A genome-wide survey of RS domain proteins

L. Boucher, C. A. Ouzounis, A. J. Enright and B. J. Blencowe

RNA 2001 7: 1693-1701

RNA (2001), *7*:1693–1701. Cambridge University Press. Printed in the USA.

DOI: 10.1017.S1355838201011074

BIOINFORMATICS

A genome-wide survey of RS domain proteins

LORRIE BOUCHER,[1] CHRISTOS A. OUZOUNIS,[2] ANTON J. ENRIGHT,[2] and BENJAMIN J. BLENCOWE[1]

[1]Banting and Best Department of Medical Research, C.H. Best Institute, University of Toronto, Toronto, Ontario, Canada
[2]The European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge Outstation, Cambridge, United Kingdom

ABSTRACT

Domains rich in alternating arginine and serine residues (RS domains) are frequently found in metazoan proteins involved in pre-mRNA splicing. The RS domains of splicing factors associate with each other and are important for the formation of protein–protein interactions required for both constitutive and regulated splicing. The prevalence of the RS domain in splicing factors suggests that it might serve as a useful signature for the identification of new proteins that function in pre-mRNA processing, although it remains to be determined whether RS domains also participate in other cellular functions. Using database search and sequence clustering methods, we have identified and categorized RS domain proteins encoded within the entire genomes of *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. This genome-wide survey revealed a surprising complexity of RS domain proteins in metazoans with functions associated with chromatin structure, transcription by RNA polymerase II, cell cycle, and cell structure, as well as pre-mRNA processing. Also identified were RS domain proteins in *S. cerevisiae* with functions associated with cell structure, osmotic regulation, and cell cycle progression. The results thus demonstrate an effective strategy for the genomic mining of RS domain proteins. The identification of many new proteins using this strategy has provided a database of factors that are candidates for forming RS domain-mediated interactions associated with different steps in pre-mRNA processing, in addition to other cellular functions.

Keywords: bioinformatics; protein motif; RNA polymerase II; splicing factor; SR protein

INTRODUCTION

Pre-mRNA splicing requires the assembly of a spliceosome consisting of four small nuclear ribonucleoprotein particles (U1, U2, U4/U6, and U5 snRNPs) and numerous non-snRNP factors (Kramer, 1996; Reed & Palandjian, 1997; Burge et al., 1999). A large number of splicing factors contain domains rich in alternating arginine and serine residues (RS domains). Originally identified in three genetically defined *Drosophila* splicing regulators, *transformer*, *transformer-2*, and *suppressor of white apricot*, RS domains were subsequently found in numerous snRNP and non-snRNP splicing factors in higher eukaryotes but, with few exceptions, were not found in yeast proteins (Fu, 1995; Manley & Tacke, 1996; Blencowe et al., 1999; Graveley, 2000). RS domain proteins can be broadly subdivided into two groups: (1) the SR family proteins, which each contain one or two RNA recognition motifs (RRMs) and a C-terminal RS domain, and (2) SR-related proteins, which have a distinct domain structure from the SR family and may or may not contain an RRM.

The RS domains of splicing factors interact and are thought to mediate the formation of cross-intron and cross-exon protein networks important for splice site selection as well as interactions that function at later steps in spliceosome assembly (Blencowe, 2000; Graveley, 2000; Smith & Valcarcel, 2000). Phosphorylation of RS domains by specific kinases is important for these interactions and differential phosphorylation of RS domains provides a means by which splicing activity can be regulated (Du et al., 1998; Xiao & Manley, 1998; Yeakley et al., 1999). Recently, RS domains have been discovered in several factors associated with transcription components including the SR-like CTD-Associated Factors (SCAFs; Corden & Patturajan, 1997) and in a 3′-end cleavage factor, Cleavage Factor-I 68 kDa subunit (CF-Im; Ruegsegger et al., 1998), suggesting that

Reprint requests to: Benjamin J. Blencowe, C.H. Best Institute, University of Toronto, 112 College Street, Room 410, Toronto, Ontario, M5G 1L6, Canada; e-mail: b.blencowe@utoronto.ca.

communication between splicing and both transcription and 3′-end processing might in some cases be mediated by interactions involving RS domain proteins (Blencowe et al., 1999; Hirose & Manley, 2000). Moreover, it has been shown that different promoters can influence the alternative splicing of a reporter pre-mRNA and that these promoter-dependent effects appear to involve specific SR family proteins (Cramer et al., 1999).

RS domains have also been shown to function in the targeting of proteins to nuclear domains (Li & Bingham, 1991; Hedley et al., 1995; Caceres et al., 1997), referred to as interchromatin granule clusters or "speckles," which concentrate many splicing components as well as hyperphosphorylated RNA polymerase II, some 3′-end processing components, and poly(A)+ RNA (Lawrence et al., 1993; Spector, 1993; Misteli & Spector, 1998). These structures are thought to function as splicing factor storage sites as well as sites where specific transcripts are processed. Overexpression of different SR protein kinases results in the dispersal of these structures (Gui et al., 1994; Colwill et al., 1996). Thus, RS domains may be important for the assembly of higher-order nuclear structures in which the components of different processes are physically juxtaposed to facilitate efficient and coordinated processing events. However, the molecular basis underlying the formation of speckle structures and their significance in relation to how different nuclear processes are integrated is not understood.

To understand in more detail the potential roles of RS domain proteins, we have utilized computational sequence analysis techniques to perform a comprehensive survey of proteins containing RS domains that are encoded within the entire genomes of *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. This survey, besides identifying previously identified SR family and SR-related proteins involved in splicing, has uncovered many new RS domain proteins that are associated functionally with different cellular processes. The results demonstrate a useful strategy for the identification of RS domain proteins, and provide a database of new factors that are candidates for forming interactions involved at different steps in the expression of RNA polymerase II transcripts, as well as in other cellular functions.

