dynamics of the proteins. The models included NColE7 and the four mutant proteins, with a sequence 446-576, ignoring the 8 amino acids

part from the GST cleavage in the experiments. The Zn^{2+}-ion in the active center coordinated a phosphate ion at its free site. The root mean square deviation (RMSD) from the reference structure at 0 ps is higher for the mutants than for NColE7 (**Fig. 5.10 a**). The lysine containing mutants are similar to each other, but the increased RMSD for the GGNG mutant indicated that this structure is more flexible. It was shown that in the KGNK and KGNG mutants the side-chain of K446, while in GGNK and GGNG the N-terminal amino group approached the phosphate ion (**Fig. 5.10. b** and **c**). This supports the experimental observation, that these positively charged groups might be able to partially take over the role of the arginine and can promote the catalytic activity. However, for GGNG only a few structures with such conformation were found, and due to the high flexibility of this chain, the N-terminus often turned out from the active site.
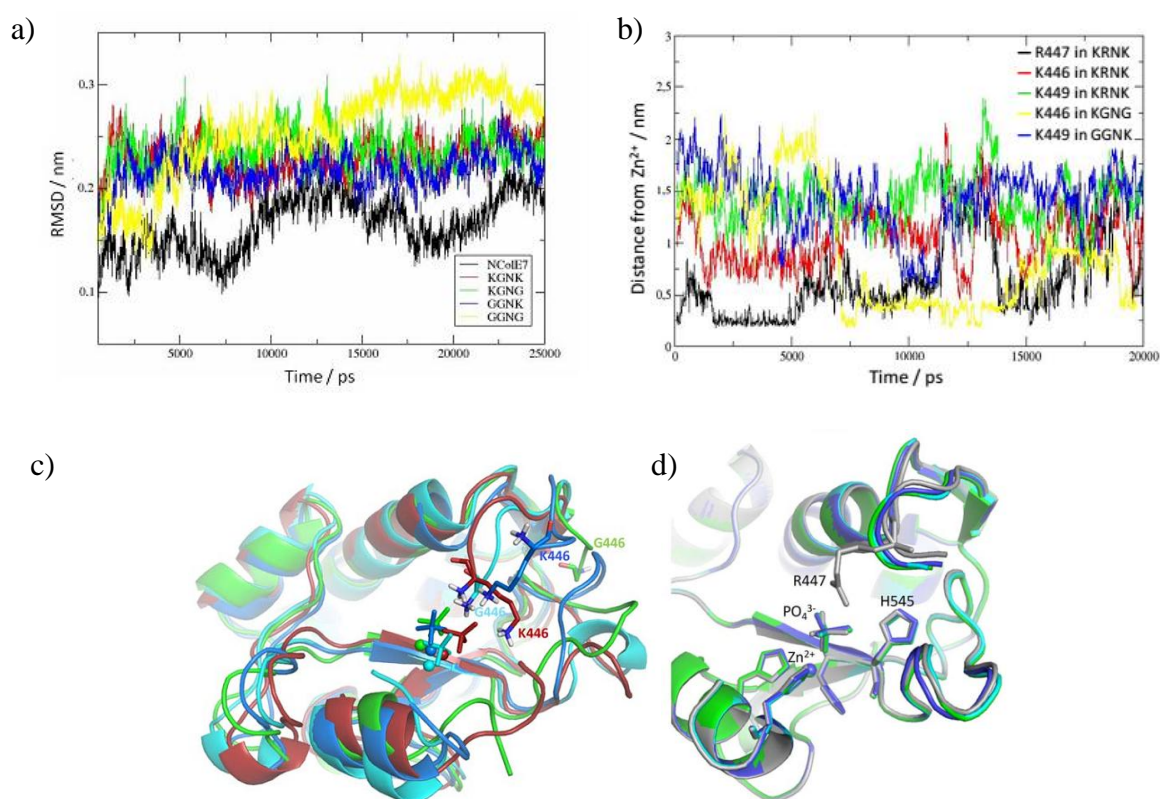
Fig. 5.10. a) *RMSD (root mean square deviation) of the studied proteins, as compared to the starting structure of the simulation* [63]. ***b)*** *Distance of the positively charged residues from the metal ion in the active center in the MD simulation of the zinc-bound proteins containing a phosphate ion in the active center.* ***c)*** *Snapshots at 20 ns of the simulations of the mutant proteins: KGNK in red, KGNG in blue, GGNK in cyan and GGNG in green. The Zn^{2+}-ions are shown by spheres and the phosphate ion by sticks.* ***d)*** *Optimized structure of the mutants and the WT NColE7. NColE7 is in grey, KGNG in blue, GGNK in cyan, and GGNG in green. Molecules were aligned by PyMOL.*

In order to get more information about the fine structure of the active site, the geometry of the proteins was optimized by semiempirical quantum chemical calculations, handling the whole molecule on the PM6 level [100, 103] in MOPAC2009 [109] (MOZYME method) with implicit water surroundings. The alignment to the optimized structure of NColE7 yielded an RMSD of 0.086 nm for KGNG, 0.087 nm for GGNK and 0.113 nm for GGNG. Small deviations from NColE7 occurred mainly at the mutated N-termini, but could also be detected

in the loop between the two β-strands of the HNH motif. The phosphate ion mimicking the scissile phosphodiester bond was also displaced (**Fig. 5.10 d**). This effect can be related to the decreased activity and supports the experimental data.

The calculations suggested, that the changes in nuclease activity might be related to changes in DNA-binding. Therefore we studied the DNA-binding of the mutants with different substrates and conditions.

5.1.4.3. DNA-binding with short dsDNA: solution study

Considering that neither the structure, nor the Zn^{2+}-binding of the mutants was changed, we hypothesized that anomalies in the DNA-binding of the mutants might be responsible for their decreased nuclease activity. It was earlier suggested, that the R447A mutant has a decreased DNA-affinity leading to the loss of 85% of nuclease activity [62]. The experiments in [62] were concluded from kinetic measurements conducted at a very low concentration ranges with an enzyme:substrate ratio close to stoichiometric. The K_M (Michaelis-Menten) constant increased to 106 nM from 27 nM suggesting the weakened DNA-affinity of the mutants. However, direct proof for weaker DNA-binding was not shown. An increased K_M (144 nM) was also obtained for the N560D mutant, even though this residue is clearly not involved in DNA-binding. We performed EMSA experiments (**Fig. 5.11**) to determine the K_d values of DNA-binding.
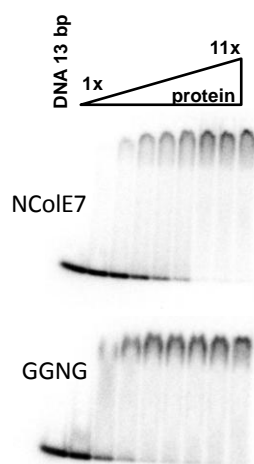
Table 5.3. The K_d constants of the protein-DNA complexes, determined from the integrated intensities of the shifted and unshifted bands of the EMSA experiment [63].

	K_d / µM
NColE7	0.16 ± 0.05
KGNK	0.15 ± 0.05
KGNG	0.32 ± 0.07
GGNK	0.30 ± 0.06
GGNG	0.30 ± 0.06

Fig. 5.11 Substrate binding of NColE7 and the GGNG mutant to a 13 bp hybridized radiolabelled oligonucleotide (0.2 µM) followed by gel mobility shift assay. The nuclease activity was inhibited by the addition of 0.5 mM EDTA. Protein concentration was varied as 0, 0.2, 0.4, 0.6, 0.8., 1.0, 1.4, 1.8, 2.2 µM for NColE7 and 0, 0.2, 0.6, 1.0, 1.4, 1.8, 2.2, 4.0, 7.0 µM for GGNG.

According to our experiments the mutants have a similar DNA affinity to the WT protein. This is in agreement with the fact that the main DNA-binding part of NColE7 is unaffected by the mutations. For the above reasons we think that the dramatic decrease in nuclease activity can not be simply explained by decreased DNA-affinity and we continued the study the DNA binding mode.

5.1.4.4. DNA-binding with short dsDNA: a QM/MM study

The DNA-binding experiments reflected the affinity of the mutant proteins, but not the fine structural details that may lead to the decrease of the nuclease activity. It is difficult to get information on atomic level about the active center of the protein, because the reaction should be frozen for the time of the experiment. We performed QM/MM calculations to identify differences in the catalytic region of the NColE7 protein and the mutant lacking the positive residues at the N-terminus (ΔN4-NColE7, sequence 450-576) in complex with DNA [110].

One of the two bases at the scissile phosphate (CYT11) is tilted in the optimized NColE7/DNA complex (**Fig. 5.12 a**) and is no longer coplanar with its guanine pair. This might indicate a strain in DNA facilitating its hydrolytic cleavage. In contrast, in the ΔN4-NColE7/DNA optimized structure the CYT11 is nearly coplanar to its pair on the opposite strand. Furthermore, in NColE7/DNA complex, the P–O5' bond length is 1.61 Å, and the P–O3' (scissile) bond is 1.65 Å, the difference which is significant in the optimized structures. The same distances in ΔN4-NColE7 change to 1.62 Å and 1.64 Å, respectively (**Fig. 5.12 b**) suggesting that the cleavage by ΔN4-NColE7 is less probable.
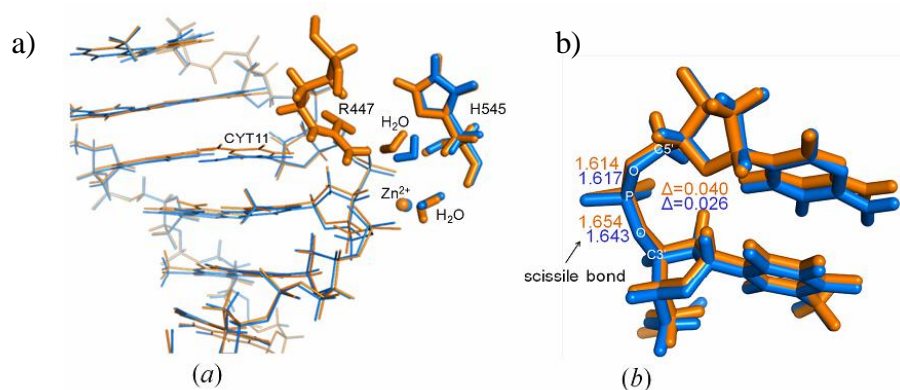
Fig. 5.12. a) *DNA binding of NColE7 (orange) and ΔN4-NColE7 (blue) in the QM/MM optimized structures* [110]. *Major differences appear in the position of the H545 and its hydrogen-bonded water molecule and in the position of the plane of CYT11.* **b**) *P–O bond lengths within the scissile phosphodiester bond in the same structures* [110].

A water molecule that was weakly bound (R(Zn–O) = 3.7 Å) in the starting structure approached the metal ion to 1.9 Å and 2.0 Å for ΔN4-NColE7 and NColE7, respectively. This is a similar type of coordination to that observed experimentally in an NColE9 crystal structure (PDB ID: 1FSJ [111]). While the Zn^{2+}-coordination is similar in the compared structures, as shown by Zn^{2+}-donor atom distances listed in **Table 5.4**, the difference between the P-O(3') and P-O(5') bond lengths of the scissile phosphate is only significant in the optimized active structure, but not in the crystal structures.

Table 5.4. *Comparison of the NColE7/DNA crystal structures with the QM/MM optimized structures* [110]. *Bond lengths are written in Å.*

Structures	Zn^{2+}-O(DNA)	P-O (3')	P-O (5')	ΔP-O (3'-5')
1PT3 (without Zn^{2+})	-	1.607	1.588	0.019
3FBD (without Zn2$^+$)	-	1.616	1.594	0.022
2IVH (H545Q)	1.769	1.604	1.596	0.008
1ZNS (H545E)	2.017	1.591	1.579	0.012
ΔN4-NColE7/DNA calc.	2.003	1.643	1.617	0.026
NColE7/DNA calc.	2.067	1.654	1.614	0.040

It must be, emphasized that all the experimental results were obtained for inactive forms of the protein, i.e. either the metal ion was absent or a functional residue was mutated. This difference between the computed NColE7/DNA and ΔN4-NColE7/DNA complexes may indicate the slight changes in the active centre that finally contribute to the loss of nuclease activity upon deletion of the N-terminal positively charged residues.

5.1.4.5. DNA-binding with short dsDNA: crystallographic study

The KGNK mutant was cocrystallized with a 18 bp DNA in the presence of EDTA to avoid any nuclease activity. Crystals were obtained in sitting drops, with a reservoir containing 0.2 M NH_4Cl (pH = 6.3) and 20% PEG3350. Some of the crystals were soaked into Zn^{2+}-acetate solutions (10-20 mM) for a short time (20-120 s) before freezing.
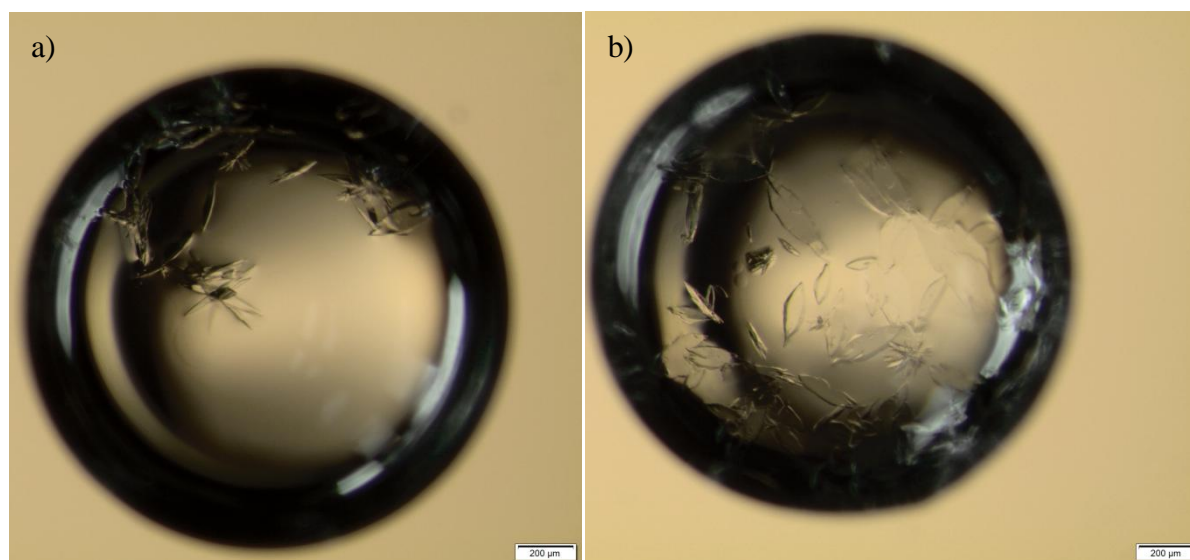
Fig. 5.13 *The 4 μl drops containing the crystals 10 days after set-up of the experiments. The scale bar shows the length of 200 μm.* **a)** *The drop containing the crystal of the KGNK/DNA complex: 3 μl of sample (124 μM dsDNA, 124 μM KGNK, 2.5 mM EDTA) mixed with 1 μl reservoir solution (0.2 M NH_4Cl pH = 6.3, 24% PEG3350).* **b)** *The drop containing the crystal of the KGNK/DNA/Zn^{2+} complex: 3 μl of sample (155 μM dsDNA, 155 μM KGNK, 2.5 mM EDTA) mixed with 1 μl reservoir solution (0.2 M NH_4Cl pH = 6.3, 28% PEG3350) Before freezing the crystal was soaked for 5 s into a solution of 45% universal cryoprotectant and 10 mM $ZnOAc_2$.*

Data sets have been collected by David Flot at the ESRF (European Synchrotron Radiation Facility) in Grenoble. The images of the drops containing the best diffracting crystals are shown in **Fig. 5.13**. and preliminary results on structure solution are shown in **Table 5.5**.

Table 5.5. *Preliminary structure refinement statistics of the KGNK/DNA complexes with and without Zn^{2+} soaking.*

	without Zn^{2+}	with Zn^{2+}
Space Group	$P\ 1\ 2_1\ 1$	$P\ 1\ 2_1\ 1$
Resolution range / Å	45.11 - 2.20	45.53 - 2.70
Unit cell a, b, c (Å)	60.99, 91.8, 102.87	61.2, 92.22, 102.34
Unit cell α, β, γ (°)	49.57, 90.00, 90.00	49.26, 90.00, 90.00
Total reflections	96679	53757
Unique reflections	26366	14792
Average multiplicity /Redundancy	3.7	3.6
Completeness (%)	95.9	98.5
R_{merge} (%)	0.106	0.180
Number of molecules per asymmetric unit	2	2
Refinement R_{work}/R_{free}	0.20/0.23	0.21/0.28
R.m.s. deviation of bonds / Å	0.016	0.014
R.m.s. deviation of angles / °	1.65	1.46
Number of protein residues	264	264
Number of DNA nucleotides	72	72
Number of water molecules	260	94
Number of bound Zn^{2+}-ions	0	2
Ramachandran plot outliers / %	0.38	1.2

Both crystals had the monoclinic $P\ 1\ 2_1\ 1$ symmetry with two protein/DNA complexes in the asymmetric unit (**Fig. 5.14**).
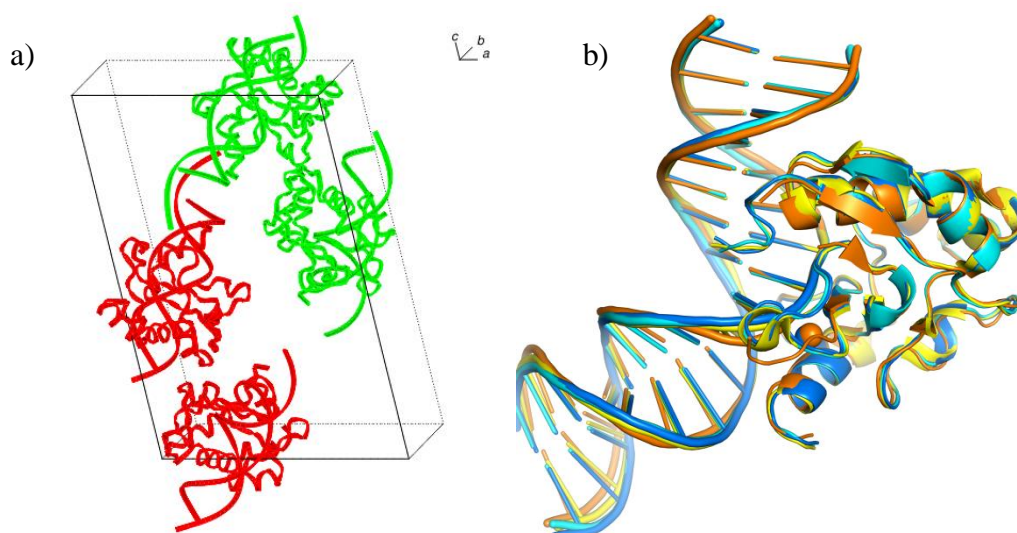
Fig. 5.14. a) *Representation of the monoclinic cell of the KGNK/DNA structure. The two red complexes include the A-B-C and D-E-F chains of the models. These are similar to each other and compose an asymmetric unit. The green complexes are another copy of the A-B-C and D-E-F chains, and can be obtained by a symmetry operation from the red ones.* **b)** *Overall structures compared to structures published earlier: light blue KGNK without Zn^{2+} (this work), dark blue KGNK with Zn^{2+} (this work), yellow NColE7 D493Q without Zn^{2+} (3FBD [48]), orange NColE7 H545Q with Zn^{2+} (2IVH [53]).*

As shown in **Fig. 5.14 b**, the overall structure of the proteins is similar to the structures published earlier. This was expected, since all NColE7 structures published so far, including mutants such as H545A, H545Q, N560D, H573A and D493Q had similar structures, regardless of complex formation with DNA or Im7.

According to structure refinement the Zn^{2+}-soaking was successful, a peak could be seen in the electron density map at the metal binding site, as shown in case of the chain *A* in **Fig. 5.15 a**. However, the occupancy is low (0.56 for chain *A* and 0.45 for chain *D* with B-factors 23.9 and 21.5, respectively). This means, that around half of the molecules in the crystal took up the Zn^{2+}-ion and the two molecules in the asymmetric unit are virtually identical.
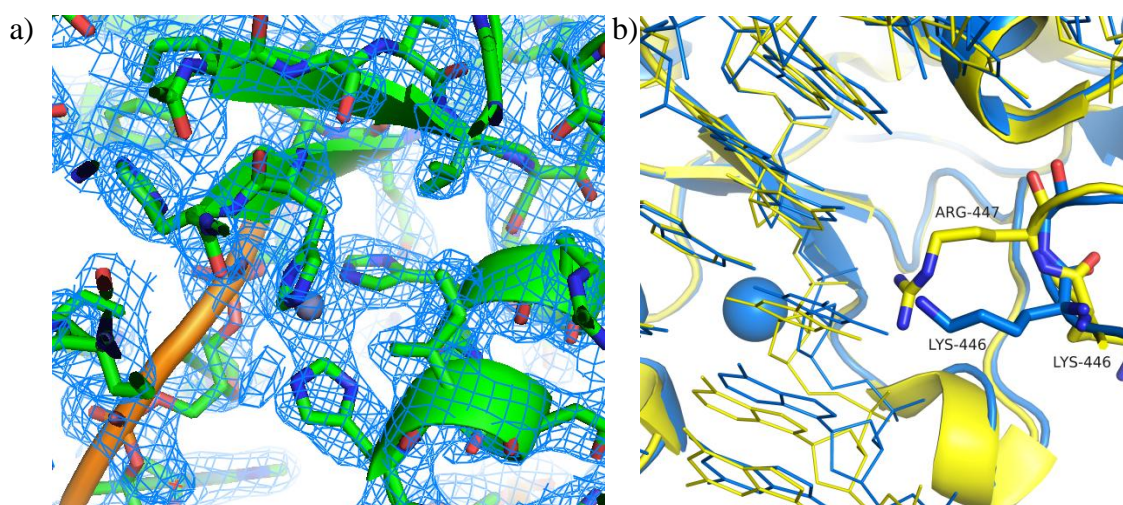
Fig. 5.15 a) *Electron density map near the metal binding HNH-motif in the KGNK/DNA crystal structure (chains A, B and C).* **b)** *Comparison of the N-termini of NColE7 D493Q without Zn^{2+} 3FBD [48] in yellow and the KGNK/DNA complex with Zn^{2+}-soaking (blue).*

As shown in **Fig. 5.15 b**, the sidechain of K446 might be oriented to point towards the phosphate group similarly approached by R447 in the WT NColE7. However, the themperature factors of this sidechain are high (ranging from ~30 at the side chain nitrogen to ~60 of the aliphatic part). In case of such high temperature factors the interpretation of the electron density is ambiguous. Nevertheless, these results confirm our computational results (Chapter 5.1.4.2) predicting that K446 approaches the active center in the absence of R447 and takes over its role.

The conformation of the DNA in the KGNK structures slightly differs from the other published structures that contain the same DNA sequence (**Fig. 5.16**).
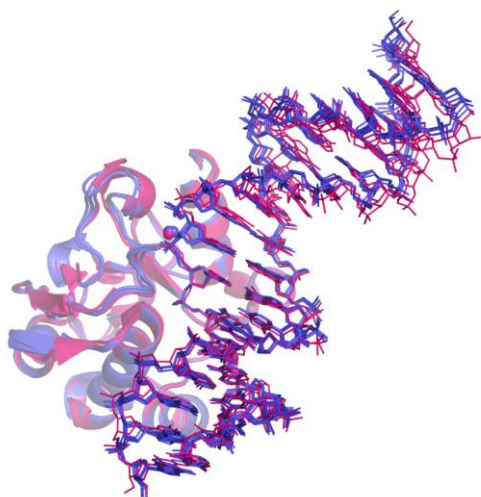
Table 5.6 *Global bend angle as determined with the 3D-Dart webserver* [112]

Structure	Bending
H545Q-NColE7, 2IVH [53]	55.2
D493Q-NColE7, 3FBD B-C [48]	43.8
D493Q-NColE7, 3FBD E-F [48]	42.9
KGNK/DNA B-C	42.8
KGNK/DNA E-F	43.8
KGNK/DNA/Zn^{2+} B-C	44.0
KGNK/DNA/Zn^{2+} E-F	41.6

Fig. 5.16 *Comparison of the KGNK/DNA(/Zn^{2+}) structure (blue) to the other two NColE7 structures (red) published with the same DNA.*

The bend angle is considerably smaller than in the H545Q-NColE7, but similar to the D493Q-NColE7. The D493Q mutation is at the protein-DNA interface that might also influence the conformation of the bound DNA. This mutant showed an increased nuclease activity compared to the WT enzyme. Therefore, the small difference in DNA-bending is probably not leading to the significant decrease in the nuclease activity of the KGNK mutant. However, the local DNA structure close to the scissile bond is different from DNA bound by NColE7. Small differences in the substrate binding may lead to severe decrease in enzymatic activity.

