Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4

Nancy Terryn[a], Leo Heijnen[b], Annick De Keyser[a], Martien Van Asseldonck[1,b], Rebecca De Clercq[a], Henk Verbakel[b], Jan Gielen[a], Marc Zabeau[b], Raimundo Villarroel[a], Taco Jesse[b], Pia Neyt[a], René Hogers[b], Hilde Van Den Daele[a], Wilson Ardiles[a], Christine Schueller[c], Klaus Mayer[c], Patrice Déhais[a], Stephane Rombauts[a], Marc Van Montagu[a], Pierre Rouzé[a,d,*], Pieter Vos[b]

[a] *Laboratorium voor Genetica, Departement Genetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium*

[b] *Keygene N.V., Postbus 216, AgroBusiness Park 90, Keyenbergseweg 6, NL-6700 AE Wageningen, The Netherlands*

[c] *Munich Information Center for Protein Sequence, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany*

[d] *Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Universiteit Gent, B-9000 Ghent, Belgium*

Received 9 January 1999

Abstract As part of the European Scientists Sequencing *Arabidopsis* program, a contiguous region (396 607 bp) located on chromosome 4 around the *APETALA2* gene was sequenced. Analysis of the sequence and comparison to public databases predicts 103 genes in this area, which represents a gene density of one gene per 3.85 kb. Almost half of the genes show no significant homology to known database entries. In addition, the first 45 kb of the contig, which covers 11 genes, is similar to a region on chromosome 2, as far as coding sequences are concerned. This observation indicates that ancient duplications of large pieces of DNA have occurred in *Arabidopsis*.

Key words: *Arabidopsis thaliana*; *APETALA2*; Genome; Sequencing

1. Introduction

In plant molecular biology and genetics, *Arabidopsis thaliana* has long been recognized as a model organism [1]. By the end of 1993, a project designated European Scientists Sequencing *Arabidopsis* (ESSA) was initiated with the aim of sequencing large fragments of the *A. thaliana* genome. Currently, a whole international team is working on the completion of that genome [2].

Within the ESSA program, most effort was concentrated initially on chromosome 4 around the *FCA* locus [3] and the *APETALA2* (*AP2*) locus. The *ap2* mutant had been described, mapped, and cloned before [4–6]. On the most recent map of Dean and Lister [7], *AP2* was found between the markers m214 and g2486 at 93.2 cM (the whole chromosome 4 being 116 cM).

Results of the sequencing and analysis of this latter region, as the result of cooperation between two laboratories, will be discussed here. This region represents the second largest contig of *Arabidopsis* being analyzed in detail, preceded by the 2 Mb *FCA* contig [3].

2. Materials and methods

2.1. Isolation and subcloning of an AP2-containing yeast artificial chromosome (YAC)

The CIC YAC clone 7A10, size 420 kb (from the CIC *A. thaliana* (L.) Heynh. ecotype Columbia library) was isolated by using AFLP (filed by Keygene N.V.) markers [8,9] derived from sequences of the *AP2* gene and subcloned in the cosmid vector pCLD04541 [10], using a partial *Sau*3AI digest. To isolate YAC-specific cosmids, a total number of 16 000 *Escherichia coli* clones were hybridized to gel-purified YAC DNA. Approximately 450 YAC-specific clones were identified. AFLP fingerprinting [9] was used to build a cosmid contig. This approach resulted in a cosmid contig of approximately 260 kb.

2.2. Isolation of bacterial artifical chromosome (BAC) clones extending the cosmid contig

The BAC clones were isolated through hybridization of the BAC library with a probe containing the inserts of all cosmids of the 260-kb contig, followed by contig building by AFLP. All AFLP markers present in the cosmid contig were also present in the BAC contig, indirectly proving colinearity of the cosmid contig with the genomic sequence. TAMU 8H13 and TAMU 10C14, which overlapped with the cosmid contig, were selected for further sequencing (Fig. 1).

2.3. Construction of cosmid and BAC subclones

DNA from cosmids and BAC clones was sheared either by sonication (Misonix Inc., Farmingdale, NY; type XL2020) or by nebulizing (Lifecare Hospital Supplies, Harborough, UK). A 1.8–2.2-kb end-repaired fraction was isolated and ligated in pUC18/*Sma*I/BAP (Pharmacia, Uppsala, Sweden) or a modified pUC19 vector (*Bam*HI/*Sal*I fragment replaced by a *Stu*I-, *Spe*I-, and *Sal*I-containing fragment). Individual colonies of the transformation were grown in 96- or 384-well microtiter plates.

*Corresponding author. Fax: (32) (9) 264 5349.
E-mail: pirou@gengenp.rug.ac.be

[1] Present address: University Hospital Nijmegen, Department of Human Genetics, Geert Grootteplein 24, NL-6500 HB Nijmegen, The Netherlands.

Abbreviations: AFLP, amplified fragment length polymorphism (is a trademark in the Benelux); BAC, bacterial artificial chromosome; EST, expressed sequence tag; YAC, yeast artificial chromosome

The genomic sequence of the contig has been submitted to the GenBank under accession numbers Z99707 and Z99708 (with 8361-bp overlap). Cognate cDNAs have accession numbers AJ002596, AJ002597, and AJ002598.

PII: S0014-5793(99)00097-6

2.4. Sequencing strategy

The sequence of the partially overlapping cosmids 3A6, 4B6, 4E12, 5C9, 2H2, 5E2, 2F3, 2H7, and BAC TAMU 10C14 was determined in a non-random approach by sequencing the cosmid or BAC ends, as well as a few random subclones. Primers were designed to isolate primer-specific clones from pools of subclones. A random sequencing approach was taken for cosmids 2C7, 3A6, 4F4, 3F11, 6B5, 5E3, 4G12, 4E3, and BAC TAMU 13H8. The ABI PRISM dye terminator cycle sequencing ready reaction kit was used mostly (Perkin-Elmer, Foster City, CA). The reaction products were analyzed on an ABI Prism 377. The sequence with the corresponding electropherograms were assembled into contigs using a home-made computer program called Sequence Assembly Facility Environment (SAFE; [11]) or the 1994 version of the Staden sequence analysis program [12]. The mean redundancy of the assembled sequence is 5.

