ELSEVIER

Gene 214 (1998) 59–66

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

# The four *cdc25* genes from the nematode *Caenorhabditis elegans*

Neville R. Ashcroft [a], Mary E. Kosinski [a], Dineli Wickramasinghe [b,1], Peter J. Donovan [b], Andy Golden [a,*]

[a] *ABL-Basic Research Program, Gene Regulation and Chromosome Biology Laboratory, PO Box B, Building 539, National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, MD 21702, USA*

[b] *ABL-Basic Research Program, Mammalian Genetics Laboratory, PO Box B, Building 539, National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, MD 21702, USA*

Received 2 February 1998; received in revised form 20 April 1998; accepted 21 April 1998; Received by E. Boncinelli

**Abstract**

During eukaryotic evolution, multicellular organisms have evolved multiple members of gene families that may display unique, partially overlapping, or redundant functions during development. More than 75% of the *C. elegans* genome has been sequenced, which represents approximately 95% of the coding sequences. This provides a unique opportunity to identify most, if not all, of the members of a given gene family. We have searched the *C. elegans* genome database for members of a key family of cell cycle regulators, the CDC25 phosphatases, and have identified four genes. The four *C. elegans* genes represent a larger family within a single organism than has been reported so far in *Drosophila*, mice and humans. An amino acid comparison revealed a high degree of similarity and identity within the phosphatase domain. This analysis also identified an expanded consensus sequence that can be used to discover new members of the CDC25 phosphatase family. However, the four *C. elegans* sequences display a few novel amino acid substitutions in the residues surrounding the invariant catalytic motif $CX_5R$. These data demonstrate the value of genome database searching for identifying new members of known gene families, understanding genetic diversity, and for studying gene structure. © 1998 Elsevier Science B.V. All rights reserved.

*Keywords:* Cell cycle; Dual-specificity phosphatase; Gene family; Genomic database; Homologs

## 1. Introduction

The dual-specificity phosphatase CDC25 was originally identified as an essential regulator of the cell cycle in *S. pombe* (Russell and Nurse, 1986; Millar et al., 1991). This phosphatase is required to drive cells through G2 into M-phase, and mutations in *cdc25* cause cell cycle arrest. Homologs of CDC25 have been identified in many other eukaryotic organisms and have also been shown to be essential regulators of the eukaryotic cell cycle (Draetta and Eckstein, 1997). Moreover, overexpression of CDC25 has been detected in a number of human cancers, suggesting a role in oncogenesis (Nagata et al., 1991; Galaktionov et al., 1995; Gasparotto et al., 1997). Although the yeasts, *S. pombe* and *S. cerevisiae*, each have a single *cdc25* gene, multicellular eukaryotes have evolved multiple *cdc25* genes. For example, *Drosophila* has two members, *string* and *twine* (Draetta and Eckstein, 1997). The STRING gene product is required predominantly for mitosis, whereas the TWINE gene product is specific for meiosis (Reed, 1995). Mice and humans have three *cdc25* genes, A, B, and C, which have implied roles at G1/S and G2/M of the cell cycle, and their activities may also be tissue-specific (Draetta and Eckstein, 1997). The relationship of the mammalian *cdc25* genes to STRING and TWINE remains unclear.

We are investigating the developmental roles of this gene family in the model organism, the hermaphroditic soil nematode, *Caenorhabditis elegans*. The pattern of early development in *C. elegans* differs from that of *Drosophila*. The cell cycles are asymmetric and asynchro-

* Corresponding author. Tel: +1 301 846 1732; Fax: +1 301 846 6911; e-mail: golden@ncifcrf.gov

[1] Present address: Molecular Biology, BASF Bioresearch Corporation, 100 Research Drive, Worcester, MA 01605, USA.

Abbreviations: aa, amino acid(s); bp, base pair(s); C, cysteine; *cdc*, cell division cycle (gene); CDC, protein product of *cdc* gene; cDNA, DNA complementary to RNA; D, aspartic acid; EST, expressed sequence tag; H, histidine; kb, kilobase(s) or 1000 bp; mRNA, messenger RNA; nt, nucleotide(s); PCR, polymerase chain reaction; R, arginine; SL, spliced leader; UTR, untranslated region; RT, reverse transcription; BLAST, basic local alignment search tool.

0378-1119/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(98)00228-5

nous from the very first embryonic cell division (Deppe et al., 1978; Sulston et al., 1983). For example, the first mitotic division gives rise to a large anterior AB somatic founder cell and a smaller posterior $P_1$ blastomere. Subsequent asymmetric cleavages generate the other somatic founder cells MS, E, C, and D, and the germline founder cell, $P_4$. During embryogenesis, the somatic founder cells all divide with unique cell cycle periodicities, thus generating asynchronous cell divisions (Deppe et al., 1978; Sulston et al., 1983). However, the cells within a given founder cell lineage divide essentially synchronously. In contrast to the proliferation of the somatic founder cells, the $P_4$ cell divides only once during embryogenesis, giving rise to the cells Z2 and Z3. These cells do not divide throughout the remainder of embryonic development, but proliferate extensively during post-embryonic development to produce more than 1000 germ cell nuclei. We are interested in how these different cell cycles are established, regulated, and maintained, and are examining the roles that the *C. elegans cdc25* genes may play in these processes.

