ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

In silico protein protein interaction between RNA POL II and CTCF Transcription Factors

[1]Maisarah Ab Samad, [1]Shaharum Shamsuddin, *[2]Daruliza Kernain Mohd Azman

[1]Advanced Molecular Biology Laboratory, School of Health Sciences, Health Campus, Universiti Sains Malaysia.

[2]Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia.

Abstract

The carboxyl terminal domain (CTD) of large subunit of RNA polymerase II (LS RNAPII) is the regulatory platform of multi-subunit enzyme in mediating transcription events. It consist of heptapeptide repeats and regulated by post translational modifications and its protein partners. CTCF is a transcription factor that interact with the CTD through its C-terminus. RNAPII is also regulated by CTCF during alternative splicing. The regulation and binding sites of the interaction is still unclear. This study aims to validate the interaction between CTCF and LS RNAPII in glioblastoma, RGBM cell line using *ex vivo* immunoprecipitation. Presence of interaction between CTCF and CTD encourage the discovery of potential binding regions based on their properties as intrinsically disorder proteins (IDPs) using bioinformatics tools. PepSite2 webs server was used to discover possible interactions between predicted molecular recognition features (MoRFs) in the C-terminus with CTD peptides retrieved from structural model of RNAPII protein complexes. All MoRFs had lowest binding score (*P*-value < 0.15) with the CTD peptides from 3D9K (phosphorylated at Ser2 and Ser5) and 3D9L (phosphorylated at Ser2). Both CTD peptides consist of functional unit of CTD which started with tyrosine. Hence, we suggested the required CTD sequence and its phosphorylation status for the interaction with C-terminus. Phosphorylation motifs in C-terminus were discovered and analysis was done by PepSite2 to predict the interaction of the motifs in unphosphorylated and phosphorylation serine/threonine residues with the CTD peptides. The adjustment of binding score and accuracy was varied among the motifs based on the protein kinases. These findings provide a preliminary and suggestive insights on the physical interactions between RNAPII and CTCF.

Keywords: *CTCF, RNAPII, CTD, MoRFs, phosphorylation*

INTRODUCTION

RNA polymerase II (RNAPII) mediates eukaryotic RNA transcription, which is the multi-subunit enzyme that transcribes protein-coding genes. The largest subunit of RNAPII or also known as Rpb1 provides the regulatory platform through its carboxyl-terminal domain (CTD) (Corden *et al.*, 1985). The CTD plays a central role in coordinating the entire transcription cycle from initiation, elongation and termination of the transcription process as well as co-transcriptional RNA maturation process (Nakahashi *et al.*, 2013). CTD consists of multiple heptapeptide repeats (consensus Tyr1–Ser2–Pro3–Thr4–Ser5–Pro6–Ser7), varying in number from 26 in yeast to 52 in vertebrates (Hsin and Manley, 2012). CTD is subjected to extensive post-translational modification predominantly phosphorylation which is critical in determining its functional role (Dahmus, 1993; Dahmus, 1995; Dahmus, 1996; Meinhart and Cramer, 2004). Hence, the “CTD code” is introduced to describe the formation of dynamic pattern in the heptads of the CTD by post-translational modifications (Buratowski, 2003). The best elucidated CTD residues are Ser2 and Ser5 in which their phosphorylation status is critical in mediating the interacting partners and the progress of transcription (Bataille *et al.*, 2012).

CTCF or CCCTC binding factor is an ubiquitous, 11 zinc finger transcription factor (Filippova *et al.*, 1996), which is known as the master weaver of genome and architectural protein by mediating chromosomal topology and boundaries(Phillips and Corces, 2009; Ong and Corces, 2014). Interaction of CTCF with LS RNAPII was discovered by Chernukhin *et al.* (2007) in which the CTCF interacts with the CTD via the C-terminus. Nevertheless, the regulation of the interactions is still unclear. Despite of this, regulation of RNPII during elongation phase is correlated with epigenetic reprogramming which is mediated by CTCF and methylation at exon 5 of CD45 locus during alternative splicing (Shukla *et al.*, 2011). Henceforth, regulation of RNPII and CTCF in the choice of exons during pre-mRNA splicing is significant in gene regulation. In addition, CTCF is also a potential tumor suppressor whereby deregulation of CTCF such as hyper methylation at CTCF binding sites and loss of CTCF binding could cause disruption of chromosomal topology and promoting cancer such as glioma genesis (Flavahan *et al.*, 2016). As alternative splicing is regulated by methylation of CTCF binding site and RNAPII, deregulation of CTCF might cause aberrant alternative splicing which is also a hallmark of cancer (Oltean and Bates, 2014). Besides that, aberrant alternative splicing also implicates pathogenesis of glioblastoma multiform which is an invasive primary brain tumor (Thorne *et al.*, 2015).

In this report, we describe the interaction of RNAPII CTD with CTCF by performing co-immunoprecipitation in glioblastoma multiform, RGBM cell line, to display the presence of interaction between the CTD and CTCF in RGBM cell line. Next, we demonstrate the *in silico* interaction via the functional features of intrinsically disordered proteins (IDPs). IDPs are protein that can be totally unstructured or being partially unstructured, consisting with intrinsically disordered regions (IDRs) in native condition (Dunker *et al.*, 2001; van der Lee *et al.*, 2014). CTCF contains IDRs at the terminal segments especially within the C-terminal domain, hence the region is reported to be disordered(Martinez and Miranda, 2010), as well as CTD of RNAPII (Corden, 1990; Meinhart and Cramer, 2004). Hence, both proteins are IDP. Functional features such as molecular recognition features (MoRFs) and short linear motifs (SLiMs) mediate the protein-protein interactions in IDRs/IDPs (Mohan *et al.*, 2006; Davey *et al.*, 2012). These features were employed to discover potential binding and phosphorylation sites for the interaction between C-terminus of CTCF and RNAPII CTD using several bioinformatics tools. The findings may provide preliminary and suggestive insights on the interaction of CTCF and RNAPII.

