Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites

Alan M. Moses\textsuperscript{a}, Muluye E. Liku\textsuperscript{b}, Joachim J. Li\textsuperscript{c}, and Richard Durbin*  

\textsuperscript{a}Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HH, United Kingdom; and Departments of \textsuperscript{b}Biochemistry and \textsuperscript{c}Microbiology and Immunology, University of California, San Francisco, CA 94143

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved September 25, 2007 (received for review February 6, 2007)

Evolutionary change in gene regulation is a key mechanism underlying the genetic component of organismal diversity. Here, we study evolution of regulation at the posttranslational level by examining the evolution of cyclin-dependent kinase (CDK) consensus phosphorylation sites in the protein subunits of the pre-replicative complex (RC). The pre-RC, an assembly of proteins formed during an early stage of DNA replication, is believed to be regulated by CDKs throughout the animals and fungi. Interestingly, although orthologous pre-RC components often contain clusters of CDK consensus sites, the positions and numbers of sites do not seem conserved. By analyzing protein sequences from both distantly and closely related species, we confirm that consensus sites can turn over rapidly even when the local cluster of sites is preserved, consistent with the notion that precise positioning of phosphorylation events is not required for regulation. We also identify evolutionary changes in the clusters of sites and further examine one replication protein, Mcm3, where a cluster of consensus sites near a nucleocytoplasmic transport signal is confined.


development of cyclin-dependent kinase regulation by turnover and lineage-specific changes of sites

The contribution of regulatory evolution to biological diversity is increasingly well appreciated (1–4). The identification of changes in transcriptional regulatory proteins (5, 6) and, more frequently, the cis-elements they recognize in noncoding DNA (reviewed in ref. 7), has provided mechanistic insight into the evolution of gene regulation.

Genes are regulated at multiple levels, however. In eukaryotes, posttranslational regulation of protein activity by phosphorylation is of particular importance (8). Although little is known in general about the evolution of this type of regulation, comparative studies of posttranslational modification sites in phosphorylase (9, 10) and fructose 1-6-bisphosphatase (11) revealed that they were not conserved between homologues.

Recent studies have applied computational approaches to databases of protein sequences to perform comparative studies on larger scales. For example, targets of protein kinase A were predicted based on conservation of consensus sites between Candida albicans and Saccharomyces cerevisiae (12). Another study examined regulation of cell-cycle proteins in four species and proposed coevolution between posttranslational regulation by phosphorylation and transcriptional regulation (13).

Phosphoregulation plays a critical role in cell-cycle control (14–16). For example, it has been found in several species that after the initiation of DNA replication, to ensure that a single round of DNA replication occurs in each eukaryotic cell cycle, a subset of the DNA replication machinery (the pre-RC) is directly inhibited by cyclin-dependent kinase (CDK) (17, 18).

Here, we examine the evolution of regulation of the pre-RC by CDKs. Several features of this system make it attractive for evolutionary analysis. First, the pre-RC proteins are found in single copy in many animals and fungi (17), so it is relatively easy to identify their orthologs in most species. Also, human CDKs have been shown to rescue yeast CDK mutations (19, 20), suggesting little change in the functional capabilities of the kinase. Finally, CDK is a proline-directed serine/threonine kinase (21) with a well defined consensus site S/T-P-X-R/K (where X is any amino acid). Evolutionary loss of the critical S/T or P is likely to preclude phosphorylation by CDK in that species.

In some cases, the specific consensus sites likely to be phosphorylated by CDK in vivo have been determined through a combination of experimental methods; we refer to these sites as “characterized.” In addition, CDK target proteins often contain multiple CDK consensus sites closely spaced in their primary amino acid sequence; we refer to these as “clusters.” Previous studies have noted that, even when clusters of characterized sites are found in orthologous pre-RC components, the individual consensus sites are not always conserved in position or number (22, 23). We refer to this as “turnover” of sites and suggest that it is consistent with regulation through mechanisms that impose loose constraints on spacing and number of phosphorylation sites (ref. 24; see Discussion).