## 2. Materials and methods

### 2.1. cDNA cloning and sequencing

A partial *cdc-25.1* cDNA was first isolated and cloned by RT-PCR. RNA was isolated from washed N2 hermaphrodites with Trizol (Gibco-BRL) according to the manufacturer's instructions. Aliquots of this RNA were reverse-transcribed and used for PCR (Titan One-Tube RT PCR System, Boehringer Mannheim) to amplify a partial *cdc-25.1* cDNA. At the time that we initiated this study, Genefinder had predicted a gene structure consisting of six exons, with a very large first intron (~1.75 kb). The RT-PCR was carried out with forward primers corresponding to sequences within the predicted exon 1 (F3: ATGGCCGCTGAAATTTGTCGG) or predicted exon 2 (F4: CTACAACTATGATACACCAACAGG). The downstream primer was complementary to the region spanning the predicted stop codon (TTATTCGGCGTCGTCAGAAATCG). Reactions using forward primer F3, and two other primers corresponding to sequences in the predicted exon 1, did not yield any product. Only reactions using forward primer F4 yielded any PCR product. The product was 1.5 kb, a size predicted by Genefinder if this message originated from the predicted exon 2. This 1.5-kb DNA fragment was then sub-cloned into pCRII (InVitrogen). This cDNA insert was used to probe a mixed-stage *C. elegans* λZAP cDNA library (Stratagene) to identify a full-length *cdc-25.1* cDNA clone. Six hundred thousand plaques were screened. Positive plaques were processed by standard methods, and the pBluescript plasmid was excised using the Rapid Excision Kit (Stratagene). Nine clones were analyzed, and were representative of the same gene. Automated sequencing of the longest clone was performed using standard methods. This clone had 10 bp of the SL1 RNA sequence at its 5′ end. The polyadenylation site (AAUGAA) was located 403 nt downstream of the stop codon. Seventeen nucleotides downstream of this AAUGAA was the polyA tail. Four EST sequences (Accession Nos C39323, C42657, C54028, C64027) from a separate cDNA database (Y. Kohara, pers. commun.; http://www.ddbj.nig.ac.jp/htmls/c-elegans/html/CE_INDEX.html) were compared with this sequence, and all were contained within this full-length clone. The 5′ sequences of all of these clones revealed that sequences from the predicted exon 1 were not contained within these cDNAs. Rather, two exons were utilized that were not originally predicted by Genefinder; they were located in what had been the predicted intron 1. Collectively, these results demonstrate that the *cdc-25.1* gene has seven exons. Thus, the partial cDNA cloned by RT-PCR actually contained sequences from exons 3 to 7. By working with the Genome Consortium prior to GenBank submission, our sequence analysis helped modify the Genefinder predictions for this gene (Accession No. AF039038). Our cDNA analyses have been incorporated into the genomic information submitted to GenBank by the Genome Consortium.

For *cdc-25.2*, 10 cDNAs were identified from a cDNA database in which the sequences of the 5′ and 3′ ends (ESTs) of the cDNA inserts were available (Y. Kohara, pers. commun.). The pBluescript plasmids containing the cDNA inserts yk109g1 (EST Accession Nos D72817 and D75756) and yk118b11 (EST Accession Nos C07232 and C08531) were excised using the Rapid Excision Kit (Stratagene). These two longest clones were sequenced, and clone yk109g1 was found to be full-length. The entire *cdc-25.2* gene spanned two cosmids, F52F10 and F16B4 (Accession No. AF039048), and a cosmid gap, which has subsequently been shown to be covered by cosmid D2011. The 5′ end of these cDNA clones did not contain any SL1 or SL2 RNA sequences. The polyadenylation site (AAUAAA) is 578 nt downstream of the predicted stop codon. Thirteen nt downstream of this AAUAAA was the polyA tail.

A cDNA of the *cdc-25.3* gene was cloned by RT-PCR using a forward primer specific for the *trans*-spliced leader RNA SL1 (GGTTTAATTACCCAAGTTTGAG), and a reverse primer from the 3′ UTR of this gene (CATGCATGGCGGCCGCCATCCTATTATCCTGGATACTAG, which included a *Not*I site at its 5′ end). This cDNA was cloned into pCRII (InVitrogen), restriction-mapped, and sequenced. A perfect AAUAAA polyadenylation signal lies 115 nt downstream of the stop codon for this gene. The reverse primer used for the RT-PCR cloning of this cDNA

annealed to sequences just 16 nt upstream of this AAUAAA.

### 2.2. Amino acid sequence alignment

The alignment shown in Fig. 3 was generated using the PILEUP, LINEUP, and PRETTYPLOT Programs from the Wisconsin Package (Version 9.1) of GCG (Madison, WI), and includes all known *cdc25* genes with the following exception: four *Xenopus cdc25* genes (Accession Nos M94263, M94264, M96857, and M96858), which are highly related to *cdc25-1* [and are likely to be from the same gene; (Izumi et al., 1992; Kumagai and Dunphy, 1992)] were not included. The complete amino acid sequence for the CDC25 proteins from *Dictyostelium*, starfish, and *Arabidopsis* are not known; only partial sequences are available.

### 2.3. Phylogenetic analysis

Phylogenetic analysis was performed using the three domains indicated in Fig. 3. Sequences from *Arabidopsis* and *Dictyostelium* were not used for this analysis since their complete amino acid sequences are not known. The phenogram was derived using the programs Clustal W V1.7 (Thompson et al., 1994) and PHYLIP (Phylogeny Inference Package) V3.5c, distributed by J. Felsenstein (Department of Genetics, University of Washington, Seattle, WA). Trees were visualized using TreeView for Macintosh V1.4 [http://taxonomy.zoology.gla.ac.uk/rod/treeview.html (Page, 1996)], and NJ-PLOT [ftp://biom3.univ-lyon1.fr/pub/mol_phylogeny/njplot/njplot (Perriere and Guoy, 1996)].