Materials And Methods

Cell line and culture conditions

Glioma RGBM? ATCC, USA and was grown in complete growth medium RPMI 1640 (GIBCO@Invitrogen, USA), supplemented with 10% fetal bovine serum and 1% Penicillin-Streptomycin in a humidified incubator containing 50 mL/L CO_2 at 37 ℃.

Co-immunoprecipitation (CO-IP)

Ex vivo protein-protein interaction was characterized via CO-IP method. For this assay, RGBM cells ($1x10^7$cells/mL) were collected, washed and lysed. The RGBM total cell lysate was precipitated with the addition of ice-cold acetone at the ratio of 1:1 to obtain a higher concentration of protein from the cell line. After incubation and centrifugation, the pellet was retrieved and resuspended in cold PBS. To investigate CTCF and RNAPII interaction *ex vivo*, the total cell lysate was precipitated with RNAPII antibody (ABCAM?). Protein complex bound to the RNAPII antibody was then incubated with protein G sepharose. The immunoprecipitation formed was resolved in SDS PAGE and probed with CTCF antibody for Western blot analysis. In order to confirm the presence of CTCF and RNAPII interaction, the reciprocal CO-IP experiment was performed in which the total cell lysate was precipitated with CTCF antibody and then probed with RNAPII antibody in the Western blot.

Prediction of MoRFs and phosphorylation motif in C-terminus

Amino acid sequence of human CTCF (UniProt IDs: P49711) and human RNAPII large subunit (UniProt IDs: P24928) were obtained from UniProt (http://www.uniprot.org/). Both CTCF and RNAPII contain IDRs. Hence, the extent of disorder across these proteins were assessed using IUPred (Dosztányi *et al.*, 2005) (http://iupred.enzim.hu/). Next, the disorder-based protein binding regions or MoRFs were also predicted by ANCHOR (Dosztányi *et al.*, 2009) (http://anchor.enzim.hu/) in both protein. In addition, $MoRF_{CHiBi\_Web}$ (Malhis and Gsponer, 2015)(http://morf.chibi.ubc.ca:8080/mcw/index.xhtml) was also applied for extended analysis and conservation of the MoRFs. The phosphorylation motifs in C-terminus were discovered using eukaryotic linear motif (ELM) resource (http://elm.eu.org/). Itis a hub for collecting, classifying and curating information about SLiMs (Dinkel *et al.*, 2015).

Prediction of interaction between C-terminus and CTD

Protein Data Bank (PDB) structures of RNAPII complex with respective protein targets were retrieved from Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). PepSite2 (Trabuco *et al.*, 2012) (http://pepsite2.russelllab.org/) requires a PDB structure in order to predict the binding position of a peptide. The binding of predicted MoRFs at the C-terminus of CTCF with the CTD peptide chain from the retrieved complexes was postulated using PepSite2.The peptide binding score was calculated in which the lower *P*-value represents the accuracy of the prediction due to the stronger signal.

Results

CTCF interacted with CTD of RNAPII *ex vivo*

This study was carried out to determine the *ex vivo* interaction between CTCF and RNAPII in the RGBM cell line via CO-IP experiment. In this assay, CTCF antibody was coupled to the protein-G-sepharose and the interaction with RNAPII was characterized in the RGBM cell line. The complex formed was resolved with SDS- PAGE and the presence of an interacting protein partner was determined with RNAPII CTD antibody.

Figure 1 displays the results from Western blot which shown the migration of RNAPII as a doublet with a molecular weight of 220 kDa and 240 kDa proteins in lane 1. Lane 2 shows the result of total proteins immunoprecipitated with CTCF antibody and probed with RNAPII antibody in western blot. The protein detected was a thick band with a molecular weight from 150-240 kDa. Based on this observation, CTCF could be a part of RNAPII complex, especially via the CTD of LS RNAPII which has been discovered by Chernukhin *et al.* (2007) to interact with the CTCF C-terminus. We attempted to elucidate interaction by predicting the site/s of C-terminus of CTCF that responsible for the interaction with the CTD of RNAPII using *in silico* analysis.

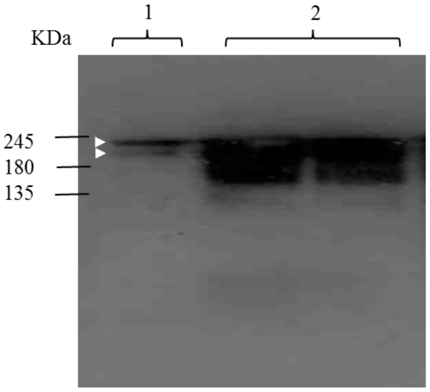

Figure 1. Western blot image of *ex vivo* interaction of CTCF and LS RNAPII. The interaction was characterized in the RGBM total cell lysate. Lane 1 shows the result of RGBM total cell lysate probed with RNAPIICTD antibody whereas lane 2 shows the result of total proteins being immunoprecipitated with CTCF antibody and probed with RNAPII antibody in Western blot.

