Identification of Linker Regions and Domain Borders of the Transcription Activator Protein NtrC from *Escherichia coli* by Limited Proteolysis, *In-Gel* Digestion, and Mass Spectrometry†

Marcus Bantscheff,‡ Verena Weiss,§ and Michael O. Glocker*,†

Faculty of Chemistry, University Konstanz, P.O. Box M732, D-78457 Konstanz, Germany, and Faculty of Biology, University Konstanz, P.O. Box M732, D-78457 Konstanz, Germany

Received April 5, 1999; Revised Manuscript Received June 7, 1999

ABSTRACT: We have developed a mass spectrometry based method for the identification of linker regions and domain borders in multidomain proteins. This approach combines limited proteolysis and *in-gel* proteolytic digestions and was applied to the determination of linkers in the transcription factor NtrC from *Escherichia coli*. Limited proteolysis of NtrC with thermolysin and papain revealed that initial digestion yielded two major bands in SDS-PAGE that were identified by mass spectrometry as the R-domain and the still covalently linked OC-domains. Subsequent steps in limited proteolysis afforded further cleavage of the OC-fragment into the O- and the C-domain at accessible amino acid residues. Mass spectrometric identification of the tryptic/thermolytic peptides obtained after *in-gel* total proteolysis of the SDS-PAGE-separated domains determined the domain borders and showed that the protease accessible linker between R- and O-domain comprised amino acids Val-131 and Gln-132 within the "Q-linker" in agreement with papain and subtilisin digestion. The region between amino acid residues Thr-389 and Gln-396 marked the hitherto unknown linker sequence that connects the O- with the C-domain. High abundances of proline-, alanine-, serine-, and glutamic acid residues were found in this linker structure (PASE-linker) of related NtrC response regulator proteins. While R- and C-domains remained stable under the applied limited proteolysis conditions, the O-domain was further truncated yielding a core fragment that comprised the sequence from Ile-140 to Arg-320. ATPase activity was lost after separation of the R-domain from the OC-fragment. However, binding of OC- and C- fragments to specific DNA was observed by characteristic band-shifts in migration retardation assays, indicating intact tertiary structures of the C-domain. The outlined strategy proved to be highly efficient and afforded lead information of tertiary structural features necessary for protein design and engineering and for structure−function studies.

To study the structure and function of a multidomain protein it is often useful to separately express and analyze the individual domains. In addition, domain swap experiments help to elucidate the specific functions of different modules, affording lead information of functional and (tertiary) structural features necessary for protein design and engineering. These routes require precise knowledge about the exact boundaries of the domains. Protein domains may be estimated by carrying out sequence alignments using sequence database entries of homologous proteins and/or of the same protein from different organisms. An alternative method to determine the exact boundaries of the domains is to perform limited proteolyses of a protein of interest and subsequent identification of the fragments. This experimental approach reveals information about the protein structure in solution and is applicable to proteins with unknown homologies. Borders of domains and linkers in multidomain proteins can be determined when proteases with broad specificities are applied (*1*). Limited proteolyses of peptide bonds cleaving at susceptible surface areas have been used to determine the flexible interdomain linker of the soluble CytR regulator (*2*) and to characterize loop regions in membrane proteins (*3*) by radiolabeling or IR spectroscopy. The combination of limited proteolysis with mass spectrometric methods has been described recently for the study of protein tertiary structures of small proteins (*4*−*6*) by creating peptides with masses well below 10 kDa. Precise molecular mass determinations enabled the identification of cleavage sites in loops connecting secondary structure elements from the peptide mixtures or from HPLC-separated fragments (*7*). Similarly, the combination of limited proteolysis with mass spectrometry was applied for probing the partly folded states of proteins (*8*) and for determining protected sites in specific protein assemblies (*9*, *10*) and in protein−DNA complexes (*11*).

The transcriptional activator protein NtrC is required for the regulation of genes necessary for the assimilation of ammonium and utilization of alternative nitrogen sources in enteric bacteria (*12*, *13*) and belongs, as a response regulator protein, to the group of "two-component signal transduction" systems (*14*, *15*). NtrC from *Escherichia coli*, a 52 kDa

* Corresponding author: Michael O. Glocker, Faculty of Chemistry, University Konstanz, P.O. Box M732, D-78457 Konstanz, Germany. Telephone: +49-7531-882690. FAX: +49-7531-883097. E-mail: Michael.Glocker@uni-konstanz.de.

† This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, Bonn, Germany) and by a grant of the DFG to H. Bujard and V. Weiss.

[1] Abreviations and symbols: ESI, electrospray ionization; HCCA, 4-hydroxy-α-cinnamic acid; MALDI, matrix assisted laser desorption/ionization; E:S, enzyme-to-substrate ratio; TFA, trifluoroacetic acid; w/w, weight-to-weight ratio; w/v, weight-to-volume ratio; *m*/*z*, mass-to-charge ratio; IPTG, isopropyl-*β*-D-thiogalactopyranosid; MWCO, molecular weight cutoff; TPCK, L-1-chlor-3-phenyl-3-(p-toluolsulfonamido)-2-butanone; FIS, factor for inversion stimulation; NifA, Nitrogen fixation protein A.

10.1021/bi990781k CCC: $18.00
Published on Web 08/03/1999

multidomain protein that contains 469 amino acid residues, was chosen to test the outlined mass spectrometric method for determining the exact boundaries of its domains. It consists of (i) an N-terminal receiver domain (R) that becomes phosphorylated (*16*, *17*), (ii) a central output domain (O) that bears ATPase activity (*18*, *19*), and (iii) a C-terminal DNA-binding domain (C) that is also required for dimerization (*20*, *21*); for review of the σ^{54} bacterial enhancer binding protein family, see refs. *22* and *23*. The particular activity properties assigned to the respective domains were determined studying individually expressed NtrC fragments of defined lengths and sizes. However, no particular attention had apparently been given to determine whether the active fragments resembled the structural domain units or not. Information about tertiary structures of NtrC domains is limited and mostly based on predictions. The C-terminal domain of NtrC has been proposed to adopt a fold similar to that of the FIS dimer based on sequence alignments and estimation of homology boxes (*21*). The architecture of the central output domain has been suggested to resemble that of classical folds found in mononucleotide-binding proteins by secondary structure prediction and fold recognition algorithms (*24*). The structure of the N-terminal domain of NtrC has been derived from structural data obtained by multidimensional NMR spectroscopy of the individually expressed receiver domain $NtrC_R$ (*25*). Information about the interdomain linker regions is also sparce. The linker sequence connecting the R- and O-domains has been assigned upon exploration of NtrC sequences from database entries using sequence pattern recognition algorithms (*26*). The high abundance of glutamine residues gave the name "Q-linker" to these interdomain partial sequences that encompass approximately 15 to 20 residues in NtrC proteins also from other organisms (*26*). To date, no information is available demonstrating that NtrC-linkers are surface exposed and/or flexible structures. Further, exact boundaries of NtrC domains and linker regions have not been determined experimentally.

