

Knowledge Discovery Techniques
for Bioinformatics, Part V-2:
Applications in Database Cleansing

Limsoon Wong



2

Plan



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

Association Rule Mining



4

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: $P(I_2 | I_1)$

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

Example

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

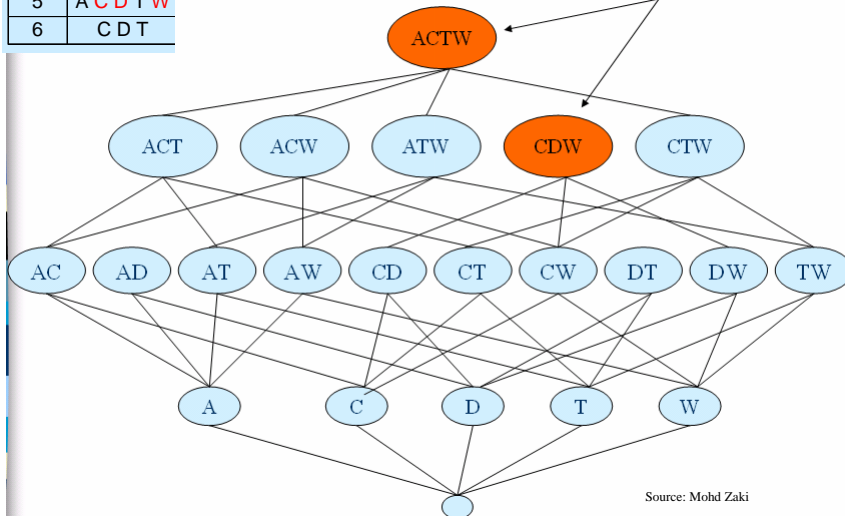
Maximal Frequent Itemsets
ACTW, CDW

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



Maximal Frequent Itemsets



Source: Mohd Zaki

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

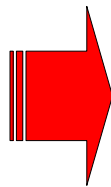


- **Maximal set**

- CDW

- **Subsets**

- CDW (3)
- CD (4)
- CW (5)
- DW (3)
- C (6)
- D(4)
- W (5)



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

Errors in Biological Databases



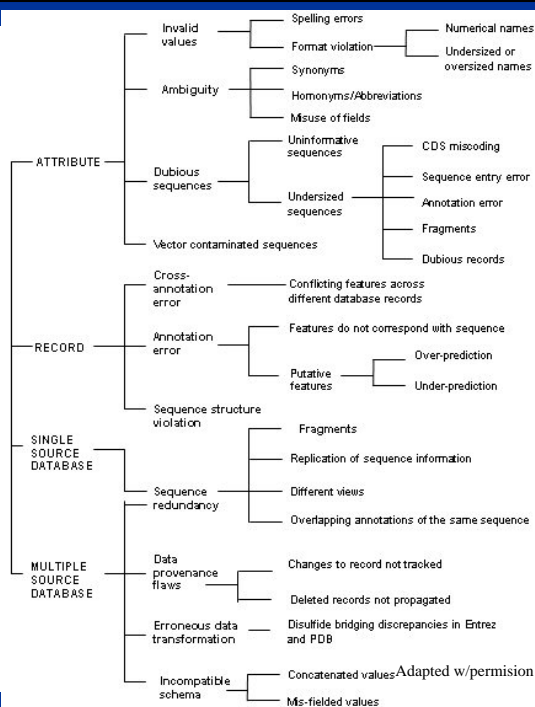
Sources of Errors, Koh et al, DBiDB, 2005



- **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 imp't 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 Justice Koh, Mong Li Lee, & Vladimir Brusic

Copyright 2006 © Limsoon Wong



Classification of Errors



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

Copyright 2006 © Limsoon Wong

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

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:AA026534 neclin-1 [Rattus norvegicus] TITLE Nectin-1 (Rattus norvegicus) TITLE Nectin-1: An Immunoglobulin-like Cell Adhesion Molecule Recruited to Cadherin-based Adhesions Junctions through Interaction with Afadin, a PDZ Domain-containing Protein gi4599354 gb AA026534.1
Cassete	Cassette	Patent Database:A76783 Sequence 11 from Patent WO9315216 CDS c1..150 /code=" gene cassette encoding intercalating Jun-zipper and linker" gi6088638 emb A76783.1 pat WO9315216 1 6088638
transmembrane	transmembrane	Swiss-Prot:P03385 Env polyprotein precursor DEFINITION Env polyprotein precursor (Contains: Surface protein (SU) (GP70); Transmembrane protein (TM) (p15E); R protein). gi119478 sp P03385 ENV_MLVMO
associated	associated	EMBL:Y18050 E.faecium ptp5 gene TITLE Modification of penicillin-binding protein 5 associated with high level ampicillin resistance in Enterococcus faecium gi1143442 emb X92687.1 EFPBPSG

Copyright © 2006 by Limsoon Wong. Adapted w/permission 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

Meaningless Seqs

Undersized sequences

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**.

