Knowledge Discovery Techniques
for Bioinformatics, Part V-2:

Applications in Database Cleansing

Limsoon Wong

Plan

- **Association Rules Mining**
- **Errors in Biological Databases**
- **Database Cleansing**

Copyright 2006 © Limsoon Wong

Association Rule Mining

What is Association Rules Mining?

- **Given a set of items/attributes, and a set of objects containing a subset of the items**
- **Find rules**

  if I_1 then I_2 (sup, conf)

  Where
  - **I_1, I_2 are sets of items**
  - **I_1, I_2 have good support: $P(I_1 + I_2)$**
  - **Rule has good confidence: $P(I_2 \mid I_1)$**

Copyright 2006 © Limsoon Wong

Association Rules Mining

- **User specifies "interestingness"**
  - Minimum support (minsup)
  - Minimum confidence (minconf)
- **Find all frequent itemsets (> minsup)**
  - Exponential Search Space
  - Computation and I/O Intensive
- **Generate strong rules (> minconf)**
  - Relatively cheap

Copyright 2006 © Limsoon Wong

Example

OID	Items
1	A C T W
2	C D W
3	A C T W
4	A C D W
5	A C D T W
6	C D T

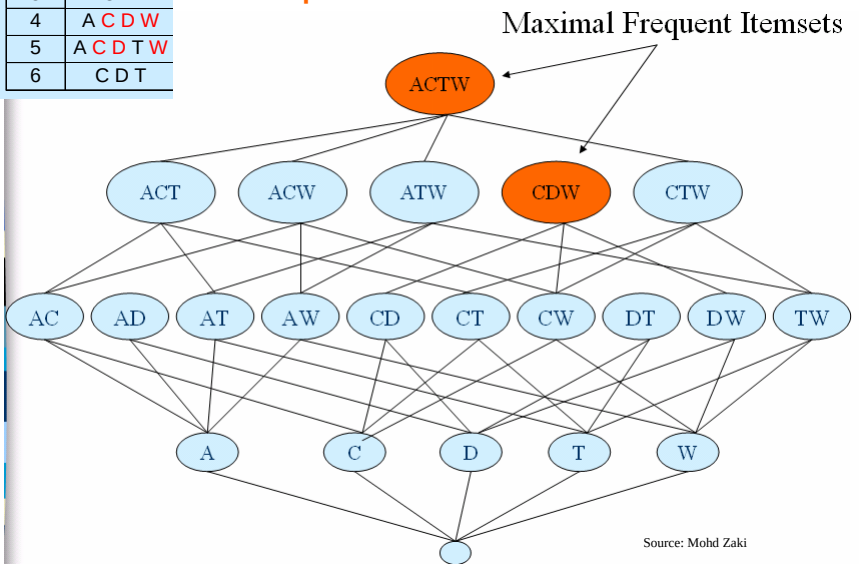
Maximal Frequent Itemsets
ACTW, CDW

Support	Itemsets
100%(6)	
83% (5)	C, CW
67% (4)	A, D, T, AC, AW, CD, CT, ACW
50% (3)	AT, DW, TW, ACT, ATW, CDW, CTW, ACTW

Copyright 2006 © Limsoon Wong

OID	Items
1	A C T W
2	C D W
3	A C T W
4	A C D W
5	A C D T W
6	C D T

Frequent Itemset Lattice

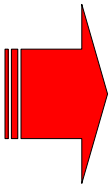


Copyright 2006 © Limsoon Wong

OID	Items
1	A C T W
2	C D W
3	A C T W
4	A C D W
5	A C D T W
6	C D T

Example: Sets to Rules

- **Maximal set**
  - CDW
- **Subsets**
  - CDW (3)
  - CD (4)
  - CW (5)
  - DW (3)
  - C (6)
  - D(4)
  - W (5)



- **CD ⇒ W, conf = 3/4 = 75%**
- **CW ⇒ D, conf = 3/5 = 60%**
- **DW ⇒ C, conf = 3/3 = 100%**
- **C ⇒ DW, conf = 3/6 = 50%**
- **D ⇒ CW, conf = 3/4 = 75%**
- **W ⇒ CD, conf = 3/5 = 60%**

Copyright 2006 © Limsoon Wong

Closed Pattern Mining Algorithms

- CLOSET, Pei et al. 2000
- CARPENTER, Pan et al. 2003
- FPclose*, Grahne & Zhu 2003
- GC-growth, Li et al. 2005
- ...