5.1.4.6. Binding of long DNA substrates

In the natural environment the substrate of the enzyme is a long DNA chain. We also performed EMSA experiments with long DNA-substrates to compare the DNA-binding properties of the mutants to that of NColE7. NColE7 is nonspecific, therefore more than one molecule can bind to the substrate. Thus smears rather than well defined bands for the protein-DNA complexes are expected (See chapter 4.3.1). We performed EMSA experiments with different length of dsDNA (188 bp, 306 bp, 777 bp), examples are shown in **Fig. 5.17**.
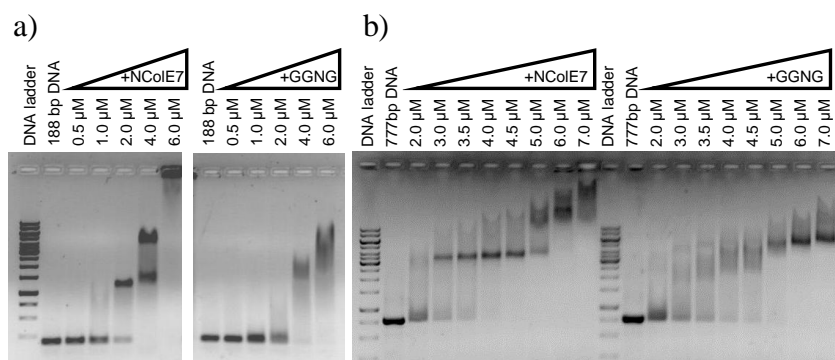
Fig. 5.17. *EMSA experiments performed with different length of substrates, e.g. 188 bp (**a**) or 777 bp (**b**) revealed differences between the WT NColE7 and the N-terminal point mutants. The sharp shifted bands of NColE7 suggest cooperative DNA-binding that is not the case with the point mutants.*

Comparing the decrease in the free DNA band intensity we can conclude, that these experiments are in agreement with the finding, that the DNA-binding affinity is similarly strong for all mutants and WT NColE7. However, there are clear differences between the gel profiles regarding the protein-DNA complexes. While for the WT NColE7 we observed sharp bands, the mutants showed smears: the less positive charge on the N-terminus, the more smeared the bands. The sharp bands seen in case of WT NColE7/DNA complexes is surprising because of the above arguments on nonspecific nucleases. We speculate, that the reason of the phenomenon could be a cooperative DNA binding process, resulting in complexes of a preferred composition. The mutations can disrupt this by either altered protein-protein interactions between molecules binding to the substrate or conformational differences in the DNA-binding mode. For example, the bending of the DNA-chain can be different. To confirm this, AFM experiments were performed. In a recent study [98] it was shown that the cooperative binding can be seen as a DNA molecule saturated by proteins, therefore higher and wider profile on the surface, while the noncooperative mutant could be seen as small spheres on the DNA chain. As shown in **Fig. 5.18**, we found the opposite phenomenon, the saturated DNA-chain could be observed for GGNG that is supposed to be noncooperative, while for the WT NColE7 individual molecules were observed. If the GGNG mutant is not able to bend DNA, but only to bind it, so that the DNA chain is accessible for further protein molecules, and the binding sites are independent. At the same time NColE7 bends DNA, and only upon the conformational change the DNA becomes accessible. However, the shape of the complex in solution could not be determined, since on the AFM sheets the molecules adopt two dimensional conformation during the drying process.
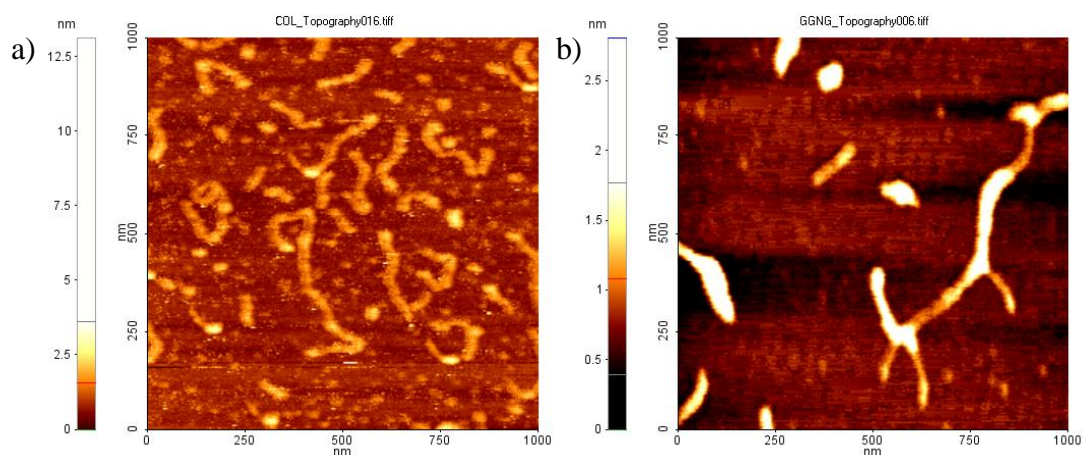
Fig. 5.18. *AFM images of NColE7 (**a**) and GGNG (**b**) with a 777 bp DNA, plated on mica sheets with MgCl$_2$.*

Our experiments with long DNA substrates draw attention to the need for careful interpretation of the experimental data with the short DNA oligonucleotides frequently used as a model of natural substrates. However, it is difficult to understand how the differences in cooperative binding influence the nuclease activity of NColE7 and of the mutants. More investigations are necessary to precisely describe this phenomenon. Since our goal is the

design of a controlled specific ZFN, our main interest is the understanding of the reaction between one enzyme and one substrate molecule.

5.1.5. Conclusions: revision of NColE7 mechanism

We performed mutational studies to get a deeper insight into the role of the N-terminus in the NColE7 nuclease. The observed order in nuclease activity is KRNK(NColE7) >> KGNK > GGNK ~ KGNG > GGNG. Despite of the dramatic effect on enzymatic activity, many features of the mutants are similar to the WT enzyme: the solution structure, Zn^{2+}-binding, DNA binding as well as the crystal structure of the KGNK mutant revealed only minor differences. Therefore, we suspect that the N-terminus may have a direct role in the catalysis.

The mechanism of NColE7 has been widely studied (Chapter 2.4.4), however the protonation of the leaving group is yet a matter of question. Considering the above observations we can not exclude that the flexible arginine side-chain mediates the proton transfer. After protonation of the leaving group it may become instantly re-protonated by H545. These two residues get close to each other and a water-mediated hydrogen bond is formed between them during the catalytic cycle (**Fig.5.19**).
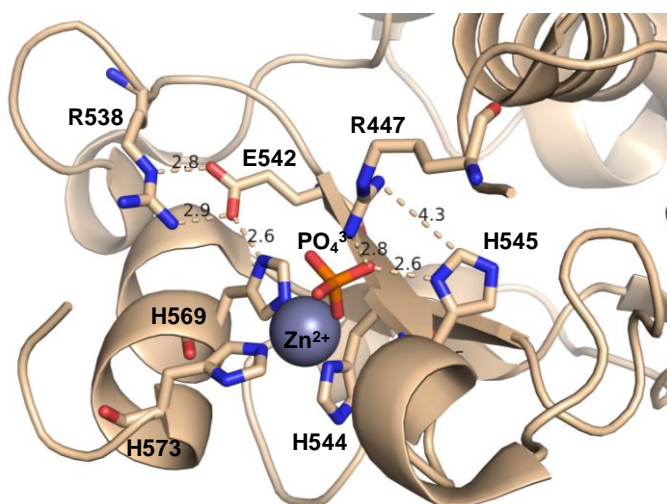
Fig.5.19 Short interatomic distances reflecting possible hydrogen bonds in the active site of NColE7 (PDB: 1M08) [110]. *The hydrogen bonds among R538, E542 and H569 (the H569 residue also binds the Zn^{2+}-ion) form a putative proton channel.*

The initial source of the proton on H545 is the water molecule, which has to become deprotonated to perform the nucleophilic attack at the partially positively charged phosphorus atom. This would be in agreement with the recently proposed shuttle mechanism according to which the leaving group is protonated by the hydrogen ion originating from the same water molecule that initiated the nucleophilic attack [58]. The above mechanism would be similar to the one found in a serine recombinase-mediated DNA cleavage [113]. Arginine is rarely mentioned in the literature to behave as an acid but the environment of an enzyme active site can significantly shift the pK_a values of critical residues. However, it can be considered plausible in biological systems that arginine acts as a base [114], especially in the presence of such a basic group as it is the 3'-alcoholate ion. This would provide a reasonable answer to the still unsolved but intriguing question about the identity of the general acid that protonates

the leaving group in NColE7. It shall also be mentioned that the multiple roles of R447 can be partially replaced by other positive charged residues, such as lysine side-chains or the N-terminal amino group in a suitable position.

The question remains why the N-terminal point mutants all retain some catalytic activity in contrast to the ΔN4-NColE7, where all these residues are missing [43]. According to molecular dynamics simulations and the crystal structure of the KGNK/DNA/Zn^{2+} complex K446 may take over the role of R447, but less efficiently due to steric reasons. On the other hand, even the presence of the backbone may promote the reaction *e.g.* by influencing the stability of the protein structure or the inhibition of reverse reaction.

Based on the results presented in this chapter we collected information that is important for the design of a controlled artificial nuclease:

1) In contrast to data published earlier [62] we concluded that the low nuclease activity resulted by the mutation of the N-terminal positively charged sidechains can not be simply explained by decrease in DNA-binding affinity of NColE7 mutants [63].
2) The positively charged residue(s) might take part in the reaction mechanism, but its position is not strictly determined [63, 110].
3) The flexibility of the N-terminal chain can be important for its function.
4) The N-terminus of NColE7 is appropriate to build an intramolecular allosteric control, because all important features of the protein (structure, metal- and DNA binding) remain unaffected by the N-terminal mutations, while the activity is decreased. [63]

5.2. THE FUNCTION OF THE N-TERMINAL CHAIN CAN BE ALTERED

After establishing the role of the N-terminus, it was intriguing, how we could influence its function in order to get ideas on how to redesign the protein for ZFN applications. As a first step, we analyzed available crystal structures and did computational modeling to identify the potentially important residues in the further parts of the N-terminal chain. We considered the residues 450-470, a long loop on the surface of the protein. This part of the sequence is interesting, because it orients the N-terminus to its position, but so far it has not been studied. Our goal was to decrease the number of interactions between the N-terminal chain and the core of the protein. The results were expected to reveal whether the N-terminal KRNK part can individually function with an altered loop of weak interactions.

The experimental conditions and computational details related to this chapter are collected in the Appendix, chapter 10.2.

5.2.1. Computational design of mutations

Semiempirical calculations were performed to predict which residues are structurally important. A virtual alanine scan was carried out within the 25 residues long N-terminal sequence of NColE7 – i.e. residues 446-470 – to obtain an estimate of the contribution of each residue to the structural stability of NColE7 [115]. The effect of the exchange of each amino acids – except for Gly and Ala – to Ala was monitored one at a time. The alignment of the optimized structures of alanine mutants to that of the native protein resulted in small differences (RMSD for the whole sequences aligned in PyMol was between 0.022 and 0.174 nm).

After the initial optimization the Φ backbone dihedral angle between the 25th and 26th residues, (470K and 471D, respectively) was altered from 290° to 30° so, that the N-terminal sequence turned away from the core of the protein (**Fig. 5.20 a**). The geometry of the unfolded structure was optimized and the contribution of the N-terminal loop to protein stability was estimated as the difference in the enthalpy of formation of the unfolded and folded forms of the protein, as given by the equation:

$$\Delta\Delta H_f = \Delta H_{f,unfolded} - \Delta H_{f,folded}$$

The value computed in this way is independent from the number of atoms in the molecule, reflecting only the interactions between the N-terminal segment and the core protein or solvent. The same procedure was carried out with all alanine mutants allowing an estimate of the contribution of each residue to the binding interactions of the N-terminal sequence.
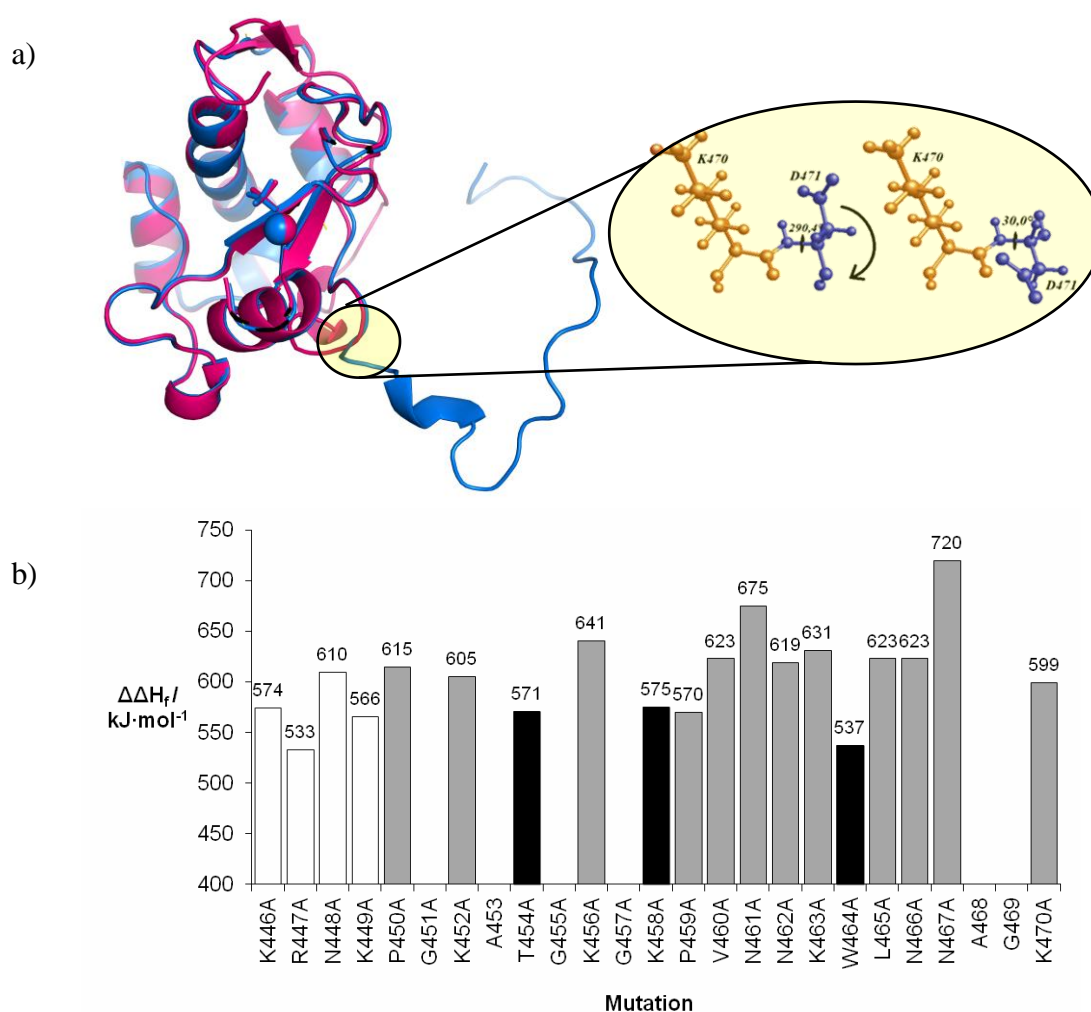
Fig. 5.20. a) *Structure of the NColE7 protein in native state (red) and with unfolded N-terminus (blue). The altered torsional angle is magnified. b) The effect of alanine mutations in the N-terminal chain on the enthalpy of formation* [115]. *If an important residue is mutated to alanine, the enthalpy of formation drops, therefore the three residues with low values (in black) were selected for further study.*

The mutation sites selected for experimental study are shown in **Fig. 5.20 b**, and their location in the NColE7 structure is shown on **Fig. 5.21 a** while their sequence is depicted in **Fig. 5.21 b**. T454 is in interaction with the loop located between the β-sheets of HNH motif. This is a structurally oriented loop, containing an essential hydrogen-bond chain that keeps H545, the general base in an optimal position for generating the water molecule for nucleophilic attack. Therefore, its mutation was expected to influence the catalytic features. K458 has interactions with V476, which is a part of the core of the protein, so this residue can be also important. W464 is located in a hydrofobic pocket, and that can stabilize this part of the protein structure.

a)

b)

```
NColE7   438-GPLGSPEFKRNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
TKW      438-GPLGSPEFKRNKPGKAAGKGAPVNNKALNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL

NColE7   SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
TKW      SKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK-576
```

Fig. 5.21. a) *The location and interactions of the amino acids selected for the experimental mutation study. T454, K458, W464 and the interacting counterparts are shown by sticks in the X-Ray structure with PDB ID: 1M08* [50]*. Distances between selected atoms are shown in Å. The HNH-motif is shown in red and the N-terminal loop in blue.* ***b)*** *Sequence of the TKW mutant, compared to NColE7. The beginning 8 residues (grey) are the result of the applied purification procedure (remaining from the GST-tag). The N-terminal loop is written in blue with the mutations in magenta. The HNH-motif is written in red.*

5.2.2. Mutations in the N-terminal chain lead to decreased nuclease activity

5.2.2.1. Protein expression and purification

The NColE7 containing the T454A, K458A and T454A mutations is called TKW in the following. The cytotoxic effect during gene cloning was observed also in this case therefore this protein was expressed and purified with the same method as the N-terminal point mutants. The gene was cloned together with that of the Im7 protein, and after the GST-affinity chromatography cation exchange at pH = 3 was performed [115]. The purified protein was validated by SDS-PAGE and MS.

5.2.2.2. Nuclease activity

Similarly to the N-terminal point mutants, TKW was shown to have a low nuclease activity (**Fig. 5.22**) [115]. Compared in LD-experiments the nuclease activity decreased in the order KGNK > TKW > KGNG.
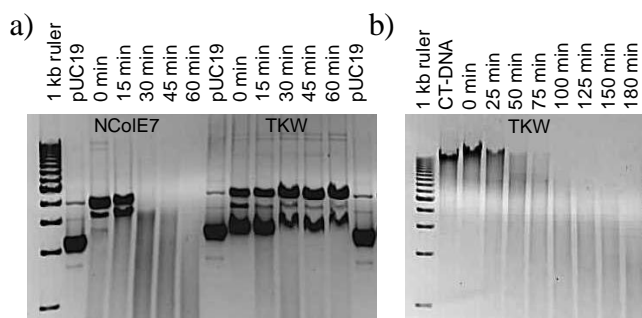
Fig. 5.22. a) *The cleavage of 74 µM (base pairs) pUC19 plasmid by 2.8 µM protein in the presence of 2.8 µM zinc(II)-acetate at 37 °C.* ***b)*** *To confirm the cleavage, reaction with 68 µM (base pairs) CT-DNA by 2.6 µM TKW and 2.6 µM zinc(II)-acetate at 37 °C was also studied. The 1 kb molecular ruler (BioRad) was used as a reference on both gels.*

The result is surprising, since none of the functionally important parts of the protein have been modified. As shown in chapter 5.1.4, the protein folding, Zn^{2+} and DNA binding affinity was not affected by the N-terminal point mutations suggesting the direct role of the N-terminal positively charged residues in the catalytic mechanism. However, the KRNK sequence is not affected in the TKW mutant, so its low nuclease activity may be attributed to other reasons.

5.2.3. The decrease in activity is related to protein structure and Zn^{2+}-binding

5.2.3.1. Protein structure

SRCD spectra of NColE7 and TKW recorded under the same conditions reflected that the mutations drastically affect the CD pattern (**Fig. 5.23 a**). The CD-spectra of proteins in the $\lambda = 170\text{-}260$ nm region is mainly determined by the secondary structure elements, such as α-helix and β-sheet. The change in the conformation of the modified loop is not expected to influence the spectrum so significantly. The fractions of the secondary structure elements estimated from the SRCD spectrum for NColE7 are as follows: 31% and 12% for α-helices and β-sheets, respectively. This is in good agreement with the values obtained from the crystal structure of NColE7 (PDB ID 1M08): 38% of α-helices and 16% of β-sheets. At the same time the significantly different SRCD spectral pattern of TKW mutant yields 13% α-helices and 21% β-sheets. Although these estimations are based on statistics (from protein CD-databases) and thus not precise, the numbers suggest that the interactions of the T454, K458 and W464 residues are widespread and their mutation affected the overall protein structure.
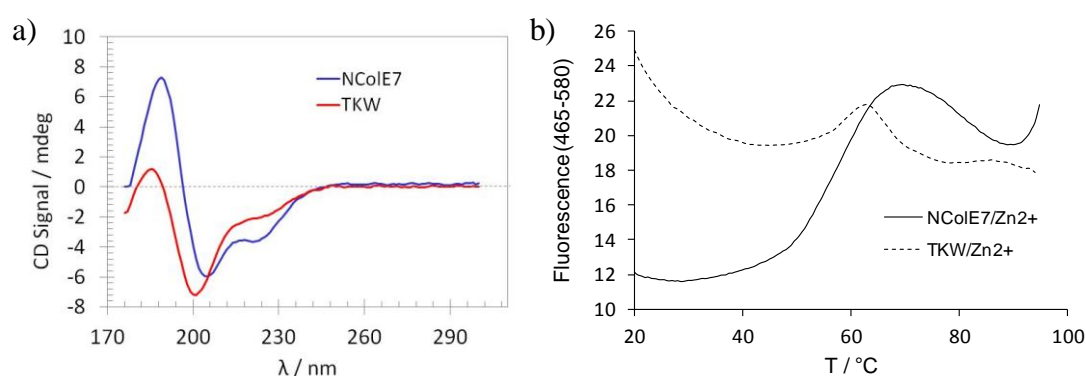
Fig. 5.23. a) *CD-spectra of the NColE7 and TKW proteins* [115]. ***b)*** *Thermal shift assay of TKW and NColE7 proteins (4 mg/ml) in the presence of Sypro Orange dye* [115]. *The melting point of NColE7 is 59 °C, as determined by the inflection point of the curve. The decreasing curve of TKW indicates that the protein is partially unfolded already at low temperature.*

The stability of the TKW and NColE7 protein structures was compared in a thermal shift assay using the Sypro Orange fluorescent dye (Invitrogen). The fluorescence of the dye is quenched by water, but once it binds to the hydrophobic regions of the unfolded protein, the fluorescence is increased. The melting point of NColE7 was 59 °C (**Fig. 5.23 b**) being the same as earlier published for colicin E9 [116]. The decrease of the fluorescence intensity in

the thermal curve of TKW supported that the protein is already partially unfolded resulting in uncertain melting point determination.

To reveal the structural changes we tried to crystallize the TKW protein. The experiments included sparse matrix screens and systematic screens, additives, alltogether more than 700 conditions. The crystals were proven to consist of protein molecules, but did not diffract the X-Ray beam. This also indicates that the protein is not misfolded but has disordered parts. **Fig. 5.24** shows some examples of the crystallization drops.