Our analysis of evolutionary changes in CDK consensus sites in pre-RC proteins reveals examples of both turnover of characterized sites in preserved clusters and lineage-specific changes in the clusters of sites. We suggest that the CDK regulation of nuclear localization of the pre-RC component Mcm3 (25) was gained on the lineage leading to S. cerevisiae after the divergence from C. albicans, and we provide experimental support for this model.

Results

Signatures of CDK Regulation in Pre-RC Proteins. To get a broad sense of the conservation of CDK regulation, we obtained...
sequences and orthologs for pre-RC proteins (see Methods) from 21 species with complete genome sequences publicly available, selected so that their phylogenetic positions were informative amongst the animals and fungi. For each protein, we identified experimentally verified CDK targets where consensus sites had been characterized ("P"s in Fig. 1 and refs. 22, 23, and 25–38) and also calculated the $S_{LR}$ statistic (see Methods), which measures the overrepresentation and spatial clustering of strong (S/T-P-X-R/K, where X is any amino acid) and weak (S/T-P) consensus matches (Fig. 1), which we have shown to be predictive of CDK regulation (39). Because the pre-RC is expected to be regulated by CDKs in all these species, a simple expectation is that the same proteins would be targets in all species. Indeed, we find proteins that have high values of $S_{LR}$ across many (Orc1, Mcm4) or all (Cd6) of the species examined, suggesting that regulation has been preserved since a common ancestor. However, other proteins (Orc2) show less consistent patterns, while some (Orc6, Mcm2, Mcm3) show lineage-specific patterns. In these cases, the changes in statistical signal could be due to either bona fide changes in regulation or incorrect classification by our statistical method, but in at least one of these cases, we see a functional difference corresponding to the statistical difference (see below).

**Turnover of Functional CDK Consensus Sites.** Even when regulation appears conserved, as has been noted in previous studies (22, 23), we found that the numbers and positions of CDK consensus sites were not always conserved. A striking example of this is the linker region of ORC1, which contains a strong cluster of CDK consensus matches in all of the animals and most of the fungi (Figs. 1 and 2a). Sites in this region are phosphorylated by CDK in *Drosophila* (23) and are involved in CDK-regulated localization and degradation of ORC1 in mammalian cells (38, 40, 41). Despite the persistence of the cluster over long evolutionary distances, examination of the numbers and positions of individual CDK consensus sites (Fig. 2a) reveals rapidly changing organization.

It is possible that this apparent turnover of sites is due simply to difficulties in comparing highly diverged amino acid sequences, or that consensus matches in clusters do not all represent functional sites and are not constrained. To rule these out, we examined the evolution of experimentally characterized consensus sites. We consider a consensus site characterized if there is some *in vivo* (including cell culture) evidence of phosphorylation and/or function in a CDK-regulated process ["P" in Fig. 1 and supporting information (SI) Table 1]. We examined these sites in alignments of orthologs from closely related species (see Methods), where most residues are unchanged and we have high-confidence in multiple alignments (84%, 74%, and 64% identical for yeast, mammals, and *Drosophila*, respectively).

We found in each clade that characterized consensus sites accumulated on average fewer substitutions than the flanking residues (rates were 20%, 60%, and 27% of flanking regions for yeast, mammals, and *Drosophila*, respectively).
CDK consensus sites on the lineage leading to MCM3, we find a dramatic statistical change in the clustered C-terminal region of or human. 

In contrast to cases like S. cerevisiae, significantly greater than the 5.37 expected if gains were randomly distributed proportional to the evolutionary distance on each branch (P = 0.0037, see Methods). For the weak consensus, we inferred 13 gains in this clade, which also greater than the 10.36 expected but is not statistically significant (P = 0.24). These data show that gains of strong consensus matches are nonrandomly distributed along the tree and suggest that the CDK-regulated shuffling of MCMs in and out of the nucleus in S. cerevisiae is due at least in part to changes in CDK consensus sites that occurred after the divergence from C. albicans (Fig. 3a).