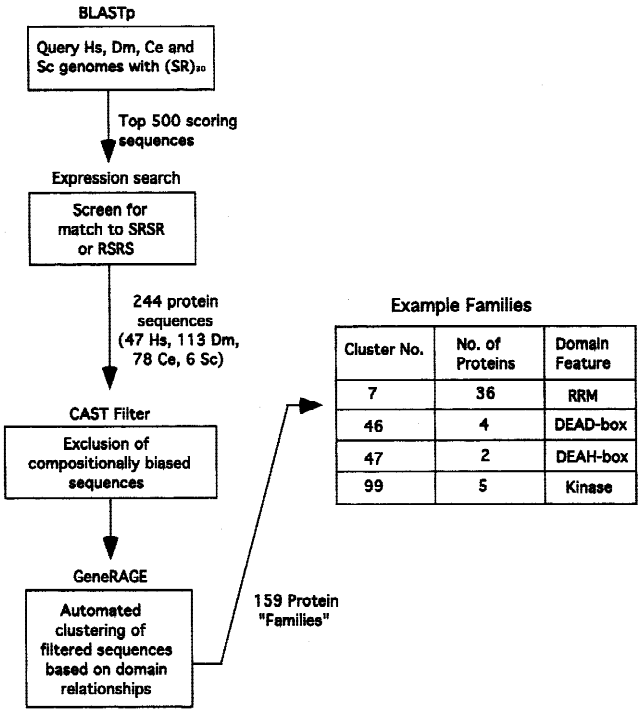

FIGURE 1. Bioinformatic strategy for the genome-wide analysis of RS domain proteins. Each step in the search and clustering procedure is indicated (refer to Materials and Methods for details). Examples of individual protein families identified by the GeneRage algorithm are shown. See Table 1 legend for key to abbreviations.

RESULTS AND DISCUSSION

Genomic mining of RS domain proteins

To establish search criteria for identifying protein sequences based on the presence of an RS domain, it was necessary to first establish an operational definition for an RS domain. To this end, we compared the sequences of several characterized SR family and SR-related splicing factors that contain short, functionally defined, RS domains in order to determine minimal features shared between these domains. All of the compared proteins contained several distributed SR or RS dipeptides and at least one stretch of two or more tandemly repeated SR/RS dipeptides (data not shown). We therefore applied search procedures for proteins containing a domain with these features. In particular, according to our filtering criteria, each protein contained a minimum of two or more consecutive SR/RS-repeat runs, usually within a region compositionally biased in SR/RS dipeptides and other motifs, often consisting of alternating arginine and/or serine residues (refer to Fig. 1 and Materials and Methods for details).

The search procedure identified 244 unique RS domain protein sequences. Among these, 47 proteins were from *H. sapiens*, 113 from *D. melanogaster*, 78 from *C. elegans*, and 6 from *S. cerevisiae*. The number of human proteins retrieved was underrepresented due to the draft nature of the human genome and possibly incorrect gene detection (Ensembl release date: June 2000; data not shown). The proteins were next clustered automatically based on their similarity outside of regions containing compositionally biased sequences (see Materials and Methods). The clustering grouped intra- and interspecies homologs as well as proteins containing significant domain similarity (see Materials and Methods and below). This step yielded 159 distinct

protein "families," of which 44 contained more than one protein and the remaining 115 contained only a single protein. The largest family (cluster 7) contained 36 distinct members, predominantly representing RS domain proteins with one or more RRM. Of the 244 RS domain proteins in the data set, 87 (11 from *H. sapiens*, 48 from *D. melanogaster*, and 28 from *C. elegans*) were represented in more than one family due to overlapping domain relationships.

A full BLASTp search of the nonredundant protein sequence database (nrdb) with each of the 244 protein sequences identified the closest relatives, although not necessarily containing an SR/RS-repeat sequence (see Materials and Methods). This identified numerous human homologs of *D. melanogaster* and *C. elegans* RS domain proteins that were absent from the *H. sapiens* Ensembl data set. The same RS domain criteria and two-step search procedure applied above (see Materials and Methods) was therefore used to determine which of the human protein homologs identified from searching the nrdb contain an RS domain. This identified 127 human RS domain proteins, of which 82 match at least one of the entries from the initial *H. sapiens* data set and 45 match at least one entry from *Drosophila* and/or *C. elegans*. All of our results, including the set of 244 proteins clustered into the 159 families, with a description of the domain structure of each protein, its closest relatives (including annotated human RS domain-containing homologs), and links to the original species databases, are available on the World Wide Web (http://maine.ebi.ac.uk:8000/sr/).

General properties of genome-wide sets of RS domain proteins

A significant feature of the data set is that the vast majority of entries are either known nuclear proteins, or else are related to nuclear-localized proteins. In fact, only 4 of the 244 protein sequences (entries 276, 288, 294, and 328) were annotated as having a possible role or sequence relationship with proteins that are nonnuclear. Moreover, the majority of the metazoan proteins in the data set are linked to the synthesis or processing of RNA polymerase II (pol II) transcripts, whereas none of the proteins are known to be associated with the biogenesis of either RNA pol I or pol III transcripts. These characteristics reduce the probability of false positives in the content of the data set, because they are consistent with previous evidence indicating a role for RS domains in the nuclear targeting of proteins, as well as for functions associated specifically with the metabolism of pre-mRNA transcripts, but not pol I or pol III transcripts (see Introduction).

