BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **245,** 53–58 (1998)
ARTICLE NO. RC988373

Protonation of the Neutral Repeats of the RNA Polymerase II CTD

Daniel P. Morris,* Robert D. Stevens,† and Arno L. Greenleaf*,[1]

Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710; and †Division of Biochemical Genetics, Department of Pediatrics, Duke University Medical Center, Durham, North Carolina 27710

Received February 25, 1998

The CTD (carboxy-terminal repeat domain) of the largest subunit of RNA Polymerase II in most eukaryotes consists of from 26 to 52 seven amino acid repeats, the consensus sequence of which is YSPTSPS. Even though this consensus repeat does not contain residues that are normally protonated under the conditions used for positive ion electrospray mass spectrometry, we find that the CTD acquires about one proton per repeat when analyzed by this procedure. We have termed this phenomenon superprotonation. Superprotonation is apparently a property of the consensus sequence as the repeat peptide, (YSPTSPS)$_4$, is superprotonated whereas other proteins and the repeat peptides (YSPTSPK)$_4$, (YSPTSPR)$_4$ and (YSPTAPR)$_4$ are not. The highly conserved nature of the contiguous consensus repeats in organisms ranging from yeast to mammals implies that the functionally significant behavior of the domain is easily perturbed. We propose that CTD superprotonation is a manifestation of a unique biophysical property that will influence and could be the basis for consensus repeat function *in vivo*. © 1998 Academic Press

The regulation of RNA Polymerase II plays a central role in many aspects of gene expression. The CTD (carboxy-terminal repeat domain) of the largest subunit of RNA polymerase II has been implicated in most of the regulatory mechanisms involving RNA polymerase II (1, 2) including assembly of the RNA Pol II holoenzyme (3, 4), regulator controlled gene activation (4-8), release from the promoter region (9, 10), efficiency of RNA elongation (11, 12) and recruitment of RNA processing components (13-22).

The CTD is highly conserved in organisms from yeast to man and consists of 26 to 52 seven amino acid repeats with the consensus sequence YSPTSPS (1). Although a shift from an unphosphorylated to a hyperphosphorylated CTD is correlated both with a transition into the elongation phase of transcription (9, 10) and with a change in the conformation of the CTD (23), little information is available to indicate what the structure of either the unphosphorylated or the phosphorylated CTD might be. When a synthetic peptide with 8 consensus repeats was analyzed by 2D NMR and Circular Dichroism, a right handed beta turn structure could be forced in 90% trifluoroethanol: however, little evidence of structure was apparent in an aqueous environment (24). The undefined structure of the repeat peptide in water seems to conflict with the nearly complete conservation of about 18 contiguous consensus repeats in organisms as evolutionarily distant as yeast and humans (1).

In the present study we identify a very unusual physical/chemical characteristic of the CTD. Even though the CTD consensus repeat, YSPTSPS, does not contain residues normally protonated under the conditions used for positive ion electrospray mass spectrometry (25), peptides containing most of the yeast CTD are highly protonated during analysis by this procedure. Indeed, species containing nearly one protonation per repeat are present. Although proteins and peptides occasionally acquire one or two positive charges beyond those expected from protonation at the Lys, Arg, His and amino terminal residues, we have been unable to find any other examples of proteins that display this level of unexpected protonation. We have termed this phenomenon superprotonation. The ability of the highly conserved consensus repeats to superprotonate reveals the presence of a very unusual negative site in this normally neutral sequence. Whatever the biophysical basis for this unusual phenonmenon, it is likely to

[1] Corresponding author: Arno L. Greenleaf, Department of Biochemistry, Duke University Medical Center, Durham, NC 27710. Fax: (919) 684-8885. E-mail: arno@biochem.duke.edu.

Abbreviations: TFA,Triflouroacetic acid; FXa, factor Xa; ESMS, electrospray mass spectrometry; RP HPLC, reverse-phase high perfomance liquid chromatography; HPLC/MS, high performance liquid chromatography/mass spectrometry

0006-291X/98 $25.00

affect the structural and biological behavior of the CTD of the largest subunit of RNA Pol II.

