NMR structure analysis and identification of the DNA binding site of the C-terminal domain of the *Bacillus subtilis* protein DnaD

Submitted for the degree of Doctor of Philosophy

Department of Molecular Biology
and Biotechnology

Farhat Yasmeen Marston

September 2010

Abstract

The *B. subtilis* protein DnaD is an essential component of the chromosome replication machinery and a global regulator of DNA architecture, as it exhibits a unique DNA remodeling activity that opens up supercoiled DNA by forming higher order nucleoprotein structures. There are no homologues of this protein in gram negative bacteria and details of its molecular structure are at present limited to the crystal structures of the N-terminal domain from *B. subtilis* and *G. Kaustophilus*, and also crystal structures of the C-terminal domain of two proteins with unknown function from *S. mutans* and *E. faecalis* (Structural Genomics projects pdb codes 2zc2 and 2i5u respectively). In this thesis, the determination of the NMR structure for the DNA binding C-terminal domain of the *B. subtilis* DnaD protein is reported. This domain is composed of five helices and an unstructured C-terminal tail, helix I-IV form a well packed hydrophobic core and helix V, which is more extensive than assumed from sequence alignments, extends away from this core fold of the protein. NMR DNA titrations, *in vitro* mutagenesis and i*n vivo* complementation experiments show that a highly conserved YxxxIxxxW motif, the final helix and a portion of the mobile C-terminal region make up the DNA binding module of *B. subtilis* DnaD and are all essential for cell viability. Sequence alignments of DnaD alone fail to identify two key elements of this DNA binding module, the poorly conserved sequences of the final helix and the C-terminal mobile segment. Sensitive Hidden Markov Models based techniques, show that the two structural domains found in DnaD are present in *B. subtilis* DnaB, a primosomal protein that also participates in replication initiation. Furthermore, these two proteins share the only strong sequence motif, the highly conserved YxxxIxxxW sequence that contributes to DNA binding.

Acknowledgements

First and foremost, I would like to express my deep appreciation and sincere gratitude to my supervisor Dr Jeremy Craven for his patience, invaluable guidance and support throughout my PhD. Jeremy provided the sequence analysis data and figures presented in chapter 7 of this thesis. I would also like to show my gratitude to Professor Panos Soultanas, for his genuine and endless enthusiasm towards the project, and for all the excellent discussions and practical advice he provided over the last few years. This project would not have been possible without his collaboration. Thanks also go to the BBSRC for funding this PhD.

I would also like to acknowledge: William Grainger for providing the very first protein sample and performing the gel shift assays experiments; Nicholas Hopcroft for providing two of the truncation constructs; and Alan Grossman for the *in vivo* data.

Many thanks also go to my advisors, Professor Jon Waltho and Dr Pat Baker, for their many helpful discussions and suggestions. Andrea Hounslow also provided invaluable help with aspects of NMR instrumentation, and many other members of the Sheffield NMR group made this PhD an enjoyable experience.

My special appreciation goes to my mother and late father, because without their love and guidance I would never have been able to get this far. Many thanks go to my brothers, sisters, bother-in-law and sister-in-laws, who have all been very supportive and fantastic at taking my mind off work to help me relax. Finally, and most importantly, I would like to thank my husband, James who has shared all of the ups and downs of this thesis and life in general.

Contents

Chapter 1: Introduction 1

1.1 General Introduction 1

1.2 Bacterial DNA replication 1

1.3 Replication Initiation 4

1.3.1 Regulation of the Initiation of DNA replication 4

1.3.2 The primosomal cascade at the *oriC* 6

1.3.3 Reconstitution of the replication folks 7

1.4 The *B. subtilis* primosomal protein DnaD 8

1.4.1 DnaD-mediated DNA remodeling 8

1.4.2 The DNA remodeling activity of DnaD opposes that of DnaB 9

1.4.2.1 The DNA remodeling effect of DnaD and DnaB is analogous to *E. coli* HU and H-NS proteins 10

1.4.3 DnaD is the link between DNA remodeling and initiation of DNA replication 10

1.4.4 DnaD consists of two domains with distinct activities 11

1.4.5 Structure of DnaD-Nd 12

1.4.6 Speculative model for scaffold formation 13

1.5 Project aim 14

1.6 Preliminary sequence analysis of DnaD-Cd 15

1.6.1 Homologous proteins and secondary structure predictions 15

Chapter 2: Materials and Methods 17

2.1 Growth Media 17

2.1.1 LB Media 17

2.1.2 LB agar 17

2.1.3 ^{15}N Minimal Media 17

2.1.4 ^{15}N-^{13}C Media 18

2.2 Antibiotic selection 18

2.3 Production of competent cells 18

2.4 Transformation 18
2.5 Plasmid amplification and purification 19
2.6 DNA cleanup 19
2.7 PCR 19
2.8 DNA sequencing 20
2.9 Agarose gel electrophoresis 20
2.10 Cloning of the DnaD-Cd215 and DnaD-Cd206 gene 21
2.11 SDS gel electrophoresis 22
2.12 Concentrating proteins 23
2.13 Buffer Exchanging 23
2.14 Protein quantification and molecular weights 23
2.15 Protein expression and purification 24
2.15.1 Protein extraction and purification buffers 24
2.15.2 Bacterial growth and protein expression 25
2.15.3 Cell disruption and protein extraction 26
2.15.4 Unfolded IMAC 26
2.15.5 Protein refolding 26
2.15.6 Folded IMAC-I 27
2.15.7 N-terminal His-tag removal 27
2.15.8 Folded IMAC-II 27
2.15.9 Gel filtration chromatography 27
2.16 CD spectroscopy 28
2.17 NMR sample preparation 28

Chapter 3: Overview of NMR Methods 29
3.1 NMR spectra 29
3.2 Assignment and structure determination 29
3.2.1 Backbone resonance assignment 29
3.2.2 Asstool: Program for backbone assignment 30
3.3 Side-chain assignment 31
3.4 NOE assignment 31

3.5 ARIA 31
3.6 CNS 32
3.7 ^{15}N Relaxation 32
3.8 DNA titrations 33
3.9 Hydrogen/Deuterium Exchange 34

Chapter 4: Optimisation of protein production and NMR sample conditions 35
4.1 A suitable candidate for structure determination 36
4.2 Increasing the quantity of protein produced 38
4.2.1 Optimisation of protein expression 38
4.2.2 Establishing a solubilisation and refolding protocol 41
4.2.3 Optimisation of protein purification 43
4.2.3.1 Removal of the N-terminal His-tag 43
4.2.3.2 Optimisation of thrombin treatment 44
4.2.3.3 Problems encountered with ammonium sulphate precipitation 45
4.2.3.4 Alternative methods to concentrate and buffer exchange proteins 46
4.2.3.5 Improving solubility of DnaD-Cd 48
4.2.4 Summary 51
4.3 Restoration of inconsistent HSQC spectra 51
4.4 Optimisation of NMR sample conditions 52
4.5 Stability of DnaD-Cd over time 54

Chapter 5: Resonance assignment and preliminary NMR structure determination of DnaD-Cd 56
5.1 Resonance assignment 57
5.1.1 Backbone assignment 57
5.1.2 Side-chain assignment 58
5.2 Preliminary NMR structure determination 59
5.2.1 Dihedral restraints 59

5.2.2 NOESY assignment 59
5.2.3 Structure calculations 60
5.3 Description of the preliminary ensemble of NMR structures for DnaD-Cd 60
5.4 Homology model for DnaD-Cd based on 2zc2 61
5.5 A potential DNA binding site 64
5.5.1 Key residues involved in the packing of DnaD-Cd are conserved 64
5.5.2 Charge distribution of DnaD-Cd 65

Chapter 6: Interaction of DnaD-Cd with DNA **66**
6.1 DNA binding of DnaD-Cd monitored by NMR 66
6.2 Elucidating the DnaD-Cd-DNA exchange regime from changes in DNA titration spectra 69
6.2.1 Protein DNA exchange regimes 69
6.2.2 DnaD-Cd-DNA binding may be under an intermediate exchange regime 69
6.3 Are regions of DnaD-Cd identified as the DNA binding site truly involved in DNA binding? 71
6.3.1 A conserved YxxxIxxxW in helix IV affects DNA binding 71
6.3.2 Helix IV alone is not sufficient for DNA binding 71
6.3.3 Helix V and part of the unstructured C-terminal region are involved in DNA binding 72
6.4 DNA-induced oligomerisation: Reconciling NMR and gel shift data 73
6.5 Helix V and the unstructured region of DnaD is essential for *B. subtilis* viability 74

Chapter 7: Domain similarity between DnaD and DnaB **76**
7.1 Merger of the DnaD-like family with the DnaB_2 family in the Pfam database 76
7.2 DnaB contains domains homologous to the DnaD-Nd and DnaD-Cd 77

7.3 DnaB contains a central DDBH2 domain, giving a domain arrangement of (DDBH1)-(DDBH2)-(DDBH2) 79
7.4 The DnaD/DnaB superfamily 79
7.5 Discussion 80

Chapter 8: Discussion and future work **83**
8.1 Discussion 83
8.2 Future work 86
References **89**

List of figures and tables

Figure 1.1:	A schematic representation of the initiation of DNA replication	Follows page	2
Figure 1.2:	A schematic representation of bi-directional replication of the bacterial chromosome	Follows page	3
Figure 1.3:	Terminator sequences in *E. coli*	Follows page	3
Figure 1.4	Schematic representation of the *oriC* region from four bacterial chromosomes	Follows page	4
Table 1.0:	The nomenclature of key proteins involved in DNA replication in *E. coli* and *B. subtilis*	Follows page	6
Figure 1.5:	Schematic model for the assembly of the *B. subtilis* PriA-dependent primosome	Follows page	8
Figure 1.6:	DnaD forms concentration dependent oligomers and nucleoprotein complexes	Follows page	8
Figure 1.7:	AFM images of plasmid DNA in the presence of DnaB alone or DnaB and DnaD together	Follows page	9
Figure 1.8:	Functional model presenting DnaD as a potential link between nucleoid remodeling and initiation of DNA replication in *B. subtilis*	Follows page	10
Figure 1.9:	The N-terminal and C-terminal domains of DnaD have distinct activities	Follows page	11
Figure 1.10:	Overview of the oligomeric state of DnaD and its two domains	Follows page	12
Figure 1.11:	Ribbon representations of the X-ray crystallography structure of DnaD-Nd	Follows page	12
Figure 1.12:	A lattice-based model for DnaD-Nd scaffold formation.	Follows page	13
Figure 1.13:	Model for DnaD-mediated scaffold formation	Follows page	13

Figure 1.14:	Ribbon representations of proteins homologous to DnaD-Cd	Follows page	15
Figure 1.15:	Secondary structure predictions from PHYRE for DnaD, 2i5u and 2zc2	Follows page	15
Figure 2.0:	Primers designed for the amplification of DnaD C-terminal domain truncations		21
Table 2.0:	Protein Parameters		24
Table 3.0:	Experimental parameters for the main NMR spectra used to assign DnaD-Cd	Follows page	30
Figure 4.0:	The first ^{1}H-^{15}N HSQC spectrum of DnaD-Cd acquired	Follows page	36
Figure 4.1:	Protein sequence for DnaD-Cd used in NMR experiments	Follows page	36
Figure 4.2:	T2 ^{15}N relaxation data for DnaD-Cd	Follows page	37
Figure 4.3:	Flow chart summarising expression and purification of DnaD-Cd before optimisation of the protocol	Follows page	38
Figure 4.4:	SDS PAGE gels showing expression of His-DnaD-Cd at 25°C and 37°C	Follows page	40
Figure 4.5:	GdHCl denaturation profile of DnaD-Cd	Follows page	41
Figure 4.6:	Comparison of HSQC spectra of protein extracted from the soluble and insoluble fraction	Follows page	43
Figure 4.7:	Suitability of the His-tag	Follows page	44
Figure 4.8:	Optimising thrombin cleavage of the His-tag	Follows page	45
Figure 4.9:	Flow chart summerising the optimized protein expression and purification protocol	Follows page	51
Figure 4.10:	Peaks from ^{1}H-^{15}N HSQC spectra restored on the addition of DTT	Follows page	52
Figure 4.11:	Optimization of NaCl concentrations	Follows page	53
Figure 4.12:	Optimisation of protein concentration	Follows page	53

Figure 4.13:	Stability of DnaD-Cd	Follows page	54
Figure 4.14:	Mass spectrum of ^{13}C-^{15}N DnaD-Cd recorded under denaturing conditions	Follows page	55
Figure 5.0:	^{1}H-^{15}N HSQC backbone assignment spectra of DnaD-Cd	Follows page	57
Figure 5.1:	An example of slices from the ^{15}N HSQC spectrum and the five 3D spectra used for resonance assignment	Follows page	57
Figure 5.2:	TALOS plots for phi and psi per residue	Follows page	59
Figure 5.3:	Example of an NOE assignment strip	Follows page	59
Figure 5.4:	Map of NOE distance restraints	Follows page	60
Figure 5.5:	Monitoring quality of the structure	Follows page	60
Figure 5.6:	Preliminary structures of DnaD-Cd	Follows page	60
Table 5.0:	Structural Statistics for the final ensemble of DnaD-Cd	Follows page	60
Figure 5.7:	Comparison of DnaD-Cd with 2zc2 from *Streptococcus mutans*	Follows page	61
Figure 5.8:	Building the final helix on the homology model of DnaD-Cd	Follows page	62
Figure 5.9:	ClustalW alignment of the sequences defined as DnaD from Figure 7.5	Follows page	62
Figure 5.10:	Defining the extent of helix V in DnaD-Cd	Follows page	62
Figure 5.11:	Hydrogen/Deuterium amide exchange plots of DnaD-Cd	Follows page	64
Figure 5.12:	The conserved YxxxIxxxW motif mapped onto a structural model of DnaD-Cd	Follows page	64
Figure 5.13:	Distribution of charge across DnaD-Cd	Follows page	65
Figure 6.0:	Gel shift assays showing DnaD-Cd forms large nucleoprotein complexes	Follows page	67
Figure 6.1:	DNA binding of DnaD-Cd monitored by NMR	Follows page	67

Figure 6.2:	Intensity changes on titrating 10 mer DNA into DnaD-Cd	Follows page	67
Figure 6.3:	NMR titration curves for the titration of 19 mer ssDNA into 0.2 mM DnaD-Cd	Follows page	67
Figure 6.4:	NMR titration curves for the titration of 10 mer ssDNA into 0.3 mM DnaD-Cd	Follows page	67
Figure 6.5:	Minor peaks show weaker DNA binding	Follows page	68
Figure 6.6:	Protein-DNA exchange regimes	Follows page	68
Figure 6.7:	Line shapes simulated on the intermediate timescale in order to reproduce DnaD-Cd titration curves that show a big drop in intensity	Follows page	70
Table 6.0:	Parameters set for each simulation	Follows page	70
Figure 6.8	The DnaD mutant proteins, Y180A and W188A are defective in DNA binding compared to wild type DnaD	Follows page	71
Figure 6.9:	Sequence of DnaD-Cd and its truncation mutants	Follows page	72
Figure 6.10:	^{15}N HSQC spectrum of DnaD-Cd196	Follows page	72
Figure 6.11:	Chemical shift perturbations in ^{15}N HSQC spectra of DnaD-Cd truncation mutants	Follows page	72
Figure 6.12:	DnaD-Cd196 is deficient in DNA binding	Follows page	72
Figure 6.13:	NMR titration curves for the titration of 10 mer ssDNA into 0.3 mM DnaD-Cd196	Follows page	72
Figure 6.14:	^{15}N HSQC spectrum of DnaD-Cd206	Follows page	72
Figure 6.15:	^{15}N HSQC spectrum of DnaD-Cd215	Follows page	73
Figure 6.16:	^{15}N HSQC titration spectra of DnaD-Cd206 and DnaD-Cd215 with 10mer DNA	Follows page	73
Figure 6.17:	NMR titration curves for the titration of 10 mer ssDNA into 0.2 mM DnaD-Cd206	Follows page	73
Figure 6.18:	Intensity changes on titrating 10 mer DNA	Follows page	73

	into DnaD-Cd215		
Figure 6.19:	DnaD-Cd215 binds DNA but DNA DnaD-Cd206 does not	Follows page	73
Figure 6.20:	Chemical shift changes during titration of 10 mer ssDNA into DnaD-Cd215	Follows page	73
Figure 6.21:	*In vivo* activity of DnaD and its mutants	Follows page	75
Figure 7.0:	Examples of bacteria and phage that have proteins identified by Pfam to have a 'DnaD-like protein motif'	Follows page	76
Figure 7.1:	HMM relationships of DnaD and DnaB	Follows page	78
Figure 7.2:	Domain similarity between DnaD and DnaB	Follows page	78
Figure 7.3:	Secondary structure and disorder predictions for DnaD and DnaB	Follows page	78
Figure 7.4:	Selecting putative DnaB or DnaD sequences	Follows page	80
Figure 7.5:	ClustalW alignment of the sequences defined as DnaB from Figure 7.4	Follows page	80

Abbreviations

AFM	Atomic force microscopy
ARIA	Ambiguous Restraints for Iterative Assignment
ATP	Adenosine-5'-triphosphate
BLAST	Basic Local Alignment Search Tool
B. subtilis	*Bacillus subtilis*
CD	Circular Dichroism (spectroscopy)
CNS	Crystallography and NMR Systems
CPMG	Carr–Purcell–Meiboom–Gill
Da	Dalton (molecular mass unit)
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic acid
DnaD-Cd	C-terminal domain of DnaD
DnaD-Cd206	C-terminal domain of DnaD from residue 129-206
DnaD-Cd215	C-terminal domain of DnaD from residue 129-215
DnaD-Nd	N-terminal domain of DnaD
dNTP	Deoxyribonucleotide mixture
ds	Double stranded
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EM	Electron microscopy
GdHCl	Guanidine hydrochloride
HPLC	High performance liquid chromatography
HSQC	Heteronuclear single quantum coherence
IMAC	Immobilized metal affinity chromotography
INEPT	Insensitive nuclei enhanced by polarisation transfer
IPTG	Isopropyl-β-D-thiogalactopyranoside
Kan^R	kanamycin-resistance
LB	Luria Bertani (media)
H-NS	Heat-stable nucleoid-structuring protein

HU	Histone-like protein
MS	Mass spectroscopy
MW	Molecular weight
NMR	Nuclear magnetic resonance (spectroscopy)
NOE	Nuclear Overhauser effect
ODx	Optical density at x nm
oriC	Origin of replication
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PHYRE	Protein homology/analogy recognition engine
PMSF	Phenylmethylsulfonyl fluoride
ppm	Parts per million
RNA	Ribonucleic acid
rpm	revolutions per minute
SDS	Sodium dodecyl sulphate
ss	Single stranded
T2	Transverse relaxation time
TOCSY	Total Correlation Spectroscopy

Chapter 1: Introduction

1.1 General Introduction

The DNA molecules in bacteria cells that encode genes essential for survival are termed chromosomes. Bacterial chromosomes are tightly packed into a structure called the nucleoid (Robinow and Kellenberger, 1994) by the introduction of supercoils (twisting of the circular double stranded helix around itself) and the binding of histone-like DNA remodeling proteins, because their length can exceed that of the cell by a factor of a 100 and more (Kim *et al.*, 2004; Lewis, 2004). In growing cells, the chromosome must be released from this supercoiled form into a relaxed state for access for crucial biological processes of transcription, regulation and replication. Such architectural changes of the chromosome are promoted by variations in the composition of DNA binding proteins in the bacterial proteome (Azam *et al.*, 1999; Schneider and Muskhelisvili, 1997), during cell growth. Studies on DNA binding proteins and their roles in DNA replication of bacterial systems have been mainly directed to the gram negative *E. coli*, but to a much lesser extent the gram positive *B. subtilis*, therefore our knowledge of such proteins of *B. subtilis* is limited (Marians, 2000; Lemon *et al.*, 2002). Among the growing list of proteins identified in *B. subtilis*, the DNA binding protein DnaD has been suggested to play a critical role in replication initiation and chromosomal remodeling (Turner *et al.*, 2004). However, the exact mechanism by which this protein functions still remains unclear.

The *B. subtilis* protein DnaD, in particular the C-terminal domain of DnaD, is the focus of this thesis. Therefore to put the work into perspective this chapter reviews the current literature regarding DnaD and outlines the aim of the project. Preliminary sequence analysis done towards the project goal is also presented in this chapter.

1.2 Bacterial DNA replication - an overview

DNA replication is a semi-conservative process whereby both strands of the parent DNA function as templates to produce two identical daughter molecules, each containing one old

and one newly synthesized strand. In bacterial cells, this process proceeds through a conserved general mechanism once every cell division cycle in response to cellular signals that are coupled to cell mass (Boye *et al.*, 1996) . The overall steps involved can be divided into three main stages known as initiation, elongation and termination.

Initiation

Replication of the bacterial chromosome initiates at a unique site called the origin of chromosomal replication (*oriC*). The *oriC* contains sequences with a higher than average AT base pair frequency known as unwinding elements and multiple highly conserved sequences, termed DnaA boxes, which are binding sites for the major replication initiator protein DnaA (Figure 1.1a). DnaA boxes show slight variation in their nucleotide sequence. Some bind DnaA with high affinity, others engage in weaker binding and require protein oligomerisation (Scharper and Messer, 1995). DnaA is an ATPase, whose nucleotide status influences interactions with the binding sites (McGarry *et al.*, 2004). At the start of initiation, DnaA binds the high affinity DnaA boxes regardless of its nucleotide state and causes DnaA to localise at the origin. Then ATP-bound DnaA interacts cooperatively with weaker affinity sites (McGarry *et al.*, 2004) and induces local unwinding of the AT rich region as the DnaA homo-oligomerise wrapping the DNA around and destabilizing the helix (Figure 1.1b). This opening of dsDNA triggers a series of events called the primosomal cascade that lead to the assembly of a multi-enzyme complex termed the primosome, which unwinds the DNA for use as a template during replication as it migrates along the DNA. The primosome also synthesizes primers necessary for initiating replication. In the early step of the cascade, DNA unwinding proteins called helicases are recruited (Figure 1.1c) to form two pre-priming complexes, one at each end of the open site of the *oriC*. As the helicases unwind DNA in both directions from the *oriC*, single strand binding protein (SSB) coats single stranded regions as they arise (Figure 1.1d) to prevent reformation of dsDNA and to hold the ssDNA in a restricted conformation. Then the RNA synthesizing proteins called primases are recruited (Figure 1.1f) and short RNA primers synthesized using the exposed base sequence as template. The addition of primase to the pre-priming complex completes the formation of the primosome.

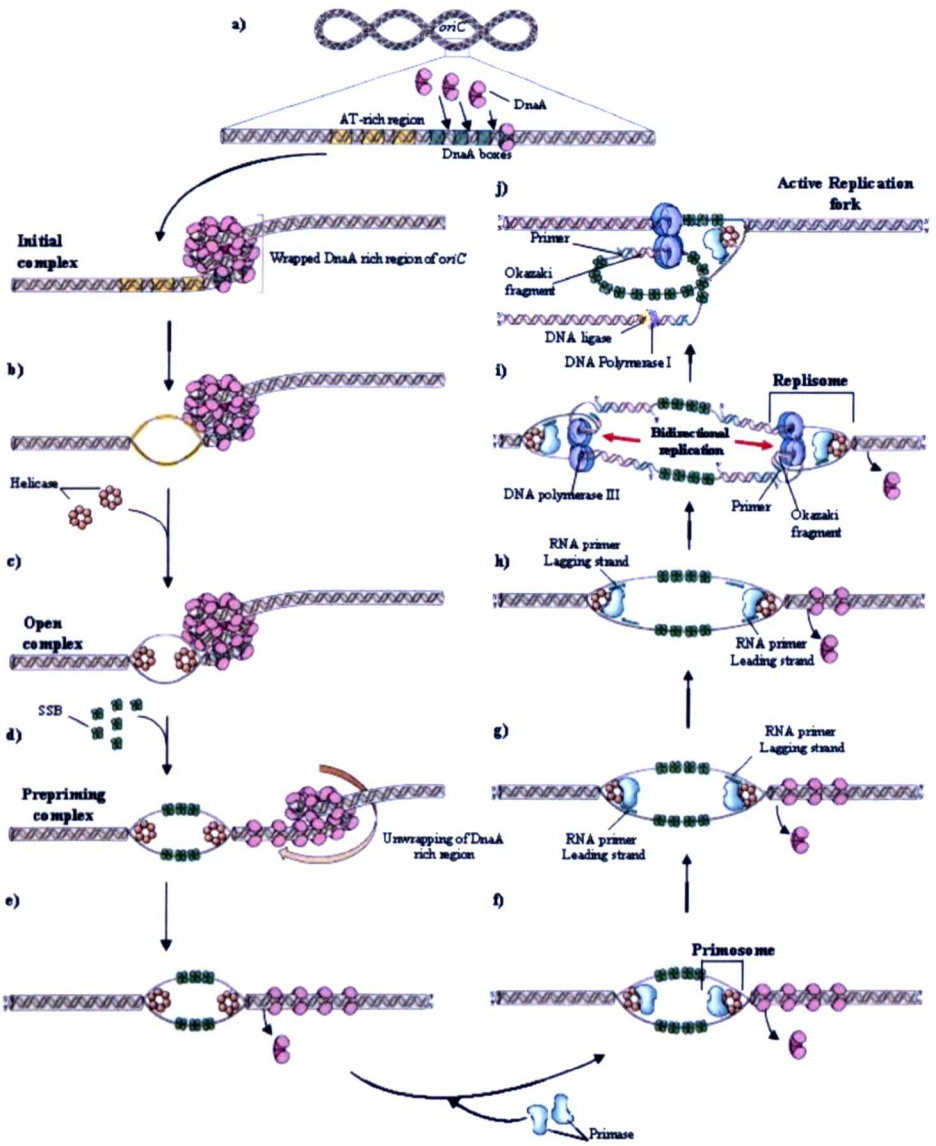


Figure 1.1: A schematic representation of the initiation of DNA replication. (a) DnaA binds co-operatively to DnaA boxes at the *oriC*. (b) The DNA duplex unwinds locally. (c) The helicase is recruited onto the single stranded DNA at each replication fork site. (d) Helicase activity unwinds DNA in both directions from the *oriC* and single stranded binding proteins (SSB) bind to single stranded DNA, keeping strands separated. (e) DnaA is displaced in the process. (f) Primases are recruited, completing the assembly of the primosomes. (g and h) RNA primers are synthesized for the leading and lagging strands. (i) Polymerase III enzymes are recruited to complete the replisomes and then each replication forks moves bi-directionally from the *oriC* replicating the entire chromosome. (j) DNA polymerase I and DNA ligase join Okazaki fragments into a continuous strand (adapted from www.virginia.edu/Heidi/chapter30/chp30.html).

Elongation

At the beginning of elongation, two DNA synthesizing enzymes called DNA polymerase are loaded alongside the primosome to form a replisome (Figure 1.1i) on the template DNA at what is termed the replication fork. DNA polymerases are multi-subunit complexes that synthesize new DNA only in a 5' to 3' direction from an existing primer. Since DNA is anti-parallel, one strand runs in a 5' to 3' direction and the other in a 3' to 5', therefore the two complementary daughter strands have to be synthesized in different ways so that replication is semi-conservative. As the primosome unwinds DNA ahead of the replication fork, DNA polymerase III moves along the 3 to 5' strand (leading strand) and synthesizes new DNA continuously from the primer synthesized when replication was initiated. The other parent strand that runs in a 5' to 3' direction (lagging strand) is synthesized from short primers in a series of fragments of 100 to 1000 nucleotides called Okazaki fragments (Figure 1.1j). This strand is looped around at the replication fork so the DNA polymerase III can move in a 5' to 3' direction relative to the strand when a sufficient stretch of the template sequence has been exposed. Downstream of this lagging strand, DNA polymerase I excises the RNA primer and replaces it with DNA. Then DNA ligase seals the remaining nick to form a continuous strand of DNA. Two replication forks proceed simultaneously in opposite directions from the *oriC* to replicate the entire circular bacteria chromosome. A larger view of the replication process is shown in Figure 1.2.

Termination

Termination of replication occurs when the bi-directionally moving replication forks meet one another at the termination region (*ter*). This termination region contains a cluster of sites called *ter* sequences that are arranged in two sets and are inversely oriented with respect to each other (Figure 1.3a). These sequences are recognition sites for sequence specific proteins that are called terminus utilization substance (Tus) in *E. coli* and replication terminator proteins (RTP) in *B. subtilis* (Hill, 1992). When these proteins bind to the *ter* sites on the double helix, the passage of the helicase approaching in one direction is blocked (Figure 1.3b), but the helicase coming from the other direction is allowed to pass by (Hill, 1992; Kamada *et al.*, 1996). The replication fork

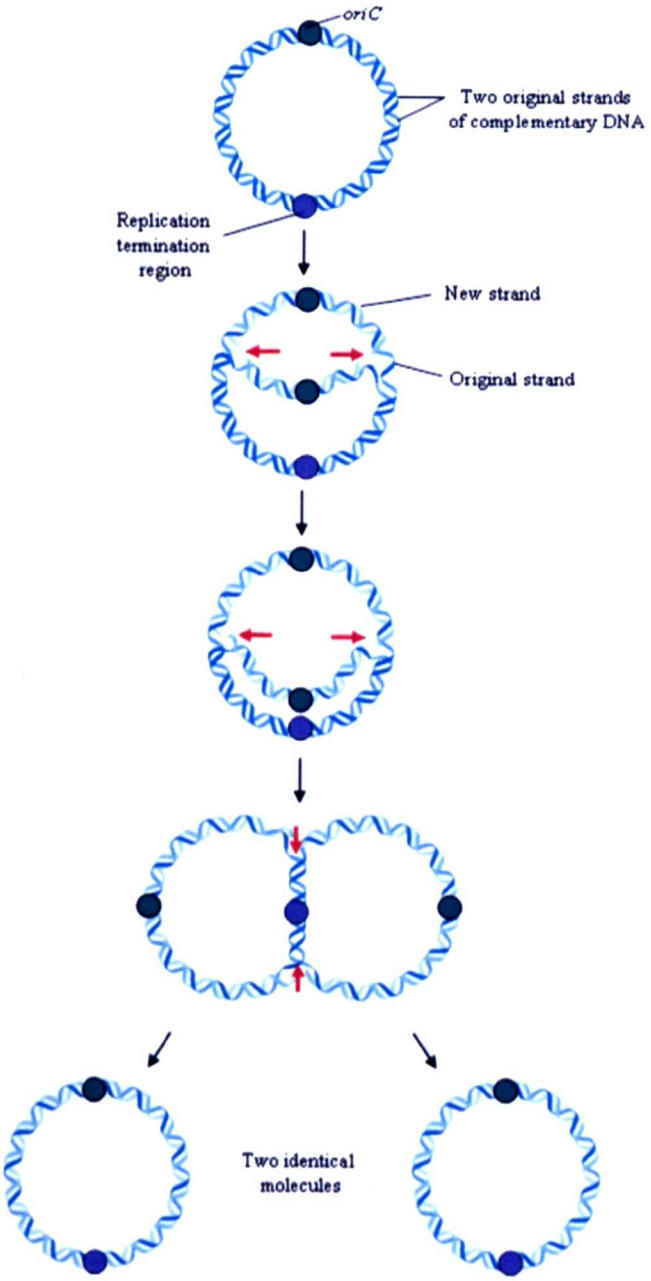


Figure 1.2: A schematic representation of bi-directional replication of the bacterial chromosome. Replication initiates at the *oriC*. Once the *oriC* has replicated, two replication forks proceed in opposite directions (red arrows) until they meet and stop at the termination region that is located half way around the chromosome. Two identical daughter molecules form, each composed of one original and one newly synthesised strand (adapted from Griffiths *et al.*, 1999).

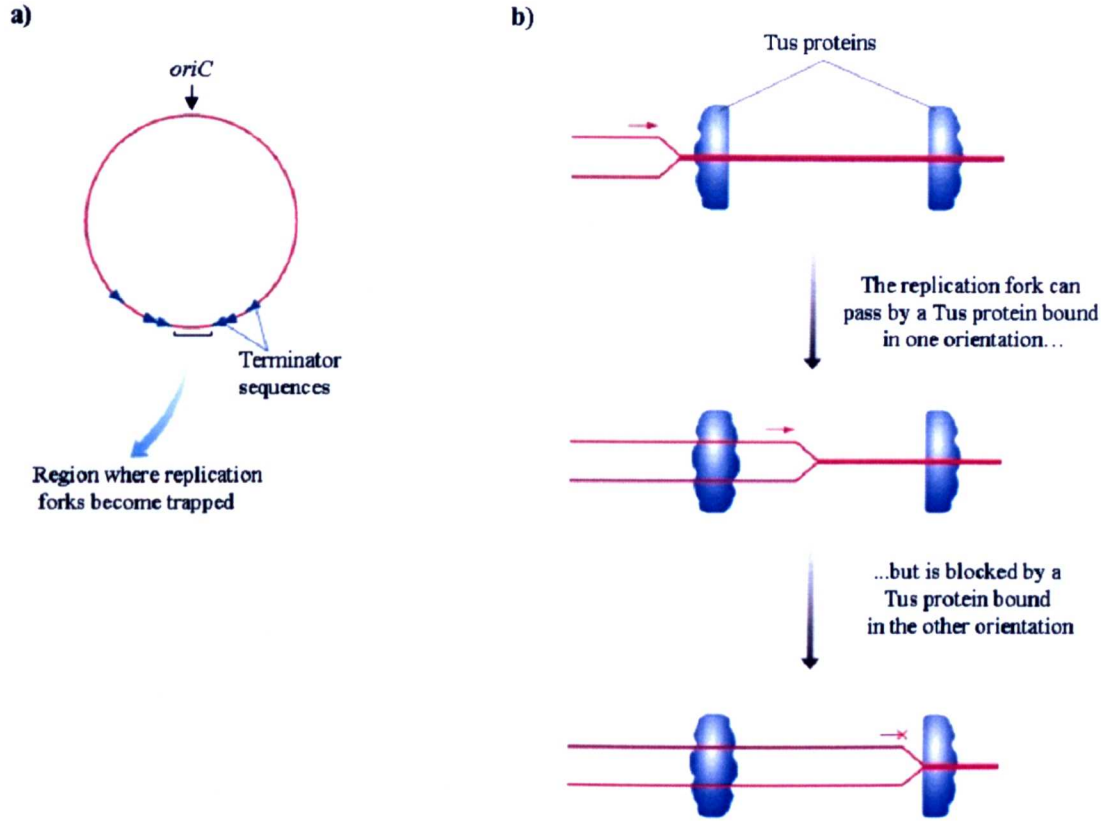


Figure 1.3: Terminator sequences in *E. coli*. (a) The six terminator sequences identified in the *E. coli* genome are shown and the direction in which the two bi-directionally moving replication forks can pass by these sequences is indicated by the blue arrow heads. (b) The replication fork passes by the left hand Tus protein, but stops when it reaches the 'impenetrable wall of β-strands' (shown as a flat surface) on the second Tus protein that faces the fork (Brown, 2006).

that penetrates the *ter* bound proteins is determined by the orientation in which the proteins bind to the *ter* sites, as the helicase that moves the fork forward is able to disrupt the structure on one side of the protein but not the other (Kamada *et al.*, 1996). Replication terminates in a short region opposite the *oriC* where both replication forks become trapped (Hill, 1992). The replisomes then disassemble and finally the two interlinked daughter molecules separate.

1.3 Replication initiation

Initiation of chromosomal replication is the key step in regulating DNA synthesis. In *B. subtilis* the general scheme of initiation presents many similarities to *E. coli* since proteins beginning (DnaA) and ending (helicases and primases) the process are conserved (Marsin *et al.*, 2001). However, several genes encoding initiator proteins homologous to *E. coli* are absent in *B. subtilis*. Several additional proteins involved in the primosomal cascade have been identified that are not present in *E. coli* showing that some disparity in the initiation strategy exists between these two bacteria.

1.3.1 Regulation of the initiation of DNA replication

A diverse range of regulatory mechanisms are used by bacteria to ensure accurate onset and timely completion of initiation of chromosomal replication. In many cases, these mechanisms are intrinsically connected to two key elements, DnaA and the *oriC* region. Throughout the bacterial kingdom, DnaA and its binding sites are well conserved (Mott and Berger, 2007). The basic structure of the *oriC* is also conserved; however the overall length, number, location and arrangement of DnaA boxes show considerable diversity (Figure 1.4) between species (Moriya *et al.*, 1994; Messer, 2002). For instance, in gram positive bacteria, DnaA boxes exist in two non-coding regions upstream and downstream of the *dnaA* gene at the *oriC* and replication is initiated only within the downstream box (Moriya *et al.*, 1992). In gram negative bacteria like *E. coli* there are no DnaA box clusters downstream of the *dnaA* gene. *B. subtilis* is slightly different to most gram positive bacteria in that two DnaA box regions flank the *dnaA* gene and are both required simultaneously for the initiation of replication (Moriya *et al.*, 1994). The basis

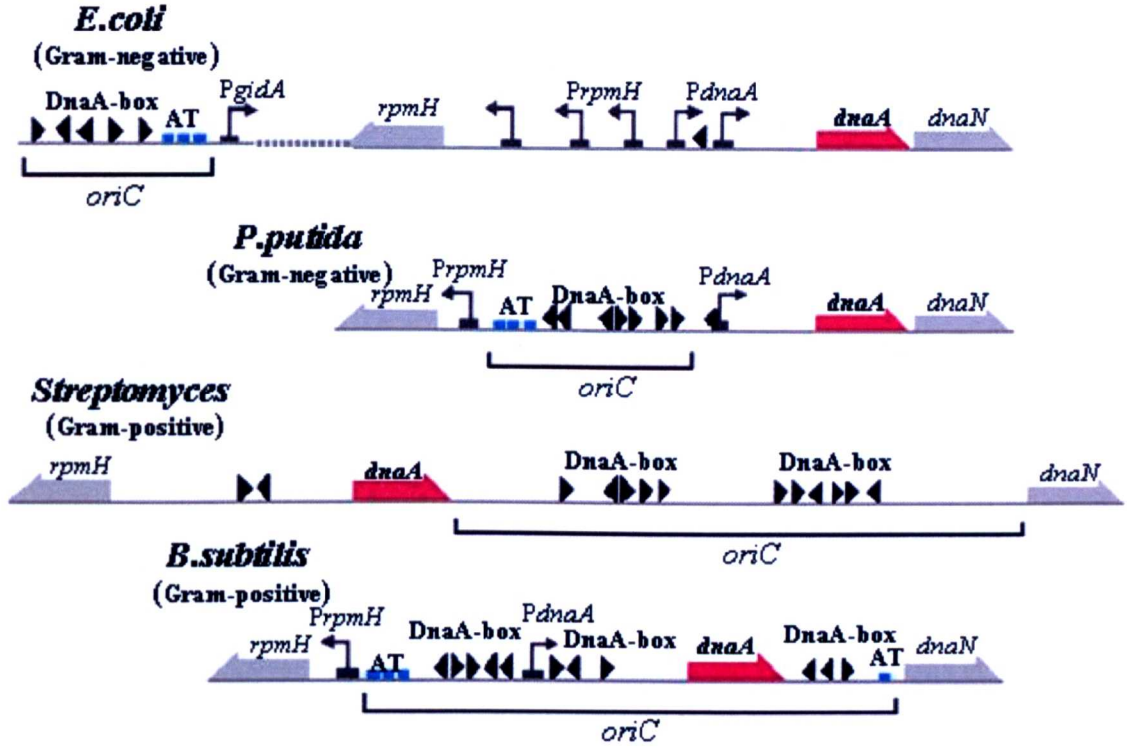


Figure 1.4: Schematic representation of the *oriC* region from four bacterial chromosomes. Black bent lines under the structure highlight the minimum region identified as the *oriC*. DnaA genes are shown in red and other genes in grey. Black arrow heads indicate the position and orientation of DnaA boxes. Blue rectangles show the location of the AT rich unwinding elements and the bent arrows highlight gene promoters within the region shown (adapted from Moriya *et al.*, 1994).

for this origin divergence is unknown, however, since the basic system for replication initiation is the same, many regulatory mechanisms identified in *E. coli* may exist in other bacteria.

Several positive and negative regulatory factors participate in regulating initiation of DNA replication in *E. coli*. The level of the ATP-bound form of DnaA (the active form) is a positive factor, as increased levels trigger local unwinding of the *oriC* (Kurokawa *et al.*, 1999). DnaA regulates the expression of its own genes as well as other genes involved in replication initiation by binding to DnaA boxes located within the promoter region (Messer and Weigel, 1997). Three key negative regulatory processes in *E. coli* inhibit extra initiation to ensure the chromosome is replicated only once per cell cycle.

One process called RIDA (Regulatory inactivation of DnaA) inactivates DnaA by converting the active initiator, DnaA-ATP, to an inactive form, DnaA-ADP, using the β subunit (DnaN) of DNA polymerase III and a protein called Had (Katayama *et al.*, 1998; Kato and Katayama, 2001; Riber *et al.*, 2006). This process has been suggested to be the major way by which *E. coli* prevents over initiation of DNA replication. Another mechanism (named datA, DnaA titration), prevents re-initiation of replication by reducing the levels of DnaA shortly after the *oriC* region has duplicated (Kitagawa *et al.*, 1998). It involves removal of the free form of the DnaA protein by a cluster of high affinity DnaA boxes within a region of the *oriC* called datA. The final regulatory process sequesters the newly replicated *oriC* to the cell membrane using a protein called seq, while reducing levels of active DnaA (Campbell & Kleckner, 1990; Lu *et al.*,1994; Ogden *et al.*, 1988).

ATP binding and hydrolysis is one regulatory mechanism likely to be common amongst bacteria that have a dnaA gene, as DnaA proteins from all bacteria have been shown to possess ATPase activity (Zakrzewska-Czerwinska *et al.*, 2007). In silico analysis on chromosomes from various bacteria show that at least two additional high affinity clusters of DnaA boxes are present in the vicinity of the *oriC*, suggesting that removal of free DnaA by DnaA box clusters is likely to be another common regulatory mechanism

(Mackiewicz *et al.*, 2004). In recent years it has been brought to light that the sequestration of the *oriC* to the membrane is not a mechanism exclusively characteristic of *E. coli*, many other bacteria have analogous systems (Kahng and Shapiro, 2001; Stephens *et al.*, 1996).

In *B. subtilis*, much less is known about regulatory mechanisms in comparison to *E. coli*. Nevertheless, it has been established that DnaA regulates expression of itself and its levels are important for determining the timing of initiation (Goranov *et al.*, 2005). As to negative regulatory mechanisms, the *E. coli* mechanism involving sequestration of the *oriC* by the protein seq is unlikely in *B. subtilis* since a protein homologous to seq does not exist in *B. subtilis* (Kunst *et al.*, 1997). However, another negative regulator called YabA has recently been found in *B. subtilis* (Hayashi *et al.*, 2005; Noirot-Gros *et al.*, 2002). The mechanism by which this protein prevents initiation has not yet been determined, but it is known that this regulator like others discussed interacts with DnaA (Noirot-Gros *et al.*, 2002; Noirot-Gros *et al.*, 2006). Overall, it seems that bacteria regulate the initiation of chromosomal replication by reducing the availability and/or activity of DnaA at the *oriC*.

1.3.2 The primosomal cascade at the *oriC*

In *E. coli*, the primosomal cascade loads the DnaB helicase into the unwound region of the *oriC*, as a stable hexameric complex with the primosomal protein DnaC (Learn *et al.*, 1997) by a direct interaction between the DnaA that is bound to DNA and DnaB in the DnaB:DnaC complex (Marszalek and Kaguni, 1994; Seitz *et al.*, 2000). It is important to note that the nomenclature of the Dna genes in *B. subtilis* differ from those in *E. coli* (Table 1.0). After loading, DnaC dissociates from the DnaB:DnaC complex by its own ATPase activity to activate the DnaB helicase (Baker and Bell, 1998). Then the DnaG primase is recruited via a helicase-primase interaction to complete the assembly of the primosome (Tougu *et al.*, 1994; Lu *et al.*, 1996; Hiasa and Marians, 1999). Finally, the primase synthesizes RNA primers required for the subsequent loading of the DNA polymerase III holoenzyme (Kelman and O'Donnell, 1995) and for initiating DNA synthesis (Bouche *et al*, 1975).

Proteins	Gene name	
	E. coli	*B. subtilis*
Initiator	*dnaA*	*dnaA*
Replicative DNA helicase	*dnaB*	*dnaC*
Helicase loader	*dnaC*	*dnaI*
Primosomal proteins		
	-	*dnaD*
	-	*dnaB*
	priA	*priA*
	priB	-
	priC	-
	dnaT	-
SSB	*ssb*	*ssb*
DNA polymerase III		
subunit α	-	*polC*
	dnaE	*dnaE*
subunit ε	*dnaQ*	-
subunit θ	*holE*	-
subunit β	*dnaN*	*dnaN*
subunit γ and τ	*dnaX*	*dnaX*
subunit δ	*holA*	-
subunit δ'	*holC*	-
subunit ψ	*holD*	-
Primase	*dnaG*	*dnaG*
DNA polymerase I	*polA*	*polA*
DNA ligase	*lig*	*yerG*

(-) Gene not present

Table 1.0: The nomenclature of key proteins involved in DNA replication in *E. coli* and *B. subtilis* (adapted from Sonenshein *et al.*, 2002).

In *B. subtilis* three other proteins, DnaB (not to be confused with the *E. coli* DnaB helicase), DnaD and DnaI are required for loading the replicative helicase DnaC, (homologue of *E. coli* DnaB) onto the *oriC* (Bruand *et al.*, 1995). DnaI is homologous to *E. coli* DnaC helicase loader (Soultanas, 2002). *B. subtilis* DnaD and DnaB have no homologues in *E. coli* but homologues of both these proteins, as well as DnaI, are found in gram-positive bacteria of low GC content (Lemon, *et al.*, 2002). The molecular events that underpin the function of DnaD, DnaB and DnaI in the primosomal cascade of *B. subtilis* are unclear but data from various experiments indicate that DnaI interacts with the helicase DnaC to form a stable DnaI–DnaC complex (Imai *et al.*, 2000; Soultanas, 2002; Velten *et al.*, 2003). DnaB forms a dual helicase loader with DnaI to load the helicase DnaC (Velten *et al.*, 2003), but also localizes at the membrane to provide a membrane attachment site for the *oriC* that is necessary for initiating chromosomal replication in *B. subtilis* (Hoshino *et al.*, 1987; Sueoka, 1998; Rokop *et al.*, 2004). The membrane attached DnaB has been suggested to regulate initiation of DNA replication by regulating the recruitment of DnaD to the membrane (Rokop *et al.*, 2004). The role of DnaD is more enigmatic. It has been implicated to act early in the cascade setting the stage for helicase recruitment, as it has been shown to interact with DnaA (Ishigo-Oka *et al.*, 2001) and also displays ssDNA and dsDNA binding activity (Marsin *et al.*, 2001). *In vitro,* it has been reported to disrupt the interaction between DnaI and DnaC to facilitate the loading of the helicase on to DNA, however it is still unknown if this is the case *In vivo* (Turner *et al.*, 2004).

1.3.3 Reconstitution of replication folks

The successful progression from initiation to elongation does not guarantee complete replication of the bacterial chromosome, as replication forks may stall or collapse at strand nicks, modified nucleotides, or other problem areas on the chromosomal DNA. Re-initiation or reassembly of stalled replication folks is often required for completing replication. The primosomal pathway discussed in section 1.3.2 initiates replication specifically at the *oriC* and is known as the DnaA primosomal pathway. At sites distinct to the *oriC*, bacteria utilize another primosomal pathway in which a protein designated

PriA coordinates the assembly of the restart primosome (also known as the PriA dependent primosome) and the replisome (Marians, 2000; Sandler and Marians, 2000; Polard *et al.*, 2002). PriA is a highly conserved protein that recognizes and binds to two distinct DNA structures. The first is a hairpin structure, known as the primosome assembly site (*pas*) and the second, a D-loop structure, which mimics products formed when damaged DNA is repaired by recombination at stalled replication fork (Marians, 1999). The DnaA and PriA pathways in *E. coli* share only one protein, the replicative helicase DnaC. In contrast to DnaA, PriA is assisted by three accessory proteins namely PriB, PriC and DnaT to load the replicative helicase as a DnaB-DnaC complex and to recruit the DnaG primase at the replication fork (Marians, 2000; Ng and Marians, 1996). The *B. subtilis* genome encodes no obvious homologues of the *E. coli* PriA partners, PriB, PriC and DnaT (Kunst *et al.*, 1997). Instead, DnaD, DnaB and DnaI proteins that are required for replication initiation at the o*riC* by DnaA also participate in replication fork reactivation in the PriA pathways (Bruand *et al.*, 2001; Marsin *et al.*, 2001). As for DnaA, DnaD interacts with PriA but shows preference for PriA-bound replication forks (Zhang *et al.*, 2005). A model proposed by Marsin *et al.*, 2001 for the assembly of the *B. subtilis* PriA-dependent primosome is presented in Figure 1.5.

1.4 The *B. subtilis* primosomal protein DnaD

1.4.1 DnaD-mediated DNA remodelling

DnaD is a 232 residue protein that exists predominately as a dimer (Turner *et al.*, 2004) but can form trimers, tetramers and large oligomers at high concentrations (Figure 1.6a). It has recently been discovered that DnaD exhibits a DNA remodelling activity that converts supercoiled plasmid into an open circular form (Turner *et al.*, 2004). The protein interacts non-specifically with plasmid DNA and oligomerises in a concentration dependent manner forming a large circular nucleoprotein scaffold of DnaD molecules within the plasmid boundary as shown in Figure 1.6b (Turner *et al.*, 2004). Gel shift assays have indicated that these nucleoprotein complexes are not a result of DnaD molecules binding along the DNA side by side (Turner *et al.*, 2004) but due to an initial DnaD-DNA complex acting as a seed, inducing the binding of more DnaD molecules

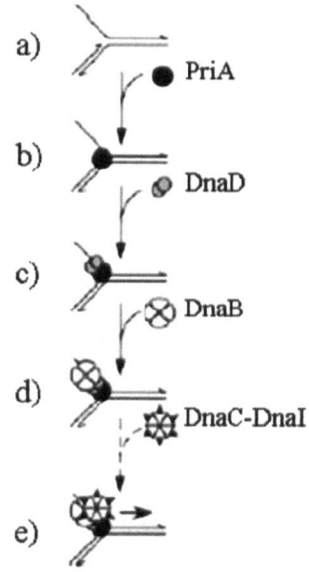


Figure 1.5: Schematic model for the assembly of the *B. subtilis* PriA-dependent primsome. (a) The expected product of recombination repair as well as stalled replication forks is a Y shaped structure that has a 5' ssDNA tail. (b) This structure is recognized and bound by PriA. (c) PriA then recruits DnaD onto the ssDNA. (d) DnaD stimulates the binding of DnaB (e) The DnaD-DnaB complex recruits the DnaC-DnaI complex onto the 5'ssDNA tail and replication restarts once DnaI dissociates, activating the helicase (adapted from Marsin *et al.*, 2001).

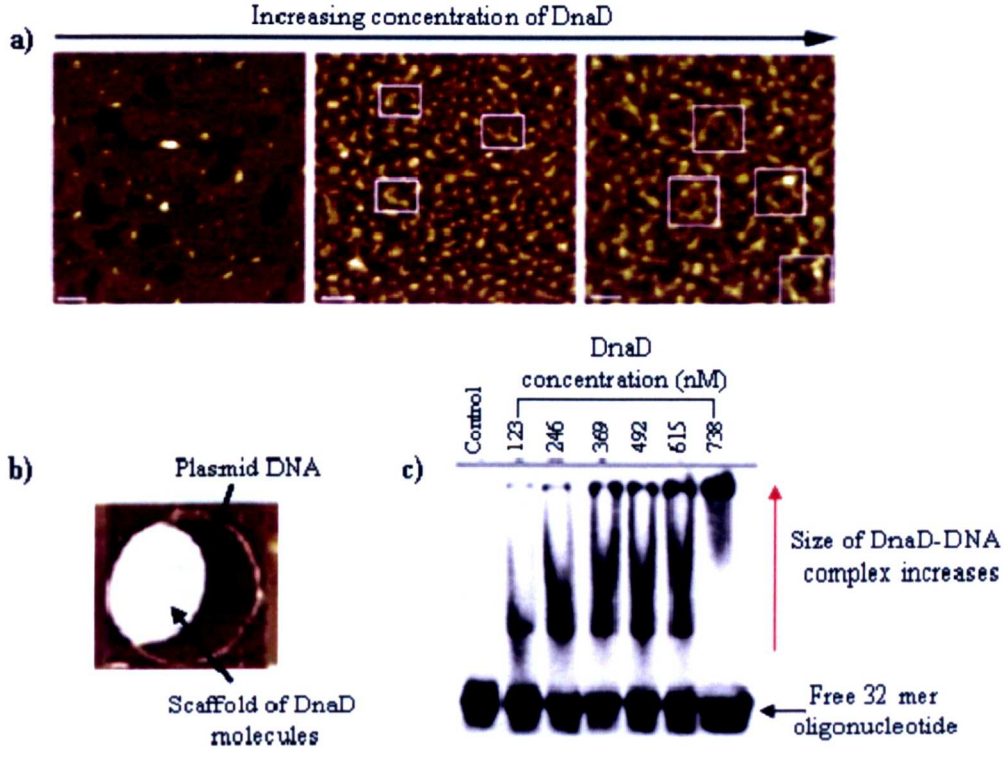


Figure 1.6: DnaD forms concentration dependent oligomers and nucleoprotein complexes. (a) AFM images showing oligomers formed by DnaD at concentrations of 0.49 nM, 4.9 nM and 49 nM. Elongated chains, semi-circular and circular structures of DnaD molecules become evident at higher concentrations. (b) An AFM image of DnaD-pBR322 complex. DNA is held in an open circular form around the scaffold of DnaD molecules. The scale bars in all AFM images represent 100 nm. (c) Gel shift assay of DnaD binding to 2.5 nM of radioactively labeled oligonucleotide. A shifted band is produced when DnaD binds to the oligonucleotide and as the concentration increases a smear starts to appear higher up the gel. At the highest concentration the shifted band remains in the well. Free DNA remains intact at the bottom of the gels as the size of DnaD complexes increases (Adapted from Turner *et al.*, 2004).

(Figure 1.6c). Once a large scaffold has formed, DnaD then remodels the supercoiled DNA by untwisting the duplex, converting the writhe into twists without changing the link number (Zhang *et al.*, 2005). It is thought such untwisting may compensate for the considerable force required to convert the intact supercoiled plasmid into an open circular form without nicking the DNA. Once the plasmid is opened up, the scaffold of DnaD molecules serves as a supportive frame work, holding the plasmid by force around the periphery keeping it in an open circular form (Turner *et al* ., 2004). Removal of this scaffold results in the plasmid collapsing back to its original supercoiled form (Zhang *et al.*, 2005). This feature has not been observed in any other DNA remodeling protein and thus unique to the remodeling activity of DnaD.

1.4.2 The DNA remodeling activity of DnaD opposes that of DnaB

Like DnaD, the *B. subtilis* primosomal protein DnaB exhibits DNA remodeling activity, however, its remodeling effect is opposite to that observed for DnaD (Zhang *et al.*, 2005). DnaD opens up supercoiled plasmid, as described in section 1.4.1, whereas DnaB binds to and compacts supercoiled plasmid (Zhang *et al.*, 2005). It forms nucleoprotein structures that are bead-like with DNA wrapped around the protein (Zhang *et al.*, 2005) as shown in Figure 1.7a. In the presence of both DnaD and DnaB proteins, supercoiled plasmids adopt a unique bi-polar conformation (Figure 1.7b) with the oligomers of the two proteins located at diametrically opposite ends (Zhang *et al.*, 2005). DnaB has been suggested to counteract the effect of DnaD, as higher concentrations of DnaD are required to start converting supercoiled plasmid into an open circular form when DnaB is present (Zhang *et al.*, 2005). However, cooperativity between the two proteins when binding to the plasmid has also been suggested (Zhang *et al.*, 2005). A direct interaction between DnaD and DnaB has not yet been detected, but it has been shown that both proteins are needed together to interact with SSB-coated ssDNA indicating that a protein-protein interaction is likely (Bruand *et al.*, 2005).

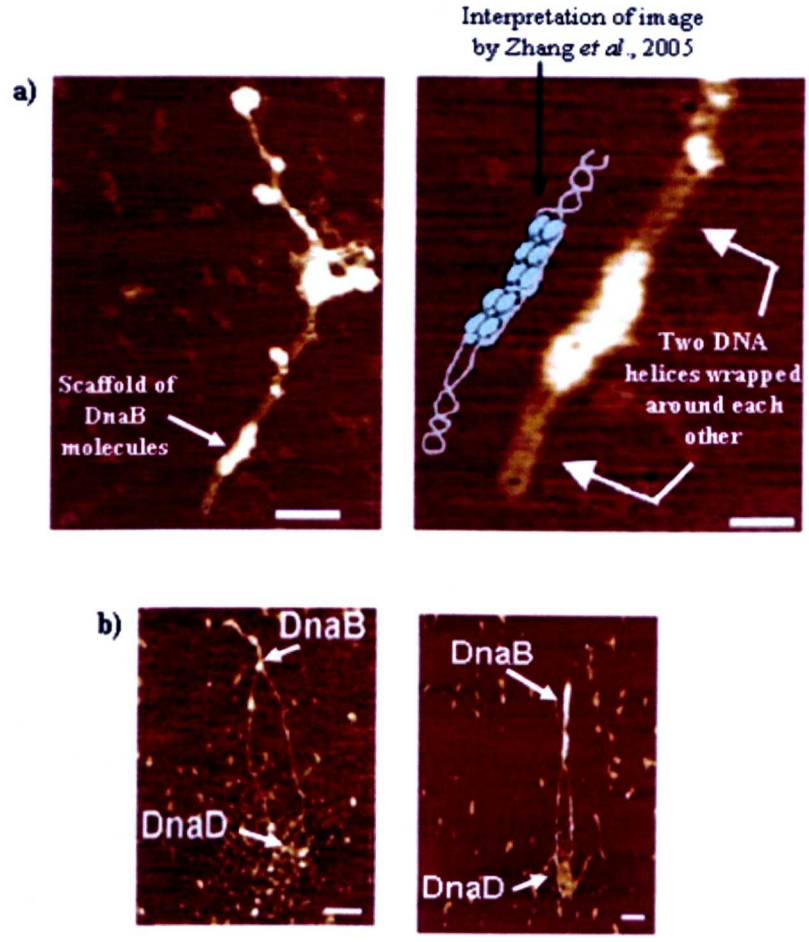


Figure 1.7: AFM images of plasmid DNA in the presence of DnaB alone or DnaB and DnaD together. (a) DnaB forms large foci at random positions along the length of DNA and causes two helices to wrap around each other, introducing the formation of supercoils, as shown in the magnified image (top right). The scale bars in the top left and right panels represent 100 nm and 50 nm respectively. (b) Bi-polar nucleoprotein structures at concentrations of DnaD higher than DnaB. Compact (DnaB-bound) and open (DnaD-bound) segments of the plasmid are visible. In both these panels, the scale bars represent 100 nm (Zhang *et al.*, 2005).

1.4.2.1 The DNA remodeling effect of DnaD and DnaB is analogous to *E. coli* HU and H-NS proteins

The effect of the DNA remodeling activities of DnaD and DnaB resemble that observed for the *E. coli* histone-like proteins HU (heat unstable nucleoid protein) and H-NS (Histone-like nucleoid structuring proteins) (Zhang *et al.*, 2005). Both HU and H-NS are non-specific DNA binding proteins and major components of the bacterial nucleoid (Schumann, 2006). They affect the overall nucleoid structure and topology, but also participate in gene regulation and DNA replication (Schumann, 2006). The HU protein binds to DNA and constrains supercoils by wrapping the DNA into nucleosome-like structures, whereas H-NS induces supercoils condensing the DNA (Dame and Goosen 2002; Schumann, 2006). Therefore, HU makes the nucleoid accessible to other proteins, as does DnaD and H-NS opposes this activity like DnaB.

1.4.3 DnaD is the link between DNA remodeling and initiation of DNA replication

At present the biological significance of the remodeling activity of DnaD is unclear. However, it has been proposed that the primary role of DnaD is its involvement in DNA architecture and its effect in the initiation of DNA replication is a secondary role resulting from its ability to alter the global architecture of the bacterial nucleoid (Zhang *et al.*, 2005). A model has been proposed to accommodate the remodeling activity of DnaD with its role in replication initiation (Turner *et al.*, 2004; Zhang *et al.*, 2005). In this model (Figure 1.8), DnaB binds to the membrane and defines the membrane attachment site. It then utilizes its DNA compaction activity to recruit large segments of the bacteria nucleoid to the membrane. Immunofluorescent studies have shown DnaB to co-localize with the *oriC* around the time of initiation and the essential role of the membrane in *B. subtilis* replication initiation is also well documented (Imai *et al.*, 2000; Rokop *et al.*, 2004). In the model, the membrane attached DnaB then recruits cytoplasmic DnaD to the *oriC* by a direct protein-protein interaction. It has been suggested that such an interaction may be regulatory, as maintaining DnaD in the cytoplasm would prevent initiation (Rokop *et al.*, 2004). Once recruited to the *oriC*

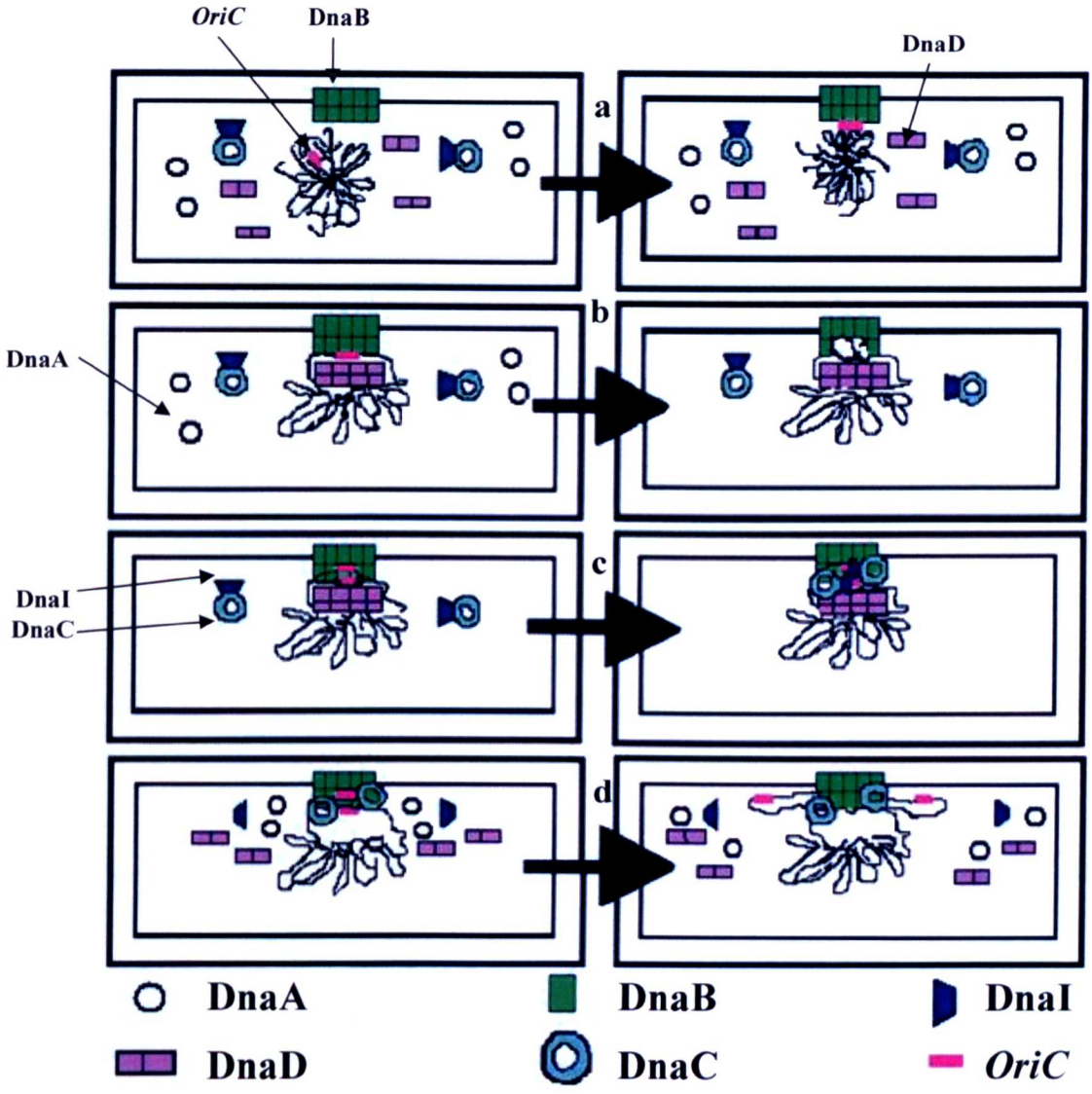


Figure 1.8: Functional model presenting DnaD as a potential link between nucleoid remodeling and initiation of DNA replication in *B. subtilis*. a) DnaB defines the membrane attachment site, and recruits supercoiled DNA to the *oriC*. DnaD interacts with DnaB and opens up the supercoiled DNA locally by forming a scaffold of DnaD molecules. b) Replication initiator proteins (DnaA) bind co-operatively to DnaA boxes at the *oriC* and the DNA duplex unwinds locally. c) DnaC is delivered onto the single strands by the helicase loader, DnaI. d) Following the delivery of DnaC, DnaI dissociates and the helicase activity of DnaC unwinds DNA in both directions from the *oriC*, displacing DnaA in the process (Zhang *et al.*, 2005).

DnaD binds to DNA and forms scaffolds in a concentration dependent manner to locally open up the supercoiled *oriC* and exposing it to the major replication initiator protein DnaA. A relatively high number of DnaD molecules estimated at 3000-5000 per cell support such an *In vivo* remodeling role (Bruand *et al.*, 2005). In the model, DnaA binds to the DnaA boxes at the exposed *oriC* stimulating unwinding, producing an open duplex. Then the DnaI helicase loader forms a complex with DnaC to mediate its delivery to the *oriC* via an interaction with DnaB. DnaD disrupts the stable interaction between DnaC and DnaI to facilitate the loading and correct positioning of the helicase onto the single stands of DNA. The primase and the polymerase is loaded sequentially to complete the replisomes, which proceed bi-directionally from the *oriC* replicating the bacterial chromosome. This model presents DnaD as the link between DNA remodeling and initiation of DNA replication (Turner *et al.*, 2004). The interaction of DnaD with DnaB appears to be the main event that couples its remodeling activity with DNA replication (Turner *et al.*, 2004).

1.4.4 DnaD consists of two domains with distinct activities

With a combination of domain prediction and standard secondary structure prediction programs, it has been identified that DnaD consists of two domains (Carneiro *et al.*, 2006): N-terminal (residues 1-128) and C-terminal (residues 129-232). The N-terminal domain (DnaD-Nd) exists predominately as a tetramer and has concentration dependent oligomerisation activity (Carneiro *et al.*, 2006). It forms large scaffold structures in the absence of DNA (Figure 1.9a) that are similar to those observed for full length DnaD bound to DNA (Turner *et al.*, 2004; Carneiro *et al.*, 2006). The appearance of these DnaD-Nd scaffolds indicates that this domain may mediate protein-protein interactions. Also, the tetrameric state of this domain compared with the dimeric state of DnaD in the absence of DNA, suggests that the DnaD-Nd has two oligomerisation patches, one of which many be masked by the C-terminal domain (DnaD-Cd) in the full length protein (Turner *et al.*, 2004; Carneiro *et al.*, 2006).

