Genetic Definition and Sequence Analysis of *Arabidopsis* Centromeres

Gregory P. Copenhaver,[1] Kathryn Nickel,[1] Takashi Kuromori,[1] Maria-Ines Benito,[2] Samir Kaul,[2] Xiaoying Lin,[2] Michael Bevan,[3] George Murphy,[3] Barbara Harris,[4] Laurence D. Parnell,[5] W. Richard McCombie,[5] Robert A. Martienssen,[5] Marco Marra,[6] Daphne Preuss[1]*

High-precision genetic mapping was used to define the regions that contain centromere functions on each natural chromosome in *Arabidopsis thaliana*. These regions exhibited dramatic recombinational repression and contained complex DNA surrounding large arrays of 180–base pair repeats. Unexpectedly, the DNA within the centromeres was not merely structural but also encoded several expressed genes. The regions flanking the centromeres were densely populated by repetitive elements yet experienced normal levels of recombination. The genetically defined centromeres were well conserved among *Arabidopsis* ecotypes but displayed limited sequence homology between different chromosomes, excluding repetitive DNA. This investigation provides a platform for dissecting the role of individual sequences in centromeres in higher eukaryotes.

Centromeres mediate chromosome segregation during mitosis and meiosis by nucleating kinetochore formation, providing a target for spindle attachment, and maintaining sister chromatid cohesion. Although centromere function in lower eukaryotes requires defined DNA sequences, identifying similar essential elements in higher eukaryotes has been a long-standing challenge (*1, 2*). Different criteria have been used to define centromeres in higher organisms. Cytogeneticists and cell biologists have used specific DNA probes and distinct DNA binding proteins to characterize the primary chromosomal constriction, delimiting regions encompassing several megabases of DNA. In contrast, geneticists and evolutionary biologists have characterized centromere activity by monitoring reduced recombination and chromosome segregation. Here, we used a genetic method, tetrad analysis, to localize regions conferring centromere activity within natural *Arabidopsis thaliana* chromosomes. This technique reveals crossovers between genetic markers and centromeres, pinpointing regions on paired homologous chromosomes that always migrate to opposite poles during meiosis I (*3*). We used tetrad analysis to monitor marker assortment on each chromosome in >1000 meioses, identifying the boundaries of all five *Arabidopsis* centromeres.

In some multicellular organisms, artificial chromosome constructs (*4*) or chromosome fragments (*5*) can recapitulate centromere functions, which suggests that specific DNA sequences are necessary. However, complete DNA sequence information is not available for these constructs, which makes it difficult to discern the contributions of individual sequence elements and chromosome context. Alternatively, centromere function in multicellular organisms also may be determined by epigenetic factors, including DNA modification, secondary structure, association with specialized chromatin components, or differential timing of replication (*6*). Here we analyzed the virtually complete sequence of two *Arabidopsis* centromeres, providing an unparalleled view of centromere composition and enabling a comprehensive analysis of the sequence motifs, DNA modifications, and structural features that contribute to centromere function.

Physical Mapping of Centromeric Regions

Previously, DNA fingerprint and hybridization analysis of two bacterial artificial chromosome (BAC) libraries enabled the assembly of physical maps covering nearly all single-copy portions of the *Arabidopsis* genome (*7*). However, repetitive DNA near the *Arabidopsis* centromeres, including 180–base pair (bp) repeats, retroelements, and middle repetitive sequences (*8, 9*), complicated efforts to anchor contiguous centromeric BAC clones (contigs) to particular chromosomes. We used genetic mapping to unambiguously assign these unanchored contigs to specific centromeres (Fig. 1), scoring polymorphic markers in 48 plants with crossovers informative for the entire genome (*3*). In this manner, we connected several centromeric contigs to the physical maps of the chromosome arms and simultaneously generated a large set of DNA markers useful for defining centromere boundaries. For chromosomes II and IV, DNA sequence analysis confirmed the structure of these contigs (*10*).

Although this analysis substantially extends the understanding of the centromeric regions, gaps in the physical maps remain at each centromere. BAC clones near these gaps have end sequences corresponding to repetitive elements that likely constitute the bulk of the DNA between contigs, including 180-bp repeats, 5*S* ribosomal DNA (rDNA), or 160-bp repeats (Fig. 1). Fluorescence in situ hybridization has shown that these repetitive sequences are abundant components of *Arabidopsis* centromeres (*8*). Genetic mapping and pulsed-field gel electrophoresis indicate that many 180-bp repeats reside in long arrays that measure between 0.4 and 1.4 Mb in the centromeric regions (*11*); sequence analysis revealed additional interspersed copies near the gaps (*10*).

Genetic Mapping of Centromere Functions

To determine which portions of the centromeric regions participate in centromere function, we used tetrad analysis, monitoring centromere marker assortment through individual meioses (Fig. 1). In *Arabidopsis*, this is possible with *quartet1* (*qrt1*), a mutation that causes the four products of male meiosis to be released as a tetrad of pollen grains (*12*). We generated plants useful for tetrad analysis by crossing *qrt* strains from the Landsberg and Columbia ecotypes and pollinating Landsberg stigmas with individual pollen tetrads from F_1 plants (*3*). Crosses typically yielded three or four progeny plants per tetrad; we analyzed the assortment of DNA polymorphisms in the progeny from >1000 tetrads.

Monitoring the position of crossovers in this population identified chromosomal regions that could be separated by recombination from centromeres (tetratype) as well as regions that always cosegregated with centromeres (ditype) (*3*). Tetratype frequencies decrease to zero at the centromere; consequently, we defined centromere boundaries as the positions that exhibited small but detectable numbers of tetratype patterns. Scoring the segregation of centromere-linked markers in about 400 tetrads localized centromeres 1 to 5 (*CEN1–CEN5*) to regions on the physical

[1]University of Chicago, Department of Molecular Genetics and Cell Biology, 1103 East 57 Street, Chicago, IL 60637, USA. [2]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [3]John Innes Centre, Colney Lane, Norwich NR4 7UJ, UK. [4]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [5]Cold Spring Harbor Laboratories, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. [6]Washington University Genome Sequencing Center, 4444 Forest Park Boulevard, St. Louis, MO 63108, USA.

*To whom correspondence should be addressed. E-mail: dpreuss@midway.uchicago.edu

map corresponding to contigs of 550, 1445, 1600, 1790, and 1770 kb, respectively. Analysis of polymorphisms corresponding to 180-bp repeats (RCEN markers) (*11*) confirmed that these repeats map within the genetically defined centromeres (*13*). In genetic units, the centromere intervals averaged 0.44 centimorgan (cM) (percent recombination = 1/2 tetratype frequency), reflecting recombination rates at least 10 to 30 times below the genomic average of 221 kb/cM (*14*).