The data set contains essentially all the previously identified SR family and SR-related proteins in *H. sapiens*, *D. melanogaster*, and *C. elegans*. For example, it contains all of the SR family or putative orthologs of SR family proteins (corresponding to: SRp20/RBP-1, SRp30a(ASF/SF2), SRp30b(SC35/PR64), SRp30c, 9G8, SRp40, SRp46, SRp55/B52, SRp75, and p54) identified in previous searches of the *D. melanogaster* and *C. elegans* genomes (Longman et al., 2000; Mount & Salz, 2000; data not shown; refer to the above Web site). However, as mentioned above, several human SR family and SR-related proteins were not retrieved from the initial searches because they were missing from the Ensembl data set used in the analysis. Nevertheless, these proteins were retrieved using the same RS domain search criteria from nrdb, and are included in the "closest relative" column in the complete data set (refer to: http://maine.ebi.ac.uk:8000/sr/; data not shown). Table 1 shows a representative set of the SR-related splicing factors that were identified in the searches, as well as putative orthologs of these proteins.

Interestingly, a large number of the data set entries are proteins not previously noted to contain RS domains, but which were identified and characterized in functional contexts other than splicing. Several of these are represented in only one species, whereas others are represented by homologs in at least two of the species surveyed. Representative examples of particular interest, sorted according to general functional criteria, are listed in Table 1 and are discussed below. The domain structures of a subset of the new SR-related proteins that are evolutionarily conserved is shown schematically in Figure 2 (see discussion below).

RS domain proteins associated with chromatin remodeling and/or transcription by RNA polymerase II

Many proteins detected in our search have been characterized as, or are related to, factors associated with chromatin regulation and/or transcription by RNA polymerase II. These include Acinus proteins (entries 189–191), a *Drosophila* bromodomain-containing protein that is 42% identical to the *S. cerevisiae* histone acetyl transferase GCN5 (entry 212), and a *Drosophila* protein that is 39% identical to HsCIR (CBF1-interacting corepressor; entry 108). CIR has been reported to mediate the association of histone deacetylase-complexes with DNA by interacting with the DNA-binding factor CBF1 (Hsieh et al., 1999). Inspection of the human homolog of DmCIR reveals that it also contains a serine and arginine-rich domain, although this did not include two tandem SR repeats and therefore was not identified in the searches (data not shown). In contrast, both *H. sapiens* and *D. melanogaster* homologs of Acinus were identified in the searches as RS domain proteins.

Hs-Acinus was originally identified as a target of Caspase-3, a cysteine protease involved in activating chromatin condensation and nuclear fragmentation dur-

TABLE 1. Selected examples of SR-related proteins identified from the database searches, sorted according to cluster assignment and functional association.[a]

Functional association	Entry number	Cluster number	Species	Motifs (other than RS)	Name of protein/homolog
Splicing (SR-related)	3	2	Dm		SRM300
	4	3	Ce		SWAP2
	5	3	Dm		SWAP2
	6	4	Hs		SWAP2
	9	6	Hs		SIP1
	23	7	Hs	RRM	RNPS1
	35	7, 64	Dm	RRM	U1-70K
	38	7, 64, 110, 127, 131, 141	Dm	RRM	U2AF-50
	44	7, 64, 127, 141	Ce	RRM	U1-70K
	76	21	Dm	SURP	SWAP1
	89	28	Dm	KH-RBD/ZNF	SF1
	90	29	Dm	PWI	SRM160
	101	38	Dm		TRA
	111	46	Ce	DEAD-BOX	U5-100K
	114	46	Dm	DEAD-BOX	U5-100K
	115	47	Ce	DEAH-BOX	HRH1
	116	47	Dm	S1-RBD/DEAH-BOX	HRH1
	244	120	Dm	RRM	TRA2
3′-end processing	97	36	Ce		FIP1
	98	36	Dm	FF	FIP1
	119	50	Ce		CF-IM 68K
	120	50	Dm		CF-IM 68K
Chromatin-associated	108	44	Dm		CIR
	190	87	Hs	RRM	ACINUS
	191	87	Dm	(RRM)	ACINUS
	212	103	Dm	BROMO	GCN5
Transcription (RNA pol II-associated)	11	6	Dm	PHD/ZNF/RING	SCAF1
	17	7	Hs		SCAF4
	25	7, 17	Hs		SCAF10/SR-CYP
	30	7, 17, 64, 127, 131	Dm	RRM	SCAF8
	63	16	Hs		SCAF9
	75	20	Ce		SRP129/SCAF11
	82	25	Ce	ZNF/RING/PHOS	FCP1a
	99	36	Dm	WW	CA150
	164	69	Ce	CYCLIN	CYCLIN L
	165	69	Dm	CYCLIN	CYCLIN L
	249	124	Dm		DSIF-P160/ SPT5
Transcription (other)	24	7, 17, 64, 127	Hs		LISCH
	85	27	Hs		CACTIN1
	87	27	Dm		CACTIN1
	133	60	Dm	ZNF/RING	NF-X1/SHUTTLECRAFT
	208	100	Hs		BTF
	291	137	Dm	PHD	ALHAMBRA
Kinases and phosphatases	203	99	Hs	KIN	PRP4-RELATED KINASE
	204	99	Hs	KIN	CLK-2 KINASE
	206	99	Dm	KIN	PITSLRE KINASE
	207	99	Dm	KIN	CRK7 KINASE
	320	148	Sc	RHOD	Ppz1p
	324	152	Sc	PHOS	Mip1p/ Cdc25p
Cell structure	168	72	Hs	B41	BAND 4.1-LIKE
	169	72	Dm	B41	BAND 4.1-LIKE
	235	114	Sc		Sla1p

[a]The entry and cluster numbers for each protein correspond to those in the complete list, which is available on the World Wide Web (http://maine.ebi.ac.uk:8000/sr/). The species to which each protein belongs as well as its constituent domains are listed. Hs: *Homo sapiens*; Dm: *Drosophila melanogaster*, Ce: *C. elegans*; Sc: *Saccharomyces cerevisiae*; Bromo: bromodomain; CYCLIN: cyclin domain; DEA(X): RNA-dependent helicase/RNA-dependent ATPase motifs; KIN: kinase motif; RBD: (KH-/S1-type) RNA-binding domain; RRM: (RNP-consensus-type) RNA recognition motif; PHD/RING: nucleic-acid-binding domains; PHOS: phosphatase motif; RHOD: rhodanese homology domain; WW, protein–protein interaction domain; ZNF: zinc finger DNA-binding domain.

