Abstract

SET-domain (SET: Su(var)3-9, E(z) and Trithorax)-containing proteins were collected through sequence searches of the available databases. After removing redundancies, the proteins belonging to three families, SU(VAR)3-9, E(Z) and TRITHORAX, were selected. Analysis of the relationship between the different members is based on pairwise alignment, compilation, and comparison of their SET-domains. The level of homology of the SET-domains defined the distribution of the proteins into families and into clades within the families. The architecture of the entire protein supported the distribution pattern built upon SET-domain similarity. Parallel cladistic and protein-architecture analyses outlined two plausible criteria for predicting function. © 2002 Published by Elsevier Science B.V.

Keywords: Chromatin modifier; Database analysis; Evolution and homology of SET-domain genes

1. Introduction

A conserved, ~130-amino-acid sequence motif was initially identified at the C-terminal ends of three gene-regulatory factors in Drosophila accounting for its name, SET: Su(var)3-9, E(z) and Trithorax (Jones and Gelbart, 1993; Dorn et al., 1993; Tschiersch et al., 1994; Stassen et al., 1995). Subsequently, SET domain was discovered in Ash, another regulator of the Trithorax group (Tripoulas et al., 1996), but the internal location of the motif in the protein molecule is an architectural feature distinguishing Ash, from the previous three. SET-domain genes were recognized in yeast, animals and plants. SET-domain genes were found in bacteria as well, but their occurrence only in pathogenic bacteria is considered a result of horizontal gene-transfer (Stephens et al., 1998). Thus, the SET domain appears to have occurred for eukaryotic functions. It is a paradigm for a motif shared by factors of both activating and repressive complexes.

Only 3 years ago, the number of recognized SET-domain genes was around 40. Based on the homology of their SET domains they were distributed into four families, SU(VAR)3-9, E(Z), ASH1 and TRITHORAX (Jenuwein et al., 1998). In the current databases, there are around 300 real and hypothetical SET-domain genes outlining new families in addition to the known four. The SET-domain genes are widely represented in the eukaryotic genomes, but information about their function is limited. For members of one family (SU(VAR)3-9) specific enzyme activity was established (Rea et al., 2000; Bannister et al., 2001; Lachner et al., 2001; Tachibana et al., 2001). The histone-specific methylase activity associated with the SET-domain of SU(VAR)3-9 and its impact upon chromatin modification was extensively reviewed (Jenuwein, 2001; Jenuwein and Allis, 2001). Equally significant were the observations that similar motifs (SET- and chromo domains) may display highly specific activities dependent upon the context of individual genes (Rea et al., 2000; Bannister et al., 2001; Lachner et al., 2001).

The goal of this paper is to systematize and comprehensively analyze currently available proteins of three families, SU(VAR)3-9, E(Z), and TRITHORAX, using the potential of a whole-genome comparative approach for gathering and interpreting knowledge existing in the databases. The approach outlined an informative picture of a significant heterogeneity inside each family, illustrated the relationship among same-family members, and a tight correlation between the level of amino acid similarity of the SET domains and the evolution of the entire architecture of any given gene.

Abbreviations: SET, Su(var)3-9, E(z) and Trithorax; DAST, domain associated with SET in Trithorax; ORF, open reading frame; DART, Domain Architecture Retrieval Tool
* Corresponding author. Tel.: +1-402-420-3993.
E-mail addresses: ralvarez@purdue.edu (R. Alvarez-Venegas), zavramova2@unl.edu (Z. Avramova).

0378-1119/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0378-1119(02)00401-8
After removing the redundancies, the currently available genes of the three families were ordered by the hierarchy of the similarity of their SET domains. Maximum Parsimony (MP) cladograms ordering homologies into a nested hierarchy and summarizing patterns of character distribution were constructed. A clade contains a set of genes that share a last
common ancestor not shared with any of the other SET-domain genes. Reported data on the function of known SET-domain genes from the three families are discussed, showing the power of a cladistic approach to analyze unknown genes. The initial discovery of a conserved SET domain in proteins of both repressive (heterochromatin, Pc-G) and activating (Trx-G) complexes, made the SET motif a paradigm for an element shared by antagonistically acting factors. It was viewed as a link mechanistically connecting counteracting activities at a molecular level. Recent results, however, demonstrated that this has been an oversimplification (Rea et al., 2000; Bannister et al., 2001; Lachner et al., 2001; Jenuwein, 2001). Our analysis suggested also that conclusions on the nature and function of newly discovered genes, drawn solely from finding of a homology hit to individual peptide motifs, might not be valid. A cladistic approach, based on single-domain homology supported by the architecture of the entire gene, may provide a tool for a more reliable prediction of function.

