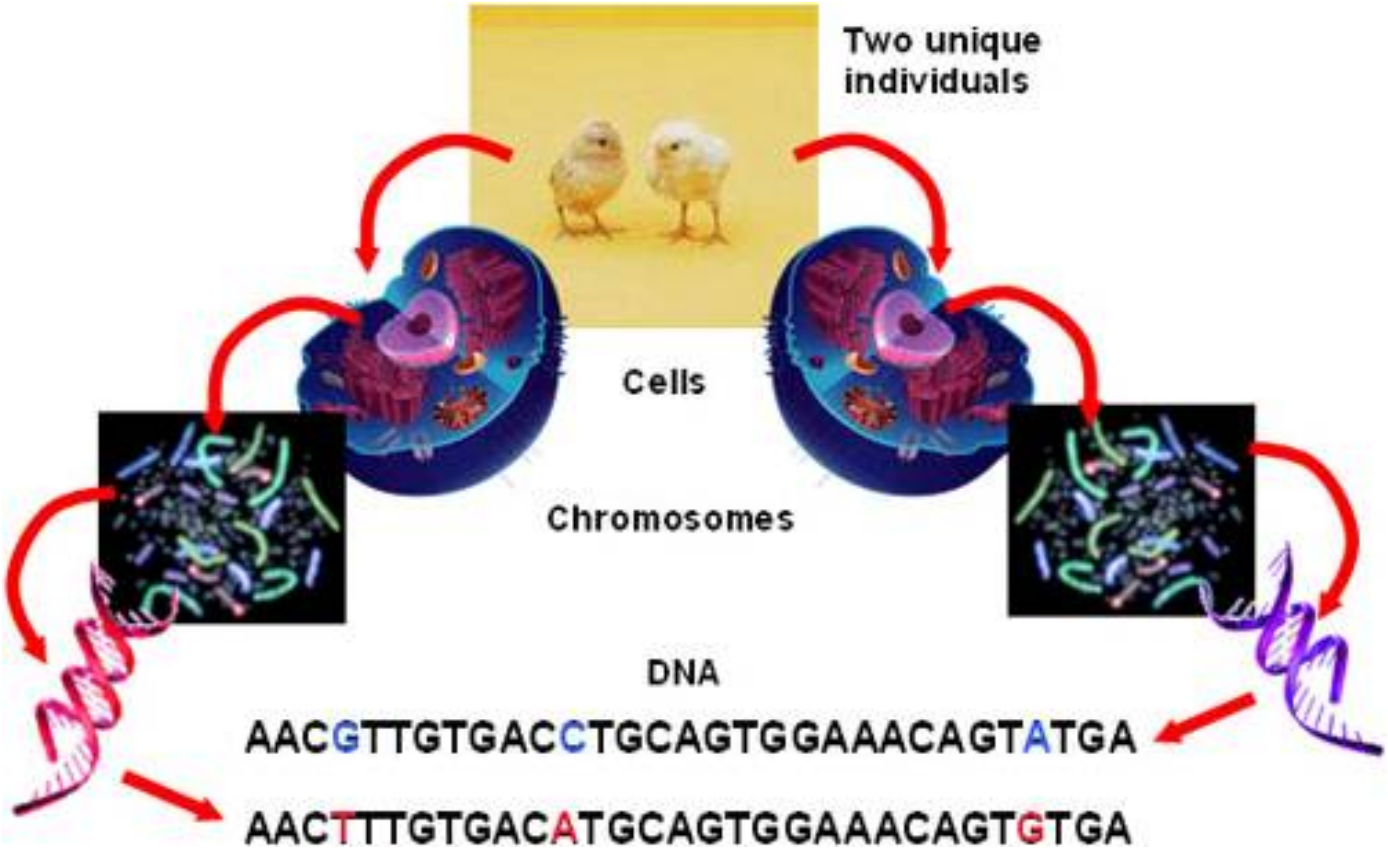
Efficient Haplotype Inference with Answer Set Programming

Esra Erdem and Ferhan Türe
Sabancı University

Haplotype Inference

- Single Nucleotide Polymorphism (SNP) is a DNA sequence variation at the single nucleotide level between one individual and another.



