Multiplexed Proteomics Mapping of Yeast RNA Polymerase II and III Allows Near-Complete Sequence Coverage and Reveals Several Novel Phosphorylation Sites

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2008

Impact Factor: 5.64 · DOI: 10.1021/ac7024283 · Source: PubMed

Anal. Chem. **2008,** *80,* 3584–3592

Multiplexed Proteomics Mapping of Yeast RNA Polymerase II and III Allows Near-Complete Sequence Coverage and Reveals Several Novel Phosphorylation Sites

Shabaz Mohammed,[†] Kristina Lorenzen,[†] Robert Kerkhoven,[†] Bas van Breukelen,[†] Alessandro Vannini,[‡] Patrick Cramer,[‡] and Albert J. R. Heck*[,†]

Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Chemistry, Utrecht University, Sorbonnelaan 16, 3584 CA Utrecht, The Netherlands, and Gene Center Munich and Center for Integrated Protein Science CiPS[M], Department of Chemistry and Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany

The multisubunit RNA polymerases (Pols) II and III synthesize mainly eukaryotic mRNAs and tRNAs, respectively. Pol II and Pol III are protein complexes consisting of 12 and 17 subunits. Here we analyzed both yeast Pol II and Pol III by multiplexed mass spectrometric analysis using various proteases and both collision induced and electron transfer dissociation. The cumulative data obtained from using the various proteases (trypsin, chymotrypsin, Glu-C and Lys-C) and the two peptide fragmentation approaches allowed us to map nearly the complete sequences of all constituents of both Pol II and III. Notably, chymotrypsin behaved equally well as and in certain circumstances better than trypsin in the context of protein coverage. Although the available high resolution structures have exposed extensive mechanistic insights into transcription, the role of post-translational modification in these processes has been addressed to a lesser extent. In our analysis of Pol II and III we detected 19 phosphorylation sites, of which 12 have not been previously reported. Identified phosphosites were mapped on the Pol II structure which provided indications that they might play a role in regulating the conformation of the clamp region and, as a consequence, interaction of Pol II with nucleic acids. The described multiplexed proteomics approach is generic and reveals that it is possible to map a protein complex to near completion while applying less than 5 μg (approximately 10 pmol) of total starting material.

The multisubunit RNA polymerases (Pols) I, II, and III synthesize eukaryotic RNA during gene transcription. Pol I and Pol II synthesize ribosomal and mainly mRNA, respectively, and Pol III transcribes small RNAs, including tRNAs, 5S rRNA, and U6 small nuclear RNA. The size and complexity of the polymerases increase from Pol II (12 subunits, 514 kDa) via Pol I (14 subunits, 588 kDa) to Pol III (17 subunits, 693 kDa). Detailed structural information is available for Pol II, including the crystal structures of the 10-subunit core enzyme,[1,2] the additional subcomplex Rpb4/7 and the complete 12-subunit enzyme.[3–5] For Pol III recent electron microscopy and mass spectrometric data gave important new insight into the structure of the complex.[6,7] In order to aid further structure and functional elucidation we set out to map, by proteomics technologies, full protein sequences including possible post-translational modifications of Pol II and Pol III.

Several strategies are available for performing a comprehensive analysis of a complex protein mixture. Common methods involve tryptic digests and reducing complexity through the use of multidimensional separation techniques such as SCX,[8] HILIC[9] or IEF [10,11] at the peptide level or IEX,[12] RP[13] or SDS−PAGE[11,14] at the protein level followed by digestion of fractions. However, since most current mass spectrometers have optimal m/z ranges for

* Author to whom correspondence should be addressed. Fax: 31 30 251 8219. E-mail: a.j.r.heck@uu.nl.

[†] Utrecht University.

[‡] Ludwig-Maximilians-Universität München.

(1) Cramer, P.; Bushnell, D. A.; Fu, J.; Gnatt, A. L.; Maier-Davis, B.; Thompson, N. E.; Burgess, R. R.; Edwards, A. M.; David, P. R.; Kornberg, R. D. *Science* **2000**, *288*, 640–649.

(2) Cramer, P.; Bushnell, D. A.; Kornberg, R. D. *Science* **2001**, *292*, 1863–1876.

(3) Armache, K. J.; Kettenberger, H.; Cramer, P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6964–6968.

(4) Armache, K. J.; Mitterweger, S.; Meinhart, A.; Cramer, P. *J. Biol. Chem.* **2005**, *280*, 7131–7134.

(5) Bushnell, D. A.; Kornberg, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6969–6973.

(6) Fernandez-Tornero, C.; Bottcher, B.; Riva, M.; Carles, C.; Steuerwald, U.; Ruigrok, R. W. H.; Sentenac, A.; Muller, C. W.; Schoehn, G. *Mol. Cell* **2007**, *25*, 813.

(7) Lorenzen, K.; Vannini, A.; Cramer, P.; Heck, A. J. R. *Structure* **2007**, *15*, 1237–1245.

(8) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. *Nat. Biotechnol.* **2003**, *21*, 532–538.

(9) Boersema, P. J.; Divecha, N.; Heck, A. J. R.; Mohammed, S. *J. Proteome Res.* **2007**, *6*, 937–946.

(10) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L. *J. Proteome Res.* **2004**, *3*, 112–119.

(11) Krijgsveld, J.; Gauci, S.; Dormeyer, W.; Heck, A. J. R. *J. Proteome Res.* **2006**, *5*, 1721–1730.

(12) Pieper, R.; Gatlin, C. L.; Makusky, A. J.; Russo, P. S.; Schatz, C. R.; Miller, S. S.; Su, Q.; McGrath, A. M.; Estock, M. A.; Parmar, P. P.; Zhao, M.; Huang, S. T.; Zhou, J.; Wang, F.; Esquer-Blasco, R.; Anderson, N. L.; Taylor, J.; Steiner, S. *Proteomics* **2003**, *3*, 1345–1364.

(13) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2199–2204.

(14) Synowsky, S. A.; van den Heuvel, R. H. H.; Mohammed, S.; Pijnappel, W.; Heck, A. J. R. *Mol. Cell. Proteomics* **2006**, *5*, 1581–1592.

10.1021/ac7024283 CCC: $40.75
Published on Web 04/17/2008

analysis, 500−4000 Th for MALDI and 300−1500 Th for ESI, trypsin digestion may not allow a complete analysis due to certain proteolytic peptides falling outside this optimal window. One way to remove all proteolytic issues is by analyzing at the protein level using a top down procedure involving electron capture dissociation (ECD) based sequencing[15,16] with a prior to analysis protein separation and fractionation.[17] Another, more easily approachable, possibility is to use a multiprotease strategy where each enzyme will provide complementary protein fragments as well as sequence overlap.[13,18–23] For instance, Schlosser et al. used a multienzyme strategy to create a cocktail of peptides that would allow the full sequence of the murine circadian protein period 2 (mPER2) to be examined by nanoLC−MS for the purpose of identifying phosphorylation sites.[20] MacCoss et al. not only used a multiprotease strategy but combined it with MuDPIT. Initial experiments focused on model proteins where near complete sequence coverages were obtained alongside several post-translational modifications.[18] A more daunting sample, which consisted of Cdc2p−TAP complexes, was also subjected to the same strategy where they identified over 200 proteins with 20 proteins attaining more than 40% sequence coverage.

