An Integrated Chemical Cross-linking and Mass Spectrometry Approach to Study Protein Complex Architecture and Function*⑤

Jie Luo‡, James Fishburn§, Steven Hahn§, and Jeffrey Ranish‡¶

Knowledge of protein structures and protein-protein interactions is essential for understanding biological processes. Chemical cross-linking combined with mass spectrometry is an attractive approach for studying protein-protein interactions and protein structure, but to date its use has been limited largely by low yields of informative cross-links (because of inefficient cross-linking reactions) and by the difficulty of confidently identifying the sequences of cross-linked peptide pairs from their fragmentation spectra. Here we present an approach based on a new MS labile cross-linking reagent, BDRG (biotin-aspartate-Rink-glycine), which addresses these issues. BDRG incorporates a biotin handle (for enrichment of cross-linked peptides prior to MS analysis), two pentafluorophenyl ester groups that react with peptide amines, and a labile Rink-based bond between the pentafluorophenyl groups that allows cross-linked peptides to be separated during MS and confidently identified by database searching of their fragmentation spectra. We developed a protocol for the identification of BDRG cross-linked peptides derived from purified or partially purified protein complexes, including software to aid in the identification of different classes of cross-linker-modified peptides. Importantly, our approach permits the use of high accuracy precursor mass measurements to verify the database search results. We demonstrate the utility of the approach by applying it to purified yeast TFIIE, a heterodimeric transcription factor complex, and to a single-step affinity-purified preparation of the 12-subunit RNA polymerase II complex. The results show that the method is effective at identifying cross-linked peptides derived from purified and partially purified protein complexes and provides complementary information to that from other structural approaches. As such, it is an attractive approach to study the topology of protein complexes. *Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.008318, 1–16, 2012.*

Most cellular processes are carried out by macromolecular complexes, and knowledge of the structure of these complexes is an essential step toward understanding how they function to control diverse cellular functions (1). Unfortunately, our ability to decipher the structure of many complexes has been hampered by the lack of robust technologies that can efficiently accomplish this goal. Although high resolution structures have been determined for many proteins and some protein complexes by x-ray crystallography, its ability to generate high resolution structures of large complexes is often limited by difficulties obtaining sufficient quantities of purified complexes, insolubility of complexes during crystallization trials, or difficulties obtaining diffraction quality crystals. When structures are obtained they often comprise only parts of the proteins because difficult areas have been removed to improve solubility or crystallization properties. Furthermore, protein crystallization typically occurs under conditions which are very different from physiological conditions, further limiting the value of this approach.

The use of site-specific cross-linking reagents and biochemical probes are also powerful approaches to investigate protein structure and the architecture of protein complexes. These approaches provide information that is indicative of the spatial proximity of amino acids or domains. Subsequently, these constraints can assist modeling of tertiary and/or quaternary structure. Unlike x-ray crystallography, site-specific cross-linking/probe approaches can be applied to large, heterogeneous complexes under physiological conditions. These approaches have been used to characterize the interaction sites of most of the components of the general transcription machinery with RNA polymerase II (Pol II)[1] (2, 3); to map the domains of the Hsp100 chaperone ClpA involved in substrate binding, unfolding, and translocation (4); to deduce the quaternary structure of ligand-gated ion channels (5); to identify the targets of transcriptional regulatory proteins (6, 7); and to study the reorganization of the *Escherichia coli* σ54-RNA polymerase-promoter DNA complex during transcription initia-

[1] The abbreviations used are: Pol II, RNA polymerase II; BDRG, biotin-aspartate-Rink-glycine; BLG, β-lactoglobulin; CID, collision-induced dissociation; DIC, *N,N'*-diisopropylcarbodiimide; PFP, pentafluorophenyl; Fmoc, *N*-(9-fluorenyl)methoxycarbonyl; SCX, strong cation exchange.

tion (8). However, there are two steps in these site-specific cross-linking/probe methods that have limited their general applicability. The first step is the modification of specific residues in the protein(s) of interest with the cross-linker/probe for each measurement. In addition to being time-consuming, the choice of residues to modify in this step can bias the identified interactions. The second step is the identification of the interacting proteins and/or specific sites that have been modified by the reagent. This requires additional time-consuming steps, such as genetic modification of the interacting protein with an epitope tag or analysis by a technology capable of providing amino acid sequence information.

Chemical cross-linking combined with MS provides a particularly promising method for inferring sites of protein-protein interactions and for mapping the topology of protein complexes because it is fast, sensitive, and data-rich (see Refs. 9 and 10 for reviews). The approach commonly entails the use of bi-functional cross-linking reagents that react with primary amine groups, which are present on the large majority of proteins because of the high frequency of lysine residues in most proteins. Unlike the approaches described above, there is no need to select and modify specific residues prior to the analysis, and identification of cross-linking sites is facilitated by tandem MS analysis with instruments capable of making thousands of high mass accuracy measurements per sample (see below).

Thus far, the chemical cross-linking/MS approach has been limited primarily to the analysis of single proteins or small complexes. This is mainly due to difficulties in the detection and identification of the informative cross-linked peptides in samples of high complexity such as those derived from large protein assemblages. Detection of informative cross-links is difficult because cross-linking experiments expand the complexity of the sample with the informative interpeptide cross-links present at low quantities relative to unmodified peptides, monolinks (reaction products in which only one functional group of the cross-linker has reacted with a protein), and loop-links (reaction products in which both functional groups of the cross-linker have reacted with residues that reside in one peptide after enzymatic digestion). The presence of a large excess of unmodified peptides, monolinks, and loop-links reduces the yield of spectra from the informative interpeptide cross-links; for larger complexes, this problem is exacerbated because of the increased complexity of the samples. Several approaches have been devised to facilitate detection of interpeptide cross-links during MS analysis. They include the use of cross-linking reagents that produce diagnostic fragmentation patterns during collision-induced dissociation (CID) (11–15), isotope-coded cross-linkers (16–18) or proteins (19), isotopic labeling of peptides derived from cross-linking reaction (20, 21), and enrichment of cross-linked products via affinity handles (22, 23).

Identification of the peptides involved in a particular cross-linked pair is challenging because cross-linked peptides typ-

ically generate fragmentation spectra that are very complex and difficult to interpret. Whereas fragmentation of a peptide produces an ion series representing the linear structure of the peptide, fragmentation of cross-linked peptides produces a set of ions representing two linear peptide structures, as well as ions that span the cross-linker. To address this challenge, two main strategies have been pursued, the first of which is the development of specialized search algorithms to identify cross-linked peptide pairs directly from their fragmentation spectra, *e.g.* PepLynx (21), Protein Prospector (24), xQuest (25), X!link (26), Xi (27), Pro-Cross-link (28), MS2Assign (29), and ASAP (30). Although this approach takes advantage of the availability of a diverse array of commercial cross-linkers, confident assignment of peptide sequences to spectra derived from cross-linked peptides continues to a pose a formidable challenge. In addition, most of these algorithms require construction of sample-specific databases to limit the number of potential peptide pairs to consider during database searching. As sample complexity increases, these algorithms lose power to make confident identifications because of the explosion of database search space that occurs when all possible peptide-peptide cross-link combinations between components in the sample are considered. As a result of these issues, a small number of cross-linked peptides are typically identified with high confidence in samples derived from large complexes, limiting the utility of the data for mapping the architecture of protein complexes. Although this approach has had limited success for the analysis of protein complex topology to date, a recent landmark report (31) of the identification of 108 high confidence linkage pairs derived from a 12-subunit RNA Pol II complex suggests that analysis of large complexes by this approach may soon become more routine.

The second strategy is the use of "MS labile" cross-linkers (13–15, 23, 32) that fragment either by in source decay or by CID (MS2) to transform cross-linked peptides into two modified peptides that can in turn be selected for CID (MS3) by data-dependent routines and identified by search algorithms such as Sequest or Mascot that are commonly used to identify linear peptides. The reduced complexity of the MS3 spectra facilitates confident peptide identification. Previously described MS labile cross-linkers have exploited the labile properties of the aspartyl-prolyl bond (D–P) (32, 33), various forms of a carbon-sulfur bond (13, 15), a urea moiety (14), and a Rink moiety (11, 23). Except for the D–P-based cross-linker, all of these reagents can produce multiple fragments (four or more) during CID of interpeptide cross-links because of the presence of two labile bonds. This can negatively impact peptide identification caused by reduced product ion intensity and duty cycle limitations. In addition, CID of interpeptide cross-links containing any of these reagents, except for the Rink-based cross-linker, produces modified peptides with different modification masses. This issue must be addressed by considering multiple amine modification masses during database searching. Like nonlabile reagents, thus far, MS labile

reagents have primarily been used to identify cross-linked peptides derived from synthetic peptides, small proteins, or small peptide-protein complexes.

In this paper, we report an approach for identifying cross-linked peptides derived from complex samples based on BDRG, a new MS labile cross-linking reagent. BDRG incorporates a biotin affinity handle and two amine-reactive pentafluorophenyl (PFP) ester groups separated by a single Rink-based bond. Importantly, fragmentation of BDRG at the Rink bond generates two ions with nearly equal masses. As such, BDRG combines the advantages of MS labile cross-linkers for confident peptide identification with an intrinsic enrichment strategy and a reduced propensity to generate multiple product ions during CID. This approach involves a protocol for the application of BDRG to identify cross-linked peptides derived from purified and partially purified protein complexes and software for identification of different classes of cross-linker-modified peptides. We demonstrate the approach by applying it to purified yeast TFIIE, a heterodimeric transcription factor complex, and to a single-step, affinity-purified preparation of the 12-subunit RNA Pol II complex. Our results show that the BDRG approach is effective at identifying cross-linked peptides derived from purified and partially purified protein complexes and provides information complementary to that from other structural approaches. As such, it provides an attractive method to study the topology of protein complexes and to infer sites of protein-protein interactions within complexes.

EXPERIMENTAL PROCEDURES

Synthesis of BDRG Molecule—All of the chemicals were purchased from Novabiochem (San Diego, CA) and Sigma-Aldrich. BDRG free acid was synthesized by conventional Fmoc solid phase peptide synthesis using Fmoc-Gly-NovaSyn TGT resin. Fmoc-Rink linker, Fmoc-Asp(O-2-PhiPr)-OH, and Biotin-ONp were coupled to the resin sequentially. The free acid form of BDRG was then cleaved from the resin using 1% TFA/dichloromethane and precipitated by the addition of pure water. The free acid was then dried under vacuum. To activate the BDRG free acid, we dissolved the molecule in 50 mM in DMF, with three molar equivalents of *N*-hydroxysuccinimide and *N,N'*-diisopropylcarbodiimide (DIC). Despite our best efforts, only ~50% of the free acid was converted to the *bis* activated form. We used this partially activated BDRG product to test the properties of the BDRG molecule and to perform initial cross-linking studies on β-lactoglobulin (BLG). Subsequently, we obtained a purified PFP-activated BDRG molecule through custom synthesis from Almac Sciences (Scotland, UK). A detailed synthesis procedure is provided in the [supplemental materials]().