The low recombination frequencies typically observed near higher eukaryotic centromeres may be due to DNA modifications or unusual chromatin states (*15, 16*). We attempted to modify these states and thus improve centromere mapping resolution by raising recombination frequencies. F_1 Landsberg/Columbia plants were treated with one of a series of compounds known to cause DNA damage, modify chromatin structure, or alter DNA modifications (*17*). We crossed tetrads from treated plants with Landsberg stigmas and recovered and analyzed progeny from 8 to 107 tetrads subjected to each treatment, yielding >600 additional tetrads. These tetrads exhibited higher recombination in regions immediately flanking the centromeres (1.6% versus 3.4% recombination in untreated and treated plants, respectively), although the sample size was insufficient to determine whether any individual treatment had a profound effect. These efforts refined the map locations of centromeres on chromosomes 2 to 5 (Fig. 1), yielding intervals spanned by contigs of 880, 1150, 1260, and 1070 kb, respectively, with all tetrads consistently localizing centromere functions to the same region.

Efforts to increase recombination yielded a large number of tetrads with crossovers near the centromeres; these crossovers clustered within a narrow region at the centromere boundaries. Five crossovers occurred over a 70-kb region near *CEN2*, and seven occurred over a 200-kb region near *CEN1*, yet no crossovers were detected in the adjacent centromeric intervals of 880 and 550 kb, respectively (Fig. 1). Thus, the centromeres were found within large domains that restrict recombination machinery activity; the transition between these domains and surrounding, recombination-proficient DNA is remarkably abrupt (Fig. 2, A and K). Although analysis of more tetrads would yield additional recombination events, the observed distribution of crossovers suggests that centromere positions would not be significantly refined.

Centromeric DNA Content

The virtually complete and annotated sequence of chromosomes II and IV allows analysis of centromeres at the nucleotide level (*10*).

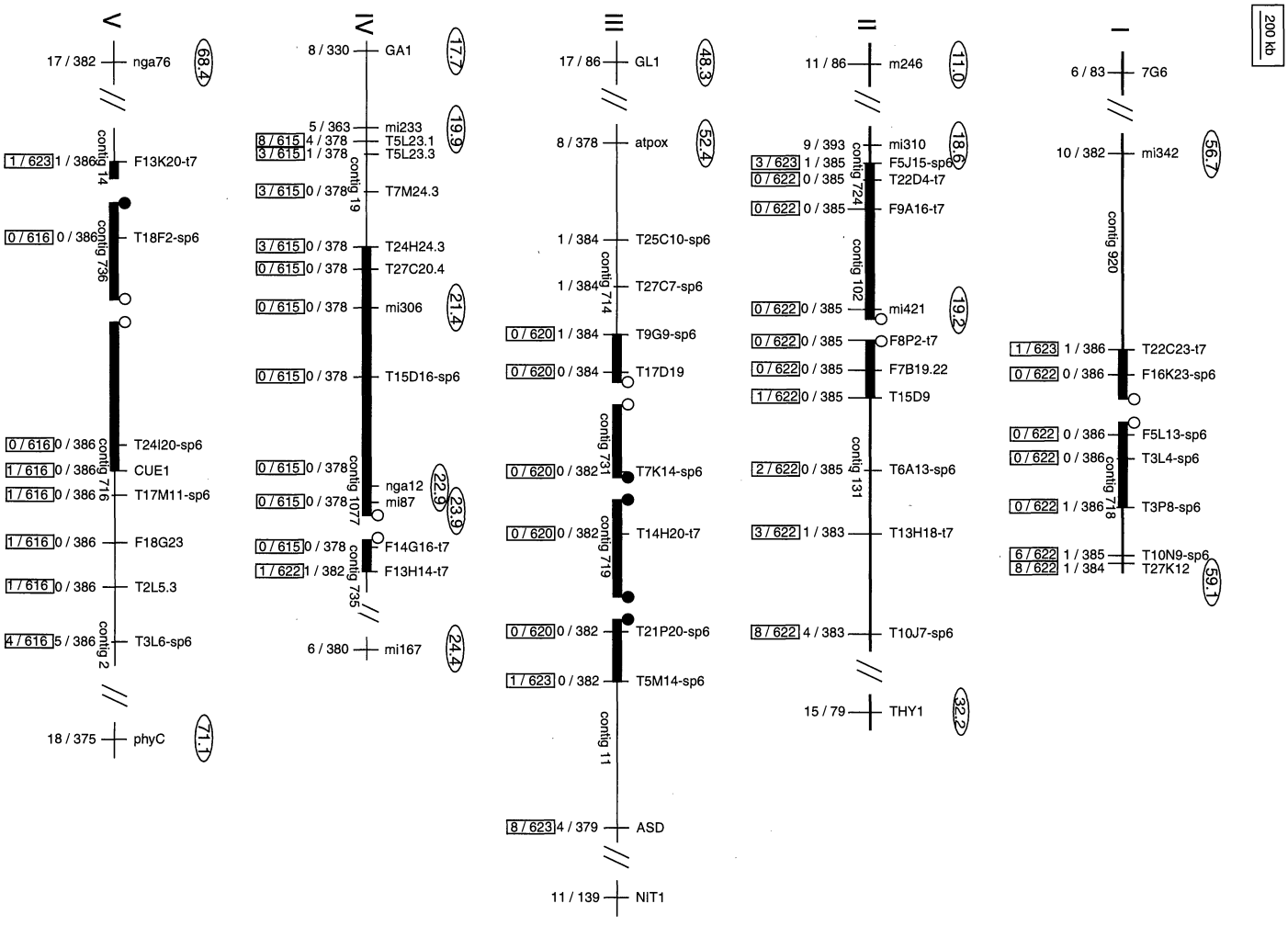

Fig. 1. Physical maps of the genetically defined *Arabidopsis* centromeres. Physical sizes were derived from DNA sequencing (chromosomes II and IV) or from estimates based on BAC fingerprinting (chromosomes I, III, and V) (*7*). Positions of markers used to confirm contig structure (above), the number of tetratype/total tetrads at those markers (below), the boundaries of the centromere (thick black bars), and the name of contigs derived from fingerprint analysis are indicated for each chromosome (*7*). For each contig, more than two genetic markers, developed from the database of BAC-end sequences (*27*), were scored. PCR primers corresponding to these sequences were used to identify size or restriction site polymorphisms in the Columbia and Landsberg ecotypes (*28*); primer sequences are available (*29*). Tetratype tetrads/total tetrads resulting from treatments that stimulate crossing over (boxes); positions of markers in centimorgans shared with the recombinant inbred (RI) map (ovals) (*14*); and sequences bordering gaps in the physical map that correspond to 180-bp repeats (open circles), 5*S* rDNA (black circles), or 160-bp repeats (gray circles) are indicated.

