Biol. Chem., Vol. 385, pp. 961–965, October 2004 • Copyright © by Walter de Gruyter • Berlin • New York. DOI 10.1515/BC.2004.125

Short Communication

A database search for double-strand containing RNAs in *Dictyostelium discoideum*

Stefan Gräf[1], Branimira E. Borisova[2], Wolfgang Nellen[3], Gerhard Steger[1] and Christian Hammann[2,*]

[1] Institut für Physikalische Biologie, Heinrich-Heine-Universität, D-40225 Düsseldorf, Germany

[2] Department of Genetics, AG Molecular Interactions, University of Kassel, D-34132 Kassel, Germany

[3] Department of Genetics, University of Kassel, D-34132 Kassel, Germany

[*] Corresponding author
e-mail: c.hammann@uni-kassel.de

Abstract

In eukaryotic cells, double-stranded RNA is degraded to 21mers and triggers RNA interference. Using a pattern description language, we have searched the EMBL database for sequences with the potential to form double strands *in cis* in *Dictyostelium discoideum*. No extended inverted repeats were found in mRNAs. However, the antisense direction of some mRNAs encoding regulatory or developmentally regulated proteins showed the ability to form double-stranded regions. In EST archives, we found potential double strands derived from a few genes, but these transcripts are not continuously encoded in the genome. Most likely, they represent hybrid molecules of sense and antisense RNAs.

Keywords: gene regulation; genome analysis; pattern search; RNAi; siRNAs.

The presence of long double-stranded RNA (dsRNA) leads in eukaryotes to silencing of genes that are homologous in sequence, by a process termed RNA interference (RNAi; Fire et al., 1998). Long dsRNA is converted into small interfering RNAs (siRNAs) by a multi-domain RNaseIII protein termed Dicer (Bernstein et al., 2001; Billy et al., 2001; Novotny et al., 2001; Zamore et al., 2000). For mammalian systems, the situation is complicated by the fact that long dsRNA provokes the PKR response, which eventually leads to apoptosis (Manche et al., 1992). However, the direct use of small duplexes of 21 nucleotide (nt)-long RNA circumvents this PKR response, allowing the application of RNAi also in mammalian systems (Elbashir et al., 2001). In mammalian systems, *D. melanogaster* and possibly *C. elegans*, siRNAs are incorporated in a protein complex termed RISC (RNA-induced silencing complex) that subsequently degrades the cognate mRNA (summarised in Tomari et al., 2004). Such a complex, however, has not been described for other organisms, such as fungi, plants or amoebae, which nevertheless do display RNA interference. For example, in *Dictyostelium discoideum* we could demonstrate the existence of an alternative mechanism acting in RNAi that requires the presence of an RNA-directed RNA polymerase (RdRP; Martens et al., 2002), leading to the interpretation that siRNAs serve as primers on the cognate mRNA, which then are extended by an RdRP to create a longer dsRNA. Similar to the situation in plants, this observation does not preclude that a RISC-like complex might exist; it is, however, not required for the functionality of RNAi in *D. discoideum*.

The functionality of the RNAi system requires that endogenous RNAs lack longer double-stranded stretches, as this would lead to degradation. Alternatively, the very presence of such double strands might be a natural mechanism of gene regulation, especially if formation or accessibility of a double strand occurred in a controlled manner.

To address the question of potential intramolecular dsRNA formation and thus the possibility of an endogenous RNAi-based regulatory system, we performed a computational analysis of the available *Dictyostelium* sequence data, as described in the legend to Figure 1. To this end, all entries containing ‘*Dictyostelium discoideum*’ in the OS (Organism Species) line were retrieved from release 78 of the EMBL sequence database (Kulikova et al., 2004) by using BioPerl (Stajich et al., 2002). Overall, 156 560 entries with 91 077 278 nt were obtained as indicated in the legend to Figure 1.