Proteins and Protein Complexes—β-Lactoglobulin was purchased from Sigma-Aldrich (L8005). Recombinant TFIIE was expressed in BL21-RIL cells from plasmid pJF23 (pETDuet-Tfa1,SUMO-Tfa2) containing a His$_6$-SUMO N-terminal tag on Tfa2 and a coexpressed untagged Tfa1. TFIIE was purified on nickel-Sepharose, the SUMO tag was removed by SUMO protease (34), and TFIIE was further purified using H-Trap Heparin using a linear gradient of 100–500 mM KCl over 30 column volumes. Purified fractions were pooled and stored at −80 °C. RNA Pol II was purified from a strain carrying a 3× FLAG epitope-tagged allele of *RPB3*. To create this strain, we first constructed plasmid pJL-HFH-1 containing the 3× FLAG epitope tag with a *CYC1* terminator and a *URA3* selectable marker, flanked by the sequences (~40 bp) identical to the start and end of the TAP tag and the *HIS3* selectable marker, respectively, present in the TAP tag yeast library from Open Biosystems. We swapped the TAP tag present in strain 6GS2 A4 containing an *RPB3-TAP-HIS3MX6* allele (Open Biosystems) with the FLAG tag by transformation of strain 6GS2 A4 with a PCR product from the pJL-HFH-1 plasmid containing the appropriate sequences followed by selection on CSM-URA plates. We grew 12 liters of the RPB3-FLAG strain in YPD (1% yeast extract, 2% peptone, 2% glucose) overnight to A_{600} of 8–10. The cells were harvested by centrifugation and lysed by glass bead beating in lysis buffer (50 mM HEPES, pH 7.9, 400 mM ammonium sulfate, 10 mM MgSO$_4$, 1 mM EDTA, 20% glycerol) with protease inhibitors. The cell lysate was then centrifuged at 20,000 × *g* for 1 h, and the supernatant was transferred to new tubes. The protein concentration was determined by Bio-Rad protein assay, and the cell lysate was frozen at −80 °C. To purify the RNA Pol II complex, we diluted 1 g of protein (from ~2 liters of cells) to 10 mg/ml with 50 mM HEPES, pH 7.9, 10 mM MgSO$_4$, 1 mM EDTA with protease inhibitors and centrifuged the diluted protein again at 20,000 × *g* for 1 h. The supernatant was then loaded onto a 3-ml anti-FLAG M2 affinity agarose (Sigma-Aldrich) column equilibrated in PBS buffer. We repeated the loading three times. Then the beads were washed extensively with lysis buffer without glycerol and wash buffer (50 mM HEPES, pH 7.9, 0.5 M NaCl, 1 mM EDTA, 1%Triton X-100, 0.1% sodium deoxycholate). After extensive washing, Pol II was eluted with 3× FLAG peptide (Sigma-Aldrich) at 0.1 mg/ml in elution buffer (50 mM HEPES, pH 7.9, 110 mM KOAc, 5 mM MgSO$_4$, 1 mM EDTA). The eluted protein was then concentrated and buffer exchanged by repeated centrifugation and resuspension with elution buffer in an Amicon Ultra-4 with 10,000 molecular weight cutoff (Millipore) to reduce the concentration of the 3× FLAG peptide. After a final centrifugation step, protein concentration was determined by Bio-Rad protein assay. We usually isolated ~200–300 μg of protein. We checked the purity of the sample and Pol II subunit composition by Coomassie-stained SDS-PAGE.

Protein Cross-linking and Sample Preparation for Mass Spectrometry—To cross-link proteins, we dissolved proteins in 100 μl of 200 mM HEPES, pH 7.9, or PBS, pH 7.5. The BDRG cross-linker is predissolved in DMF at 20 mM and added directly to the cross-linked proteins at 0.1 to 0.5 mM final concentration. A microprecipitate typically forms because of the insolubility of the cross-linker in aqueous solution. The reaction is performed at room temperature for 30 min to 2 h with occasion disturbance. Then 10 μl of 1 M NH$_4$HCO$_3$ is added to quench the reaction, and the proteins are precipitated by adding 5 volumes of acetone and incubating overnight at −20 °C. The precipitated proteins are dissolved in 1 M urea and digested with trypsin (1:20 w/w) overnight at 37 °C. The peptides are then purified by C$_{18}$ chromatography (The Nest Group, Inc.), and BDRG-modified peptides are enriched using an avidin cartridge (Invitrogen) as follows. The avidin column was first washed twice with 500 μl of elution buffer (0.4% TFA, 30% ACN) at 200 μl/min and then equilibrated with 1 ml of 1× PBS, pH 7.5. The peptides were dissolved in 200 μl of 1× PBS, pH 7.5, and loaded onto the avidin column at 50 μl/min. The flow through was reapplied to the column. Then the column was washed twice with 500 μl of 1× PBS, pH 7.5; twice with 500 μl of wash buffer (50 mM ammonium bicarbonate, 20% methanol), and once with 500 μl of 20% methanol at the rate of 400 μl/min. The peptides were eluted with 800 μl of elution buffer at 50 μl/min and dried in a SpeedVac. No BDRG-modified peptides were observed in the avidin flowthrough by LC-MS analysis. More than 90% of the identified peptide spectra from the avidin-enriched fraction correspond to peptides modified by BDRG. The avidin-enriched peptides can be directly analyzed by LC-MS3, or they can be further fractionated by strong cation exchange chromatography (SCX). SCX was performed using a 2-mm × 1-cm guard cartridge (Idex Health & Science, Oak Harbor,

WA), packed with Partisphere SCX (5 μm, 200 Å, Whatman). The flow rate was set at 50 μl/min using a syringe pump (Harvard Apparatus, Holliston, MA). The column was equilibrated in buffer A (0.5% acetic acid, 2% ACN). Peptides were eluted stepwise with the following buffers: 90 mM ammonium acetate, 0.5% acetic acid, 2% ACN; 120 mM ammonium acetate, 0.5% acetic acid, 8% ACN; and 250 mM ammonium acetate, 0.5% acetic acid, 12% ACN.

Mass Spectrometry and Data Analysis—Peptide samples were analyzed by reversed phase HPLC electrospray ionization LC-MS using an LTQ-Orbitrap (Thermo Scientific). The HPLC column (75 μm $\times$ 15 cm) was packed in house with C_{18} resin (Magic C_{18} AQ, 5 μm; Michrom BioResources, Auburn, CA). The peptides were resolved by running a gradient of buffer A (0.1% formic acid) to buffer B (0.1% formic acid, 99.9% ACN) as follows: 5–20% buffer B over 15 min, 20–35% buffer B over 60 min, and 35–80% buffer B over 10 min. A fixed flow rate of 350 nl/min was used.

In general, MS1 scans were acquired in the Orbitrap with a resolution of 60,000 from 400 to 1800 m/z. Multistage tandem MS scans were acquired in the LTQ using data-dependent acquisition based on the intensity of the ions observed in the preceding scan. For each Orbitrap MS1 scan, 5×10^5 ions were accumulated over a maximum time of 500 ms. For each LTQ multistage tandem MS scan, 5×10^3 ions were accumulated over a maximum time of 250 ms. The normalized collision energy for CID was set at 35%. A CID isolation window of 2 m/z was used. For each MS1 scan, we performed upto nine multistage tandem MS scan events using parallel mode with dynamic exclusion of 3 min for each ion selected for MS2. These scans included three MS2 scans on the three most intense ions in the MS1 scan. For each MS2 scan, two MS3 events were triggered based on the two most intense ions in the MS2 scan. For analysis of the avidin flow through samples, five MS2 scans on the five most intense ions in the preceding MS1 scan were used.

RAW files from the LTQ-Orbitrap were converted to mzXML by ReAdW (version 4.3.1) with default parameters. The MS2 and MS3 scan events were searched separately by Sequest-pvm v.27 (runsequest-c2 for MS2 (default) and runsequest-c3 for MS3 search) against a bovine database (20060720, 9396 entries) or a yeast protein database (Saccharomyces Genome Database, 20060126, 7588 entries) with differential modifications of 340 Da on lysine and 16 Da on methionine. The precursor peptide mass tolerance was set at 3 Da, the fragment ion mass tolerance was set at 1 Da, the number of tryptic termini was set to 1, and the maximum number of missed cleavage sites was set at 3. Peptide identification was achieved by processing the search results with the Trans-Proteomic Pipeline. For any ion to be identified as a BDRG-modified peptide (monolinks, loop-links, or cross-links), at least one MS2 or MS3 spectrum from the same precursor is required to be identified as a Lys-modified peptide. A 5% error rate as determined by PeptideProphet (35) was used for each analysis. For the Pol II sample, we searched the cross-linking data against the yeast database, a reversed yeast database and the human IPI database (ipi.HUMAN.v 3.71, 86745 entries) to define the false positive rate for identification of the modified peptides. At an error rate of 5%, one spectrum was assigned to a Lys-modified peptide in the yeast reverse database search, seven spectra (three unique peptides) were assigned to Lys modified peptides in the human database search, whereas 1906 spectra (216 unique peptides) were assigned to Lys-modified peptides in the yeast database search. Therefore, the false positive rate for identification of modified peptides is less than 0.3% at a 5% error rate cutoff. For the Pol II sample, a confined search against the 102 proteins identified in the sample was also performed. 2940 spectra (272 unique peptides) were assigned to Lys-modified peptides at a 5% error rate. If all of these assignments are correct, the sensitivity (or false negative rate) of the whole proteome database search would be ~65–70%. We noticed that the search was especially biased against short peptides. Short peptides are difficult to confidently identify because they produce relatively few fragment ions that can be used for database searching. In addition, the 340-Da mass modification can be especially problematic for short peptides where it contributes a significant portion of the peptide mass without contributing sequence-specific fragment ions that can be used for database searching.

The peptides identified with Lys modifications were output as an interact.pep.xls file. A perl script (RunBDRGlink.pl, available upon request) used the peptide and protein identities from this file and the MS1 precursor mass and scan numbers from the corresponding mzXML files to infer whether the identified BDRG-modified peptides were monolinks, cross-links, or loop-links as described in supplemental Fig. 1.

