CONTRIBUTIONS to SCIENCE, **1** (3): 313-322 (2000)
Institut d'Estudis Catalans, Barcelona

Genome fluidity. The case of plants

Josep M. Casacuberta and Pere Puigdomènech*
Departament de Genètica Molecular, Institut de Biologia Molecular de Barcelona (CID-CSIC)

Abstract

The publication of complete sequences of large genomes is becoming routine. The analysis of these data provides a general outlook of genome structures, which helps us to understand how a genome is built up. A genome is obviously not a completely random polymer, but nor is it a fixed, optimised structure. Plant genomes seem to be particularly fluid, which allows large differences in size and organisation to occur in closely related species. Here we analyse some aspects of plant genome structure and comment on several mechanisms that contribute to genome varability, in particular the function of mobile elements that are abundant and active components in the plant genomes.

Resum

La publicació de la seqüència completa de grans genomes està esdevenint una rutina. L'anàlisi d'aquestes dades proporciona una visió general de les estructures dels genomes, la qual cosa ens ajuda a entendre com estan formats els genomes. Òbviament un genoma no és un polímer completament a l'atzar, però tampoc ens apareix com una estructura rígida i optimitzada. Els genomes de les plantes semblen especialment fluids i, per tant, hi ha diferències importants en la longuitud i l'organització, fins i tot entre espècies molt properes. En aquest article analitzarem alguns aspectes de l'estructura dels genomes de plantes i discutim macanismes que contribueixen a generar variabilidad en els genomes, en particular, la funció dels elements mòbils que són components abundants i actius en els genomes de les plantes.

Key words: Plant genome, mobile elements, polyploidy

Genomes and their paradoxes

A genome is a nucleotide polymer including stretches containing information that can be reproduced and expressed. The polymer is found inside all living cells and the information they contain determine the organization, physiology and fate of all organisms. However, it remains unclear as to what proportion of a genome is informative and what proportion is random, neutral or parasitic. It is certainly not true that a genome is a completely random polymer of the four nucleotides but it is also probably untrue that a genome is a fixed, perfectly optimised structure where each nucleotide has an exactly specified function. For the first time these questions can be studied at the level of the primary structure because long genomic stretches and complete genomes are becoming available.

*Author for correspondence: Pere Puigdomènech, Departament de Genètica Molecular, Institut de Biologia Molecular de Barcelona (CID-CSIC). Jordi Girona, 18. 08034 Barcelona, Catalonia (Spain). Tel. 34 934006129. Fax: 34 932405904. Email: pprgmp@cid.csic.es

The publication of complete sequences of large genomes is becoming routine. Beginning with viruses, then moving to mixoplasma, bacteria and yeast, we now have the first complete animal genome, that of the nematode *Caenorhabditis elegans*. The insect genome *Drosophila melanogaster* and the genome sequence of the plant *Arabidopsis thaliana* are finished and the genome of *Homo sapiens* will be soon available in databases. Probably rice, mouse, pine and a long etcetera will follow as well as comparative analyses of populations within a species. We are entering new era for Biology where researchers in the life sciences will have access to tools with unprecedented possibilities. The result should be that the function of an increasing number of genes will be discovered. Our knowledge of biological functions, the foundation of genetic variability and the analysis of diseases is entering a new period where the dream of biologists, the analysis of an organism as a whole, will become feasible.

Genome projects are producing large data-banks on genes and helping the analysis of gene functions. But these projects are also providing a general outlook of genome structures, allowing us to compare the arrangement of

genes and to undestand of how a genome is built up. The composition of intergenic regions, which is assumed to have no coding properties, is now being clarified. In fact, a correlation was found between genome size and biological complexity. If the size of a genome is plotted in relation to biological evolution it appears that genomes are larger in the most complex organisms. Yeast has more genes than bacteria, invertebrates more genes than yeasts and mammals more genes than insects or nematodes.

However, a couple of exceptions to this rule have been found. These exceptions are amphibians and plants. This is known as the C number paradox. In both cases a great heterogeneity in genome size is observed that does not correspond to any parameter that could be related to the complexity of the species. Moreover, the data on gene number that are becoming available for these species indicate that genome size is not related to the number of genes either. Plants are an extreme example of large genomic variability. When plant genomes are examined by mapping it is found that in some cases even the order of genes is conserved among evolutionary related species that may have genomes of very different sizes (see Table 1). This is the case of Gramineae,which show a high degree of conservation between species whose genome size differ by a factor of 10.

Table 1. DNA content in different plant species (in Mbp)

Arabidopsis thaliana	145
Prununs armeniaca	294
Citrus sinensis	367
Oryza sativa	419
Cucumis melo	454
Sorghum bicolor	748
Licopersicum esculentum	907
Zea mays	2.292
Pisum sativum	3.947
Hordeum vulgare	4.873
Allium cepa	15.290
Triticum aestivum (6x)	15.966
Fritillaria assyriaca	110.000

The sequencing of large portions of the genome is allowing us to understand the paradox. The data indicate that one of the main explanations for the different genome lengths is the presence of repetitions of DNA sequences. These repeated elements in most cases are mobile through the genome. In fact they are an important source of variability in plant genomes, which gives us a picture of which enables us to see genomes as highly fluid structures. Another element that creates large changes in genome size is polyploidy, especially that leads to in plants. The possible impact of these different levels of genome variability in plants will be the object of discussion in this contribution.

