

Multiple Cross and Inbred Strain Haplotype Mapping of Complex-Trait Candidate Genes

Yeong-Gwon Park, Robert Clifford, Kenneth H. Buetow, and Kent W. Hunter¹

Laboratory of Population Genetics, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Identifying complex-trait candidate genes after initial low-resolution mapping has proven to be a difficult and labor-intensive undertaking, usually requiring years to develop and analyze congenic strains. As a result, to date, few complex-trait genes have been discovered. Recently it was suggested that SNP haplotype analysis in inbred strains might be useful for mapping of complex traits. In this study, we have combined medium-resolution haplotype mapping with multiple experimental cross-mapping experiments to reduce the number of potential candidate genes in a complex-trait candidate interval. Coincident mapping of a modifier gene in multiple experimental crosses using different inbred strains is consistent with the common inheritance of a modifier allele. A haplotype map was developed in four inbred strains of mice used in our complex-trait mapping crosses across the proximal 10 cM of proximal Chromosome 19 to identify haplotype blocks that segregate appropriately. Only ~23 out of >400 genes met this criteria. This strategy coupled with tissue and expression arrays, as well as our recently described common pathway analysis to reduce the number of high-priority candidates, may provide a rapid, efficient method to identify and prioritize complex-trait candidate genes without requiring construction of congenic mouse strains.

Identification of the genetic basis for complex traits or phenotypic modification has been an extremely arduous task to date. Although >1000 modifier loci have been mapped in the mouse genome, only a handful have been linked to specific gene polymorphisms (Korstanje and Paigen 2002). This is owing primarily to the laborious and expensive process of making interval-specific congenic mice to isolate candidate regions in a fixed genetic background, followed by generation of subcongenic animals for high-resolution mapping.

As a result, investigators are increasingly attempting to develop novel methods for high-resolution mapping of modifier genes that circumvent this process. These strategies include mapping in outbred (Nagase et al. 2001) or HS strains of mice (Mott and Flint 2002). Other investigators have suggested making F_1 crosses between the members of recombinant inbred (RI) strains, which allows repeated interrogation of a fixed genotype to reduce nongenetic variance while increasing the mapping resolution of the panel (Williams et al. 2001). Performing multiple mapping experiments using different inbred strain partners, followed by microsatellite-based haplotype mapping of the progenitor strains, has also been suggested as a method of refining the initial modifier gene mapping (Hitzemann et al. 2000).

Recently it has been suggested that *in silico* SNP haplotype analysis might be a useful strategy for mapping complex traits (Grupe et al. 2001). Although a controversial idea (Chesler et al. 2001; Darvasi 2001), it is possible that combining the developing mouse inbred strain SNP databases (Lindblad-Toh et al. 2000; Grupe et al. 2001) with multiple experimental crosses may provide an important tool for investigators to narrow the large list of potential candidate genes to a manageable, prioritized list for further analysis. For example,

coincident mapping of modifier loci to the same chromosomal region for a particular trait using different inbred strain partners is consistent with inheritance of a common allele. Examination of the haplotype structure across the candidate region might reveal regions that segregate appropriately with the phenotypic modification and therefore may harbor the causative polymorphism. If the conserved haplotype blocks are sufficiently small, this method has the potential of significantly reducing the high-priority candidate genes in the 10–20-cM regions initially defined in preliminary complex-trait analysis. The high-priority candidate list can then be further reduced by tissue or microarray experiments to identify genes expressed in tissues of interest or in appropriate biological pathways, or in pathways shared among independent QTL candidate peaks affecting the same phenotypic trait (Cozma et al. 2002). This “systems biology” approach may have the capacity to significantly accelerate modifier gene discovery, at least in some cases, by circumventing the need for generation of congenic and subcongenic animals.

RESULTS

To assess this possibility, we have constructed a medium-resolution haplotype map across the proximal 8 Mb of mouse Chromosome 19 in four strains of mice. These four strains, DBA/2J, FVB/NJ, AKR/J, and NZB/B1NJ, were used to map a metastatic efficiency modifier locus to the proximal 10 cM of the chromosome (Hunter et al. 2001). Two separate mapping experiments were used; an intercross between the AKXD recombinant inbred panel and FVB/N-TgN(MMTV-PyVT)^{634Mul}, and an [FVB/NJ × (NZB/B1NJ × FVB/N-TgN(MMTV-PyVT)^{634Mul})] backcross. Analysis of the crosses revealed that the DBA/2J and NZB/B1NJ alleles suppressed metastatic efficiency. Candidate haplotype intervals therefore would include those regions where DBA/2J and NZB/B1NJ shared a common haplotype and differed from both AKR/J and FVB/NJ.