We examined the sequence composition within the genetically defined centromere boundaries and compared it with adjacent pericentromeric regions (Fig. 2). Analysis of two centromeres facilitates comparisons of sequence patterns and identification of conserved sequence elements. *CEN2* and *CEN4* are particularly well suited for this analysis; they reside on structurally similar chromosomes with 3.5-Mb rDNA arrays on their distal tips, regions measuring 3 and 2 Mb, respectively, between the rDNA and centromeres, and 16- and 13-Mb regions on their long arms (*10, 18*).

Repeat Abundance in Centromeric Regions

All higher eukaryotes examined to date contain repetitive DNA in their centromeric regions. If this DNA is sufficient for centromere functions, its abundance outside a genetically defined centromere is likely to be

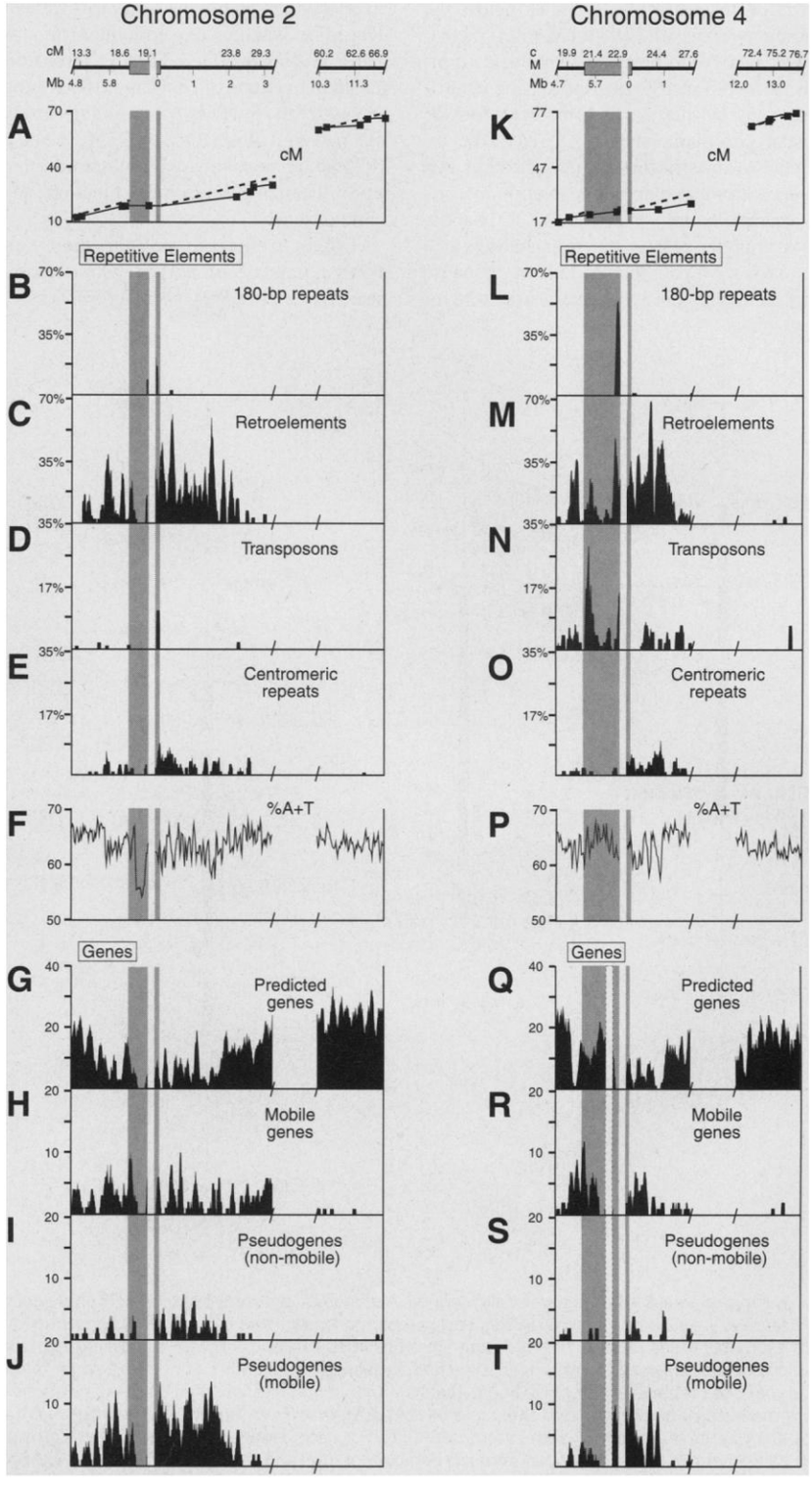

Fig. 2. Properties of the chromosome II and IV centromeric regions. (Top) Drawing of genetically defined centromeres (gray shading: *CEN2*, left; *CEN4*, right), adjacent pericentromeric DNA, and a distal segment of each chromosome, scaled in megabases as determined by DNA sequencing (gaps in gray shading correspond to gaps in the physical maps) (*10*). Physical distances (megabases) starting at the telomeres of the short arm and also at the centromeric gaps are shown, as are centimorgan positions (RI map). (Bottom) The density of each feature is plotted relative to chromosomal position (megabases). (**A** and **K**) Centimorgan positions of markers on the RI map (solid squares) compared with the genomic average of 1 cM/221 kb (dashed line). One crossover within *CEN4* occurred in the RI mapping population (*14*), perhaps reflecting a difference between male meiotic recombination (this study) and recombination in female meiosis. (**B** to **E** and **L** to **O**) Percent DNA occupied by repetitive elements in a 100-kb sliding window at 10-kb intervals: (B and L) 180-bp repeats; (C and M) sequences with similarity to retroelements, including del, Ta1, Ta11, copia, Athila, LINE, Ty3, TSCL, 106B (Athila-like), Tat1, LTRs, and Cinful (*10*); (D and N) sequences with similarity to transposons, including Tag1, En/Spm, Ac/Ds, Tam1 MuDR, Limpet, MITES, and mariner (*10*); (E and O) previously described centromeric repeats including 163A, 164A, 164B, 278A, 11B7RE, mi167, pAT27, 160- and 500-bp repeats, and telomeric sequences (*8, 9*). (**F** and **P**) Percent A+T in a 50-kb sliding window at 25-kb intervals. (**G** to **J** and **Q** to **T**) Number of predicted genes or pseudogenes in a 100-kb sliding window at 10-kb intervals. Predicted genes (G and Q) and pseudogenes (I and S) typically not found on mobile DNA elements; predicted genes (H and R) and pseudogenes (J and T) often carried on mobile DNA, including reverse transcriptase, transposase, and polyproteins (*10*). Annotation was obtained from GenBank records, from the AGAD database, and by BLAST comparisons with the database of repetitive *Arabidopsis* sequences (*24, 30*); annotation in progress is noted (dashed lines) (*10*). Although updates to annotation records may change individual entries, the overall structure of the region will not be significantly altered.

dramatically reduced, thus preventing the formation of multiple centromeres along the chromosome arm. The *Arabidopsis* 180-bp repeat sequences have such properties. They were found in the gaps of each centromeric contig (Figs. 1 and 2, B and L), with few repeats and no long arrays elsewhere in the genome (*10, 11*). DNA hybridization experiments have shown that other repeats, including retrotransposons, middle repetitive elements, and telomeric sequences map near *Arabidopsis* centromeres (*8, 9*). The annotated sequence of chromosomes II and IV identified regions with homology to these repeats (*10*), within both the functional centromeres and the adjacent regions (Fig. 2, B to E and L to O).