In this study, we applied limited proteolyses with thermolysin, papain, and subtilisin followed by micropreparative SDS-PAGE of the separated domains and *in-gel* digestion with trypsin (*27*−*30*) in order to identify domains (tertiary structural units) and flexible linker regions of NtrC. Our results described herein define the flexible part of the N-terminal linker region (Q-linker) and further enabled to determine the hitherto unknown C-terminal linker region that connects the O- and C-domains in NtrC. Thus, the outlined strategy proved to be highly efficient and afforded lead information of tertiary structural features necessary for protein design and engineering and for structure−function studies.

MATERIALS AND METHODS

Materials. Trypsin, thermolysin, papain, and subtilisin were purchased from Sigma (Sigma, Deisenhofen, Germany). Solvents were HPLC-grade (Merck KG, Darmstadt, Germany).

NtrC Expression and Purification. NtrC was purified from extracts of a 9-L culture of *E. coli* FI8208 (*31*) carrying plasmid pFI20 grown at 37 °C and induced with 1 mM IPTG. PFI20 is a pB*R*322 derivative that carries *ntrC* under the control of the *ptac* promoter and *lacI*^{q} (*32*). Purification was performed as described previously (*33*, *34*), with the exception that after the heparin-Sepharose column a DEAE-Sepharose column (CL-6B Sepharose, Pharmacia Biotech, Freiburg, Germany) in Tris[HCl] buffer with a linear NaCl gradient was applied yielding in highly purified NtrC (34 mg total). The concentration of NtrC was estimated from the absorbance at 280 nm, using $A^{1\%} = 9.1$ as calculated from the amino acid composition (*35*).

Limited Proteolysis of NtrC. Limited proteolysis of NtrC with thermolysin and subtilisin, respectively, was conducted for 2 h at 37 °C. NtrC was dissolved in 50 mM Tris-HCl, pH 7.5, that contained 50 mM KCl, 5 mM $CaCl_2$, and 20% glycerol. Papain digestions were performed for 2 h at 37 °C. NtrC was dissolved in 40 mM Tris-HCl, pH 7, that contained 40 mM KCl, 4 mM $MgCl_2$, 0.8 mM EDTA, 2 mM DTT, and 20% glycerol. Enzyme/substrate ratios were varied from 1:6912 to 1:54 (*w*/*w*). A time-course of digestion ranging from 10 min to 2 h was performed using the enzyme/substrate ratio of 1:864 (w/w). In all experiments, final NtrC concentration was 0.68 mg/mL. Sample solutions were dialyzed against 2% acetic acid/methanol (9:1, v/v) using a cellulose membrane (MWCO 3500; Carl Roth GmbH & Co, Karlsruhe, Germany) prior to mass spectrometric analyses.

Gel Electrophoresis of NtrC Fragments. SDS-polyacrylamide gel electrophoresis was performed according to previous descriptions (*36*, *37*) using a BioRad Mini-Protean II system (Biorad, Munich, Germany). Staining was carried out with 0.02% Coomassie Brilliant Blue G-250 in 10% acetic acid.

In-Gel Digestion of NtrC Fragments. Tryptic digestions in the SDS gel matrix were carried out according to a previously described procedure (*28*, *30*) with the following modifications. Staining and destaining of the gel were restricted to 30 min in order to minimize fixing of the protein in the gel. Protein containing parts were excised and the gel pieces were washed for 20 min in 500 μL 60% acetonitrile under rigid shaking. After the supernatant was removed, the gel pieces were dried in a vacuum centrifuge and subsequently incubated in 50 μL 50 mM NH_4HCO_3, pH 8, for 20 min. This procedure was repeated twice. The gel pieces were then incubated for 20 min in 30−50 μL 50 mM NH_4HCO_3, pH 8, that contained 12.5 ng/μL TPCK treated trypsin. To remove excess trypsin, the supernatant was discarded and subsequently 50 μL of 50 mM NH_4HCO_3, pH 8, was added. Proteolytic digestion was continued for 12 h at 37 °C under gentle shaking. Peptides were eluted from the gel pieces by adding 200 μL of 60% acetonitrile and rigid shaking for 2 h. The supernatant was lyophilized and redissolved in CH_3-CN/0.1% TFA (2:1, v/v) prior to MALDI-MS analyses.

MALDI-MS Instrumentation and Acquisition Conditions. MALDI-MS analyses were carried out with a Bruker Biflex linear time-of-flight spectrometer (Bruker-Franzen, Bremen, Germany) equipped with a SCOUT source, video system, UV nitrogen laser (337 nm), and a dual microchannel plate detector. The acceleration voltage was set to 20−25 kV. Mass calibration was carried out using hen eggwhite lysozyme (HEL) and insulin as external and internal standards, respectively. Spectra were processed by means of the X-MASS data system. A 1 μL sample of the solution was mixed with 0.8 μL of a saturated solution of α-cyano-4-hydroxy cinnamic acid (HCCA) dissolved in CH_3CN/0.1% TFA (2:1, v/v) directly on the target. After evaporation of

the solvent, the matrix/analyte mixture was recrystallized from 1.5 μL CH_3CN/0.1% TFA (2:1, v/v) prior to data acquisition.

Nano-ESI-MS Instrumentation and Acquisition Conditions. ESI-MS spectra were recorded on a Vestec A 201 single quadrupole mass spectrometer (PerSeptive Biosystems, Framingham, MA), fitted with a self-built nanoESI source; gold-coated borosilicate capillaries were prepared as reported elsewhere (*38*). The voltage at the capillary tip was adjusted to 1.1−1.3 kV with a declustering potential typically set to 20 V. A volume of 0.5−2 μL was loaded by dipping the capillary into the sample solution.

Polyacrylamide Gel Electrophoresis Migration Retardation Assays. The "Lp binding site" (*39*) is a short oligonucleotide that was synthesized using standard solid-phase synthesis protocols (MWG-Biotech GmbH; Ebersberg, Germany). The sequences of the complementary oligonucleotides are 5′−CGAT***TGCACTAAAATGGTGCA***ATCGATTCACATCGAGGACGTGGACGTAT−3′ and 5′−CGATACGTCCACGTCCTCGATGTGAATCGAT***TGCACCATTTTAGTGCA***AT−3′ (The NtrC binding site is printed in bold-italic letters). The oligonucleotides were annealed by heating the complementary strands at 100 μM concentration in TE-buffer (10 mM Tris-HCl pH 8, 1 mM EDTA) for 15 min at 80 °C and cooling slowly to 4 °C. The protein−DNA complexes were formed by incubating 20 μL of a solution containing 800 nM Lp binding site and 2 μM NtrC or 2 μL limited digestion mixture (E:S = 1:864) in TBE-buffer (19.8 mM Tris, 19.8 mM boric acid, 0.2 mM EDTA, pH 8.4) at room temperature for 20 min. After addition of bromphenolblue (5 μL, 0.25% w/v) and sucrose (5 μL, 30% w/v), native TBE-gel-electrophoresis was carried out according to descriptions (*40*) using 8% polyacrylamide gels. Staining was performed with ethidiumbromide.