Fig. 5.24 a) *An example of the commonly observed shape of protein precipitate (30 mg/ml protein, 0.2M lithium sulfate, 0.1M phosphate-citrate pH 4.2, 20%w/v PEG1000)* ***b)*** *An example of the protein and PEG concentration based optimization screen of the condition 30 mg/ml protein, 0.2 M lithium sulfate, 0.1 M sodium acetate of pH 4.5 and 30 w/v% PEG 8000* ***c)*** *The protein was crystallized with the Izit dye in the same condition as in point b). The crystals adsorbed the dye which means they consist of protein molecules.*

5.2.3.2. Metal-ion binding

SRCD spectra of TKW in the presence of increasing Zn^{2+} concentration showed that Zn^{2+} has an impact on the structure of the protein, but not even ten equivalents of the metal ion could saturate it (**Fig. 5.25**). Nevertheless, the CD pattern approached that of the NColE7-Zn^{2+} complex. A higher excess of Zn^{2+}-addition could not be measured due to precipitation. The increased sensitivity of the CD-spectrum on metal binding implies that the structural changes included the metal-binding active centre.
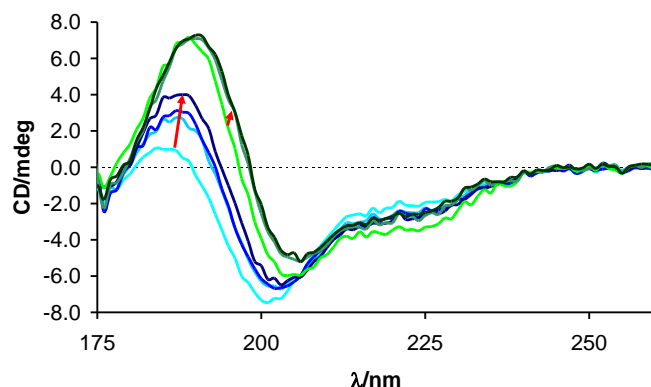
Fig. 5.25. *SRCD spectra of the NColE7 (green) and TKW mutant (light blue) proteins (31,6 μM in 20 mM HEPES, pH 7.4) and the titration of the proteins with Zn^{2+}-ions (the latter curves were corrected for dilution)* [115]. *The arrows show the effect of increasing Zn^{2+}:protein ratio from 0:1 to 10:1.*

The differences in Zn^{2+}-binding were also well reflected in the ESI-MS spectra [115]. The observed molecular weight for NColE7 corresponded to ~100 % holo form after the addition of one equivalent Zn^{2+}-ion. However, the mass spectra of TKW with one equivalent of Zn^{2+} showed a dual distribution with only around one third of the protein in the holo form

(**Fig. 5.26 a**) The highly charged peaks were also observed in case of NColE7 (data not shown) indicating that a fraction of the protein molecules got partially unfolded in the ionization process. However, while in NColE7 even the highly charged molecules contained the Zn^{2+}-ion, in case of TKW these peaks (around the *m/z* value 872, **Fig. 5.26 b**) corresponded to the apo form of the protein.

a)

b)
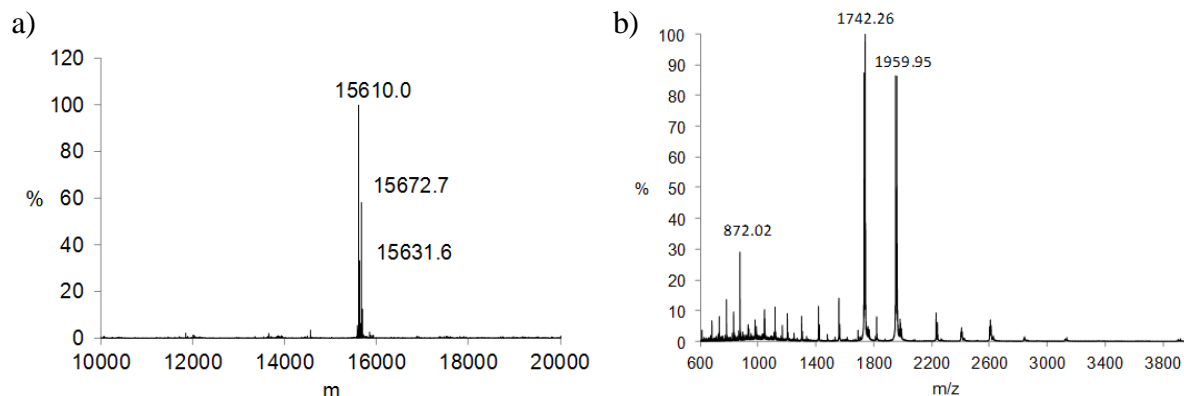
Fig. 5.26. a) *Mass spectrum of the TKW protein in the presence of one equivalent zinc(II)-acetate, calculated from the m/z spectrum, including both the high and low m/z value peaks* [115]. ***b)*** *A part of the nano ESI-MS m/z spectrum of 20 μM TKW mutant protein incubated with 20 μM zinc(II)-acetate before measurement* [115]. *The calculated mass from the most intense peaks at high m/z values corresponded to the holo form of the protein, while the envelope in the 600-1400 Da range allowed for calculation of the apo protein mass. This means that (i) the structure of the apo protein is probably less stable and thus highly charged, and (ii) a fraction of the protein did not contain Zn^{2+}-ion.*

The weak Zn^{2+}-binding of TKW was confirmed by isothermal microcalorimetric titration (**Fig. 5.27** and **Table 5.7**). TKW also binds Zn^{2+} with 1:1 stoichiometry, however the K_d value increased by three orders of magnitude to 7.7 ± 0.4 μM, as compared to of 16.6 ± 3.5 nM of NColE7 [115]. The enthalpic change was also lower for the mutant protein, than for NColE7.
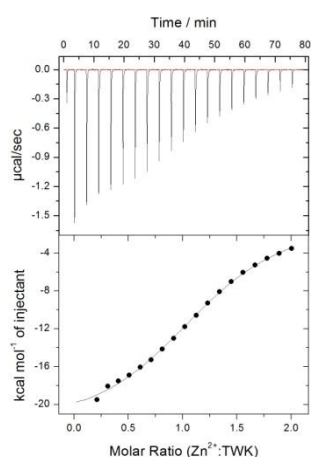
Table 5.7. *Data obtained for the one site model curve fit of ITC curves. The indicated errors refer to fitting errors.*

	NColE7	TKW
n	1.19 ± 0.002	1.24 ± 0.01
K (M^{-1})	$(6.03 \pm 1.23) \times 10^7$	$(1.30 \pm 0.7) \times 10^5$
ΔH (kcal/mol)	-32.9 ± 0.2	-22.7 ± 0.3
ΔS (cal/mol/K)	-74.7	-52.8

Fig. 5.27. *Titration of the TKW protein by $ZnCl_2$* [115].

5.2.3.3. DNA-binding

The binding of the DNA substrate was investigated by native PAGE gel mobility shift assay of a 13 base pairs long PCR fragment. Similarly to the N-terminal point mutants around the same concentrations of the native and the mutant protein were necessary to shift the DNA band (**Fig. 5.28 a**). The calculated K_d vales related to the DNA/protein complexes were 0.16 μM for NColE7 and 0.14 ± 0.05 μM for TKW [115]. This indicated that surprisingly, there was no difference in the strength of DNA binding. The EMSA experiments with large DNA (**Fig. 5.28 b**) revealed that the cooperativity of DNA binding is lost also in the case of this mutant.
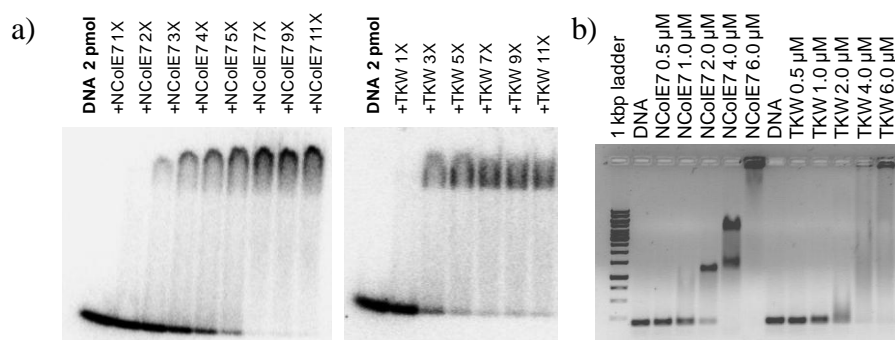
Fig. 5.28 a) *Substrate binding of NColE7 and TKW proteins to a 13 bp hybridized oligonucleotide (0.2 μM) followed by gel mobility shift assay. The nuclease activity was inhibited by the addition of 0.5 mM EDTA. **b)** Substrate binding of NColE7 and TKW proteins to a 188 bp dsDNA substrate (50 μM/bp) in presence of 0.5 mM EDTA.*

5.2.4. Induced folding promotes nuclease activity

Considering the low activity and widespread changes in the conformation of the TKW mutant, it is surprising that only minor difference was found in DNA-binding affinity. A possible explanation was obtained by SRCD spectroscopy (**Fig. 5.29**). The spectrum of the NColE7/DNA complex was identical to the sum of the component spectra [115]. Thus, no significant conformational change was detected in agreement with the identical crystal structures of the protein in the presence and absence of DNA. Although the folding of TKW has significantly been affected by the mutations, the SRCD spectrum of the TKW/DNA complex showed very similar pattern to that of the NColE7/DNA complex. This suggested that an NColE7-like structure of the TKW mutant has been induced by the interaction with DNA. Similar induced folding of a truncated staphylococcal nuclease has been observed upon addition of its inhibitor to the unfolded protein [117].
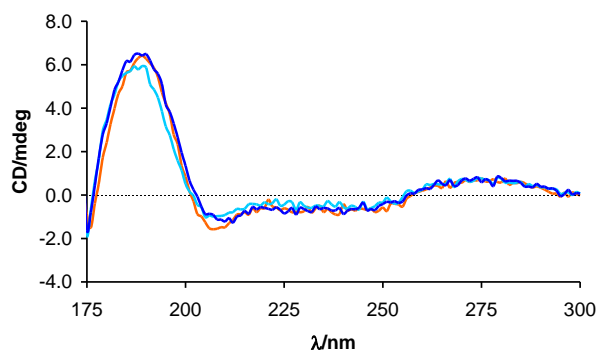
Fig. 5.29 The SRCD spectra of NColE7 (orange) and TKW (31,6 μM in 20mM HEPES, pH = 7.4) in the presence of DNA (light blue in the absence of Zn^{2+}-ions and dark blue with one equivalent of Zn^{2+}-ions) [115]. All curves were corrected for dilution.

The effect of the addition of one equivalent Zn^{2+}-ions to the protein/DNA complexes on the SRCD spectral pattern was comparable for both the NColE7 and TKW proteins showing that the DNA-binding mediated the preformed structure of the HNH motif, and thereby contributed to the stronger Zn^{2+}-binding.

5.2.5. Further mutants reveal the importance of the W464 residue

To determine wich mutation(s) contribute to the special features of the TKW mutant we expressed and purified double and single point mutants (TK, TW, W) with the same methodology as described for TKW. It turned out, that it is the W464A mutation that causes severe changes in the structure of the protein, and this correlated with the nuclease activity and Zn^{2+}-binding (**Fig. 5.30**).
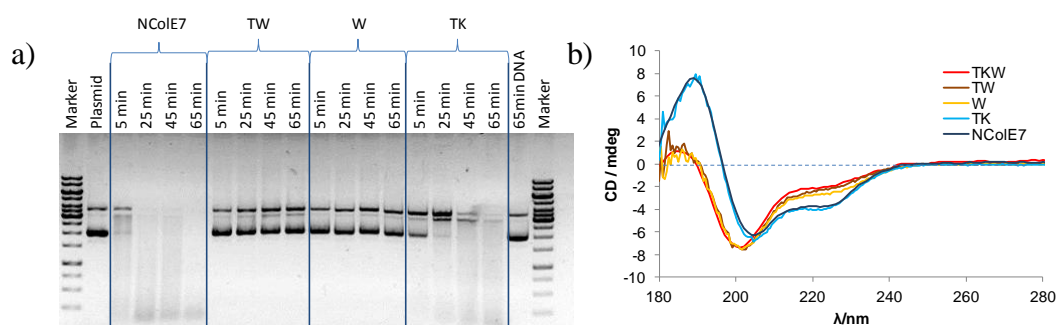
Fig. 5.30 a) *Nuclease activity assay performed with the double and single point mutants.* **b)** *Comparison of the CD-spectra of mutants with different combinations of the T, K and W mutations (16 μM, without Zn^{2+}-addition).*

Although the severe changes are related to the W464A mutation, the nuclease activity of the TK mutant also slightly decreased as compared to NColE7 indicating, that the interactions of these residues contribute to the function of the protein.

5.2.6. *Structure induction and preorganization of the Zn^{2+}-binding site*

The induction of the TKW structure by the DNA substrate is an important finding concerning the redesign of NColE7. It means that certain structural changes can be influenced and that provides opportunity for the regulation not only based on the mechanism of the enzyme but also on structural changes. The phenomenon was further studied with the Im7 protein. As discussed in Chapter 2.4.1, Im7 is a natural inhibitor protein of NColE7, hindering the substrate binding site by the formation of a stable complex. The N-terminal chain is not involved in the NColE7-Im7 interface, therefore it was interesting whether the complex formation with Im7 could improve the structure. As **Fig.5.31** shows, Im7 was able to perform a structural induction similarly to the substrate DNA.
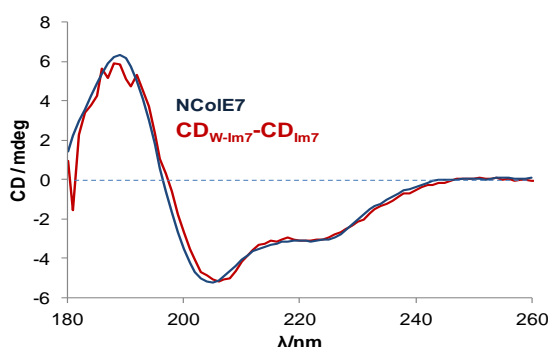
Table 5.8. *Zn^{2+}-binding of the mutant proteins with and without Im7.*

mutant	K_d (protein)	K_d (protein/Im7)
NColE7	16.6±3.5nM	61±18 nM
TKW	11 ± 1 μM	33±23 nM
TK	230±40 nM	not measured
TW	51±3 μM	not measured
W	5.6±0.3 μM	55±25 nM

Fig.5.31 *CD-spectra of NColE7 (black) and the W mutant in complex with the Im7 protein (red), the latter spectrum was obtained by subtracting the Im7 spectrum from the W/Im7 spectrum.*

As **Table 5.8** shows, all mutants having the W464A mutation have a decreased Zn^{2+}-binding affinity, while TK retained the strong metal binding property. This means that exactly those mutants have weak metal-binding affinity that also have a partially unfolded structure. These results confirm that the structural changes also affect the metal binding site. This also means that NColE7 has a preorganized metal-binding site that is disrupted in the W464A mutants. If the structure is induced, the preorganization occurs, confirmed by ITC measurements performed in the presence of Im7 protein.

5.2.7. *Conclusions: regulation of NColE7 via preorganization of the metal-binding site*

The TKW mutations (T454A/K458A/W464) are located in the N-terminal loop of NColE7, a part of the protein that is not directly involved in DNA-binding or cleavage, and is not part of any stable secondary structure elements. The dramatic effect on nuclease activity was therefore surprising, and could be explained by the severe structural changes. As shown by CD- and ITC-measurements on Zn^{2+}-binding, these changes also influenced the metal-binding active center. The preorganized metal binding site in NColE7 is unstable in TKW and can only partially be rescued by the addition of Zn^{2+}-ions. The DNA-mediated folding of

TKW leading to the preorganization of the metal binding site is a feature that could be exploited in the redesign of NColE7. The induction of the conformation was also achived by the complex formation with the Im7 protein. The induction of the structure was also followed by rescued Zn^{2+}-binding affinity.

The above phenomena were attributed to the W464A mutation out of the three mutation sites in the TKW mutant. The role of Trp in the stabilization the secondary structure of proteins is known [118]. An example is the staphylococcal nuclease [119]. The CD spectral changes of the W140A mutant were reported emphasizing the importance of the Trp residues in its folding. The deletion of 13 residues (including W140) from the C-terminus of this nuclease resulted an unfolded protein, however its folding could be induced under certain conditions, and it is active in presence of Ca^{2+} [117]. In the NColE7 protein the W464 residue is located in a hydrophobic pocket close to the HNH-motif (**Fig. 5.32**). This can explain the importance of this residue in the stability and function of the protein.
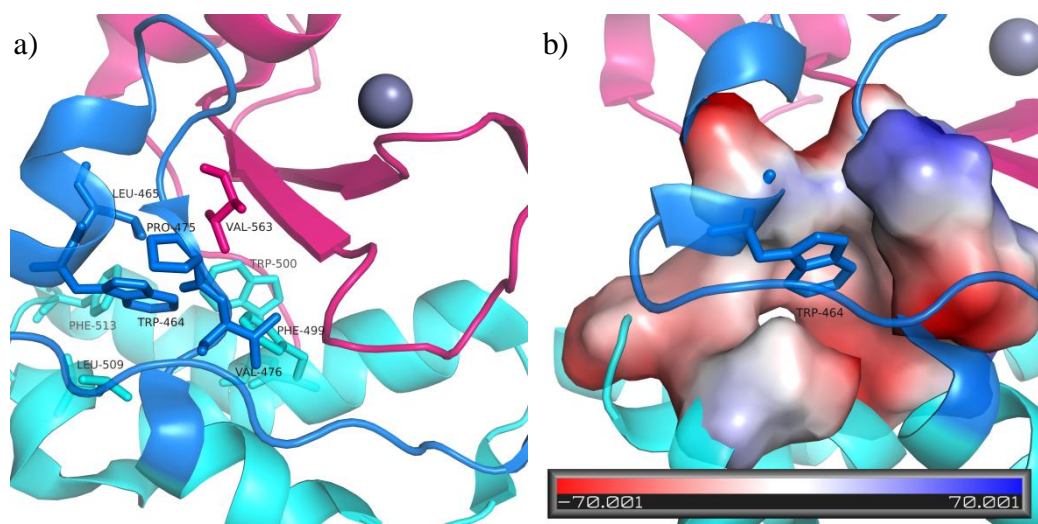
Fig. 5.32 a) The location of W464 in a hydrophobic pocket formed by L465, P475, F499, W500, L509 and F513 showed on the 1M08 [50] structure. b) The vacuum electrostatics surface of the same residues around W464, generated by PyMOL.

The following conclusions could be drawn from the results about the T454A, K458A and W464A mutations, regarding the redesign of the enzyme:

1) The W464 residue is essential for the stability and folding and therefore for the effective nuclease activity of the NColE7 protein.
2) The N-terminal region, especially with W464 is important for the stabilization of the HNH-motif and thus for the preorganization of the Zn^{2+}-binding site [115]. However, this function can be substituted by substrate or Im7-binding [115].
3) If only other interactions are missing between the N-terminal loop and the core of the protein (TK mutant), the function of the protein is only slightly affected.

5.3. COMPUTATIONAL DESIGN OF NOVEL ZFN-S

Learning from the different point mutants the next step was to establish a strategy for the design of an NColE7-based ZFN [108]. As it was shown in **Fig. 3.1 a**, the idea is to join an N-terminally truncated and thus inactive NColE7 molecule to three ZF motifs, whereas the missing N-terminus is to be joined to the other end of the ZF array. In this case the N-terminal unit would control the protein: it can cleave DNA after the specific binding occured, bringing together the N- and C-terminal parts of NColE7. The dets of the computational modeling procedure are described in the Appendix, chapter 10.3.

5.3.1. Set-up of the initial models

The initial models were constructed based on the following criteria:

- The three zinc-finger proteins bind DNA in the major groove;
- NColE7 is also a major groove binding protein, however, its catalytic centre, the HNH-motif, acts in the minor groove;
- The complex has to allow for the design of intramolecular allosteric regulation, which supposes that the protein takes one turn around DNA and the HNH motif is to be separated from the regulatory element by the zinc-finger motifs.
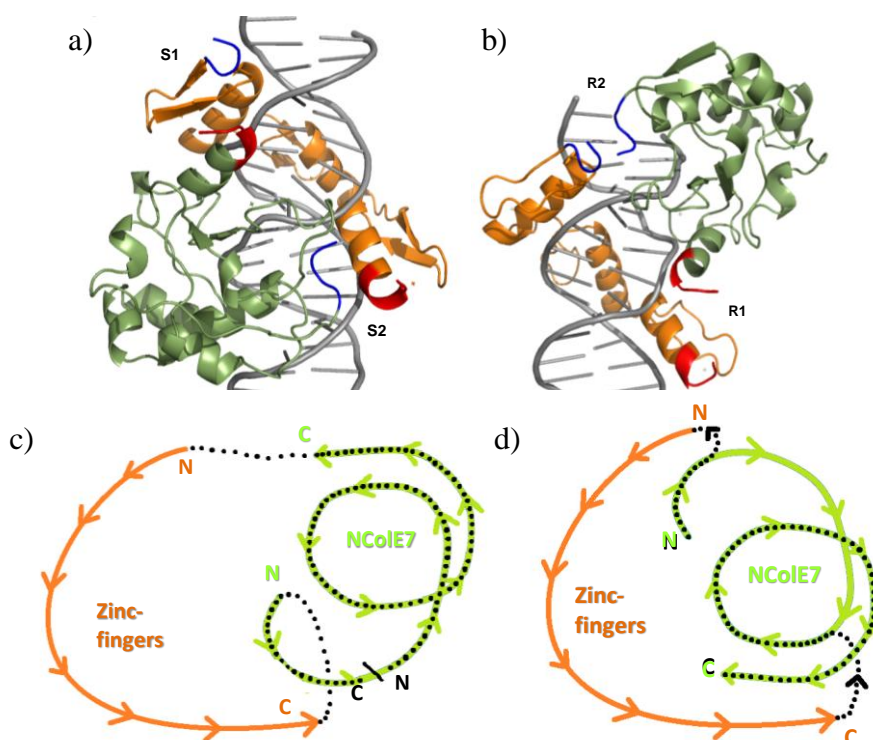
Fig. 5.33. *Two possible complexes for the design of a ZFN, in straight (**a**) and reverse (**b**) orientations* [108]. *NColE7 in green and ZF protein consisting of 3 fingers in orange. The N- and C-termini of the proteins are marked in blue and red, respectively. The panels **c)** and **d)** show the fusion strategies that are applied to form one polypeptide chain, in the straight and reverse approach, respectively. The arrows show the N→C termini direction of the new sequences.*

The proteins can fulfill these criteria in two possible orientations (**Fig. 5.33 a** and **b**). In the "straight" orientation the N-terminus of the ZF is close to the C-terminus of the nuclease thus, the linkers can be designed to bind the termini of the proteins to form a continuous circular

sequence (**Fig. 5.33 c**). In this case new termini can be chosen virtually at any position of the NColE7 sequence. In contrast, in the "reverse" orientation the N-terminus of the ZF is close to the N-terminus of the nuclease, and the linkers have to connect the two proteins at other parts than their termini as two N-termini or two C-termini can not be fused (**Fig. 5.33 d**). The fusion points of the sequences were selected based on sterical proximity of the NColE7 and ZF parts in the initial reverse model.

5.3.2. *Linker design*

Linkers were designed for both models, and the best linkers were selected based on the estimates of the free energy of the linkers according to FoldX [120]. The four final models were constructed using all possible combinations of the best linkers, their sequences are shown in **Fig. 5.34**.