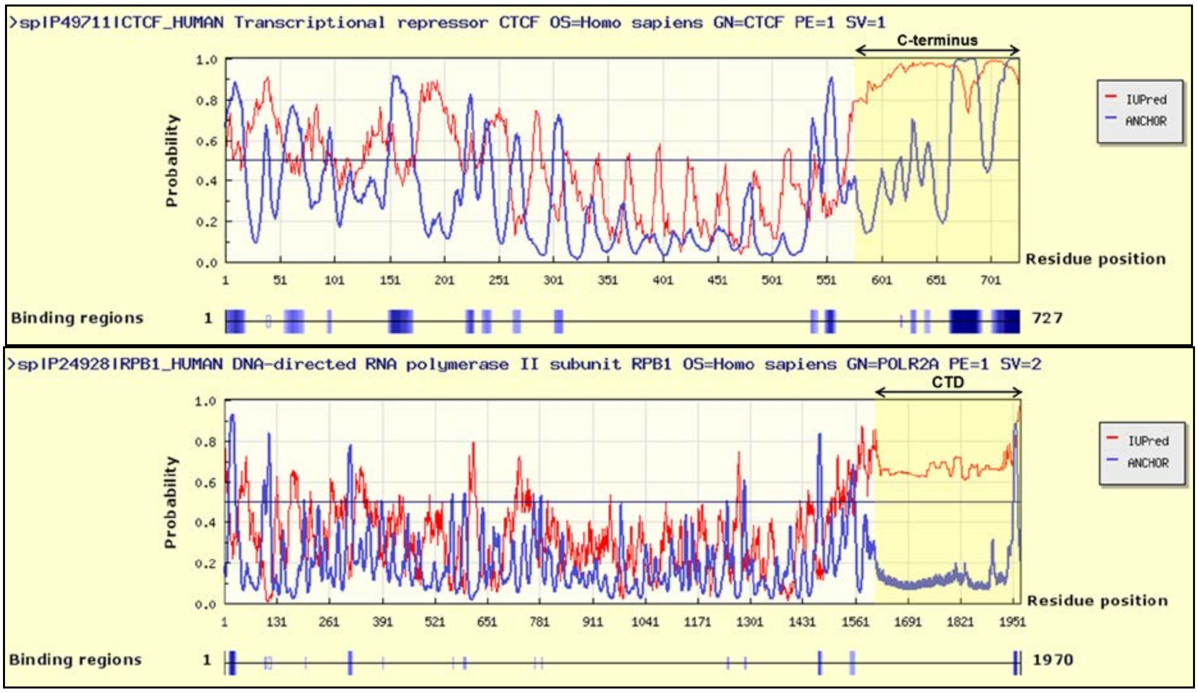

Figure 2. The disorder propensities and disorder-based binding regions in CTCF and RNA RNAPII. The disorder threshold is indicated as a thin line (at score =0.5) in all plots to show a boundary between disorder (>0.5) and order (<0.5) for IUPred and binding (>0.5) and non-binding (<0.5) regions by ANCHOR. Both C-terminus of CTCF and RNA RNAPII CTD were disordered and four binding regions were discovered in the C-terminus.

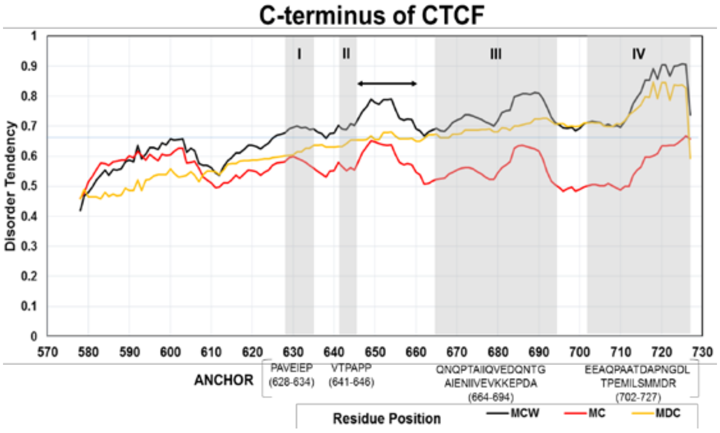

Figure 3A. Plot of MoRFs within C-terminus of CTCF by MoRF$_{\text{CHIBI\_Web}}$. Four MoRFs or binding regions were predicted by ANCHOR in the C-terminus as shown in the shaded regions. Through MoRF$_{\text{CHIBI_Web,}}$several parameters were also measured. MCW (in black) represents overall MoRF prediction propensity score. MC (in red) is the prediction score for local physiochemical properties of the amino acid sequence. MDC (in orange) is the prediction score which is based on protein disorder prediction and conservation information. Suggested cut-off value is 0.66 (Malhis et al., 2015). All predicted MoRFs had MCW score >0.66. MoRF$_{\text{CHIBI_Web}}$also predicted the presence of additional MoRFs as indicated by the arrow. The residues at third and fourth MoRFs were well conserved (MDC score >0.66).

Evaluation of IDRs in CTCF C-terminus and CTD of RNAP II

Both C-terminus and CTD are unstructured in native condition. *In silico* analysis was conducted to elucidate the functional features between these unstructured regions using the disorder-function paradigm (van der Lee *et al.*, 2014). The evaluation of disorder propensities by IUPred are displayed in Figure 2 in which both CTCF C-terminus and CTD of RNAPII were totally disordered, thus agreeing with the previous reports (Corden, 1990; Meinhart and Cramer, 2004; Martinez and Miranda, 2010).As both regions were unstructured, the MoRFs within IDRs could mediate the interactions between CTCF and RNAPII. Thus, the disorder-based protein binding regions or MoRFs were discovered within the IDRs through ANCHOR.

Discovery and examination of MoRFs in C-terminus of CTCF

There were four MoRFs were predicted in the C-terminus of CTCF (Figure 3A). Additionally, the sequences were also subjected to MoRF$_{\text{CHiBi}}$ web to evaluate the MoRFs prediction parameters which are MCW (based on propensity score), MC (based on local physiochemical properties of the amino acid sequence) and MDC (based protein disorder prediction and conservation information). All predicted MoRFs by ANCHOR had high MCW based on the suggested cut-off value (0.66), depicting accuracy of the regions being MoRFs (Malhis *et al.*, 2015). Notwithstanding, there were other regions with high MCW score in the MoRF$_{\text{CHiBi_}}$web plot. This is might due to the high specificity and sensitivity of MoRF$_{\text{CHiBi_}}$web compared to ANCHOR (Malhis *et al.*, 2015).