⇒ **We have efficient algorithms for mining association rules**

Copyright 2006 © Limsoon Wong

Errors in Biological Databases

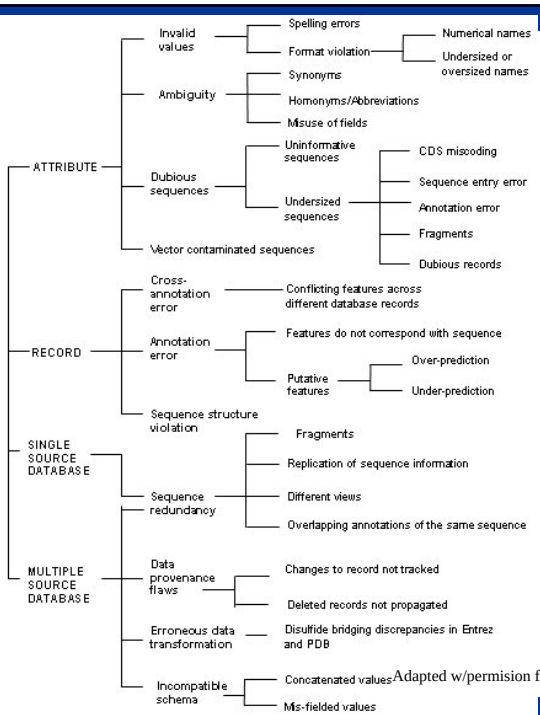
Sources of Errors, Koh et al, *DBiDB*, 2005

- **11 types & 28 subtypes of data artifacts**
  - Critical artifacts (vector contaminated sequences, duplicates, sequence structure violations)
  - Non-critical artifacts (misspellings, synonyms)
- **> 20,000 seq records in public contain artifacts**
- **Identification of these artifacts are impt for accurate knowledge discovery**

- **Sources of artifacts**
  - Diverse sources of data
    - **Repeated submissions of seqs to db's**
    - **Cross-updating of db's**
  - Data Annotation
    - **Db's have diff ways for data annotation**
    - **Data entry errors can be introduced**
    - **Different interpretations**
  - Lack of standardized nomenclature
    - **Variations in naming**
    - **Synonyms, homonyms, & abbrevn**
  - Inadequacy of data quality control mechanisms
    - **Systematic approaches to data cleaning are lacking**

Adapted w/ permission from Judice Koh, Mong Li Lee, & Vladimir Brusic

Copyright 2006 © Limsoon Wong



Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

Classification of Errors

Copyright 2006 © Limsoon Wong

- ATTRIBUTE
  - Invalid values
    - **Spelling errors**
    - Format violation
  - Ambiguity
  - Dubious sequences
  - Vector contaminated sequence
- RECORD
  - Cross-annotation error
  - Annotation error
  - Sequence structure violation
- SINGLE SOURCE DATABASE
  - Sequence redundancy
  - Data Provenance flaws
- MULTIPLE SOURCE DATABASE
  - Erroneous data transformation
  - Incompatible schema

Spelling Errors

- Usually typo errors
- Occurs in different fields of the record
- We identified 569 possible misspelled words affecting up to 20,505 nucleotide records in Entrez.

Misspellings	Corrections	Context of the misspellings
Immunoglobin	Immunoglobulin	GenBank:AAD26534 nectin-1 [Rattus norvegicus] TITLE Nectin/PRR: An Immunogloblin-like Cell Adhesion Molecule Recruited to Cadherin-based Adherens Junctions through Interaction with Afadin, a PDZ Domain-containing Protein gi|4590334|gb|AAD26534.1
Cassete	Cassette	Patent Database:A76783 Sequence 11 from Patent WO9315210 CDS <1..150 /note="gene cassete encoding intercalating jun-zipper and linker" gi|6088638|emb|A76783.1||pat|WO|9315210|11[6088638]
tranmembrane	transmembrane	Swiss-Prot:P03385 Env polyprotein precursor DEFINITION Env polyprotein precursor [Contains: Surface protein (SU) (GP70); Tranmembrane protein (TM) (p15E); R protein]. gi|119478|sp|P03385|ENV_MLVMO
asociated	associated	EMBL:Y18050 E.faecium pbp5 gene TITLE Modification of penicillin-binding protein 5 asociated with high level ampicillin resistance in Enterococcus faecium gi|1143442|emb|X92687.1|EFPBP5G