Haplotype Inference

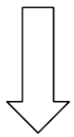
- Single Nucleotide Polymorphism (SNP) is a DNA sequence variation at the single nucleotide level between one individual and another.
- SNPs can be used
  - to diagnose diseases earlier,
  - to discover patterns of inheritance that affect health, and
  - to avoid giving drugs to patients likely to have side effects.

Haplotype Inference

- Single Nucleotide Polymorphism (SNP) is a DNA sequence variation at the single nucleotide level between one individual and another.
- SNPs can be used
  - to diagnose diseases earlier,
  - to discover patterns of inheritance that affect health, and
  - to avoid giving drugs to patients likely to have side effects.
- Haplotype inference is to infer haplotypes (SNPs of maternal and paternal chromosomes), from genotypes (mixed SNP data).

Haplotype Inference

Chromosome, paternal: ataggtccCtatttccaggcgcCgtatacttcgacgggActata
Chromosome, maternal: ataggtccGtatttccaggcgcCgtatacttcgacgggTctata



Genotype → G/C C/C A/T

Haplotype Inference

Genotype →	G/C	C/C	A/T
Haplotype 1 →	C	C	A
Haplotype 2 →	G	C	T
or			
Haplotype 1 →	G	C	A
Haplotype 2 →	C	C	T

Haplotype Inference by Pure Parsimony (HIPP)

- HIPP: Given a set of genotypes, find the minimal set of haplotypes that explain each of the genotypes.
- The decision version of HIPP (i.e., deciding that a set of k haplotypes that explain the given genotypes exists) is NP-hard (Gusfield, 2003).

Answer Set Programming (ASP)

- Theoretical basis: answer set semantics (Gelfond & Lifschitz, 1988)
- A method to solve combinatorial search problems (Marek & Truszczynski, 1999; Niemelae, 1999; Lifschitz, 1999)
- Systems for computing answer sets:
  - Smodels (Helsinki University of Technology, 1996)
  - Dlv (Vienna University of Technology, 1997)
  - Cmodels (University of Texas at Austin, 2002)
  - Pbmodels (University of Kentucky, 2005)
  - Clasp (University of Potsdam, 2006)

HIPP

Mathematical Description

- Haplotype: vector of sites, each site 0 or 1
  Genotype: vector of sites, each site 0,1, or 2

- Two haplotypes h_1 and h_2 *explain* a genotype g if for every site j the following hold:

$$\textit{ambiguous site} \begin{cases} \text{if } g[j] = 2 \text{ then } h_1[j] = 0 \text{ and } h_2[j] = 1 \\ \qquad\qquad\qquad\quad \text{or } h_1[j] = 1 \text{ and } h_2[j] = 0 \end{cases}$$

$$\textit{resolved site} \begin{cases} \text{if } g[j] = 1 \text{ then } h_1[j] = 1 \text{ and } h_2[j] = 1 \\ \text{if } g[j] = 0 \text{ then } h_1[j] = 0 \text{ and } h_2[j] = 0 \end{cases}$$