## 3. Results

### 3.1. Identification of C. elegans genes belonging to the cdc25 gene family

The entire genome of *C. elegans* is being sequenced by the Genome Consortium (Wilson et al., 1994) and was 78% complete as of 20 March 1998 (http://genome.wustl.edu/gsc/gschmpg.html). This represents approximately 95% of the coding sequences of *C. elegans* (J. Spieth, pers. commun.). We have searched the *C. elegans* database for homologs using TBLASTN (Altschul et al., 1990; Gish and States, 1993), and have identified four genes predicted to code for proteins with homology to the CDC25 family of phosphatases. These searches were performed on a monthly basis, and no new *cdc25*-like sequences have been identified since February 1997. The four genes found were originally sequenced from cosmids and given the names K06A5.7a, F16B4.8, ZK637.11, and R05H5.2; all are now in the GenBank Database. The exon/intron structures of these genes were predicted by Genefinder (P. Green and L. Hillier, in preparation), a program used by the Genome Consortium. In agreement with *C. elegans* nomenclature, we have adopted the following names for these four *cdc25* genes: *cdc-25.1* (K06A5.7a), *cdc-25.2* (F16B4.8), *cdc-25.3* (ZK637.11), and *cdc-25.4* (R05H5.2). We have confirmed and modified the exon/intron structure predictions for three of the four genes by cDNA sequence analyses. The structures for *cdc-25.1*, *cdc-25.2*, *cdc-25.3*, and *cdc-25.4* are shown in Fig. 1. These genes do not appear to have arisen by any obvious gene duplication as their exon/intron structures vary greatly, and they map to four separate linkage groups. The *cdc-25.3* gene is coded for by three exons, whereas *cdc-25.2*, *cdc-25.1*, and *cdc-25.4* are encoded by 6, 7, and 8 exons, respectively. Remarkably, a common splice donor site at the beginning of the phosphatase domain is shared by *cdc-25.2*, *cdc-25.3* and *cdc-25.4*, whereas *cdc-25.2* and *cdc-25.4* also share a common splice acceptor site near the end of the phosphatase domain (Fig. 1). The overall amino acid similarity between the four predicted *C. elegans* proteins is weak. However, within an 118-amino-acid region of the phosphatase domain, there is 31% amino acid identity and 47% similarity among the four *C. elegans* sequences. When comparing the same domain between CDC-25.2 and CDC-25.3, this identity is 53%.

### 3.2. cDNA cloning and analyses

To confirm the predictions made by Genefinder, we cloned and sequenced cDNAs corresponding to three of the four genes. A full-length *cdc-25.1* cDNA was isolated from a mixed-stage *C. elegans* cDNA library. Our sequence analysis of this cDNA clone was used to modify the predictions made by Genefinder (see Section 2.1). This message was *trans*-spliced to the short spliced leader RNA, SL1, which is commonly found in *C. elegans* mRNAs (Krause and Hirsh, 1987) (Fig. 2). For *cdc-25.2*, 10 cDNAs were identified from a cDNA database in which the ESTs were reported (see Section 2.1). The two longest clones were sequenced: yk109g1 and yk118b11. Clone yk109g1 was found to be full-length. The 5′ end of these cDNA clones did not contain any SL1 or SL2 RNA sequences. Despite the presence of a consensus splice acceptor site (TTTCAG) 47 bp upstream of the predicted ATG codon (Fig. 2), no SL1 or SL2 *trans*-spliced mRNAs have been detected by RT-PCR analysis. The *cdc-25.3* cDNA was cloned by RT-PCR using a forward primer specific for the *trans*-spliced leader RNA SL1, and a reverse primer from the 3′ UTR of this gene (see Section 2.1). The mRNA for this gene was spliced from its three exons, as predicted by Genefinder, and was also SL1 *trans*-spliced (Figs. 1 and 2). Expression of the *cdc-25.4* gene could not be detected. No signal has been detected by Northern analysis or by a variety of RT-PCR and nested PCR protocols using mRNAs from mixed-stage animals. In addition, RT-PCR failed to detect a cDNA fragment