MATERIALS AND METHODS

Overexpression and purification of GST-yCTD. The fusion protein GST-yCTD was placed under the control of an IPTG inducible promotor in pGEX3X (Pharmacia) as previously described (26). Expression and glutathione Sepharose 4B purification of GST-yCTD were done as recommended by Pharmacia (their reference XY-058-00-01) with the modifications previously indicated (26). Purification of full sized GST-yCTD was done by loading the protein fractions from the glutathione Sepharose 4B column directly onto a 2 cm (5 ml) Ni^{2+} -NTA-agarose (Qiagen) column equilibrated with column buffer (5 mM Imidazole, 500 mM NaCl, 1 mM PMSF, and 10 mM Tris pH 7.8). The column was washed with 6× 5 ml of column buffer and 6× 5 ml of wash buffer (60 mM imidazole, 500 mM NaCl, 1 mM PMSF, and 10 mM Tris pH 7.8) and eluted with 1 ml fractions of elution buffer (200 mM imidazole, 500 mM NaCl, 1 mM PMSF, and 10 mM Tris pH 7.8) (26).

FXa[2] digestion of the GST-yCTD fusion protein. Digestions of GST-yCTD were done by thawing aliquots of the fusion protein (1.2 mg/ml) in 10 mM Tris pH 7.8, adding $CaCl_2$ to 2 mM and FXa (New England Biolabs) to 20 μg/ml and incubating for 1 hr at room temperature. Digestions of the partially proteolyzed fusion protein which did not bind the nickel column were done similarly in 0.2 mM $CaCl_2$ and 10 mM NaCl.

HPLC and mass spectrometry. Fragments from FXa digests of the GST-yCTD were separated by RP-HPLC on a 2.1x 150 mm C4 column (214TP5215, Vydac) with matching guard column. At loading the column was equilibrated with 90% solution A (0.025% TFA) and 10% solution B (0.025% TFA in 90% acetonitrile). The column was eluted using an ISCO microbore HPLC syringe pump at 50 μl per minute using the following gradient: 10% B from 0 to 10 min., linear gradient to 90% B from 10 to 40 min. and 90% B from 40 to 50 min.

Mass measurements were made on a Micromass-VG Quattro BQ (Altrincham, UK) triple quadrupole mass spectrometer equipped with a pneumatically assisted electrostatic ion source operating at atmospheric pressure and in positive ion mode. Synthetic peptides (Chiron Mimotopes, Raleigh, NC) reconstituted to 1 mg/ml in 50% aqueous acetonitrile containing 1% formic acid and RP-HPLC fractions were analyzed by loop injection into a stream of 50% aqueous acetonitrile flowing at 10 μl/min. Spectra were acquired in multi channel acquisition mode (MCA) from m/z 350-1600 or 700-1600 (scan time 5 sec.) for synthetic peptides or HPLC fractions, respectively.

For HPLC/MS analysis, the effluent from the column was split evenly into two streams. One stream was delivered to the ion source of the mass spectrometer and the other to a UV detector monitoring at 220 nm. Spectra were acquired in continuum mode from m/z 300-1600 (scan time 5 sec.).

The mass scale was calibrated with horse heart myoglobin (16951.48 Da) with a resolution corresponding to a peak width at half height of 1.0 Da for m/z 893. For experiments with synthetic peptides, the declustering potential was minimized so that the highest charge state could be observed. The mass spectra were transformed to a molecular mass scale using software (VG MassLynx) supplied by the manufacturer.

RESULTS

Purification of GST-yCTD and yCTD. Analysis of the CTD both as a protein domain and as a substrate for kinases required a homogenous CTD polypeptide. Because the CTD of RNA Pol II is very susceptible to proteolytic digestion, we introduced a GST at the amino-terminal end of the yeast CTD and a 6XHis at the carboxy-terminal end of the yCTD (Fig. 1A). Following overexpression in E. coli, this GST-yCTD fusion protein could be purified by consecutive use of the affinity tags at each end (Fig. 1B). A FXa cleavage site between the GST and the yCTD allowed cleavage of the GST group from the yCTD which includes the 6XHis tag.

