Duplicate Detection in Biological Data using Association Rule Mining

Judice L.Y.Koh[1,2], Mong Li Lee[2], Asif M. Khan[1], Paul T.J. Tan[1] and Vladimir Brusic[1]

[1]Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
(+65)-68748281
{judice,dsif,tjtan,vladimir}@i2r.a-star.edu.sg

[2]National University of Singapore
School of Computing
Lower Kent Ridge road
Singapore 119260
(+65)-68742905
leeml@comp.nus.edu.sg

ABSTRACT

Recent advancement in biotechnology has produced a massive amount of raw biological data which are accumulating at an exponential rate. Errors, redundancy and discrepancies are prevalent in the raw data, and there is a serious need for systematic approaches towards biological data cleaning. This work examines the extent of redundancy in biological data and proposes a method for detecting duplicates in biological data. Duplicate relations in a real-world biological dataset are modeled into forms of association rules so that these duplicate relations or rules can be induced from data with known duplicates using association rule mining. Our approach of using association rule induction to find duplicate relations is new. Evaluation of our method on a real-world dataset shows that our duplicate association rules can accurately identify up to 96.8% of the duplicates in the dataset at the accuracy of 0.3% false positives and 0.0038% false negatives.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining, scientific databases*.

General Terms

Management

Keywords

Data cleaning, Association rules, Biological database

1. INTRODUCTION

Public sequence databases contain one of the most frequently accessed information on the Web. Databases such as GenBank [4] and Swiss-Prot [6] provide descriptions of common biological entities (genes, proteins, among others), and are intensively utilized by molecular biologists for their research. Over the last two decades, the rapid development of biological databases has been driven by an explosive growth of data due to high throughput sequencing and automation in genetics and proteomics. For example, the most recent statistical report of GenBank shows that the number of GenBank sequences has doubled over two years from the 14,976,310 sequences in 2001 to 30,968,418 sequences in 2003.

At the same time, numerous computational methods and algorithms, particularly in data mining, have been developed to extract the hidden knowledge in these data that is relevant to our understanding of the biological systems. Data mining results can be highly sensitive to the noise (errors and missing values) in the training datasets, therefore demanding that the datasets to contain only high quality data which are correct, accurate, consistent and concise. But in reality, public sequence data are incomplete, noisy, erroneous and highly redundant. The classes of errors contributing to the low quality of the public sequence databases are discussed in [18].

The protein or DNA sequences submitted by biologists from numerous sequencing centers and laboratories around the world to the public sequence databases are subjected to various sources of redundancy:

1. The same sequence may be submitted by the biologist to more than one database without cross-referencing these records.
2. The sequence is submitted more than once to a same database.
3. Annotations of the same sequence are submitted separately by different research groups.
4. Fragments and partial entries of the same protein or DNA sequence may be stored in different database records.

Biological data duplicates are varying representations of the same protein or DNA sequences in different database records. They provide hints of the redundancy in biological datasets. For example, record 1 and 2 in figure 1 refer to the same protein found separately in a PIR [3] and a Swiss-Prot database records. The example is likely resulted from the submission of the same protein sequence to both PIR and Swiss-Prot without cross-referencing to each other records. In this paper, we devise a method for determining the biological data duplicates.

Fields	Record 1	Record 2
Locus ID	P34180	S22388
Definition	Phospholipase A2, neutral precursor (Ammodytin I2) (Phosphatidylcholine 2-acylhydrolase).	phospholipase A2 (EC 3.1.1.4) ammodytin I2 precursor - western sand viper.
Database	swissprot: locus	pir: locus S22388;

source	PA2N_VIPAA, accession P34180;	
Organism	Vipera ammodytes ammodytes	Vipera ammodytes ammodytes
Sequence	MRTLWIVAVCLIGVE GNLYQFGNMIFKMTK KSALLSYSNYGCYCG WGGKGKPQDATDRC CFVHDCCYGRVNGC DPKLSIYSYSFENGDI VCGGDDPCLRAVCEC DRVAAICFGENLNTY DKKYKNYPSSHCTET EQC	MRTLWIVAVCLIGVE GNLYQFGNMIFKMTK KSALLSYSNYGCYCG WGGKGKPQDATDRC CFVHDCCYGRVNGC DPKLSIYSYSFENGDI VCGGDDPCLRAVCEC DRVAAICFGENLNTY DKKYKNYPSSHCTET EQC

Figure 1. Duplicate protein records. Record 1 and 2 are protein sequences from different databases.