¹Corresponding author.

E-MAIL hunterk@mail.nih.gov; FAX (301) 435-8963.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.786403>.

High-resolution haplotype analysis of human Chromosome 21 indicated that human haplotypes varied between 5 and ~100 kb in length (Patil et al. 2001). It was anticipated, owing to the derivation of most laboratory inbred strains from relatively few progenitors, that the mouse haplotype blocks would probably be larger than humans. Therefore, ~500-bp amplicons were designed approximately every 50 kb to try to identify a large fraction of the potential haplotype blocks. An ~800-kb region ~5 Mb distal (4653–5690 kb) of the centromere containing only putative repetitive elements was excluded from the SNP discovery. A second gene-poor region (6000–6678 kb) containing only 1 predicted gene was also excluded from analysis. In addition, the most telomeric 1.5 Mb of the candidate region consisting of a large number of olfactory receptors, which were considered as unlikely metastasis efficiency modifier genes, was also excluded.

Initially, 17 microsatellite markers and 84 PCR amplicons were assayed for polymorphisms in the four inbred strains. Three microsatellites (~18%) and 31 of the amplicons (~37%) were nonpolymorphic in the four inbred strains assayed. The average spacing of the polymorphisms in the assayed regions was ~115 kb. Potential haplotype blocks were defined as sets of polymorphisms that were present in two of the inbred strains.

To assess the approximate size of the haplotype blocks, selected regions were subjected to further analysis. Additional amplicons were generated and sequenced in the *Ntlk-FosL1* region to determine the length of the haplotype block. Amplicons consisted of inter- and intragenic noncoding regions as well as exons and their flanking intronic regions in the *Rela* and *HtaTip* loci. Analysis of polymorphisms in the four strains demonstrated the presence of at least one haplotype that extends ~150 kb, and possibly as many as three haplotype blocks if the *HtaTip* NZB/B1NJ polymorphism is a common polymorphism.

A significant fraction of the polymorphisms observed were present in only one of the four strains, indicating that they may be rare polymorphisms unique to that strain. This possibility was tested by screening additional strains of mice for the presence of the SNPs. Using the CIDR database (<http://www.cidr.jhmi.edu/mouse/index.html>), a phylogenetic tree was constructed to identify strains that clustered near FVB/NJ and NZB/B1NJ, and therefore might be likely to share haplotype structures. Subsets of SNPs were selected to specifically test whether a subset of the unique SNPs observed in the original four strains was present in the phylogenically related strains. As can be observed in Figure 1, all of the 8 SNPs originally identified in NZB/B1NJ were replicated, indicating that these are common SNPs representing true haplotype blocks. In addition, additional haplotype blocks were identified in NZW/LacJ and BUB/BnJ, indicating that there were at least three common founder haplotypes for this region of Chromosome 19 in the progenitors of the common laboratory inbred strains.

DISCUSSION

Unlike the NZB/B1NJ haplotypes, only half of the FVB/NJ-identified SNPs replicated in the expanded set of inbred mice. Several possibilities might explain this. First, the SNPs might be unique to FVB/NJ, having arisen after the foundation of this strain. Although we cannot exclude this possibility with the current data, the juxtaposition of the FVB/NJ SNPs indicates that this may represent a real haplotype block. Second,

although the mouse strains selected for screening were chosen based on their phylogenetic relationship, only two of the strains were Swiss-derived, as is FVB/NJ. Screening additional Swiss-derived mice might reveal that these SNPs are part of a conserved haplotype block.

Interestingly, examination of the SNP map indicates that the rate of mutation may not be identical across this region of the genome. The majority of the amplicons contained a single SNP or a single small indel. However, in a 300-kb interval of the chromosome (~1827–2128 kb), all of the amplicons show multiple polymorphisms in the strains tested. It is possible that this is simply caused by random chance and the number of strains screened at these loci. The juxtaposition of these loci and the lack of similar results at other loci screened in the larger set of animals, however, imply that the increased polymorphism rate may have a biological basis.