N4-ZF-C105 (reverse)

KRNKPGKATGKGKPVNNKWLNNAGKDGGDAFMEKPYKCPECGKSFSQSSNLQKHQRTHTGEKPY
KCPECGKSFSQSSDLQKHQRTHTGEKPYKCPECGKSFSRSDHLSRHQRTHQGRFERGDYGLGSP
VPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPELSKQFSRNNNDRMKVGKAPKTRTQDVSGKRT
SFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK

N4-ZF-C45 (reverse)

KRNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
SKQFSRNNNDRMKVGKAPKTRTGGDAFMEKPYKCPECGKSFSQSSNLQKHQRTHTGEKPYKCPE
CGKSFSQSSDLQKHQRTHTGEKPYKCPECGKSFSRSDHLSRHQRTHQFGTGAWMLLGQDVSGKR
TSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK

N46-ZF-C45 (reverse)

KRNKPGKATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPEL
SKQFSRNNNDRMKVGKAPKTRTGGGDGGDQMEKPYKCPECGKSFSQSSNLQKHQRTHTGEKPYK
CPECGKSFSQSSDLQKHQRTHTGEKPYKCPECGKSFSRSDHLSRHQRTHQFGTGAWMLLGQDVS
GKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGK

C123-ZF-N7 (straight)

ATGKGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFKSFDDFRKKFWEEVSKDPELSKQFSRN
NNDRMKVGKAPKTRTQDVSGKRTSFELHHEKPISQNGGVYDMDNISVVTPKRHIDIHRGKQGYE
RLGMEKPYKCPECGKSFSQSSNLQKHQRTHTGEKPYKCPECGKSFSQSSDLQKHQRTHTGEKPY
KCPECGKSFSRSDHLSRHQRTHQFEGGSKRNKPGK

Fig. 5.34. *Sequences of the designed zinc-finger nucleases* [108]. *The zinc-finger proteins are in blue, while NColE7 is divided into three parts: the original N-terminus (in red), the middle part not used in the model (in grey, crossed out) and the original C-terminal part (in green). The linkers are shown with black bold letters. The HNH motif of NColE7 is underlined. The N4-ZF-C105, N4-ZF-C45 and N46-ZF-C45 models are in the reverse orientation, while C123-ZF-N7 is the straight joint of protein sequences. For explanation of the names see the text.*

The sequence of zinc-fingers are unchanged in the models (**Fig. 5.35**), except for cutting the last one or two residues when it was required for the linker design. NColE7 is divided into two parts: the N-terminus (N*X*) and the C-terminus (C*Y*), where *X* and *Y* refer to the number of residues involved in the model and $X + Y$ is always less than the total number of amino acids in NColE7. The zinc-finger sequence is inserted between these two parts.

Accordingly, the "straight" models are named as C*Y*-ZF-N*X*, while the "reverse" models are named N*X*-ZF-C*Y*.
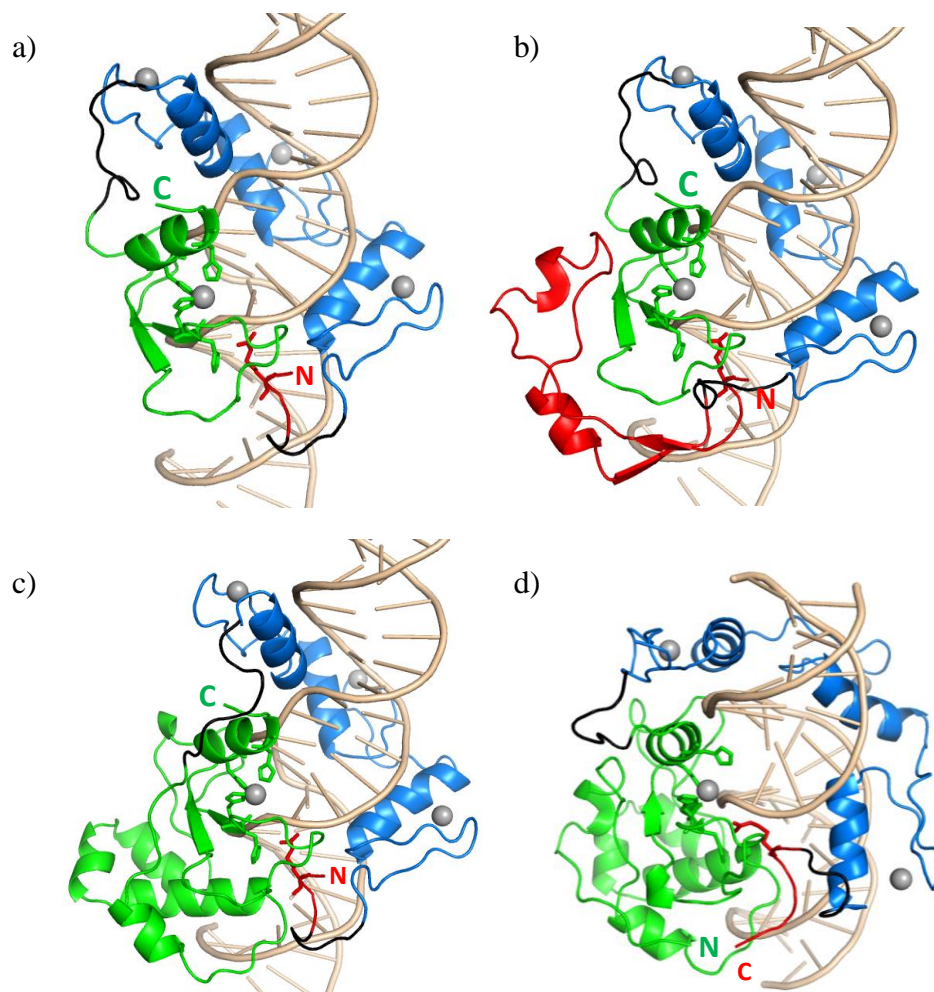
Fig. 5.35. *Designed ZFN structures, as starting points of MD simulations* [108]. *The N-terminal part of NColE7 is shown in red, the C-terminal part in green, the ZF-s in blue, and the linkers are shown in black. The grey spheres indicate Zn^{2+}-ions.* ***a)*** *N4-ZF-C45 (reverse)* ***b)*** *N46-ZF-C45 (reverse)* ***c)*** *N4-ZF-C105 (reverse)* ***d)*** *C123-ZF-N7 (straight). The C- and N-termini of the zinc-finger nucleases are marked by "C" and "N", respectively.*

5.3.3. Molecular dynamics investigation

10-ns MD simulations were performed to study the stability and structural features of the designed proteins with and without DNA. Some examples of the analysis are shown below.

5.3.3.1. Overall structural stability

The overall change of the structure compared to the starting structures is measured by the atom-positional root-mean-square deviation (RMSD) as a function of time (**Fig. 5.36**).
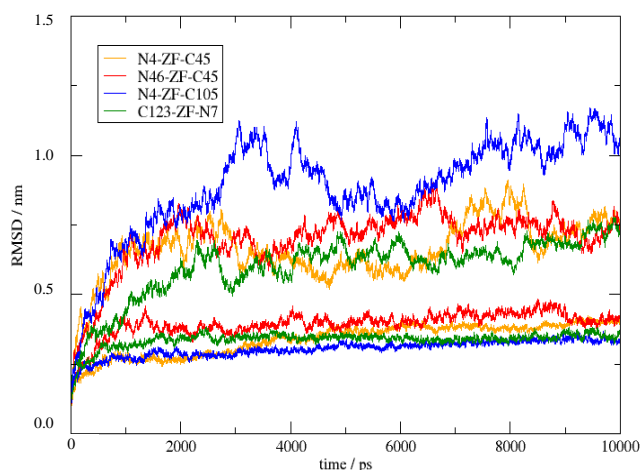
Fig. 5.36 *Atom-positional root-mean-square deviation (RMSD) of the ZFN models in the simulations in complex with DNA (lower curve in each color) and in the control simulations of the proteins without the DNA (higher curves in each color)* [108].

The proteins in complex with DNA all converged towards values around 0.4 nm in the simulations, while in the control simulations without DNA they had more severe structural changes up to 0.9 nm. The ZFN-s containing the major part of NColE7 as a continuous sequence, *i.e.* N4-ZF-C105 and C123-ZF-N7 had RMSD values that stayed closest to those in the reference simulation of NColE7, as expected. None of the free ZFN proteins reached equilibrium within 10 ns, because the cyclic shaped molecules continued to change their tertiary structure. This could be detected also by the changes in the radius of gyration (average distance from the center of mass, *e.g.* **Fig. 5.37**) attributed to the flexible linkers and not to the protein domains themselves.
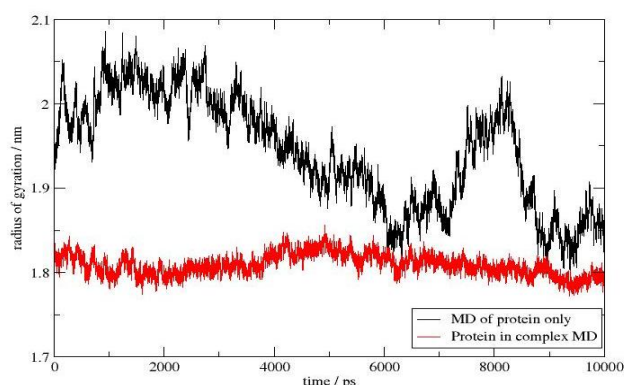
Fig. 5.37 *Radius of gyration in the simulation of N4-ZF-C45 in complex in the presence and in the absence of DNA.*

The zinc-fingers behave similarly in all four models: they had a stable structure as seen in snapshots and low atom-positional root-mean-square fluctuations (RMSF < 0.2 nm). In contrast, the nuclease part without the DNA substrate was less stable. The DNA structure was stabilized in a similar manner by all four ZFN proteins.

5.3.3.2. HNH-motif and potential catalytic features

For the designed ZFN proteins not only the overall structural stability, but also the proper function of the catalytic centre, i.e. the HNH motif is essential. The atom-positional root-mean-square fluctuations (RMSF) of the HNH residues are compared in **Fig. 5.38**. N4-

ZF-C45 had the least stable HNH-motif both in the DNA-bound and free form. This is in agreement with the previous study on the HNH motif itself [121], where it was shown that its isolated structure is not stable and its zinc-binding is weak. The HNH-motif of C123-ZF-N7 has the most similar features to that of NColE7 – as expected, since this protein has the most complete nuclease domain. Despite the small deviations, all four models retained the β-sheets and the α-helix in the HNH-motif.
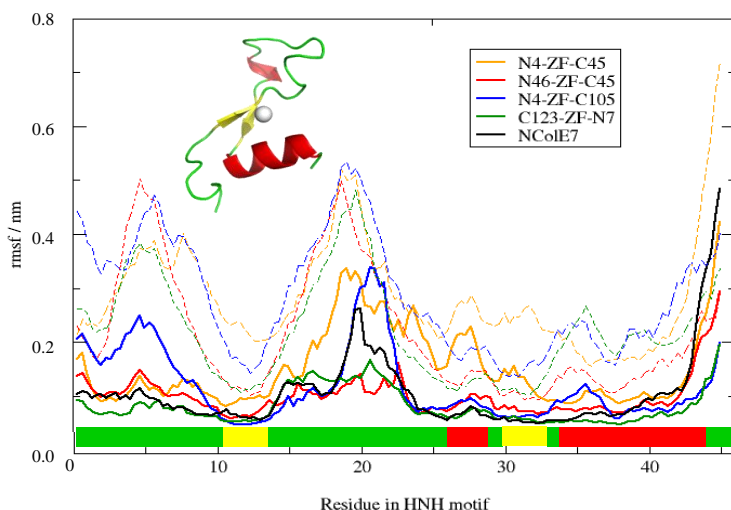
Fig. 5.38. Atom-positional root-mean-square fluctuations (RMSF) of residues in the HNH motif of ZFN-s and NColE7 as a reference [108]. The residue numbers are shifted to 1-45 in all models for comparison, corresponding to the HNH residues 532-576 in the NColE7 numbering. Dashed lines indicate the control simulations for proteins without DNA. The secondary structural elements of HNH motif are marked above the x axis, showing the loops in green, α-helices in red and β-sheets in yellow.

According to our expectations the stabilization of the nuclase structure could be induced by the substrate, as it was found for the TKW mutant of NColE7. The MD simulations confirm this, since the structure of HNH-motif is stable in the presence of DNA, while it changes in the free protein. This was also confirmed by DSSP [122] secondary structure analysis, as shown in **Fig. 5.39**. The secondary structure of the HNH motif was stable in the simulations of ZFN-s with DNA substrate.
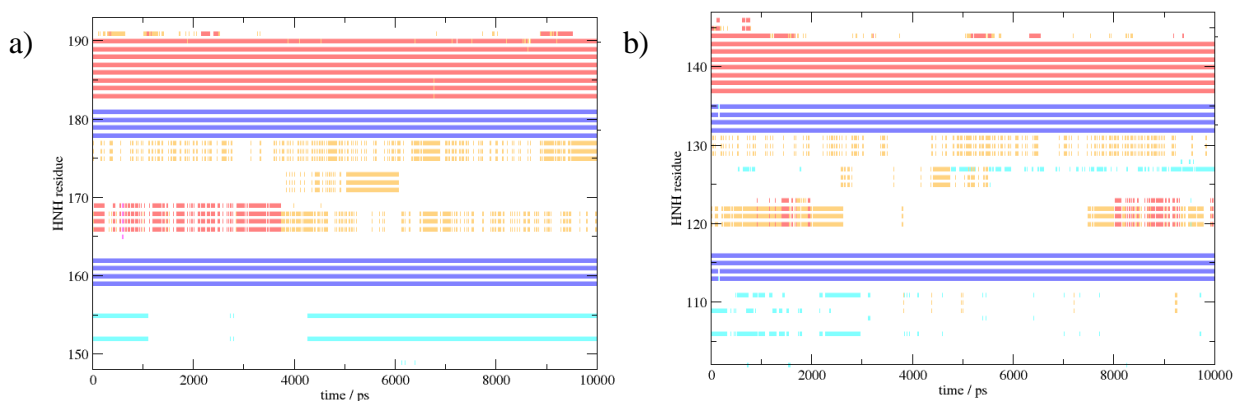
*Fig. 5.39 DSSP analysis [123] of the HNH-motif (C45-part) in the N4-ZF-C45 protein in presence (**a**) and absence (**b**) of the DNA substrate shows that DNA-binding influences the structure of HNH-motif. The 3-type helix is in yellow, the 4-type helix in red, β-strands are in blue and β-bridges in light blue.*

During the catalytic reaction the Zn^{2+}-ion of the nuclease domain binds the oxygen of the DNA scissile phosphate group. The Zn^{2+}-O distance was found to be stable as a function of time for each of the models except for N46-ZF-C45, where the coordination bond was lost (**Fig. 5.40**).
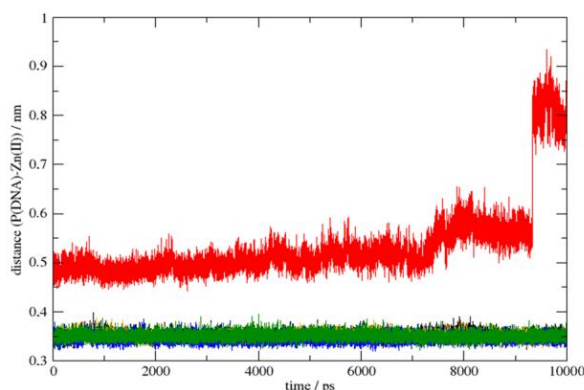
Fig. 5.40 Catalytic distances as a function of time: Zn^{2+}-O (at the scissile phosphate).

5.3.3.3. Protein-DNA interactions

The protein-DNA interaction energies were compared in**Table 5.9**. The N4-ZF-C105 model shows the strongest interaction with DNA. This result is expected among the reverse models, since this model contains the most intact NColE7 domain. Even the straight model has weaker interaction with the substrate, which can be explained by the more flexible loop in N4-ZF-C105 and the C-terminal position of HNH motif, similarly to the case of natural NColE7.

model	average interaction energy (kJ/mol)
N46-ZF-C45	-6235 ± 39
N4-ZF-C45	-6109 ± 33
N4-ZF-C105	-7369 ± 68
C123-ZF-N7	-6384 ± 64

Table 5.9 Interaction energies calculated as the sum of van der Waals and electrostatic interaction energies between the proteins and DNA molecule [108].

The control mechanism of the designed ZFN would be assured by the positively charged sequence, originally found at the N-terminus of NColE7 ("KRNK"). Particularly, R447 is important, as it stretches out its side-chain towards the scissile phosphodiester group. The analysis of the distances between these two residues as a function of time showed that N46-ZF-C45 is less probable to exert nuclease activity (**Fig 5.41**). It is important to note that all the other models showed an Arg-P distance suitable for an allosteric control mechanism. Concerning this distance, the N4-ZF-C105 model is the most optimal, confirming that it is preferable to keep the controlling loop at the N-terminus and the catalytic centre at the C-terminus.
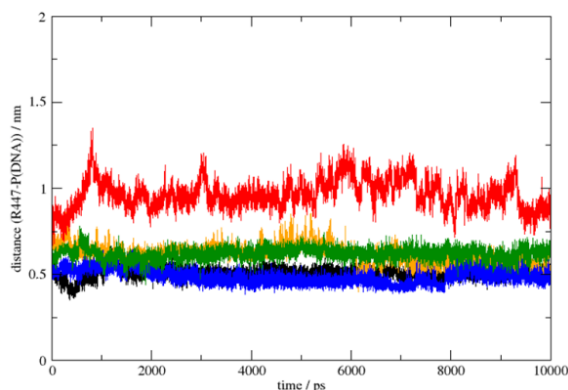
Fig. 5.42 a shows the number of H-bonds between the atoms of the residue corresponding to R447 in NColE7 and DNA atoms. C123-ZF-N7 has the most similar dynamics of R447-DNA interactions to NColE7, showing three hydrogen bonds with a reasonable probability at the end (7-10 ns) of the simulation. The Arg in the KRNK part of N4-ZF-C105 and N4-ZF-C45 also shows an extensive H-bond network. Notably, the Arg in N4-ZF-C45 binds to the same phosphate group as NColE7 in the beginning, but changes to the neighboring phosphate after 6500 ps (**Fig. 5.42 b**) indicating that the truncation of the nuclease structure has high impact on the function. N46-ZF-C45 forms Arg-DNA H-bonds only a few times during the simulation, which is the result of more severe structural changes after dividing the nuclease over two domains of the constructed ZFN.
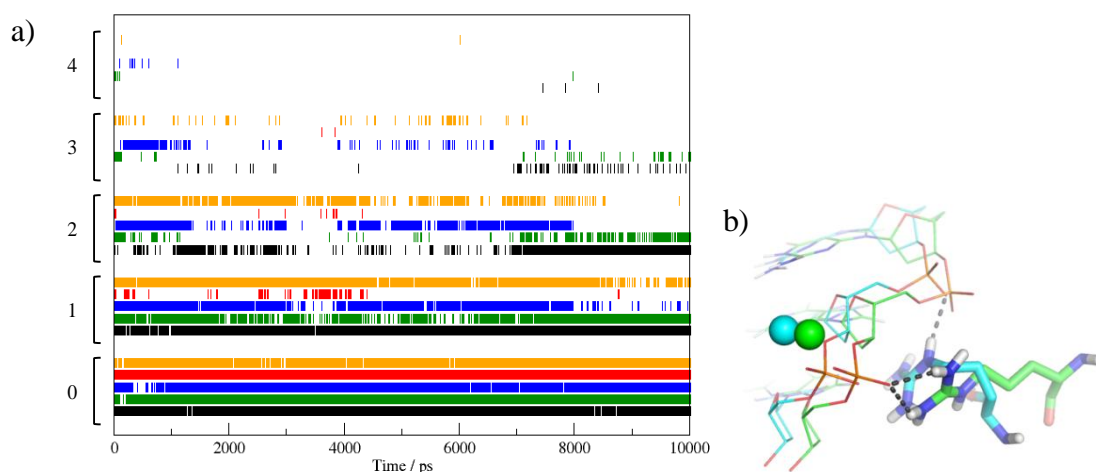
Fig. 5.42. a) *The number of H-bonds between the arginine corresponding to R447 in NColE7 and the DNA atoms in the different models as a function of time* [108]. *Numbers on the y axis indicate how many hydrogen bonds were found at a given time step. NColE7 is in black, N4-ZF-C45 in orange, N46-ZF-C45 in red, N4-ZF-C105 in blue and C123-ZF-N7 in green.* **b)** *The interactions of the controlling Arg at 2000 and 8500 ps in the N4-ZF-C45 simulation.*

5.3.4. Conclusions: four designed NColE7-based ZFN models

In the process of the design of NColE7-based ZFNs a novel type of control was introduced. Namely, NColE7 was divided into two parts using its special features: the HNH-motif at the C-terminus of the protein is only functional in the presence of the positively charged N-terminal sequence. These parts are linked to different ends of the ZF array. Therefore, the newly modeled nucleases should be able to exhibit their hydrolytic activity only upon ZF binding to the specific DNA recognition sequence and as a result, the controlling loop and the HNH motif become in suitable proximity. By establishing such a control, we avoid the risk that partially degraded proteins maintain nonspecific activity, leading to cytotoxicity.

Four ZFN models of this kind have been computationally designed and subjected to molecular dynamics simulations [108]. Three models (N4-ZF-C45, N4-ZF-C105 and N46-ZF-C45) had, similarly to NColE7, the allosteric activator sequence at the N-termini, the HNH-motif based parts at the C-termini and the ZF array in between. A fourth model (C123-ZF-N7) rather placed the allosteric sequence at the C-terminus and the HNH-motif at the N-terminus. This latter model simply connected the termini of the NColE7 to the zinc-fingers. This approach allows for easy redesign, however, its drawback is, that the termini are exchanged. As a result of this the distance between the Arg (corresponding to R447 in NColE7) and DNA increased and the number of H-bonds involving Arg decreased during the simulation. Also, the structure of the isolated HNH motif was less stable in comparison to the oppositely built model, ie. N4-ZF-C105 as predicted by DSSP. Consequently, the ZFN molecules keeping the original termini seem to be more promising. The most complete model among these three is the N4-ZF-C105, which exhibits almost all the interactions predicted to be important in the mutation experiments. All features of this model were similar to NColE7 with a small difference in the loop between the β-strands of its HNH-motif, which was more flexible based on the RMSF profile and the characteristic H545-N560 distance. These changes probably arose from the interaction of the loop both with the bulky part of NColE7 and with the newly designed linker. The change of the loop structure could hamper the nuclease activity, since this part plays an important role in the reaction, promoting the generation of the nucleophilic OH⁻. The nuclease in the N4-ZF-C45 model is truncated and contains only the N-terminal KRNK sequence and the HNH-motif. It was previously shown experimentally, that the HNH motif itself does not fold correctly, and the lower stability can also be seen in the control simulations for proteins without DNA. However, this model may still be functional as the interaction with DNA could induce the correct structure, since it behaves well in the simulations with DNA. The N46-ZF-C45 model is tempting in terms of strict conformational regulation of the ZFN, as the two independent parts of NColE7 have to fold together while the ZF-s bind around the specific DNA. However, even though the protein is stable, the catalytic properties of this model in the simulations were significantly worse than in the other models.

5.4. EXPERIMENTAL STUDY OF THE DESIGNED ZFN-S

The three straight models (N4-ZF-C105, N46-ZF-C45 and N4-ZF-C45) were selected for experimental study. The experimental protocols related to this chapter can be found in the 10.4 chapter in the Appendix.

5.4.1. Expression and purification of the designed proteins

The genes of the designed proteins were constructed using the pQE70 plasmid for the NColE7 part and the gene of the ZF protein in the 1MEY [12] structure. The genes were inserted into the pET-21a vector (Novagen) between the NdeI and EcoRI restriction enzyme sites. No sequences encoding for purification tags were fused in order to obtain the protein with the precise amino acid sequence, without any additional resudies, making the purification procedure more complicated.