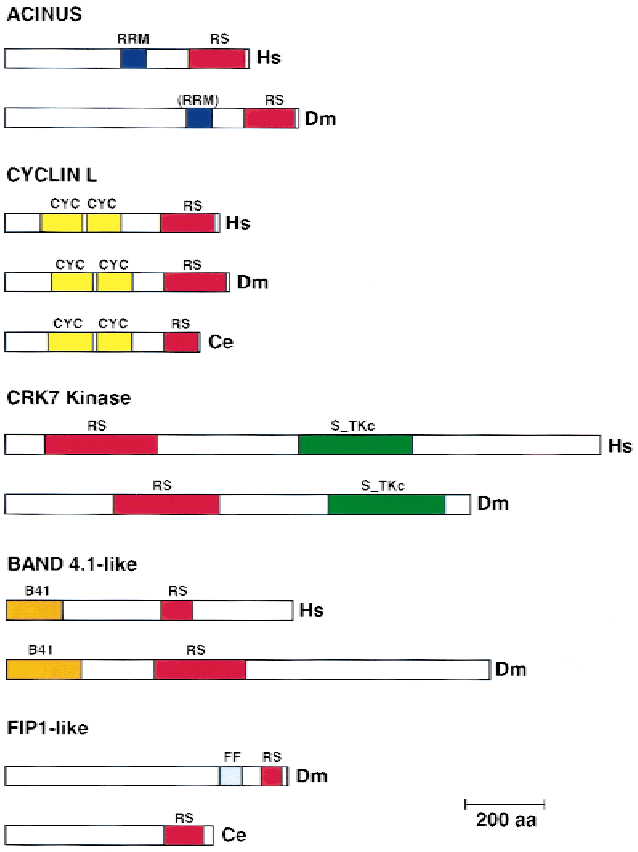

FIGURE 2. Domain architecture of selected examples of new SR-related proteins that are evolutionarily conserved. Refer to Table 1 for a key to the identity of the different motifs depicted. The proteins and constituent motifs are shown approximately to scale, where the RS domain encompasses the most SR/RS dipeptide-rich region in each protein.

ing apoptosis, although a role for this protein during normal cellular growth has not been determined (Sahara et al., 1999). Interestingly, besides an RS domain, both *H. sapiens* and *D. melanogaster* homologs of Acinus proteins contain an RRM. Hs-Acinus also contains a SAF-A/B, Acinus and PIAS (SAP) motif, which, in the scaffold attachment factors (SAF-A and -B), is implicated in binding to nuclear matrix or scaffold attachment regions (SARs) that border transcriptionally active chromosome domains (Aravind & Koonin, 2000). Intriguingly, several other SAP motif-containing proteins also have RNA-binding domains and/or links with RNA processing (Aravind & Koonin, 2000). For example, SAF-A (also known as hnRNP-U) is an RGG-box RNA binding protein that, like Acinus, is a target of caspase cleavage during apoptosis. Moreover, SAF-B has been reported to interact with both RNA polymerase II and several RS domain proteins (Nayler et al., 1998). The identification of a conserved RS domain in Acinus-related proteins, together with the previous observation that Hs-Acinus contains an RRM, suggests that these proteins might function in pre-mRNA processing as well as in mediating apoptosis-induced chromatin condensation.

Metazoan RS domain proteins with links to transcription and the cell cycle

In addition to the identification of previously identified RS domain proteins associated with the carboxyl-terminal domain (CTD) of RNA pol II such as SCAFs (see Introduction, Table 1, and data set entries 11, 17, 25, 30, 63, 75), the searches identified several RNA pol II-associated factors not previously recognized as RS domain proteins. These include the *C. elegans* homolog of the CTD phosphatase FCP1a (entry 82), the *Drosophila* homolog of Hs-CA150 (entry 99), which is a factor involved in HIV-1 Tat-mediated activation of transcription as well as transcriptional elongation (Sune et al., 1997), and the *Drosophila* homolog of Sc-Spt5 (entry 249; Kaplan et al., 2000), which is also involved in transcriptional elongation by RNA pol II (Wada et al., 1998). Interestingly, *H. sapiens*, *D. melanogaster*, and *C. elegans* homologs of a recently identified cyclin-related protein (designated CYCLIN L) were identified as RS domain proteins (entries 164, 165; see Fig. 2). CYCLIN L contains extensive similarity to other cyclins, including two copies of the "cyclin" domain, which is also found in the transcription factor TFIIB and the Retinoblastoma protein. Although not functionally characterized, CYCLIN L is closely related to CYCLIN K, which is associated with RNA pol II and is thought to modulate the activity of cyclin-dependent kinase-9 (CDK9), which phosphorylates the RNA pol II CTD (Edwards et al., 1998).

Other factors implicated in transcription identified as RS domain proteins include the following: Hs-LISCH, a protein with 89% identity to a liver-specific bHLH-Zip transcription factor (refer to entry 24); Hs-BTF (entry 208), an apoptosis-promoting transcriptional repressor that interacts with Bcl-2-related proteins (Kasof et al., 1999); Dm-Alhambra protein (entry 291), which is related to the mammalian transcription factors AF-10 and AF-17; the *Drosophila* homolog of Hs-NF-X1 (entry 133), a cytokine-inducible transcription factor implicated in the control of immune responses (Stroumbakis et al., 1996); and both the human and *Drosophila* homologs of Cactin (entries 85–87). Cactin was originally isolated by its interaction with the Rel family transcription factor Dm-I-kappaB (Cactus) and, like Dm-NF-X1, is important for *Drosophila* development (Lin et al., 2000).