In addition, conservation of the MoRF residues than average IDR residues indicates the involvement of the functional residues in binding (Meszaros *et al.*, 2007). Based on the analysis by MoRF$_{\text{CHiBi_}}$web, the third and fourth predicted MoRFs had high MDC value (>0.66), depicting the regions to be well-conserved. Nevertheless, the cut-off value is not static and could not be used as the definite value for categorical prediction (Malhis *et al.*, 2015). Hence, all predicted MoRFs could still mediate the interactions with targeted proteins.

MoRFs from CTD of RNAPII complex structural model

The CTD in RNAP II is composed of up to 52 heptapeptiderepeats (Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7) which are disordered and important in mediating the interactions with other proteins. On the contrary, only one MoRF were predicted by ANCHOR within the CTD as shown in Figure 3B. The protein sequence was also subjected to MoRF$_{\text{CHiBi\_Web}}$ to analyse the MoRFs in the CTD. Figure 3B shows high MCW (>0.66) near the 37th to 52th heptapeptide while ANCHOR predicted the 52th heptapeptide to be disordered binding sire. As MoRF$_{\text{CHiBi\_Web}}$ has higher accuracy and sensitivity compared to ANCHOR, there could be possibilities for ANCHOR to erroneously treat some true MoRF residues with high prediction scores as non-MoRF residues (Malhis *et al.*, 2015). Hence, all heptapeptides could also be MoRFs in the CTD.

Based on previous study by Hsu *et al.* (2013), three MoRFs were discussed which mediate the interactions involving the CTD of RNAPII. They were examined from three structural model of RNAPII complex with CTD small phosphatase 1 (PDB: 2GHQ), protein 1 of cleavage and polyadenylation factor I (PDB: 1SZA) and mRNA capping enzyme alpha subunit (PDB: 1P16). However, only MoRF from the first complex is derived from human RNA polymerase II. Thus, we retrieved other complexes from PDB and 10 complexes were found from the PDB as listed in Table 1. Each interacting proteins were bound with the peptide or MoRFs in the disordered CTD (Hsu *et al.*, 2013). The peptides in the CTD complexes were characterized by phosphorylation of serine residues (Ser2, Ser5 and/or Ser7).Human Scp1 mutant interacted to singly (Ser5) and doubly phosphorylated (Ser2/Ser5) peptides spanning both a single and double CTD repeat (Zhang *et al.*, 2006). Binding of SCAF8 via CID towards the CTD of RNAPII is mediated by heptapeptide repeats with phosphorylated Ser2, Ser5 and Ser7 (Becker *et al.*, 2008). Another complex is the interaction of RPRD1A and RPRD1Bvia CTD interaction domains (CIDs) with CTD repeats phosphorylated at Ser2, and Ser7(Ni *et al.*, 2014). Interestingly, all the CTD peptides were the MoRFs that positioned near 28th to 30th heptapeptide repeats in CTD (Figure 3B).

Hypothetical interaction between MoRFs and ELM of CTCF C-terminus with RNAPII CTD

As both regions are unstructured, *in silico* analysis via structural modelling is unreliable when predicting the interactions between the IDRs. Currently there is no *in silico* tool can be used to predict the interaction between IDRs as their structure do not have binding surface in native condition. PepSite2 is a web server that could predict the peptide-mediated interactions from solved protein structure models (Trabuco *et al.*, 2012). However in this study, we attempted to predict the interactions of C-terminus of CTCF with the CTD peptide chain from the protein complexes in Table 1 as there is no solved protein structure for the CTD and RNAPII. As PepSite2 prediction is limited to 10-mer peptides, the analysis was performed in a sliding window of 10 residues to assess the binding of the third and fourth MoRFs of C-terminus with the CTD peptides.

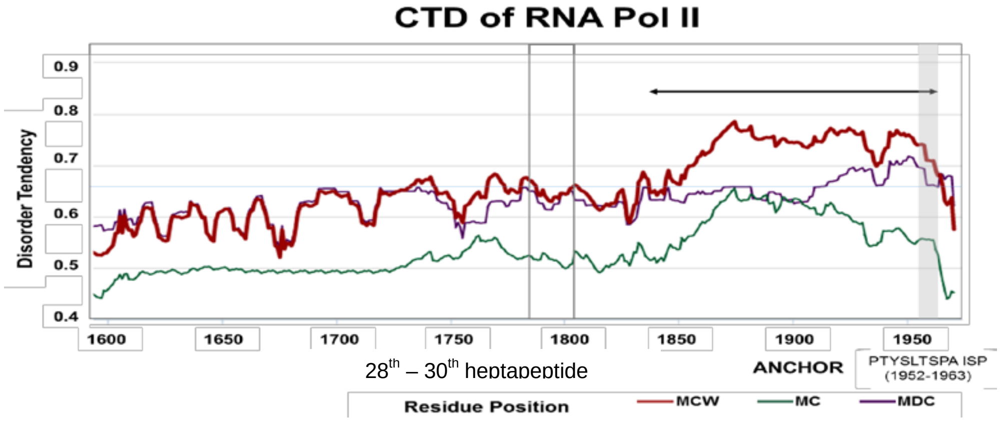

Figure 3B. Plot of MoRFs within CTD of RNAPII by MoRF$_{\text{CHIBI\_Web}}$. One MoRFwas predicted by ANCHOR as shown in the shaded regions. Suggested cut-off value is 0.66 (Malhis *et al.*, 2015).MoRF$_{\text{CHIBI_Web}}$predicted high MCW (>0.66) near the 37th to 52th heptapeptide as shown by the arrow. Meanwhile, the CTD peptides from RNAPII-protein complexes were the MoRFs that positioned near 28th to 30th heptapeptide repeats.