Searches for intramolecular dsRNA were performed by a modified version of PatScan (Dsouza et al., 1997) in forward and reverse strands of each entry in overlapping mode and with outputting bracket-dot notation, indicating the secondary structure. All search patterns, given in the legend to Figure 1, describe a long hairpin with 25 to 50 basepairs (bp) in the helix and different loop sizes (3 to 5 nt, 6 to 10 nt, 11 to 20 nt, 21 to 30 nt, 31 to 40 nt, 41 to 50 nt). The majority of the overall 15 836 matched sequences in the database entries were derived from expressed sequence tag (EST) archives (see Figure 1).

Since the same match can be found in a database entry by more than one search pattern, we pooled all hits from the different patterns and filtered out those that were contained completely in another hit. This reduced the number of hits to 3170. The *Dictyostelium* genome is characterised by an AT content of nearly 80% (Glockner et al., 2002) with a strong accumulation of A and T residues in non-transcribed and non-coding regions. Although we cannot rule out that some of these sequences may play a regulatory role, the majority does most likely not generate any transcripts. To avoid hits with low or no significance, sequences with an AT content of more than 86%, sequences with more than 10 consecutive As or Ts and helices containing exclusively AU or GU base pairs were filtered out. This resulted in 1932 remaining hits. When analysed by Clustal (Jeanmougin et al., 1998), no consensus could be obtained for the double strands contained in these hits (data not shown). The lengths of the double strands of these database entries were not uniformly distributed. One cluster, which consists of about 1/3 of all hits, had helices in the narrow range of 25 to 29 bp as shown in Figure 2A; this is considered to be the minimal target size for Dicer and most dsRNA binding proteins. The majority of hits, however, showed the ability to form double strands of 50 bps or longer. This peculiar distribution is only in part due to the search parameters: since the cut-off was set to 25 bp, we do not know how abundant shorter helices are. We also have not investigated in detail the length distribution of double strands longer than 50 bp; there is evidence, however, for a second cluster near 54 bp (data not shown).

About 90% of the filtered hits were derived from EST archives. EST database entries are difficult to handle because of the uncertainties concerning the correct orientation of the ESTs (Lehner et al., 2002). In the context