The identification of several RS domain proteins as factors that are functionally associated with transcription initiation and elongation suggests them as possible additional candidates for mediating the coupling of transcription to pre-mRNA processing (see Introduction and below). Moreover, the identification of a conserved RS domain in CYCLIN L suggests it as a candidate for potentially linking pre-mRNA processing with transcription as well as the cell cycle. However, it is also interesting to consider that the RS domains of these proteins might facilitate the formation of protein–

protein interactions that are important for transcription and/or the cell cycle, independent of splicing.

RS domain proteins associated with 3′-end processing

Previous isolation of a cDNA encoding the 68-kDa subunit of the human pre-mRNA cleavage factor (CF-Im-68) identified it as an RS domain protein (Ruegsegger et al., 1998). In the present study, both the *Drosophila* and *C. elegans* homologs of CF-Im-68 (entries 119, 120) were also identified as RS domain proteins, demonstrating that the RS domain is a highly conserved feature of this cleavage factor. Interestingly, *Drosophila* and *C. elegans* proteins that are related to Fip1p (entries 97 and 98), a factor involved in the polyadenylation step of 3′-end formation in *S. cerevisiae* (Preker et al., 1995), were also both identified as RS domain proteins (Fig. 2). Like CF-Im-68, these metazoan proteins represent candidates for linking splicing and 3′-end formation via RS domain-mediated interactions, although a function of these domains more specifically associated with 3′-end processing is another possibility.

Metazoan RS domain proteins that are kinases and phosphatases

Our search procedure has resulted in the identification of several kinases (mostly found in cluster 99), including members of the Clk/Sty kinase (CDC28/cdc2-like) family (e.g., entry 204), which previously were reported to contain RS domains and shown to function in the phosphorylation of serine and threonine residues within RS domains (Stojdl & Bell, 1999). Other entries correspond to human kinases not previously recognized as RS domain proteins. These include CDK-like proteins, with homology to MAPKs, that have over 90% sequence identity to hPRP4 (entries 203, 205; Huang et al., 2000). Human PRP4, like ScPrp4p (which lacks an RS domain), has been identified as a component of U4/U6 snRNP although its precise function is not known (Horowitz et al., 1997; Teigelkamp et al., 1998). SR family proteins are known to be important for the recruitment of U4/U6 and U5 snRNPs (as a U4/U6.U5 tri-snRNP particle) to pre-mRNA during spliceosome formation (Roscigno & Garcia-Blanco, 1995). It is therefore interesting to speculate that the RS domain of the hPRP4-related proteins might function in mediating interactions with SR family and/or other RS domain proteins that are important for the recruitment of the U4/U6.U5 tri-snRNP to splicing complexes. Moreover, differential phosphorylation by the kinase activity of these proteins could regulate these interactions.

Human and *Drosophila* homologs of another Cdc-2-like kinase, designated as CRK7, were both identified as RS domain proteins (refer to entry 207; Fig. 2). Although the function of this kinase is not known, HsCRK7 has been reported to colocalize with the SR family protein SC35 in nuclear speckles, suggesting that it may have a function associated with pre-mRNA processing (J. Pines, unpubl. data; refer to GenBank accession number AAF36401). Several of the metazoan RS domain proteins we have identified are known phosphatases or are regulatory subunits of phosphatases (e.g., CeFCP1a mentioned above and entries 88, 180, and 237) although in general the RS domains in these proteins are short and poorly conserved (data not shown). Interestingly, two of the six *S. cerevisiae* RS domain proteins in the list, Ppz1p and Mih1p, are known phosphatases having functions associated with cell cycle progression (entries 320, 324; see discussion of *S. cerevisiae* RS domain proteins below).

RS domain proteins with links to cell structure

Several of the new RS domain proteins have been characterized as, or are related to, actin-binding proteins with functions associated with the cytoskeleton, but which also have a nuclear connection. One of these, yeast Sla1p, is discussed below. Other examples are human and *Drosophila* entries that are closely related to Band 4.1 family proteins (entries 168, 169; Fig. 2). In erythrocytes, Band 4.1 proteins are thought to facilitate anchoring of the cytoskeleton to the outer cell membrane by promoting the interaction of spectrin with actin. However, these proteins have also been detected in the nucleus where they concentrate with splicing factors in nuclear speckles (Krauss et al., 1997). Moreover, Band 4.1 proteins have been detected in association with splicing complexes and several splicing factors, including SR family proteins (Lallena et al., 1998). Based on numerous findings of actin, actin-like proteins, as well as actin-binding proteins including Band 4.1 in the nucleus, it has been proposed that a nuclear homolog of the actin cytoskeleton exists and that such a structure might facilitate the organization of processes including splicing (Rando et al., 2000). Our findings represent the first evidence for a structural relationship between splicing factors and Band 4.1 family proteins. Thus, consistent with the previous evidence for an association between Band 4.1 family proteins and SR family proteins, it is possible to speculate that specific Band 4.1-related proteins interact directly with one or more of these splicing factors via their RS domains.

S. cerevisiae RS domain proteins

Although splicing factors in yeast and vertebrates are generally highly conserved, *S. cerevisiae* lacks homologs of SR family proteins and many SR-related proteins. Moreover, several vertebrate splicing factors contain RS domains that are missing from the *S. ce-*

revisiae orthologs of these proteins. A homolog of mammalian SR protein kinases (SRPKs) has been identified, Sky1p, which can phosphorylate the RS domains of metazoan SR proteins (Yeakley et al., 1999). Npl3p, an hnRNP-like nucleocytoplasmic shuttling protein that binds to poly(A)+ RNA, and which is implicated in mRNA export, has been identified as a substrate of Sky1p (Siebel et al., 1999). Npl3p contains an arginine/glycine-rich region that includes several dispersed SR dipeptides (Birney et al., 1993), one of which is phosphorylated by Sky1p (Yun & Fu, 2000; Gilbert et al., 2001). However, it does not contain two tandem SR/RS dipeptides and therefore was not identified in our analysis. It was therefore of considerable interest to determine whether the genome of *S. cerevisiae* encodes other SR-like proteins with more extensive SR/RS dipeptide-repeat stretches that might be substrates for Sky1p.

