Identification of cross-linked peptides from complex samples

Bing Yang[1,2,6], Yan-Jie Wu[3,6], Ming Zhu[2,4,6], Sheng-Bo Fan[3,6], Jinzhong Lin[2], Kun Zhang[3], Shuang Li[2], Hao Chi[3], Yu-Xin Li[2], Hai-Feng Chen[3], Shu-Kun Luo[2], Yue-He Ding[2], Le-Heng Wang[3], Zhiqi Hao[5], Li-Yun Xiu[3], She Chen[2], Keqiong Ye[2], Si-Min He[3] & Meng-Qiu Dong[2]

We have developed pLink, software for data analysis of cross-linked proteins coupled with mass-spectrometry analysis. pLink reliably estimates false discovery rate in cross-link identification and is compatible with multiple homo- or hetero-bifunctional cross-linkers. We validated the program with proteins of known structures, and we further tested it on protein complexes, crude immunoprecipitates and whole-cell lysates. We show that it is a robust tool for protein-structure and protein-protein–interaction studies.

Chemical cross-linking of proteins coupled with mass spectrometry analysis (CXMS) can provide valuable information about protein folding and protein-protein interaction[1]. In theory CXMS can help determine the overall architecture of a large protein complex by identifying direct binding partners within the complex and localizing the binding interface. Protein samples need not be highly purified for cross-linking as they must be for crystallography. However, this potential has yet to be realized for several reasons. Protease digestion of cross-linked samples produces highly complex mixtures containing regular, mono-linked, loop-linked and interlinked peptides (**Supplementary Fig. 1**). Interlinked peptides, the most informative category with respect to protein folding and protein-protein interaction, are particularly difficult to detect because they are the least abundant type in the digest[1]. Moreover, the fragmentation spectra of interlinked peptides are too complex for conventional database search algorithms, which consider only one peptide per spectrum. In cross-link spectra, single-cleavage products of two peptides (some fragments remain cross-linked and some do not) are mixed with double-cleavage products on either or both peptides as well as the fragments resulting from linker breakage (**Supplementary Fig. 2**). Although algorithms and experimental strategies have been developed to analyze cross-link spectra, they take into consideration only a subset of the fragment ions or require the use of special cross-linkers that break in a controlled way to release intact peptides[1]. These special cross-linkers are not readily available and have not been tested extensively[2–4]. Other challenges include how to search a large database containing thousands of proteins and how to estimate false discovery rate (FDR). Because cross-linking involves two peptides, the search space increases with the square of the number of candidate sequences (n^2). For example, the cross-link search space for an *Escherichia coli* lysate is 10,000 times larger than the human protein database for linear peptide identification (**Supplementary Table 1**). No algorithm has been able to search a database larger than that of *E. coli* for cross-links. Last, software programs are needed to annotate cross-link spectra for human inspection.

We use a cross-linking method featuring a readily available, amine-specific cross-linker, BS^3 (**Supplementary Fig. 1b**)[5]. Having optimized all the component processes, including sample preparation, HPLC and mass spectrometry (MS) (Online Methods and **Supplementary Figs. 3–5**), we find that higher-energy collisional dissociation (HCD) is the most effective MS method and that the use of a heavy isotope–labeled cross-linker to accompany the light one (such as $[d_0]/[d_4]$-BS^3), which produces characteristic doublet peaks for cross-linked peptides in MS1 spectra (**Supplementary Fig. 1b–g**), is optional.

We designed our software program pLink specifically for the analysis of cross-linked peptide data, including an algorithm to estimate FDR (**Fig. 1**, **Supplementary Figs. 6–15**, **Supplementary Tables 2–6** and **Supplementary Note**). The FDR calculation is based on *in silico* cross-linking of forward (F) and reversed (R) peptide sequences, computed as the number of identified F-R and R-F cross-links subtracted by the number of identified R-R cross-links, then divided by the number of identified F-F cross-links (details in **Supplementary Note**). pLink can identify regular peptides, mono-linked, loop-linked and interlinked peptides in one search against a database as large as the human protein database (>90,000 proteins). At 5% FDR, pLink achieved >95% accuracy, >90% sensitivity and >95% specificity when tested using a large annotated data set (>40,000 spectra) against a large database (≥6,000 forward sequences and ≥6,000 reversed sequences) (**Fig. 1b**). Automated annotation and display of cross-link spectra are realized through pLink itself and another program called pLabel, which facilitates spectrum labeling by users (**Supplementary Fig. 16**). We tested our tools using purified proteins, protein complexes, *in vivo* immunoprecipitates, *E. coli* lysates and *Caenorhabditis elegans* lysates.

[1]College of Biological Sciences, China Agricultural University, Beijing, China. [2]National Institute of Biological Sciences, Beijing, Beijing, China. [3]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [4]College of Life Sciences, Beijing Normal University, Beijing, China. [5]Thermo Fisher Scientific, San Jose, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to M.-Q.D. (dongmengqiu@nibs.ac.cn) or S.-M.H. (smhe@ict.ac.cn).