RESULTS

The BDRG Cross-linking Approach—To localize sites of protein-protein interactions in protein complexes and to map the topology of complexes, we have developed an MS-based approach to identify cross-linked peptides derived from protein complexes. It is based on a new, homo-bifunctional, MS labile cross-linking reagent called BDRG (Fig. 1*A*) that we designed with the following features: 1) BDRG has two PFP ester groups that react with primary amines, and to a lesser extent with secondary amines, at pH 8 (36). Accordingly, when applied to proteins, the reagent will react primarily with the ε-amino group of lysine residues and the α-amino group of N termini. Compared with *N*-hydroxysuccinimide activated esters, PFP activated esters exhibited higher reactivity toward amine groups and greater stability in aqueous solution (37). 2) BDRG contains a biotin moiety that is used to enrich cross-linker-modified peptides by avidin affinity chromatography. 3) BDRG contains a single MS labile Rink linker group between the two PFP groups, which permits the separation of two cross-linked peptides during CID. Importantly, depending on the orientation of the cross-linker joining the two peptides, either peptide can be left modified with either the BD or RG moiety upon fragmentation at the Rink bond. However, by design, these modifications differ in mass by only 1 Da. Thus, the two modified forms of a peptide are not resolved during LTQ analysis, and only one mass modification needs to be considered during database searching. These two features, a single MS labile bond and the production of two cross-linker-derived moieties of nearly equal mass upon fragmentation, reduce the number of fragment ions observed in the MS2 spectrum compared with other MS labile approaches. This facilitates identification of cross-linked peptides by minimizing duty cycle constraints, maximizing the abundance of the peptide contributing to each MS3 spectrum and simplifying database searching. 4) BDRG has a maximum theoretical cross-linking distance of 27.0 Å, which should provide spatial constraints that are useful for modeling the architecture of complexes.

A flow chart describing the BDRG approach is presented in Fig. 1*B*. The approach begins with a cross-linking reaction of a purified or partially purified protein complex with BDRG. The
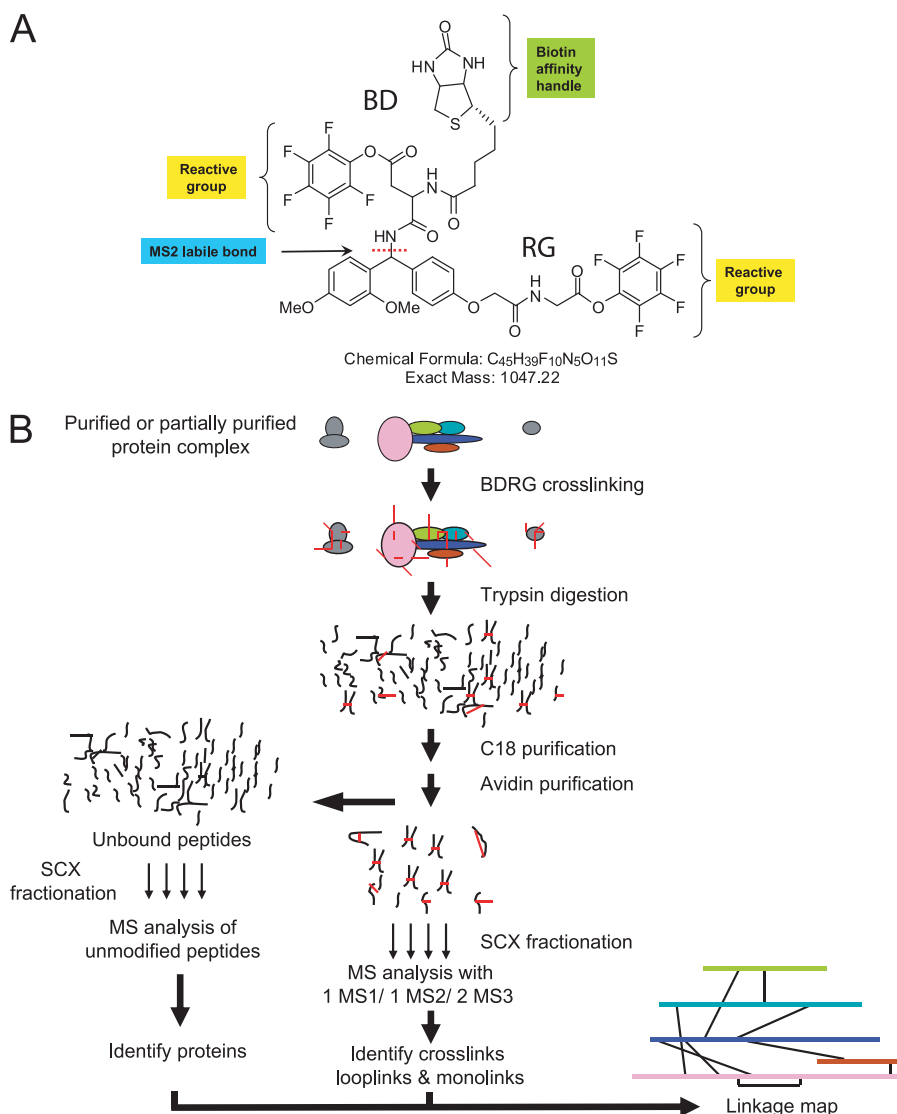
FIG. 1. **Overview of the BDRG cross-linking approach for mapping the architecture of protein complexes.** *A*, structure of the BDRG cross-linker. BDRG is composed of a biotin affinity handle (B), a PFP ester modified aspartate reside (D), a Rink group (R), and a PFP ester modified glycine reside (G). The MS labile bond is indicated by the *dashed line*. *B*, schematic of the BDRG approach for mapping the architecture of protein complexes. See text for details.

cross-linked sample is then precipitated with acetone to remove BDRG that has not reacted with proteins, resuspended in buffer, and digested under denaturing conditions with trypsin. After trypsin digestion, the sample is purified by C_{18} chromatography. Next, cross-linker-modified peptides are enriched by avidin affinity chromatography and then fractionated by SCX chromatography to further reduce sample complexity. Each fraction is analyzed using a high mass accuracy LTQ-Orbitrap using the following method: one MS1 event in the Orbitrap followed by one data-dependent MS2 event in the ion trap on the most abundant ion in the previous MS1 spectrum and then two data-dependent MS3 events in the ion trap on the two most abundant ions in the MS2 spectrum (Fig. 2*A*). MS2 and MS3 spectra are then used to search an appropriate database with Sequest to identify potential BDRG-modified peptides (see below). The avidin flow through fraction is also fractionated by SCX chromatography and analyzed by LC-MS1-MS2 followed by Sequest database

searching to identify the proteins present in the original sample.

Strategy to Identify BDRG Cross-linked Peptides—Because BDRG contains an MS labile Rink linker group between the two amine reactive PFP groups (Fig. 1*A*), CID of BDRG results primarily in the generation of two fragment ions: the BD ion (formula, $C_{14}H_{20}N_4O_4S$; monoisotopic mass, 340.12; average mass, 340.40), containing the biotin and aspartate moieties, and the RG ion (formula, $C_{19}H_{19}NO_5$; monoisotopic mass, 341.12; average mass, 341.36), containing the Rink linker and glycine moiety. During MS2 analysis of interpeptide cross-links, cleavage also occurs predominantly at the Rink labile bond, resulting in the generation of BD- and RG-modified peptide ions (Fig. 2*A*). CID of these peptide ions followed by database searching of their MS3 spectra is then used to identify each peptide. We note that each peak corresponding to the modified peptide ions may be composed of a mixture of the peptide modified by either the BD or the RG moiety, which
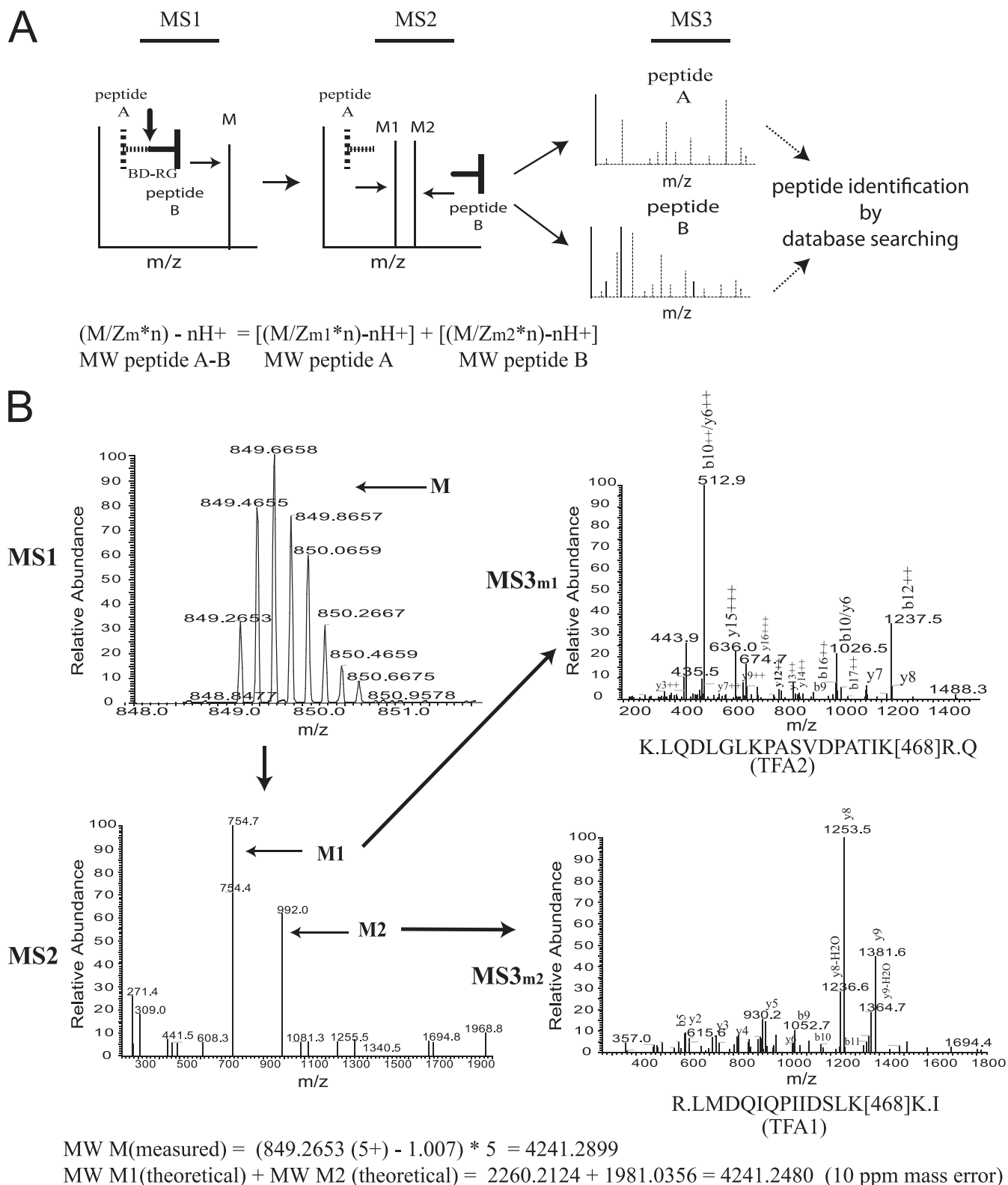
$$(M/Z_m*n) - nH+ = [(M/Z_{m1}*n)-nH+] + [(M/Z_{m2}*n)-nH+]$$

MW peptide A-B   MW peptide A   MW peptide B

K.LQDLGLKPASVDPATIK[468]R.Q
(TFA2)

R.LMDQIQPIIDSLK[468]K.I
(TFA1)

MW M(measured) = (849.2653 (5+) - 1.007) * 5 = 4241.2899

MW M1(theoretical) + MW M2 (theoretical) = 2260.2124 + 1981.0356 = 4241.2480 (10 ppm mass error)

FIG. 2. **Identification of BDRG cross-linked peptides.** *A*, schematic describing the identification of BDRG cross-linked peptides by LC-MS3 analysis. Peptide A is cross-linked to peptide B by BDRG. The MS labile bond is indicated by the *arrow*. *M* is the ion corresponding to cross-linked peptide A-B. *M1* and *M2* are the product ions corresponding to peptide A and B, respectively, after MS2 fragmentation of cross-linked peptide A-B. *m/z*, mass to charge ratio; *n*, the integer number of charges on the analyte; H+, the mass of a proton = 1.007; *MW*, molecular weight of the analyte. *B*, example of MS data that led to the identification of an intermolecular cross-link in TFIIE. The detection of

would not be resolved by the LTQ. Because the BD and RG moieties have similar masses, a database search can be performed to identify the modified peptides using a single differential mass modification on lysine of 340 Da. The two identified peptides predict a theoretical mass for the intact cross-linked peptide pair, which was measured at high accuracy during MS1; comparison of the measured and theoretical masses provides additional verification that the cross-linked peptide pair has been correctly identified.