The genes were transformed to *E.coli* BL21(DE3) cells and protein expression was induced in 500 ml bacterial culture. A three step purification procedure was applied. First the raw extract was subjected to a gravity column filled with the P11 cellulose phosphate-based high capacity cation exchanger resin and eluted with a salt gradient.
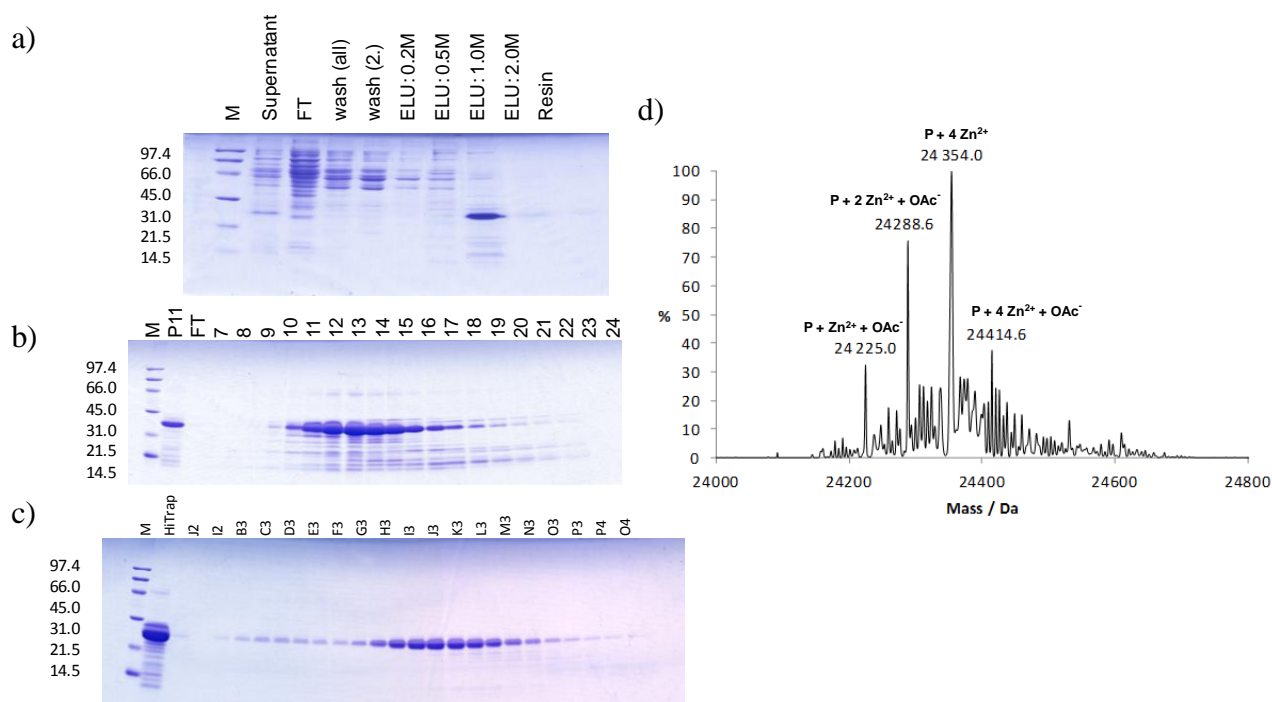
Fig. 5.43. *SDS gels following the purification of the N4-ZF-C105 protein (24 kDa). **a)** P11 resin in gravity column.The band around 24 kDa can be seen in the supernatant of cell extracts, that was used for purification. The flow through and wash fractions did not contained significant amount of the target protein indicating that it bound to the column. In the stepwise elution the fraction of 1 M NaCl contained the N4-ZF-C105 protein, together with some impurities (the other bands at both lower and higher MW ranges.)   **b)** HiTrap cation exchange chromatography fractions starting with the 1 M elution fraction of the P11 column ("P11"). Fractions 11-16 from the gradient elution were collected andsubjected to the gel filtration column. **c)** Gel filtration fraction was started from the collected sample marked as "HiTrap". The fractions H3-N3 contained the N4-ZF-C105 protein with no detectable impurity. These fractions were used for further experiments investigating the features of the protein. **d)** The mass spectrum of N4-ZF-C105 protein. The low intensity peaks correspond to Na^+-adducts.*

The fraction containing the target protein (1M NaCl) was diluted 10× and applied to another cation exchange column (HiTrap column, dextran coupled to cross-linked agarose matrix) in FPLC. The fractions of interest were collected and further purified with gel-filtration in order to separate the proteins with similar charge but different size. An example of the purification procedure is shown in **Fig. 5.43**.

According to ESI-MS measurements the molecular weights of the purified proteins match well with the expected masses, based on the sequence containing the methionin at the N-terminus (**Table 5.10**, an example of the spectra is shown in **Fig. 5.43 d**).

Table 5.10 *The measured and expected molecular weight of the proteins according to the highest peak in the mass spectra. All expected masses were calculated for proteins with an additional starting methionin, that is a result of bacterial expression.*

	Expected mass / Da	Measured mass / Da	Most abundant species
N4-ZF-C105	24354.6	24354.0	Protein + 4 Zn^{2+}
N4-ZF-C45	17063.1	17062.0	Protein + 3 Zn^{2+}
N46-ZF-C45	21528.1	21527.4	Protein + 3 Zn^{2+}

5.4.2. DNA-binding affinity and specificity

The DNA-binding of the purified protein was studied with a 21 base pairs (bp) long ^{32}P-labelled DNA containing the recognition sequence (GAGGCAGAA) [12] of the zinc-finger protein. According to the competition assay with unlabelled specific and nonspecific 21 bp DNA fragments the ZF-s maintained their DNA-binding ability (**Fig. 5.44**) and specificity (**Fig. 5.45**) after the fusion with the NColE7 segments.
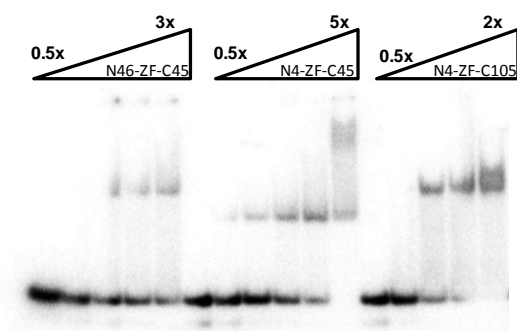
Fig. 5.44. Gel mobility shift of a ^{32}P-labelled 21 bp specific DNA (s.DNA*, 0.2 μM) containing the GAGGCAGAA target sequence in the presence of increasing ZFN protein concentrations.
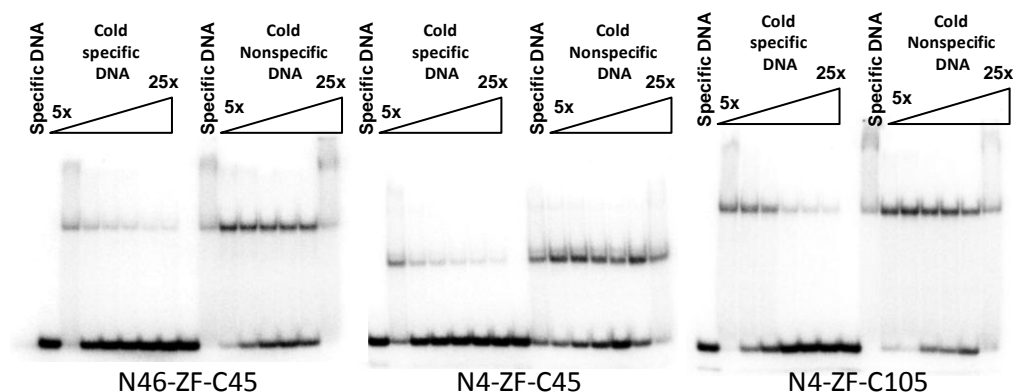
Fig. 5.45 *Competition assay performed by adding an increasing amount of unlabelled specific (s.DNA) or nonspecific (ns.DNA) 21 bp DNA from 1 μM to 5 μM to the solution containing 0.2 μM ^{32}P-labelled DNA and 0.3 μM protein.*

5.4.3. Nuclease activity

The nuclease activity was investigated with the pGEX-6P-1 vector including the target DNA sequence cloned into the SmaI site. According to the computational model the protein can access the DNA at one strand only, therefore only a single strand nicking is expected at the recognition site. A nick on the supercoiled plasmid DNA leads to the opencircular form, which was observed within 46 hours (**Fig. 5.46**). The three proteins showed different activities.
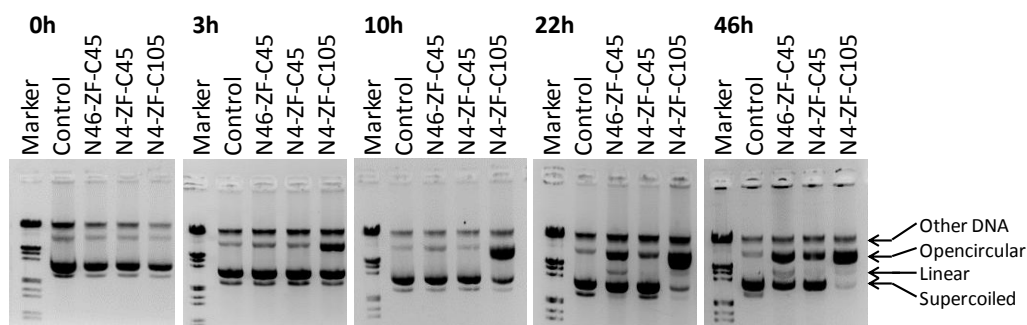
Fig. 5.46 *Cleavage of the pGEX-6P-1 vector containing the recognition sequence of the zinc-finger domain (32 nM) by the ZFN-s (850 nM). The reaction mixtures contained 2 μM ZnSO$_4$ and 50 mM HEPES at pH = 7.9 and were incubated up to 46 hours at 37 °C. Aliquots taken at the denoted times were run in agarose gel electrophoresis experiments.*

The highest nuclease activity was found for N4-ZF-C105, where after 10 hours most of the supercoiled DNA-molecules were converted to open circular form. The intensity of the linear form did not significantly increase, suggesting that the reaction is specific. In a nonspecific reaction upon accumulation of the cleavage events two nicks close to each other in the opposite strands would lead to the linear form of the plasmid, similarly to the NColE7 experiments, as shown in chapter 5.1.3. A significantly lower, but yet notable nuclease activity was found also with the other two proteins.

To confirm the specificity of nicking activity primer extension experiments were performed (**Fig. 5.47**). The reaction mixture incubated for 46 hours was used as a template for the extension of the selected primers. The extension of the primer on the template ends. Thus, the primer extension leads to a unique single stranded fragment, with a length determined by the location of the nicked site in case of specific nicking. Since each reaction contains one of the primers only, the reaction is quantitative. After running the reaction mixture on a gel, a band can be observed only if a significant amount of the plasmid molecules is nicked at the same site. The pGEX-6P-1 vector cleaved by the EcoRI restriction endonuclease was used as a positive control, resulting in a 98 or 86 bp fragment in the reactions with the 3' or 5' primers, respectively. Based on the computational model in case of the ZFN-s a 101 bp DNA fragment is expected in the reaction with the 3'-pGEX sequencing primer, as a result of nicking the DNA at the recognition site (GAGGCAGAA). On the opposite strand, there is a sequence almost identical with the recognition sequence of the ZF (ATCTGCCTC), the underlined base differs from the recognition site), that also may facilitate DNA-binding and cleavage which can be detected in the reaction with the 5'-primer. As shown in **Fig. 5.47 b**, both strands were nicked by N4-ZF-C105 and the fragments of the expected size could be obtained.
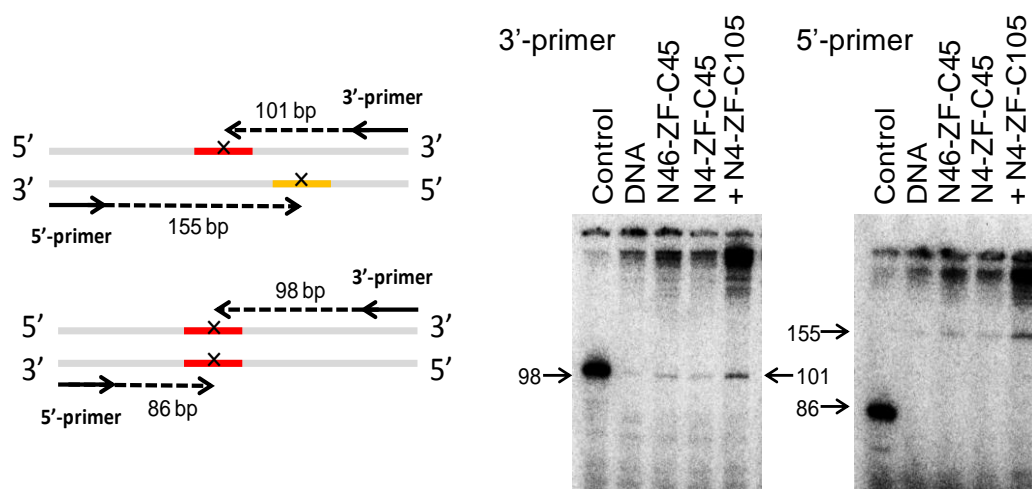
Fig. 5.47 a) Scheme of the primer extension reaction. The template is obtained from 46 h reaction of the ZFN and the recognition sequence (GAGGCAGAA, red line) containing plasmid DNA. The yellow line is a polymorph of the recognition sequence differing only in one base (ATCTGCCTC). In the reaction the primer is elongated until there is a nick on the strand. b) Scheme of the control primer extension reaction performed by the pGEX plasmid cleaved by EcoRI. c) Primer extension experiments performed with the 5'-pGEX and the 3'-pGEX radiolabelled sequencing primers, according to the scheme shown in Fig. 4. a and b.As a positive control reaction the cleavage of the empty pGEX-6P-1 vector was carried out by the EcoRI restriction endonuclease (Control) and the uncut plasmid was used for negative control (DNA).

5.4.4. Conclusions: a new ZFN-design strategy

We expressed and purified three designed NColE7-based ZFN-s. The most promising model is N4-ZF-C105. Previously we have shown that the ΔN25-NColE7 is inactive, and has an altered solution structure and weak DNA and Zn^{2+}-binding affinity [31]. Here, the C105 part of the ZFN equals to ΔN26-NColE7, and therefore it is also expected to be inactive. As shown with the TKW mutations in the N-terminal chain, W464 has an impact on structural stability, however the active structure of the protein can be induced by the substrate binding. In N4-ZF-C105 W464 is missing and such an activation of the nuclease could occur resulting in the specific single strand cleavage. The specific nicking activity observed on the strand of DNA as the computational models suggest confirms this theory, however a wider investigation of the features of this model is planned.

This is the first example of ZFNs containing an inactive nuclease domain that is activated by the specific substrate binding during the reaction. The recognition sequence of 9 bp by a three finger ZF array does not provide enough specificity in human genome. However, the nick of DNA at the off-target sites may be corrected by the repair mechanism of the cell. In the presence of a suitable template sequence the homologous recombination might be induced even by a single strand nick on DNA [31]. Furthermore, the strategy introduced here can be generally applied in the design of chimeric nucleases with any kind of DNA-binding protein domains: good examples being the transcription activator-like effector (TALE) motifs. As long as the DNA binding domains do not interact with the recognition sequence, the NColE7-based nuclease is inactivated, preventing the DNA from the harmful nonspecific cleavages. Thus the chimeric nuclease activated at the binding site can offer a safe option for gene editing.

6. SUMMARY AND FUTURE OUTLOOK

My PhD thesis was aimed at designing a novel, controllable ZFN by the fusion of NColE7 to a three-ZF array. This required detailed knowledge on the nuclease domain itself, with an emphasis on the role of its N-terminus intended to be used as a control unit. NColE7 is a Zn^{2+}-binding metalloenzyme, the metal-ion being essential for the catalytic activity. The affinity of the enzyme and its mutants towards Zn^{2+}, as well as the changes in the coordination sphere of the metal ion are tools for fine-tuning of the catalytic activity. As such it represents a challenging bioinorganic chemistry research, in which the understanding of the catalytic center was another question to answer.

As it was previously shown, the truncated ΔN4-NColE7, lacking the "KRNK" sequence from the N-terminus is inactive [43]. To understand the role of this sequence, in the first part of my work I expressed and purified the NColE7 protein and its four point mutants, containing the KGNK, KGNG, GGNK or GGNG sequence at the N-terminus, instead of the KRNK sequence [63]. The mutants retained the cytotoxicity leading to problems in the gene construction, but this could be solved by the coexpression and of the Im7 protein and separation later in the purification procedure [63]. The cytotoxicity suggested that the nuclease activity is not lost upon these mutations [108]. This was confirmed by agarose gel electrophoresis and linear dichroism spectroscopic monitoring of the catalytic experiment [63]. However, the remaining nuclease activity was only minor, increasing reaction times by several orders of magnitude. I studied the proteins in order to explain the dramatic change in the enzymatic activity, which turned out to be a question difficult to answer. According to CD-measurements the solution structure of the proteins is not significantly altered. The Zn^{2+}-affinity was determined by ITC, and it was also found to be unaffected [63]. These were expected results, since the mutations involve only the flexible terminus of the protein. The most intriguing question was the DNA-binding. Even though the number of positive charges at the N-terminus was decreased, the DNA-binding affinity was not significantly affected. The DNA-binding part of NColE7 is not this loop but rather the central helical part interacting with the major groove of DNA. The crystal structure of KGNK with DNA revealed that there are only minor changes near the active site. K446 may get close to the catalytic center, similarly to R447. This is in agreement with MD simulations suggesting that K446 can partially take over the role of R447 and with the observation that out of the four mutants KGNK is the most active [63]. The other mutants have even lower, but yet detectable activity in contrast to the truncated ΔN4-NColE7 being inactive. The role of R447 in the mechanism is still a matter of question [110]. The stabilization of the negatively charged transition state during the nucleophile attack by an OH^--ion is possible however the Zn^{2+}-ion can play this role alone in other enzymes. On the other hand, R447 may take part in the inhibition of the reverse reaction. There is also a possibility that R447 is a part of a proton transfer chain to protonate the leaving alcoholate ion. Nevertheless, the flexibility and possibility to substitute it with another sidechain suggested that the N-terminus could be used to control the function of NColE7.

In the redesign of NColE7 we planned to separate the N-terminus and to use it as a controlling segment. Therefore, it was important to get information on how this part of the protein is kept in its place. According to published crystal structures the N-terminus is continued in a long loop of ~25 residues. We performed a computational alanine scan in order to identify the most important residues in this loop that may have important interactions with other parts of the protein [115]. The three selected residues were T454, K458 and W464. Accordingly, the triple alanine mutant "TKW" and also other mutants, TK, TW and W were expressed and purified [115]. CD-experiments revealed that W464 has an essential role in the structural integrity of the protein. All mutants containing the W464A exchange (TKW, TW and W) had low activity and weak Zn^{2+}-binding indicating that the metal binding center is destabilized. The folding of the protein could be induced by the binding of DNA or the Im7 inhibitor protein. This was accompanied by the rescue of Zn^{2+}-binding. This proves that the preorganization of the metal binding site is essential for NColE7. The DNA-induced structure is an extraordinary feature that we attempted to utilize in the design the controlled enzyme.

After collecting informations on NColE7 the next step was to computationally design the NColE7-based ZFN molecules [108]. The procedure was started from the crystal structures of a three ZF-motif (PDB ID 1MEY [12]) in complex with DNA and that of NColE7 (PDB ID 3FBD [48]) in complex with DNA. In the initial models the two DNA-molecules were aligned in a way that the two proteins bind close to each other. After that short linker sequences were designed to fuse the two molecules. A progrm improved for the modelling of missing loops in protein crystal structures (LoopX) was used to find the backbone of the linkers. Then sidechains were selected based on energies estimated by the FoldX program. The final models contained some part of the N-terminus (N*X*) fused to one terminus of the ZF array, and some part of the C-terminus (C*Y*) fused to the opoosite terminus of the ZF-s. The "straight" N*X*-ZF-C*Y* models included N4-ZF-C105, N4-ZF-C45 and N46-ZF-C45. The fourth, "reverse" model was C123-ZF-N7. NColE7 itself is a 131 residues long chain, and the models were lacking different parts from the middle of the NColE7 sequence. The four models were subjected to 10 ns MD simulations. The analysis of the trajectories pointed out that as expected, the more complete NColE7 is the more stable the ZFN. The catalytic center of the three straight models was shown to be more stable in presence than in absence of DNA, in agreement with the experimental results of TKW. W464 is missing from N4-ZF-C105 and N4-ZF-C105. The N46-ZF-C45 model has a large interface between the N- and C-terminal parts of NColE7 and it was stable in the simulation, however important catalytic residues changed their location that might be not advantageous for the activity. C123-ZF-N7 contains all the NColE7 residues except for only one. It behaved well in the simulations however, it is a question whether the separation of only 7 residues at the N-terminus is enough for strict regulation. Nevertheless the models were promising and we decided to continue experiments with the three straight models.

The genes encoding the proteins were constructed and the proteins were expressed and purified. ESI-MS was applied to validate the proteins. N4-ZF-C105 was detected in a major fraction with four Zn^{2+}-ions suggesting that the three ZF-s and also the NColE7 catalytic center contain the metal ion, while the other two proteins with only three metal ions. This is in

agreement with the expectation that the truncated NColE7 domain has a weak Zn^{2+}-binding. The DNA-binding was investigated in EMSA experiments and as it turned out, the strength and specificity of the binding of ZF-s was retained in the models. Plasmid cleavage tests showed that the N4-ZF-C105 protein is able to perform single strand cleavages at the specific site. The chain that is cleaved is in agreement with the expectations based on the computational models. Since we know from previous experiments that the C105 part of NColE7 is inactive, the observed cleavage may be the result of an induction mechanism.

This is the first case to report an NColE7-based ZFN. So far, also no controlled ZFN has been published. The control mechanism introduced here relies on the activation of the enzyme upon the specific binding of ZF-s to DNA. This could inhibit off-target cleavages leading to a safe ZFN. Our results provide a proof of concept for this design strategy, and might open a new direction in the design of DNA-editing enzymes. In the future we plan to study the control mechanism in N4-ZF-C105 in detail, and to optimize the models to get more effective and controlled ZFN-s.