RECEIVED 12 FEBRUARY; ACCEPTED 28 MAY; PUBLISHED ONLINE 8 JULY 2012; DOI:10.1038/NMETH.2099

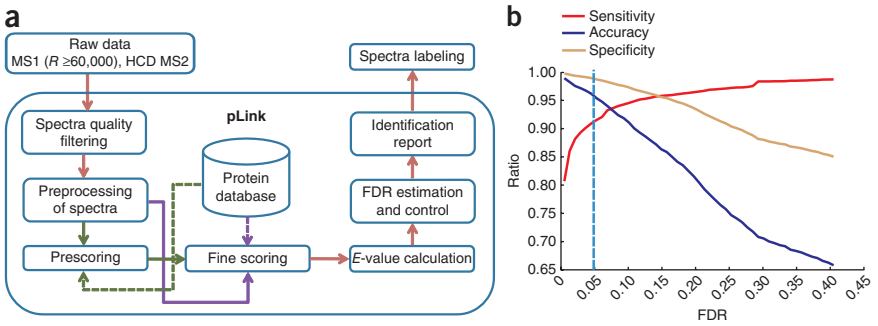

Figure 1 | The pLink program. (**a**) Software schema. (**b**) Performance of pLink in identification of interlinked peptides.

Analysis of a glutathione S-transferase homodimer (GST, ~25 kDa × 2) cross-linked with 1/1 [d_0/d_4]-BS^3 identified eight pairs of cross-links with three or more spectral observations for each (**Fig. 2a**, **Supplementary Fig. 17** and **Supplementary Table 7**). As lysine has a flexible 6-Å side chain and BS^3 has an 11.4-Å spacer arm, the two Cα atoms of cross-linked lysines should be within a 24-Å distance. According to the crystal structure of a GST homodimer, all except one pair have a Cα-Cα distance of less than 24 Å. Most or all of the cross-links appear to be intrasubunit ones, but the Lys124-Lys112 cross-link can be formed between two subunits, too. This result shows that cross-linking typically captures the conformation of the protein faithfully.

We next analyzed a copurified recombinant Cbf5-Nop10-Gar1-Nhp2 complex (CNGP, ~80 kDa) using [d_0]-BS^3, which is the protein part of box H/ACA ribonucleoproteins involved in pseudouridine synthesis. The crystal structure of the Cbf5-Nop10-Gar1 complex has been determined, and the fourth protein Nhp2 was modeled based on its homolog in Archaea[6]. A total of 15 interlinks (179 spectral copies) were identified from two CXMS experiments (**Supplementary Table 8**; those with more than three spectral copies are shown in **Fig. 2b**). Twelve of them are structurally reasonable (**Fig. 2b**). The other three appear to be incompatible with the structural model; they might result from nonspecific cross-linking of native proteins, denatured or aggregated proteins, or alternative protein conformations (**Supplementary Note**).

To estimate the extent of nonspecific cross-linking, we mixed ovalbumin, BSA and a purified F15E11.13-F15E11.14 heterodimer at different ratios while keeping the total protein concentration the same. These three proteins or complexes do not interact with each other, so cross-links identified between them represent nonspecific interactions, whereas cross-links identified within a protein or complex can be either specific or nonspecific. We found only a couple of spectra identifying nonspecific cross-links, compared to hundreds of spectra for within-protein (or complex) cross-links (**Supplementary Fig. 5**). These results suggest that vast majority of identified cross-links are specific and can be used to determine the three-dimensional structures of proteins or protein interactions.

To probe unknown protein-complex structures, we analyzed the yeast UTP-B complex. UTP-B is a six-protein complex (550 kDa, assuming one copy each of Utp1, Utp6, Utp18, Utp12, Utp13 and Utp21) essential for maturation of the small-unit ribosome[7–9]. No structure is available for UTP-B or any of its components. Except for Utp6, the other five proteins contain one or two predicted WD domains. We detected 71 high-quality interlinked

