

Déjà vu — A study of duplicate citations in Medline

Data and text mining

BIOINFORMATICS


Vol. 24 no. 2 2008, pages 243–249

ebasünnis käitumine teaduses

(scientific misconduct)

- The Office of Science and Technology Policy defines research misconduct as *'fabrication, falsification or plagiarism in proposing, performing or reviewing research, or in reporting research results'*.
- The National Library of Medicine (NLM) defines a duplicate publication as one that *'substantially duplicates another article without acknowledgement'*.
(<http://www.nlm.nih.gov/pubs/factsheets/errata.html>)
- plagiarism and repeated publication of the same data
- waste of time and energy for authors, reviewers and readers

Duplikaat publikatsioonid

- NLM annotated 607 records in Medline with the publication type 'Duplicate Publication'
 - 409 included abstracts, enabling us to classify 171 (42%) as true duplicate publications.
 - The remainder were errata, updates or comments
-  Martinson et al. 2005 studied 3234 NIH funded research
 - 1.4% of the respondents admitted to plagiarism
 - 4.7% to multiple publications of the same data.
- Schein and Paladugu 2001 noted that, 'Almost 1 in every 6 original articles published in leading surgical journals represents some form of redundancy'

Andmed, programmid ja tulemused

- Medline'i andmebaasi
 - pealkiri ja abstract
 - Täisartikleid
- text-similarity search tool eTBLAST
<http://invention.swmed.edu/etblast/index.shtml>
- web-accessible database Déjà vu, at
<http://spore.swmed.edu/dejavu>

eTBLAST

- web tool for searching from electronic literature databases such as Medline (Lewis et al., 2006)
- Each query is formed by a title and an abstract, from which eTBLAST removes the stopwords
- computes a quantitative similarity score
- this score has no upper bound
- The citation with the highest similarity score is always the self identity and is referred to as Rank 1

<http://invention.swmed.edu/etblast/index.shtml>

Training and experimental data sets

Four non-overlapping sets of queries were prepared:

- a benchmarking dataset from the **171 known** and **visually-verified** Medline duplicate pairs
- a set of **5313 randomly-selected** Medline citations, all of which included both a title and abstract
- **twelve sets of 5000** Medline records, **60 000** total, that included both titles and abstracts, selected randomly from each of the last 12 years
- a set of **5465** Medline records that also have **full text** available in PubMed Central

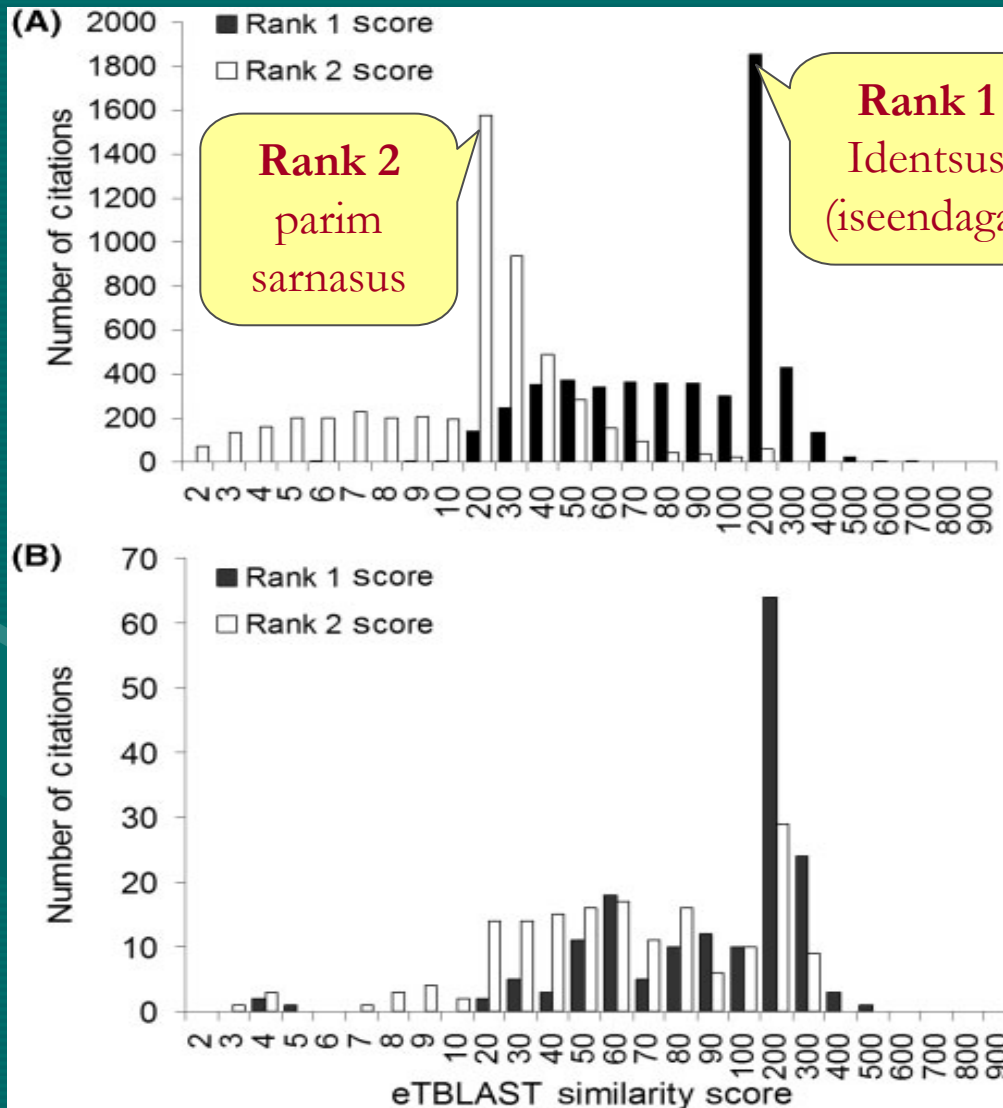
Manual classification of highly similar citations