Recently electron transfer dissociation (ETD) has emerged as a new method for peptide sequencing,[24,25] possessing complementary features to collision induced dissociation (CID). ETD prefers higher charge states and therefore more basic peptides. The technique shows signs of maturity with elegant phosphopeptide sequencing data generated.[13,26] The latter study also used multiple enzymes for a more comprehensive analysis although emphasis was placed on phosphorylation with all peptides being subjected to TiO_2 based phosphopeptide enrichment.[27,28]

Here we report a multiple-protease strategy, resulting in four complex peptide digests that are separated by reversed phase (RP) nanoLCMSMS, whereby each of the peptides is subjected to ETD with a supplemental collisional activation step (ETcaD) and CID sequencing by a linear ion trap mass spectrometer, leading to eight individual analyses. For comparison, we also analyzed the same samples by nanoLC FT-ICR using only CID MS/MS. An in-house tool was built that allows filtering on peptide scores followed by displaying the observed sequence coverage with a heat map indicating the confidence of the proteomics mapping of each of the amino acids in every protein of both Pol II and III as obtained in each of the eight individual experiments and combinations thereof.

By combining all these experiments, not consuming more than 5 μg of sample in total, we were able to achieve near to 100% sequence coverage of all individual protein constituents involved in the Pol II and Pol III complexes. Not only did we reach our goal of achieving comprehensive analysis, we identified a significant number of novel phosphorylation sites. The sites of Pol II were mapped on its structure, suggesting a role in regulating the conformation of the clamp region and, as a consequence, interaction of Pol II with nucleic acids.

EXPERIMENTAL SECTION

Purification of Pol II and Pol III. Pol II and Pol III core were purified as described in refs 29 and 7, respectively. The additional subunits Rpb4 and 7 were overexpressed and purified as described.[29] The RNA pol purifications were of very high purity as evidenced by the fact that all RNA pol subunits were detected at the top of the list of identified proteins. Next to RNA pol subunits a few elongation factor components were detected as well as some of the usual background, abundant, heat shock proteins. A list of all identified proteins (and their identified peptides) is given in the Supporting Information presented in the scaffold file.

Mass Spectrometry. Two microgram aliquots of the purified protein complex were resuspended in 50 mM NH_4HCO_3 pH 8.0. Reduction and subsequent alkylation were performed with 45 mM DTT for 30 min at 56 °C and 100 mM iodoacetamide for 30 min at room temperature in the dark, respectively. Subsequently, 80 ng (1/25 w/w) of protease (endoproteinases Lys-C, trypsin, Glu-C, chymotrypsin (all Roche-Diagnostics, Netherlands)) was added and digestion was performed overnight at 37 °C. The sample was then acidified using 5% formic acid. The digestions were subsequently analyzed by nanoLC-LTQ-FT-MS (Thermo, San Jose, CA) and by nanoLC-LTQ-XL-MS (Thermo, San Jose, CA) at a material level of 0.5 μg. An Agilent 1100 series LC system was equipped with a 20 mm Aqua C18 (Phenomenex, Torrance, CA) trapping column (packed in-house, i.d., 100 μm; resin, 5 μm) and a 250 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH, Ammerbuch, Germany) analytical column (packed in-house, i.d., 50 μm; resin, 3 μm). Trapping was performed at 5 μL/min for 10 min and washed with solvent A (0.6% acetic acid in water), and elution was achieved with a gradient of 0−32% B (0.6% acetic acid in 80/20 acetonitrile/water) in 60 min, 32−40% B in 5 min, 40−100% B in 2 min and 100% B for 2 min leading to a total analysis time of 90 min. The flow rate was passively split from 0.4 mL/min to 100 nL/min when performing the elution analysis. Nanospray was achieved using a distally coated fused silica emitter (New Objective, Cambridge, MA) (o.d., 360 μm; i.d., 20 μm, tip i.d. 10 μm) biased to 1.8 kV. In the case of the LTQ-FT, the mass spectrometer was operated in the data dependent mode to automatically switch between MS and MS/MS. Survey full scan MS spectra were

(15) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.
(16) Han, X. M.; Jin, M.; Breuker, K.; McLafferty, F. W. *Science* **2006**, *314*, 109–112.
(17) Garcia, B. A.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L. *Nat. Methods* **2007**, *4*, 487–489.
(18) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900–7905.
(19) Fischer, F.; Poetsch, A. *Proteome Sci.* **2006**, *4*, 1–12.
(20) Schlosser, A.; Vanselow, J. T.; Kramer, A. *Anal. Chem.* **2005**, *77*, 5243–5250.
(21) Distler, A. M.; Kerner, J.; Peterman, S. M.; Hoppel, C. L. *Anal. Biochem.* **2006**, *356*, 18–29.
(22) Wu, S. L.; Kim, J.; Hancock, W. S.; Karger, B. *J. Proteome Res.* **2005**, *4*, 1155–1170.
(23) Running, W. E.; Ravipaty, S.; Karty, J. A.; Reilly, J. P. *J. Proteome Res.* **2007**, *6*, 337–347.
(24) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
(25) Good, D. M.; Coon, J. J. *Biotechniques* **2006**, *40*, 783–789.
(26) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E. P.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2193–2198.
(27) Pinkse, M. W. H.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. R. *Anal. Chem.* **2004**, *76*, 3935–3943.
(28) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. D. *Mol. Cell. Proteomics* **2005**, *4*, 873–886.
(29) Edwards, A. M.; Darst, S. A.; Feaver, W. J.; Thompson, N. E.; Burgess, R. R.; Kornberg, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 2122–2126.