Mass spectrometry on the yCTD. When FXa cleaved GST-yCTD was subjected to RP HPLC, major peaks eluted at 36 and 46% acetonitrile. The mass spectrum of the peptide in the first peak transforms to give to give a mass of 21,501.9 Da, in agreement with the expected

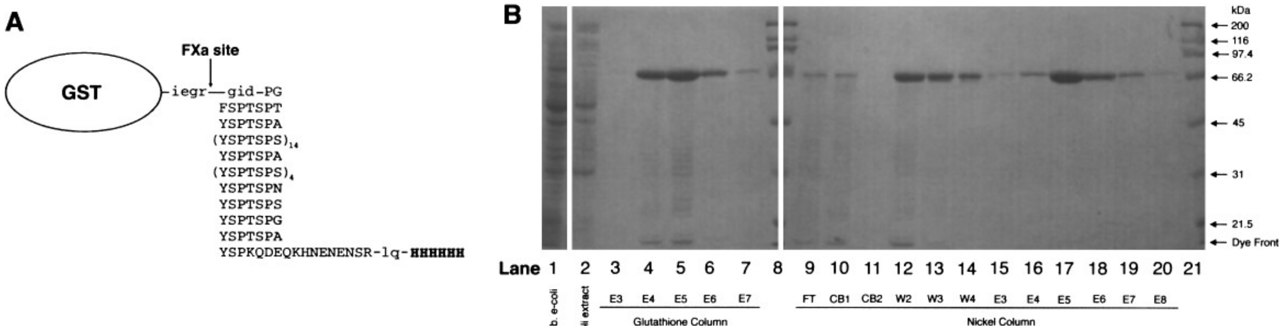

FIG. 1. GST-yCTD and its purification. Fig. 1A shows a diagram of the fusion protein consisting of the amino-terminal GST, the FXa cleavage site, three linking amino acids, the CTD (capital letters) , two linking amino acids and the 6×His tag. Protein-containing fractions from the steps in the purification of GST-yCTD are shown analyzed on 12% Laemmli gel in Fig. 1B (stained with Coomassie Blue). Lanes 1 and 2 contain 0.00025% of the triton solubilized E. coli and 0.00025% of the supernatant following centrifugation of the solubilized E-coli (E. coli extract), respectively. Other sample lanes contain 0.2% of the indicated fraction. Lanes 3 to 7 contain the glutathione elution fractions (1ml), lane 9 contains the flow though from loading glutathione elution fractions 4 through 7 on a 2 cm height 5 ml nickel column, lanes 10 and 11 contain column buffer wash fractions (5 ml), lanes 12 to 14 contain the 60 mM imidazole wash fractions (5 ml), while lanes 15 to 20 contain the 200 mM imidazole elution fractions (1 ml). Lane 8 and 21 contain molecular weight markers.

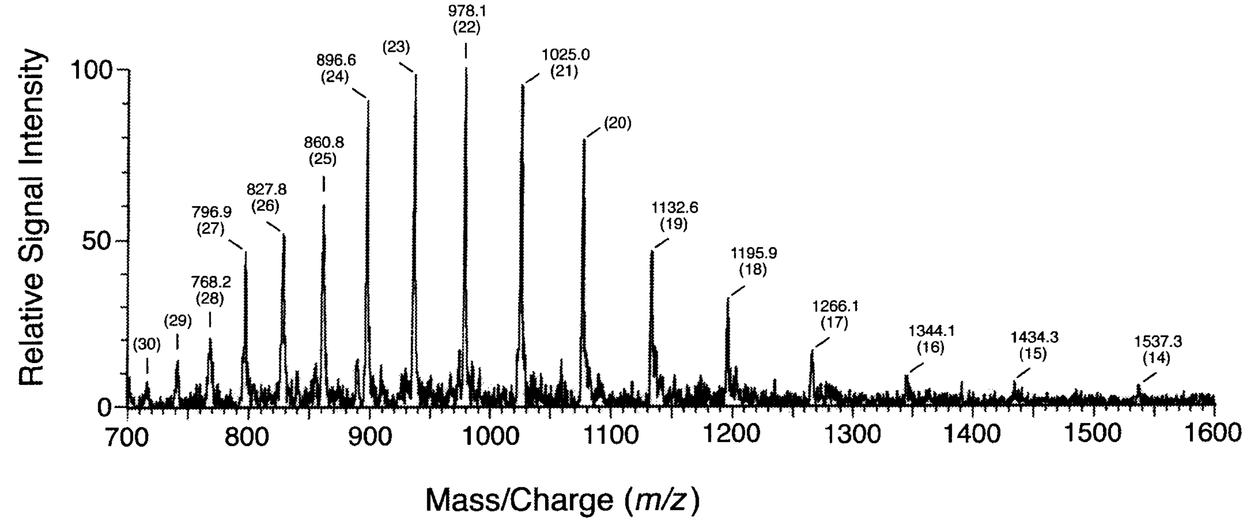

FIG. 2. Mass spectrum on the yCTD. Raw data from analysis of HPLC purified yCTD shows the mass to charge ratio of the ionized yCTD molecules. The protonation state of each species is indicated in parenthesis below the mass to charge ratio. Even though the yCTD (including the 6×His tag) contains only 11 sites that would usually be protonated, it is associated with from 14 to 30 protons.