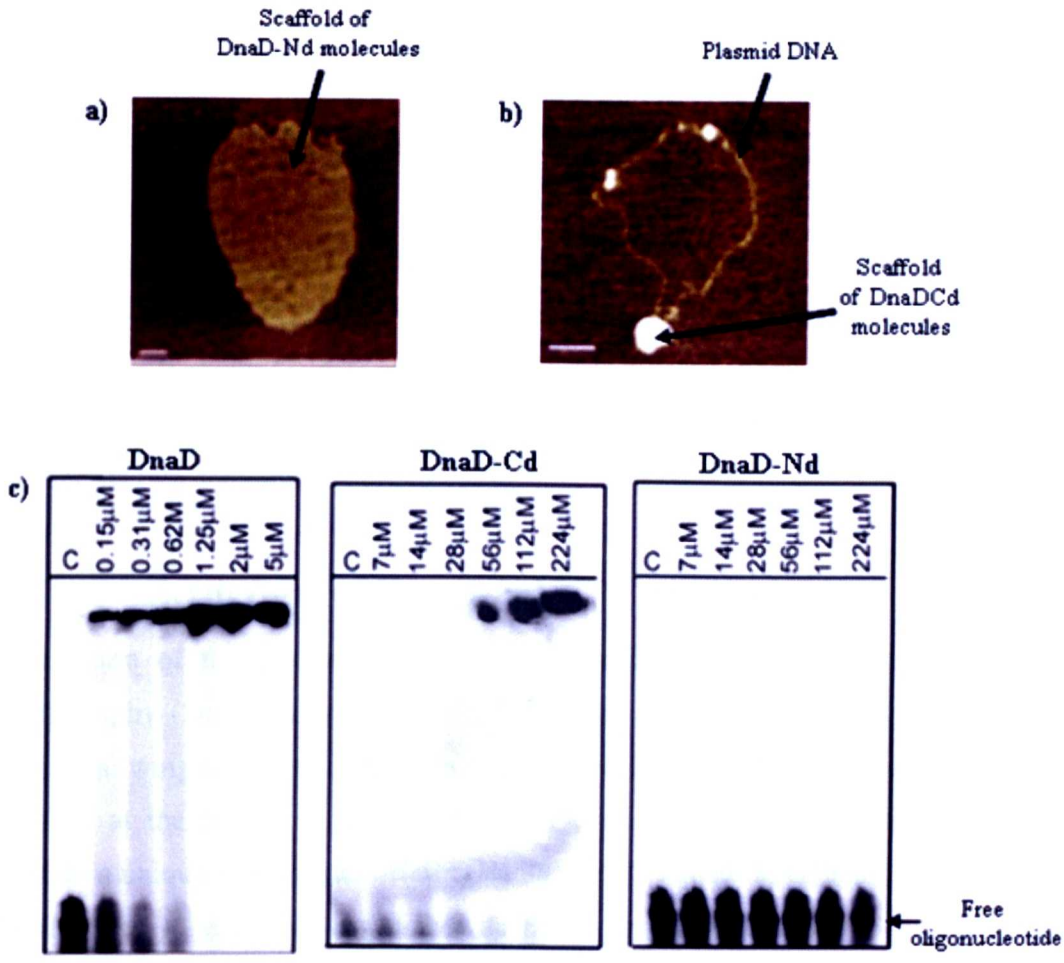


Figure 1.9: The N-terminal and C-terminal domains of DnaD have distinct activities. AFM images of the scaffolds formed by (a) DnaD-Nd and (b) DnaD-Cd bound to plasmid. The scale bars in both images represents 200 nm. (c) Gel shift assays of DnaD and its two domains binding to a 54 mer single stranded oligonucleotide at 2.5 nM. DnaD and DnaD-Cd bind DNA forming nucleoprotein complexes that are too big to enter the gel. DnaD-Nd does not bind DNA. Lane C shows the control reaction which is oligonucleotide in the absence of protein (Carneiro *et al.*, 2006).

The C-terminal domain (DnaD-Cd) is monomeric (Carneiro *et al.*, 2006). It binds DNA and exhibits a separate DNA–dependent oligomerisation activity by which it forms large complexes that are distinct from the scaffolds, formed by both DNA bound DnaD and the DnaD-Nd alone (Turner *et al.*, 2004; Carneiro *et al.*, 2006). This domain does not demonstrate DNA-remodeling activity when separate or when mixed with the DnaD-Nd (Carneiro *et al.*, 2006), suggesting that the two domains must be linked within the same polypeptide, as the sum of their activities is required for DNA remodeling. Furthermore, higher concentrations of DnaD-Cd than DnaD are required for comparable binding to DNA in gel shift assays (Turner *et al.*, 2004; Carneiro *et al.*, 2006), supporting that these domains many function co-operatively. Figure 1.10 summarises oligomerisation of DnaD and its two domains.

1.4.5 Structure of DnaD-Nd

The structure of the N-terminal domain of DnaD has been determined by X-ray crystallography (Schneider *et al.*, 2007; Schneider *et al.*, 2008; Huang *et al.*, 2008), revealing a winged Helix-Turn-Helix fold (WH fold), with additional polypeptide extensions at the N and C termini (Figure 1.11a). The extension at the N-terminus comprises a unique Helix-Strand-Helix (H1'-S1'-H2') structural element, while a single helix (H3') of 17 residues long makes up the C-terminal extension (Schneider *et al.*, 2008). Both these extensions participate in dimerisation and tetramerisation interactions. In the dimer (Figure 1.11b and c), the β-strand (S1') of the N-terminal extension from each monomer, pairs up forming a two stranded β-sheet (Schneider *et al.*, 2008). The β-sheet of two dimers come together in the tetramer (Figure 1.11d) forming a unique 2 x 2 stranded β-sheet that has not been observed in any other oligomerisation protein of known structure. In the structure, the cross over observed between the H1' helix of the N-terminal extension of one dimer and the H3' of the C-terminal extension of another dimers, suggests that this region may contribute to the tetramerisation interface, in addition to the β-sheet tetramerisation motif (Schneider *et al.*, 2008).

The wings and helices of the WH fold remain exposed in the tetrameric assembly (Figure 1.11d). In particular, helix H3 which is exposed an all four corners of the

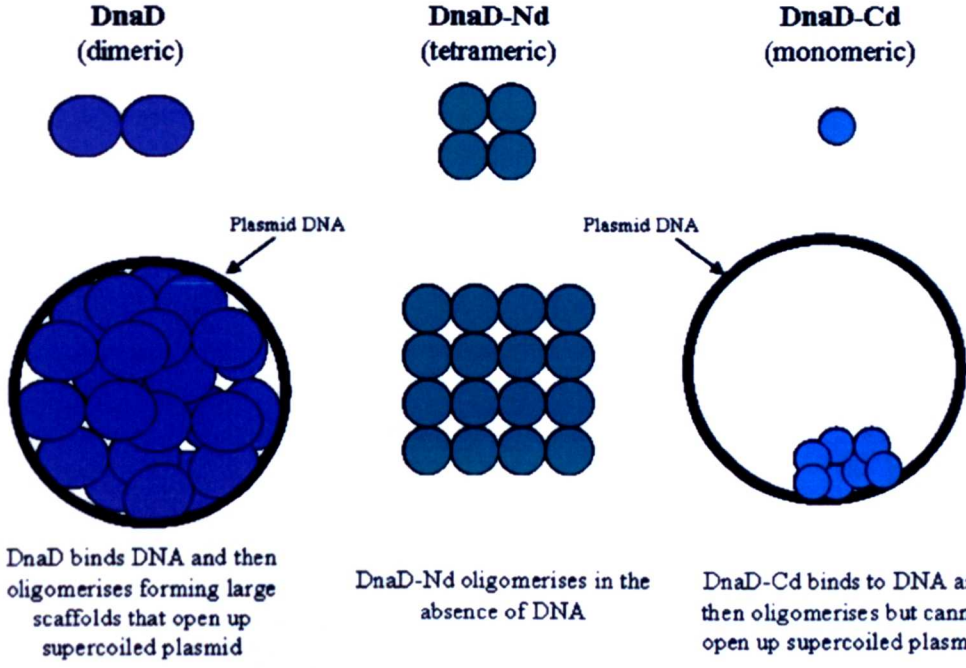


Figure 1.10: Overview of the oligomeric state of DnaD and its two domains. DnaD molecules (purple circles) exists predominantly as dimers, while DnaD-Nd (green circles) and DnaD-Cd molecules (blue circles) exist as tetramers and monomers respectively. In the presence of plasmid DNA, DnaD oligomerise forming scaffolds that open up the supercoiled plasmid. DnaD-Nd oligomerises in the absence of DNA and DnaD-Cd only oligomerises when bound to DNA.

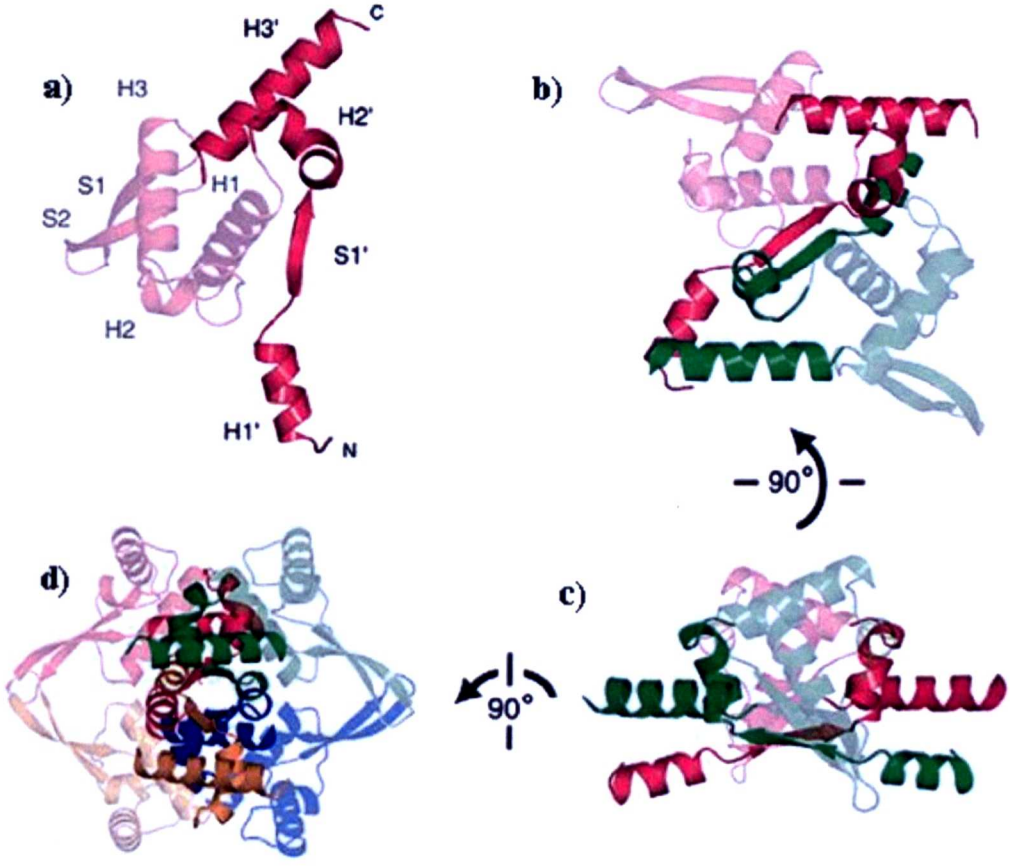


Figure 1.11: Ribbon representations of the X-ray crystallography structure of DnaD-Nd. A (a) monomer, (b and c) dimer and (d) tetramer of DnaD-Nd is shown. The elements of the WH fold are in semitransparent mode and the N- and C-terminal extensions are highlighted (Schneider *et al*, 2008). The two monomers in the dimer are shown in red and green. The monomers in the second dimer of the tetramer are shown in yellow and blue. The dimer in (c) has been rotated 90 degrees relative to the previous image to give a side view (Schneider *et al*., 2008).

tetramer (Schneider *et al.*, 2008). Proteins with a WH motif use helix H3 most commonly, but not exclusively for making contact with the major grove of DNA (Gajiwala *et al.*, 2000). DnaD-Nd does not have a DNA binding function as discussed in section 1.4.5, therefore the H3 may engage in protein-protein interactions as seen for ESCRT-II complex (Hierro *et al.*, 2004; Teo *et al.*, 2004). The H3' helix of the C-terminal extension is also significantly exposed in the tetramer. Structure-directed mutagenesis of the DnaD-Nd combined with AFM imaging have indicated that the WH fold and H3' helix of the C-terminal extension are both likely to participate in cross tetramer interactions that are important for scaffold formation (Schneider *et al.*, 2008).

1.4.6 Speculative model for scaffold formation

Schneider *et al.*, 2008 have proposed a structural model for the formation of DnaD mediated scaffold, based on the activity of the full length protein, the individual domains and the crystal structure of DnaD-Nd. They have also assumed that some of the interactions that mediate scaffold formation are present in the crystal lattice thus, the arrangement of DnaD-Nd molecules seen in the lattice have been considered when constructing the model (Figure 1.12a). In the crystal lattice, contact between DnaD-Nd tetramers is made almost exclusively by the H3 helix and the wings (Figure 1.12b). The H3' helix of the C-terminal extension is not involved in crystal contacts. This may be because a particular set of interactions dominate, giving the lattice order and the DnaD-Nd molecules directionality. It also possible that a particular or more common conformation of the wings, which is associated with a high level of flexibility, has been selected by crystallisation conditions and thus affects the direction and orientation of molecular packing. In the lattice, there is sufficient space between the DnaD-Nd tetramers to accommodate four DnaD-Cd molecules which would be essential for full length DnaD scaffold formation.

In view of the data thus far Schneider *et al.*, 2008 propose that two conformers of the protein exist (Figure 1.13). These conformations named, D2 and D2* differ only by the orientation of their C-terminal domain relative to the N-terminal domain. This difference makes one conformation, D2, incapable of forming structures that are larger than dimers.

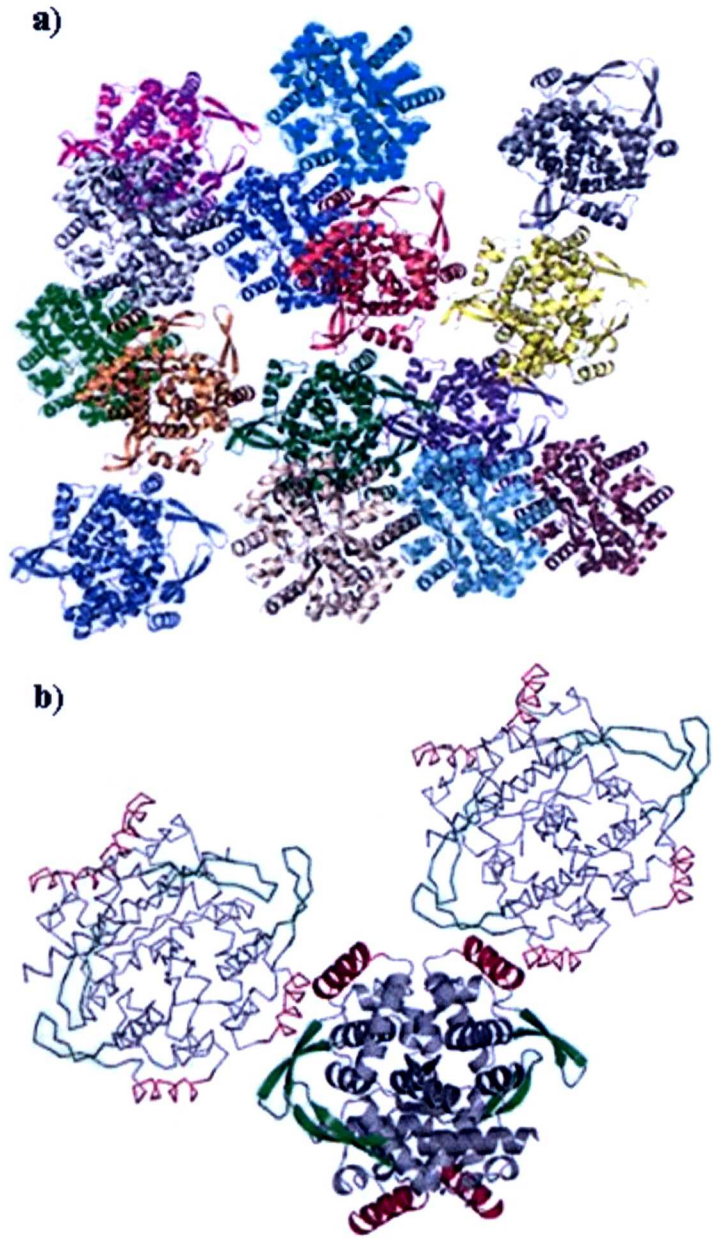


Figure 1.12: A lattice-based model for DnaD-Nd scaffold formation. (a) The packing of DnaD-Nd molecules along a plane of the crystal lattice is shown. Overlapped molecules have been removed to make it easier to view association between molecules. Each coloured molecule represents a different tetramer. (b) Three tetramers from the crystal lattice are shown to highlight the inter-tetramer interaction involving H3 (red helices) and the wing hairpin (green β-sheets) (Schneider *et al.*, 2008).

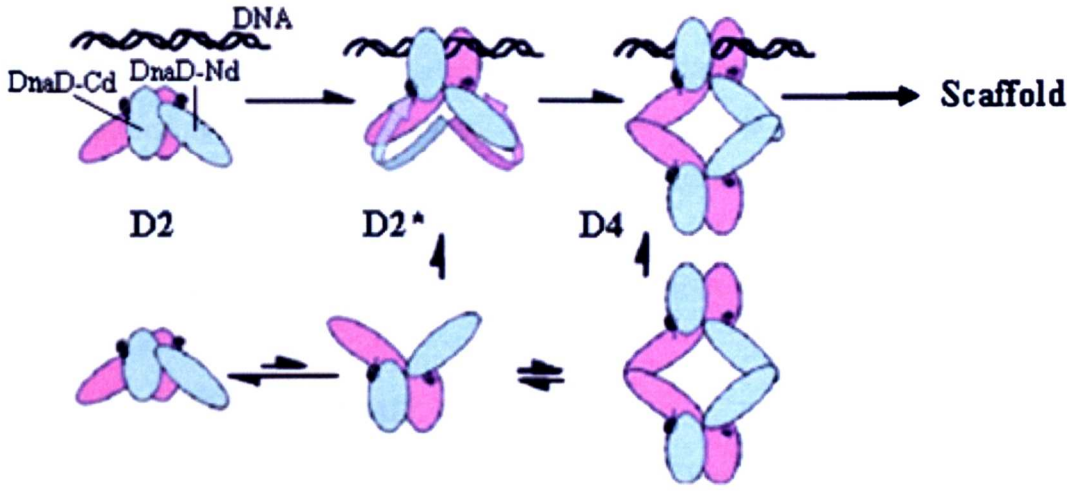


Figure 1.13: Model for DnaD-mediated scaffold formation. As a dimer, DnaD exists in two conformations, D2 and D2*. Only the D2* conformation can form the tetramer (D4) and only the tetramer forms the scaffold. In the absence of DNA, D2 dominates, as the equilibrium shifts from D2* to D2. When DNA is introduced, both dimers interact with the DNA. D2* becomes stabilized but D2 undergoes a conformational change converting it to D2*(shown as curly arrows). This D2*-DNA complex attracts another D2* to form a tetramer (D4-DNA). This tetramer draws in more D4 molecules forming the scaffold (adapted from Schneider *et al.*, 2008).

Therefore, tetramers (D4) and higher order oligomeric structures can only form through the association of D2*. In the absence of DNA, D2, D2* and D4 all exist, but D2 dominates. As DNA becomes available the C-terminal domain of D2 interacts with the DNA and undergoes a conformational change that converts it to the more stable form D2*. An increase in the concentration of D2* encourages the formation of the D4-DNA species. It is this D4-DNA that has been suggested to act as a seed for the formation of scaffolds observed within the boundaries of plasmid DNA.

1.5 Project aim

The structures of proteins help in the understanding of biological processes at the atomic level. In some cases they are used to help explain or interpret previous experiments and in other cases they are used to suggest alterations in natural systems or guide modifications in a more controlled fashion to answer yet other questions. The recently determined X-ray crystallographic structure of the DnaD-Nd (see section 1.4.6) has led to a better understanding of the molecular details that underpin DnaD-mediated scaffold formation. Elucidating the structure of the DNA binding DnaD-Cd may further enhance the current understanding of this process through guiding site directed mutagenesis studies aimed at revealing the role of the individual structural elements of this domain in DNA binding. Furthermore, it may provide a framework for the interpretation of the mutagenesis and other data. Combining the information from the structure of both domains may possibly provide insight into the structure function relationship of the DNA remodeling role of DnaD. It may also lead to a better understanding of how remodeling of the bacterial nucleoid is associated with initiation of DNA replication in *B. subtilis* and possibly contribute towards elucidating the mechanism by which DnaD links these two processes.

Progress towards determining the structure of DnaD-Cd by X-ray crystallography was made (Schneider *et al.*, 2007). Crystals that were of diffraction-quality (2.9Å) were grown successfully, however, failures of all attempts to reproduce these crystals led to the decision of using an alternative method. Solution state NMR is a well established technique used for structure determination of proteins, RNA and DNA. It is often used

as an alternative to X-ray crystallography in situations where proteins cannot be crystallized. Besides investigating the structure, NMR is commonly used to study dynamics and binding properties of molecules. Currently, this technique is limited to small to medium size proteins, but since DnaD-Cd is quite small (13 kDa), NMR seemed to be a suitable method not only for structure determination but also to study DNA binding properties of this domain.

The aims of this project were to:

- Primarily, determine the structure of DnaD-Cd using solution state NMR.
- Investigate the binding of DnaD-Cd to DNA by NMR DNA titration.
- Investigate the structural similarity between the *B. subtilis* DnaD and DnaB protein. (This was not part of the initial project aims, but came about following the merger of the DnaD-like family into the DnaB_2 family).

1.6 Preliminary sequence analysis of DnaD-Cd

1.6.1 Homologous proteins and secondary structure predictions

Two structures, one from *Enterococcus faecalis* (pdb code 2I5U, released 2006) and the other from *Streptococcus mutans* (pdb code 2ZC2, released 2007) deposited in the Protein Data Bank by the Midwest Center for Structural Genomics, show sequence homology to the C-terminal domain of *B. subtilis* (Pair wise identities of 25% and 41% respectively to residues 129-232 of *B. subtilis* DnaD-Cd). Both structures are composed of four helices (I-IV), with 2zc2 consisting of an additional short helix (V) that spans four residues, L194 to V197 (Figure 1.14). The structure of 2zc2 corresponds to a small region of the protein encoded by gene *SMU_1465C* which has been predicted by the Phyre fold recognition server (Kelly & Sternberg, 2009) to contain an N-terminal DnaD domain. No such domain has been detected in the full length 2i5u sequence, making 2zc2 the closest model to DnaD-Cd (Figure 1.15).

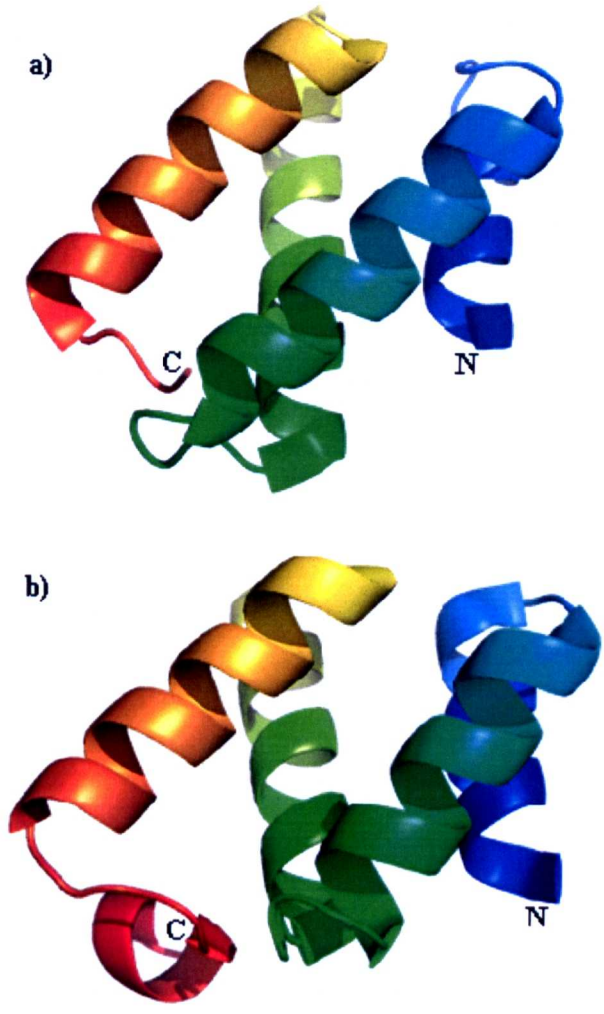


Figure 1.14: Ribbon representations of proteins homologous to DnaD-Cd. Structures of (a) 2i5u and (b) 2zc2 are coloured from blue to red from the N- to the C-termini respectively.

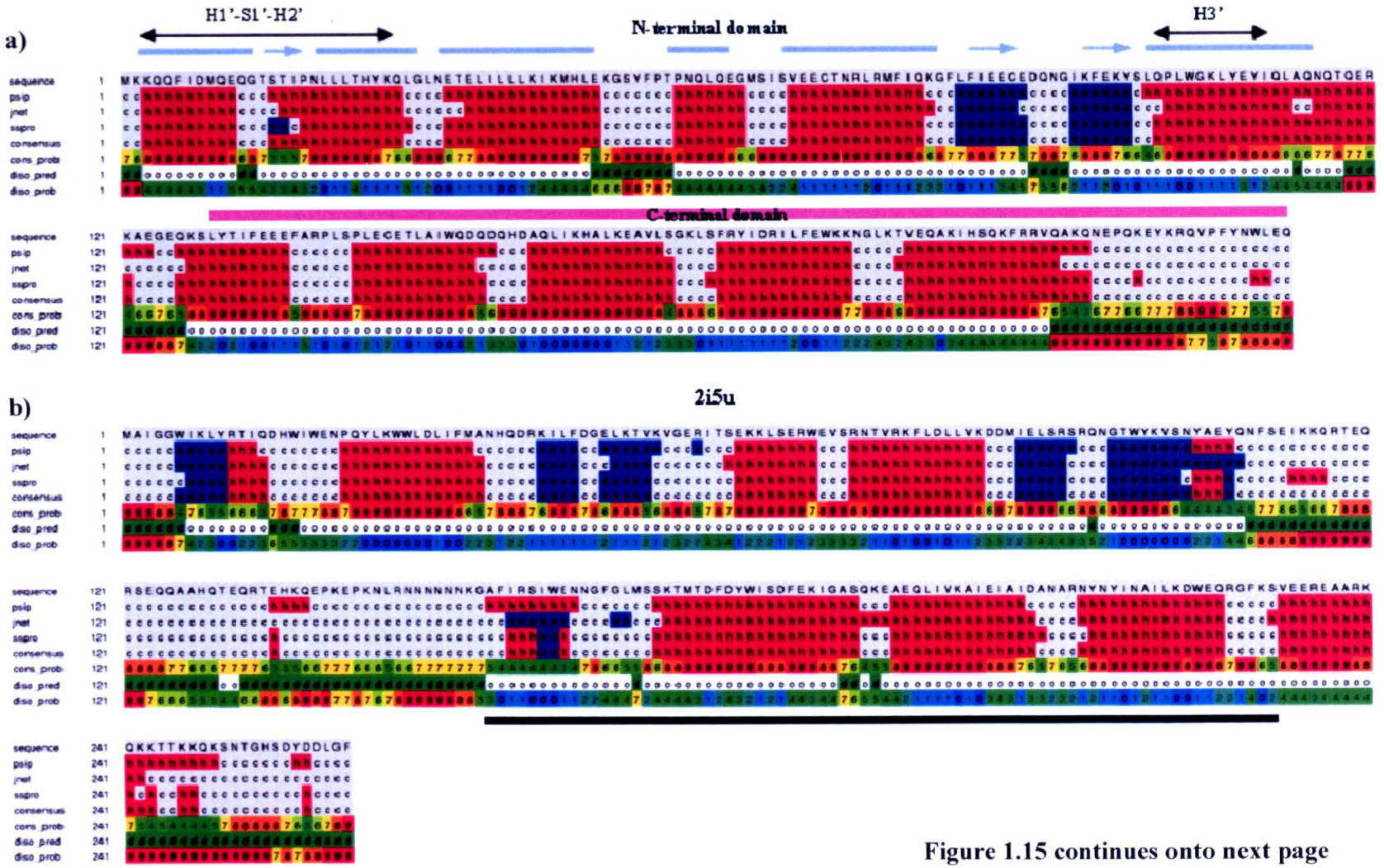


Figure 1.15 continues onto next page

Figure 1.15 Continued from previous page

c)

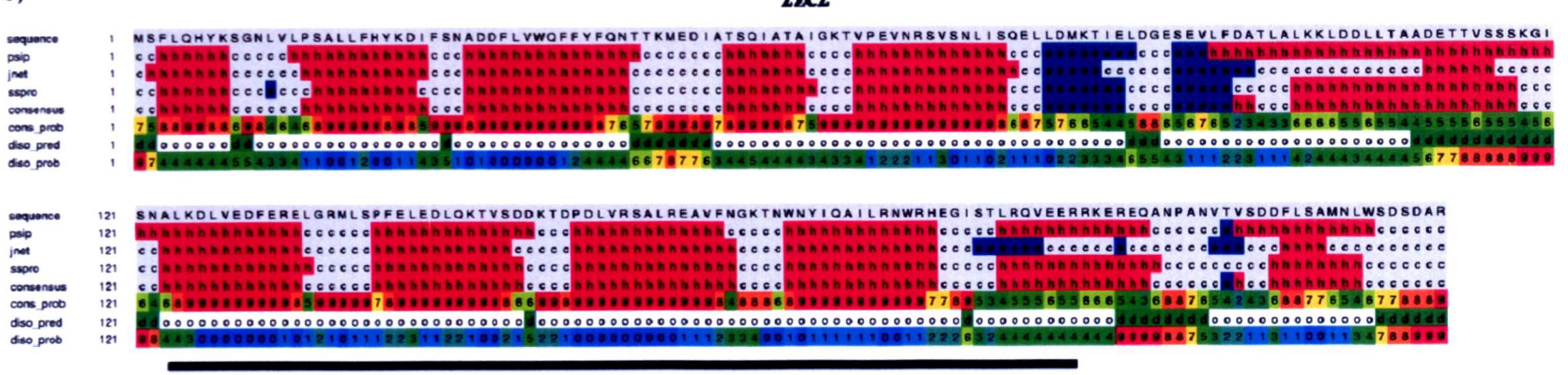


Figure 1.15: Secondary structure predictions from PHYRE for (a) DnaD, (b) 2i5u and (c) 2zc2. Phyre obtains a consensus prediction from three secondary structure predictions (psipred, Jnet and sspro). Below the consensus prediction, a measure of confidence in the prediction (cons_prob) is shown. The bottom two lines in each block give a prediction of order (o) or disorder (d), along with the measure of confidence in this prediction (disco_prob). Structural elements of the N- domain and the location of the C-terminal domain in DnaD are shown along the top of the alignment in blue and magenta respectively. The region of the alignment for 2zc2 and 2i5u that the structures in Figure 1.11 correspond to are shown by the black line.

Secondary structure predictions for the DnaD-Cd region (residues 129-232) by the Phyre server suggest that this domain consists of five helices with a disordered region at the extreme C-terminus, which extends for approximately 23 residues. The lengths of all these helices, except helix V, are consistent with the 2zc2 structure. Helix V in DnaD-Cd, is predicted to extend as far as residue R207, whereas this helix terminates earlier, at V197 (equivalent to A199 in *B. subtilis*. DnaD) in 2zc2. PSI-BLAST sequences alignment (Altschul *et al.*, 1997; http://blast.ncbi.nlm.nih.gov/Blast.cgi) obtained using DnaD-Cd (residues 129-232) for input as the query, shows good conservation up to residue 197, consistent with the terminating boundary of 2zc2. Sequences after this point in the alignment diverge. As a result of this uncertainty in the terminating boundary of helix five, a construct consisting of the entire DnaD-Cd (residues 129-232) was assessed for structure determination (see chapter 4).

Chapter 2: Materials and methods

This chapter describes methods for various techniques used during the course of this work. Several methods stipulated were used repeatedly; any changes employed are detailed in the relevant sections. All chemical reagents used were of analytical grade and purchased from Sigma Aldrich, Novagen, BDH and New England Biolabs unless otherwise stated. Chromatography columns and resins were purchased from GE Healthcare.

2.1 Growth media

2.1.1 Luria-Bertani (LB) media:

10 g Bacto tryptone
5 g Yeast extract
10 g NaCl

The volume was made up to 1 Litre with Milli-Q™ water and then autoclaved to sterilise.

2.1.2 Luria-Bertani (LB) agar:

Bacto-agar was added to LB media (15 g/L). This was adjusted to pH 7.0 with 5 M NaOH and then autoclaved to sterilise.

2.1.3 ^{15}N Minimal media:

6 g Na_2HPO_4
3 g KH_2PO_4
0.5 g NaCl
0.614 g $MgSO_4.7H_20$

The volume was made up to 1 Litre with Milli-Q™ water and autoclaved to sterilize. This was then supplemented with 1 g of (^{15}N) NH_4Cl and 4 g ^{12}C glucose dissolved in 35 ml of water. Biotin (10 mg) and Thiamine (10 mg) were also added as solid

powder.

2.1.4 ^{15}N-^{13}C Media:

Double labelled media was prepared by substituting 4 g of ^{12}C glucose in ^{15}N minimal media with 1.5 g of ^{13}C glucose. The quantities of all other components remained the same.

2.2 Antibiotic selection

All plasmid vectors used contained a *Kan*^{R} gene. A 10 mg/ml sock solution of kanamycin was made in Milli-Q^{TM} water and filtered with a 0.2 μm filter before storage at -20°C in 1 ml aliquots. When required, these stocks were gently thawed and added to media or agar at a final concentration of 30 μg/ml.

2.3 Production of competent cells

Cells (XL1 or BL21 (DE3)) were streaked on LB agar plates and incubated at 37°C overnight. LB media (10 ml) was inoculated with a single colony and grown overnight. More LB media (100 ml) was inoculated with 10 μl of this overnight culture and incubated at 37°C. At an OD_{600} of ~0.6, cultures were stored on ice for 10 min and then centrifuged at 5000 rpm (~2292 x g) for 10 min. The cell pellets were re-suspended in 10 ml of 0.1 M $CaCl_2$, re-pelleted and then re-suspended in a smaller volume of 4 ml. DMSO was added to the cells before and after incubation on ice, for 15 min. Cells were stored in aliquots of 200 μl at -80°C.

2.4 Transformation

Plasmid DNA (2 μl-10 μl) was added to 200 μl of gently thawed XL1 or BL21 (DE3) competent cells and mixed by pipetting. The reactions were incubated for 30 min on ice. These were heat shocked for 45 sec at 42°C and then immediately incubated on ice for 2 min. LB media (200 μl) was then added and each reaction incubated at 37°C for 30 min. These transformations were finally plated on Kan^R (30 μg/ml) agar plates and incubated at 37°C overnight.

2.5 Plasmid amplification and purification

XL1 competent cells were transformed. These were plated on Kan^R (30 μg/ml) agar and incubated overnight at 37°C. Single colonies were used to inoculate10 ml of Kan^R (30 μg/ml) LB media which again was incubated at 37°C overnight. Plasmid DNA was extracted and purified from the XL1 cells with the QIAprep Kit (Qiagen). All plasmids were eluted in 50 μl of water and stored at -20°C.

2.6 DNA clean-up

DNA was extracted and purified with the QIAquick gel extraction kit. This was also used to cleanup DNA from enzymatic reactions.

2.7 PCR

For the amplification of a particular region of pET28a-DnaD-Cd plasmid DNA, the following components were added to a sterile eppendorf and made up to 50 μl with sterile water.

5 μl	10x Reaction buffer
1 μl	pET28a-DnaD-Cd template
1 μl	Forward primer (20 pmol)
1 μl	Reverse primer (20 pmol)
5 μl	dNTP
1 μl	Turbo Polymerase
30 μl	sterile water

Reactions were run in a Techne Progene™ thermocycler with the following setup.

Step	Temperature (°C)	Time (sec)	Number of cycles
Initial strand separation	95	15	1
Strand separation	95	15	30
Primer annealing	42.5	10	
Strand extension	72	20	
Final extension	72	5	1

PCR reactions were run on a 1% agarose gel to check for the presence of amplified product and to estimate the amount of DNA.

2.8 DNA sequencing

Plasmid samples were sequenced at the core genomic facility in house. For complete coverage, sequencing was done in both directions; forward from the T7 promoter and in reverse from the T7 terminator. The data were analysed with FinchTv.

2.9 Agarose gel electrophoresis

Markers:

New England Biolabs 100 bp and 1 kbp ladders were used as size indicators. These were run on 1% agarose gels alongside DNA samples. The amount of DNA was estimated from comparison to the band intensity of markers.

Buffers:

50x TAE buffer:

4.84 g Trishydroxymethylamine, 10 ml Glacial acetic acid and 1 mM EDTA

6x loading dye:

0.25% bromophenol blue and 40% sucrose in water.

Electrophoresis:

Agarose (0.5 g) was dissolved in 50 ml of 1 x TAE buffer to make a 1% gel. This was poured into a sealed cast after the addition of gel red visualisation dye (5 µl). Samples were prepared with 1 x loading dye and a maximum volume of 30 µl loaded. Gels were electrophoresed at 70 V for 90 min in 1x TAE buffer and imaged with a UV IMAGER.

2.10 Cloning of the DnaD-Cd215 and DnaD-Cd206 gene

The DnaD-Cd215 and DnaD-Cd206 constructs were created by site directed mutagenesis (SDM) using the QuickChange II Site directed mutagenesis Kit. The SDM procedure uses PCR to create mutant plasmids with staggered nicks. During PCR, the dsDNA template (pET28a DnaD-Cd vector) is firstly denatured. Then two synthetic primers, each complementary to opposite strands of the template vector and designed with a stop codon at the desired site are annealed and extended without displacement. The primers designed for each construct are shown in Figure 2.0. Following PCR, the reaction mixture is treated with the endonuclease DpnI to digest methylated template DNA. Reactions are run on a 1% agarose gel to confirm the presence of PCR product. Then, XLI cells are transformed with 10 µl of the reaction and incubated at 37°C overnight, on Kan^R agar plates. The plasmid in the single colonies is purified for sequencing as stipulated in section 2.5.

a) Forward Primer (N-terminal)
5'- GAATGAATAACAAAAAGAGTATAAAAGG -3'

Reverse Primer (C-terminal)
5'- CTTTTTGTTATTCATTCTGCTTTGC - 3'

b) Forward Primer (N-terminal)
5'- CAAAAATTCTAACGTGTACAAGCAAAG -3'

Reverse Primer (C-terminal)
5'- GTACACGTTAGAATTTTTGGCTGTG - 3'

Figure 2.0: Primers designed for the amplification of DnaD C-terminal domain truncations. A stop codon (red text) is introduced to create a) DnaD-Cd215 and b) DnaD-Cd206.

2.11 SDS gel electrophoresis

SDS PAGE markers:

Bio-rad pre-stained SDS PAGE markers were used as size indicators. These have been calibrated on a Bis-Tris gel.

Myosin	194,813 Da
B–galactosidase	109,648 Da
Bovine serum albumin	56,779 Da
Ovalbumin	41,491 Da
Carbonic anhydrase	27,750 Da
Soyabean trypsin inhibitor	20,710 Da
Lysozyme	15,389 Da
Aprotinin	6,573 Da

SDS PAGE buffers:

Upper buffer:

0.5 M Tris/HCl, 0.4% (w/v) SDS, pH 6.8

Lower buffer:

1.5 M Tris/HCl, 0.4% (w/v) SDS, pH 6.8

Running buffer:

25 mM Tris/HCl, 0.19 M glycine, 0.1% (w/v) SDS, pH 8.3

Loading buffer:

50 mM Tris/HCl, 100 mM DTT, 2% (w/v) SDS, 0.1% (w/v) bromophenol blue, 10% (v/v) glycerol.

Stain:

0.4% (w/v) Coomassie blue R, 8% (v/v) acetic acid, 46% (v/v) methanol, 10% (v/v) acetic acid, 30% (v/v) methanol.

16% Polyacrylamide resolving gel:

2.5 ml lower buffer, 4 ml 40% (w/v) acrylamide/Bis (29:1), 3.5 ml dH_2O, 100 µl 10% (w/v) APS and 10 µl TEMED.

4% Polyacrylamide stacking gel:

2 ml upper buffer, 1.125 ml 40% (w/v) acrylamide/Bis (29:1), 6.375 ml dH_2O, 110 µl 10% (w/v) APS and 11 µl TEMED.

Electrophoresis:

SDS PAGE was performed using a Bio-Rad mini protein II apparatus with the capacity to run two gels simultaneously. Gels were prepared with a 16% resolving gel below a 4% ploy acrylamide stacking gel, into which a suitably sized comb was inserted, to create wells. Samples were prepared with 50% loading buffer, of which 2-20 µl was loaded. Gels were electrophoresed for approximately an hour at a constant voltage of 200 V in 1x running buffer. These were stained in Comassie Brilliant blue for an hour and then destained in a methanol/acetic acid solution on a slowly rocking platform. The resulting gels were visualised under a light box.

2.12 Concentrating proteins

Dilute protein samples were concentrated using a 5000 MWCO or 3000 MWCO VIVA spin concentrator. Firstly, the membrane was washed with Milli-Q^{TM} water by centrifugation at 6000 rpm (~3300 x g). Then the protein was applied and the volume reduced to achieve the desired concentration.

2.13 Buffer exchanging

Samples with large volumes were concentrated to approximately 1 ml. The desired buffer was then added and the sample concentrated again. This was repeated several times to remove unwanted buffer.

2.14 Protein quantification and molecular weight

All theoretical extinction coefficients and molecular weights shown in Table 2.0 were

calculated from the amino acid sequences by ExPASy (Gasteiger *et al.*, 2005). These were used for the quantification of protein at $A_{280\ nm}$.

Table 2.0: Protein Parameters

Protein	Extinction coefficient (M^{-1} cm^{-1})	Mr
DnaD-Cd232		
Unlabelled ^{15}N labelled ^{15}N-^{13}C labelled	22460	12990.8 13152.8 13743.8
DnaD-Cd196		
Unlabelled ^{15}N labelled	13980	8447.7 8547.7
DnaD-Cd206		
Unlabelled ^{15}N labelled	13980	9645.0 9761.0
DnaD-Cd215		
Unlabelled ^{15}N labelled	13980	10755.3 10790.3

2.15 Protein expression and purification

2.15.1 Protein extraction and purification buffers

All buffers were filtered and degassed using a vacuum pump. These were further treated with argon to remove traces of dissolved oxygen. DTT was only added to buffers once fully degassed.

Lysis Buffer:

100 mM phosphate (pH 7.4), 0.5 M NaCl, 10% w/v sucrose, 1 mM DTT and 20 mM imidazole.

Combined phosphate buffer ($Na_xH_xPO_4$) 0.5M (5X stock):

The pH of this combined buffer is dependent on the ratio of 1 M NaH_2PO_4 and 1 M Na_2HPO_4. Stock solutions (1 M) at pH 7.4 and pH 6.5 were made. These were diluted to 0.5 M in Milli-Q™ water. For a 1 M stock at pH 7.4: ~113 ml of 1 M NaH_2PO_4 was added to ~487 ml of 1 M $Na_2H_2PO_4$. For a 1 M stock at pH 6.5: ~210

ml of 1 M NaH_2PO_4 was added to ~300 ml of 1 M $Na_2H_2PO_4$.

Stripping buffer:

40 mM phosphate (pH 7.4), 0.5 M NaCl and 50 mM EDTA.

Binding buffer:

100 mM phosphate (pH 7.4), 0.5 M NaCl, 20 mM imidazole and 1 mM DTT.

5 M GdHCl binding buffer:

100 mM phosphate (pH 7.4), 0.5 M NaCl, 20 mM imidazole, 5 M GdHCl and 1 mM DTT.

Elution buffer:

100 mM phosphate (pH 7.4), 0.5 M NaCl, 500 mM imidazole and 1 mM DTT.

5 M GdHCl elution buffer:

100 mM phosphate (pH 7.4), 0.5 M NaCl, 500 mM imidazole, 5M GdHCl, and 1 mM DTT.

Gel filtration buffer:

100 mM phosphate (pH 6.5), 150 mM NaCl and 5 mM DTT.

2.15.2 Bacterial growth and protein expression

BL21 DE3 cells were transformed and incubated overnight at 37°C. Kan^R LB (10 ml) was inoculated with a single colony and again incubated at 37°C overnight, while shaking (200 rpm). Cells were pelleted (10 ml) by centrifugation at 6000 rpm (~3300 x g) for 20 min. These were resuspended in 1 ml of LB media and used to inoculate a 2 L flask containing 500 ml of Kan^R media. Cultures were grown to an OD of ~ 0.8 at 37°C with shaking (200 rpm). These were induced with 1 mM IPTG to express protein overnight. Cells were harvested by centrifugation at 5000 rpm (~4650 x g) for 20 min at 4°C and stored at -80°C, after resuspention in lysis buffer.

2.15.3 Cell disruption and protein extraction

PMSF (1 mM) and an EDTA free protease inhibitor cocktail, which specifically inhibits cysteine proteases, aspartic proteases, metalloproteases and aminopeptidases was added at 50 µl per litre of bacteria to gently thawed cell paste. This was sonicated using a MSE Soniprep 150 sonicator for 5x 40 seconds at the maximum amplitude setting of 20 microns. The soluble fraction was separated from the insoluble by centrifugation at 10,000 rpm (~12100 x g) at 4°C for 30 min, and stored at 4°C for purification. Inclusion bodies that make up the insoluble fraction were re-solubilised with 5 M GdHCl binding buffer, to recover more expressed protein. This solubilised lysate was sonicated (4x 20 sec) and then clarified by centrifugation at 10,000 rpm (~12100 x g) for 30 min at 4°C. The resulting supernatant was retained. This treatment of the insoluble fraction was repeated, in order to extract the maximum amount of protein.

2.15.4 Unfolded Immobilized metal affinity chromatography (IMAC)

A 10 ml His-Trap™ nickel based affinity column was equilibrated with 5 M GdHCl binding buffer and then the supernatant retained from re-solubilisation of the pellet loaded. The column was washed with 5 M GdHCl binding buffer and protein eluted using a 20–500 mM imidazole gradient, in 5 M GdHCl elution buffer. Fractions containing DnaD-Cd were pooled for refolding.

2.15.5 Protein refolding

Pooled fractions were transferred to a dialysis bag of appropriate capacity. This was placed in a beaker containing 4 litres of binding buffer, and stored overnight at 4°C while stirring gently. The following morning, the binding buffer was changed and stored for a minimum of 2 hours. Protein aggregates were removed by centrifugation at 10,000 rpm (~12100 x g) for 30 min at 4°C before loading onto the next purification column.

2.15.6 Folded Immobilized metal affinity chromatography- I (IMAC-I)

Protein from the soluble fraction was incubated with Dnase I and 10 mM magnesium chloride at 25° C for 30 min. This was loaded on a 10 ml His-Trap™ nickel based affinity column that had been pre-equilibrated with binding buffer. The column was washed with 60 ml of binding buffer, before eluting protein with a 20-500 mM imidazole gradient in elution buffer. A 20 µl sample from each fraction collected was analysed by SDS PAGE. Fractions containing reasonably pure DnaD-Cd were pooled and dialysed to remove imidazole, in preparation for His-tag removal. All these steps, excluding the incubation with Dnase I, were also used to purify the refolded protein. After this step in the purification protocol the method for purifying soluble and refolded protein was identical.

2.15.7 N-terminal His-tag removal

Protein samples were incubated with restriction grade thrombin at 1.5 units per mg of protein. These were incubated while shaking at 25°C overnight.

2.15.8 Folded Immobilized metal affinity chromatography II (IMAC-II)

Thrombin treated samples were loaded on a 10 ml His-Trap™ column pre-equilibrated with binding buffer. The column was washed with approximately 60 ml of binding buffer during which His-tag cleaved protein was collected. Fractions containing His-tag cleaved DnaD-Cd were pooled and concentrated to 5 ml for further purification.

2.15.9 Gel filtration chromatography

DnaD-Cd (5 ml) was loaded on a superdex 75 gel filtration column that had been pre-equilibrated with 450 ml of gel filtration buffer. DnaD-Cd was eluted in the same buffer. Fractions containing monomeric DnaD-Cd were pooled and concentrated to a final concentration of approximately 0.6 mM. This was then flash frozen in 500 µl

aliquots with liquid nitrogen, and stored at -80°C.

2.16 CD spectroscopy

Two stock solutions, one with 6 M GdHCl and one without, were prepared with 20 μM DnaD-Cd. These were combined in varying amounts to make 27 samples of DnaD-Cd in a GdHCl range of 0-6 M. On preparation, samples were immediately incubated to equilibrate at 25°C for 2 hours, after which CD spectra were acquired using a Jasco 11800 spectropolarimeter. Cuvetts (1 mm) containing 250 μl of sample were placed in a thermostated cell holder and secured by a spacer. The temperature was adjusted to 25°C and CD measurements of sample or buffer were recorded at wavelengths 190-250 nm. Parameters were set as: Sensitivity 100 mdeg; Data pith 1nm; Scanning mode continuous; Band width 1 nm; scanning speed 20 nm/min; Response 16 s and cell length 0.1 cm. The data obtained, in millidegree was plotted as mean residue ellipticity ($degcm^2dmol^{-1}$).

2.17 NMR sample preparation

All NMR samples were prepared with 10% D_2O for locking on the deuterium frequency. Argon was used to remove oxygen from tubes and an extra 5 mM DTT added to discourage aggregate formation.

Chapter 3: Overview of NMR methods

This chapter briefly describes the background to the NMR experiment used during the course of the work. Spectral parameters are also detailed in the relevant section of this chapter.

3.1 NMR spectra

All NMR spectra were recorded on Bruker DRX spectrometers operating at a proton frequency of 600 or 800 MH_Z at 298K. Spectrometers were controlled by UNIX work stations running Topspin (Bruker). The data was transferred to LINUX workstations and processed with Felix 2007 (Accelrys, CA, USA).

3.2 Assignment and Preliminary NMR structure determination

3.2.1 Backbone resonance assignment

The assignment of backbone HN, NH, CO, Cα and Cβ resonances were accomplished by the five experiments listed below in the order they were examined.

HSQC (Bodenhausen *et al.*, 1980; Piotto *et al.*, 1992)	**correlates HN to NH**
HNCO (Grzesiek and Bax, 1992; Kay *et al.*, 1994)	**correlates HN_i to CO_{i-1}**
HN(CA)CO (Clubb *et al.*, 1992)	**correlates HN_i to CO_i and $[CO_{i-1}]$**
HNCA (Grzesiek and Bax, 1992; Kay *et al.*, 1994)	**correlates HN_i to $C\alpha_i$ and $[C\alpha_{i-1}]$**
CBCA(CO)NH (Grzesiek and Bax, 1993; Muhandiram and Kay, 1994)	**correlates HN_i to $C\alpha_{i-1}$ and $C\beta_{i-1}$**
HNCACB (Muhandiram and Kay, 1994; Wittekind and Mueller, 1993)	**correlates HN_i to $C\alpha_i$, $C\beta_i$, $[C\alpha_{i-1}]$ and $[C\beta_{i-1}]$**

The name of the experiment shows the route by which polarization was transferred,

and in brackets are the atoms for which no frequency information was obtained. The subscripts refer to the residue number correlated, for example, HN_i to CO_{i-1} means the HN of a residue correlates to the CO of the preceding residue (i-1). In square brackets, weaker secondary correlations are highlighted.

The different pulse sequences of these 3D heteronuclear experiments, as well as the HSQC are constructed of a combination of INEPT (Insensitive Nucleus Enhancement Polarization Transfer) blocks which allows polarization to transfer between different nuclei. In all these experiments, magnetisation is initially generated by excitation of ^{1}H nuclei for enhanced sensitivity and then sequentially transferred via INEPT blocks to less sensitive hetero-nuclei (^{13}C or ^{15}N) for frequency labelling before transferring the magnetisation back to ^{1}H for signal detection. The selectivity in transfer of polarization is achieved by the use of selective carbon pulses. Therefore, at the end of each experiment information about the N, H and the selected carbon frequencies was obtained after Fourier transform. Acquisition parameters for the backbone spectra are shown in Table 3.0.

3.2.2 Asstool: Program for backbone assignment

The assignment of spin systems to residues within the protein sequence was performed with the program asstool (Reed *et al.*, 2003). This program builds up a sequence of matching spin systems using the Monte Carlo stimulated annealing method which compares 'self' chemical shifts from a spin system to 'preceding' shifts from all other spin systems. It also matches the carbon chemical shifts (Cα, Cβ, Co) of spin system to amino acid types. Asstool typically performs 30 separate runs iteratively. In each run, spin systems are randomly assigned to a residue in the protein sequence and scores are calculated for the chemical shift matches of 'self' and 'preceding' residue. Scores are also calculated for chemical shift similarity between assigned residue type and characteristic random coil amino acid shifts. Each run is performed in an iterative manner where assignments are randomly swapped and scores recalculated until a stable score is obtained for three consecutive iterations. On completion of all 30 runs, the output from each run is collated to produce a list of spin systems corresponding to each residue in the protein sequence.

Table 3.0: Experimental parameters for the main NMR spectra used to assign DnaD-Cd.

Experiment	Spec. (MHZ)	D1				D2				D3			
		Nuc.	Sw (Hz)	No pts	Acq (ms)	Nuc.	Sw (Hz)	No pts	Acq (ms)	Nuc.	Sw (Hz)	No pts	Acq (ms)
Backbone													
HSQC	600	^{1}H	9615.3	2304	119.8	^{15}N	2128.4	412	120.3				
HNCO	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}C	1509.0	100	33.1
HNCACO	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}C	1509.0	100	33.1
HNCA	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}C	3621.7	128	17.7
CBCACONH	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}C	9959.7	90	4.5
HNCACB	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}C	9959.8	140	7.0
Sidechain													
HBHACONH	600	^{1}H	9615.3	2048	106.5	^{15}N	1581.1	72	22.8	^{13}H	6001.5	150	25.0
CCH TOCSY	800	^{1}H	8389.2	2048	122.1	^{13}C	3470.9	80	11.5	^{13}C	11311.8	300	13.3
HCCH TOCSY	800	^{1}H	8389.2	2048	122.1	^{13}C	3470.9	80	11.5	^{1}H	8389.2	300	17.9
Structural													
$^{13}C^{15}N$-edited NOESY	800	^{1}H	12820.5	2048	79.9	^{13}C	4629.6	64	6.9	^{1}H	11204.5	372	16.6
						^{15}N	2111.0	64	15.1				

Abbreviated headings; Spec: ^{1}H frequency of spectrometer used, Nuc: nuclei recorded, Sw: Spectral width, No pts: number of points and Acq: acquisition time.

3.3 Side-chain assignment

Resonance assignments for side-chain α and β protons were carried out using a HBHACONH (Grzesiek and Bax, 1993; Muhandiram and Kay, 1994) spectrum. The assignments of other aliphatic carbons and their associated proteins were determined from CCH and HCCH TOCSY spectra (Kay *et al.*, 1993). These 3D TOCSY experiments use cross-polarisation to allow correlation of a CH pair with all carbons along the side-chain in the CCH TOCY or with all the protons in the HCCH TOCSY. The majority of remaining resonances, mainly the aromatic side-chains of Trp, Tyr, Phe and His and amides of Asp and Glu were assigned from ^{1}H-^{1}H through space connectivity information in the NOESY spectra. Acquisition parameters for the side-chain assignment spectra are shown in Table 3.0.

3.4 NOE assignment

^{1}H-^{1}H distance restraints were derived form the ^{13}C and ^{15}N edited HSQC NOESY spectra (Marion *et al.*, 1989). In these experiments a ^{13}C or ^{15}N HSQC type spectrum is extended into a third dimension; the NOE dimension. Each cross peak originates from an NOE between two protons that possess dipolar coupling. These NOEs are created when protons exchange magnetization in a distance dependant fashion during the mixing time. A signal is only observed if the distance between the protons is less than approximately 5Å. The^{15}N edited NOESY detects NOEs to a NH from any other proton in the molecule, where as the ^{13}C NOESY detects NOEs to a CH group. These spectra provide information on the chemical shifts of the two protons involved and also on the attached N or C. The intensities of the NOE cross peaks are related to the distance of the interacting spins by: $1/r^{6}$ where r is the distance between the two interacting spins. Acquisition parameters for the spectra are shown in Table 3.0.

3.5 ARIA

ARIA (1.2) was used to perform an automated assignment of NOEs and monitor structure determination (Rieping *et al.*, 2007). This software generates ambiguous NOE distance restraints from a list of NOE peaks in the ^{15}N and ^{13}C NOESY spectra, provided by the user. It accepts unambiguous NOE restraints and dihedral angle

information to apply unmodified throughout the structure calculation protocol. An external structure calculation engine, CNS is launched by ARIA to perform calculations and generate structures (see section 3.6). Initially, 20 structures are calculated by applying all distances restraints supplied. Distances from the 7 lowest energy structures are then used to reject violating restraints from the ambiguous assignments of NOESY cross peaks. Another set of 20 structures are then calculated with the modified ambiguous restraint list, the 7 best structures are selected and the NOE assignment further refined. In this way many rounds of calculations are performed, and conformers analysed for improving restraints, until an ensemble of structures with high convergence is achieved. The quality of the final ensemble of structures is evaluated within ARIA by analysing their energy and the agreement of residues with allowed regions of the Ramachandran plot.

3.6 CNS

CNS was launched by ARIA to carry out structure calculations for all 9 iterations. Sets of structures were generated using a simulated annealing protocol. Initially, a string of amino acids defined by the protein sequence was taken and subjected to the experimentally derived NOEs and dihedral angles restraints. The movement of atoms was simulated at high temperature to allow free movement through conformational space. Then the temperature was slowly cooled and Van der Waals forces introduced. This reduced the velocity of atoms, gradually restraining them into their low energy conformations.

3.7 ^{15}N Relaxation

In NMR experiments, the relaxation of a particular nucleus is influenced by fluctuations in the magnetic field experienced by the nucleus. Such fluctuations are primarily generated when the whole molecule changes its orientation, as it tumbles in solution. Relaxation is also affected by localised motion present at that site; therefore all nuclei in a given protein do not have the same relaxation time. The relaxation of ^{15}N amides depends mostly on the amide proton motion and is generally used as a probe to gain information on the structure, areas of local mobility and the correlation time of the protein. Typically, a series of amide correlated spectra are collected to

measure ^{15}N longitudinal relaxation (T1): the return of magnetisation to the z axis and ^{15}N transverse relaxation (T2): the decay of x-y magnetization to zero. In this study only T2 relaxation experiments were carried out since these alone were sufficient to provide information on the overall correlation time and areas of local mobility for DnaD-Cd. In the pulse sequences of these T2 experiments, a refocused INEPT block was used to transfer magnetisation of the amide proton to its attached nitrogen in a similar way to the HSQC spectrum. The ^{15}N magnetisation was then refocused and a CPMG (Carr, Purcell, Meiboom, Gill) sequence that was composed of a series of spin echo elements applied, to spin lock the magnetisation in the transverse plane. During the length of this sequence, relaxation due to ^{15}N T2 occurred. The surviving transverse magnetisation was chemical shift encoded and then transferred back to the amide proton for acquisition. Typically, a series of these T2 experiments are run with a number of spin echo elements that vary the delay. For DnaD-Cd, T2 spectra with relaxation intervals of 0, 17, 33.9, 50.9, 67.8, 84.8, 118.7 and 135.7 ms were acquired. Intensities for each amide resonance were plotted against relaxation intervals and fit to the following equation;

$$\mathbf{I = Ae^{-t/[T]}}$$

Where T is the T2, A is a constant and I is the decrease in intensity upon delay (t). The rate of decay i.e. T2 relaxation rate was measured and plotted against residue number upon completion of resonance assignment.

3.8 DNA Titrations

DNA titrations of DnaD-Cd were performed using a 2 mM stock of a 10 mer (5'-GTTAATTGCTC-3') or 19 mer (5'-CGATCAATAAGGAGTCGCC-3') ssDNA exchanged into identical conditions to the protein. At each DNA concentration ^{1}H 1D, ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC spectra were acquired. Using in house macros, intensity and chemical shift changes were followed over the HSQC series and then plotted separately against DNA concentration for every assigned residue. Only the plots of intensity changes were further analysed as they were much more informative than those of chemical shift changes. All overlapped peaks were excluded from the analysis and Cross-peak intensities were corrected for the volume change before being fit to the following:

$$I = I_0 e^{-t/[b]} + I\infty$$

Where I_0 is the amplitude, -t is the concentration at which the signal plateaus, b is the DNA concentration (%) and $I\infty$ is the intensity value at the maximum concentration of DNA added. There was no obvious theoretical reasoning for the use of this model, it was simply preferred to other complex models involving possibly multiple kds, self association and complex relaxation effects.

Titrations with ^{15}N labelled DnaD-Cd196, DnaD-Cd206 and DnaD-Cd215 were later carried out in a similar manner. For all truncations intensity changes in ^{1}H 1Ds and ^{1}H-^{15}N HSQC were followed and plotted as described above.

3.9 Hydrogen/Deuterium Exchange

Amide protons are labile and readily exchange with their surrounding solvent. However, when these protons are buried within the structure or involved in stable secondary and tertiary structure elements, they become protected. Hydrogen/Deuterium exchange experiments enable the exposure of individual residues in a protein to be estimated from the protection against exchange, which is imposed by the protein structure i.e. the protection factor. Therefore, to determine the exposure of the helices of DnaD-Cd, in particular helix V, a ^{15}N labelled sample was diluted in a 50% D_2O containing buffer. ^{15}N HSQC spectra were recorded at 3.0, 5.0, 7.5, 10.3, 13.2, 15.6, 18.0 and 20.5 hours. Intensities of each amide resonance were plotted against the time at which spectra were recorded and compared to intensities of a control spectrum in which the protein was diluted with an equal volume of water buffer.

Chapter 4: Optimisation of protein production and NMR sample conditions

An efficient structure determination by NMR requires a highly purified ^{15}N labelled sample that remains stable for at least a week. Inhomogeneous preparations, impurities or even aggregation of the protein can severely impair the process. Therefore, the first step in any NMR study involves establishing whether the protein is amenable to NMR analysis by assessing the quality of the sample and behavior of the protein. One spectrum in particular, the HSQC spectrum is commonly used to rapidly provide such an assessment, as it maps the backbone amides of a protein according to their proton and nitrogen chemical shifts. Since the chemical shift of each amide is very sensitive to its environment, the distribution of peaks and number of signals observed in this spectrum can help to guide the decision as to whether a protein is a suitable candidate for structure determination under the chosen conditions. Generally, proteins that give a HSQC spectrum showing the number of peaks expected from the amino acid sequence and signals that are sharp as well as dispersed are classified as 'ideal' candidates. Spectra with some but not all of these features are classified as 'promising'. Usually, the buffer composition of proteins giving such spectra is optimized by adjusting temperature, pH and ionic strength. In some cases alternative buffers are also trialed. Once a satisfactory HSQC spectrum is obtained, the protein is then doubled labeled with ^{15}N and ^{13}C for commencing the structure determination process.

As the first step towards structure determination, the suitability of DnaD-Cd for NMR analysis was assessed. This chapter explains the observations that led to establishing DnaD-Cd as a suitable candidate and describes the strategy employed to overcome the poor protein yield that would have resulted in termination of this project. Details about screening for optimal conditions to improve spectra quality and information about stability of DnaD-Cd are also discussed in this chapter.

4.1 A suitable candidate for structure determination

To assess the quality of spectra DnaD-Cd generates, the ^{15}N HSQC spectrum shown in Figure 4.0 was collected using a ^{15}N labelled sample provided by William Grainger (Nottingham University). Examination of this spectrum showed good signal dispersion in the ^{1}H dimension, suggesting the protein was well-folded. Approximately 160 peaks were counted. The backbone amide of each residue should give rise to a signal in the HSQC spectrum, except the prolines as they are devoid of a proton bound nitrogen and therefore cannot transfer magnetisation. Out of the 160 peaks counted, 105 were expected to come from backbone amides in the protein. Side-chain amides of asparagine and glutamine also give rise to peaks in HSQC spectra. These can be distinguished from backbone amides since they are represented by two peaks in the ^{1}H dimension with identical nitrogen chemical shifts. The 3 asparagine and 11 glutamine residues in the sequence of DnaD-Cd, shown in Figure 4.1 would account for about 28 cross peaks. The three signals around 10 ppm were characteristic for the side-chain amide group of tryptophans and thus it was likely that each peak came from one of the three tryptophans in the protein. At this stage of the project, it was not possible to elucidate the cause of the remaining 21 peaks. However, extra peaks in a spectrum usually indicate impurities or sample inhomogeneity. Conformational changes and intra-molecular interactions between structural elements can also affect the number observed. In some cases such events attenuate signals and in others several signals are observed for one residue. Any one or a combination of these events could be considered as possible causes for the extra peaks observed.

Further analysis of the HSQC spectrum revealed a non-uniform distribution of peak intensities. Many sharp and intense signals were clustered around the centre of the ^{1}H dimension, these were characteristic to unfolded regions of a protein. Signals dispersed across the spectrum were less intense and more varied. The line widths of these signals were also broader than those in the central region of the spectrum, as expected for peaks corresponding to a folded region of the protein.

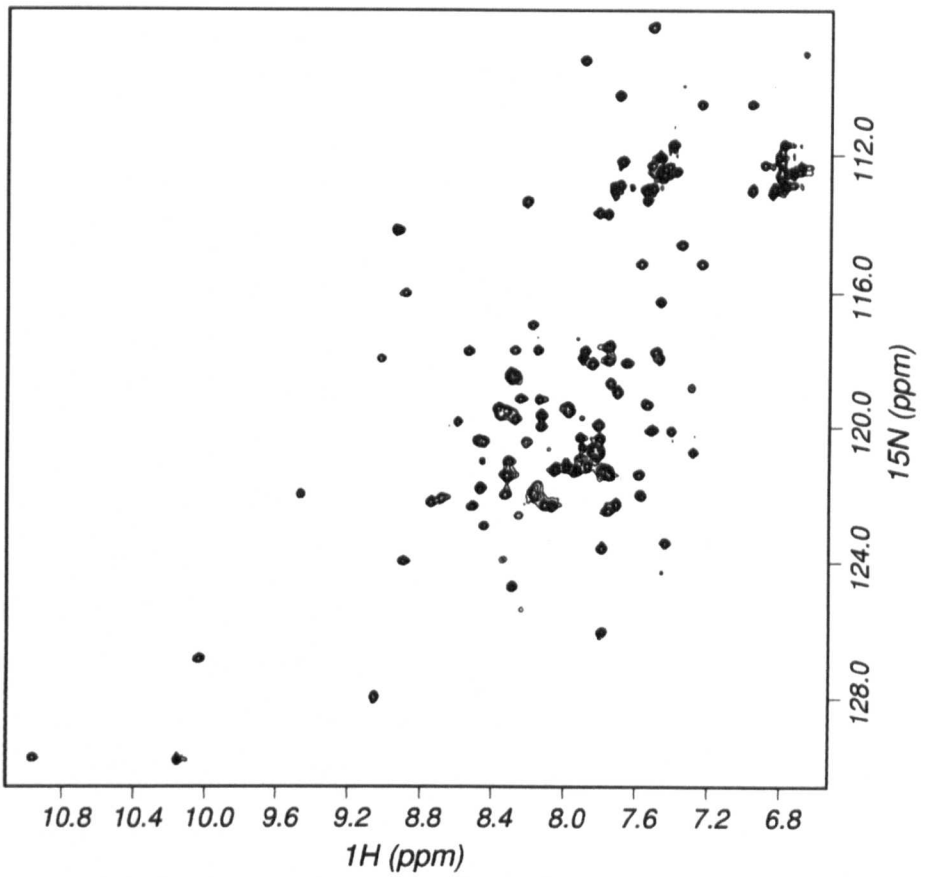


Figure 4.0: The first ^{1}H-^{15}N HSQC spectrum of DnaD-Cd acquired. Good signal dispersion is observed in the ^{1}H dimension, however a greater number of peaks (160) than expected are present. 108 cross-peaks should correspond to backbone amides, 28 to side-chain amides of asp and glu and 3 to side-chains of trp. The remaining 21 peaks were unaccounted for, but may result from the presence of impurities or protein degradation.

130 140 150 160 170 180
GSHMLY TIFEEEFARP LSPLECETLA IWQDQDQHDA QLIKHALKEA VLSGKLSFRY IDRI LFEW
190 200 210 220 230
KK NGLKTVEQAK IHSQKFRRVQ AKQNEPQKEY KRQVPFYNWL EQ

Figure 4.1: Protein sequence for DnaD-Cd used in NMR experiments. The 104 residues in black are for DnaD-Cd from *B. subtilis*. In green are residues that remain after cleavage of the His-tag.