We carried out an analysis of scorpion toxins in SCORPION, a fully referenced database of 221 scorpion toxins [20] to assess the extent of redundancy in biological data. The SCORPION records compiled from public database sources GenBank/GenPept [4], Swiss-Prot, EMBL [10], DDBJ [17], TrEMBL [6], PIR [3] and PDB [8] using keyword searches, were found to be overlapping to various degrees. Among the 143 duplicated scorpion toxins found by manual inspection, 27 toxins are replicated in any two different databases and 13 were replicated across any five different databases (Figure 2A).

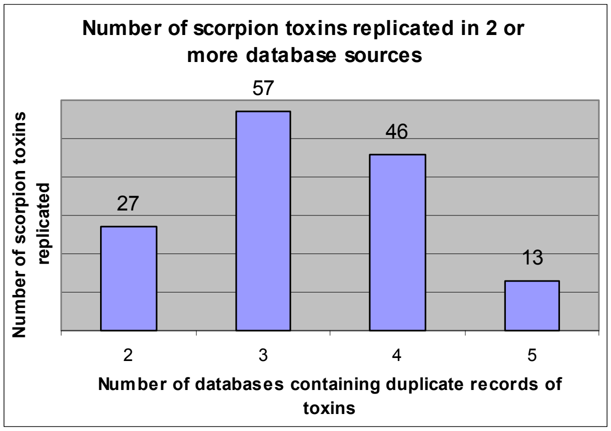


Figure 2A. Number of scorpion toxins replicated across two or more databases.

Figure 2B reports the number of duplicate scorpion toxin records found in each database. The extent of redundancy of data records in databases is most extensive in GenBank; 135 records are replicated in other databases. But again, the GenBank database contains the largest and most widely used pool of biological sequences. PDB duplicates refer to database records describing varying 3D structural views of the same protein sequence. These records have different orientations or conformations of the same protein sequence. For example, Entrez [21] records 1DJT_A and 1DJT_B are separate spatial organisations of the same scorpion toxins.

To ensure non-redundancy in the SCORPION dataset, duplicate records were deleted and redundant or partial records were merge-joined by the biologist manually.

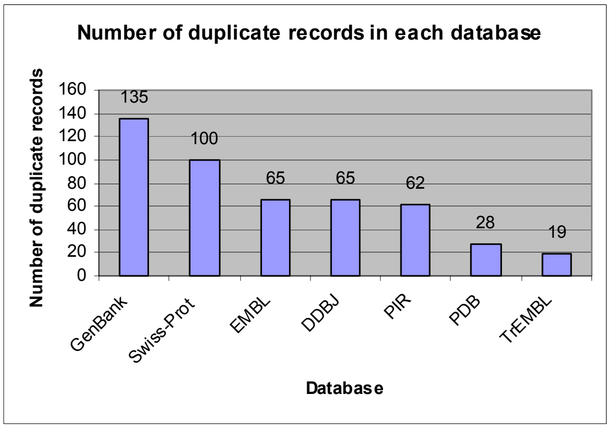


Figure 2B. Number of duplicate records in each database.

In this paper, we explore the matching criteria for comparing pairs of biological records and develop models of the duplicate relations in biological data using association rules mining. Our method of modeling and mining duplicate relations using association rule mining is new.

The rest of the paper is organised as follows: In section 2, we present some background knowledge of the problem and related works. We present our materials and methods in section 3. In section 4, we discuss results of our experiments and we conclude in section 6.

2. RELATED WORK

2.1 Data cleaning

The process of detecting and removing database defects and duplicates is referred to as data cleaning. Early works in data cleaning have focused on the merge/purge problem [13, 14]. The merge/purge problem addresses the fundamental issue that database records existing in disparate forms may refer to the same real world object and in the detection of these inexact duplicates in a database. Most data-cleaning methods are intended for managing customer information, and the data cleaning marketplace is largely focused on the cleaning of address and name lists for various marketing tasks. The merge/purge method, for example, has been commercialized into the TDMSUITE [24] and the Sagent solutions [15] which enable mailers to de-duplicate their mailings.

Duplicate relations can be learned from the datasets. Techniques that explore the adaptive learning of similarity measurements for string for duplicate detection are explored in [5, 9]. These techniques are appropriate for datasets which can be uniquely identified by a key field which is a string, such as the identifier names. For example, they can be used to detect the duplicates in figure 3 by matching only the name field “J.Koh” and “Judice Koh”. In the case of biological data, a record has more than one key field and we cannot rely on a single field match. Also, a biological record contains fields of other types which cannot be matched by string similarity measurements.