2.5. Sequence comparisons

Sequence analysis was done initially by the Martinsried Institute for Protein Sequences (MIPS) center (Martinsried, Germany) and refined by the bioinformatics team of the Laboratory of Genetics (Ghent, Belgium). To this end, the sequence was cut into pieces of 7000 bp with 1000-bp overlaps. Each piece was submitted to a BLASTX/non-redundant protein and BLASTN/non-redundant DNA search as well as a BLASTN/EST search (mostly on the Beauty BCM server). From the resulting files, the homologues with the highest score and a reliable annotation were looked for.

When a homologue was found with a high score ($P(N) < E-100$) and with homology over its whole length, its protein sequence was followed to check whether the transcript did not show any frameshifts or stop codons and whether the intron borders were correct. When any doubt arose, NetGene2 predictions were used [13,14].

For genes with weak homologies, different prediction programs were used according to the specific problems: NetStart when the start codon was not obvious, GeneMark [15] together with GenScan [16] for exon predictions, GeneMark for exon frame predictions, GenScan to check that small exons were not missed, and NetGene2 for intron border prediction. When an expressed sequence tag (EST) was found through BLASTN/EST (most of the time for the 5′ or 3′ parts of the genes), the EST was used to complement the information obtained by BLASTX.

When no homologues were found through BLASTX, and no ESTs were available, we relied upon prediction programs, namely GenScan and GeneMark together to locate potential exons, NetGene2 for the intron borders, and NetStart [17] for the position of the start codon. In all cases, the coding sequence (CDS) was reconstructed, translated, and submitted to BLASTP, for a last homology search and check for gaps.

Updated FASTA searches done at MIPS on the genes in this contig can be found at http://speedy.mips.biochem.mpg.de/arbi/data/ap2_contig.html. Full annotations as well as various genomic features can also be found at the Ghent site (http://spider.rug.ac.be/public/seq/ap-2.html).

3. Results and discussion

3.1. General overview of the contig

Fig. 1 gives an overview of the clones used for the sequencing program. Initially cosmid clones were used, but during the project BAC clones became available and these were used to continue the contig sequencing (see Section 2).

The 103 predicted genes that are present in this region are summarized in Tables 1 and 2 and in Fig. 2. Table 1 provides information on their putative function, highest related entry in the public databases as well as EST sequences that correspond to the putative genes whereas in Table 2 the genes are classified based on their homology.

In conclusion, the putative role of approximately half of the genes could be established by sequence similarity to known genes. These genes have been classified into 15 classes according to their putative cellular role that will be used to describe genes identified in the genome program [3]. This list is preliminary and new categories and subcategories will be added as more of the genome is sequenced (updated version available at http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html).

3.2. Annotation and re-annotation process

After the first MIPS automatic annotation, manual re-annotation showed discrepancies on about four genes out of

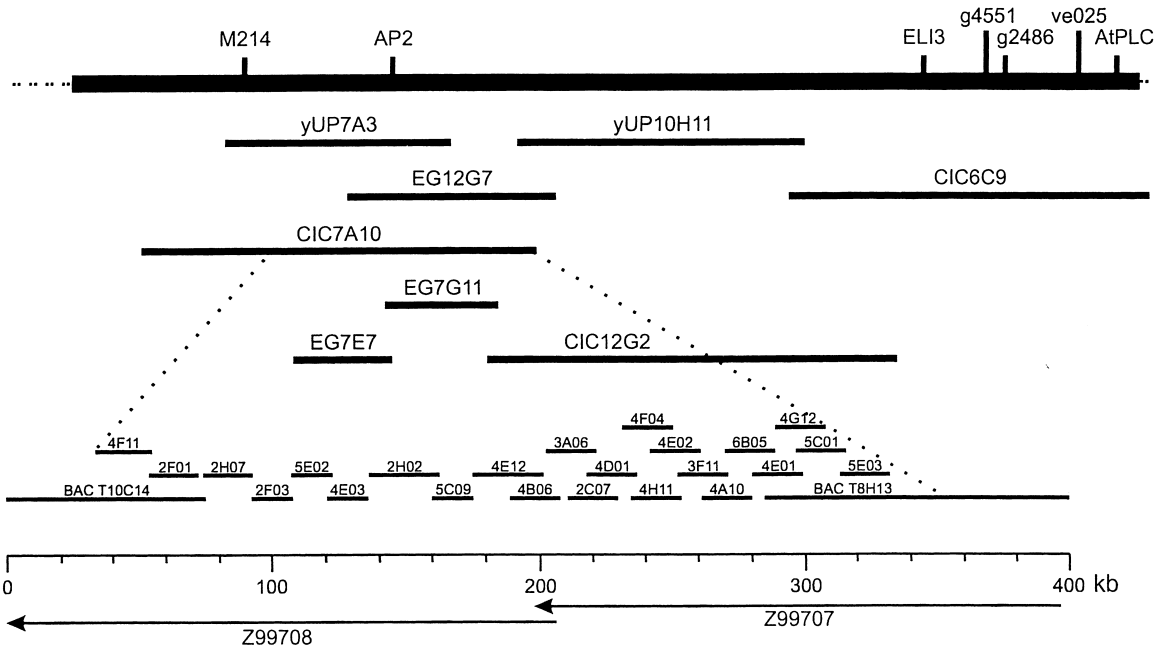

Fig. 1. Chromosomal location of the *AP2* gene. Overview of the cosmids and BACs used for sequencing. The upper line represents the chromosomal region (approximately 10 cM) with some known markers in the region. The sequenced cosmids and BAC clone are indicated as well as the region covered by the two database submissions covering this region (Z99707 and Z99708). The sequence Z99707 (206 440 bp) contains 50 genes (1–50) with an 8361-bp overlap with Z99708 (198 555 bp) that contains 53 genes (51–103).