2. The architecture of the proteins and SET-similarity distribution patterns

The criterion distributing the proteins to one of the three families was the relatedness of their SET-peptides. This cladistic approach revealed well-supported internal heterogeneity within each family. The architecture of members of various subgroups (built as predicted by the SMART, PFAM and DART programs) is discussed in relation to their homology clustering pattern.

Comparing the entire genomes of Saccharomyces cerevisiae (as a model for a unicellular) and of Caenorhabditis elegans (as a multicellular example), Chervitz et al. (1998) suggested that functions common for unicellular and multicellular species most probably use similar genes. Consequently, similar genes found in the genomes of a unicellular and a multicellular organism are most probably involved in similar functions defined as ‘core’ functions (Chervitz et al., 1998). Therefore, core-function genes code for proteins that have been conserved during the transition to multicellularity, while novel functions related to multicellularity use novel genes. This hypothesis will be the framework within which the data will be analyzed.

2.1. Relationship between the members of the SU (VAR)3-9 family

The members of this family grouped into six clades (Fig. 1). Clades I and II are the only clusters containing genes that have been cloned and for which a function has been established (see below). The remaining clades contain only putative proteins.

The clustering pattern is supported by the architecture of the proteins. Each particular clade has members with similar architecture, while those in sister-groups have the same architecture. Thus, a unique architectural feature of the proteins of clade I is the presence of a chromo domain, absent from the members of all other clades. Within clade I, the group of the invertebrate proteins carry a motif found in translation elongation factors EF-1α and EF-TU (Tschiersch et al., 1994; Stassen et al., 1995; Krauss and Reuter, 2000) absent from the human and mouse proteins. The biochemical role of these motifs is not clear but it is evident that this particular combination has evolved for insect-specific functions. Clade II groups animal proteins containing several copies of ANK repeats (Bork, 1993). ANK repeats found in bacteria, yeast, plants, and metazoa display both nuclear and transmembrane localization (Bork, 1993).
28


The human G9A is an MHC III gene with a histone methylase activity (Tachibana et al., 2001).

Clade III contains Arabidopsis genes, called SUVAR-Related (SUVR) because their SET domain was found to be closer to the SET of the human G9A than to the SUVARH39 (Baumbusch et al., 2001). The components of clade III, SUVR1, 2 and 4, do not contain other architectural elements beside the SET- and a cysteine-rich adjacent (Pre-SET) domain. The latter is a signature feature for the SU(VAR)3-9 family. It is interesting to note that the only member of the family from the genome of C. elegans (AAC71154) has a structure similar to the plant representatives, not the animal, and that by the degree of relatedness of its SET domain it is placed between the animal (clade II) and plant (clade III) proteins.

Clade IV grouped three unusual members: two apparently truncated Arabidopsis genes and an ORF encoding a putative human chimeric protein. SUVR 3 and 5 are the only members of the family that do not contain the pre-SET domain. The putative human protein carries regions with homology to Transposase 1 and a Reverse Integrate (Robertson and Zumpano, 1997). The subdivision of the five SUVR members (Baumbusch et al., 2001) into a cluster (carrying SUVR1, 2, and 4) and another (carrying SUVR 3 and 5), is supported by a difference in their molecular structure (compare clades IV and VI). Further, the plant components of clade VI segregate into two sub-groups. Despite their high similarity, the two clusters differ in the presence/absence of AT-hooks. A signature feature for the clade is the presence of a conserved domain, named G9a in the SMART program. It needs to be clarified that this G9a domain is identical to the YDG domain, described recently as a novel motif revealed in the Arabidopsis homologs (Baumbusch et al., 2001). A function for G9a (YDG) has not been established but its assembly with the SET domain has evolved as a plant-specific feature.

A well-supported clade carrying animal proteins (V) was resolved between the plant clades IV and VI. The members of clade V carry a split SET domain and a methyl CpG-binding domain (MBD) as signature features. It is notable that MBD is present also in the Drosophila protein, despite the apparent lack of methylated CpGs in the genome. In addition, the vertebrate, but not the fly, proteins contain two conserved (Tudor) domains. The Tudor domain, present in one or several copies, is found in eukaryotic proteins that colocalize with RNA and with complexes associated with single-stranded DNA (Ponting, 1997; Callebaut et al., 1997). A function for the Tudor domain has not been established but it does not bind directly DNA/RNA and appears to function as protein-interacting motif within nucleic acid/protein complexes (Hirose et al., 2000; Selenko et al., 2001).