Fig. 2*B* shows an example of the MS analysis of an inter-peptide cross-link derived from the heterodimeric yeast TFIIE complex. A 5+ ion, *M*, with an *m/z* of 849.26 is observed in the high resolution MS1 scan. CID of this ion results primarily in the generation of two ions, M1 and M2, with *m/z* values of 754.7 and 992.0, respectively, in the MS2 spectrum (*lower left*). These two ions are then selected for MS3 analyses. Sequest searches of the MS3 spectra against a yeast protein database (7588 entries) result in the unambiguous identification of peptides corresponding to the two subunits of TFIIE. The data are consistent with BDRG linkage of the two peptides at the lysine residues labeled *K[468]*, and the combined neutral masses of these two modified peptides equals the neutral mass corresponding to the precursor peptide ion's measured *m/z*.

Other types of BDRG cross-linked products can be easily distinguished by inspection of the data and the search results. A monolink generates a diagnostic reporter ion during MS2 analysis with an *m/z* of 358/359 that corresponds to a BD or RG fragment, after hydrolysis of the PFP ester, and a peptide ion with a 340-Da modification on a lysine residue. The mass of the monolinked peptide, identified from its MS3 spectra, will be 358/359 Da less than the mass of the precursor peptide. Thus, this 358/359-Da mass difference is used to identify monolinked peptides (Fig. 3*A*). We also observed water loss at the unlinked end of the cross-linker during electrospray, which results in a 340-Da mass difference between the precursor ion and the monolinked peptide ion observed during MS3 analysis. CID of BDRG loop-links results in extensive fragmentation along their peptide backbone (Fig. 3*B*), presumably because the energy used to break the Rink labile bond remains trapped in the peptide ion. Similar behavior was also observed for loop-links derived from other cross-linkers containing a single labile bond, such as the D–P cross-linker (33). The fragmentation behavior of BDRG loop-links permits their identification as doubly modified peptides by database searching of their MS2 spectra (Fig. 3*B*). This behavior is in contrast to that of protein interaction reporter loop-links,

which instead generate a reporter ion and a modified peptide ion during CID, requiring acquisition of MS3 spectra for their identification by database searching (23).

A perl script has been developed for automated analysis of BDRG cross-link data using peptide identification information from a PeptideProphet output file, the measured accurate masses of the cross-linker-modified peptide ions from MS1 spectra, the calculated masses of the cross-linker-modified product ions, and scan number information to generate a list of the identified peptide sequences, the locations of the cross-linker-modified residues, and the type of cross-link (supplemental Fig. 1).

Identification of Cross-linked Peptides in β-Lactoglobulin—To evaluate the performance of the BDRG cross-linking strategy, we initially applied it to analyze the topology of purified bovine BLG. After incubating 100 μg of BLG with BDRG in 100 μl of PBS buffer, pH 7.5 (0.2 mM final BDRG concentration, cross-linker to protein concentration ratio, 5:1), for 1 h, the cross-linked sample was quenched by ammonium bicarbonate, acetone-precipitated, and trypsin-digested. Cross-linker-modified peptides were then purified by avidin affinity chromatography, and the sample was analyzed using an LTQ-Orbitrap as described in Fig. 1*B*. The MS2 and MS3 spectra were used to search a bovine protein database to identify monolinks, loop-links, and cross-links. We identified 10 monolinks, 6 loop-links, and 13 cross-links in BLG (Table I and supplemental Table 1). These results compare favorably to previously reported results from studies that used either the DSS (25) or BS[3] (21) cross-linker in which four cross-links were identified. The only previously identified cross-link that we did not observe was a link between Lys-47 and Lys-69. This could be due to the fact that we did not reduce the disulfide bond between residue 66 and 160 prior to trypsin digestion. We mapped the cross-links/loop-links onto the three-dimensional structure of the protein (Protein Data Bank code 1bsy) and measured the distances between the nitrogen atoms involved in the linkages (Table I and supplemental Fig. 2*A*). The cross-links clustered in two distinct locations on the surface of the protein separated by Lys-8, and the cross-linked distances ranged from 5.6 to 30.5 Å (median, 17.07 Å). The first cluster consists of Lys-75, Lys-77, Lys-83, and Lys-91. The distances between these residues are less than 20 Å. The second cluster consists of Lys-135, Lys-138, and Lys-141, with distances smaller than 12 Å. Lys-8 cross-links to residues within both clusters. Interestingly, BLG can exist as a homodimer in solution at pH 6.2–8.2 with three distinct interprotein contacts (38). One of the interfaces in-

---

a 5+ peptide ion, *M*, with *m/z* of 849.26 in a high resolution Orbitrap MS1 scan triggers an MS2 scan event. During MS2, this ion is isolated and subjected to CID. This results primarily in fragmentation at the Rink bond, which liberates the two cross-linker-modified peptide ions, M1 and M2. Next, each product ion, M1 and M2, is isolated and subjected to CID to produce MS3 spectra that are used to identify its peptide sequence and the site of cross-linking by sequence database searching. The site of cross-linking is indicted by the Lys-468 residue. 468 corresponds to the mass in Daltons of the BD- or RG-modified lysine residue. The sum of the theoretical masses of the two identified cross-linker-modified peptides, M1 and M2, is calculated and compared with the measured mass of the precursor peptide, M.
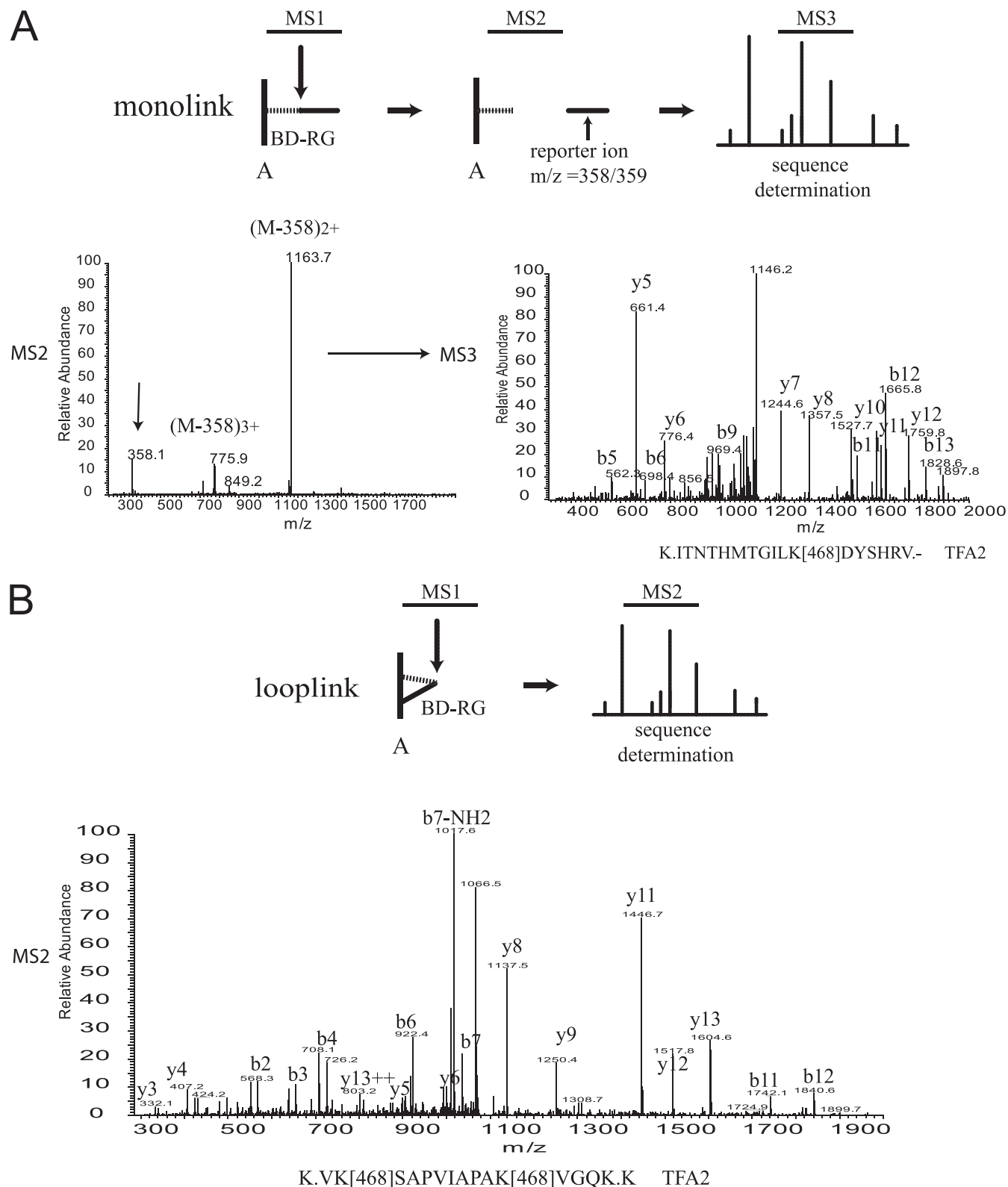
FIG. 3. **Identification of BDRG derived monolinks (A) and loop-links (B).** *A*, schematic describing the identification of a BDRG-derived monolink and MS data that led to the identification of a monolinked peptide from Tfa2. An ion with *m/z* = 358 is a reporter for monolinks. During MS2 analysis, monolinks will mainly generate product ions with a loss of 358 from the precursor ion. MS3 analysis of the 2+ ion led to identification of the indicated BDRG-modified peptide from Tfa2. The site of BDRG modification is indicted by the Lys-468 residue.

volves the side chain of Lys-8 acting as a "key" inserted into a "lock" from the neighboring molecule (Protein Data Bank code 2AKQ) (38). We noticed that at this interface the distance between Lys-83 and Lys-135 is 11.63 Å, and the distance between Lys-75 and Lys-135 is 18.16 Å, whereas Lys-75/Lys-83 and Lys-135 are located at two separate regions in the BLG monomer at distances of 30.52 and 27.65 Å, respectively. Therefore, it is possible that the identified cross-linked peptides consisting of K.IPAVFK[468]IDALNENK.V (Lys-83) and R.TPEVDDEALEK[468]FDK.A (Lys-135), and R.TPE-VDDEALEK[468]FDK.A (Lys-135) and K.IIAEK[468]TK.I (Lys-75) result from cross-linking at this interface. However, because BLG is a homodimer, we cannot distinguish between intra- and interprotein cross-linking for these sites. It is also possible that the cross-link involving Lys-8 and Lys-75 results from an interprotein interaction. Nonetheless, the results demonstrate that the BDRG method is effective at identifying residues that are in close proximity in monomeric proteins.