Table 1 List of RNAPII complexes with respective interacted proteins with the sequence, position and phosphorylation status of CTD peptide sequence in each complexes and chain. Abbrevitions: Scp1; Human synaptonemal complex protein 1, RPRD1A/1B; Regulation Of Nuclear Pre-MRNA Domain Containing 1A/1B, SCAF8; SR-Related CTD-Associated Factor 8.

#	RNAPII complex	Interacted protein	Chain	CTD peptide sequence	Position	Phosphorylation	References
1	2GHT	CTD-specific phosphatase Scp1	C	SPTSP	1798-1802	Ser-2/Ser-5	(Zhang *et al.*, 2006)
			D	SYSPTSPS	1796-1803		
2	2GHQ		C	SPTSP	1798-1802	Ser-5	
			D	PSYSPTSPS	1795-1803		
3	4JXT	human RPRD1A CID	B	PSYSPTSPSYS	1788-1798	Ser-7	(Ni *et al.*, 2014)
4	4Q94	human RPRD1B CID	C	PSYSPTSPSYS	1788-1798	Ser-2	
5	4Q96		C	PSYSPTSPSYS	1788-1798	Unphosphorylated	
6	3D9K	RNA processing factor SCAF8	Y	PSYSPT	1795-1800	Ser-2/Ser-5	(Becker *et al.*, 2008)
			Z	YSPTSPSYS	1790-1798		
7	3D9L		Y	YSPTSPSYSP	1790-1799	Ser-2	
			Z	PSYSPTSP	1795-1802		
8	3D9M		Y	YSP	1797-1799	Ser-5	
			Z	PSYSPTSPS	1795-1803		
9	3D9N		Y	PSYSPTSP	1795-1802	Ser-2/Ser-7	
			Z	PSYSPTSPS	1795-1802		
10	3D9O		Z	PSYSPTSPS	1795-1803	Unphosphorylated	

Based on the analysis, PepSite2 predicted lowest peptide binding score between the MoRFs and the CTD peptide in 3D9K-z and 3D9L-y complexes with score p-value < 0.15 as shown in Table 2. The CTD peptide from these complexes contained the double repeat peptides that started with Tyr1 and included the following repeat with Tyr8-Ser9 in 3D9K-z and Tyr8-Ser9-Pro10 in 3D9L-y. For the third and fourth MoRFs, the first ten residues displayed significant peptide binding score to the CTD peptides from both complexes. The lower the peptide binding score represents the accuracy of the prediction due to the strong signal. Even though the predictions could not produce any reliable score, this method performs with the efficiency of 55% of correct prediction (Petsalaki *et al.*, 2009).

Phosphorylation of motifs might implicate the interaction of C-terminus to CTD peptide

Post-translational modification especially phosphorylation is thought to regulate the interaction and binding affinity of CTCF to CTD of LS RNAPII (Chernukhin *et al.*, 2007). Hence, we would like to discover potential phosphorylation motifs in C-terminus that might implicate the interaction with the CTD peptide from both complexes.

ELM resource was used to retrieve the information about SLiMs in CTCF, particularly the phosphorylation motifs in the C-terminus. Four phosphorylation motifs were discovered which were MOD_CK1, MOD_CK2, MOD_NEK2_1 and MOD_ProDKin_1. MOD_CKI and MOD_CK2 are phosphorylation motifs for casein kinase 1 (CK1) and 2 (CK2) while the other two sites are targeted by Never in mitosis A (NimA)-related kinases (NEK) and Proline-directed kinases (ProDKin). Nevertheless, only the phosphorylation sites by CK2 within the C-terminal region are experimentally validated (Klenova *et al.*, 2001). Like MoRFs, SLiMs could also mediate the formation of protein complexes(Davey *et al.*, 2012). Hence, the motifs were subjected to the PepSite2 to predict the potential interaction with the CTD peptides.

Phosphorylated serine and threonine were also substituted with the residues in the motifs. The peptide binding scores for unphosphorylated and phosphorylated motifs in C-terminus with CTD peptide in 3D9K-z and 3D9L-y were tabulated in Table 3. The scores involved binding that include the unphosphorylated or phosphorylated serine/threonine residues. Based on the analysis, motifs that been phosphorylated by CK1, CK2 and NEK displayed reduction of *p*-value score compared to unphosphorylated motifs when interact with CTD peptide from the two complexes. As low peptide binding score indicates a better accuracy for the interaction to occur, phosphorylation of the motifs might contribute to the accuracy of the interaction. On the other hand, phosphorylation of threonine by ProDKin showed increment in the peptide binding score, causing the accuracy of the interaction become reduced. Here, phosphorylation of threonine in the second MoRF results with the increasing of these compared to the unphosphorylated status, thus decreasing the accuracy of the interaction.

Table 2 Prediction of interaction between MoRFs in C-terminus with CTD peptide from 3D9K-z and 3D9L-y complexes by PepSite2. These interactions presented lowest peptide binding score compared to the interactions with CTD peptide from other complexes. The lower the *P*-value indicate the accuracy of the interactions.

	MoRFs in C-terminus of CTCF	**Position**	**Peptide binding score (P-value)**	
			3D9K-z	**3D9L-y**
I	PAVEIEP	628-634	0.1279	0.1088
II	VTPAPP	641-646	0.07398	0.08498
III	QNQPTAIIQVEDQNTGAIENIIVEVKKEPDA	664-694	01972	0.197
IIIa	qnqptaiiqv	664-653	0.07652	0.04308
IIIb	edqntgaien	654-663	0.1937	0.2042
IIIc	iivevkkepd	664-673	0.3216	0.3437
IV	EEAQPAATDAPNGDLTPEMILSMMDR	702-727	0.1398	0.1766
IVa	eeaqpaatda	702-711	0.0757	0.08458
IVb	pngdltpemi	712-720	0.176	0.1965
IVc	lsmmdr	721-727	0.1676	0.2487

Table 3 Prediction of interaction between phosphorylation motifs in C-terminus with CTD peptide from 3D9K-z and 3D9L-y complexes by PepSite2. Four phosphorylation motifs which were MOD_CK1, MOD_CK2, MOD_NEK2_1 and MOD_ProDKin_1. Phosphorylation of the serine residues by CK1, CK2 and NEK2 were predicted to reduce the binding score and improve the accuracy, in contrast to the phosphorylation by MOD_ProDKin_1. The scores denoted by (*) were not the first rank match. Each binding involved interactions with the serine and threonine residues.