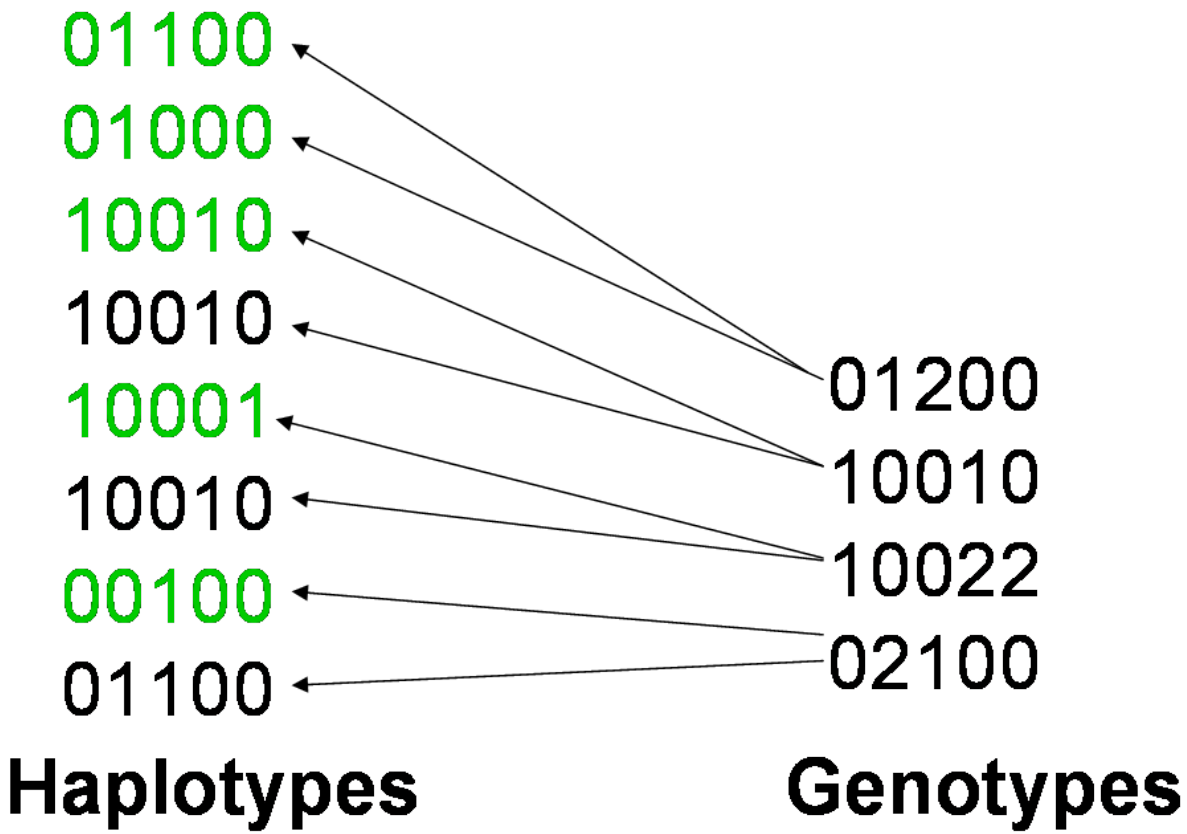
HIPP-DEC: Decision version of HIPP

Given a set G of n genotypes, and a positive integer k, decide whether there is a set H of at most k unique haplotypes such that each genotype in G is explained by two haplotypes in H.