Table 1
Overview of the 103 predicted genes in the 397-kb contig

No.	S[a]	Function/ product	Position	Best homologue accession	Name	Species	BLASTP score	EST	BLASTN/EST score	Class[b]	Category[c]
1	W	Cytochrome P450	<1–560	P24465	cytochrome P450	avocado	1.6E-25	H76015	2.1E-76	3	12
2	W	Cytochrome P450	978–2731	P24465	cytochrome P450	avocado	7.6E-60	H76015	3.6E-65	2	12
3	W	Cytochrome P450	3115–5137	P24465	cytochrome P450	avocado	1.5E-67	AA395149	1.3E-149	2	12
4	W	Cytochrome P450	5994–7943	P24465	cytochrome P450	avocado	1.6E-21	T43640	5.9E-47	3	12
5	W	Cytochrome P450	8851–11 532	P24465	cytochrome P450	avocado	1.0E-68	F13573	6.4E-104	2	12
6	C	Unknown	12 186–12 879					N38505	9.9E-131	5	13
7	W	Unknown	17 504–17 755	AC002391	unknown	*Arabidopsis*				4	13
8	W	Unknown	18 322–20 936					D15421 RICE	5.7E-22	6	13
9	W	Cation-transporting ATP-ase	21 332–25 697	P30336	Cu-transporting ATPase	mouse	3.3E-14	Z33731	2.0E-104	3	7,22
10	C	Myb-related	26 120–27 082	Z54136	Myb-related	*Arabidopsis*	4.2E-69	T43211	5.5E-35	2	4.19.04
11	W	Kinase-like receptor	37 499–39 895	P33543	receptor kinase TMK1	*Arabidopsis*	4.5E-50	H76836	1.4E-135	3	10,01
12	C	Unknown	42 632–43 066					H76663	4.6E-34	5	13
13	C	Pseudo-gene	50 097–51 300	S04132	photosystem II oxygen-evolving complex	pea		D49160 RICE	8.2E-22	3	2,2
14	C	Cold acclimation protein	51 715–52 578	U73216	cold acclimation	*Triticum aestivum*	4.4E-56	AA650647	1.3E-35	2	11,05
15	C	Putative histone binding	53 662–55 308	A25680	nuclear histone binding	African clawed frog	3.3E-07			3	13
16	W	Thioredoxin	56 454–57 848	P73920	thiodisulfide inter-change protein	*Synechocystis*	4.4E-23	AA394562	1.9E-150	3	2,2
17	C	Putative cell division regulator	58 172–60 529	U80043	misato gene	*Drosophila melano-gaster*	3.5E-15	R64911	5.4E-110	3	3,22
18	C	Unknown	61 577–62 335 or 63 009							6	13
19	W	Glu-t-RNA	66 821–66 868	K00193	Glu-tRNA	*Drosophila melano-gaster*				3	4,19
20	W	Unknown	67 029–69 104					F15442	2.1E-30	6	13
21	W	Pectinesterase	70 560–72 864	Z94058	pectinesterase	tomato	1.6E-123	H76007	1.3E-106	2	1,05 9,01
22	W	Hydroxynitrile lyase	73 627–74 699	Z34271	pir7a	rice	3.3E-39	AA651482	2.5E-26	2	11,02
23	W	Pseudo-gene hydroxynitrile lyase	74 956–75 569	Z34271	pir7b	rice	2.5E-27	R30024	9.5E-20	3	11,02
24	C	Unknown	75 724–78 117	U63839	nucleoporin	rat	2.3E-02	AA231701	6.9E-18	6	13
25	C	Unknown	78 561–80 761	AC001229	unknown	*Caenorhabditis elegans*	0.0E+00	AA394956	2.9E-117	4	13
26	W	Unknown	81 487–83 341					AA651330	6.3E-01	6	13
27	W	Unknown	85 821–88 511	Z69793	unknown	*Caenorhabditis elegans*	1.2E-08	T21702	7.3E-19	4	13
28	C	Unknown	89 125–90 346							6	13
29	C	Unknown	93 476–90 968							6	13
30*	W	Patatin homologue	100 626–102 768	P15478	patatin	tobacco	9.2E-57	H35955	2.3E-128	2	6,2
31	W	Patatin homologue	103 950–106 111	AJ002596	patatin	tobacco	4.7E-228	H35956	2.6E-106	3	6,2
32	W	Patatin homologue	108 042–110 423	AJ002596	patatin	tobacco	1.1E-153	H35957	7.5E-29	3	6,2
33	C	Methionyl-amino-peptidase-like	110 599–112 509	AL008883	methionyl amino-peptidase	yeast	8.8E-70	AA394671	6.1E-23	2	6,07
34	C	Unknown	113 337–115 534							6	13
35	C	Unknown	120 506–121 725					AA394779	2.0E-154	5	13
36	C	Caltractin-like	122 304–123 263	P41210	caltractin	*Atriplex nummularia*	2.0E-58	AA395741	6.5E-136	2	9,04 3,25