ATPase Assay. ATPase activity was assayed as described (*41*). The assay mixture contained the following components in a total volume of 800 μL: 50 mM Tris, pH 7.5, 150 mM KCl, 5 mM $MgCl_2$, 0.1 mM EDTA, 30 mM phosphoamidate, 0.5 mM phosphoenolpyruvate, 0.2 mM NADH, 8 units pyruvate kinase, 20 units lactate dehydrogenase, 0.1 μg/μL BSA, 250 nM NtrC. The assay was performed at ambient temperature using a 1 mL cuvette with a Shimadzu UV1202 UV-vis spectrometer (Shimadzu Corp. Kyoto, Japan). The reaction was initiated by addition of 8 μL ATP (100 mM). The solution was mixed in a cuvette for 10−15 s, and the rate of decrease in absorbance at 340 nm was monitored in 10 s intervals for 40 min.

RESULTS

Limited Proteolysis of NtrC, SDS-PAGE Separation, and Mass Spectrometric Molecular Weight Determination of Domains. Limited proteolyses of NtrC from *E. coli* performed for 2 h with small amounts of thermolysin (E:S approximately 1:7000 to 1:1500; w/w) yielded distinct bands by SDS-PAGE that migrated at 14 kDa, 37 kDa, and 52 kDa, respectively (Figure 1). From the apparent masses, these bands were tentatively assigned as the N-terminal receiver domain (R), the still covalently linked output and C-terminal domains (OC), and undigested NtrC, respectively. With larger amounts of thermolysin (E:S approximately 1:900 to 1:100; w/w), the 52 kDa band disappeared completely and

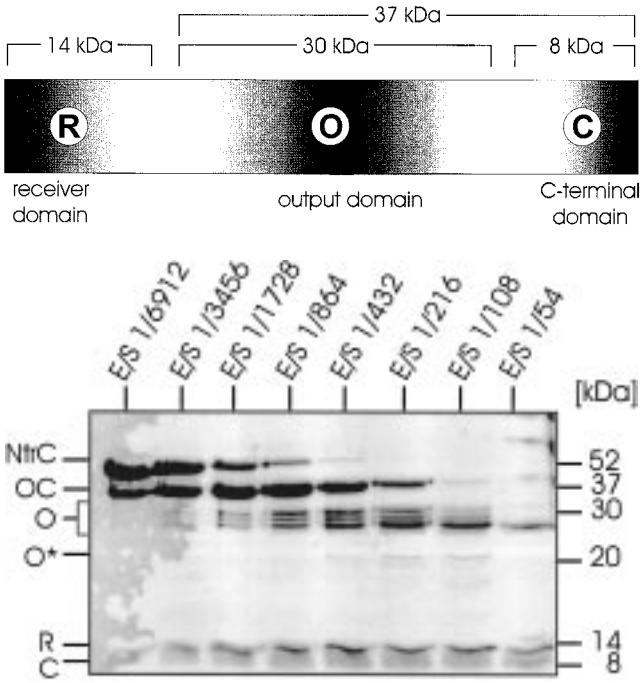

FIGURE 1: SDS−PAGE analysis of limited thermolytic digestion of NtrC. NtrC was digested with increasing enzyme-to-substrate ratio (from left to right). Fragments were separated in individual bands and tentatively assigned to corresponding domains according to their apparent masses. NtrC, intact protein; OC, Output-C-terminal domain; O, region of output domain fragments; O*, core fragment of the output domain; R, receiver domain; C, C-terminal domain. The apparent molecular masses are given at the right. A schematic representation of NtrC, indicating the fragments obtained by limited proteolyses and corresponding molecular masses, is given on top.

an additional band migrating at 8 kDa (C-domain, C) appeared together with a triplet of bands at approximately 30 kDa (Output-domain, O). When the NtrC to thermolysin ratio was adjusted to ca. 1:50 (w/w), bands were visible corresponding to the R-, C-, and O-domains together with a barely detectable band at 20 kDa that may represent a thermolysin-resistant core fragment of the O-domain (O*). These results showed that the O-domain was further truncated, while R- and C-domains remained stable under the applied limited proteolysis conditions. Hence, we conclude that thermolysin cleaved NtrC initially at susceptible bonds possibly located in the linker regions, leaving the tightly folded domains mostly undigested. Precise molecular masses of the thermolytic fragments were determined by MALDI-MS and ESI-MS (Table 1) from the fragment mixtures enabling unambiguous domain assignments. The MALDI mass spectra of the fragment mixtures showed the presence of singly and multiply charged ions for NtrC and fragments in the mass range between 5000 and 60 000 Da with compositions comparable to those observed by SDS−PAGE. As an example, the spectrum obtained after 30 min with E:S of 1:864 is shown (Figure 2). In addition, the spectra revealed that the 14 kDa band consisted of at least two products. The minor product showed an ion signal at *m*/*z* 14 392 corresponding to the N-terminal receiver domain (aa 1−130; M_r = 14 390.5). The major product with an ion signal at *m*/*z* 14 616 represented the N-terminal receiver domain carrying two more amino acid residues at its C-terminus (aa 1−132; M_r = 14 618). Similarly, at *m*/*z* 8332, an ion signal was

Table 1: Mass Spectrometric Molecular Weight Determination of NtrC Domains Obtained by Limited Proteolysis

protease	fragment	molecular mass: SDS−PAGE (kDa)	MALDI-MS (Da)	calculated (Da)	sequence range
thermolysin[a]	R	14	14 616[d]	14 617.8	M1−Q132
			14 493[d]	14 490.6	M1−V131
			14 392[d]	14 390.5	M1−N130
	OC	37	37 655[d]	37 655.1	L133−E469
	O*	20	20 246[d]	20 248.3	I140−R320
			20 949	20 947.0	L133−R320
	C	8	9035[d]	9034.2	V390−E469
			8332	8331.5	M397−E469
papain[b]	R	14	14 617	14 617.8	M1−Q132
			14 392	14 390.5	M1-N130
	C	8	8864	8864.0	E392−E469
			8735	8734.9	S393−E469
			8544	8546.7	S395−E469
			8460	8459.6	Q396−E469
			8332	8331.5	M397−E469
subtilisin[c]	RO	44	43 734[d,e]	43 726	M1−T394
	C	8	8864	8864.0	E392−E469
			8736[d]	8734.9	S393−E469
			8547[d]	8546.7	S395−E469
			8333[d]	8331.5	M397−E469

[a] E:S = 1:864. [b] E:S = 1:1080. [c] E:S = 1:4320. [d] Confirmed by ESI-MS. [e] Average mass of major compound in the fragment mixture.