Fields	Record 1	Record 2
Name	J.Koh	Judice Koh
ID	744147-H	744147
Contact no.	97999914	63884783

Figure 3. Duplicate customer records.

Little work has been done on biological data cleaning and it is usually carried out in proprietary or ad-hoc manner, sometimes even manual. Systematic processes are lacking. From among the few examples, [23] uses stringent selection criteria to select 310 complete and unique records of *Homo sapiens* splice sites from the 4300 raw records in EMBL database. Although rigorous elimination of data is effective in removing redundancy, it may result in loss of critical information. In another example, a sequence structure parser is used to find missing or inconsistent features in records using the constraints of gene structure [19]. The method is only limited to detecting violations of the gene structure.

2.2 Association Rule Mining

Association rules mining or induction is commonly used in *market basket analysis* to find items frequently bought together by shoppers. The first algorithm for mining frequent item sets is the *Apriori* was used for market basket analysis [2]. For example, if amazon.com discovers that shoppers who buy the book "Data Mining: Concepts and Techniques" usually buy another book "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", they can arrange an offer for an offer package of these 2 books to increase their competitiveness in sale. The rules are induced from items that are most frequently occurred together, known as the *frequent item set*. A rule *"Buy(A) ^ Buy(B) → Buy(C)"* indicates that a customer who buys item A and item B buys C, with the interestingness of this rule measured from probabilities of *support* and *confidence*. The support is the percentage of transactions in amazon.com that contain A, B and C. The confidence is the percentage of transactions that contain A and B also containing C. Intuitively, the *antecedent* of this rule refers to the pattern or the set of frequent items, that is *associated with* a phenomenon which is the *consequent*.

Association rule mining has been applied to data cleaning to detect outliers but not in detecting duplicates. A method for cleaning data of using ordinal association rules is presented in [12]. Patterns of ordinal relationships among the fields in the dataset are mined to determine the common orderings of field values in a dataset. Deviations from these orderings are identified as "dirty data". Also, interval association rule induction is used in [11] to detect outlier values. Both methods are suitable for eliminating the outliers in datasets which do not follow the frequent patterns of field values. They cannot be used for detecting duplicates. Also they can only be applied to datasets with only numerical fields. Biological datasets contains few numerical fields.

3. PROBLEM FORMULATION

3.1 Duplicate Relation Models

Each duplicate relation or overall similarity of two records in a biological dataset is determined from the similarities of selected record fields. Thus, duplicate relations can be represented by a conjunctive clause of the value requirements of selected fields or *matching criteria*. Duplicate relations of this form are also known as the merging rules. An example of a duplicate relation model is the rule that two records with (1) identical protein or DNA sequence, (2) are of the same length, and (3) belong to the same species, are duplicates. The rule can be represented as:

Identical protein ^ same length ^ same species → duplicate

The conjunctive clause can be translated into a set of restricted values on each of the matching criteria, which can be calculated by applying data type specific similarity functions (S for sequence similarity, N for numerical ratio and M for Boolean matching) on the sequence, sequence length and species fields respectively.

S(Seq)=1.0 ^ N(Seq Length)=1.0 ^ M(species)=1 → duplicate

If we encode the matching values as items, the rule takes the form of an association rule and we can easily apply association rule mining to induce models of the duplicate relations from dataset of known duplicates.

SE1.0 ^ LE1.0 ^ SP1 → duplicate

The antecedents of this association rule are restricted values of the three criteria and the consequent is the duplicate relation.