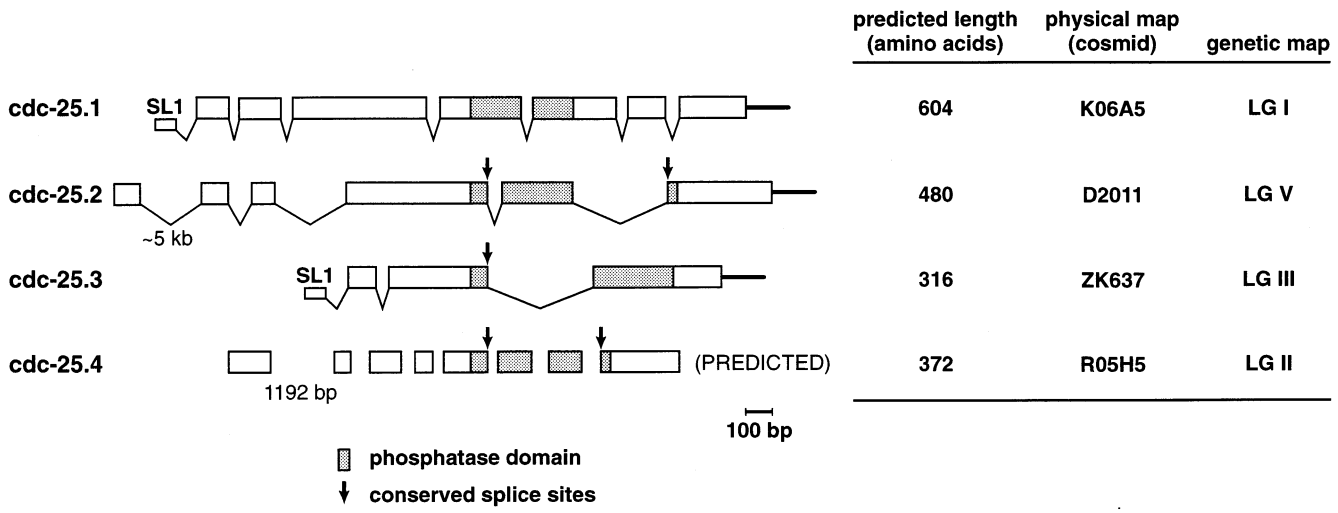


Fig. 1. Gene structure, amino acid length, and physical and genetic map positions of the four *C. elegans cdc-25* genes. This figure depicts the gene structure of the four *C. elegans cdc-25* genes based on the Genefinder Program predictions and cDNA sequencing. The structure shown for *cdc-25.4* is based solely on the Genefinder prediction. Indicated to the right are the number of predicted amino acid residues encoded by each gene, the cosmid from which each gene was originally sequenced by the Genome Consortium, and the linkage group (LG) on which each gene resides. The structures are to scale except for the size of intron 1 of *cdc-25.2* and *cdc-25.4*, and the 22-bp SL1 *trans*-spliced leader RNA. The arrows mark the positions of conserved splice sites (see Section 3.1).

**cdc-25.1**

genomic sequence TTTTAGTTGAGGGCTGTGCATCTCTGATG
cDNA sequence SL1-TTGAGGGCTGTGCATCTCTGATG

**cdc-25.2**

genomic sequence TTTCAGAAAACTCCAAACACGTGGGCACAGTCATTTAATCTAGAGTATTTATTATG
cDNA sequence GAAAACTCCAAACACGTGGGCACAGTCATTTAATCTAGAGTATTTATTATG

**cdc-25.3**

genomic sequence ATTCAGGTATACCTAGTTCAGTCTAGCTCCAAGTCTCCTGATG
cDNA sequence SL1-GTATACCTAGTTCAGTCTAGCTCCAAGTCTCCTGATG

Fig. 2. Nucleotide sequence of the 5′ UTRs of the *cdc-25* cDNAs and the corresponding genomic sequences. Shown are genomic sequences upstream of the predicted initiator methionine residue in *cdc-25.1*, *cdc-25.2*, and *cdc-25.3*. Underlined in the genomic sequences are the splice acceptor site sequences, whose consensus sequence is TTTCAG (Blumenthal and Steward, 1997). Shown below each genomic sequence are the cloned cDNA sequences. The methionine codon is underlined in each of the cDNA sequences. Both *cdc-25.1* and *cdc-25.3* are SL1 *trans*-spliced to the expected splice sites. The *cdc-25.2* cDNA does not appear to be SL1 or SL2 *trans*-spliced by cDNA sequencing and RT-PCR analyses.

using mutant populations that have a high incidence of males. The *cdc-25.4* mRNA may be transiently expressed in a few cells at a specific developmental stage. Analysis of RNAs isolated from synchronous cultures, at specific developmental stages, under various growth conditions, may be required to detect *cdc-25.4* expression. Although expression of this gene has not yet been detected, there is no evidence to suggest that it is a pseudogene. Based on the predicted genomic structure and sequence, it lacks the features characteristic of *C. elegans* pseudogenes (Ward et al., 1988; Heschl and Baillie, 1989), such as premature termination codons and frameshift mutations. All of the amino acid sequence analyses in this report, with the exception of CDC-25.4, are based on the predicted amino acid sequences from cloned cDNAs. The amino acid sequence shown for CDC-25.4 is based solely on the Genefinder predictions used by the Genome Consortium.

### 3.3. Amino acid comparisons with CDC25 proteins from other organisms

The CDC25 phosphatases are the most conserved among the tyrosine phosphatase families (Fauman and

Saper, 1996). Therefore, stringent BLAST searches can be performed to identify CDC25 homologs. The CDC25 family is distinguished by three invariant residues (D, C, and R) and the sequences that flank the active site motif ($CX_5R$) in the phosphatase domain (Fauman and Saper, 1996). Comparison of the phosphatase domains of the four *C. elegans* CDC25 proteins with CDC25 sequences from 16 other species reveals a high degree of amino acid identity and conservation (Fig. 3). The phosphatase domain can be divided into three highly conserved regions that are separated by two short linker regions of variable length. The first region, comprised of 21 amino acids (boxed in Fig. 3), has only three variant residues among all family members. The consensus of this region can be defined as IIDCRYPYEYxGGHIxGAxNL; the aspartic acid residue (D; arrow in Fig. 3) serves as a general acid (Denu et al., 1995), and is found in all CDC25 proteins. In this highly conserved domain of 21 residues, CDC-25.3 and CDC-25.4 differ from the consensus by three, whereas CDC-25.1 and CDC-25.2 differ by four. The CDC25 sequences of *S. pombe* and *S. cerevisiae* differ from the consensus by three and five residues, respectively. This 21-amino-acid region is separated from a second conserved domain by a variable region that shares a considerable sequence similarity among the vertebrate sequences.

