A.K. *Mol. Cell* **11**, 1537–1548 (2003).
3. Aviv, T. *et al. Nat. Struct. Biol.* **10**, 614–621 (2003).
4. Wang, C. & Lehmann, R. *Cell* **66**, 637–647 (1991).
5. Bergsten, S.E. & Gavis, E.R. *Development* **126**, 659–669 (1999).
6. Dahanukar, A., Walker, J.A. & Wharton, R.P. *Mol. Cell* **4**, 209–218 (1999).
7. Smibert, C.A., Lie, Y.S., Shillinglaw, W., Henzel, W.J. & Macdonald, P.M. *RNA* **5**, 1535–1547 (1999).
8. Smibert, C.A., Wilson, J.E., Kerr, K. & Macdonald, P.M. *Genes Dev.* **10**, 2600–2609 (1996).
9. Gavis, E.R., Lunsford, L., Bergsten, S.E. & Lehmann, R. *Development* **122**, 2791–2800 (1996).
10. Dahanukar, A. & Wharton, R.P. *Genes Dev.* **10**, 2610–2620 (1996).
11. Crucs, S., Chatterjee, S. & Gavis, E.R. *Mol. Cell* **5**, 457–467 (2000).
12. Schultz, J., Ponting, C.P., Hofmann, K. & Bork, P. *Protein Sci.* **6**, 249–253 (1997).
13. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536–540 (1995).
14. Kim, C.A., Gingery, M., Pilpa, R.M. & Bowie, J.U. *Nat. Struct. Biol.* **9**, 453–457 (2002).
15. Kim, C.A. *et al. EMBO J.* **20**, 4173–4182 (2001).
16. Conti, E. & Kuriyan, J. *Structure Fold. Des.* **8**, 329–338 (2000).
17. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. *Cell* **94**, 193–204 (1998).
18. Eklof Spink, K., Fridman, S.G. & Weis, W.I. *EMBO J.* **20**, 6203–6212 (2001).
19. Huber, A.H. & Weis, W.I. *Cell* **105**, 391–402 (2001).
20. Graham, T.A., Weaver, C., Mao, F., Kimelman, D. & Xu, W. *Cell* **103**, 885–896. (2000).
21. Wang, X., McLachlan, J., Zamore, P.D. & Hall, T.M. *Cell* **110**, 501–512 (2002).
22. Wang, X., Zamore, P.D. & Hall, T.M.T. *Mol. Cell* **7**, 855–865 (2001).
23. Edwards, T.A., Pyle, S.E., Wharton, R.P. & Aggarwal, A.K. *Cell* **105**, 281–289 (2001).
24. Fribourg, S., Gatfield, D., Izaurralde, E. & Conti, E. *Nat. Struct. Biol.* **10**, 433–439 (2003).
25. Lau, C.K., Diem, M.D., Dreyfuss, G. & Van Duyne, G.D. *Curr. Biol.* **13**, 933–941 (2003).
26. Shi, H. & Xu, R.M. *Genes Dev.* **17**, 971–976 (2003).
27. Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. & Morikawa, K. *Proc. Natl. Acad. Sci. USA* **97**, 8257–8262 (2000).
28. Roe, S.M. *et al. Mol. Cell* **2**, 361–372 (1998).
29. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H. & Nagai, K. *Nature* **372**, 432–438 (1994).

The CTD code

Stephen Buratowski

How does the C-terminal domain (CTD) of RNA polymerase II interact specifically with multiple targets? A recent paper describing the structure of this domain with a mRNA capping enzyme guanylyltransferase suggests that the CTD is a contortionist that, upon post-translational modification, adopts different configurations specifically recognized by its partners.

The gene expression field has been experiencing a period of remarkable integration. Eukaryotic RNA polymerase II (RNAPII) produces mRNA, but that is only its most basic function. While transcribing, RNAPII also scans for DNA damage and modifies the surrounding chromatin. Through protein-protein interactions, RNAPII also acts as a platform for several mRNA processing factors that modify the mRNA as it is being synthesized. One particularly important component for these interactions is the C-terminal domain (CTD) of the RNAPII largest subunit. The CTD couples transcription with histone methylation, mRNA splicing, and polyadenylation, but its best-characterized direct interaction is with the mRNA capping enzyme. A recent report in *Molecular Cell* presents the crystal structure of the CTD bound to the capping enzyme guanylyltransferase (Cgt1), extending our understanding of this interaction to the atomic level[1]. This and other studies lend insight into how transcription by RNAPII is linked to so many other processes.