Undersized protein sequences in major databases

Sequence Length	DDBJ	EMBL	GenBank	PDB	Swiss Prot	PIR
1	218	528	1015	0	3	0
2	171	364	383	0	15	0
3	42	15	25	0	12	0

Undersized nucleotide sequences in major databases

Sequence Length	DDBJ	EMBL	GenBank	PDB
1	73	228	6	3
2	66	108	40	9
3	113	108	45	24
4	81	77	55	87
5	233	104	0	0

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

7

Invalid values

ATTRIBUTE

- 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

```

exon      51..504      /gene="aya7"
intron    505..592      /gene="aya7"
exon      593..713      /gene="aya7"
intron    714..700      /gene="aya7"
exon      701..872      /gene="aya7"
intron    873..954      /gene="aya7"
exon      955..1180     /gene="aya7"
intron    1181..1284     /gene="aya7"
exon      1285..1605     /gene="aya7"
intron    1606..1785     /gene="aya7"
exon      1740..4277     /gene="aya7"

exon      789..1444     /gene="rbp7+"
intron    1445..1550     /gene="rbp7+"
exon      1514..1818     /gene="rbp7+"
intron    1819..1819     /gene="rbp7+"

```

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

rbp7+ RNA polymerase II subunit in GENBANK record AF055916 has overlapping exon 1 and exon 2.

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

Invalid values

ATTRIBUTE

- 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

```

! :AA039643 alpha toxin TX15 - [gi11692003]
LOCUS      AA039643      85 aa      linear      IMF 01-SEP-2001
DEFINITION alpha toxin TX15 [Dubious matenonii].
ACCESSION  AA039643
VERSION   AA039643.1 GI:11692003
SEQUENCE  Join AF181917 accession AF181917.1
SOURCE    Mesobius matenonii [Dubious matenonii]
ORGANISM  Mesobius matenonii
           Escherichia, Bacteroidetes, Chalcidacea, Arachnida, Scorpiones,
           Buthidae, Buthinae, Mesobuthus.
REFERENCE  1 [exon] 1 to 81
AUTHOR    Zhu, D.Y., Li, W.J., Song, J.C., Liu, R., Jiang, D.H. and Mao, I.
TITLE     New novel proteinase of Mesobius matenonii scorpion alpha-toxin
JOURNAL   Toxicon 39 (12), 1853-1861 (2000)
MEDLINE  18531861
PROVIDE   1 [exon] 1 to 81
          2 [exon] 1 and 20001.1.
TITLE     Dioxin Subunit in
          Subunit 130-1308-1999) Virology Department, Wuhan University,
          Luojia Mountain, Wuhan, Hubei 430072, China
METHOD    conceptual translation.
COMMENT   Location/Qualifiers
          source          /organism="Mesobius matenonii"
                       /db_xref="GenBank:AA039643"
          exon            /protein_type="toxin gland"
          CDS             /protein="alpha toxin TX15"
          ORIGIN          /code="gen"
          1 metvlelala llympevew vqplalddn. capyqenay oadckhnga svpypqwy
          41 ypnacwylk pdkyirvpp hknpp
          //
          Revised August 8, 2002.
          http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=
          -protein&list_uids=11692003&dopt=GenPept