HIPP-DEC

An Example

k=5

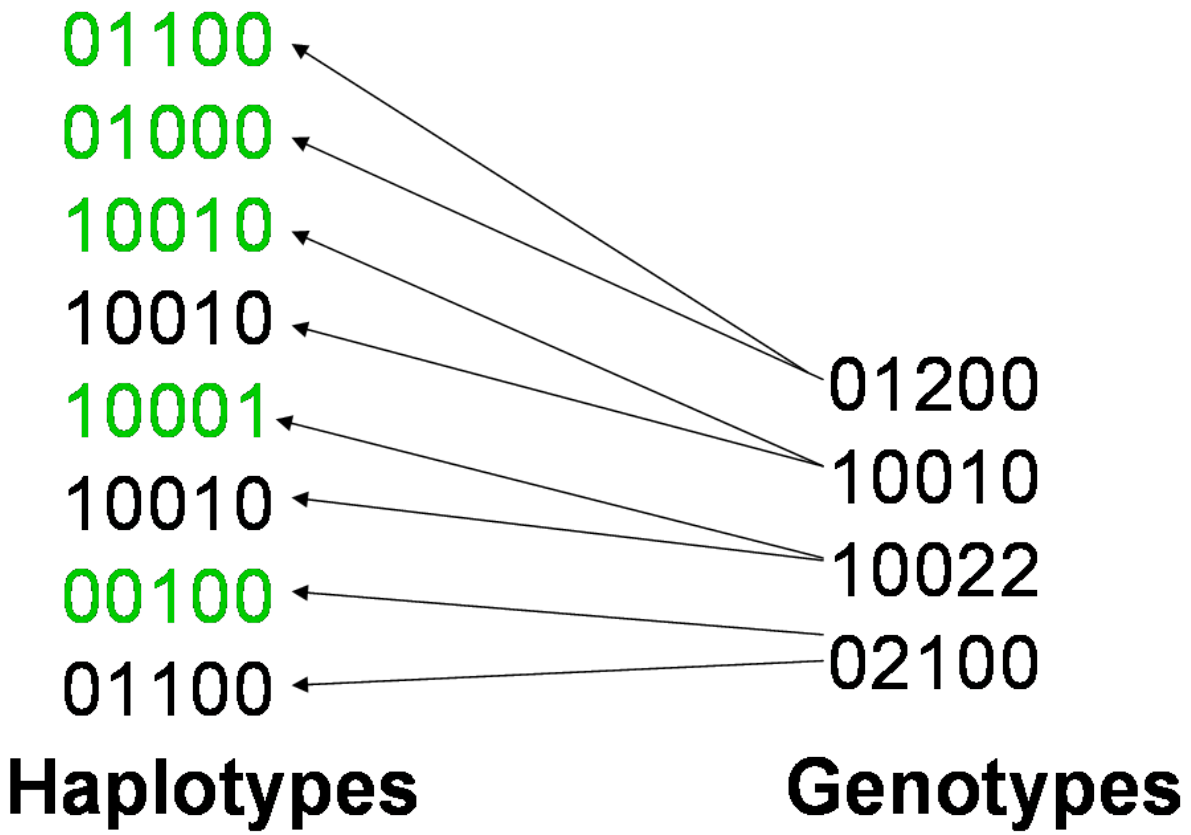


HIPP-DEC

An Example

k=5

H={01100,01000,10010,10001,00100}
G={01200,10010,10022,02100}



Solving HIPP in ASP

Representing Genotypes and Haplotypes

- Genotypes are described by atoms of the form $amb(g, j)$
  - $amb(g, j) \in X$ iff $g[j] = 2$
  - $\neg amb(g, j) \in X$ iff $g[j] = 1$
  - $amb(g, j), \neg amb(g, j) \notin X$ iff $g[j] = 0$
- Haplotypes are described by atoms of the form $h(i, j)$
  - $h(i, j) \in X$ iff $i[j] = 1$
  - $h(i, j) \notin X$ iff $i[j] = 0$

Representing Genotypes and Haplotypes

Genotype 3

```
10022
```

is represented by the atoms:

```
-amb(1,1). amb(1,4). amb(1,5).
```

01200
10010
10022
02100
G

Haplotype 1

```
01100
```

is represented by the atoms:

```
h(1,2). h(1,3).
```

01100
01000
10010
10011
10000
00100
H

HIPP-DEC

Assumptions and Constraints

A1 H is a set that contains $2n$ haplotypes, $h_1, \ldots, h_{2n}$, and

A2 every genotype g_i in G is explained by two haplotypes, h_{2i} and $h_{2i\text{-}1}$, in H.

C1 For every genotype g in G, for every ambiguous site j of g, the values of the j'th sites of these haplotypes are different.

C2 For every genotype g in G, for every resolved site j of g, the values of the j'th site of these haplotypes are $g[j]$.

C3 There are at most k unique haplotypes in H.