Polyploidies in plant genomes

Whole genome duplications have been supposed to be essential in large evolutionary steps, for instance in the origin of vertebrates (see [1] for a recent review). The existence of duplicate genes or even of full metabolic or developmental pathways allows new functions to be created as the essential ones are covered by one of the copies. For example, the nematode and *Drosophila,* which have a relativly small genome, also have a number of genes that is lower approximately by a factor of four than that found in vertebrates. Interestingly, while large genome duplications have been frequent in plant genome evolution, higher plants seem to have maintained a similar number of genes. The number of genes found in *Arabidopsis* (around 20000) is probably very similar to that in rice, which has a genome four times larger, and does not different greatly from that found in the nematode or in *Drosophila*. This number would appear to be that required for the construction of a higher organism. It would seem that the genome duplications that allowed the large evolutionary jumps in the animal kingdom to occur have not been produced in plants. It might also have to be considered that the majority of angiosperms have a relatively recent origin. Flowering plants appeared when dinosaurs were the dominant species on our planet. It would be of particular interest to compare the genetic changes that produce this essential jump in plant evolution.

More than 50% of all plant species are polyploid or have undergone periods of polyploidy in their evolutionary history [2]. Even some species that were supposedly diploids have been found to be ancient polyploids. An interesting example is maize, which is a natural alotetraploid. When analyzing gene coding for metabolic pathways, it was found that in maize they were often duplicated. This finding was confirmed when molecular markers became available. Many probes hybridized in two different chromosomes of maize. When syntheny, the condition by which cereal genomes can be alligned considering only a limited number of large chromosome rearrangements, was discovered, it became clear that maize had to be thought of as containing two genomes. Present hypotheses argue that maize was formed around 11 million years ago from two species, one of which was very similar to sorghum [3]. However, at the same time many single genes are found in maize and in the case of some multigene families, such as α-tubulins, they appear to have the same number of members in maize as in Arabidopsis, for instance [4]. Therefore, the number of genes in maize does not seem to have increased and in some cases the duplications have somehow been suppressed. We will be able to confirm this hypothesis when the maize genome is analyzed in detail and compared to other genome structures such as rice or Arabidopsis. Man has created new types of polyploids during the breeding process. Two good examples are wheat, which in its cultivated form is either a tetraploid or an hexaploid formed by two or three genomes from different species, and sugarcane, which is currently an artificial hexadecuploid.

But even in plants that are diploids, different levels of polyploidy can be found in different cell types. When the DNA content of different cell types from the organs of typical diploid plants is analyzed, normally a distribution of multi-

ples of diploids is found. In a normal higher plant cells containing diploid, tetraploid, octuploid cells or even cells of higher ploidy may coexist. In fact, the mechanism producing this phenomenon, endoreduplication, i.e. an uncoupling between DNA replication and cell division, is now under study. It is thought to be a gene regulation mechanism used in plants to produce differentiation in some cell types.

Why have plants not profited from these large genome duplications to make evolutionary jumps as happened for instance in the origin of vertebrates? It seems that polyploidy is such a trivial mechanism in plants that it cannot be used as an evolutionary opportunity. On the other hand, it can also be argued that plants may have a special ability to tolerate particularly large genomes, therefore there has not been a strong selective pressure against these duplications during evolution and, thus, these events have been maintained even in cases in which they have not been used to create new functions. On the contrary, perhaps the price to pay for having a large genome is too high for vertebrate genomes without a major benefi in terms of evolution, thus explaining that these duplication events are associated with qualitative important jumps in the evolution of those genomes.

Transposable elements: an introduction

The presence of genetic elements with the ability to transpose within the genome was first proposed by Barbara McClintock [5], in 1947, based on her work on chromosome breaks in maize, but the idea was not accepted by the scientific community until the molecular characterisation of the first transposable element (TE) from bacteria was reported [6]. Since then, TEs have been found in virtually all organisms where they constitute an important fraction of their genome, up to 10% of the *Drosophila melanogaster* genome and 35% of the human genome [7]. As the insertion of these elements, which can be as great as 15 Kb, is potentially a highly mutagenic event, their function in the genome has been, and still is, very controversial. McClintock's ideas of TEs being major actors of genomic fluidity and chromosomal rearrangements as a response facing situations of stress (see for example [8]) have been countered by those who think that the vast majority of TE insertions are probably deleterious and that TEs are merely parasitic or selfish [9,10]. In the last few years it has been shown that the mobilisation of some TEs can have beneficial effects on the genome and that a few TEs can play essential roles in some genomes. Nevertheless, it is hard to imagine a benefit gained from the transposition of most TEs than explains the success of these elements in colonising the genomes of virtually all organisms. However, what is underiable is the tremendous impact that the presence of these mobile elements has in genomes and the influence of TEs in genome evolution. In the following sections, we briefly comment on the mutagenic capacity of plant TEs; we also report some examples of possible roles displayed by TEs and evidence of their contribution to the evolution of plant genomes.

TEs and the evolution of plant genomes

Most transposable elements can be grouped into two classes based on their structure and mechanism of transposition. Class I elements, also known as retrotransposons, transpose via an RNA intermediate, while class II elements transpose via a DNA molecule. Most class I elements code for a reverse transcriptase enzyme that serves to synthesize a new DNA copy from the RNA intermediate while class II elements usually code for a transposase that catalyses the cleavage of the TE from its original position and the insertion into a new genome position. Nevertheless, non-autonomous class I and class II elements that do not code for the enzymes needed for transposition can be transactivated by fully active TE. Transposition of class I elements is replicative and the number of elements increases exponentially with transposition, while mobilisation of class II elements is usually conservative, maintaining the number of elements present in the genome. In the last few years a new class of TE sharing the structural characteristics of both class I and class II elements has been described. These elements, named MITES (Miniature Inverted-repeated Transposable Elements), transpose by an unknown mechanism and therefore remain unclassified (see Fig. 1 for a scheme).