The CTD is a simple repetition (27–52 repeats, depending upon the organism) of the heptapeptide sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. No analogous domain exists on the related RNAPI and RNAPIII enzymes, and the CTD is completely dispensable for RNA polymerization. The CTD is highly phosphorylated *in vivo*, and many proteins are thought to bind to this domain. Interacting partners include the Mediator complex that regulates transcription initiation, several histone methyltransferases, the capping enzyme that modifies the 5′ end of mRNA, and the polyadenylation factors that modify the 3′ end (for reviews, see refs. 2,3). How does such a simple sequence interact with so many targets? Rather than carrying all these factors throughout the transcription cycle, the CTD interacts dynamically with each factor at the appropriate time. A series of different phosphorylations and conformation changes generates configurations specific for binding of particular factors. In essence, there is a CTD 'code' that specifies the position of RNAPII in the transcription cycle.

The two major CTD phosphorylations occur at distinct points in the transcription cycle. The serine in the fifth position (Serine 5) is phosphorylated by the basal transcription factor TFIIH near the promoter, and genetic and biochemical data show that capping enzyme is recruited by this modification[4–7]. The structure of the *Candida albicans* guanylyltransferase (Cgt1)-CTD phosphopeptide complex illustrates how a CTD code can be read. The peptide used contains four heptad repeats, each phosphorylated at serine 5, but only seventeen residues (two repeats) are visible in the structure. The phosphopeptide binds in a cleft on the nucleotidyl transferase domain, with an extended β-like conformation containing one turn at proline 6. The phosphates on two serine 5 residues from adjacent repeats bind in positively charged pockets and act as electrostatic anchors to either end of the binding cleft. In addition to serine 5, the tyrosine and two prolines within each repeat also make specific contacts with Cgt1. These interactions are consistent with mutagenesis data reported by Fabrega *et al.*[1], as well as with previous biochemical and genetic studies.

Serine 2 is phosphorylated during elongation by a different kinase. There are suggestions that polyadenylation factors may interact specifically with the serine 2 phosphorylated form of the CTD. Therefore, the two phosphorylations help distinguish early and late phases of transcription[7].

In addition to phosphorylation, a CTD code probably also includes *cis-trans* isomerization at the two prolines that follow the phosphorylated serines. The proline isomerase Pin1/Ess1 acts at prolines preceded by a phosphorylated residue and has been implicated in mRNA 3′ end formation (ref. 8 and references therein). It is informative to compare the CTD (serine 5-P)–capping enzyme structure with that of Pin1 bound to the CTD phosphorylated at both serine 2 and serine 5 (ref. 9). Whereas the capping enzyme-bound CTD has a β-like configuration, the Pin1-bound CTD is more like a type II polyproline helix. In both structures, the prolines are in the *trans* configuration. Pin1 binds to the CTD (at least in part) via its WW domain, a motif found in several other CTD-binding

The author is in the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: steveb@hms.harvard.edu

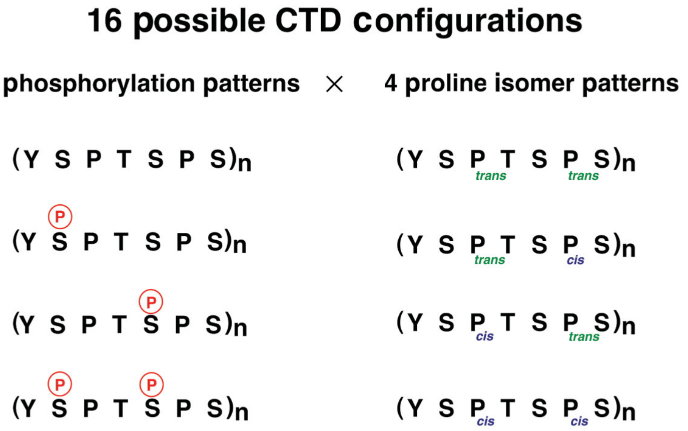

Figure 1 A possible CTD code. Possible phosphorylation sites are denoted by circled P (red). Two prolines can adopt either the *cis* or *trans* configuration.

proteins[10], and it will be important to determine if these other interactions occur via similar molecular contacts.