```

```

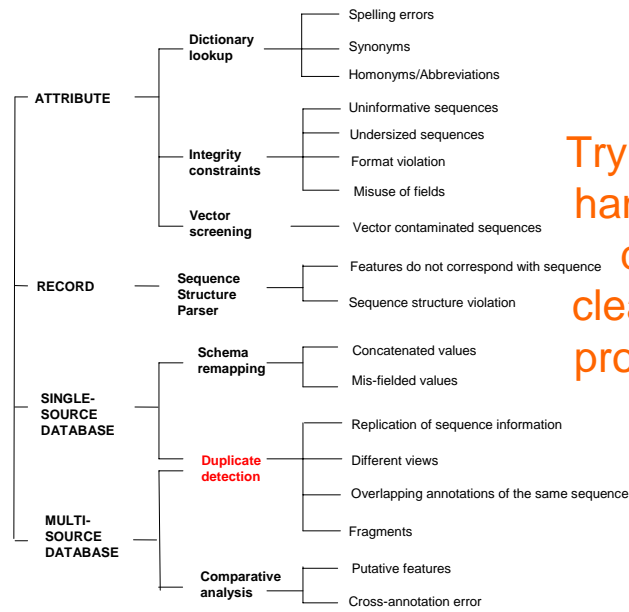
! :AA039642 alpha toxin TX15 - [gi11692003]
LOCUS      AA039642      85 aa      linear      IMF 01-SEP-2001
DEFINITION alpha toxin TX15 [Dubious matenonii].
ACCESSION  AA039642
VERSION   AA039642.1 GI:11692003
SEQUENCE  Join AF181917 accession AF181917.1
SOURCE    Mesobius matenonii [Dubious matenonii]
ORGANISM  Mesobius matenonii
           Escherichia, Bacteroidetes, Chalcidacea, Arachnida, Scorpiones,
           Buthidae, Buthinae, Mesobuthus.
REFERENCE  1 [exon] 1 to 81
AUTHOR    Zhu, D.Y., Li, W.J., Song, J.C., Liu, R., Jiang, D.H. and Mao, I.
TITLE     New novel proteinase of Mesobius matenonii scorpion alpha-toxin
JOURNAL   Toxicon 39 (12), 1853-1861 (2000)
MEDLINE  18531861
PROVIDE   1 [exon] 1 to 81
          2 [exon] 1 and 20001.1.
TITLE     Dioxin Subunit in
          Subunit 130-1308-1999) Virology Department, Wuhan University,
          Luojia Mountain, Wuhan, Hubei 430072, China
METHOD    conceptual translation.
COMMENT   Location/Qualifiers
          source          /organism="Mesobius matenonii"
                       /db_xref="GenBank:AA039642"
          exon            /protein_type="toxin gland"
          CDS             /protein="alpha toxin TX15"
          ORIGIN          /code="gen"
          1 metvlelala llympevew vqplalddn. capyqenay oadckhnga svpypqwy
          41 ypnacwylk pdkyirvpp hknpp
          //
          Revised August 8, 2002.
          http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=
          -protein&list_uids=11692003&dopt=GenPept

```

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

8

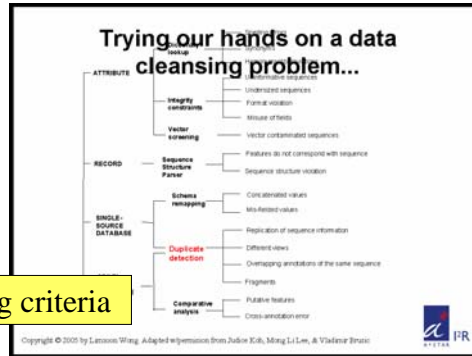
Database Cleansing



Trying our hands on data cleansing problems



Association Rule Mining for De-duplication



Select matching criteria

Compute similarity scores from known duplicate pairs

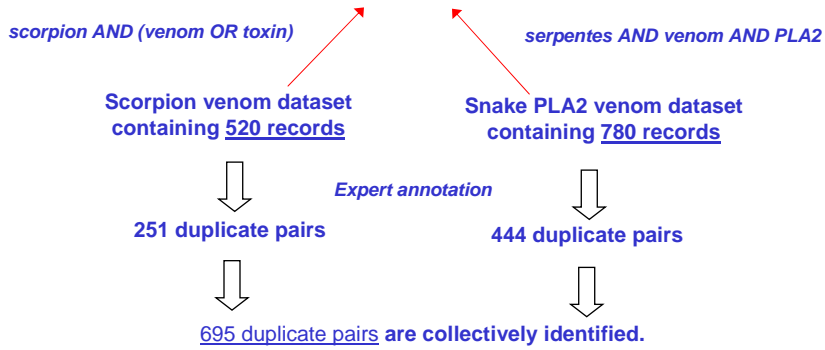
Generate association rules

Detect duplicates using the rules

Dataset



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





1. $E(\text{Accession})$ = String edit distance of the accessions.

2. $N(\text{Seq Length})$ = numerical ratio of the length of two sequences.

3. $E(\text{Def})$ = String edit distance of the definitions.