The analysis presented here is based on the hypothesis that the colocalization of the metastasis modifier loci in the independent crosses was due to the inheritance of alleles from common progenitors. Although our genetic data are consistent with this hypothesis, there is a major caveat. If, instead of a single locus or tightly linked group of polymorphic loci that were inherited from a common progenitor, there were two different linked QTLs in the modifying strains, one of which, for example, is the basis of the metastasis suppression in DBA/2J and the other in NZB/B1NJ, then this strategy would not be effective. However, it might be useful to reduce the number of genes to be considered as candidates. By comparing between two strains instead of all four, for example, AKR/J and DBA/2J, haplotype blocks that are shared between the strains in a QTL candidate region could be eliminated from initial consideration. For this example it would reduce the number of genes approximately twofold. The utility of this method, however, will be directly determined by the mouse strains in question. As shown in Figure 1, comparison of FVB/NJ and NZB/B1NJ would not result in a significant reduction of potential candidate genes.

Examination of the haplotype blocks also indicates that there are five regions that meet the criteria for candidate-gene consideration. These regions encompass ~25 genes, according to the Celera genome database. This is a 16-fold reduction of the number of potential candidate genes (>400) in the original candidate region, not including the >100 olfactory receptor genes excluded from consideration. Haplotype mapping may, therefore, significantly reduce the number of genes in a QTL candidate region for candidate-gene analysis, but significant numbers of genes may still need to be examined. Previously Belknap et al. (2001) suggested that combinations of techniques may be useful to provide strong evidence that a particular gene may be the genetic basis for a QTL effect. Similarly, different techniques and strategies may serve as additional filters to reduce the number of high-priority candidate genes. These techniques may include using microarray analysis to identify genes in interesting haplotype blocks that are expressed in tissues of interest or that are differentially expressed between inbred strains of interest; pathway analysis, the identification of biochemical pathways that have members associated with independent QTL peaks affecting the same trait (Cozma et al. 2002); and surveys of literature for genes lying in the haplotype blocks of interest that were previously known to affect the trait of interest. Applying these filters on top of the haplotype block mapping may reduce the number of candidate genes to a handful that could subsequently be screened for polymorphisms and subjected to in