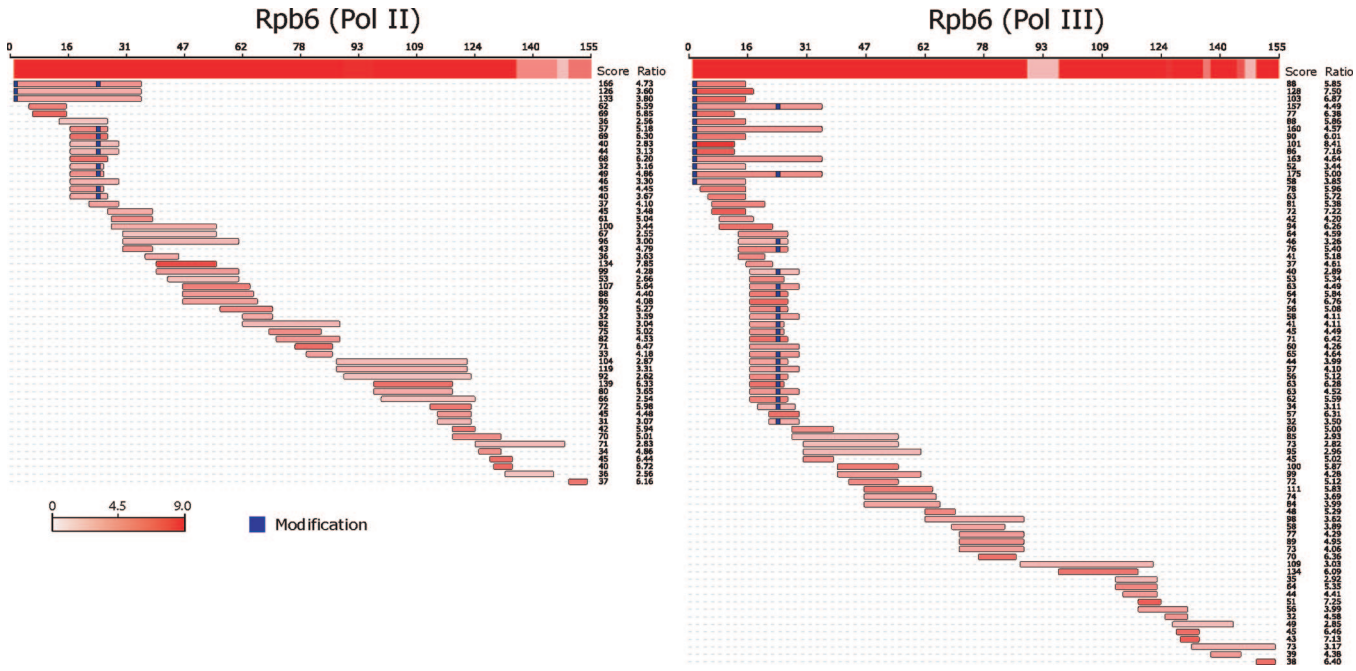

Figure 1. Sequence coverage for Rpb6 from Pol II and Pol III as obtained over all individual experiments using different proteases and activation methods. A white to red color scheme is applied to highlight confidence in identification. The color of each peptide is determined by the Mascot score divided by the number of amino acids in the identified peptide. White represents a Mascot score of 0 while red represents a Mascot score >9, per amino acid. For the overall sequence coverage the scores per amino acid in all different peptides as sequenced in all different experiments were summed. Peptides that were found to be phosphorylated or contained an oxidized methionine were considered to be unique.

acquired from m/z 350 to m/z 1500 in the FT-ICR with a resolution of $R = 100{,}000$ at m/z 400 after accumulation to a target value of 2,000,000 in the linear ion trap. The two most intense ions were fragmented in the linear ion trap using collisionally induced dissociation at a target value of 10,000. In the case of the LTQ-XL, the mass spectrometer was operated in the data dependent mode to automatically switch between MS and MS/MS ETcaD and MS/MS CAD. Survey full scan MS spectra were acquired from m/z 350 to m/z 1500 in the LTQ after accumulation to a target value of 30,000 in the linear ion trap. The two most intense ions were fragmented in the linear ion trap at a target value of 10,000.

Spectra were processed with Bioworks 3.3 (Thermo, Bremen, Germany), and the subsequent data analysis was carried out using the Mascot (version 2.1.0) software platform (Matrix Science, London, U.K.). For the initial screen, the LTQ-FT analyses were searched against the Swiss protein database (version 51.6) with the appropriate enzyme allowing 2 missed cleavages, carbamidomethyl (C) as fixed modification and oxidation (M), phosphorylation (ST) and N-acetylation (protein N terminus) as variable modifications. The peptide tolerance was set to 20 ppm and the MS/MS tolerance to 0.9 Da. The proteins identified were used to create a new database to be used for analysis of the LTQ-XL ETcaD and CAD Data. The LTQ results were searched against the new database with the appropriate enzyme allowing 2 missed cleavages and appropriate instrument configuration. For Glu-C the number of miscleavages allowed was 5 due to the enzyme possessing poor efficiency under these digestion conditions. Carbamidomethyl (C) was used as a fixed modification, and oxidation (M), phosphorylation (ST) and N-acetylation (protein N terminus) were used as variable modifications. The peptide tolerance was set to 3 Da and the MS/MS tolerance to 1.5 Da. All data was compiled into a scaffold database which can be downloaded (https://bioinformatics.chem.uu.nl/supplementary/mohammed_RNAPol/) and freely interrogated using the scaffold viewer.

Software and Visualization. A software program was developed implementing a graphical user interface (GUI, JAVA 1.6 Netbeans IDE 5.5) for the integration and filtering of the peptide identifications from the multiple parallel LC/MS/MS experiments. Multiple Mascot output files (i.e., *.dat-files) together with the corresponding protein sequences database (i.e., fasta-files) were used as input. Peptide identifications were combined and mapped onto the proteins followed by calculating the sequence coverage per protein. Before mapping, peptides can be and were filtered based on several criteria, such as the mascot peptide score. As we found that the Mascot peptide score filter was not very reliable when considering larger peptides (i.e., above 2500 Da), we chose to introduce an extra filter by dividing the peptide score by the number of amino acids present in the peptide. The latter value enabled us to filter out large peptides which tend to have better scores by chance but are often ambiguous peptide identifications. The final filter values used were Mascot peptide score >30 with at least a score contribution of 2.5 per amino acid. As an additional filter all identical peptides within a single experiment were removed, accepting only the peptide with the highest Mascot score. To attain the overall sequence coverage picture, the scores of the same amino acid in different unique sequences are algebraically summed. Peptides that are phosphorylated or contain an oxidized methionine at a particular site are considered to be unique. In the heat maps produced protein coverage is now indicated per individual amino acid using the score/amino acid ratios described above for all the underlying peptides. Since phosphosite localization by Mascot is still an imprecise process, the amino acid is only considered "detected" with no consideration