A good approximation of the size of a protein can be obtained from the correlation time, which is a measurement of the time required for a molecule subject to Brownian motion to rotate through a significant fraction of 360°. It is dependent on the size, shape, and dynamics of the molecule, as well as the bulk physical characteristics of the solvent. Therefore the correlation time is directly related to the volume and molecular weight of the protein. When the correlation time is combined with information about peak intensities and line widths, it can also be a good indicator as to whether aggregation takes place under the chosen conditions. For example, a dimeric protein will tumble with characteristics of a protein twice its molecular weight and therefore the correlation time would be larger than that of the monomeric species. However, other factors such as unfolded regions can also slow the tumbling. This only becomes problematic when unfolded regions are large, as the tumbling is slowed severely. Since relaxation of the ^{15}N nuclei can be informative of the overall correlation time and areas of local mobility such as unfolded regions, T2 relaxation measurements were recorded for DnaD-Cd. Experimental and data handling details are given in section 3.7. The transverse relaxation time for each peak in the HSQC spectrum, excluding peaks from side-chain amide, were extracted and plotted in ascending order (Figure 4.2). Inspection of these T2 rates revealed that approximately 20 resonances came from residues in a highly mobile region of the protein, as indicated by long T2 relaxation times. This was consistent with the secondary structure prediction by PHYRE (Figure 1.14) of 19 unstructured residues at the extreme C-termini of DnaD-Cd. The remaining resonances were in a well structured region of the protein, as suggested by shorter relaxation times of approximately 60 ms. Since the ^{15}N relaxation time in structured regions of a protein is often dominated by the overall tumbling, it is frequently used to calculate the correlation time. For DnaD-Cd a correlation time of 12 ns was obtained, which was longer than for a protein of 13 kDa. This may be a combination of the effect of the unstructured C-terminal portion and a degree of transient self- association.

From this assessment it was concluded that the quality of the HSQC spectrum was good and DnaD-Cd was a 'promising' candidate for structure determination. Alternative buffers, protein concentrations and temperatures were screened before commencing

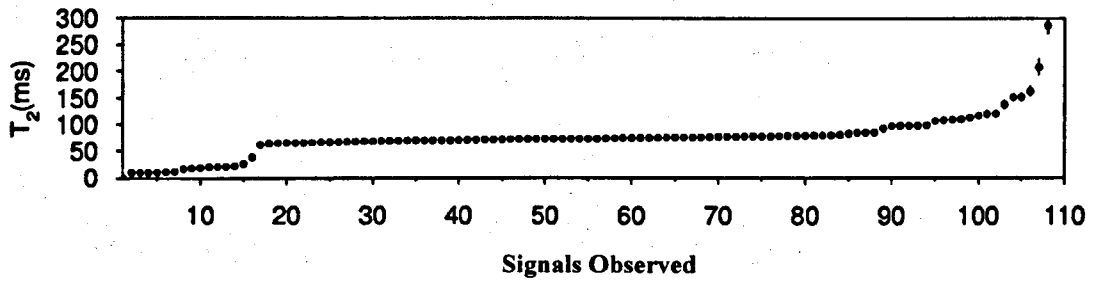


Figure 4.2: T2 ^{15}N relaxation data for DnaD-Cd. The T2 relaxation times (y-axis) for signals observed in ^{15}N HSQC spectra (x-axis) are plotted in ascending order. An intermediate relaxation time of 60 ms is obtained for most resonances signals, suggesting the protein is well structured. The 20 signals with T2 > 70 ms probably correspond to residues within the unstructured region predicted by PHYRE (Figure 1.14), since greater mobility is indicted by longer T2 values. The first 16 signals with T2 < 50 ms may correspond to artefacts in the spectrum.

structure determination, to establish conditions in which degradation and transient self association/aggregation of DnaD-Cd was reduced (See section 4.4 for NMR sample condition optimisation).

4.2 Increasing the quantity of protein produced

Several ^{15}N labelled samples of relatively high concentration (approx. 0.5 mM to 1 mM) were required to screen DnaD-Cd for optimal NMR conditions. The initial expression and purification protocol provided by William Grainger (Nottingham University) gave a single 500 µl sample at 0.6 mM (A total of 4 mg of protein) from 4 litres of media, which is a poor quantity for such growth. Figure 4.3 summarises the expression and purification protocol used to produce the sample mentioned above. This protocol was originally established for protein crystallisation work and thus, the purity of the sample rather than the quantity was of greater importance. At this stage of the project, a poor yield of protein was not of concern as the cost of labelling with ^{15}N is only slightly more than that of preparing an unlabelled sample. However, since DnaD-Cd had been selected for structure determination, at least one ^{15}N and ^{13}C labelled sample was required to record multidimensional spectra for resonance assignment. Labelling with ^{13}C is an order of magnitude more expensive than ^{15}N alone, so a low protein yield from a large growth would not be acceptable.

A number of strategies have been described in the literature to improve the yield of recombinant proteins. Amongst them, optimisation of the expression and purification protocol is one most commonly used. Therefore, in order to improve the yield of DnaD-Cd, the expression and purification protocol was assessed to identify stages that could be optimised.

4.2.1 Optimisation of protein expression

Proteins are generally expressed in bacterial cells with the aim to extract large amount of soluble protein. However, over expression is not always accepted by the metabolic system of the host cell, and can result in the accumulation of protein in inclusion bodies

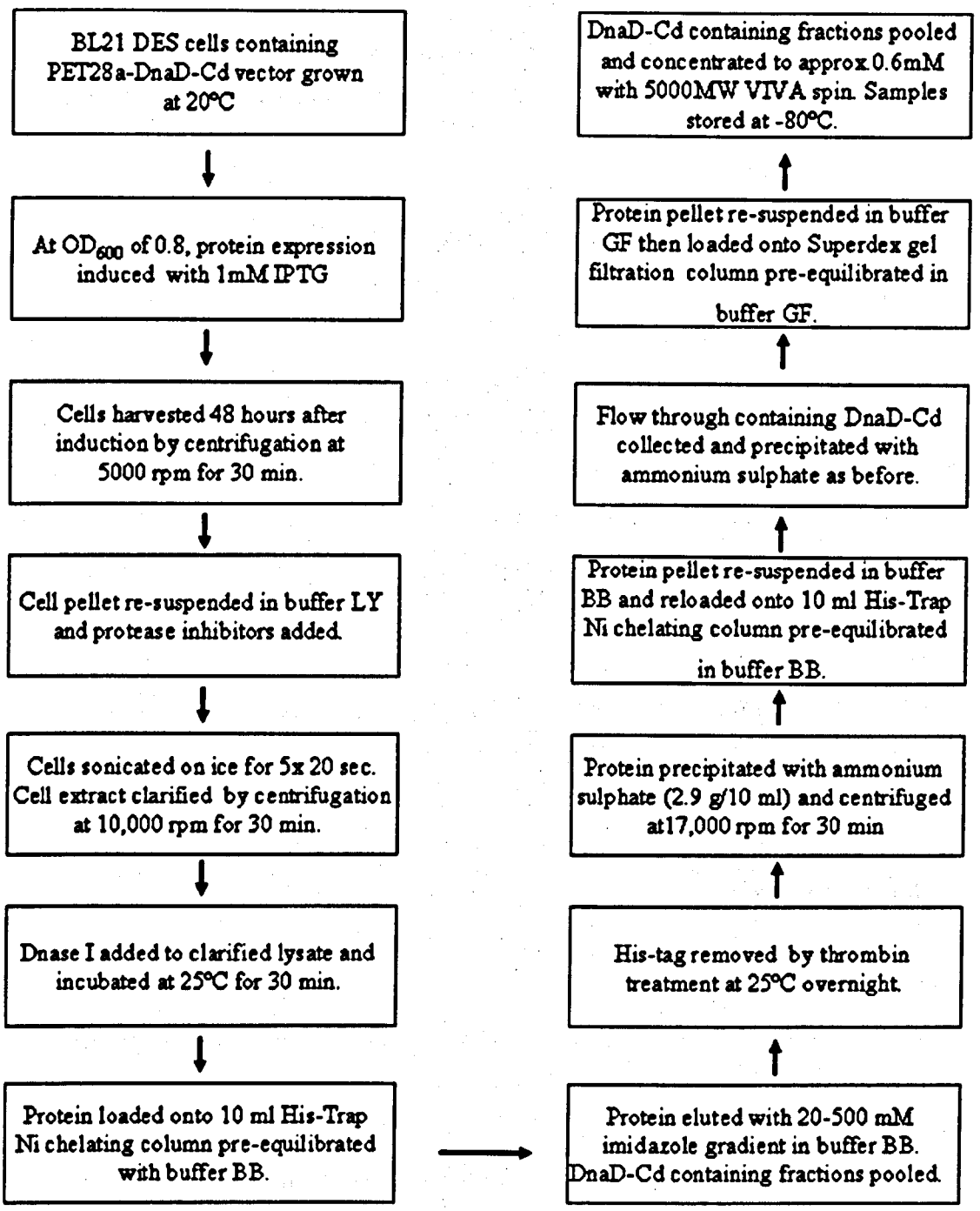


Figure 4.3: Flow chart summarising expression and purification of DnaD-Cd before optimisation of the protocol. The composition of the various buffers used is as follows: Buffer LY; 100 mM phosphate pH 7.4, 0.5 M NaCl, 20 mM Imidazole and 10% w/v sucrose. Buffer BB; 100 mM phosphate pH 7.4, 0.5 M NaCl and 20 mM Imidazole. Buffer GF; 100 mM phosphate pH 7.4, and 50 mM NaCl.

as insoluble aggregates. The temperature at which proteins are expressed has an impact on the ratio of soluble and insoluble protein. In most cases, the accumulation of soluble protein is favoured at lower temperatures (Sørensen and Mortensen, 2005).

The initial expression of DnaD-Cd at a low temperature of 20°C produced a quantity of total protein that was not enough to meet our requirements. Furthermore, half of this total protein expressed had segregated into inclusion bodies. In such a situation, reducing the temperature may not be beneficial as the quantity of protein would remain an issue even though the solubility may improve. Conversely, increasing the temperature may increase expression thus improve the overall yield, however the amount of protein partitioned into inclusion bodies may also increase. Expressing proteins in inclusion bodies is often considered undesirable, as the proteins are devoid of biological activity thus requiring solubilisation and refolding procedures to recover them in a functionally active form. Despite this drawback, expressing in inclusion bodies has its advantages. For example, the difference in size of inclusion bodies compared to other cellular contaminants enables easy isolation. Also, a larger quantity of protein can be expressed in shorter periods of time and be protected from proteolytic attack by cellular proteases. Because of these advantages, overexpressing in inclusion bodies is increasingly becoming the method of choice and is already widely used in the commercial production of many proteins (Walsh *et al.*, 2001). With the option of developing a protein refolding protocol, it was no longer crucial to improve solubility of DnaD-Cd during expression. However, it was still necessary to increase the overall yield. Therefore, an expression trial was done at 25°C and 37°C in LB media. Cells were grown to an OD_{600} of 0.8, and then induced with 1 mM IPTG. Samples (2 x 1 ml) of the cell culture were collected every hour for 8 hours. A sample was also taken at 24 hours. The total protein expression at each time point was compared by SDS PAGE. To prepare samples for the SDS analysis the cell cultures (1 ml) collected at each time point were centrifuged (13,000 rpm (~17,900 x g) for 10 min) to pellet cells. The resulting supernatant was removed and the pellet re-suspended in 200 µl of lysis buffer. Loading buffer (10 µl) was added to 20µl of the re-suspended cells and then heated for 1 min at 60 °C before immediately loaded 15 µl on SDS PAGE gels. Comparison of the SDS PAGE band

intensities inferred that the total protein expressed increased with expression time at both 25°C and 37°C, however, more protein was produced at 37°C at any point in time (Figure 4.4a). From this expression trial it could not be determined whether the increased production of protein was due to a greater number of cells being present or because more protein was being produced per cell, as the OD of samples was not compared. Since the main purpose of the expression trial was to establish an expression time and temperature that would give a good protein yield, the reasons behind the increased protein production were not looked into.

It was expected that a portion of the total protein expressed would partition into the insoluble fraction of the cell, therefore the amount of protein in the two cellular factions was compared using SDS PAGE. To separate the soluble from the insoluble fractions, the 1 ml cell culture samples collected at each expression time point were centrifuged (13,000 rpm (~17,900 x g) for 10 min) to pellet cells. The resulting supernatant was removed and the pellet re-suspended in 300 µl of lysis buffer. This re-suspended pellet was sonicated (2 x 5 sec) and then centrifuged (13,000 rpm (~17,900 x g) for 5 min). The resulting supernatant, which was the soluble fraction of the cell, was transferred to a clean tube and the pellet which was the insoluble faction re-suspended in a volume of H_2O equivalent to that of the supernatant. Loading buffer (10 µl) was added to 20 µl of the soluble and insoluble fractions before loading 20 µl on the SDS PAGE gels. Comparison of the SDS PAGE band intensities of the two fractions revealed that a portion of the protein had partitioned into inclusion bodies, as expected, however this process appeared to be influenced by both expression time and temperature (Figure 4.4b). For example, at 6 hour after induction, more overexpressed protein had partitioned into inclusion bodies at 37°C than at 25°C. Therefore, reducing the expression time at higher temperatures would improve the yield of soluble protein. At 25°C and 37°C, significantly more soluble protein was produced from 4 hours after induction compared to the overall expression at 20°C. It had not yet been established whether refolding DnaD-Cd from inclusion bodies was possible, so an optimal time and temperature to independently express soluble and insoluble protein were selected. The

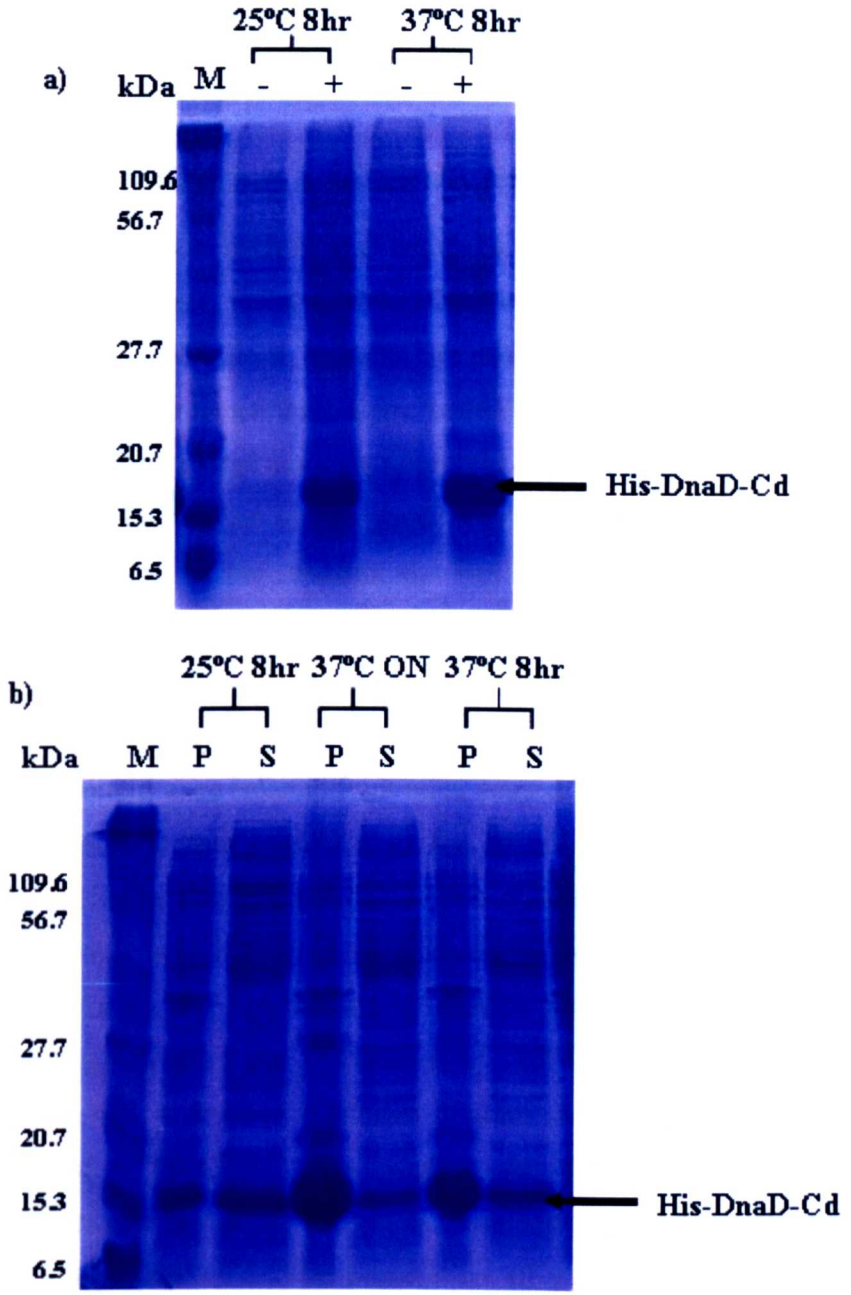


Figure 4.4: SDS PAGE gels showing expression of His-DnaD-Cd at 25°C and 37°C. (a) The total cellular protein expressed immediately before (-) and 8 hours after (+) induction with IPTG at 37°C and 25°C is shown. Equivalent volumes of sample (15 µl) were loaded in each lane. (b) The amount of overexpressed protein partitioned into the soluble (S) and insoluble (P) factions is shown for expression at 37°C 8 hours, 37°C overnight (ON) and 25°C 8 hours. In each lane, 20 µl of sample was loaded. Lane M shows the Mw markers, the sizes of which are as indicated in kDa.

optimal expression time and temperature was 8 hours at 25°C to yield soluble protein and 24 hours at 37°C to yield insoluble protein.

4.2.2 Establishing a solubilisation and refolding protocol

Recovering active protein from *E. coli* inclusion bodies is a four step process: Firstly, the inclusion bodies are isolated from the cell; secondly, protein aggregates are solubilised; thirdly, solubilised proteins are refolded and then finally the refolded proteins are purified. Amongst these steps, solubilisation and refolding are the two most important for recovering large amounts of protein and thus require careful consideration. In particular, the solubilisation agent should be carefully chosen, as different agents vary in the effect they have on the structure and flexibility of the denatured protein, which subsequently affects the efficiency of refolding (Tsumoto *et al.*, 2003).

As the first step in the recovery of active DnaD-Cd, inclusion bodies containing overexpressed protein were isolated from other smaller components in the lysed cells, by centrifugation at 5000 rpm (~4650 x g) for 30 min. The resulting pellet was then solubilised with buffer containing GdHCl. This solubilisation agent denatures proteins into a more flexible and disordered state compared to some other agents like SDS. It supposedly works by directly binding to peptide bonds, unfolding the protein in a concentration dependent manner (Monero *et al.*, 1994). High concentrations of 6-7 M are usually required to achieve sufficient binding (Monero *et al.*, 1994). Initially, a concentration of 6 M GdHCl was used in the solubilisation buffer. Later, CD experiments were carried out to establish the minimum GdHCl concentration required to completely unfold DnaD-Cd. The preparation of samples for CD spectroscopy is detailed in section 2.16. CD spectra were collected for GdHCl concentrations ranging from 0 M to 6.0 M and then overlaid for comparison (Figure 4.5a). The spectrum at 0 M GdHCl showed two distinct negative peaks (222 nm and 208 nm) which were characteristic for a well folded α-helical protein. As the concentration of GdHCl increased, the signal for these two peaks also increased (more positive), as expected for a protein that was being unfolded. At concentrations around 4 M GdHCl, spectra characteristic for random coils were produced inferring that the protein had mostly likely

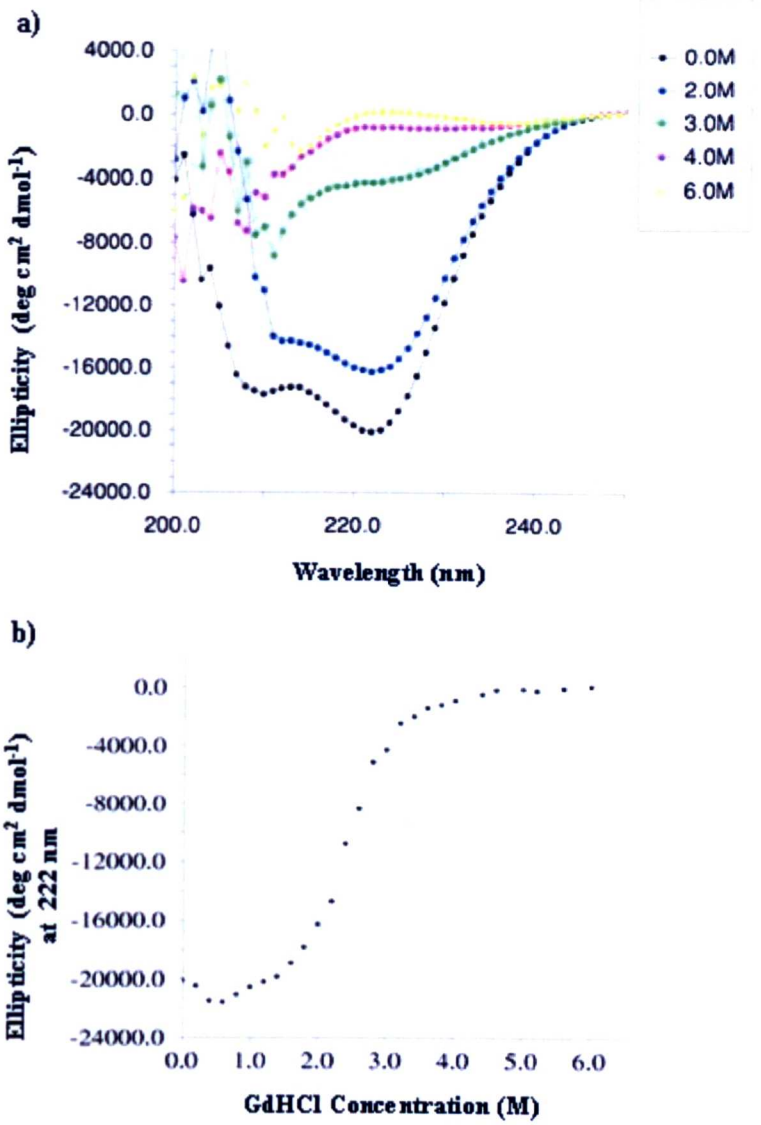


Figure 4.5: CD spectra and a denaturation profile of DnaD-Cd at different GdHCl concentrations. (a) CD spectra of DnaD-Cd were recoded at 25°C for GdHCl concentrations ranging from 0 M to 6 M in buffer consisting of 100 mM phosphate (pH 6.5), 150 NaCl and 5 mM DTT. Spectra at the low GdHCl concentrations show peaks at 208 nm and 222 nm that are characteristic to α- helices. As the concentration of GdHCl increases these peaks become less defined generating curves characteristic to random coils. For clarity of this image, only selective spectra from the range of GdHCl concentrations tested have been overlaid. (b) The GdHCl denaturation curve of DnaD-Cd at only 222 nm is shown. The CD signal around 208 nm becomes increasingly poor with the increasing GdHCl concentration and thus a denaturation curve at this wavelength could not be constructed. The denaturation curve at 222 nm shown infers that DnaD-Cd remains compact and folded at the low denaturant concentration. As the denaturant concentration is increased the protein undergoes an unfolding transition. The protein is completely unfolded by 5 M GdHCl.

completely unfolded. Figure 4.5b shows the denaturation profile for DnaD-Cd constructed by extracting values at 222 nm for all GdHCl concentrations tested. This profile clearly shows the transition of DnaD-Cd from a folded to an unfolded state. Based on these CD results, the GdHCl concentration in the solubilisation buffer was reduced to 5 M.

Alongside GdHCl, the reducing agent, DTT was added to the solubilisation buffer. In a concentrated solution of protein, reducing agents help prevent the formation of inter and intra molecular interactions, by keeping cystine residues reduced. The DTT in a solution of unfolded DnaD-Cd, should prevent disulphide bonds forming between the single cystine in each unfolded protein chain.

Once the inclusion bodies were solubilised, the solution was clarified by centrifugation and then refolded by reducing the GdHCl concentration. The way in which the denaturant concentration is reduced can affect the efficiency of refolding, as this along with the solvent conditions determines the rate of folding verses misfolding and aggregation (Tsumoto *et al.*, 2003). In recent years, a number of methods have been described for reducing denaturant concentrations to refold proteins from inclusion bodies (Tsumoto *et al.*, 2003). Amongst these, dialysis is one most commonly used in small scale operation. In dialysis, solubilised proteins are enclosed into a porous membrane and placed in a large volume of refolding buffer. Proteins are gradually exposed to descending concentrations of denaturant. As the denaturant concentration decreases to the concentration of the surrounding buffer the protein refolds into its native structure. To refold DnaD-Cd, the solubilised protein was dialysed against a refolding buffer (100 mM phosphate pH 7.4, 50 mM NaCl and 1 mM DTT) at 4°C overnight with constant stirring. Soon after initiating dialysis, the protein solution became increasingly turbid; by the end large aggregates were visible. SDS PAGE analysis revealed that these aggregates were mainly composed of DnaD-Cd. Protein aggregation is one of the reasons for poor recovery of refolded protein (Maxwell *et al.*, 2003: Singh and Panda 2005). Besides the method of refolding and sample conditions mention above, impurities in the refolding mixture can also influence aggregation (Maxwell *et al.*, 2003). Removal

of contaminants from solubilised inclusion bodies prior to refolding has been shown to reduce aggregation and thereby increase the yield of refolded proteins (Maxwell *et al.*, 2003). To reduce aggregation of DnaD-Cd during refolding, inclusion bodies containing overexpressed protein were extracted, solubilised and clarified as before. The clarified solution was then loaded on a His-Trap nickel chelating column that had been equilibrated in buffer SB (100 mM phosphate pH 7.4, 0.5 M NaCl, 20 mM Imidazole, 5 M GdHCl and 1 mM DTT). The column was washed with buffer SB and then the protein eluted with a 20 mM -500 mM imidazole gradient in buffer SB. Fractions containing DnaD-Cd were pooled and subsequently refolded by dialysis against the refolding buffer, as described above. At the end of dialysis, a clearer solution with fewer aggregates was observed, compared to the solution refolded without prior purification. These aggregates were removed by centrifugation and the clear solution purified according to the protocol described in Figure 4.3. A ^{15}N HSQC spectrum was then acquired with the purified protein, to verify the success of refolding. The spectrum showed signal dispersion in the proton dimension comparable to the spectrum of protein from the soluble fraction suggesting that the protein had successfully refolded into a similar conformation (Figure 4.6). However, the chemical shift of some crosspeaks has changed and a few new intense signals had appeared. Furthermore, some very low intensity signals were observed predominately in the central region of the spectrum typical for small degradative products and unfolded polypeptides. Overall the refolding procedure was successful for DnaD-Cd.

4.2.3 Optimisation of protein purification

Quantification of protein at every step of the purification protocol, showed a loss of protein at every stage. Approximately 80% of total DnaD-Cd had been lost by the end of purification. Two procedures; thrombin treatment for removal of the His-tag and ammonium sulphate precipitation were identified to account for approximately 60% of this loss. Therefore these procedures in particular were investigated.

4.2.3.1 Removal of the N-terminal His-tag

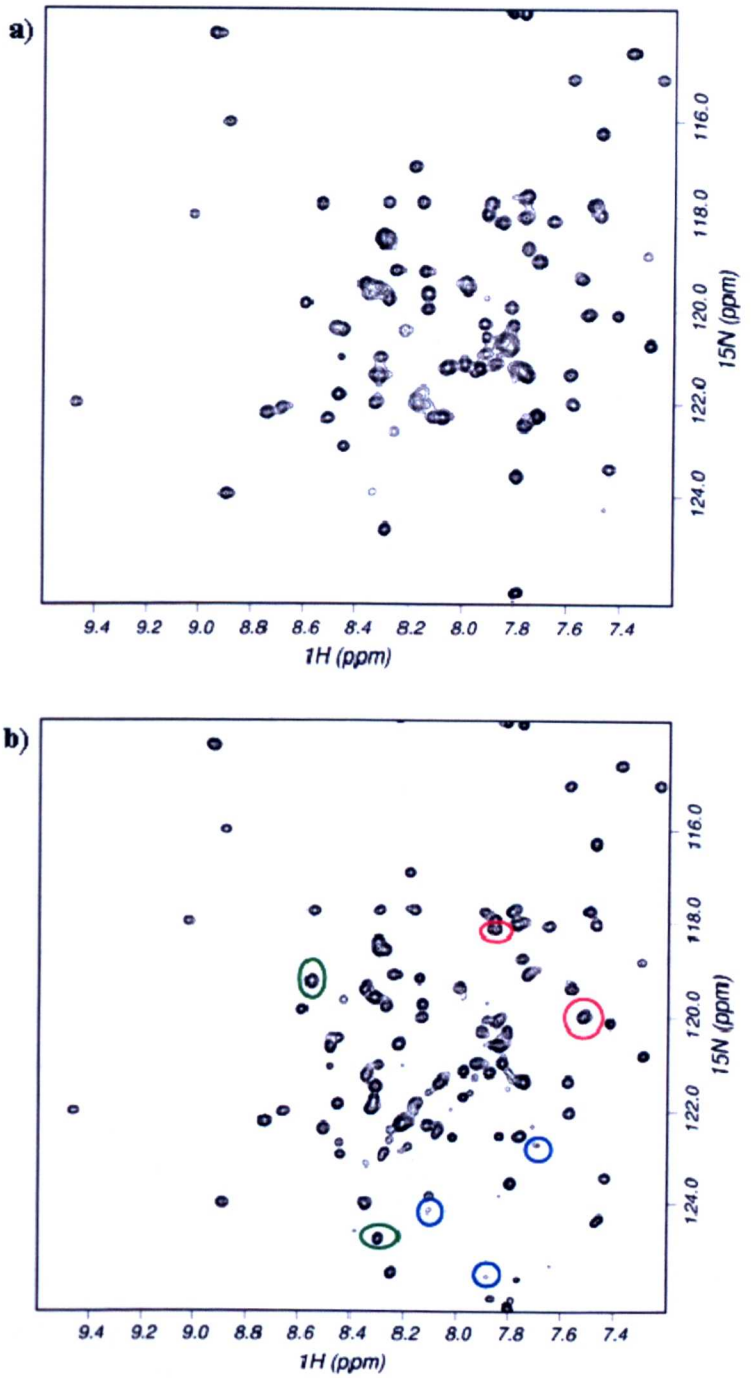


Figure 4.6: Comparison of HSQC spectra of protein extracted from the soluble and insoluble fraction. (b) Refolded DnaD-Cd shows ^{1}H signal dispersion similar to (a) protein purified from the soluble fraction. However, some signals have shifted (red circles), some new intense signals have appeared (green circles) and a few new signals that are of very low intensity have also appeared (blue circles).

The N-terminal His-tag provided an efficient tool for purifying DnaD-Cd without prior knowledge of its biochemical properties. However, this tag was removed during purification as the unstructured nature of such affinity tags may be disadvantageous toward the quality of the HSQC spectra for reasons briefly explained in section 4.1. Nevertheless, leaving the tag on the protein would eliminate the need for thrombin treatment and also reduce the number of purification steps, thereby improve on the loss previously endured. To establish whether cleavage of the His-tag was necessary, the quality of the ^{15}N HSQC spectra collected with and without the tag was compared (Figure 4.7). In spectra from the His-tag protein the line widths were slightly larger and the signals less intense. This may be due to the presence of the additional 12 residues from the His-tag slowing the tumbling of His-DnaD-Cd. Thus, removing the His-tag undoubtedly gave spectra that were of better quality and possibly more suitable for structure determination.

4.2.3.2 Optimisation of thrombin treatment

Thrombin is a site specific endoprotease that cleaves between Arg and Gly in the recognition sequence LeuValProArgGlySer. This enzyme exhibits non-specific cleavage under many conditions, in general 1 unit of thrombin per milligram of target protein is sufficient for efficient cleavage in 1 x thrombin cleavage buffer, at 20°C for 16 hours. However, the efficiency of cleavage under these conditions many vary since the His-tag cleavage site in each protein can be presented differently.

For cleavage of the His-tag on DnaD-Cd, the recommended amount of thrombin mentioned above was used per milligram of protein. Thrombin cleavage buffer was added to the reaction already consisting of 100 mM phosphate pH 7.4, 500 mM salt and 192 mM imidazole before incubation at 25°C overnight while shaking. Under these conditions the solution became increasing turbid over time, by the end of the incubation period precipitate was clearly visible. In order to assess the efficiency of His-tag cleavage under these conditions, samples of DnaD-Cd were extracted before and after thrombin treatment and compared with SDS PAGE. Furthermore, the precipitate that became increasingly visible throughout the treatment reaction was examined.

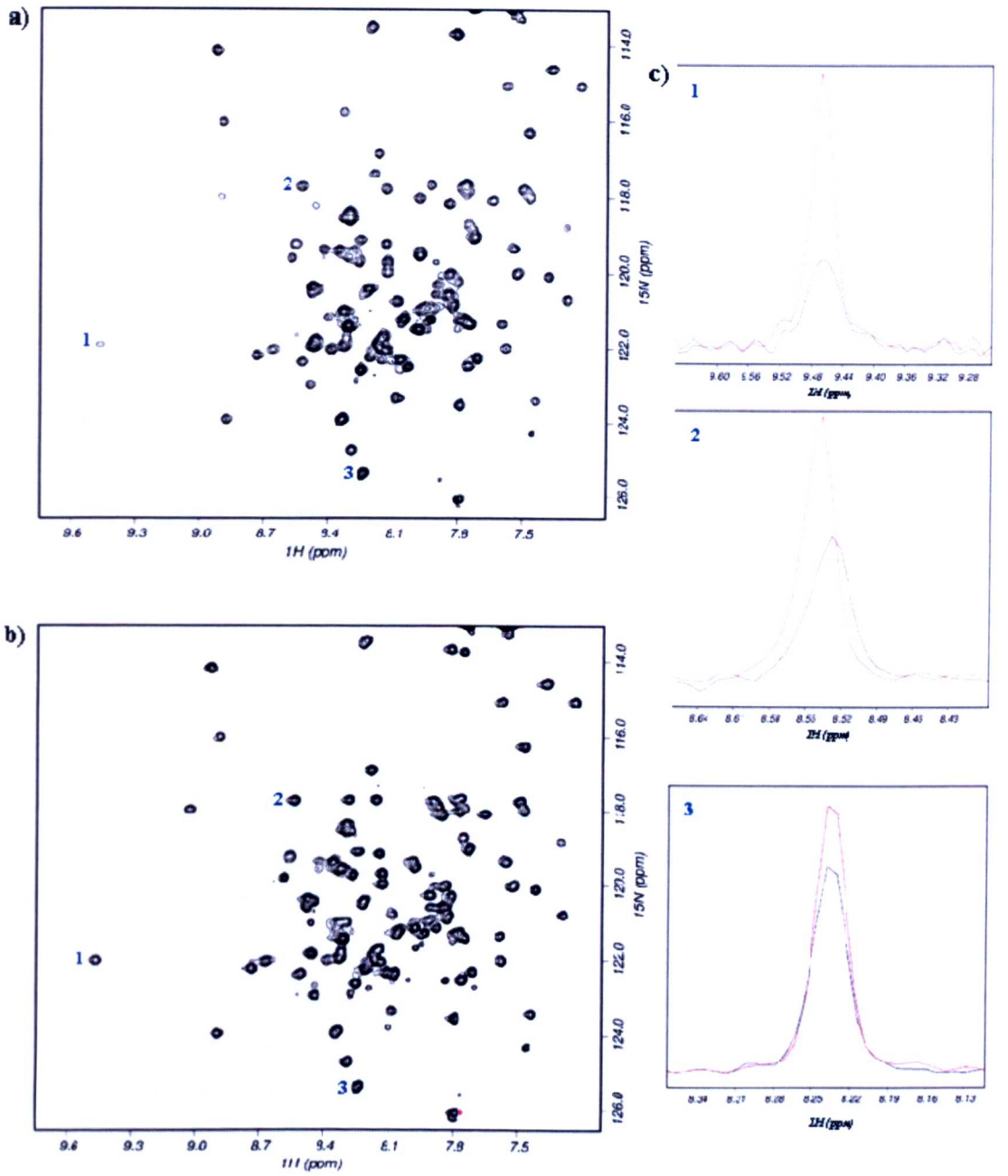


Figure 4.7: Suitability of the His-tag. A region of the ^{1}H ^{15}N HSQC spectrum (a) with and (b) without the His-tag is shown. Resonance signal numbered in blue are shown in (c) as 1D's. Peaks in black are from spectra with the His-tag and in red from spectra without the tag. The protein concentration in both spectra is the same. Spectra are also plotted at the same contour level.

Comparison of the band intensity and position on the SDS PAGE gel revealed insufficient cleavage, as the His-tag had only been cleaved on approximately 50% of the total DnaD-Cd. The position of an intense band observed for the precipitate inferred it was predominately composed of DnaD-Cd. Undoubtedly the poor activity of thrombin and precipitate formed accounted for the poor protein yield at this step of the protocol.

The effect of a number of commonly used components and conditions on the activity of thrombin have been analysed by Novagen. Amongst these, imidazole was one component shown to reduce the activity, as in the presence of this molecule five times more enzyme was required to achieve > 95% cleavage of their test protein. In order to improve the activity of thrombin for removal of the His-tag on DnaD-Cd, imidazole was removed by dialysis before initiating cleavage as described before. On completion of the incubation period this solution appeared less turbid in comparison to the reaction in the presence of imidazole. A significantly smaller protein pellet was obtained when clarifying the solution and subsequently a smaller and less intense band corresponding to DnaD-Cd was observed on the SDS PAGE gel. Besides improving the solubility of DnaD-Cd the removal of imidazole had also enhanced the activity of thrombin, as two times more protein had been cleaved than before. However, despite this a large amount of protein still remained tagged. Therefore, to further improve the efficiency of cleavage the thrombin concentration and incubation times were optimised. DnaD-Cd was incubated with 0.5 units, 1 unit or 1.5 units of thrombin per milligram of protein, at 25°C for 24 hours while shaking. During the incubation period samples were extracted every hour for 10 hours and then at 24 hours. SDS PAGE analysis of these samples showed the amount of His-tag cleaved protein to increase with time at all thrombin concentrations, however, more protein was cleaved at the higher concentrations at any point in time. Adequate removal of the His-tag from DnaD-Cd was only obtained with 1.5 units of thrombin per mg of protein at the maximum incubation time of 24 hours (Figure 4.8).

4.2.3.3 Problem encountered with ammonium sulphate precipitation

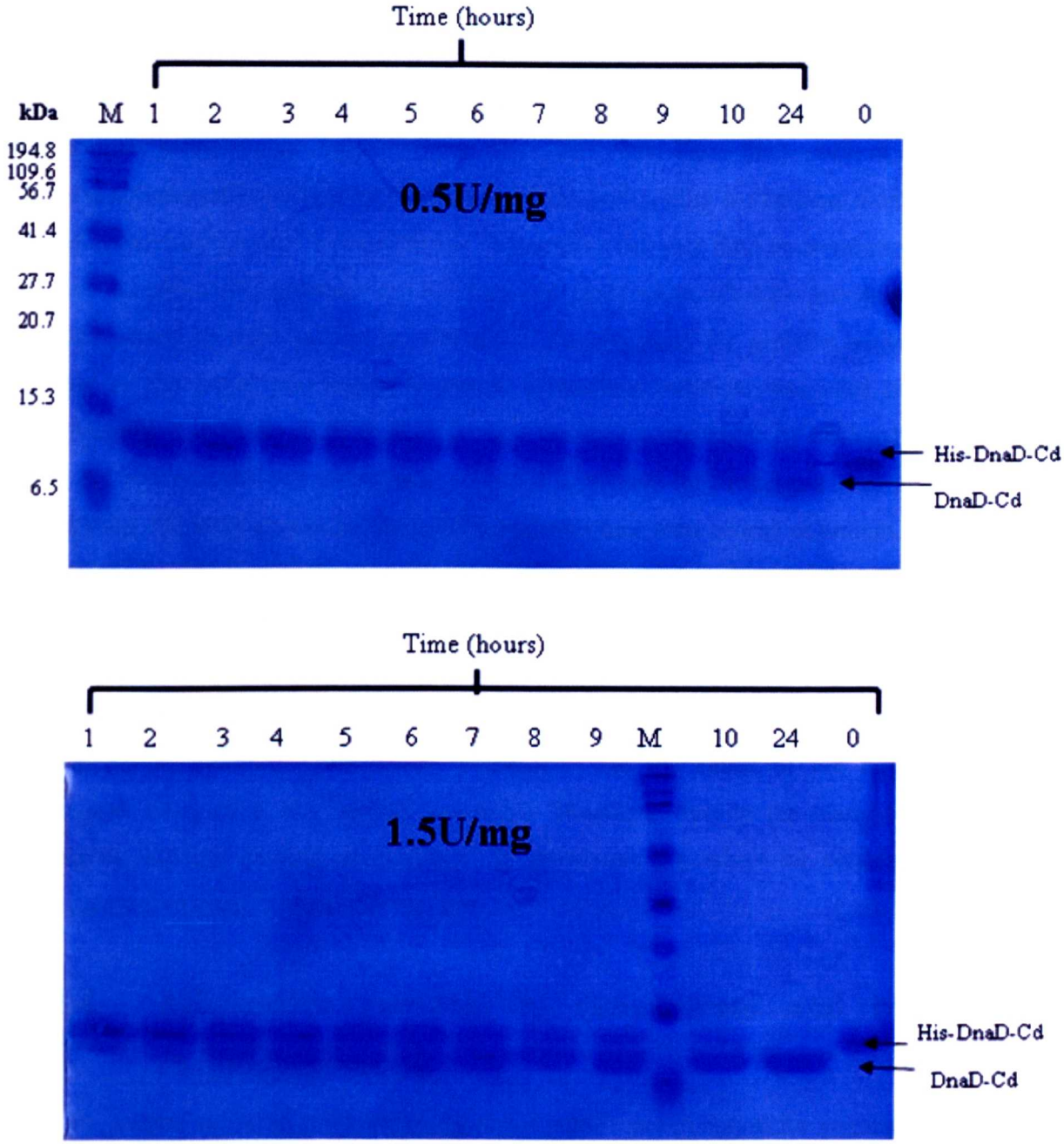


Figure 4.8: Optimising thrombin cleavage of the His-tag. SDS PAGE gels showing cleavage of the His-tag from DnaD-Cd using 0.5U/mg or 1.5U/mg of thrombin. Samples were extracted at the time indicated. Mw markers are shown in Lane M and, are labelled in the top image.

Ammonium sulphate is commonly used to precipitate protein at cold temperatures without denaturation. It serves as a useful method for rapidly concentrating dilute solutions and exchanging buffers. In early stages of purification this method can be used to fractionate proteins since larger proteins precipitate first. Thus, the protein of interest can be simultaneously purified and concentrated by establishing the optimal ammonium sulphate concentration for its precipitation. Also, since salt at a high concentration is an effective bacteriostatic, proteins can be stored as ammonium sulphate precipitates for extended period of time, which is particularly advantageous for proteins sensitive to freezing. In the initial purification protocol, ammonium sulphate was used twice to precipitate solutions containing DnaD-Cd; once, to remove imidazole by exchanging the buffer and a second time to concentrate a solution of protein for further purification with gel filtration. On both occasions the precipitated solutions were stored overnight. When required the protein precipitate was collected by centrifugation and then redissolved into the desired volumes and buffer. Disappointingly, it was never possible to completely solubilise the protein pellet, some aggregates were always visible. SDS PAGE analysis revealed that these aggregates were mainly composed of DnaD-Cd, which implies that the inability to redissolve the protein precipitate was the cause of the large loss experienced every time this method was used. Therefore, despite the many benefits ammonium sulphate precipitation offered, an alternative procedure for concentrating protein and exchanging buffers was required.

4.2.3.4 Alternative methods to concentrate and buffer exchange proteins

There are several methods that completely avoid protein precipitation for concentrating solution or exchanging buffers. Amongst these ultrafiltration is one that can be used to both concentrate and buffer exchange protein. In ultrafiltration, the protein solution is placed in a chamber above a membrane that is fitted in the ultrafiltration device and has a molecular weight cut-off smaller than the protein of interest. An external force such as centrifugal or suction pressure is then applied which forces the solution through the membrane, retaining particles larger than the molecular weigh cut-off on the surface. As the volume of the solution in the chamber decreases the protein concentration

subsequently increases. Once concentrated, the protein can then be buffer exchanged by repeatedly filling the chamber with the desired buffer and concentrating, until the level of exchange needed has been achieved. Gas pressurised cells and centrifugal concentrators are two rapid and gentle ultrafiltration devices commonly used, that differ by the external force required to filter molecules through the membrane. In the gas pressurised cell, high pressure exerted by compressed nitrogen forces molecules through the membrane, whereas centrifugal concentrators use centrifugal force generated overtime during centrifugation. Fouling of the membrane is a problem commonly encountered with ultracentrifugation devices designed with horizontal membrane, when solutions laden with particulates are forced towards the bottom. One centrifugal concentrator in particular, called the VIVA spin, has been designed with a twin vertical membrane and a thin channel filtration chamber to minimise this problem. When centrifugal force is applied to the VIVA spin, particulates are pulled away from the membrane into a 50 µl concentrate pocket at the bottom of the device. In this way aggregates are prevented from settling along the entire length on the membrane thereby allowing solutions to be rapidly concentrated even when laden with particulate.

Another gentle way of concentrating proteins that avoids precipitation is evaporation. For evaporation, the protein solution is exposed to a flow of compressed air in order to gradually eliminate water and thereby reduce the volume. In addition to concentrating protein, evaporation of the solution also increases the concentration of buffer components. Therefore before reducing the volume it is necessary to transfer the protein into a buffer whose components will be at the desired concentration in the final volume.

With the intention of finding an alterative method to ammonium sulphate precipitation for concentrating solutions containing DnaD-Cd, the suitability of evaporation and ultrafiltration was tested. A sample of pure DnaD-Cd was split into three aliquots. One aliquot of protein was concentrated in a gas pressurized cell purchased from amicon and another with a VIVA spin centrifugal concentrator. The last aliquot was dialyzed to reduce the salt and phosphate concentration, and then exposed to a fast flow of compressed air, to concentrate the protein by evaporation. Changes in solution turbidity

due to protein aggregation were monitored as the samples were reduced to approx. 1 ml and the size of the protein pellets collected on clarification of these solutions were compared. Out of the three procedures trialed, reducing the volume by evaporation showed the greatest change in turbidity and also gave the largest pellet on clarification. Quantification of the amount of protein by absorbance at 260 nm and comparison of signal intensity in ^{1}H 1D spectra, showed that 65% of the protein had been lost when concentrated by evaporation. Samples concentrated in the gas pressurized cell and VIVA spins showed much smaller losses. Under the chosen conditions, ultrafiltration was clearly a more suitable method for concentrating DnaD-Cd than evaporation. Since the VIVA spin was much faster at concentrating the protein than the gas pressurized cell, for reasons discussed above, it was the device chosen for concentrating DnaD-Cd.

The VIVA spin could also be used to buffer exchange proteins, therefore its suitability for buffer exchanging DnaD-Cd was tested. A concentrated sample of pure protein was repeatedly diluted, re-concentrated and quantified once aggregates had been removed. Comparison of the amount of protein showed the loss in a single round of buffer exchange to be twice that of concentrating alone. This further increased every time the protein was diluted and concentrated. In contrast, dialysis had shown to be effective in removing small molecules, without substantial loss of protein, during refolding and thrombin treatment. Given that the protein solutions were dilute when such a result was obtained; it was assumed that concentrating repeatedly rather than the device itself was responsible for the poor recovery of protein when buffer exchanging with the VIVA spin. This problem of protein loss through aggregate formation when concentrating had to be addressed to make the VIVA spin a sufficient method for buffer exchanging DnaD-Cd. As it was unknown if this was possible, dialysis was used to remove imidazole during purification while strategies to improve solubility of DnaD-Cd were looked into.

4.2.3.5 Improving solubility of DnaD-Cd

The extreme sensitivity of proteins to solution conditions often makes it difficult to concentrate them up to the necessary level without causing aggregation or precipitation,

when conditions are not optimal. To circumvent this problem a number of buffer additives and cosolvents that may improve solubility and stability of proteins have been identified (Bondos and Bicknell, 2003). However, the list of these possible additives is long thus it is not easy to determine which cosolvent at what concentration would be effective on a particular protein. Therefore, knowledge about the function and behavior of the protein or a member of its family can be useful for selecting additives that are likely to succeed. For novel proteins, where such information is unknown, the general approach taken is to systematically screen various additives at concentrations within the range suggested in the literature.

Several studies on DnaD have shown that the C-terminal domain binds DNA (Turner *et al.*, 2004; Carneiro *et al.*, 2006). It has repeatedly been demonstrated that nucleic acid binding proteins often require high concentrations of salt to increase solubility and prevent aggregation (Bondos and Bicknell, 2003). Therefore, in an attempt to increase solubility of DnaD-Cd, NaCl was selected as the first additive to trial. Three dilute protein samples in 100 mM sodium phosphate buffer were concentrated in the presence of 50 mM, 0.2 M or 0.5 M NaCl. Determination of protein concentration at approx. 1 ml revealed samples to be more concentrated and consist of fewer aggregates at higher levels of NaCl. Therefore, the NaCl concentration was increased from 0.2 M to 0.5 M in all purification buffers.

It was also known that DnaD-Cd contains a single cysteine, however its spatial location within the protein remained unidentified without a structure. If this cysteine was position on the surface of the protein it may contribute to the problem of aggregation, when involved in forming non-native inter molecular interaction. In which case, supplementing the buffer with the additive DTT may prevent such bond forming and thereby significantly reduce aggregation. With the assumption that the cystine was on the protein surface, the effect of DTT on DnaD-Cd was tested. Dilute samples of DnaD-Cd were incubated with 1 mM, 3 mM, 5 mM, 10 mM and 15 mM DTT for approx. 10 min and then concentrated to give a final volume of 750 µl. Comparison of sample turbidity as the volume was reduced suggested DTT was effective in reducing

aggregation since the clarity of samples at all concentration of DTT had improved. Determination of protein concentration at the final volume confirmed that samples concentrate better in the presence of DTT.

Amongst the extensive list of possible additives, a single set of amino acids namely L-Arg and L-Glu had been reported to increase solubility and long term stability of several proteins, when added simultaneously to dilute samples at a concentration of 50 mM (Golovanov *et al.*, 2004). They have also been shown to prevent degradation and precipitation of samples over time. The precise way in which these amino acids act to improve solubility is unknown, however it has been speculated that both L-Arg and L-Glu bind to and mask oppositely charged residues on the surface of the protein, while the aliphatic side chains of these residues cover the exposed hydrophobic regions adjacent to the charged residues they bind. HSQC spectra of several proteins tested by Golovanov and coworkers have shown the chemical shifts and line widths of resonance signals remain unaffected in the presence of these amino acids, suggesting that the protein retains its structure. Also, the 50 times more intense signals observed for Arg and Glu do not interfere with the resonance assignment process because unlabeled signals are filtered out when multidimentional spectra are acquired with ^{15}N and ^{13}C labeled samples.

The reported success of this single set of additive in situations identical to DnaD-Cd, where solubility become a problem when concentrating samples up to a high level, made it an attractive additive to trial. Thus, dilute samples of DnaD-Cd were concentrated in the presence or absence of 50 mM L-Arg and L-Glu to a volume of approximately 1 ml. The amount of precipitate observed while the volume reduced was compared and at the final volume the concentration determined. For samples concentrated in the presence of L-Arg and L-Glu the precipitate formed and concentration determined was equivalent to samples concentrated in the absence of these additives. Clearly, L-Arg and L-Glu had not improved the solubility of DnaD-Cd. It was not possible to check these samples by NMR as the protein precipitated when 10% D_20 was added.

4.2.4 Summary

The initial expression and purification protocol for DnaD-Cd gave a single 500 µl sample at 0.6 mM (A total of 4 mg of protein) from 4 L of culture. Following the optimization of expression and purification, eight 500 µl samples of DnaD-Cd at 0.6 mM were purified from a 1 L culture (A total of 32 mg of protein). The major change to the protocol that enabled such a significant improvement in protein yield was introducing a solubilisation and refolding step for extracting insoluble protein from inclusion bodies. Optimizing the His-tag Cleavage stage and replacing ammonium sulphate precipitation, which was used for concentrating and buffer exchanging proteins, with ultracentrifugation and dialysis also contributed to the improved yield. Protein aggregation is always an issue with DnaD-Cd, particularly when concentrating samples. Increasing the NaCl concentration in all purification buffers but also adding the buffer additive DTT reduced aggregation and thus also contributed to an increased protein yield. Overall, the new protein expression and purification protocol enabled enough protein to be purified making labeling with ^{13}C and ^{15}N feasible. An overview of the new protein expression and purification protocol is shown in Figure 4.9.

4.3 Restoration of inconsistent HSQC spectra

When NMR samples of refolded DnaD-Cd were made up from different protein preparations, the quality of HSQC spectra varied. In the spectra that were of the poorest quality many resonance signals had shifted, the intensity of these had reduced and line widths increased. Also, a few signals were completely obliterated. These changes observed were indicative of transient self association. The least intense and completely obliterated signals usually correspond to residues in or around the interacting region. Chemical shift changes of only a few resonance signals generally indicate minor conformation changes. The reduction in signal intensity and an increase in the line width over the entire spectrum are characteristic of larger proteins and reflect the slower tumbling of the molecule as it forms larger structures. Since all sample of DnaD-Cd that were prepared following the optimised expression and purification protocol were treated from expression to preparation of the NMR sample in an identical manner, handling of

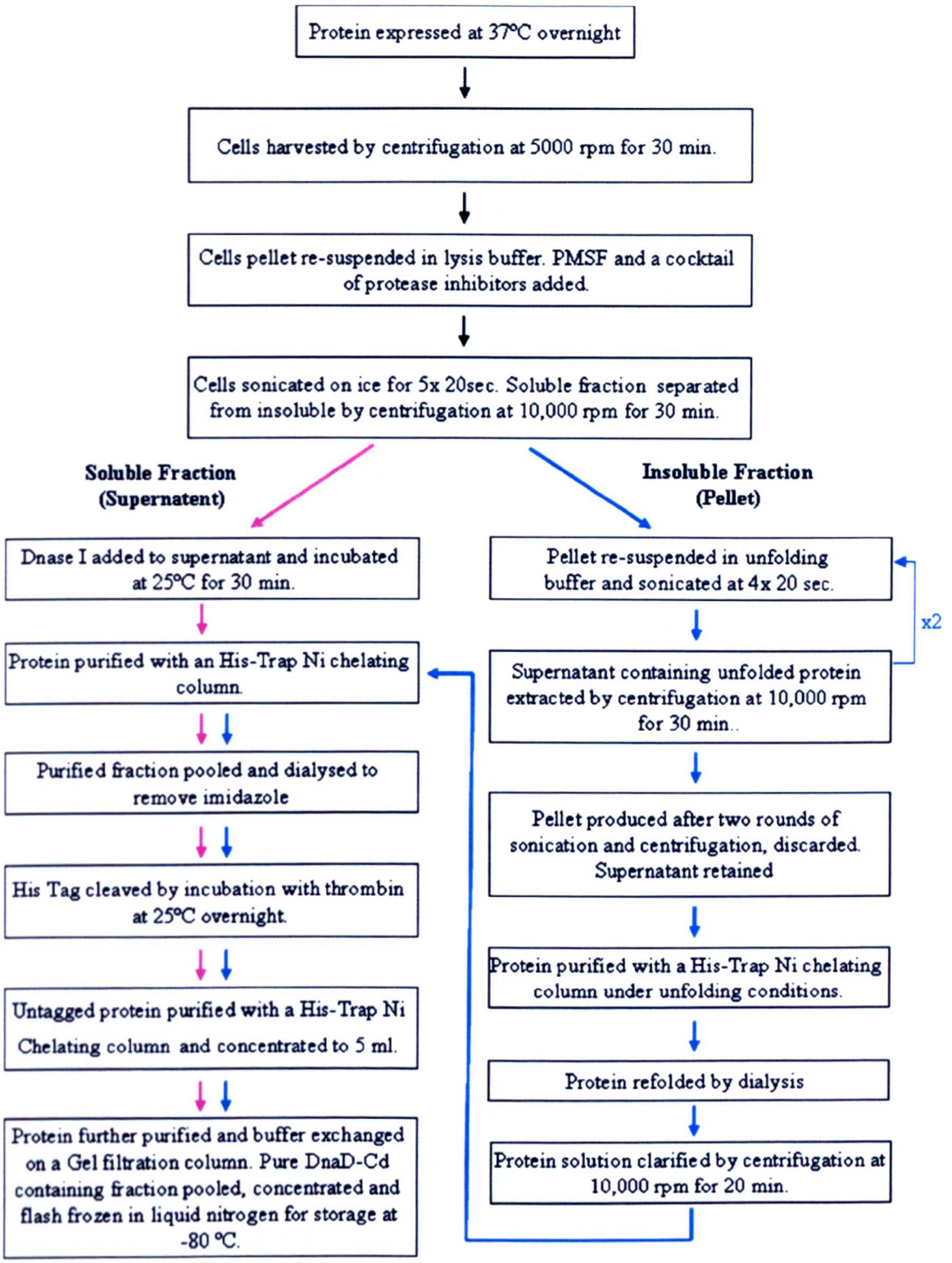


Figure 4.9: Flow chart summarising the optimised protein expression and purification protocol.

For a detailed description of each step see section 2.15.

the protein could not be responsible for the discrepancies observed in spectra. However, the final buffer consisted of 5 mM DTT; the half life of this reducing agent is highly pH dependent, at pH 6.5 it has an approximate half-life of 40 hr when stored at 20°C. Once prepared, NMR sample were stored at 4°C until required and all spectra were acquired at 25°C. If the levels of reduced DTT had dropped significantly during the storage period, protein self association would be likely. Since levels of reduced and oxidised form of DTT could be easily monitored with a 1D spectrum at 4 ppm and 5 ppm respectively, the peak corresponding to reduced DTT was compared from different preparations. As expected, samples that gave the poorest quality HSQC spectra showed significantly lower levels of DTT. To establish whether these spectra could be restored by increasing levels of this reducing agent, samples were systematically treated with concentrations of DTT ranging from 1 mM to 10 mM, and then HSQC spectra were acquired. On addition of DTT at concentrations above 5 mM, the resolution and sensitivity of spectra had markedly improved. Shifted peaks had retuned to their monomeric chemical shift positions. Line widths of peaks and intensities of all signals were now comparable to spectra of good quality. Changes described are shown in Figure 4.10 for selected signals. These observations suggest that DTT had caused a marked reduction in self association, beneficial to the quality of NMR spectra. Thus, to avoid such inconsistencies in future spectra, all DnaD-Cd samples were treated with 5 mM DTT prior to acquisition. The equilibrium of oxidised: reduced DTT was monitored before and after the triple resonance experiments were record for structure determination.

4.4 Optimisation of NMR Sample conditions

^{15}N labeled samples of DnaD-Cd were analyzed under a series of different buffer conditions, to ensure the best signal dispersion and uniformity was achieved in HSQC spectra. First, the pH of a 0.5 mM sample with solvent conditions of 100 mM phosphate (pH 7.4), 50 mM salt and 5 mM DTT, was systematically reduced. At each pH, HSQC spectra were acquired at 298K and the signal intensity and appearance of cross peaks assessed. On decreasing the pH by one unit to pH 6.5, a marginal increase in signal intensity and minute change in the number of cross peaks was observed. However, a reduction to pH 6 resulted in sample precipitation. Testing the effect of a pH above 7.5

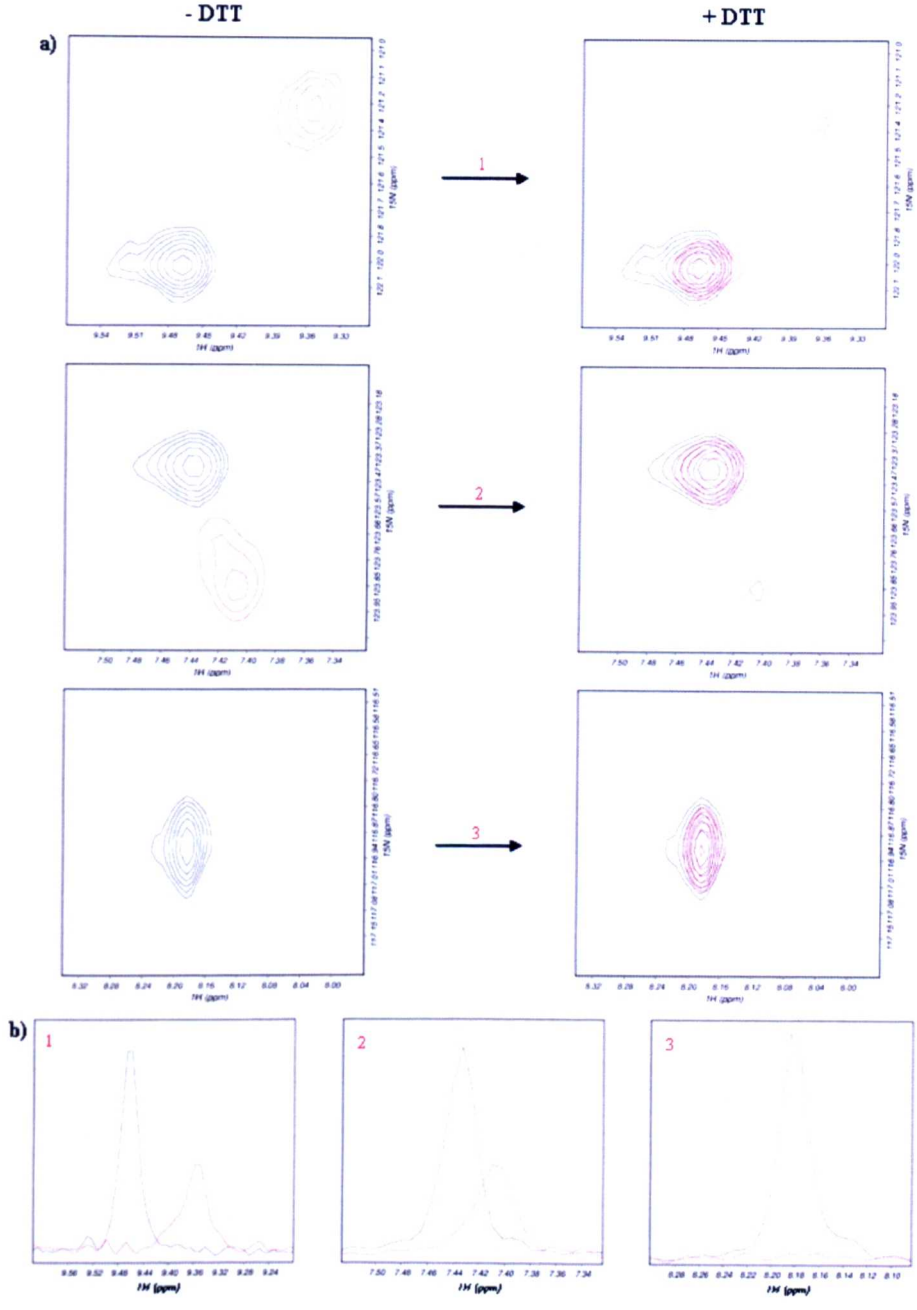


Figure 4.10: Peaks from ^{1}H-^{15}N HSQC spectra restored on the addition of DTT. (a) In black are signals from protein purified from the soluble fraction. Overlaid in red are resonance signals of refolded protein in the presence or absence of 5 mM DTT. The changes in signal intensity on the addition of DTT are shown in (b) as 1D's. Peaks in red are from spectra in the absence of DTT and in black from spectra with DTT.

was not necessary because NMR experiments for structure determination are generally preformed at a pH below 7, to reduce the loss of data through proton solvent exchange (Reid, 1997). Therefore, a pH of 6.5 was selected for the final NMR buffer.

Additional HSQC spectra were then acquired at 298K, 303K, 308K, 313K, 318K and 323K, with a second sample consisting of 100 mM phosphate (pH 6.5), 50 mM salt and 5 mM DTT. At temperatures between 298K and 308K, a large difference in signal intensity, dispersion or uniformity was not observed. However, above 308K, spectra showed characteristics of a partly unfolded protein and, thus, the lowest temperature was chosen. Another sample, identical to the second, was then prepared to establish the lowest NaCl concentration that would maintain the solubility of DnaD-Cd, because high levels have been shown to compromise the sensitivity of the cryogenic probes (Kelly *et al.*, 2002) and thereby affect spectra quality. Since NaCl concentrations within the range 100 - 150 mM are typically used as compromise between these factors, HSQC spectra were collected in the presence of 50 mM, 150 mM and 200 mM NaCl at 298K. Comparison of these spectra showed signals to be a fraction more intense and better resolved at 150 mM NaCl than at 50 mM NaCl (Figure 4.11). No further improvement was observed on increasing the salt concentration to 200 mM. Therefore, a concentration of 150 mM NaCl was selected for NMR samples. The final step of the purification protocol was modified, so that pure DnaD-Cd was eluted from the gel filtration column in the desired NMR buffer of 100 mM phosphate (pH 6.5), 150 mM NaCl and 5 mM DTT (see section 2.15.10).

Finally, the highest concentration at which the majority of DnaD-Cd remained monomeric and gave spectra with good sensitivity and resolution was established. A 1 mM sample of DnaD-Cd in 100 mM phosphate (pH 6.5), 150 mM NaCl and 5 mM DTT was systematically diluted to give a protein concentration of 0.8 mM, 0.5 mM or 0.2 mM. At each concentration, HSQC spectra were acquired and the signal intensity and line widths compared. Protein at 0.5 mM gave rise to signals that were more intense than at a concentration of 0.2 mM, however the gain in signal intensity observed was not proportional to the increase in concentration (Figure 4.12). Also, the line widths of many

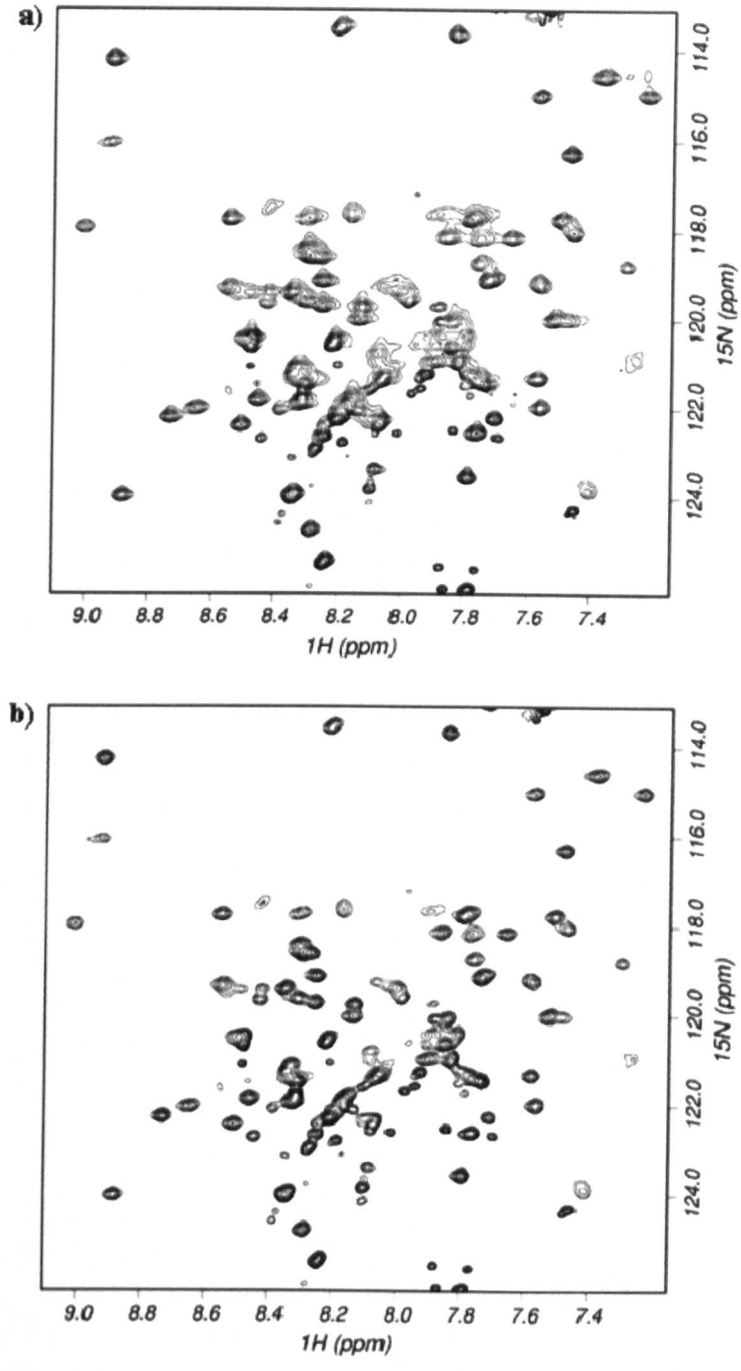


Figure 4.11: Optimization of NaCl concentrations. Increasing salt concentration from 50 mM (a) to 150 mM (b) generates peaks with higher signal intensity and better resolution. All other components in the sample used to acquire these spectra were identical. Both spectra are processed and displayed identically and thus are directly comparable.