Table 1 (*continued*)
Overview of the 103 predicted genes in the 397-kb contig

No.	S[a]	Function/ product	Position	Best homologue accession	Name	Species	BLASTP score	EST	BLASTN/EST score	Class[b]	Category[c]
37	C	Unknown	123 922–125 057							6	13
38	C	Heat shock factor	125 978–127 024	U68017	HS factor HSF4	*Arabidopsis*	2.2E-188	R90161	0.0E+00	1	11,05 4.19.04
39	W	Unknown	130 672–133 998							6	13
40	W	Unknown	136 224–137 857							6	13
41	C	Ribonucleoprotein	138 243–140 333	S40774	ribonucleoprotein	*Xenopus*	7.8E-22	T43343	1.6E-05	3	4,22
42	W	Kinase	143 838–147 989	AB000798	protein kinase similar to NPK1	*Arabidopsis*	1.3E-23	Z26660	8.9E-32	3	10.04.04
43	C	Putative nicotinate phosphoribosyl-transferase	148 171–150 743	Z99120	nicotinate phospho-ribosyltransferase	*Bacillus subtilis*	5.3E-47	N65739	3.2E-93	3	13
44	C	Unknown	151 299–153 514	AC002330			2.9E-07	C21906	3.1E-08	6	13
45	C	AP2	164 541–166 683	U12546	apetala2	*Arabidopsis*	1.6E-283			1	4.19.04
46	W	Unknown	174 463–176 665	D63999	dehydrogenase	*Synechocystis*	1.8E-126			3	13
47*	C	TINY-like	178 076–178 665	AJ002598	TINY	*Arabidopsis*	2.6E-182	AA404810	4.0E-115	2	4.19.04
48	W	Unknown	185 868–188 050					Z46429	4.7E-76	5	13
49	W	Cysteine proteinase	191 501–192 989	P25251	cysteine proteinase	rape	7.8E-198	R84153	4.3E-81	2	6,13
50	C	Putative homeotic protein	193 958–198 256	A57632	BEL1	*Arabidopsis*	6.3E-80			2	4.19.04
51	W	Unknown	8363–11 049	AC003000			6.7E-89	N65197	2.4E-124	5	13
52	W	Unknown	13 796–16 202	S57377	probable membrane protein	yeast	4.4E-13	H77118	1.8E-104	5	13
								AA605507	3.9E-120		
53	C	Unknown	16 736–17 917					H77079	7.2E-11	6	13
54	C	Unknown	19 236–20 105					T45881	9.3E-147	5	13
55	C	Unknown	21 622–22 727	U41558	unknown	*Caenorhabditis elegans*	7.5E-11	L46423	1.3E-48	4	13
56	C	Geranylgeranyl-pyro-phosphatase synthase	24 988–26 103	P34802	geranylgeranyl-pyro-phosphatase synthase	*Arabidopsis*	1.0E-179	L37477	1.0E-08	1	20,2
57	W	Ubiquitin-conjugating protein	27 468–28 378	P52491	ubiquitin-conjugating protein	yeast	4.0E-34	T41550	0.0E+00	3	6,07
58	C	Unknown	31 311–33 255	AJ001694	membrane protein	*Thermotoga maritima*	3.1E-08			4	13
59	W	Unknown	35 403–36 625					R64792	1.0E-162	5	13
60	W	Glucosyltransferase	38 024–39 398	Q40287	UTP-glucose glucosyl-transferase	cassava	9.6E-106			2	1,05
61	C	Aminopepidase	39 636–42 894	AF038591	X-pro aminopeptidase	rat	5.9E-135	N96008	3.0E-68	2	6,13
62	C	Phycoerythrin-like protein	43 401–44 974	D45900	Ledi-3	*Lithospernum erythrorhizon*	1.50E-49	R64949	1.0E-158	2	12
63	W	Putative homeotic protein	53 303–54 968	X94947	homeotic protein	tomato	2.0E-23			3	4.19.04
64	W	G-box binding GBF1	57 865–59 767	X63894	G-box binding GBF1	*Arabidopsis*	1.0E-162	H37673	3.0E-43	1	4.19.04 9,1
65	C	Pseudo-gene	59 949–61 848	Q00765	TB2	human	2.0E-20			3	13
66	C	Scarecrow homologue	62 097–63 557	U62798	scarecrow	*Arabidopsis*	1.0E-23	N65163	1.0E-175	3	4.19.04
67	W	Putative storage protein	69 280–71 174	Z54364	vicilin	*Matteuccia struthiopteris*	4.0E-38	Z48554	4.0E-70	3	6,2
68	W	Splicing factor	72 008–75 478	S20250	splicing factor	human	8.0E-70	N96704	1.0E-112	3	4,22
69*	W	Putative salt inducible	75 900–77 138	AJ002597	putative salt inducible	*Arabidopsis*	0.0E+00	R65492	1.0E-168	3	11,05
70	W	Transport protein	80 134–81 937	Q10286	sugar transporter	beetroot	1.6E-45			3	7,07
71	W	Unknown	82 336–83 267					H36382	3.0E-12	6	13
72	W	Putative transcription initiation factor	84 089–86 333	P29053	transcription initiation factor IIB	African clawed frog	5.0E-26			3	4,19

Table 1 (*continued*)
Overview of the 103 predicted genes in the 397-kb contig

No.	S[a]	Function/ product	Position	Best homologue accession	Name	Species	BLASTP score	EST	BLASTN/EST score	Class[b]	Category[c]
73	W	Unknown	91 169–92 429	U88068	sec14	rice	5.0E-21	R65425 H76207	1.0E-120 3.0E-74	5	13
74	W	Unknown	93 093–97 527							6	13
75	W	Unknown	99 953–100 709	(zinc finger domain)						3	4.19.01
76	W	Unknown	102 341–104 070	U83881	hydrolase	*Pseudomonas*				3	13
77	C	Putative LEA protein	104 647–105 947	P20075	LEA	soybean	2.7E-17	Z29854	1.0E+131	3	12
78	W	MADS box protein	107 427–108 470	Y12776		*Picea abies*	2.0E-33			3	4.19.04
79	C	Unknown	108 955–111 658	Z48583	unknown	*Caenorhabditis elegans*	1.0E-83			4	13
80	C	Unknown	114 568–114 956							6	13
81	C	Unknown	116 627–116 959							6	13
82	W	Unknown	121 900–124 218							6	13
83	C	Unknown	124 653–126 204	HLH motif				AA395050	1.0E-138	5	13
84	W	Unknown	127 848–129 498	D90906	hypothetical protein	*Synechocystis*	7.0E-55			4	13
85	W	Unknown	134 183–136 498	Q07283	trichohyalin	human	3.0E-08			3	13
86	C	Unknown	140 167–140 841							6	13
87	W	Unknown	143 111–143 479					N97118	0.0E+00	5	13
88	C	Unknown	144 810–147 488	P24859	Sec14 (PI/PC) transport protein	*Kluyveromyces lactis*	4.0E-47	AA395164	1.0E-08	3	13
89	C	Putative acyltransferase	148 494–151 020	X95641	serine C-palmitoyl transferase	mouse	2.0E-93	R90586	8.0E-81	3	1,06
90	W	Unknown	153 143–154 490							6	13
91	W	Unknown	156 519–156 956							6	13
92	W	MAP kinase	158 205–159 373	Q39027	MAP kinase 7	*Arabidopsis*	0.0E+00	T46244	1.0E-64	2	10.04.04
93	W	Unknown	160 512or 161 813 160 186 or 162 618					T88343	1.0E-64	6	13
94	W	Peroxidase	163 701–164 970	X98804	peroxidase	*Arabidopsis*	1.0E-109	F13569	1.0E+113	2	12
95	W	Putative ribosomal protein	165 361–165 900	P36212	L12 ribosomal protein	*Arabidopsis*	2.0E-11	Z18520 AA042150	1.0E-107 0.0E+00	3	5,01
96	C	Ubiquitin-conjugating protein	166 728–167 649	P42743	ubiquitin-conjugating enzyme	*Arabidopsis*	7.0E-68	AA728692	6.0E-85	3	2,13
97	C	Actin-interacting protein	169 146–172 037	P46681	AIP2 protein	yeast	1.0E-144	AA042437	1.0E-70	2	9,04
98	W	Unknown	172 841–174 872	Q09316	unknown	*Caenorhabditis elegans*	1.0E-121			4	13
99	W	Cytochrome P450	177 617–181 645	S55379	cytochrome P450	*Arabidopsis*	2.0E+91	N96214	0.0E+00	2	20,1
100	C	Unknown	182 008–182 368							6	13
101	W	β-Galactosidase precursor	188 484–192 778	P45582	β-galactosidase precursor	tomato	7.6E-256	Z37275	1.0E-179	2	1,05
102	W	Acid phosphatase	193 761–195 881	AL022141	acid phosphatase	*Arabidopsis*	7.5E-120			2	10.04.07
103	W	Unknown	196 431 > 198 459							6	13

