Incremental Maintenance of Biological Databases Using Association Rule Mining

Kai-Tak Lam[1,a], Judice L.Y. Koh[2,3,b], Bharadwaj Veeravalli[1], and Vladimir Brusic[4]

[1] Department of Electrical & Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[3] School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 119260
[4] Australian Centre for Plant Functional Genomics, School of Land and Food Sciences, and the Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia
[a]althorz@hotmail.com, [b]judice@i2r.a-star.edu.sg

Abstract. Biological research frequently requires specialist databases to support in-depth analysis about specific subjects. With the rapid growth of biological sequences in public domain data sources, it is difficult to keep these databases current with the sources. Simple queries formulated to retrieve relevant sequences typically return a large number of false matches and thus demanding manual filtration. In this paper, we propose a novel methodology that can support automatic incremental updating of specialist databases. Complex queries for incremental updating of relevant sequences are learned using Association Rule Mining (ARM), resulting in a significant reduction in false positive matches. This is the first time ARM is used in formulating descriptive queries for the purpose of incremental maintenance of specialised biological databases. We have implemented and tested our methodology on two real-world databases. Our experiments conclusively show that the methodology guarantees an F-score of up to 80% in detecting new sequences for these two databases.

1 Introduction

In-depth analysis about a specific subject in molecular biology, specifically those associated with the structural and functional properties of a particular group of sequences typically requires access to an extensive knowledge base which may take the form of a specialist database. By integrating subject specific molecular information from public data sources such as GenBank and Swiss-Prot with data analysis tools, a specialist database facilitates the extraction of new knowledge of the topic under study for its users. Some examples of specialist databases include svNTX – a database of functionally classified snake neurotoxins [1], APD – an antimicrobial peptides database with their functional classification [2], Aminoacyl-tRNA synthetases database (AARS) – a database of AARS enzymes that carry out specific esterification of tRNAs [3], svPLA$_2$ – a database of snake PLA2 venoms [4], and a food allergen sequence database for assessing potential allergenicity in transgenic food [5].

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 140 – 150, 2006.

With biotechnological advancement and high-throughput sequencing, new sequences are rapidly accumulating in the public data sources. On the other hand, specialist databases created by the researchers are easily out-dated given the exponential growth of new data in the data sources. Frequent updating using simple queries (keywords and sequence searches) of the specialist databases are handicapped by the high number of chance matches and the need to filter them manually.

In this paper, we apply text and data mining techniques together with motif identification techniques to formulate complex queries that in turn, can be used to search the public data sources for the purpose of updating these specialist databases. The use of complex queries which are "machine-learned" from a given specialist database, as opposed to user-defined simple queries, reduces the number of chance matches in the database updating process.

With accurate retrieval of new records of high relevance, the method is a crucial step towards enabling automatic incremental updating of any specialist database. Particularly in a biological data warehousing system such as BioWare which comprises of a number of specialist databases organised around different topics [6], a general method for incremental updating of any specialist database reaps great benefits to its users.

We will first present a brief review of Association Rule Mining (ARM) which is predominantly used in the mining of frequent patterns in a database. In section 2, we present our proposed automated queries formulation method. In section 3, we evaluate the performance and present the results for two specialised biological databases. And we explore other applications of this method in the concluding section 4.

1.1 Association Rule Mining

Since its introduction in 1993, ARM [7] has been widely used for market basket analysis in the finance sector, and in other applications such as the identification of gene expression in bioinformatics domain [8].

In ARM, the *support* of each itemset - percentage of occurrences in the database is computed. Itemsets with supports higher than the user-specified *minimum support* (*minsup*) are identified as *frequent itemsets.* Many association rule mining algorithms are variants of Apriori [9], which employs a bottom-up, breadth-first search enumerating every frequent itemset.

In this paper, we are interested in identifying a special type of frequent itemsets, known as the *maximal frequent itemsets* (MFS) which contains the maximum number of items that achieve the *minsup*. This group of frequent itemsets has the property that any addition of item into the MFS will cause the support of the set to fall below the *minsup*.