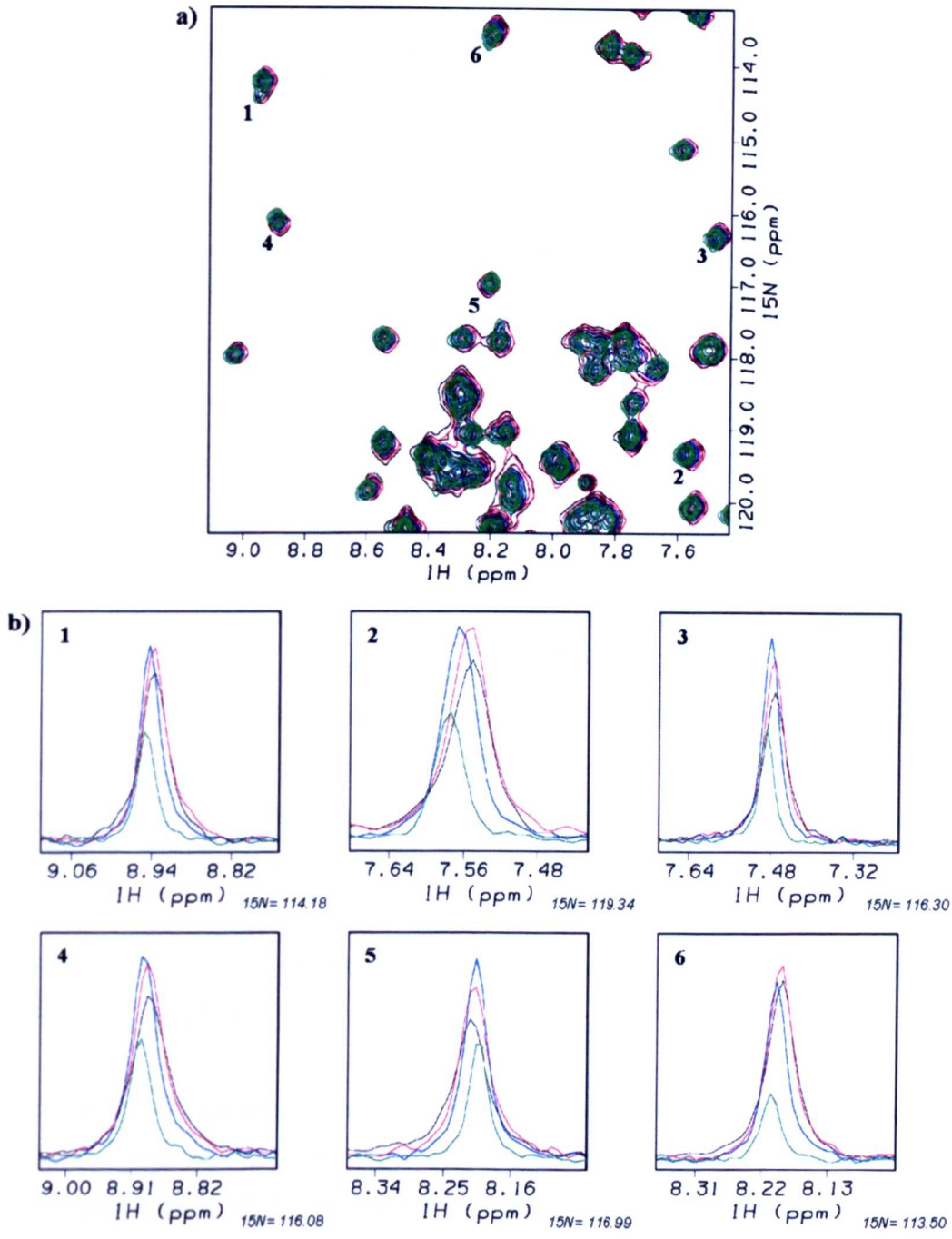


Figure 4.12: Optimisation of protein concentration. (a) A region of ^{1}H-^{15}N HSQC spectra of DnaD-Cd at various protein concentrations tested overlaid. (b) Resonance signals numbered in image a are shown as 1D peaks in b. Signals from protein at a concentration of 1 mM, 0.8 mM, 0.5 mM and 0.2 mM are in black, red, blue and green respectively.

peaks in this spectrum were broader, indicating that a portion of the protein had aggregated. At 0.8 mM signals became less intense and even broader compared to the spectrum at 0.5 mM. However, the least intense signals with the largest line widths were observed in spectra of protein at 1 mM. These observations at the higher concentration of DnaD-Cd suggested that the increased proximity of molecules at such concentration increases their propensity to aggregate and form large assemblies that are undetectable by NMR. As a compromise between signal intensity and aggregation propensity a concentration of 0.5 mM was selected for structure determination.

4.5 Stability of DnaD-Cd over time

The storage of proteins for any length of time can pose stability problems. For long term storage, proteins are generally lyophilised into a dehydrated powder or frozen at -80°C in an appropriate buffer. When stored for short periods, they are usually refrigerated at 4°C with all the components necessary to stabilise the protein, such as reducing agents, additives, antibacterial agents and protease inhibitors. Since problems had been experienced with re-dissolving lyophilised DnaD-Cd, the protein was flash frozen with liquid nitrogen in 1 ml aliquots, for long term storage at -80°C. When a sample was needed for data collection, a single aliquot of protein was thawed and then stored indefinitely at 4°C rather than re-freezing, to avoid stresses exerted by repeated freeze thawing. DTT was added to these samples on a weekly basis to maintain levels of the reduced form needed to prevent protein self association and sodium azide was added to inhibit bacterial growth.

The quality of a sample was monitored over 5 months to establish how long samples could be stored at 4°C before they became unsuitable to use. ^{15}N HSQC spectra were acquired on days 1, 10, 15, 31, 35, 38 and 140 to compare the signal intensity, dispersion and the number of peaks. In spectra collected on day 1, several low intensity signals were observed to partially overlap major peaks dispersed throughout the spectrum (Figure 4.13). These minor signals in spectra collected on day 10 increased in intensity becoming more resolved, while the intensity of the major peaks decreased. The signal dispersion in this spectrum remained unchanged, but additional peaks appeared in the

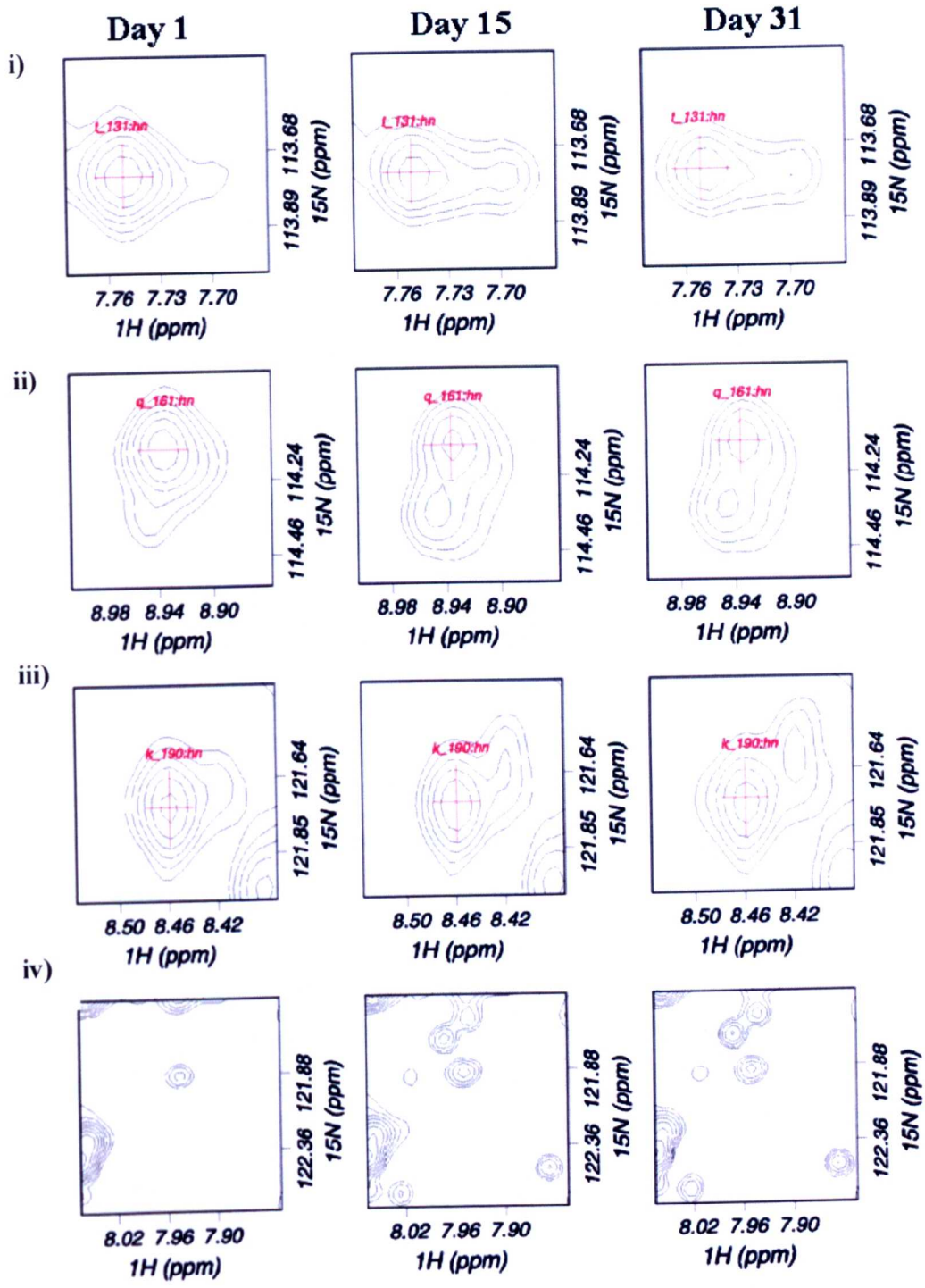


Figure 4.13: Stability of DnaD-Cd. Signals for residue (i) T131, (ii) Q161 and (iii) K190 in ^{15}N HSQC spectra collected on day 1, 15 and 31 are shown. The major peaks are marked with a red cross. (iv) A region of each spectrum that is typical for unfolded peptides is enlarged.

central region and the intensity of crosspeaks previously identified as degradative products had increased, suggesting that the protein had degraded further. This pattern of changes continued in spectra collected on day 15 and 31, however, after day 31, the intensity of both major and minor peaks continuously decreased. In the NMR tube, protein precipitate had become increasingly visible after day 31 and was probably accountable for the loss in intensity observed. Analysis of the sample after day 140 by mass spectrometry confirmed that the protein had degraded, as four long and many short peptides were detected. Figure 4.14 shows the mass spectroscopy trace of DnaD-Cd after day 140. The most intense peak at 13730.2Da corresponded to full length (residues 129-232) ^{15}N and^{13}C labeled DnaD-Cd. The other three less intense peaks of masses 10191.1, 1204.2 and 12848.78 were presumably product of DnaD-Cd degradation. It was likely that the protein was degraded from the C-terminus as removal of an unstructured region would not significantly affect the conformation of the protein and thereby the chemical shift and dispersion of amide resonances would remain almost unchanged. The masses of the less intense peaks in the mass spectrum corresponded nicely with the following polypeptides: Peak of 10191.1 Da may correspond to residue 129-206, peak of 1204.2 Da may correspond to 129-221 and peak of 12848.78 Da may correspond to 129-224. The minor signals observed overlapping the major peaks could come from any one of these polypeptides. DnaD-Cd was clearly being degraded at some point between expression and NMR sample preparation, but since the amount of degraded protein was less than 10% of the total full length protein on day 1 of data collection it was not a major issue. However, the progression of degradation was a problem as the increasing intensity of the minor peaks could result in uncertainty when assigning resonances and thus hinder structure determination. In order to avoid such problems samples more than 10 days old were not used to collect spectra for the resonance assignment procedure.

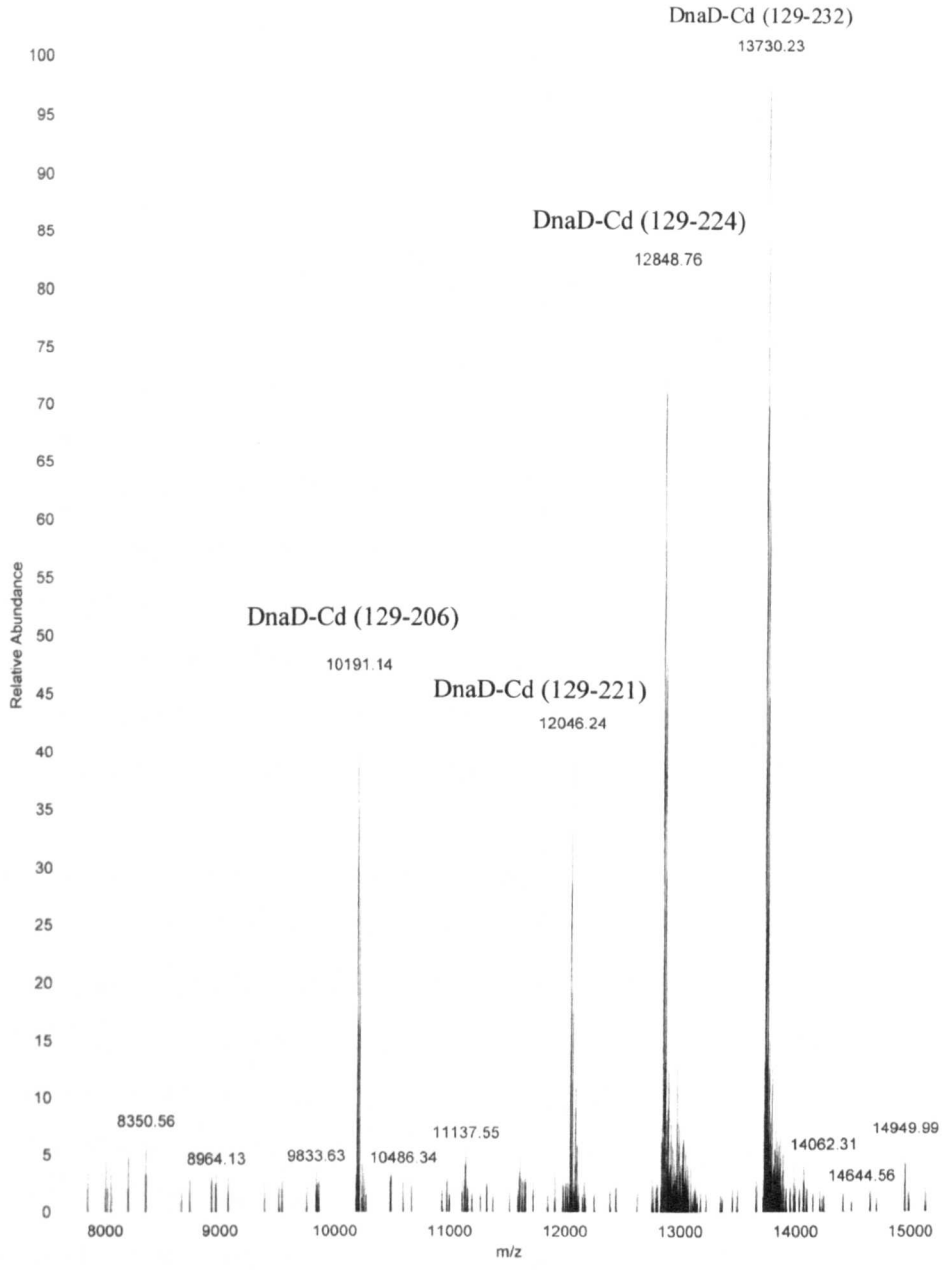


Figure 4.14: Mass spectrum of ^{13}C-^{15}N DnaD-Cd recorded under denaturing conditions. This sample contains mainly full length DnaD-Cd (129-232), however smaller amounts of shorter DnaD-Cd polypeptides are also present as indicated (Data provided by William Grainger, Nottingham University).

Chapter 5: Resonance assignment and preliminary NMR structure determination of DnaD-Cd

The general strategy for the determination of structures in solution by NMR consists of a number of steps, starting with assigning the different NMR resonance signals observed in spectra to particular nuclei in the molecule, a process known as resonance assignment. This is followed by collecting 3D information in the form of restraints on torsion angles and distances that link different atoms within a residue or most importantly between different residues in the polypeptide chain. Once enough restraints are obtained, computer programs such as ARIA are used to derive structures that satisfy as many of the restraints as possible, in addition to general properties of proteins such as bond lengths and angles. The end result of such computational supported structure calculation procedures is an ensemble of structures that represent the protein.

This chapter describes the process by which the preliminary ensemble of NMR structures for DnaD-Cd was determined. First the task of resonance assignment which was divided into backbone and side-chain assignment is detailed, and then the procedure for deriving 3D restraints and performing structure calculations is described. The preliminary DnaD-Cd ensemble is also presented. It was not possible to derive an ensemble of structures that were of 'high resolution' because enough NOEs could not be assigned in NOESY spectra to use as distance restraints in structure calculation. Besides poor sensitivity and spectral overlap, cross peaks corresponding to a minor conformation of DnaD-Cd resulting from degradation of the protein at its C-terminus were most likely responsible for the unresolvable ambiguities in the identification of both sequential and long range NOEs. In order to calculate an improved ensemble of structures representing DnaD-Cd, the entire resonance assignment procedure would need to be repeated with a version of the protein that gives better spectra, i.e. a construct without the C-terminal unstructured region. Given that the structure, 2zc2 which shows sequence homology to DnaD-Cd appeared in the protein data bank midway through this project, a homology

model was constructed for DnaD-Cd based on the 2zc2 structure and checked against the NMR data. This chapter describes the construction of this homology model but also, discusses residue conservation and charge distribution across DnaD-Cd, which point to a region of the protein that may potentially be a DNA binding site.

5.1 Resonance Assignment

5.1.1 Backbone assignment

A ^{15}N HSQC spectrum and five 3D triple resonance (^{1}H, ^{15}N, ^{13}C) spectra were used for the assignment of backbone HN, NH, CO, Cα and Cβ resonances. First, the 161 NMR resonance signals observed in the ^{1}H ^{15}N HSQC spectrum shown in Figure 5.0 were systematically assigned a number, to define the NH peak recorded until residues they corresponded to were identified. Then the chemical shift positions of each NH peak were analyzed in the five 3D triple resonance spectra, namely HNCO, HNCACO, HNCA, CBCACONH, and HNCACB to assign chemical shift values to their corresponding Co_{i-1}, Co, $C\alpha_{i-1}$, Cα, $C\beta_{i-1}$ and Cβ groups. Nuclei correlated in each spectrum and acquisition parameters are detailed in section 3.2.1. An example of slices from the triple resonance spectra are shown in Figure 5.1. The group of nuclei associated with a particular amide is referred to as a spin system and therefore the number assigned to that amide is the spin system number.

Signals picked in the ^{15}N HSQC spectrum that corresponded to amide side-chains or to products of degradation were eliminated from the backbone assignment process. Cross peaks corresponding to side-chain amides were easily identified as signals for these peaks were not observed in the five triple resonance backbone assignment spectra. Also, the side-chain amides of asparagines and glutamine usually give rise to two resonance signals with the same nitrogen but different proton chemical shifts and the side-chain amide of tryptophan generally has a chemical shift around 10 ppm. Signals corresponding to the small polypeptide products of degradation are usually of low intensity and mostly located at a ^{1}H chemical shift of around 8 ppm, which is typical for random coil shifts. Cross peaks corresponding to minor forms of the protein were also of

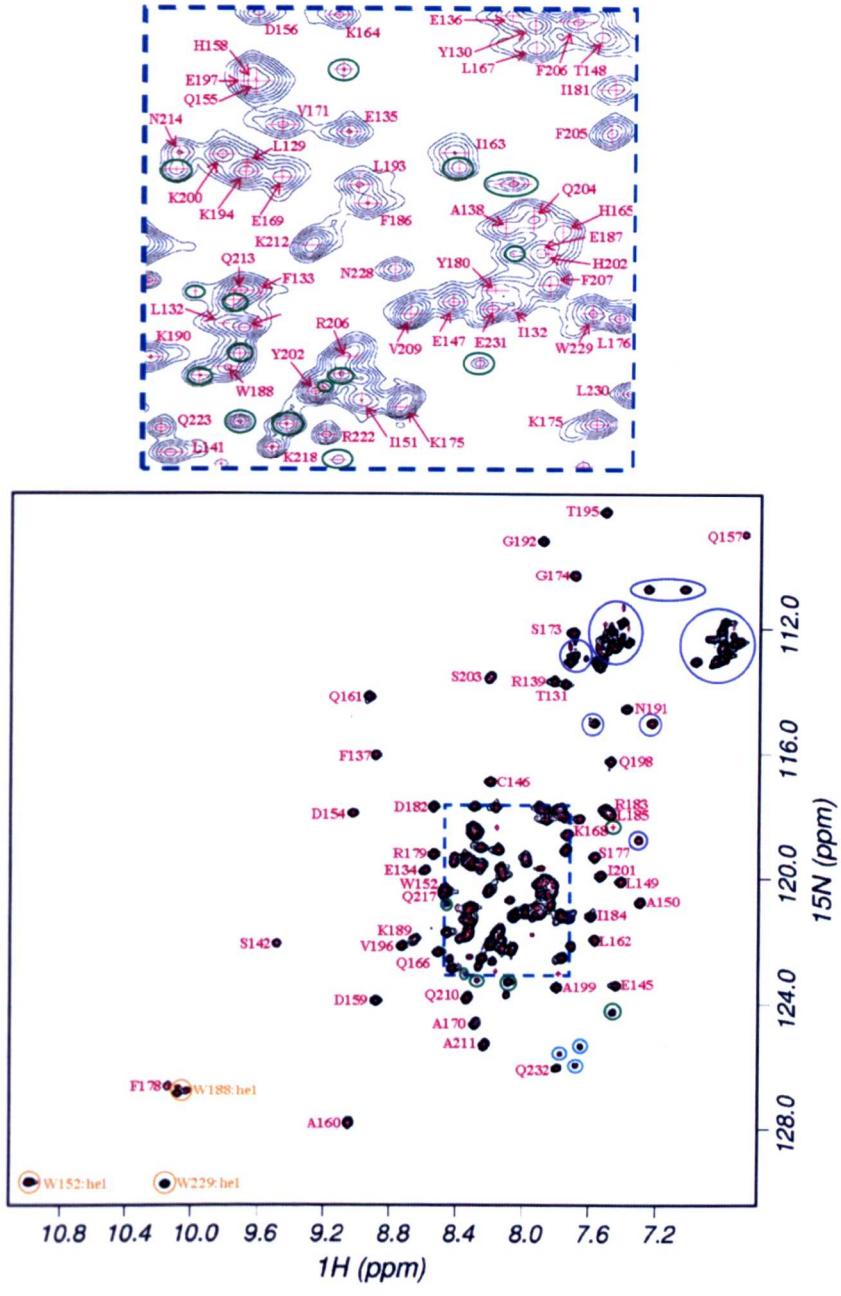


Figure 5.0: ^{1}H-^{15}N HSQC backbone assignment spectra of DnaD-Cd. Resonance signals that were given a spin system number are highlighted with a red cross. Signals from side-chains are circled in purple. Signals from degradative products and peaks not assigned to a specific amide are circled in green. Assignments from backbone experiments are shown in red and side-chain signals deduced from the NOESY spectra in orange. Signals circled in cyan possibly result from the C-terminal end. The dashed line in blue highlights a region of the spectrum that has been enlarged and shown above the main spectrum.

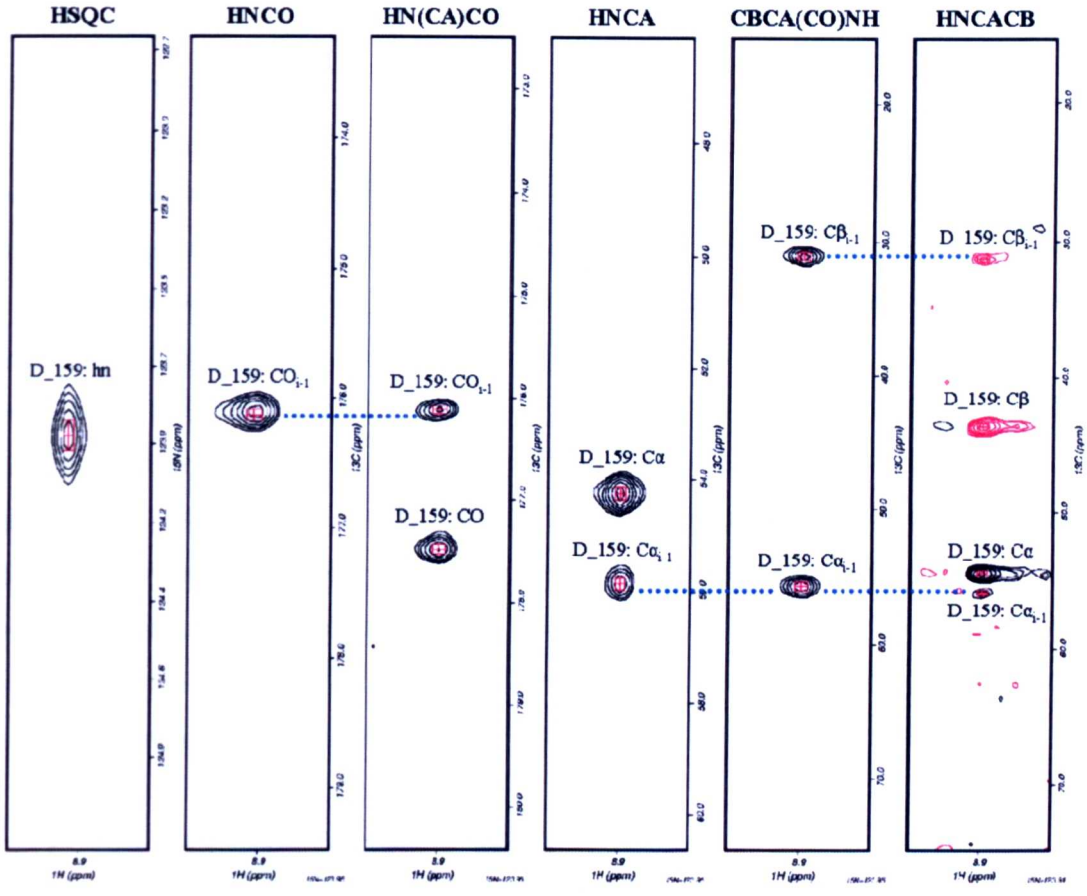


Figure 5.1: An example of slices from the ^{15}N HSQC spectrum and the five 3D spectra used for resonance assignment. The slices for spin system 137 are displayed in the order they were examined. Resonances are labeled with assignment deduced by Asstool, and signals in different spectra that are assigned to the same nucleus are highlighted with a blue dashed line.

low intensity, but positioned either very close to or partially overlapping their corresponding major peak. In total, 53 spin systems were removed from the backbone assignment process.

Backbone assignment was performed with the remaining 108 spin systems by chemical shift matching with the program asstool. Section 3.2.2 describes backbone assignment by asstool in detail. Asstool was configured to complete thirty iterations for each assignment attempt. After each attempt the output was checked to confirm that the program had correctly matched chemical shifts of 'self' with 'preceding' residues. The spin systems that were known to be assigned to a high degree of certainty were fixed in the round of assignments that followed. Triple resonance spectra were repeatedly referred to for confirming chemical shifts assigned or to identify missing peaks. By the final run of asstool, backbone assignment was successfully completed for all non-prolyl residues except for L144, F226, Y227 and residues from the His-tag that remained after thrombin cleavage, as these were matched with more than two spin systems.

5.1.2 Side-chain assignment

For DnaD-Cd residues 129 to 232, 604 side-chain atoms were assigned out of the 883 side-chain atoms that should have been assigned. As a starting point for side-chain assignment, Cα and Cβ chemical shifts derived from the backbone assignment were used to locate strips of resonance signals, at their chemical shift values in HBHACONH, HCCH and CCH TOSCY spectra. Each amino acid type was then taken in turn and assigned based on the characteristic patterns of chemical shifts. The Hα and Hβ side-chains were assigned mainly from the HBHACONH spectra, when this was not possible the CCH and HCCH TOCSY spectra were used. All other side-chains except the aromatic ring of phenylalanine, tyrosine and tryptophan were also assigned from TOCSY spectra. Resonances for neighbouring groups of the aromatic rings were looked for in ^{15}N edited NOESY, and an aromatic ^{13}C HSQC spectrum was used to locate signals that possibly corresponded to aromatic side chains in the ^{15}N NOESY spectra. Many resonances that remained unassigned were heavily overlapped. For experimental details and spectra parameters see section 3.3.

5.2 Preliminary NMR structure determination

5.2.1 Dihedral restraints

Backbone phi and psi torsion angles were predicted by TALOS from N, CO, Hα, Cα, and Cβ chemical shifts data (Cornilescu *et al.*, 1999). This program searches a database constructed using the well defined parts of X-ray crystallographic structures of 2.2Å resolution. It compares their deposited secondary shifts with the N, CO, Hα, Cα, and C_β chemical shifts assigned for a given protein. Data from three consecutive residues is used simultaneously to select the 10 best matches for phi and psi from the database, for the central residue. Matches that indicate consistent values for phi and psi and lie in allowed regions of the Ramachandran plot are classified as 'good'. Their average and standard deviation are used as predictions. When matches are inconsistent, no predictions are made for the central residues and subsequently declared 'ambiguous'. Out of the 97 predictions made for DnaD-Cd, 58 were classified as 'good'. These 'good' restraints were introduced into structure calculations. Dihedral angle plots for residues 127-231 are shown in Figure 5.2. These plots indicate the presence of five helices: I, L129-A138; II, L145-D157; III, A161-S173; IV, F179-N192 and V, V197-F206. Residues after V206 have plots typical for random coil.

5.2.2 NOESY assignment

NOE peaks in the ^{15}N and ^{13}C edited NOESY spectra were assigned to derive distance restraints for structure calculations. Spectral parameters are detailed in section 3.4. To begin with intra-residue peaks for each residue were assigned by chemical shift matching. Then short to medium range NOEs that were characteristic to helices were assigned. These included NOEs between adjacent amide protons d_{NN} (i,i+1) and NOEs between the HN and Hα proton in the third consecutive residue $d_{\alpha N}$ (i,i+3). In some cases NOEs between the HN and Hβ $d_{\beta N}$ (i, i+1) of the neighbouring residue were assigned. Last of all, unambiguous long range NOEs were assigned to define the overall fold. This gave a total of 450 intra- and inter-residue NOEs to introduce in structure calculation. An example of an NOE assignment strip is shown in Figure 5.3.

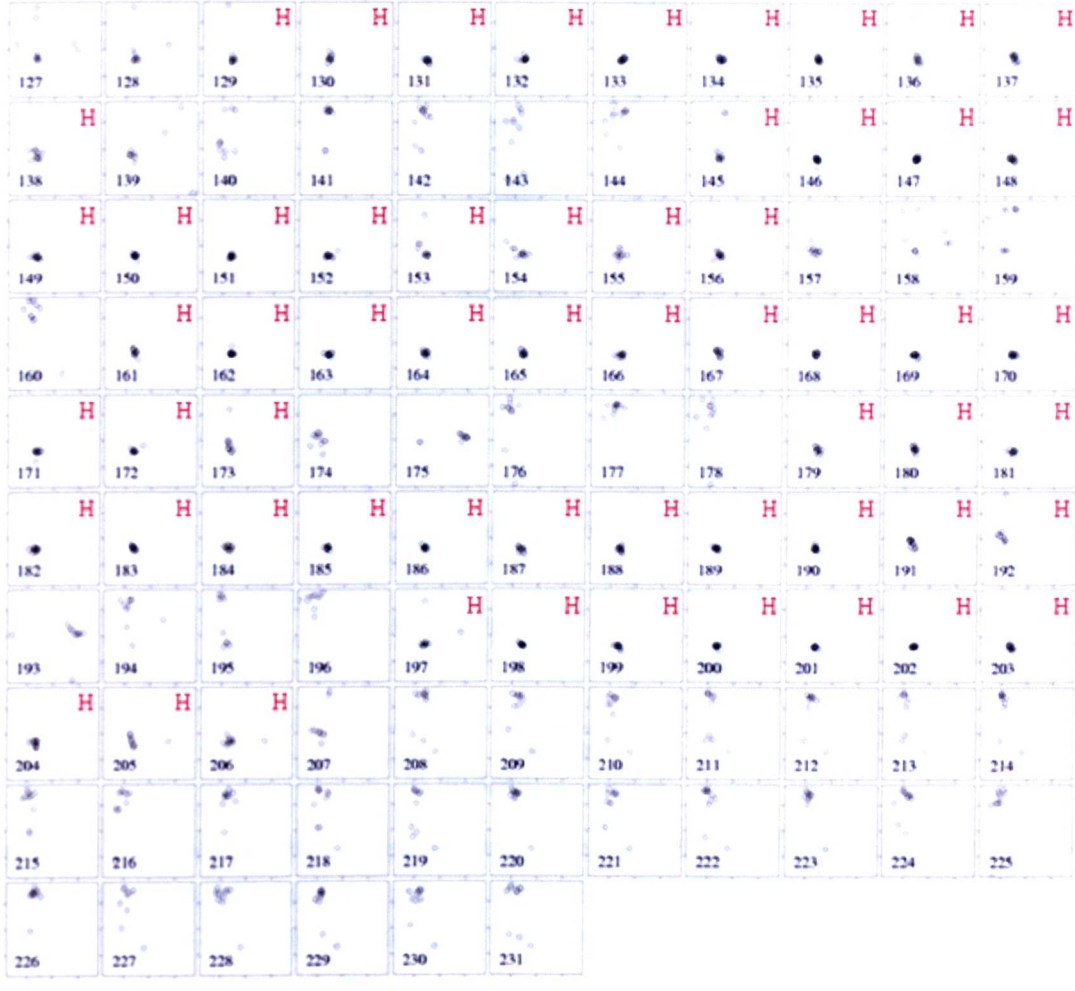


Figure 5.2: TALOS plots for phi and psi per residue. The phi (x) and psi (y) angles of the ten best matches are shown for every residue as open circles. In each plot, the x and y axes cover -180° to +180°. Residues with angles typical for helices are annotated with an 'H' in red.

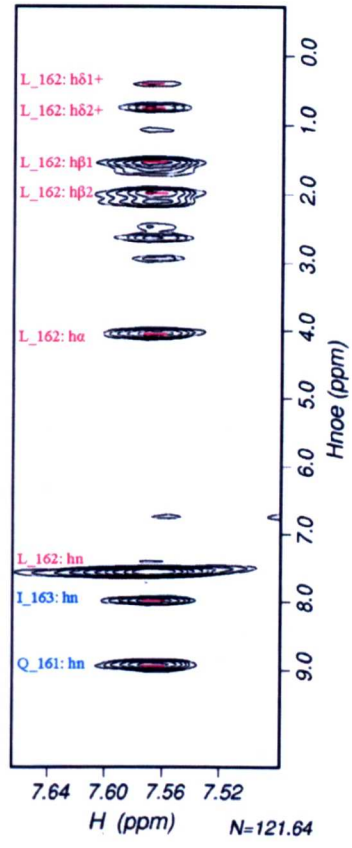


Figure 5.3: Example of an NOE assignment strip. The HN NOE strip of Leu 162 from the ^{15}N edited NOESY is displayed. All intra residue peaks assigned are labelled in red and sequential NOEs to the HN of Gln 161 (d_{NN} (i, i-1)) and Ile 163 (d_{NN} (i, i-1)) are labelled in blue. Long range NOEs from Leu 162 have not been assigned.

5.2.3 Structure Calculations

Structure calculations were carried out with ARIA (1.2) using a set of unambiguous NOE distance restraints and backbone dihedral angle restraints, that were classified as 'good' by TALOS. Section 5.2.1 discusses the prediction of backbone dihedral angles by TALOS in detail and section 3.5 describes the ARIA package. In the initial structure calculations, all ambiguous distance restraints derived from NOESY data by ARIA itself were also included. From these ambiguous restraints, ARIA gradually classified 31 NOEs as unambiguous. These were manually checked in the NOESY spectra after each set of calculations, before reintroducing them as unambiguous. Only 21 of the 31 NOEs were confirmed as unambiguous, the remaining 10 were from the heavily overlapped region of the NOESY spectra, and corresponded to the region of the protein predicted as unstructured. Numerous rounds of calculations were carried out to refine the structure and after each set violating NOEs were removed. The energy and RMSD of the ensemble of structures was monitored as a guide to quality. At the end of the final calculation 486 NOE restraints defined the ensemble of seven structures; the distribution of these is shown in Figure 5.4. These were of low energy and had RMSD values in structured regions of around 1.5 Å. (Figure 5.5). Also, their backbone angles were within the allowed regions of the Ramachandran plot. The final ensemble of structures is shown in Figure 5.6 and the statistics in Table 5.0.

5.3 Description of the preliminary ensemble of NMR structures for DnaD-Cd

The average structure of the preliminary ensemble of DnaD-Cd structures is composed of five helices; I L129-F133; II C146-D156; III A160-V171; IV Y180-K190 and V V196-I201 respectively. The number of helices is consistent with TALOS analysis of backbone chemical shift data. The first four helices in the average structure pack together into a tight bundle forming a hydrophobic core and the firth helix extends away from this main core of the protein. Continuing from helix V is a disordered C-terminal tail comprising residues 202-232. Three of the five helices (helix II, III, and IV) make up the most well define region of the protein with RMSD of less than 1.5 Å. The most

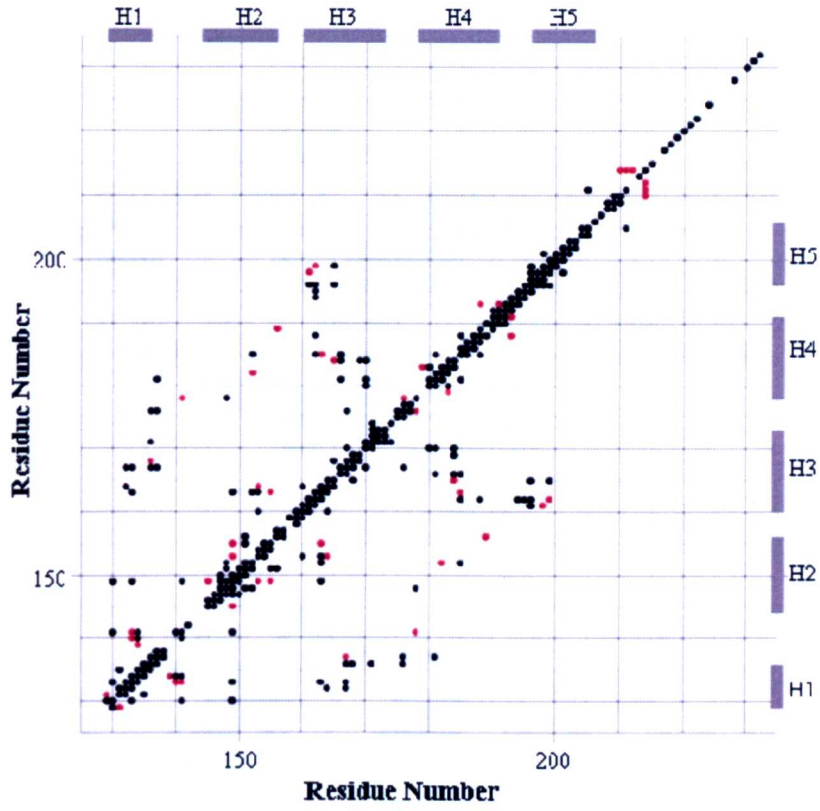


Figure 5.4: Map of NOE distance restraints. Each dot represents an NOE restraint, the position of these shows which residues are specially connected. In black are unambiguous NOEs assigned through NOESY assignment and in red are NOEs identified during structure calculation by ARIA. The grey rectangles along the top and right hand side of the map represent helices (H1, H2, H3, H4 and H5) that make up DnaD-Cd.

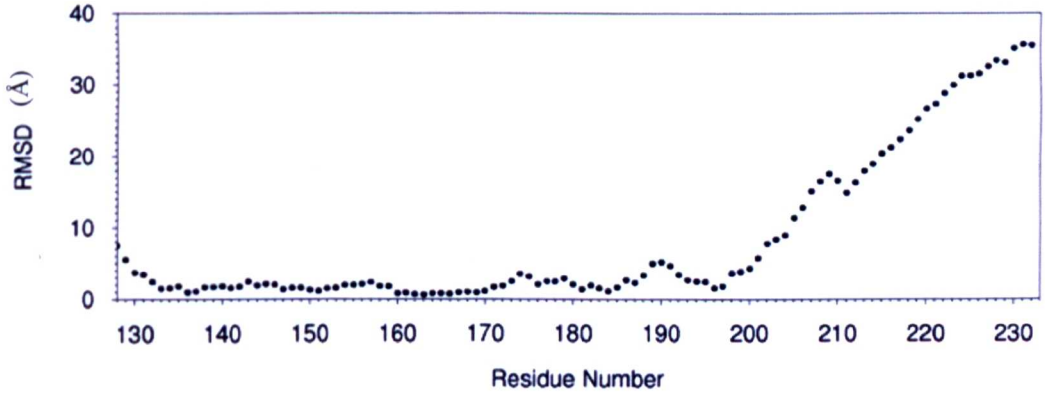


Figure 5.5: Monitoring quality of the structure. The average RMSD values of backbone atoms are displayed in sequence order for the final ensemble of structures.

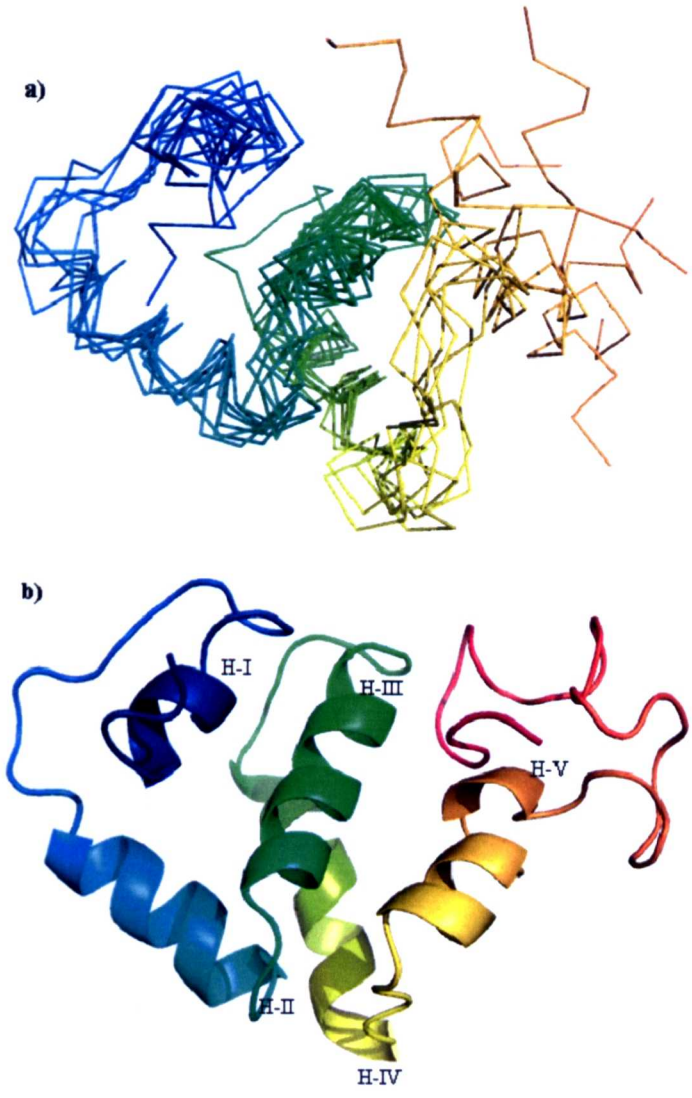


Figure 5.6: Preliminary structures of DnaD-Cd. (a) The preliminary ensemble of seven structures is shown for residues 129-206. Helices are numbered on the cartoon representation of the (b) average structure (residue 129-232) of the preliminary ensemble. Both images are colored from the N- to the C-terminus in blue to orange for clarity of the individual helices and are in a similar orientation.

Table 5.0: Structural Statistics for the final ensemble of DnaD-Cd

NOE Distances	
Intra-residue	290
Sequential	85
Medium range (2< |i-j| <4)	23
Long range (|i –j| >4)	88
Dihedral restraints	112
Procheck	
Favored regions (%)	70.9
Additionally allowed regions (%)	22.6
Generously allowed regions (%)	4.7
Disallowed regions (%)	1.9

poorly defined helix, helix V lacks a definite direction and terminating boundary, as seen from the ensemble in Figure 5.6. This reflects the lack of NOEs in this region as seen in Figure 5.4 and is consistent with the large RMSD of approx. 3Å.

5.4 Homology model for DnaD-Cd based on 2zc2

It was not found to be possible to refine the preliminary ensemble of NMR structures of DnaD-Cd because only a limited number of NOEs could be assigned unambiguously in NOESY spectra to use as distance restraints in structure calculation. The major obstacles were poor spectral sensitivity and crosspeak overlap. Further problems were related to the presence of additional peaks from minor conformations of the protein which result when DnaD-Cd is truncated at the C-terminal end. In order to calculate a more precise structure for DnaD-Cd, the entire resonance assignment procedure could possibly be repeated with a construct that would give better quality spectra such as a construct of DnaD-Cd with a C-terminal truncation at residue E215. A ^{15}N HSQC spectrum for the DnaD-Cd215 construct made later on in this project for DNA titration studies is shown in Figure 6.15.

While work towards structure determination for DnaD-Cd was being carried out, a structure from *Streptococcus mutant* (pdb code 2ZC2) which shows sequence homology to residues 129-232 of DnaD-Cd (Pairwise sequence identity of 41%) was deposited in the protein data bank. The construct in the 2zc2 structure corresponded to residues L129-I201 of *B. subtilis* DnaD, with the final structured residue corresponding to A199. An alignment of DnaD-Cd with the sequence from *Streptococcus mutant*, corresponding to the structure 2zc2 is shown in Figure 5.7b. Overlaying the structure of 2zc2 with the average structure of the preliminary ensemble for DnaD-Cd showed that these two proteins share the same fold (Figure 5.7a). Helices I-IV were particularly similar in length and orientation. Since a construct of DnaD-Cd that would give improved spectra was not available to immediately repeat the resonance assignment procedure to obtain a better model than the preliminary ensemble of structure for DnaD-Cd, a homology model was built based on structure of 2zc2 using the SWISS-MODEL server. SWISS-MODEL uses the amino acid sequence of the target protein to select template structures

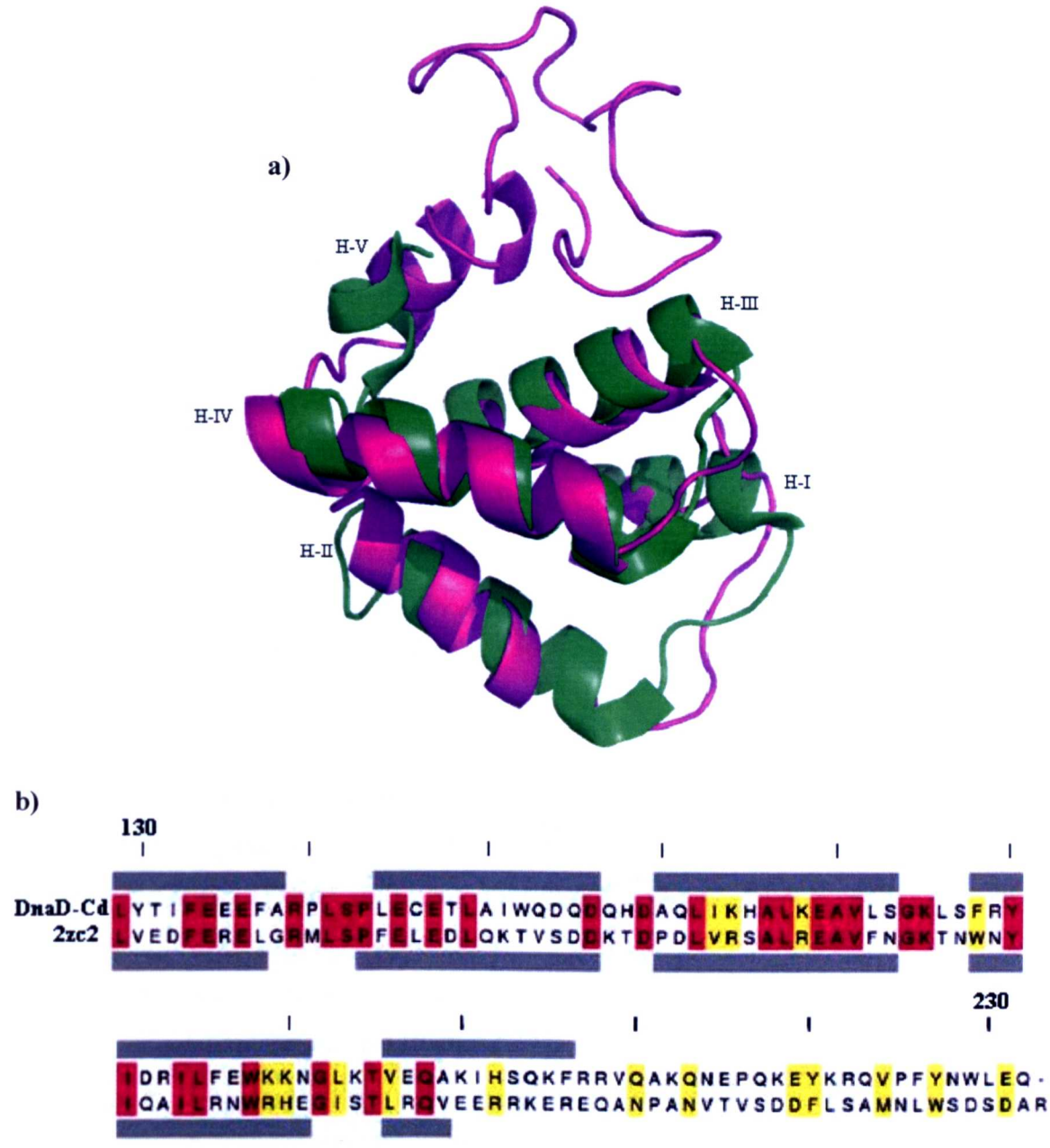


Figure 5.7: Comparison of DnaD-Cd with 2zc2 from *Streptococcus mutans*. (a) The average preliminary structure for DnaD-Cd (magenta) and the structure of 2zc2 (green) are overlaid. The C-terminal ends of the helices are labeled with helix number. (b) An alignment of DnaD-Cd with the sequence from *Streptococcus mutans*, corresponding to the structure 2zc2 is shown. The extent of helices in the two structures is shown in grey above and below the sequences. Identical residues are colored red, and similar residues are colored yellow. The tickmarks denote every tenth residue starting from residue 130.

from the PDB that have sequence identity of more than 25% with the target sequence. For DnaD-Cd, SWISS-MODEL selected the structure of 2zc2 as the only template. To build the homology model it first performed a local pair-wise alignment of the 2zc2 and DnaD-Cd sequences. It then generated the core of the model using the backbone atom positions of the 2zc2 structure. The side-chain atoms were modeled in and an energy minimization procedure was carried to regularize joining fragments and to identify parts of the model with conformational errors. Finally, the model was evaluated. The 3D homology model for DnaD-Cd is shown in Figure 5.8i. More than 50 long-range ($|i-j| > 4$) NOEs that defined the preliminary ensemble of structures corroborated the overall fold of the domain based on 2zc2. Examples of clearly assignable long range NOEs that defined the core fold were from the aromatic ring of F133 to the methyl groups of L149, I163, and L167, and from the methyl group of A166 to the methyl groups of I181 and L185. Residues that were part of the well packed hydrophobic core: residues F133, F137, L141, L149, I163, A166, L167, F178, I181, and L185 correspond (with a two residue shift in numbering) to the core residues F131, L135, L139, F147, V161, A164, L165, W176, I179 and L183 in the structure of 2zc2.

The length of helix V in the homology model for DnaD-Cd was as long as suggested by the ClustalW alignment of sequences selected to represent the extended family of putative DnaD-Cd (Figure 5.9), but considerably shorter then suggested by the following:

- Secondary structure predictions for DnaD by Phyre suggest helix V terminates at F206 (Figure 1.14a).
- TALOS analysis of the backbone chemical shift data indicates that helix V spanned residues V196-F206 (Figure 5.2 and Figure 5.10b)
- ^{15}N T_2 relaxation data inspected on a per residue basis suggested that the protein is predominantly disordered from residue 206 onwards (Figure 5.10d)

The amide strips of the ^{15}N edited NOESY spectra for residues between V196 and Q213 were re-analyzed for $d_{NN}(i,i+1)$ NOEs, but due to signal overlap such NOEs could not be

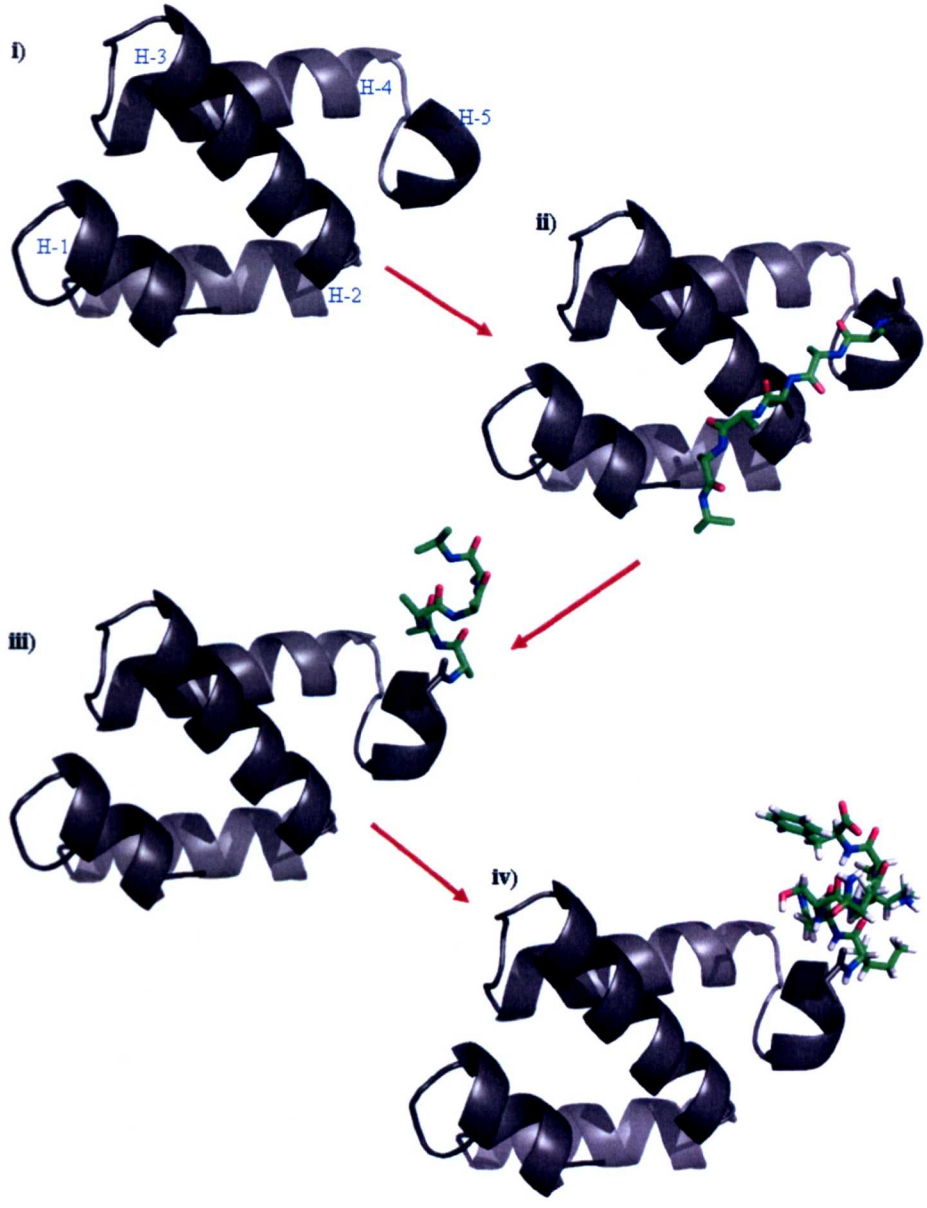


Figure 5.8: Building the final helix on the homology model of DnaD-Cd. (i) The model of DnaD-Cd was built using swiss-model from the coordinates of 2zc2. (ii) The six additional residues (A199-F206) were modeled in using XPLOR. (iii) Helical dihedral angles were imposed on the six new residues using in-house python scripts. (iv) Side-chains were added and steric clashes removed using XPLOR. The homology model is represented in cartoon (gray) and the additional six residues in stick, with the Carbon, Hydrogen, Nitrogen and Oxygen atoms colored in green, white, blue and red respectively.

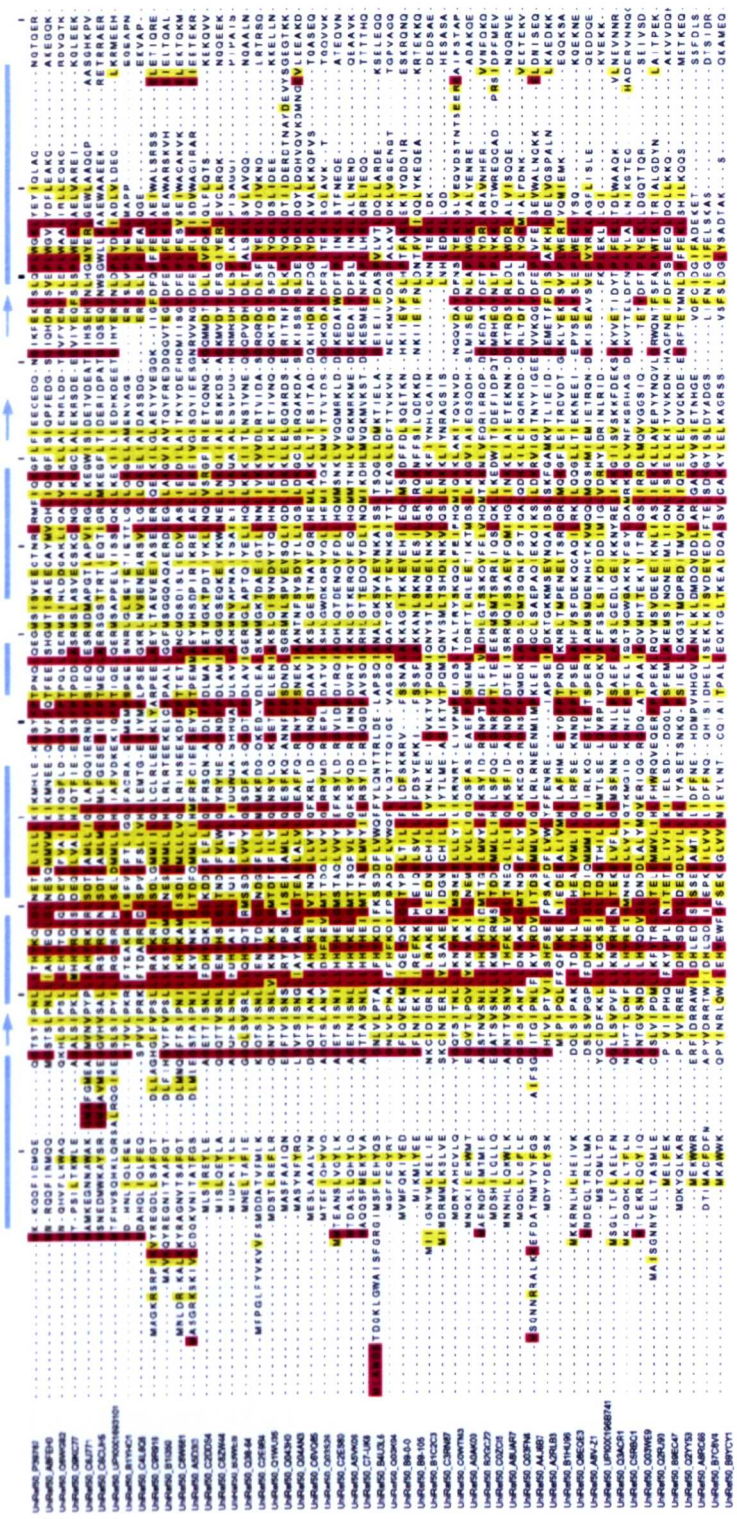


Figure 5.9 continues onto the next page

Figure 5.9 continued from the previous page

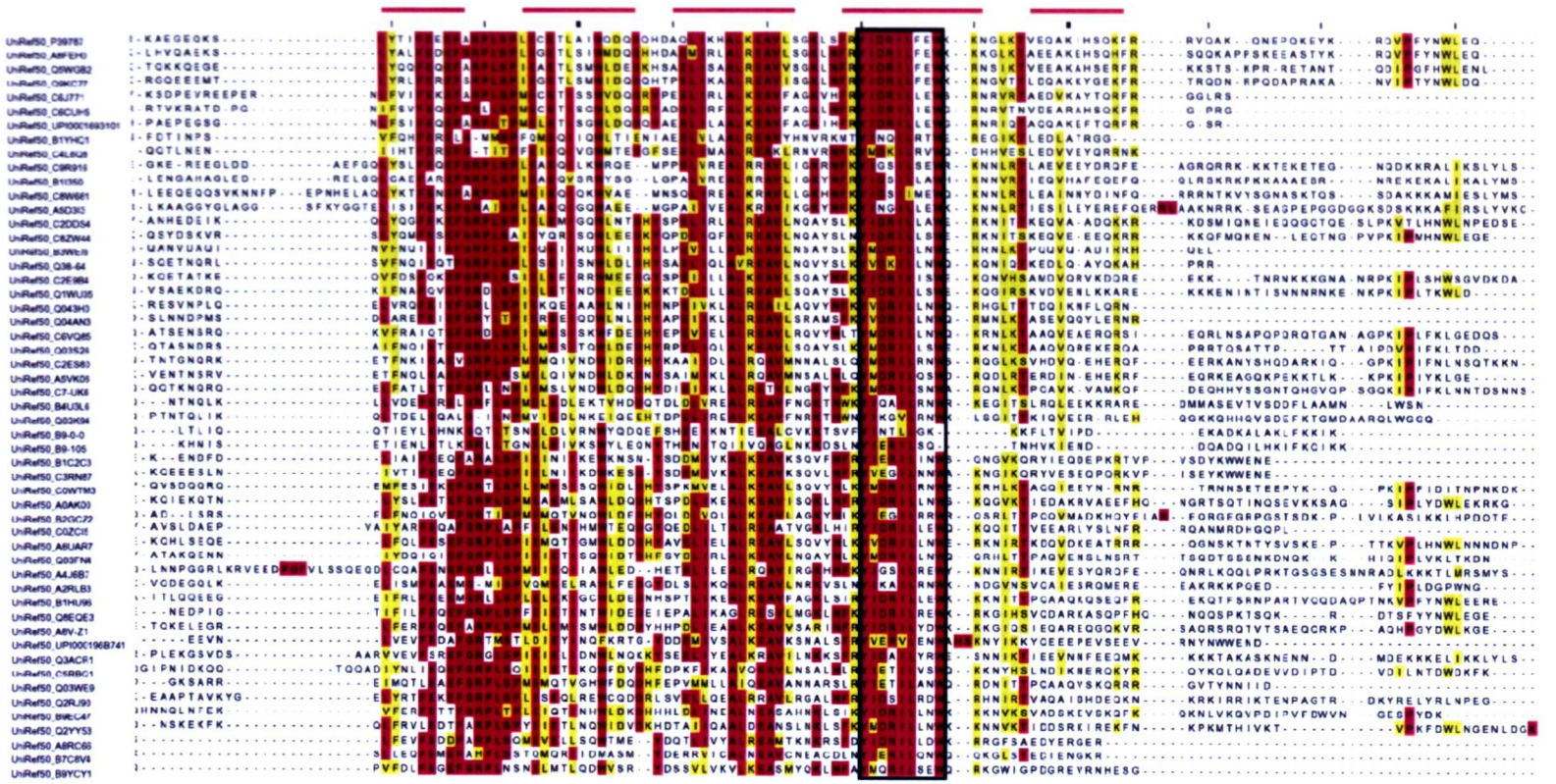


Figure 5.9: ClustalW alignment of the sequences defined as DnaD from Figure 7.5. Residues that are identical in at least 50% of sequences (excluding gaps) are colored in red and residues similar across at least 50% of the sequences in yellow. Sequences are marked with their Uniprot identifier. *B. subtilis* DnaD is the first sequence in the alignment (P39787). Secondary structure of the DnaD-Nd (blue) and DnaD-Cd (red) is shown along the top of the alignment. The location of the YxxxIxxxW motif is shown boxed in black. The tickmarks denote numbering of the *B. subtilis* DnaD sequence (thin/thick ticks at 10/50 residue intervals).

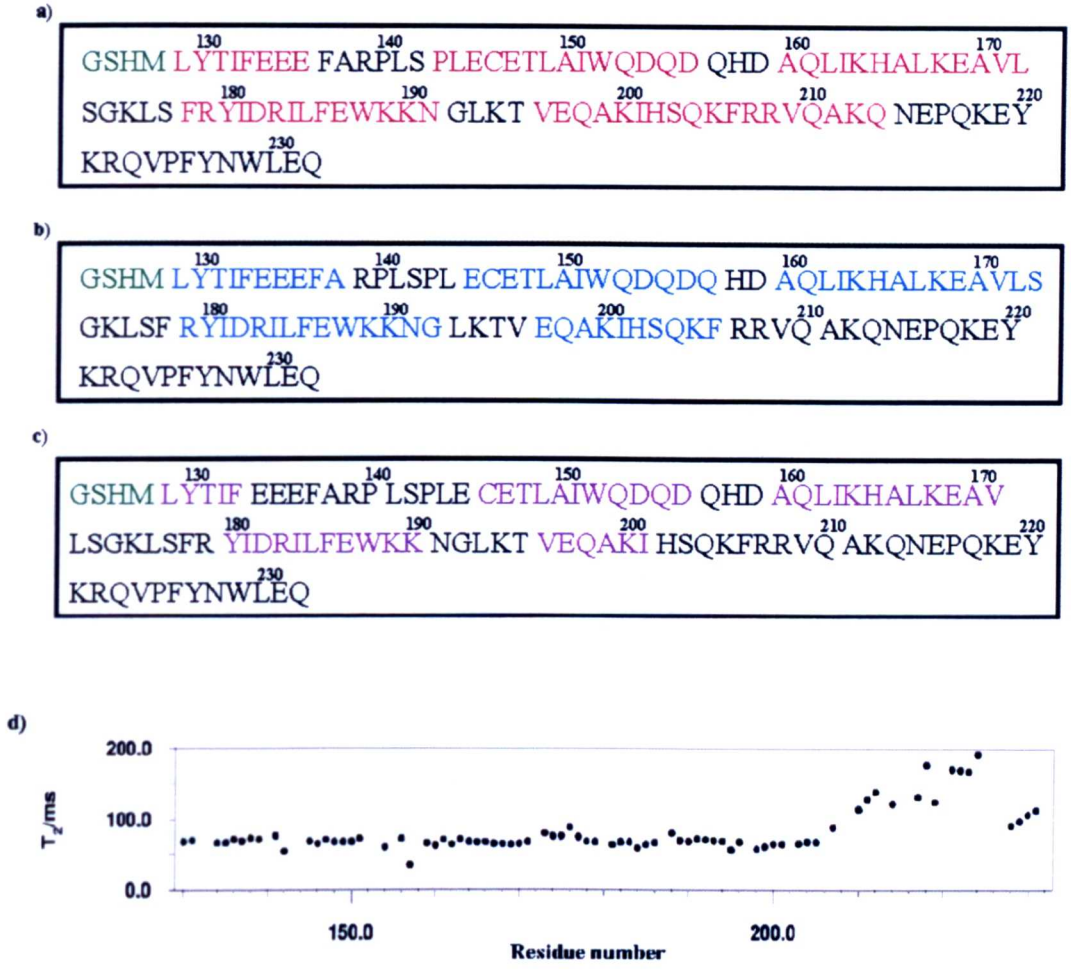


Figure 5.10: Defining the extent of helix V in DnaD-Cd. Helix boundaries predicted by (a) Phyre, (b) TALOS and (c) ARIA are highlighted in red, blue and magenta respectively. Residues remaining after cleavage of the N-terminal His tag are in green. (d) The ^{15}N relaxation data for residues 129 to 206 is shown. Uniform T2 values are obtained up to residue 206. The elevated values beyond this point indicate that the boundary between structure and disorder is approximately at residue 206.The reduced value of T2 for the four data points at the extreme C-terminus suggest some degree of constraint or conformational broadening, but there is no evidence for regular structure in the NOEs or chemical shifts for these residues.

identified. Secondary structure predictions for 2zc2 and 2i5u (Figure 1.14) by Phyre had suggested that helix V was considerably more extensive than presented in their X-ray crystallographic structures. Therefore, it was likely that this helix was at least as long as inferred by the NMR data for the *B. subtilis* protein in many, and probably all, DnaD-Cd domains. In light of all the data above, helix V of the homology model for DnaD-Cd was extended to the length indicated by the NMR data (F206). To start with the Cα atoms for the six additional residues were added with 9999.999 as the coordinates in the pdb of the homology model of DnaD-Cd that was generated by SWISS-MODEL. Then this modified pdb was run in X-PLOR to model in the backbone and side chains of the six additional residues as random coil. To do this, X-PLOR reads the modified pdb, realises that it has these extra residues, and from its topology file knows which atoms it needs to add in besides Cα. The energy minimisation routine within X-PLOR places the atoms in so that the covalent geometry is satisfied. Using an in-house python script, dihedral angles of -57° and -47° that are typical for helices were imposed on these additional residues to fold helix V into its secondary structure. Finally, any remaining steric clashes were removed using the rigid body energy minimisation script within X-PLOR (Brunger, 1993).