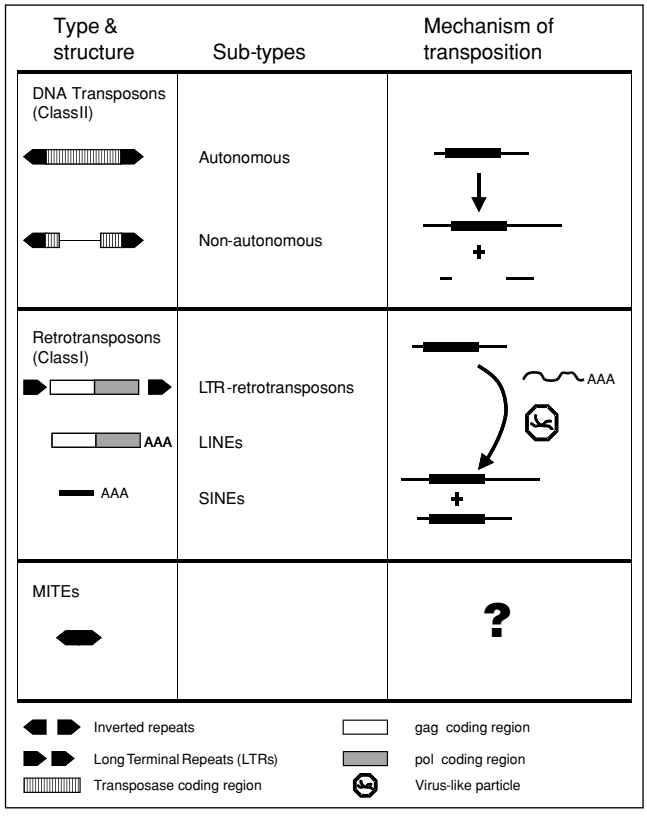

Figure 1. Scheme of the structure and mechanism of transposition of the principal classes of mobile elements.

Representatives of the different groups of class I and class II elements have been found in all eukaryote genomes, but retrotransposons and MITEs seem to have been especially successful in colonising plant genomes. As an example, the retrotransposon BARE-1 is present in more than

50000 copies in the genome of barley [11], as is the Tourist family of MITEs in the maize genome [12]. However, TEs have not proliferated at high levels in all plant genomes. While large genomes, such as maize (2000 Mb, approximately), contain a very high number of retrotransposons (up to 70-85% of the maize genome), small genomes like the one from *Arabidopsis thaliana* (120 Mb) have a very low content of TEs (4-7% of the genome). The enormous variability of plant genome size, even between closely related species, seems therefore to correlate with a high variability of copy number of the mobile elements contained.

The analysis of large stretches of genomic sequences begins to shed some light on how TEs have contributed to the evolution of plant genomes. The sequencing of 280 Kb of the *adh* locus in maize has shown that the intergenic region is constituted by nested arrays of retrotransposons accounting for more than 60% of the DNA [13]. These authors have recently shown that all these retrotransposons have been integrated within the past 3 million years [14], which suggests that the genome of maize could have doubled in size within this short period. How are so many elements incorporated so quickly, and why are they tolerated? As explained above, it is assumed that maize is an allotetraploid that was formed by the interspecific cross of two different parental genomes 11 million years ago [3], and it has been suggested that the resulting redundancy of genetic loci could have allowed a high transposition activity by buffering its deleterious effects [15]. The newly inserted transposons could, in turn, have provided new safe target sites for future transpositions, increasing exponentially the number of potential sites for integration with time, and explaining why retrotransposons are found in maize genome forming nested arrays of elements [14]. The success of TEs in colonising plant genomes could thus be related to the prevalence of polyploidy in plants. Indeed, as we have already mentioned, more than 50% of all plant species are polyploid or have undergone periods of polyploidy in their evolutionary history [2]. The differences in genome size between closely related plant species could thus simply be the result of a different TE amplification history.

However, while the amplification of TEs seems to be much more frequent than their elimination in plants, a recent report has shown that the loss of retrotransposons by recombination between their Long Terminal Repeats (LTRs) could be an important mechanism for reducing genome size. This could explain the genome size differences between closely related species of the *Hordeum* genus [16]. Plant genome size could therefore be the result of an equilibrium between forces that tend to genome obesity, such as the activity of TEs, and others that counter the former by elimination of repeated sequences.

Impact of retrotransposons in plant genome

Mobilisation of TEs is often a deleterious process. This is particularly true in the case of retrotransposons, as the insertion of these long elements is often irreversible. There are many examples of mutations caused by TEs in nature and, in fact, it is the ability to generate mutations that revealed the existence of mobile elements and has allowed scientists to characterise them. The analysis of insertional mutations in a single gene, the *Waxy* gene of maize, which gives rise to a viable and easily visualised phenotype, resulted in the first characterisation of a plant transposon, the maize *Ac/Ds* element [17], and has subsequently allowed the characterisation of TEs of different types, as retrotransposons [18] and MITEs [12]. On the other hand, many classical mutations selected from naturally occurring phenotypes by plant breeders on the basis of their improved agricultural quality are caused by TEs. This effect has been demonstrated for the brown midrib3 *(bm3*) mutation of maize, which was shown to increase the digestibility of forage maize. Molecular analysis of *bm3* varieties showed that the phenotype results from a retrotransposon insertion [19] or a deletion occurring in the structural gene coding for caffeic acid O-methyltransferase, a key enzyme in the lignin biosynthesis pathway. Furthermore, the high mutagenic capacity of TEs has been used with great succes as a tool to produce insertional mutants in all eukaryotic model systems. Manipulation of the structure of these elements has led to controlled mutagenic systems that greatly facilitate the cloning of the genes whose mutation causes a particular phenotype.