Importantly, we have used the accurate mass measurements of precursor ions acquired in the Orbitrap to validate the database search results. The theoretical masses of all the cross-linked peptide precursors corresponding to two component peptides identified by database searching of the two MS3 spectra are within 20 ppm of the measured precursor masses for all of the cross-linked peptide pairs identified in the BLG study (Table I). Because the accurate precursor masses are not used during the peptide identification process, they can be used to evaluate the validity of the identified cross-linked peptides, thus providing further confidence in identification of cross-linked peptides by the BDRG strategy (see "Discussion").

Application of the BDRG Approach to Study the Architecture of TFIIE—Next we applied the BDRG approach to study the architecture of the yeast general transcription factor TFIIE. TFIIE exists primarily as a heterodimer in solution composed of the Tfa1 and Tfa2 subunits. The TFIIE large subunit Tfa1 contains an N-terminal winged helix domain and an adjacent zinc-binding domain (Fig. 4). The TFIIE small subunit Tfa2 contains two tandem winged helix domains (39, 40). NMR structures of three conserved TFIIE domains have been determined (39, 41, 42), but only a low resolution EM structure of the TFIIE heterodimer is available (43). This is likely due to difficulties in obtaining diffraction quality crystals of the complete complex. Thus, it is particularly important to develop alternative approaches to study the architecture of complexes such as TFIIE.

100 μg of purified TFIIE (supplemental Fig. 3) was cross-linked with BDRG (final concentration, 0.2 mM; cross-linker to protein concentration ratio, 20:1) in 100 μl of HEPES buffer,

pH 7.9, for 1 h and analyzed by the BDRG approach as described above. We detected nine interprotein cross-links and seven intra-protein cross-links (supplemental Fig. 6). The cross-linking results are summarized in Fig. 4 and supplemental Table 2. The combined theoretical masses of all cross-linked peptide precursors corresponding to two component peptides identified by database searching are within 20 ppm of the measured precursor peptide masses (supplemental Table 2). We observed 28 spectra that matched to 19 unique loop-linked peptides and 269 spectra from 49 unique monolinked peptides. We observed loss of water and oxidation of some monolinks (supplemental Table 2), which was not observed with other types of cross-linker-modified peptides. This could be due to the high abundance of monolinks, making these forms easier to observe.

Several of our findings show that the BDRG method provides important information about the architecture of TFIIE. Consistent with structural information, Tfa2 residues involved in an intraprotein cross-link (Lys-158–Lys-165) are 11 Å distant in the NMR structure of the conserved first winged helix domain of human TFIIEβ (39). The limited amount of TFIIE structural information, outside of the conserved domains, precludes us from mapping the other identified cross-links onto a structural model of TFIIE. However, our results indicate that two regions of Tfa1 bracketing the zinc-binding domain (residues 90–110 and a region encompassing amino acid 187) lie in close proximity to residues 277–304 at the C terminus of Tfa2. These two regions of Tfa1 overlap perfectly with the regions of human TFIIEα that are essential for interaction with human TFIIEβ (44), strongly suggesting that our cross-linking identified the region of Tfa1 involved in dimerization. Tfa2 residues including Lys-277, Lys-284, Lys-294, and Lys-303 cross-link to Tfa1 Lys-187, and these Tfa2 residues lie adjacent to the Tfa2 second winged helix domain previously implicated in dimerization based on deletion mapping studies of human TFIIEβ (45). We also identified Tfa1 intraprotein cross-links on either side of the Tfa1 zinc-binding domain involving Lys-101–Lys-171 and Lys-110-Lys-187, indicating that these regions are spatially close to each other. Our combined results suggest that the regions on either side of the zinc-binding domain are brought into close proximity by the zinc-binding domain to form the Tfa1 dimerization surface, which binds to a C-terminal region of Tfa2.

The BDRG analysis of TFIIE identified Lys residues involved in monolinks, loop-links, and cross-links (Fig. 4, *black ovals*), indicating that these residues are accessible to the cross-linker in solution. On the other hand, we noticed that there are regions in both Tfa1 and Tfa2 that contain several Lys residues that were not observed in a BDRG-modified form in any

---

B, schematic describing the identification of a BDRG-derived loop-link and MS data that led to the identification of a loop-linked peptide from Tfa2. Loop-linked peptide ions generate extensive fragmentation patterns during MS2 analysis. Database searching of the MS2 spectra led to identification of the indicated BDRG-modified peptide from Tfa2.

TABLE I

Sequences of identified BDRG cross-linked and loop-linked peptides from β-lactoglobulin and the distances between modified lysine residues

The site of cross-linking is indicated by K[468]. The corresponding positions of the modified Lys residues within the protein sequence are shown in parentheses. Prob. indicates the PeptideProphet probability for correct peptide identification.

| Peptide 1	Prob.	Peptide 2	Prob.	Precursor m/z	Precursor z	Theoretical mass	Δ mass (ppm)	Distance (Å)
Cross-links								
R.TPEVDDEALEK[468]FDK.A (135)	0.74	-.LIVTQTMK[468]GLDIQK.V (8)	1	976.2326	4	3900.9054	−1.6	17.03
K.IDALNENK[468]VLVLDTDYKK.Y (91)	1	-.LIVTQTMK[468]GLDIQK.V (8)	1	872.2599	5	4356.2638	−0.8	19.66
K.IPAVFK[468]IDALNENK.V (83)	0.8	R.TPEVDDEALEK[468]FDK.A (135)	1	972.2233	4	3884.8707	−2.3	30.52
K.TK[468]IPAVFK.I (77)	0.98	K.IDALNENK[468]VLVLDTDYK.K (91)	1	1182.2782	3	3543.8212	−2.8	12.88
K.TK[468]IPAVFK.I (77)	0.99	K.ALK[468]ALPMHIR.L (141)	0.96	683.6251	4	2730.4755	−2.4	19.36
K.IIAEK[468]TK.T (75)	0.95	Q.TMK[468]GLDIQK.V (8)	1	629.3294	4	2513.2911	−1.9	19.6
K.TK[468]IPAVFK.I (77)	0.99	-.LIVTQTMK[468]GLDIQK.V (8)	1	793.1834	4	3168.6968	1.7	19.6
-.LIVTQTMK[468]GLDIQK.V (8)	1	K.ALK[468]ALPMHIR.L (141)	0.99	683.9713	5	3414.8231	−1.7	18.38
K.ALK[468]ALPMHIR.L (141)	0.99	R.TPEVDDEALEK[468]FDK.A (135)	0.99	1155.2362	3	3462.6841	0.3	11.47
R.TPEVDDEALEK[468]FDK.A (135)	0.99	K.IIAEK[468]TK.T (75)	0.94	1039.5044	3	3115.4949	−1.7	27.65
K.IDALNENK[468]VLVLDTDYKK.Y (91)	0.97	K.IIAEK[468]TK.T (75)	0.98	893.7199	4	3570.8532	−1.4	17.07
-.LIVTQTMK[468]GLDIQK.V (8)	1	K.IIAEK[468]TK.T (75)	0.99	767.9154	4	3067.6339	−1.2	19.6
Loop-links								
K.IPAVFK[468]IDALNENK[468]VLVLDTDYKK.Y (83–91)	0.99			1142.5986	3	3424.763	2.7	6.42
K.TK[468]IPAVFK[468]IDALNENK.V (77–83)	1			1240.6403	2	2479.246	7.7	10.47
K.FDK[468]ALK[468]ALPMHIR.L (138–141)	1			740.3751	3	2218.107	−2.2	7.26
K.IIAEK[468]TK[468]IPAVFK.I (75–77)	1			1069.0734	2	2136.133	−0.9	12.1
K.IDALNENK[468]VLVLDTDYK[468]K. (91–100)	0.90			924.1252	3	2769.357	−1.8	29.76
R.TPEVDDEALEK[468]FDK[468]ALK.A (135–138)	0.97			876.4116	3	2626.215	−1.4	5.58
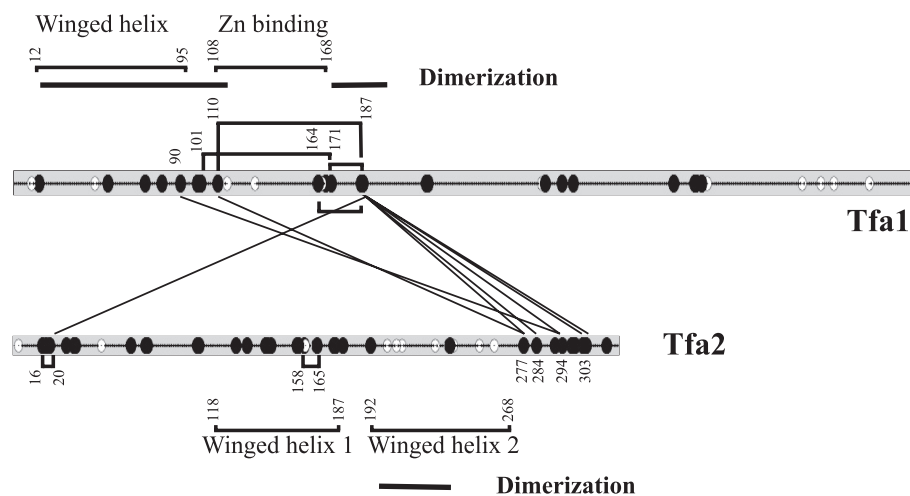
FIG. 4. **Linkage map of cross-linked residues identified from BDRG cross-linking of the heterodimeric complex TFIIE.** *Black lines* connect identified cross-linked lysine residues. *Black ovals* indicate lysine residues that were identified as BDRG-modified residues, suggesting that they are exposed to the cross-linker. *White ovals* indicate lysine residues that were not identified in a BDRG-modified form by MS analysis. The locations of previously described domains are also indicated.

of the identified peptides. The first region of these "clustered unmodified Lys residues" is between residues 194 and 268 (Fig. 4, *white ovals*) of Tfa2. This region corresponds to the dimerization region in human TFIIEβ (45). It is possible that our inability to detect BDRG-modified peptides corresponding to this region is attributable to the limited accessibility of the Lys residues caused by dimerization. The second region of clustered unmodified Lys residues is located at the C terminus of Tfa1. The corresponding region of the human TFIIEα subunit has been shown to interact with both p53 and the p62 subunit of TFIIH (44, 46, 47). Our data suggest that this region of Tfa1 may be inaccessible in the TFIIE complex. Indeed, we identified peptides containing BDRG-modified Lys residues that map to this region as well as the dimerization domain of Tfa2 after denaturing the sample prior to cross-linking (data not shown). A compelling model is that the C-terminal region of Tfa1 is buried in the TFIIE complex, and it becomes exposed upon interacting with Pol II to recruit TFIIH to the preinitiation complex.