2.2. Function of SU(VAR)3-9 family of members

A recent breakthrough was the discovery that SET-peptides posses protein-methylating enzyme activity (Rea et al., 2000). However, only genes of the Su(var)3-9 type encode protein methylase activity that selectively modifies lys9 of histone H3. The enzyme activity was mapped to the SET domain and the adjacent cys-rich regions. Neither the E(z) nor the Trithorax SET peptides have this activity (Rea et al., 2000; Bannister et al., 2001; Lachner et al., 2001; Nakayama et al., 2001). The SET-domain of SUV39H1 has weak sequence homology with plant protein methyltransferases (Rea et al., 2000). Some of the six homologous plant sequences have been classified as potential histone lysine transferases. Only one has been functionally characterized and was found to lack histone methylase activity (Klein and Houtz, 1995).

Until recently, the only known SET-proteins with methylase activity belonged to clade I. Histone methyltransferase activity was very recently reported for the SET domain of the human G9a from clade II (Tachibana et al., 2001). In addition to lys9 of histone H3, the G9a could modify lys27 and histone H1. It remains to be shown whether genes from other clades of the same family possess this activity.

The discovery that methylated lys9 of His3 creates a specific binding site for the chromodomain of HP1 (Bannister et al., 2001; Lachner et al., 2001) provided new insights into the mechanism of heterochromatin formation and maintenance. These issues have been recently reviewed (Jenuwein, 2001 and references therein) and will not be covered here. A unique structural feature of clade I genes is the presence of a chromodomain. Ever since the discovery of a chromodomain in Polycomb, the chromo motif has been considered a molecular link relating mechanistically the silencing at the homeotic gene loci and the silencing by heterochromatin (Cavalli and Paro, 1998). Mammalian chromodomain proteins and their function have also been recently exhaustively reviewed (Jones et al., 2000) and will not be discussed here. Only the chromodomains of HP1 and of Swi6 could bind specifically the methylated Lys9, while the chromo motifs of Su(var)3-9, Mi-2, or Polycomb could not (Bannister et al., 2001; Lachner et al., 2001). The possible role of the chromo domains of these latter proteins remains to be established but it is clearly different from binding methylated tails of histone H3. Thereby, even conserved peptide motifs may display highly specific functions subjected to the unique influence of the neighboring elements. The discovery that E(z) lacks histone-methylating activity, on the one hand, and that Polycomb is unable to bind the methylated Hs lys9, on the other hand, suggests that current ideas that the silencing activity of the Pc-G complex could be mechanistically similar to the formation of heterochromatin need to be revisited.

The SU(VAR) SET domain, itself, is regulated via a specific phosphorylation during G1/S transition. Forced expression of the gene antagonizes cell growth and the phosphorylation balance may be mediated by a factor, Sbf (Firestein et al., 2000). The product of the Schizosaccharomyces pombe CIC4 gene is important for centromeric integrity, for gene repression and for heterochromatin formation.
2.4. Function of E(Z) family members

The product of the Drosophila gene E(z) is a paradigm for a SET-domain protein from the Polycomb group (Pc-G). Genes homologous to E(z) were found in C. elegans, in mammals and in plants. A common structural feature for the E(z) family members is the presence of a cystein-rich region, CXC, immediately upstream of the SET domain. The family resolves into two well-supported clades (Fig. 2). In clade II, the Drosophila E(z) is separated from the mammalian homologs. The six mammalian members split further into two sister-groups. The architecture of the proteins supports this clustering pattern: only one sister-group carries SANT domains. The SANT domain was defined initially as a DNA binding motif (Aasland et al., 1996) but some may bind proteins (Cutler et al., 1998).

The other clade includes plant proteins with remarkable architectural similarity among themselves and to E(z). However, the plant representatives do not carry one of the two conserved regions (domain I). The SET-domains of two Arabidopsis proteins (CLF and AAD09108) are apparently closer to each other than to the SET of MEDEA (MEA). It is interesting that the cladistic approach positioned the plant MEA and the C. elegans Mes-2 on terminal branches, not clustered with any of the subgroups. Both proteins are maternally transmitted factors involved in embryogenesis and germ-line functions (Grossniklaus et al., 1998; Kiyosue et al., 1999; Luo et al., 1999; Kelly and Fire, 1998; Korf et al., 1998; Holdeman et al., 1998).

2.5. The TRITHORAX Family

The product of the Drosophila gene E(z) is a paradigm for a SET-domain protein from the Polycomb group (Pc-G). Genes homologous to E(z) were found in C. elegans, in mammals and in plants. A common structural feature for the E(z) family members is the presence of a cystein-rich region, CXC, immediately upstream of the SET domain. The family resolves into two well-supported clades (Fig. 2). In clade II, the Drosophila E(z) is separated from the mammalian homologs. The six mammalian members split further into two sister-groups. The architecture of the proteins supports this clustering pattern: only one sister-group carries SANT domains. The SANT domain was defined initially as a DNA binding motif (Aasland et al., 1996) but some may bind proteins (Cutler et al., 1998).