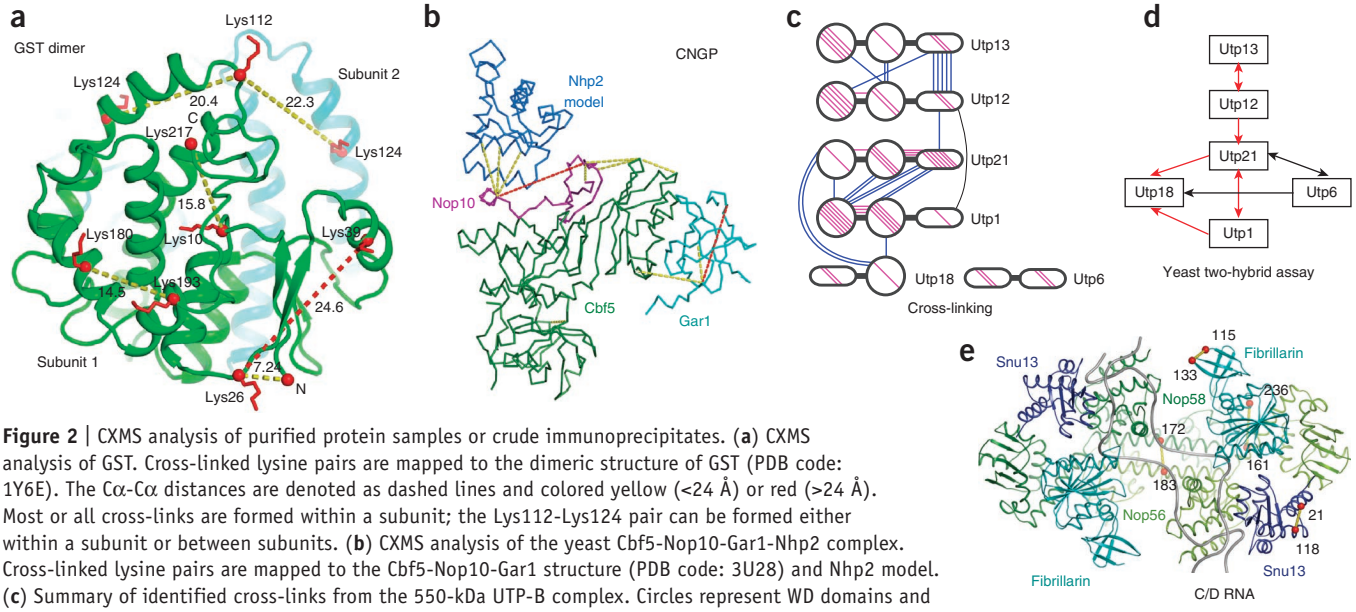

Figure 2 | CXMS analysis of purified protein samples or crude immunoprecipitates. (**a**) CXMS analysis of GST. Cross-linked lysine pairs are mapped to the dimeric structure of GST (PDB code: 1Y6E). The Cα-Cα distances are denoted as dashed lines and colored yellow (<24 Å) or red (>24 Å). Most or all cross-links are formed within a subunit; the Lys112-Lys124 pair can be formed either within a subunit or between subunits. (**b**) CXMS analysis of the yeast Cbf5-Nop10-Gar1-Nhp2 complex. Cross-linked lysine pairs are mapped to the Cbf5-Nop10-Gar1 structure (PDB code: 3U28) and Nhp2 model. (**c**) Summary of identified cross-links from the 550-kDa UTP-B complex. Circles represent WD domains and ovals represent other domains found in UTP-B subunits. Cross-links are denoted by lines. Intramolecular (magenta) and intermolecular cross-links (blue) are colored if they are consistent with yeast two-hybrid (Y2H) interactions but are black if they are not. (**d**) Y2H results between the UTP-B subunits. Y2H interactions from bait to prey are indicated by arrows and colored in red if detected in CXMS analysis, or in black if not. (**e**) CXMS analysis of the FIB-1 complex immunoprecipitated from *C. elegans*. Four identified cross-linked Lys pairs are mapped to an archaeal C/D RNA-protein complex structure (PDB code: 3PLA) and denoted by yellow dashed lines and Cα spheres. Nop56 and Nop58 are assumed to form a heterodimer. The residue numbers of *C. elegans* proteins are indicated.

lysine pairs (1,337 spectral copies), of which 50 are intramolecular and 21 intermolecular (**Fig. 2c** and **Supplementary Table 9**). The intramolecular cross-links occur mainly within the predicted individual domains. A few cross-links occur between adjacent domains in Utp12, Utp21 and Utp1, suggestive of intramolecular domain-domain contacts. The protein-protein interaction map of UTP-B deduced from our data is mostly consistent with that obtained previously by yeast two-hybrid (Y2H) assays[10], indicating that the cross-linking data primarily report direct protein interactions (**Fig. 2c**,**d**). We also observed one cross-link between Utp1 and Utp12, and this interaction was not detected by Y2H. Large numbers of cross-links between Utp13 and Utp12 and between Utp21 and Utp1 suggest extensive binding interfaces. Collectively, these data provide valuable distance constraints at the residue level to model the structure of individual proteins and the complex.

To find out whether we could combine CXMS with immunoprecipitation to detect direct interactions, we immunoprecipitated GFP-tagged *C. elegans* fibrillarin (FIB-1)[11] and treated the immunoprecipitate with BS^3. The methyltransferase fibrillarin is a component of C/D RNA-protein complexes (RNPs) responsible for site-specific ribose-2′-*O*-methylation of RNA. In eukaryotes, a C/D RNP consists of four proteins—fibrillarin, Nop56, Nop58 and Snu13—and a C/D RNA. We identified all four proteins, two intramolecular cross-links (in Snu13 and fibrillarin) and two intermolecular cross-links (fibrillarin-Nop56 and Nop56-Nop58) (**Fig. 2e** and **Supplementary Table 10**). All four cross-links are consistent with a structural model of worm C/D RNP constructed from the crystal structure of an archaeal counterpart[12,13]. The eukaryotic Nop56 and Nop58 proteins have long been suspected to form a heterodimer[12]. Our data provide *in vivo* evidence to support this prediction.