[a]S, strand.
[b]The homology classification can be found in Table 3.
[c]The functional category to be found at the MIPS web site (http://muntjac.mips.biochem.de/arabi/fca/gene/funcat_table.html). In the last column putative cognate ESTs are indicated (> 95% identity); * were fully sequenced and submitted to the GenBank as cognate cDNAs with accession numbers AJ002596–AJ002598. The numbering is done according to the EMBL/GenBank entries Z99707 (genes 1–50) and Z99708 (genes 51–103).

Table 2
Classes of similarities to genes

Class	BLAST E value	Type of matching protein	Number	Predicted function
1	identical	same protein	4	4
2	$<10^{-50}$	known protein	23	23
3	between 10^{-10} and 10^{-50}	known protein	33	28
4	$<10^{-10}$	hypothetical protein	8	–
5	$>10^{-10}$	none, but has cognate EST match	11	–
6	<150	none, no cognate EST match	24	–
		Total	103	55

five. In addition, the re-annotation pointed to a few frameshifts caused by sequence errors that have been corrected this way. Similar issues have been raised for annotation of bacterial genomes [18], but to a lesser extent. As the task of annotation of a higher eukaryote genome is much more difficult because of the larger size of the genome, the lesser gene content, the split gene structure, and less homologies to known database entries, such a result is not surprising. It is clearly a warning for caution when using the present-day annotations, and it indicates that complete re-annotation of the genome will be needed.

3.3. Statistical analysis of the contig

As can be deduced from Table 1, 59 genes can be found on the Watson and 44 on the Crick strand. Gene density is quite high (3.85 kb/gene) compared to that of the *FCA* region [3] (4.8 kb/gene) and the mean density of 4.1 kb/gene reported for 6.7 Mb sequenced on chromosome 5 [19]. Nevertheless, regions of 1.2 Mb on chromosome 5 have been reported with a similarly high gene density of 3.84 kb/gene [20]. In total, 42% of the genes are highly similar to *Arabidopsis* EST sequences ($>95\%$ similarity).

Table 3 represents a statistical analysis done on both strands of the 400-kb contig in terms of occurrence and size of genes, introns, and exons. Intergenic regions cover approximately 53% of the contig, whereas introns represent 16% and coding sequences 31%. As a mean, the first intron is longer than the others and the first and last exons longer than the middle ones. It is noteworthy that local as well as strand heterogeneities were observed. For example, genes in the first half of the sequence have smaller introns than those in the second half of the contig, whereas exons have a similar aver-