The other clade includes plant proteins with remarkable architectural similarity among themselves and to E(z). However, the plant representatives do not carry one of the two conserved regions (domain I). The SET-domains of two Arabidopsis proteins (CLF and AAD09108) are apparently closer to each other than to the SET of MEDEA (MEA). It is interesting that the cladistic approach positioned the plant MEA and the C. elegans Mes-2 on terminal branches, not clustered with any of the subgroups. Both proteins are maternally transmitted factors involved in embryogenesis and germ-line functions (Grossniklaus et al., 1998; Kiyosue et al., 1999; Luo et al., 1999; Kelly and Fire, 1998; Korf et al., 1998; Holdeman et al., 1998).

2.4. Function of E(Z) family members

The Drosophila E(z) gene, and its mammalian homologs, have been implicated in gene-repression (Jones and Gelbart, 1990; Cardoso et al., 1998; Van Lohuizen et al., 1998; Laible et al., 1997). Because the genes of clade I of the SU(VAR)3-9 family act also as repressors, the presence of the SET motif has been considered a common molecular basis for the assembly of repressive complexes (Laible et al., 1997; Pirrotta, 1998). The ability of EZH2 and E(z) to act as dose-dependent modifiers of PEV and to restore disrupted TPE in yeast (caused by deletion of the SET domain of SET1), supported this view (Laible et al., 1997). However, the lack of methylase activity of the E(z) SET-peptide clearly contradicts an assumption that the repressive function of Su(var)3-9 and of E(z) might be achieved via mechanisms similar to heterochromatin formation.

E(z) belongs to the Polycomb Group. The products of Pc-G genes (about 14 are cloned and characterized) may form several distinct complexes through various protein combinations (Richard et al., 1998; Shao et al., 1999 and references therein). Among the best-studied interactions is the specific interaction of E(z) and of Esc. It is remarkable that in all cases studied, the two proteins bind through a region at the N-end of the E(z) molecule, despite the apparent lack of a highly conserved N-region in the homologs from different species (Jones et al., 1998; Spillane et al., 2000; Yadegari et al., 2000).

Biochemical evidence has indicated that Esc and E(z), and EED and EZH2/EZH1 (the respective mammalian homologs) form a complex required at the initiation stage of repression. This complex, called Pc-Gi (initiation) complex has been required to stably maintain an already initiated repression pattern (Van Lohuizen et al., 1998; Richard et al., 1998; Shao et al., 1999). Repression by Pc-Gi is achieved via histone deacetylation: EED binds specifically HDAC1 and 2 and deacetylates histone H4 (Van der Vlag and Otte, 1999). These results connect directly repression by Pc-Gi with nucleosome modification.

Both E(z) and Esc are expressed early in embryogenesis and continue to be expressed throughout development. They might accomplish the bridging of the initially repressive function of gap-genes (Van Lohuizen et al., 1998). It is interesting that the only two interacting Pc-G proteins reported in C. elegans are homologs of E(z) and Esc, involved in germ-line maintenance (Kelly and Fire, 1998; Korf et al., 1998; Holdeman et al., 1998). Likewise, the only known interactions between Pc-G proteins in plants are the genetic and the direct interaction of the products of MEDEA (MEA) and Fertilization independent endosperm (FIE) (respective Arabidopsis homologs of E(z) and Esc) (Grossniklaus et al., 1998; Kiyosue et al., 1999; Luo et al., 1999; Spillane et al., 2000; Ohad et al., 1999). Apparently, the Pc-Gi complex is an ancient regulatory mechanism existing before the separation of the two kingdoms. It will be noted that a binding partner for CLF, the only other cloned plant E(z) homolog, has not been established yet. It will be noted also that E(z) and Esc are not obligatory partners and that E(z) does not necessarily function always through Esc. Thus, the N-terminal region of the human ENX1 binds the product of a different gene, the protooncogene Vav, implicating the complex in hematopoietic signal transduction (Hobert et al., 1996).