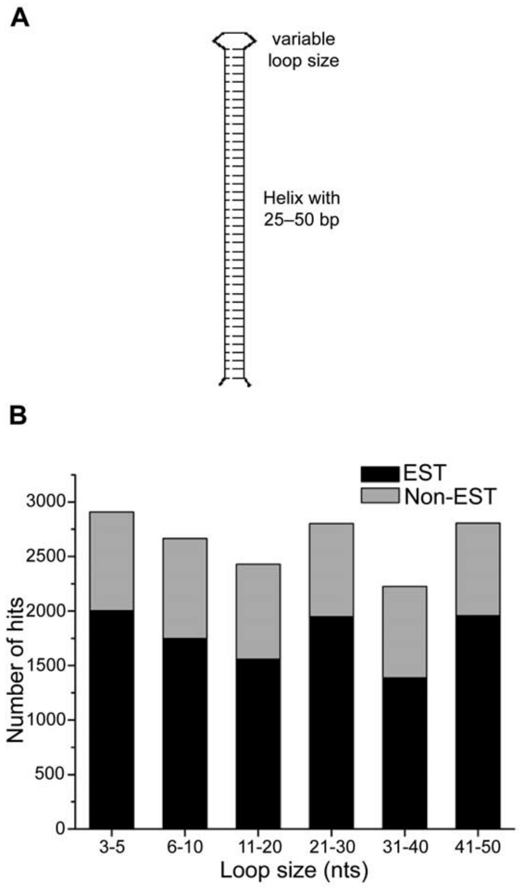

Figure 1 Distribution of nucleotide sequences with the ability to form double strands encoded in the *Dictyostelium* genome.
Results are displayed after filtering out overlapping hits and grouped by the size of the hairpin loop closing the double strand. *Dictyostelium* sequence data were retrieved from EMBL divisions estp6inv (invertebrate EST), htgp6inv (high throughput genome), inv (invertebrates) and org (organelles) resulting in overall 156 560 entries with 91 077 278 nucleotides. EMBL release 78 contains genomic sequence data of chromosome 2, sequences of published genes and EST data.
(A) Entries were searched for sequences able to form hairpins with perfect helices of at least 25 bp and hairpin loops of different lengths by a modified version of PatScan (Dsouza et al., 1997). For base pairs, Watson-Crick as well as wobble pairs were allowed [r1=(au,ua,gc,cg,gu,ug)]. The probability of hits in a random genome with a frequency of A:T:C:G = 4:4:1:1, which is the nucleotide frequency of *Dictyostelium* (Glockner et al., 2002), is stated as p-value. Pattern 1 (3–5 loop nt): p1=25…50 3…5 r1~p1[0,0,0], p-value: 1 match per 5.07×10^8 nt; pattern 2 (6–10 loop nt): p1=25…50 6…10 r1~p1[0,0,0], p-value: 1 match per 3.04×10^8 nt; pattern 3 (11–20 nt): p1=25…50 11…20 r1~p1[0,0,0], p-value: 1 match per 1.52×10^8 nt; pattern 4 (21–30 loop nt): p1=25…50 21…30 r1~p1[0,0,0], p-value: 1 match per 1.52×10^8 nt; pattern 5 (31–40 loop nt): p1=25…50 31…40 r1~p1[0,0,0], p-value: 1 match per 1.52×10^8 nt; pattern 6 (41–50 loop nt): p1=25…50 41…50 r1~p1[0,0,0], p-value: 1 match per 1.52×10^8 nt. (B) Matches from EST archives are shown as black bars and matches from all other archives as grey bars.

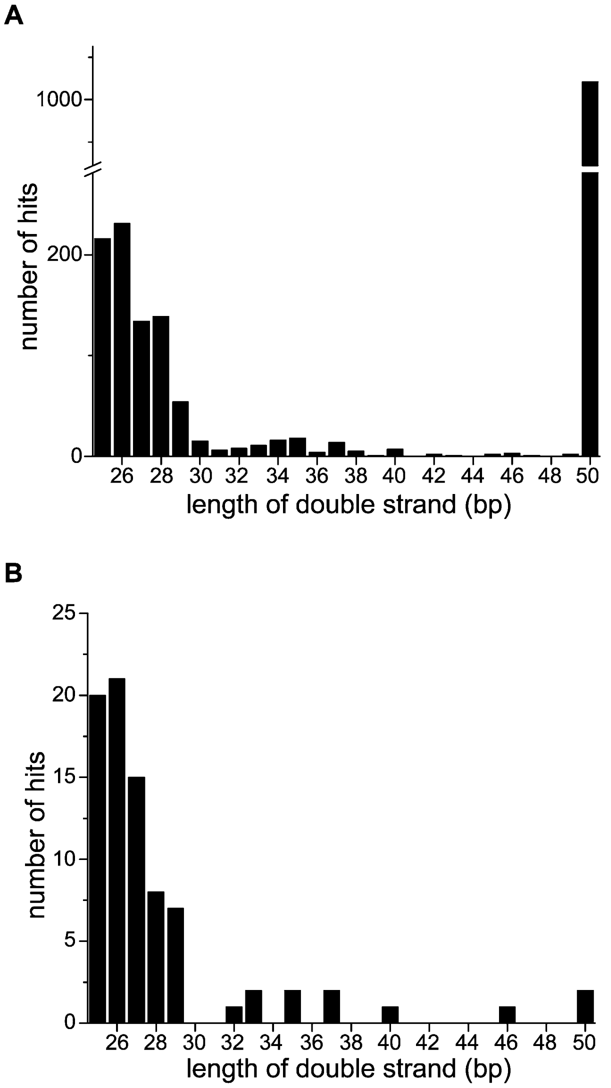

Figure 2 Distribution of the lengths of double-stranded regions.
(A) Length distribution of double strands in the merged and filtered hits from the database search, including EST data. The number of hits with 50 bp also includes hits with longer helices.
(B) Length distribution of double strands in the unique sequences, excluding EST data.