average mass for the yCTD including the 6XHis tag (21,503.5 Da). The mass spectrum of this peptide (Fig. 2), shows a charge state distribution from 14 to 30, implying there are at least 30 protonation sites. However, the protein has only eleven sites that are normally protonated in the positive ion ESMS experiment. These data strongly suggest that the yCTD has the ability to superprotonate under the conditions used in ESMS experiments. The second peak (46% acetonitrile) contained many species, none of which could be clearly identified by mass spectrometry; however, SDS PAGE analysis indicated it contained the cleaved GST (data not shown).

To investigate further the superprotonation phenomenon, a FXa digest was done on material that did not bind to the nickel column and presumably did not contain the 6XHis tag. Four CTD peptides were identified by HPLC/MS. These species were cleaved at the FXa site, included all of the CTD repeats, lacked the 6XHis tag and were superprotonated (Fig. 3). The smallest peptide, missing the 6XHis tag and 9 amino acids of the nonrepeat carboxy-terminal region (see Fig. 1), contains only one lysine and the amino terminus yet still displays as many as 24 protonation sites (Fig. 3C).

Mass spectrometry on synthetic repeat peptides. To determine if a smaller number of repeats could be superprotonated, the synthetic peptide, $(YSPTSPS)_4$, was subjected to ESMS (Fig. 4). This peptide, which contains only one usual protonation site, gave a mass spectrum with primarily doubly and triply protonated species. In contrast, $(YSPTSPK)_4$, which is based on the consensus sequence of the plasmodium CTD, has a mass spectrum that consists primarily of the $[M+3H^+]/3$, $[M+4H^+]/4$ and $[M+5H^+]/5$ ions as would be expected based upon amino acid composition (Fig. 4). The peptides, $(YSPTSPR)_4$ and $(YSPTAPR)_4$, have essentially the same charge state distribution as $(YSPTSPK)_4$ (data not shown).

DISCUSSION

As a step toward obtaining information on the CTD and its phosphorylation sites we created a GST-yCTD fusion protein with a carboxy-terminal 6XHis tag. A FXa cleavage site between the amino-terminal GST and the yCTD sequence allowed removal of the GST moiety. When the molecular mass of HPLC-purified yCTD was measured by ESMS, it was found to be in agreement with the theoretical average molecular mass. Surprisingly, under the mildly acidic conditions used in the ESMS experiment (pH ca. 2), the mass spectrum showed that the yCTD acquired protons in excess of the eleven expected on the basis of the number of Lys, Arg, His and amino-terminal residues (25). Indeed, the mass spectral data show that most yCTD peptides exhibit between 8 and 16 additional protonation sites beyond the eleven usual sites. Although protonation at one or two extra locations is occasionally observed during mass spectrometry of proteins and peptides, the extent of superprotonation of the yCTD is to the best of our knowledge without precedence.

Several additional lines of evidence demonstrate that the ability of the yCTD to superprotonate resides in the repeated sequence of the CTD. Peptides that contained the yCTD but lacked the 6×His tag and portions of the nonconsensus carboxyl-terminal sequence were all superprotonated. The superprotonation of these peptides indicates the 6×His tag is not responsible for