To our knowledge, cross-link identification from cell lysates has been attempted only twice before, both times with *E. coli*[14,15]. The most comprehensive CXMS analysis carried out so far yielded 71 *E. coli* interlinks from two experiments[14]. We identified 394 interlinks from BS^3-treated *E. coli* lysates. Structural information is available from the Worldwide Protein Data Bank (PDB, http://www.wwpdb.org/) to evaluate 208 cross-links; 178 (85.6%) are compatible with the structures of corresponding proteins and complexes in the PDB (**Supplementary Fig. 18a**,**b** and **Supplementary Table 11**). Protein-protein interactions represented by 124 interlinks included known interactions (10 from PDB; 5 from bacteriome.org[16]) and previously unreported ones (**Supplementary Table 11**). Five out of eight novel interactions were tested positive by Y2H (**Supplementary Fig. 18c**).

From an even more complex sample, whole *C. elegans* lysates, we identified 39 interlinked peptides (5% FDR) by searching the entire *C. elegans* protein database (~25,000 proteins) (**Supplementary Fig. 18d** and **Supplementary Table 12**). The reduction in cross-link identification is likely due to the large search space.

pLink also works with cross-linkers besides BS^3, including DSS (amine-amine), EDC (amine-carboxyl), AMAS and sulfo-GMBS (amine-sulfhydryl) (**Supplementary Tables 13–16**). A careful comparison of pLink and xQuest[14] showed a significant improvement of cross-link identification by pLink (**Supplementary Tables 17–20** and **Supplementary Note**).

pLink and pLabel are available from http://pfind.ict.ac.cn/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors wish to express gratitude to A.F. Hühmer and D. Horn (Thermo Fisher Scientific) for discussion; L.-L. Du (National Institute of Biological Sciences, Beijing (NIBS)) for critical suggestions; S.J. Lo (Chang Gung University) for the *fib-1 ::GFP* strain; E.-Z. Shen, H.-Q. Wang, X.-D. He and G.-H. Cai (NIBS) for help with microscopy, GFP immunoprecipitation and peptide experiments; and R. Aebersold and T. Walzthöni (ETH Zurich) for help with xQuest search. We thank Z. Yuan, C. Liu, R.-X. Sun, Y. Fu and other members of the pFind team for discussion and support. This work was funded by the Ministry of Science and Technology of China (863 grant 2007AA02Z1A7 and 973 grant 2010CB835203 to M.-Q.D.; 863 grant 2008AA022310 and 973 grant 2010CB835402 to K.Y.; 973 grants 2012CB910602 and 2010CB912701 to S.-M.H.), the CAS Knowledge Innovation Program (KGCX1-YW-13) and an ICT basic research grant to S.-M.H. We thank the municipal government of Beijing for funds allocated to NIBS.

AUTHOR CONTRIBUTIONS

B.Y. performed most wet-lab experiments and data analysis, and contributed to manuscript preparation; Y.-J.W. developed pLink; M.Z. performed peptide experiments (with B.Y.), non-BS^3 cross-linking experiments and Y2H experiments; S.-B.F. fixed software bugs and maintained the system; J.L., S.L. and S.-K.L. purified the UTP-B, CNGP and F15E11.13-F15E11.14 protein complexes, respectively; K.Z. and L.-Y.X. developed pLabel for cross-links; H.C., H.-F.C. and L.-H.W. contributed to pLink development; Y.-X.L. contributed to data analysis; Y.-H.D. contributed to non-BS^3 cross-linking experiments; Z.H. and S.C. contributed to MS analysis; K.Y. provided crucial samples and contributed to data interpretation and manuscript preparation; S.-M.H. directed software development and revised the manuscript; M.-Q.D. conceived the study, directed wet-lab experiments and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2099. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Singh, P., Panchaud, A. & Goodlett, D.R. *Anal. Chem.* **82**, 2636–2642 (2010).
2. Zhang, H. *et al. Mol. Cell. Proteomics* **8**, 409–420 (2009).
3. Petrotchenko, E.V., Serpa, J.J. & Borchers, C.H. *Mol. Cell. Proteomics* **10**, M110.001420 (2011).
4. Kao, A. *et al. Mol. Cell. Proteomics* **10**, M110.002212 (2011).
5. Hermanson, G.T. *Bioconjugate Techniques* 2nd edn Ch. 4, 241–243 (Academic, London, 2008).
6. Li, S. *et al. Genes Dev.* **25**, 2409–2421 (2011).
7. Dragon, F. *et al. Nature* **417**, 967–970 (2002).
8. Grandi, P. *et al. Mol. Cell* **10**, 105–115 (2002).
9. Krogan, N.J. *et al. Mol. Cell* **13**, 225–239 (2004).
10. Champion, E.A., Lane, B.H., Jackrel, M.E., Regan, L. & Baserga, S.J. *Mol. Cell. Biol.* **28**, 6547–6556 (2008).
11. Lee, L.W., Lo, H.W. & Lo, S.J. *Gene* **455**, 16–21 (2010).
12. Aittaleb, M. *et al. Nat. Struct. Biol.* **10**, 256–263 (2003).
13. Lin, J. *et al. Nature* **469**, 559–563 (2011).
14. Rinner, O. *et al. Nat. Methods* **5**, 315–318 (2008).
15. Xu, H., Hsu, P.H., Zhang, L., Tsai, M.D. & Freitas, M.A. *J. Proteome Res.* **9**, 3384–3393 (2010).
16. Su, C. *et al. Nucleic Acids Res.* **36**, D632–D636 (2008).