	Phosphorylation motifs	**Position**	**ELM**	**Peptide binding score (P-value)**	
				3DKz	**3D9Ly**
I	KMRSKKE	601-607	MOD_CK2	0.2058	0.2958
	KMRjKKE	*S-604*		0.1909	0.2315
II	EDSSDSE	607-613	MOD_CK2	0.3279	0.2484
	edSsdse	S-609		0.279	0.193
	edsSdse	S-610		0.279	0.2289
	edssdSe	S-612		0.266	0.2041
III	SSDSENA	609-615	MOD_CK1	0.1888	0.2463
	Ssdsena	S-609		0.1899	0.2052
	sSdsena	S-610		0.2153	0.1778
	ssdSena	S-612		0.1531	0.2003
IV	VTPAPP	628-634	MOD_ProDKin_1	0.09865*	0.08498
	vTpapp	641-646		0.1289	0.0941
V	QPVTPAP	639-645	MOD_ProDKin_1	0.03843*	0.03393
	qpvTpap	T-642		0.0432	> 0.1
VI	GDLTPEM	714-720	MOD_ProDKin_1	0.1207	0.1689*
	gdlTpem	T-717		0.1341	0.2008*
VII	MILSMM	720-725	MOD_NEK2_1	0.2317*	0.2921
	milSmm	S-723		0.2078	0.2566

DISCUSSION

Endogenous largest subunit of RNAPII, LS RNAPII (Rpb1) expression was detected using anti-RNAPII antibody which was raised against the CTD region. In this experiment, the endogenous expression of RNAPII was detected as a doublet with molecular weights of 220 and 240 kDa in SDS-PAGE. These two subunits of RNAPII were labeled as hypophosphorylated (LS RNAPIIa) and hyperphosphorylated (LS RNAPIIo) respectively. The phosphorylation of RNAPII occurs at the CTD that contained multiple phosphorylation sites (Dahmus, 1995; Dahmus, 1996).

The earlier study proved the interaction of LS RNAPII with CTCF through its CTD (Chernukhin *et al.*, 2007). Previous reports have stated that CTD appeared to frame itself to its binding partner, adopting different conformations during the interaction. The flexibility of CTD binding to its target sequences, combined with post-translational modification by phosphorylation, provides a way for the CTD to interact with multiple structure dissimilar partners.

Post-translational modification especially phosphorylation is thought to regulate the interaction of CTCF with CTD of LS RNAPII. Based on the CO-IP result, both forms of LS RNAPII were retained by CTCF which could be due to the lack of post-translational modifications in the C-terminus (Chernukhin *et al.*, 2007). Notably, phosphorylation of C-terminus with casein kinase 2(CK2) is found to reduce the binding affinity with LS RNAPII (Chernukhin *et al.*, 2007). In addition, CTCF mediates RNAPII pausing during alternative splicing of CD45 and implicates the RNAPII elongation dynamics (Shukla *et al.*, 2011). Elongation event is carried out by LS RNAPII0 which is in hyperphosphorylated form where phosphorylation of Ser2 and Thr4 phosphorylation would occur in the CTD (Hsin and Manley, 2012). Henceforth, the interaction between the C-terminus could also be determined by the phosphorylation status of CTD. Nevertheless, the mechanisms of this interaction is still unclear.

In order to deduce the potential binding sites within the C-terminus and CTD, *in silico* analysis was conducted based on their features as unstructured domain. Four MoRFs were predicted in the C-terminal domain which gave the lowest peptide binding score (*p*-value <0.15) and more accuracy with the phosphorylated CTD peptide from RNAPII complex with SCAF8 (structural model: 3D9K-z and 3D9L-y). In addition, the CTD peptide in 3D9K-z was phosphorylated at Ser2(Ser(P)2) while the peptide was phosphorylated at Ser2and Ser5 (Ser(P)2 and Ser(P)5) in 3D9L-y.Ser(P)2 indicates elongation while both Ser(P)2 and Ser(P)5 could mediate the late elongation phase during transcription cycle(Mayfield *et al.*, 2016). In addition, SCAF8 is bound to three consecutive phosphoserine residues Ser(P)2-Ser(P)5-Ser(P)9 (3D9K-z) in the CTD of elongating RNAPII with highest affinity(Becker *et al.*, 2008). Henceforth, it could be assumed that the interaction could occur in the similar manner which correlate with the implication of CTCF to the dynamic elongation phase by RNAPII (Shukla *et al.*, 2011).

In contrast with the CTD peptides from other complexes, the CTD peptide from 3D9K-z and 3D9L-y complexes contained the double repeat peptide that started with Tyr1 and included the following repeat with Tyr8-Ser9 in 3D9K-z and Tyr8-Ser9-Pro10 in 3D9L-y. We would like to assume that the potential MoRFs in C-terminus might require to the double peptide region that starts with Tyr1 to interact with the CTD. This region serves as functional unit of the CTD which comprises one full heptapeptide repeat including the next four residues of the following repeat (Eick and Geyer, 2013).Nevertheless, the conformation of CTD peptides were fixed with the binding-site environment in the globular protein partners (Petsalaki *et al.*, 2009). In both models, the CTD peptide is bound to the CID which is the scaffold for CTD-binding proteins. The CID has well-defined structure which consists of eight α-helixes that are arranged into a right-handed super helix (Meinhart and Cramer, 2004). This will be the limitation of the accuracy of the interaction analysis using PepSite2.