The searches identified six *S. cerevisiae* ORFs, of which three encode putative proteins (ORFPs: YGR278W [entry 103], YMR014W [entry 319], YDR474C [entry 325]) and three encode known proteins (Sla1p [entry 235], Ppz1p [entry 319] and Mih1p/Cdc25p [entry 324]; see Table 1). Of the six protein sequences, only YGR278W has a candidate RS domain-containing metazoan homolog, the putative protein product of the *C. elegans* ORF Q17336 (entry 102). Curiously, the RS domain in the *C. elegans* protein sequence is at the N-terminus whereas it resides at the C-terminus in the *S. cerevisiae* sequence.

Similar to Band 4.1 proteins described above, Sla1p is an actin-binding protein that is important for the organization of the cortical cytoskeleton (Holtzman et al., 1993). Deletion of *sla1* is synthetically lethal with a null mutation in *anc1*, which was originally isolated in a screen for yeast genes that enhance defects in actin functions (Welch & Drubin, 1994). Anc1p is a nuclear protein that has been identified as a component of the SWI/SNF chromatin-remodeling complex (Cairns et al., 1996). It is therefore possible that Sla1p, like Band 4.1 proteins, has a nuclear association, as well as a structural role in association with the actin cytoskeleton.

The other two characterized *S. cerevisiae* RS domain proteins, Mih1p and Ppz1p, are both phosphatases implicated in cell cycle progression. Mih1p, a homolog of *Schizosaccharomyces pombe* Cdc25, is a tyrosine phosphatase that controls the level of phosphorylation (at tyrosine 19) of the cell cycle regulatory kinase Cdc28p (Russell et al., 1989). The level of Mih1p is important for controlling the timing of nuclear division in yeast cells that have failed to produce a bud and are under control of a morphogenesis checkpoint (Sia et al., 1996).

Ppz1p has been implicated in the G1 to S transition during the cell cycle and its integrity is also important for the tolerance of yeast cells to salt (Clotet et al., 1999). Interestingly, Sky1p is also implicated in ion homeostasis; deletion of SKY1 renders yeast cells tolerant to salt, whereas overexpression of Sky1p results in increased sensitivity to salt (Erez & Kahana, 2001). This increased sensitivity depends on the integrity of Ppz1p, indicating that Sky1p and Ppz1p may function in the same pathway (Erez & Kahana, 2001). Based on this observation and the identification of Ppz1p as an RS domain protein in the present study, it is possible to speculate that the activity of Ppz1p might be regulated by phosphorylation of its RS domain by the action of Sky1p. For example, the differential phosphorylation of Ppz1p by Sky1p could function to regulate ion homeostasis. Moreover, it is also possible that Sky1p, the homolog of which in *S. pombe* (Dsk1) has been implicated in the regulation of mitosis (Takeuchi & Yanagida, 1993), might also function to control the activity of Ppz1p and/or Mih1p during cell cycle progression in *S. cerevisiae*.

Conclusions and evolutionary considerations

The systematic identification of RS domain proteins using computational sequence-analysis techniques in this study has revealed an array of new metazoan factors that are functionally associated with processes in addition to splicing, including chromatin structure/remodeling, transcription, cell cycle, cell structure, and ion homeostasis. The RS domains of these factors could function to support protein–protein interactions in these processes, as well as roles in subcellular targeting and/or the communication between different processes. The number of RS domain proteins identified varied significantly between species. Although the genome of *Drosophila* contains a smaller number of predicted protein-encoding genes than that of *C. elegans* (approximately 14,000 versus 19,000, respectively) more RS domain proteins were identified in *Drosophila*. It has been proposed that the greater organismal complexity of *Drosophila* compared to *C. elegans* might be accounted for in part by it having a greater number of protein isoforms generated by alternative splicing (reviewed in Black, 2000; Graveley, 2001). The higher number of RS domain proteins in relation to the total number of genes in *Drosophila* is consistent with the possibility that it has a greater requirement for regulated splicing. Conversely, the identification of only six ORFs in *S. cerevisiae* encoding RS domain proteins is consistent with this organism lacking a general requirement for the recognition of poorly conserved splice sites, and for the regulation of splice site selection. It is possible that the RS domains of the *S. cerevisiae* proteins identified in this study, such as the phosphatases Ppz1p and Mih1p, evolved to provide other functions, including the regulation of ion homeostasis and cell cycle progression.

MATERIALS AND METHODS

Refer to Figure 1 for a schematic outline of the procedures used. A BLAST search against all predicted protein ORFs from the entire genomes of *H. sapiens*, *D. melanogaster*, *C. elegans*, and *S. cerevisiae* was initially performed using an artificial RS domain sequence consisting of 30 tandemly repeated SR dipeptides as a query to extract sequences that are compositionally biased in alternating serine and arginine residues. To select only the protein sequences containing an RS domain consisting of one or more stretches of at least two tandem SR/RS dipeptides, a regular expression search requiring a perfect match to either SRSR or RSRS was applied to the 500 top-scoring sequences from the initial similarity search. This filtering procedure yielded 244 protein sequences, which were subsequently clustered into related groups ("families") using the GeneRAGE algorithm (Enright & Ouzounis, 2000). GeneRAGE uses protein similarity information from BLAST to automatically generate protein families based on the detected domain structures of the input sequences. During the clustering procedure, compositionally biased regions in the sequences, including the RS domains, were masked using the CAST algorithm (Promponas et al., 2000). The clustering resulted in 159 distinct families containing a total of 331 entries. Each of these entries corresponded to 1 of the 244 protein sequences, sorted into one or more of the clusters depending on the relationship between its domain structure and other proteins in the data set. All 244 of the protein sequences were used as queries (masked with CAST-cutoff score 40) against the nonredundant protein sequence database to identify related proteins (NCBI BLAST version 2.0; E-value cutoff 10e-06). Human homologs identified in the BLAST searches were screened using the procedures described above to identify those that contain an RS domain (these are indicated with an asterisk in the database posted at http://maine.ebi.ac.uk:8000/sr/). Known sequence motifs in each protein were detected using the SMART system (Ogiwara et al., 1996).