Helix V, which was built into the homology model of DnaD-Cd extended away from the four-helix core of the protein. Long range NOEs that would constrain this helix were searched for in the ^{15}N and ^{13}C NOESY spectra, but such NOEs could not be found, suggesting that helix V did not pack against the rest of the protein. The DnaD-Cd truncation mutant DnaD-Cd196, which lacked helix V, supported the peripheral nature of this final helix, because it produces a protein that was well folded, as inferred from the small crosspeak perturbations in its HSQC spectrum (Figure 6.10). Furthermore, two strongly up-field shifted signals from the methyl groups of residues I184 and L185 at -0.60 and -0.85 ppm respectively in the 1D spectrum of DnaD-Cd were still present in the 1D spectrum of DnaD-Cd196. The shift for I184 arises from close proximity to W188 and thus suggested that deleting helix V does not significantly perturb the arrangement of the tight bundle of four helices. Consistent with a peripheral nature, hydrogen-deuterium exchange data suggested that helix V was more exposed than the rest of the

protein, as this had more rapid amide hydrogen-deuterium exchange rates than the other helices. Exchange for all residues in Helix V was complete after 2.5 hours at 298K, pH 6.5, while intensity was still observed for residues in other helices at that time (Figure 5.11).

5.5 A potential DNA binding site.

5.5.1 Key residues involved in the packing of DnaD-Cd are conserved

ClustalW alignment of sequences from the extended family of putative DnaD-Cd (Figure 5.9) showed that residues involved in the packing of the overall structure were highly conserved, in particular a YxxxIxxxW motif. Six hydrophobic residues L185 (99%), Y180 (99%), A170 (97%), I184 (92%), W188 (91%) and A166 (91%) were the most highly conserved in the alignment (The mean absolute conservations value for each residue is shown in brackets). Three of these residues, Y180, I184 and W188 were part of the YxxxIxxxW motif. These were located on a more exposed face of helix IV, whereas the other residues, L185, A166 and A170 seem to be packed into the core of the protein, anchoring this helix. The position of the six highly conserved residues is shown on the structural model in Figure 5.12. The solvent accessible area was calculated for each residue in DnaD-Cd with the MSMS package (Sanner *et al.*, 1996) embedded within the program chimera (Pettersen *et al.*, 2004). The solvent accessible area is defined as the area swamped out by a sphere of 1.4 Å radius representing a water molecule, while it is rolled over the surface of the protein. It basically describes the area over which contact between the protein and solvent can occur. A solvent accessible area of 64.0 $Å^2$, 24.2 $Å^2$ and 40.3 $Å^2$ was obtained for residues, Y180, I184 and W188 respectively, suggesting that these particular residues were not highly exposed on helix IV. However, residues R183, K189 and K190, which are also part of helix IV, were highly accessible to the solvent as values of 177.4 $Å^2$, 166.7 $Å^2$ and 178.9 $Å^2$ respectively were obtained. Out of the five helices in DnaD-Cd, the most solvent accessible residues lie on helix IV. Figure 5.12d shows solvent accessibility of each residue.

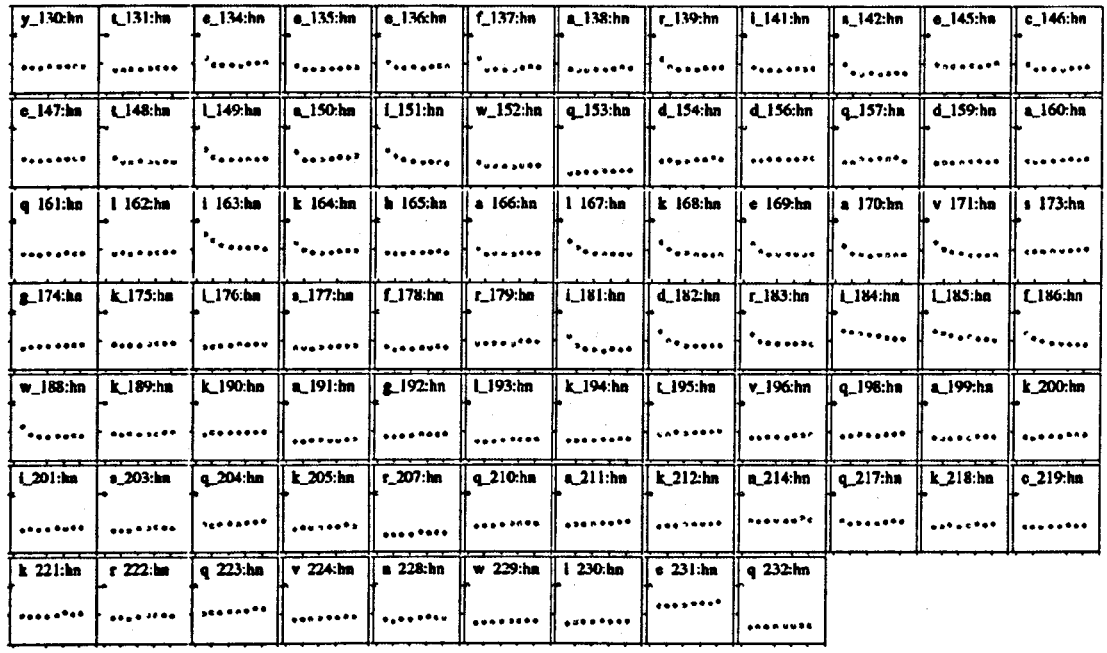


Figure 5.11: Hydrogen/Deuterium amide exchange plots of DnaD-Cd. Hydrogen/Deuterium amide exchange data plotted as peak heights of the HSQC crosspeaks for each residue relative to its height in the control spectrum (y-axis), versus time in hours (x-axis). The x-axis covers a time of 0 to 25 hours and the y-axis covers an intensity of 0 to 1.5. The first data point was extracted from the control spectrum in which the protein was diluted with water, and the points that follow were taken from ^{15}N HSQC spectra acquired at 3.0, 5.0, 7.5, 10.3, 13.2, 15.6, 18.0 and 20.5 hours.

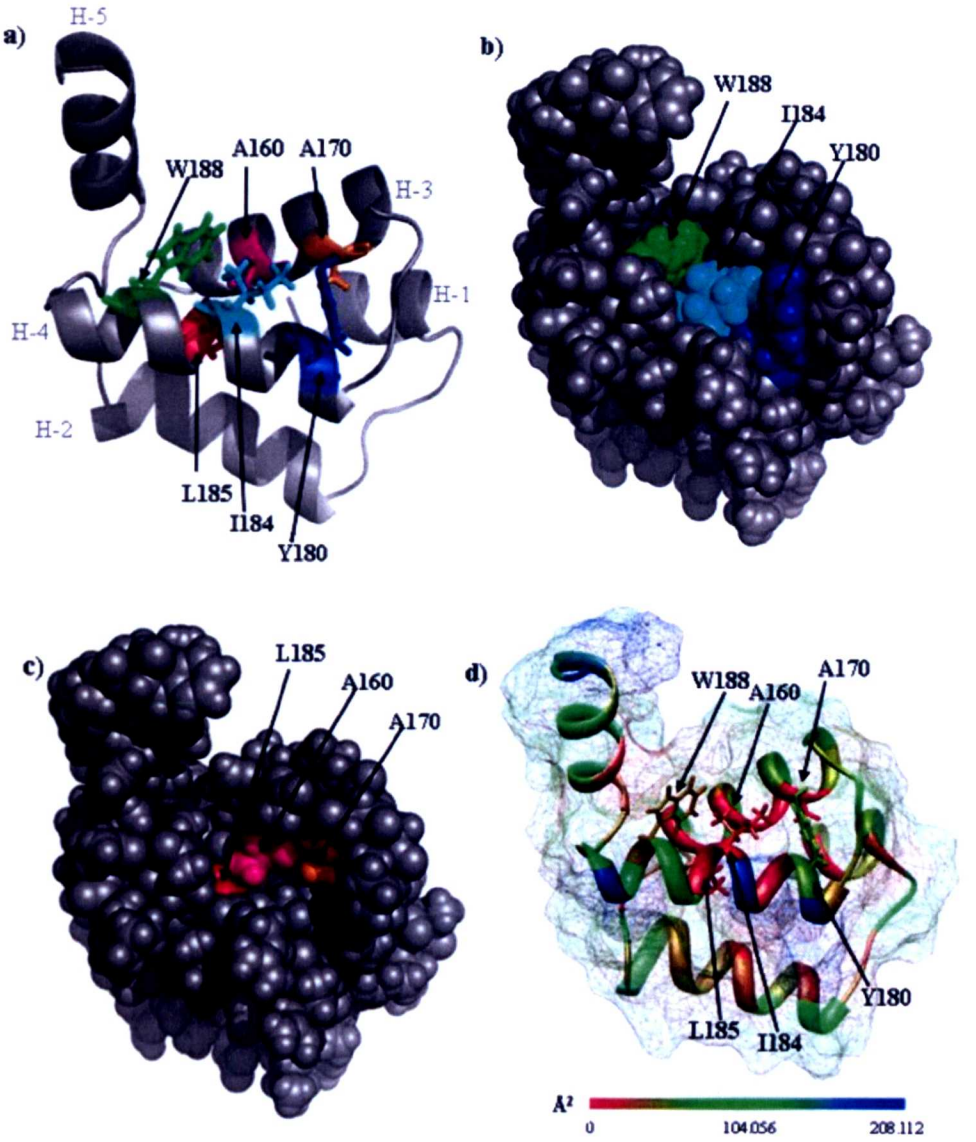


Figure 5.12: The conserved YxxxIxxxW motif mapped onto a structural model of DnaD-Cd. (a) A ribbon and (b) spacefilling representation of the DnaD-Cd model showing the location of the six most highly conserved residues, A160, A170, Y180, I184, L185 and W188 in pink, orange, blue, cyan, red and green respectively. Residues I184, Y180 and W188 have been removed in (c).These images were created using Pymol. (d) A mesh surface (in semitransparent mode) and ribbon representation of DnaD-Cd showing solvent accessibility of each residue. The residues are colored as in the key below the model. Side chains of the six highly conserved residues are shown in sticks on the ribbon representation. The solvent accessible surface area was calculated with the MSMS package embedded within the program chimera, which was used to create this image.

5.5.2 Charge distribution of DnaD-Cd

DNA-binding proteins often have positively charged clusters by which they bind to negative charges of DNA. Since DnaD-Cd is a non-specific DNA binding protein (Marsin *et al.*, 2001), at least one cluster of positively charged residues was expected on its exposed surface. Therefore, to try and identify the potential DNA binding sites of DnaD-Cd, the predominant distribution of charge amongst sequences of putative DnaD-Cd and on the surface of the DnaD-Cd itself was examined. The fraction (f_{NG}) of basic and acidic residues was plotted for each position in the ClustalW alignment of putative DnaD-Cds shown in Figure 5.9. A distinctly bipolar distribution of charge was observed across the alignment (Figure 5.13a). Residues corresponding to helices I, II and part of helix III of DnaD-Cd were mainly acidic, while residues after this point, in particular between residues 196 to 192 and 203 to 215 were predominately basic. On the exposed surface of DnaD-Cd (Figure 5.13b) distinct clusters of basic residues was observed. These were around the highly conserved YxxxIxxxW motif strongly suggesting that it may be part of the DNA binding region.

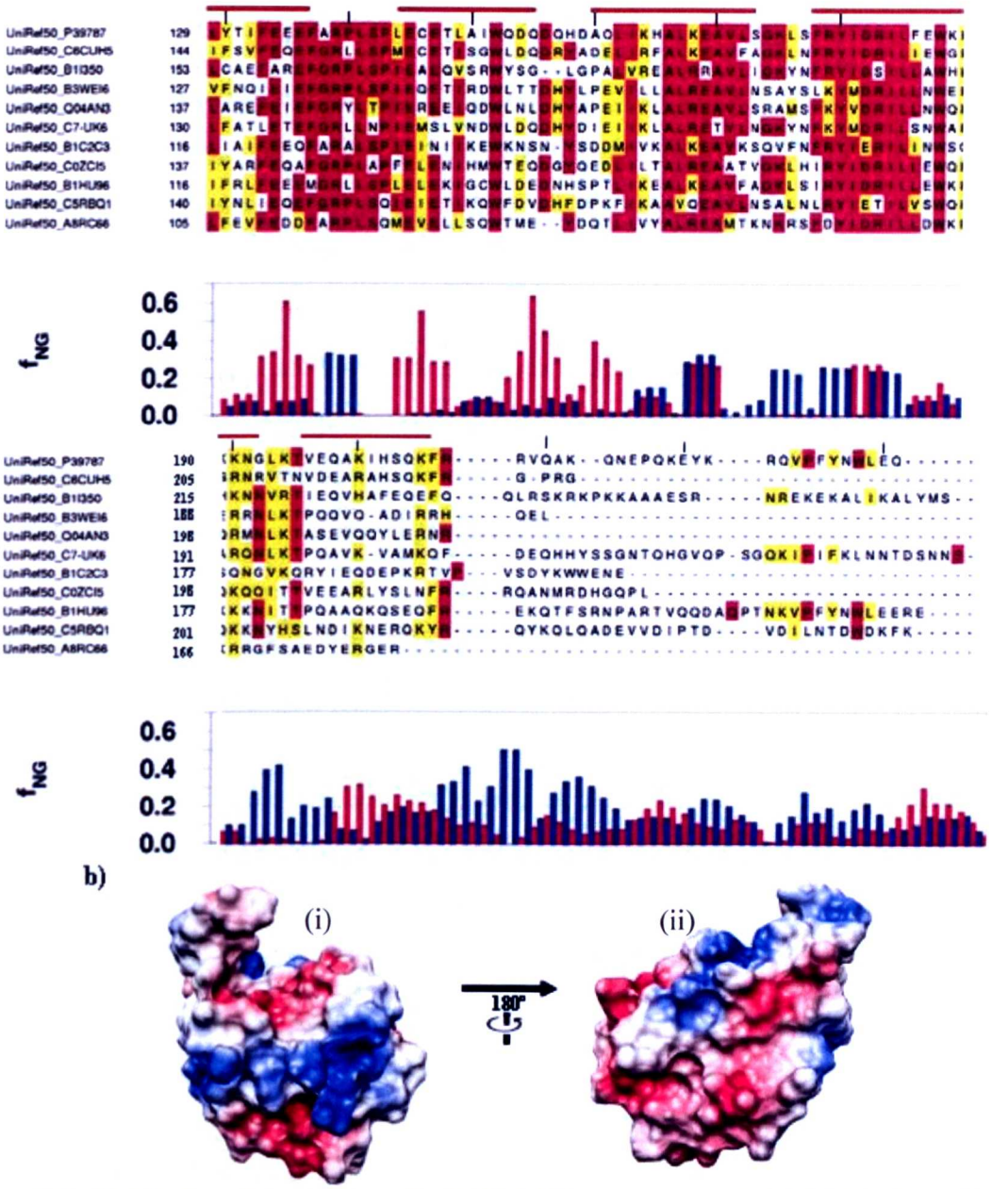


Figure 5.13: Distribution of charge across DnaD-Cd. (a) The sequences are 1, 6, 11, 16 etc drawn from the C-terminal part of the alignment in Figure 5.9. Secondary structure of DnaD-Cd is shown along the top and every tenth residue, starting from residue 130 is tickmarked. The location of the YxxxIxxxW motif is shown by the black box. Below the alignment, the fraction, f_{NG}, of basic (blue) and acidic (red) residues is shown, expressed as a fraction of the total non-gapped residues at each position in the alignment of the C-terminal region of DnaD in Figure 5.9. The profile was smoothed using a three-residue window. (b) A model of DnaD-Cd showing basic (blue) and acidic (red) residues across the surface. The electrostatic surface was calculated using the program DelPhi and visualized using the program Chimera. The model (i) is in the same orientation as in Figure 5.11 and (ii) is rotated as indicated.

Chapter 6: Interaction of DnaD-Cd with DNA

Ligand binding plays an important part in the dynamics of proteins and thus can be crucial to the functioning of macromolecular systems (Reid, 1997). Protein-ligand interactions may be accomplished by conformational changes in the ligand, protein or in both. NMR is a particularly useful technique for deriving information on binding sites, as any change in the environment of a protein is reflected by the resonance signals observed in spectra. For instance, when a ligand binds to a protein, the peaks may move, broaden, change in intensity or even disappear completely. Groups showing such changes are likely to be close to or even part of the ligand binding site and thus can help map the protein-ligand interface.

The DNA binding of DnaD is non-specific (Marsin *et al.*, 2001; Turner *et al.*, 2004). A combination of gel shift and AFM data have established that the DNA binding activity of DnaD is confined to its C-terminal domain, however, the molecular details of this interaction remain unknown (Carneiro *et al.*, 2006). This chapter focuses on the binding of DnaD-Cd to DNA at the residue level. It discusses NMR titration data of DnaD-Cd and its truncation mutants which together led to the identification of elements that are part of the DNA binding module. It also presents supporting data from gel shift assays carried out in the laboratory of Panos Soultanas at Nottingham University. Finally, *In vivo* experiments performed in the laboratory of Alan Grossman at the Massachusetts Institute of Technology in Cambridge showing that the identified DNA binding module of DnaD is essential for *B. subtilis* cell viability are discussed in this chapter.

6.1 DNA binding of DnaD-Cd monitored by NMR

In order to investigate how DnaD-Cd binds to DNA, a 10 mer and 19 mer single stranded (ss) DNA were separately titrated into ^{15}N-^{13}C labelled DnaD-Cd up to a DNA:protein stoichiometry of 1:1. At each DNA concentration, 1D proton, 2D ^{15}N

HSQC and 2D ^{13}C HSQC spectra were collected. Using in house macros, signal intensity and chemical shift changes were followed. Given that DnaD-Cd had been observed in gel shift assays to form large complexes with the 10 mer and 19 mer ssDNA (Figure 6.0) that were used for NMR titrations, it was suspected that on the addition of DNA to DnaD-Cd, the intensity of peaks across the entire NMR spectrum would progressively decrease to complete obliteration, as increasing amounts of protein was sequestered into DNA:DnaD-Cd complexes. As a result of this sequestration, chemical shifts for DNA bound DnaD-Cd would not be observed as large complexes are undetectable by NMR due to the effect on signal relaxation. Thus, it was likely that little or even no useful information would be extracted from DNA titrations. Surprisingly, the attenuation of signal across the NMR spectrum on titrating 10 mer (Figure 6.1) or 19 mer ssDNA with DnaD-Cd was far from uniform. During the titration with 10 mer DNA (Figure 6.2a), the cross peaks for some residues (e.g. Y130) gradually decreased in intensity, but reached a plateau at approximately 30% of the original intensity. Other residues either showed a much more severe drop in intensity (e.g. A211 and A170) or remained almost unchanged (e.g. V224). A similar effect on the cross peaks intensity was observed during the titration with 19 mer DNA (Figure 6.3), however, the overall intensity drop was less pronounced, but consistent with the lower protein concentration used. The protein concentration used in both NMR and gel shift experiments is critical for the interpretation of the results and is addressed in detail in section 6.4, which discusses reconciliation of the NMR and gel shift data.

Chemical shift changes were also observed in HSQC spectra for both 10 mer and 19 mer titration sets, however these were less than 0.05 ppm (for a weighted combined ^{1}H and ^{15}N amide shift), even for residues showing a significant drop in intensity. For some peaks, the chemical shift changes were easily followed up to a stoichiometry approaching 1.1, whereas other peaks, particularly those showing a large intensity change, were more difficult to track because the signal became rapidly attenuated, with some cross peaks attenuated below detection levels. The overall chemical shift changes observed for the titration with 10 mer DNA is shown in Figure 6.1b.

As the chemical shift changes were very small, only signal intensity changes were

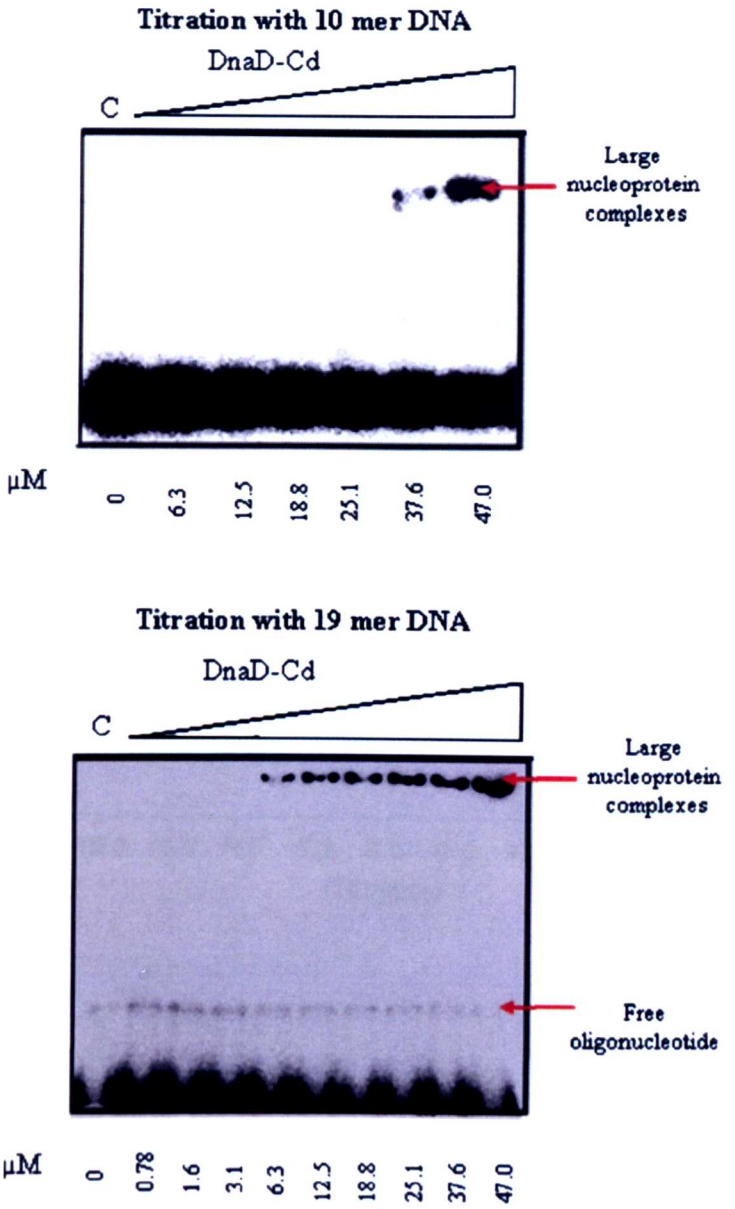


Figure 6.0: **Gel shift assays showing DnaD-Cd forms large nucleoprotein complexes.** DnaD-Cd binds 10 mer and 19 mer ssDNA forming complexes that are too big to enter the gel. DNA binding reactions were carried out in 75 mM NaCl, 87.5 mM Tris pH 7.5, 1.75 mM EDTA 5 mM MgCl and 0.77 mM DTT with 2.5 nM 10 mer or 19 mer ssDNA at increasing concentrations of proteins, as indicated. The control reactions which were 2.5 nM radioactively labeled 10 mer or 19 mer ssDNA in the absence of protein are shown in lane C (Image provided by Panos Soultanas).

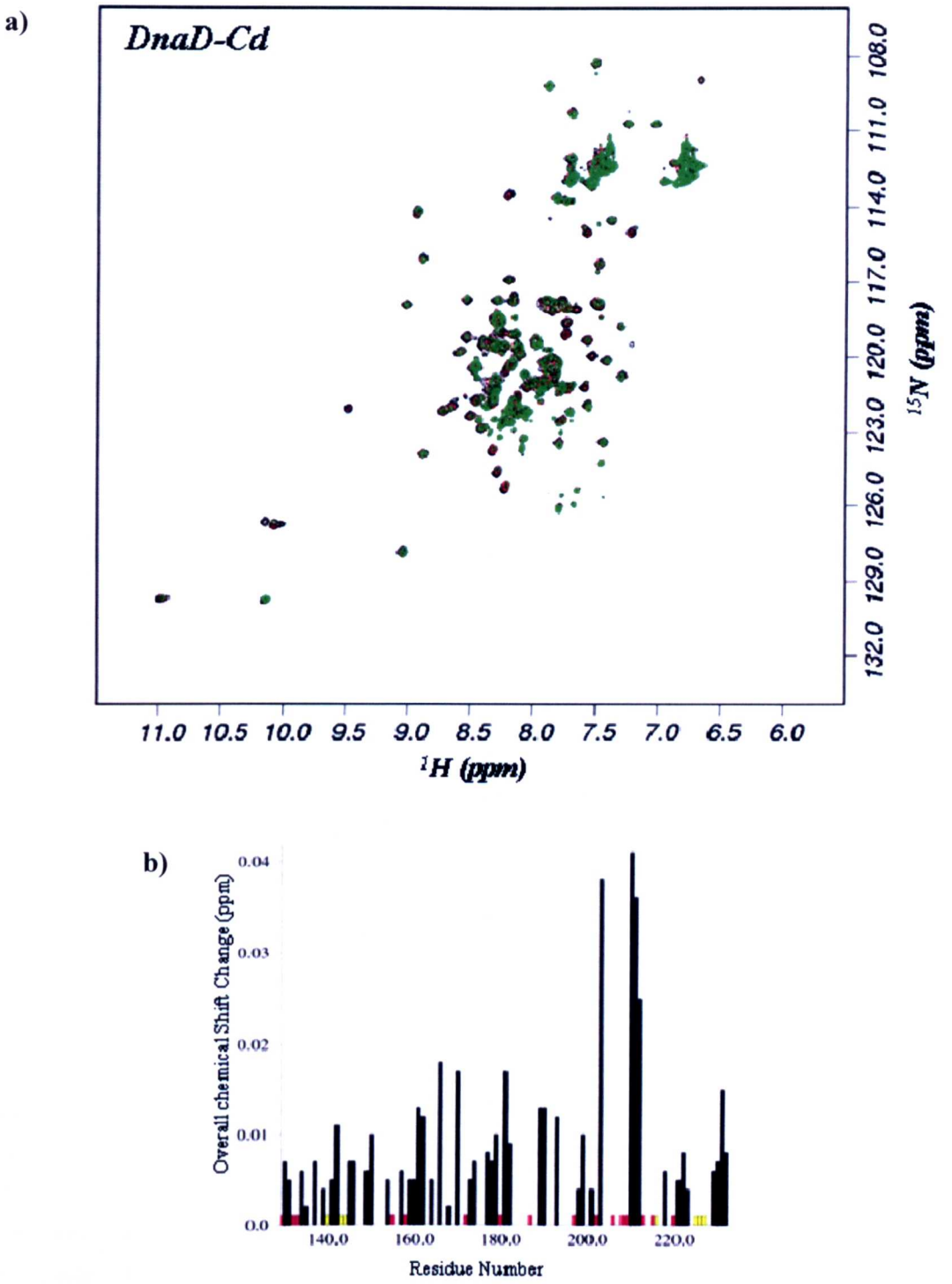


Figure 6.1: DNA binding of DnaD-Cd monitored by NMR. (a) ^{15}N HSQC spectra of DnaD-Cd titrated with 10 mer DNA. Spectra at 0 μM (black), 60 μM (red) and 120 μM (green) DNA are overlaid. (b) The overall amide chemical shift change is plotted per residue. The bars in black show the overall chemical shift change for assigned residues. The bars in red highlight residues with heavily overlapped or noisy crosspeaks and bars in yellow highlight unassigned residues. No bar means no chemical shift change.

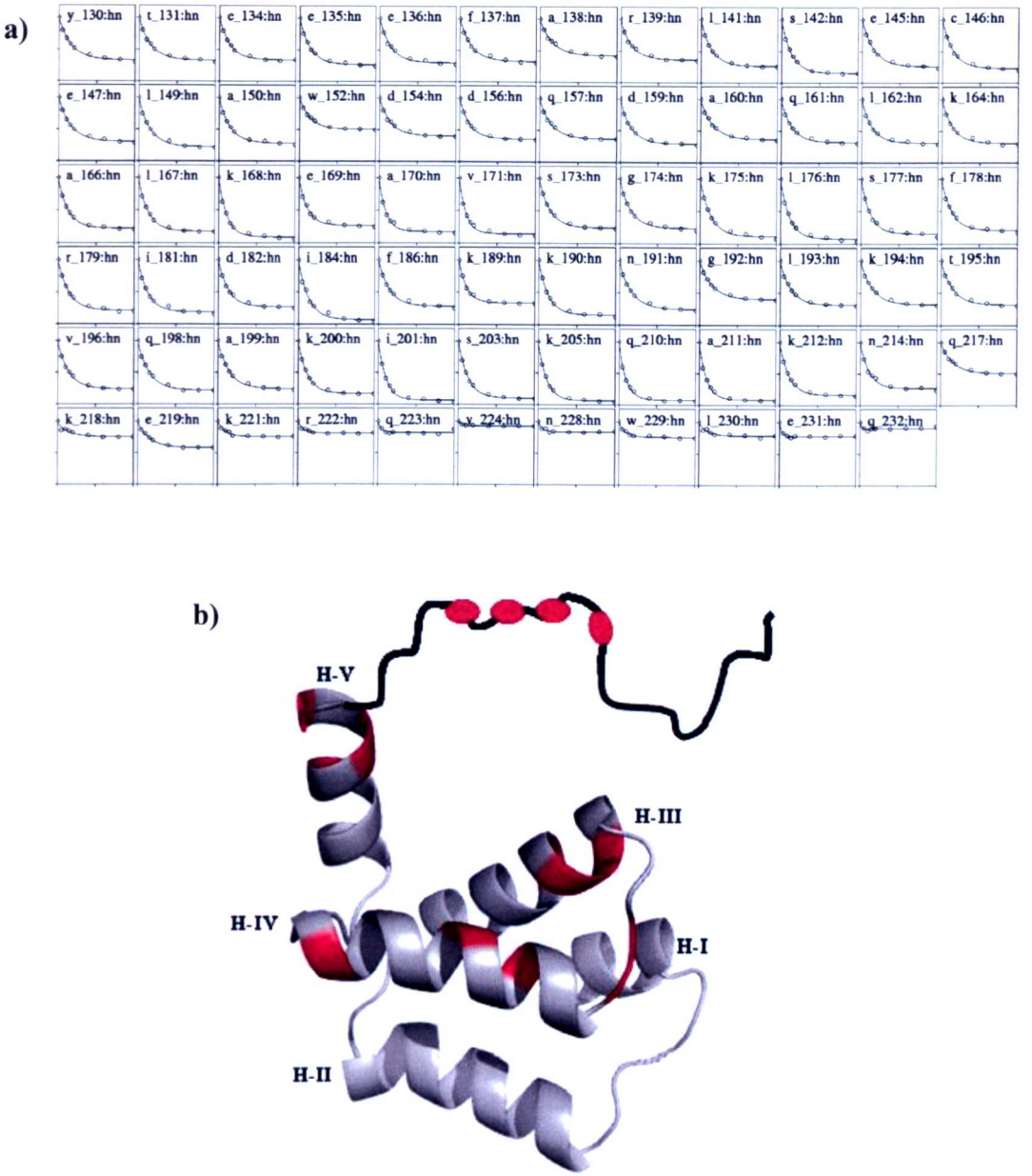


Figure 6.2: Intensity changes on titrating 10 mer DNA into DnaD-Cd. (a) NMR titration curves for the titration of 10 mer ssDNA into 0.3 mM DnaD-Cd. The data are plotted as peak heights of the HSQC crosspeaks for each residue relative to its height at the start of the titration (y-axis), versus DNA concentration (x-axis). Data points were extracted from 2D ^{15}N HSQC spectra for DNA concentrations of 0 μM, 15 μM, 30 μM, 45 μM, 60 μM, 120 μM, 180 μM, 240 μM and 300 μM. In each graph, the x axis covers DNA concentrations from 0 to 300 μM, and the y axis covers ratios from 0 to 1.2. The curves are fit to a decaying exponential plus y-offset shown in section 3.8. Titration curves are sorted by residue position and graphs from overlapped or unassigned peaks have been removed. (b) The model of DnaD-Cd showing, in red residues for which crosspeak intensity was reduced by more than 80% after the addition of 1:1 10 mer DNA to DnaD-Cd. These are residues 168, 170, 171, 176, 177, 181, 184, 190, 201, 203, 205, 210, 211, 212 and 214. The mobile portion of the protein is depicted in sketch form.

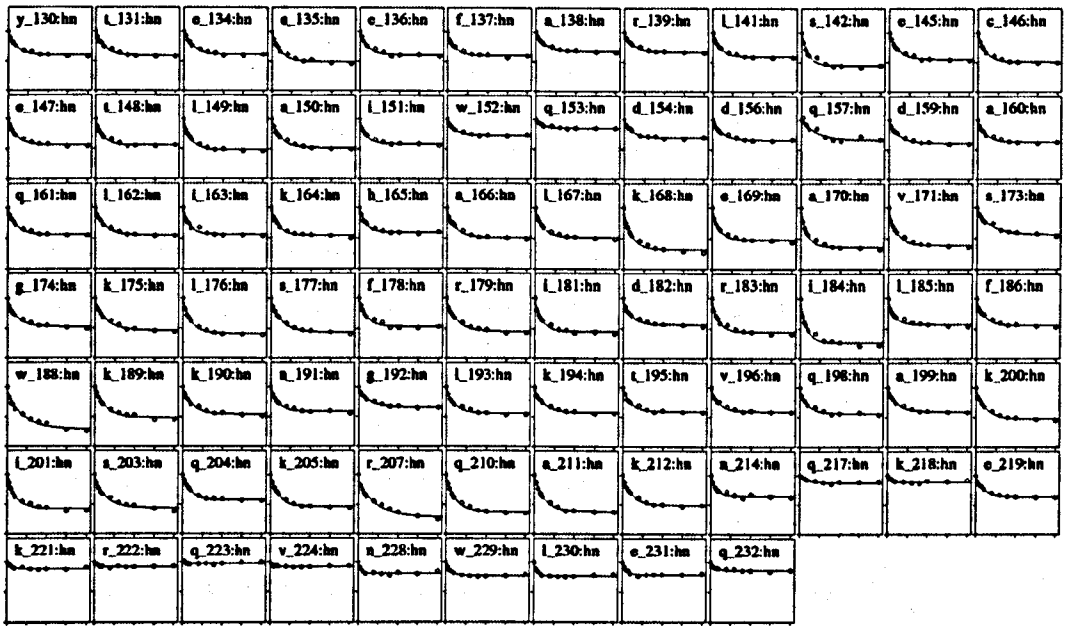


Figure 6.3: NMR titration curves for the titration of 19 mer ssDNA into 0.2 mM DnaD-Cd. The data are plot as in Figure 6.2. Data points were extracted from ^{15}N HSQC spectra for DNA concentrations of 0 μM, 10 μM, 20 μM, 30 μM, 40 μM, 80 μM, 120 μM, 160 μM, 200 μM 300 μM and 400 μM. In each graph, the x-axis covers DNA concentrations from 0 to 410 μM, and the y-axis covers ratios from 0 to 1.4.

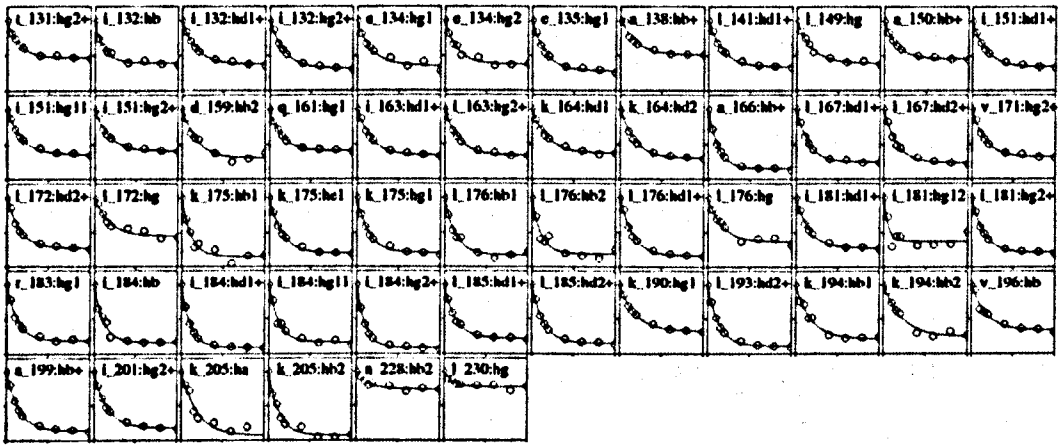


Figure 6.4: NMR titration curves for the titration of 10 mer ssDNA into 0.3 mM DnaD-Cd. The data was extracted from ^{13}C HSQC spectra and plotted as in Figure 6.2. In each graph, the x-axis covers DNA concentrations from 0 to 300 μM, and the y-axis covers ratios from 0 to 1.2.

analysed in detail to understand binding of DnaD-Cd with DNA. In the 10 mer titration set, the signal decay to approximately 30% intensity was interpreted to be due to an increase in the overall tumbling time of the protein as a result of protein-DNA complex formation and possibly inter-complex interactions. A certain stoichiometry for DnaD-Cd versus DNA could not be established from this data, but a drop of 30% intensity was compatible with the formation of small DNA:DnaD-Cd complexes, no larger than a total of 2-3 protein and/or DNA chains.

Of the mobile residues in DnaD-Cd, the most C-terminal of these exhibited a negligible decrease in intensity. These presumably retained sufficient mobility in the complex to be unaffected by the increase in overall tumbling time. Residues that showed much greater reduction in crosspeak intensity than caused by the increase in tumbling time provided most information. These were most likely affected by conformational exchange in the complex as explained in section 6.2.2 scenario1, and thus would reveal the DNA binding region. Locating such residues on the model of DnaD-Cd (Figure 6.2b) identified the region of the protein that most likely binds DNA to include Helices III, IV (which contains the highly conserved YxxxIxxxW motif discussed in section 5.5.1), Helix V and a large portion of the mobile C-terminal tail (up to residue 215). The ^{13}C HSQC titration spectra did not contribute significantly greater detail towards the binding site, as the effect observed on the addition of DNA were interpreted as discussed above for ^{15}N HSQC spectra. Intensity curves for ^{13}C HSQC spectra are shown in Figure 6.4.

When assessing the quality of ^{15}N-^{13}C samples of DnaD-Cd, minor peaks corresponding to truncated DnaD-Cd were observed (see Figure 4.12). In DNA titration HSQC spectra these minor peaks became more prominent before their gradual disappearance, as the signal of the partially overlapping major peaks drops with increasing DNA concentrations (Figure 6.5). This implied that the polypeptide to which these minor peaks correspond to may have a much lower affinity for DNA than DnaD-Cd.

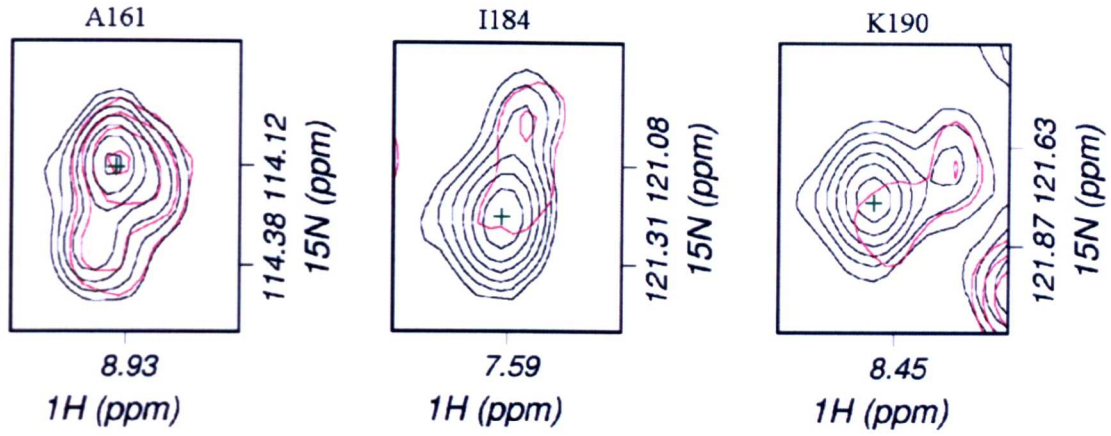


Figure 6.5: Minor peaks show weaker DNA binding. Resonance signals are shown from titration spectra at 0 μM (black) and 120 μM (red) DNA. The major peaks are marked with a green cross.

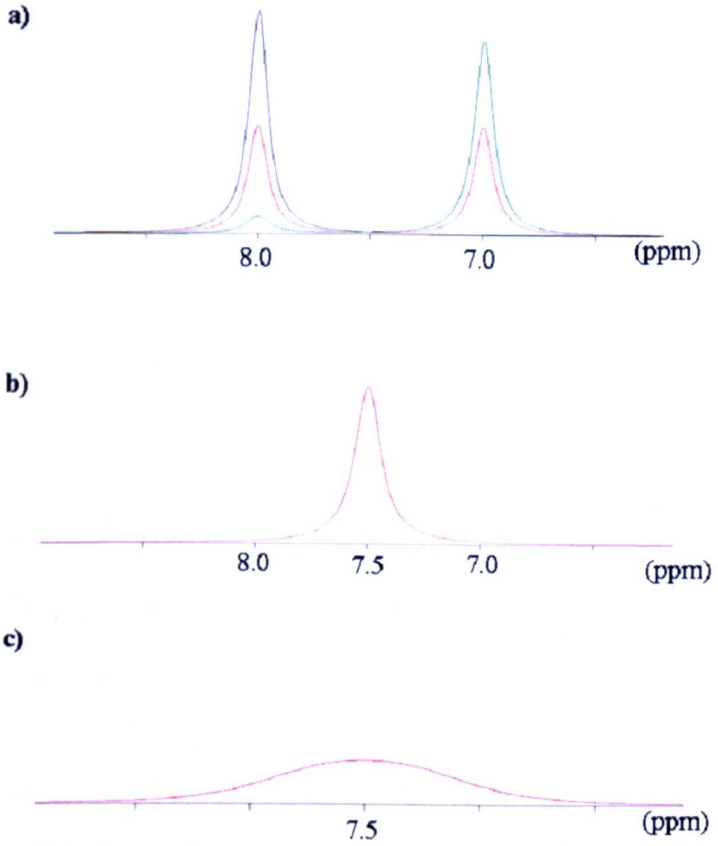


Figure 6.6: Protein-DNA exchange regimes. Line shapes simulated for (a) slow (b) fast and (c) intermediate exchange. The chemical shift values of the free and bound states were set at 8 ppm and 7 ppm respectively. The peak in black represents 100% free protein. Peaks in red represent equal populations of the two states and peaks in green represent 90% of the protein bound to DNA.

6.2 Elucidating the DnaD-Cd-DNA exchange regime from changes in DNA titration spectra.

6.2.1 Protein-DNA exchange regimes

Exchange regimes determine how a spectrum of a protein-DNA mixture changes during a titration. There are three regimes that are usually considered, the first, slow exchange applies when the rate of exchange between the DNA free and bound state is much less than the difference in chemical shift. Under this regime a distinct peak for each state, with intensities proportional to their population would be observed (Figure 6.6a). In the second regime, fast exchange, the rate of exchange would be much greater than the difference in chemical shift. As the DNA concentration increases through the titration, a progressive change in chemical shift towards the bound state would be observed (Figure 6.6b). The slow and fast exchange regimes are the most useful to obtain detailed information about binding. The most difficult to analyse, intermediate exchange, applies when the rate of exchange is approximately equal to the difference in chemical shift. An increase in the population of the bound state during the titration results in broadening of peaks. Therefore, when equal population of the two states exist, a broad peak that spans the chemical shift of the free and bound state of the protein is observed (Figure 6.6c).

6.2.2 DnaD-Cd-DNA binding may be under an intermediate exchange regime

During the DNA titration of DnaD-Cd with 10 mer or 19 mer DNA no new peaks were observed in HSQC spectra, but also resonances corresponding to residues exhibiting a severe drop in intensity showed a very small and insignificant shift (<0.05 ppm). Thus, the possibility of a simple slow exchange or fast exchange process taking place was eliminated. However, a slight broadening of crosspeaks was observed, suggesting that the free and bound forms of the protein may be in intermediate exchange. In order to confirm such a timescale of exchange, intensity and line width changes observed in titration spectra were recreated using an in house

line shape simulation program. This program required line widths for the DNA free state as input, which were determined from experimental resonance peaks. Line widths for the DNA bound state and the difference in frequency between the two states was also required. These were unknown, as peaks for the bound form were not observed. Values for these parameters were set to satisfy two scenarios that may explain the absence of signal for the bound form of the protein.

Scenario 1: Existence of multiple conformations

DNA binding to DnaD-Cd in different registers could give rise to multiple states of the DNA bound form. Simulations were set to assume that these multiple conformations were in exchange with each other, or that the populations of these multiple forms were very small with similar chemical shift values (~0.01 ppm difference). The line widths for the bound state was set to be very broad and the chemical shift difference was set at either 300 Hz (0.5 ppm) for clear separation or 60 Hz (0.01 ppm) for separation that was undetectable. As shown in Figure 6.7a and b, in this situation the peak for the bound form would be very broad. So, if peaks were present for DnaD-Cd-DNA they would be indistinguishable from the signal noise.

Scenario 2: An incomplete titration on the intermediate timescale

This second scenario explored the possibility of an incomplete titration. It was assumed that only one conformation of the bound protein exists. The number of DNA molecules DnaD-Cd binds at any one time was unknown; therefore it was possible that the protein was not 100% bound at the maximum DNA concentration used. Simulations set to capture this scenario had a smaller line width (40 Hz) for the bound form of the protein and the chemical difference between the two states was set as before. As shown in Figure 6.7 c and d, the peak for the DNA bound protein would begin to sharpen as the DNA concentration is successively increased.

Titrations of DnaD-Cd215 with 10 mer DNA, up to a protein:DNA stoichiometry of 1.2 implies that scenario 1 of conformational exchange was more plausible, as new peaks did not appear in the HSQC spectra, despite the DNA concentration being twice that of the protein. Basically, the changes (line broadening, signal intensity and

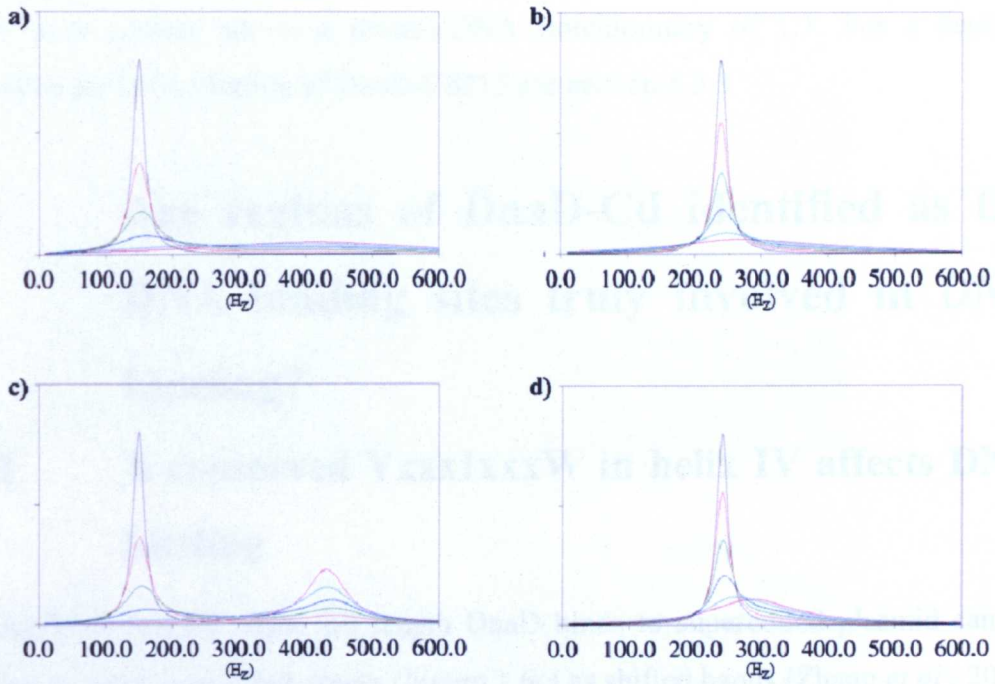


Figure 6.7: Line shapes simulated on the intermediate timescale in order to reproduce DnaD-Cd titration curves that show a big drop in intensity. Line simulations for (a and b) scenario 1 and (c and d) scenario 2. Simulations were run with a large (300 Hz) and a small (60 Hz) chemical shift difference between the free and the bound forms of the protein. Resonances for the free protein were set at 150 Hz or 250 Hz. The peaks in black represent 100% free protein. In red, green, blue, cyan and magenta the population of protein that is DNA bound is 20%, 40%, 60% and 90% respectively. Parameters set for each simulation are shown in Table 6.0.

Table 6.0: Parameters set for each simulation.

Parameters	Simulation			
	a	b	c	d
DNA dissociation rate	200 s^{-1}	60 s^{-1}	200 s^{-1}	60 s^{-1}
Line width for DNA free protein	20 Hz	20 Hz	20 Hz	20 Hz
Line width for DNA bound protein	300 Hz	300 Hz	40 Hz	40 Hz
Chemical shift difference	300 Hz	60 Hz	300 Hz	60 Hz

chemical shifts) observed were very similar to those in DnaD-Cd titration spectra, which were carried out to a protein:DNA stoichiometry of 1:1. For a detailed discussion on DNA titration of DnaD-Cd215 see section 6.3.3

6.3 Are regions of DnaD-Cd identified as the DNA binding sites truly involved in DNA binding?

6.3.1 A conserved YxxxIxxxW in helix IV affects DNA binding

The scaffolds formed when full length DnaD binds to supercoiled plasmid can be detected in agarose gel shift assays (Figure 1.6c) as shifted bands (Zhang *et al.*, 2005; Carneiro *et al.*, 2006). Therefore, to test the involvement of the highly conserved YxxxIxxxW motif in DNA binding, the DNA binding behaviours of two single mutants of full length DnaD, Y180A and W188A were examine using gel shift assays. The mutant proteins were created and purified as described in Carneiro *et al.*, 2006 in the laboratory of Panos Soultanas, University of Nottingham. Throughout the purifications, both Y180A and W188A behaved the same as the wt DnaD indicating that there was nothing different/strange about their stability or overall fold. Gel shift assays were also performed in the laboratory of Panos Soultanas, University of Nottingham. In these assays, the Y180A and W188A DnaD mutant proteins were observed to shift plasmid up the agarose gel but at higher concentrations relative to native DnaD, suggesting that they were both defective in DNA binding (Figure 6.8a). Consistent with this observation, comparative gel shift assays showed that Y180A and W188A were also defective in binding to dsDNA and ssDNA, although the defect was more pronounced with the Y180A than W188A mutant (Figure 6.8b). In light of these observations it was concluded that both Y180 and W188 were involved in DNA-binding.

6.3.2 Helix IV alone is not sufficient for DNA binding

In order to examine whether a DnaD polypeptide up to and including Helix IV but

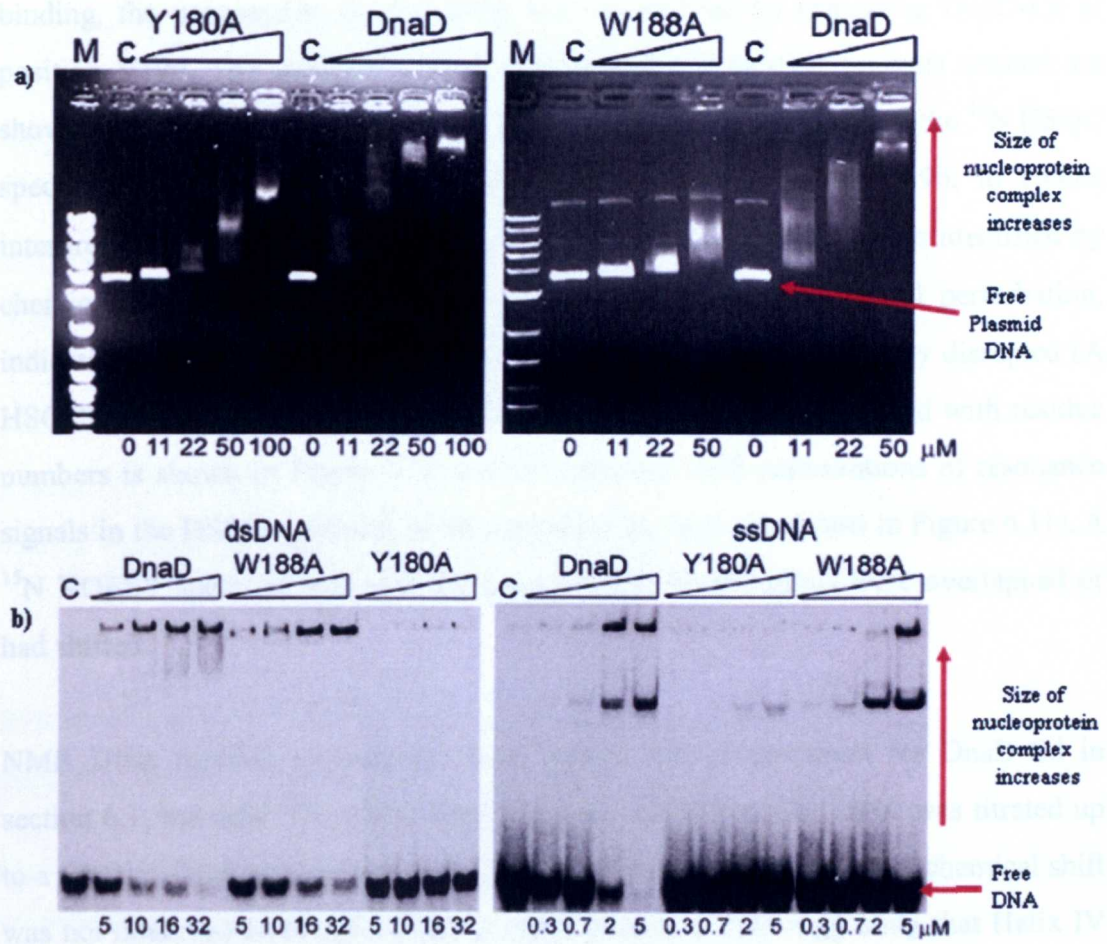


Figure 6.8: The DnaD mutant proteins, Y180A and W188A are defective in DNA binding compared to wild type DnaD. (a) Arose gel shift assays of DnaD, Y180A and W188A complexes with supercoiled pBSK plasmid. Binding reactions were carried out with increasing concentrations of proteins, as indicated in 50 mM Tris, pH7.5, 2 mM EDTA, 1 mM DTT, 350 mM NaCl and 4.4 µM pBSK. Lane M indicates the 1Kb DNA ladder from New England Biolabs. The DNA ladder band sizes are: 10.0 Kb, 8.0 Kb, 6.0 Kb, 5.0 Kb, 4.0 Kb, 3.0 Kb, 2.0 Kb, 1.5 Kb, 1.0 Kb and 0.5 Kb Lane C indicates the control reaction which is pBSK plasmid. (b) Polyacrylamide gel shift assays directly comparing the binding of DnaD and the Y180A, W188A mutant proteins to ssDNA and dsDNA. The oligonucleotide Y180AF (5'-CTCAGTTTCCGCGCGATTGACCG GATT-3') was used as ssDNA and the complementary Y180AR (5'-AATCCGGTCAATCGCGCGGAAACTG AG-3') was used to prepare the dsDNA. Binding reactions were carried out in 100 mM NaCl, 50 mM Tris pH 7.5, 1 mM EDTA, 1 mM DTT, 2.2% v/v glycerol with 0.625 nM ssDNA or 0.125 nM dsDNA probes at increasing concentrations of proteins, as indicated (Image provided by Panos Soultanas).

lacking Helix V and the unstructured C-terminal region was adequate for DNA binding, the polypeptide DnaD-Cd196 was constructed by truncating DnaD-Cd at position V196. The sequences of all the DnaD-Cd truncation mutants created are shown in Figure 6.9. The assignment of amide resonance signals from the ^{15}N HSQC spectra of DnaD-Cd was transferred to ^{15}N HSQC of DnaD-Cd196, to enable interpretation of NMR data on a per residue bases. Most signals were identified by chemical shift matching since the spectrum showed relatively small perturbation, indicating that the core structure of DnaD-Cd196 had not been greatly disrupted (A HSQC spectrum of DnaD-Cd196 where resonance signals are labelled with residue numbers is shown in Figure 6.10 and the chemical shift perturbations of resonance signals in the HSQC spectrum of all truncation mutants are shown in Figure 6.11). A ^{15}N NOESY spectrum was also used, particularly when signals were overlapped or had shifted.

NMR DNA titration experiments were carried out as described for DnaD-Cd in section 6.1, but only ^{15}N HSQC spectra were collected and the DNA was titrated up to a protein: DNA stoichiometry of 1:2. A change in signal intensity or chemical shift was not observed across this titration series (Figure 6.12a) suggesting that Helix IV alone was insufficient for DNA binding, despite it being part of the DNA-binding module. Perhaps, elements of the DNA binding module work cooperatively and thus the entire module is needed to enable DNA binding of DnaD-Cd. Titration curves for DnaD-Cd196 are shown in Figure 6.13. Gel shift assay with ssDNA and dsDNA further confirmed this DNA binding deficiency of DnaD-Cd196, as bands representing nucleoprotein complexes were not observed (Figure 6.12b).

6.3.3 Helix V and part of the unstructured C-terminal region are involved in DNA binding

Two more truncations of DnaD-Cd, DnaD-Cd206 and DnaD-Cd215 were constructed to confirm that Helix V and part of the unstructured C-terminus up to residue 215 participated in DNA binding. As for the truncation mutant, DnaD-Cd196, amide resonance signals in ^{15}N HSQC spectra of DnaD-Cd206 and DnaD-Cd215 were identified. A HSQC spectrum where resonance signals are labelled with residue

DnaD-Cd

GSHM LYTIFEEEFA RPLSP LECETLAIWQDQDQ HD AQLIKHALKEAVLS
GKLSF RYIDRILFEWKKN GLKT VEQAKIHSQKF RRVQ AKQNEPQKEY
KRQVPFYNWLEQ

DnaD-Cd215

GSHM LYTIFEEEFA RPLSP LECETLAIWQDQDQ HD AQLIKHALKEAVLS
GKLSF RYIDRILFEWKKN GLKT VEQAKIHSQKF RRVQ AKQNE

DnaD-Cd206

GSHM LYTIFEEEFA RPLSP LECETLAIWQDQDQ HD AQLIKHALKEAVLS
GKLSF RYIDRILFEWKKN GLKT VEQAKIHSQKF

DnaD-Cd196

GSHM LYTIFEEEFA RPLSP LECETLAIWQDQDQ HD AQLIKHALKEAVLS
GKLSF RYIDRILFEWKKN GLKT V

Figure 6.9: Sequence of DnaD-Cd and its truncation mutants. The five helixes are highlighted in blue and residues that remain after cleavage of the His-tag in green.

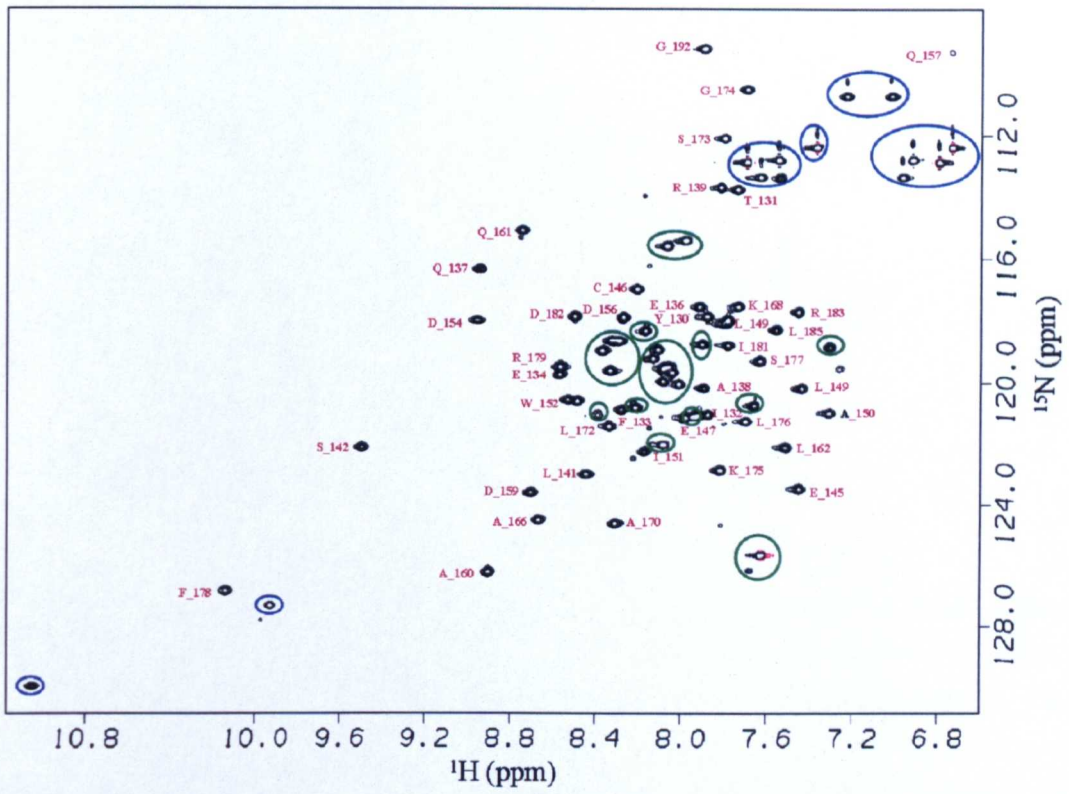


Figure 6.10: ^{15}N HSQC spectrum of DnaD-Cd196. Resonance signals labelled in red were identified by comparison to a ^{15}N HSQC spectrum of DnaD-Cd and with the help of a ^{15}N edited NOESY spectrum. Signal that could not be deduced are circled in green and signals from side-chains are circled in blue.

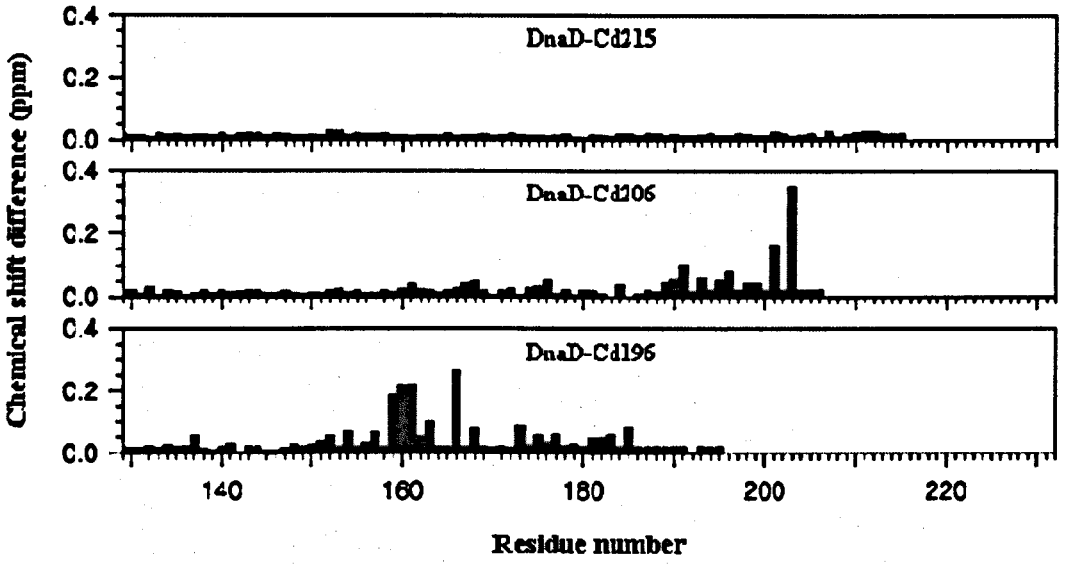


Figure 6.11: Chemical shift perturbations in ^{15}N HSQC spectra of DnaD-Cd truncation mutants. The difference in backbone amide chemical shits assignments between DnaD-Cd and the truncation mutants are plotted per residues. Bars in black show the chemical shift difference for residues whose amide resonance signal were assigned in spectra of the truncation mutants, bars in red show the chemical shift difference for residues whose amide resonance signal were not assigned in spectra of the truncation mutants and bars in green show residues for which the amide resonance was not assigned during backbone assignment of DnaD-Cd.

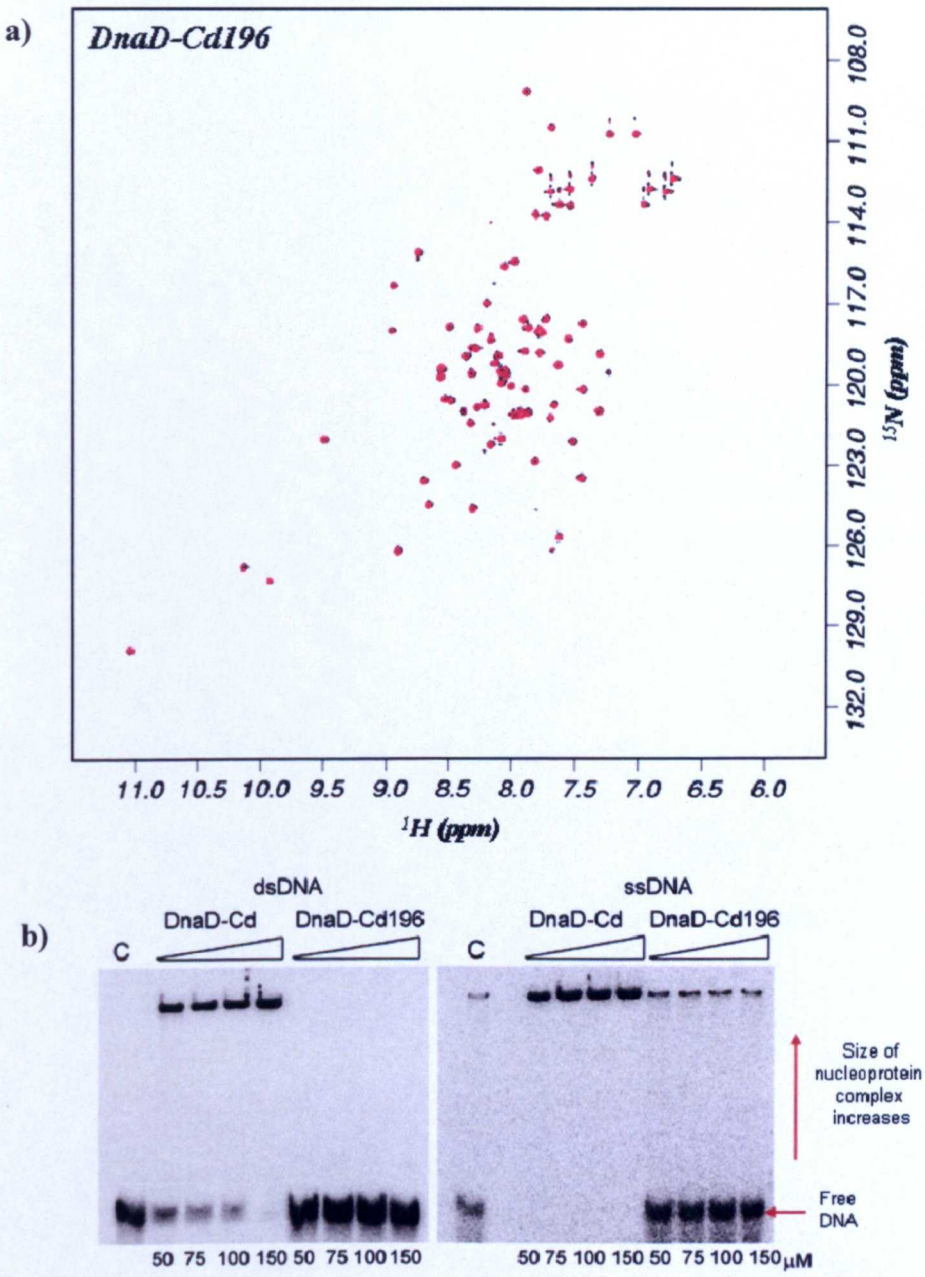


Figure 6.12: DnaD-Cd196 is deficient in DNA binding. (a) ^{15}N HSQC spectra of 0.3 mM DnaD-Cd196 titrated with 10 mer DNA. Spectra at 0μM (black) and 300μM (red) DNA are overlaid. (b) Gel shift assays of DnaD-Cd and DnaD-Cd196 with ss and dsDNA. Binding reactions were carried out in 30 mM NaCl, 50 mM Tris pH 7.5, 1 mM EDTA and 1 mM DTT with 0.625 nM ss or dsDNA at increasing concentrations of DnaD-Cd or DnaD-Cd196, as indicated.