7. REFERENCES

1. Yang W (2011) Nucleases: diversity of structure, function and mechanism. Q Rev Biophys 44:1-93

2. Hsia K, Li C, Yuan H (2005) Structural and functional insight into sugar-nonspecific nucleases in host defense. Curr Opin Struct Biol 15:126-134

3. Dupureur CM (2008) Roles of metal ions in nucleases. Curr Opin Chem Biol 12:250-255

4. McCall KA, Huang C, Fierke CA (2000) Function and mechanism of zinc metalloenzymes. J Nutr 130:1437S-46S

5. Gregory DS, Martin AC, Cheetham JC, Rees AR (1993) The prediction and characterization of metal binding sites in proteins. Protein Eng 6:29-35

6. Lee YM, Lim C (2008) Physical basis of structural and catalytic Zn-binding sites in proteins. J Mol Biol 379:545-553

7. Hunt JA, Ahmed M, Fierke CA (1999) Metal binding specificity in carbonic anhydrase is influenced by conserved hydrophobic core residues. Biochemistry 38:9054-9062

8. Rouet P, Smih F, Jasin M (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. Proc Natl Acad Sci U S A 91:6064-6068

9. Pingoud A, Silva GH (2007) Precision genome surgery. Nat Biotechnol 25:743-744

10. Gyurcsik B, Czene A (2011) Towards artificial metallonucleases for gene therapy: recent advances and new perspectives. Future Med Chem 3:1935-1966

11. Y. G. Kim, J. Cha, S. Chandrasegaran (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc Natl Acad Sci U S A 93:1150-1160

12. Kim CA, Berg JM (1996) A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. Nat Struct Biol 3:940-945

13. Jantz D, Amann BT, Gatto GJ,Jr, Berg JM (2004) The design of functional DNA-binding proteins based on zinc finger domains. Chem Rev 104:789-799

14. Dhanasekaran M, Negi S, Imanishi M, Sugiura Y (2007) DNA-Binding ability of GAGA zinc finger depends on the nature of amino acids present in the beta-hairpin. Biochemistry 46:7506-7513

15. Mandell JG, Barbas CF,3rd (2006) Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. Nucleic Acids Res 34:W516-23

16. Sakai-Kato K, Umezawa Y, Mikoshiba K, Aruga J, Utsunomiya-Tate N (2009) Stability of folding structure of Zic zinc finger proteins. Biochem Biophys Res Commun 384:362-365

17. Yan W, Imanishi M, Futaki S, Sugiura Y (2007) Alpha-helical linker of an artificial 6-zinc finger peptide contributes to selective DNA binding to a discontinuous recognition sequence. Biochemistry 46:8517-8524

18. Anand P, Schug A, Wenzel W (2013) Structure based design of protein linkers for zinc finger nuclease. FEBS Lett 587:3231-3235

19. Laity JH, Dyson HJ, Wright PE (2000) DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers. J Mol Biol 295:719-727

20. Wah DA, Hirsch JA, Dorner LF, Schildkraut I, Aggarwal AK (1997) Structure of the multimodular endonuclease FokI bound to DNA. Nature 388:97-100

21. Wah DA, Bitinaite J, Schildkraut I, Aggarwal AK (1998) Structure of FokI has implications for DNA cleavage. Proc Natl Acad Sci U S A 95:10564-10569

22. Mani M, Smith J, Kandavelou K, Berg JM, Chandrasegaran S (2005) Binding of two zinc finger nuclease monomers to two specific sites is required for effective double-strand DNA cleavage. Biochem Biophys Res Commun 334:1191-1197

23. Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim YG, Chandrasegaran S (2001) Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. Mol Cell Biol 21:289-297

24. Shimizu Y, Bhakta MS, Segal DJ (2009) Restricted spacer tolerance of a zinc finger nuclease with a six amino acid linker. Bioorg Med Chem Lett 19:3970-3972

25. Bibikova M, Golic M, Golic KG, Carroll D (2002) Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics 161:1169-1175

26. Bibikova M, Beumer K, Trautman JK, Carroll D (2003) Enhancing gene targeting with designed zinc finger nucleases. Science 300:764

27. Carroll D (2011) Genome engineering with zinc-finger nucleases. Genetics 188:773-782

28. Tebas P, Stein D, Tang WW, Frank I, Wang SQ, Lee G, Spratt SK, Surosky RT, Giedlin MA, Nichol G, Holmes MC, Gregory PD, Ando DG, Kalos M, Collman RG, Binder-Scholl G, Plesa G, Hwang WT, Levine BL, June CH (2014) Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. N Engl J Med 370:901-910

29. Nagy Z, Soutoglou E (2009) DNA repair: easy to visualize, difficult to elucidate. Trends Cell Biol 19:617-629

30. Porteus M (2010) Testing a three-finger zinc finger nuclease using a GFP reporter system. Cold Spring Harb Protoc 2010:pdb.prot5531

31. McConnell Smith A, Takeuchi R, Pellenz S, Davis L, Maizels N, Monnat RJ,Jr, Stoddard BL (2009) Generation of a nicking enzyme that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. Proc Natl Acad Sci U S A 106:5099-5104

32. Metzger MJ, McConnell-Smith A, Stoddard BL, Miller AD (2011) Single-strand nicks induce homologous recombination with less toxicity than double-strand breaks using an AAV vector template. Nucleic Acids Res 39:926-935

33. Cornu TI, Thibodeau-Beganny S, Guhl E, Alwin S, Eichtinger M, Joung J, Cathomen T (2007) DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. Mol Ther 16:352-358

34. Catto LE, Ganguly S, Milsom SE, Welsh AJ, Halford SE (2006) Protein assembly and DNA looping by the FokI restriction endonuclease. Nucleic Acids Res 34:1711-1720

35. Szczepek M, Brondani V, Buchel J, Serrano L, Segal DJ, Cathomen T (2007) Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. Nat Biotechnol 25:786-793

36. Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA, Gregory PD, Pabo CO, Rebar EJ (2007) An improved zinc-finger nuclease architecture for highly specific genome editing. Nat Biotechnol 25:778-785

37. Guo J, Gaj T, Barbas CF,3rd (2010) Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. J Mol Biol 400:96-107

38. Sollu C, Pars K, Cornu TI, Thibodeau-Beganny S, Maeder ML, Joung JK, Heilbronn R, Cathomen T (2010) Autonomous zinc-finger nuclease pairs for targeted chromosomal deletion. Nucleic Acids Res 38:8269-8276

39. Kim H, Kim JS (2014) A guide to genome engineering with programmable nucleases. Nat Rev Genet 15:321-334

40. Chan SH, Bao Y, Ciszak E, Laget S, Xu SY (2007) Catalytic domain of restriction endonuclease BmrI as a cleavage module for engineering endonucleases with novel substrate specificities. Nucleic Acids Res 35:6238-6248

41. Chan SH, Stoddard BL, Xu SY (2011) Natural and engineered nicking endonucleases--from cleavage mechanism to engineering of strand-specificity. Nucleic Acids Res 39:1-18

42. Imanishi M, Negi S, Sugiura Y (2010) Non-FokI-based zinc finger nucleases. Methods Mol Biol 649:337-349

43. Czene A, Németh E, Zóka IG, Jakab-Simon NI, Körtvélyesi T, Nagata K, Christensen HEM, Gyurcsik B (2013) The role of the N-terminal loop in the function of the colicin E7 nuclease domain. J Biol Inorg Chem 18:309-321

44. Chak K, Kuo W, Lu F, James R (1991) Cloning and Characterization of the Cole7 Plasmid. J Gen Microbiol 137:91-100

45. Liao C, Hsiao K, Liu Y, Leng P, Yuen HS, Chak K (2001) Processing of DNase Domain during Translocation of Colicin E7 across the Membrane of Escherichia coli. Biochem Biophys Res Commun 284:556-562

46. Hsia K, Chak K, Liang P, Cheng Y, Ku W, Yuan HS (2004) DNA Binding and Degradation by the HNH Protein ColE7. Structure 12:205-214

47. Ko T, Liao C, Ku W, Chak K, Yuan HS (1999) The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. Structure 7:91-102

48. Wang Y-, Wright J, D., Doudeva LGH-, Lim C, Yuan HS (2009) Redesign of high-affinity nonspecifid nucleases with altered sequence preference. J Am Chem Soc 131:17345-17353

49. Cascales E, Buchanan SK, Duche D, Kleanthous C, Lloubes R, Postle K, Riley M, Slatin S, Cavard D (2007) Colicin biology. Microbiol Mol Biol Rev 71:158-229

50. Cheng Y, Hsia K, Doudeva LG, Chak K, Yuan HS (2002) The Crystal Structure of the Nuclease Domain of Colicin E7 Suggests a Mechanism for Binding to Double-stranded DNA by the H–N–H Endonucleases. J Mol Biol 324:227-236

51. Sui M, Tsai L, Hsia K, Doudeva L, Ku W, Han G, Yuan H (2002) Metal ions and phosphate binding in the H-N-H motif: Crystal structures of the nuclease domain of ColE7/Im7 in complex with a phosphate ion and different divalent metal ions. Protein Sci 11:2947-2957

52. Doudeva L, Huang D, Hsia K, Shi Z, Li C, Shen Y, Cheng Y, Yuan H (2006) Crystal structural analysis and metal-dependent stability and activity studies of the ColE7 endonuclease domain in complex with DNA/Zn2+ or inhibitor/Ni2+. Protein Sci 15:269-280

53. Wang Y, Yang W, Li C, Doudeva LG, Yuan HS (2007) Structural basis for sequence-dependent DNA cleavage by nonspecific endonucleases. Nucleic Acids Res 35:584-594

54. Huang H, Yuan HS (2007) The Conserved Asparagine in the HNH Motif Serves an Important Structural Role in Metal Finger Endonucleases. J Mol Biol 368:812-821

55. Levin KB, Dym O, Albeck S, Magdassi S, Keeble AH, Kleanthous C, Tawfik DS (2009) Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nat Struct Mol Biol 16:1049-1055

56. Chen C, Pettitt BM (2011) The binding process of a nonspecific enzyme with DNA. Biophys J 101:1139-1147

57. Ku W, Liu Y, Hsu Y, Liao C, Liang P, Yuan H, Chak K (2002) The zinc ion in the HNH motif of the endonuclease domain of colicin E7 is not required for DNA binding but is essential for DNA hydrolysis. Nucleic Acids Res 30:1670-1678

58. Bueren-Calabuig JA, Coderch C, Rico E, Jimenez-Ruiz A, Gago F (2011) Mechanistic insight into the catalytic activity of betabetaalpha-metallonucleases from computer simulations: Vibrio vulnificus periplasmic nuclease as a test case. Chembiochem 12:2615-2622

59. Pommer AJ, Cal S, Keeble AH, Walker D, Evans SJ, Kühlmann UC, Cooper A, Connolly BA, Hemmings AM, Moore GR, James R, Kleanthous C (2001) Mechanism and cleavage specificity of the H-N-H endonuclease colicin E9. J Mol Biol 314:735-749

60. Walker DC, Georgiou T, Pommer AJ, Walker D, Moore GR, Kleanthous C, James R (2002) Mutagenic scan of the H-N-H motif of colicin E9: implications for the mechanistic enzymology of colicins, homing enzymes and apoptotic endonucleases. Nucleic Acids Res 30:3225-3234

61. Garinot-Schneider C, Pommer AJ, Moore GR, Kleanthous C, James R (1996) Identification of Putative Active-site Residues in the DNase Domain of Colicin E9 by Random Mutagenesis. J Mol Biol 260:731-742

62. Shi Z, Chak K, Yuan H (2005) Identification of an essential cleavage site in ColE7 required for import and killing of cells. J Biol Chem 280:24663-24668

63. Németh E, Körtvélyesi T, Thulstrup PW, Christensen HEM, Kozísek M, Nagata K, Czene A, Gyurcsik B (2014) Fine tuning of the catalytic activity of colicin E7 nuclease domain by systematic N-terminal mutations. Protein Science 23:1113-1122

64. Maté MJ, Kleanthous C (2004) Structure-based Analysis of the Metal-dependent Mechanism of H-N-H Endonucleases. J Biol Chem 279:34763-34769

65. Li CL, Hor LI, Chang ZF, Tsai LC, Yang WZ, Yuan HS (2003) DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. EMBO J 22:4014-4025

66. Walker P, Leong L, NG P, Tan S, Waller S, Murphy D, Porter A (1994) Efficient and Rapid Affinity Purification of Proteins using Recombinant Fusion Proteases. Bio-Technology 12:601-605

67. Richey J (1982) FPLC: a comprehensive separation technique for biopolymers. Am Lab 14:104-129

68. Madadlou A, O'Sullivan S, Sheehan D (2011) Fast protein liquid chromatography. Methods Mol Biol 681:439-447

69. Karas M, Bahr U, Dulcks T (2000) Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. Fresenius J Anal Chem 366:669-676

70. Yu X, Wojciechowski M, Fenselau C (1993) Assessment of metals in reconstituted metallothioneins by electrospray mass spectrometry. Anal Chem 65:1355-1359

71. Kaltashov IA, Bobst CE, Zhang M, Leverence R, Gumerov DR (2012) Transferrin as a model system for method development to study structure, dynamics and interactions of metalloproteins using mass spectrometry. Biochim Biophys Acta 1820:417-426

72. van den Bremer ET, Jiskoot W, James R, Moore GR, Kleanthous C, Heck AJ, Maier CS (2002) Probing metal ion binding and conformational properties of the colicin E9 endonuclease by electrospray ionization time-of-flight mass spectrometry. Protein Sci 11:1738-1752

73. Leavitt S, Freire E (2001) Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr Opin Struct Biol 11:560-566

74. Wiseman T, Williston S, Brandts JF, Lin LN (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem 179:131-137

75. Kelly SM, Jess TJ, Price NC (2005) How to study proteins by circular dichroism. Biochim Biophys Acta 1751:119-139

76. Miles AJ, Wallace BA (2006) Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. Chem Soc Rev 35:39-51

77. Whitmore L, Wallace BA (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. Biopolymers 89:392-400

78. Miles AJ, Hoffmann SV, Tao Y, Janes RW, Wallace BA (2007) Synchrotron Radiation Circular Dichroism (SRCD) spectroscopy: New beamlines and new applications in biology. J Spectroscopy 21:245-255

79. Vasák M, Kagi JHR, Sigl H (1973) Metal Ions in Biological Systems. In: . Dekker, New York, pp 213

80. Buxbaum E (2011) Determination of Protein Concentration. In: . Springer US, pp 183-186

81. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T (1995) How to measure and predict the molar absorption coefficient of a protein. Protein Sci 4:2411-2423

82. Asherie N (2004) Protein crystallization and phase diagrams. Methods 34:266-272

83. Page R, Grzechnik SK, Canaves JM, Spraggon G, Kreusch A, Kuhn P, Stevens RC, Lesley SA (2003) Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. Acta Crystallogr D Biol Crystallogr 59:1028-1037

84. Newman J, Egan D, Walter TS, Meged R, Berry I, Ben Jelloul M, Sussman JL, Stuart DI, Perrakis A (2005) Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. Acta Crystallogr D Biol Crystallogr 61:1426-1431

85. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 275:1-21

86. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF,Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535-542

87. Marrington R, Dafforn T, Halsall D, Rodger A (2004) Micro-volume Couette flow sample orientation for absorbance and fluorescence linear dichroism. Biophys J 87:2002-2012

88. Norden B, Kubista M, Kurucsev T (1992) Linear Dichroism Spectroscopy of Nucleic-Acids. Q Rev Biophys 25:51-170

89. Hicks MR, Rodger A, Thomas CM, Batt SM, Dafforn TR (2006) Restriction enzyme kinetics monitored by UV linear dichroism. Biochemistry 45:8912-8917

90. Gololobov GV, Chernova EA, Schourov DV, Smirnov IV, Kudelina IA, Gabibov AG (1995) Cleavage of supercoiled plasmid DNA by autoantibody Fab fragment: application of the flow linear dichroism technique. Proc Natl Acad Sci U S A 92:254-257

91. Matsuoka Y, Nielsen PE, Norden BJ (1984) On the structure of active chromatin. A flow linear dichroism study of chromatin fractionated by nuclease digestion. FEBS Lett 169:309-312

92. Lyubchenko YL (2011) Preparation of DNA and nucleoprotein samples for AFM imaging. Micron 42:196-206

93. Zheng J, Li Z, Wu A, Zhou H (2003) AFM studies of DNA structures on mica in the presence of alkaline earth metal ions. Biophys Chem 104:37-43

94. Rivetti C, Guthold M, Bustamante C (1996) Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. J Mol Biol 264:919-932

95. Lyubchenko YL, Shlyakhtenko LS (2009) AFM for analysis of structure and dynamics of DNA and protein-DNA complexes. Methods 47:206-213

96. Bao Q, Chen H, Liu Y, Yan J, Droge P, Davey CA (2007) A divalent metal-mediated switch controlling protein-induced DNA bending. J Mol Biol 367:731-740

97. Seong GH, Kobatake E, Miura K, Nakazawa A, Aizawa M (2002) Direct atomic force microscopy visualization of integration host factor-induced DNA bending structure of the promoter regulatory region on the Pseudomonas TOL plasmid. Biochem Biophys Res Commun 291:361-366

98. Rajavelu A, Jurkowska RZ, Fritz J, Jeltsch A (2012) Function and disruption of DNA methyltransferase 3a cooperative DNA binding and nucleoprotein filament formation. Nucleic Acids Res 40:569-580

99. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173-1213

100. Rezac J, Fanfrlik J, Salahub D, Hobza P (2009) Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. J Chem Theory Comput 5:1749-1760

101. Korth M, Pitonak M, Rezac J, Hobza P (2010) A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods. J Chem Theory Comput 6:344-352

102. Stewart JJP (1996) Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. Int J Quantum Chem 58:133-146

103. Stewart JJP (2009) Application of the PM6 method to modeling proteins. J Mol Model 15:765-805

104. Ryde U, Nilsson K (2003) Quantum chemistry can locally improve protein crystal structures. J Am Chem Soc 125:14232-14233

105. Rulisek L, Ryde U (2006) Structure of reduced and oxidized manganese superoxide dismutase: a combined computational and experimental approach. J Phys Chem B 110:11511-11518

106. Ryde U (1996) The coordination of the catalytic zinc in alcohol dehydrogenase studied by combined quantum-chemical and molecular mechanics calculations. J Comput Aided Mol Des 10:153-164

107. Ryde U, Olsson MHM (2001) Structure, strain, and reorganization energy of blue copper models in the protein. International Journal of Quantum Chemistry 81:335-347

108. Németh E, Schilli GK, Nagy G, Hasenhindl C, Gyurcsik B, Oostenbrink C (2014) Design of a colicin E7 based chimeric zinc-finger nuclease. J Comp-Aided Mol Des 28:841-850

109. Stewart JJP (2008) . MOPAC2009, Colorado Springs, Stewart Computational Chemistry, http://openmopac.net

110. Czene A, Tóth E, Németh E, Otten H, Poulsen JN, Christensen HEM, Rulíšek L, Nagata K, Larsen S, Gyurcsik B (2014) A new insight into the zinc-dependent DNA-cleavage by the colicin E7 nuclease: a crystallographic and computational study. J Biol Inorg Chem, DOI: 10.1039/c4mt00195h

111. Kühlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C (2000) Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. J Mol Biol 301:1163-1178

112. van Dijk M, Bonvin AM (2009) 3D-DART: a DNA structure modelling server. Nucleic Acids Res 37:W235-9

113. Keenholtz RA, Mouw KW, Boocock MR, Li NS, Piccirilli JA, Rice PA (2013) Arginine as a general acid catalyst in serine recombinase-mediated DNA cleavage. J Biol Chem 288:29206-29214

114. Silva PJ, Schulz C, Jahn D, Jahn M, Ramos MJ (2010) A tale of two acids: when arginine is a more appropriate acid than H3O+. J Phys Chem B 114:8994-9001

115. Németh E, Körtvélyesi T, Kožíšek M, Thulstrup PW, Christensen HEM, Nagata M, Asaka, Kyosuke, Gyurcsik B (2014) Substrate binding activates the designed triple mutant of the colicin E7 metallonuclease. J Biol Inorg Chem, DOI: 10.1007/s00775-014-1186-6

116. Mosbahi K, Walker D, Lea E, Moore G, James R, Kleanthous C (2004) Destabilization of the colicin E9 endonuclease domain by interaction with negatively charged phospholipids - Implications for colicin translocation into bacteria. J Biol Chem 279:22145-22151

117. Flanagan J, M., Kataoka M, Shortle D, Engelman D, M. (1992) Truncated staphylococcal nuclease is compact but disordered. Proc Natl Acad Sci U S A 89:748-752

118. Samanta U, Pal D, Chakrabarti P (2000) Environment of tryptophan side chains in proteins. Proteins 38:288-300

119. Hu HY, Wu MC, Fang HJ, Forrest MD, Hu CK, Tsong TY, Chen HM (2010) The role of tryptophan in staphylococcal nuclease stability. Biophys Chem 151:170-177

120. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Res 33:382-388

121. Gyurcsik B, Czene A, Jankovics H, Jakab-Simon NI, Ślaska-Kiss K, Kiss A, Kele Z (2013) Cloning, purification and metal binding of the HNH motif from colicin E7. Protein Expr Purif 89:210-218

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

123. Sreerama N, Woody RW (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. Anal Biochem 287:252-260

124. Lin Y, Liao C, Liang P, Yuan H, Chak K (2004) Involvement of colicin in the limited protection of the colicin producing cells against bacteriophage. Biochem Biophys Res Commun 318:81-87

125. Fukada H, Takahashi K (1998) Enthalpy and heat capacity changes for the proton dissociation of various buffer components in 0.1 M potassium chloride. Proteins 33:159-166

126. Miles AJ, Janes RW, Brown A, Clarke DT, Sutherland JC, Tao Y, Wallace BA, Hoffmann SV (2008) Light flux density threshold at which protein denaturation is induced by synchrotron radiation circular dichroism beamlines. J Synchrotron Radiat 15:420-422

127. Hess B, Kutzner C, van dS, Lindahl E (2008) GROMACS 4:‰ Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 4:435-447

128. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656-1676

129. Olsson MHM, Sҩndergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput 7:525-537

130. Klamt A, Schüürmann G (1993) Cosmo - a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. J Chem Soc Perkin Trans 2:799-805

131. Rulisek L, Solomon EI, Ryde U (2005) A combined quantum and molecular mechanical study of the O2 reductive cleavage in the catalytic cycle of multicopper oxidases. Inorg Chem 44:5612-5628

132. Srnec M, Aquilante F, Ryde U, Rulisek L (2009) Reaction mechanism of manganese superoxide dismutase studied by combined quantum and molecular mechanical calculations and multiconfigurational methods. J Phys Chem B 113:6074-6086

133. Klusak V, Barinka C, Plechanovova A, Mlcochova P, Konvalinka J, Rulisek L, Lubkowski J (2009) Reaction mechanism of glutamate carboxypeptidase II revealed by mutagenesis, X-ray crystallography, and computational methods. Biochemistry 48:4126-4138

134. Rokob TA, Rulísek L (2012) Curvature correction for microiterative optimizations with QM/MM electronic embedding. Journal of Computational Chemistry 33:1197-1206

135. Treutler O, Ahlrichs R (1995) Efficient molecular numerical integration schemes. J Chem Phys. doi: http://dx.doi.org/10.1063/1.469408

136. Perdew JP, Burke K, Ernzerhof M (1996) Generalized Gradient Approximation Made Simple. Phys Rev Lett 77:3865-3868

137. Schäfer A, Horn H, Ahlrichs R (1992) Fully optimized contracted Gaussian basis sets for atoms Li to Kr. J Chem Phys. doi: http://dx.doi.org/10.1063/1.463096

138. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999-2012

139. Lee MC, Duan Y (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. Proteins 55:620-634

140. Reuter N, Dejaegere A, Maigret B, Karplus M (2000) Frontier Bonds in QM/MM Methods: A Comparison of Different Approaches. J Phys Chem A 104:1720-1735

141. Soares TA, Hunenberger PH, Kastenholz MA, Krautler V, Lenz T, Lins RD, Oostenbrink C, van Gunsteren WF (2005) An improved nucleic acid parameter set for the GROMOS force field. J Comput Chem 26:725-737

142. Beaten L (2010) Reconstruction of protein structures from polypeptide fragment libraries. Thesis, Switch Laboratory, Vrije Universiteit Brussel

143. Schmid N, Christ CD, Christen M, Eichenberger AP, van Gunsteren WF (2012) Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. Comput Phys Commun 183:890-903

144. Berendsen HJC, Postma JPM, van Gunsteren WF, Dinola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:8-3684-3690

145. Ryckaert J, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Chem Phys 23:327-341

146. Eichenberger AP, Allison JR, Dolenc J, Geerke DP, Horta BAC, Meier K, Oostenbrink C, Schmid N, Steiner D, Wang D, van Gunsteren WF (2011) GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories. J Chem Theory Comput 7:3379-3390

8. ACKNOWLEDGEMENTS

9. PUBLICATION LIST

Publications related to the dissertation

1. **E. Németh**, T. Körtvélyesi, P.W. Thulstrup, H.E.M. Christensen, M. Kožíšek, K. Nagata, A. Czene, B. Gyurcsik: Fine tuning of the catalytic activity of colicin E7 nuclease domain by systematic N-terminal mutations.
   Protein Sci., 23, 1113‑1122 (2014)
   IF = 2,861

2. A. Czene, E. Tóth, **E. Németh**, H. Otten, J.-C.N. Poulsen, H.E.M. Christensen, L. Rulíšek, K. Nagata, S. Larsen, B. Gyurcsik: A new insight into the zinc-dependent DNA-cleavage by the colicin E7 nuclease: a crystallographic and computational study.
   Metallomics, (2014); DOI: 10.1039/c4mt00195h
   IF = 3,978

3. **E. Németh**, T. Körtvélyesi, M. Kožíšek, P.W. Thulstrup, H.E.M. Christensen, M. Asaka, K. Nagata, B. Gyurcsik: Substrate binding activates the designed triple mutant of the colicin E7 metallonuclease.
   J. Biol. Inorg. Chem., (2014); DOI 10.1007/s00775-014-1186-6
   IF = 3,164

4. **E. Németh**, G.K. Schilli, G. Nagy, C. Hasenhindl, B. Gyurcsik, C. Oostenbrink: Design of a colicin E7 based chimeric zinc-finger nuclease.
   J. Comp-Aid. Mol. Des., 28, 841‑850 (2014)
   IF = 2,782

5. **E. Németh**, M. Asaka, K. Nagata, C. Oostenbrink, B. Gyurcsik: Specific nicking of DNA by a colicin E7 based zinc finger nuclease - a new design strategy.
   Angew. Chem. Int. Ed. (2014) - submitted

   ΣIF = 12,785

Further publications

1. A. Czene, E. Németh, I.G. Zóka, N.I. Jakab-Simon, T. Körtvélyesi, K. Nagata, H.E.M. Christensen and B. Gyurcsik: On the role of the N-terminal loop in the function of colicin E7 nuclease domain.
   J. Biol. Inorg. Chem., 18, 309-321 (2013)
   IF = 3,164