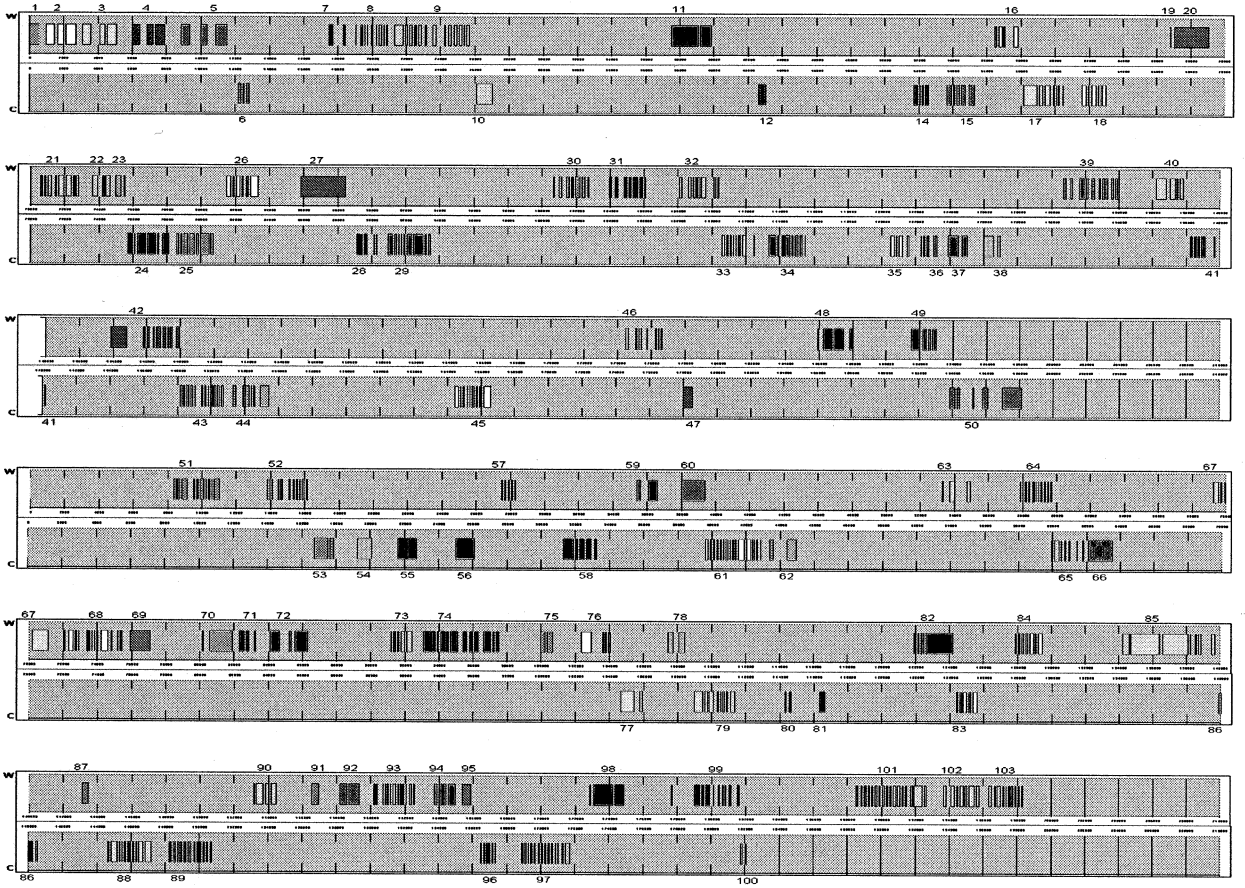

Fig. 2. Schematic overview of the exons of the 103 genes present in the 379-kb contig. Genes are marked by numbers; each number represents a new gene. Gene 13 is not indicated.

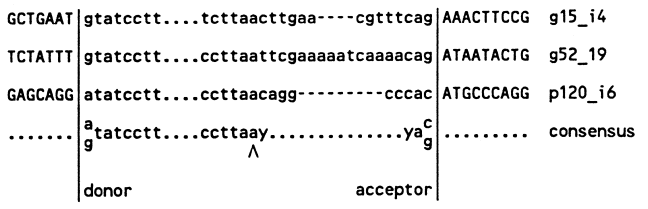

Fig. 3. U12 class introns. g15_i4, U12-type intron #4 of gene 15 from this contig; g52_i9, U12-type intron #9 of gene 52 from this contig; p120_i6, U12-type intron #6 from the human *p120* gene. The intron sequence is given in lowercase. ∧ marks the branch point.

age size (data not shown). In addition, both exons and introns on the Watson strand are larger than those on the Crick strand.

3.4. Finding of two U12-type introns

The 102 protein-encoding genes were found or predicted to contain 429 introns. All of these introns, except two, have the consensus sequences of classical U2-type introns. The remaining two are the last intron (i4) from gene 15, which is distantly related to the histone-binding protein from *Xenopus*, and the last intron (i9) from gene 52, which is clearly similar to a small yeast gene family encoding membrane proteins of unknown function. Interestingly, a paralogue of gene 52 was found in another contig from chromosome 4 (al022224/al021637), but this gene does not seem to have any intron. These introns, although having GT-AG borders, display the distinctive features of U12-type introns with their very conserved donor and branch sites and the typical short distance between the potential branch point and the acceptor site with no polypyrimidine tract, as shown in Fig. 3. Such introns were initially found in animals where they have been shown to need specific snRNAs for their splicing. More recently, they have been found in plants too [21,22] and are considered to be of rare occurrence. The finding of two members out of a total 429 may give a first estimation of their actual frequency in the *Arabidopsis* genome. For the time being, these introns are not correctly predicted by any gene prediction program.

3.5. Gene clustering

There are three clusters of tandem gene repeats in the contig. A cluster of five *P450* genes is found at its 5′ extremity. These genes have the same genomic structure (three exons) and their coding sequences are very similar to each other (74–83% similarity between copies 1–4, 60% between them and copy 5), their best homologue being an elicitor-induced *P450* from licorice. There is another *P450* gene (gene 99) at the other end of the contig that differs from the genes in the cluster (low homology, 10 exons). The clustering of these five *P450* genes suggests that they might originate from a common ancestor late in evolution and could probably be involved in different steps of the same pathway (secondary metabolism, defence, etc.).

Three patatin genes (genes 30, 31, and 32) are also found clustered. Patatin is the major storage protein of potato and homologues have already been found in other species, but not yet in *Arabidopsis*. The patatins found in this cluster are similar to each other, the first and second copy being very close (90%). The third copy is more distant (67–69% similarity to the others) and differs specifically at the N-terminus, suggesting a different subcellular localization of this member.

The third cluster is an imperfect tandem repeat of a gene encoding a protein with strong similarity to hydroxynitrile lyase, an enzyme that produces cyanide by hydrolyzing cyanogen glucosides, which are secondary metabolites produced in a narrow range of plants. Only the first repeat contains a complete and potentially functional gene with three exons. The gene in the second repeat seems truncated after the first exon, and would thus be a pseudo-gene. Paralogues of this gene have been found in the *FCA* region [3]; the finding of such a gene is unexpected, because *Arabidopsis* has not been reported to produce cyanogen glucosides. It would be interesting to check whether these genes are functional and induced by pathogens and/or predator attack, and to examine which substrate their products would hydrolyze.