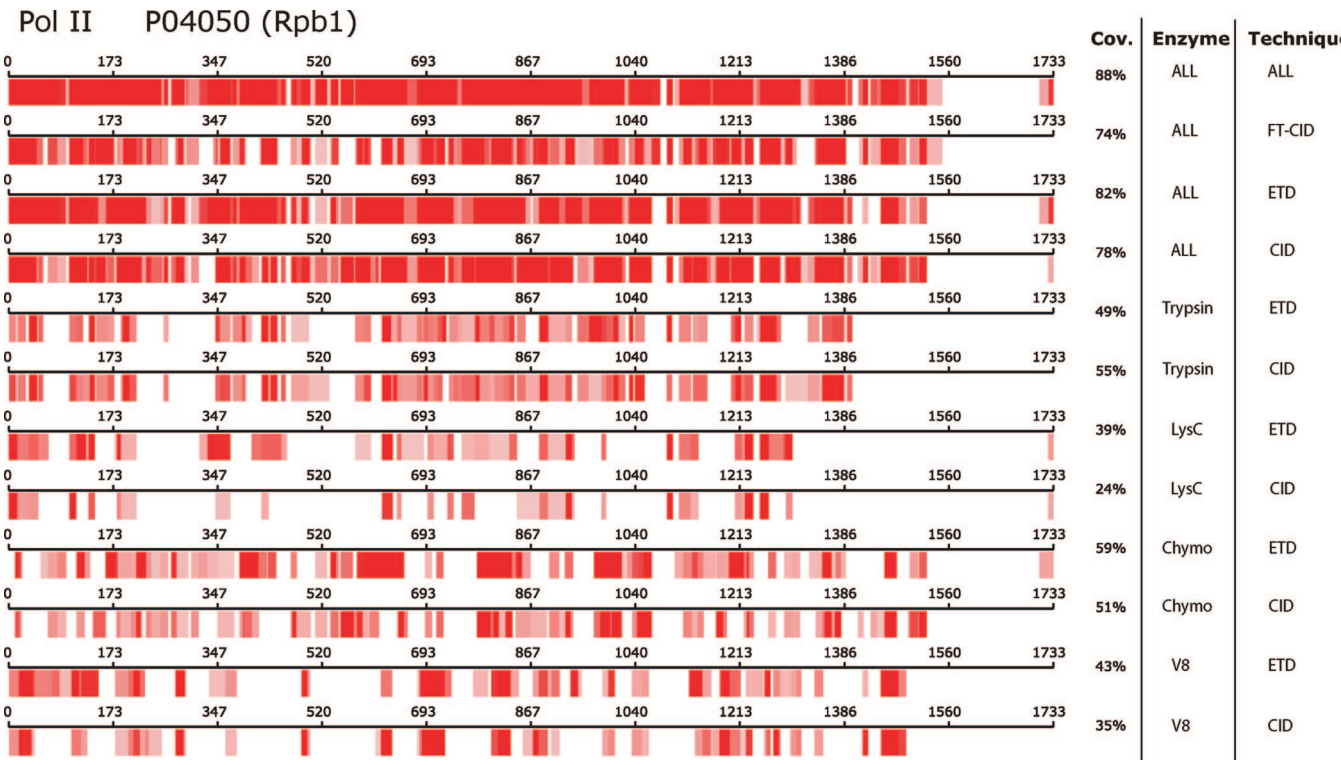

Figure 2. Sequence coverage achieved for Pol II Rbp1 using the indicated enzymes and activation methods for peptide sequencing. A white to red color scheme is applied to highlight confidence in peptide/amino acid identification. Each unique peptide (from all analyses) is broken down to its constituent amino acids, and an equal fraction of the total peptide Mascot score is applied to each amino acid. If certain amino acids are present in more than one peptide with unique sequence, then the amino acid scores are summed. White represents a Mascot score of 0 while red represents a Mascot score >9, per amino acid.

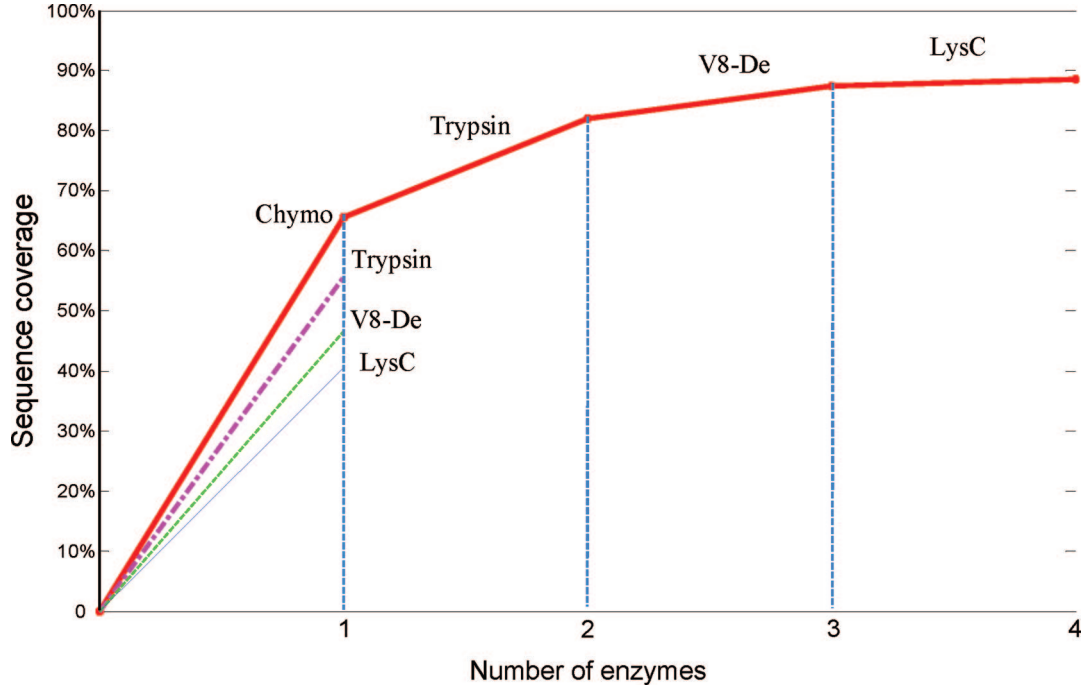

Figure 3. Graph indicating the shortest route to optimal sequence coverage for both Pol II and III (thick continuous red line) with the four enzymes used: chymotrypsin, trypsin, V8 and LysC. In a single experiment chymotrypsin performed best (first segment continuous red line), closely followed by trypsin (pink dot line). Applying these two enzymes alone, the coverage was on average 82%. As discussed earlier, LysC contributes the least to the coverage.

placed on its phosphorylation status, and thus, the heat maps do not consider or display phosphorylation sites. All phosphosites suggested by Mascot are manually evaluated, and the results are presented in Table 1 and in the supplementary data set 2 (Supporting Information). These values were transformed into a white to red color gradient in linearized RGB space interpolating