These results show that the BDRG approach is effective at detecting inter- and intraprotein cross-links in protein complexes, the cross-links identify residues that are spatially close to one another, and the information is useful for mapping protein complex topology. Identification of clustered cross-linker-modified or unmodified Lys residues can also provide information about the relative accessibility of these residues in individual proteins or in protein complexes.

Architecture of RNA Polymerase II—We next used the BDRG approach to analyze the architecture of the RNA Pol II complex. Pol II is a large complex composed of 12 subunits with a total mass greater than 500 kDa. We isolated yeast Pol II from a strain carrying a 3× FLAG-tagged Rpb3 subunit in a single step using immobilized anti-FLAG antibodies. The complex is ~70–80% pure as estimated by Coomassie-

stained SDS-PAGE and mass spectrometry analysis (supplemental Fig. 4). Approximately 100 μg of Pol II was cross-linked with BDRG final concentration (0.4 mM; cross-linker to protein concentration ratio, 200:1) in 100 μl of HEPES buffer, pH 7.9, and analyzed by LC-MS. We identified six interprotein linkage pairs and 17 intraprotein linkage pairs from 39 total spectral pairs (supplemental Table 3 and Fig. 7). These linkage pairs involve five Pol II subunits. In addition, we identified 160 unique monolinked peptides from 1189 spectra, corresponding to 25 yeast proteins. 138 of these monolinked peptides, from 1150 spectra, corresponded to 10 Pol II subunits. No monolinked peptides were observed for Rpb11 and Rpb12. We also identified 24 unique loop-linked peptides from 84 spectra, which corresponded to eight proteins. Not surprisingly, most of the cross-linker modified peptides map to Rpb1 and 2, which together account for over 90% of the total mass of Pol II and occupy much of the Pol II surface.

Upon mapping the location of the cross-linked resides onto the structure of Pol II (Protein Data Bank code 1i50) (48), it is clear that the identified cross-linked residues are in close proximity to one another (Fig. 5). The cross-links mapped primarily to two regions. One of the regions is centered at the interface consisting of the Rpb1 clamp, Rpb5, and Rpb6. Lys-72 of Rpb6, contained in peptide K.EK[468]AIPK.D, was cross-linked to Lys-15 of Rpb1, contained in peptide R.TVK[468]EVQFGLFSPEEVR.A, and Lys-171 of Rpb5 in peptide R.LK[468]ESQPR.I. The distances between these residues are 15.5 and 10.0 Å, respectively (Fig. 5A). Because there is no structure reported for the N terminus of Rpb6 (from residues 1–69), there could be even closer contacts between Rpb6 and Rpb5 and between Rpb6 and Rpb1 through the N terminus of Rpb6. However, because this region is deficient in lysine residues, it is difficult to detect cross-links corresponding to this region by our approach. The other region is located
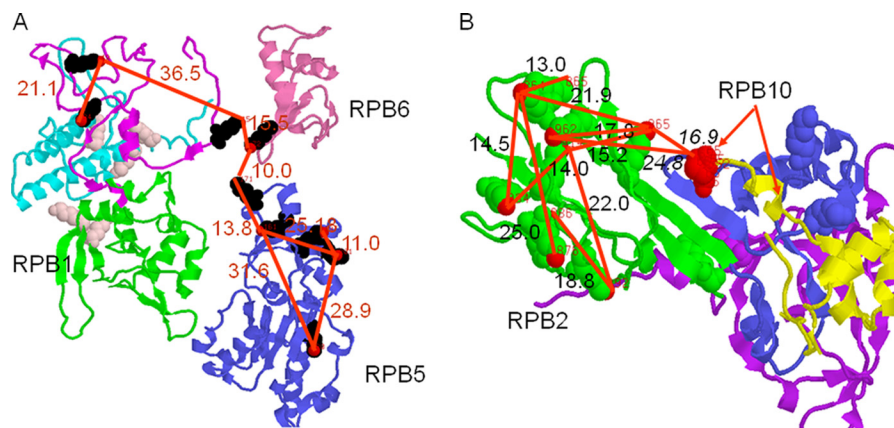
FIG. 5. **Summary of identified BDRG cross-links for RNA polymerase II.** *A*, cross-linking between the Rpb1 clamp, Rpb5, and Rpb6. The Rpb1 regions shown are the clamp core 1 (residues 1–95) in *pink*, the clamp head (residues 96–234) in *green*, and clamp core 2 (residues 235–346) in *blue*. BDRG identified cross-links between lysine residues (shown in *black*) are indicated by *red lines*. Distances (Å) between cross-linked lysine residues are indicated by *red numbers*. *B*, cross-linking between Rpb2 and Rpb10. The Rpb2 regions shown are hybrid-binding region 1 (residues 750–852) in *blue*, the wall region (residues 853–973) in *green*, and the hybrid-binding region 2 (residues 974–1127) in *magenta*. The crystal structure of Pol II lacks the five C-terminal amino acids of Rpb10 ending at Pro-65, spacefilled in *red*. As a result, the cross-linking distances that we present involving Lys-68 are approximations. Distances (Å) between cross-linked lysine residues are indicated by *black numbers*.

at the wall region of Rpb2, where we observed cross-linking between Lys-68 in Rpb10 with Lys-962 and Lys-965 in Rpb2 (Fig. 5*B*). Lys-68 is at the C terminus of Rpb10 contained in the tryptic peptide R.YNPLEK[468]RD.-. Because the crystal structure of Pol II lacks the five C-terminal amino acids of Rpb10 ending at Pro-65, we used the distances between Pro-65 and Lys-965 and Lys-962 of Rpb2 to evaluate the validity of our observed cross-links. The distances between Pro-65 and Lys-965 and Lys-962 of Rpb2 are 16.9 and 24.8 Å, respectively. The results provide further support for the utility of the BDRG approach for identifying residues which are in close proximity in protein complexes.

Outside of these two regions, we identified one cross-linked peptide involving Lys-1102 in Rpb1 in peptide R.LK[468]EILNVAK.N, which resides close to the trigger loop (residues 1070–1090), and Lys-507 in fork loop 2 in Rpb2 in peptide R.DGK[467]LAK.P (48). The free Pol II structure (Protein Data Bank code 1i50) lacks structural information for fork loop 2 between residues 503–508 in Rpb2. Fork loop 2 is structured in the Pol II elongation complex (Protein Data Bank codes 1Y1W and 2E2I) (49, 50). Fork loop 2 plays a direct role in maintaining the transcription bubble by blocking the propagation of noncoding DNA at the transcription register +3 site. In the Pol II elongation complex, Rpb2 Lys-507 and Rpb1 Lys-1102 are separated by a noncoding base, and a direct distance of 12.32 Å (Protein Data Bank code 2E2I) to 13.91 Å (Protein Data Bank code 1Y1W). The fact that the Rpb1 Lys-1102 is cross-linked to Rpb2 Lys-507 and the importance of flexible fork loop 2 in maintaining the transcription bubble led us to investigate the functional importance of Rpb1 Lys-1102. Lys-1102 is close to the trigger loop in conserved region G, and the amino acids flanking Lys-1102 (GVPRL<u>KE</u>) are identical in human and yeast Rpb1. The trigger loop is involved in

substrate selection and polymerase catalysis. However, the role of α helix 37 encompassing Lys-1102 (GVPRL*K*ELIN) has been largely ignored. Upon re-examining the crystal structure of elongating Pol II (Protein Data Bank code 2E2I), we noticed that Lys-1102 directly clashes with the noncoding DNA base at the +3 position (Fig. 6). α helix 37 is parallel to the dissociated noncoding DNA backbone. α helix 37 and the following loop (residues 1107–1114 in Rpb1) form a palm-like structure holding the noncoding DNA, with Lys-1102 and Asn-1110 functioning as two fingers guiding the noncoding DNA (Fig. 6). Lys-1102 also functions to block the noncoding DNA extending toward the bridge helix, which binds to the template DNA and active center. Taking these results together, we postulate that Lys-507 in fork loop 2, Lys-1102 in the α helix 37, and Asn-1110 define an exit path for the noncoding DNA.

DISCUSSION

Chemical cross-linking combined with mass spectrometry provides a particularly attractive method to obtain spatial information on proteins and protein complexes because it is data-rich, sensitive, and fast, but difficulties involving the detection and identification of the informative cross-linked peptides have limited its use. Here we report a strategy to enrich and confidently identify cross-linked peptides using BDRG, a new homo bifunctional cross-linking reagent. Enrichment of cross-linked peptides is achieved via a biotin affinity handle, and confident identification of cross-linked peptides is achieved by MS3 analysis of the modified peptide ions that are generated by fragmentation of BDRG at its Rink bond. Although other MS-labile cross-linkers have been described (11, 13–15, 23, 33), BDRG is the only cross-linker that contains an affinity handle along with a single MS-labile bond. The presence of one labile bond is advantageous because it
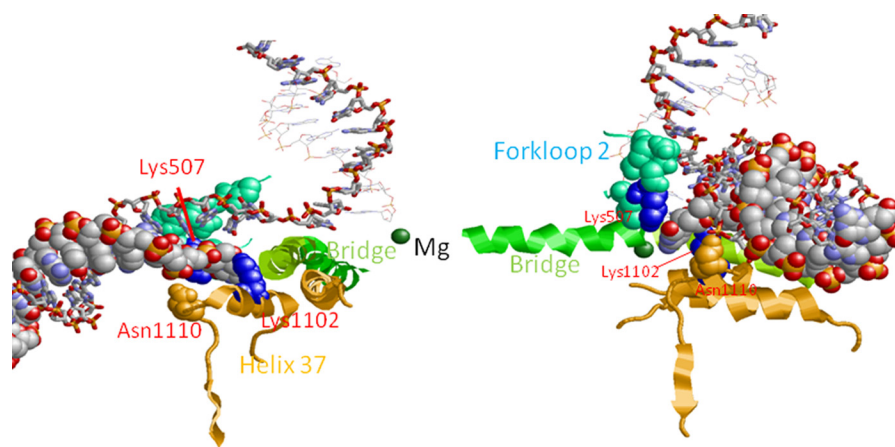
FIG. 6. **Structure of the Pol II elongation complex (Protein Data Bank code 2E2I).** The RNA is shown as a framed structure. The template DNA is shown as a stick structure. The noncoding DNA is shown as a spacefill structure. The two BDRG-identified cross-linked lysine residues, Lys-507 (Rpb2) and Lys-1102 (Rpb1), are spacefilled in *blue*. Forkloop 2 is shown in *cyan*. The bridge helix is *green*. The *green ball* is the Mg ion at the active center.

reduces the number of fragmentation products generated during CID. This facilitates identification of desired product ions by 1) avoiding reduction in their abundance and 2) promoting data-dependent selection of these ions for MS3 analysis. Furthermore, unlike all other MS-labile cross-linkers except for the Rink-based reagents, fragmentation at the labile bond in BDRG generates product ions with nearly equal modification masses. As a result, only one modification mass need be considered during database searching of BDRG-modified peptides. The combination of an inherent enrichment strategy and confident identification of cross-linked peptides makes the BDRG approach attractive for mapping the topology of protein complexes. Indeed, we have successfully applied BDRG to study the architecture of large and small protein complexes, and when possible, we validated the results by mapping the identified cross-links onto the known structures. Importantly, this work represents a major advance in structural characterization of large protein complexes because the successful application of cross-linking/MS to study the topology of partially purified, large complexes has only been described in one other recent report (13). Furthermore, unlike other cross-linking/MS studies of large complexes in which samples were extensively fractionated prior to MS analysis (13, 31), BDRG samples were either analyzed directly or only separated into a few fractions after biotin enrichment.