3.6. Genome duplication

During the process of gene search, several genes at the 5′ extremity of the contig were observed to have homologues located in the AC002391 contig from chromosome 2. To

Table 3
Statistical analysis of the *AP2* contig

Total	Number	Mean per gene	Total size (bp)	Mean size (bp)	%	W strand[c] (+)	C strand[c] (−)
Genic regions							
RNA CDS	1		71				
Protein CDS[a]	100		121 289	1214	**31.2**	56*	44*
Without introns	15		13 707	914		7*	8*
With introns	85		107 692	1270		49*	36*
Introns[a,b]							
Total	429	**5.05**	63 530	148	**16.2**	160	133
First	85		16 785	197		248	129
Rest	344		46 745	136		137	134
Exons[a,b]							
Total	526	**6.19**	107 692	205		230	180
First	85		25 961	305		314	293
Rest	346		55 625	161		176	141
Last	85		26 106	307		364	229
Intergenic regions							
	102		208 556	2045	**52.6**		

[a]Statistics on full-length genes.
[b]Means refer to intron-containing genes; for all genes, the mean numbers of introns and exons per gene are 4.29 and 5.26, respectively.
[c]W and C strand columns refer to sizes, except for figures followed by an asterisk, where they represent numbers. The region analyzed is the whole 396-kb contig, except for gene 13, for which the prediction is uncertain (a pseudo-gene).

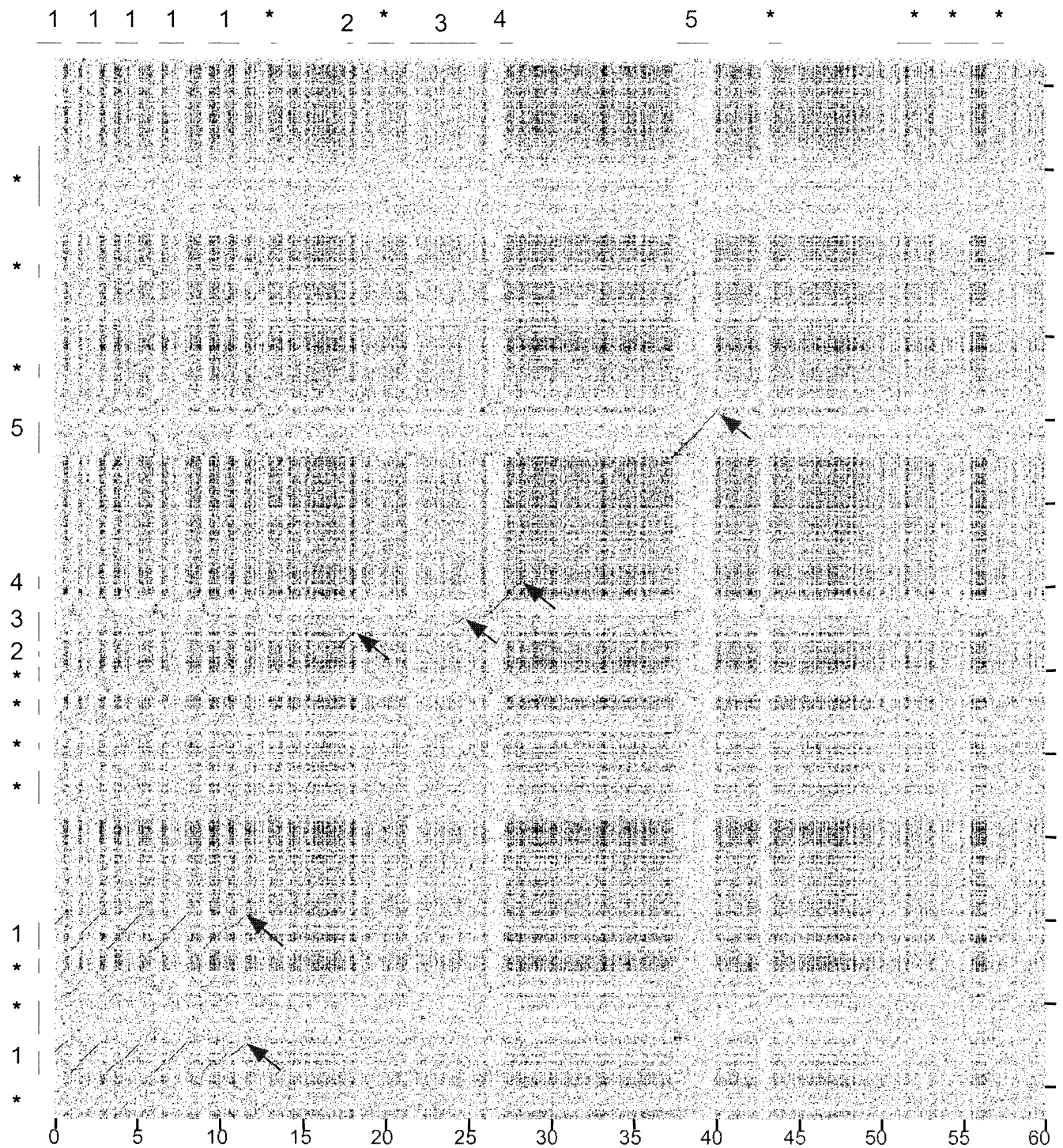

Fig. 4. Dot-plot comparison of a region on chromosome 2 (AC002391, bp 15 000–75 000) with the first 60 kb of the sequence described here (Z99707). Genes are numbered as in Table 1. Asterisks indicate non-homologous genes, the arrows inside the plot mark the regions of homologous genes.

check the gene arrangement, a dot-plot comparison of both contigs was performed using the GCG compare/dot-plot and the dotter [23] programs. As seen in Fig. 4, the first 45 kb (Watson strand) of the contig show similarities to a region of comparable size of the AC002391 contig (Crick strand, 77 000–32 000). The extent of the duplication is perhaps larger, because no sequence is available yet on the 5′ side of the *AP2* contig described here. The homology is patchy, being mostly restricted to the coding sequence of the homologous genes. Among the 12 genes found in this region on the contig, nine have a counterpart in the chromosome 2 contig, the order and the orientation of the genes being conserved. This situation strongly suggests an ancient duplication of this region in the *Arabidopsis* genome. This duplication must indeed be ancient, because most non-coding sequences (introns, intergenic sequences) are not conserved and because the two copies