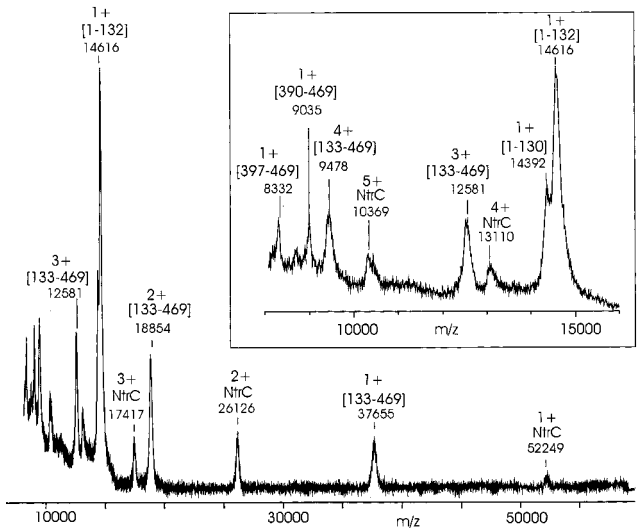

FIGURE 2: MALDI-MS analysis of limited thermolytic digestion of NtrC. Singly and multiply charged ion signals are present for still intact NtrC and for thermolytic fragments. Numbers in parentheses refer to amino acid positions in intact NtrC. E:S = 1:864. HCCA was used as matrix.

observed for the C-terminal domain (aa 397−469; M_r = 8332) and a signal at *m*/*z* 9035 was indicative for the C-terminus, including additional amino acid residues at its N-terminus (aa 390−469; M_r = 9034). The sample also showed a singly charged ion signal for NtrC at *m*/*z* 52 249 (M_r = 52 254) and for the OC-fragment at *m*/*z* 37 655 (OC; aa 133−469; M_r = 37 655) together with multiply charged ion signals.

Limited proteolysis of NtrC with papain again revealed digestion sites at Asn-130 and Gln-132, yielding N-terminal products with molecular masses of 14 617 and 14 392, respectively. In addition, multiple C-terminal products with molecular masses between 8332 and 8864 indicated cleavages at residues Glu-392 up to Met-397 (Table 1). Cleavages in this area were also determined mass spectrometrically when subtilisin was used for limited proteolysis experiments, showing that this region was highly susceptible, indicating a linker region. Subtilisin did not cleave between the R- and the O-domain when added in small amounts (E:S = 1:4 000), consistent with the described substrate specificity of this protease (*8*).

Table 2: MALDI-MS Identification of Tryptic Peptides from *In-Gel* Digested NtrC Domains[a]

sequence range	domain	$[M+H]^+$ calcd	$[M+H]^+$ exp. of fragments: NtrC	R	OC	O	O*	C
1−46	R	4980.7		4976				
4−16	R	1461.6	1461	1461				
17−21	R	702.8	702	702				
47−56	R	1129.3	1130	1129				
57−67	R	1146.5	1146	1146				
71−81[b]	R	1335.7		1335				
71−117	R	5242.0	5244	5242				
95−117[b]	R	2622.9		2622				
118−129	R	1485.6	1485	1485				
118−130[d]	R	**1599.7**		**1600**				
130−174[c]		4714.4	4713					
133−152[d]	O	2146.4			2147	2147		
160−189	O	3153.6	3154			3156	3157	
203−224	O	2406.7	2405		2406	2407	2407	
225−251	O	3010.3	3009		3010	3010	3011	
255−263	O	1069.2	1069		1069	1070	1069	
264−275	O	1260.5	1261		1260	1261	1261	
276−287	O	1394.5	1395		1394	1395	1395	
293−300	O	1120.3	1120		1119	1120	1120	
301−314	O	1713.1	1715		1714	1713	1714	
324−331	O	942.1	942		942	942		
339−350	O	1351.5	1352		1351	1353		
351−358	O	913.1	912		913	913		
390−412[d]		**2592.8**						**2593**
397−412[d]	C	**1890.1**						**1890**
416−431	C	1808.9	1808		1808			1810
432−445	C	1577.8	1580		1580			1580
440−450	C	1263.4	1263					1263
451−456	C	701.8	701		701			702
451−462	C	1415.7	1416		1416			

[a] After SDS-PAGE separation of thermolytic NtrC fragments, selected peptides are listed. [b] Derived from nonspecific activity of trypsin, confirmed by gas-phase Edman sequencing. [c] Peptides contain sequence parts of linker and corresponding domain. [d] Numbers printed in bold represent peptides produced by thermolytic and tryptic cleavage.

Identification of Domain Borders and Linker Sequences by In-Gel Digestion and MALDI-MS Peptide Mapping. Precise mass spectrometric molecular weight determinations can be used to assign small domains derived from limited proteolyses for which either the N- or the C-terminus is known as is the case for the C- and R-domain in NtrC, respectively. However, a given nominal mass in the range of e.g. 15 kDa and above, can match with numerous partial sequences from the central part of a protein. Thus, precise sequence assignment of large fragments from the interior part of a protein as obtained by limited proteolysis with proteases of broad substrate specificity solely based on experimentally determined masses is rendered impossible. To identify the domain borders and linker sequences, we subsequently subjected the SDS−PAGE-separated thermolytic fragments to *in-gel* total proteolyses with trypsin followed by mass spectrometric peptide mapping. The protein migrating as a 52 kDa band yielded peptides that could be assigned to the entire NtrC sequence (Table 2). By contrast, tryptic peptides from the 14 kDa fragment were exclusively derived from the R-domain and the fragment migrating as a

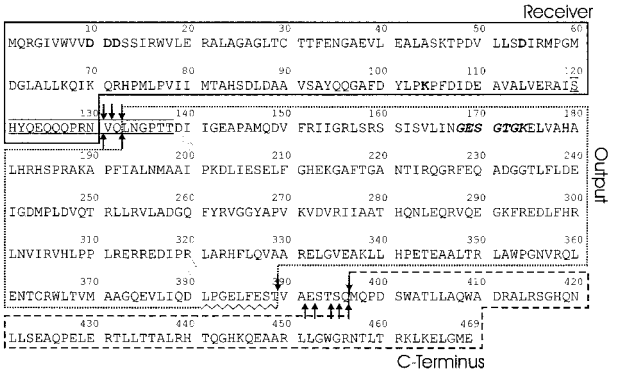

FIGURE 3: NtrC sequence and cleavage sites from limited proteolyses. The domains are boxed. Receiver, strait line; Output domain, dotted line; C-terminus, hatched line. Arrows depict cleavge sites of limited proteolysis with thermolysin (↓) and papain (↑), respectively. Amino acid residues that are important for phosphorylation of the receiver domain are printed in bold. The ATP-binding motif in the O-domain is printed in italic letters. The underlined sequence stretch represents the predicted Q-linker region (*26*). The zigzag line indicates nondetermined borders. Start and end of the core fragment (O*) of the O-domain are denoted.