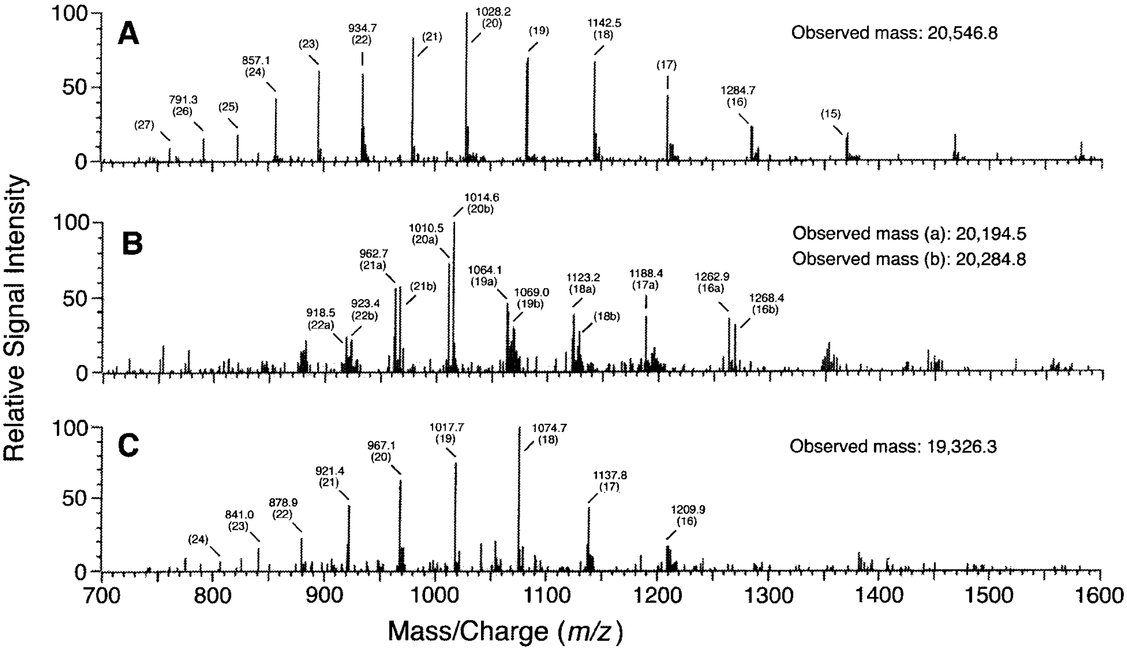

FIG. 3. Analysis of the yCTD lacking the 6×His tag and various lengths of the nonrepeat carboxy-terminal CTD region. Products from a FXa cleavage of partially proteolyzed GSTyCTD were analyzed by on-line HPLC/Mass spectrometry. Four species were identified with masses consistent with yCTD species missing carboxy-terminal residues. The species have average masses consistent with the calculated mass of the yCTD missing the carboxy-terminal amino acids QHHHHH (20552.5 Da, Fig. 3A), RLQHHHHHH (20283.0 Da, Fig. 3B), SRLQHHHHHH (20196.0 Da, Fig. 3B) and KHNENENSRLQHHHHHH (19330.1 Da, FIG. 3C). The protonation state of each species is indicated in parenthesis below the mass to charge ratio.

this phenomenon and suggests the carboxy-terminal nonconsensus residues are also not required. In addition, the mass spectrum for each of these peptides shows charge states with at least 20 unexpected protonation sites. This approximates the number of consensus repeats in these peptides.

To demonstrate that the YSPTSPS sequence was sufficient for superprotonation, the repeat peptide, $(YSPTSPS)_4$, was analyzed by ESMS. Since all of the side chains in this peptide are neutral, the only expected protonation site is the amino-terminus. Nevertheless, the mass spectrum for this peptide consisted primarily of doubly and triply protonated forms. This is in contrast to the spectra for $(YSPTSPK)_4$, $(YSPTSPR)_4$ and $(YSPTAPR)_4$, which displayed charge state distributions in accordance with each peptide's amino acid composition. Thus the sequence, YSPTSPS, is sufficient to induce superprotonation, although the fact that $(YSPTSPS)_4$ had only 1 or 2 and not 3 or 4 extra protonation sites suggests that multiple repeats must be contiguous for superprotonation to occur.

The fact that the sequence, YSPTSPK, apparently does not superprotonate is interesting as this heptamer is the consensus repeat found in the Plasmodium CTD and is also common in the carboxy-terminal 18 repeats of the mammalian CTD (1). In this regard it is noteworthy that the YSPTSPK-rich region of the mouse CTD is unable to substitute for the YSPTSPS-rich region of the mouse CTD (27) Similarly, the YSPTSPS-rich but not the YSPTSPK-rich regions of the mammalian CTD can be used to replace the yeast CTD (28). The *in vitro* superprotonation observed within the YSPTSPS but not the YSPTSPK repeats may be related to this difference in *in vivo* behavior.