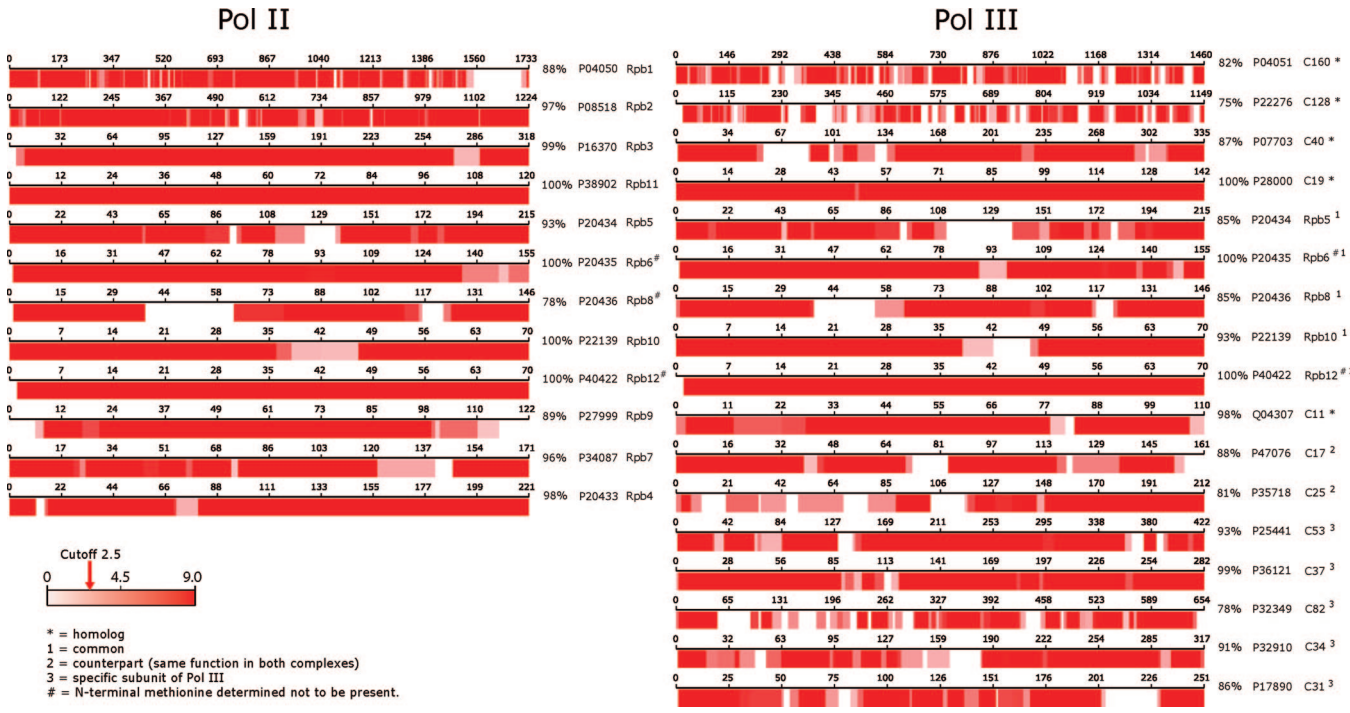

Figure 4. Graphical representation of the total sequence coverage achieved for each protein in Pol II and Pol III. A similar to Figure 2 white to red color scheme is applied to highlight confidence in identification. * denotes homologous proteins in Pol II and Pol III. 1 denotes identical proteins in Pol II and Pol III. 2 denotes counterpart proteins in Pol II and Pol III, i.e., same function in both complexes. 3 denotes specific subunit of Pol III. # denotes N-terminal methionine determined not to be present.

Table 1. Identified Pol II and Pol III Phosphorylation Sites[a]

	Accession Number	Phosphosite	Best Sequenced Peptide	Mascot Score	Method	Remark
Pol II Rbp1 (B220)	P04050	Ser 348	(R)IRGNLMGKRVDFpSARpTVISGDPNLELDQVGVPK(S)	81	Tryp – LTQ (CID)	Not reported S1293 reported in (1) and T621 in (3)
Pol II Rbp1 (B220)	P04050	Thr 351	(R)IRGNLMGKRVDFpSARpTVISGDPNLELDQVGVPK(S)	81	Tryp – LTQ (CID)	Not reported S1293 reported in (1) and T621 in (3)
Pol II Rbp1 (B220)	P04050	Thr 1471	(Y)MPEQKITEIEDGQDGGVpTPY(S)	110	Chym– LTQ-FT (CID)	
Pol II Rbp2 (B150)	P08518	Ser 156	(E)LIAEEpSEDDSE(S)	49	V8– LTQ-FT (CID)	Not previously reported. Rpb2 has been shown to interact with RCK1 (4)
Pol II Rbp3 (B3)	P16370	Ser 32	(R)EASKDNVDFILSNVDLAMANpSLRR(V)	53	Tryp – LTQ (ETD)	
Pol II Rbp6 (ABC 23)	P20435	Ser 24	(E)DFDVEHFpSDEE(T)	69	V8– LTQ-FT (CID)	Not reported. Stoichiometrically phosphorylated, conserved between Pol II and Pol III, Rpb6 interacts with CMK1 (4)
Pol III C160	P04051	Thr 9	(E)VVVpSETPKRIKGLE(F)	53	V8– LTQ (ETD)	
Pol III C53	P25441	Ser 119	(K)SEGpSGSSLVQK(G)	58	Tryp – LTQ-FT (CID)	Not reported. S182,228, 232 & T323 reported (1,2,3)
Pol III C53	P25441	Ser 178	(R)NLIEDEDDGEpSEK(S)	51	Tryp– LTQ-FT (CID)	Reported (1)
Pol III C53	P25441	Ser 224	(R)EIQEALpSEKPTR(E)	62	Tryp – LTQ (ETD)	Reported (1,2)
Pol III C53	P25441	Thr 347	(K)KNIKKKDpTKDALSTRELAGK(V)	47	Lys-C– LTQ (ETD)	
Pol III C82	P32349	Ser 394	(K)RSGpSNAAASLPSK(K	81	Lys-C – LTQ (ETD)	Reported (2,3), additionally S392,394 reported (2,3)
Pol III C37	P36121	Ser 52	(E)NGTNSAIANEQEEKpSEE(V)	71	V8 – LTQ (CID)	
Pol III C37	P36121	Thr 61	(E)VKAEDDpTGEEEE(D)	61	V8 – LTQ (ETD)	Reported (3)
Pol III AC19	P28000	Thr 15	(K)KTATEVpTPQEPK(H)	48	Tryp– LTQ (CID)	Reported (3)
Pol III AC19	P28000	Thr 33	(K)HIQEEEEQDVDMpTGDEEQEEEPDREK(I)	115	Tryp – LTQ (CID)	Reported (1)
Pol III C31	P17890	Thr 101	(K)RKPNIILDEDDpTNDGIERYSDK(Y)	98	Lys-C– LTQ (ETD)	
Pol III C31	P17890	Ser 189	(K)LKELAEDVDDApSTGDGAAK(G)	107	Lys-C– LTQ (ETD)	Reported (1,3) T190 reported in (3)
Pol III Rpb6 (ABC 23)	P20435	Ser 24	(E)DFDVEHFpSDEE(T)	62	V8 – LTQ-FT (CID)	Not reported. Stoichiometrically phosphorylated, conserved between Pol II and Pol III, Rpb6 interacts with CMK1 (4)

[a] In the last column it is indicated whether the site was identified earlier in large scale phosphoproteomics studies and if there is an interaction with a kinase reported.

between 0 and 9.0, thereby creating a confidence heat map. All scores above 9.0 are shown in full red. From the program images as well as tabular data with detailed peptide information can be exported. A stand-alone version of the software is available and can be downloaded from https://bioinformatics.chem.uu.nl/supplementary/mohammed_RNAPol/.