Two distinct loci encode the mammalian homologs, EZH1 and EZH2, with complementary expression profiles during development (Cardoso et al., 1998; Abel et al., 1996). There are no data regarding the function of their SET domains but only the SET domain of EZH2, and not of EZH1, can specifically bind a region in the XNPATR-X gene product. The SET domains of the two proteins differ by only 8 a.a. suggesting a basis for their functional specificity (Cardoso et al., 1998). This difference, along with the presence of SANT domains in the EZH2 homologs, is also reflected by their classification into separate SET-homology clusters (Fig. 2).
Fig. 2. SET-domain homology distribution of members of the E(Z) family. Conditions and abbreviations are as in Fig. 1. Proteins entered: Mm, mouse (BBA25018); MmEzh1 (AAC52655); Ezh1, human (NP_001982); Ezh2, human (NP_004447); Ezh2 Mm, mouse (NP_031997); Os, rice (BAB1944); clf At, CURLY LEAF of Arabidopsis (CAA71599); At, Arabidopsis (AAD09108); medea At, MEDEA of Arabidopsis (AAC39445); Ce, C. elegans (AAC27124).
Fig. 3. SET-domain homology distribution of members of the TRITHORAX family. Conditions and abbreviations are as in Fig. 1. ePHD is a motif defined as extended PHD finger (Baumbusch et al., 2001); Luteo is a motif from the SMART site with unclear function; RRM is an RNA recognition motif (see SMART). Proteins entered: At ATX1-5, Arabidopsis homologs of Trithorax (AF329273, AAF29390, CAB71104, ALL2215, BAA97320, respectively); Ce, C. elegans (T24864); ALR Dm, All-Leukemia Related from Drosophila (T12687); Hs, human (AAF74766); ALL Tr, Leukemia related from T. rubripes, (AAC34383); Hs, human (NP_003473); Dv, D. virilis (T13857); TRX Dm, D. melanogaster Trithorax (P20659); Hs, human (NP_055542); MLL Tr, Mixed leukemia lineage of T. rubripes (AAC41377); Gg, G. gallus (CAA09454); Hs, human (A44265); HRX Mm, mouse (P55200); Sp, S. pombe (PAB1652); SET1Sc, SET-domain protein1 in S. cerevisiae (P38827); Ce, C. elegans (AAA21163); Dm, D. melanogaster (AAF45425); Hs, human (BAA83028).
3. Clade I contains two members from unicellular species (yeast), proteins from a nematode, *Drosophila*, and humans. A unique feature of clade I proteins is that they carry the SET domain as an only architectural motif. An apparent exception is the protein from the fission yeast with a peptide defined as RNA-recognition motif. For only one protein, the yeast SET1, functions have been reported (see below). Because the architecture of the proteins from the multicellular species within the clade is similar to SET1, it could be expected that these proteins might be involved in ‘core’ eukaryotic functions (Chervitz et al., 1998) conserved during the transition to metazoans. It is the only clade among all three SET-domain families containing proteins with such type of functions.

Proteins found only in animals constitute two well-supported clades, II and III. Clade II dissolves into two sister-groups containing either *Drosophila* or vertebrate proteins. Notable differences between the members of clades II and III are the split DAST boxes, the different number, and the positioning of the PHD-fingers. The bromo- and the CXXC conserved domains, appearing only in the vertebrate proteins support the sister-group segregation within clade II.

The *Drosophila trithorax* and the human *HRX*, are the only proteins from clade II for which function has been studied (see below). The members of clade III share similarly positioned SET, DAST, PHD finger and HMG peptides. The two invertebrate proteins are shorter and show variance in the position of the individual architectural elements.

Clade IV is monophyletic including five *Arabidopsis* genes found in the current Database. The *Arabidopsis* proteins are of a smaller size and are the only cluster containing a PWWP-domain. The role of this motif has not been established but PWWP-domains are found in regulatory factors and in de novo methyltransferases (Stee et al., 2000; Okano et al., 1999). Remarkably, the PWWP-domain is not present in any other member of the family, suggesting that this particular combination has evolved in the *trithorax* genes acquired for plant-specific functions. Major distinctions among the *Arabidopsis* Trithorax proteins is that only ATX1 and ATX2 contain Tudor- and DAST domains, supporting their placement into a sister group separated from the other three Trithorax homologs. It is interesting to note also that only ATX1 and ATX2, from the entire family, contain Tudor which is found also in some members of the SU(VAR)3-9 family. The importance of this combination for the function of the proteins from the two clades is not clear.

### 2.6. Function of TRITHORAX family members

The Trithorax group, Trx-G, contains proteins connected in a genetic network in *Drosophila* that may counteract the repressive effects of the Pc-G. The interactions between various individual members of the Pc-G/Trx-G complexes have been actively studied and reviewed (Pirrotta, 1998; Brock and van Lohuizen, 2001). The involvement of the Pc-G/Trx-G complexes in a variety of processes such as centromeric and telomeric silencing, cell-cycle regulation, disease and cancer (Sharples and DePinho, 1999; Van Lohuizen, 1999) have defined the Pc-G/Trx-G proteins as major players unifying global regulation at cellular and organismal levels.