4. $M(\text{DB})$ = Same source of data from PDB.

5. $M(\text{DB})$ = Records from the same database source.

6. $M(\text{Species})$ = Records belongs to same (1) or different (0) species and genus.

7. $M(\text{Ref})$ = ratio of similar references shared by two records.

8. $M(\text{Feature})$ = ratio of similar bond and site features in two records.

9. $S(\text{Seq})$ = similarity of the sequences by blkseq

```

LOCUS       P22550             121 aa             11aaace   VRI 15-09A-2004
DEFINITION Phospholipase A2 homolog (Myristic II).
ACCESSION   P22550
VERSION     P22550 GI:17423156
DBSOURCE    swissprot: locus P225_ATM06, accession P22550;
            class: standard
            created: Feb 25, 2001.
            sequence updated: Feb 25, 2002.
            annotation updated: Mar 15, 2004.

seqifs (non-sequence databases): MS09P11165, InterProIPM01211,
PfamPF0068, PfamSF000128, ProdomP0000202, SMARTSM00025,
PROSITEPS0119, PROSITEPS0118
KEYWORDS    Tania; Cytolysis.
SOURCE      Atropaidea nummifera
ORGANISM    Atropaidea nummifera
            Eukaryota; Metazoa; Chordata; Crustacea; Vertebrata; Mollusca;
            Cephalopoda; Squamata; Sclerozoa; Scyphozoa; Cnidaria;
            Viridiplantae; Charophyta; Atropaidea.
REFERENCE   1 (residues 1 to 121)
            Angula, V., Ghoshal-Deo, I., Klupe-Green, A., Passani, I. D. and
            Lomonaco, R.
            Structural characterization and phylogenetic relationships of
            myristic II from Atropaidea (Cnidaria) nummifera snake venom, a
            Lys48 phospholipase A(2) homologue
            Int. J. Biochem. Cell Biol. 34 (10), 1269-1276 (2002)
            2112349
            12123577
            PMID:
            REFSEQ:
            TIGR00-Venom
            [FUNCTION] Hypotonic and catalytic protein that lacks PA2 enzymatic
            activity.
            [SUBMITTER] Homologous.
            [SOURCE] ATROPAIDEA nummifera.
            [MISC] ATROPAIDEA nummifera does not bind calcium as one of the calcium binding
            ligands in the Asp-Tyr in position 42.
            [SIMILARITY] Belongs to the phospholipase A2 family. Group II
            subfamily.
FEATURES
            Location/Qualifiers
            source          1..121
                            /organism="Atropaidea nummifera"
                            /db_xref="taxon:44738"
            protein         1..121
                            /product="Phospholipase A2 homolog"
            bond            bond(28,44)
                            /bond_type="disulfide"
                            /note="By similarity."
            bond            bond(42,85)
                            /bond_type="disulfide"
                            /note="By similarity."
            site            47
                            /site_type="active"
                            /note="By similarity."
            site            58
                            /site_type="active"
                            /note="By similarity."
            ORIGIN
            1 atpqlwml qbtgkaaps yfygnagvy gaegpkdat dectohktz ykltatasp
            61 tdayrwmk klvtegnwv slkqecck avallatd dptvtytr pklctktdk
            121 e
    
```

Features to Match

Adapted w/permission from Justice 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 Q9NG8 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%)

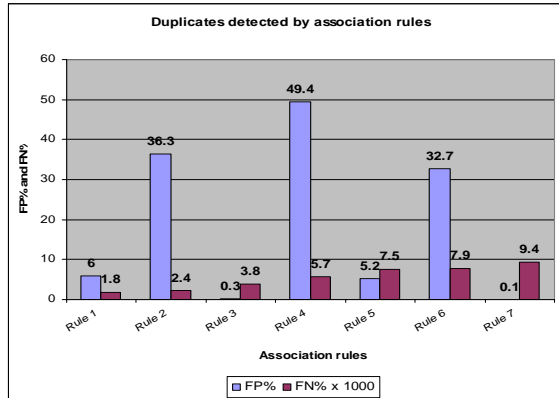
Copyright 2006 © Limsoon Wong

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.



Association rules	FP%	FN% x 1000
Rule 1	6	1.8
Rule 2	36.3	2.4
Rule 3	0.3	3.8
Rule 4	49.4	5.7
Rule 5	5.2	7.5
Rule 6	32.7	7.9
Rule 7	0.1	9.4

Adapted w/permission 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