HIPP-DEC

Formulation

For every haplotype *H* and for every site *j*, a value is generated

```
{h(H,J)} :- haplo(H),site(J).
```

HIPP-DEC

Formulation

C1 For every genotype *g* in *G*, for every ambiguous site *j* of *g*, the values of the *j*'th sites of these haplotypes are different.

```
:- amb(G,J), h(2*G,J), h(2*G-1,J).
:- amb(G,J), not h(2*G-1,J), not h(2*G,J).
```

HIPP-DEC

Formulation

C2 For every genotype *g* in *G*, for every resolved site *j* of *g*, the values of the *j*'th site of these haplotypes are equal to *g*[*j*].

```
:- -amb(G,J), not h(2*G-1,J).
:- -amb(G,J), not h(2*G,J).
:- not -amb(G,J), not amb(G,J), h(2*G-1,J).
:- not -amb(G,J), not amb(G,J), h(2*G,J).
```

HIPP-DEC

Formulation

C3 There are at most k unique haplotypes in H.

HIPP-DEC

Formulation

C3 There are at most k unique haplotypes in H.

Different haplotypes:

```
diffhapp(H1,H2) :- 1{h(H1,J), h(H2,J)}1,
        haplo(H1;H2), H1<H2.
```

Unique haplotypes:

```
unique(1).
unique(H) :- H-1{diffhapp(H1,H):haplo(H1)},
        haplo(H), H>1.
```

HIPP-DEC

Formulation

C3 There are at most k unique haplotypes in H.

Different haplotypes:

```
diffhapp(H1,H2) :- 1{h(H1,J), h(H2,J)}1,
        haplo(H1;H2), H1<H2.
```

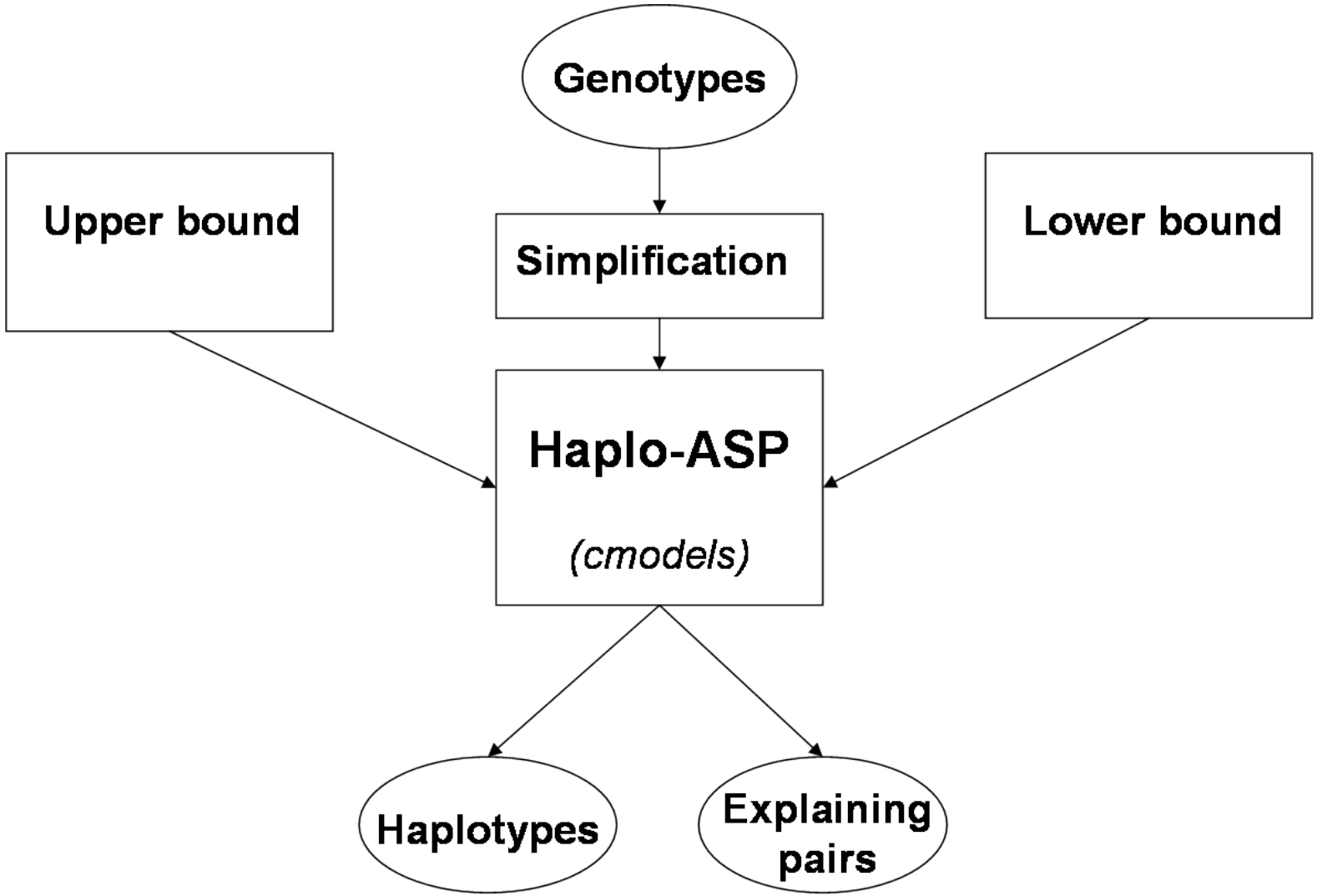
Unique haplotypes:

```
unique(1).
unique(H) :- H-1{diffhapp(H1,H):haplo(H1)},
        haplo(H), H>1.
```