2 Proposed Methodology

A specialist sequence database can be characterized by the textual features as well as sequence features of the sequence records it contains. Textual features refer to the key terms present in the textual attributes of the database records. For example, the term "phospholipase A2" occurs in a significant number of records in svPLA$_2$, the

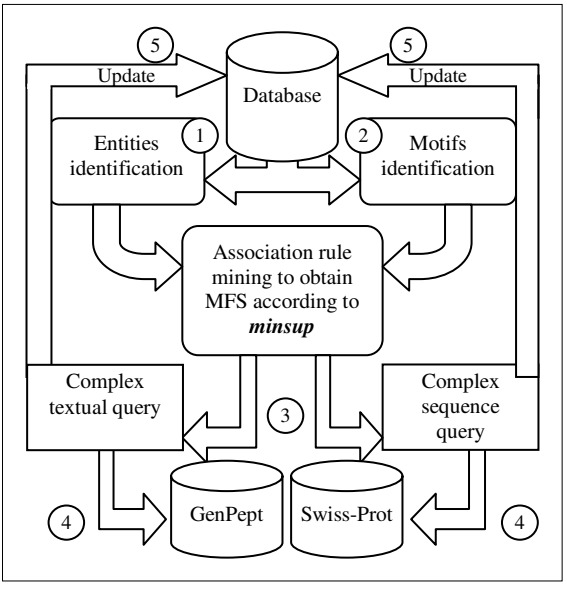


Fig. 1. Flow chart of the proposed methodology. The number in circles denotes the order of the steps required for the formulation of complex queries and how these queries are used in updating the specialist database.

snake PLA_2 venom database. And sequence features, more commonly referred to as motifs, are sequences that characterised certain biochemical functions. One example of sequence features is the so-called zinc finger motif, CXX(XX) CXXXXXXXXXXXXHXXXH, which is found in widely varying families of DNA-binding proteins.

In this paper, we explore the extraction of both textual and sequence features from a given database to formulate complex queries which in turn, are used for incremental retrieval of new sequences of high relevance to the database. Figure 1 shows the primary steps in our proposed methodology. First, textual (entities) and sequence (motifs) features, and their supports in the sequence records, are extracted from the input database or dataset through the identification engines. Each feature corresponds to an item in the frequent itemsets mining process, and each record is a transaction. Using the *Apriori* program [10] and given a user-defined *minsup*, the MFS of the database are generated. Each MFS is formulated, using the Boolean operators, into a complex query which is utilised to search for the new sequences in the data sources. The method is evaluated based on the relevance of the new sequences retrieved.

2.1 Entities Identification

For entities identification, a biological named-entities recogniser (BNER) [11] is used to identify biologically significant words and phrases, along with the complementary elimination of stop-words. Selected textual fields of the database, including reference title, species, and taxonomy, are parsed by the entities identification engine.

We conduct a comparative study of two publicly available BNER programs - PowerBioNE (PBNE) published by Zhou *et. al.*, [12], and ABNER developed by Settles [13]. The efficacy of the BNER programs depends on its *Corpus*. PBNE uses the GENIA Corpus V3.0 [14] which contains 2000 MEDLINE abstracts of 360K words. ABNER, on the other hand, uses the NLPBA corpus [15], a variant of the GENIA corpus, and the BioCreative corpus [16].

In our experiment with the two specialist databases, ABNER performed consistently better than PBNE (results not shown, but is available on request). The entities extracted are sorted and arranged according to their occurrences in the databases.

2.2 Motifs Identification

For motifs identification, we utilize the MEME system for the detection of motifs [17]. MEME is an unsupervised learning algorithm based on a combination of Hidden Markov models, expectation maximization (EM), an EM-based heuristic for choosing the starting point for EM, a maximum likelihood ratio-based heuristic for determining the best number of model free parameters, multistart for searching over possible motif widths, and greedy search for finding multiple motifs.

Since motifs are highly specific biological patterns, their identification, and the generation of queries from these motifs, are typically unique for different databases or datasets.