The paradigm for a Trx-G member is the *Drosophila* Trithorax. Identified genetically as a positive regulator of homeotic genes (Mazo et al., 1990). Mutations in SET of *trithorax* are embryonic lethal (Mortin et al., 1992) but the role of most of the individual building domains has not been established yet. The capacity of Trithorax to both self-associate and to specifically bind other proteins suggested that it may act within a complex or in separate, but interacting, complexes. The SET of the *Drosophila* TRX can bind the histones (Katsani et al., 2001), while the yeast SET1 can associate with a complex of seven proteins (COMPASS) important for cell growth (Milles et al., 2001). HRX binds INI (products of the human homologs of *trithorax* and *snf5*, respectively), providing a direct link between the Trx-G and the SWI/SNF (Rozenblatt-Rosen et al., 1998). Trithorax and Ash1, two members from the Trx-G, can interact through their SET domains (Rozovskaya et al., 1999, 2000). MLL, a mammalian homolog of Trithorax, can dynamically regulate a target *Hox* gene and counteracts the effect of *bmi-1*, a *Psc* homolog (Hanson et al., 1999). Phosphorylation has been suggested as a mechanism for the specific regulation of the HRX SET via myotubularin-related proteins (Cui et al., 1998).

A recently discovered novel signaling pathway involving the SET of the yeast SET1 opened yet another path to cell-function (Schramke et al., 2001). SET1 can act also as a dual-function regulator involved in repressive functions (Nislow et al., 1997). Some mammalian homologs of Trithorax can play repressive roles as well. Several types of translocation-induced leukemia result from a fusion of HRX with various partners (Sharples and DePinho, 1999; Van Lohuizen, 1999; Rozovskaya et al., 1999, 2000). The fact that many genes are upregulated in these mixed-lineage leukemia has suggested that the function of the native protein had been to repress them (Van Lohuizen, 1999).

The N-amino acid regions of the HRX and the *HRX* protein but is conserved in the molecules of the human homolog (Hanson et al., 1999). The HDAC complex via their CXXC-basic domains to deacetylate the targeted nucleosomes. The *Drosophila* and the human trithorax homologs provide an example of proteins belonging in the same
family (even in the same clade II) that still may play different, even antagonistic roles. It will be noted that the level of sequence similarity of their SET domains has classified them in two well-separated sister-groups. This fact suggests that in order to predict a function for an unknown gene (when based on analogy with a known gene) it has to belong into the same sister-group as a gene with a known function.

3. SET-domain factors and gene regulation in plants

The first plant SET-domain genes, CURLY LEAF (CLF) and MEDEA (MEA), were recognized as Arabidopsis homologs of E(z) (Goodrich et al., 1997; Grossniklaus et al., 1998; Kiyosue et al., 1999; Luo et al., 1999) (see Fig. 2). The product of CLF represses AGAMOUS, a flower homeotic gene and may act during flower development to promote fate determination (Goodrich et al., 1997). The similarities between CLF and E(z), both structural and functional, provided the first evidence for an involvement of a plant SET-domain gene in the repression of a developmental function (Goodrich et al., 1997). MEA is a maternally inherited factor involved in embryogenesis (Grossniklaus et al., 1998; Kiyosue et al., 1999; Luo et al., 1999). The parent-of-origin effects might be explained by different mechanisms but an important conclusion drawn from these studies was that, at a cellular level, mutation of the genes in Arabidopsis cause upregulation of replication and proliferation (Luo et al., 1999; Yadegari et al., 2000; Vielle-Calzada et al., 2000). This conclusion agrees well with recent data from animal systems suggesting a link between Pc-G repression and cell cycle regulation (Brock and van Lohuizen, 2001; Hanson et al., 1999).

MEA functions in cooperation with FIE, a homolog of Esc (Kiyosue et al., 1999; Luo et al., 1999; Ohad et al., 1999; Spillane et al., 2000; Yadegari et al., 2000). Thus, both the genetic and the direct interactions between MEA and FIE indicate that a functional complex, homologous to the Pc-Gi, has been preserved in the evolution of animals and plants. A partner interacting with CLF has not been discovered yet. The question of whether CLF is capable to bind an Esc homolog or whether it functions within a complex remains open.

The first examples of plant genes homologous to the animal trithorax genes have been reported recently in Arabidopsis (Alvarez-Venegas and Avramova, 2001a; Baumbusch et al., 2001). Mutant atrx1 Arabidopsis plants have impaired growth, flower and ovule defects, suggesting pleiotropic developmental functions for ATX-1 (Alvarez-Venegas et al., submitted). While the function of the Arabidopsis trithorax homologs has not been established yet, three important issues deserve attention. First, it is not known yet whether the plant TRITHORAX family members function within a complex of synergistically acting factors similar to the Trx-G complex of Drosophila. Second, the belonging of ATX-1 and ATX-2 in the TRITHORAX family (Fig. 3) does not automatically imply that they would have activating functions. Lastly, it is not known whether the plant trithorax homologs could counteract the repressive effect of any Pc-G gene. In Drosophila, the proteins of the Pc-G/Trx-G complexes are antagonistically involved in homeotic gene regulation. It is not known whether similar antagonistic complexes exist and function in plants. These are important issues that await further studies.