ONLINE METHODS

Chemicals. Acetonitrile, formic acid, ammonium bicarbonate and ammonium acetate were purchased from J.T. Baker. [d_0]- and [d_4]-bis(sulfosuccinimidyl) suberate (BS^3), disuccinimidyl suberate (DSS), 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC), sulfo-NHS, *N*-(α-maleimidoacetoxy) succinimide ester (AMAS) and *N*-(γ-maleimidobutyryl-oxy) sulfosuccinimide ester (sulfo-GMBS) were from Pierce. Dimethylsulfoxide (DMSO), HEPES, MES and other general chemicals were from Sigma-Aldrich.

Purified proteins. Ovalbumin and BSA were purchased from Sigma-Aldrich. Recombinant GST was purified using glutathione beads (GE Healthcare). *C. elegans* F15E11.13-F15E11.14 heterodimer was coexpressed in *E. coli* BL21(DE3) with an N-terminal 6×His-SMT3 or SMT3 tag, copurified on HisTrap column and further purified by gel filtration after cleavage of the SMT3 tag by protease Ulp1. The yeast H/ACA Cbf5-Nop10-Gar1-Nhp2 (CNGP) complex was prepared as previously described[6]. UTP-B was expressed by using the Bac-to-Bac baculovirus expression system (Invitrogen). Utp18 was fused to a C-terminal 6×His tag, and the other proteins were untagged. The UTP-B complex was purified through HisTrap, anion exchange and gel filtration chromatography (GE Healthcare). A single-chain anti-GFP antibody (nanobody) was produced as described before[17]. We cross-linked purified nanobody to NHS-activated Sepharose 4 Fast Flow beads (GE Healthcare) following the manufacturer's instruction.

Cross-linking and LC-MS analysis of synthetic peptides. A total of 38 peptides (**Supplementary Table 2**) were synthesized at GL Biochem. The sequences were based on preliminary CXMS results of the UTP-B protein complex using a database search strategy as described in ref. 18. Each peptide was dissolved in 20 mM HEPES, pH 7.5, to 5 mM concentration. Equal amounts of [d_0]- and [d_4]-BS^3 were dissolved in DMSO to 10 mM concentration. Peptides containing cysteine were alkylated with iodoacetamide. For two-peptide cross-linking, 4 µl of each peptide and 1.5 µl of cross-linker were incubated at room temperature (RT) for 2 h and terminated with 0.5 µl of 400 mM ammonium bicarbonate at RT for 20 min. BS^3 was chosen because it has been extensively tested and is known to have good specificity and reactivity, and this cross-linker has a stable isotope-labeled version readily available. Cross-linking with a 1:1 mixture of [d_0]- and [d_4]-BS^3 produces cross-linked peptides that appear in full MS scans as two approximately co-eluting peaks of equal intensity and that are 4.0251-Da apart (**Supplementary Fig. 1b–g**). This feature helps confirm the attachment of the cross-linker to a peptide.

After a 1,000-fold dilution, 5 µl of each cross-linking reaction (741 in all, encompassing all possible two-peptide combinations) was analyzed on LTQ-Orbitrap-ETD (Thermo-Fisher Scientific) coupled to an Agilent 1200 HPLC pump. Cross-linking isoforms can be separated (**Supplementary Fig. 4**). Each run took 35 or 37 min. Each full scan (R = 100,000) was followed by three sets of HCD-CID-ETD triple-play MS2 scans. Singly and doubly charged precursors were excluded.

CXMS analysis of purified proteins or protein complexes. GST: the cross-linking condition was optimized (**Supplementary Fig. 3**). In a 20-µl reaction, 10 µM purified GST (25 kDa, in 50 mM HEPES, pH 7.5, 100 mM KCl) was cross-linked at RT for 1 h with a 1:1 mix of [d_0]- and [d_4]-BS^3 at 0.5 mM (0.25 mM each), which corresponded to a 1:50 protein:cross-linker molar ratio or roughly 1 µg BS^3/µg protein. Higher cross-linker concentration and longer reaction time were avoided to minimize over–cross-linking artifacts. The reaction was terminated at RT with 20 mM ammonium bicarbonate for 20 min. For the cross-linking experiment in **Supplementary Tables 19** and **20**, only [d_0]-BS^3 was used.