ACKNOWLEDGMENTS

We thank Debbie Field, Deming Xu, Mike Tyers, and members of the Blencowe laboratory for helpful comments and suggestions on the manuscript. This work was supported in part by an Operating Grant and Scholar Award from the Canadian Institutes for Health Research to B.J.B.

Received June 5, 2001; returned for revision July 11, 2001; revised manuscript received September 7, 2001

REFERENCES

Aravind L, Koonin EV. 2000. SAP—a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem Sci 25*:112–114.

Birney E, Kumar S, Krainer AR. 1993. Analysis of the RNA-recognition motif and RS and RGG domains: Conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res 21*:5803–5816.

Black DL. 2000. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell 103*:367–370.

Blencowe BJ. 2000. Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci 25*:106–110.

Blencowe BJ, Bowman AL, McCracken S, Rosonina E. 1999. SR-related proteins and the processing of messenger RNA precursors. *Biochem Cell Biol 77*:277–291.

Burge C, Tuschl T, Sharp PA. 1999. *Splicing of precursors to mRNAs by the spliceosomes.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Caceres JF, Misteli T, Screaton GR, Spector DL, Krainer AR. 1997. Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J Cell Biol 138*:225–238.

Cairns BR, Henry NL, Kornberg RD. 1996. TFG/TAF30/ANC1, a component of the yeast SWI/SNF complex that is similar to the leukemogenic proteins ENL and AF-9. *Mol Cell Biol 16*:3308–3316.

Clotet J, Gari E, Aldea M, Arino J. 1999. The yeast ser/thr phosphatases sit4 and ppz1 play opposite roles in regulation of the cell cycle. *Mol Cell Biol 19*:2408–2415.

Colwill K, Pawson T, Andrews B, Prasad J, Manley J, Bell J, Duncan P. 1996. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J 15*:265–275.

Corden JL, Patturajan M. 1997. A CTD function linking transcription to splicing. *Trends Biochem Sci 22*:413–416.

Cramer P, Caceres JF, Cazalla D, Kadener S, Muro AF, Baralle FE, Kornblihtt AR. 1999. Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell 4*:251–258.

Du C, McGuffin ME, Dauwalder B, Rabinow L, Mattox W. 1998. Protein phosphorylation plays an essential role in the regulation of alternative splicing and sex determination in *Drosophila*. *Mol Cell 2*:741–750.

Edwards MC, Wong C, Elledge SJ. 1998. Human cyclin K, a novel RNA polymerase II-associated cyclin possessing both carboxy-terminal domain kinase and Cdk-activating kinase activity. *Mol Cell Biol 18*:4291–4300.

Enright AJ, Ouzounis CA. 2000. GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics 16*:451–457.

Erez O, Kahana C. 2001. Screening for modulators of spermine tolerance identifies sky1, the SR protein kinase of *Saccharomyces cerevisiae*, as a regulator of polyamine transport and ion homeostasis. *Mol Cell Biol 21*:175–184.

Fu X-D. 1995. The superfamily of arginine/serine-rich splicing factors. *RNA 1*:663–680.

Gilbert W, Siebel CW, Guthrie C. 2001. Phosphorylation by Sky1p promotes Npl3p shuttling and mRNA dissociation. *RNA 7*:302–313.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA 6*:1197–1211.

Graveley BR. 2001. Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet 17*:100–107.

Gui J-F, Lane WS, Fu X-D. 1994. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature 369*:678–682.

Hedley M, Amrein H, Maniatis T. 1995. An amino acid sequence motif sufficient for subnuclear localization of an arginine/serine-rich splicing factor. *Proc Natl Acad Sci USA 92*:11524–11528.

Hirose Y, Manley JL. 2000. RNA polymerase II and the integration of nuclear events. *Genes & Dev 14*:1415–1429.

Holtzman DA, Yang S, Drubin DG. 1993. Synthetic-lethal interactions identify two novel genes, SLA1 and SLA2, that control membrane cytoskeleton assembly in *Saccharomyces cerevisiae*. *J Cell Biol 122*:635–644.

Horowitz DS, Kobayashi R, Krainer AR. 1997. A new cyclophilin and the human homologs of yeast Prp3 and Prp4 form a complex associated with U4/U6 snRNPs. *RNA 3*:1374–1387.

Hsieh JJ, Zhou S, Chen L, Young DB, Hayward SD. 1999. CIR, a corepressor linking the DNA binding factor CBF1 to the histone deacetylase complex. *Proc Natl Acad Sci USA 96*:23–28.

Huang Y, Deng T, Winston BW. 2000. Characterization of hPRP4 kinase activation: Potential role in signaling. *Biochem Biophys Res Commun 271*:456–463.

Kaplan CD, Morris JR, Wu C, Winston F. 2000. Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in *D. melanogaster*. *Genes & Dev 14*:2623–2634.

Kasof GM, Goyal L, White E. 1999. Btf, a novel death-promoting transcriptional repressor that interacts with Bcl-2-related proteins. *Mol Cell Biol 19*:4390–4404.

Kramer A. 1996. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Ann Rev Biochem 65*:367–409.

Krauss SW, Larabell CA, Lockett S, Gascard P, Penman S, Mohandas N, Chasis JA. 1997. Structural protein 4.1 in the nucleus of human cells: Dynamic rearrangements during cell division. *J Cell Biol 137*:275–289.