This model predicts that the region homologous to the Sc-Mcm3-CTR from species outside this clade would not confer regulated localization to a GFP reporter construct. We therefore inserted the homologous region of C. albicans Mcm3 into such a construct (Fig. 3b) and tested its localization in S. cerevisiae in cells arrested in G1 (by alpha factor) or G2 (by nocodazole). Although the S. cerevisiae construct showed nuclear localization in the G1 but not the G2 arrest (Fig. 3d, compare iv with viii), confirming a functional difference in this region of the protein between these species. To further resolve the evolutionary events that lead to regulated localization of MCMs in S. cerevisiae, we performed similar experiments, using the C-terminal region of Mcm3 from Candida glabrata (Fig. 3e, ii and vi) and Kluveromyces lactis (Fig. 3e, iii and vii) and found that these showed regulated nuclear localization, consistent with the origin of this regulation in the ancestor of the S. cerevisiae clade.
Taken together, our experiments and sequence analysis of the Mcm3 C terminus are consistent with the model that the functional CDK consensus sites that regulate nuclear localization arose in the common ancestor of the S. cerevisiae clade after divergence from C. albicans. It is important to note that the Sc-Mcm3-CTR contains other regulatory sequences (ref. 25 and SI Fig. 5a), including a Crm1-dependent export signal, which also appeared at that time, and a basic nuclear localization signal, which is shared by all ascomycetes and may be important for the observed nuclear localization of the MCMs in S. pombe (46). Consistent with this model, we identify a basic nuclear localization signal, but no leucine rich export signals in the homologous region of the C. albicans protein (SI Fig. 5b and c). To rule out the possibility that there were cryptic export signals or CDK sites further downstream of the region we defined as homologous to the Sc-Mcm3-CTR, we also performed all of the experiments, using the entire C terminus from each species and found similar results (data not shown).

**Discussion**

Inhibition of the pre-RC by CDKs to prevent rereplication is an ancient feature of the eukaryotic cell cycle (17). Our results suggest that, even though this regulatory logic is preserved, its mechanistic implementation can evolve rapidly.

For example, we found that, on average, 16% (11–21% ± SE) of characterized CDK consensus sites in pre-RC components in budding yeast, human, and Drosophila are not conserved in alignments of closely related species. In ORC1, the presence of polymorphisms in the human population suggests that the reshuffling of regulatory regions continues.

Traditional models of phosphoregulation invoke allosterically driven conformational changes as a consequence of phosphorylation, which presumably require modification at precise positions in the protein structure. More recent analyses of phosphoregulation suggest alternative regulatory paradigms involving multiple phosphorylation sites that do not need to be conserved (24). Clusters of multiple phosphorylation sites can modulate interactions (25, 47–49) or provide specific dynamic properties (50–52) and these mechanisms may not depend on the specific locations or numbers of sites (24).

Consistent with this model, we found statistical evidence for constraint at the level of the cluster of consensus sites in the linker region of ORC1, despite weak constraint at the amino acid level. In clusters, when new consensus sites appear via point mutations, constraints on the ancestral sites may be relaxed, allowing them to accumulate destructive substitutions. Interestingly, this stabilizing selection model was first proposed for transcriptional enhancer elements in DNA, where, despite little similarity in primary sequence, orthologous enhancers could drive similar expression patterns by preserving clusters of transcription factor binding sites (53, 54).

In addition to turnover of consensus sites in conserved clusters, we found cases of entire clusters that are not conserved over evolution. We observed lineage-specific accumulation of consensus sites in the C terminus of S. cerevisiae Mcm3, which we showed was associated with functional differences in localization of a reporter construct (Fig. 3). We also note that the C-terminal cluster of consensus sites in yeast Cdc6 (29) shows a similar pattern, appearing even more recently (SI Fig. 5d). Because CDK inhibits the pre-RC through multiple regulatory mechanisms (35), we suggest that new mechanisms may evolve without drastic negative consequences. Thus, a possible explanation for these lineage-specific changes is “regulatory network turnover” (55), in which interactions are gained and lost in the context of a preserved regulatory logic.