3.2 Matching Criteria

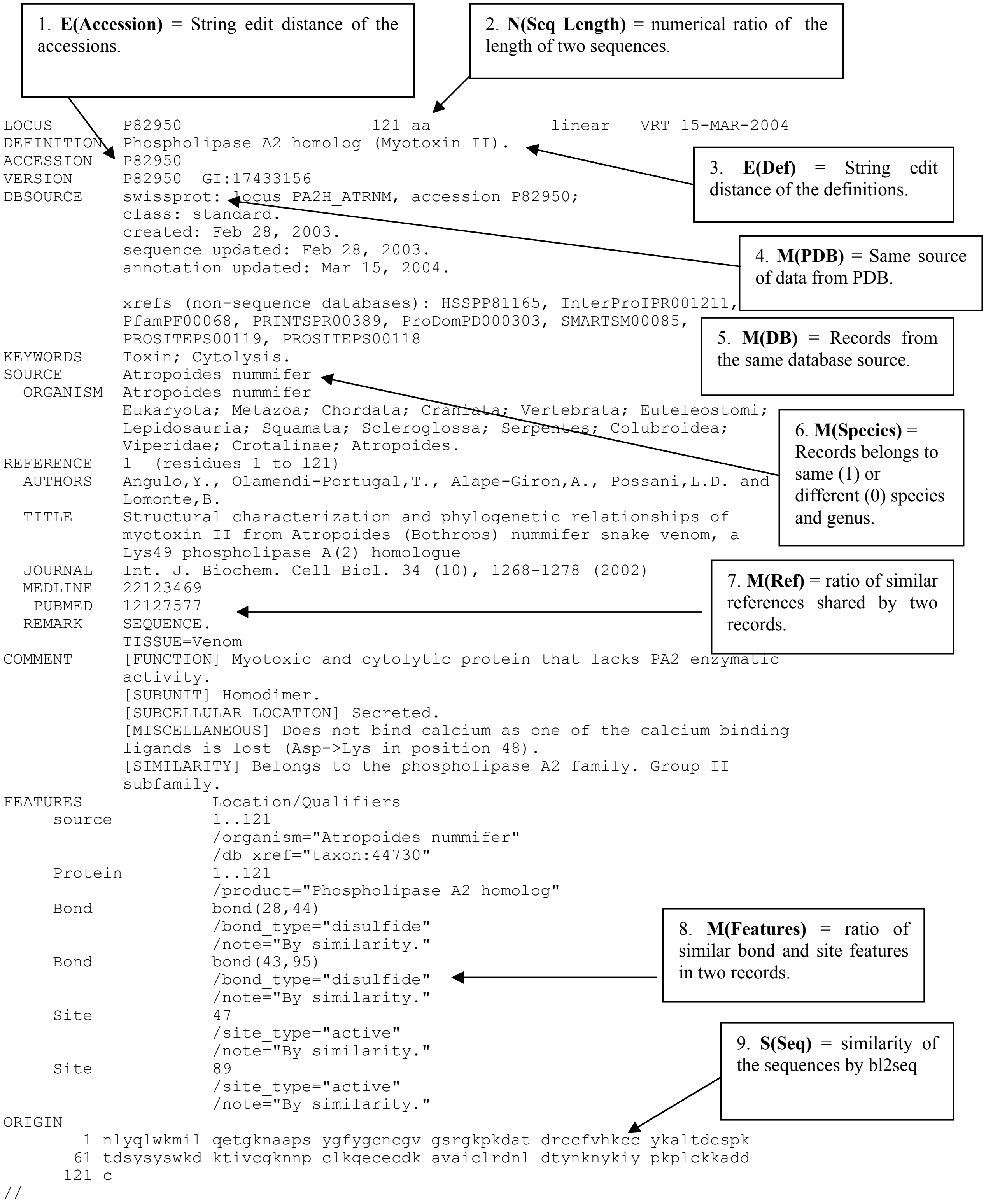
Biological records from Entrez are compared across a set of nine matching criteria (Figure 4). Because a biological record contains three main types of fields: (1) Protein and DNA sequences, (2) categorical fields, and (3) free-text strings, varying mechanisms for comparing the different fields or criteria are used, and we refer to them as the similarity functions. These functions measure the degree of similarity of corresponding fields.

Protein or DNA sequences are matched using their percentage identity scores computed from BLAST 2 sequences (bl2seq) algorithm [22]. Bl2seq utilize the gapped BLAST 2.0 algorithm [1] to align and compare pair-wise DNA-DNA or protein-protein sequences, and the percentage identity scores reflect the degree of similarity of the two sequences. We denote the sequence similarity function as **S**.

Categorical fields contain values belonging to a fixed value-set. For example, the organism fields in Entrez records are derived from the taxonomy of the organisms, which are fixed values already established in the GenBank database. Same field values are scored as either 1 (belongs to same category) or 0 (belongs to different category), and we denote the Boolean matching as similarity function **M.**

The third type of data fields are the free-text strings. The most common method for comparing string is the Edit distance or Levenshtein distance [16]. The edit distance computes the minimum number of edit operations (insertions, deletions, and substitutions) of single characters that are needed to transform from one string to another, and we denote the edit distance by **E**.

Figure 5 shows an example of the similarity scores of the ORIGIN sequence field, the ORGANISM field and the DEFINITION of two scorpion venom records in Entrez.



1. **E(Accession)** = String edit distance of the accessions.

2. **N(Seq Length)** = numerical ratio of the length of two sequences.

3. **E(Def)** = String edit distance of the definitions.