RESULTS AND DISCUSSION

General Results and Assimilation of Data. Initial focus was on the quality of RNA Pol II and III purifications and choices of proteases. Therefore, we analyzed Pol II and Pol III tryptic digests on an LTQ-FT mass spectrometer followed by a search against the Swis-Prot database. These results indicated a relatively

clean purification where all 12 and 17 components of the Pol II and Pol III complexes, respectively, were confidently identified (see scaffold file for full list of protein identifications found using the Swis-Prot database, https://bioinformatics.chem.uu.nl/supplementary/mohammed_RNAPol/ ftcid sprot search). After applying a Mascot score cutoff of 30 ($p < 0.01$), the average sequence coverage was 39%, although certain proteins were covered by not more than 10% (supplementary Figure 1, Supporting Information). Although the analyte mixtures represent reasonably complex digests where undersampling by the mass spectrometer can be an issue, there are other potential obstacles that may hamper obtaining higher sequence coverage including inappropriate size of the tryptic peptide, incomplete digestion, poor MS responses and poor CID fragmentation.

In order to tackle such issues a strategy was designed using a number of alternative enzymes to create complementary pools of proteolytic peptides. Moreover, sequential analysis by CID and electron transfer dissociation (ETD) was applied to allow a broader range of peptides to be successfully sequenced. In present-day proteomics trypsin has been the enzyme of choice for two main reasons. First, trypsin digestion leads to highly selective protein hydrolysis; and second, tryptic peptides will contain a basic residue at the C-terminus, which is beneficial for sequencing by CID.[30,31] However, the introduction of ETD, which has a slightly different set of criteria for optimal peptide sequencing, may require a reanalysis of preferred enzyme for protein identification. In order to add to the protease discussion and to achieve comprehensive sequence coverage we chose and applied a set of proteases: trypsin, chymotrypsin, V8 (primarily a Glu-C) and Lys-C. Two microgram aliquots of sample were digested with one of the above-mentioned proteases, and approximately 0.5 μg (approximately 1 pmol) of material was subjected to nanoLC−MS−MS analysis, where each peptide−ion was subjected to both CID and ETD peptide sequencing. To accept a peptide as being identified we applied a minimum MASCOT score of 30 per peptide and a MASCOT score/amino acid value of 2.5. The latter criterion was applied to filter out large peptides with high peptide MASCOT scores that possessed poor spectra with dubious annotation (see Experimental Section).

A program was developed to filter and assimilate multiple mascot results files (*.dat), providing graphical outputs for peptide identifications and protein sequence coverage. Figure 1 represents such graphical output files for the identical proteins Rpb6 from Pol II (on the left) and Rpb6 from Pol III (on the right). In these files each unique peptide identified is aligned and mapped on the protein sequence, whereby the color of each amino acid indicates its identification confidence score. At the top the full protein sequence is shown, where the color indicates the confidence of identification which is obtained by summing the results over all parallel analyses (for details see Experimental Section). As can be observed a similar proteome map is achieved for Rpb6 originating from Pol II or III. Identical sections of Rpb6 from Pol II and Pol III are observed to have similar levels of sequencing success. An alternative way of analyzing the data is represented in Figure 2, which shows the breakdown of the contribution for each enzyme and activation technique as obtained by each of the eight individual experiments for Rpb1 (Pol II) highlighting the complementary nature of the analyses. This figure also emphasizes the need to perform multiple analyses to achieve a comprehensive analysis. Similar diagrams for the other Pol proteins are shown in the supplementary Figure 2 (Supporting Information). Additionally, supplementary data set 1 (Supporting Information) displays the individual protein sequence coverage for each enzyme using ETD, CID and FT-CID.

An unanticipated finding of our multienzyme, multiactivation approach was that chymotrypsin rivals the commonly used protease trypsin in terms of peptide identifications and protein coverage for both CID and ETD. CID spectra of chymotryptic peptides are usually not as easily interpretable as tryptic peptides since they lack the characteristic dominant y series. However, our data with chymotryptic peptides indicate that database sequence matching incomplete b and y series can be equally sufficient for confident peptide identifications. V8 produced a slightly lower number of peptide identifications for Pol II and Pol III for both CID and ETD, but did attain the highest number of phosphopeptides identifications. Lys-C provided a slightly different picture with peptides possessing three charges or more being more dominant—possibly ideal for ETD. However, the overall number of peptides identified with Lys-C was lower relative to the other enzymes. A possible cause may be found in the fact that Lys-C produces larger peptides, where noncovalent interactions can hamper peptide fragmentation and thus successful sequencing.[32,33] The combination of ETD on an orbitrap or other high-resolution mass spectrometer would likely improve our Lys-C results, for instance through the use of initial charge state screening and an appropriate "supplemental" activation[33] to allow dissociation of noncovalently attached ETD product ions.

Figure 3 shows the average sequence coverage per protein analyzed for each of the used proteases. In a single experiment we found chymotrypsin most successful. If two experiments are performed, we found, somewhat unexpectedly, that trypsin is the best partner to achieve highest sequence coverage. Figure 4 highlights that it is necessary to employ all enzymes to reach near complete sequence coverage, although in our experiment Lys-C did not provide a significant addition in sequence coverage. Combining all experimental data we obtain a near complete proteome sequence mapping of all protein subunits of Pol II and Pol III as visualized in Figure 4. A few proteins still show some gaps in the sequence coverage. For instance, the Rpb1 sequence coverage map contains a gap in its C-terminal domain (CTD) that is formed by 26−27 heptapeptide repeats of the sequence YSPTSPS spanning over 180 amino acids not targeted by any of the four used proteases. Such a peptide will be difficult to detect and sequence considering its size and its lack of basic residues. Interrogation of MS data did not indicate any candidate unidentified peaks that could suggest the presence of the CTD region, which may mean it was lost during sample preparation or introduction into the MS. In C31 an unobserved stretch of sequence consists almost exclusively of acidic residues which

(30) Sakurai, H.; Mitsuzawa, H.; Kimura, M.; Ishihama, A. *Mol. Cell. Biol.* **1999**, *19*, 7511–7518.
(31) Huang, Y. Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *Anal. Chem.* **2005**, *77*, 5800–5813.
(32) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399–1406.
(33) Hakansson, K.; Chalmers, M. J.; Quinn, J. P.; McFarland, M. A.; Hendrickson, C. L.; Marshall, A. G. *Anal. Chem.* **2003**, *75*, 3256–3262.