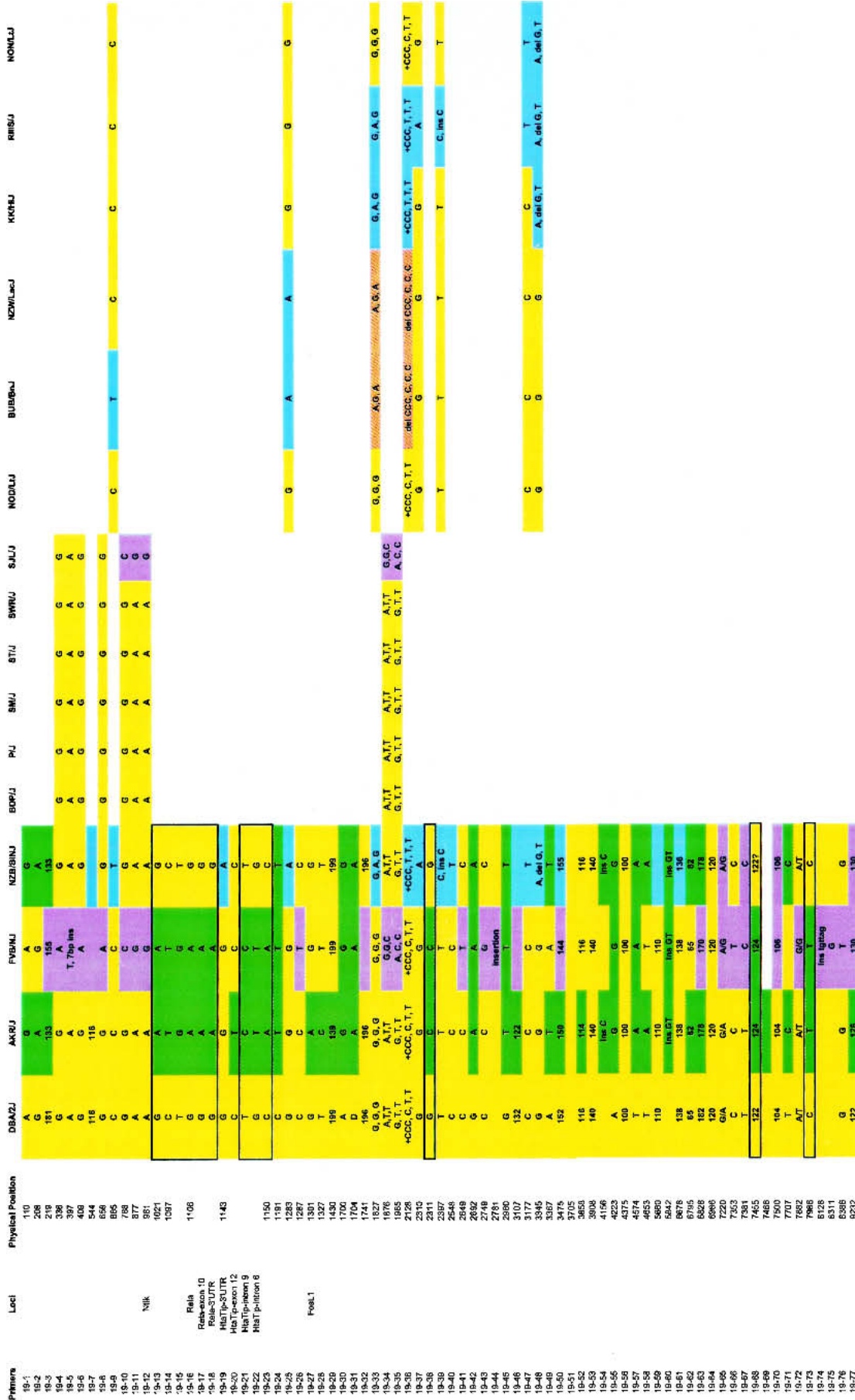


Figure 1 Haplotype map of proximal Chromosome 19. The mouse strains screened are indicated across the top of the figure. Primer pairs are listed in the left-hand column. Haplotypes are indicated by colors. The polymorphisms present in DBA/2J (yellow), AKR/J (green), FVB/NJ (purple), NZB/B1NJ (light blue), BUB/B6L (light green), BM/J (light purple), ST/J (light blue), SWR/J (light purple), SJL/J (light blue), NOD/LtJ (light green), BUB/B6L (light purple), K1K1/J (light blue), R1R1/J (light blue), and NOD/LtJ (light green) are indicated. The polymorphisms observed are shown in each cell. (Del) Deletion, (ins) insertion. When multiple polymorphisms were observed, commas separate the polymorphisms. Microsatellite-based polymorphisms are shown as PCR product size, in base pairs. For those microsatellites that were significantly different in size, as determined by agarose gel electrophoresis, the exact size was not determined, so no base pair size was included in the figure. The physical position column indicates the distance, in kilobases, from the centromere according to the Celera database. The haplotype blocks that segregate appropriately according to the coinheritance hypothesis are boxed.

vitro and in vivo analysis for their role in the phenotype of interest.

Based on this combined strategy, we have applied these filters to our haplotype data. Pathway analysis has been performed on the data obtained from the initial metastasis efficiency mapping studies (Hunter et al. 2001) using the publicly available gene lists. Three molecular pathways were observed to have members present in each of the QTL candidate regions. The Chromosome 19 candidate region contains ~20 genes of known function that are members of these molecular pathways, and thus might be considered candidates for further analysis. Combining this information with expression array data (T. Qiu, G.V.R. Chandramouli, N.W. Alkharouf, K.W. Hunter, and E.T. Liu, in prep.), as well as literature searches to identify pathways known to be associated with metastatic progression reduces the candidate genes to two. Sequence analysis of these genes has revealed a polymorphism in a conserved domain in one of the candidates that would be likely to have functional consequences (data not shown). In vitro and in vivo experiments are presently underway to evaluate this gene as the Chromosome 19 metastasis efficiency modifier gene *Mtes1*.

In conclusion, these data indicate that using multiple crosses to find shared putative modifier alleles, combined with haplotype maps, may be a useful method for candidate modifier gene identification. The ability to perform the multiple cross-haplotype analyses will clearly improve as the SNP density across large numbers of inbred strains increases. At present, because it is not clear what the exact size of a mouse haplotype block is, the density of SNPs to achieve saturation is not known. Our data are consistent with the existence of a large number of haplotype blocks of >80–100 kb. However, it is not unlikely that we may have missed some haplotype blocks, including some that fit the criteria used to select candidate genes, either from insufficient density of SNPs or by not screening putatively gene-poor regions. Nevertheless, we believe that as the SNP maps are expanded across more inbred strains, haplotype mapping in inbred strains from multiple cross experiments may prove a valuable tool for rapidly identifying and prioritizing candidate genes for more detailed analysis.