Copyright © 2006 by Limsoon Wong. Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

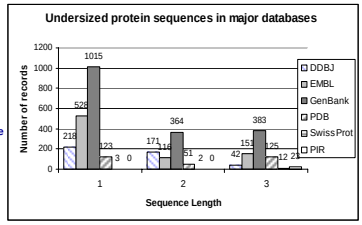
- ATTRIBUTE
  - Invalid values
    - Uninformative sequences
    - **Undersized sequences**
  - Ambiguity
  - Dubious sequences
  - Vector contaminated sequence
- RECORD
  - Cross-annotation error
  - Annotation error
  - Sequence structure violation
- SINGLE SOURCE DATABASE
  - Sequence redundancy
  - Data provenance flaws
- MULTIPLE SOURCE DATABASE
  - Erroneous data transformation
  - Incompatible schema

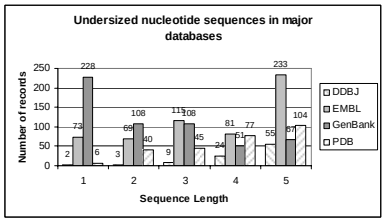
Meaningless Seqs

Among the 5,146,255 protein records queried using Entrez to the major protein or translated nucleotide databases , 3,327 protein sequences are shorter than four residues (as of Sep, 2004).

- In Nov 2004, the total number of undersized protein sequences increases to 3,350.

- Among 43,026,887 nucleotide records queried using Entrez to major nucleotide databases, 1,448 records contain sequences shorter than six bases (as of Sep, 2004).
- In Nov 2004, the total number of undersized nucleotide sequences increases to 1,711.



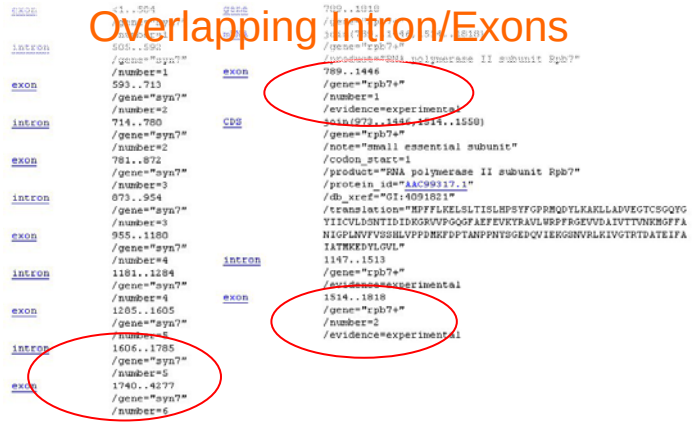


Copyright © 2006 by Limsoon Wong. Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

- ATTRIBUTE
  - Invalid values
  - Ambiguity
  - Dubious sequences
  - Vector contaminated sequence
- RECORD
  - Cross-annotation error
  - Annotation error
  - Sequence structure violation
- SINGLE SOURCE DATABASE
  - Sequence redundancy
  - Data Provenance flaws
- MULTIPLE SOURCE DATABASE
  - Erroneous data transformation
  - Incompatible schema

Overlapping intron/exon

Overlapping Intron/Exons



Syn7 gene of putative polyketide synthase in NCBI TPA record BN000507 has overlapping intron 5 and exon 6.

rpb7+ RNA polymerase II subunit in GENBANK record AF055916 has overlapping exon 1 and exon 2.

Copyright © 2006 by Limsoon Wong. Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

- ATTRIBUTE
  - Invalid values
  - Ambiguity
  - Dubious sequences
  - Vector contaminated sequence
- RECORD
  - Cross-annotation error
  - Annotation error
  - Sequence structure violation
- SINGLE SOURCE DATABASE
  - Sequence redundancy
  - Data provenance flaws
- MULTIPLE SOURCE DATABASE
  - Erroneous data transformation
  - Incompatible schema