The second conserved domain includes the active site motif, $HCX_5R$, which has been proposed to form a phosphate binding loop (Barford et al., 1994; Stuckey et al., 1994; Su et al., 1994; Jia et al., 1995). The cysteine and arginine residues of this motif (arrows in Fig. 3) are essential for biochemical activity and mutation of these residues inactivate CDC25 (Guan and Dixon, 1991; Kumagai and Dunphy, 1992; Zhang et al., 1994; Xu and Burke, 1996). A well-conserved region of 51–53 amino acids surrounds the $HCX_5R$ motif (Fig. 3). A short variable region follows this domain in which the *C. elegans* sequences are two to eight amino acids longer than other CDC25 sequences. The third conserved domain, with the consensus sequence CxPxxYxxM [part of Fauman and Saper's CH2-B consensus (Fauman and Saper, 1996)], is absolutely conserved in all CDC25 proteins. The functional significance of these four invariant residues is unknown. Future mutagenesis experiments may allow the biochemical and biological relevance of some of these conserved amino acids to be examined. In this analysis, we have expanded the consensus sequence for the CDC25 family and shown that a large number of residues have been conserved throughout evolution.

Although the four *C. elegans* CDC25s conform to the consensus sequence for CDC25 phosphatases, some notable differences exist. A few amino acid substitutions can be found in the region surrounding the active site motif, $HCX_5R$. The most obvious changes in the catalytic domains of the *C. elegans* sequences are three residues that flank the invariant cysteine residue (designated zero in Fig. 3) of the active site, and are as follows: position −4, proline in place of other nonpolar residues; position −1, tyrosine in place of histidine; position +4, glutamine in place of serine. These differences are also present in another nematode CDC25, based on the partial sequence of a *cdc-25* cDNA clone from *Pristionchus pacificus* (GenBank Accession No. AA191781). The effect of these substitutions on structure and activity are unknown. However, upon mutation of histidine at position −1 in human CDC25A and -B (Xu and Burke, 1996) or in *Drosophila* STRING (Gautier et al., 1991) or in *Yershinia* phosphatase (Zhang and Dixon, 1993), no loss in catalytic activity was observed. Of the 27 CDC25 sequences examined, only six vary at this position: murine CDC25C (Leu) and hamster CDC25 (Leu), and all four *C. elegans* CDC25s (Tyr).

Phylogenetic analysis of the phosphatase domain sequences suggests that the CDC25 proteins from *C. elegans* are more closely related to each other than they are to the CDC25 proteins of other organisms (Fig. 4), and that *cdc-25* gene duplication most likely occurred during nematode evolution. It appears that the *C. elegans* sequences diverged from the *Drosophila* and vertebrate sequences early in the evolution of the *cdc25* gene family. Because of this early divergence, it is difficult to establish the relatedness of the individual *C. elegans* sequences with the mammalian isoforms A, B, and C, or with TWINE and STRING of *Drosophila*.

## 4. Discussion

### *4.1.* C. elegans *has four* cdc25 *genes*

Our data demonstrate that the *C. elegans* genome contains four *cdc25* genes, whereas in mice and humans, only three homologs have been identified to date. The murine and human *cdc25* genes, in addition to the second *Drosophila* homolog, *twine*, were identified by PCR using degenerate primers. These approaches rely on designing degenerate oligonucleotides from contiguous stretches of conserved amino acids. The PCR cloning approach, while powerful, can fail to identify genes that have a weak homology. Systematic analysis of genomic databases, however, relies on the ability to recognize conserved motifs. Our identification of four *cdc25* genes in *C. elegans* suggests that the *cdc25* gene family in multicellular eukaryotes may be more complex than previously thought. Such analyses may aid in our understanding of genetic diversity and genetic redundancy.