CNGP: in a 20 µl reaction, 10 µM CNGP protein complex (78 kDa, in 50 mM HEPES, pH 8.0, 500 mM NaCl) and 1 mM [d_0]-BS^3 were incubated at RT for 1 h. The reaction was terminated and digested as above.

UTP-B: reaction conditions were similar to those of GST but with 1 mM BS^3-[d_0] and 1 mM BS^3-[d_4].

To assess nonspecific cross-linking, ovalbumin, BSA and F15E11.13-F15E11.14 (referred to as the BOF mix) were mixed at 1:18:1, 4:15:1, 9.5:9.5:1, 15:4:1 or 18:1:1 while the total protein concentration was kept constant at 20 µg/µl in a 20 µl reaction in 50 mM HEPES, pH 8.3, 150 mM NaCl. Each sample was cross-linked with 200 µg [d_0]-BS^3 and 200 µg [d_4]-BS^3 at RT for 1 h. The reaction was stopped by 100 mM ammonium bicarbonate.

DSS, AMAS or sulfo-GMBS cross-linking reactions: 10 µM CNGP complex was cross-linked with 1 mM DSS, AMAS or Sulfo-GMBS in 50 mM HEPES, pH 8.0, 500 mM NaCl (10 µl) at 25 °C for 1 h before it was quenched with 20 mM ammonium bicarbonate for 20 min.

EDC cross-linking reactions: 1 mM EDC and 10 µM CNGP complex in 100 mM MES, pH 6.0, 500 mM NaCl (10 µl) were incubated at 25 °C for 15 min before the pH was adjusted to 7.2 with 0.5 µl 1 M HEPES, pH 8.3. Then sulfo-NHS was added to 2.5 mM. The reaction was stopped 2 h later with 0.4 µl 500 mM Tris, pH 8.5, for 20 min.

CXMS analysis of FIB-1 IP from *C. elegans*. A transgenic *C. elegans* strain expressing GFP-tagged *fib-1* under the *fib-1* promoter[11] was cultured following standard protocols[19]. Worms were washed three times with Lysis Buffer 1 (75 mM HEPES, pH 7.5, 1.5 mM EGTA, 1.5 mM $MgCl_2$, 150 mM KCl, 15% glycerol, 0.075% NP-40, 1 mM PMSF, 1× EDTA-free Complete Protease Inhibitor Cocktail (CPIC, Roche)) and transferred to two 1.5-ml screw-capped centrifuge tubes (150 µl packed worms per tube). After the addition of 150 µl Lysis Buffer 1 and 450 µl prechilled glass beads, worms were homogenized using the FastPrep system (MP Biomedicals) at 6.5 m/s, 30 s per pulse for three pulses, with 5 min of cooling on ice between pulses. Homogenates were cleared by centrifugation at 4 °C at 14,000 r.p.m. for 30 min. Before IP, 20 µl nanobody beads were precleaned with 1 ml of 0.1 M Glycine, pH 2.6, three times, and equilibrated with Wash Buffer 1 (50 mM HEPES, pH 7.5, 1 mM EGTA, 1 mM $MgCl_2$, 100 mM KCl, 10% glycerol, 0.05% NP-40, 1 mM PMSF, 1× CPIC). We incubated 20 µl nanobody beads with 400 µl worm lysate at 4 °C for 2 h, then we washed them twice with 50 mM HEPES, pH 7.5, 100 mM KCl, 1× CPIC. The supernatant was removed as much as possible without disturbing the beads. Then 50 µg [d_0]-BS^3 was added and allowed to react for 2 h at RT. The beads were washed three times with Wash Buffer 1 containing 300 mM KCl and another two times using Wash Buffer 1 without protein inhibitor. Immunoprecipitated proteins were digested on beads with trypsin.

Cross-linking of *E. coli* and *C. elegans* lysates. *E. coli* OP50 cell lysates (~3 mg proteins in 75 mM HEPES, pH 8.3, 150 mM NaCl, 1 mM PMSF, 1× CPIC) were digested with 2 units of DNase I (New England Biolabs) and 10 µg RNase A (TIANGEN) at RT for 30 min and filtrated using an Amicon Ultra-4 10k unit. The final dilution factor for small molecules was 125-fold. For cross-linking, 20 µl lysate was incubated at RT with 200 µg $[d_0]$-BS^3 (experiment 1) or 200 µg $[d_0]$/$[d_4]$-BS^3 (experiment 2) for 2 h before ammonium bicarbonate was added to 100 mM to stop the reaction. The *C. elegans* lysates were prepared in the same way using $[d_0]$-BS^3.