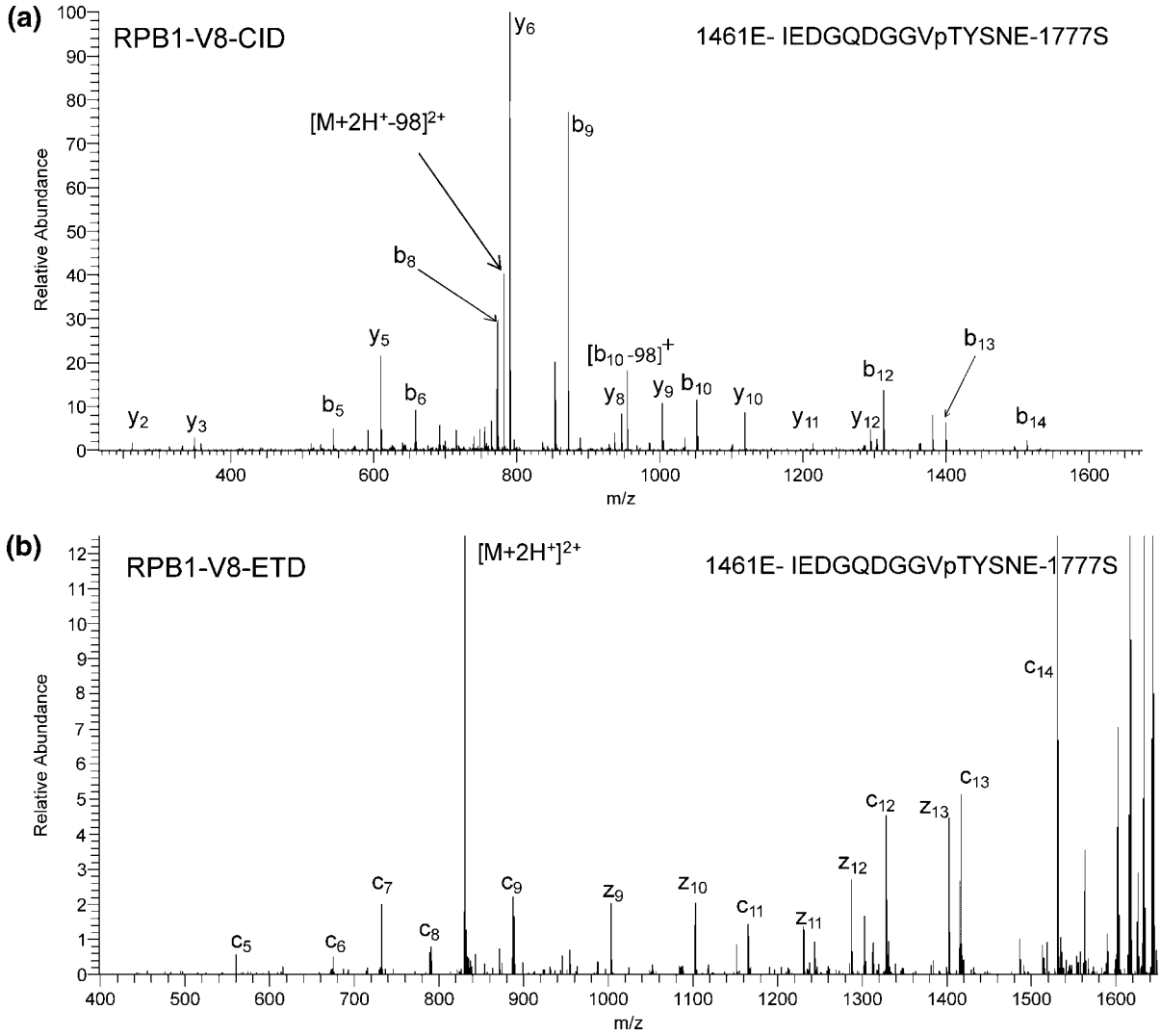

Figure 5. Tandem MS spectrum of the phosphopeptide IEDGQDGGVpTYSNE (Rpb1) originating from a V8 digest of the Pol II complex sequenced by using (a) CID, mascot score 106 and (b) ETD, mascot score 79. Both spectra were recorded on the LTQ ion trap, whereby the latter was obtained using supplemental collision activation.

seem to be, by the here used proteases and conditions, noncleavable. Included in all analyses was a search for phosphorylated peptides and acetylated protein N-termini. Most phosphorylated peptides were identified in more than one of the individual experiments. All identified and validated phosphosites are listed in Table 1 where it states the site location within the protein and in which proteolytic peptide it was detected that had the best Mascot score. Corresponding annotated spectra, additional confirmatory annotated spectra are available in supplemental data set 2 (Supporting Information). Below, we focus in more detail on these phosphorylations and provide more detailed context for two Pol proteins: Rpb1 of Pol II and Rpb6, which is present and identical in Pol II and Pol III.

Rpb1. Rpb1 has a mass of approximately 192 kDa and represents the biggest subunit of Pol II. It is one of the ten core proteins and is homologous to C160 in Pol III. By applying different enzymes and activation methods we were able to obtain a near complete sequence coverage with the exception of the CTD domain (see above). Rpb1 represented a protein for which chymotrypsin provided the highest sequence coverage obtained by a single enzyme (59%). However, typical for the total data set, the majority of the phosphorylation sites in Rpb1 were identified by using V8 as the enzyme for digestion. Rpb1 has been the subject of intensive research both in structural and in biochemical ways. Its structure is known,[3] and the phosphorylation status of the CTD governs the transcription cycle.[34] However, despite extensive studies on the phosphorylation of the CTD of Rpb1 and other large scale phosphorylation studies[26,35,36] we still detected three previously unidentified phosphorylation sites in Rpb1 that do not lie within the CTD: S348, T351 and T1471. Figure 5 shows CID and ETD spectra originating from V8 peptides used to identify T1471. This phosphopeptide and all the other phospho identifications were mapped onto the structural model of Pol II as shown in Figure 6. T1471 lies in a region that links the core of Rpb1 to the CTD. Of particular interest, both phosphosites S348 and T351 are located adjacent to the so-called switch 2 at the base of the polymerase clamp, a region that undergoes conformational

(34) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E. P.; Coon, J. J. *Anal. Chem.* **2007**, *79*, 477–485.
(35) Dahmus, M. E. *Biochim. Biophys. Acta* **1995**, *1261*, 171–182.
(36) Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villen, J.; Elias, J. E.; Gygi, S. P. *J. Proteome Res.* **2007**, *6*, 1190–1197.

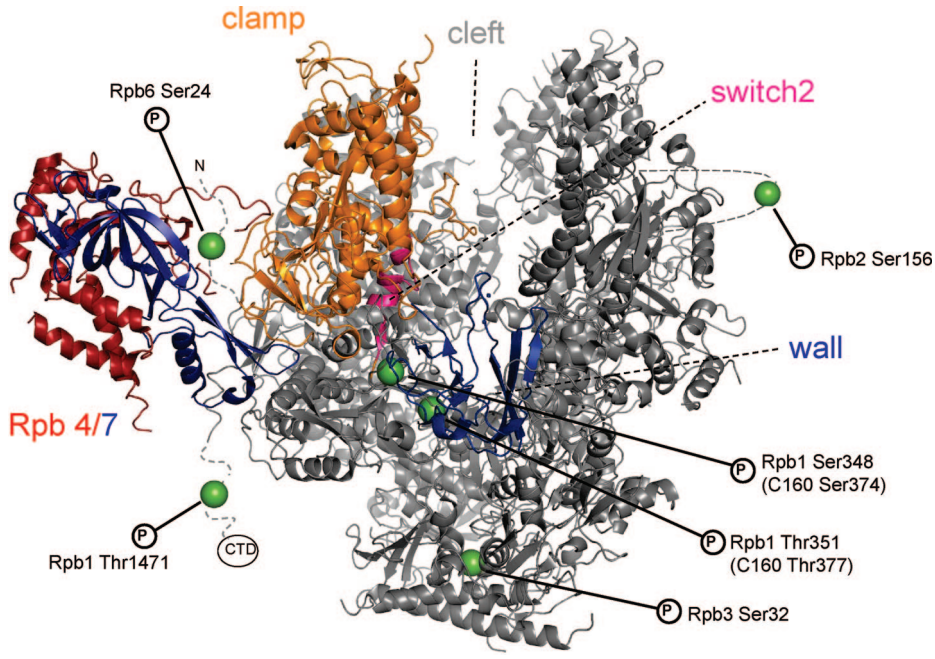

Figure 6. Structural model of Pol II highlighting the locations of the identified phosphosites.