Formation of complexes involving IDPs are radically diverged from the complexes formed by ordered proteins (Uversky and Dunker, 2013). Aside of common insight of disorder-to-order transition that occur after the interaction of IDRs with globular protein, structure formation might not occur after interaction between IDRs whereby some of the residues would remain unstructured after binding due to the lack of electron density in the crystal structures (Mohan *et al.*, 2006). In this case, there is possibility that the unstructured C-terminus of CTCF could interact with the unstructured CTD without specific structure conformation in the final complex. As reviewed by Uversky and Dunker (2013), there are three alternative mechanisms for the interactions to occur between the disordered regions and one of them is via electrostatic interactions. It is thought that electrostatic interactions keep the motifs of an IDP to the binding sites of its disordered protein partner, mediating the interactions between the IDPs (Borg *et al.*, 2007). More importantly, electrostatic interaction between a dynamic IDP and its partner is correlated with the binding affinity mediated by phosphorylation motifs (Borg *et al.*, 2007).

As phosphorylation is thought to mediate the regulation of CTCF and LS RNAPII interaction and binding affinity, the conformational ensemble might be altered by phosphorylation adjacent to the binding motif. Not to mention that post-translational modification, especially phosphorylation can be classified as a functional switch which plays major role in major part in modulating the conformational ensemble and interactions of IDPs (Wright and Dyson, 2015). In this study, phosphorylation sites were found frequently located in the C-terminus by the ELM resource and while CTD is regulated extensively by phosphorylation (Dahmus, 1993; Dahmus, 1995; Dahmus, 1996). Functional phosphorylation of C-terminus by CK2 is associated for reduction of CTCF activity (Klenova *et al.*, 2001) while another study reported its role in switching the CTCF as repressor to activator (El-Kady and Klenova, 2005). Based on the result, phosphorylation by CK2 at the motifs resulted with the decreasing of peptide binding score indicating better accuracy of the interaction. Although CK2 phosphorylation is assumed to reduce the binding of CTCF to RNAPII (Chernukhin *et al.*, 2007), there is no clear explanation about the finding. As interactions between IDRs could be transient, the downstream experiments on these complexes would be challenging and there is been a tendency to over-interpret the results of in-cell experiments (Gibson *et al.*, 2015).

Ser612 resulted with the lowest peptide binding score compared to other phosphorylated residues when interacted with Ser(P)2 and Ser(P)5 CTD peptide (3D9K-z). Ser612 has been identified as a critical residue in the functional regulation by phosphorylation whereas Ser604 is not critical for CTCF function (El-Kady and Klenova, 2005). On the other hand, binding of motifs with phosphorylated Ser609 by CK2to the CTD peptide with Ser(P)2 (3D9L-y) displayed low binding score compared to other residues. Nonetheless, phosphorylation of Ser609 could also be regulated by Ser612 (El-Kady and Klenova, 2005). Besides CK2, phosphorylated residues by CK1 and NEK2 also resulted with reduction of binding score whereas phosphorylation of ProDKin in C-terminus increased the *P*-value. Unfortunately, phosphorylation by other protein kinases is unknown and no experimental validation is carried out so far. As modifications of and IDR/IDP by different kinases can result in different signaling outputs (Wright and Dyson, 2015), it is imperative to validate the functional phosphorylation of these motifs in C-terminus.

CONCLUSION

This study is our attempt to use peptide docking tool, PepSite2 to predict possible interactions between IDRs in C-terminus of CTCF and CTD of RNAPII. With *ex vivo* interaction between the domains was confirmed in this study and previous study by Chernukhin *et al.* (2007), it is important to elucidate the regulations and mechanisms of this interaction. Here, we obtain preliminary insight on the possibilities for C-terminus to interact with the functional unit of CTD via the linear motifs and/or the MoRFs. Notably, both phosphorylation status of the CTD functional unit and the modifications in the C-terminus of CTF are critical to the binding affinity of this protein-protein interactions. As PepSite2 could only analyses the interaction with the fixed-position peptides, our results are imperfect to infer the accuracy of the interaction as other factor such as electrostatic interaction could play major role in IDPs interaction. Nevertheless, this *in silico* analysis serves as the initiation to gain insight about the regulation of CTCF and RNAPII interactions based on their functional features as IDPs.

Acknowledgement

This work was funded by USM Short term grant (grant number: 304.CIPPM.6315051).

Conflict of Interest

No conflict of interest in this study.