2. T. Körtvélyesi, K. E. Kövér, E. Németh, B. Penke, L. Szilágyi, M. Kasperkiewicz, A. Orosz: Study of the Binding Peptide-Derivatives on HSP90 Protein which have Anti-Cancer and Autoimmune Inhibitor Effects,
   Int. J. Drug. Des. Discov., 5, 1308-1317 (2014)

Conference proceedings with impact factor

1. E**. Németh**, C. Oostenbrink, B. Gyurcsik: Design of a novel type of enzymatic control in NColE7-based zinc finger nucleases.
   J. Biol. Inorg. Chem., 19, S833‑S852 (2014 - S838–S839); DOI 10.1007/s00775-014-1164-z
   IF = 3,164

Publications in national journals

1. Németh Eszter: Áramlásos lineáris dikroizmus spektroszkópia alkalmazása nukleázok tanulmányozására. Magyar Kémikusok Lapja, 2014. február

Presentations at international conferences

1. **E. Németh**, B. Gyurcsik, A. Czene, I.N. Simon, I.Gy. Zóka, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Allosteric control in the nuclease domain of colicin E7. *QBIC-3 3rd Quantum Bioinorganic Chemistry Conference*, Cesky Krumlov, Czech Republic, 25-28 June, (2011) – poster

2. B. Gyurcsik, A. Czene, I.N. Simon, **E. Németh**, I.Gy. Zóka, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Allosteric control in the nuclease domain of colicin E7, International Workshop on Metal Containing Drugs, Szeged, Hungary, 30-31 August (2011) – lecture

3. **E. Németh**, B. Gyurcsik, T. Körtvélyesi, P. W. Thulstrup, H.E.M. Christensen: Computational design, expression and spectroscopic study of a mutant NColE7 metallonuclease. *XII International Symposium on Inorganic Biochemistry, Collaboration and Beyond*, Wrocław, Poland, August 28 – September 1 (2013) - lecture

4. **E. Németh**, T. Körtvélyesi, P. W. Thulstrup, H.E.M. Christensen, B. Gyurcsik: Fine tuning of the catalytic activity by the positively charged residues at the N-terminus of the ColE7 nuclease. *6th Central Europe Conference - Chemistry towards Biology,* Trieste, Italy, September 10-13 (2013) – lecture

5. **E. Németh**, C. Oostenbrink, B. Gyurcsik: Design of a novel type of enzymatic control in NColE7-based zinc finger nucleases. *EUROBIC12, 12th European Biological Inorganic Chemistry Conference*, Zürich, Switzerland, August 24-28 (2014) – poster

6. **E. Németh**, J. Kopniczky, E. Balázs, B. Gyurcsik: Cooperative DNA-binding of the colicin E7 nuclease domain. *Significance of Knotted Structures for Function of Proteins and Nucleic Acids - Biophysical Society meeting*, Warsaw, Poland, September 17-21 (2014) - poster

7. **E. Németh**, M. Asaka, K. Nagata, C. Oostenbrink, B. Gyurcsik: Redesign of the colicin E7 nuclease domain into a controlled and specific artificial enzyme.*Tsukuba Global Science Week 2014*, Tsukuba, Japan, September 28-30 (2014) – invited lecture

Coauthor at international conferences

1. B. Gyurcsik, A. Czene, N.I. Simon, **E. Németh**, I.G. Zóka, G. Nagy, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Study on the HNH motif as a possible component of a controlled zinc-finger nuclease. *International Conference on Gene Targeting*, Vienna, Austria, 9-12th February (2011), - lecture

2. B. Gyurcsik, A. Czene, N.I. Simon, **E. Németh**, I.G. Zóka, G. Nagy, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Study on the HNH motif as a possible component of a controlled zinc-finger nuclease. *New Trends in Coordination, Bioinorganic and Applied Inorganic Chemistry; XXIII. International Conference on Coordination and Bioinorganic Chemistry*, Smolenice, Slovak Republic, 5-10th June (2011), - lecture

3. B. Gyurcsik, A. Czene, N.I. Simon, **E. Németh**, I.G. Zóka, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Study on the HNH motif as a possible component of a controlled zinc-finger nuclease. *4th European Conference on Chemistry for Life Sciences (4ECCLS)*, Budapest, Hungary, August 31- September 3 (2011) - lecture

4. B. Gyurcsik, A. Czene, N.I. Simon, **E. Németh**, I.G. Zóka, T. Körtvélyesi, H.E.M. Christensen, K. Nagata: Design of a novel artificial nuclease based on the HNH. *ISABC11 - 11th International Symposium on Applied Bioinorganic Chemistry*, Barcelona, Spain, 2-5 December (2011)-poster

5. Czene, **E. Németh**, I.G. Zóka, N.I. Simon, B. Gyurcsik, H.E.M. Christensen, K. Nagata: Mutant Colicin E7 proteins reveal the conditions for allosteric control of the enzymatic action. *ISABC11 - 11th International Symposium on Applied Bioinorganic Chemistry*, Barcelona, Spain, 2-5 December (2011)-poster

6. B. Gyurcsik, A. Czene, **E. Németh**, I.G. Zóka, E. Endreffy, H.E.M. Christensen, K. Nagata: Targeting the breakpoint in duchenne muscular dystrophy. *Nobel75, 75th Anniversary of Albert Szent-Györgyi's Nobel Prize Award*, Szeged, Hungary, 22-25 March (2012)-lecture

7. B. Gyurcsik, A. Czene, **E. Németh**, I.G. Zóka, E. Endreffy, H.E.M. Christensen, K. Nagata: Targeting the breakpoint in Duchenne Muscular Dystrophy *ISMEC 2012, International Symposia on Metal Complexes,* Lisbon, Portugal, June 18th - 22nd, ISMEC Acta, Vol. 2, p.111 (2012)-poster

8. B. Gyurcsik, A. Czene, **E. Németh**, I.G. Zóka, E. Endreffy, H.E.M. Christensen, K. Nagata: Targeting the breakpoint in Duchenne Muscular Dystrophy by an artificial metallonuclease. *EUROBIC11, 11th European Biological Inorganic Chemistry Conference*, Granada, Spain September 12-16 (2012) - poster

9. B. Gyurcsik, A. Czene, **E. Németh**, I.Gy. Zóka, E. Endreffy, H.E.M. Christensen, K. Nagata: Design of a novel artificial metallonuclease for targeting the breakpoint in Duchenne Muscular Dystrophy. *The 3rd Leading Graduate Schools International Conference*, Tsukuba, Japan, November 1-2 (2012) - invited lecture

10. K. Borsos, **E. Németh**, B. Gyurcsik: The interaction between a truncated mutant of NColE7 nuclease and its immunity protein. *13th Edition of Academic Days Timisoara - New trends and strategies in the chemistry of advanced materials with relevance in biological systems, technique and environmental protection*, Timisoara, Romania, June 13-14 (2013) - poster

11. B. Gyurcsik, A. Czene, **E. Németh**, I.G. Zóka, E. Endreffy, H.E.M. Christensen, K. Nagata: The nuclease domain of colicin E7 as a potential template for artificial nuclease design. *13th Edition of Academic Days Timisoara - New trends and strategies in the chemistry of advanced materials with relevance in biological systems, technique and environmental protection*, Timisoara, Romania, June 13-14 (2013) - plenary lecture

12. B. Gyurcsik, **E. Németh**, P. W. Thulstrup, M. Kožíšek, H.E.M. Christensen, K. Nagata: Fine tuning of the allosteric activation in the nuclease domain of colicin E7. *Global Science Week 2013 - 'Integration of Chemistry and Life Science'*, Tsukuba, Japan, October 2-4 (2013) – invited lecture

13. **E. Németh**, G. Schilli, C. Koncz, A. Czene, M. Kožíšek H.E.M. Christensen, K. Nagata, B. Gyurcsik: The effect of the metal ions on the active centre of colicin E7 nuclease. *ISMEC 2014, International Symposium on Metal Complexes*, Pavia, Italy, June 08-12 (2014) - lecture

14. **E. Németh**, C. Oostenbrink, M. Asaka, K. Nagata, B. Gyurcsik: Novel zinc finger nucleases based on colicin E7 nuclease domain. *CTB 2014 - 7th Central Europe Conference Chemistry Towards Biology*, Katowice, Poland, September 9-12 (2014) - lecture

Presentations at national conferences

1. Roósz B., **Németh E.**, Hoffmann A. E., Körtvélyesi T.: Kukurbit[7]uril "host-guest" komplexeinek elméleti vizsgálata. KeMoMo-QSAR szimpózium, Szeged (2010) - poszter

2. **Németh E.**, Keserű Gy. M.: A Caco-2 permeabilitás in silico predikciója. A Magyar Tudomány Ünnepe, Szeged (2010) - előadás

3. **Németh E.**, Keserű Gy. M.:A Caco-2 permeabilitás in silico predikciója, XXXIII. Kémiai Előadói Napok, Szeged (2010) - előadás

4. **Németh E.**, Zóka I.Gy., Czene A., Jakab-Simon I.N., Gyurcsik B., Körtvélyesi T., Nagata K., Christensen H.E.M.: A colicin E7 fehérje nukleáz doménjének vizsgálata. *MKE, 1. Nemzeti konferencia*, Sopron 2011. május 22-25 - előadás

5. **Németh E.**, Gyurcsik B., Körtvélyesi T: Allosztérikus kontroll a Colicin E7 fehérje nukleáz doménjében. *XXXIV. Kémiai Előadói Napok*, Szeged, 2011. November 2-4 - előadás

6. **Németh E.**, Schilli G., Gyurcsik B., Körtvélyesi T.: Pontmutációk tanulmányozása a colicin E7 fehérje nukleáz doménjében. *46. Komplexkémiai Kollokvium és az MTA Koordinációs Kémiai Munkabizottság ülése,* 2012. május 21-23, Mátrafüred - előadás

7. **Németh E.**, Gyurcsik B., Körtvélyesi T: A colicin E7 metallonukleáz domén finomhangolása. *XXXVI. Kémiai Előadói Napok*, 2013. október 28-30. Szeged - előadás

8. **Németh E.**, Czene A., Gyurcsik B.: Egy fehérjealapú mesterséges metallonukleáz szabályozásának lehetősége. *MTA Koordinációs Kémiai és Peptidkémiai Munkabizottság közös ülése*, 2013. október 28, Budapest - előadás.

Coauthor at national conferences

1. Zóka I. Gy., Gyurcsik B., Czene A., **Németh E.**: A colicin E7 nukleáz N-terminális végén rövidített mutáns fehérjéinek vizsgálata. *XXXIV. Kémiai Előadói Napok*, Szeged, 2011. November 2-4 - előadás

2. Gyurcsik B., Czene A., **Németh E.**, Tóth E., Zóka I.Gy.: Aktiváló allosztérikus szabályozás kialakításának lehetősége egy Colicin E7 alapú mesterséges metallonukleázban. *46. Komplexkémiai Kollokvium és az MTA Koordinációs Kémiai Munkabizottság ülése,* 2012. május 21-23, Mátrafüred - előadás

3. **Németh E.**, Czene A., Borsos K., P.W. Thulstrup, Gyurcsik B.: A colicin E7 metallonukleáz aktivitásának hangolása az N-terminális végen kialakított pontmutációk révén. *47. Komplexkémiai Kollokvium és az MTA Koordinációs Kémiai Munkabizottság ülése*, 2013. május 29-31, Mátraháza - előadás.

4. Schilli G., **Németh E.**, Gyurcsik B.: Egy mutáns NColE7 metallonukleáz előállítása *XXXVI. Kémiai Előadói Napok*, 2013. október 28-30. Szeged - előadás

10. APPENDIX

10.1. METHODS FOR THE N-TERMINAL POINT MUTANTS

10.1.1. Cloning and toxicity experiments

The pQE70 plasmid containing the genes of NColE7 and the Im7 immunity protein (a generous gift from prof. K.-F. Chak, Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei, Taiwan) served as a template for the amplification of DNA segments including the gene of the native and mutated NColE7. The primers applied in PCR are collected in Table S1. The obtained fragments were cloned into a pGEX-6P-1 vector (GE Healthcare BioSci.) within the EcoRI and XhoI cloning sites. The plasmids were transformed into *E. coli* DH10B cells and spread on an LB/Amp (10 μg/ml ampicillin) plates. The colonies (usually 1-3 colony / 100 μl transformed cell) were cultivated in LB/Amp (10 μg/ml ampicillin) solution, which was then sedimented and the plasmids were purified with the QIAGEN Mini Plasmid Purification kit. The purified plasmids containing the gene of the mutant proteins were used as templates in a further PCR with the pGEX sequencing primers, and the products were sequenced. Inactive mutants were selected by *Escherichia coli* DH10B cells. Since NColE7 itself is toxic for the cells, in native bacteria it is coexpressed with its cognate immunity protein (Im7) [50, 51]. Here in the absence of Im7 gene the cloning procedure implied the selection of genes of nontoxic proteins.

10.1.2. Cloning, protein expression and purification

The pQE70 plasmid containing the NColE7 and Im7 immunity protein genes was a generous gift from prof. K.-F. Chak, Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei, Taiwan [44, 124]. From this template the primers applied in PCR (collected in Table S1.) amplified DNA segments including the genes of the native or mutated NColE7 as well as the Im7 protein. The obtained fragments were cloned into a pGEX-6P-1 vector (GE Healthcare BioSci.) providing an N-terminal glutathione-S-transferase (GST) affinity fusion tag. The plasmids were transformed into *E. coli* DH10B cells for DNA cloning and into *E. coli* BL21 (DE3) cells for protein production in 3×650 ml LB/Amp culture. At OD_{600} 0.6-0.7 the protein expression was induced with isopropyl β-*D*-1-thiogalactoside (IPTG, final concentration 0.1 mM), and the incubation was continued for further two hours at 25 °C. Cells were then sedimented by centrifugation and the pellets were resuspended in 50 ml PBS buffer (0.14 M NaCl, 2.7 mM KCl, 10.0 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.3) by sonication. The soluble fractions were loaded on a GST affinity chromatography column (GSTPrepFF16/10, GE Healthcare BioSci.). The fusion proteins were cleaved on column with Human rhinovirus C3 protease [66] - sold as PreScission protease by GE Healthcare - to remove the GST tag. 20 ml of 10 μM protease in PBS was loaded on the column and the reaction was continued overnight at 4 °C or 2 hours at room temperature followed by the elution with PBS. The fractions containing the nuclease mutants in complex with the immunity protein as well as the protease were collected. To disrupt the interaction between the NColE7 mutants and the Im7 protein the pH was adjusted to 3.0 after

a 3× dilution with a 20 mM Gly/HCl buffer. The components were then separated on a Sepharose SP FF 16/10 cation exchange column with a binding buffer 20 mM Gly/HCl pH = 3.0 and a gradient of 0-2 M NaCl in 30×column volume (CV). The immunity protein was eluted at pH = 8.0 with PBS conating 2M NaCl. The fractions of the nucleases or Im7 were concentrated by Amicon ultrafilter with 5 kDa cutoff, and the buffer was exchanged to 20 mM HEPES, pH = 7.7.

10.1.3. Nano-electrospray ionization mass spectrometry

Mass spectra were obtained on a LCT Premier (Waters) instrument equipped with a Nanoflow Electrospray Ionization (nano-ESI) source and a time-of-flight (TOF) analyzer. The instrument was operated in positive ion mode and it was calibrated using 100 mg/ml CsI in 50% 2-propanol in the *m/z* range from 600 to 12000. Samples were sprayed from middle size Au/Pd-coated borosilicate glass capillary needles (Proxeon) loaded with 3 µl protein solution. The protein concentration was between $10 - 20$ µM in 500 mM ammonium acetate (Sigma) buffer. For the study of metal binding, 1 equivalent Zn^{2+}-acetate was added to the protein samples before measurement. The de-salting of the protein solution and buffer exchange to the volatile buffer was done using Micro BioSpin chromatography column (BioRad). The needle voltage was typically around 1200 V and 50 V cone voltage was applied, with a cone gas maintained at 20 L/h and the source temperature was maintained at 50 °C. The recorded *m/z* data were deconvoluted using the MassLynx^{TM} v4.1 (Waters) software equipped with the MaxEnt1 algorithm. The high charge states of the multiply charged spectrum, ranging from +10 to +17, were used to calculate the apparent mass.

10.1.4. Isothermal titration calorimetry

Isothermal calorimetric titrations were performed on a MicroCal Auto ITC-200 (GE) instrument. The protein samples (~ 50 µM) were prepared by 12 hour dialysis in 7000 MWCO Thermo Scientific Slide-A-Lyzer casettes, against 20 mM cacodylate buffer, pH = 7.0. $ZnCl_2$ was dissolved in the same buffer. The ionization enthalpy of cacodylate – applied as buffer – is close to zero (-0.47 kcal/mol [125]), so the contribution of protonation/deprotonation processes, if there were any present, was negligible in the observed ΔH. The dilution heat of $ZnCl_2$ with the buffer was determined for each experiment and the integrated data of dilution heats were subtracted from the corresponding data of protein titrations. A control titration of the KGNK mutant with plain buffer was done showing no significant effects (data not shown). The enthalpy change during the titrations of 200 µl protein solutions with 2 µl aliquots of 400 µM $ZnCl_2$ up to 40 µl (spacing of 240 s) can be consequently attributed to the metal binding or competition processes. Instead of degassing the samples before titration the plates were shortly centrifuged.

10.1.5. Gel mobility shift assays

In the protein-DNA binding studies the concentration of the 13 base pair oligonucleotide was 0.2 µM, and the protein final concentration ranged between 0 and 7 µM. The ^{32}P-radiolabelled DNA was hybridized from one single oligonucleotide with

complementary sequence at each end forming a loop. The samples were run on 6% native PAGE at 4 °C . The reaction mixture contained 4 mM NaCl, 4 mM HEPES buffer (pH = 7.9) and 100 µM EDTA to inhibit DNA digestion. For DNA cleavage studies 450 ng/well 2686 base pairs pUC19 plasmid DNA was applied, and the protein concentration was 2.8 µM. Zn^{2+}-ions were added to proteins prior the reaction in 1:1 molar ratio. 10 µl of the reaction was loaded onto an ethidium bromide containing 1% agarose gel. The electrophoresis was performed in a Bio-Rad Wide Mini Sub Cell® GT system at 6.7 V/cm in TAE buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA, pH = 8.0). For comparison 6 µl of the Bio-Rad 1 kbp marker DNA was also loaded to the gel.

10.1.6. Circular dichroism spectroscopy

The Synchrotron Radiation Circular Dichroism (SRCD) spectra were recorded at the SRCD facility at the CD1 beamline at the Institute for Storage Ring Facilities (ISA), University of Aarhus, Denmark [78, 126]. All spectra were recorded with 1 nm steps and a dwell time of 2 s per step, in 100.4 µm quartz cells (SUPRASIL, Hellma GmbH, Germany). The concentration of the protein solutions was 3.2×10^{-5} M in 10 mM HEPES, pH = 7.7.

10.1.7. Linear dichroism spectroscopy

Flow linear dichroism spectra were measured on a Jasco-815 CD spectrophotometer equipped for linear dichroism spectroscopy (LD), using a microvolume Couette flow device as described in [87]. An additional quartz lens was mounted to allow for focusing onto the sample in the Couette cell, which was positioned as close to the photomultiplier as possible in the J-815 sample compartment. In the Couette cell, an outer quartz cylinder rotates, and an inner quartz rod is static, the annular gap is 0.25 mm giving a combined light path of 0.5 mm. Water circulation through the metal block of the flow device thermostated the cell to 298.0 K. The sample volume was 70 µl and 3000 rpm rotation was applied. The optical bandwidth was 1 nm and the spectra were recorded in continuous mode between 190 and 400 nm with 50 nm/min scanning speed, 1 s integration time, 0.5 nm data pitch, and with 15 L/min nitrogen flow. LD of double stranded DNA samples yields a characteristic negative signal at the absorbance maximum of the DNA bases (*ca.* 260 nm) since the transition moments of the base π- π* transitions are located in the plane of the bases, and as these are oriented orthogonally with respect to the axis of DNA helix [88].

Protein-DNA binding was studied with a 130 µM (final concentration calculated for base pairs) *calf thymus* CT-DNA sample. The mixture contained 60 µM EDTA, 17 mM KH_2PO_4, 2.4 mM HEPES (the pH of both buffers was adjusted to 7.7) and 0-5 µM protein. The solutions were incubated for 10 min before recording the spectra. The incubation time did not influence the results.

The cleavage of 130 µM (bp) CT-DNA by different mutants was followed for 8 hours (1 spectrum/h). The proteins (40 µM) were incubated with one equivalent of $Zn(Ac)_2$ (40 µM) for 30 min at room temperature before the reaction started. The final concentration of both the protein and Zn^{2+}-ions was 0.5 µM. 720 µl of reaction mixture was incubated at 37 °C and in

each hour 70 µl aliquot was taken for LD test. All solutions were in mixed buffers of 17 mM KH_2PO_4 and 2.4 mM HEPES, the pH of both buffers was adjusted to 7.7. The spectra were smoothed with the means-movement method, convolution width of 11.

10.1.8. Atomic force microscopy

Scanning force microscopy images were taken under ambient conditions in air, with a PSIA XE-100® Atomic Force Microscope. The probe tips were purchased from NT-MDT® (type NSG01) and had a nominal radius of curvature of ≈10 nm and cone angle ≈10° at the apex. Lateral forces during scanning were minimized by operating the AFM in the intermittent contact (tapping) mode. The oscillation amplitude and setpoint parameters were adjusted to ensure low (≤ 1 nN) contact forces during imaging. The samples for AFM experiments contained 25 or 100 nM of the 777 bp DNA fragment, occasionally 50 nM protein (NColE7 or GGNG), 2 mM $MgCl_2$ and 4 mM HEPES (pH = 7.7). 20 µl aliquot was dropped onto the mica plate (diameter 1 cm) and dried on air for 10 min, followed by repeated washing with 5×20 µl water, then completely dried on air.

10.1.9. Computational methods

10.1.9.1.MD simulations

Initial conformation of NColE7 and the shortened mutants was obtained from a crystal structure 1M08 [50]. The original N-terminus of NColE7 was restored by an M446K mutation. In all models the proteins had uncapped termini (i.e. NH_3^+ and COO^- groups).

Molecular dynamics (MD) calculations were performed with GROMACS 4.05,[127] with the force field Gromos 53a6 [128]. The ionizable residues were charged according to the default pK_a values at pH = 7.2 detected by PropKa 3.0 [129]. Each protein was placed in a cubic box with edge size of *ca.* 8 nm, and solvated by explicit SPC/E water model containing about 16000 equilibrated water molecules. The system was neutralized with Cl^- ions replacing water molecules according to the electrostatic potential of the system points. Energy minimization was carried out with the steepest descent method. 200 ps position restrained dynamics was performed in NVT ensemble to equilibrate the system (solvate and generate initial velocities with Maxwell distribution) including explicit water molecules. 25 ns MD simulations were performed in the NPT ensemble with periodic box conditions. The integration step was 2 fs. The temperature was set to 300 K and isotropic Berendsen p-coupling and T-coupling was applied. For Coulomb interactions PME was applied with 0.9 nm cut-off for electrostatic and 1.6 nm for van der Waals interactions. The dielectric constant was set to 1.0. The LINCS constraint algorithm was used. Trajectories were analyzed starting at 500 ps.

10.1.9.2. Semi-empirical quantum chemical calculations

Semi-empirical quantum chemical computations were performed on the proteins including a Zn^{2+} and a phosphate ion with the PM6 method implemented in MOPAC2009 [99, 102, 103, 109]. Localized molecular orbitals were applied by the MOZYME [102]

model. The solvation was considered by COSMO method [130] with the dielectric constant of 78.4. The geometry optimization was carried out by the L-BFGS method after the initial minimization of the hydrogen positions. The gradient norm was set to 1.0 kcal/mol/Å. The thermodynamic parameters were computed with the PM6 method, then recalculated in one SCF cycle with the PM6-DH2 correction.