4. **M(PDB)** = Same source of data from PDB.

5. **M(DB)** = Records from the same database source.

6. **M(Species)** = Records belongs to same (1) or different (0) species and genus.

7. **M(Ref)** = ratio of similar references shared by two records.

8. **M(Features)** = ratio of similar bond and site features in two records.

9. **S(Seq)** = similarity of the sequences by bl2seq

```
LOCUS       P82950                  121 aa            linear   VRT 15-MAR-2004
DEFINITION  Phospholipase A2 homolog (Myotoxin II).
ACCESSION   P82950
VERSION     P82950  GI:17433156
DBSOURCE    swissprot: locus PA2H_ATRNM, accession P82950;
            class: standard.
            created: Feb 28, 2003.
            sequence updated: Feb 28, 2003.
            annotation updated: Mar 15, 2004.

            xrefs (non-sequence databases): HSSPP81165, InterProIPR001211,
            PfamPF00068, PRINTSPR00389, ProDomPD000303, SMARTSM00085,
            PROSITEPS00119, PROSITEPS00118
KEYWORDS    Toxin; Cytolysis.
SOURCE      Atropoides nummifer
  ORGANISM  Atropoides nummifer
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Lepidosauria; Squamata; Scleroglossa; Serpentes; Colubroidea;
            Viperidae; Crotalinae; Atropoides.
REFERENCE   1  (residues 1 to 121)
  AUTHORS   Angulo,Y., Olamendi-Portugal,T., Alape-Giron,A., Possani,L.D. and
            Lomonte,B.
  TITLE     Structural characterization and phylogenetic relationships of
            myotoxin II from Atropoides (Bothrops) nummifer snake venom, a
            Lys49 phospholipase A(2) homologue
  JOURNAL   Int. J. Biochem. Cell Biol. 34 (10), 1268-1278 (2002)
  MEDLINE   22123469
   PUBMED   12127577
  REMARK    SEQUENCE.
            TISSUE=Venom
COMMENT     [FUNCTION] Myotoxic and cytolytic protein that lacks PA2 enzymatic
            activity.
            [SUBUNIT] Homodimer.
            [SUBCELLULAR LOCATION] Secreted.
            [MISCELLANEOUS] Does not bind calcium as one of the calcium binding
            ligands is lost (Asp->Lys in position 48).
            [SIMILARITY] Belongs to the phospholipase A2 family. Group II
            subfamily.
FEATURES             Location/Qualifiers
     source          1..121
                     /organism="Atropoides nummifer"
                     /db_xref="taxon:44730"
     Protein         1..121
                     /product="Phospholipase A2 homolog"
     Bond            bond(28,44)
                     /bond_type="disulfide"
                     /note="By similarity."
     Bond            bond(43,95)
                     /bond_type="disulfide"
                     /note="By similarity."
     Site            47
                     /site_type="active"
                     /note="By similarity."
     Site            89
                     /site_type="active"
                     /note="By similarity."
ORIGIN
        1 nlyqlwkmil qetgknaaps ygfygcncgv gsrgkpkdat drccfvhkcc ykaltdcspk
       61 tdsysyswkd ktivcgknnp clkqececdk avaiclrdnl dtynknykiy pkplckkadd
      121 c
//
```

Figure 4. Matching criteria in Entrez records.

	1910194A	P45639	Score
ORIGIN	MCMPCFTTD HQMARKCDD CCGGKGRGK CYGPQCLCR	MCMPCFTTD HQMARKCDD CCGGKGRGK CYGPQCLCR	1
ORGANISM	Leiurus quinquestriatus quinquestriatus	Leiurus quinquestriatus quinquestriatus	1
DEFINITION	chlorotoxin.	Chlorotoxin	0.92

Figure 5. Similarity scores of Entrez records 1910194A and P45639.

4. MATERIALS AND METHODS

Our proposed framework for finding the duplicates in biological data is shown in Figure 6. We first select matching criteria for comparing record pairs. Selective attributes based on these matching criteria of the record pairs are compared using varying similarity functions which depend on the data types of the attributes. The similarity values for each pair of records in the training data are computed and used to generate the association rules that describe the duplicates. These association rules can be used to detect duplicates in biological datasets.