Finally, we note that the accretion of regulatory motifs in the Mcm3 C terminus is analogous to the evolutionary gain of transcription factor binding sites in enhancers (56). In extending this model to phosphorylation sites, we suggest that the cooptation of a new target into an existing regulatory network by acquisition
of motifs for preexisting, trans-acting factors is a general mechanistic basis for evolutionary increases in regulatory specificity and, perhaps, organismal complexity.

Methods

Proteins, Orthologs, and Clustering of CDK Sites. For the animal and yeast genomes used in Fig. 1 i and iii, protein sequences and ortholog assignments were obtained from the TreeFam database (57) and Yeast Gene Order Browser (58), respectively. To assign orthologs for the species not included in these databases (Fig. 1 ii), we obtained amino acid sequences from J. Stajich (University of California, Berkeley, CA; http://fungal.genome.duke.edu).

We then aligned the fungal and animal orthologs (from TreeFam or Yeast Gene Order Browser), using T-Coffee software (59), created profile-hidden Markov models, and searched the additional genomes for matches to these profiles, using the HMMer package (http://hmmer.janelia.org, using the -forward option). We took the top hit as the ortholog in each case, except for CDC6, where the top hit was the same as the top hit for ORC1 in one of the fungi, so we took the second hit. Where a protein was present in multiple copies in a species (e.g., CDC7 in S. pombe), we excluded that protein for that species from further analyses (gray box in Fig. 1). If the HMMer e-value was >0.001 or the protein was truncated relative to other orthologs, we deemed the ortholog low confidence (gray box in Fig. 1).

For each protein in each species, we computed $S_{LR}$, a log likelihood ratio statistic, which measures clustering and enrichment of motifs in a sequence. Briefly, this statistic compares the likelihood of the observed motifs and their spacing under a model that includes clusters to that under the genomic background frequency or a model, including clusters of weak sites only (for details, see ref. 39). We computed the background frequencies of these motifs in each of the genomes studied. We reported the analysis shown in Fig. 1 by using other statistical measures and found similar results (SI Fig. 6).

Alignments of Closely Related Species. We obtained ortholog assignments and protein sequences for each of the characterized CDK targets from budding yeast in S. paradoxus, S. mikatae, and S. bayanus from SGD (60), from human in mammals from TreeFam or from Drosophila from 12 Drosophilidae (V. Iyer, D. Pollard, and M. Eisen, personal communication). These were aligned with T-Coffee, and truncated orthologs were removed, except in the case of mammalian CDT1, where only the N-terminal region was available. Alignments of all of the characterized sites are available as SI Dataset 1. To compute the $dn/ds$, we obtained coding DNA sequences and inserted the gaps from the protein alignments into these. For the linker region of ORC1 (which we took to be amino acids 196–470 in the human sequence), we used paml (61) to compute maximum-likelihood branch lengths with either an unknown $dn/ds$ or $dn/ds$ fixed at 1, assuming the phylogeny (((human,chimp),macaque),(mouse,rat),dog,cow). We compared two twice the difference in likelihoods to a chi-squared distribution with one degree of freedom. Human SNPs and alternative mouse transcripts for ORC1 were obtained from Ensembl (version 41; ref. 62). We note that $dn/ds$ for the clusters of CDK sites were higher on average than the whole proteins, with ORC1 showing the highest value (data not shown).

Simulations of Orc1 Evolution. To obtain the distribution of the difference of the mean and standard deviation of $S_{LR}$ for the ORC1 linker used the following procedure. We extracted the amino acid alignment and used paml (61), using the mammalian phylogeny described above to obtain the maximum-likelihood estimates for the branch lengths (in amino acid substitutions per site) and to reconstruct the ancestral sequence. We then used the ROSE sequence evolution software (63) to simulate (with default parameters for protein evolution) along the estimated tree starting from the ancestral sequence. Finally, we computed the average and standard deviation of the $S_{LR}$ in the simulated sequences for the extant species.