37 kDa band produced peptides that derived from the O- and the C-domains only. The 8 kDa fragment afforded tryptic peptides only from the C-domain. Finally, the 30 kDa and 20 kDa fragments afforded peptides from the O-domain, substantiating that the fragments migrating at 20 kDa band represented a core fragment (O*) of the O-domain. All three bands migrating as a triplet at 30 kDa showed the same N-terminal peptide ion signals but differed at the C-termini (data not shown), representing C-terminal truncations of the O-domain. Lys-C digestion of the HPLC-separated O*-fragment afforded an N-terminal peptide ion signal at *m*/*z* 5377 (aa 140−189; M_r = 5378.2) and an ion signal at *m*/*z* 3580 for the C-terminal peptide aa 293−320 (M_r: 3581.2) in agreement with mass spectrometric peptide mapping of tryptic peptides and MALDI-MS molecular mass determination (cf. Table 1).

The borders of the domains and linkers were determined from peptides that were cleaved by thermolysin at one end and by trypsin at the other end. Double-digested peptides from the linker region between the R- and the OC-domain afforded ion signals at *m*/*z* 1600 (aa 118−130; M_r = 1599.7) and at *m*/*z* 2147 (aa 133−152; M_r = 2146.4), respectively. Thus, the protease susceptible region comprised amino acids Val-131 and Gln-132. The same cleavage sites in this linker region were identified mass spectrometrically using papain as protease (data not shown). Protease accessible residues in this region were found within the Q-linker sequence only, in agreement with tryptic digestion of NtrC that showed a highly susceptible cleavage site at Arg-129 (*42*). It is noteworthy to mention that the predicted Q-linker region in NtrC from *E. coli* has been proposed to comprise amino acid residues 120−138 (underlined in Figure 3) (*26*). This is a far longer stretch than that determined experimentally to be accessible to limited proteolysis.

Additional cleavages obtained by limited proteolysis of NtrC with thermolysin split the OC-fragment into the O- and the C-domain at accessible amino acid residues ranging from Thr-389 to Gln-396 (Table 2). Double-digested peptides were obtained by subsequent *in-gel* tryptic digestion and resulted in an ion signal at *m*/*z* 1890 (aa 397−412; M_r = 1890.1)

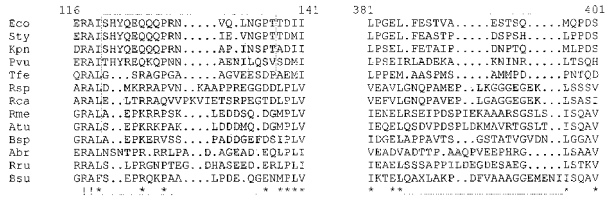

FIGURE 4: Sequence alignments of "Q-linkers" and "PASE-linkers" from thirteen related NtrC proteins. The entire sequences of the NtrC proteins were aligned using the ClustalW WWW service at the European Bioinformatics Institute (*59*). Shaded boxes contain linker sequences. Proteobacteria groups, organisms, and database entry numbers (in parentheses) are given. α: Rsp, *Rhodobacter sphaeroides* (X78533); Rca, *Rhodobacter capsulatus* (X72382); Rme, *Rhizobium meliloti* (M15810); Atu, *Agrobacterium tumefaciens* (J03678); Bsp, *Bradyrhizobium sp.* (M14227); Abr, *Azospirillum brasilense* (Z37984); Rru, *Rhodospirillum rubrum* (U30377); Bsu, *Brucella suis* (Y13236). *β*: Tfe, *Thiobacillus ferrooxidans* (L18975). *γ*: Sty, *Salmonella typhimurium* (X85104); Kpn, *Klebsiella pneumoniae* (X02617); Pvu, *Proteus vulgaris* (X68129); Eco, *Escherichia coli*.

that corresponded to the N-terminal peptide of the C-domain and a signal at *m*/*z* 2593 (aa 390−412; M_r = 2592.8) that was derived from the C-domain, including linker residues. Limited proteolyses with papain and subtilisin afforded cleavages in the same region (Figure 3), thus identifying the susceptible linker region that connected the O- with the C-domain.

Determination of Amino Acid Compositions in Linkers that Connect O-Domains with C-Domains in NtrC Response Regulator Proteins. Having experimentally identified the linker sequence that connected the O-domains with the C-domain in NtrC from *E. coli*, we aligned the related amino acid sequences of thirteen NtrC proteins from different proteobacteria that were available from the SWISS PROT database. The linker regions were found to resemble sequences in which abundance of conserved amino acid residues was low and in which significant sequence gaps appeared (Figure 4). Sequence homologies within the linker regions were evident within proteobacteria groups α, *β*, and *γ*, respectively, but little or no sequence similarities were found between the groups. The investigated linkers spanned sequence lengths from 11 to 20 amino acid residues. Proline, alanine, glutamic acid, and serine residues were found to be present in high abundance by comparing the amino acid residue compositions of these sequence elements with that of the entire NtrC proteins and with unrelated proteins (Table 3). Proline and alanine residues are found in abundance in many linkers (*43*), substantiating that the investigated regions represented linker sequences. The high abundance of glutamic acid together with the very low abundance of arginine residues may explain the estimated acidic character of these linkers. Further, the abundance of serine residues in the linker sequences was found to be particularly high (2.5-fold higher abundance than average when compared to NtrC proteins and 1.6-fold higher abundance than average when compared to unrelated proteins; Table 3) when compared with NtrC proteins. Hence, we use the name "PASE-linker" for the partial structures that connect the O-domains with the C-domains in NtrC proteins, in analogy with the Q-linker (*26*) that connects the R-domains with the O-domain.

Structure−Function Correlation of Fragments Obtained by Limited Proteolysis. Densitometric intensities of thermolytic fragments of NtrC were obtained from SDS−PAGE-

Table 3: Amino Acid Preferences for Residues Contained in Linker Oligopeptides Connecting O- and C-Domains in NtrC Proteins

	composition (%)			ratio (PASE-linkers/proteins)	
amino acid	unrelated proteins[a]	NtrC proteins	PASE-linkers	unrelated proteins	NtrC proteins
P	5.1	5.4	11.0	2.2	2.0
A	8.3	9.9	14.0	1.7	1.4
S	6.9	4.4	11.0	1.6	2.5
E	6.2	7.1	10.0	1.6	1.4
G	7.2	7.1	9.5	1.3	1.3
M	2.4	2.8	3.0	1.3	1.1
D	5.3	6.2	6.5	1.2	1.0
T	5.8	5.5	6.0	1.0	1.1
Q	4.0	4.3	3.5	0.9	0.8
N	4.4	3.1	3.5	0.8	1.1
K	5.7	3.3	4.0	0.7	1.2
V	6.6	6.5	4.5	0.7	0.7
I	5.2	5.3	3.0	0.6	0.6
F	3.9	2.7	2.0	0.5	0.7
H	2.2	2.0	1.0	0.5	0.5
R	5.7	9.1	3.0	0.5	0.3
L	*b*	12.1	4.0	*b*	0.3
Y	3.2	1.9	0.0	0.0	0.0
C	1.7	0.3	0.0	0.0	0.0
W	1.3	1.0	0.0	0.0	0.0

[a] Frequency of occurrence of amino acid residue in sequences of 1021 unrelated proteins (*58*). [b] No data available.