Trypsin digestion. All samples except FIB-1 IP were precipitated by four volumes of acetone at −20 °C for 30 min. Precipitated proteins were dried in air and resuspended in 8 M urea, 100 mM Tris, pH 8.5. After reduction with 5 mM TCEP for 20 min and alkylation with 10 mM iodoacetamide for 15 min in the dark, samples were diluted to 2 M urea with 100 mM Tris, pH 8.5, and digested with trypsin (at 50:1 protein:enzyme ratio) at 37 °C for 16 h in the presence of 1 mM $CaCl_2$ and 20 mM methylamine. Digestion was stopped by adding formic acid to 5% final concentration. The FIB-1 immunoprecipitate was digested on beads by resuspending 20 µl beads in 50 µl 8 M urea, 100 mM Tris, pH 8.5, and processed as above.

Fractionation of *E. coli* and *C. elegans* peptides by strong cation exchange. The tryptic digest of cross-linked *E. coli* or *C. elegans* lysate was off-line fractionated using a 250 µm (ID) two-phase column. This column contained a 2-cm-long reverse phase (RP) section (3 µm, 125 Å, Luna C18 resin from Phenomenex) upstream of a 2-cm-long Strong Cation Exchange (SCX) section 5 µm, 120 Å SCX resin from Whatman) and a frit at the end.

Peptides digested from 50 µg proteins were directly loaded onto the column. After desalting with Buffer A (5% ACN, 0.1% FA), peptides were eluted from RP to SCX resin with Buffer B (80% ACN, 0.1% FA). Five SCX fractions were collected (eluted with 5 µl 0.10, 0.25, 0.50, 1.0, 1.0 M ammonium acetate), and one-fifth of each was analyzed.

Mass spectrometry analysis and pLink search. Two types of column setup—high-flow and low-flow—were used, each consisting of an analytical column with a pulled tip and a precolumn with a frit. A high-flow (~200 nl/min) analytic column is 8 cm long, 100 µm ID packed with 3 µm, 125 Å Luna C18 resin, and the precolumn was 2 cm long and 250 µm ID, with 10 µm, 90 Å Jupiter C12 resin (Phenomenex). A low-flow (~30 nl/min) analytic column is a 10 cm long, 50 µm ID column packed with 3 µm, 125 Å Luna C18 resin, and the precolumn was 8 cm long, 75 µm ID and packed with 10 µm, C18 resin (Yamamura Chemical Research Institute).

The GST, FIB-1 IP and BOF mix samples were analyzed using high-flow columns; the *E. coli* and *C. elegans* SCX fractions were analyzed using low-flow columns; the CNGP and UTP-B samples were analyzed using both high- and low-flow columns. From parallel comparisons, we found that, with respect to the number of cross-link identifications, high- and low-flow columns give similar results, but signal intensity is higher using low-flow columns. Because low-flow columns tend to be clogged more often, we recommend that they be used when the amount of sample is extremely low. The high-flow column setup is recommended for most situations.

An Agilent 1200 quaternary pump was interfaced with either LTQ-Orbitrap-ETD or LTQ-orbitrap Velos. For a typical gradient: at 100 or 200 µl/min, a 70- or 120-min run starts with a 45-min or 90-min linear gradient from 100% buffer A (0.2 mM acetic acid in water) to 35% buffer B (70% ACN, 0.2 mM acetic acid) then continues with a 15-min gradient from to 35% to 100% buffer B. This is followed by a 5-min isocratic flow of 100% buffer B and a 5-min gradient from 100% to 0% buffer B and then a final wash with 10-min or 15-min buffer A. The output flow was split using a microTee to ~200 nl/min through the column.

On LTQ-Orbitrap Velos, each full scan (R = 100,000) was followed by ten HCD scans at R = 7,500 and NCE = 45; +1, +2 and unknown-charge-state ions were excluded from MS2 scans; monoisotopic screening was disabled. Dynamic exclusion repeat count, exclusion list and exclusion duration were set to 1, 50 and 60 s (or 1, 200 and 150 s for *E. coli* experiment #2). Minimal signal threshold for MS2 was 5,000.

On LTQ-Orbitrap-ETD, each full scan (R = 100,000) was followed by 5 HCD scans at R = 7,500 and NCE = 40; precursors of +1, +2 or unknown charge state were excluded; monoisotopic screening was disabled. Dynamic-exclusion repeat count, repeat duration, exclusion list and exclusion duration were set to 2, 30, 200 and 30 s for GST, FIB-1 IP and UTP-B or 1, 30, 200 and 60 s for *E. coli* cross-link experiment #1 *and C. elegans* samples. Minimal signal threshold for MS2 was 5,000. The MS method for CNGP was the same as that for FIB-1 IP except that NCE was set to 45.