REFERENCES

1. Bataille, A. R., Jeronimo, C., Jacques, P.-É., Laramée, L., Fortin, M.-È., Forest, A., Bergeron, M., Hanes, S. D. & Robert, F. (2012). A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Molecular cell,* **45(2),** 158-170.
2. Becker, R., Loll, B. & Meinhart, A. (2008). Snapshots of the RNA processing factor SCAF8 bound to different phosphorylated forms of the carboxyl-terminal domain of RNA polymerase II. *J Biol Chem,* **283(33),** 22659-69.
3. Borg, M., Mittag, T., Pawson, T., Tyers, M., Forman-Kay, J. D. & Chan, H. S. (2007). Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci U S A,* **104(23),** 9650-5.
4. Buratowski, S. (2003). The CTD code. *Nature Structural & Molecular Biology,* **10(9),** 679-680.
5. Chernukhin, I., Shamsuddin, S., Kang, S. Y., Bergstrom, R., Kwon, Y. W., Yu, W., Whitehead, J., Mukhopadhyay, R., Docquier, F., Farrar, D., Morrison, I., Vigneron, M., Wu, S. Y., Chiang, C. M., Loukinov, D., Lobanenkov, V., Ohlsson, R. & Klenova, E. (2007). CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol Cell Biol,* **27(5),** 1631-48.
6. Corden, J. L. (1990). Tails of RNA polymerase II. *Trends in biochemical sciences,* **15(10),** 383-387.
7. Corden, J. L., Cadena, D. L., Ahearn, J. M. & Dahmus, M. E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proceedings of the National Academy of Sciences,* **82(23),** 7934-7938.
8. Dahmus, M. E. (1993). The role of multisite phosphorylation in the regulation of RNA polymerase II activity. *Progress in nucleic acid research and molecular biology,* **48,** 143-179.
9. Dahmus, M. E. (1995). Phosphorylation of the C-terminal domain of RNA polymerase II. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression,* **1261(2),** 171-182.
10. Dahmus, M. E. (1996). Phosphorylation of mammalian RNA polymerase II. *Methods in enzymology,* **273,** 185-193.
11. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H. & Gibson, T. J. (2012). Attributes of short linear motifs. *Mol Biosyst,* **8(1),** 268-81.
12. Dinkel, H., Van Roey, K., Michael, S., Kumar, M., Uyar, B., Altenberg, B., Milchevskaya, V., Schneider, M., Kühn, H. & Behrendt, A. (2015). ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Research***,** gkv1291.
13. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics,* **21(16),** 3433-3434.
14. Dosztányi, Z., Mészáros, B. & Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics,* **25(20),** 2745-2746.
15. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M. & Hipps, K. W. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling,* **19(1),** 26-59.
16. Eick, D. & Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem Rev,* **113(11),** 8456-90.
17. El-Kady, A. & Klenova, E. (2005). Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS letters,* **579(6),** 1424-1434.
18. Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J. & Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology,* **16(6),** 2802-2813.
19. Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L. & Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature,* **529(7584),** 110-114.
20. Gibson, T. J., Dinkel, H., Van Roey, K. & Diella, F. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun Signal,* **13(1),** 42.
21. Hsin, J.-P. & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development,* **26(19),** 2119-2137.
22. Hsu, W. L., Oldfield, C. J., Xue, B., Meng, J., Huang, F., Romero, P., Uversky, V. N. & Dunker, A. K. (2013). Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci,* **22(3),** 258-73.
23. Klenova, E. M., Chernukhin, I. V., El-Kady, A., Lee, R. E., Pugacheva, E. M., Loukinov, D. I., Goodwin, G. H., Delgado, D., Filippova, G. N. & León, J. (2001). Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Molecular and cellular biology,* **21(6),** 2221-2234.

24. Malhis, N. & Gsponer, J. (2015). Computational identification of MoRFs in protein sequences. *Bioinformatics,* **31(11),** 1738-1744.
25. Malhis, N., Wong, E. T., Nassar, R. & Gsponer, J. (2015). Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. *PLoS One,* **10(10),** e0141603.
26. Martinez, S. R. & Miranda, J. (2010). CTCF terminal segments are unstructured. *Protein Science,* **19(5),** 1110-1116.
27. Mayfield, J. E., Burkholder, N. T. & Zhang, Y. J. (2016). Dephosphorylating eukaryotic RNA polymerase II. *Biochim Biophys Acta,* **1864(4),** 372-387.
28. Meinhart, A. & Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature,* **430(6996),** 223-226.
29. Meszaros, B., Tompa, P., Simon, I. & Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol,* **372(2),** 549-61.
30. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K. & Uversky, V. N. (2006). Analysis of molecular recognition features (MoRFs). *Journal of molecular biology,* **362(5),** 1043-1059.
31. Nakahashi, H., Kwon, K.-R. K., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S. & Yamane, A. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell reports,* **3(5),** 1678-1689.
32. Ni, Z., Xu, C., Guo, X., Hunter, G. O., Kuznetsova, O. V., Tempel, W., Marcon, E., Zhong, G., Guo, H., Kuo, W. H., Li, J., Young, P., Olsen, J. B., Wan, C., Loppnau, P., El Bakkouri, M., Senisterra, G. A., He, H., Huang, H., Sidhu, S. S., Emili, A., Murphy, S., Mosley, A. L., Arrowsmith, C. H., Min, J. & Greenblatt, J. F. (2014). RPRD1A and RPRD1B are human RNA polymerase II C-terminal domain scaffolds for Ser5 dephosphorylation. *Nat Struct Mol Biol,* **21(8),** 686-95.
33. Oltean, S. & Bates, D. (2014). Hallmarks of alternative splicing in cancer. *Oncogene,* **33(46),** 5311-5318.
34. Ong, C. T. & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet,* **15(4),** 234-46.
35. Petsalaki, E., Stark, A., Garcia-Urdiales, E. & Russell, R. B. (2009). Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput Biol,* **5(3),** e1000335.
36. Phillips, J. E. & Corces, V. G. (2009). CTCF: master weaver of the genome. *Cell,* **137(7),** 1194-211.
37. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. & Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature,* **479(7371),** 74-9.
38. Thorne, A. H., Cavenee, W. K. & Furnari, F. B. (2015). Alternative RNA Splicing in the Pathogenesis of GBM. *Medical Research Archives***(1)**.
39. Trabuco, L. G., Lise, S., Petsalaki, E. & Russell, R. B. (2012). PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Res,* **40(Web Server issue),** W423-7.
40. Uversky, V. N. & Dunker, A. K. (2013). The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. *F1000 Biol Rep,* **5,** 1.
41. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J. & Jones, D. T. (2014). Classification of intrinsically disordered regions and proteins. *Chemical reviews,* **114(13),** 6589-6631.
42. Wright, P. E. & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology,* **16(1),** 18-29.
43. Zhang, Y., Kim, Y., Genoud, N., Gao, J., Kelly, J. W., Pfaff, S. L., Gill, G. N., Dixon, J. E. & Noel, J. P. (2006). Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Mol Cell,* **24(5),** 759-70.