Each highly similar duplicate pair identified by eTBLAST was manually verified by at least two authors of this study and classified the putative duplicates into :

- Duplicate/Different Authors (DA),
- Duplicate/Same Authors (SA),
- Duplicate/Update/Same Journal (SJ),
- Duplicate/Update/Different Journal (DJ),
- Duplicate Medline Issue (MI),
- Duplicate/Other, errata, false positive or no abstract

In the course of this study we manually read and classified nearly **5000** citations and approximately **250** of their associated electronically available full text articles that had been categorized as highly similar by eTBLAST

Histograms of the frequency distributions of the Rank 1 and Rank 2 scores

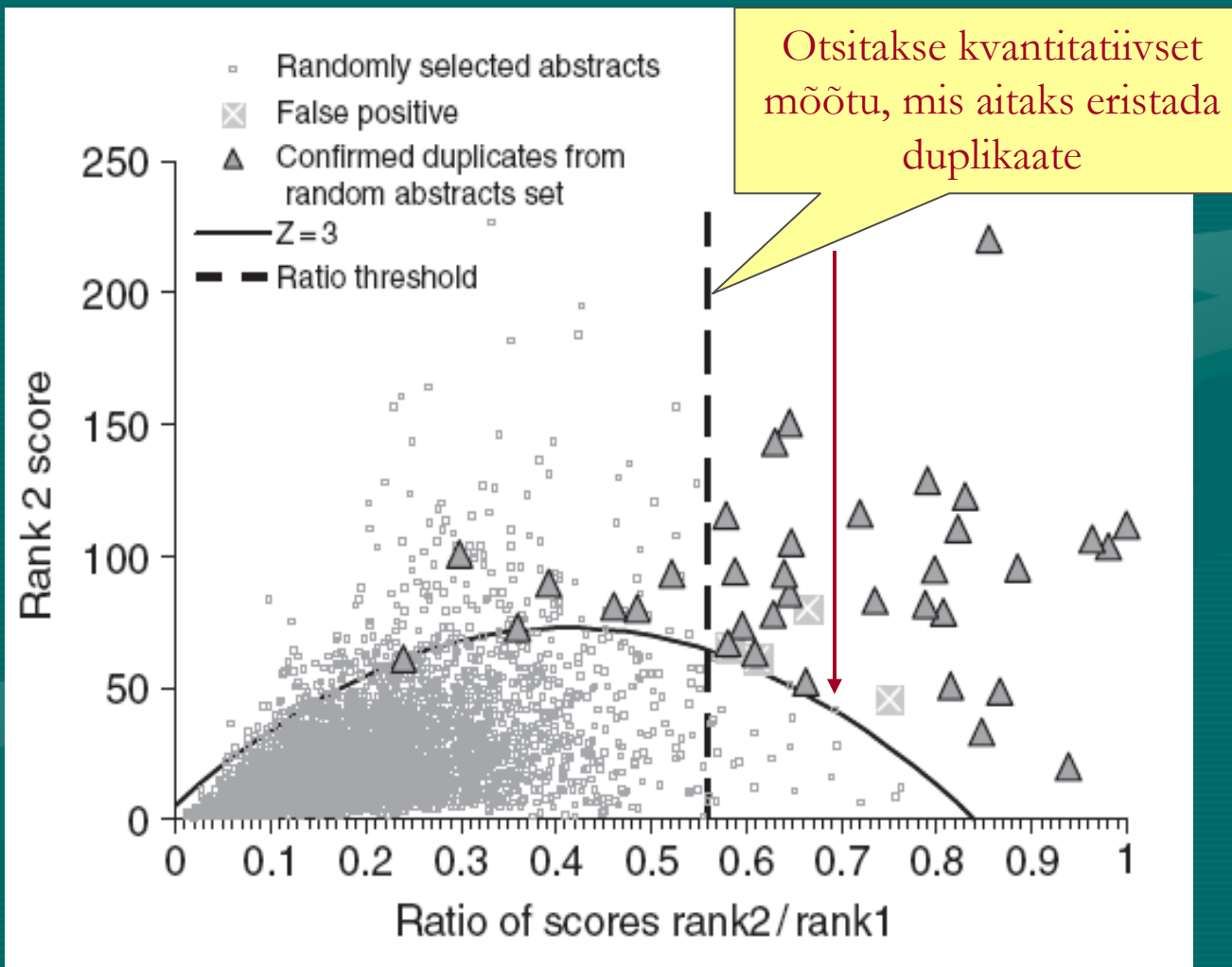


- 1A 5313 random Medline citations

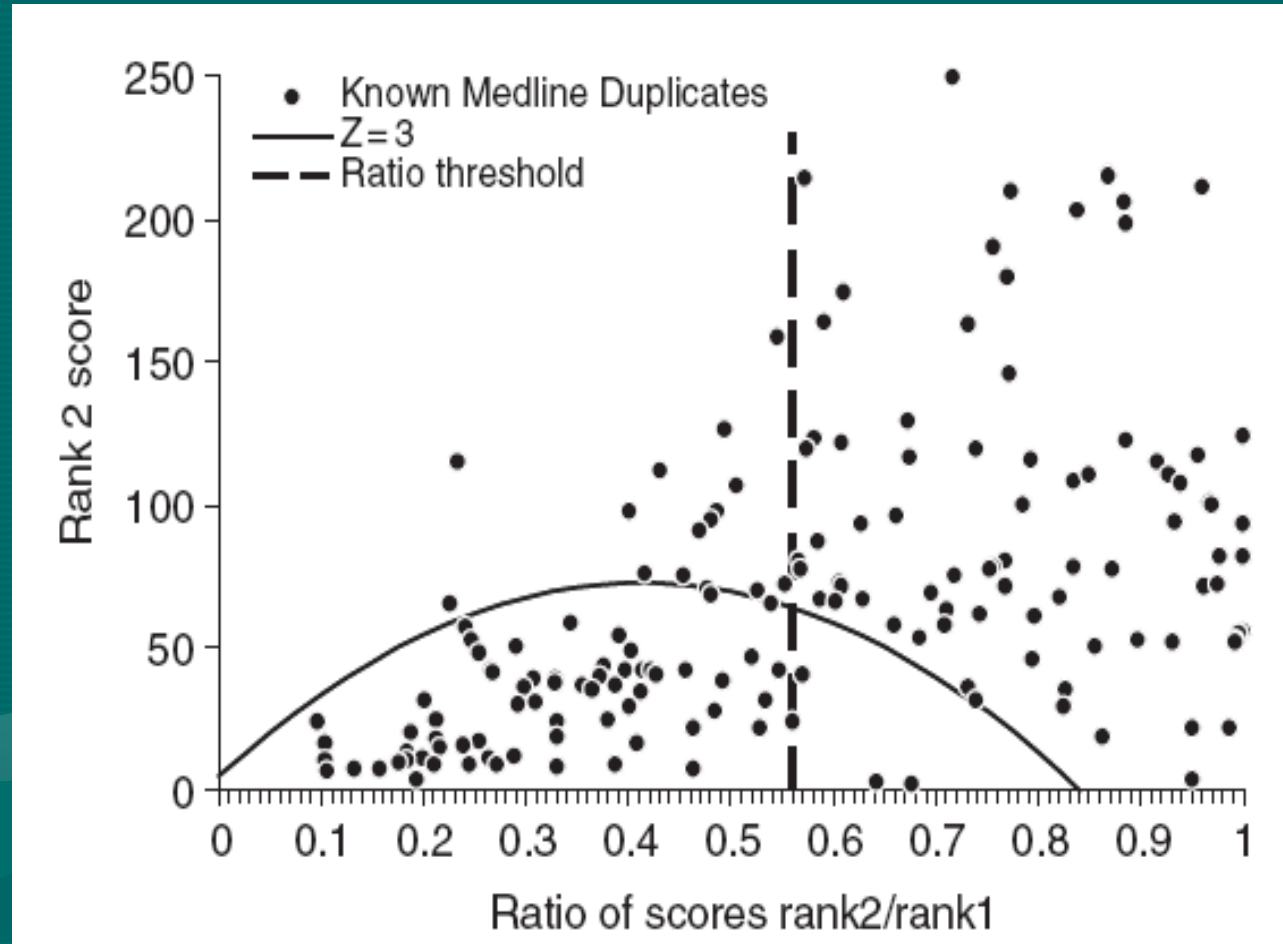
- 1B for the 171 known duplicates

This figure suggests that the Rank 2/Rank 1 score ratio may distinguish duplicate and non-duplicate pairs.

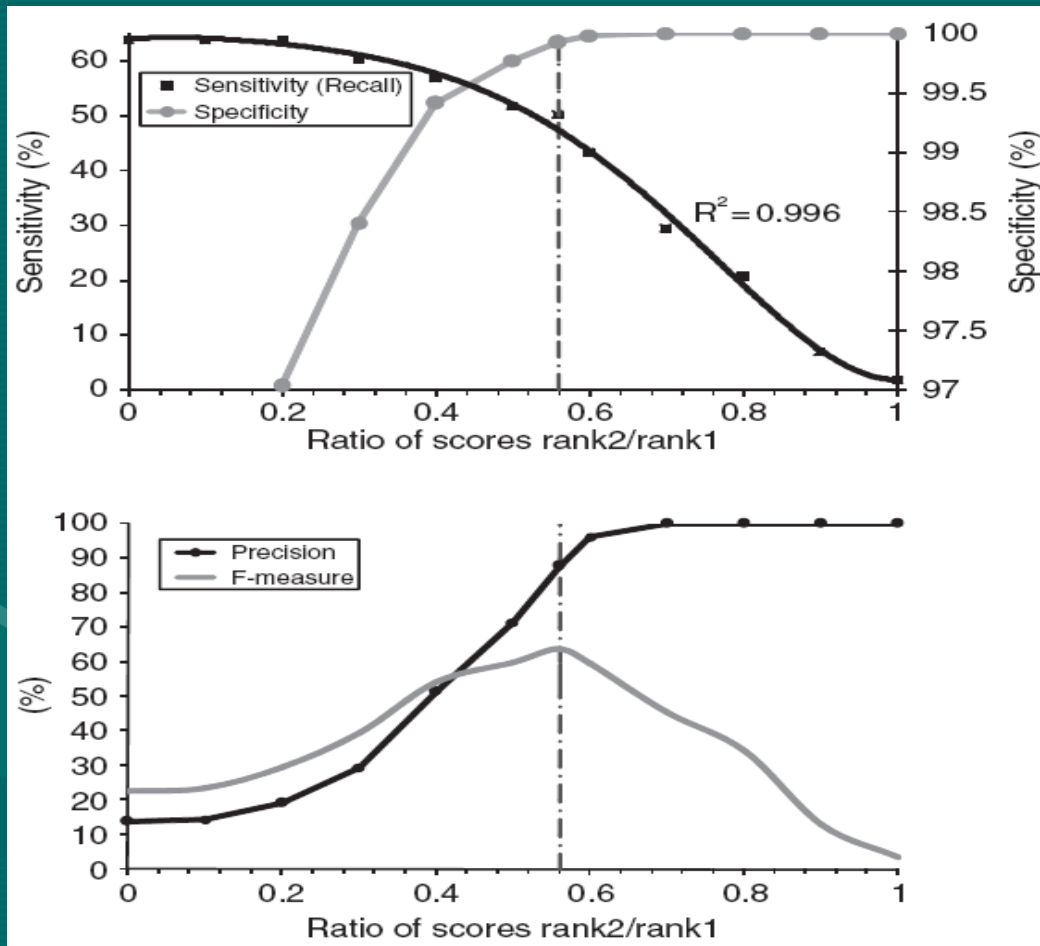
The results of searching Medline with 5313 random citations as queries