For cross-link identification, exclusion of +1 and +2 precursors is crucial. Sometimes further exclusion of +3 precursors can lead to more cross-link identifications. Notably, enabling monoisotopic screening invariably resulted in a reduction in cross-link identification (data not shown). For samples cross-linked with a 1:1 mix of $[d_0]$- and $[d_4]$-BS^3, we examined whether triggering MS2 data acquisition with specified mass tags (mass delta = 4.0251, partner intensity range = 50–100%, MS2 on both partners) would help with cross-link identification. We found that it did not, with or without exclusion of +2 precursors. Isotopic data-dependent MS2 did not help with identification either (data not shown).

pLink search parameters: precursor mass tolerance 50 p.p.m., fragment mass tolerance 20 p.p.m., cross-linker $[d_0]$-BS^3 or $[d_0]$/$[d_0]$-BS^3 (cross-linking sites K and protein N terminus, isotope shift 4.0247 Da, xlink mass-shift 138.0680796, monolink mass-shift 156.0786442), fixed modification C 57.02146, peptide length mininum 4 amino acids and maximum 100 amino acids per chain, peptide mass minimum 400 and maximum 10,000 Da per chain, enzyme trypsin, two missed cleavage sites per chain (four per cross-link). The wormpep216 protein database was used for *C. elegans* lysates. The *E. coli* protein sequences for the K-12 substrain MG1655 downloaded from NCBI on 2011-12-01 were used for *E. coli* lysates. The *E. coli* protein sequences for the strain OP50 downloaded from NCBI on 2010-06-28 were used for performance tests of pLink. Other protein sequences (such as GST, BSA, CNGP and UTP-B) were either downloaded from NCBI or provided by researchers who made the recombinant proteins.

Verification of protein-protein interactions by yeast two-hybrid (Y2H) assays. To verify novel protein-protein interactions

doi:10.1038/nmeth.2099

suggested by interlinked peptides from *E. coli* lysates (**Supplementary Table 11**), we selected ten pairs more or less randomly and tested them with Y2H assays. Each selected ORF was amplified by PCR and cloned into the Clontech vectors pGBKT7, a Gal4 DNA-binding domain–based bait vector (BD) and pGADT7, a Gal4 activation domain–based prey vector (AD). All the constructs were verified by sequencing. Cloning was successful for eight test pairs: (i) #52, AD-AAC73576.1 + BD-NP_415483.2 and BD-AAC73576.1 + AD-NP_415483.2 (AAC73576.1 is adenylate kinase and NP_415483.2 is methylglyoxal synthase); (ii) #69, BD-AAA58136.1 + AD-YP_025307.1 (AAA58136.1 is translation elongation factor EF-Tu and AD-YP_025307.1 is multidrug efflux system transporter); (iii) #70, BD-AAA58136.1 + AD-AAC74522.1 (AAA58136.1 is translation elongation factor EF-Tu and AAC74522.1 is polyhydroxybutyrate (PHB) synthase, an ABC transporter periplasmic binding protein homolog); (iv) #71, BD-AAA58136.1 + AD-YP_026243.1 (AAA58136.1 is translation elongation factor EF-Tu and YP_026243.1 is predicted von Willebrand factor containing protein); (v) #91, AD-AAA97042.1 + BD-NP_416801.2 (AAA97042.1 is GroEL protein and NP_416801.2 is predicted inner membrane protein); (vi) #98, AD-AAC73200.1 + BD-AAC75219.1 (AAC73200.1 is lipid II flippase, an integral membrane protein involved in stabilizing FstZ ring during cell division and AAC75219.1 is inner membrane protein, UPF0324 family); (vii) #110, AD-AAA69093.1 + BD-AAC74589.1 (AAA69093.1 is phosphoglycerate kinase and AAC74589.1 is autoinducer 2-binding protein); (viii) #115, AD-NP_416518.2 + BD-AAC73708.1 (NP_416518.2 is low-affinity putrescine importer and AAC73708.1 is universal stress protein UP12); (ix) positive control, AD-SV40 large T antigen + BD-p53; and (x) negative control, AD-NP_416518.2 + BD-p53.

Each pair of bait and prey vectors was cotransfected into two-hybrid strain AH109 bearing *GAL1::HIS3* and *GAL2::ADE2* reporter genes. Each transformant was cultured on plates without leucine and tryptophan (SD –LW). After 2 d, all of the plates were replica-plated onto SD –LW plates and three other types of plates selecting for positive interactions: SD –LWH (no leucine, tryptophan or histidine), SD –LWHA (no leucine, tryptophan, histidine or adenine); and SD –LWH3AT plates (SD –LWH supplemented with 2.5 mM 3-amino-1,2,4-triazole). Colonies were selected for activation of the *HIS3* reporter gene. Protein-protein interactions suggested by interlink numbers 69, 70, 71, 98 and 115 were tested positive on SD –LWH plates. Y2H results are indicated above in parentheses.

17. Kubala, M.H., Kovtun, O., Alexandrov, K. & Collins, B.M. *Protein Sci.* **19**, 2389–2401 (2010).
18. Panchaud, A., Singh, P., Shaffer, S.A. & Goodlett, D.R. *J. Proteome Res.* **9**, 2508–2515 (2010).
19. Brenner, S. *Genetics* **77**, 71–94 (1974).