Replication of sequence information

- Different views
- Overlapping annotations of the same sequence

Seqs w/ Identical Info

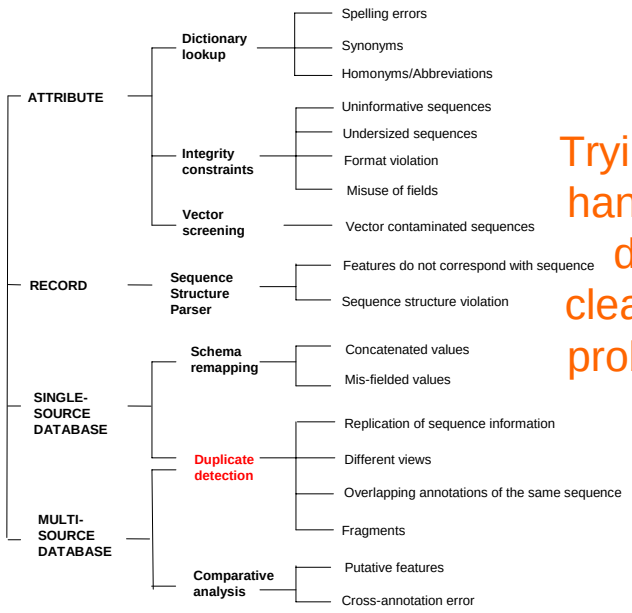
- Submission of the same sequence to different databases
- Repeated submission of the same sequence to the same database
- Initially submitted by different groups
- Protein sequences may be translated from duplicate nucleotide sequences

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=11692005&dopt=GenPept

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=11692005&dopt=GenPept

Copyright © 2006 by Limsoon Wong. Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

Database Cleansing

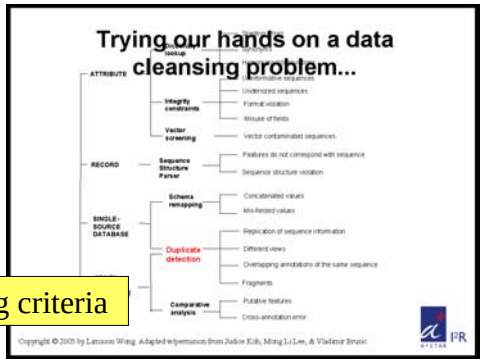


Trying our hands on data cleansing problems

Copyright 2006 © Limsoon Wong

19

Association Rule Mining for De-duplication



Select matching criteria

↓

Compute similarity scores from known duplicate pairs

↓

Generate association rules

↓

Detect duplicates using the rules

Copyright 2006 © Limsoon Wong

20

Dataset

Entrez (GenBank, GenPept, SwissProt, DDBJ, PIR, PDB)

scorpion AND (venom OR toxin)