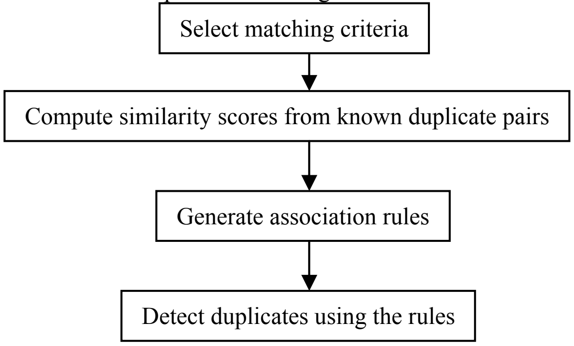

Select matching criteria
Compute similarity scores from known duplicate pairs
Generate association rules
Detect duplicates using the rules

Figure 6. Duplicate detection framework.

The training dataset contains the similarity scores of pairs of records across the nine criteria. To generate the items from the scores, we encode the values with field labels (Figure 7A). For quantitative values such as the sequence similarity scores which range from 0 to 1.0, the values are partitioned into equiwidth bins of 0.1. Hence, sequence similarity score item "SQ0.95" becomes "SQ0.9".

The implementation of the Apriori algorithm in [7] is used for inducing the association rules from the training. Figure 7B shows the frequent itemsets or rules generated from the apriori algorithm. We select the rules with support measures above the threshold of 90%, and from among these rules, we determine the rule with the lowest FP% and FN% (the definition of these measurements are given in 5.2) as the best rule.

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

Figure 7A. Field labels from each pair of duplicates in training dataset.

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%)
AC0.0 DB0 LE1.0 PD0 SQ1.0 (43.9%)
RF1.0 DB0 SP1 LE1.0 PD0 (42.7%)

Figure 7B. Frequent itemsets from association rule mining. The values in the brackets are the support measures of the association rules.

5. EXPERIMENT

5.1 DATASET

We evaluate our method on real-world sequence annotations from the Entrez retrieval system which contain protein records from various database sources, including sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to PIR, SWISS-PROT, PRF, and PDB.

Two set of records SD1 and SD2 are combined to form the training dataset. SD1 is the scorpion venom dataset containing 520 records retrieved from Entrez using the keywords "scorpion AND (venom OR toxin)". SD2 is the snake PLA2 venom dataset containing 780 records retrieved from Entrez using the keywords "serpentes AND venom AND PLA2". The duplicates among these 1300 records are annotated separately by two biological domain experts. SD1 contains 251 duplicate pairs. SD2 contains 444 duplicate pairs. 695 duplicate pairs are collectively identified.

5.2 Performance measurements

We compare the performance of using association rules induced from the training dataset to detect duplicates in the 1300 records of SD1 and SD2 with user or domain rules. The eight user rules are defined manually by domain experts based on their understanding of the biological data.

Performance of the rules are evaluate using two measures, the false negative percentage (FN%) and the false positive percentage (FP%). The false negatives are the number of true record linkage pairs which are not identified by the rules. FN% equals $100.N_{FN}/|R|_{dup}$ where N_{FN} is the number of FNs and $|R|_{dup}$ is the total number of duplicate record linkages in the relation.

In *Proceedings of the Second European Conference on Data Mining and Text Mining in Bioinformatics*, held in conjunction with ECML/PKDD in Pisa, Italy. 24. September, 2004.

Similarly, the false positives are the number of non-record linkages identified by the rules. FP% equals $100.N_{FP}/|R|$ where N_{FP} is the number of distinct record pairs mis-identified by the rules and |R| is the number of records in the relation. Good rules are indicated by low FN% and low FP%.

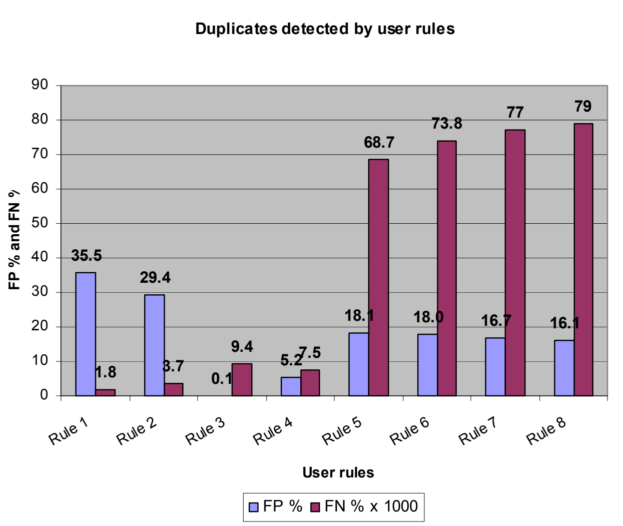
5.3 Results

Figure 8 shows the performance of detecting duplicates using user defined rules. The rule giving 1 false positive (FP% = 0.1) and 79 false negatives (FN% = 0.0094) indicate that a pair of annotated sequences duplicate is best identified by the association rule that these sequences are identical S(Seq)=1.0, of the same sequence length N(Seq length)=1.0, belongs to the same species M(Species) and are from different data sources M(DB) = 0:

S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(Species)=1 ^ M(DB)=0 → Duplicate

Another rule giving lower FN% (63 false negatives) of 7.5 but higher FP% (36 false positives) of 5.2 indicates that duplicate relation has little dependency on the species of which the sequence belongs to.