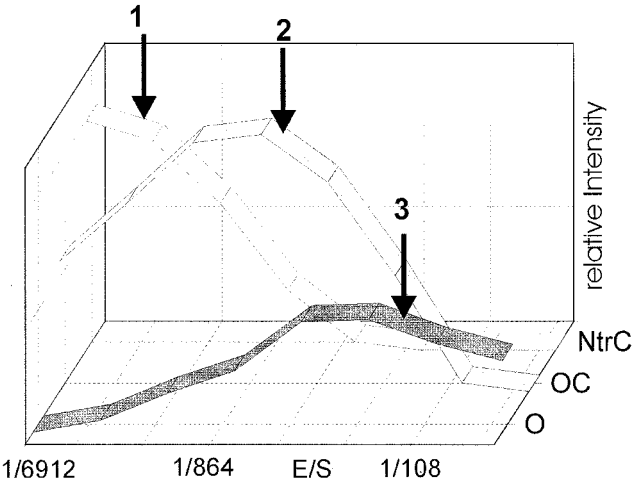

FIGURE 5: Densitometric intensities of thermolytic fragments of NtrC as functions of enzyme-to-substrate ratios. NtrC fragments obtained by limited proteolysis were separated by SDS−PAGE (cf. Figure 1). White, NtrC; gray, OC-fragment; black, O-fragments. Vertical arrows indicate primary (1), secondary (2), and tertiary (3) cleavages. Primary and secondary clevage sites occurred in linker sequences.

separated bands (cf. Figure 1) and were plotted as functions of enzyme-to-substrate ratios (Figure 5). The results showed that the densitometric intensity of the OC-band increased at the expense of that corresponding to intact NtrC. This can be explained by initial cleavages (primary cleavages, 1) of thermolysin cutting NtrC into the R- and OC-fragments at susceptible residues within the Q-linker. Subsequent cleavages caused a decrease in the OC-band intensity and a simultaneous increase of the intensity for the O-domain band. This is in agreement with the proteolytic separation of the O-domain from the C-domain (secondary cleavages, 2) occurring at amino acid residues within the PASE-linker. Later cleavage events (tertiary cleavages, 3) caused diminished intensities of the O-domain by further truncation (e.g., in loop structures).

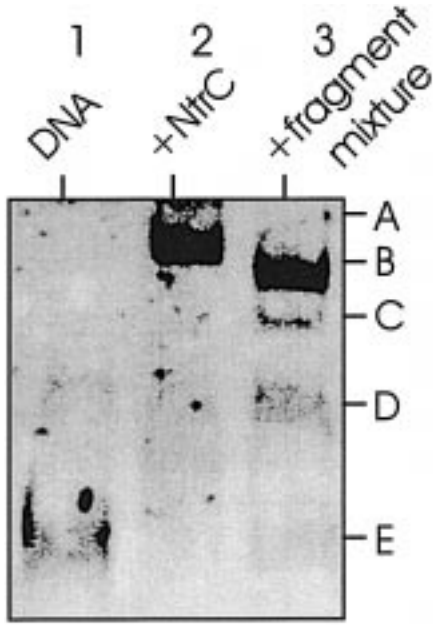

FIGURE 6: Polyacrylamide gel electrophoresis migration retardation assay. The duplex DNA (Lp-binding site) encompasses a strong binding site for NtrC dimers from the *glnLp* promoter of the *glnAntrBC* operon (*39*, *60*). Lane 1, DNA alone; Lane 2, DNA plus NtrC; Lane 3, DNA plus thermolytic fragments from NtrC. **A**, DNA complexed with NtrC; **B**, **C**, and **D**, DNA complexed with OC- and C-fragments, respectively; **E**, DNA alone.

ATPase activity was lost completely after removal of the R-domain from the OC-fragment, consistent with results from previous reports (*44*). The observed low ATPase activity in the tested samples (e.g., after thermolytic digestion with E:S = 1:864; cf. Figure 1) correlated to the still present residual amounts of undigested NtrC in the mixtures (data not shown). By contrast, DNA-binding studies using polyacrylamide gel electrophoresis migration retardation assays with samples that predominantly contained intact OC- and C-domain fragments showed characteristic band shifts after addition of a short double-stranded oligonucleotide (50 bp; cf. Materials and Methods) (*45*) that contained a strong binding site for NtrC dimers (Figure 6). A strong band may be attributed to a complex consisting of OC-fragments and duplex-DNA (B in Figure 6) as this band was observed at a slightly different location than that for the complex that consisted of duplex-DNA and intact NtrC (A in Figure 6). A further band observed in this sample (C) is in agreement with the assumption that OC- and C-fragments were bound simultaneously to the duplex-DNA. The fastest migrating band (D) may be attributed as a complex consisting of duplex-DNA and C-domains, as this band was still migrating differently than free DNA (E). From these results, we conclude that the C-domain of these proteolytically derived fragments exhibits an intact tertiary structure even after removal of the R-domain.

DISCUSSION

Knowledge of surface amino acid residues is of great benefit for understanding protein function, particularly for proteins of unknown structures. A well-established methodology for the investigation of partial peptides of proteins is the combination of proteolytic degradation with mass spectrometry (*46*, *47*), e.g. for disulfide bond assignments

(*48*, *49*). Here, hydrolytic cleavage is preferentially carried out under such conditions that particular peptide bonds are cleaved with a highly specific protease e.g. trypsin, creating a specific "peptide map". The local amino acid sequence satisfies the specificity requirements of the selected protease leading to defined products that are predictable from the amino acid sequence. By contrast, limited proteolysis using proteases with broad substrate specificity presents an effective method for determining exposed surface sites independent from the substrate specificity of the protease (*1*, *7*). In addition, protease susceptible protein conformations may be a result of dynamic processes so that flexibility of particular protein regions determines initial cleavage sites (*50*). Independent of their substrate specificity range, proteases represent molecular recognition systems that obey particular structure requirements. Substrate models for proteases have shown efficient cleavage when 10 and more residues from a given protein structure were able to locally unfold and to assume the suitable template conformation that was necessary for formation of the enzyme−substrate complex. The most susceptible nicksite in proteins was characterized to make the fewest interactions with the rest of the protein (*1*). All of the mentioned prerequisites for tertiary structure dependent limited proteolysis are thought to be present in linker regions of multidomain proteins, such as in the response regulator protein NtrC. Mass spectrometry plays a key role for the identification of the proteolytic fragments.