Nevertheless, the mobilisation of TEs could also have beneficial effects for the host genome. A paradigmatic case is that of retrotransposons HetA and TART of *Drosophila*. *Drosophila* chromosomes do not present the typical telomeric repeats at their end that serve to preserve genetic information to be lost during replication. Instead, Drosophila telomers are composed of tandem arrays of the HetA and TART retrotransposons [20]. The lack of telomerase activity in this species is thus compensated by the specific and repeated insertion of retrotransposons at the end of its chromosomes. The target site specificity of these elements makes them not only non-deleterious, but absolutely essential for the maintainance of the integrity of the host genome.

Plant chromosomes do have typical telomeric repeats, but a high number of retrotransposon copies have been found in another structural region of the chromosome: the centromer. Indeed, the centromers of the chromosomes of several grasses [21,22], as well as those of Arabidopsis [23], seem to contain tandem arrays of different types of retrotransposons. This centromeric location could be of special evolutive significance. It has been recently shown that in an interspecific mammalian hybrid, the atypically extended centromers of the chromosomes of one of the parents are composed primarily of a highly amplified retrotransposon, and it has been proposed that the repeated insertion of this element could help to differentiate the homologous chromosomes, facilitating the fertility of the hybrid [24]. As interspecific crosses are commonplace in plant genome evolution, the centromeric presence of retroelements in plants may indicate a similarly important role of plant retroelements in facilitating the preservation of interspecific hybrids. However, as most plant retrotransposons do not display a high target site specificity, being dis-

persed in the genome, a possible centromeric function does not seem to be the general case for these elements.

While a direct beneficial role for each TE sitting in the plant genome is difficult to imagine, it has been proposed that a more subtle beneficial effect on gene regulation could be the reason for its evolutionary preservation. Retrotransposons and MITEs are frequently found close to coding regions in most plant species [25], and it has been suggested that their insertion could have modified the transcriptional regulation of genes during evolution. Nevertheless, to date there is not a single clear example of a TE sequence playing the role of transcriptional promoter or terminator of a normal plant gene.

An active role for TEs in plant genome evolution was first proposed by McClintock some 40 years ago (see for example [8]). Her theory was that TEs worked to rearrange the genome under severe stress conditions to produce individuals that could better respond to that particular stress. The fact that many TEs, and particularly retrotransposons [26] are activated in stress conditions fits well with this idea. Nevertheless, it is hard to understand an immediate effect of transposition in generating fitter individuals. An interesting idea is that, perharps, the «positive» role of TEs in plant genome evolution does not necessarily imply a direct and immediate positive role on fitness. It may well be that a short-term deleterious effect of TEs on the genome could allow a long-term positive role. McFadden and Knowles [27] have proposed that the irreversible deleterious mutations caused by the insertion of TEs could be a singular way to escape to evolutionary stasis. Transposition could cause irreversible deleterious mutations forcing a species, which is no longer adapted to the media, to evolve rapidly. These rapid evolutionary events could thus be at the origin of speciation processes. In this sense, it is interesting to note that most TEs contain subfamilies that are genera or species-specific, suggesting that bursts of transposition are often associated with speciation processes (see for example [28,29]). A recent survey of the evolution of the *Emigrant* MITE from Arabidopsis [30] has shown that this element is actively transposing accompanying the recent spread of this plant species and the definition of the ecotypes, suggesting a role for this element in an ongoing speciation process (Casacuberta et al, unpublished).

It thus seems that, by means of a positive and direct action on plant genes, a more general role in genome structure or simply by causing irreversible deleterious mutations that promote escape from evolutionary stasis, TEs are major players in eukaryote genome evolution. The participation of TEs seems to be specially relevant in the case of plants, perhaps due to their ability to tolerate important variations of genome size.

Instability within plant protein repetitive gene sequences

While, in general, the non-coding fraction of the genome displays a higher degree of variability compared to that shown by coding regions, large degree of variability can also be observed within regions coding for special classes of plant proteins. This is the case of the genes coding for the highly repetitive proteins found in the storage compartments and the cell wall. The sequencing of these protein types was a difficult and almost impossible task before DNA techniques could be applied due to their repetitive nature and, in fact, all the information we have on these proteins comes from recombinant DNA studies. On the one hand, these are often insoluble proteins, while on the other hand, there are complex mixtures of polypeptides that are difficult to purify and analyze. In the last twenty years, our knowledge of plant protein sequences has expanded enormously and it is now possible to compare sequences and to attempt to draw conclusions on the formation of these sequences and their stability.

One example comes from the work of our group on maize storage proteins. The genes coding for storage proteins were among the first to be cloned due to their importance and to the high abundance of their mRNA at the moment of grain filling within the developing seed. The protein sequences encoded by these genes have a number of distinct features that correspond to the need for the appropriate deposition of the proteins in the dry seed and for their function as a source of amino acids in the germinating plantlet. For this reason these proteins probably have loose sequence requirements that include a tight folding, hydrophobic character and a highly biased amino acid composition. Many of the plant storage proteins have a repetitive sequence, one example of these proteins was the first sequence cloned in Catalonia, namely the cDNA coding for two storage protein from maize then called glutelins and now classified in the group of γ-zeins [31].