S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(DB)=0 → Duplicate



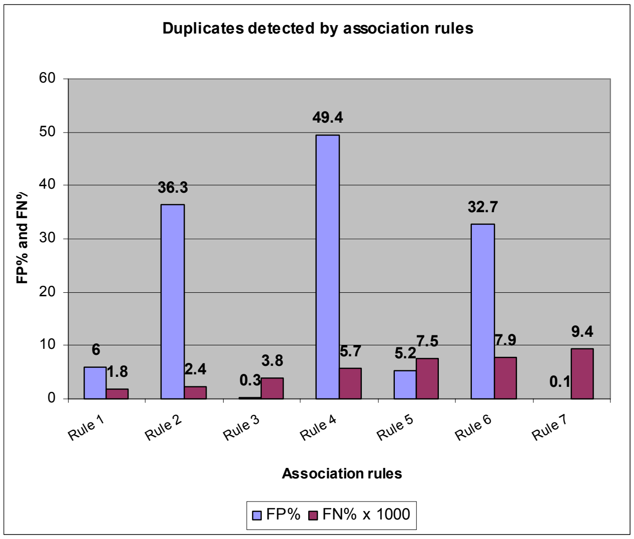
Rule 1	S(Seq)=1.0 ^ N(Seq length)=1.0
Rule 2	S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(Species)=1
Rule 3	S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(Species)=1 ^ M(DB)=0
Rule 4	S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(DB)=0
Rule 5	S(Seq)=1.0 ^ N(Seq length)=1..0 ^ M(Species)=1 ^ E(Def)=0.5
Rule 6	S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(Species)=1 ^ E(Def)=0.6
Rule 7	S(Seq) ^ N(Seq length)=1.0 ^ M(Species) ^ E(Def)=0.7
Rule 8	S(Seq)=1.0 ^ N(Seq length)=1.0 ^ M(Species)=1 ^ E(Def)=0.8

Figure 8. FP% and FN% using user rules for duplicate detection.

Rules using the definition fields for duplicate detection show high degree of false negative and the FN% increases with more rigid requirement for similar definitions E(Def)=0.5 to 0.8. Hence, the definition field of the sequence annotations is not a critical determinant of record similarity. In reality, standardized naming convention for proteins is lacking and hence, there is no restriction in defining a protein, giving rise to diverse names and definitions.

The best rule induced from association rule mining gives 2 false positives (FP% = 0.3) and 38 false negatives (FN% = 0.0038) (Figure 9). This rule is supported by 96.8% of the training record pairs, indicating the 96.8% of the training record pairs have identical sequence S(Seq)=1.0, of the same sequence length N(Seq length)=1.0, belongs to the same species M(Species) and are both not PDB records M(PDB) = 0:

S(Seq)=1 ^ N(Seq Length)=1 ^ M(Species)=1 ^ M(PDB)=0 → Duplicate



Rule 1	S(Seq)=1 ^ N(Seq Length)=1 ^ M(PDB)=1 (99.7%)
Rule 2	S(Seq)=1 ^ M(PDB)=0 ^ M(Species)=0 (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%)

Figure 9. FP% and FN% using association rules for duplicate detection. The values in brackets refer to the support measurements.

The experiment with duplicate detection using user-defined rules and association rules indicate that association rule mining can detect duplicates more effectively than the user rules.

The derivation of user rules for identifying duplicates requires good understanding of the data, and the combinations of criteria used are based on the user's knowledge of biological domain. With our method, the rules are automatically generated. Intuitively, this means that the domain understanding of the duplicates is mined from the training dataset rather than defined by the users. In the last experiment with duplicates, we have shown that best association rule (2 false positives, 38 false negatives) of the duplicate relation is more effective in identifying duplicates than the user rules (1 false positive, 79 false negatives). From the result, we also deduce that the key fields or criteria that can be used for comparing sequence annotations are (1) Sequence similarity, (2) Sequence length, (3) Species, and (4) Data sources.