Our results showed that initial proteolytic cleavages had taken place within the interdomain linkers (cf. Figures 1 and 5) and the O-domain had been further truncated by later cleavage events. Consequently, controlled proteolysis may provide a suitable preparative approach for obtaining functionally active subfragments for structure−function studies of proteins as an alternative to the more commonly used approach of "domain liberation" by recombinant methods. The individually expressed R-domain of NtrC that comprised amino acid residues 1−124 (*25*) as well as a "12.5 kDa-band" obtained by tryptic proteolysis, resembling an amino terminal fragment of NtrC (*16*, *51*), have been shown to serve as substrates to phosphorylation by NtrB. Similarly, an individually expressed C-terminal fragment of NtrC containing the 90 C-terminal amino acid residues (i.e., amino acid residues 380−469) was found to be capable of binding to DNA (*20*), consistent with the results described herein. Finally, it has been demonstrated that the isolated central domain of NtrC-homologous proteins NifA and DctD is sufficient for the activation of transcription at σ^{54}-dependent promoters (*52*−*54*). In summary, all of the as yet described active receiver fragments of NtrC were either comprising the coresponding domain sequences and matching the herein determined domain boundaries or contained longer sequence stretches. By contrast, NtrC constructs that lacked the O-domain and parts of the N-terminal sequence of the C-domain showed DNA-binding activity (*55*). The active O-domain fragments mostly extended the herein determined domain borders. It is interesting to note that the O-domain encompassing secondary structure elements of NtrC were predicted (*24*), using the partial sequence comprising amino acid residues 141−376, well within the here determined O-domain borders.

Knowledge about domain boundaries is of importance for the creation of functional fusion proteins by recombinant methods. Recently, two fusion proteins (Fus1 and Fus2), comprising NtrC-domains, were cloned so that each of them contained the N-terminus of NtrC and the C-terminus of NifA. In the case of Fus1, the fusion position was chosen N-terminally to the detected "PASE-linker" region at residue Gln-379. This fusion protein was inactive in vivo and in vitro. By contrast, in the case of Fus2, the fusion position was chosen within the "PASE-linker" at residue Ser-393. Fus2 showed ATPase activity in vitro and was active in vivo as a transcriptional activator of a σ^{54}-dependent promoter carrying binding sites for NifA (Weiss et al., manuscript in preparation).

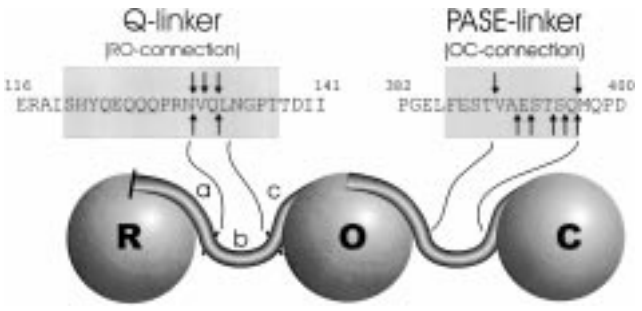

FIGURE 7: Schematic representation of domains and linker sequences of NtrC from *E. coli*. The three domains of NtrC (R, O, and C) are depicted as spheres and linkers as connecting ribbons. These are divided into three parts: ***a***, linker part that structurally belongs to the N-terminal domain; ***b***, structurally independent and, hence, protease susceptible hinge region; ***c***, linker part that structurally belongs to the C-terminal domain. Amino acid sequences (top) are given for the "Q-linker" (connecting R-domain with O-domain) and the "PASE-linker" (connecting O-domain with C-domain) in NtrC from *E. coli* (shaded boxes). Arrows depict cleavge sites of limited proteolysis with thermolysin (↓) and papain (↑), respectively.

Linkers may generally be composed of structurally distinct parts (Figure 7). At least three components may be distinguished in a linker sequence. An N-terminal stretch that structurally belonged to the N-terminal domain from which the linker originates, a structurally independent and, hence, protease susceptible central hinge region, and a part that structurally belonged to the C-terminal domain forms its terminus. The fact that cleavage sites obtained by limited proteolysis of NtrC with thermolysin, papain, subtilisin, and trypsin (*42*) were located within the central region of the Q-linker sequence (*26*) is in agreement with this structure model. The domain-associated linker parts may be highly flexible permitting local unfolding prior to proteolytic cleavage. A recently published X-ray crystallographically determined tertiary structure of NarL, a transcription activator consisting of two domains, lacked the linker part as this sequence stretch was not resolvable from the diffraction map (*56*, *57*), possibly because of its high flexibilty.

Analyzing the amino acid composition of linker regions of 13 related NtrC proteins showed that proline and alanine residues were found to be present in high abundance in these structurally flexible elements (cf. Table 3). These residues are found in abundance, e.g. in "PRO type" linkers by determining composition ratios to unrelated proteins (*43*). Most interestingly, a high abundance of serine residues was found in the linker connecting the O-domains with the C-domain. Serine residues may play a determinant role for the structural characteristics of these linkers as they are able to form hydrogen bonds with associated amino acid residues

besides undergoing polar interactions with the solvent (*43*). The amino acid composition further suggests that these linkers possess an acidic character due to the high abundance of glutamic acid residues. By contrast, the polarity pattern of Q-linkers shows accumulation of positively charged amino acid residues in the N-terminal part and negatively charged residues in the C-terminal part (cf. Figure 4) giving rise to a dipolar charge pattern. Summarizing these features and in analogy with the Q-linker that connects the R-domain with the O-domain (*26*), we use the name "PASE-linker" for the partial structures that connect the O-domains with the C-domains in NtrC proteins.

As demonstrated in this study, the combination of limited proteolysis with micropreparative techniques and mass spectrometry, such as *in-gel* digestion−MS coupling, MALDI-MS peptide mapping, and nano-ESI-MS provides a highly efficient tool for the rapid identification of domain borders and linker-regions in multidomain proteins, affording lead information of tertiary structural features necessary for protein design and engineering and for structure−function studies.

ACKNOWLEDGMENT

We are grateful to Dr. Michael Przybylski and to Dr. Winfried Boos in whose laboratories parts of this work were carried out. We thank I. Mettke for excellent technical assistance.