The 171 citations in Medline with a Publication Type 'Duplicate Publication' after removing errata



Determine thresholds



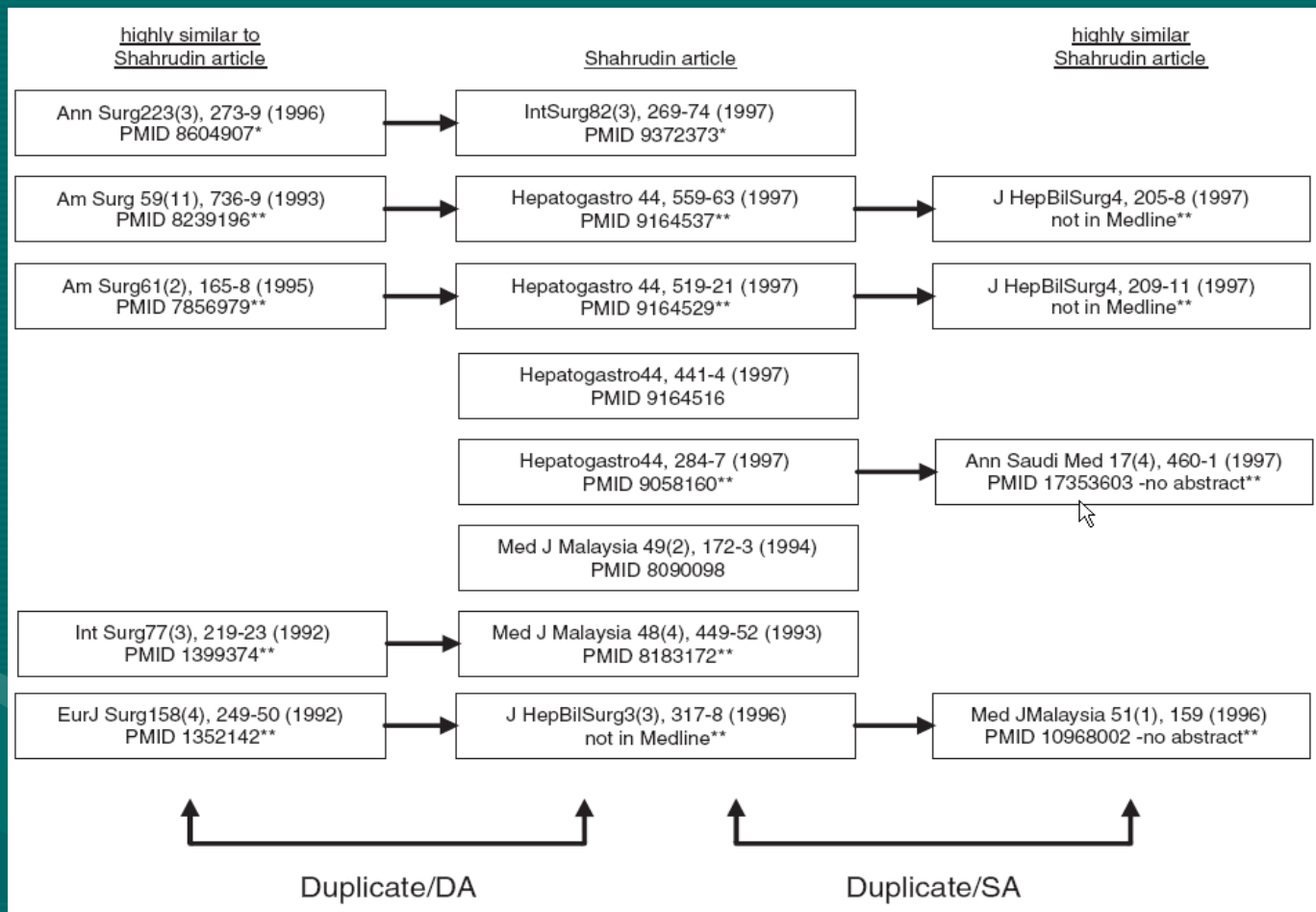
The Rank 2/Rank 1 score ratio threshold was determined from inspecting the sensitivity and specificity curves. A ratio of 0.56 corresponds to the highest **F-measure** as **the best compromise between precision and recall**