of our search, a lack of knowledge of the directionality of ESTs is detrimental, as a matched sequence might show a perfect long double strand only due to the presence of inter-dispersed GU and UG RNA base pairs. In the reverse complementary orientation of such a sequence, no double strand could be formed due to the presence of A and C nucleotides at the corresponding positions. Because of unknown directionality of ESTs, all perfect matches to predicted or known coding sequences in the Sequencing Center Gene Predictions section of Dictybase (http://www.dictybase.org/) were excluded from further consideration.

Surprisingly, EST hits still remained that could not be directly connected to annotated sequences in the *Dictyostelium* genome: clustered on a small number of genes, we found sequences with the potential to form double strands that were only partially identical to mRNAs. These sequences came from cyclase-associated protein (CAP), spore coat protein SP96, a histidine kinase, and four further ORFs. The EST entries related to these proteins were unusual in that they showed a double strand that was not genomically encoded, and thus could best be described as a hybrid RNA consisting of a sense and an antisense stretch. This may be explained by template swapping of the reverse transcriptase from the mRNA to the newly synthesised cDNA strand. However, in RNA folding studies of appropriate sequences using the secondary structure prediction program Mfold (Zuker, 2003) we did not observe highly structured parts in the relevant positions of the mRNAs that could explain such a template swapping. In this context, it is worth noting that also in humans novel sense-antisense RNA-hybrid structures were observed recently by RACE experiments on endogenous troponin I antisense RNA (Bartsch et al., 2004).

We next investigated the matched sequences from all archives except ESTs. The 186 hits were partially redundant, and contained identical double-strand sequences for example from DNA transposons. These had not been filtered out, since they had different surrounding sequences in the genome. Of the 82 unique sequences, none was in sense orientation within a coding region; however, 45 corresponded to the antisense orientation of protein encoding regions, 16 matched 3′- or 5′-untranslated regions (UTRs) and 5 corresponded to introns or intron-exon boundaries with the loop of the hairpin frequently containing the splice site. The remaining hits were derived from the *Dictyostelium* extrachromosomal palindromic ribosomal RNA gene sequence, from *Dictyostelium* DNA transposons (DDT-A, DDT-B, DDT-S, Thug-S; Glockner et al., 2001), and 11 sequences were located in presumably non-coding chromosomal regions. For one of the 45 assumed antisense RNAs, the psvA antisense transcript, we have previously shown that it is expressed and serves a regulatory function (Hildebrandt and Nellen, 1992).

The size distribution of the double strand in the remaining 82 sequences is shown in Figure 2B. It is obvious that the predominant size family of 50 bp and more disappeared while the cluster between 25 bp and 29 bp remained. The long double-stranded regions in Figure 2A are therefore mostly derived from ESTs. Some are most likely artifacts from reverse priming while others may represent genuine sense-antisense-hybrid molecules (Bartsch et al., 2004).

Table 1 Annotated proteins from *Dictyostelium discoideum* with double-stranded regions in antisense orientation of the mRNA.

Database entry[a]	Name[b]	Length of ds (bp)	Protein function[c]
DDB0184759	Roco11	25	Tyrosine kinase
DDB0185108	LvsD	25	BEACH domain protein
DDB0201665	Roco10	25	Tyrosine kinase
DDB0167949	DrnA	26	Ribonucelase III, dicer homologue
DDB0191393	CAR2	26	cAMP receptor
DDB0191441	WarA	26	Transcription factor
DDB0185054	GbpA	27	cGMP-specific phosphodiesterase
DDB0185194	DokA	27	Histidine kinase
DDB0191346	PIK4	27	1-phosphatidylinositol 4-kinase
DDB0169468	DG1040	28	CCR4-Not complex component, Not1
DDB0184761	Roco9	28	Tyrosine kinase
DDB0188821	GP138D	28	IPT/TIG domain protein
DDB0215390	DymB	28	Dynamin GTPase
DDB0191480	Roco8	29	Protein kinase
DDB0220020	DhkM	30	Histidine kinase
DDB0185116	PSVA	32	Pre-spore vesicle protein
DDB0215384	DhkD	32	Double histidine kinase
DDB0001464	GP138C	34	Cell surface glyco-protein gp138
DDB0203479	DhkF	34	Histidine kinase
DDB0001698	GerD	35	Germination protein
DDB0001804	SP96	50	Spore coat protein