We have not yet identified the location of the protonation sites in the consensus repeats. Within each repeat, potential protonation sites include the five amide and the two imide bonds as well as the three serine, one threonine and one tyrosine OH groups. Of the chemical groups mentioned, the carbonyl oxygens of amide bonds are the most negative, normally displaying a pKa of about −1.0 [p. 6 in (29)]. The inability of these groups to protonate in most polypeptides suggests the superprotonation event involves either multiple coordination sites or a large increase in the electronegativity of a single site. Determining the exact nature of the protonation sites will require additional experimental efforts, but it is clear that the phenomenon of superprotonation is rare and almost certainly arises from the sequence of the CTD consensus repeat.

While the biological implications of CTD superprotonation remain to be determined, one possibility is that the protonation site represents a cation binding site. The absence of an appropriate cation could certainly account for the surprising lack of structure that has been observed for the consensus repeats of this

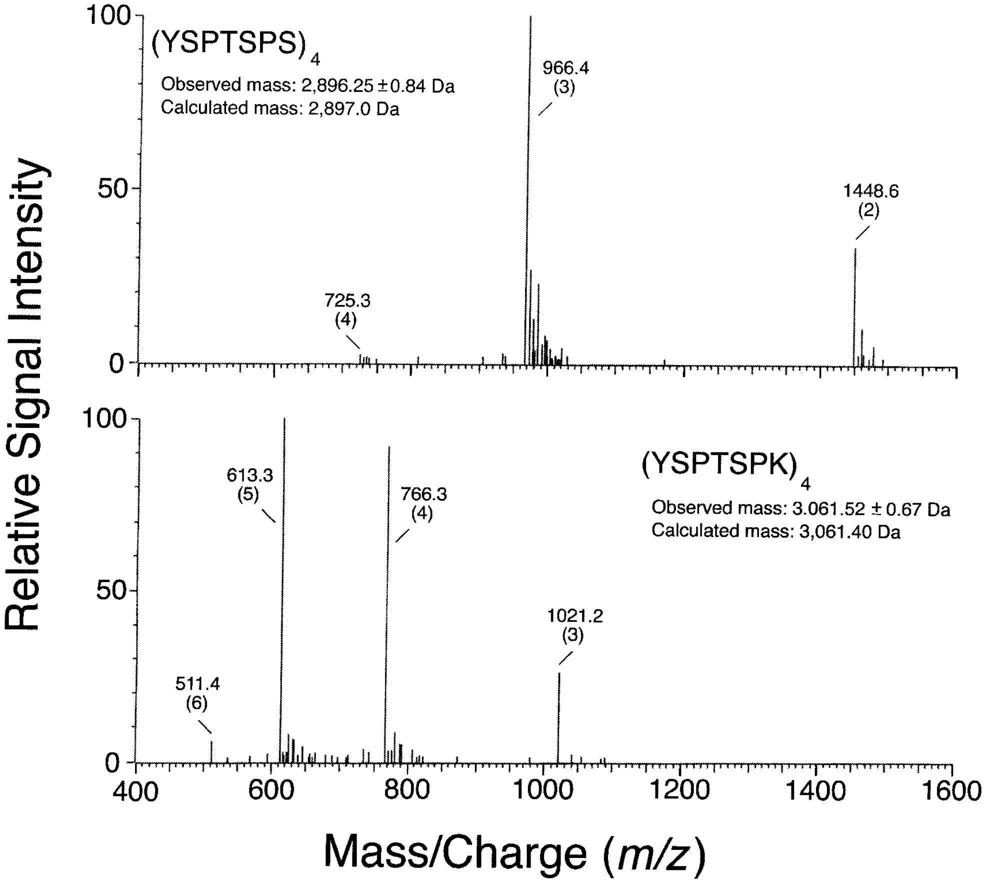

FIG. 4. Mass spectrometric analysis of a synthetic peptide with 4 consensus repeats reveals superprotonation. Even though the peptide (YSPTSPS)$_4$ contains only one usual protonation site (the amino terminus) almost all of the observed species were protonated either 2 or 3 times (top panel). In contrast a control peptide, (YSPTSPK)$_4$, which contains 4 Lys residues in addition to the amino terminus, exhibits very little unexpected protonation (bottom panel).

highly conserved domain (24). The existence of a protonation site in the neutral repeats of the CTD of RNA Pol II reflects an apparently unique biophysical property of this domain. This property may underlie the biological requirement for the consensus repeat.