6. CONCLUSION

With rapid growth of public biological data and fast development of computational methods based on mining of these data, achieving high quality datasets is becoming increasingly important for effective data mining. In this paper, we presented a novel method for data cleaning, specifically in duplicate detection, using association rule mining.

The paper achieved preliminary contributions to biological data cleaning. It explores scoring functions and criteria for matching sequence records. Also, it introduces a new method for modeling duplicate relations using association rules. The method is evaluated with rules defined manually by domain experts. The duplicate detection rules identified from this paper can be used for cleaning any protein sequence annotations.

This work focuses on the duplicate detection in a representative biological dataset using the Apriori method for association rule mining. Our future work in improving the duplicate detection method for large scale datasets will use this result as a basis.

7. REFERENCES

[1] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389-3402, 1997.

[2] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of Data*, Washington, D.C., United States, 1993, 207-216.

[3] Barker, W.C., Garavelli, J.S., Hou, Z., Huang, H., Ledley, R.S., McGarvey, P.B., Mewes, H.W., Orcutt, B.C., Pfeiffer, F., Tsugita, A., Vinayaka, C.R., Xiao, C., Yeh, L.S., Wu,C. Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res. 29*, 29-32, 2001.

[4] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. GenBank: update. *Nucleic Acids Res. 32*, 23-26, 2004.

[5] Bilenko, M. and Mooney, R.J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, D.C., United States, 2003, 39-48.

[6] Boeckmann, B., *et al*. The SWISS-PROT protein knowledge base and its supplement TrEMBL. *Nucleic Acids Res. 31*, 365–370, 2003.

[7] Borgelt, C. and Kruse, R. Induction of association rules: Apriori implementation. *14th Conference on Computational Statistics*, 2002.

[8] Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J. and Berman, H.M. The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res. 32*, 223-225, 2004.

[9] Cohen, W.W. and Richman, J. Learning to Match and Cluster Large High-Dimensional Data Sets For Data Integration. In Proceedings of the eighth ACM *SIGKDD internation conference on Knowledge discovery and data mining,* Edmonton, Alberta, Canada, 475-480, 2002.

[10] Kulikova, T., *et al*. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res. 32*, 27-30, 2004.

[11] Lu, R., Lee, M.L., Hsu, W. Using interval association rules to identify dubious values. 5th International Conference on Web-Age Information Management, 2004.

[12] Marcus, A., Maletic, J.I., Lin, K. (2001) Ordinal association rules for error identification in data sets. In *Proc. of the Tenth International Conference on Information and Knowledge Management (ACM CIKM 2001).*

[13] Mauricio, A.H. and Stolfo, J.S. (1995) The Merge/Purge Problem for Large Databases. In *Proceedings of the 1995 ACM SIGMOD Conference on Management of Data,* San Jose, California, United States, 1995, 127–138.

[14] Mauricio, A.H.and Stolfo, J.S. Real-world Data is dirty: Data Cleansing and the Merge/Purge Problem. *Data Mining and Knowledge Discovery*, 2(1), 1998, 9-27.

[15] Merge/Purge by Proxy from Sagent Solution. http://www.whitehat.com/whitehatpapers.cfm

[16] Mills, D.L. A new algorithm to determine the Levenshtein distance between t*wo* strings. *Proc. Workshop on String Matching*, 1979.

[17] Miyazaki, S., Sugawara, H., Gojobori, T., Tateno, Y. DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res. 31*(1), 13-16, 2003.

[18] Müller, H., Naumann, F. and Freytag, J-C. Data quality in genome databases. *Proceedings of the International Conference on Information Quality (IQ 2003)*, Boston, October 2003.

[19] Overton,C.G. and Haas,J. Case-Based Reasoning Driven Gene Annotation. *Computational Methods in Molecular Biology.* Elsevier Science, 1998.

[20] Srinivasan,K.N., Gopalakrishnakone,P., Tan,P.T., Chew,K.C., Cheng,B., Kini,R.M., Koh,J.L., Seah,S.H., Brusic,V. SCORPION, a molecular database of scorpion toxins. *Toxicon*, 40, 23-31, 2002.

[21] Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, 266, 141–162, 1996.

[22] Tatusova,T.A. and Madden,T.L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, 174, 247–250, 1999.

[23] Thanaraj, T.A. A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures. *Nucleic Acids Res. 27*(13), 2627-2637, 1999.

[24] Triplex Merge/Purge program, Direct Marketing Corporation. http://www.tdmc.com/merge_purge.html