Like other MS labile cross-linkers, BDRG facilitates acquisition of fragmentation spectra of cross-linked peptides that can be confidently interpreted by standard database search algorithms using a full species-specific protein sequence database. In contrast, interpretation of the complex fragmentation spectra derived from peptides cross-linked with nonlabile reagents requires specialized search algorithms. Furthermore, to improve throughput and confidence in peptide assignment, the search is typically performed against a database restricted to the proteins of interest and the most abundant proteins in the sample (27, 31). However, a restricted data-

base may eliminate the correct interpretation from consideration (if the spectrum is derived from a protein not in the restricted set) and provides less information about the false positive distribution, making it more difficult to separate true positives from false positives. As the purity of the protein complex is relaxed and the sample increases in complexity, the advantages of restricting the database diminish, but the disadvantages remain. For example, MS analysis of the immunopurified Pol II sample used in our studies identified a total of 90 proteins in addition to the 12 Pol II subunits. Many of these non-Pol II proteins are highly expressed proteins that commonly co-purify with target proteins in affinity purification schemes. Even though Pol II subunits are the most abundant proteins in the sample, restriction of the database to Pol II subunits increases the probability that spectra derived from the co-purifying proteins are incorrectly assigned to Pol II subunits.

Whereas the false discovery rate of the Pol II analysis was quite low (<1%), we found there was a high false negative rate. We noticed that short peptides (7–10 amino acids long) were the main source of false negatives. It is likely that the relatively large differential mass modification of 340 Da on Lys that is used in the database search inadvertently biases the search against shorter peptides. This is further complicated by the presence of cross-linker-derived fragment ions in the MS3 spectra that can be misinterpreted by the search algorithm as fragment ions from a different peptide. We found that after initial searches against the whole proteome database, a confined search using a sample-specific database helped to identify some of the false negatives (see "Experimental Procedures"). Strategies to decrease the false negative rate may help to identify more cross-linked peptides.

Because the LTQ was used to measure the *m/z* values of the fragment ions, we used a wide peptide mass tolerance of 3 Da during database searching. We also used a 3-Da window to judge whether the combined masses of the peptides corresponding to the two product ions from potential cross-linked

peptides were similar to the precursor mass during an initial screening process. However, our search results suggest that the combined theoretical masses of cross-linked peptide pairs are within 20 ppm of the precursor mass. Thus, the accurately measured precursor masses provide important additional evidence to evaluate the validity of the BDRG identified cross-linked peptides. Importantly, this strategy is fundamentally different from methods that utilize specialized search algorithms to directly interpret MS2 spectra of cross-linked peptides in which accurate precursor masses are used during the peptide identification process to limit the size of the database to be searched (21, 25, 27). Because the accurate mass measurements are only used during the verification process and not during peptide identification, comparison of the theoretical mass of the cross-linked peptide precursor corresponding to two component peptides identified by database searching of the two MS3 spectra with the measured mass of the observed precursor peptide provides an additional validation step in the BDRG approach that is not available in other approaches.

It is useful to compare our BDRG-based analysis of the Pol II complex, which identified 23 linkage pairs with the analysis by Chen *et al.* (31), which employed the commercially available BS[3] cross-linker and identified 108 high confidence linkage pairs. Chen *et al.* used a highly purified preparation of Pol II, a relatively hydrophilic, sulfo-*N*-hydroxysuccinimide activated cross-linking reagent with an 11.4 Å spacer arm, extensive sample fractionation and MS analysis time, and a cross-link identification strategy that involves the use of in-house-developed software to match MS2 spectra to candidate sequences in a sample-specific database. In comparison, we used a partially purified Pol II preparation containing ~90 co-purifying proteins, a relatively hydrophobic, PFP-activated cross-linker with a 27 Å spacer arm, affinity enrichment followed by minimal fractionation and MS analysis time, and a cross-link identification strategy that involves the use of a common database search algorithm to match MS3 spectra to candidate sequences in a whole proteome database. It is likely that the quantitative and qualitative differences in the cross-links identified in the two studies are due to differences in sample purity, the physicochemical properties of the cross-linkers employed, the extent of sample fractionation and analysis time, and the cross-link identification strategies. It is important to note that most (16 of 23) cross-links observed with BDRG were not among the 108 cross-links observed by Chen *et al*. This indicates that the two methods provide information that is complementary to one another and underscores the utility of developing approaches that employ cross-linkers with a diverse range of reactivities and physicochemical characteristics.

One important difference between the two approaches that warrants special attention is the relative purity of the samples that were analyzed. Although fewer cross-links were identified in our study compared with the study by Chen *et al.*, our sample was approximately eight times more complex than the sample used by Chen *et al.* This is an important point. The ability to confidently identify cross-links derived from samples of high complexity, such as partially purified protein complexes or the many large macromolecular assemblages that carry out essential cellular functions, is a critical unmet need in biology. It is often extremely difficult to purify protein complexes to homogeneity, and, as a result, approaches that can handle the complexity found in partially purified samples are in high demand. Furthermore, because elucidation of global protein-protein interaction networks and protein interactions within large macromolecular assemblages are high priorities in the systems biology community, it is critical to develop cross-linking approaches that are not limited by sample complexity. In general, cross-linking approaches that rely on interpretation of MS2 spectra derived from cross-linked peptides are limited to the analysis of rather simple mixtures (9). The xQuest strategy, which requires the use of isotopically labeled cross-linkers (25), and the Protein Prospector approach (24) are notable exceptions. This limitation is due primarily to the explosion in database search space that occurs when all possible cross-link combinations in a sample are considered, which in turn can hinder the ability of search algorithms to distinguish true positives from false positives. As shown here, the BDRG approach provides a way to confidently identify cross-links derived from samples of high complexity using database search algorithms that are commonly used for peptide identification and represents a step toward meeting this critical need.

In addition to sample purity, differences in the chemical properties of the cross-linkers and the methodologies employed in the two studies likely account for some of the quantitative and qualitative differences in the results. For example, compared with the BS[3] cross-linker used by Chen *et al.*, BDRG is relatively hydrophobic, requiring organic solvent for resuspension. Because the organic concentration is reduced to ~5% during cross-linking reactions, we suspect that the hydrophobicity and/or insolubility of BDRG may limit its cross-linking efficiency. A second issue is related to the use of the biotin group. We have noticed in these studies and in other studies that employ biotin modifying reagents that biotin-modified samples can cause reduced flow during microcapillary chromatography, which we believe is due to insolubility issues. In addition, we observed that some monolinked species were modified by 16 Da, which may be due to oxidation of the biotin moiety. Finally, whereas the specificity of the avidin step is quite high, it is possible that some sample loss occurs during this step. To address these issues, new versions of BDRG are being designed that are more hydrophilic and contain new affinity enrichment handles.

A third issue relates to the effectiveness of our MS3-based identification strategy. Although CID of BDRG cross-linked peptides results primarily in the generation of two BDRG-modified peptide ions, which typically appear as the two most intense ions in MS2 spectra, additional fragment ions may also be

observed, and these ions can reduce the effectiveness of our MS3-based identification strategy. These additional fragment ions may be due to the presence of multiple charge states of the BDRG-modified peptide ions or fragmentation along the peptide backbone. Ions resulting from peptide backbone fragmentation are typically less intense than the desired BDRG-modified ions and therefore will not be chosen for MS3 analysis. Indeed these ions are rarely identified in our searches. This issue may be addressed in the future by optimizing the collision energy for MS2. Additional ions may also arise because of the presence of co-eluting peptides that fall into the CID isolation window. This issue can be further minimized by additional sample complexity reduction steps. Although it is possible that the efficiency of identification may be affected by the presence of additional ions in the MS2 spectra, they do not affect the confidence of cross-linked peptide identification. It is expected that improvements in cross-linker design and advances in MS instrumentation/methods will lead to even more effective BDRG-based cross-linking approaches.

Many of the BDRG cross-linked residues that we identified are located close to unstructured and flexible domains of proteins or protein complexes; for example, the C terminus of the Rpb10, the N terminus of Rpb6, and the forkloop 2 of Rpb2. These examples demonstrate that the BDRG approach can place spatial constraints on distances between unstructured or mobile regions in proteins and complexes. Possible explanations for this behavior could be that the unstructured regions are readily accessible to the cross-linker and/or they form cross-links more efficiently than more rigid regions. Flexible regions impose less stringent requirements on the orientation of the Lys residues to be cross-linked by the cross-linker. Importantly, these regions are often missing from crystal structures. In addition, in the TFIIE study, we found clusters of Lys residues whose lack of observed modifications could indicate their inaccessibility to the cross-linker. Thus, the method presented here can provide structural information that is complementary to information provided by crystal structure studies and is suitable for the study of large, partially purified complexes.

In summary, we have developed a new chemical cross-linking/MS approach that permits the confident identification of cross-linked peptides derived from purified and partially purified protein complexes. Application of the approach to study the architecture of the heterodimeric TFIIE complex and the 12-subunit Pol II complex indicates that the method is effective at identifying regions of proteins that are in close proximity within the complexes. Although the current study did not yield cross-link data that is sufficient to comprehensively map the architecture of Pol II, it did provide a large number of distance constraints that are useful for understanding subunit organization. This is particularly exciting given the dearth of reports on the successful application of MS-based cross-linking approaches for mapping the architecture of large complexes. Furthermore, the identification of cross-links between regions of Pol II that are not resolved in the crystal structure demonstrates the complementarity of this approach to high resolution structural approaches. We expect that improvements in cross-linker design, MS instrumentation, and computational approaches to analyze the data will yield a robust approach for studying the architecture of large complexes in the near future. Integration of the distance constraints obtained from this approach with data from other structural approaches should enable the generation of models of large complexes that will provide significant insights into the mechanistic basis underlying their function.