Sequences resembling retrotransposons are rare on *Arabidopsis* chromosome arms, yet these elements are abundant in centromeric regions (*10*). In a 4.3-Mb sequenced region that includes *CEN2* and a 2.7-Mb sequenced region that includes *CEN4*, retrotransposon homology accounted for >10% of the DNA sequence, with maxima of 62% and 70%, respectively (Fig. 2, C and M). Sequences with similarity to transposons or middle repetitive elements occupied a similar zone but were less common (11% and 29% maximum density for chromosomes II and IV, respectively) (Fig. 2, D, E, N, and O). Finally, low-complexity DNA, including microsatellites, homopolymer tracts, and A+T–rich isochores are enriched in *Drosophila* and *Neurospora* centromeres (*19*) but not within *Arabidopsis* centromeres. Near *CEN2*, simple repeat sequence densities were comparable to those on chromosome arms, occupying 1.5% of the sequence within the centromere and 3.2% in the flanking regions and ranging from 20 to 319 bp in length (71-bp average). Except for an insertion of mitochondrial DNA at *CEN2* (*10*) the DNA in and around the centromeres did not contain any large regions that deviate significantly from the genomic average of 64% A+T (Fig. 2, F and P) (*20*).

Repetitive Elements and Centromere Functions

Unlike the 180-bp repeats, all other repetitive elements near *CEN2* and *CEN4* were less abundant within the genetically defined centromeres than in the flanking regions. The high concentration of repetitive elements outside the functional centromere domain indicates that they are insufficient for centromere activity. Thus, identifying segments of the *Arabidopsis* genome enriched in these repetitive sequences does not pinpoint regions that provide centromere function; a similar situation may occur in other higher eukaryotic genomes.

Repetitive DNA flanking the centromeres may play an important role, forming an altered chromatin conformation that nucleates or stabilizes centromere structure. Alternatively, other mechanisms could result in the accumulation of repetitive elements near centromeres. Although evolutionary models predict that repetitive DNA accumulates in regions of low recombination (*16*), many *Arabidopsis* repetitive elements are more abundant in the recombinationally active pericentromeric regions than in the centromeres themselves. Instead, retroelements and other transposons may insert preferentially into regions flanking centromeres or may be eliminated from the rest of the genome at a higher rate.

Abundance of Genes in Centromeric Regions

Expressed genes are located within 1 kb of essential centromere sequences in *Saccharomyces cerevisiae*, and multiple copies of tRNA genes reside within an 80-kb fragment necessary for centromere function in *Schizosaccharomyces pombe* (*21*). In contrast, genes are thought to be relatively rare in the centromeres of higher eukaryotes, although there are notable exceptions. The *Drosophila light*, *concertina*, *responder*, and *rolled* loci all map to the centromeric region of chromosome 2, and translocations that remove *light* from its native heterochromatic context inhibit gene expression (*22*). In contrast, many *Drosophila* and human euchromatic genes become inactive when they are inserted near a centromere (*22*). Thus, genes that reside near centromeres likely have special control elements that allow expression. The sequences of *Arabidopsis CEN2* and *CEN4* provide a powerful resource for understanding how gene density and expression correlate with centromere position and associated chromatin.

Annotation of chromosomes II and IV identified many genes within and adjacent to *CEN2* and *CEN4* (*10*) (Figs. 2 and 3). The abundance of mobile elements resulted in a relatively high frequency of reverse transcriptase, retroviral polyprotein, and transposase genes compared with chromosome arms (Fig. 2, H and R). However, other genes typically not associated with transposable elements were also predicted in the centromeric regions (Fig. 2, G and Q). The density of predicted genes on *Arabidopsis* chromosome arms averaged 25 per 100 kb (*20*), and in the repeat-rich regions flanking *CEN2* and *CEN4* this decreased to 9 and 7 genes per 100 kb, respectively. Many predicted genes also were found within the recombination-deficient, genetically defined centromeres. Within *CEN2*,

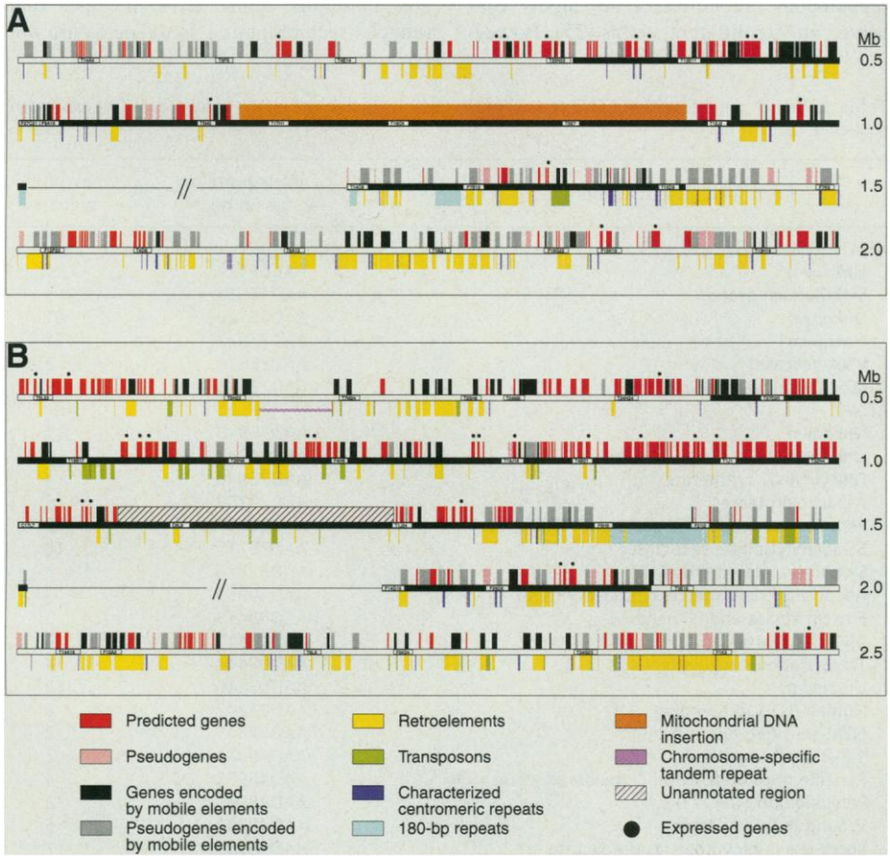

Fig. 3. Sequence features at *CEN2* (**A**) and *CEN4* (**B**). Central bars depict annotated genomic sequence of indicated BAC clones (*10*); black, genetically defined centromeres; white, regions flanking the centromeres; //, gaps in physical maps. Sequences corresponding to genes and repetitive features, filled boxes (above and below the bars, respectively), are defined as in Fig. 2.

there were 5 predicted genes per 100 kb; *CEN4* was strikingly different, with 12 genes per 100 kb.

