ACCURACY OF FOUR HEURISTICS FOR THE FULL SIBSHIP RECONSTRUCTION PROBLEM IN THE PRESENCE OF GENOTYPE ERRORS

Dmitry Konovalov
School of Information Technology, James Cook University, Townsville, QLD 4811, Australia

www.kingroup.org

The full sibship reconstruction (FSR) problem

$$X_i = \left((x_{i1}, x'_{i1}), (x_{i2}, x'_{i2}), \ldots, (x_{iL}, x'_{iL}) \right)$$

- Each locus is described by an unordered pair of alleles and *L* is the total number of loci which are assumed to be unlinked.
- Microsatellite markers

Microsatellites

- “Microsatellites, or Simple Sequence Repeats (SSRs) are short stretches of repeated DNA, found in most genomes, which show exceptional variability in humans and most other species.
  - This variability has made SSRs the genetic marker of choice for the vast majority of applications, including the analysis of genetic structure and parentage testing, and for investigating evolutionary links between species and populations…” *http://www.bio.bris.ac.uk/conted/microsatellite%20markers.htm*

“Microsatellites Revolution”

- “Most biologists know about the **revolution** occurring in conservation biology, molecular ecology and population genetics as a result of the widespread adoption of microsatellites in population studies…” Luikart & England (1999)

The FSR problem

- Missing parental genotypes
- Large search space
- Genotyping errors

Partition space size

- X={(a/b),(1/3),(A4/A2),…}
- $\{X_1\},\{X_2\},\{X_3\}$
  $\{X_1,X_2\},\{X_3\}$
  $\{X_1\},\{X_2,X_3\}$
  $\{X_1,X_3\}\{X_2\}$
  $\{X_1,X_2,X_3\}$
- 115975 possible partitions for 10 genotypes

An FSR algorithm

- Scoring function
  - How to compare different partitions.
- Search method
  - How to search the partition space.
- Exact solution has not been found

Accuracy of an FSR algorithm

- Generate known partition *A*
  - e.g. 5x10 – 5 families of 10 full-siblings each
- Reconstructed partition *B*
  - e.g. (6,4,5,5,5)
- Compare *A* to *B*

Accuracy-error

- the percentage of **incorrectly** assigned individuals, e.g.
  - D(5x10, {6,4,5,5,5})=1
  - error=1/50=2%
- Accuracy is the percentage of **correctly** assigned individuals, e.g.
  - accuracy=1-error=100-2%=98%

Genotyping (typing) error

- mutation
- human error
- Polymerase Chain Reaction (PCR) misprinting

Error model

- *n* – genotypes, *L* – loci
- *e* – locus error rate
  - e/2 – allele error rate
- *m=enL* – number of “mutated” alleles
  - e.g. n=100, e=2%, L=12, m=24
  - *"A quarter of genotypes with error”* (Hoffman & Amos 2005)

SIMPSON algorithm

$$S = \frac{1}{n(n-1)} \sum_{k=1}^{r} g_k (g_k - 1) = -\frac{1}{(n-1)} + \frac{1}{n(n-1)} \sum_{k=1}^{r} g_k^2$$

- Scoring function
  - SIMPSON index
  - Mendelian exclusion principle
- Search method
  - random walk

Modified SIMPSON (MS) algorithm

- Scoring function: SIMPSON
- Search method
  - Ascending order of genotype distances [error in the paper: should be for each individual]
  - Build all possible groups and keep *w* (window size) of them
  - Known complexity: $O(n^3)$ (Konovalov *et al.* 2005)

sample genotypes

- [0], grp=0, 3/0, 4/1, 0/1, 0/1, 2/3, 4/3, 0/1, 4/3, 1/1, 4/2
- [1], grp=1, 3/1, 0/1, 0/3, 4/3, 4/0, 4/4, 1/1, 1/4, 4/1, 3/2
- [2], grp=1, 3/1, 1/1, 0/1, 3/2, 4/0, 2/0, 2/1, 1/4, 4/1, 2/2
- [3], grp=1, 3/4, 1/2, 0/1, 4/3, 3/0, 4/0, 2/1, 1/4, 3/1, 2/2
- [4], grp=0, 1/1, 3/4, 0/1, 0/1, 1/3, 4/0, 3/1, 4/3, 0/4, 4/4
- [5], grp=0, 1/1, 3/4, 0/1, 0/1, 2/3, 1/3, 3/3, 4/3, 0/1, 4/2