4. Evolution of individual architectural motifs

4.1. Distribution of conserved domains between unicellular and multicellular eukaryotes

The SET-domain is an abundant motif that has apparently assembled for eukaryotic functions (Stephens et al., 1998). The occurrence and the expansion of the fraction of the SET-domain genes is compared with the expansion of the genome sizes of a unicellular (S. cerevisiae) and three multicellular species for which all-genome data are currently available (Table 1). The ratio of the overall number of ORFs in the genomes of yeast to the ORFs in C. elegans, Drosophila and Arabidopsis thaliana is 1:2.7:2:4, respectively. However, with the transition to multicellularity, the SET-domain containing genes have proliferated significantly, expanding four-to eightfold. The fraction of the genes belonging to a particular SET-domain family is different in the yeast and in the multicellular species. Thus, in the genome of S. cerevisiae there are no genes belonging to the Su(var)3-9 and the E(z) families, while one gene belongs to the Tithorax family. This fact may suggest that functions

Table 1
Evolution of some individual architectural domains

<table>
<thead>
<tr>
<th>S. cerevisiae</th>
<th>C. elegans</th>
<th>D. melanogaster</th>
<th>A. thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ORFa</td>
<td>6269 (a)</td>
<td>~17,300 (b)</td>
<td>~13,600 (c)</td>
</tr>
<tr>
<td>SETb</td>
<td>8</td>
<td>37</td>
<td>42</td>
</tr>
<tr>
<td>Su(var)</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>E(z)</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TrxSET</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>DAST</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

a ORF numbers are as estimated in: (a) Cherry et al., 1998; (b) Reboul et al., 2001; (c) Gopal et al., 2001; (d) AGI, 2000.

b The distribution (number) of SET-domains is from SMART (SM0317).
exercised by the Trithorax-type SET-motif have occurred earlier than functions linked to the SET domains of the other two families. It has to be noted, though, that the majority of the SET-genes, from both yeast and higher eukaryotes, belongs to families outside the analyzed three (Table 1). With the exception of the ASH family, all cloned and studied genes are from the Su(var)3-9, E(z) and Trithorax families, providing our major source of current knowledge.

Within the multicellular species, the fraction of a particular type of SET-proteins varies as well. In general, it is proportional to the increase in overall genome size (Table 1). Notable is the preferential proliferation of the Arabidopsis genes belonging to the SU(VAR)3-9 family, as compared to the worm and the fly. The majority of the Arabidopsis Su(var)3-9 homologs (SUVH) are intronless and some of them might represent pseudogenes. The predominant fraction of the Arabidopsis SUVH genes, however, is transcribed (Baumbusch et al., 2001). These authors have suggested that the high number of these genes might be a result of two processes, extensive genome duplications and retrotransposition. Whether this tendency would be observed in other plants remains to be seen when complete genome information becomes available for other species.

Most SET-domain proteins have a modular structure composed of various architectural motifs. A combinatorial approach to gene assembly creates enormous possibilities for variation and precision required to meet the specific needs of a living system. Only two putative Arabidopsis proteins, SUVR3, 5, and clade I-proteins from the TRITHORAX family, do not carry architectural elements other than the SET domain. The evolutionary history of the individual motifs assembled within a protein may be different. Thus, the ANK motifs, the GTP-EFTU, the SANT domain, are present also in bacteria (Bork, 1993; Tschiersch et al., 1994; Aasland et al., 1996). The chromodomain is considered an ancient motif because a peptide with similar three-dimensional structure has been discovered in Archea (Ball et al., 1997). Therefore, the finding of these motifs in bacteria, yeast, animals and plants, indicates that they have occurred early in evolution and that they have roles common for the functioning of both prokaryotes and eukaryotes.

Most of the architectural elements (the AT-hooks, the PHD-fingers, the PWWP-, the Tudor-, and the bromo-domains) are not present in bacteria but are present in S. cerevisiae. Such motifs, therefore, may have arisen to meet the specific needs of eukaryotes.

With the appearance of multicellular organisms, ‘core’ functions have been conserved (reflected by the finding of similarly organized genes), while new genes have evolved for multicellular needs. Such genes are expected to be present in multicellular species and to be absent from the genomes of unicellulars (Chervitz et al., 1998; Aravind and Subramanian, 1999). Based on this hypothesis, it may be concluded that most proteins from the three SET-domain families have evolved for functions related to multicellularity. Clade I, of Trithorax is a notable exception.