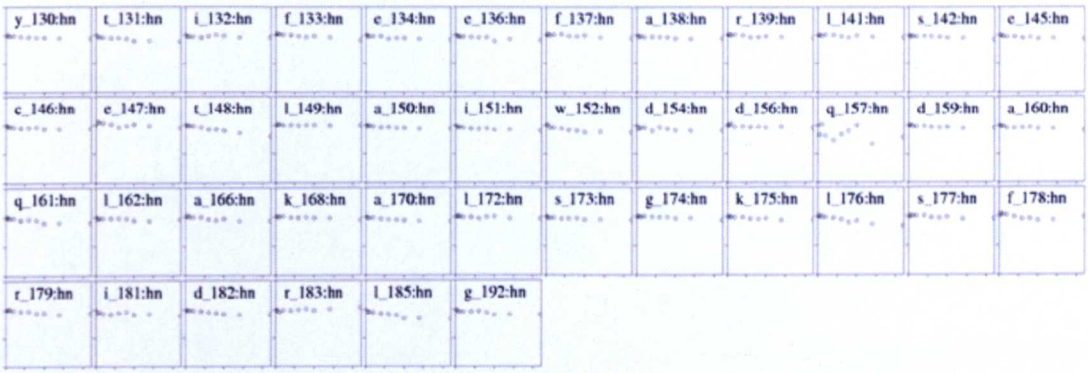


Figure 6.13: NMR titration curves for the titration of 10 mer ssDNA into 0.3 mM DnaD-Cd196. The data are plot as in Figure 6.2, but additional points at DNA concentrations of 420 μM and 615 μM were extracted from ^{15}N HSQC spectra. In each graph, the x-axis covers DNA concentrations from 0 to 615 μM, and the y-axis covers ratios from 0 to 1.2

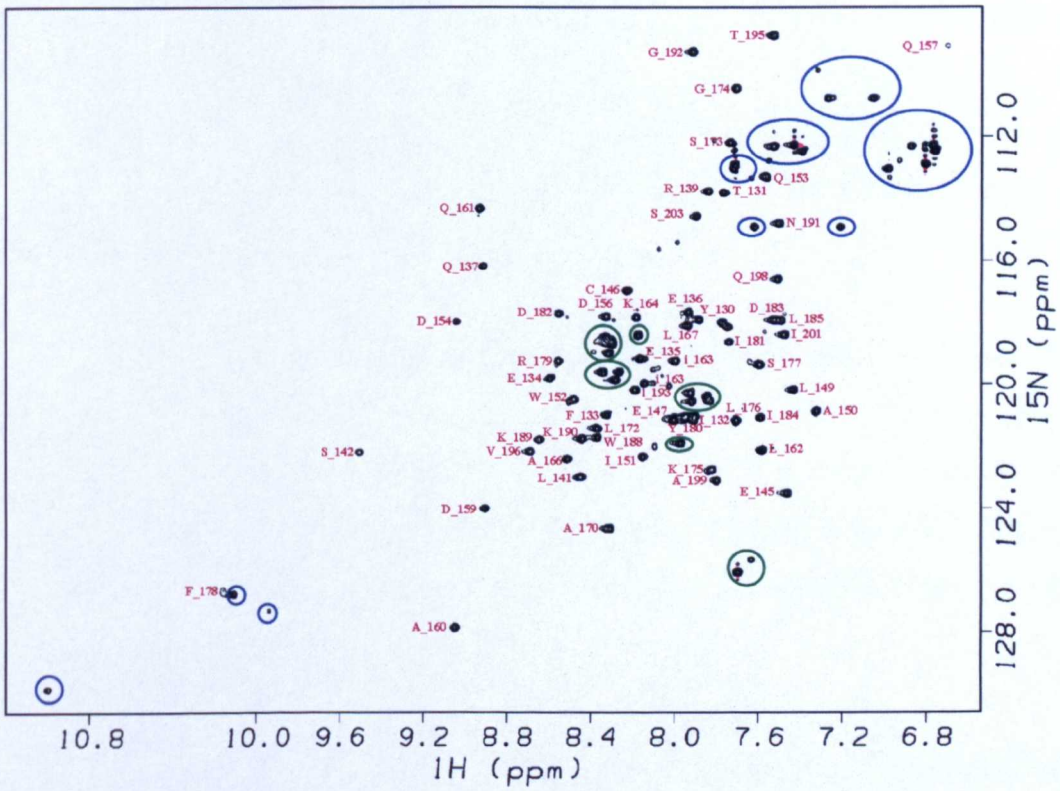


Figure 6.14: ^{15}N HSQC spectrum of DnaD-Cd206. The colour scheme in this spectrum is as in 6.10.

numbers is shown in Figure 6.14 for DnaD-Cd206 and Figure 6.15 for DnaD-Cd215. NMR DNA titrations were carried out as described for DnaD-Cd196 in section 6.3.2. For DnaD-Cd206, signal intensity or chemical shift changes were not observed throughout the titration series (Figure 6.16a and Figure 6.17), implying that it does not bind DNA. Furthermore, gel shift assays showed that DnaD-Cd206 was deficient in DNA binding (Figure 6.19a). In contrast to DnaD-Cd206, titration curves for DnaD-Cd215 (Figure 6.18a) were similar to those of DnaD-Cd. The intensity attenuation was slightly greater and probably attributable to the superior tumbling of the apo-DnaD-Cd215, which reflected in the significantly improved spectra. The mean ^{15}N T_2 of apo-DnaD-Cd and apo-DnaD-Cd215 were 60 ms and 100 ms respectively. The pattern of chemical shift changes was also very similar; with residues towards the C-terminal end showing the greatest shift although still less than 0.05 ppm (Figure 6.20). Gel shift assays, verified that DnaD-Cd215 had DNA binding activity, but also established that it was considerably weaker than full length DnaD-Cd (Compare Figure 6.12b and Figure 6.19a). Taken together, these data showed that helix V and part of the unstructured C-terminus (residues 206-215) was essential for DNA binding and together with YxxxIxxxW motif in helix IV made up the complete DNA-binding module of DnaD.

^{15}N relaxation experiments were performed before and after titrating 10 mer DNA with DnaD-Cd215 in order to estimate the size of the protein:DNA complexes formed. A T_2 of 100 ms for apo-DnaD-Cd215 had reduced to 50 ms for DNA bound DnaD-Cd215 (Figure 6.19b). This reduction implies that DNA-protein complexes that form are small and possibly no more than 2-3 DnaD-Cd215 and /or DNA molecules.

6.4 DNA-induced oligomerisation: Reconciling NMR and gel shift data

Gel shift assays show that DnaD-Cd forms very large nucleoprotein complexes on binding DNA whereas, NMR DNA titration data implies that the size of such complexes is much smaller. This discrepancy between gel shift and the NMR experiments could be explained if more DnaD-Cd ("P") rather than oligonucleotide

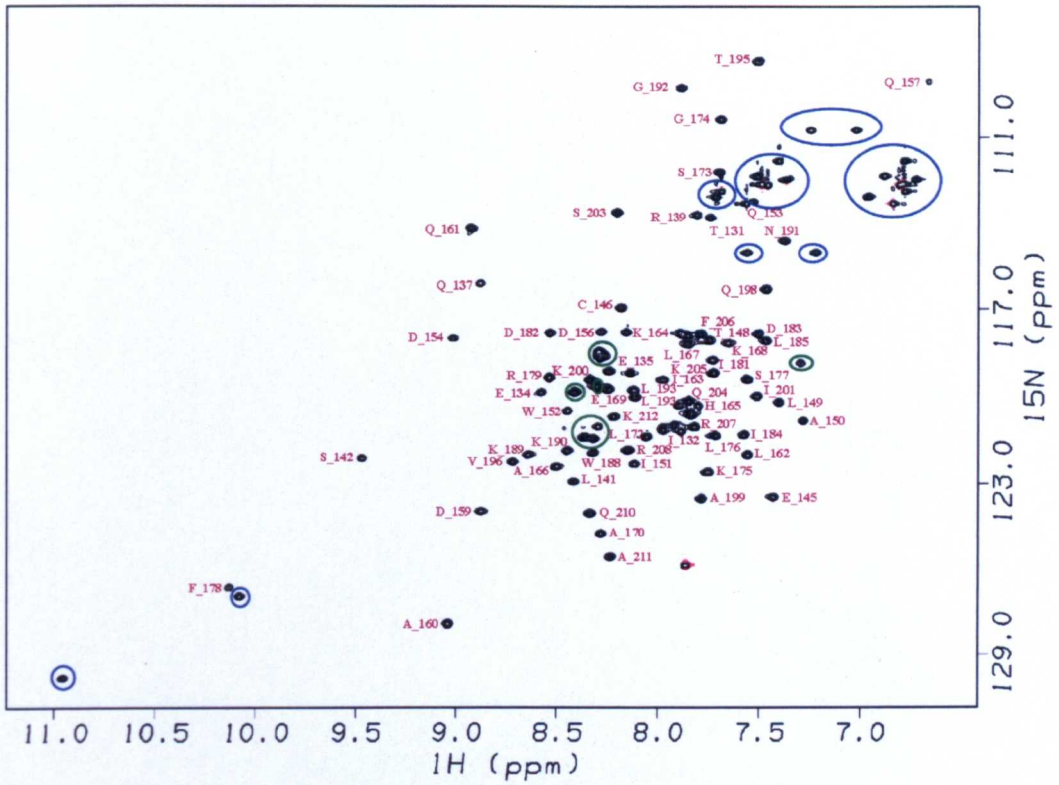


Figure 6.15: ^{15}N HSQC spectrum of DnaD-Cd215. The colour scheme in this spectrum is as in 6.10.

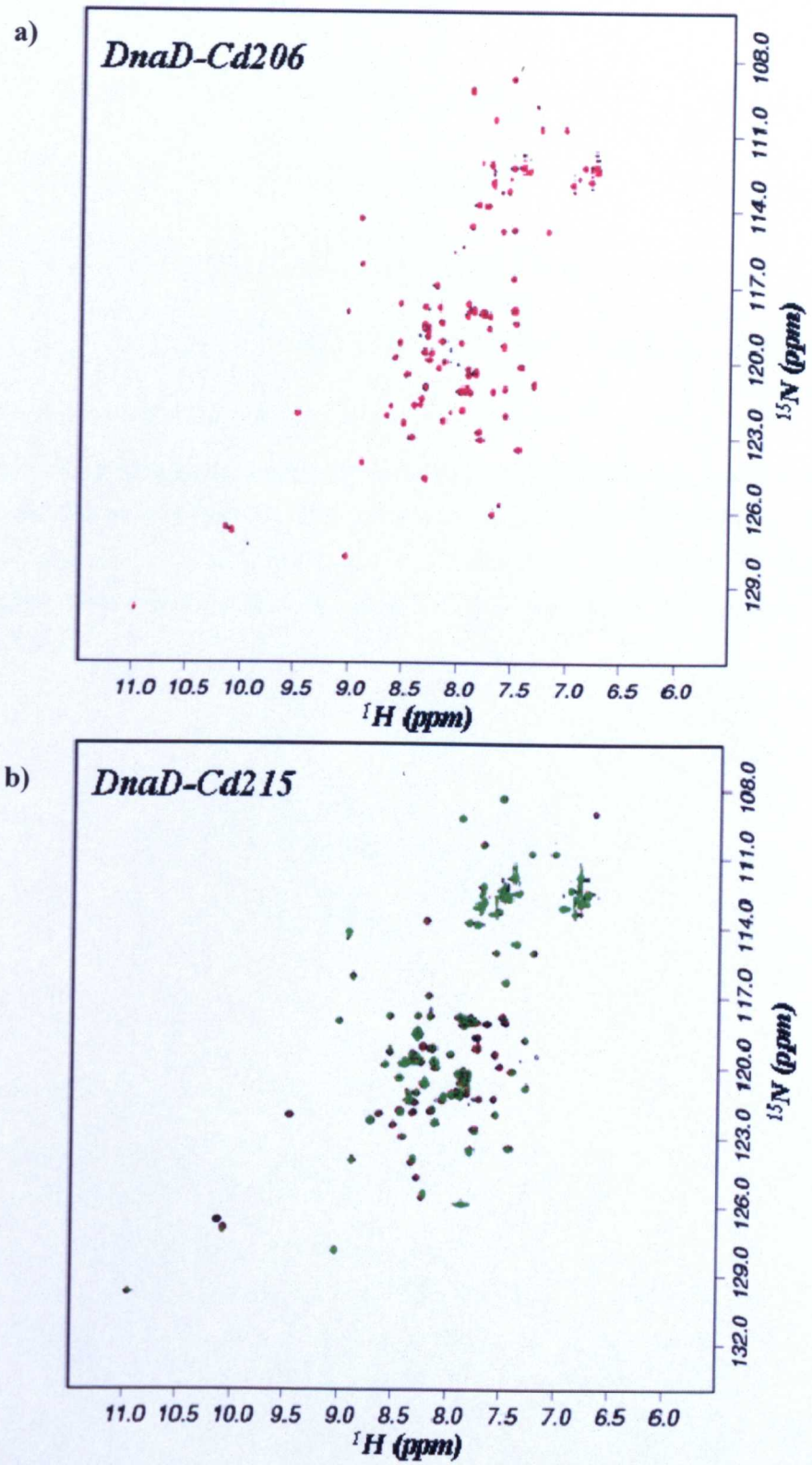


Figure 6.16: ^{15}N HSQC titration spectra of DnaD-Cd206 and DnaD-Cd215 with 10mer DNA. The colours of the overlaid spectra are as in Figure 6.8a and Figure 6.1a for DnaD-Cd206 and DnaD-Cd215 respectively.

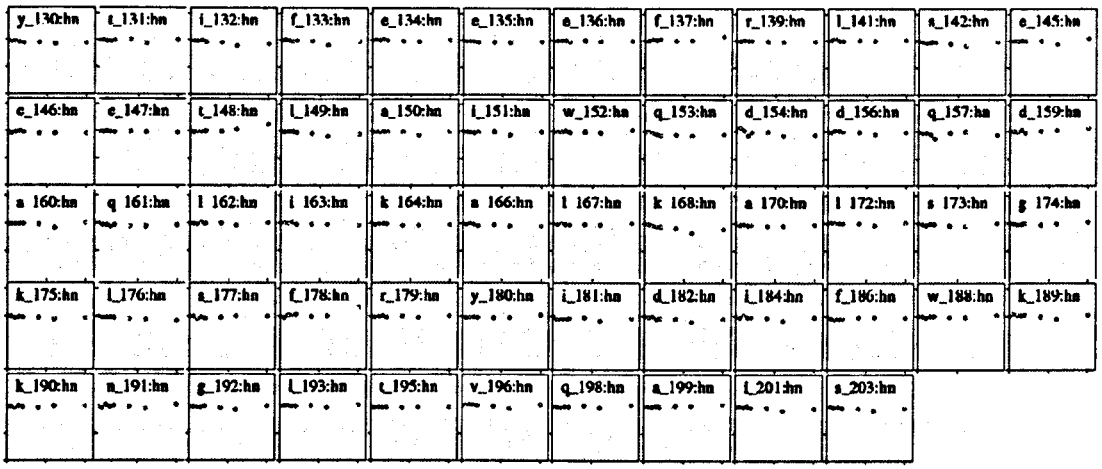
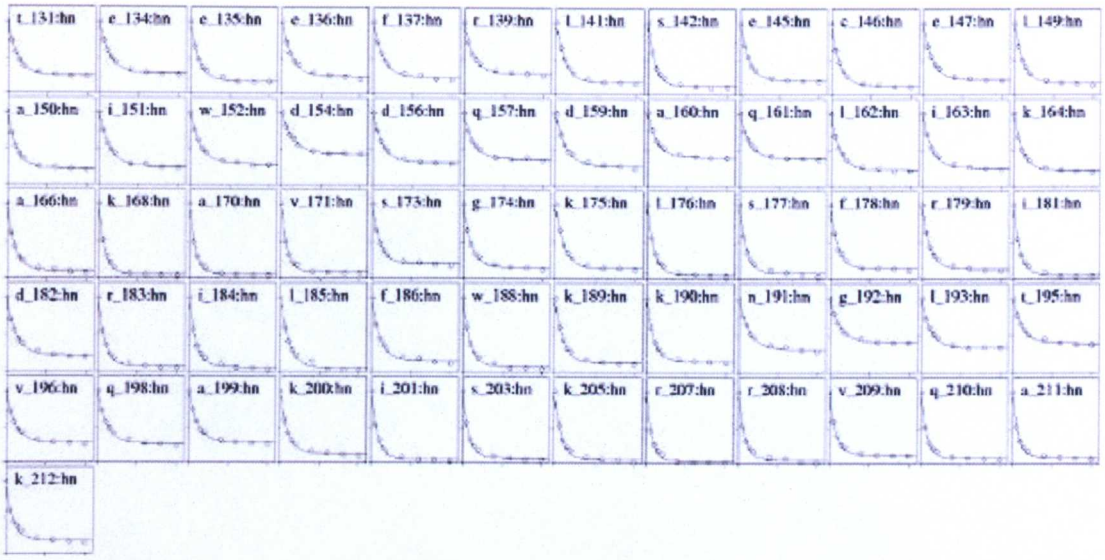
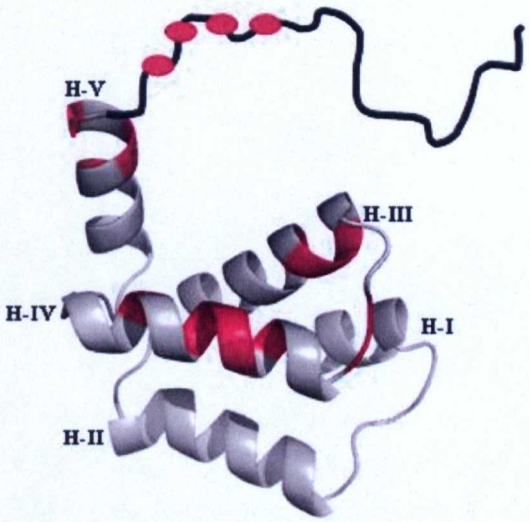


Figure 6.17: NMR titration curves for the titration of 10 mer ssDNA into 0.2 mM DnaD-Cd206. The data are plot as in Figure 6.2. Data points were extracted from ^{15}N HSQC spectra for DNA concentrations of 0 μM, 17.5 μM, 35 μM, 52.5 μM, 70 μM, 1400 μM, 210 μM, 280 μM and 350 μM. In each graph, the x-axis covers DNA concentrations from 0 to 385 μM, and the y-axis covers ratios from 0 to 1.6.

a)



b)



130 140 150 160 170

LYTIFEEEFARPLSPLECETLAIWQDQDQHDAQLIKHALKEAVLS

180 190 200 210

GKLSFRYIDRILFEWKKNGLKTVEQAKIHSQKFRRVQAKQNE

Figure 6.18: Intensity changes on titrating 10 mer DNA into DnaD-Cd215 (a) NMR ^{15}N HSQC intensity titration curves for titration of 10 mer DNA into 0.4 mM DnaD-Cd215. The data and axis for each graph are plotted as in Figure 6.2a. The slightly larger intensity drops in this titration set compared to DnaD-Cd is probably attributable to the superior tumbling of the apo-protein (The mean ^{15}N T_2 of DnaD-Cd is ~ 60 mM and of DnaD-Cd215 is ~ 100 ms). (b) The model of DnaD-Cd and sequence of DnaD-Cd215 showing, in red residues for which crosspeak intensity was reduced by more than 90% after the addition of 1:1 10 mer DNA to DnaD-Cd215. These are residues 168, 170, 171, 176, 177, 181, 183, 184, 185, 188, 201, 203, 205, 207, 208, 210 and 211. On the model, the mobile portion of the protein is depicted in sketch and the C-terminal end on the helices is labeled. The black boxes on the sequence highlight helix boundaries.

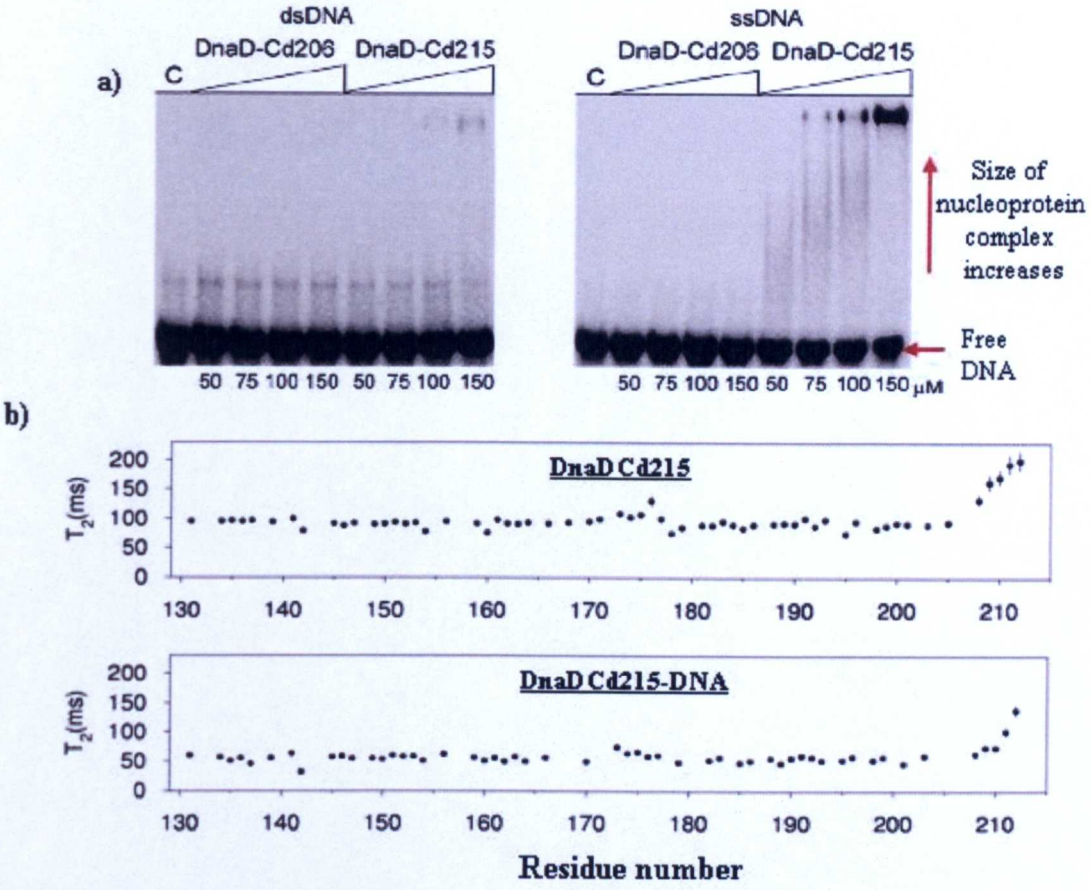


Figure 6.19: DnaD-Cd215 binds DNA but DNA DnaD-Cd206 does not (a) Gel shift assays of DnaD-Cd, DnaD-Cd206 and DnaD-Cd215 with ssDNA and dsDNA. Binding reactions were carried out as described in Figure 6.8b with increasing concentrations of DnaD-Cd206 and DnaD-Cd215, as indicated. (b) ^{15}N T_2 relaxation plots for apo-DnaD-Cd215 and DNA bound DnaD-Cd215. The reduced ^{15}N T_2 value of 50 ms for DNA bound DnaD-Cd215 implies that DNA-protein complexes are composed of no more than 2-3 DnaD-Cd215 and/or DNA molecules.

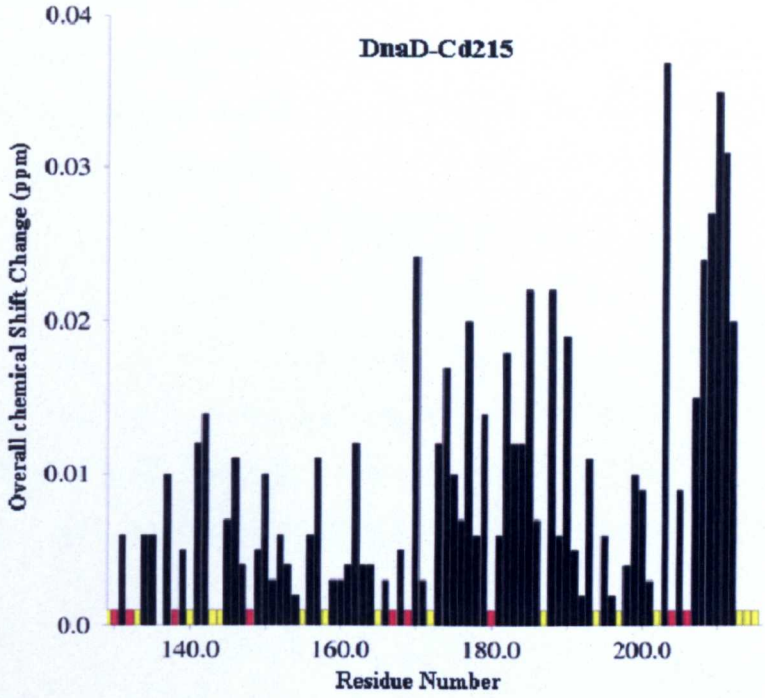


Figure 6.20: Chemical shift changes during titration of 10 mer ssDNA into DnaD-Cd215. The overall amide chemical shift change observed in ^{15}N HSQC spectra of DnaD-Cd215 on addition of DNA is plotted per residues. Bars in black show the overall chemical shift change for residues whose amide resonance signal were assigned, bars in red show residues that were heavily overlapped and bars in yellow show residues for which the amide resonance was not assigned during backbone assignment of DnaD-Cd, but also residues that were not identified by chemical shift matching and comparison to spectra of DnaD-Cd.

DNA ("O") was present in the large complexes formed in gel shift. For instance, assume that one molecule of DNA binds with a different number of protein molecules to give OP, OP_2 ... OP_n. At equilibrium, the equilibria of the individual species will satisfy the following:

$$[OP_n].[P]^m/[OP_{m+n}]=K_{m,n}$$

Where $K_{m,n}$ is the dissociation constant for the equilibrium between the OP_n and OP_{n+m} species. This would further imply that the ratio of the two species which differ in the number of bound protein molecules (*m*), OP_n and OP_{n+m} is inversely proportional to the free protein concentration [P] to the power *m* (see below).

$$[OP_n]./[OP_{m+n}]=K_{m,n}/[P]^m$$

Therefore, since the free protein in gel shift assays is increasingly in excess over the DNA, the [P] may be sufficiently high that the larger species dominate. As the free protein in NMR experiments becomes depleted on adding DNA, the [P] may be sufficiently low that smaller species dominate.

6.5 Helix V and the C-terminal unstructured region of DnaD is essential for *B. subtilis* viability

Several *in vivo* complementation experiments were carried by Wiep K. Smith (In the laboratory of Alan Grossman at the Massachusetts Institute of Technology in Cambridge, USA) in order to establish whether the mutant DnaD proteins, dnaDY180A, dnaDW188A and dnaDV196-stop, which show altered DNA binding activity *in vitro* were able to function *in vivo* to support cell growth. In the first set of complementation experiments, each mutant protein was assessed for its ability to rescue the growth defect of a *dnaD23ts* strain of *B. subtilis* at the non-permissive temperature. (For experimental details see Marston *et al.*, 2010) From the assessment of growth on LB agar, it was found that *dnaDY180A and dnaDW188A* complemented the growth defect of the *dnaD23ts* strain at the non-permissive temperature, whereas

dnaD-V196-STOP did not (Figure 6.21a). It was unknown whether the complementation of the *dnaD23ts* strain by *dnaDY180A* and *dnaDW188A* was due to true complementation or a result of mixed multimers forming with DnaD23 protein. To clarify this uncertainty, a deletion-insertion mutation of *dnaD* was introduced into strains expressing each of the different *dnaD* alleles (For experimental details see Marston *et al.*, 2010). From the assessment of growth on LB agar, it was found that strains expressing *dnaDY180A* and *dnaDW188A* sustained growth (Figure 6.21b), but strains expressing *dnaD-V196-STOP* did not. Furthermore, the DnaDY180A and DnaDW188A mutant protein had altered activity *in vivo,* even though they were able to complement mutations in *dnaD* that cause loss of function. Taken together, these data indicated that the DnaDY180A and DnaDW188A mutant protein which have reduced DNA binding ability i*n vitro* were likely to have reduced activity *in vivo*, but nonetheless supported cell growth. The inability of dnaD-V196-STOP protein to restore growth implied that an essential function of DnaD, most likely the DNA binding activity, was located in the C-terminal region, after V196.

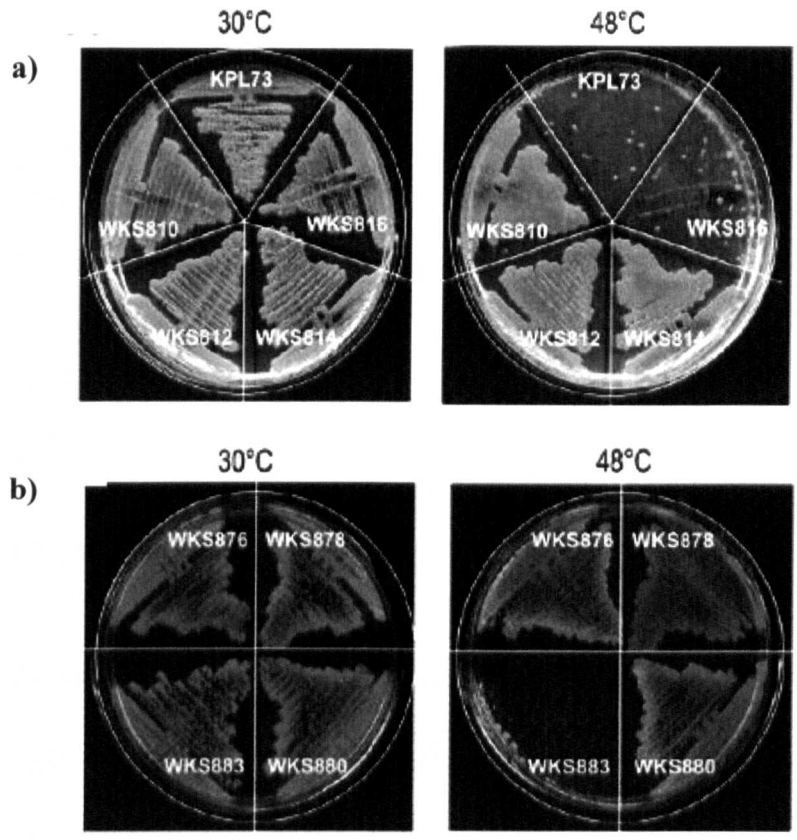


Figure 6.21: *In vivo* activity of DnaD and its mutants. (a) At the non-permissive temperature (48°C), truncated DnaD-V196-STOP does not complement the growth defect of *dnaD23ts*. The *dnaD23ts* strains carrying an ectopic inducible wild type *dnaD* (WKS810), *dnaDY180A* (WKS812), *dnaDW188A* (WKS814) or *dnaD-V196-STOP* (WKS816) were grown overnight on LB agar at temperatures as indicated. The control was a *dnaD23ts* strain without inducible *dnaD* (KPL73). (b) DnaD-Y180A and DnaD-W188A remain unaffected when the native copy of *dnaD* is knocked out. Strains carrying an inducible wild type DnaD (WKS876), Y180A (WKS878), W188A (WKS880) and DnaD23 (WKS883) sustain growth at 30°C and thus are not affected by the deletion of native *dnaD*. LB agar plates were incubated overnight at 30°C or 48°C, as indicated. The WKS883 strain (DnaD23) is shown as control (Marston *et al.*, 2010).

Chapter 7: Domain similarity between DnaD and DnaB

As discussed in chapter 1, the *B. subtilis* protein DnaB has a DNA remodelling activity like DnaD, however its remodelling effect is opposite to that of DnaD. The DnaD protein opens up super-coiled plasmid, whereas DnaB compact's super-coiled plasmid. This chapter discusses the bioinformatics analysis which supports a sequence link between DnaD and DnaB. It also explains how sets of sequences representing DnaD and DnaB related proteins were selected. The functional similarity between the two proteins is also discussed in this chapter. The sequence analysis and figures presented in this chapter were prepared by Dr Jeremy Craven.

7.1 Merger of the DnaD-like family with the DnaB_2 family in the Pfam database

The Pfam database is a large collection of conserved protein families and domains with sequence coverage of 74% (Finn *et al.*, 2008; http://pfam.sanger.ac.uk). Each protein family is represented by a multiple sequence alignment generated from a hidden Markov model (HMM), which is a table that encodes the probabilities of different amino acid being present at each position in the domain. HMMs are also used by Pfam to search sequence databases for homologues to improve alignment of existing families.

In Pfam release 18.0 (October 2006), the C-terminal domain of DnaD was listed as a conserved 'DnaD-like protein motif' under the Pfam DnaD family (Accession No. PF04271). This motif was encoded on a 72 residue pattern corresponding to residues 129-201 of *B. subtilis* DnaD. It was found in a number of bacterial and phage proteins, many with unknown function. Some of these proteins had a single DnaD-like domain (Figure 7.0), while others had two domains in tandem. The exceptions were proteins like the orf4 protein from phage 7201 of *Streptococcus thermophilus,* which had another domain associated with replication initiation called the RepA-like domain, linked to its DnaD-like domain.

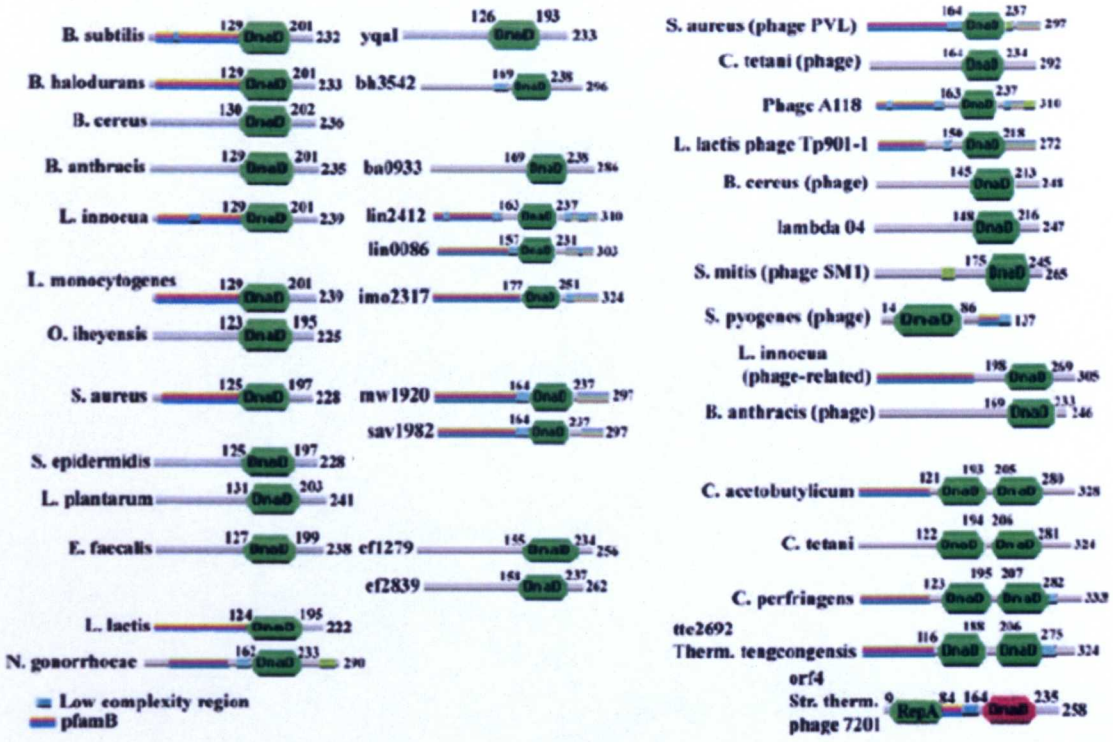


Figure 7.0: Examples of bacteria and phage that have proteins identified by Pfam to have a 'DnaD-like protein motif'. The 'DnaD-like protein motif' is found in all DnaD proteins, examples of bacterial species with a DnaD protein are shown in the left column. Proteins of unknown function and phage proteins that have such a motif are shown in the middle and right column respectively. Proteins with two DnaD-like domains in tandem or with this domain linked to another domain distinct form the DnaD-like domains are shown at the bottom of the right column (Carneiro *et al.*, 2006).

In the recent Pfam release 24.0 (October 2009), the DnaD family 22.0 has been merged with the existing DnaB_2 family (Accession No. PF07261), which encodes a 77 residue pattern that is found on a domain in a region towards the C-terminus of *B. subtilis* DnaB and many other bacterial replication initiation proteins. Since evidence for this merge is undocumented and a sequence link between the C-terminal domains of DnaD and DnaB had not been suggested in the literature before the release of Pfam 24.0, this was investigated.

7.2 DnaB contains domains homologous to DnaD-Nd and DnaD-Cd

BLAST searches against the NCBI nr protein database were performed with DnaD and DnaB as the query sequence to establish whether the sequence link between DnaD and DnaB suggested by Pfam really exists. Despite using separate domains and the full length sequence, the search with DnaD did not yield any significant matches to proteins annotated as DnaB and vice versa for searches with DnaB. However, in iterative PSI-BLAST sequence searches done for DnaD-Cd described in section 1.6.1, sequences homologous to DnaB were detected, but at e-values below typical inclusion thresholds. Pairwise BLAST comparison of the two proteins also hinted a sequence link, as a 25 residue segment which spanned residues K175-A199 of DnaD showed 44% identity. But, much clearer evidence for such a link was obtained using the HHpred server (http://toolkit.tuebingen.mpg.de/HHpred). This protein server detects remote homology between sequences. It first builds a multiple sequence alignment based upon the query sequence using PSI-BLAST and then calculates a HMM for the alignment. It compares this HMM with HMMs built for alignments based on other sequence families such as sequences with structure in the PDB, in order to hypothesise the function, structure, and domain composition of the query protein. These HMM-HMM comparisons are a much more powerful method for detecting distinct relationships than sequence-sequence comparisons (performed by BLAST) or profile-sequence comparisons (performed by PSI-BLAST), as conserved features of two families of sequences are compared rather than features in any one sequence from a particular family. The submission of full length DnaB as the query

to HHpred, yielded the top hit for residues 300-393 to the DnaD-like domain 2zc2 (PDB code 2zc2 e=3.4e-07), which is closely homologous to DnaD-Cd (See section 1.6.1). A much weaker hit was obtained for residues 38-112 to the DnaD N-terminal domain (PDB code 2v79, e=1.1).

In order to verify that the presence of both domains did not influence the selection of these hits, in particular the weaker one, two portions of DnaB spanning residues 1-147 (MALDY....FFTH) and 300-472 (SLDG...YSAY) were submitted to HHpred separately. Residues 306-472 yielded strong hits to the PDB structures 2zc2 and 2i5u (e-values of 10-12 and 10-9, respectively), whilst residues 1-147 yielded the structure of DnaD-Nd as the top scoring hit (e-value of 0.1). The alignments between query and hit HMMs are shown for these two searches in Figure 7.1a and 7.1b respectively.

Other HHpred hits for the N-terminal domain search that had scores marginally lower than the structure of DnaD-Nd were mainly helix-turn-helix motifs, in particular winged helix-turn-helix domains like the DnaD-Nd. This therefore raised the question of whether DnaD-Nd being the top hit was other than a coincidence. If, however, the secondary structure predictions of DnaD and DnaB made by Phyre are compared, it is apparent that DnaB contains very similar elements to DnaD either side of the region aligned by HHpred (Figure 7.3). In particular a helix-strand-helix motif is predicted at the N-terminus of DnaB. These helix-strand-helix elements (H1'-S1'-H2') in DnaD are involved in dimer and tetramer formation (Schneider *et al.*, 2008). Since, DnaB also forms tetramers via its N-terminal domain (Grainger *et al.*, 2010) is seem plausible to assume that DnaD-Nd being the top hit is due to a true link between the two domains rather that just a coincidence. This bioinformatics analysis supports the merger of the pfam DnaD-like domain and the DnaB_2 family based on the C-terminal domains of the DnaD and DnaB proteins, but also discovers that the N-terminal domain of DnaB is sequentially similar to that of DnaD.

To avoid confusion, these two regions of homology are termed DDBH1 and DDBH2 domains (DnaD DnaB Homology 1 and 2) hear after. The relative positions of these domains in DnaD and DnaB are shown in Figure 7.2.

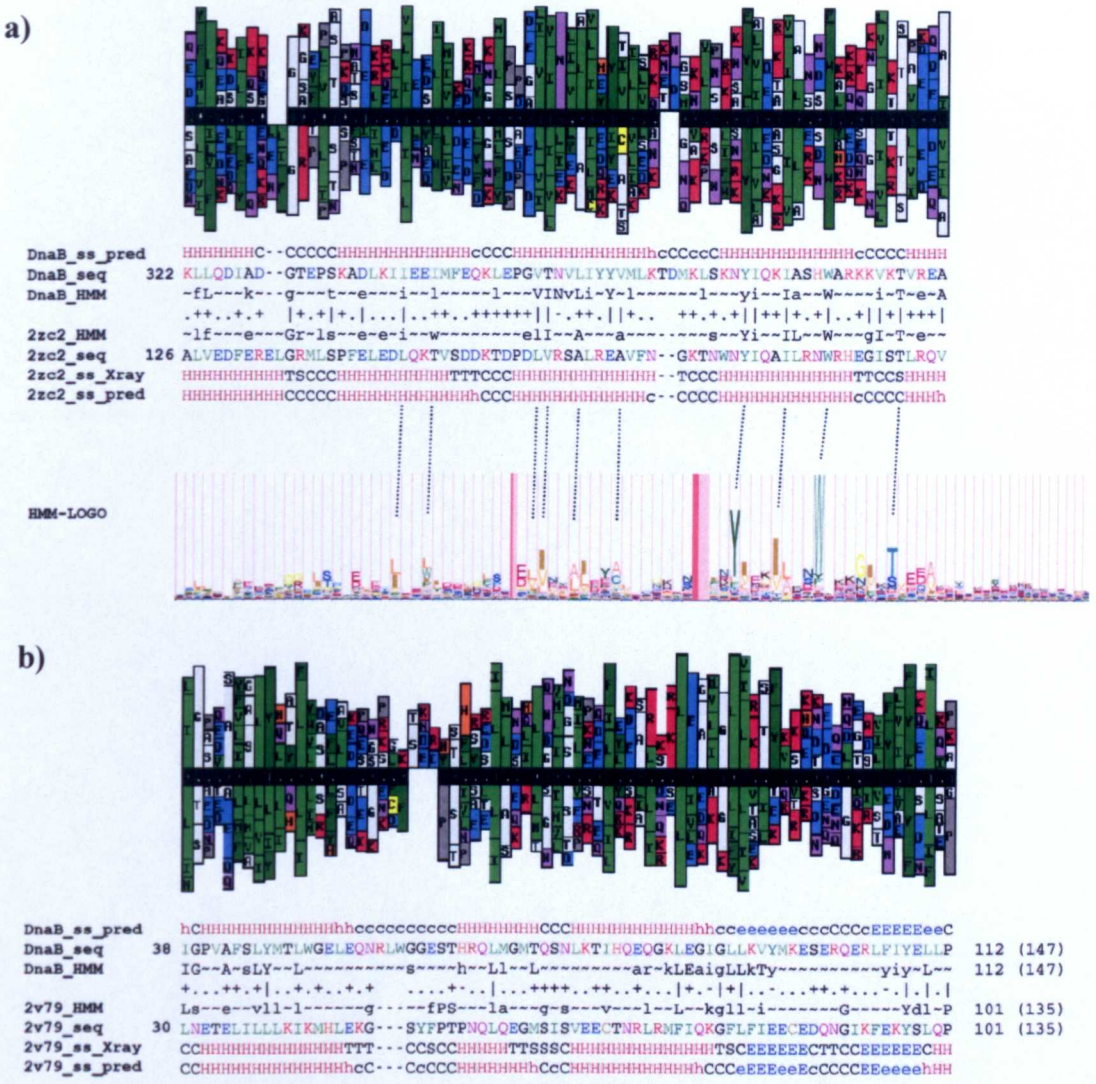


Figure: 7.1: HMM relationships of DnaD and DnaB. The alignments returned by HHpred for (a) residues 306-472 and (b) 38-112 of DnaB with their corresponding top PDB hit are shown. The abbreviations used in these alignments are as follows: ss_pred: secondary structure prediction by PSIPRED; seq: sequence; HMM: HHpred representation of HMM where upper and lower case letters denote amino acids with >60% and >40% probability respectively. Below the top alignment is a HMM LOGO created using the logomat-M server (http://www.sanger.ac.uk/cgi-bin/software/analysis/logomatm.cgi) for the DDBH2 domain with the residue types that have the highest probability between the two HMMs highlighted by dashed lines. Above both alignments are histograms of the DnaB and the PDB HMMs. The bars in these histograms are proportional to the probability of a certain amino acid type at each position. Bars for probabilities < 10% are omitted (Marston *et al.*, 2010).

Figure: 7.2: Domain similarity between DnaD and DnaB. The secondary structure of DnaD derived experimentally (see chapter 5) and the secondary structure of DnaB predicted by PHYRE is shown. Regions for which HHpred detected homology are colour coded. The estimated extent of the DDBH1 and DDBH2 domains in the two proteins are shown above or below the domain graphics by the double headed arrows (Marston *et al.*, 2010).

Figure 7.3: Secondary structure and disorder predictions for DnaD and DnaB. The data contained in these PHYRE predictions are as explained in Figure 1.14. The secondary structure shown above the main text of the predictions is colored as in Figure 7.2. The location of the YxxxIxxxW motif is shown boxed. For DnaD, the flanking H1'-S1'-H2' and H3' regions are also shown. Note that PHYRE poorly predicts S1' for DnaD and the length of the first strand of the beta-hairpin predicted for DnaB is also vague. The thin and thick tickmarks are at 10 and 50 residue intervals respectively (Marston *et al.*, 2010).

7.3 DnaB contains a central DDBH2 domain, giving it a domain arrangement of (DDBH1)-(DDBH2)-(DDBH2)

Full length DnaB has been shown to resolve into three domains on trypsin proteolysis; an N-terminal domain, a central domain and a C-terminal domain (Grainger *et al.*, 2010). Since the HHpred search with full length DnaB yielded hits only for the N-and C-terminal domains, residues 146-305 (THPA...AEED) corresponding to the central domain were submitted to this HHpred server. Surprisingly, the top scoring hit for this sequence was the structure of 2zc2 again, although with a poor e-value of 2.3. Usually, a hit with an e-value as such would be considered insignificant, but taking into account that the scoring of HHpred could not be influenced by the strongly matching domain in DnaB because a portion of the sequence which did not overlap this domain was used, made it a very convincing find. Therefore, the central domain of DnaB is a divergent DDBH2 domain and thus another piece of the link between the DnaD and DnaB sequences, which has not been picked up by the Pfam server.

Based on the sequence analysis, a domain structure for DnaB was proposed and is shown in Figure 7.2. This figure shows that the majority of the secondary structure elements predicted for DnaB are analogous to those that make up the N- and C-terminal domains of DnaD. Furthermore, these predications are consistent with the pattern of conservation across the *B. subtilis* DnaB sequences. A region at the extreme C-terminus of DnaB (residues 443-463) predicted as helical has been left unclassified in this domain graphic. Also, two conserved regions (*B. subtilis* DnaB residues 155-163 and 425- 433) that have been predicted as ordered but have uncertain secondary structure are not marked on this domain graphic.

7.4 The DnaD/DnaB superfamily

In order to collect a large set of sequences that would represent DnaD and DnaB related proteins, the Uniprot50 sequences from Pfam that show 50% maximum sequence identity were iteratively searched with the C-terminal domains of firstly,

DnaD (residues 129-232) and then of DnaB (residues 306-472) as the query, using jackhammer from the hmmer3 package. Jackhammer uses a HMM based on the query sequence to carry out its first search and then builds a new HMM from sequences returned by the search. Using this new HMM, the search is repeated until sequence coverage is obtained. Searches with DnaD-Cd and DnaB-Cd, yielded 313 and 318 sequences respectively, which were strongly overlapping. Such an overlap of sequences is expected when two query sequences are truly homologous and the search method performs an extremely thorough search for homologues. On removal of these overlapping sequences, a combined set of 326 sequences was obtained that represented DnaD-Cd and DnaB-Cd. Using ClustalW, the full length sequences corresponding to this combined set were aligned and then the pairwise identities of each sequence in this alignment was individually calculated to the sequence of DnaD or DnaB. The pairwise identities are plotted in Figure 7.4. These plots clearly showed that a subset of sequences from this combined set had a similarity to DnaD above the baseline value, and a separate subset had a similarity to DnaB, again above the baseline value. Furthermore, the lengths of these sequences in these two subsets showed good regularity. Therefore, these subsets of sequences were selected as the working set of DnaD and DnaB orthologues. Sequence alignments of the DnaD and DnaB sequences are shown in Figure 5.9 and Figure 7.5 respectively.

A HMM was built from the alignment of the full combined set of 326 DnaD and DnaB sequences for the region that corresponded to the DDBH2 domain in DnaD, using the hammer3 package (Figure 7.1). The HMM-LOGO, which is a graphical interpretation of HMMs that highlights regions distinguishing homologous subfamilies from each other, showed that the YxxxIxxxW motif known to participate in DNA binding in DnaD and a highly conserved threonine (or serine) positioned seven residues C-terminal of this motif were most dominant in this full combined set of sequences. The significance of these features in the DnaD and DnaB proteins are addressed in the discussion below.

7.5 Discussion

The merger of the DnaD-like domain family with the DnaB_2 families in Pfam is

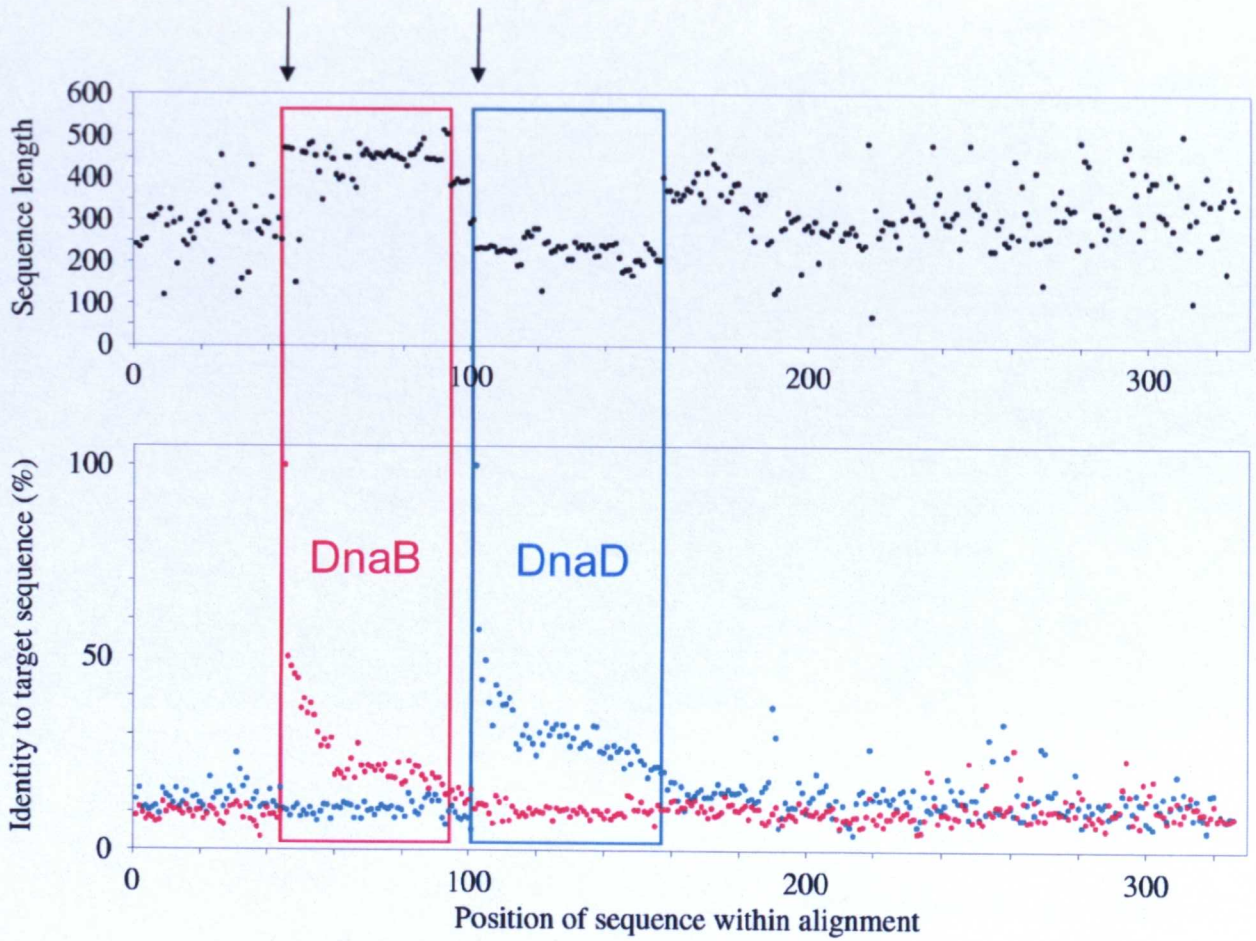


Figure 7.4: Selecting putative DnaB or DnaD sequences. Analysis of the ClustalW alignment of the combined collection of 326 sequences is show graphically. The top graph shows the length of all sequences in the alignment. The bottom graph shows the % identity to *B. subtilis* DnaB (red) and *B. subtilis* DnaD (cyan). In both graphs, the data is ordered according to position (row) within the alignment. The position of the DnaB proteins is marked with the arrows (Marston *et al.*, 2010).

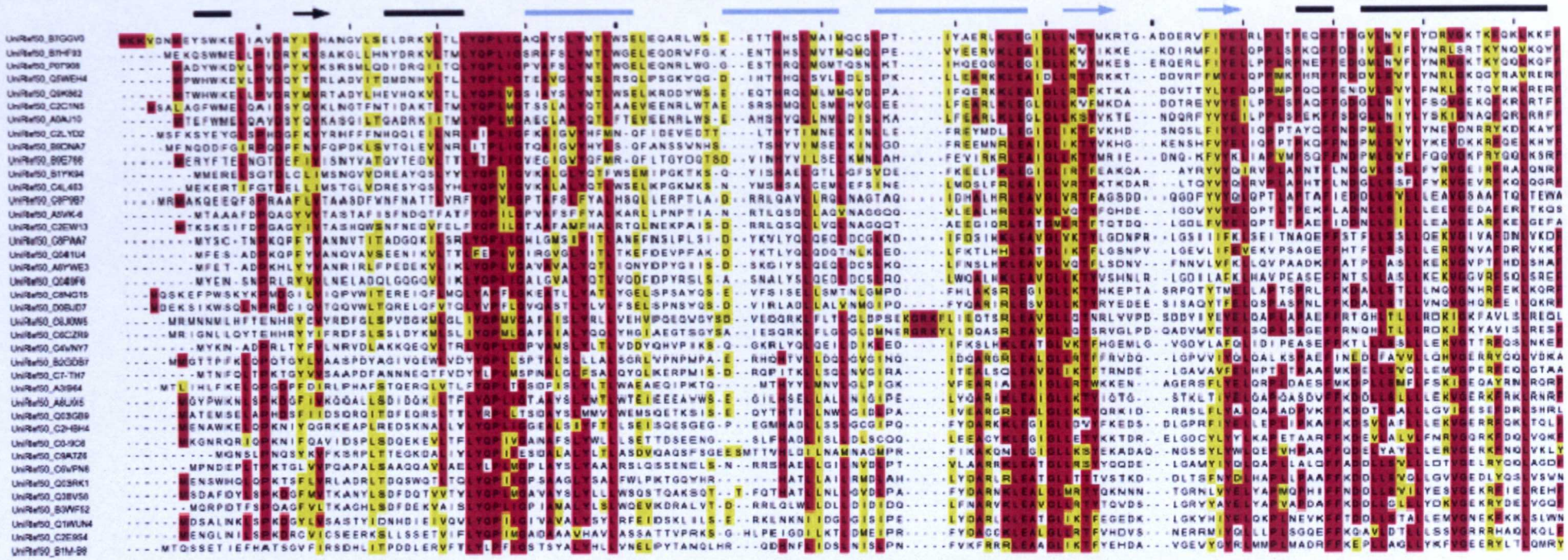
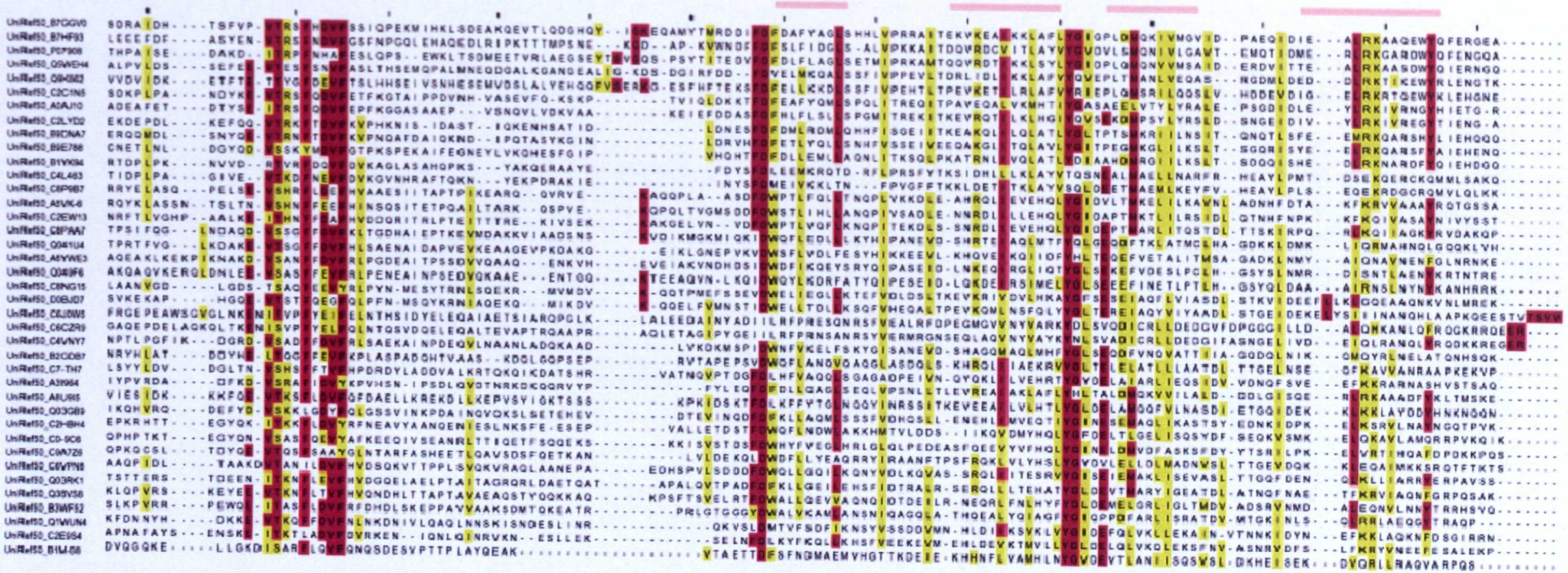


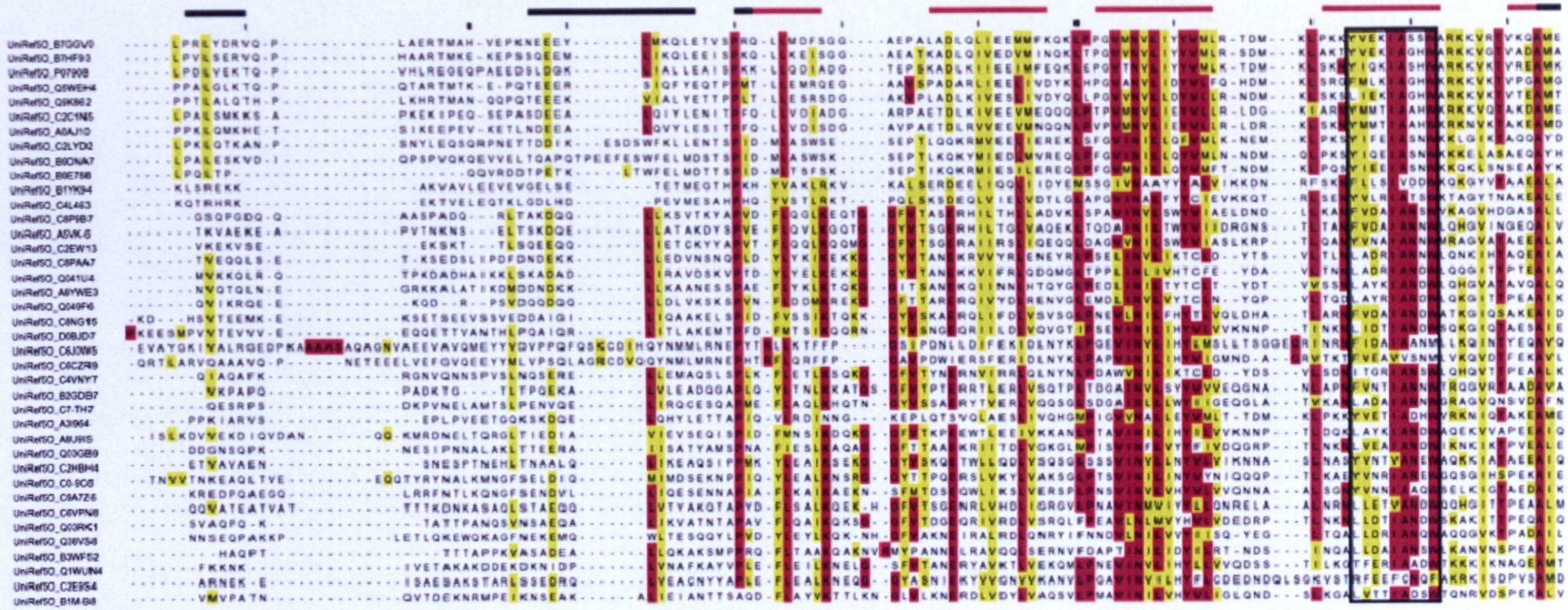
Figure 7.5 continues onto next page

Figure 7.5 continued from previous page



Continues onto next page

Figure 7.5 continued from previous page



Continues onto next page

Figure 7.5 continued from previous page

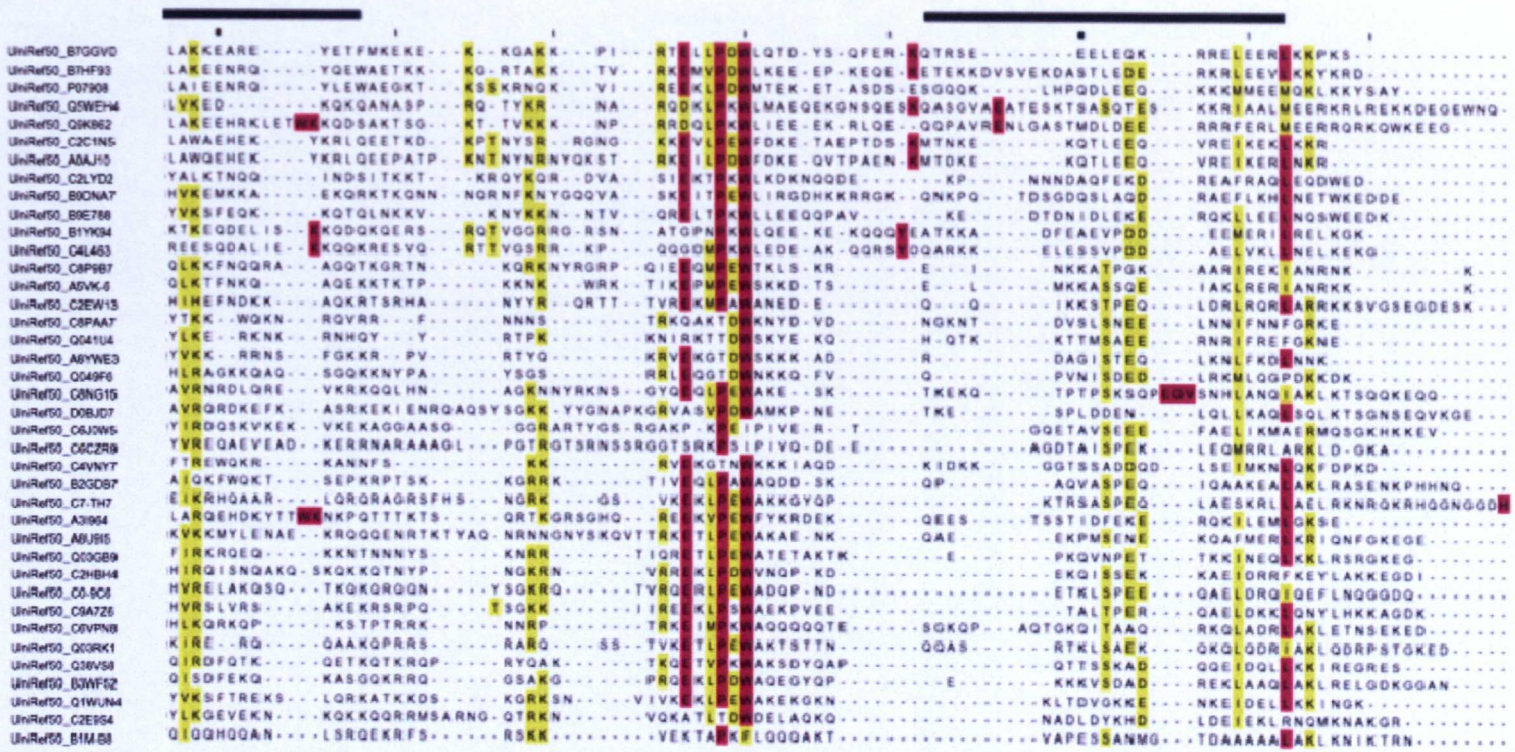


Figure 7.5: ClustalW alignment of the sequences defined as DnaB from Figure 7.4. The colour scheme of the alignment and the tickmarks are as in Figure 5.9. The secondary structure shown along the top of the alignment is as in Figure 7.3. The location of the YxxxIxxxW motif is shown boxed. DnaB is the third sequence in this alignment (P07908).

based only on the sequence similarity between the C-terminal domains of the DnaD and DnaB proteins. Sequence analysis using sensitive HMM based methods discover that DnaB is modular, comprising the domain structure (DDBH1)-(DDBH2)-(DDBH2), where the similarity of the second DDBH2 domain is much higher to DnaD DDBH2 than that of the first. This sequence similarity of the DnaB DDBH1 and central DDBH2 domain to that of the domains in DnaD have not previously been identified.

The structural similarity between the DnaD and DnaB proteins relates to their functional properties which are also very similar. One common property between DnaD and DnaB is that their DDBH1 domains form dimers that associate into tetramers (Carneiro *et al.*, 2006; Grainger *at al.*, 2010). In DnaD, a helix-strand-helix (H1-S1-H2) element in the DDBH1 domain that extends from the N-terminus of the winged helix fold facilitates dimer and tetramer formation whereas a single helix (H3) which extends from the C-terminus of the same winged helix fold participates in inter-tetramer interactions, essential for scaffold formation (Schneider *et al.*, 2008). Since these two structural elements are conserved in the DDBH1 domain of DnaB, it is likely that this domain in DnaB has the ability to oligomerise like the DnaD DDBH1 domain.

Another common property between the DnaD and DnaB proteins is that the DDBH2 domains are monomeric but have an oligomerisation activity that is induced by binding to DNA (Carneiro *et al.*, 2006: Grainger *et al.*, 2010). The combined data from NMR DNA binding studies and gel shift experiments (discussed in chapter 6) have shown that the DNA binding module in the DnaD DDBH2 domain comprises the particularly highly conserved YxxxIxxxW motif, the relatively unstable helix V and part of the unstructured C-terminal region. ClustalW alignments have shown that this YxxxIxxxW motif is also conserved amongst the putative DnaB sequences. Furthermore, a region in DnaB comparable to the region covered by the DNA binding modules in DnaD has recently been shown to participate in DNA binding and DNA–induced oligomerisation (Grainger *et al.*, 2010). This region in DnaB includes residues 365–400 of the second DDBH2 and residues 400–428 of the C-terminal extension (Grainger *et al.*, 2010). Clearly, the DNA-binding and DNA-

induced oligomerisation ability of the DDBH2 domain in the DnaD and DnaB proteins seems to be structurally and functionally preserved.