serpentes AND venom AND PLA2

Scorpion venom dataset containing 520 records

Snake PLA2 venom dataset containing 780 records

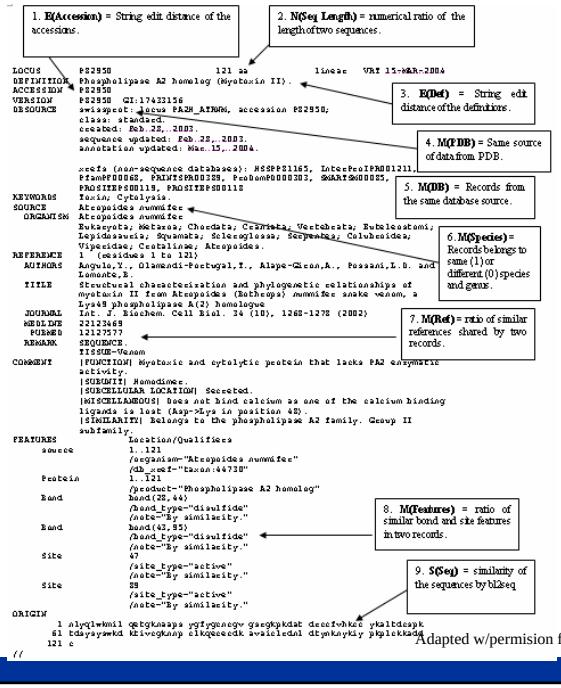
Expert annotation

251 duplicate pairs

444 duplicate pairs

695 duplicate pairs **are collectively identified.**

Copyright 2006 © Limsoon Wong



Features to Match

Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

Copyright 2006 © Limsoon Wong

Association Rule Mining

Similarity scores of known duplicate pairs

AAG39642 AAG39643 AC0.9 LE1.0 DE1.0 DB1 SP1 RF1.0 PD0 FT1.0 SQ1.0
AAG39642 Q9GNG8 AC0.1 LE1.0 DE0.4 DB0 SP1 RF1.0 PD0 FT0.1 SQ1.0
P00599 PSNJ1W AC0.2 LE1.0 DE0.4 DB0 SP1 RF1.0 PD0 FT1.0 SQ1.0
P01486 NTSREB AC0.0 LE1.0 DE0.3 DB0 SP1 RF1.0 PD0 FT1.0 SQ1.0
O57385 CAA11159 AC0.1 LE1.0 DE0.5 DB0 SP1 RF0.0 PD0 FT0.1 SQ1.0
S32792 P24663 AC0.0 LE1.0 DE0.4 DB0 SP1 RF0.5 PD0 FT1.0 SQ1.0
P45629 S53330 AC0.0 LE1.0 DE0.2 DB0 SP1 RF1.0 PD0 FT1.0 SQ1.0

Association rule mining ⇩

Frequent item-set with support

LE1.0 PD0 SQ1.0 **(99.7%)**
SP1 PD0 SQ1.0 **(97.1%)**
SP1 LE1.0 PD0 SQ1.0 **(96.8%)**
DB0 PD0 SQ1.0 **(93.1%)**
DB0 LE1.0 PD0 SQ1.0 **(92.8%)**
DB0 SP1 PD0 SQ1.0 **(90.4%)**
DB0 SP1 LE1.0 PD0 SQ1.0 **(90.1%)**
RF1.0 SP1 LE1.0 PD0 SQ1.0 **(47.6%)**
RF1.0 DB0 LE1.0 PD0 SQ1.0 **(44.0%)**

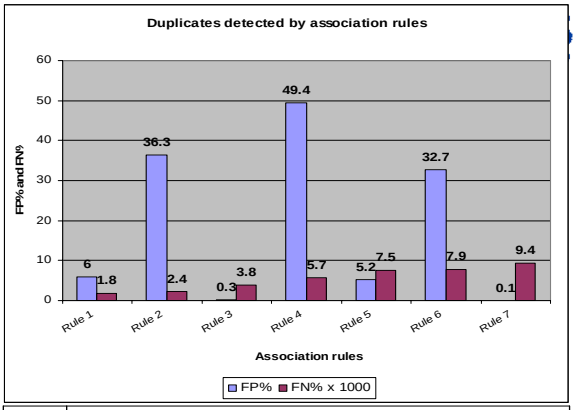
Copyright 2006 © Limsoon Wong

Results

Rule 1. Identical sequences with the same sequence length and not originated from PDB are 99.7% likely to be duplicates.

Rule 2. Identical sequences with the same sequence length and of the same species are 97.1% likely to be duplicates.

Rule 3. Identical sequences with the same sequence length, of the same species and not originated from PDB are 96.8% likely to be duplicates.



Rule 1	S(Seq)=1 ^ N(Seq Length)=1 ^ M(PDB)=0 (99.7%)
Rule 2	S(Seq)=1 ^ M(PDB)=0 ^ M(Species)=1 (97.1%)
Rule 3	S(Seq)=1 ^ N(Seq Length)=1 ^ M(Species)=1 ^ M(PDB)=0 (96.8%)
Rule 4	S(Seq)=1^ M(PDB)=0 ^ M(DB)=0 (93.1%)
Rule 5	S(Seq)=1 ^ M(Seq Length)=1 ^ M(PDB)=0 ^ M(DB)=0 (92.8%)
Rule 6	S(Seq)=1 ^ M(Species)=1 ^ M(PDB)=0 ^ M(DB)=0 (90.4%)
Rule 7	S(Seq)=1 ^ N(Seq Length)=1 ^ M(Species)=1 ^ M(PDB)=0 ^ M(DB)=0 (90.1%)

Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

Copyright 2006 © Limsoon Wong

Acknowledgements

- **Mohamed Zaki**
- **Judice Koh**
- **Vladimir Brusic**
- **Mong Li Lee**

Copyright 2006 © Limsoon Wong

References

- Koh et al., “A Classification of Biological Data Artifacts”, DBiBD, 2005
- R. Agrawal, et al., “Mining association rules between sets of items in large databases”, SIGMOD, 1993
- J. Han, et al., “Mining frequent patterns without candidates generation”, SIGMOD, 2000
- H. Li, et al., “Relative risk and odds ratio: A data mining perspective”, PODS, 2005

Copyright 2006 © Limsoon Wong