Sensitivity & selectivity

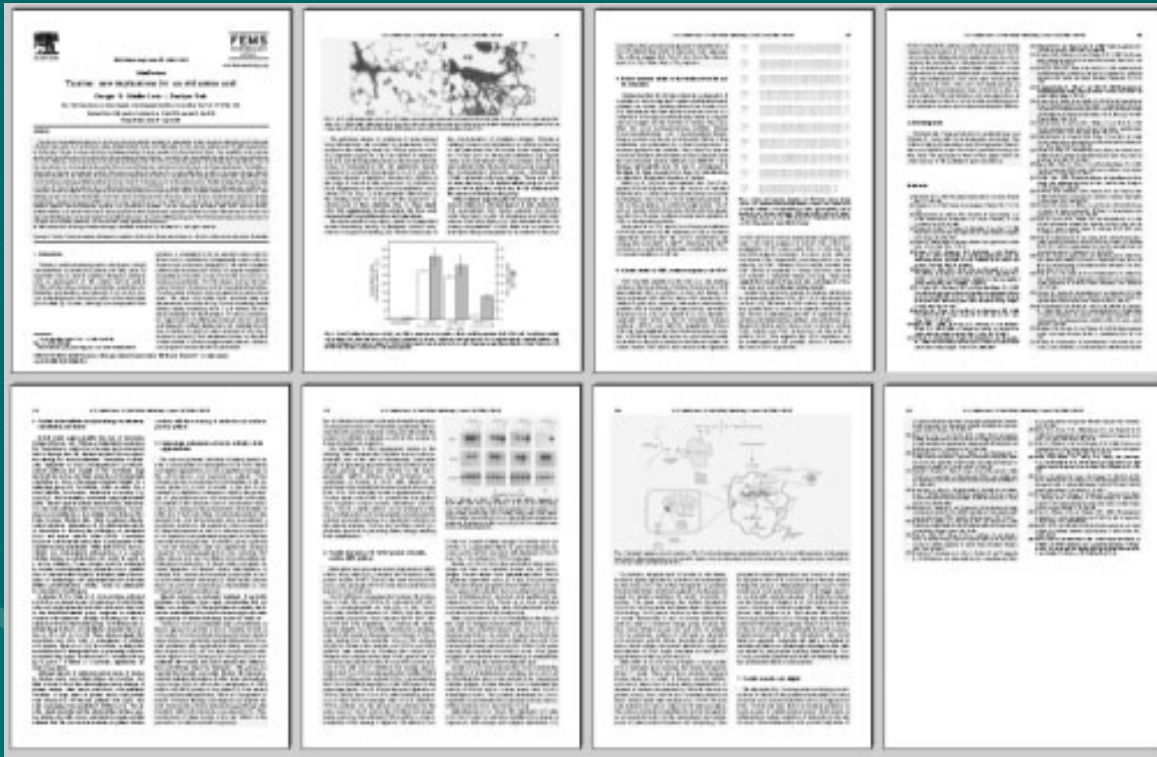
Table 1. Duplicate algorithm statistics averaged on the 12 year time series (60 000 searches)

Characteristics	Mean \pm SD (%)
Sensitivity (or Recall)	50.3 \pm 4.0
Specificity	99.8 \pm 0.1
Positive predictive value (or Precision)	87.8 \pm 10.9
Negative predictive value	99.3 \pm 0.4

Multiple duplicate publications by Shahrudin, Mohd-Dun



Original and Duplicate/SA-classified articles share many elements



- 75% of the text,
- two of the five figures
- 90% of the references were identical

The original paper (G Schuller- Levis and E Park, ‘Taurine: new implications for an old amino acid’. PMID 14553911) was accepted for publication five days after the submission of the duplicate (G Schuller-Levis and E Park, ‘Taurine and its chloramine: modulators of immunity, a mini-review’

The **Déjà vu** results database

1. **browse** Déjà vu entries with no specific search method. Each entry links to the scientific citation along with full text whenever freely available;
2. **search** Déjà vu content by authors, title word, abstract word, year and comment word;
3. **view** Déjà vu results in a particular category or identified by a particular ‘discovery method’ (eTBLAST or manual);
4. **provide comments** in order to contest a record or submit a potential duplication that will be reviewed by authors of this manuscript.

Summary

- Uuriti 62 213 juhuslikult valitud Medline'i artiklit:
 - **0.04%** of the citations with no shared authors were highly similar and are thus potential cases of **plagiarism**
 - **1.35%** with shared authors were sufficiently similar to be considered a **duplicate**
 - Extrapolating, this would correspond to **3500** and **117 500** duplicate citations in total, respectively