A conserved peptide domain was recognized in the molecular structure of the Arabidopsis ATX1 and ATX2 (Alvarez-Venegas and Avramova, 2001) (Fig. 3). A PSI-BLAST search with a section of about 150 amino acids, between amino acids 365–514 of ATX1 run against the non-redundant section of the GenBank at the NCBI, has identified 30 proteins with a threshold E value of 0.001. Twenty-four of these belong to the Trithorax family and six to putative genes of unknown origin. This observation suggested the name of the motif: Domain Associated with SET in Trithorax (DAST). This domain contains two smaller phenylalanine-tyrosine rich (FYR) domains identified earlier in three Arabidopsis genes unrelated to trithorax (Balcuinas and Ronne, 2000). We suggested to refer to this domain as DAST because it is a signature feature for the proteins of the TRITHORAX family: all members, except those of clade I and the three Arabidopsis genes in clade IV contain DAST. It is split in the genes of clade II. The occurrence of DAST in the databases is summarized in Table 1. There are no DAST motifs in the genome of S. cerevisiae. DAST appears in the nematode as a single copy associated with SET, while in the fly and in Arabidopsis, DAST has taken up functions outside the Trithorax function. The arrangement of the DAST with respect to the other motifs in Arabidopsis is plant-specific (Fig. 3). A function for DAST is not known yet but the finding of DAST in animal and in plant proteins, while absent from the genomes of yeast and bacteria, suggests that this motif has evolved later and, possibly, in connection to functions required by multicellularity.

4.2. Evolution of regulatory mechanisms in animals and plants

The finding of plant SET-domain genes involved in developmental and homeotic gene regulation has interesting implications for the evolution of animals and plants. According to conventional theories, plants and animals have diverged from a unicellular ancestor (Baldauf and Palmer, 1993; Wainright et al., 1993; Wang et al., 1999). Separation from a unicellular ancestor would indicate that plants and animals have independently achieved multicellularity and the mechanisms regulating it. It may be expected, therefore, that different mechanisms (genes) would govern the balance between proliferation/differentiation, homeotic gene regulation and the control of development. In plants, organ development is not restricted to the embryonic stage and differentiation and organogenesis occur throughout the life span of the organism. This implies that plant and animals might have evolved differently their genes required for regulating these functions. However, evidence is beginning to accumulate that plants and animals might share common regulatory mechanisms. The homology between genes regulating stem cell maintenance in Arabidopsis and Drosophila (Moussian et al., 1998; Benfey, 1999) and the recognition of plant homeotic and developmental regulators as homologs of E(z), Esc and Trithorax, provided unexpected examples of
shared mechanisms. The finding of DAST in the genomes of both plants and animals, while absent in yeast, is compatible with a hypothesis that plants and animals might have had a multicellular ancestor (Doolittle et al., 1996).

5. Perspective

For an unknown gene, a function similar to an already established one might be expected when at least one of two criteria were met: the architecture of the unknown gene is the same as the architecture of the gene with known function; in cases when only one conserved element is considered, cladistic analysis of its sequence similarity should be expected when at least one of two needs of multicellular systems. The SET-domain genes of eukaryotes, while others have assembled for multicellular history: some have occurred in bacteria, some in unicellular yeasts, and one protein in the TRITHORAX family. The S. cerevisiae protein does not contain other elements besides SET in its molecules. A combinatorial approach to gene assembly creates enormous possibilities for variation and precision required to meet the specific needs of multicellular systems. The SET-domain genes of multicellular species may have evolved via three different pathways (Chervitz et al., 1998): (a) assembly of new motifs (DAST), (b) reshuffling of already existing motifs (the Arabidopsis genes from the TRITHORAX family) and (c) duplications of individual modules (PHD fingers, ANK repeats, Tudor, SET, SANT domains, etc.), followed by divergence and combinations with new motifs for achieving a finer precision of function. The phylogenetic relationship and cladistic patterns closely parallel the molecular architecture of the proteins. The individual elements building the members of the families may have different evolutionary history: some have occurred in bacteria, some in unicellular eukaryotes, while others have assembled for multicellular-specific functions. The homology-clustering pattern is congruent with the presence/absence of the accompanying motifs. Thus, the whole-protein molecule architecture supports the respective segregation into (sub)classes based on the degree of similarity of one peptide. A cladistic approach may provide a tool for predicting function.

Acknowledgements

We are indebted to J.-M. Deragon (Univ. Blaise Pascal) and to A. Mazo (T. Jefferson University) for helpful critical comments on an earlier version of the manuscript.

References


Dorn, R., Kraus, V., Reuter, G., Saumweber, H., 1993. The enhancer of...


Okano, M., Bell, D.W., Haber, D.A., Li, E., 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247–257.