There is strong evidence that several predicted centromeric genes are transcribed. The phosphoenolpyruvate phosphate translocator gene (*CUE1*) defines one *CEN5* border; mutations in this gene cause defects in light-regulated gene expression (*23*). Within the sequenced portions of *CEN2* and *CEN4*, 17% (27/160) of the predicted genes share >95% identity with cloned cDNAs (expressed sequence tags), with threefold more matches in *CEN4* than in *CEN2* (Table 1) (*10, 24*). Twenty-four of these genes have multiple exons, and four correspond to single-copy genes with known functions (Table 1). To investigate whether the remaining 23 genes are uniquely encoded at the centromere, we queried the database of annotated genomic *Arabidopsis* sequence. With two exceptions, no homologs with >95% identity were found elsewhere in the 80% of the genome that has been sequenced (Table 1). The number of independent cDNA clones corresponding to a single-copy gene provides an estimate of gene expression levels. On chromosome II, predicted genes highly similar to entries in the cDNA database (>95% identity) matched an average of four independently derived cDNA clones (range, 1 to 78). Within *CEN2* and *CEN4*, 11 of 27 genes exceeded this average (Table 1). Finally, most genes encoded at *CEN2* and *CEN4* were not related to each other, nor did they correspond to genes predicted to play a role in centromere functions; instead they have diverse roles (Table 1).

Many genes in the *Arabidopsis* centromeric regions appear to be nonfunctional because of early stop codons or disrupted open reading frames, whereas few pseudogenes reside on the chromosome arms (*10*). Although many of these pseudogenes exhibit similarity to mobile elements, some correspond to genes that typically are not mobile (Fig. 2, I, J, S, and T). Within the genetically defined centromeres there were 1.0 (*CEN2*) and 0.7 (*CEN4*) nonmobile pseudogenes per 100 kb; the repeat-rich regions bordering the centromeres had 1.5 and 0.9 genes per 100 kb, respectively. The distributions of pseudogenes and transposable elements are overlapping, which suggests that DNA insertions in these regions contributed to gene disruptions.

Model for Centromere Expansion

Arabidopsis centromeres contain a central region of 180-bp repeats surrounded by moderately repetitive DNA with dramatically reduced recombination. Flanking this genetically defined centromere are regions with normal recombination levels that are highly enriched in mobile elements. The abundance of repetitive sequences suggests that insertion events have contributed to substantial structural change of the centromeric regions over evolutionary time. Modern *Arabidopsis* centromeres potentially evolved from smaller domains composed of unique or low-copy sequences, and the accumulation of insertions presumably generated the larger, more repetitive regions observed today. In this view, integration of mobile elements could separate domains important for centromere function.

There may be constraints that limit centromere growth. For example, essential centromeric domains may require an arrangement suitable for assembly into higher order structures. In addition, mechanisms that inhibit crossing over in centromeric regions are likely required to prevent unequal sister chromatid exchanges that cause imbalances in critical DNA elements (*16*). Reduced recombination might be achieved by delaying replication or pairing of centromeric DNA or by limiting access of the recombination machinery through specialized chromatin structures. Thus, centromere structure is likely shaped by the action of mobile DNA element insertions balanced by selective pressures that maintain centromere function.

Conservation of Centromeric DNA

Saccharomyces cerevisiae centromeres on homologous chromosomes are highly conserved among different strains (*25*). However, the abundance of mobile DNA elements at *Arabidopsis* centromeres could contribute to a high sequence divergence among ecotypes. To investigate the conservation of *CEN2* and *CEN4*, we designed polymerase chain reaction (PCR) primer pairs corresponding to unique regions in the Columbia sequence and surveyed the centromeric regions of Landsberg and Columbia at about 20-kb intervals (Fig. 4). We obtained amplification products of the same length in both ecotypes for most primer pairs (80%), which indicates that the amplified regions are highly similar (Fig. 4), and 5% of the products revealed a size polymorphism. In the remaining cases, primer pairs amplified Columbia but not Landsberg DNA, even at low stringencies. In these regions, additional primers were designed to determine the extent of nonhomology. In addition to a large insertion of mitochondrial DNA in *CEN2* (*10*), we identified two other nonconserved regions (Fig. 4). Because this DNA is absent from the Landsberg centromeres, it is unlikely to be required for centromere function; consequently, the relevant portion of the centromeric sequence is reduced to 577 kb (*CEN2*) and 1250 kb (*CEN4*). Extensive sequence conservation between Landsberg and Columbia centromeres indicates that reduced recombination frequencies are not

Table 1. Predicted genes within *CEN2* and *CEN4* that correspond to the cDNA database. EST = expressed sequence tag.

Putative function	GenBank protein accession no.	No. of EST matches*
CEN2		
Unknown	AAC69124	1
SH3 domain protein	AAD15528	5
Unknown	AAD15529	1
Unknown†	AAD37022	1
RNA helicase‡	AAC26676	2
40*S* ribosomal protein S16	AAD22696	9
CEN4		
Unknown	AAD36948	1
Unknown	AAD36947	4
Leucyl tRNA synthetase	AAD36946	4
Aspartic protease	AAD29758	6
Peroxisomal membrane protein (PMP22)§	AAD29759	5
5′-adenylylsulfate reductase§	AAD29775	14
Symbiosis-related protein	AAD29776	3
Adenosine triphosphate synthase γ chain 1 (APC1)§	AAD48955	3
Protein kinase and EF hand	AAD03453	3
ABC transporter	AAD03441	1
Transcriptional regulator	AAD03444	14
Unknown	AAD03446	12
Human PCF11p homolog	AAD03447	6
NEM-sensitive fusion protein	AAD17345	2
1,3-β-glucan synthase	AAD48971	2
Pyridine nucleotide-disulfide oxidoreductase	AAD48975	4
Polyubiquitin (UBQ11)§	AAD48980	72
Wound-induced protein	AAD48981	6
Short-chain dehydrogenase/reductase	AAD48959	7
SL15†	AAD48939	2
WD40-repeat protein	AAD48948	2

*Independent cDNAs with >95% identity. †Related gene present in noncentromeric DNA. ‡Potentially associated with a mobile DNA element. §Characterized gene (*32*).

the result of large regions of nonhomology but instead are a property of the centromeres themselves.