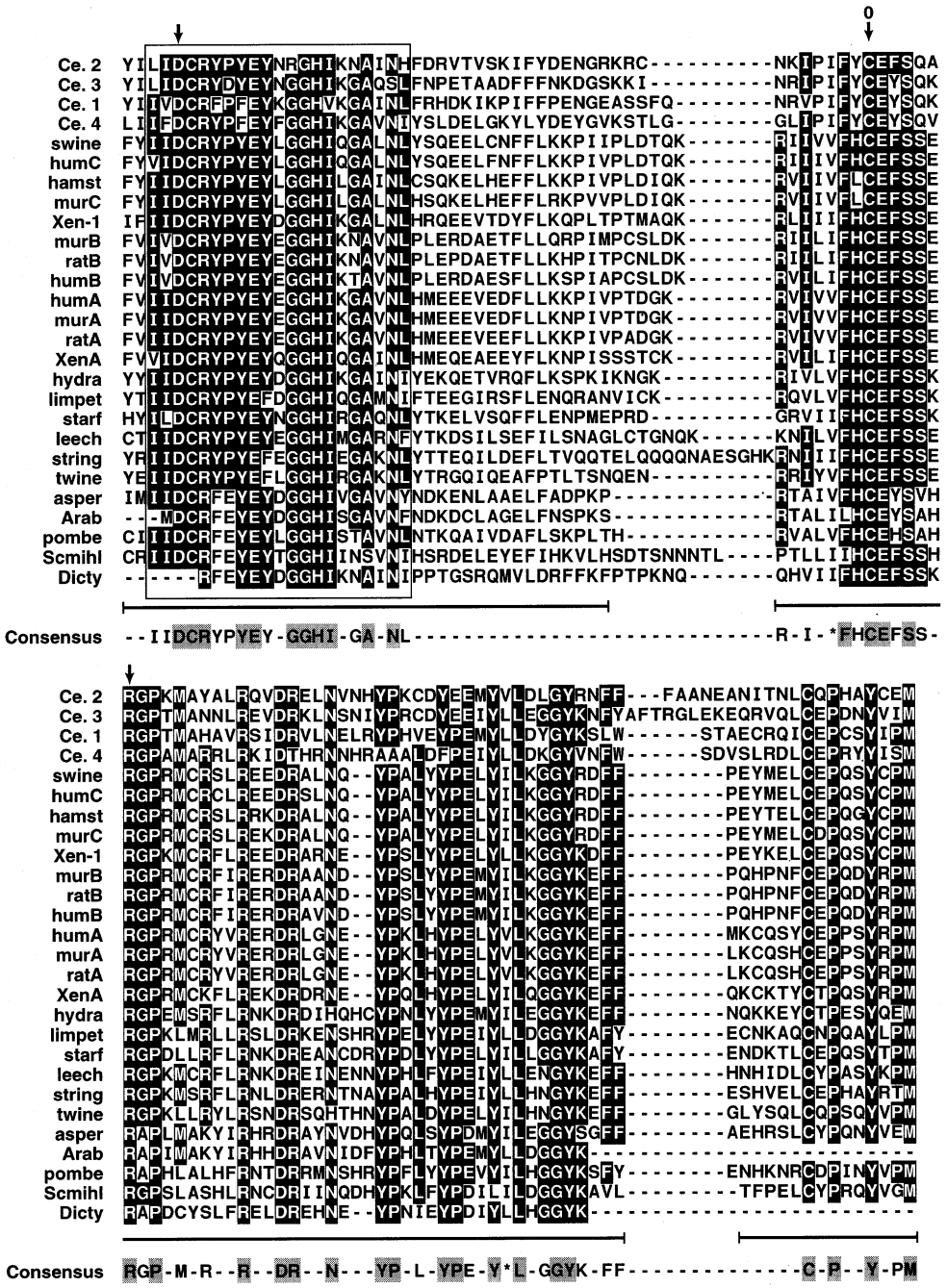


Fig. 3. Amino acid sequence alignment of the phosphatase domains of 27 CDC25 family members. The arrows above the alignment mark the D, C, and R residues known to be essential for CDC25 phosphatase activity. The '0' also marks the invariant cysteine residue at the core of the $HCX_5R$ catalytic motif. The box indicates the 21-amino-acid domain that contains the invariant D residue. Amino acids that are found in at least 18 of the 27 CDC25 sequences are highlighted in black. A consensus sequence is shown below the alignment at positions where at least 18 of the 27 sequences are identical. Shaded residues in this consensus are identical in at least 25 of the 27 CDC25 sequences shown. An asterisk marks the positions in which all 27 sequences have a conservative amino acid substitution (I, L, V). The bars underneath the alignment mark the three domains that were used together to generate the phylogram shown in Fig. 4. Gaps (—) were inserted to optimize the alignment. Genbank Accession Nos (or references) are included in parentheses following each protein name. The following abbreviations are used in this figure: Ce.2: *Caenorhabditis elegans* (nematode) CDC-25.2 (AF039048), Ce.3: *Caenorhabditis elegans* CDC-25.3 (P30634, Z11115), Ce.1: *Caenorhabditis elegans* CDC-25.1 (AF039038), Ce.4: *Caenorhabditis elegans* CDC-25.4 (Z48795), swine: *Sus scrofa* (pig) CDC25 (X78317), humC: *Homo sapiens* (human) CDC25C (M34065), hamst: *Mesocricetus auratus* (hamster) CDC25 (D10878), murC: *Mus musculus* (mouse) CDC25M1 (L16926; P48967), Xen-1: *Xenopus laevis* (frog) CDC25-1 (M94262), murB: *Mus musculus* CDC25M2 (S93521), ratB: *Rattus norvegicus* (rat) CDC25B (D16237), humB: *Homo sapiens* CDC25B (M81934), humA: *Homo sapiens* CDC25A (M81933), murA: *Mus musculus* CDC25M3 (U27323), ratA: *Rattus norvegicus* CDC25A (D16236), XenA: *Xenopus laevis* CDC25A (D82960), hydra: *Hydra vulgaris* (hydra) CDC25 (Laubrock and Schaller, 1996), limpet: *Patella vulgata* (limpet) CDC25 (van der Kooij et al., 1996), starf: *Asterina pectinifera* (starfish) CDC25 (K. Tachibana and T. Kishimoto, pers. commun.; Okumura et al., 1996), leech: *Helobdella robusta* (leech) CDC25 (U25089), string: *Drosophila melanogaster* (fruitfly) STRING (M24909), twine: *Drosophila melanogaster* TWINE (X69018), asper: *Aspergillus nidulans* (aspergillus) nimT (X64601), Arab: *Arabidopsis thaliana* (plant) CDC25 (Millar et al., 1991), pombe: *Schizosaccharomyces pombe* (fission yeast) CDC25 (M13158), Scmih1: *Saccharomyces cerevisiae* (budding yeast) MIH1 (J04846), Dicty: *Dictyostelium discoideum* (slime mold) CDC25-1 (Millar et al., 1991).