Generally, the winged helix-turn-helix motif in winged helix proteins that are DNA binding only form interactions with dsDNA (Gajiwala *et al.*, 2000). However, this is not the case for DnaB or DnaD, as a polypeptide fragment of DnaB comprising residues 1–300 which essentially spans the (DDBH1)-(DDBH2) domains binds exclusively to ssDNA (Grainger *et al.*, 2010) and residues 1–128 that make up the DDBH1 DnaD domain completely lack the ability to bind DNA (Carneiro *et al.*, 2006). The role of the DDBH1 domains in both DnaD and DnaB therefore seems to be to participate in protein-protein interactions leading to dimer and tetramer formation and not to bind DNA. Moreover, the ssDNA binding ability of residues 1–300 of the (DDBH1)-(DDBH2) domains in DnaB is most probably the activity of the DDBH2 domain and not that of the DDBH1 domain.

DDBH2 domains are also found in numerous bacterial and phage proteins, many of which have unknown functions (Carneiro *et al.*, 2006). As in DnaB, quite a few of these proteins have two tandem DDBH2 domains, whilst other proteins have their DDBH2 domains fused with other domains such as the orf4 protein from phage 7201 of *S. thermophilus* that has a RepA-like domain associated with initiation of DNA replication fused with its DDBH2 domain (Carneiro *et al.*, 2006). Similar modular architectures have been observed in a putative replication initiation protein from phage A118 of *Listeria monocytogenes* and in Protein 20 from *Lactobacillus plantarum* (Rokop *et al.*, 2004). Therefore, the modular character of the DDBH1 and DDBH2 domains and their structural and/or functional properties are likely to be conserved across many bacterial and phage proteins beyond the DnaD and DnaB primosomal protein.

Chapter 8: Discussion and future work

8.1 Discussion

Selecting a suitable construct of DnaD-Cd for NMR based studies

Before work towards NMR based studies of structure determination and DNA binding could commence, a construct of DnaD-Cd had to be selected, ideally one that correctly spanned the structural and functional region of the domain. When this project was initiated, the only available model for DnaD-Cd was that of the 2i5u structure, in which there are four helices (I-IV) that bundle together, forming a hydrophobic core. The last residue of the fourth helix is equivalent to V197 of *B. subtilis* DnaD-Cd. Thirteen months into the project another structure, 2zc2, showing greater sequence homology to DnaD-Cd than 2i5u became available. Like 2i5u, the 2zc2 structure is composed of four major helices (I-IV), but has an additional very short helix V that spanned 4 residues (L194 to V197, equivalent to V196 to A199 in *B. subtilis* DnaD). The function of the proteins to which these two structures correspond to is unknown and thus it could not be said whether these structures spanned the functional regions of the domain. Secondary structure predictions suggested that DnaD-Cd was composed of five helices and a 23 residue unstructured region, however helix V was much longer than in the available models. In contrast, residue conservation in sequence alignments implied that DnaD-Cd was structured up to residue V197 as in 2zc2. Due to such discrepancies as to the proper extent of a structural and functional C-terminal domain the first construct studied by NMR included the whole of the C-terminal tail.

Preliminary NMR structure determination and homology model construction

Using standard resonance assignment and structure calculation methods, a preliminary ensemble of NMR structures composed of five helices and an unstructured C-terminal tails was calculated for DnaD-Cd. The overall fold of the ensemble made up of helices I- IV was very similar to the structures 2i5u and 2zc2. As in 2zc2, helix V was highly exposed as it extended away from the core of the protein. However, it was longer than that of 2zc2 but shorter than indicated by

TALOS analysis of backbone chemical shift data and ^{15}N T2 relaxation data of DnaD-Cd. Undoubtedly the NMR data supported the secondary structure predictions that helix V was considerably more extensive than suggested by the structures 2zc2 and 2i5u, or by sequence alignments. However, it was not possible to refine the preliminary ensemble of structures and therefore carry out a full NMR structure determination as only a limited number of NOEs could be unambiguously assigned due to the insufficient quality of the ^{15}N and ^{13}C NOESY spectra. As noted above, the closest available model to DnaD-Cd was that of the X-ray structure of 2zc2. Since X-ray structures are generally more precise than NMR structures but more importantly since the NOEs that determined the overall fold in the preliminary ensemble of NMR structures for DnaD-Cd agreed with the fold in 2zc2, a hybrid homology model was constructed based on 2zc2. Furthermore, the final helix was extended in this model as a regular helix to a length indicated by the NMR data.

Even though the construct of the DnaD C-terminal domain that included the entire C-terminal tail was not ideal for structure determination, as only a low resolution structure was obtained, it proved to be important for elucidating the complete DNA binding module because a shorter construct lacking the C-terminal tail would not have been sufficient as discussed below.

The DNA binding module of DnaD-Cd

When *B. subtilis* DnaD-Cd interacts with DNA, it oligomerises forming large complexes on the DNA (Carneiro *et al.*, 2006). It was, therefore, expected that titrating even a small amount of DNA into DnaD-Cd would gradually, but uniformly decrease the signal observed in spectra to complete obliteration, as the size of the DNA:Protein complexes increased. Furthermore, the chemical shift changes would increasing become difficult to detect as signals for the DNA:Protein complex would most likely become very broad and possibly indistinguishable form the signal noise due to the effect large complexes have on signal relaxation. On the whole, it was expected that little or even no information regarding the DNA binding site in DnaD may be extracted from these experiments. Surprisingly, the loss of signal intensity was differential across spectra and did not become completely obliterated at the end of the titration series, which was terminated at a protein DNA stochiometary of 1:1 in

the case of DnaD-Cd and 1:2 for DnaD-Cd215. Moreover, chemical shift changes were observed, although they were very small, (All shift change were < 0.05 ppm) and difficult to track, particularly for cross-peaks with a substantial intensity drop. Differential line broadening was also observed, which is quite common for a protein undergoing conformational exchange on binding oligonucleotides, and is generally attributed to the possibility of a variety of binding registers. Line shape simulations confirmed that conformation broadening is a plausible explanation for this data, but also showed that the very small chemical shift changes and line broadening observed is consistent with exchange between the DNA free and the DNA bound forms of the protein on the slow side of intermediate exchange.

The observation of quite small complexes in the NMR titrations and rather large complexes in the gel shift experiments were most likely a result of the different relative concentrations of DNA and protein used, if the protein:DNA complexes are made up of a larger number of protein than DNA molecules. A detail analysis of the complexes is really needed to resolve this issue. However, if this explanation is true then it would raise several questions: particularly, what prevents DnaD-Cd domains oligomerising in the absence of DNA, and what limits the size of the aggregates formed? It also implies that the full explanation of the action of DnaD may be more intricate than described in the most recent speculative model (Zhang *et al.*, 2005) that shows a second oligomerisation interface on the N-terminal domain being unmasked by the C-terminal domain when it binds DNA.

Despite the difference in complex size, the combined data from NMR DNA titrations and gel shift experiments showed that the DNA binding module in DnaD-Cd is comprised of a highly conserved YxxxIxxxW motif (in the penultimate helix) that forms an exposed patch on the surface of the protein, the relatively unstable final helix and the unstructured C-terminus tail up to residue 215. The absence of any one element makes DnaD-Cd deficient in DNA binding and thus high order oligomerisation. Furthermore, *In vivo* complementation experiments showed that this DNA binding module in DnaD is essential for *B. subtilis* cell viability.

A sequence link between the DnaD and DnaB proteins

Sequence analysis using sensitive Hidden Markov Model-based techniques revealed that the *B. subtilis* proteins DnaD and DnaB share similar domain organizations. DnaD has the domain structure (DDBH1)-(DDBH2) whereas DnaB has the domain structure (DDBH1)-(DDBH2)-(DDBH2). Solving the domain structures of DnaD and DnaB facilitates the rationalisation of their common properties. The DDBH1 domain is composed of a conserved N-terminal helix-strand-helix element that forms β-sheets in dimers and tetramers, and a C-terminal helix that is responsible for cross-tetramer oligomerisation. The DDBH2 domain has a C-terminal DNA-binding site containing a penultimate helix with the conserved motif YxxxIxxxW, and also the last helix, plus a section of the C-terminal unstructured region. DDBH2 is more strongly conserved than DDBH1 between DnaD and DnaB. However, whilst the highly conserved motif of the DDBH2 is part of a well conserved protein core, the helix loosely associated to it, and the unstructured regions are both poorly conserved. When coupled, these two domains achieve DNA binding and high order oligomerisation.

8.2 Future work

Refine preliminary ensemble of NMR structures of DnaD-Cd

A full structure determination for DnaD-Cd could not be performed due to the insufficient quality of spectra. However, since the DnaD-Cd truncations, DnaD-Cd206 and DnaD-Cd215, have significantly improved tumbling characteristics, in comparison to DnaD-Cd, giving spectra with sharper and better resolved crosspeaks, any one of these two constructs may prove helpful in improving the preliminary ensemble of NMR structures. Furthermore, the complete absence of the mobile C-terminal tails in the case of DnaD-Cd206 or even part of the tail in the case of DnaD-Cd215, will not only reduces the number of cross-peaks in the central region of their corresponding spectra improving on ambiguities in the NOE assignment but may also significantly reduce the likelihood of protein degradation, as suggested by the truncation of DnaD-Cd no further than V206.

Determine the orientation of helix V of DnaD-Cd relative to the main core fold

The NMR data suggests that helix V of DnaD-Cd does not pack against the main core fold of the protein. The fact that no NOEs were found in NOESY spectra to properly constrain this final helix, means that the orientation of helix V with respect to the core fold of the protein remains unknown. Since NOEs can usually be detected only between protons that are < 5Å apart giving short range or local distance restraints, the presence of two pairs of NOEs at for example opposite ends of the protein, would not provide any information on how they are related to each other. In contrast to NOEs, residual dipolar couplings (RDC) can provide both short and long range distance information as well as information on angles formed by a vector connecting the two atoms (Lipsitz and Tjandra, 2004). RDCs are complimentary to the NOE data and often utilized as restraints in structure calculations to help refine or resolve discrepancies between NMR and X-ray crystallography structures of the same molecule. Therefore, RDCs could be used to determine orientation of helix V relative to the main core fold of DnaD-Cd.

X-ray crystallographic structure of the DnaD-Cd, DnaD-Cd-DNA and DnaD-DNA

The comparison of the NMR structure of a protein to its X-ray structure may enable detection of differences in conformational features for which no constraints were available or more importantly, may reveal complementary insights into the function and activity of the protein. For comparison of the NMR structure of DnaD-Cd derived once the preliminary ensemble of NMR structure is refined, X-ray crystallographic structure determination may be pursued. Furthermore, attempts to crystallise DnaD-Cd-DNA and DnaD-DNA complexes may be made. Since previous crystallization trials of DnaD-Cd have failed to generate reproducible crystals for unknown reasons, future trials will be carried out with the truncated construct DnaD-Cd215, as it contains all elements of the DNA binding module.

DNA titrations with dsDNA

DnaD-Cd binds both ssDNA and dsDNA non-specifically (Marsin *et al.*, 2001: Turner *et al.*, 2004). NMR DNA binding studies were performed with only ssDNA, since the gel shift assays with truncations of DnaD-Cd inferred that the binding site

for dsDNA and ssDNA were most likely identical. To confirm this hypothesis, NMR DNA titrations will be carried out with dsDNA. Moreover, the DNA binding affinity will be determined, since gel shifts also show that dsDNA binds weaker than ssDNA.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-402.

Azam, T.A., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999) Growth phase-dependent variations in the protein composition of the *Escherichia coli* nucleoid. *J. Bacteriol.* **181:** 6361–6370.

Baker, T.A., and Bell, S.P. (1998) Polymerases and the replisome: machines within machines. *Cell* **92:** 295–305.

Bodenhausen, G., and Ruben, D.J. (1980). Natural abundance ^{15}Nitrogen-NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69:** 185-189.

Bondos, S., and Bicknell, A. (2003) Detection and prevention of protein aggregation before, during, and after purification. *Anal. Biochem.* **316**: 223–231.

Bouche, J.P., Zechel, K., and Kornberg, A. (1975) dnaG gene product, a rifampicin-resistant RNA polymerase, initiates the conversion of a single-stranded coliphage DNA to its duplex replicative form. *J. Biol. Chem.* **250:** 5995–6001.

Boye, E., Stokke, T., Klecknert, N., and Skarstad, K., (1996) Coordinating DNA replication initiation with cell growth: Differential roles for DnaA and SeqA proteins. *Proc. Natl. Acad. Sci. USA* **93:** 12206-12211.

Brown, T.A. (2006) Genomes. 2 nd ed, Wiley-Liss.

Bruand, C., Ehrlich, S.D., and Janniere, L.J. (1995) Primosome assembly site in *Bacillus subtilis*. *EMBO J.* **14:** 2642–2650.

Bruand, C., Farache, M., McGovern, S., Ehrlich, S.D., and Polard, P. (2001) DnaB, DnaD and DnaI proteins are components of the *Bacillus subtilis* replication restart primosome. *Mol. Microbiol.* **42:** 245–255.

Bruand, C., Velten, M., McGovern, S., Marsin, S., Serena, C., Ehrlich, S.D., and Polard, P. (2005) Functional interplay between the *Bacillus subtilis* DnaD and DnaB proteins essential for initiation and re-initiation of DNA replication. *Mol. Microbiol.* **55:** 1138–1150.

Campbell, J.L., and Kleckner, N. (1990). *E. coli oriC* and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell* **62:** 967–979.

Carneiro, M.J., Zhang, W., Ioannou, C., Scott, D.J., Allen, S., Roberts, C.J., and Soultanas, P. (2006) The DNA-remodelling activity of DnaD is the sum of oligomerization and DNA-binding activities on separate domains. *Mol. Microbiol.* **60:** 917–924.

Clubb, R.T., Thanabal, V., and Wagner, G. (1992). A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue $^{1}H^{N}$, ^{15}N, and $^{13}C'$ chemical shifts in ^{15}N-^{13}C-labelled proteins. *J. Magn. Reson.* **97:** 213-217.

Cornilescu, G., Delaglio, F., and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13:** 289–302.

Dame, R.T., and Goosen, N. (2002) HU: promoting or counteracting DNA compaction? *FEBS Lett.* **529:** 151–156.

Errington, J., and Daniel, R.A. (2002) Cell division during growth and sporulation (In) *Bacillus subtilis* and its Closest Relatives: From Genes to Cells. Sonenshein, A.L., Hoch, J.A. and Losick, R. (eds). Washington, DC: ASM Press. pp.73–86.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, J.S., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res.* **36:** D281-288.

Gajiwala, K.S., Chen, H., Cornille, F., Roques, B.P., Reith, W., Mach, B., and Burley, S. K. (2000) Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* **403:** 916–921.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A.(2005) Protein Identification and Analysis Tools on the ExPASy Server;(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607.

Goranov, A.I., Katz, L., Breier, A.M., Burge, C.H.B., and Grossman, A.D. (2005) A transcriptional response to replication status mediated by the conserved bacterial replication protein DnaA. *Proc. Natl. Acad. Sci. USA* **102:** 12932–12937.

Golovanov, A.P., Hautbergue, G.M., Wilson, S.A., and Lian, L. (2004) A Simple Method for Improving Protein Solubility and Long-Term Stability. *J. Am. Chem. Soc.* **126:** 8933–8939.

Grainger, W.H., Machón, C., Scott, D.J., and Soultanas, P. (2010) DnaB proteolysis *in vivo* regulates oligomerization and its localization at *oriC* in *Bacillus subtilis*. *Nucleic Acids Res.* **38:** 2851-64.

Griffiths, A.J.F., Jeffrey, M.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W. M. (1999) An introduction to genetic analysis. W. H. Freeman and Co.

Grzesiek, S., and Bax, A. (1992). Improved 3D triple-resonance NMR techniques applied to a 31-kDa protein. *J. Magn. Reson.* **96:** 432–440.

Grzesiek, S., and Bax, A. (1993). Amino-acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J. Biomol. NMR*, **3:** 185–204.

Hayashi, M., Ogura, Y., Harry, E.J., Ogasawara, N., and Moriya, S. (2005). *Bacillus subtilis* YabA is involved in determining the timing and synchrony of replication initiation. *FEMS Microbio.l Lett.* **247:** 73–79.

Hiasa, H., and Marians, K.J. (1999) Initiation of bidirectional replication at the chromosomal origin is directed by the interaction between helicase and primase. *J. Biol. Chem.* **274:** 27244–27248.

Hierro, A., Sun, J., Rusnak, A.S., Kim, J., Prag, G., Emr, S.D., and Hurley, J.H. (2004). Structure of the ESCRT-II endosomal trafficking complex. *Nature* **431:** 221–225.

Hill, T.M. (1992). Arrest of bacterial DNA replication. *Annu. Rev. Microbiol.* **46:** 603–633.

Hoshino, T., McKenzie, T., Schmidt, S., Tanaka, T., and Sueoka, N. (1987) Nucleotide sequence of *Bacillus subtilis dnaB*: a gene essential for DNA replication initiation and membrane attachment. *Proc. Natl. Acad. Sci. USA* **84:** 653–657.

Huang, C.Y., Chang, Y.W., and Chen, W.T. (2008) Crystal structure of the N-terminal domain of *Geobacillus kaustophillus* HTA426 DnaD protein. *Biochem. Biophys. Res. Comm.* **375:** 220-224.

Imai, Y., Ogasawara, N., Ishigo-Oka, D., Kadoya, R., Daito, T., and Moriya, S. (2000) Subcellular localization of Dnainitiation proteins of *Bacillus subtilis*: evidence that chromosome replication begins at either edge of nucleoids. *Mol. Microbiol.* **36:** 1037–1048.

Ishigo-Oka, D., Ogasawara, N., and Moriya, S. (2001) DnaD protein of *Bacillus subtilis* interacts with DnaA, the initiator protein of replication. *J. Bacteriol.* **183:** 2148–2150.

Kahng, L.S., and Shapiro, L. (2001) The CcrM DNA methyltransferase of Agrobacterium tumefaciens is essential, and its activity is cell cycle regulated. *J. Bacteriol.* **181:** 3065–3075.

Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N., and Morikawa, K. (1996). Structure of a replication-terminator protein complexed with DNA. *Nature* **383:** 598–603.

Katayama, T., Kubota, T., Kurokawa, K., Crooke, E., and Sekimizu, K. (1998). The initiator function of DnaA protein is negatively regulated by the sliding clamp of the *E. coli* chromosomal replicase. *Cell* **94:** 61–71.

Kato, J., and Katayama, T. (2001). Hda, a novel DnaA-related protein, regulates the replication cycle in *Escherichia coli. EMBO J.* **20:** 4253–4262.

Kay, L. E., Xu, G.Y., Singer, A.U., Muhandiram, D.R., and Formankay, J.D. (1993). A gradient-enhanced HCCH TOCSY experiment for recording side-chain ^{1}H and ^{13}C correlations in H_20 samples of proteins. *J. Magn. Reson.* **B101:** 333-337.

Kay, L. E., Xu, G. Y., and Yamazaki, T. (1994). Enhanced-sensitivity triple-resonance spectroscopy with minimal H_2O saturation. *J. Magn. Reson.* Ser. A **109:**129-133.

Kelly, A.E, Ou, H.D., Withers, R., and Dötsch, V. (2002) Low-conductivity buffers for high-sensitivity NMR measurements. *J. Am. Chem. Soc.* **124:** 12013-9.

Kelley, L.A, Sternberg, M.J (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4:** 363-371.

Kelman, Z., and O'Donnell, M. (1995) DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. *Annu. Rev. Biochem.* **64:** 171–200.

Kim, J., Yoshimura, S.H., Hizume, K., Ohniwa, R.L., Ishihama, A., and Takeyasu, K. (2004) Fundamental structural units of the *Escherichia coli* nucleoid revealed by atomic force microscopy. *Nucleic Acids Res.* **32:** 1982–1992.

Kitagawa, R., Ozaki, T., Moriya, S., and Ogawa, T. (1998) Negative control of replication initiation by a novel chromosomal locus exhibiting exceptional affinity for *Escherichia coli* DnaA protein. *Genes Dev.* **12:** 3032–3043.

Kornberg, A. (1988) DNA Replication, *J. Biol. Chem.* **263:** 1-4.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390:** 249–256.

Kurokawa, K., Nishida, S., Emoto, A., Sekimizu, K., and Katayama, K. (1999). Replication cycle-coordinated change of the adenine nucleotide- bound forms of DnaA protein in Escherichia coli. *EMBO J.* **18:** 6642–6652.

Learn, B.A., Um, S.J., Huang, L and McMacken, R. (1997) Cryptic single-stranded-DNA binding activities of the phage lambda P and *Escherichia coli* DnaC replication initiation proteins facilitate the transfer of E-coli DnaB helicase onto DNA. *Proc. Natl. Acad. Sci. USA* **94:**1154–1159.

Lemon, K.P., Moriya, S., Ogasawara, N., and Grossman, A. (2002) Chromosome replication and segregation. (In) *Bacillus Subtilis* and its Closest Relatives: from Genes to Cells. Sonenshein, L., Hoch, J.A., and Losick, R. (eds). Washington, DC: ASM Press. pp.73–86.

Lewis, P.J. (2004). Bacterial subcellular architecture: recent advances and future prospects. *Mol. Microbiol.* **54:** 1135-50.

Li, Y., Kurokawa, K., Matsuao, M., Fukuhara, N., Murakami, K., and Sekimizu, K. (2004) Identification of temperature sensitive *dnaD* mutants of *Staphylococcus aureus* that are defective in chromosomal DNA replication. *Mol. Gen. Genom.* **271:** 447–457.

Lipsitz, R.S., and Tjandra, N. (2004) Residual dipolar couplings in NMR structure analysis. *Annu. Rev. Biophys. Biomol. Struct.* **33:** 387–413

Lu, M., Campbell, J.L., Boye, E., and Kleckner, N. (1994). SeqA: a negative modulator of replication initiation in *E. coli*. *Cell* **77:** 413–426.

Lu, Y.B., Ratnakar, P.V., Mohanty, B.K., and Bastia, D. (1996) Direct physical interaction between DnaG primase and DnaB helicase of *Escherichia coli* is necessary for optimal synthesis of primer RNA. *Proc. Natl. Acad. Sci. USA* **93:**12902–12907.

Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M.R., and Cebrat, S. (2004) Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res.* **32:** 3781–3791.

Marsin, S., McGovern, S., Ehrlich, S.D., Bruand, C., and Polard, P. (2001) Early steps of *Bacillus subtilis* primosome assembly. *J. Biol. Chem.* **276:** 45818–45825.

Marston, F.Y., Grainger, W.H., Smits, W.K., Hopcroft, N.H., Green, M., Hounslow, A.M, Grossman, A.D., Craven, C.J, and Soultanas, P. (2010) When simple sequence comparison fails: the cryptic case of the shared domains of the bacterial replication initiation proteins DnaB and DnaD. *Nucleic Acids Res.* 1–13.

Marians, K.J. (1999) PriA: at the crossroads of DNA replication and recombination. *Prog Nucleic Acids Res. Mol. Biol.* **63:** 39–67.

Marians, K.J. (2000) PriA-directed replication fork restart in *Escherichia coli*. *Trends Biochem. Sci.* 25: 185–189.

Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M., and Clore, G.M. (1989) Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-^{15}N Hartmann-HAHN-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1.beta. *Biochemistry.* **28:** 6150-6156.

Marszalek, J., and Kaguni, J. M. (1994) DnaA protein directs the binding of DnaB protein in initiation of DNA replication in *Escherichia coli*. *J. Biol. Chem.* **269:** 4883-4890.

Maxwell, K.L., Bona, D., Liu, C.S., Arrowsmith C.H., and Edwards, A.M. (2003). Refolding out of guanidine hydrochloride is an effective approach for high-throughput structural studies of small proteins. *Protein Sci.* **12:** 2073-2080.

Messer, W., and Weigel, C. (1997) DnaA initiator–also a transcription factor. *Mol. Microbiol.* **24:** 1–6.

Messer, W. (2002) The bacterial replication initiator DnaA. DnaA and *oriC*, the bacterial mode to initiate DNA replication. *FEMS Microbiol. Rev.* **26:** 355–374.

McGarry, K.C., Ryan, V.T., Grimwade, J.E., and Leonard, A.C. (2004) Two discriminatory binding sites in the *Escherichia coli* replication origin are required for DNA strand opening by initiator ATP–DnaA. *Proc. Natl Acad. Sci. USA* **101:** 2811–2816.

Monera, O.D., Kay, C.M., Hodges, R.S. (1994). Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interaction. *Protein Sci.* **3:** 1984-1991.

Moriya, S., Firsheln, W., Yoshikawa, H., and Ogasawara, N. (1994) Replication of a *Bacillus subtilis oriC* plasmid in vitro. *Mol. Microbiol.* **12:** 469-478.

Moriya, S., Imai, I., Hassan, A.K.M., and Ogasawara, N. (1999) Regulation of Initiation of *Bacillus subtilis* Chromosome Replication. *Plasmid* **41:** 17–29.

Mott, M.L., and Berger, J.M. (2007) DNA replication initiation: mechanisms and regulation in bacteria. *Nat. Rev. Microbiol.* **5:** 343-54.

Muhandiram, D.R., and Kay, L.E (1994). Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity *J. Magn. Reson.*, Ser. B **103:** 203-216

Noirot-Gros, M.F., Dervyn, E., Wu, L.J., Mervelet, P., Errington, J., Ehrlich, S.D., and Noirot, P. (2002). An expanded view of bacterial DNA replication. *Proc. Natl. Acad. Sci. USA* **99:** 8342–8347.

Noirot-Gros, M.F., Velten, M., Yoshimura, M., McGovern, S., Morimoto, T., Ehrlich, S. D., Ogasawara, N., Polard, P., and Noirot, P. (2006). Functional dissection of YabA, a negative regulator of DNA replication initiation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **103:** 2368–2373.

Ng, J.Y., and Marians, K.J. (1996) The ordered assembly of the phiX174- type primosome. Isolation and identification of intermediate protein-DNA complexes. *J. Biol. Chem.* **271**:15642–15648.

Ogden, G.B., Pratt, M.J., and Schaechter., M. (1988). The replication origin of the *Escherichia coli* chromosome binds to cell membranes only when hemimethylated. *Cell* **54**: 127–135.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera, a visualization system for exploratory research and analysis. *J Comput Chem.* **13**: 1605-12.

Piotto, M., Saudek, V., and Sklenar, V. (1992). Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR.* **2**: 661-665

Polard, P., Marsin, S., McGovern, S., Velten, M., Wigley, D.B., Ehrlich, S.D., and Bruand, C. (2002) Restart of DNA replication in Gram-positive bacteria: functional characterization of the *Bacillus subtilis* PriA initiator. *Nucleic Acids Res.* **30**: 1593–1605.

Reed, M.A.C., Hounslow, A.M., Sze, K.H., Barsukov, I.G., Hosszu, L.L.P., Clarke, A.R., Craven, C.J. and Waltho, J.P. (2003) Effects of domain dissection on the folding and stability of the 43 kDa protein PGK probed by NMR. *J. Mol. Biol.* **330**: 1189–1201.

Reid, D.G. (1997) Protein NMR Techniques (In) Methods in Molecular Biology vol. 601, ed. Reid, Humana Press, Totowa, N.J.

Rieping, W., Habeck, M., Bardiaux, B., Bernard A., Malliavin T.E., Nilges M. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinf.* **23**: 381-382.

Riber, L., Olsson, J.A., Jensen, R.B., Skovgaard, O., Dasgupta, S., Marinus, M.G., and Lobner-Olesen, A. (2006) Hda mediated inactivation of the DnaA protein and dnaA gene autoregulation act in concert to ensure homeostatic maintenance of the *Escherichia coli* chromosome. *Genes Dev.* **20**: 2121–2134.

Robinow, C., and Kellenberger, E. (1994) The bacterial nucleoid revisited. *Microbiol. Rev.* **58**: 211–232.

Rokop, M.E., Auchtung, J.M., and Grossman, A.D. (2004) Control of DNA replication initiation by recruitment of an essential initiation protein to the membrane of *Bacillus subtilis*. *Mol. Microbiol.* **52:** 1757–1767.

Sandler, S. J., and Marians, K. J. (2000). Role of PriA in replication fork reactivation in *Escherichia coli*. *J. Bacteriol.* **182:** 9–13.

Sanner, M., Olson, A. J., and Spehner, J. C. (1996). Reduced Surface: an efficient way to compute molecular surfaces. *Biopolymers.* **38:** 305-320

Schaper, S., and Messer, W. (1995) Interaction of the initiator protein DnaA of *Escherichia coli* with its DNA target. *J. Biol. Chem.* **270:** 17622–17626.

Schneider, R., and Muskhelisvili, G. (1997) FIS modulates growth phase-dependent topological transitions of DNA in *Escherichia coli*. *Mol. Microbiol.* **26:** 519–530.

Schneider, S., Zhang, W., Soultanas, P., and Paoli, M. (2008) Structure of the N-Terminal Oligomerization Domain of DnaD Reveals a Unique Tetramerization Motif and Provides Insights into Scaffold Formation, *J. Mol. Biol* **376:** 1237–1250.

Schumann, W. (2006) Dynamics of the bacteria chromosome: Structure and function. Wiley-VCH. pp33-34.

Seitz, H., Weigel, C., and Messer, W. (2000) The interaction domains of the DnaA and DnaB replication proteins of *Escherichia coli*. *Mol. Microbiol.* **37:** 1270–1279.

Singh, S.M., and Panda, A.K. (2005). Solubilisation and refolding of bacterial inclusion body proteins. *J. Biosci. Bioeng.* **99:** 303-310.

Sonenshein, A.L., Hoch, J.A and Losick, R. (2002) *Bacillus subtilis* and its closest relatives: from genes to cells. Washington, DC: ASM Press.

Sørensen, H.P., and Mortensen, K.K (2005). Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microbial Cell Factories* **4:**1-8

Soultanas, P. (2002) A functional interaction between the putative primosomal protein DnaI and the main replicative DNA helicase DnaB in *Bacillus*. *Nucleic Acids Res.* **30:** 966–974.

Schneider, S., Carneiro, M.J.V.M., Ioannou, C., Soultanas, P., and Paoli, M. (2007). Crystallization and X-ray diffraction analysis of the DNA-remodelling protein DnaD from *Bacillus subtilis*. *Acta. Crystallogr. Sect. F*, **F63:** 110–113.

Stephens, C., Reisenauer, A., Wright, R., and Shapiro, L. (1996) A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. *Proc. Natl. Acad. Sci. USA* **93:** 1210–1214.

Sueoka, N. (1998) Cell membrane and chromosome replication in *Bacillus subtilis*. *Prog Nucleic Acids Res. Mol. Biol.* **59:** 35–53.

Teo, H., Perisic, O., Gonzalez, B., and Williams, R. L. (2004). ESCRT-II, an endosome-associated complex required for protein sorting: crystal structure and interactions with ESCRT-III and membranes. *Dev. Cell* **7:** 559–569.

Tougu, K., Peng, H., and Marians, K.J. (1994) Identification of a domain of *Escherichia coli* primase required for functional interaction with the DnaB helicase at the replication fork. *J. Biol. Chem.* **269:** 4675–4682.

Tsumoto K,E., Ejima, D., Kumagai, I., and Arakawa, T. (2003). Practical considerations in refolding proteins from inclusion bodies. *Protein Expression Purif.* **28:** 1-8.

Turner, I.J., Scott, D.J., Allen, S., Roberts, C.J., and Soultanas, P. (2004) The *Bacillus subtilis* DnaD protein: a putative link between DNA remodelling and initiation of DNA replication. *FEBS Lett* **577:** 460–464.

Velten, M., McGovern, S., Marsin, S., Ehrlich, S.D., Noirot, P., and Polard, P. (2003) A two-protein strategy for the functional loading of a cellular replicative DNA helicase. *Mol. Cell* **11:** 1009–1020.

Walsh, G. (2003) Biopharmaceuitical benchmark. Nat .Biotechnology., **21:** 865-870.

Wittekind, M., and Mueller, L. (1993). HNCACB, a high-sensitivy 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta-carbon resonances in proteins. *J. Magn. Reson. Series B.* **B101:** 201-205.

Zakrzewska-Czerwin, J., Jakimowicz, D., Zawilak-Pawlik, A., and Messer, W. (2007) Regulation of the initiation of chromosomal replication in bacteria. *FEMS Microbiol. Rev.* **31:** 378–387.

Zhang, W., Carneiro, M.J.V.M., Turner, I.J., Allen, S., Roberts, C.J., and Soultanas, P. (2005) The *Bacillus subtilis* DnaD and DnaB proteins exhibit different DNA remodeling activities. *J. Mol. Biol.* **351:** 66–75.

www.virginia.edu/Heidi/chapter30/chp30.html

http://blast.ncbi.nlm.nih.gov/Blast.cgi

http://pfam.sanger.ac.uk

Nucleic Acids Research Advance Access published June 28, 2010

Nucleic Acids Research, 2010, 1–13
doi:10.1093/nar/gkq465

When simple sequence comparison fails: the cryptic case of the shared domains of the bacterial replication initiation proteins DnaB and DnaD

Farhat Y. Marston[1], William H. Grainger[2], Wiep Klaas Smits[3], Nicholas H. Hopcroft, Matthew Green[1], Andrea M. Hounslow[1], Alan D. Grossman[3], C. Jeremy Craven[1,*] and Panos Soultanas[2,*]

[1]Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, [2]School of Chemistry, Centre for Biomolecular Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK and [3]Department of Biology, Building 68-530, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received March 22, 2010; Revised May 8, 2010; Accepted May 11, 2010

ABSTRACT

DnaD and DnaB are essential DNA-replication-initiation proteins in low-G+C content Gram-positive bacteria. Here we use sensitive Hidden Markov Model-based techniques to show that the DnaB and DnaD proteins share a common structure that is evident across all their structural domains, termed DDBH1 and DDBH2 (DnaD DnaB Homology 1 and 2). Despite strong sequence divergence, many of the DNA-binding and oligomerization properties of these domains have been conserved. Although eluding simple sequence comparisons, the DDBH2 domains share the only strong sequence motif; an extremely highly conserved YxxxIxxxW sequence that contributes to DNA binding. Sequence alignments of DnaD alone fail to identify another key part of the DNA-binding module, since it includes a poorly conserved sequence, a solvent-exposed and somewhat unstable helix and a mobile segment. We show by NMR, *in vitro* mutagenesis and *in vivo* complementation experiments that the DNA-binding module of *Bacillus subtilis* DnaD comprises the YxxxIxxxW motif, the unstable helix and a portion of the mobile region, the latter two being essential for viability. These structural insights lead us to a re-evaluation of the oligomerization and DNA-binding properties of the DnaD and DnaB proteins.

INTRODUCTION

Bacillus subtilis DNA replication initiates at a defined chromosomal origin, *oriC*, where the highly conserved master replication initiator protein DnaA binds to recruit sequentially DnaD, DnaB and the DnaI–DnaC complex (1–6). The hexameric replicative ring helicase DnaC [the homologue of *Escherichia coli* DnaB (beware of the confusing nomenclature of bacterial primosomal proteins: *B. subtilis* DnaB is a distinct protein from the helicases typified by *E. coli* DnaB)] is loaded in an ATP-dependent reaction onto the locally melted AT-rich region by the DnaB–DnaI dual helicase-loading system (7–9). The DnaD, DnaB and DnaI proteins are also components of the replication restart primosome at stalled and collapsed replication forks outside *oriC* (3,10).

Bacillus subtilis DnaD exhibits a complex range of protein oligomerization and DNA recognition behaviors. It binds single stranded (ss) and double stranded (ds) DNA (2), remodels supercoiled DNA by eliminating writhe through increasing negative twist while maintaining an unchanged linking number (11–14). Atomic Force Microscopy indicates that upon binding to DNA, DnaD forms large scaffolds which open up supercoiled DNA (11,12). The isolated DnaD N-terminal domain (DnaD-Nd; DnaD residues 1–128) readily forms aggregates (and crystallises) in the absence of DNA and this has led to the suggestion that the DnaD C-terminal domain (DnaD-Cd; DnaD residues 129–232) masks the aggregation surface of DnaD-Nd: upon binding of DnaD-Cd to DNA this surface is unmasked and oligomerization of DnaD-Nd is triggered (15,16). The

*To whom correspondence should be addressed. Tel: +44 114 222 4323; Fax: +44 114 272 2800; Email: c.j.craven@shef.ac.uk
Correspondence may also be addressed to Panos Soultanas. Tel: +44 115 951 3525; Fax: +44 115 846 8002; Email: panos.soultanas@nottingham.ac.uk

DNA-remodelling activity of DnaD is the sum of a scaffold-forming activity in its N-terminal domain, DnaD-Nd (15,16), and DNA-binding and DNA-induced oligomerization activities in its C-terminal domain, DnaD-Cd (15).

Bacillus subtilis DnaB also binds ss and dsDNA (2), remodels supercoiled and linear DNA by lateral compaction (12), cooperates as a helicase co-loader with DnaI (the homologue of *E. coli* DnaC) (7,9), acts as an origin attachment protein (17–21), regulates the recruitment of DnaD to the membrane-attached *oriC* (22) and participates in regulation of replication initiation synchrony (23). Trypsin proteolysis resolves *B. subtilis* DnaB into three domains: an N-terminal domain (residues 1–184), a central domain (residues 185–296) and a C-terminal domain (residues 297–432) with an unstructured region at the C-terminal end (residues 434–472) (9). Residues 1–300 (including the N-terminal and central domains) mediate dimerization, tetramerization and ssDNA binding, and residues 365–428 (in the C-terminal domain) mediate higher order oligomerization and ss and dsDNA binding (24). The proteolytically sensitive C-terminus of DnaB seems to regulate higher order oligomerization and association with *oriC in vivo* (24).

Although they do not appear to interact with each other in a yeast two-hybrid assay (22,25,26) the two proteins are needed together to bind to SSB-coated ssDNA, indicating a potential cryptic protein–protein interaction (3). Furthermore, the DnaB-S371P and the DnaD-A166TΔD154Q155 mutant proteins interact with wild-type DnaD and DnaB, respectively, in yeast two-hybrid assays, indicating that a conformational change may be required to expose the cryptic interaction interface (3,22,27). A genetic link between the two proteins has been established. The *dnaB75* (*dnaBS371P*) suppresses the temperature sensitive phenotypes caused by *dnaD23ts* and *dnaB134ts* (3,22,27). It is clear, therefore, that there is a close functional relationship between the two proteins. Work here indicates that this functional relationship is underpinned by a sequence and structural relatedness.

The structure of the N-terminal domain of *B. subtilis* DnaD has been determined by X-ray crystallography (16,28,29), revealing a winged helix-turn-helix fold. Whilst proteins with such a motif commonly, but not exclusively, have a DNA-binding function, this is not the case for the DnaD N-terminal domain, which instead facilitates formation of extensive protein scaffolds (14–16). The DNA-binding properties of DnaD are confined to the C-terminal domain (14,15). An extension at the N-terminus comprising a helix-strand-helix (H1′-S1′-H2′) mediates dimerization and tetramerization via two-stranded β-sheets formed by pairing of the strands in these extensions from two and four molecules, respectively. Helix (H3′) attached at the C-terminal end of the winged helix-turn-helix core appears to be involved in higher order cross-tetramer oligomerization (16).

Two structures of proteins homologous to the C-terminal domain of *B. subtilis* DnaD have been deposited by the Midwest Center for Structural Genomics: a homolog from *Streptococcus mutans* UA 159 (PDB code 2ZC2) and that from *Enterococcus faecalis* (PDB code 2I5U), with pair-wise identities to *B. subtilis* DnaD of 41% and 25%, respectively. Neither of these proteins has been otherwise characterized. The *S. mutans* protein is a fragment of a protein encoded by the gene *SMU 1465C*. Using the Phyre fold recognition server (30), the *SMU 1465C* encoded protein is predicted to contain the N-terminal DnaD domain whereas no such domain is detectable in the full length 2i5u sequence (data not shown but compare predicted secondary structure for DnaD, 2zc2 and 2i5u in Supplementary Figures S1 and S2). The 2zc2 structure, therefore, forms the strongest basis for the construction of models of the full length *B. subtilis* DnaD. The construct in the 2zc2 structure corresponds to residues L129–I201 of *B. subtilis* DnaD, with the final structured residue corresponding to A199 (Figure 1A).

Here, we demonstrate a structural link between DnaD and DnaB, which enables the formulation of a unified understanding of their properties. We show that the two structural domains found in DnaD (which we term DDBH1 and DDBH2 for DnaD-DnaB Homology Domain 1 and 2) are also present in DnaB; sensitive domain detection methods reveal that DnaB has the domain structure (DDBH1)-(DDBH2)$_2$. The similarity is considerably higher for the second DDBH2 domain, the proteins sharing a particularly highly conserved motif YxxxIxxxW, which forms an exposed patch on the surface of the protein. The structured part of the DDBH2 domain contains a more extensive final helix than hitherto assumed. We show that the YxxxIxxxW motif, the final helix and a portion of the unstructured C-terminal tail of DnaD all contribute to DNA binding, and are essential for *B. subtilis* viability.

MATERIALS AND METHODS

Bioinformatics

In addition to resources referenced in the text, we made extensive use of Jalview (31) (http://www.Jalview.org/), ClustalW (32), hmmer3 (33) (http://hmmer.janelia.org/) and cd-hit (34). Other manipulation of data was carried out using in-house python scripts on Linux workstations.

The model of *B. subtilis* DnaD was built from the coordinates of 2zc2 using swissmodel (35) (http://swissmodel.expasy.org/). The additional residues were modelled in using X-PLOR (36) to add random coil residues up to residue F206. Helical dihedral angles were then imposed on residues A199–F206 using an in-house python script, and finally side-chain clashes were removed using X-PLOR.

NMR

NMR spectra were acquired at 298 K on a Bruker DRX series 600-MHz spectrometer equipped with triple-resonance cryoprobe. Spectra were processed and analyzed using FELIX (FELIX NMR Inc.). Assignments of backbone resonances were obtained from ^{15}N-HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB and CBCA(CO)NH spectra. The assignment was completed using FELIX for peak picking and the asstools package as previously described

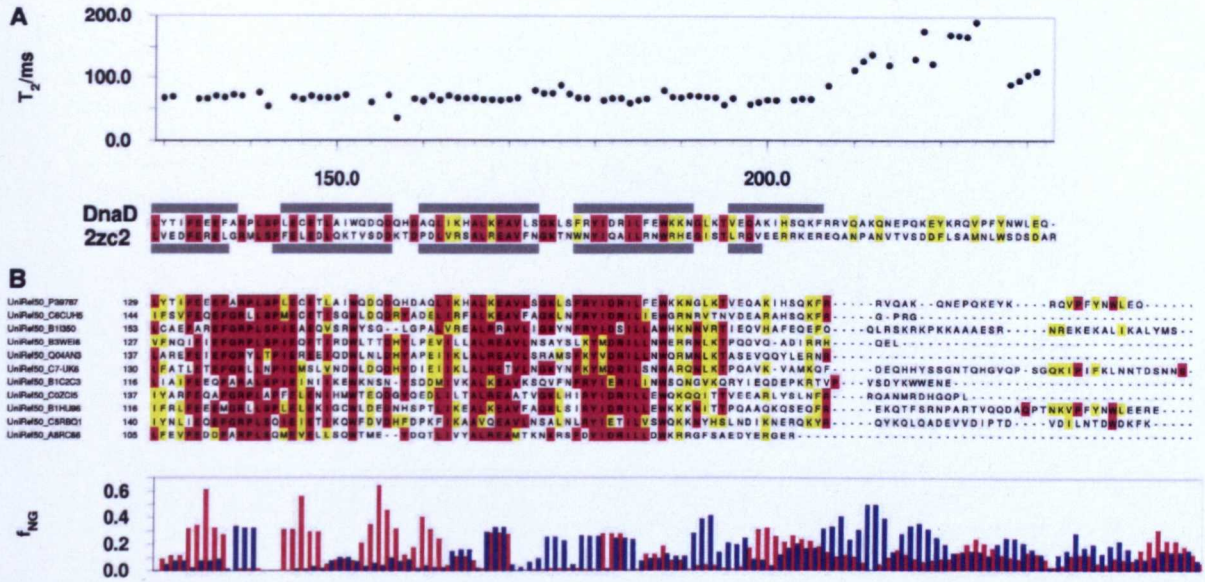


Figure 1. Defining the extent of the C-terminal domain of DnaD. (**A**) NMR ^{15}N relaxation (T_2) data, which defines regions of mobility in solution. The backbone from residue 129 to 206 shows uniform T_2 values, with elevated values beyond residue 206 indicating that the boundary between structure and disorder is approximately at residue 206. The reduced value of T_2 for the four data points shown at the C-terminus suggest some degree of constraint or conformational broadening, but there is no evidence for regular structure in the NOEs, chemical shifts or DNA-binding intensity changes for these residues. An alignment of *B. subtilis* DnaD-Cd with the sequence from *S. mutans*, corresponding to the structure 2zc2 is also shown. The extent of helices in the two structures is shown in grey above and below the sequences. Identical residues are coloured red, and similar residues are colored yellow. (**B**) ClustalW alignment of DnaD sequences. The sequences are sequences 1, 6, 11, 16, etc. drawn from the collection in Supplementary Figure S3. Residues that are identical in at least 50% of sequences (excluding gaps) are colored red. Residues that are similar across at least 50% of sequences are colored yellow. Sequences are marked with their Uniprot identifier. Below the alignment the fraction, f_{NG}, of basic (blue) and acidic (red) residues is shown, expressed as a fraction of the total non-gapped residues at each position in the alignment in Supplementary Figure S3. The profile was smoothed using a three-residue window. Across all parts of this figure the data are aligned vertically by residue number of *B. subtilis* DnaD, except where extensive gapping prevents this beyond residue 215.

(37). For side-chain assignments, CCH-TOCSY and HCCH-TOCSY spectra were employed, and NOEs were assigned in a 100-ms NOESY-($^{15}N/^{13}C$)-HSQC experiment. Samples were prepared in 100 mM sodium phosphate buffer (pH 6.5), 150 mM NaCl, 10 mM DTT and 0.3 mM protein. To provide an optimal sensitivity, 0.3 mM protein was used: higher concentrations led to poorer line shape and sensitivity presumably due to increased transient self-association. DNA titrations were performed using a 2 mM stock of a 10-mer ssDNA (5′-GTTATTGCTC-3′) exchanged into identical buffer conditions to those of the protein solution. Cross-peak intensities were corrected for the volume change during titrations. Amide exchange experiments were initiated by diluting the protein with an equal volume of D_2O buffer. Exchange was detected by comparing intensities to a control spectrum in which the protein was diluted with an equal volume of H_2O buffer. For the measurement of ^{15}N T_2 values, the Bruker pulse program hsqct2etf3gpsi (Avance version 04/01/15) was employed which essentially implements the sequence of Farrow *et al.* (38). CPMG delays of 0, 17, 34, 51, 68, 85, 119, 136 and 189 ms were employed. The correlation time was estimated from the ^{15}N T_2 by simple application of the Lipari–Szabo formula (38) with S^2 set to 0.85 for the rigid backbone, $\tau = 40$ ps and ^{15}N CSA = 160 p.p.m. The 2D ^{15}N-^{1}H correlation spectra were typically acquired for 2.5 h per spectrum (DNA titrations and amide exchange) or 3.5 h per spectrum (T_2).

Mutagenesis and protein purifications

Truncated and full length DnaD-Cd constructs encoding 129-232 (*dnaD-Cd*), 129–177 (*dnaD-Cd177*), 129–196 (*dnaD-Cd196*), 129–206 (*dnaD-Cd206*) and 129–215 (*dnaD-Cd215*) of *B. subtilis DnaD* were generated by PCR using pET22b-DnaD-Cd (*B. subtilis* DnaD residues 129–232) (15) as substrate and cloned as NdeI-HindIII fragments in pET28a and introducing a stop codon immediately prior the HindIII site. Site-directed mutagenesis was carried out with the QuickChange II-E mutagenesis kit (Stratagene). All constructs were verified by DNA sequencing and all oligonucleotides in this study were purchased from Eurogentec.

Full-length DnaD proteins were purified as described before (11,15). DnaD-Cd 129–232, DnaD-Cd 129–215, DnaD-Cd 129–206 and DnaD-Cd 129–196 were prepared by a modified version of these protocols to increase yield for NMR. Proteins were overexpressed in BL21 (DE3) *E. coli* at 37°C in LB or M9 medium with $^{15}NH_4SO_4$ or $^{13}C_6$ D-glucose as the sole carbon and nitrogen sources. Large proportions of the DnaD-Cd proteins were observed in the insoluble fractions, which were solubilized using 5 M guanidine hydrochloride. The proteins were purified using nickel affinity

chromatography, followed by refolding by dialysis. The refolded proteins were passed down the nickel column for a second time before removing the His-tag with thrombin, after which the cleaved tag was removed by a third-nickel affinity step. Finally, the proteins were further purified by gel filtration (Superdex S75). The resulting thrombin cleaved proteins possess the additional residues GSHM at the N-terminus.

Gel shift assays

Agarose gel shift assays were carried out with supercoiled pBSK plasmid as described before (12). Polyacrylamide gel shift assays were carried out with 0.625 nM ssDNA or dsDNA and various concentrations of proteins (DnaD, Y180A, W188A, DnaD-Cd196, DnaD-206 and DnaD-Cd215), as described before (16). The oligonucleotide Y180AF (5′-CTCAGTTTCCGCGCGATTGACCG GATT-3′) was used as ss probe and the complementary Y180AR (5′-AATCCGGTCAATCGCGCGGAAACTG AG-3′) was used to prepare the ds probe. Binding reactions were incubated for 20 min at RT and then resolved through 8% v/v native polyacrylamide gels.

Bacillus subtilis strains

Bacillus subtilis strains used are listed in Table 1. All are isogenic with JH642 and contain the *trpC2* and *pheA1* alleles (39). *dnaD23* is a temperature-sensitive allele that prevents replication initiation at the non-permissive temperature (3,10,40). The transposon insertion Tn*917*ΩHU151 is linked to *dnaD* (22,41,42).

The *dnaD* alleles were fused to the LacI-repressible IPTG-inducible promoter Pspank(hy) and integrated into the chromosome at the non-essential *amyE* or wild-type *dnaD* sites, the *dnaD* open reading frame, including the ribosome-binding site and followed by a TAA-stop codon was amplified by PCR using primers oWKS-178 (5′-CCCAAGCTTAGTAAGAGGTGTGAC GATGAAAAAACAGCAATTTATTGATA-3′) and oWKS-179 (5′-ACATGCATGCTTATTGTTCAAGCC AATTGTAAAAAGG-3′) and chromosomal DNA from a wild-type *B. subtilis* strain (JH642) as template. For the *dnaDY180A* and *dnaDW188A* mutations, PCR was carried out with the same primers but using the pET22b-derived expression plasmids carrying the mutant alleles. For the C-terminally truncated *dnaD*, the coding sequence including ribosome-binding site was amplified using primers oWKS-178 and oWKS-180 (5′-ACATG CATGCTTACACAGTTTTAAGCCCATTTTTCTTCC A-3′), which introduces a TAA stop codon after V196. The rest of the coding sequence of *dnaD* is not present in this construct. PCR products were digested with HindIII and SphI, and cloned into similarly digested pDR111 (43) putting it downstream of the IPTG inducible promoter Pspank(hy). Inserts were sequenced to verify the *dnaD* alleles. Pspank(hy)-*dnaD* constructs were introduced into the *B. subtilis* via double crossover at the non-essential *amyE* locus. Chromosomal DNA from these strain(s) was used to transform a *dnaD23ts* (KPL73) or *oriN oriC*⁻ strain (MMB208).

A deletion mutation in *dnaD* was constructed using the long-flanking homology PCR approach (44). Fragments ~1 kb upstream and downstream of *dnaD* were amplified using primers oWKS-204 (5′-GAAGCTGCGGCAATGG CTTTAG-3′), oWKS-205 (5′-AACGTATTTCCCGCTG TCCTCC-3′), oWKS-206 (5′-TACCGCACAGATGCG TAAGGAGCGTCACACCTCTTACTGTAAAGG-3′) and oWKS-207 (5′-TAATAATGAGATAATGCCGACT GTACTAAAGTGAAAAGGTGACAATCGTG-3′), resulting in two products with regions of homology to a chloramphenicol resistance cassette (*cat*). These products were used as megaprimers in an expand Long Template PCR reaction (Roche) with plasmid pGEMcat (45) as template, followed by a second Expand Long Template PCR using primers oWKS-204 and oWKS-205. Of purified PCR product of the expected size, 30 µl was transformed into strains carrying ectopic copies of *dnaD* or its variants, and double cross-over replacement of the *dnaD* gene with the *cat* cassette in the transformants was verified by PCR and digest.

Table 1. *Bacillus subtilis* strains

Strain	Relevant genotype (reference)
KPL73	*dnaD23*(ts)-Tn917ΩHU151 (*mls*) (22,41,42,)
WKS802	*amyE*::{Pspank(hy)-*dnaD, spc*}
WKS804	*amyE*::{Pspank(hy)-*dnaDY180A, spc*}
WKS806	*amyE*::{Pspank(hy)-*dnaDW188A, spc*}
WKS808	*amyE*::{Pspank(hy)-*dnaD-V196-STOP, spc*}
WKS810	*dnaD23*(ts)-Tn*917*ΩHU151 (*mls*); *amyE*::{Pspank(hy)-*dnaD, spc*}
WKS812	*dnaD23*(ts)-Tn*917*ΩHU151 (*mls*); *amyE*::{Pspank(hy)-*dnaDY180A, spc*}
WKS814	*dnaD23*(ts)-Tn*917*ΩHU151 (*mls*); *amyE*::{Pspank(hy)-*dnaDW188A, spc*})
WKS816	*dnaD23*(ts)-Tn*917*ΩHU151 (*mls*); *amyE*::{Pspank(hy)-*dnaD-V196-STOP, spc*}
MMB208	*spoIIIJ::oriN oriC-S (kan tet)* (51)
WKS849	*spoIIIJ::oriN oriC-S (kan tet); amyE*::{Pspank(hy)-*dnaD, spc*}
WKS865	*amyE*::{Pspank(hy)-*dnaD23, spc*}
WKS886	*dnaD23*(ts)-Tn*917*ΩHU151 (mls); amyE::{Pspank(hy)-*dnaD23, spc*}
WKS876	*amyE*::{Pspank(hy)-*dnaD*, spc} Δ*dnaD* (*cat*)
WKS878	*amyE*::{Pspank(hy)-*dnaDY180A*, spc} Δ*dnaD* (*cat*)
WKS880	*amyE*::{Pspank(hy)-*dnaDW188A*, spc} Δ*dnaD* (*cat*)
WKS883	*amyE*::{Pspank(hy)-*dnaD23, spc*} Δ*dnaD* (*cat*)

RESULTS

DnaB contains domains homologous to the DnaD-Nd and DnaD-Cd domains, defining DDBH1 and DDBH2 domains

Pair-wise sequence identity between DnaD and DnaB is insignificant, at only 16% for the full-length proteins in a ClustalW alignment. Evidence for an extensive relationship between DnaB and DnaD was obtained using HHpred (46) (http://toolkit.tuebingen.mpg.de/HHpred). HHpred first performs a PSI-BLAST search based upon the query sequence and forms a Hidden Markov Model (HMM) from the resulting alignment. A HMM is primarily a table of probabilities of different amino acid types being present at each position within a domain (47). The power of HHpred is that it then compares this pattern of

probabilities with HMMs built for alignments based on other sequence families. For example, it is possible to search against HMMs for alignments based on all the sequences with structures in the PDB. This allows rather distant relationships to be detected since it is the conserved features of two families of sequences which are compared rather than features in any one sequence from a particular family. The top hit using full length DnaB as the query was for residues 300–393 to the sequence of the DnaD-like C-terminal domain (PDB code 2zc2, $E = 3.4e\text{-}07$), and a much weaker hit for residues 38–112 to the 3D structure for the DnaD N-terminal domain (PDB code 2v79, $E = 1.1$).

The list of lower scoring HHpred hits for the N-terminal domain search was dominated by helix-turn-helix motifs (in particular winged helix-turn-helix domains), and the hit to the DnaD-Nd structure was only marginally stronger than these. If, however, one looks either side of the region aligned by HHpred to DnaD-Nd, one finds that DnaB and DnaD contain very similar secondary structure elements (Supplementary Figure S1). In particular, a helix-strand-helix motif is predicted at the N-terminus of DnaB. In DnaD, the putatively corresponding H1′-S1′-H2′ elements are key to the formation of both the dimer and the tetramer (16). A dimeric arrangement would also account for the lack of a possible partner for the predicted S1′ strand in a monomeric DnaB.

We term these two regions of homology DDBH1 and DDBH2 domains (DnaD DnaB Homology 1 and 2). The relative positions of these domains in DnaD and DnaB are shown in Figure 2. We note that concurrently with our work the Pfam *DnaD* family was merged into the *DnaB_2* family (48). The evidence for the merge is undocumented. *DnaB_2* (and the former *DnaD*) both encode domains similar in extent to DDBH2.

Another piece of evidence linking the DnaD and DnaB sequences emerged from HHpred analysis of the sequence of residues 146–305 of *B. subtilis* DnaB. This is the sequence between the regions of homology identified above. The top scoring hit was to the 2zc2 structure again, albeit with a poor *E*-value of 2.3. In the absence of other indications this hit would be considered extremely marginal, but in the emerging picture of a structural link between DnaD and DnaB it is very compelling. The fact that DnaD and DnaB contain other strongly matching domains does not influence the scoring of HHpred when a portion of sequence not overlapping these domains is used as the query. Therefore, the central domain of DnaB is a divergent DDBH2 domain.

Figure 2. Domain structure of DnaD and DnaB. The figure shows the experimentally determined secondary structure of DnaD and the predicted secondary structure of DnaB (Supplementary Figure S1). The secondary-structure elements are color coded to show the regions for which HHpred detects homology. The estimated extent of the DDBH1 and DDBH2 domains in the two proteins is shown.

The scheme in Figure 2 shows that domains found in DnaD account for the vast majority of the secondary structural elements predicted for *B. subtilis* DnaB, with intervening linkers predominantly predicted as unstructured (Supplementary Figure S1) and is also consistent with the pattern of conservation across the *B. subtilis* DnaB sequences (Supplementary Figure S4). Additionally, there is a predicted helical region (*B. subtilis* DnaB residues 443–463), which we have left unclassified. There are also two conserved regions of predicted order but uncertain secondary structure (*B. subtilis* DnaB residues 155–163 and 425–433) which we have not marked in our domain graphics.

The DnaD/DnaB superfamily

To collect as large a representative set of DnaD and DnaB related proteins as possible we searched the 50% maximum sequence identity Uniprot50 sequence collection using the C-terminal domains (residues 129–232 of DnaD and residues 306–472 of DnaB) as seeds in two separate iterative searches using jackhmmer from the hmmer3 package. The searches yielded 313 and 318 sequences, respectively. The two sets of sequences were strongly overlapping with the combined non-redundant set containing 326 sequences. This overlap is expected if the two query sequences indeed are homologous, and the search procedure is obtaining a sufficiently exhaustive search for homologues. The full-length sequences corresponding to the combined set were then aligned using ClustalW, and using this alignment the pair-wise identities of each sequence to DnaD or DnaB were individually calculated (Supplementary Figure S5). A subset of the sequences clearly has a similarity to DnaD well above the baseline value, and a separate subset has a similarity to DnaB above the baseline value. The lengths of these sequences also show good consistency in these two subsets (Supplementary Figure S5). We define these two subsets as our working set of DnaD and DnaB orthologues. Sequence alignments of these two sets of sequences are shown in Supplementary Figures S3 and S4. A HMM was built (using hmmer3) corresponding to the DDBH2 domain of DnaD in an alignment of the full 326 sequence set of DnaD and DnaB related proteins (Figure 3). Note the prominent YxxxIxxxW motif and the highly conserved threonine (or serine) residue seven residues C-terminal of this motif. The significance of these features is discussed below.

What is the extent of a 'structural' and 'functional' C-terminal domain of DnaD?

As a preliminary to NMR-based studies of DNA binding, we aimed to define a construct that correctly spanned the functional part of the DnaD C-terminal domain; we show further below that the correct definition of such a domain is not straightforward and previous structural studies have used constructs that are too short. As noted above, the closest available model is the 2zc2 structure, in which there are four major helices (I–IV) and a very short helix V

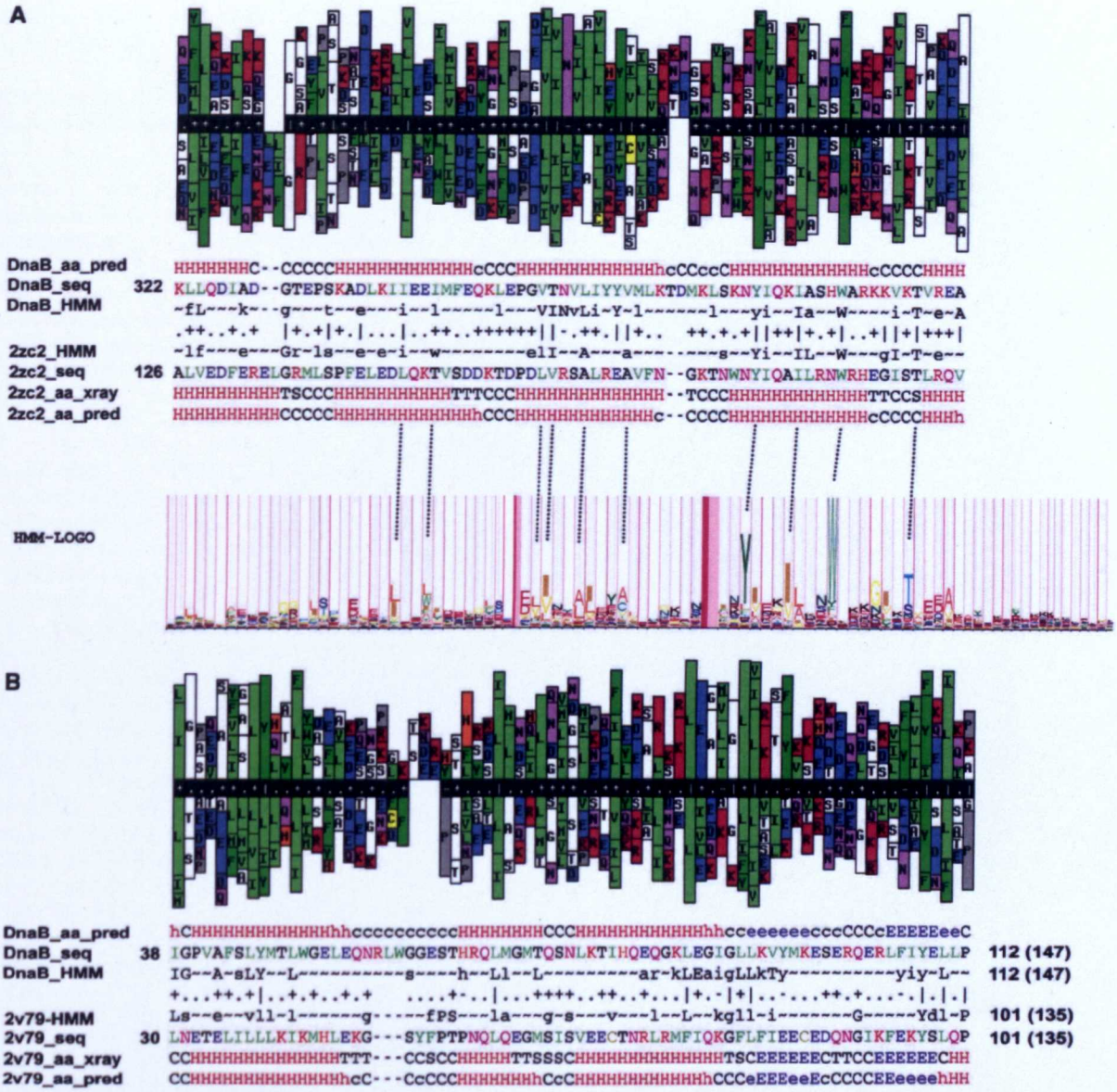


Figure 3. HMM relationships between DnaD and DnaB. (**A**) Top HHpred PDB hit obtained with a query of residues 306–472 of DnaB. The hit is the 2zc2 structure which is closely homologous to the *B. subtilis* DnaD-Cd. The HMM-LOGO for the DDBH2 domain of the combined sequence set of 326 sequences is shown below the HHpred alignment, with the correspondence of the highest probability residue types between the two HMMs shown by dashed lines. (**B**) Top HHpred PDB hit obtained with a query of residues 1–147 of DnaB. The hit is the 2v79 X-ray structure of the *B. subtilis* DnaD-Nd domain. In the HHpred, alignment parts of (A) and (B) the following abbreviations are used: ss_pred, secondary-structure prediction by PSIPRED; seq, sequence; HMM, HHpred representation of HMM where upper and lower case letters denote amino acids with >60 and >40% probability, respectively. Above the alignment are histograms of the DnaB and the PDB HMMs. Bars are proportional to probability of a certain amino acid type at each position, with bars omitted for probabilities <10%. The alignments are images returned by HHpred. The HMM-LOGO was created using the logomat-M server (http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi).

which spans four residues (from L194 to V197, equivalent to V196 to A199 in *B. subtilis* DnaD). Consistent with this, alignment of sequences obtained using PSI-BLAST (49) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with residues 129–232 of *B. subtilis* DnaD as the query yielded an alignment in which good conservation was observed up to approximately *B. subtilis* DnaD residue 197, but was highly divergent after that point, with most sequences continuing for at least 30 residues further (data not shown, but see Figure 1B and Supplementary Figures S3 and S4 for an alignment of a more refined set of sequences based on our analyses below). In contrast, secondary structure prediction programs used by the Phyre server suggested a much longer helix V, extending as far as residue R207 (Supplementary Figure S1), with the remainder of the C-terminal polypeptide disordered. In light of this uncertainty as to the proper extent of a structural and functional C-terminal domain, the first construct that we

studied below by NMR included the whole of the C-terminal tail.

NMR analysis of the structure of the DnaD C-terminal domain

The backbone NMR assignment of *B. subtilis* DnaD residues 129–232 was completed using standard methods, with all non-prolyl residues assigned except residues L144, F226, Y227 and residues from the His-tag remaining after thrombin cleavage. The spectra were relatively poor for a protein of this size, with ^{15}N relaxation data suggesting a correlation time of ~12 ns (Figure 1A). This may be a combination of the effect of the unstructured C-terminal portion, and a degree of transient self association. TALOS analysis (50) of the backbone chemical shift data indicates the presence of five helices: I L129-A138; II L144-D156; III A160-S173; IV F178-N191; V V196-F206 (Figure 1A). ^{15}N T_2 data also confirm that the protein is predominantly disordered from residue 206 onwards (Figure 1A). Therefore, the experimental data clearly support the secondary structure predictions that helix V is considerably more extensive than suggested by available crystal structures 2zc2 and 2i5u, or by sequence alignments.

Analysis of the 3D ^{15}N- and ^{13}C-NOESY HSQC spectra yielded >50 long-range ($|i-j| > 4$) NOEs that corroborate the overall fold of the domain based on the 2zc2 structure, however the quality of the spectra was insufficiently high to allow full NMR structure determination. Despite a detailed search of the spectra, no NOEs could be found to properly constrain the final helix V suggesting that this highly hydrophilic helix is not closely packed against the main core fold of the protein. We, therefore, constructed a hybrid homology model based on 2zc2, with the final helix extended as a regular helix to a length indicated by the NMR data (Figure 4A and C). Secondary-structure predictions by Phyre (Supplementary Figure S2) suggest that this helix would be considerably longer in both the 2zc2 and 2i5u full length sequences; it is, therefore, likely that helix V is as long as observed in the *B. subtilis* protein in many, and probably all, DDBH2 domains.

Helices I–IV form a packed structure which corresponds to the well conserved region of the alignments in Figure 1B and Supplementary Figure S3. The structure has a well packed hydrophobic core: residues F133, F137, L141, L149, I163, A166, L167, F178, I181 and L185 correspond (with a two residue shift in numbering) to the core residues in the 2zc2 structure, F131, L135, L139, F147, V161, A164, L165, W176, I179 and L183, respectively. Examples of clearly assignable long-range NOEs that define the core fold were from the aromatic ring of F133 to the methyl groups of L149, I163 and L167, and from the methyl group of A166 to the methyl groups of I181 and L185.