Sequence Similarity Between *CEN2* and *CEN4*

Discerning the rules that govern centromere function will likely require analysis of both primary DNA sequence and higher order structures. As a first step, we searched for previously unidentified sequence motifs shared between *CEN2* and *CEN4*, excluding retroelements, transposons, characterized centromeric repeats, and coding sequences resembling mobile genes (*10*). After masking additional repetitive sequences, including homopolymer tracts and microsatellites, we compared contigs of 417 kb (*CEN2*) and 851 kb (*CEN4*) with BLAST (*25*).

This comparison showed that the complex DNA within the centromere regions is not highly homologous; only 16 DNA segments in *CEN2* matched 11 regions in *CEN4* with >60% identity (Fig. 5). These homologous sequences comprise a total of 17 kb (4%) of *CEN2*, have an average length of 1017 bp, and have an A+T content of 65%. On the basis of their similarity, we sorted the matching sequences into groups including two families that contain eight sequences each (AtCCS1 and AtCCS2), three sequences from a small family (AtCCS3), and four sequences found once within each centromere (AtCCS4-AtCCS7), one of which (AtCCS6) corresponds to predicted CEN2 and CEN4 proteins with similarity throughout their exons and introns (Fig. 5). Searches of the *Arabidopsis* genomic sequence database demonstrated that AtCCS1 to AtCCS5 are moderately repetitive sequences that appear in centromeric and pericentromeric regions; the remaining sequences are present only in the genetically defined centromeres.

Similar comparisons of all 16 *S. cerevisiae* centromeres defined a consensus consisting of a conserved 8-bp CDEI motif, an A+T–rich 85-bp CDEII element, and a 26-bp CDEIII region with seven highly conserved nucleotides (*25*). In contrast, surveys of the three *S. pombe* centromeres revealed conservation of overall centromere structure but no universally conserved motifs (*2*). Additional tests will be required to determine whether the conserved sequences we identified in the *Arabidopsis* centromeres contribute to function.

Conclusions

By combining genetic analysis with investigation of DNA sequence, we have defined chromosomal regions with specific properties. They confer meiotic centromere activity and are dramatically deficient in recombination. Structurally, they are composed of moderately repetitive DNA and a core of 180-bp repeats embedded in a highly repetitive pericentromeric region. Further analysis of these regions will yield insights into the role of specific binding proteins, assembly of unique chromatin structures, and altered patterns of DNA modification, replication, and pairing. Moreover, these studies provide a platform for identification of the minimal sequence that provides centromere function. Such sequences might be spread across the entire genetically defined region, be concentrated at a discrete point, or exist as redundant copies within the centromere (*26*).

The centromeres of other multicellular eukaryotes, like those of *Arabidopsis*, may harbor numerous expressed genes that specify important functions. Investigating how genes are maintained in recombination-deficient, repeat-rich regions will improve the understanding of genome evolution. Obtaining the sequences of centromeres from a diverse array of organisms will elucidate the general mechanisms that govern centromere function.

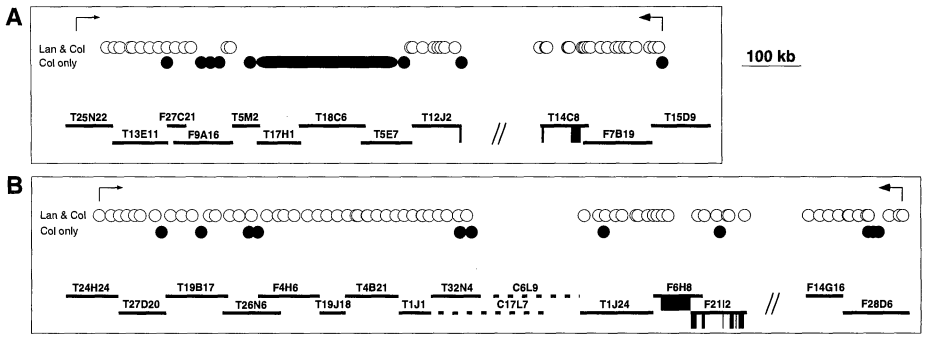

Fig. 4. Conservation of centromere DNA. BAC clones (bars) sequenced in *CEN2* (**A**) and *CEN4* (**B**) are indicated; arrows denote boundaries of the genetically defined centromeres. PCR primer pairs yielding products from only Columbia (filled circles) or from both Landsberg and Columbia (open circles); BACs encoding DNA with homology to the mitochondrial genome (gray bars); 180-bp repeats (gray boxes); unannotated DNA (dashed lines); and gaps in the physical map (double slashes) are shown.

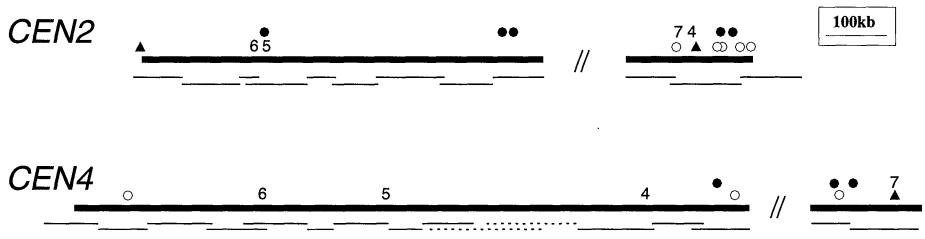

Fig. 5. Sequences common to *CEN2* and *CEN4*. Genetically defined centromeres (bold lines), sequenced BAC clones (thin lines), and unannotated BAC clones (dashed lines) are displayed as in Fig. 4. Repeats AtCCS1 (*A. thaliana* centromere conserved sequence) and AtCCS2 (closed and open circles, respectively); AtCCS3 (triangles); and AtCCS4-7 (4–7) are indicated (GenBank accession numbers AF204874 to AF204880) and were identified with BLAST 2.0 (*31*).