Distance matrix

- [0][...]={0.0, 1.2, 1.2, 1.0, 0.9, 0.7}
- [1][...]={1.2, 0.0, 0.7, 0.8, 1.4, 1.5}
- [2][...]={1.2, 0.7, 0.0, 0.6, 1.3, 1.4}
- [3][...]={1.0, 0.8, 0.6, 0.0, 1.3, 1.4}
- [4][...]={0.9, 1.4, 1.3, 1.3, 0.0, 0.6}
- [5][...]={0.7, 1.5, 1.4, 1.4, 0.6, 0.0}
- AvrDist of lower triangle=1.067

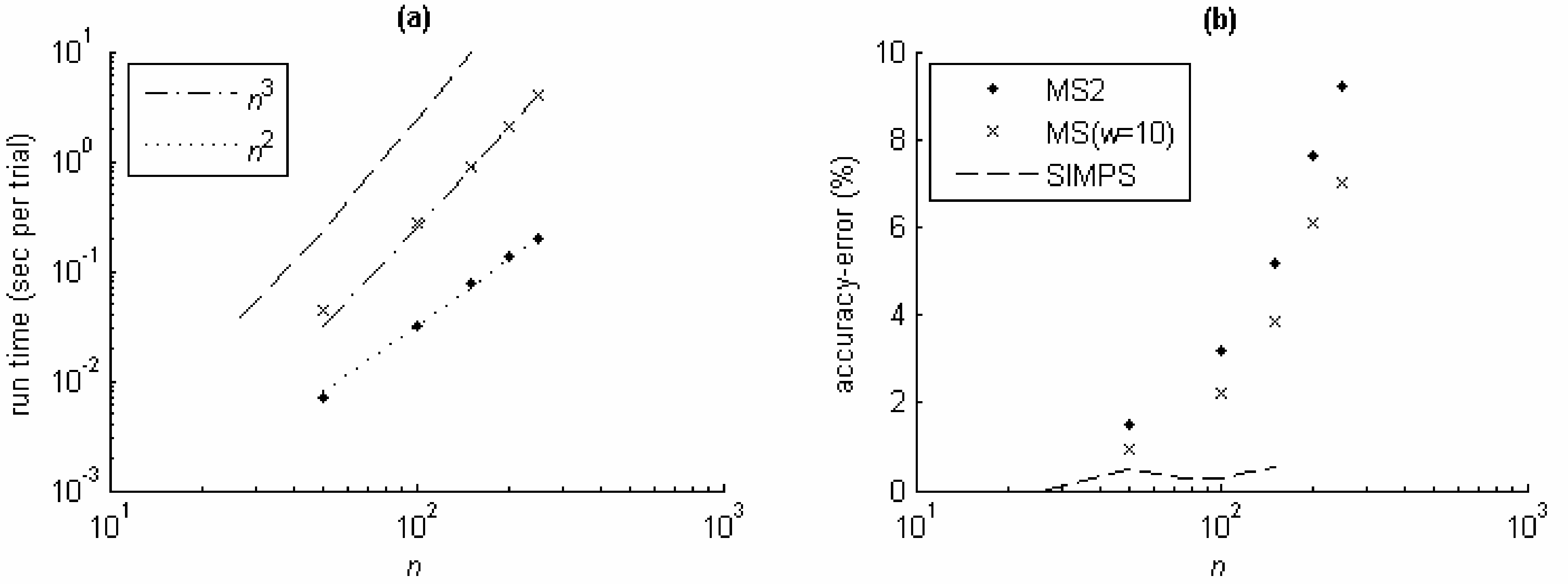
Sorted distance matrix

- AvrDist=1.067
- (3, 2, 0.6)(1, 2, 0.7)(0, 2, 1.2)(4, 2, 1.3)(5, 2, 1.4)
- (2, 3, 0.6)(1, 3, 0.8)(0, 3, 1.0)(4, 3, 1.3)(5, 3, 1.4)
- (5, 4, 0.6)(0, 4, 0.9)(2, 4, 1.3)(3, 4, 1.3)(1, 4, 1.4)
- (4, 5, 0.6)(0, 5, 0.7)(2, 5, 1.4)(3, 5, 1.4)(1, 5, 1.5)
- (5, 0, 0.7)(4, 0, 0.9)(3, 0, 1.0)(1, 0, 1.2)(2, 0, 1.2)
- (2, 1, 0.7)(3, 1, 0.8)(0, 1, 1.2)(4, 1, 1.4)(5, 1, 1.5)
  - sorted within each row
  - then rows are sorted by lowest distance
  - access order={1, 1, 1} {0, 0, 0}(group ids shown)