[a]Database entries were obtained by blasting the matched sequences from the EMBL database search against Dictybase (http://dictybase.org).
[b]Names of data base entries from Dictybase are given.
[c]Protein function information was taken from Dictybase.

Importantly, no mRNA was found that matched in sense any of the search patterns. A closer inspection of the 45 double-stranded regions in antisense orientation to mRNAs showed that these did not occur preferentially at specific positions within RNAs, but were randomly distributed over the length of the RNA. Again, alignment of double-stranded sequences did not result in any consensus sequence.

Out of the 45 hits, 21 represent known proteins (Table 1). Surprisingly, the vast majority of the mRNAs corresponding to these (antisense) matches encoded regulatory proteins such as kinases and only a few encoded putative structural components that are expressed in *Dictyostelium* development (Table 1). These proteins also include one of the Dicer homologues in *Dictyostelium*, drnA. The remaining unknown ORFs were analysed for conserved domains using InterProScan (Zdobnov and Apweiler, 2001). While 10 hits remained undefined, 14 putative proteins again contained domains predicted to be involved in regulation (Table 2).

The observation that no double-stranded regions longer than 24 bp could be detected in mRNAs was rather

Table 2 Predicted protein coding sequences with conserved protein domains from *Dictyostelium discoideum* that show double-stranded regions in antisense orientation of the mRNA.

Database entry[a]	Length of ds (bp)	Conserved protein domains[b]	E-value[c]
DDB0203342	25	Zinc finger/ATP binding	7e-35 (865 aa)
DDB0203430	25	TPR-like	1.8e-28 (300 aa)
DDB0205889	26	HIS2-CDC14	2e-20 (110 aa)
DDB0168394	27	SNF1/AMP-ACTIVATED kinase	1e-143 (405aa)
DDB0168507	27	Lipin_N	1.1e-49 (106aa)
DDB0167766	28	RhoGAP domain	6.4e-35 (150 aa)
DDB0168995	28	Light-mediated development DET1	1e-151 (260 aa)
			1e-132 (250 aa)
DDB0206578	28	Armadillo repeat fold	9.1e-17 (30–140 aa), 7 times
DDB0168109	29	Serine/threonine protein kinase	4.4e-101 (300 aa)
DDB0168784	29	Zinc finger	1e-44 (130 aa)
DDB0217114	29	Protein kinase domain	3.3e-24 (270 aa)
DDB0217510	30	Nucleotidyltransferase	4.2e-36 (200aa)
DDB0217750	34	Ankyrin repeat	1.2e-61 (350 aa)
		Cyclin-like F-box	1.9e-08 (45 aa)
DDB0167287	47	Murine GABA_A receptor	1e-172 (360 aa)

[a] Database entries were obtained by blasting the matched sequences from the EMBL database search against Dictybase (http://dictybase.org).

[b] Conserved protein domains or signatures were obtained by analysing amino acid sequences in InterProScan (Zdobnov and Apweiler, 2001).

[c] The E-value denotes the stringency with which the protein domain or signatures were found, and in brackets over what length of amino acid stretch.

striking. This indicated a stringent evolutionary selection against dsRNA that could otherwise elicit the RNAi response and result in mRNA degradation.

Some of the hybrid RNAs consisting of sense and antisense parts of coding genes could by no means be explained by reverse priming artefacts or template switching. Especially in the light of the recently described hybrid RNAs from the mammalian troponin gene (Bartsch et al., 2004), these may represent a new class of regulatory molecules or regulation intermediates.