γ-zeins were cloned from a c-DNA library by screening of an antibody raised in rabbits against the protein purified from maize flour. The antibody reacted in western blots against two polypeptides of 28 and 16 kDa respectively in whole protein extracts of maize seeds. According to this result, the screening produced not only a protein sequence but two different protein sequences were identified among the cDNA clones isolated. Interestingly the two clones did not hybridize at DNA level. After sequencing the two cDNAs it was shown that they corresponded to proteins showing a high degree of similarity in a protein domain that was rich in proline and repetitive in nature but a large disimilarity in parts of their sequence. In 28kDa γ-zein seven repeats of a PPPVHL sequences exist while in 16 kDa γ-zein only three of these repeats are observed. It has been shown that the different domains of the protein have a function in the targeting of the protein within special compartments of the endosperm storage cells [32].

The comparison of the two protein sequences of the γ-zein proteins allowed interesting relations to be established between the two maize sequences [33]. From the comparison, four distinct types of sequence variability could be deduced: the first was the duplication of the sequence itself. The great similarity allowed a common origin for the two proteins to be proposed. Interestingly it was later shown that the

gene coding for the 28 kDa γ-zein is unstable, some maize varieties have two copies of the gene in tandem while others have only one copy. The second source of variablity arises from the duplication of proline-rich sequences, this mechanism will be discussed below. The third kind of variability in the sequence of the γ-zein protein family is the duplication of small sequences. This phenomenon is mainly observed in the 3'-end of the mRNA. In maize this kind of duplication is often attributed to the visit of transposons in the genes. Finally, the fourth source of sequence variablity observed is the well-known appearance of point mutations.

Repetitive proline-rich proteins in the plant cell wall

The mechanisms that cause variability in storage proteins have also been found when comparing many gene sequences analyzed since then in maize. A particularly interesting example was found in a highly repetitive sequence from the maize cell wall. Plant cell wall proteins constitute one of the most characteristic examples of plant protein sequences. They are supposed to have a structural function that in some cases can be related to reactions of the plants to pathogen attack through reinforcement of the cell wall. Most of the cell wall proteins are very repetitive in nature and generally either rich in proline or in glycine [34]. The most characteristic example of these proteins is the group of extensins described from dicotyledonous species. A number of genes coding for proteins of this type have been cloned in maize in our group. These are the genes coding for HRGP (Hydroxyproline-rich glycoprotein), the most abundant proteins to be extracted from the maize cell wall, the gene coding for HyPRP a gene expressed specifically in the embryo [35] and having a hybrid type of protein, proline-rich and hydrophobic and ZmPRP an extreme example of protein formed by homogeneous proline-rich repeats and which seems to take part in the formation of the secondary cell wall. The best closely studied of these genes is that coding for HRGP.

HRGP genes have been cloned from maize [36], teosinte [37], sorghum [38] and rice [39] (Caelles et al, 1992). The HRGP protein is present in all the cells of the plant while the mRNA is an excellent marker of tissues active in proliferation in the vegetative [40] and embryonic tissues [41] of maize. This information has allowed the gene sequences of both the coding and the adjacent regions to be compared. The results fit with those already described for γ-zein sequences with interesting additional observations. When the promoter, coding and 3' transcribed but non-translated segments of the gene are compared in terms of nucleotide substitutions the most variable region appears to be the coding region. However the kind of variability observed in the regions is not the same and in fact insertions or deletions appear to be more frequent than single nucleotide changes. This observation may be an indication of the large activity of mechanisms producing variability at least in the maize genome. In the 5' regions, duplications are observed that leave untouched a number of boxes that may correspond to elements of the promoter of the gene. The changes observed in the 3' region, including the single intron typical of these genes, are mostly due to small duplications. In the coding regions, as was observed in the case of γ-zeins, many changes occured but always as blocks of some of the repetitive units (see Fig. 2). In fact, the molecular weight of the protein in the different species analyzed varies considerably.

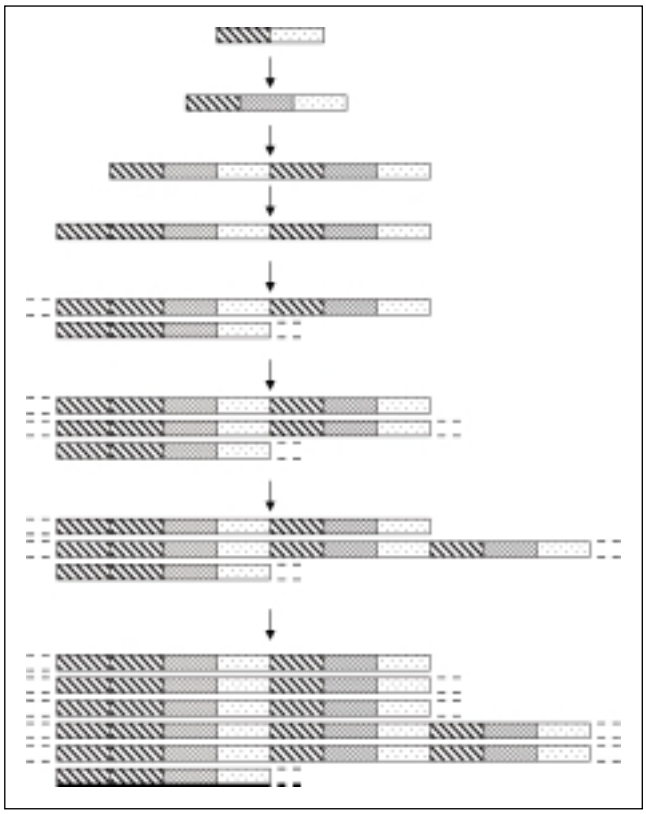

Figure 2. In the figure a pathway for the production of the maize HRGP protein is proposed. The initial protein contains the first elements that, by successive duplications, produce the protein sequence as it is now found in one of the maize varieties.