METHODS

The haplotype map was developed using a variety of resources. Based on the Celera database (<http://www.celera.com>), the approximate physical position and order of the MIT microsatellite markers relative to the gene order in the first 10 Mb of the chromosome were identified. Relative sizes of the microsatellite markers were determined by analysis on either 4% agarose gels or 6% acrylamide gels. Putative SNPs in a number of genes were identified using the CGAP GAI SNP discovery tools (<http://lpgws.nci.nih.gov:82/perl/snp2ref>), primers were designed, and potential SNPs were analyzed by direct sequencing. Additional SNPs were developed by designing primers from the genome sequence, followed by sequence analysis in the four strains of interest. For sequencing, PCR products were purified with QIAGEN PCR purification kits, and double-strand sequencing was performed with a Perkin Elmer BigDye Dye Terminator sequence kit. Analysis was performed on a Perkin Elmer 3100 Automated Fluorescent Sequencer. Sequences were compiled and analyzed with the computer software packages PHRED and PHRAP (Gordon et al. 1998) to identify SNPs.

ACKNOWLEDGMENTS

These data were generated through the use of the Celera Discovery System and Celera's associated databases.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Belknap, J., Hitzemann, R., Crabbe, J., Phillips, T., Buck, K., and Williams, R. 2001. QTL analysis and genome-wide mutagenesis in mice: Complementary genetic approaches to the dissection of complex traits. *Behav. Genet.* **31**: 5–15.
- Chesler, E.J., Rodriguez-Zas, S.L., and Mogil, J.S. 2001. In silico mapping of mouse quantitative trait loci. *Science* **294**: 2423.
- Cozma, D., Lukes, L., Rouse, J., Qiu, T.H., Liu, E.T., and Hunter, K.W. 2002. A bioinformatics-based strategy identifies c-Myc and Cdc25A as candidates for the Apmt mammary tumor latency modifiers. *Genome Res.* **12**: 969–975.
- Darvasi, A. 2001. In silico mapping of mouse quantitative trait loci. *Science* **294**: 2423.
- Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**: 195–202.
- Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J.K., Klein, R.F., Ahluwalia, M.K., Higuchi, R., and Peltz, G. 2001. In silico mapping of complex disease-related traits in mice. *Science* **292**: 1915–1918.
- Hitzemann, R., Demarest, K., Koyner, J., Cipp, L., Patel, N., Rasmussen, E., and McCaughan Jr., J. 2000. Effect of genetic cross on the detection of quantitative trait loci and a novel approach to mapping QTLs. *Pharmacol. Biochem. Behav.* **67**: 767–772.
- Hunter, K.W., Broman, K.W., Voyer, T.L., Lukes, L., Cozma, D., Debies, M.T., Rouse, J., and Welch, D.R. 2001. Predisposition to efficient mammary tumor metastatic progression is linked to the breast cancer metastasis suppressor gene *Brms1*. *Cancer Res.* **61**: 8866–8872.
- Korstanje, R. and Paigen, B. 2002. From QTL to gene: The harvest begins. *Nat. Genet.* **31**: 235–236.
- Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Lavolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., et al. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24**: 381–386.
- Mott, R. and Flint, J. 2002. Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. *Genetics* **160**: 1609–1618.
- Nagase, H., Mao, J.H., de Koning, J.P., Minami, T., and Balmain, A. 2001. Epistatic interactions between skin tumor modifier loci in interspecific (*spretus/musculus*) backcross mice. *Cancer Res.* **61**: 1305–1308.
- Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- Williams, R., Threadgill, D., Airey, D., Gu, J., and Lu, L. 2001. RIX mapping: A demonstration using CXB RIX hybrids to map QTLs modulating brain weight in mice. *Soc. Neurosci. Abst.* **27**: program 587. 3.

WEB SITE REFERENCES

- <http://lpgws.nci.nih.gov:82/perl/snp2ref>; CG AP GAI SNP discovery tools.
- <http://www.celera.com>; Celera database.
- <http://www.cidr.jhmi.edu/mouse/index.html>; CIDR database.

Received September 9, 2002; accepted in revised form October 30, 2002.