References and Notes

1. J. M. Craig, W. C. Earnshaw, P. Vagnrelli, *Exp. Cell Res.* **246**, 249 (1999).
2. L. Clarke, *Curr. Opin. Genet. Dev.* **8**, 212 (1998).
3. G. P. Copenhaver, W. E. Browne, D. Preuss, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 247 (1998).
4. J. J. Harrington *et al.*, *Nature Genet.* **15**, 345 (1997); M. Ikeno *et al.*, *Nature Biotechnol.* **16**, 431 (1998).
5. T. D. Murphy and G. H. Karpen, *Cell* **82**, 599 (1995); E. Kaszas and J. A. Birchler, *EMBO J.* **15**, 5246 (1996).
6. G. H. Karpen and R. C. Allshire, *Trends Genet.* **13**, 489 (1997).
7. M. Marra *et al.*, *Nature Genet.* **22**, 265 (1999); T. Mozo *et al.*, *Nature Genet.* **22**, 271 (1999).
8. M. Murata, J. S. Heslop-Harrison, F. Motoyoshi, *Plant J.* **12**, 31 (1997); J. S. Heslop-Harrison, M. Murata, Y. Ogura, T. Schwarzacher, F. Motoyoshi, *Plant Cell* **11**, 31 (1999); A. Brandes, H. Thompson, C. Dean, J. S. Heslop-Harrison, *Chrom. Res.* **5**, 238 (1997).
9. D. A. Wright *et al.*, *Genetics* **142**, 569 (1996); A. Konieczny, D. F. Voytas, M. P. Cummings, F. M. Ausubel, *Genetics* **127**, 801 (1991); M.-L. Chye, K.-Y. Cheung, J. Xu, *Plant Mol. Biol.* **35**, 893 (1997); Y.-F. Tsay, M. J. Frank, T. Page, C. Dean, N. M. Crawford, *Science* **260**, 342 (1993); E. J. Richards, H. M. Goodman, F. M. Ausubel, *Nucleic Acids Res.* **19**, 3351 (1991); C. R. Simoens, J. Gielen, M. Van Mantagu, D. Inze, *Nucleic Acids Res.* **16**, 6753 (1988); T. Pelissier, S. Tutois, S. Tourmente, J. M. Deragon, G. Picard, *Genetica* **97**, 141 (1996).
10. X. Lin *et al.*, *Nature*, in press; R. Wambutt *et al.*, *Nature*, in press.
11. E. K. Round, S. K. Flowers, E. J. Richards, *Genome Res.* **7**, 1053 (1997).
12. D. Preuss, S. Y. Rhee, R. W. Davis, *Science* **264**, 1458 (1994).
13. Polymorphisms associated with the 180-bp repeats were analyzed by pulsed-field gel electrophoresis as described in (*11*). Segregation of these polymorphisms in tetrads with informative crossovers confirmed complete linkage of a 180-bp repeat array at each centromere.
14. C. Somerville and S. Somerville, *Science* **285**, 380 (1999); http://nasc.nott.ac.uk/new_ri_map.html (August 1999).
15. K. H. A. Choo, *Genome Res.* **8**, 81 (1998); J. Puechberty, *Genomics* **56**, 247 (1999); M. M. Mahtani and H. F. Willard, *Genome Res.* **8**, 100 (1998).
16. B. Charlesworth, C. H. Langley, W. Stephan, *Genetics* **112**, 947 (1986); B. Charlesworth, P. Sniegowski, W. Stephan, *Nature* **371**, 215 (1994).

17. F_1 Landsberg *qrt1*/Columbia *qrt1* plants were grown under 24-hour light in 1-inch (2.54-cm) square pots and treated with methanesulfonic acid ethyl ester (0.05%), 5-aza-2′-deoxycytidine (25 or 100 mg/liter), Zeocin (1 μg/ml), methanesulfonic acid methyl ester (75 parts per million), *cis*-diamminedichloroplatinum (20 μg/ml), mitomycin C (10 mg/liter), *N*-nitroso-*N*-ethylurea (100 μM), *n*-butyric acid (20 μM), trichostatin A (10 μM), or 3-methoxybenzamide (2 mM). Plants were watered and flower-bearing stems were immersed in these solutions. Alternatively, plants were exposed to 350-nm ultraviolet light (7 or 10 s) or heat shock (38° or 42°C for 2 hours). Pollen tetrads from these plants were used to pollinate Landsberg stigmas 3 to 5 days after each treatment; the F_1 plants were then subjected to additional treatments (up to five times per plant, every 3 to 5 days).
18. G. P. Copenhaver and C. S. Pikaard, *Plant J.* **9**, 259 (1996).
19. X. Sun, J. Wahlstrom, G. H. Karpen, *Cell* **91** 1007 (1997); E. B. Cambareri, R. Aisner, J. Carbon, *Mol. Cell. Biol.* **18**, 5465 (1998).
20. M. Bevan, I. Bancroft, H.-W. Mewes, R. Martienssen, R. McCombie, *Bioessays* **21**, 110 (1999).
21. R. M. Kuhn, L. Clarke, J. Carbon, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1306 (1991).
22. G. H. Karpen, *Curr. Opin. Genet. Dev.* **4**, 281 (1994); A. R. Lohe and A. J. Hilliker, *Curr. Opin. Genet. Dev.* **5**, 746 (1995).
23. H.-M. Li, K. Culligan, R. A. Dixon, J. Chory, *Plant Cell* **7**, 1599 (1995).
24. http://www.tigr.org/tdb/at/agad/.
25. U. Fleig, J. D. Beinhauer, J. H. Hegemann, *Nucleic Acids. Res.* **23**, 922 (1995).
26. G. P. Copenhaver and D. Preuss, *Curr. Opin. Plant Biol.* **2**, 104 (1999).
27. http://www.tigr.org/tdb/at/abe/bac_end_search.html.
28. C. J. Bell and J. R. Ecker, *Genomics* **19**, 137 (1994); A. Konieczny and F. M. Ausubel, *Plant J.* **4**, 403 (1993).
29. http://genome-www.stanford.edu/Arabidopsis/aboutcaps.html.
30. http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html; http://nucleus.cshl.org/protarab/AtRepBase.htm.
31. http://blast.wustl.edu.
32. B. Tugal, M. Pool, A. Baker, *Plant Physiol.* **120**, 309 (1999); J. F. Gutierrez-Marcos, M. A. Roberts, E. I. Campell, J. L. Wray, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13377 (1996); N. Inohara *et al.*, *J. Biol. Chem.* **266**, 7333 (1991); J. Callis, T. Carpenter, C-W. Sun, R. D. Vierstra, *Genetics* **139**, 921 (1995).
33. Supported in part by grants from the National Science Foundation, the U.S. Department of Agriculture, the Consortium for Plant Biotechnology Research, and the David and Lucile Packard Foundation. M.-I.B., S.K., X.L., M.B., G.M., B.H., L.D.P., W.R.M., R.A.M., and M.M. are members of the *Arabidopsis* Genome Initiative. We thank L. Mets, R. Esposito, S. Rounsley, J. A. Mayfield, and K. C. Keith for helpful discussions; R. Stein, D. Jurcin, P. Ridley, and E. Bent for technical assistance; and M. Spielman and S. Streatfeild for sharing genetic markers.