Another proline-rich protein, ZmPRP, offers an extraordinary example of repetitive sequence [42]. Its backbone is formed, besides a signal peptide, of an extremely repetitive sequence where the element PEPK is present in more than 80% of the polypeptide. The final protein is an alternating polypeptide of basic and acid amino acids that may result in a highly insoluble fiber. This protein has been localized in the secondary cell wall where it may take part in the formation of this highly impermeable and resistent structure.

The image that emerges from the analysis of maize repetitive proteins described to date is that different mechanisms acting on the variability of both the coding and the adjacent regions are active in of genes of this type. The results as a whole indicate that in a large genome such as that of maize these mechanisms, probably point mutation, transposon insertion and excision, homologous recombination and unequal crossing-over, are highly active. It is difficult to say

whether centuries of man is intervention with this species has accumulated mechanisms of high efficiency in the production of variability in this species. Comparions between large genomes, which will be possible in the next decade, should enable us to answer this question. In any case the presence of repetitive sequences is the consequence of such an action and the origin of phenomena such as chromosome reorganization. The fact that we observe in different parts of the genome the action of different mechanisms may be an effect of the distinct kinds of selective pressure acting on them. It is clear that a short duplication may be deletereous in the coding region while a transposon visit in the 3' region probably has no effect in the stability of the mRNA.

Plant introns

Many plant genes have introns in the same way as genes from other eukaryotes have. Their presence within genes gives rise to the same kind of questions regarding the origin and evolution of intervening sequences as the presence of these elements in animal systems do. However, plants introns have specific features. For instance, attempts to assay the processing of plant introns by animal cells or viceversa have always resulted in negative data indicating that the mechanisms of splicing in animal or plant cells are different. This finding may result from minor details in the processing machinery in the same way as transcription factors belong in general to the same families of proteins in plant and animal kingdoms but, as a rule, plant or animal promoters do not function in heterologous systems.

In general plant introns are shorter than animal introns. Examples of plant introns of more than 2 kb have been described but the average to date is between 200 and 300 bp. This may be due to the small number of genes analyzed at this level in plants and the fact that the most complex plant genome analyzed to date, that of *Arabidopsis thaliana*, is a very compact genome with small introns [43] . However, this rule seems to hold true for plants.

When comparing the position of plant introns in the genes that have homologues in animal species very little concordance is found. In large gene families with proteins having similar functions in eukaryotes, such as tubulins, it is not possible to uncover a general rule that indicates that an ancestral common species had all the introns, because the number of introns would be clearly excessive. In any case, putting together these data it is difficult to find any relation between the position of intervening sequences and any possible structural domain of the protein, as has been proposed for the origin of introns. The consequence of this observation is that introns may appear as well as disappear during the evolution of species, constituting another source of genetic variability.

The mobility of introns might be related to a class of abundant plant parasites, the viroids. Viroids are one of the shortest type of the active oligonucleotide molecule. In animal systems the best known example is the agent of hepatitis δ. Viroids are small double stranded circular RNA molecules and they have sequence features that assimilate themselves to processed introns. For this reason it seems that viroids are either parasitic introns, that they are parasite RNA molecules using the mechanisms of intron processing, or that introns are domesticated remarks of small RNA molecules that were active in the origins of our present cells. If introns are created and destroyed throughart the evolution of species they are also a mechanism of production of genetic variability that may have a use, for instance in the production of new protein sequences through differential splicing.

General and final considerations

The image emerging from the discipline of Genomics is that the genome of plants is not a fixed and immobile structure but a highly variable construction comparable to an ecosystem. In this «jungle», the various individual elements, including transposable elements, fight to survive and struggle to reproduce. The «fittest» occupy the largest area of the territory while genes seem to be overwhelmed by the action of these fertile cousins and may even be killed by them. Of course all comparisons have a limit and the struggle for life within a genome has to collaborate to produce the fittest possible individual in its own struggle for life in the world of living organisms. If this is not so the element itself disappears.

Genomes, therefore, are not just a linear succession of genes. Genes are surrounded by mobile elements, they are the subject of different mechanisms that continuously vary their sequence and they are also interrupted by introns. The origin of the different non-coding genome structures is still the subject of heated debated. It is clear that pathogenic elements such as retrovirus or viroids have similarities with retrotransposons and with introns, respectively. It is also true that probably in the case of mobile elements, and beyond doubt in the case of introns via differential splicing, they are used as gene regulation mechanisms. It might be that these elements are the remarks of ancient pre-genome molecules that have been converted into constituents of the genome where they continue to have their own independent cycles and where they establish either an equilibrium with the host genome or they become pathogens, managing to survive an equilibrium with the species. Alternatively, it might be that mobile elements are genes, collections of genes or pieces of genes that have escaped from the normal functioning of gene regulation and become autonomous at some stage, or even pathogenic.