changes in the transition to a transcribing complex and contact DNA−RNA hybrids in the active site. Therefore, the phosphorylation status of these sites can play a direct role in regulating the conformation of the clamp region and, as a consequence, interaction of Pol II with nucleic acids.[3–5]

Rpb6. Common to both polymerases, Rbp6 has been shown to play a role in the binding of the Rpb4/7 heterodimer in Pol II.[37] Nevertheless, the regulation mechanism of the possible release or attachment of the two additional subunits is unknown. Rpb4 and Rpb7 are counterparts to C17 and C25 in Pol III. The heterodimer remained partly intact in dissociation experiments on both polymerases in native tandem mass spectrometry, confirming that Rpb4 and Rpb7 are located closely to each other, and probably strongly physically bound.[7] In those experiments Rpb6 was detected with two masses, separated by 80 Da, in a ratio of about 2:3, most likely caused by an abundant single phosphorylation on Rpb6. Through our multienzyme approach we were able to determine the site as being S24. This phosphosite was identified using different enzymes and with both activation methods for Rpb6 in both Pol II and Pol II complexes.

Further RNA Pol Phosphorylation. Including the phosphorylation sites mentioned above we found in total 19 sites in the two Pol complexes excluding the CTD of Rpb1, the largest subunit of Pol II. Seven of them were found in Pol II and thirteen in Pol III (see Table 1). Six of these sites have been reported previously in large scale phosphoproteomics studies.[26,35,36,38] Interestingly, most of the phosphorylation sites are in proteins, which are unique for Pol II or Pol III, and only one was found in a subunit that is shared by both Pol II and Pol III (i.e., the above-mentioned site in Rpb6).

The level of phosphorylation is dependent on many factors like cell cycle growth condition, activity in transcription etc. Evidently, due to the purification and sample handling as well as detection methods some phosphorylation sites may still be missed. However, all but one of the amino acids in identical protein subunits were detected as unphosphorylated in this analysis. Protein phosphorylation often has a regulatory function; yet, the two protein complexes fulfill different roles in the transcription machinery and share the same environment in the nucleus. The locations of the identical subunits are similar in both Pol II and Pol III and are thus likely to be susceptible to phosphorylation at similar sites as found for the subunit Rpb6. Since Pol II and Pol III transcribe different classes of genes, it is unlikely that phosphorylation on these identical/common proteins would allow this differential behavior and, perhaps, this is a reason why little phosphorylation is observed since these subunits are unlikely to be kinase targets. As mentioned above, the exception is Rpb6, where an abundant, almost stoichiometric, phosphorylation is observed. Rpb6 has been shown to play a role in the binding of a homologue heterodimer (Rpb4 and Rpb7 which are homologous to C17 and C25 in Pol III) to the two polymerases. The available crystal structure[3] reveals that the N-terminal tail of Rpb6, where the observed phosphosite resides, is not resolved and would be expected to lie on the surface of the complex. These features underline the N-terminus as a possible target for a kinase.

Further analysis of the remaining identified phosphorylation sites in the homologous proteins of the two polymerases was performed with the primarily tool being sequence comparison

(37) Ptacek, J.; Devgan, G.; Michaud, G.; Zhu, H.; Zhu, X.; Fasolo, J.; Guo, H.; Jona, G.; Breitkreutz, A.; Sopko, R.; McCartney, R. R.; Schmidt, M. C.; Rachidi, N.; Lee, S. J.; Mah, A. S.; Meng, L.; Stark, M. J.; Stern, D. F.; De Virgilio, C.; Tyers, M.; Andrews, B.; Gerstein, M.; Schweitzer, B.; Predki, P. F.; Snyder, M. *Nature* **2005**, *438*, 679–684.

(38) Gerber, J.; Reiter, A.; Steinbauer, R.; Jakob, S.; Kuhn, C. D.; Cramer, P.; Griesenbeck, J.; Milkereit, P.; Tschochner, H. *Nucleic Acids Res.* **2008**, *36*, 793–802.

using the BLAST algorithm.[39] S156 is located in a flexible loop and can likely be a target of kinase activity. All other sites were not deemed to be conserved.

CONCLUSION

Here we used a multiplexed proteomics approach using four different proteases and two different peptide activation methods to attempt a complete mapping of all proteins of the yeast RNA polymerase II and III. Our approach allowed a near complete sequence coverage of each protein, consuming in total less than 5 μg (approximately 10 pmol) of starting material, and provided informative insights in the strength and weaknesses of different proteases and activation methods. As an additional benefit, our near comprehensive mapping of the two Pol complexes enabled us to reveal several novel phosphorylation sites. Mapping these new sites onto the crystal structure suggests that, for some, a role may exist in regulating the conformation of the clamp region and, as a consequence, interaction of RNA with nucleic acids.

ACKNOWLEDGMENT

This work was supported by The Netherlands Proteomics Centre (www.netherlandsproteomicscentre.nl).

SUPPORTING INFORMATION AVAILABLE

Supplementary Figure 1: Bar graphs indicating the sequence coverage achieved for each protein in the two polymerase complexes through using Trypsin proteolysis and LTQ-FTICR based CID. Supplementary Figure 2: The sequence coverage achieved for each Pol II and Pol III protein using the indicated enzymes and activation methods for peptide sequencing. A white to red color scheme is applied to highlight confidence in peptide/amino acid identification. Each unique peptide (from all analyses) is broken down to its constituent amino acids and an equal fraction of the total peptide Mascot score is applied to each amino acid. If certain amino acids are present in more than one peptide with unique sequence then the amino acid scores are summed. White represents a Mascot score of 0 while red represents a Mascot score >9, per amino acid. Supplementary data set 1: An Excel sheet with embedded graphs representing the results for each enzyme and peptide activation technique specifically the sequence coverage gained of each protein from the two polymerase complexes. Supplementary data set 2: Raw spectra and corresponding annotation for all phosphosites and sequences mentioned in Table 1. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review November 27, 2007. Accepted March 19, 2008.

AC7024283

(39) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.