differ by several insertions/deletions of various genes. One such deletion is found in the chromosome 2 counterpart of gene 9 for which homologies are found with several metal (Cu, Cd) transporters. On chromosome 4, gene 9 is predicted with 13 exons, encoding a 819-amino acid protein. On chromosome 2, the corresponding gene is truncated on its 5′ side and is predicted to have three exons encoding a 221-amino acid protein. The *P450* cluster described above is the 5′-most part of the duplicated region. Whereas the chromosome 4 copy contains five *P450* members in tandem repeat, only two *P450* members are found in the chromosome 2 counterpart, separated by an insertion of two foreign genes and flanked by a *P450* of different origin. The genetic distances between the different *P450* members suggests that the duplication of the *P450* genes on chromosome 2, and at least the duplications responsible for the four first copies on chromosome 4, occurred after the duplication of the 45-kb region itself. Although the *Arabidopsis* genome is of small size, duplication of individual genes appears to be frequent. Duplication on a larger scale has been suggested from mapping data [24], but, to our knowledge, this is the first demonstration of a duplication of a large region in *Arabidopsis*. The completion of the *Arabidopsis* genome sequence will tell us to which extent these duplications occur. A comparative analysis of the repeats in various ecotypes and neighbor species will inform us when this happened in the *Arabidopsis* evolutionary history.

Acknowledgements: The authors wish to acknowledge the ABRC Stock Center in Ohio for the distribution of ESTs, M. Bevan for coordinating the ESSA program, H.W. Mewes, S. Klostermann, and N. Chalwatzis from MIPS for computer data analysis, Peter Breyne and Koen Goethals for critical reading of the manuscript, M. De Cock for help preparing it, and Rebecca Verbanck for the figures. This work was supported by the EU (Biotech ERBB102-CT93-0075 and ERBB104-CT96-0338) and by the Flemish government (VLAB-COT).

References

[1] Goodman, H.M., Ecker, J.R. and Dean, C. (1995) Proc. Natl. Acad. Sci. USA 92, 10831–10835.
[2] Arabidopsis Genome Initiative (1997) Plant Cell 9, 476–478.
[3] Bevan, M., Bancroft, I., Bent, E., Love, K., Piffanelli, P., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., Drost, L., Ridley, P., Hudson, S.-A., Patel, K., Murphy, G., Wedler, H., Wedler, E., Wanbutt, R., Weitzenegger, T., Pohl, T., Terryn, N., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Lecharny, A., Kreis, M., Lao, N., Kavanagh, T., Hempel, S., Kotter, P., Entian, K.-D., Rieger, M., Scholfer, M., Funk, N., Muller-Auer, S., Silvey, M., James, R., Montfort, A., Pons, A., Puigdomenech, P., Douka, A., Voukelatou, E., Milioni, D., Hatzopoulos, P., Piravandi, E., Obermaier, B., Hilbert, H., Duesterhoeft, A., Moores, T., Jones, J., Eneva, T., Palme, K., Benes, V., Rechman, S., Ansorge, W., Cooke, R., Berger, C., Delseny, M., Volckaert, G., Mewes, H.-W., Schueller, C. and Chalwatzis, N. (1998) Nature 391, 485–488.
[4] Koornneef, M., van Eden, J., Hanhart, C.J., Stam, P., Braaksma, F.J. and Feenstra, W.J. (1983) J. Hered. 74, 265–272.
[5] Bowman, J.L., Smyth, D.R. and Meyerowitz, E.M. (1989) Plant Cell 1, 37–52.
[6] Jofuku, K.D., den Boer, B.G.W., Van Montagu, M. and Okamuro, J.K. (1994) Plant Cell 6, 1211–1225.
[7] Dean, C. and Lister, C. (1996) in: Weeds World: The International Electronic Arabidopsis Newsletter (Anderson, M., Ed.), Vol. 3 (iii), pp. 1–6, Nottingham Arabidopsis Stock Centre, Nottingham.
[8] Zabeau, M. and Vos, P. (1993) European Patent Application EP 534858A1.
[9] Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. and Zabeau, M. (1995) Nucleic Acids Res. 23, 4407–4414.
[10] Bancroft, I., Love, K., Bent, E., Sherson, S., Lister, C., Cobbett, C., Goodman, H.M. and Dean, C. (1997) in: Weeds World: The International Electronic Arabidopsis Newsletter (Anderson, M., ed.), Vol. 4 (ii), pp. 1–9, Nottingham Arabidopsis Stock Centre, Nottingham.
[11] Déhais, P., Coppieters, J. and Van Montagu, M. (1996) Arch. Physiol. Biochem. 104, B38.
[12] Staden, R. (1996) Mol. Biotechnol. 5, 233–241.
[13] Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. (1996) Nucleic Acids Res. 24, 3439–3452.
[14] Tolstrup, N., Rouzé, P. and Brunak, S. (1997) Nucleic Acids Res. 25, 3159–3163.
[15] Borodovsky, M. and MacIninch, J.D. (1993) Comput. Chem. 17, 123–133.
[16] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[17] Pedersen, A.G. and Nielsen, H. (1997) in: Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. and Valencia, A., Eds.), pp. 226–233, American Association for Artificial Intelligence Press, Menlo Park, CA.
[18] Galperin, M.Y. and Koonin, E.V. (1998) In Silico Biol. 1, 0007 (http://www.bioinfo.de/isb/1998/01/0007/).
[19] Nakamura, Y., Sato, S., Kaneko, T., Kotani, H., Asamizu, E., Miyajima, N. and Tabata, S. (1997) DNA Res. 4, 401–414.
[20] Kaneko, T., Kotani, H., Nakamura, Y., Sato, S., Asamizu, E., Miyajima, N. and Tabata, S. (1998) DNA Res. 5, 131–145.
[21] Wu, H.-J., Gaubier-Comella, P., Delseny, M., Grellet, F., Van Montagu, M. and Rouzé, P. (1996) Nature Genet. 14, 383–384.
[22] Sharp, P.A. and Burge, C.B. (1997) Cell 91, 875–879.
[23] Sonnhammer, E.L.L. and Durbin, R. (1995) Gene 167, GC1–10.
[24] Kowalski, S.P., Lan, T.-H., Feldmann, K.A. and Paterson, A.H. (1994) Genetics 138, 499–510.