In any case, the visit of the different types of mobile elements, together with point mutation and recombination, is a major cause of variability within genomes. Genes appear isolated in the middle of this tempestous ocean suffering from time to time the attack of these elements and as a result produce classical mutations. Plant genomes thus appear as an equilibrium that each species has reached between the different mechanisms that produce variability. Among these

mechanisms, at least in plants, the movement of transposons or the reproduction of retrotransposons are among the most important. However, to reach these equilibria the features of the cellular machinery of the plant controlling systems such as DNA replication, DNA repair, transcription or recombination that continuosly act on the maintenance and dynamics of the genome are also essential. The participation of all these factors in a complex manner may explain why there appears to be no rule for understanding the size of the genome of the different plant species.

In the same way as a species increases its chanches of survival by optimizing the use of its own pool of variability, the equilibrium that a genome reaches is the result of the previous state of the species genome. Sudden changes are essential for the definition of a new species and the accumulation of a large number of mutations and chromosome rearrangements may produce new gene properties that may be important for the definition and survival of a species. These large genome changes produced by mobile elements may also be a way of isolating a population so that it becomes a species. Genome duplication does not seem to be an important evolutionary driving force in plants and here it is proposed that this is why genome duplication is a normal mechanism of cell differentiation in plants.

Man is now one of the decisive forces in the evolution of species on our planet. He acts directly in the protection of endangered species, in the shuffling of species among different continents and the extinction of parasites. He also acts indirectly on the surface and atmosphere of the planet. Man has been acting indirectly on the genome of many species through the systematic breeding of domesticated plant and animal species. And he is beginning to act directly through transgenesis. Precise knowledge of the structure and dynamics of genomes is essential in this process. In particular it is essential in trying to predict how our species will influence both the equilibrium between species and finally our species itself. Directly or indirectly, this is our task (or our fate) and our responsibility in the near future.

Acknowledgements

The authors are indebted to CICYT (grant BIO97-0729) for financial support. This study has been carried out within the framework of the Centre de Referència de Biotecnologia de la Generalitat de Catalunya.

References

[1] Meyer, A., Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell. Biol.* 11, 699-704.

[2] Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264, 421-4.

[3] White, S., Doebley, J. (1998) Of genes and genomes and the origin of maize. *Trends Genet.* 14,327-32.

[4] Montoliu, L., Rigau, J., Puigdomènech, P. (1992). Analysis by PCR of the number of homologous genomic sequences to α-tubulin in maize *Plant Sci.* 84, 179-185.

[5] McClintock, B. (1947) Cytogenetic studies of maize and neurospora. Carnegie Inst. Washington Year Book 46, 146-152.

[6] Shapiro, J.A. (1969) Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*. *J. Mol. Biol.* 40, 93-105.

[7] Labrador, M., Corces, V.G. (1997) Transposable element-host interactions: regulation of insertion and excision. *Annu. Rev. Genet. 31*,381-404.

[8] McClintock, B. (1984) The significance of responses of the genome to challenge. *Science 22*6, 792-801.

[9] Doolitle, W.F., Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601-603.

[10] Orgel, L.E., Crick, F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604-607.

[11] Suoniemi, A., Anamthawat-Jonson, K., Arna, T., Schulman, A.H. (1996) Retrotransposon BARE-1 is a major dispersed component of the barley (Hordeum vulgare) genome. *Plant Mol. Biol.* 30, 1321-29.

[12] Bureau, T., Wessler, S. (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes *Plant Cell* 4, 1283-94.

[13] SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765-8.

[14] SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nature Genet.* 20, 43-5.

[15] Voytas, D.F., Naylor, G.J.P. (1998) Rapid flux in plant genomes. *Nature Genet.* 20, 6-7.

[16] Vicient, C., Suoniemi, A., Anamthawat-Jonson, K., Tanskanen, J., Beharav, A., Nevo, E., Schulman, A.H. (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. *Plant Cell* 11, 1769-84.

[17] Fedoroff, N., Wessler, S., Shure, M. (1983) Isolation of the transposable maize controlling elements Ac and Ds. *Cell* 35, 235-42.

[18] Varagona, M.J., Purugganan, M., Wessler, S. (1992) Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. *Plant Cell* 4, 811-20.[19] Vignols, F., Rigau, J., Torres, M.A., Capellades, M., Puigdomènech, P. (1995) The brown midrib3 (bm3) mutation in maize occurs in the gene encoding caffeic acid O-methyltransferase. *Plant Cell* 7, 407-16.

[20] Pardue, M.L., Danilevskaya, O.N., Lowenhaupt, K., Slot, F., Traverse, K.L. (1996) Drosophila telomeres: new views on chromosome evolution. *Trends Genet.* 12, 48-52.

[21] Miller, J.T., Dong, F., Jackson, S.A., Song, J., Jiang, J.

(1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics* 150, 1615-23.

[22] Presting, G.G., Malysheva, L., Fuchs, J., Schubert, I. (1998) A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J.* 16, 721-8.

[23] Pelissier, T., Tutois, S., Tourmente, S., Deragon, J.M., Picard, G. (1996) DNA regions flanking the major Arabidopsis thaliana satellite are principally enriched in Athila retroelement sequences. *Genetica* 97, 141-51.

[24] O'Neill, R.J., O'Neill, M.J., Graves, J.A. (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* 393, 68-72.

[25] Wessler, S., Bureau, T., White, S. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Devel.* 5, 814-21.

[26] Grandbastien, M.A. (1998). Activation of retrotransposons under stress conditions. *Trends Plant Sci.* 3, 181-187.