Modified SIMPSON (MS2) algorithm

- Scoring function: SIMPSON
- Search method
  - **MS's** access order
  - Add to the largest valid sib group
- The fastest known FSR algorithm: $O(n^2)$

Figure 1. *O(n²) vs. O(n³)*



on 3GHz PC

r - groups with 5 full-siblings each: *n=5r*

SIMPSON (T=n^3)

Descending Ratio algorithm

- Scoring function
  - Goodnight & Queller (1999) pairwise likelihoods
  - Primary hypothesis inside a group
  - null hypothesis between groups
- Search method
  - Descending order of likelihood ratio
  - Build all, keep the best

SIMPSON-assisted Descending Ratio (SDR)

- SDR
  - MS’s access order
  - Finish with DR
- “better” DR for full-sibling groups
- Issues:
  - presence of parents, half-siblings, cousins

Figure 2. No genotyping errors

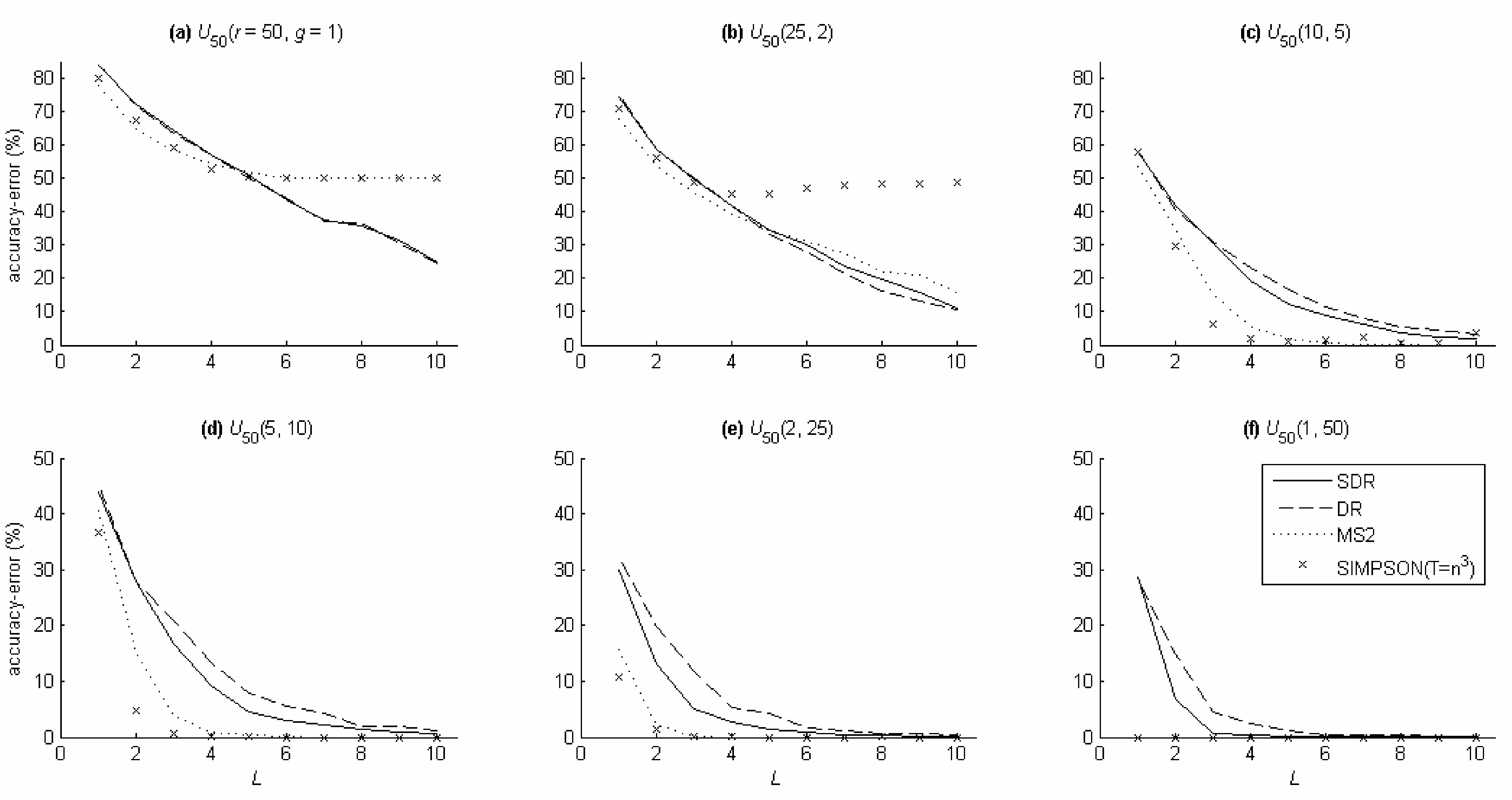


Figure 3. With 2% locus errors

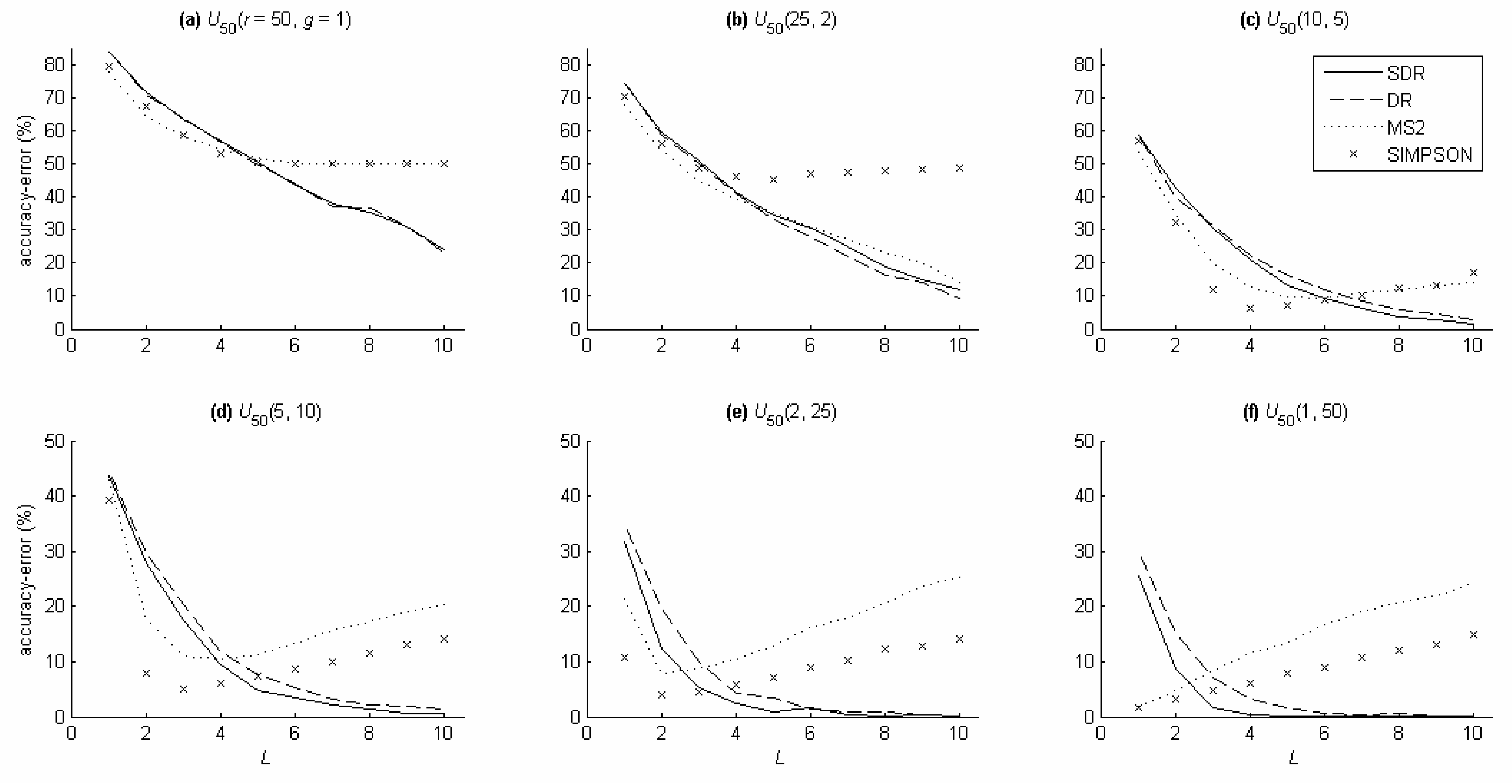
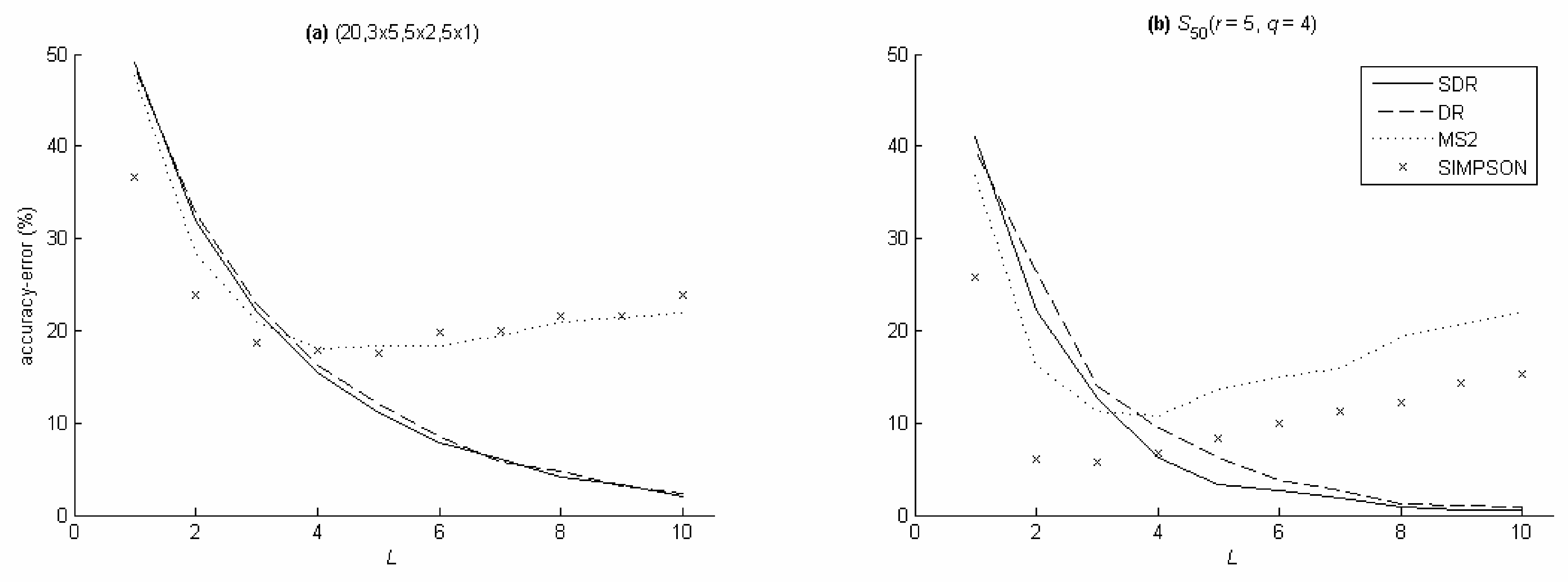


Figure 4. Non-uniform



Future direction

- Biological applications:
  - Partition *A* is **unknown**
- Reconstructed partition *B*
- Confidence level of *B* ?
  - Useful for biologists
  - Speed is essential to allow bootstrap, etc.

Unanswered questions

- Why DR is less accurate than Mendelian exclusion? Pairwise or overall likelihoods, or both?
- Is there an exact p-time solution for the Mendelian exclusion formulation of the FSR problem?
- How to calculate confidence levels?
- What is the best currently available FSR algorithm?
- Presence of half-siblings, parents, cousins, etc.