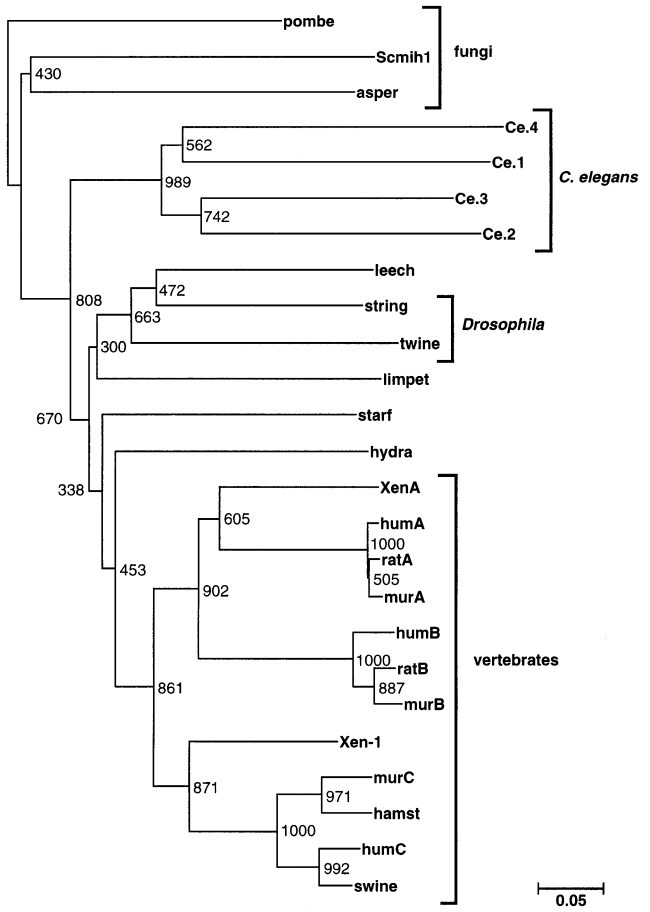


Fig. 4. Phenogram of 25 CDC25 phosphatase domains from 15 different species. This tree is based on the analysis of the three domains indicated in Fig. 3. See the legend of Fig. 3 for definitions of abbreviations. Numbers on the branches indicate bootstrap values for 1000 trials. The *S. pombe* CDC25 sequences were used as an outgroup.

### 4.2. *Why are there four* cdc25 *genes in* C. elegans*?*

This genomic analysis raises the question of why multiple *cdc25* genes are expressed in *C. elegans*. The unique mode of *C. elegans* embryogenesis may require the activity of multiple *cdc25* genes. Unlike *Drosophila*, the earliest mitoses are not synchronous. The five somatic founder cells have unique doubling times leading to asynchronous proliferation of embryonic cells. Perhaps these different cycling rates require different CDC25 activities. Furthermore, the germline precursor cells, Z2 and Z3, do not divide during embryogenesis, but proliferate extensively during larval and adult development to populate the gonad with germ cell nuclei. Individual *cdc25* genes may regulate germline proliferation and entry into meiosis. It may also be possible that one or more of these *cdc25* genes are required for male-specific development, dauer-specific development, or development in the wild (versus the laboratory). Perturbation of *cdc25* gene expression in *C. elegans* may elucidate the unique, and/or overlapping roles that these four genes play during development and gametogenesis.

## Acknowledgement

We thank the Genome Sequencing Centers at Washington University in St. Louis and the Sanger Centre in Cambridge for communication of DNA sequence data prior to publication. We also gratefully acknowledge all of the scientists involved in the *C. elegans* Genome Consortium without whom this project would have been arduous. Special thanks to J. Spieth at the St. Louis Genome Sequencing Center for all of his help and discussions concerning the *C. elegans cdc25* gene family and this manuscript. Thanks also to Y. Kohara for his cDNA clones and access to his EST database, to R. Durbin and J. Thierry-Mieg for making much of our analysis possible using ACeDB, to K. Tachibana and T. Kishimoto for their partial starfish CDC25 sequence, to T. Burglin, J. Dalgaard, G. Smythers, and R. Stephens for computer analysis advice, to M. Wilson and R. Hoch for excellent technical assistance, and to A. Cacace, D. Chase, J. Schumacher, A. Valsamakis, and R. Yip for critical evaluation of this manuscript. This research was sponsored by the National Cancer Institute, DHHS, under contract with ABL.

## References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Barford, D., Flint, A.J., Tonks, N.K., 1994. Crystal structure of human protein tyrosine phosphatase 1B. Science 263, 1397–1404.

Blumenthal, T., Steward, K., 1997. RNA Processing and Gene Structure. In: Riddle, D.L., Blumenthal, T., Meyer, B.J., Priess, J.R. (Eds.), *C. ELEGANS* II. Cold Spring Harbor Press, Cold Spring Harbor, NY, pp. 117–145.

Denu, J.M., Zhou, G., Guo, Y., Dixon, J.E., 1995. The catalytic role of aspartic acid-92 in a human dual-specific protein-tyrosine-phosphatase. Biochemistry 34, 3396–3403.

Deppe, U., Schierenberg, E., Cole, T., Krieg, C., Schmitt, D., Yoder, B., von Ehrenstein, G., 1978. Cell lineages of the embryo of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 75, 376–380.