REFERENCES

1. Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003) From words to literature in structural proteomics. *Nature* **422,** 216–225
2. Wagner, S. D., Kugel, J. F., and Goodrich, J. A. (2010) TFIIF facilitates dissociation of RNA polymerase II from noncoding RNAs that lack a repression domain. *Mol. Cell. Biol.* **30,** 91–97
3. Chen, H. T., Warfield, L., and Hahn, S. (2007) The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat. Struct. Mol. Biol.* **14,** 696–703
4. Hinnerwisch, J., Fenton, W. A., Furtak, K. J., Farr, G. W., and Horwich, A. L. (2005) Loops in the central channel of ClpA chaperone mediate protein binding, unfolding, and translocation. *Cell* **121,** 1029–1041
5. Langosch, D., Thomas, L., and Betz, H. (1988) Conserved quaternary structure of ligand-gated ion channels: The postsynaptic glycine receptor is a pentamer. *Proc. Natl. Acad. Sci. U.S.A.* **85,** 7394–7398
6. Reeves, W. M., and Hahn, S. (2005) Targets of the Gal4 transcription activator in functional transcription complexes. *Mol. Cell. Biol.* **25,** 9092–9102
7. Fishburn, J., Mohibullah, N., and Hahn, S. (2005) Function of a eukaryotic transcription activator during the transcription cycle. *Mol. Cell* **18,** 369–378
8. Burrows, P. C., Severinov, K., Buck, M., and Wigneshweraraj, S. R. (2004) Reorganisation of an RNA polymerase-promoter DNA complex for DNA melting. *EMBO J.* **23,** 4253–4263
9. Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., and Aebersold, R. (2010) Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **9,** 1634–1649
10. Sinz, A. (2006) Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev.* **25,** 663–682
11. Tang, X., Munske, G. R., Siems, W. F., and Bruce, J. E. (2005) Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. *Anal. Chem.* **77,** 311–318
12. Back, J. W., Hartog, A. F., Dekker, H. L., Muijsers, A. O., de Koning, L. J., and de Jong, L. (2001) A new cross-linker for mass spectrometric analysis of the quaternary structure of protein complexes. *J. Am. Soc. Mass Spectrom.* **12,** 222–227
13. Kao, A., Chiu, C. L., Vellucci, D., Yang, Y., Patel, V. R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S. D., and Huang, L. (2011) Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10,**

10.1074/M110.002212

14. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M., and Sinz, A. (2010) Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82,** 6958–6968

15. Lu, Y., Tanasova, M., Borhan, B., and Reid, G. E. (2008) Ionic reagent for controlling the gas-phase fragmentation reactions of cross-linked peptides. *Anal. Chem.* **80,** 9279–9287

16. Petrotchenko, E. V., Olkhovik, V. K., and Borchers, C. H. (2005) Isotopically coded cleavable cross-linker for studying protein-protein interaction and protein complexes. *Mol Cell Proteomics* **4,** 1167–1179

17. Seebacher, J., Mallick, P., Zhang, N., Eddes, J. S., Aebersold, R., and Gelb, M. H. (2006) Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. *J. Proteome Res.* **5,** 2270–2282

18. Schmidt, A., Kalkhof, S., Ihling, C., Cooper, D. M., and Sinz, A. (2005) Mapping protein interfaces by chemical cross-linking and Fourier transform ion cyclotron resonance mass spectrometry: Application to a calmodulin/adenylyl cyclase 8 peptide complex. *Eur. J. Mass Spectrom.* **11,** 525–534

19. Taverner, T., Hall, N. E., O'Hair, R. A., and Simpson, R. J. (2002) Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry. *J. Biol. Chem.* **277,** 46487–46492

20. Back, J. W., Notenboom, V., de Koning, L. J., Muijsers, A. O., Sixma, T. K., de Koster, C. G., and de Jong, L. (2002) Identification of cross-linked peptides for protein interaction studies using mass spectrometry and 18O labeling. *Anal. Chem.* **74,** 4417–4422

21. Zelter, A., Hoopmann, M. R., Vernon, R., Baker, D., MacCoss, M. J., and Davis, T. N. (2010) Isotope signatures allow identification of chemically cross-linked peptides by mass spectrometry: A novel method to determine interresidue distances in protein structures through cross-linking. *J. Proteome Res.* **9,** 3583–3589

22. Sinz, A., Kalkhof, S., and Ihling, C. (2005) Mapping protein interfaces by a trifunctional cross-linker combined with MALDI-TOF and ESI-FTICR mass spectrometry. *J. Am. Soc Mass Spectrom.* **16,** 1921–1931

23. Chowdhury, S. M., Munske, G. R., Tang, X., and Bruce, J. E. (2006) Collisionally activated dissociation and electron capture dissociation of several mass spectrometry-identifiable chemical cross-linkers. *Anal. Chem.* **78,** 8183–8193

24. Chu, F., Baker, P. R., Burlingame, A. L., and Chalkley, R. J. (2010) Finding chimeras: A bioinformatics strategy for identification of cross-linked peptides. *Mol. Cell. Proteomics* **9,** 25–31

25. Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L. N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008) Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **5,** 315–318

26. Lee, Y. J., Lackner, L. L., Nunnari, J. M., and Phinney, B. S. (2007) Shotgun cross-linking analysis for studying quaternary and tertiary protein structures. *J. Proteome Res.* **6,** 3908–3917

27. Maiolica, A., Cittaro, D., Borsotti, D., Sennels, L., Ciferri, C., Tarricone, C., Musacchio, A., and Rappsilber, J. (2007) Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* **6,** 2200–2211

28. Gao, Q., Xue, S., Doneanu, C. E., Shaffer, S. A., Goodlett, D. R., and Nelson, S. D. (2006) Pro-Cross-link: Software tool for protein cross-linking and mass spectrometry. *Anal. Chem.* **78,** 2145–2149

29. Schilling, B., Row, R. H., Gibson, B. W., Guo, X., and Young, M. M. (2003) MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically cross-linked peptides. *J. Am. Soc. Mass Spectrom.* **14,** 834–850

30. Young, M. M., Tang, N., Hempel, J. C., Oshiro, C. M., Taylor, E. W., Kuntz, I. D., Gibson, B. W., and Dollinger, G. (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 5802–5806

31. Chen, Z. A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J. C., Nilges, M., Cramer, P., and Rappsilber, J. (2010) Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29,** 717–726

32. Soderblom, E. J., and Goshe, M. B. (2006) Collision-induced dissociative chemical cross-linking reagents and methodology: Applications to protein structural characterization using tandem mass spectrometry analysis. *Anal. Chem.* **78,** 8059–8068

33. Soderblom, E. J., Bobay, B. G., Cavanagh, J., and Goshe, M. B. (2007) Tandem mass spectrometry acquisition approaches to enhance identification of protein-protein interactions using low-energy collision-induced dissociative chemical cross-linking reagents. *Rapid Commun. Mass Spectrom.* **21,** 3395–3408

34. Herbig, E., Warfield, L., Fish, L., Fishburn, J., Knutson, B. A., Moorefield, B., Pacheco, D., and Hahn, S. (2010) Mechanism of Mediator recruitment by tandem Gcn4 activation domains and three Gal11 activator-binding domains. *Mol. Cell. Biol.* **30,** 2376–2390

35. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392

36. Hermanson, G. T. (2008) *Bioconjugate Techniques*, Academic Press, Orlando, FL

37. Kida, S., Maeda, M., Hojo, K., Eto, Y., Nakagawa, S., and Kawasaki, K. (2007) Studies on heterobifunctional cross-linking reagents, 6-Maleimidohexanoic acid active esters. *Chem. Pharm. Bull.* **55,** 685–687

38. Qin, B. Y., Bewley, M. C., Creamer, L. K., Baker, H. M., Baker, E. N., and Jameson, G. B. (1998) Structural basis of the tanford transition of bovine β-lactoglobulin. *Biochemistry* **37,** 14014–14023

39. Okuda, M., Watanabe, Y., Okamura, H., Hanaoka, F., Ohkuma, Y., and Nishimura, Y. (2000) Structure of the central core domain of TFIIEbeta with a novel double-stranded DNA-binding surface. *EMBO J.* **19,** 1346–1356

40. Geiger, S. R., Lorenzen, K., Schreieck, A., Hanecker, P., Kostrewa, D., Heck, A. J., and Cramer, P. (2010) RNA polymerase I contains a TFIIF-related DNA-binding subcomplex. *Mol. Cell* **39,** 583–594

41. Meinhart, A., Blobel, J., and Cramer, P. (2003) An extended winged helix domain in general transcription factor E/IIE alpha. *J. Biol. Chem.* **278,** 48267–48274

42. Okuda, M., Tanaka, A., Arai, Y., Satoh, M., Okamura, H., Nagadoi, A., Hanaoka, F., Ohkuma, Y., and Nishimura, Y. (2004) A novel zinc finger structure in the large subunit of human general transcription factor TFIIE. *J. Biol. Chem.* **279,** 51395–51403

43. Jawhari, A., Uhring, M., De Carlo, S., Crucifix, C., Tocchini-Valentini, G., Moras, D., Schultz, P., and Poterszman, A. (2006) Structure and oligomeric state of human transcription factor TFIIE. *EMBO Rep.* **7,** 500–505

44. Ohkuma, Y., Hashimoto, S., Wang, C. K., Horikoshi, M., and Roeder, R. G. (1995) Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha. *Mol. Cell. Biol* **15,** 4856–4866

45. Okamoto, T., Yamamoto, S., Watanabe, Y., Ohta, T., Hanaoka, F., Roeder, R. G., and Ohkuma, Y. (1998) Analysis of the role of TFIIE in transcriptional regulation through structure-function studies of the TFIIEb subunit. *J. Biol. Chem.* **273,** 19866–19876

46. Di Lello, P., Miller Jenkins, L. M., Mas, C., Langlois, C., Malitskaya, E., Fradet-Turcotte, A., Archambault, J., Legault, P., and Omichinski, J. G. (2008) p53 and TFIIEalpha share a common binding site on the Tfb1/p62 subunit of TFIIH. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 106–111

47. Okuda, M., Tanaka, A., Satoh, M., Mizuta, S., Takazawa, M., Ohkuma, Y., and Nishimura, Y. (2008) Structural insight into the TFIIE-TFIIH interaction: TFIIE and p53 share the binding region on TFIIH. *EMBO J.* **27,** 1161–1171

48. Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292,** 1863–1876

49. Kettenberger, H., Armache, K. J., and Cramer, P. (2004) Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol. Cell* **16,** 955–965

50. Meyer, P. A., Ye, P., Suh, M. H., Zhang, M., and Fu, J. (2009) Structure of the 12-subunit RNA polymerase II refined with the aid of anomalous diffraction data. *J. Biol. Chem.* **284,** 12933–12939