Reconstruction of Ancestral Mcm3 CDK Matches. Because we wanted to reconstruct the ancestral organization of CDK matches in Mcm3 over longer evolutionary distances where we were no longer confident in the alignment of individual residues, we devised the following parsimony method. First, we obtained protein predictions for six additional Ascomycete genomes (http://fungal.genome.duke.edu/), assigned orthologs as above, made a multiple alignment of the protein sequences, using T-Coffee, and used paml to obtain maximum-likelihood estimates of the branch lengths for the tree topology shown in Fig. 3A. We then searched the aligned sequences for matches to the CDK consensus and created an “alignment” of CDK consensus matches by treating any CDK match within five amino acid residues as another in a different species as “aligned.” For Mcm3, this yielded 31 aligned “columns,” where there was a match to either the strong or weak CDK consensus in at least one species. Based on this, we used the “classical parsimony” algorithm (64) to reconstruct the ancestral states, either “strong match,” “weak match,” or “background” and infer the number of gains and losses for strong and weak matches along each branch.

Although the current view supports the clade containing K. lactis, Ashbya gossypii, Kluyveromyces waltii, and Saccharomyces kleyveri as a sister to the clade containing S. cerevisiae (65, 66), the placement of the species (Fig. 3A) is not yet conclusively established (66). We therefore repeated the analysis using a multifurcation at this node and found similar results regarding the asymmetry, but observed variation in the estimates of CDK consensus gain and loss events on each branch (data not shown).

To calculate the expected number of gains in the Saccharomyces clade under the hypothesis of symmetrically distributed changes, we assume the number of branch positions is large relative to the number of matches and that gains of matches are rare (no multiple hits). The expected number of gains in a subclade $c$ is then Poisson with mean $= n_g c$, where $t$ is the sum of the branch lengths (tree length), $t_c$ is the sum of the branches in the clade $c$, and $n_g$ is the number of gains inferred along the whole tree. To calculate the ancestral values of $S_{LR}$ we reconstructed the ancestral positions of each column of aligned matches by recursively assigning to each ancestor the average position of the matches in its children.

Construction of GFP Reporters and Localization Assays. We obtained genomic DNA for C. albicans, C. glabrata, and K. lactis from D. Galgoczy (University of California, San Francisco) and A. Johnson (University of California, San Francisco) and for A. gossypii from A. Gladfelder (Dartmouth College, Hanover, NH). We amplified the region homologous to the S-Mcm3-CTR or the entire C terminus by PCR (Phusion; Finnzymes, Espoo, Finland), using primers (IDT Technologies, Coralville, IA) that introduced ClaI or EcoRI restriction sites into the 5′ ends of the PCR product. Primer sequences are available on request. These PCR products were inserted between the ClaI and EcoRI sites in the plasmid pML104, a gal inducible TRP1 integrating plasmid containing the S. cerevisiae Mcm2 nuclear localization signal and three tandem copies of GFP (25). All constructs were confirmed by sequencing (MClab, South San Francisco, CA). Plasmids were transformed into YJL310 (67), grown, arrested and photographed as described in figures 4, 5, 6B, 8, and 9 of ref. 25. The cell-cycle arrests were confirmed by scoring the fraction budded for >60 cells for each strain under each condition. The GFP localization panels shown were “representative,” and observations were confirmed by scoring the fraction showing nuclear staining for >60 cells for each construct under each condition.

Moses et al.
We thank Dr. Dave Morgan, Dr. Seth Grant, Dr. Avril Coghlan, and Dr. Jean-Karim Heriche, and Dave Galgoczy for helpful discussions; Dave Galgoczy, Dr. Alexander Johnson, and Dr. Amy Gladfelter for genomic DNA; and Dr. Avril Coghlan for comments on the manuscript. This work was supported by National Institutes of Health Grants SF31CA110268-03 (to M.E.L.) and R01 GM59704 (to J.J.L.). A.M.M. is a Sanger Postdoctoral Fellow. R.D. and the Wellcome Trust Sanger Institute are funded by the Wellcome Trust.