C3:

```
:- k+1 {unique(H):haplo(H)}.
```

Haplo-ASP


Genotypes
Upper bound
Simplification
Lower bound
Haplo-ASP
(cmodels)
Haplotypes
Explaining pairs

Experimental Results

- HIPP instances: 334 instances (294 real, 40 artificial)
  - abcd: 90
  - ace: 90 (Angiotensin Converting Enzyme)
  - ibd: 90 (Inflammatory Bowel Disease)
  - hapmap: 24 (HapMap Project)
  - uniform: 20
  - nonuniform: 20
- Systems: Haplo-ASP, SHIPs, RPoly, Hapar
- Results

Experimental Results

- HIPP instances: 334 instances (294 real, 40 artificial)
- Systems: Haplo-ASP, SHIPs, RPoly, Hapar
  - SHIPs
    based on SAT (Lynce&Marques-Silva, 2006)
  - RPoly
    based on Pseudo-Boolean Optimization (Graca et al, 2007)
  - Hapar
    based on a branch & bound algorithm (Wang&Xu, 2003)
- Results

Experimental Results

- HIPP instances: 334 instances (294 real, 40 artificial)
- Systems: Haplo-ASP, SHIPs, RPoly, Hapar
- Results
  - Haplo-ASP solved more number of problems compared to SHIPs and RPoly.
  - RPoly is faster for many problems.

Experimental Results

Group	# of problems	# of problems solved		
		SHIPs	Haplo-ASP	RPoly
hapmap	24	24	23	23
abcd	90	90	90	90
ace	90	90	90	90
ibd	90	78	89	88
uniform	20	20	20	20
non-uniform	20	20	20	20
Total	334	322	332	331

Variations of HIPP

- Domain specific knowledge (e.g., observed patterns)

  site 2 of each haplotype is 1:

```
:- not h(H,2), haplo(H).
```

Variations of HIPP

- Domain specific knowledge (e.g., observed patterns)

  site 2 of each haplotype is 1:

  ```
  :- not h(H,2), haplo(H).
  ```

- Haplotype Inference with Present-Absent Genotype data (HIPAG)
  - Tested on 17 Killer cell Immunoglobulin-like Receptor (KIR) genes for Caucasian population
  - Haplo-ASP: exact solution, 76.8% accuracy
  - Haplo-IHP: approximation, 73.2% accuracy

Conclusion

- A novel formulation of HIPP and its variations
  - Genotypes with missing information
  - Present-absent genotype data
  - Haplotype patterns for some gene families
- New methods for haplotype inference problems
  - Lower/upper bound computation
  - Simplification of genotypes
  - Accuracy check
- Haplo-ASP: the only system that can solve HIPP and its variations
  - Solves more number of problems compared to SHIPs and RPoly.
  - Finds more accurate results compared to Haplo-IHP.