Lallena MJ, Martinez C, Valcarcel J, Correas I. 1998. Functional association of nuclear protein 4.1 with pre-mRNA splicing factors. *J Cell Sci 111*:1963–1971.

Lawrence J, Carter K, Xing X. 1993. Probing functional organization within the nucleus: Is genome structure integrated with RNA metabolism? *Cold Spring Harbor Symp Quan Biol 58*:807–818.

Li H, Bingham PM. 1991. Arginine/serine-rich domains of the *su(w^a)* and *tra* RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. *Cell 67*:335–342.

Lin P, Huang LH, Steward R. 2000. Cactin, a conserved protein that interacts with the *Drosophila* IkappaB protein cactus and modulates its function. *Mech Dev 94*:57–65.

Longman D, Johnstone IL, Caceres JF. 2000. Functional characterization of SR and SR-related genes in *Caenorhabditis elegans*. *EMBO J 19*:1625–1637.

Manley J, Tacke R. 1996. SR proteins and splicing control. *Genes & Dev 10*:1569–1579.

Misteli T, Spector DL. 1998. The cellular organization of gene expression. *Curr Opin Cell Biol 10*:323–331.

Mount SM, Salz HK. 2000. Pre-messenger RNA processing factors in the *Drosophila* genome. *J Cell Biol 150*:F37–F44.

Nayler O, Stratling W, Bourquin JP, Stagljar I, Lindemann L, Jasper H, Hartmann AM, Fackelmayer FO, Ullrich A, Stamm S. 1998. SAF-B protein couples transcription and pre-mRNA splicing to SAR/MAR elements. *Nucleic Acids Res 26*:3542–3549.

Ogiwara A, Uchiyama I, Takagi T, Kanehisa M. 1996. Construction and analysis of a profile library characterizing groups of structurally known proteins. *Protein Sci 5*:1991–1999.

Preker PJ, Lingner J, Minvielle-Sebastia L, Keller W. 1995. The FIP1 gene encodes a component of a yeast pre-mRNA polyadenylation factor that directly interacts with poly(A) polymerase. *Cell 81*:379–389.

Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA. 2000. CAST: An iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics 16*:915–922.

Rando OJ, Zhao K, Crabtree GR. 2000. Searching for a function for nuclear actin. *Trends Cell Biol 10*:92–97.

Reed R, Palandjian L. 1997. *Spliceosome assembly*. Oxford: IRL Press.

Roscigno R, Garcia-Blanco M. 1995. SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA 1*:692–706.

Ruegsegger U, Blank D, Keller W. 1998. Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell 1*:243–253.

Russell P, Moreno S, Reed SI. 1989. Conservation of mitotic controls in fission and budding yeasts. *Cell 57*:295–303.

Sahara S, Aoto M, Eguchi Y, Imamoto N, Yoneda Y, Tsujimoto Y. 1999. Acinus is a caspase-3-activated protein required for apoptotic chromatin condensation. *Nature 401*:168–173.

Sia RA, Herald HA, Lew DJ. 1996. Cdc28 tyrosine phosphorylation and the morphogenesis checkpoint in budding yeast. *Mol Biol Cell 7*:1657–1666.

Siebel CW, Feng L, Guthrie C, Fu XD. 1999. Conservation in budding yeast of a kinase specific for SR splicing factors. *Proc Natl Acad Sci USA 96*:5440–5445.

Smith CW, Valcarcel J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem Sci 25*:381–388.

Spector D. 1993. Macromolecular domains within the cell nucleus. *Ann Rev Cell Biol 9*:265–315.

Stojdl DF, Bell JC. 1999. SR protein kinases: The splice of life. *Biochem Cell Biol 77*:293–298.

Stroumbakis ND, Li Z, Tolias PP. 1996. A homolog of human transcription factor NF-X1 encoded by the *Drosophila* shuttle craft gene is required in the embryonic central nervous system. *Mol Cell Biol 16*:192–201.

Sune C, Hayashi T, Liu Y, Lane WS, Young RA, Garcia-Blanco MA. 1997. CA150, a nuclear protein associated with the RNA polymerase II holoenzyme, is involved in Tat-activated human immunodeficiency virus type 1 transcription. *Mol Cell Biol 17*:6029–6039.

Takeuchi M, Yanagida M. 1993. A mitotic role for a novel fission yeast protein kinase dsk1 with cell cycle stage dependent phosphorylation and localization. *Mol Biol Cell 4*:247–260.

Teigelkamp S, Achsel T, Mundt C, Gothel SF, Cronshagen U, Lane WS, Marahiel M, Lührmann R. 1998. The 20kD protein of human [U4/U6.U5] tri-snRNPs is a novel cyclophilin that forms a complex with the U4/U6-specific 60kD and 90kD proteins. *RNA 4*:127–141.

Wada T, Takagi T, Yamaguchi Y, Ferdous A, Imai T, Hirose S, Sugimoto S, Yano K, Hartzog GA, Winston F, Buratowski S, Handa H. 1998. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & Dev 12*:343–356.

Welch MD, Drubin DG. 1994. A nuclear protein with sequence similarity to proteins implicated in human acute leukemias is important for cellular morphogenesis and actin cytoskeletal function in *Saccharomyces cerevisiae*. *Mol Biol Cell 5*:617–632.

Xiao SH, Manley JL. 1998. Phosphorylation-dephosphorylation differentially affects activities of splicing factor ASF/SF2. *EMBO J 17*:6359–6367.

Yeakley JM, Tronchere H, Olesen J, Dyck JA, Wang HY, Fu XD. 1999. Phosphorylation regulates in vivo interaction and molecular targeting of serine/arginine-rich pre-mRNA splicing factors. *J Cell Biol 145*:447–455.

Yun CY, Fu XD. 2000. Conserved SR protein kinase functions in nuclear import and its action is counteracted by arginine methylation in *Saccharomyces cerevisiae*. *J Cell Biol 150*:707–718.