Draetta, G., Eckstein, J., 1997. Cdc25 protein phosphatases in cell proliferation. Biochim. Biophys. Acta 1332, M53–M63.

Fauman, E.B., Saper, M.A., 1996. Structure and function of the protein tyrosine phosphatases. Trends Biochem. Sci. 21, 413–417.

Galaktionov, K., Lee, A.K., Eckstein, J., Draetta, G., Meckler, J., Loda, M., Beach, D., 1995. CDC25 phosphatases as potential human oncogenes. Science 269, 1575–1577.

Gasparotto, D., Maestro, R., Piccinin, S., Vukosavljevic, T., Barzan, L., Sulfaro, S., Boiocchi, M., 1997. Overexpression of CDC25A and CDC25B in head and neck cancers. Cancer Res. 57, 2366–2368.

Gautier, J., Solomon, M., Booher, R.N., Bazan, J.F., Kirschner, M., 1991. cdc25 is a specific tyrosine phosphatase that directly activates $p34^{cdc2}$. Cell 67, 197–211.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nature Genet. 3, 266–272.

Guan, K.L., Dixon, J.E., 1991. Evidence for protein-tyrosine-phosphatase catalysis proceeding via a cysteine-phosphate intermediate. J. Biol. Chem. 266, 17026–17030.

Heschl, M.F.P., Baillie, D.L., 1989. Identification of a heat-shock pseudogene from *Caenorhabditis elegans*. Genome 32, 190–195.

Izumi, T., Walker, D.H., Maller, J.L., 1992. Periodic changes in phosphorylation of the *Xenopus* cdc25 phosphatase regulates its activity. Mol. Biol. Cell 3, 927–939.

Jia, Z., Barford, D., Flint, A.J., Tonks, N.K., 1995. Structural basis for phosphotyrosine peptide recognition by protein tyrosine phosphatase 1B. Science 268, 1754–1758.

Krause, M., Hirsh, D., 1987. A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. Cell 49, 753–761.

Kumagai, A., Dunphy, W.G., 1992. Regulation of the cdc25 protein during the cell cycle in *Xenopus* extracts. Cell 70, 139–151.

Laubrock, A., Schaller, H.C., 1996. Cloning and expression of a cdc25 homologue from hydra. Eur. J. Cell Biol. 71, 319–323.

Millar, J., McGowan, C., Jones, R., Sadhu, K., Bueno, A., Richardson, H., Russell, P., 1991. *cdc25* M-phase inducer. Cold Spring Harbor Symp. Quant. Biol. 56, 577–584.

Nagata, A., Igarashi, M., Jinno, S., Suto, K., Okayama, H., 1991. An additional homolog of the fission yeast cdc25+ gene occurs in humans and is highly expressed in some cancer cells. New Biol. 3, 959–968.

Okumura, E., Sekiai, T., Hisanaga, S., Tachibana, K., Kishimoto, T., 1996. Initial triggering of M-phase in starfish oocytes: a possible novel component of maturation-promoting factor besides cdc2 kinase. J. Cell Biol. 132, 125–135.

Page, R.D.M., 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comp. Appl. Biosci. 12, 357–358.

Perriere, G., Guoy, M., 1996. WWW-Query: An on-line retrieval system for biological sequence banks. Biochemie 78, 364–369.

Reed, B.H., 1995. *Drosophila* development pulls the strings of the cell cycle. BioEssays 17, 553–556.

Russell, P., Nurse, P., 1986. *cdc25*$^{+}$ functions as an inducer in the mitotic control of fission yeast. Cell 45, 145–153.

Stuckey, J.E., Schubert, H.L., Fauman, E., Zhang, Z.-Y., Dixon, J.E., Saper, M.A., 1994. Crystal structure of *Yersinia* protein tyrosine phosphatase at 2.5A and the complex with tungstate. Nature 370, 571–575.

Su, X.-D., Taddel, N., Stefani, M., Ramponi, G., Nordlund, P., 1994. The crystal structure of a low-molecular-weight phosphotyrosine protein phosphatase. Nature 370, 575–578.

Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N., 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev. Biol. 100, 64–119.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

van der Kooij, A., Nederbragt, A.J., Goedemans, H.J., van Loon, A.E., 1996. The *stringlike* genes of the limpet *Patella vulgata*. Gene 172, 261–265.

Ward, S., Burke, D.J., Sulston, J.E., Coulson, A.R., Albertson, D.G., Ammons, D., Klass, M., Hogan, E., 1988. Genomic organization of major sperm protein genes and pseudogenes in the nematode *Caenorhabditis elegans*. J. Mol. Biol. 199, 1–13.

Wilson, R. et al., 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature 368, 32–38.

Xu, X., Burke, S.P., 1996. Roles of active site residues and the $NH_2$-terminal domain in the catalysis and substrate binding of human Cdc25. J. Biol. Chem. 271, 5118–5124.

Zhang, Z.Y., Dixon, J.E., 1993. Active site labeling of the *Yersinia* protein tyrosine phosphatase, the determination of the pKa of the active site cysteine and the function of the conserved histidine 402. Biochemistry 32, 9340–9345.

Zhang, Z.Y., Wang, Y., Wu, L., Fauman, E.B., Stuckey, J.A., Schubert, H.L., Saper, M.A., Dixon, J.E., 1994. The Cys(X)5Arg catalytic motif in phosphoester hydrolysis. Biochemistry 27, 15266–15270.