23 September; accepted 15 November 1999

Collisional Breakup in a Quantum System of Three Charged Particles

T. N. Rescigno,[1] M. Baertschy,[2] W. A. Isaacs,[3] C. W. McCurdy[2,3]

Since the invention of quantum mechanics, even the simplest example of the collisional breakup of a system of charged particles, $e^- + H \rightarrow H^+ + e^- + e^-$ (where e^- is an electron and H is hydrogen), has resisted solution and is now one of the last unsolved fundamental problems in atomic physics. A complete solution requires calculation of the energies and directions for a final state in which all three particles are moving away from each other. Even with supercomputers, the correct mathematical description of this state has proved difficult to apply. A framework for solving ionization problems in many areas of chemistry and physics is finally provided by a mathematical transformation of the Schrödinger equation that makes the final state tractable, providing the key to a numerical solution of this problem that reveals its full dynamics.

Electron-impact ionization of atoms and molecules is one of the most basic phenomena in low-energy collision physics. It is the fundamental mechanism for ion formation in mass spectroscopy and is responsible for forming and sustaining low-temperature plasmas that are used in applications ranging from fluorescent lighting to the processing of silicon chips. These collisions are governed by none of the selection rules that limit optical excitation, primarily because the incident electron cannot be distinguished from those of the target. Thus, electron impact stands as one of the most efficient means for exciting and ionizing atoms and molecules.

It seems almost incredible that even the simplest example of an electron impact–initiated breakup problem, the ionization of a hydrogen atom in a collision with an electron, has resisted solution until now. Although the Schrödinger equation has been known for more than 70 years, there has been no framework that allowed its complete solution for this case. In contrast, the bound states of the helium atom, another system with only two electrons, were computed accurately in the 1950s. That work established a framework that allowed the development of modern quantum chemistry as a practical discipline. The theoretical framework demonstrated here provides a basis for developing practical methods to treat ionizing collisions of electrons with atoms and molecules.

[1]Lawrence Livermore National Laboratory, Physics Directorate, Livermore, CA 94551, USA. [2]Department of Applied Science, University of California, Davis, Livermore, CA 94550, USA. [3]Lawrence Berkeley National Laboratory, Computing Sciences, Berkeley, CA 94720, USA.

The Quantum Mechanics of Three Charged Bodies

Although the analytic solution of the wave function for the isolated hydrogen atom played a pivotal role in establishing the new quantum theory during the early part of this century, no corresponding solutions exist for systems with three or more charged particles. Indeed, the nonrelativistic quantum mechanics of two-electron atoms has a long history, beginning with the work of Hylleraas (*1*) on bound states in the 1930s that culminated with Pekeris's (*2*) accurate determination of the bound states of helium in the late 1950s.

Scattering problems are intrinsically more difficult. It was not until 1961 that the simplest collision problem in a two-electron system, scattering of an electron by a hydrogen atom without energy exchange, was solved numerically by Schwartz (*3*) with Kohn's variational principle (*4*). Since then, the effort to solve the problem of collisions in which energy is transferred into excitation of states with quantum numbers n and l [$e^- + H(1s) \rightarrow e^- + H(nl)$] has produced benchmark calculations of excitation probabilities and angular distributions for excited bound states of the hydrogen atom. In the case in which only probabilities for excitation of the target atom are required, the traditional approach has been to expand the unknown solution of the Schrödinger equation in terms of the known wave functions of the target—the so-called "close-coupling" method. The initial applications of this method were confined to low energies at which only a few target states could be excited (*5*).

The next major hurdle to overcome was the extension of such studies to collision energies above that needed to ionize the target where a continuously infinite number of final states is possible. The convergence of the close-coupling method was convincingly and dramatically illustrated by Bray and Stelbovics (*6*) in 1993, who showed that a "convergent" close-coupling method could be developed for calculating elastic and excitation probabilities. They replaced the true ionized states of the hydrogen atom with a finite set of "pseudostates" and systematically increased their number until convergence was achieved. Using these ideas, they performed the first accurate computations of the total probability for ionization. Their work completed another chapter on the dynamics of two-electron systems, but not the final chapter. Attempts to use this approach to predict

References and Notes

[3] **Assaying Genome-Wide Recombination and Centromere Functions with Arabidopsis Tetrads**

Gregory P. Copenhaver; William E. Browne; Daphne Preuss

Proceedings of the National Academy of Sciences of the United States of America, Vol. 95, No. 1. (Jan. 6, 1998), pp. 247-252.

Stable URL:

http://links.jstor.org/sici?sici=0027-8424%2819980106%2995%3A1%3C247%3AAGRACF%3E2.0.CO%3B2-U

[9] **Identification of a Mobile Endogenous Transposon in Arabidopsis thaliana**

Yi-Fang Tsay; Mary J. Frank; Tania Page; Caroline Dean; Nigel M. Crawford

Science, New Series, Vol. 260, No. 5106. (Apr. 16, 1993), pp. 342-344.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819930416%293%3A260%3A5106%3C342%3AIOAMET%3E2.0.CO%3B2-Y

[12] **Tetrad Analysis Possible in Arabidopsis with Mutation of the QUARTET (QRT) Genes**

Daphne Preuss; Seung Y. Rhee; Ronald W. Davis

Science, New Series, Vol. 264, No. 5164. (Jun. 3, 1994), pp. 1458-1460.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819940603%293%3A264%3A5164%3C1458%3ATAPIAW%3E2.0.CO%3B2-F

[14] **Plant Functional Genomics**

Chris Somerville; Shauna Somerville

Science, New Series, Vol. 285, No. 5426. (Jul. 16, 1999), pp. 380-383.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819990716%293%3A285%3A5426%3C380%3APFG%3E2.0.CO%3B2-S

LINKED CITATIONS

[21] **Clustered tRNA Genes in Schizosaccharomyces pombe Centromeric DNA Sequence Repeats**

Robert M. Kuhn; Louise Clarke; John Carbon

Proceedings of the National Academy of Sciences of the United States of America, Vol. 88, No. 4. (Feb. 15, 1991), pp. 1306-1310.

Stable URL:

http://links.jstor.org/sici?sici=0027-8424%2819910215%2988%3A4%3C1306%3ACTGISP%3E2.0.CO%3B2-9

[32] **Three Members of a Novel Small Gene-Family from Arabidopsis thaliana Able to Complement Functionally an Escherichia coli Mutant Defective in PAPS Reductase Activity Encode Proteins with a Thioredoxin-Like domain and ``APS Reductase'' Activity**

Jose F. Gutierrez-Marcos; Michael A. Roberts; Edward I. Campbell; John L. Wray

Proceedings of the National Academy of Sciences of the United States of America, Vol. 93, No. 23. (Nov. 12, 1996), pp. 13377-13382.

Stable URL:

http://links.jstor.org/sici?sici=0027-8424%2819961112%2993%3A23%3C13377%3ATMOANS%3E2.0.CO%3B2-B