Solution studies of unbound CTD peptides suggest that it is largely unstructured. However, phosphorylation can influence the propensity of the CTD peptides to form β-turns and type II polyproline helices[11]. It remains to be determined whether phosphorylation changes the equilibria between the *cis* and *trans* forms of the CTD prolines. Just considering the possible patterns of phosphorylation and proline configurations (**Fig. 1**), sixteen distinct states can be specified within a single CTD repeat. Each state is potentially a specific recognition site for an interacting factor. The number gets much larger if higher order patterns containing multiple repeats are considered. Therefore, by regulating the timing of phosphorylation events during transcription, a great deal of information can be encoded in the CTD. Further complexity could come from other reported covalent modifications of the CTD (which include ubiquitylation, glycosylation, and phosphorylation of other residues within the repeat) and nonconcensus repeats found in most organisms.

Another interesting structure with implications for coupling between transcription and mRNA processing is that of the complete 12-subunit RNAPII[12,13]. Until recently, the available high-resolution structures of RNA polymerase II were missing the Rpb4 and Rpb7 subunits. The Rpb4/7 dimer is located near the channel where RNA exits the polymerase, close to the predicted location of the CTD. The Rpb4/7 subcomplex is not required for RNA polymerization. RNAPII lacking Rpb4/7 can form transcription initiation complexes, yet these complexes fail to initiate transcription[14]. Rpb7 contains an oligonucleotide binding (OB) fold and a ribonucleoprotein (RNP) fold, domains seen in some single-stranded nucleic acid binding proteins. It has been suggested that Rpb7 could interact with the nascent mRNA, although there is not yet any data to show that this occurs during transcription. Rpb4/7 could also potentially modulate the interaction between the CTD and its modifying enzymes.

Even with only two structures of CTD interactions available[1,9], it is apparent simplicity of the RNAPII CTD is deceptive. Multiple conformations and modifications allow it to interact with many distinct targets. It remains to be seen whether the interaction sites of the targets will fall into recognizable classes (such as the WW domain) or will be widely divergent. Many of the interactions may involve 'induced fit' of the CTD to its target, so more structures will be required to determine the rules that govern the readout of the CTD code.

1. Fabrega, C., Shen, V., Shuman, S. & Lima, C.D. *Mol. Cell* **11**, 1549–1561 (2003).
2. Hampsey, M. & Reinberg, D. *Cell* **113**, 429–432 (2003).
3. Maniatis, T. & Reed, R. *Nature* **416**, 499–506 (2002).
4. Rodriguez, C.R., Cho, E.J., Keogh, M.C., Moore, C.L., Greenleaf, A.L. & Buratowski, S. *Mol. Cell Biol.* **20**, 104–112 (2000).
5. Schroeder, S.C., Schwer, B., Shuman, S. & Bentley, D.L. *Genes Dev.* **14**, 2435–2440 (2000).
6. Pei, Y., Hausmann, S., Ho, C.K., Schwer, B. & Shuman, S. *J. Biol. Chem.* **276**, 28075–28082 (2001).
7. Komarnitsky, P., Cho, E.J. & Buratowski, S. *Genes Dev.* **14**, 2452–2460 (2000).
8. Myers, J.K., Morris, D.P., Greenleaf, A.L. & Oas, T.G. *Biochemistry* **40**, 8479–8486 (2001).
9. Verdecia, M.A., Bowman, M.E., Lu, K.P., Hunter, T. & Noel, J.P. *Nat. Struct. Biol.* **7**, 639–643 (2000).
10. Sudol, M., Sliwa, K. & Russo, T. *FEBS Lett.* **490**, 190–195 (2001).
11. Bienkiewicz, E.A., Moon Woody, A. & Woody, R.W. *J. Mol. Biol.* **297**, 119–133 (2000).
12. Armache, K.J., Kettenberger, H. & Cramer, P. *Proc. Natl. Acad. Sci. USA* **100**, 6964–6968 (2003).
13. Bushnell, D.A. & Kornberg, R.D. *Proc. Natl. Acad. Sci. USA* **100**, 6969–6973 (2003).
14. Orlicky, S.M., Tran, P.T., Sayre, M.H. & Edwards, A.M. *J. Biol. Chem.* **276**, 10097–10102 (2001).