10.1.9.3. QM/MM calculations on protein – DNA complexes

Protein setup for QM/MM calculations

The protein setup was performed based on the crystal structures of the mutant D493Q-NColE7 (3FBD, [48]) and 1M08 [50]. Residue D493 was restored in an orientation obtained by manual superposition with the corresponding parts of the 2IVH [53] structure containing a H545Q mutant NColE7 in complex with a 18 bp DNA substrate and a Zn^{2+}-ion. The last five base pairs of the DNA molecule (not interacting with the protein in our model) and S445 were deleted. As in the 3FBD structure there was no Zn^{2+}-ion, it was positioned into the metal binding site of the HNH motif, using the Zn^{2+}-containing 1M08 structure as the template. The metal ion binding H544, H569 and H573 were protonated at the uncoordinated N^{ε}, N^{δ} and N^{δ} nitrogens, respectively, whereas H545 at the N^{ε} to be suitable to activate the nucleophilic water molecule.

The construction of the QM/MM system followed the recommended protocol [106, 107] that is almost identical to those used in our previous QM/MM studies [131-133] The quantum system (System 1) consisted of 169 atoms for NColE7 and 156 for ΔN4-NColE7 (**Fig. 10.1**) including the Zn^{2+}-ion, H544, H545, H569, H573, V555 and the sidechain of R447 (only for NColE7). The MM part was divided into System 2 (surrounding of the quantum region) that is allowed to move in the QM/MM minimizations and System 3 (the rest of the protein) that is kept fixed.
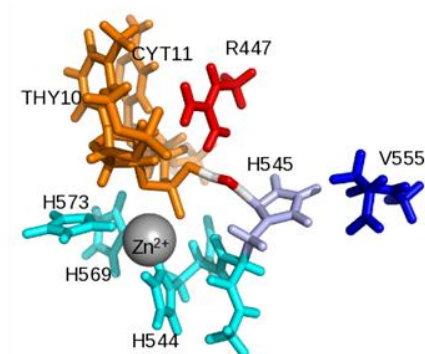
Fig. 10.1. *Quantum region (system 1) of the NColE7/DNA model, containing 169 atoms. At the dangling bonds hydrogen caps are used. The quantum region of ΔN4-NColE7/DNA (156 atoms) is the same, except for the missing R447 residue.*

The protein-DNA complex was solvated by water molecules (keeping also those resolved in the crystal structure) in the sphere with a radius $R = 35$ Å from its centre (~ 4000 water molecules in total). An additional water molecule was inserted into the putative position for the catalytic water, between H545 and the scissile phosphate. Total of 42 K^+ and 32 Cl^- ions were added to neutralize the system and to achieve a salt concentration of 0.3 M.

The starting geometry for QM/MM calculations was obtained using *sander* module of an AMBER package. The minimization was carried out with heavy atoms fixed at their crystallographic position, whereas hydrogen atoms (added to the DNA-protein complex by the *Leap* module of Amber) and water molecules added to the model were allowed to relax. The oxygen atoms of catalytic and those water molecules resolved in the crystal structure were kept fixed. After initial minimization, 180 ps simulated annealing (*T* initial 353 K, *T* final 0 K, no pressure scaling, NVT ensemble) with the same set of fixed atoms as described above was carried out. Finally, the resulting structure was subjected to an additional molecular mechanics minimization. No periodic boundary conditions were used. The water molecules were kept inside the sphere by a harmonic potential with a default force constant to prevent their dissociation in the course of the MD simulation (simulated annealing).

The initial QM/MM structure of the ΔN4-NColE7 was prepared using the above procedure after deletion of the KRNK amino acid string from the N-terminus of NColE7.

Construction of the QM/MM system

The quantum system consisted of 169 atoms for NColE7 and 156 for ΔN4-NColE7. The Zn^{2+}-ion, the catalytic water molecule, the side chain of R447 (including only the positively charged guanidinium and the neighbouring -CH_2- group), the metal binding histidines and the catalytic H545, as well as, V555 (deemed important for positioning the H545 residue) was included in the model. In all cases the truncation involved "fairly inert" non-polar C-C bonds to prevent any artifact in the QM/MM calculations. Hydrogen-link approach was used to cap the dangling bonds in the QM part.

Beside the quantum region (denoted also as System 1), the standard approach in the QM/MM protocol is the division of the MM part into System 2 (surrounding of the quantum region) that is allowed to move in the QM/MM minimizations and System 3 (the rest of the protein) that is kept fixed. The cut-off for selection of System 2 was $R = 6$ Å (on per-residue basis), *i.e.* all residues containing any of their atoms 6 Å or less from the quantum system were included in the System 2. The System 2 consisted of 1439 atoms (54 residues, 8 DNA bases and 3 K^+ and 2 Cl^- ions) and 124 water molecules. The rest of the protein (System 3) consisted of 13719 atoms (75 residues, 16 DNA bases and 39 K^+ and 30 Cl^- ions and the remaining water molecules 3973).

Description of the QM/MM procedure

The QM/MM calculations were carried out employing a modified version of the ComQum program [106, 107, 134]. The Turbomole 6.3 program [135] was used for the quantum chemical calculations carried out using DFT method and implying Perdew-Burke-Ernzerhof (PBE) functional [136] and Ahlrich's def2-SVP basis set [137]. Resolution of the identity (density-fitting) was used to expedite the DFT calculations. MM calculations were carried out in the AMBER 8 program package (sander module) and the ff03 [138, 139] force field (parm99 set in Amber).

In this approach, the protein and solvent are split into three subsystems: (1) The QM region (system 1) contains the most interesting atoms and is relaxed by QM/MM forces. (2) System 2 consists of all of the residues within 6Å of any atom in system 1 and is relaxed by a full MM minimization in each step of the QM/MM geometry optimization. (3) Finally, system 3 contains the remaining part of the protein and surrounding solvent molecules and is kept fixed at the original (crystallographic) coordinates. In the quantum chemical calculations, the QM system is represented by a wave function, whereas all of the other atoms are represented by an array of partial point charges, one for each atom, taken from Amber libraries. Thereby, the polarization of the quantum chemical system by the surroundings is included in a self-consistent manner. In the MM calculations for the QM/MM forces and energies, all of the atoms are represented by the Amber force field. When there is a bond between systems 1 and 2 (a junction), the quantum region is truncated by hydrogen atoms, the positions of which are linearly related to the corresponding carbon atoms in the full system (the hydrogen-link approach [106, 140]). To avoid overpolarization of the quantum system, point charges on atoms in MM region bound to junction atoms are zeroed and the remaining charges on the truncated amino acid are adjusted to keep the fragment neutral. The total energy is calculated as

$$E_{tot} = E_{QM} + E_{MM123} - E_{MM1} \tag{1}$$

Here, E_{QM} is the QM energy of the quantum system truncated by the hydrogen atoms, excluding the self-energy of the surrounding point charges. E_{MM1} is the MM energy of the quantum system, still truncated by hydrogen atoms, but without any electrostatic interactions. Finally, E_{MM123} is the classical energy of all the atoms in the system with normal atoms at the junctions and with the charges of the quantum system set to zero (to avoid double counting of the electrostatic interactions). By the use of this approach, errors caused by the truncation of the quantum system should cancel out. The calculated forces are the gradient of this energy, but owing to the differing junctions in the various calculations, they have to be corrected using the chain rule. The geometry optimizations were carried out in two steps. First, systems 2 and 3 were frozen and only the quantum system was optimized (this structure will be referred to as Protein_Fixed). Second, both systems 1 and 2 were allowed to relax. In the MM optimization of system 2, the charges on the quantum atoms were updated in each iteration of the QM/MM optimization. This optimization was performed with the looser convergence criterion of 10^{-4} a.u. for the change in energy and 10^{-2} a.u. for the maximum norm of the Cartesian gradient (0.26 and 50 kJ.mol^{-1} Å^{-1})). Then, system 2 was frozen again, and the geometry optimization was continued with default convergence criteria (10^{-6} and 10^{-3} a.u.).

10.2. METHODS FOR THE "TKW" MUTANTS

10.2.1. Computational alanine scan

Chain *A* from the crystal structure 1M08 of the Zn^{2+}-bound NColE7 [50] was applied as the starting geometry for NColE7. Alanine mutations were built in the optimized NColE7 structure manually and the resulted models were optimized as well. All proteins had uncapped

termini (i.e. NH_3^+ and COO^- groups). Semiempirical quantum chemical computations were performed with the PM6 method [99, 103] implemented in MOPAC2009 [109]. Localized molecular orbitals were applied by the MOZYME [102] method. The solvation was considered by COSMO method [130] with the dielectric constant of 78.4. The geometry optimization was carried out by the L-BFGS method after the initial minimization of the hydrogen positions. The gradient norm was 1.0 kcal/mol/Å. The thermodynamic parameters were computed with the PM6 method, and recalculated in one SCF cycle with the correction PM6-DH2 [100, 101].

10.2.2. Cloning, protein expression and purification

The gene of the triple mutant protein was amplified by PCR from the pQE70 plasmid (a generous gift of Prof. K.-F. Chak, Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei, Taiwan - [50]) containing the gene of NColE7 and the Im7 protein. The PCR fragment was cloned into a pGEX-6P-1 vector that provides an N-terminal GST fusion. The plasmid containing the target gene was transformed into *E. coli* DH10B cells and *E. coli* BL21 (DE3) cells for cloning and protein expression, respectively. The procedure described earlier was applied for protein purification. All the proteins were stored in 20 mM HEPES, at pH = 7.7.

10.2.3. Monitoring of DNA binding and cleavage by gel electrophoresis

In the protein-DNA binding studies the concentration of the 13 base pair oligonucleotide was 0.2 µM, and the protein final concentration ranged between 0 and 7 µM. The ^{32}P-radiolabelled DNA was hybridized from one single oligonucleotide with complementary sequence at the ends forming a loop upon self hybridization. The samples were run on 6% native polyacrylamide gel electrophoresis (PAGE) at 4 °C . The reaction mixture contained 4 mM NaCl and 4 mM HEPES buffer (pH = 7.9). The DNA cleavage reaction was inhibited by 100 µM EDTA chelating the metal ion that is essential for catalytic activity [57].

The cleavage of 74 µM (base pairs) pUC19 plasmid by proteins (2.8 µM) in the presence of 2.8 µM zinc(II)-acetate was followed by agarose gel electrophoresis. Incubation time before electrophoretic run was 0-60 min at 37 °C. The CT-DNA cleavage assay contained DNA (68 µM, base pairs), protein (2.6 µM) and zinc(II)-acetate (2.6 µM). Incubation time was varied between 0 and 180 min at 37 °C. 10 µl aliquots were loaded onto an ethidium bromide containing 1% agarose gel with 2 µl of 6× DNA Loading Dye (Thermo Sci.). The electrophoresis was performed in a Bio-Rad Wide Mini Sub Cell® GT system at 6.7 V/cm in TAE buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA, pH = 8.0). For comparison 6 µl of the Bio-Rad 1 kbp marker DNA was also loaded to the gel.

10.2.4. Nano-electrospray ionization mass spectrometry

Mass spectrometric (MS) measurements were performed on a LCT Premier (Waters) instrument equipped with a nanoflow electrospray ionization (nano-ESI) source and a time-of-flight (TOF) analyzer in positive ion mode as described earlier. The recorded *m/z* data were

deconvoluted using the MassLynx^{TM} v4.1 (Waters) software equipped with the MaxEnt1 algorithm. The high charge states of the multiply charged spectrum, ranging from +10 to +17, were used to calculate the apparent mass.

10.2.5. Isothermal titration calorimetry

Isothermal calorimetric titrations were performed with the MicroCal Auto-iTC$_{200}$ (GE Healthcare Life Sciences) instrument. The protein samples (~50 μM) were dialyzed for 12 hours in 7000 MWCO Thermo Scientific Slide-A-Lyzer casettes against 20 mM cacodylate buffer, pH = 7.0. The concentration of the protein after dialysis was determined by HPLC amino acid analysis. ZnCl$_2$ was dissolved in the same buffer (c = 400 μM). The dilution heat of ZnCl$_2$ with the buffer was determined for each experiment and the integrated data of dilution heats were subtracted from the corresponding data of protein titrations. 200 μl of protein solutions were titrated with 2 μl aliquots of 400 μM ZnCl$_2$ up to 40 μl and a spacing of 240 s. Prior to titrations the plates were centrifuged to avoid bubbles in the samples.

10.2.6. Fluorescence based thermal shift assay

Thermal stability of the proteins was measured in a ROCHE LightCycler 480 instrument. The Sypro Orange (Invitrogen) fluorescent dye was applied in 625× dilution to detect protein unfolding. The protein concentration was 0.4 mg/ml with a twofold molar excess of ZnCl$_2$, in 20 mM cacodylate buffer, pH = 7.0. Samples were heated from 20 to 95 °C, with 4.40 °C/s.

10.2.7. Linear dichroism spectroscopy

Flow linear dichroism spectra were measured on a Jasco-815 CD spectrometer equipped for linear dichroism spectroscopy (LD) as described in [87]. CT-DNA cleavage was followed in reaction mixtures containing 130 μM (base pairs) CT-DNA and 5 μM TKW with 5, 25 or 50 μM zinc(II)-acetate in a microvolume Couette flow device. The samples contained 17.7 mM KH$_2$PO$_4$ and 2.5 mM HEPES buffer, pH = 7.7. 70 μL aliquots of the sample incubated at 37 °C were used to record LD spectra at room temperature. Control experiments were conducted with CT-DNA in the presence or absence of zinc(II)-acetate, TKW, EDTA or Im7. Protein binding was studied with a 130 μM (final concentration calculated for base pairs) CT-DNA sample. The reaction mixture contained 60 μM EDTA to inhibit nuclease digestion, 17 mM KH$_2$PO$_4$, 2.4 mM HEPES (the pH was adjusted to 7.7) and 0-5 μM protein. The solutions were incubated for 10 min before recording the spectra. Longer incubation times did not influence the results.

10.2.8. Circular dichroism spectroscopy

The Synchrotron Radiation Circular Dichroism (SRCD) spectra were recorded at the CD1 beamline of the storage ring ASTRID at the Institute for Storage Ring Facilities (ISA), University of Aarhus, Denmark [78, 126]. Camphor-sulfonic acid served as a calibration material for the instrument. All spectra were recorded with 1 nm steps and a dwell time of 2 s per step, using a 100.4 μm quartz cell (SUPRASIL, Hellma GmbH, Germany), in the wavelength range of 175-260 nm. The concentration of the protein solutions was adjusted to

3.2×10^{-5} M in 10 mM HEPES, pH = 7.7. From raw spectra the water baseline was subtracted.

10.3. COMPUTATIONAL DESIGN OF THE NOVEL ZFN-S

10.3.1. Construction of the initial models

The DNA/NColE7/ZF complexes were built by structural alignments based on the DNA chains of the DNA/NColE7 (2IVH, [53]) and DNA/ZF (1MEY, [12]) structures in PyMol. In the alignment process all possible combinations of 5 or 8 base pair long parts of the sequence on both DNA chains were superimposed and the DNA originating from the DNA/NColE7 complex was removed. The structures were energy minimized using the GROMOS 45A4 force field [141] and the models with the lowest energy were selected. Out of the two DNA structures the specific site of the zinc fingers was kept in the final structure, and prolonged by repeating selected parts. Thus the initial models contained a DNA molecule with a sequence GAACTATGAGGCAGAACT ("straight" models) or AACTATGAGGCAGAACTATGAGG ("reverse" models) in complex with the ZF-s and NColE7 in close proximity, so that they could be joined with a suitable linker sequence.

10.3.2. Linker design

The design of the linkers was carried out with the software package LoopX [142], an algorithm, designed to graft loop backbone coordinates into protein crystal. The underlying database of loop backbone coordinates (N, Cα, O, C) was constructed by *in silico* digestion of 14525 protein crystal structures from the ASTRAL95 dataset (less than 95% sequence identity). Linkers exhibiting appropriate length, end-to-end distance and RMSD of the respective anchor residues were identified from this database and included in the model as polyalanine chains. A position scan was carried out with FoldX [120], where all residues in a linker were mutated to the other 19 amino acids and changes in the free energy of folding were estimated. The sequences providing the energies of all designed loops are listed in Tables S2 and S3. Cysteins were excluded, because of their reactivity. Prolines were also excluded, as the computational methods seem to overestimate its stability due to covalent contributions; initially, one or more prolines were found in most of the linkers. Introducing extra prolines to the linker would introduce unwanted steric constraints to the structure. The structures were refined to avoid steric clashes, and the stability was estimated by FoldX.

The four final models were constructed using all possible combinations of the best linkers. The sequence of zinc fingers are unchanged in the models, except for cutting the last one or two residues when it was required for the linker design. NColE7 is divided into two parts: the N-terminus (N*X*) and the C-terminus (C*Y*), where *X* and *Y* refer to the number of residues involved in the model and *X* + *Y* is always less than the total number of amino acids in NColE7. The zinc finger sequence is inserted between these two parts. Accordingly, the "straight" models are named as C*Y*-ZF-N*X*, while the "reverse" models are named N*X*-ZF-C*Y*.

10.3.3. MD simulations

MD simulations were carried out using the GROMOS11 suite of simulation programs [143], with the GROMOS force field, parameter set 45A4 [141]. The energy minimized starting structures of proteins, DNA molecules and complexes were centered in periodic rectangular boxes with a minimal solute-to-wall distance 0.9 nm. The box was filled with ~12000 SPC water molecules with a minimal solute to solvent distance of 0.23 nm and subsequently minimized with the steepest descent algorithm, with a treshold of 0.1 kJ mol^{-1} to relax unfavorable interactions. Na$^+$ or Cl$^-$-ions were added to neutralize the system. Further water molecules at randomly selected positions were exchanged to Na$^+$ and Cl$^-$ -ions to achieve a NaCl concentration of 0.2 M. The system was thermalized in 5 discrete simulation steps of 20 ps at increasing temperatures, followed by another 20 ps at 298 K. The MD simulations were run for 10 ns at 298 K (relaxation time 0.1 ps) using the weak-coupling method [144]. The pressure was maintained at 1 atm using the weak-coupling method and a relaxation time of 0.5 ps and an estimated isothermal compressibility of 4.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$. All bond lengths were constrained to their optimal value by the SHAKE algorithm [145] with a relative geometric accuracy of 10^{-4}. During all calculations the Zn^{2+}-ions were kept fixed using harmonic distance restraints: a Zn^{2+}-N(His) interaction with an optimal bond length of 0.209 nm and a force constant of 14710 kJ mol^{-1} nm^{-2}, while the Zn^{2+}-S(Cys) interactions with an optimal bond length of 0.231 nm and a force constant of 18150 kJ mol^{-1} nm^{-2}. The analysis of trajectories was done by various tools of the GROMOS++ suite [146].

10.4. EXPERIMENTS WITH THE NOVEL ZFN MODELS

10.4.1. Gene construction

The genes were constructed by Béla Gyurcsik. The gene of the C105 NColE7 part of the N4-ZF-C105 protein was amplified by PCR from the pQE70 plasmid [44] (a generous gift of Prof. K.-F. Chak, Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei, Taiwan) by using the oligonucleotides Mod4R2/2F 5'-cgagaggggtgattatggcttaggttctcctgttccagatc-3' as the forward (responsible for creating an overlapping sequence with the gene of the ZF protein described below) and ColC-SE-R 5'-ccgg*aattc*ttatttacctcggtgaatatcaatatg-3' as the reverse primers. The DNA sequence of the ZF protein was synthesized from overlapping oligonucleotides by Integrated DNA Technologies (gBlocks^{TM} Gene Fragment process) based on the sequence provided in 1MEY Protein Data Bank entry.[12] The primer Mod3R1/2F: 5'-aaacggaataagggtggtgacgcattcat-ggaaaaaccgtataaatgccc-3' was applied to create the new N-terminus including the 446-KRNK-449 sequence of NColE7 (according to the original colicin E7 numbering) and the GGDAF linker sequence, while the primer Mod4R2/2R: 5'-cataatcacccctctcgaa-acgaccctgatgggtacgctgat-3' was responsible to fuse the C-terminal linker sequence GRFERGDYG to the ZF protein. The overlapping DNA sequences encoding for C105 NColE7 and ZF were fused in an extension reaction and finally the Col-N-R1/2F: 5'-aggc*catatg*aaacggaataagggtggtgac-3' and the ColC-SE-R primers were applied to amplify the

full length sequence. This was inserted into the pET-21a vector (Novagen) between the NdeI and EcoRI restriction enzyme sites (underlined sequences). The insert DNA sequence contained a C-terminal stop codon (formatted to italic in the reverse primer sequence).

10.4.2. Protein purification

The genes encoding the proteins were transformed to *E. coli* BL21(DE3) cells and the colonies were grown overnight at 37 °C. The colonies were picked and inoculated to 5 ml LB(Amp) solution and grown until $OD_{600} = 0.6$. 2×2.5 ml of the culture was used to inoculate 2×250 ml LB/Amp solution. The protein production was induced by IPTG (final concentration 0.1 mM), followed by incubation on 22 °C for further 12 hours. The cultures were centrifuged by Avanti-HP25(Beckman, 3000 G, 15 min, 4 °C) and resuspended in 25 ml cold PBS. After 4×30 s sonication the sample was centrifuged (20 min, 10000 G, 4 °C). The sonication and centrifugation steps were repeated with the supernatant. The collected 50 ml supernatant was used for protein purification.

The supernatant was applied to a gravity column containing 10 ml P11 resin slurry. After washing the column with 30 ml PBS, a four step elution was carried out with 20-20 ml of a 20 mM HEPES solution (pH = 7.9) containing 0.2, 0.5, 1.0, 2.0 M NaCl. The 3rd fraction (1 M NaCl) was diluted $10\times$ with the 20 mM HEPES buffer and applied to a HiTrap SP XL 1 ml column equilibrated with a binding buffer of 20 mM HEPES (pH = 7.9). After washing the column with 3 CV (column volume) a 30 CV gradient was performed with a NaCl concentration 0-2M. 5 ml of the fractions containing the target protein were collected and purified further with gel filtration on a HiTrap 16/60 Sephacryl S-200 HR column with a buffer containing 20 mM HEPES (pH = 7.9) and 200 mM NaCl.

10.4.3. DNA-binding experiments

DNA-binding experiments were performed with a 21 bp specific oligonucleotide (0.2 μM) with protein concentrations 0, 0.1, 0.2, 0.3, 0.4 μM. The reaction mixture contained 20 mM HEPES (pH = 7.9) and 20 mM NaCl, 1 mM $ZnSO_4$ and 10% glycerol. 0.1% of the DNA was labelled by kination with a $[\gamma\text{-}^{32}P]$ ATP. The total volume of each sample (10 μl) was loaded to a 6% native Tris-glycine gel, and visualized by Typhoon FLA 7000 (GE Healthcare). The competition assays were performed with 0.2 μM specific ^{32}P labelled DNA, 0.3 μM protein and an increasing amount of unlabelled specific or nonspecific DNA from 1 μM to 5 μM.

10.4.4. Plasmid DNA cleavage experiments

The reaction mixture contained 32 nM DNA (with the ZF recognition sequence inserted into pGEX-6P-1 or the original pGEX-6P-1 - GE Healthcare vector), 0.85 μM protein, 2 μM $ZnSO_4$, 50 mM HEPES (pH = 7.9) and 40 mM NaCl. The reaction was incubated at 37 °C and 4 μl aliquots were used for gel electrophoresis at each time point. Before loading to the gel SDS was added to a final concentration 0.1 m/V%.

10.4.5. Primer extension experiments

A reaction mixture contained 10 ng of DNA digested with designed artificial ZFN for 46 hours, 2 mM dNTP, 100 fmol of [γ-^{32}P] labelled primer, standard Taq reaction buffer and 1.5 unit of Taq polymerase (New England BioLabs). The reaction was incubate at 98 °C for 1 minute, after that, the reaction was incubated at 98°C for 20 seconds, at 50°C for 30 seconds, at 72°C for 20 seconds, and this sequence was repeated 10 times. The reaction was terminated by phenol/chloroform extraction and ethanol precipitation. The synthesized fragment was separated on 8% polyacrylamide gel containing 8.3 M urea, and was visualized by Typhoon FLA 7000. The primer sequence was the following:

5'-primer, 5'-GGGCTGGCAAGCCACGTTTGGTG-3',
3'-primer, 5'-CCGGGAGCTGCATGTGTCAGAGG-3'.