ACKNOWLEDGMENTS

We thank Arijit Dasgupta and Wen Chen for their technical help in the production and purification of the fusion protein. This work was funded by National Institutes of Health grant GM40505.

REFERENCES

1. Corden, J. L. (1990) *Trends. Biochem. Sci.* **15,** 383–387.
2. Koleske, A. J., and Young, R. A. (1995) *Trends Biochem. Sci.* **20,** 113–116.
3. Thompson, C. M., Koleske, A. J., Chao, D. M., and Young, R. A. (1993) *Cell* **73,** 1361–1375.
4. Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H., and Kornberg, R. D. (1994) *Cell* **77,** 599–608.
5. Allison, L. A., and Ingles, C. J. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 2794–2798.
6. Scafe, C., Chao, D., Lopes, J., Hirsch, J. P., Henry, S., and Young, R. A. (1990) *Nature* **347,** 491–494.
7. Liao, S. M., Taylor, I. C., Kingston, R. E., and Young, R. A. (1991) *Genes Dev.* **5,** 2431–2440.
8. Gerber, H. P., Hagmann, M., Seipel, K., Georgiev, O., West, M. A., Litingtung, Y., Schaffner, W., and Corden, J. L. (1995) *Nature* **374,** 660–662.
9. O'Brien, T., Hardin, S., Greenleaf, A. L., and Lis, J. T. (1994) *Nature* **370,** 75–77.
10. Dahmus, M. E. (1995) *Biochim. Biophys. Acta.* **1261,** 171–182.
11. Marshall, N. F., Peng, J., Xie, Z., and Price, D. H. (1996) *J. Biol. Chem.* **271,** 27176–27183.
12. Lee, J. M., and Greenleaf, A. L. (1997) *J. Biol. Chem.* **272,** 10990–10993.
13. Greenleaf, A. L. (1993) *Trends Biochem. Sci.* **18,** 117–119.
14. Bregman, D. B., Du, L., van der Zee, S., and Warren, S. L. (1995) *J. Cell Biol.* **129,** 287–298.
15. Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R. V., Gentile, C., Gebara, M., and Corden, J. L. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 6975–6980.
16. Kim, E., Du, L., Bregman, D. B., and Warren, S. L. (1997) *J. Cell Biol.* **136,** 19–28.
17. Du, L., and Warren, S. L. (1997) *J. Cell Biol.* **136,** 5–18.

18. McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M., and Bentley, D. L. (1997) *Nature* **385,** 357–361.

19. Steinmetz, E. J. (1997) *Cell* **89,** 491–495.

20. Yue, Z., Maldonado, E., Pillutla, R., Cho, H., Reinberg, D., and Shatkin, A. J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 12898–12903.

21. Cho, E. J., Takagi, T., Moore, C. R., and Buratowski, S. (1997) *Genes Dev.* **11,** 3319–3326.

22. McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Program, A. E., Shuman, S., and Bentley, D. L. (1997) *Genes Dev.* **11,** 3306–3318.

23. Zhang, J., and Corden, J. L. (1991) *J. Biol. Chem.* **266,** 2297–2302.

24. Cagas, P. M., and Corden, J. L. (1995) *Proteins* **21,** 149–160.

25. Wang, G., and Cole, R. B. (1997) *in* Solution, Gas–Phase, and Instrumental Parameter Influences on Charge–State Distributions in Electrospray Ionization Mass Spectrometry (Cole, R. B., Ed.), pp. 137–174, John Wiley and Sons, Inc., New York.

26. Morris, D. P., Lee, J. M., Sterner, D. E., Brickey, W. J., and Greenleaf, A. L. (1997) *Methods* **12,** 264–275.

27. Bartolomei, M. S., Halden, N. F., Cullen, C. R., and Corden, J. L. (1988) *Mol. Cell. Biol.* **8,** 330–339.

28. Corden, J. L., and Ingles, C. J. (1992) *in* Carboxy-terminal Domain of the Largest Subunit of Eukaryotic RNA Polymerase II (McKnight, S. L., and Yamamoto, K. R., Eds.), pp. 81–107, Cold Spring Harbor Press, Cold Spring Harbor.

29. Creighton, T. E. (1984) Proteins: Structures and Molecular Principles, W. H. Freeman and Co., New York.