The peripheral nature of helix V is further supported by the observation that deletion of this helix in the DnaD-Cd196 truncated mutant (see below) yielded a readily purifiable protein which is well folded as judged by the relatively small perturbations of its HSQC spectrum compared to the DnaD-Cd232 construct (Supplementary Figure S8D). Furthermore, the strongly up-field shifted signals from the methyl groups of residues I184 and L185 (at −0.60 and −0.85 p.p.m., respectively, in DnaD-Cd) are still observed in 1D spectra of DnaD-Cd196. This shift for I184 arises from close proximity to W188 thus the detailed arrangement of this crucial area of sequence is not perturbed significantly by the deletion of helix V. In contrast further truncation, deleting helix IV, yielded a protein that was insoluble and probably unfolded. Consistent with its exposure, helix V also has markedly more rapid amide hydrogen–deuterium exchange rates than the other helices: exchange is complete for all residues in helix V after 2 h at 298K, pH 6.5, in contrast to the other helices in which observable intensity is still observed at that time (data not shown). The high conservation of *B. subtilis* DnaD T195 (T193 in 2zc2; T389 in *B. subtilis* DnaB) (Supplementary Figures S3 and S4) suggests that its role may be as a helix capping residue for helix V, which may help to fix the orientation of this helix.

Across the extended family of putative DnaD-Cd orthologues represented in Supplementary Figure S3 (and the subset shown in Figure 1B), the YxxxIxxxW motif (and key residues involved in its packing into the overall structure) is extremely highly conserved. The six most highly conserved hydrophobic residues in the alignment are Y180 (100%), L185 (98%), W188 (96%), I184 (94%), A166 (92%) and A170 (92%). The mean absolute conservation of the hydrophobic residues in this alignment in the structured part of the C-terminal domain is 62% (SD = 26%). The structural model shows that residues Y180, I184 and W188 lie on an exposed face of helix IV, whereas L185, A166 and A170 appear to be more packed into the core of the protein, anchoring this helix (Figure 4A).

DNA binding monitored by NMR

To investigate how DnaD-Cd binds to DNA, we titrated a 10-mer ssDNA into $^{15}N/^{13}C$ labeled DnaD-Cd and collected 1D proton, 2D ^{15}N-HSQC and 2D ^{13}C-HSQC spectra. Given the observation from gel shifts that very large complexes were formed between DnaD-Cd and DNA (15), it was anticipated that on adding DNA to DnaD-Cd the effect on the NMR spectrum might be to simply reduce the intensity of peaks across the spectrum, as increasing amounts of protein were sequestered into DNA:DnaD-Cd complexes. Surprisingly, the effects on peak intensities observed were far from uniform as shown in Figure 4B and Supplementary Figure S6. (We return to the reconciliation of the NMR and gel shift data in the discussion.) For some residues (e.g. E134) the crosspeaks decreased in intensity, but reached a plateau at ~30% of the original intensity; for some residues this drop in intensity was much more severe (e.g. I184); and for some residues, the crosspeaks were effectively unchanged (e.g. W229). We interpret the decay to ~30% intensity to be due to an increase in the overall tumbling time of the protein–DNA complex. The drop to 30% intensity is only compatible with the formation of rather small DNA:DnaD-Cd complexes, certainly no larger than

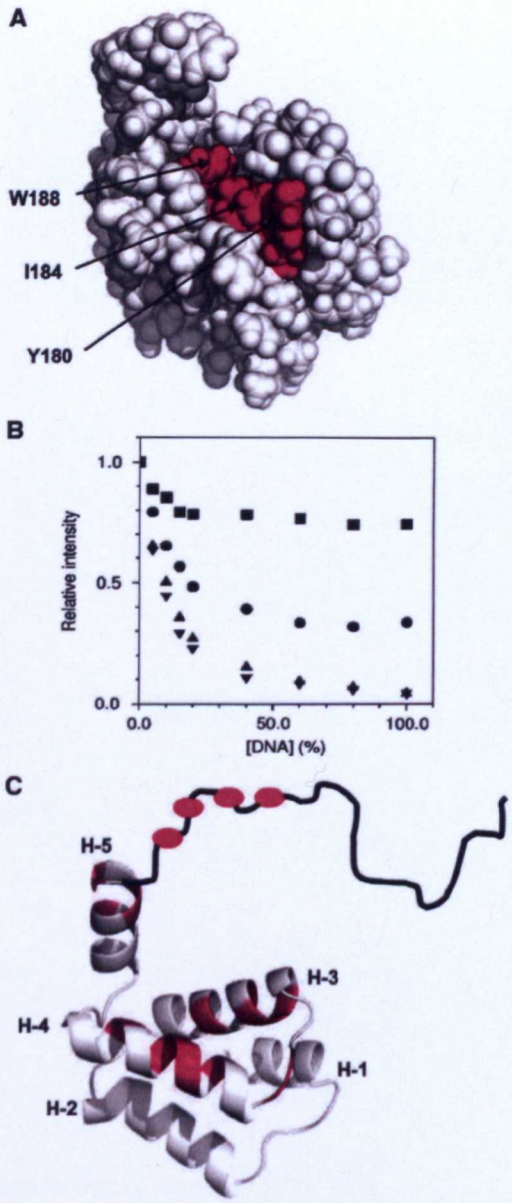


Figure 4. The conserved YxxxIxxxW motif and DNA-binding data mapped onto a structural model of DnaD-Cd. (**A**) A space filling representation of the DnaD-Cd model, with the six most highly conserved residues in DnaD-Cd (A166, A170, Y180, I184, L185 and W188) colored red. Three of these residues (W188, Y180 and I184) form an exposed hydrophobic patch, whereas A166, A170 and L185 appear to anchor this segment to the main hydrophobic core of the protein. (**B**) Examples of three types of behavior observed on titration of 10 mer DNA into DnaD-Cd. HSQC crosspeaks for the majority of residues (such as E134, circles) show a moderate intensity reduction to ~30% of initial value. Some residues show a marked intensity reduction (residue I184, triangles; residue K205, inverted triangles). Some residues (residue W229, squares) show only a very small reduction in intensity. DNA concentration is expressed as fraction of protein concentration. Data for all residues are presented in Supplementary Figure S6. (**C**) Localization of DNA binding by perturbation of ^{15}N HSQC crosspeaks. The structure is shown in cartoon representation in the same orientation as (A), with the C-terminal end of helices labeled. The mobile portion is depicted in sketch form. Residues for which the crosspeak intensity is reduced by >90% after addition of 1:1 10-mer DNA to DnaD-Cd215 are colored red. These are residues 168, 170, 171, 176, 177, 181, 183, 184, 185, 188, 201, 203, 205, 207, 208, 210 and 211. This figure is based on data for DnaD-Cd215 sample since it gave superior HSQC spectra (presumably due to superior tumbling properties). A closely similar subset of residues is selected with an 80% cutoff in the titration with DnaD-Cd. The slightly greater drop in intensity in the DnaD-Cd215 titration probably arises from the increased sharpness of the lines for the apo-protein in this construct.

perhaps a total of 2–3 protein and/or DNA chains. The residues that show negligible decrease in intensity are the most C-terminal of the mobile residues in the apo-protein, and presumably retain sufficient mobility in the complex to be unaffected by the increase in the overall tumbling time.

The most information comes from the residues that show much greater cross-peak attenuation than caused by the increase in tumbling time. These residues must be affected by conformational exchange in the complex, and thus reflect on the DNA-binding region. The greatest such changes in intensity are represented on the 3D structure of DnaD in Figure 4C. This identifies the region of the protein that binds DNA to include residues in helices III and IV (which contains the highly conserved YxxxIxxxW motif), residues in helix V and, most significantly, a large portion of the mobile region (up to residue 215).

^{13}C-HSQC spectra showed effects in line with these interpretations of the ^{15}N-HSQC spectra (Supplementary Figure S7), but did not contribute significantly greater detail. Only very small chemical shift changes (<0.05 p.p.m. for a weighted combined ^{1}H and ^{15}N amide shift) were observed in the ^{15}N HSQC spectra (Supplementary Figure S8), even for residues showing substantial intensity changes.

A conserved YxxxIxxxW in helix IV affects DNA binding

To test the involvement of the highly conserved YxxxIxxxW motif in DNA binding, residues Y180 and W188 were mutated to alanines in full-length DnaD and the DNA-binding properties of the mutant proteins were examined. Native DnaD binds to supercoiled pBSK and forms scaffold complexes that can be detected as shifted bands in agarose gels (12,15). The Y180A and W188A mutant proteins were able to shift pBSK in gel shift assays but at higher concentrations relative to native DnaD indicating a DNA-binding defect (Figure 5A). Direct verification was obtained with comparative gel shift assays showing that Y180A and W188A exhibited defective DNA-binding activities with ds and ssDNA (Figure 5B). This defect was more pronounced with the Y180A protein. We conclude that Y180 and W188 are directly involved in DNA binding.

To examine whether a DnaD polypeptide up to and including helix IV, but lacking helix V and the unstructured C-terminal region, was sufficient for DNA binding we truncated DnaD-Cd at position V196 (DnaD-Cd196). DnaD-Cd binds DNA but is unable to form scaffold structures detectable by plasmid shifts and its DNA-binding

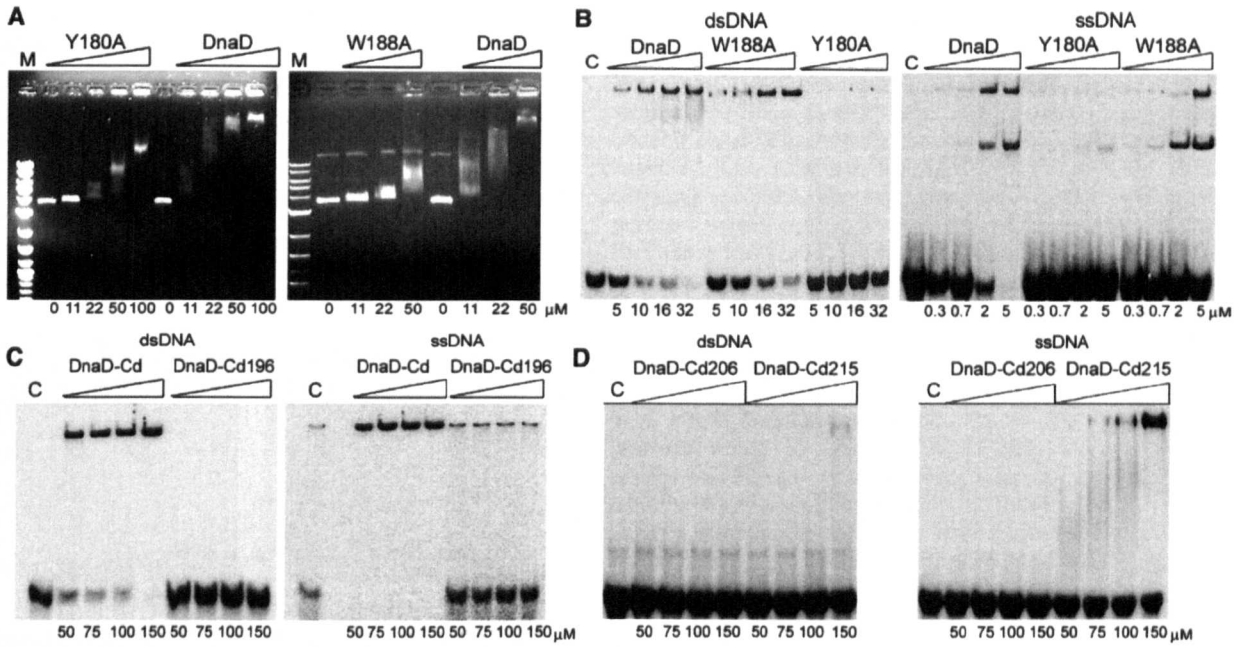


Figure 5. DNA binding of DnaD-Cd and truncation mutants. (**A**) Y180A and W188A have reduced DNA-binding activity. Agarose gel shift assays using DnaD, Y180A and W188A complexes and supercoiled pBSK plasmid. Comparative experiments were carried out with increasing concentrations of proteins, as indicated. Both mutant proteins were defective in DNA binding compared to wild-type DnaD. (**B**) Direct comparison of the DNA-binding activities of full length DnaD and the mutant Y180A, W188A proteins. Gel shift assays were carried out in 100 mM NaCl, 50 mM Tris pH 7.5, 1 mM EDTA, 1 mM DTT, 2.2% v/v glycerol with 0.625 nM ssDNA or 0.125 nM dsDNA probes at increasing concentrations of proteins, as indicated. (**C**) DnaD-Cd196 is deficient in DNA binding. DNA-binding reactions were carried out with increasing concentrations of DnaD and DnaD-Cd196 proteins as indicated in 30 mM NaCl, 50 mM Tris pH 7.5, 1 mM EDTA, 1 mM DTT and 0.625 nM ss or dsDNA probes, as indicated. (**D**) DnaD-Cd215 binds DNA but DnaD-Cd206 does not. DNA-binding reactions were carried out with increasing concentrations of DnaD-Cd206 and DnaD-Cd215 proteins as in described in panel (C).

activity can therefore only be detected by classical gel shift assays (15). DnaD-Cd196 did not bind DNA in gel shift assays (Figure 5C) and NMR DNA titration experiments (Supplementary Figure S8D), suggesting that helix IV by itself is not sufficient for DNA binding, despite the fact that it is a part of the DNA-binding module.

Helix V and an unstructured C-terminal region containing residues 206–215 are involved in DNA binding

Results of the NMR DNA titrations suggested that helix V and part of the unstructured C-terminus up to residue 215 also participate in DNA binding. We constructed two more truncated versions of DnaD-Cd, the DnaD-Cd206 and DnaD-Cd215 polypeptides. Gel shift assays demonstrated that DnaD-Cd206 did not bind DNA (Figure 5D), whereas DnaD-Cd215 did (though considerably weaker than full length DnaD-Cd (compare Figure 5B and C). NMR DNA titration experiments confirmed that DnaD-Cd206 did not bind DNA but DnaD-Cd215 did (Supplementary Figures S6 and S8). The combined data show that helix V and part of the unstructured C-terminus (residues 206–215) are essential for DNA binding and together with YxxxIxxxW motif in helix IV constitute the complete DNA–binding module of DnaD.

Helix V and the C-terminal unstructured region of DnaD are essential for *B. subtilis* viability

Bacillus subtilis DnaD is essential for viability. To determine if the mutant DnaD proteins with altered DNA-binding activity *in vitro* are functional *in vivo* and able to support cell growth, we fused the mutant alleles (*dnaDY180A*, *dnaDW188A* and *dnaDV196-stop*) to the IPTG-inducible promoter Pspank(hy), introduced these alleles into an ectopic site on the chromosome (*amyE*), and tested for the ability to complement loss of function mutations in *dnaD*. As controls, we fused the wild-type *dnaD* and the temperature sensitive *dnaD23ts* mutation (22,41,42) to Pspank(hy). Using western blotting with polyclonal anti-DnaD antibodies, we verified that induction by IPTG led to overproduction of protein, demonstrating that these proteins are not inherently unstable *in vivo* (data not shown).

We assessed whether ectopic copies of wild-type or mutant *dnaD* could rescue the growth defect of a *dnaD23ts* strain at the non-permissive temperature. Since Pspank(hy) is not fully repressed, these experiments were carried out in the absence of IPTG and relied on the basal, leaky, expression of the *dnaD* alleles from Pspank(hy). We found that wild-type *dnaD*, *dnaDY180A* and *dnaDW188A*, but not *dnaD-V196-STOP*, were capable of restoring growth at the non-permissive temperature as judged by growth on LB agar (Figure 6A). The lack of

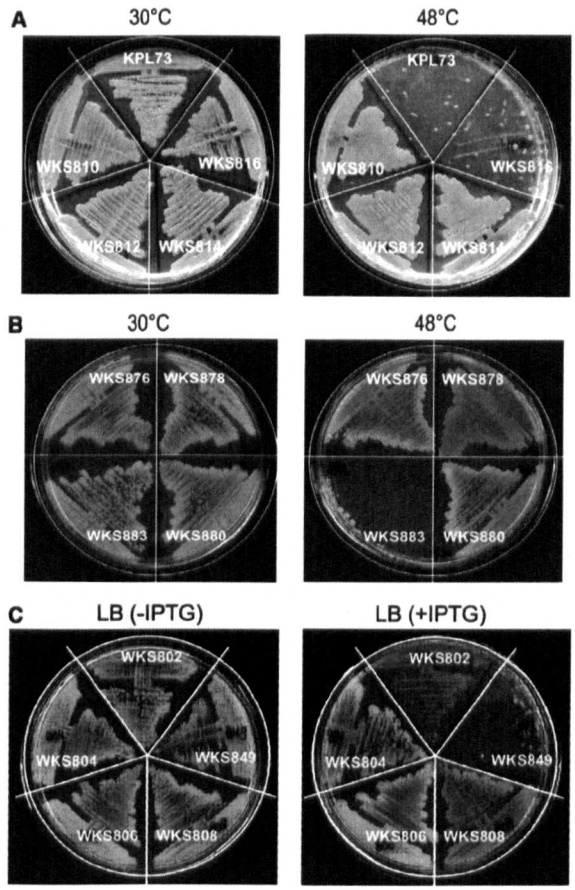


Figure 6. Biological significance of DnaD binding to DNA. (**A**) Truncated DnaD-V196-STOP does not rescue growth of *dnaD23ts* at the non-permissive temperature. The *dnaD23ts* strains carrying an ectopic inducible wild-type *dnaD* (WKS810), *dnaDY180A* (WKS812), *dnaDW188A* (WKS816) or *dnaD-V196-STOP* (WKS818) were grown on LB agar. A *dnaD23ts* strain without inducible *dnaD* (KPL73) is shown as control. Plates were incubated overnight at 30°C or 48°C as indicated. (**B**) DnaD-Y180A and DnaD-W188A sustain growth. Strains carrying an inducible wild-type DnaD (WKS876), Y180A (WKS878), W188A (WKS880) and DnaD23 (WKS883) can support growth at 30°C when the native copy of *dnaD* is knocked out. LB agar plates were incubated over night at the indicated temperature. As a control, 48°C is shown for WKS883. (**C**) Overproduction of wild-type DnaD causes growth inhibition. Strains were grown on LB agar plates, without or with 1-mM IPTG over night at 37°C. WKS802-808 contain an ectopic inducible *dnaD* (or mutant thereof) in an otherwise wild-type background, whereas WKS849 contains a heterologous origin and is inactivated for *oriC*. Information on all strains is provided in Table1.

complementation by *dnaD-V196-STOP* indicates that an essential function of DnaD, presumably DNA-binding activity, is located in the C-terminal region, after V196. We also verified that an ectopic inducible copy of *dnaD23* did not complement a *dnaD23ts* mutation (data not shown), consistent with previous findings (3).

The complementation of *dnaD23ts* by *dnaDY180A* and *dnaDW188A* could be due to true complementation or to the formation of mixed multimers with DnaD23 protein. To test if the *dnaDY180A* and *dnaDW188A* alleles truly function to support cell growth, we introduced a deletion–insertion mutation of *dnaD* into strains expressing each of the different *dnaD* alleles from Pspank(hy). We found that the native *dnaD* could be deleted in strains expressing wild-type *dnaD*, *dnaDY180A* and *dnaDW188A* (Figure 6B), but not *dnaD-V196-STOP*. As expected, knocking out the native *dnaD* from the strain carrying the inducible *dnaD23ts* rendered this strain temperature sensitive (Figure 6B). These results demonstrate that the Y180A and W188A missense mutants, that have reduced DNA-binding activity *in vitro* (Figure 5A and B), can support growth whereas the C-terminally truncated DnaD, that has little or no DNA-binding activity, cannot. In line with these observations, we were not able to obtain transformants with a plasmid that would result in a truncation of the *dnaD* gene at the native locus (corresponding to DnaD-V196-STOP), whereas it was readily introduced in a merodiploid strain harboring an ectopic copy of wild-type *dnaD* (data not shown).

Even though the missense mutants DnaDY180A and DnaDW188A were able to complement loss of function mutations in *dnaD*, they had altered function *in vivo*. In the course of analyzing the phenotypes of the Pspank(hy)-*dnaD* fusions, we found that overexpression of wild-type DnaD upon the addition of 1 mM IPTG caused a growth defect as judged by growth on LB agar plates (Figure 6C, strain WKS802). This inhibitory effect was not due to DnaD acting on the chromosomal origin of replication, *oriC*. In a strain containing the heterologous origin of replication *oriN* and a deletion in *oriC* (51), overexpression of DnaD (with 1 mM IPTG) still caused a severe defect in cell growth (Figure 6, strain WKS849). In contrast, overexpression of DnaDY180A, DnaDW188A and DnaD-V196-STOP in otherwise wild-type cells did not cause an obvious defect in growth as judged on LB agar plates (Figure 6C). Since these mutant proteins have altered DNA-binding activity *in vitro*, we suggest that the inability to inhibit growth when overexpressed is due to decreased DNA binding *in vivo*.

We also compared other growth characteristics of strains expressing either *dnaDY180A* or *dnaDW188A* with those of strains expressing wild-type DnaD. These experiments were carried out in cells deleted for the endogenous wild-type *dnaD* (*ΔdnaD*) and expressing each of the *dnaD* alleles from Pspank(hy). We tested for growth on agar plates made with defined minimal medium and succinate as carbon source. The mutants were able to grow under these conditions indicating that they support viability under rapid and slow growth conditions. The mutants also grew and sporulated on agar plates with Difco sporulation medium (data not shown). Finally, we measured doubling times in liquid LB medium and minimal succinate medium. The wild-type and *dnaDW188A* mutant had similar doubling times, but the *dnaDY180A* mutant had a slightly longer doubling time indicating a small defect in cell growth (Table 2). Taken together, these data indicate that the DnaDY180A and DnaDW188A, with a reduced ability to bind to DNA, likely have reduced activity *in vivo*, but nonetheless support cell growth, viability and sporulation.

Table 2. Growth rate of *dnaD* mutants

Strain	Doubling time (minutes ± SE)[a]	
	LB	Succinate
WKS876 (wt)	32.0 ± 0.5	64.6 ± 3.5
WKS878 (Y180A)	35.5 ± 0.7	75.21 ± 2.3
WKS880 (W188A)	30.8 ± 0.3	65.72 ± 2.0

[a]Growth rates of cells deleted for the native *dnaD* (*ΔdnaD*) and harboring an ectopic copy of wild-type *dnaD* (WKS876), *dnaDY180A* (WKS878) or *dnaDW188A* (WKS880), each under control of Pspank(hy), were grown in liquid LB or defined minimal medium with succinate as carbon source at 37°C in the absence of IPTG. The culture doubling time is indicated ± the standard error of the mean ($n = 3$).

DISCUSSION

The structural similarities between DnaD and DnaB correlate with many functional properties of the two proteins. A common property of the DDBH1 domains of both DnaD and DnaB is that they form dimers which associate into tetramers (15,24). In the DDBH1 domain of *B. subtilis* DnaD, dimer–tetramer association is mediated by a helix-strand-helix (H1′-S1′-H2′) element flanking the N-terminus of the winged helix-turn-helix core while a helix (H3′) flanking the C-terminus of the helix-turn-helix core is involved in higher order cross-tetramer oligomerization (16). As both structural elements are preserved in DnaB, the same oligomerization properties are likely to be also conserved in the DDBH1 domain of DnaB.

The DDBH2 domains of DnaD and DnaB are monomeric but oligomerize upon binding to DNA (15,24). Our data show that the C-terminal part of the *B. subtilis* DnaD DDBH2 domain and part of its disordered C-terminal extension are involved in DNA binding. The equivalent region of *B. subtilis* DnaB, encompassing part of the C-terminal region of its second DDBH2 (residues 365–400) and its C-terminal extension (residues 400–428), has also been shown experimentally to bind DNA and to mediate DNA–induced higher order oligomerization (24). DNA-binding and DNA-induced oligomerization appear to be functionally and structurally conserved in the DDBH2 domain in both proteins. Furthermore, in alignments of DnaD sequences the segment corresponding to helices I and II is predominantly acidic, whilst the segment comprising the DNA-binding module (helices IV, V, and mobile residues to approximately *B. subtilis* DnaD residue 215) is predominantly basic (Figure 1B), consistent with the latter having a DNA-binding function. This acidic/basic charge distribution is closely mimicked in DnaB (data not shown).

DNA-binding winged helix-turn-helix motifs bind exclusively to dsDNA (52). However, a polypeptide fragment of *B. subtilis* DnaB (residues 1–300), encompassing the (DDBH1)-(DDBH2) fragment, binds ssDNA and not dsDNA (24) and a polypeptide fragment of *B. subtilis* DnaD (residues 1–128), encompassing the DDBH1 domain, does not bind DNA (14,15). It therefore appears that the function of the DDBH1 domains in both DnaD and DnaB is not to bind DNA but to mediate dimerization–tetramerization. From the above we predict that the ssDNA-binding activity of the *B. subtilis* DnaB fragment encompassing (DDBH1)-(DDBH2) resides on the DDBH2 domain. Hence, the middle DDBH2 domain of *B. subtilis* DnaB has likely evolved to bind exclusively ssDNA and lost the ability to bind dsDNA.

DDBH2 motifs are found in many bacterial and phage proteins of unknown functions (15). In several of these proteins two tandem DDBH2 domains are present, equivalent to the tandem DDBH2 domains found in *B. subtilis* DnaB, while in other proteins DDBH2 domains are fused in putative phage initiation factors. In the orf4 protein from phage 7201 of *S. thermophilus* the DDBH2 domain is fused to a RepA-like domain associated with initiation of DNA replication (15). Similar modular architectures have been observed in a putative replication initiation protein from phage A118 of *Listeria monocytogenes* and in Protein 20 from *Lactobacillus plantarum* (22). Therefore, the modular character of the DDBH1 and DDBH2 motifs and their structural/functional properties appear to be conserved across many bacterial and phage proteins beyond the DnaD and DnaB primosomal proteins.

Bacillus subtilis DnaD-Cd oligomerizes upon binding to DNA (15). We, therefore, expected that titration of DNA into DnaD-Cd would lead to a progressive, complete and uniform decrease of signal intensity in NMR experiments, and would thus lead to little useful information. However, the attenuation of signals was far from uniform, and did not lead to complete loss of signal intensity. The observation of conformational broadening on binding oligonucleotides is in contrast quite common, and generally ascribed to the possibility of a variety of binding registers. Calculations show that the observation of very small shift changes and differential broadening in a titration is compatible with exchange (substantially on the slow side of intermediate exchange) between the free protein and a bound form in which peaks corresponding to certain residues are severely broadened (data not shown). The observation of large complexes in the gel shift experiments, and the rather small complexes suggested in the NMR experiments, may be a consequence of the different relative concentrations of DNA and protein in these two experiments if the complexes contain a superstoichiometric ratio of protein to DNA. Detailed analysis of these complexes will be needed with alternative methods to resolve this issue.

In summary, the bacterial replication initiation proteins DnaD and DnaB share similar domain organizations. Elucidation of their domain structures greatly helps to unify and explain their similar properties. The DDBH1 domain contains a conserved N-terminal helix-strand-helix element that forms β-sheets in dimers–tetramers, and a C-terminal helix that mediates cross-tetramer oligomerization. The DDBH2 domain has a C-terminal DNA-binding module comprised of a conserved YxxxIxxxW motif (in the penultimate helix), the last helix and part of the unstructured region at the C-terminus. Thus the DDBH2 domain (which is the most strongly conserved domain between the two

proteins) displays the curious contrast of a very well conserved motif, built on a well conserved protein core structure, coupled to a loosely associated helix and an unstructured region, both of the latter being ill conserved. Together these elements achieve DNA binding and high order oligomerization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

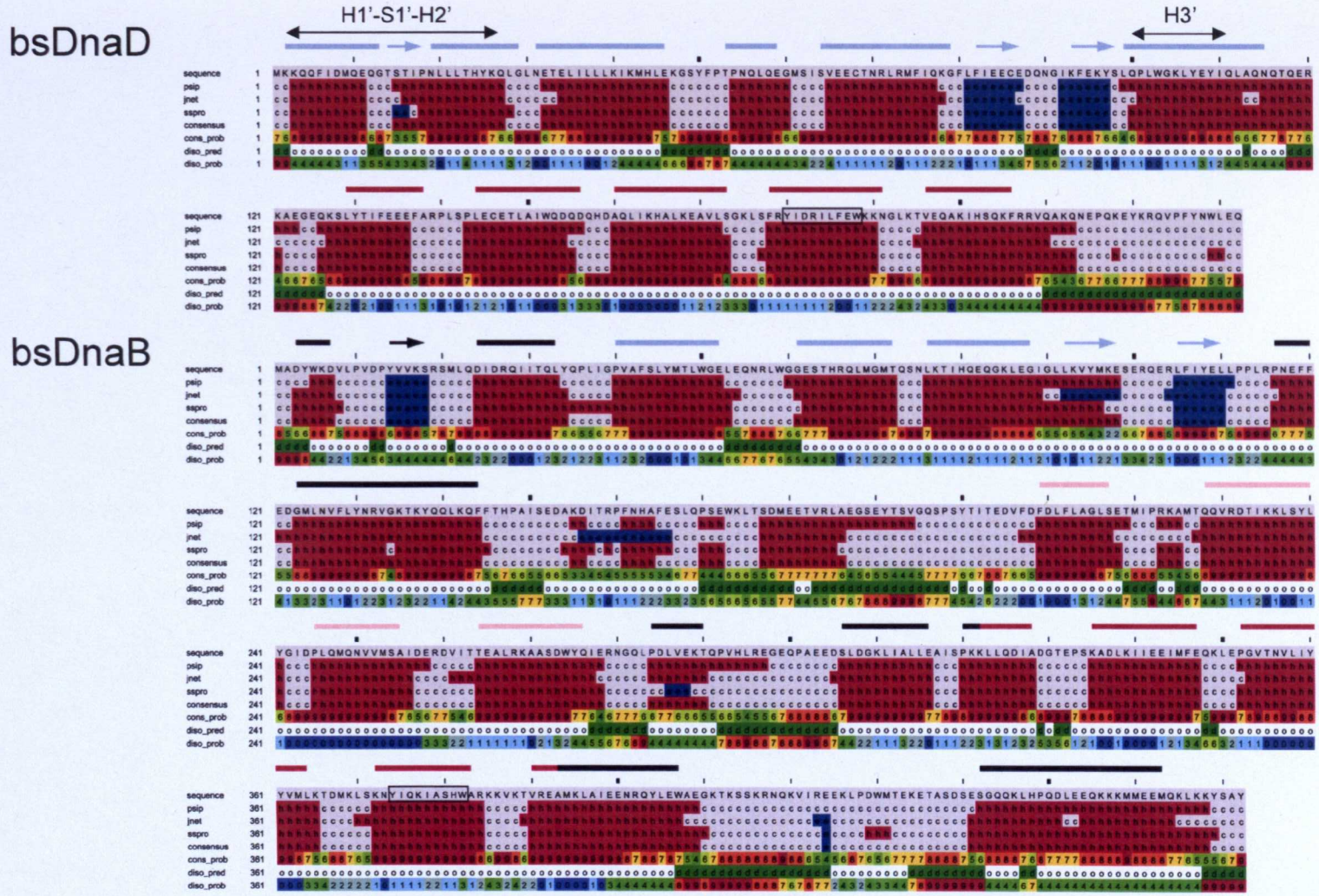
Biotechnology and Biological Sciences Research Council (BBSRC) (grant BB/E006450/1 to P.S.); BBSRC Ph.D. studentship (to F.Y.M.); National Institute of Health (grant GM41934 to A.D.G.); Rubicon fellowship from the Netherlands Organization for Scientific Research (to W.K.S.). Funding for open access charge: BBSRC.

Conflict of interest statement. None declared.

REFERENCES

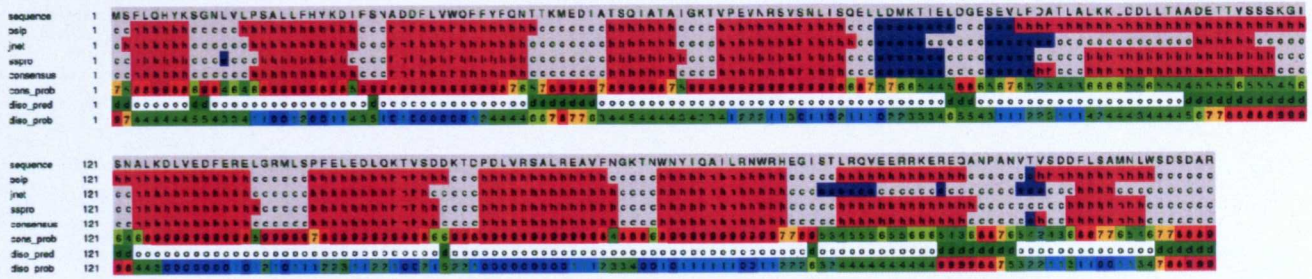
1. Bruand,C., Ehrlich,S.D. and Janniere,L. (1995) Primosome assembly site in *Bacillus subtilis*. *EMBO J.*, **14**, 2642–2650.
2. Marsin,S., McGovern,S., Ehrlic,H.S.D., Bruand,C. and Polard,P. (2001) Early steps of *Bacillus subtilis* primosome assembly. *J. Biol. Chem.*, **276**, 45818–45825.
3. Bruand,C., Velten,M., McGovern,S., Marsin,S., Serena,C., Ehrlich,S.D. and Polard,P. (2005) Functional interplay between the *Bacillus subtilis* DnaD and DnaB proteins essential for initiation and re-initiation of DNA replication. *Mol. Microbiol.*, **55**, 1138–1150.
4. Mott,M.L. and Berger,M. (2007) DNA replication initiation: mechanisms and regulation in bacteria. *Nat. Rev. Microbiol.*, **5**, 343–354.
5. Zakrzewska-Czerwinska,J., Jakimowicz,D., Zawilak-Pawlik,A. and Messer,W. (2007) Regulation of the initiation of chromosomal replication in bacteria. *FEMS Microbiol. Rev.*, **31**, 378–387.
6. Smits,W.K., Goranov,A.I. and Grossman,A.D. (2010) Ordered association of helicase loader proteins with the *Bacillus subtilis* origin of replication *in vivo*. *Mol. Microbiol.*, **75**, 452–461.
7. Velten,M., McGovern,S., Marsin,S., Ehrlich,S.D., Noirot,P. and Polard,P. (2003) A two-protein strategy for the functional loading of a cellular replicative DNA helicase. *Mol. Cell*, **11**, 1009–1020.
8. Ioannou,C., Schaeffer,P.M., Dixon,N.E. and Soultanas,P. (2006) Helicase binding to DnaI exposes a cryptic DNA-binding site during helicase loading in *Bacillus subtilis*. *Nucleic Acids Res.*, **34**, 5247–5258.
9. Nunez-Ramirez,R., Velten,M., Rivas,G., Polard,P., Carazo,J.M. and Donate,L.E. (2007) Loading a ring: Structure of the *Bacillus subtilis* DnaB protein, a co-loader of the replicative helicase. *J. Mol. Biol.*, **367**, 764–769.
10. Bruand,C., Farache,M., McGovern,S., Ehrlich,S.D. and Polard,P. (2001) DnaB, DnaD and DnaI proteins are components of the *Bacillus subtilis* replication restart primosome. *Mol. Microbiol.*, **42**, 245–255.
11. Turner,I.J., Scott,D.J., Allen,S., Roberts,C.J. and Soultanas,P. (2004) The *Bacillus subtilis* DnaD protein: a putative link between DNA remodeling and initiation of DNA replication. *FEBS Lett.*, **577**, 460–464.
12. Zhang,W., Carneiro,M.J.V.M., Turner,I.J., Allen,S., Roberts,C.J. and Soultanas,P. (2005) The *Bacillus subtilis* DnaD and DnaB proteins exhibit different DNA remodeling activities. *J. Mol. Biol.*, **351**, 66–75.
13. Zhang,W., Allen,S., Roberts,C.J. and Soultanas,P. (2006) The *Bacillus subtilis* primosomal protein DnaD untwists supercoiled DNA. *J. Bacteriol.*, **188**, 5487–5493.
14. Zhang,W., Machón,C., Orta,A., Phillips,N., Roberts,C.J., Allen,S. and Soultanas,P. (2008) Single-molecule atomic force spectroscopy reveals that DnaD forms scaffolds and enhances duplex melting. *J. Mol. Biol.*, **377**, 706–714.
15. Carneiro,M.J., Zhang,W., Ioannou,C., Scott,D.J., Allen,S., Roberts,C.J. and Soultanas,P. (2006) The DNA-remodelling activity of DnaD is the sum of oligomerization and DNA-binding activities on separate domains. *Mol. Microbiol.*, **60**, 917–924.
16. Schneider,S., Zhang,W., Soultanas,P. and Paoli,M. (2008) Structure of the N-terminal oligomerization domain of DnaD reveals a unique tetramerization motif and provides insights into scaffold formation. *J. Mol. Biol.*, **376**, 1237–1250.
17. Winston,S. and Sueoka,N. (1980) DNA membrane association is necessary for initiation of chromosomal and plasmid replication in *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA*, **77**, 2834–2838.
18. Hoshino,T., McKenzie,T., Schmidt,S., Tanaka,T. and Sueoka,N. (1987) Nucleotide sequence of *Bacillus subtilis dnaB*: a gene essential for DNA replication initiation and membrane attachment. *Proc. Natl Acad. Sci. USA*, **84**, 653–657.
19. Sueoka,N. (1998) Cell membrane and chromosome replication in *Bacillus subtilis*. *Prog. Nucleic Acids Res.*, **59**, 34–53.
20. Watabe,K. and Forough,R. (1987) Effects of temperature sensitive variants of the *Bacilus subtilis dnaB* gene on the replication of low-copy number plasmid. *J. Bacteriol.*, **169**, 4141–4146.
21. Sato,Y., McCollum,M., McKenzie,T., Laffan,J., Zuberi,A. and Sueoka,N. (1991) *In vitro* type II binding of chromosomal DNA to membrane in *Bacillus subtilis*. *J. Bacteriol.*, **173**, 7732–7735.
22. Rokop,M.E., Auchtung,J.M. and Grossman,A.D. (2004) Control of DNA replication initiation by recruitment of an essential initiation protein to the membrane of *Bacillus subtilis*. *Mol. Microbiol.*, **52**, 1757–1767.
23. Li,Y., Kurokawa,K., Reutimann,L., Mizumura,H., Matsuo,M. and Sekimizu,K. (2007) DnaB and DnaI temperature-sensitive mutants of *Staphylococcus aureus*: evidence for involvement of DnaB and DnaI in synchrony regulation of chromosome replication. *Microbiol.*, **153**, 3370–3379.
24. Grainger,W.H., Machón,C., Scott,D.J. and Soultanas,P. (2010) DnaB proteolysis *in vivo* regulates oligomerization and its localization at *oriC* in *Bacillus subtilis*. *Nucleic Acids Res.*, **38**, 2851–2864.
25. Ishigo-Oka,D., Ogasawara,N. and Moriya,S. (2001) DnaD protein of *Bacillus subtilis* interacts with DnaA, the initiator protein of replication. *J. Bacteriol.*, **183**, 2148–2150.
26. Noirot-Gros,M.F., Dervyn,E., Wu,L.J., Mervelet,P., Errington,J., Ehrlich,S.D. and Noirot,P. (2002) An expanded view of bacterial DNA replication. *Proc. Natl Acad. Sci. USA*, **99**, 8342–8347.
27. Rokop,M.E. and Grossman,A.D. (2009) Intragenic and extragenic suppressors of temperature sensitive mutations in the replication initiation genes *dnaD* and *dnaB* of *Bacillus subtilis*. *PLoS ONE*, **4**, e6774.
28. Schneider,S., Carneiro,M.J., Ioannou,C., Soultanas,P. and Paoli,M. (2007) Crystallization and X-ray diffraction analysis of the DNA-remodelling protein DnaD from Bacillus subtilis. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **63**, 110–113.
29. Huang,C.Y., Chang,Y.W. and Chen,W.T. (2008) Crystal structure of the N-terminal domain of *Geobacillus kaustophillus* HTA426 DnaD protein. *Biochem. Biophys. Res. Comm.*, **375**, 220–224.
30. Kelley,L.A. and Sternberg,M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.
31. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
32. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) ClustalW and ClustalX version 2. *Bioinformatics*, **23**, 2947–2948.
33. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

34. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
35. Arnold,K., Bordoli,L., Kopp,J. and Schwede,T. (2006) The SWISS-MODEL workspace: A web-based environment for protein structure homology modeling. *Bioinformatics*, **22**, 195–201.
36. Brünger,A.T. (1992) *X-PLOR Version 3.1*. Yale University Press, New Haven, Connecticut.
37. Reed,M.A., Hounslow,A.M., Sze,K.H., Barsukov,I.G., Hosszu,L.L., Clarke,A.R., Craven,C.J. and Waltho,J.P. (2003) Effects of domain dissection on the folding and stability of the 43 kDa protein PGK probed by NMR. *J. Mol. Biol.*, **330**, 1189–11201.
38. Farrow,N.A., Muhandiram,R., Singer,A.U., Pascal,S.M., Kay,C.M., Gish,G., Shoelson,S.E., Pawson,T., Forman-Kay,J.D. and Kay,L.E. (1994) Backbone dynamics of a free and a phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochem.*, **33**, 5984–6003.
39. Perego,M., Spiegelman,G.B. and Hoch,J.A. (1988) Structure of the gene for the transition state regulator, abrB: regulator synthesis is controlled by the *spo0A* sporulation gene in *Bacillus subtilis*. *Mol. Microbiol.*, **2**, 689–699.
40. Karamata,D. and Gross,J.D. (1970) Isolation and genetic analysis of temperature-sensitive mutants of *B. subtilis* defective in DNA synthesis. *Mol. Gen. Genet.*, **108**, 277–287.
41. Goranov,A.I., Katz,L., Breier,A.M., Burge,C.B. and Grossman,A.D. (2005) A transcriptional response to replication status mediated by the conserved bacterial replication protein DnaA. *Proc. Natl Acad. Sci. USA*, **102**, 12932–12937.
42. Lemon,K.P., Kurtser,I., Wu,J. and Grossman,A.D. (2000) Control of initiation of sporulation by replication initiation genes in *Bacillus subtilis*. *J. Bacteriol.*, **182**, 2989–2991.
43. Britton,R.A., Eichenberger,P., Gonzalez-Pastor,J.E., Fawcett,P., Monson,R., Losick,R. and Grossman,A.D. (2002) Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of *Bacillus subtilis*. *J. Bacteriol.*, **184**, 4881–4890.
44. Wach,A. (1996) PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *S. cerevisiae Yeast*, **12**, 259–265.
45. Youngman,P., Poth,H., Green,B., York,K., Olmedo,G. and Smith,K. (1989) Methods for genetic manipulation, cloning, and functional analysis of sporulation genes in *Bacillus subtilis*. In Smith,I., Slepecky,R. and Setlow,P. (eds), *Regulation of Procaryotic Development*. American Society for Microbiology, Washington, DC, pp. 65–87.
46. Söding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
47. Schuster-Böckler,B., Schultz,J. and Rahmann,S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
48. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,J.S., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
49. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
50. Cornilescu,G., Delaglio,F. and Bax,A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
51. Berkmen,M.B. and Grossman,A.D. (2007) Subcellular positioning of the origin region of the *Bacillus subtilis* chromosome is independent of sequences within oriC, the site of replication initiation, and the replication initiator DnaA. *Mol. Microbiol.*, **63**, 150–165.
52. Gajiwala,K.S. and Burley,S.K. (2000) Winged helix proteins. *Curr. Opin. Struct. Biol.*, **10**, 110–116.

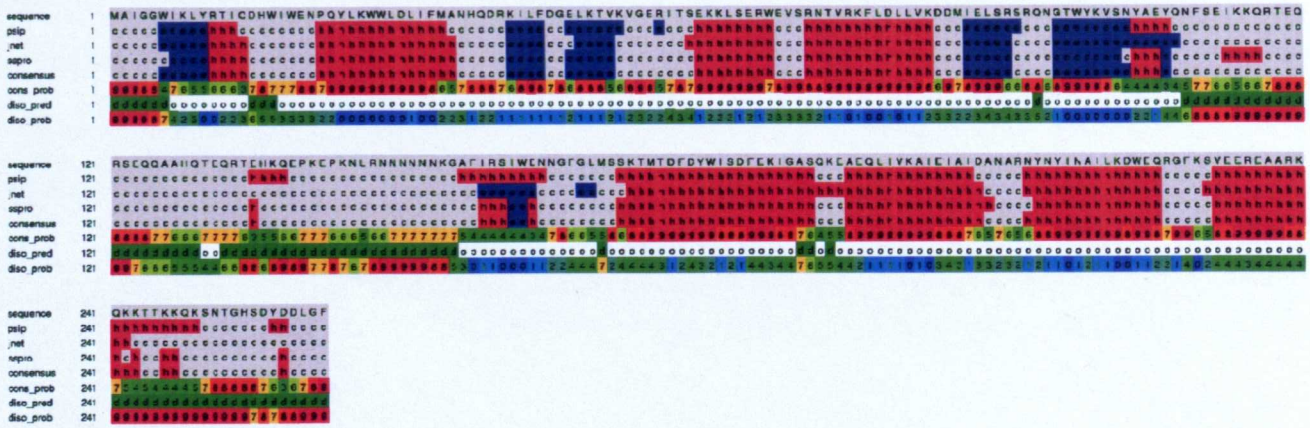


Supplementary Figure 1. Secondary structure and disorder predictions from Phyre for DnaD and DnaB. Phyre uses three secondary structure predictions (psipred, Jnet and sspro) to obtain a consensus prediction. Below the consensus prediction is shown a measure of the confidence in the prediction. The bottom two lines of each block give a prediction of order (o) or disorder (d), along with a measure of the confidence in the prediction. The location of the YxxxIxxxW motif is shown boxed. The secondary structures coloured as in Figure 3 of the main text are shown above the Phyre predictions. For DnaD this is experimentally determined, whereas for DnaB the secondary structure is derived from our analysis of the Phyre predictions. The flanking H1'-S1'-H2' and H3' regions are also shown for DnaD. Note that crystallographically determined S1' is poorly predicted for DnaD. Also, in DnaB the prediction of the first strand of the beta-hairpin of the winged helix-turn-helix motif is equivocal, but note the good match of HMMs in this region, Figure 2B. This figure was created using an in-house script to convert the HTML content from Phyre to PostScript. The thin and thick tickmarks are at 10/50 residue intervals.

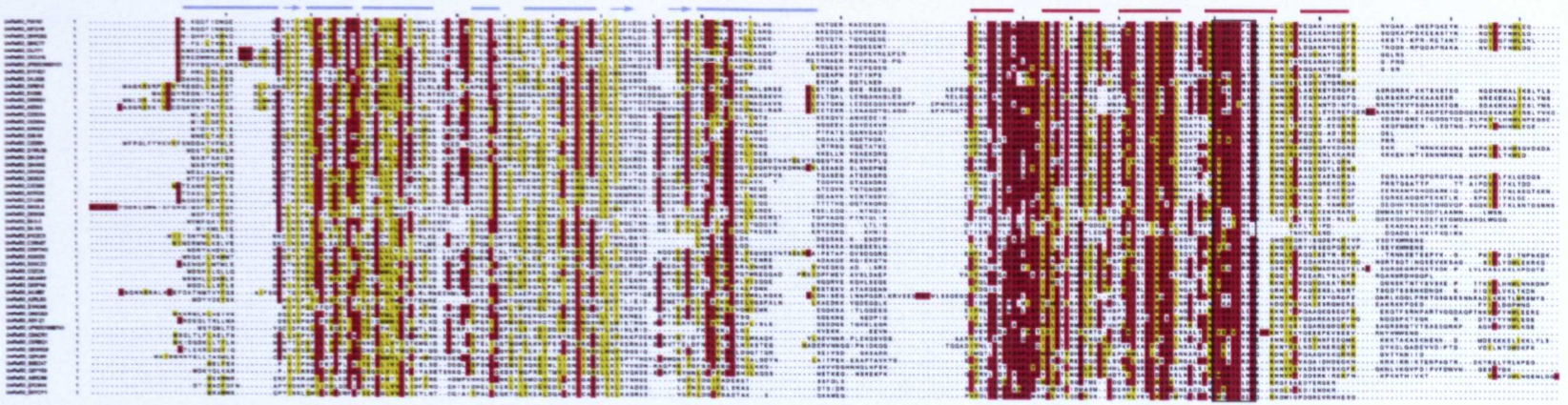
2zc2



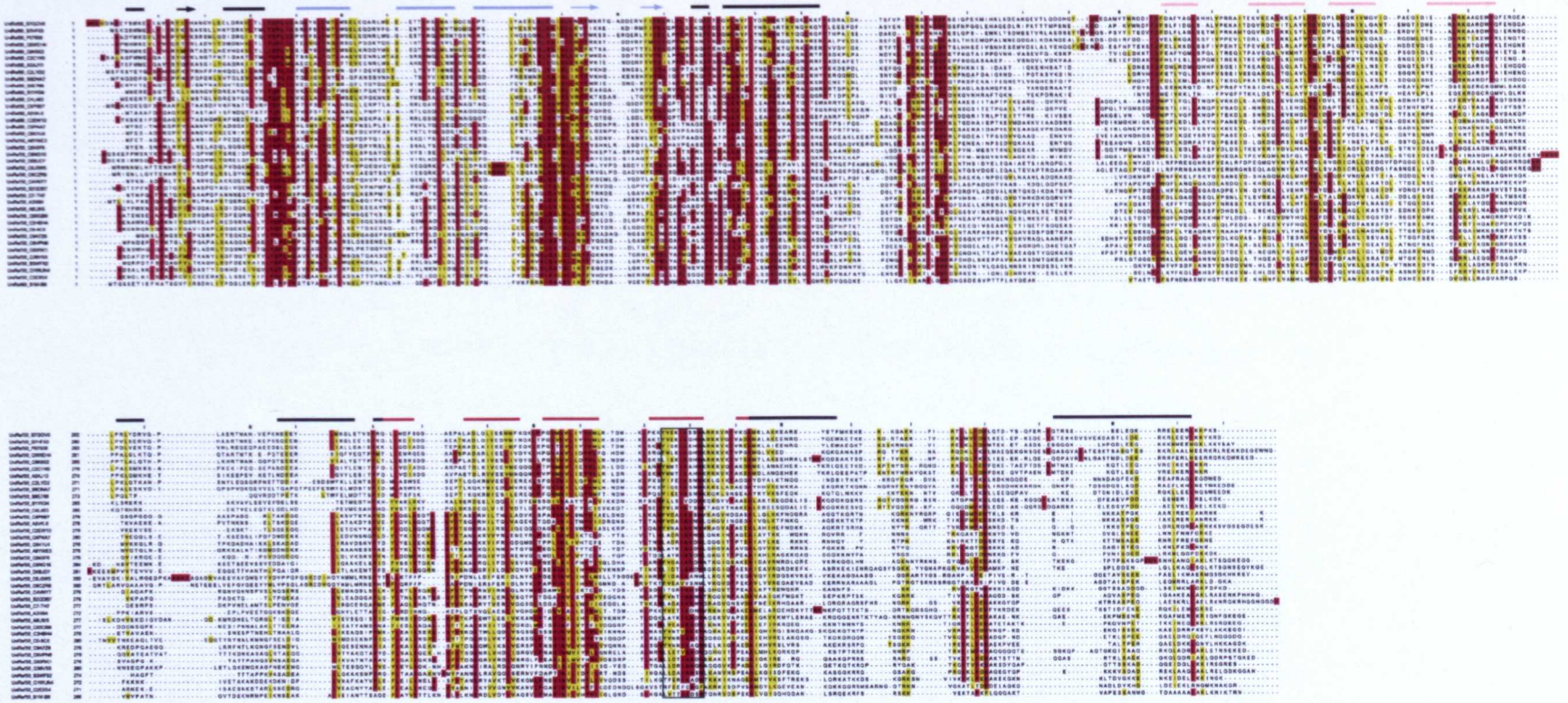
2i5u



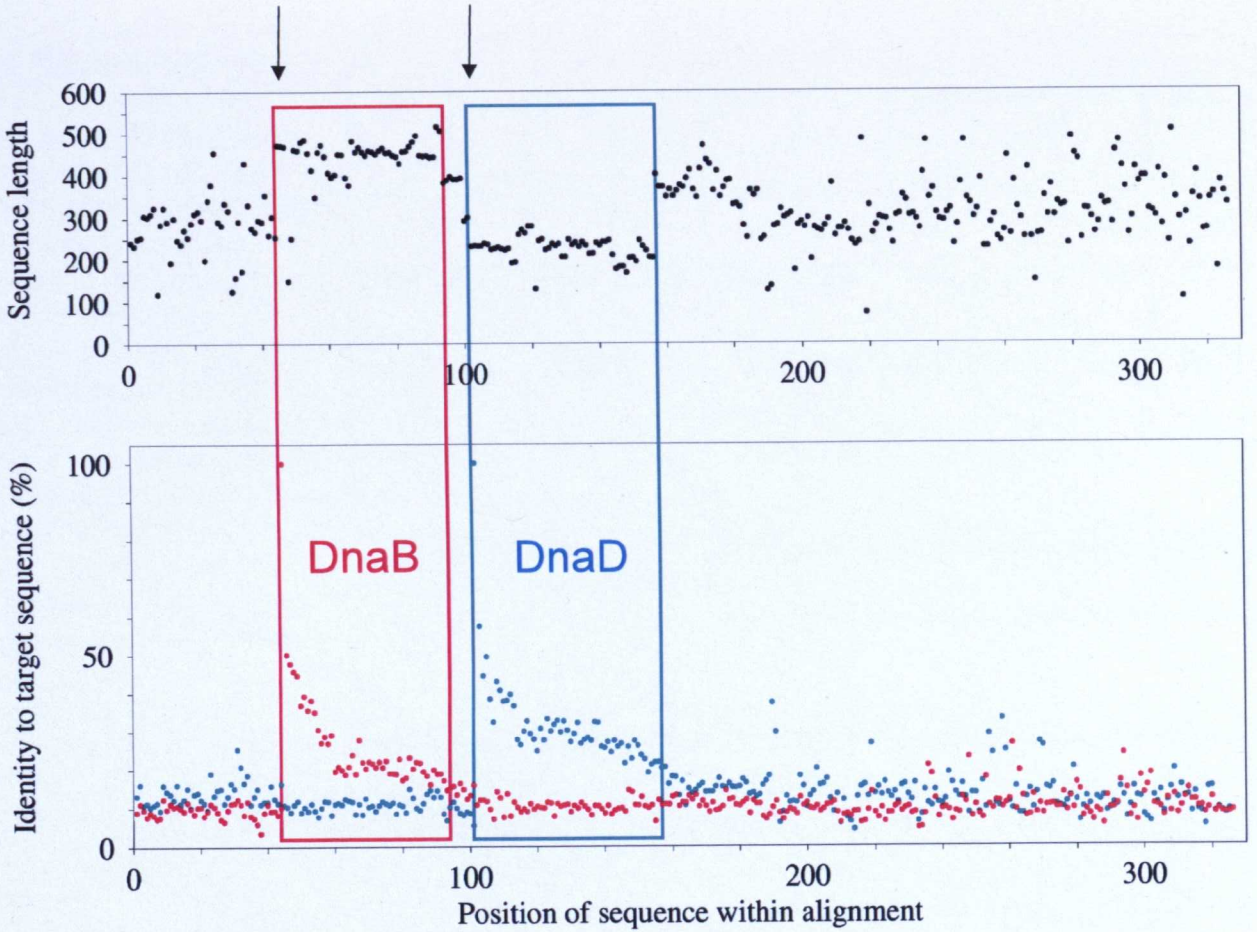
Supplementary Figure 2. Secondary structure and disorder predictions from Phyre for the full length proteins containing the domains in the 2zc2 and 2i5u structures. For details of the data contained in the Phyre predictions see legend to Supplementary Figure 1.



Supplementary Figure 3. Clustalw alignment of the sequences defined as DnaD from Supplementary Figure 5. The colour scheme of the alignment is as in Figure 1B. Secondary structure coloured according to homology to DnaD from Figure 3 is shown along the top of the alignment. The location of the YxxxIxxxW motif is shown boxed. *B. subtilis* DnaD is the first sequence in the alignment (P39787). The tickmarks denote numbering of the B. subtilis DnaD sequence (thin/thick ticks at 10/50 residue intervals).

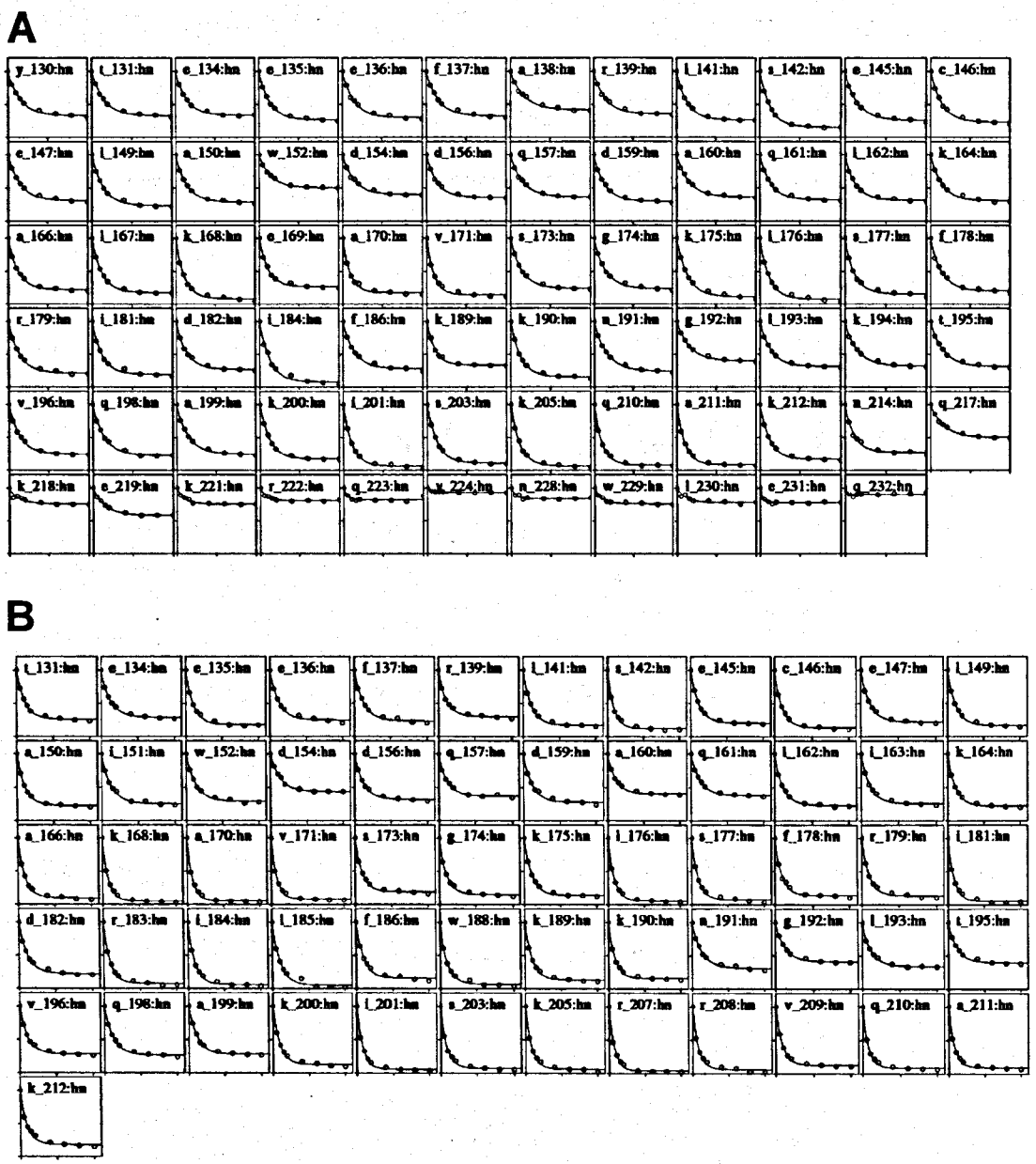


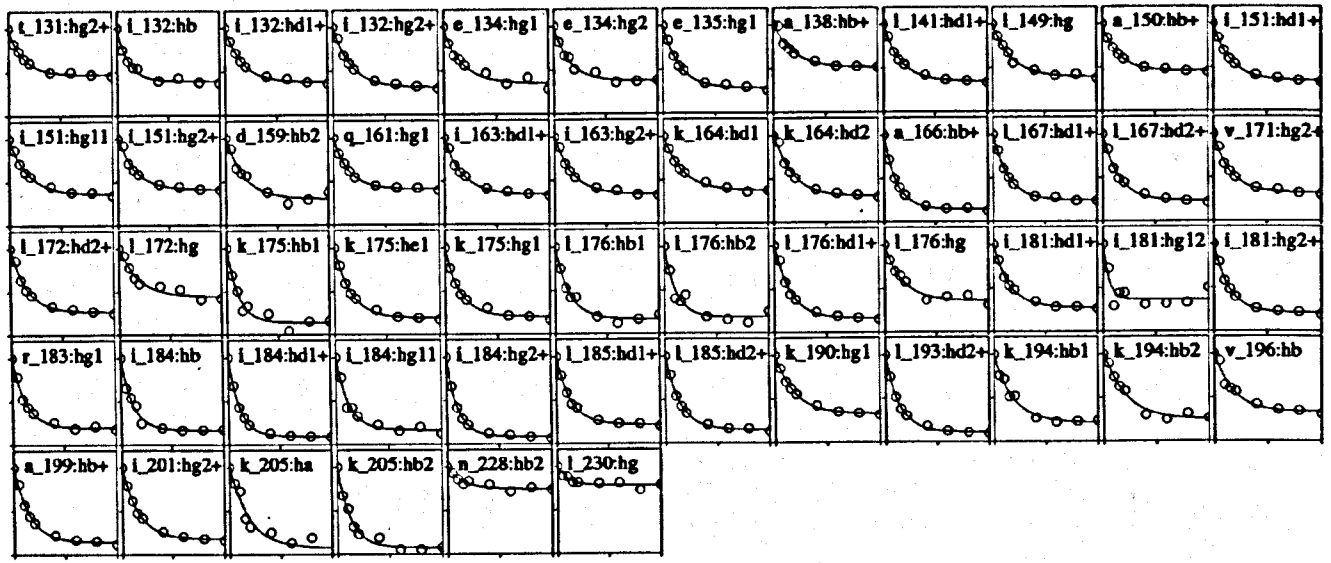
Supplementary Figure 4. Clustalw alignment of the sequences defined as DnaB from Supplementary Figure 5. The colour scheme of the alignment is as in Figure 1B. Secondary structure coloured according to homology to DnaD from Figure 3 is shown along the top of the alignment. The location of the YxxxIxxxW motif is shown boxed. *B. subtilis* DnaB is the third sequence in the alignment (P07908). The tickmarks denote numbering of the B. subtilis DnaB sequence (thin/thick ticks at 10/50 residue intervals).



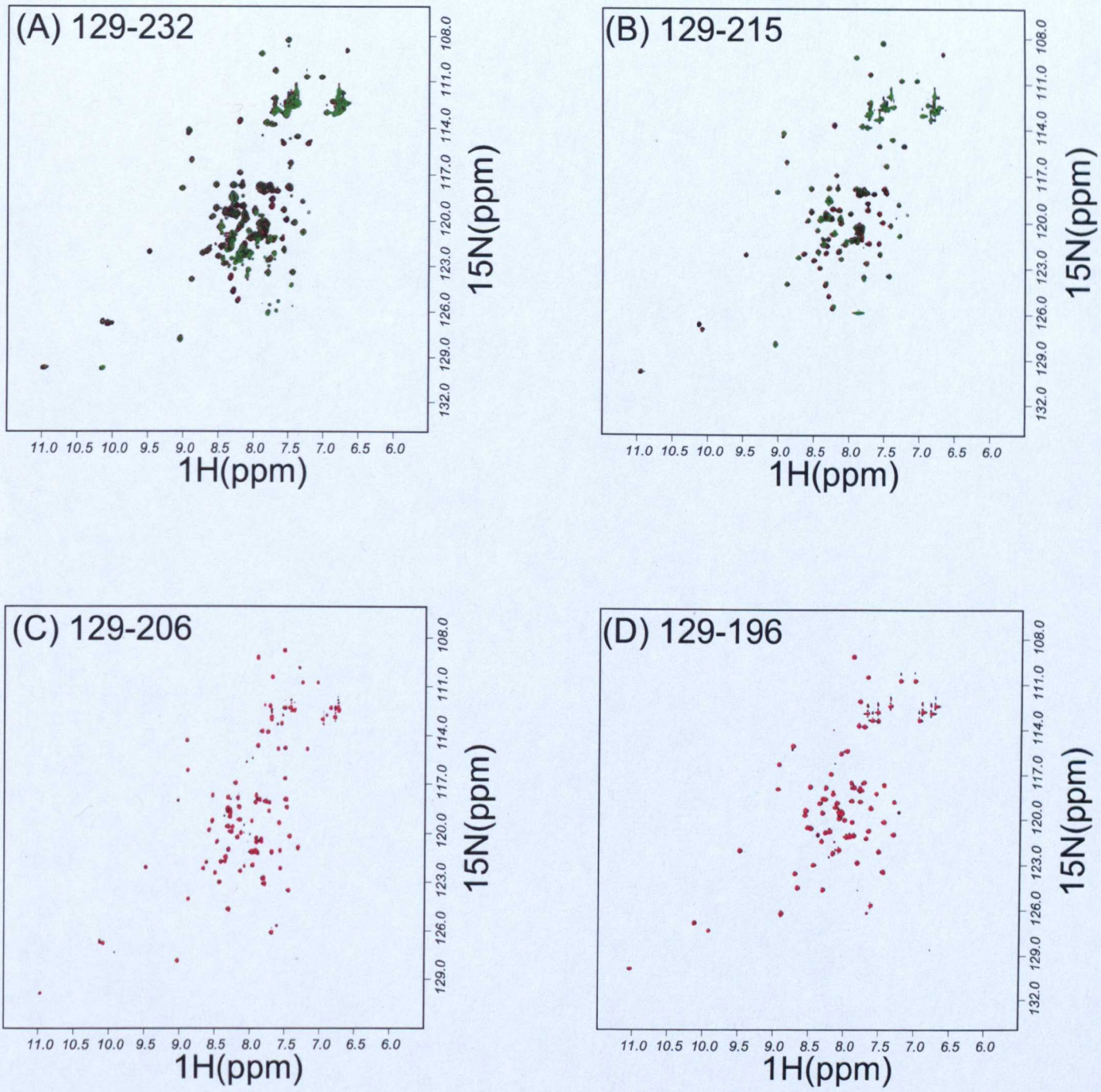
Supplementary Figure 5. Selection of putative DnaB or DnaD sequences. The two graphs are analyses of the clustalw alignment of the combined collection of 326 sequences as described in the main text. In both graphs the data is ordered according to position (row) within the alignment.The upper graph shows the lengths of the sequences. The lower graph shows % identity to B. subtilis DnaB (red) and B. subtilis DnaD (cyan). The arrows mark the positions of the *B. subtilis* proteins. These two proteins are first within their clusters simply because they were positioned first in the file of sequences to be aligned.

Supplementary Figure 6. NMR ^{15}N-HSQC intensity titration curves for titration of 10mer DNA into (**A**) DnaD-Cd232, and (**B**) DnaD-Cd215. The data are plotted as peak heights of the HSQC crosspeaks for each residue relative to its height at the start of the titration, versus fractional DNA concentration. In each graph, the x axis covers DNA concentrations from 0 to 1:1 stoichiometry, and the y axis covers ratios from 0 to 1.2. The curves are fits to a decaying exponential plus y-offset, which empirically was found to adequately interpolate the data. There is no apparent theoretical justification for this particular choice. We preferred this simple choice to any more elaborate models involving possibly multiple Kds, self association, and complex relaxation effects. The fitted y-offset was used to determine selection of residues in Figure 4C
The protein concentration in 0.3 mM in (**A**) and 0.4 mM in (**B**). Data are shown for all residues for which the HSQC crosspeak could be clearly resolved. The slightly larger intensity drops in the DnaD-Cd215 sample are probably attributable to the superior tumbling of the apo-protein, reflected in substantially improved spectra.





Supplementary Figure 7. ^{13}C HSQC intensity titration curves (for resolved and assignable crosspeaks) corresponding to the ^{15}N HSQC titration curves in Supplementary Figure 6A.



Supplementary Figure 8. ^{15}N HSQC spectra of *Bacillus subtilis* DnaD C-terminal domain constructs titrated with 10mer DNA. (A) DnaD residues 129-232 apo (black), +20% DNA (red), +40% (green). (B) DnaD residues 129-215 apo (black), +20% DNA (red), +40% (green). (C) DnaD residues 129-206 apo (black), +100% DNA (red). (D) DnaD residues 129-196 apo (black), +100% DNA (red).