Our observation that antisense transcripts of several genes have the potential to form substantial double strands could be serendipitous, especially since antisense transcription of these genes has only been shown experimentally in one case. It is, however, intriguing that none of the housekeeping genes but only genes somehow involved in cellular regulation display this feature. *Cis-* and *trans*-acting antisense RNAs have increasingly attracted attention (Lavorgna et al., 2004). Our analysis of putative intramolecular dsRNA formation should further contribute to this rapidly evolving new field of research in gene regulation.

Acknowledgments

This work was supported by grants of the Deutsche Forschungsgemeinschaft to G.S. (Ste465/4) and to W.N. (Ne285/8) and by the Zentrale Forschungförderung (ZFF) of the University of Kassel to C.H. B.E.B. was supported by an Otto-Braun-Stipend.

References

Bartsch, H., Voigtsberger, S., Baumann, G., Morano, I., and Luther, H. P. (2004). Detection of a novel sense-antisense RNA-hybrid structure by RACE experiments on endogenous troponin I antisense RNA. RNA *10*, 1215–1224.

Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Billy, E., Brondani, V., Zhang, H., Muller, U., and Filipowicz, W. (2001). Specific interference with gene expression induced by long, double-stranded RNA in mouse embryonal teratocarcinoma cell lines. Proc. Natl. Acad. Sci. USA *98*, 14428–14433.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet. *13*, 497–498.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494–498.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature *391*, 806–811.

Glockner, G., Eichinger, L., Szafranski, K., Pachebat, J. A., Bankier, A. T., Dear, P. H., Lehmann, R., Baumgart, C., Parra, G., Abril, J. F. et al. (2002). Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. Nature *418*, 79–85.

Glockner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M. A., Cox, E., Eichinger, L., Noegel, A. A., and Rosenthal, A. (2001). The complex repeats of *Dictyostelium discoideum*. Genome Res. *11*, 585–594.

Hildebrandt, M. and Nellen, W. (1992). Differential antisense transcription from the *Dictyostelium* EB4 gene locus-implications on antisense-mediated regulation of messenger RNA stability. Cell *69*, 124–204.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. Trends Biochem. Sci. *23*, 403–405.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R. et al. (2004). The EMBL Nucleotide Sequence Database. Nucleic Acids Res. *32* (database issue), D27–30.

Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C. M., and Casari, G. (2004). In search of antisense. Trends Biochem. Sci. *29*, 88–94.

Lehner, B., Williams, G., Campbell, R. D., and Sanderson, C. M. (2002). Antisense transcripts in the human genome. Trends Genet. *18*, 63–65.

Manche, L., Green, S. R., Schmedt, C., and Mathews, M. B. (1992). Interactions between double-stranded RNA regulators and the protein kinase DAI. Mol. Cell. Biol. *12*, 5238–48.

Martens, H., Novotny, J., Oberstrass, J., Steck, T. L., Postlethwait, P., and Nellen, W. (2002). RNAi in *Dictyostelium*: the role of RNA-directed RNA polymerases and double-stranded RNase. Mol. Biol. Cell *13*, 445–453.

Novotny, J., Diegel, S., Schirmacher, H., Mohrle, A., Hildebrandt, M., Oberstrass, J., and Nellen, W. (2001). *Dictyostelium* double-stranded ribonuclease. Methods Enzymol. *342*, 193–212.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H. et al. (2002). The Bioperl toolkit: perl modules for the life sciences. Genome Res. *12*, 1611–1618.

Tomari, Y., Du, T., Haley, B., Schwarz, D. S., Bennett, R., Cook, H. A., Koppetsch, B. S., Theurkauf, W. E., and Zamore, P. D. (2004). RISC assembly defects in the *Drosophila* RNAi mutant armitage. Cell *116*, 831–841.

Zamore, P. D., Tuschl, T., Sharp, P. A., and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25–33.

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Received July 5, 2004; accepted August 10, 2004