[27] McFadden, J., Knowles, G. (1997) Escape from evolutionary stasis by transposon-mediated deleterious mutations. *J. Theor. Biol.* 186, 441-447.

[28] Lenoir, A., Cournoyer, B., Warwick, S., Picard, G., Deragon, J.M. (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. *Mol. Biol. Evol.* 14, 934-41.

[29] Vernhettes, S., Grandbastien, M.A., Casacuberta, J.M. (1998) The evolutionary analysis of the Tnt1 retrotransposon in Nicotiana species reveals the high variability of its regulatory sequences. *Mol. Biol. Evol.* 15, 827-36.

[30] Casacuberta, E., Casacuberta, J.M., Puigdomènech, P., Monfort, A. (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of Arabidopsis thaliana: characterisation of the Emigrant family of elements. *Plant J.* 16, 79-85.

[31] Prat, S., Cortadas, J., Puigdomènech, P., Palau, J. (1985) Nucleic acid (cDNA) and amino acid sequences of the maize endosperm glutelin-2. *Nucleic Acid Res.* 13, 1493-1504.

[32] Torrent, M., Geli, M.I., Ruiz-Avila, L., Canals, J., Puigdomènech, P., Ludevid, M.D. (1994) Role of structural domains for maize gamma-zein retention in Xenopus oocytes. *Planta* 192, 512-518.

[33] Prat, S., Pérez-Grau, L., Puigdomènech, P. (1987) Multiple variability in the sequence of a family of maize endosperm proteins. *Gene* 52, 41-49.

[34] José, M., Puigdomènech, P. (1993) Structure and expression of genes coding for structural proteins of the plant cell wall. Tansley Review No. 55. *New Phytologist* 124, 259-282.

[35] José, M., Ruiz-Avila, L., Puigdomènech, P. (1992) A Maize embryo-specific Gene encodes a proline-rich and hydrophobic Protein. *Plant Cell* 4, 413-423

[36] Stiefel, V., Ruíz-Avila, L., Raz, R., Vallés, M.P., Gómez, J., Pagès, M., Martínez-Izquierdo, J.A., Ludevid, M.D., Langdale, J.A., Nelson, T., Puigdomènech, P. (1990) Expression of maize cell wall hydroxyproline-rich glycoprotein gene in early leaf and root vascular differentiation. Plant Cell 2, 785-793.

[37] Raz, R., José, M., Moya, A., Martínez-Izquierdo, J.A., Puigdomènech, P. (1992) Different mechanisms generating sequence variability are revealed in distinct regions of the hydroxyproline-rich glycoprotein gene from maize and related species. *Mol. Gen. Genet.* 233, 252-259.

[38] Raz, R., Crétin, C., Puigdomènech, P., Martínez-Izquierdo, J.A. (1991) The sequence of a hydroxyproline-rich glycoprotein gene from Sorghum vulgare. *Plant Mol. Biol.* 16, 365-367.

[39] Caelles, C., Delseny, M., Puigdomènech, P. (1992) The hydroxyproline-rich glycoprotein gene from *Oryza sativa*. *Plant Mol. Biol.* 18, 617-619.

[40] Ludevid, M.D., Ruíz-Avila, L., Vallés, M.P., Stiefel, V., Torrent, M., Torné, J.M., Puigdomènech, P. (1990) Expression of a cell wall protein genes in dividing and wounded tissues of *Zea mays*. *Planta* 180, 524-529.

[41] Ruiz-Avila, L., Burgess, S., Stiefel, V., Ludevid, M.D., Puigdomènech, P. (1992) Accumulation of cell wall hydroxyproline-rich glycoprotein gene mRNA is an early event in maize embryo cell differentiation. *Proc. Natl. Acad. Sci.* USA 89, 2414-2418.

[42] Vignols, F., José-Estanyol, M., Caparrós-Ruiz, D., Rigau, J., Puigdomènech, P. (1999) Involvement of a Maize Proline-Rich Protein in Secondary Cell Wall Formation as Deduced from its Specific mRNA Localization. *Plant Mol. Biol.* 39, 945-952.

[43] The European Union Arabidopsis Genome Sequencing Consortium (including Casacuberta, E., Monfort, A., Puigdomènech, P.), the Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402, 769-777.

About the authors

The authors belong to the Departament de Genètica Molecular, Institut de Biologia Molecular de Barcelona*, CSIC. The work in the Department is devoted in the main to Plant Molecular Biology. In fact it was in this Department that the first sequence ever cloned in Catalonia, a maize cDNA, was isolated and published in 1985. The group has followed the evolution in the field becoming interested in gene structure and expression, analysis in transgenic plants and more recently in the results of genomic research. In particular, they are interested in gene function and evolution.*

Josep M. Casacuberta is a Chemistry graduate (UAB) and Doctor in Sciences (UAB). He conducted postdoctoral research in the Laboratoire de Biologie Cellulaire, INRA, Versailles, France and was Professor of Plant Physiology at the University of Paris VII. He has been Científic Titular *del CSIC since 1998.*

Pere Puigdomènech is a Physics graduate (UB), Doctor in Sciences (USTL, Montpellier) and Doctor in Biology (UAB). He held postdoctoral posts in Portsmouth Polytechnic (UK) and Max-Planck-Institut für Molekulare Genetik (Berlin, Germany). He was Associate Professor at the Department of Biochemistry (UAB). He has been Professor d'investigació *del CSIC since 1990 and Director of the Institut de* Biologia Molecular de Barcelona *del CSIC.*