REFERENCES

1. Hubbard, S. J. (1998) *Biochim. Biophys. Acta 1382*, 191−206.
2. Jorgensen, C. I., Kallipolitis, B. H., and Valentin-Hansen, P. (1998) *Mol. Microbiol. 27*, 41−50.
3. Echabe, I., Dornberger, U., Prado, A., Goni, F. M., and Arrondo, J. L. (1998) *Protein Sci. 7*, 1172−1179.
4. Zappacosta, F., Pessi, A., Bianchi, E., Venturini, S., Sollazzo, M., Tramontano, A., Marino, G., and Pucci, P. (1996) *Protein Sci. 5*, 802−813.
5. Fontana, A., Zambonin, M., Polverino de Laurato, P., De Filippis, V., Clementi, A., and Scaramella, E. (1997) *J. Mol. Biol. 266*, 223−230.
6. Hu, Y., Craig, F., and English, A. M. (1996) *Inorg. Chim. Acta 242*, 261−269.
7. Bantscheff, M., Weiss, V., and Glocker, M. O. (1998) *Eur. Mass Spectrom. 4*, 279−285.
8. Fontana, A., Polverino de Laurato, P., De Fillipis, V., Scaramella, E., and Zambonin, M. (1997) *Fold Des. 2*, R17−R26.
9. Scaloni, A., Miraglia, N., Orru, S., Amodeo, P., Motta, A., Marino, G., and Pucci, P. (1998) *J. Mol. Biol. 277*, 945−958.
10. Macht, M., Fiedler, W., Kürzinger, K., and Przybylski, M. (1996) *Biochemistry 35*, 15633−15639.
11. Cohen, S. L., Ferre, D., R., A. A., Burley, S. K., and Chait, B. T. (1995) *Protein Sci. 4*, 1088−1099.
12. Magasanik, B. (1996) in *Escherichia coli and Salmonella* (Neidhardt, F. C., Ed.) pp 1344−1356, ASM Press, Washington, D. C.
13. Reitzer, L. J. (1996) in *Escherichia coli and Salmonella* (Neidhardt, F. C., Ed.) pp 391−407, ASM Press, Washington, D. C.
14. Parkinson, J. S., and Kofoid, E. C. (1992) *Annu. Rev. Genet. 26*, 71−112.
15. Stock, J. B., Surette, M. G., Levit, M., and Park, P. (1995) in *Two-component Signal Transduction* (Hoch, J. A., & Silhavy, T. J., Ed.) pp 25−51, ASM Press, Washington, D. C.
16. Keener, J., and Kustu, S. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 4976−4980.
17. Sanders, D. A., Gillece-Castro, B. L., Burlingame, A. L.; Koshland Jr., D. E. (1992) *J. Bacteriol. 174*, 5117−5122.
18. Flashner, Y., Weiss, D. S., Keener, J., and Kustu, S. (1995) *J. Mol. Biol. 249*, 700−713.
19. Weiss, D. S., Batut, J., Klose, K. E., Keener, J., and Kustu, S. (1991) *Cell 67*, 155−167.
20. Porter, S. C., North, A. K., Wedel, A. B., and Kustu, S. (1993) *Genes and Development 7*, 2258−2273.
21. Klose, K. E., North, A. K., Stedman, K. M., and Kustu, S. (1994) *J. Mol. Biol. 241*, 233−245.
22. Porter, S. C., North, A. K., and Kustu, S. (1995) in *Two-Component Signal Transduction* (Hoch, J. A., & Silhavy, T. J., Ed.) pp 147−158, ASM Press, Washington, D. C.
23. Morett, E., and Segovia, L. (1993) *J. Bacteriol. 175*, 6067−6074.
24. Osuna, J., Soberon, X., and Morett, E. (1997) *Protein Sci. 6*, 543−555.
25. Volkman, B. F., Nohaile, M. J., Amy, N. K., Kustu, S., and Wemmer, D. E. (1995) *Biochemistry 34*, 1413−1424.
26. Wootton, J. C., and Drummond, M. H. (1989) *Protein Eng. 2*, 535−543.
27. Bantscheff, M., Weiss, V., and Glocker, M. O. (1999) *4th Int. Symp. Mass Spectrom. Health Life Sci.*, in press.
28. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Saglio-cco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., Mann, M. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 14440−14445.
29. Patterson, S. D., and Aebersold, R. (1995) *Electrophoresis 16*, 1791−1814.
30. Mortz, E., Vorm, O., Mann, M., and Roepstorff, P. (1994) *Biol. Mass Spectrom. 23*, 249−261.
31. Fiedler, U., and Weiss, V. (1995) *EMBO J. 14*, 3696−3705.
32. Fiedler (1996) *Ph.D. thesis, University of Konstanz.*.
33. Reitzer, L. J., and Magasanik, B. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 5554−5558.
34. Weiss, V., and Magasanik, B. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 8919−8923.
35. Gill, S. C., and von Hippel, H. P. (1989) *Anal. Biochem. 182*, 319−326.
36. Schägger, H., and v. Jagow, G. (1988) *Anal. Biochem. 173*, 201−205.
37. Laemmli, U. K. (1970) *Nature 227*, 680−685.
38. Fligge, T., Bruns, K., and Przybylski, M. (1998) *J. Chromatogr. B Biomed. Sci. Appl. 706*, 91−100.
39. Weiss, V., Claverie, M. F., and Magasanik, B. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 5088−5092.
40. Sambrooke, J., Fritsche, E. F., and Maniatis, T. (1989) *Molecular Cloning* Cold Spring Harbor Laboratory Press, New York.
41. Norby, J. G. (1988) *Methods Enzymol. 156*, 116−119.
42. Farez-Vidal, E., Wilson, T. J., Davidson, B. E., Howlett, G. J., Austin, S., and Dixon, A. (1996) *Mol. Microbiol. 22*, 779−788.
43. Argos, P. (1990) *J. Mol. Biol. 211*, 943−958.
44. Drummond, M. H., Contreras, A., and Mitchenall, L. A. (1990) *Mol. Microbiol. 4*, 29−37.
45. Mettke, I., Fiedler, U., and Weiss, V. (1995) *J. Bacteriol. 177*, 5056−5061.
46. Glocker, M. O., Kalkum, M., Yamamoto, R., and Schreurs, J. (1996) *Biochemistry 35*, 14625−14633.
47. Przybylski, M., Glocker, M. O., Nestel, U., Schnaible, V., Blüggel, M., Diederichs, K., Weckesser, J., Schad, M., Schmid, A., Welte, W., and Benz, R. (1996) *Protein Sci. 5*, 1477−1489.
48. Glocker, M. O., Arbogast, B., Schreurs, J., and Deinzer, M. L. (1993) *Biochemistry 32*, 482−488.
49. Spiess, C., Happersberger, H. P., Glocker, M. O., Spiess, E., Rippe, K., and Ehrmann, M. (1997) *J. Biol. Chem. 272*, 22125−22133.

50. Neurath, H. (1980) in *Protein Folding* (Jaenicke, R., Ed.) pp 501−523, Elsevier/North-Holland Biomedical Press, Amsterdam.
51. Kamberov, E. S., Atkinson, M. R., Feng, J., Chandran, P., and Ninfa, A. J. (1994) *Cell. Mol. Biol. Res. 40*, 175−191.
52. Berger, D. K., Narberhaus, F., and Kustu, S. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 103−107.
53. Huala, E., and Ausubel, F. M. (1989) *J. Bacteriol. 171*, 3354−3365.
54. Huala, E., Stigter, J., and Ausubel, F. M. (1992) *J. Bacteriol. 174*, 1428−1431.
55. Chen, P. and Reitzer, L. J. (1995) *J. Bacteriol. 177*, 2490−2496.
56. Baikalov, I., Schröder, I., Kaczor-Grzeskowiak, M., Grzeskowiak, K., Gunsalus, R. P., and Dickerson, R. E. (1996) *Biochemistry 35*, 11053−11061.
57. Baikalov, I., Schroder, I., Kaczor-Grzeskowiak, M., Cascio, D., Gunsalus, R. P., and Dickerson, R. E. (1998) *Biochemistry 37*, 3665−3676.
58. McCaldon, P., and Argos, P. (1988) *Proteins 4*, 99−122.
59. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res. 22*, 4673−4680.
60. Ueno, N. S., Mango, S., Reitzer, L. J., and Magasanik, B. (1984) *J. Bacteriol. 160*, 379−384.

BI990781K