2.3 Query Formulation

Maximal frequent feature sets in the database are mined through the use of *Apriori* with the parameters: minimum number of items per set is 2, maximum number of items per set is 25. The features within each MFS are combined using "AND" and different MFS are consolidated using "OR". Figure 2 shows an example of the complex query extracted from a snake venom database.

phospholipase AND chordata AND colubroidea AND craniata AND euteleostomi AND lepidosauria AND metazoa AND scleroglossa AND serpentes AND squamata AND vertebrate (96.2%)

"phospholipase a" AND phospholipase AND vertebrata AND squamata AND serpentes AND scleroglossa AND metazoa AND lepidosauria AND euteleostomi AND craniata AND colubroidea AND chordate (96%)

CNPKLDTYSYSCxNG AND RPWWSYADYGCYCGWGGSGTPVDALDRCCFVHDCC YGKAEK (86%)

RFVCECDRAAAICFADNPYTYN AND RPWWSYADYGCYCGWGGSGTPVDALDRCCFVHDCC YGKAEK (85.3%)

Fig. 2. Example combinatory query formulated from svPLA2 database

2.4 Updating of Database

As shown in Figure 1, the incremental maintenance of a specialist database is ideally an iterative process of query formulation, searching of new sequences in the data sources, and updating them to the database. Users have the option of using queries from entities identification, motif identification, or a combination of both. In our experiment, we concluded that any of these three approaches reduces the false positive (irrelevant) records retrieved and thus the number of records that need to be filtered manually.

One of the unique strengths of this methodology lies in the combined use of motifs and textual entities in characterizing a sequence database or dataset in a simple and efficient manner.

2.5 Performance Metrics

The performance of our proposed methodology is quantified and measured by the *Precision, Recall and, the F-score* metrics:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2}$$

$$\text{F-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

where, *TP,FP* and *FN* are true positives, false positives and false negatives respectively.

In our experiments, the *TP*s of a formulated query refer to retrieved records from the data sources using the query which are also found in the original database. *FN*s are database records not retrieved. And *FP*s are non-database records retrieved using the query.

Precision measures the fraction of records retrieved by the complex queries which are relevant to the input database. Recall measures the fraction of relevant records which are retrieved. F-score combines them into a single value for purpose of comparison.

3 Performance Evaluation and Discussions

The databases used for performance evaluation are Snake Venom PLA_2 ($svPLA_2$) [4] and Food Allergen [5]. The $svPLA_2$ database contains 289 functionally annotated, non-redundant $svPLA_2$ toxins used for the studying of the pharmacological effect of these toxins and for supporting detailed structure-function analysis. Sequences in the $svPLA_2$ database were retrieved from GenPept and Swiss-Prot using the simple query "serpentes AND phospholipase OR pla2", followed by manual filtering to remove chance matches and fragmented sequences.

The Food Allergen database contains 633 unique protein sequences that are used in the analysis of allergenicity in transgenic food. This database is used in the monitoring of the possible allergic reactions towards genetically modified food.

In our experiments, the original entries in the specialist database are used as the set of TP. The retrieved records are first filtered by the date at which the database was most recently updated. After which, these filtered records are compared with the entries in the specialist database. The matched records are the *TP*s while the unmatched records are the *FP*s. Those entries in the database that do not have a match are the *FN*s.

3.1 Snake Venom PLA_2 Database

We compare the performance of the complex queries to the simply query used by the biologists when constructing the database. Generally, a higher F-score indicates that the query is more efficient in identifying sequences in the data sources relevant to snake PLA_2 venoms. In addition, we are interested to find out if a combinatory approach of using both textual and motif features result in a more accurate query, compared to using only textual or sequence-based queries.

Complex Textual Query. A total of 1268 key terms are identified using ABNER program. As the optimal *minsup* is unknown, the precision and recall are computed at varying *minsup* values. As shown in Figure 3a, complex textual query at the *minsup* of 96% has an F-score of 74%, a great improvement over the original query of 50%. From Table 1, it can be seen that the number of records retrieved using the complex textual query is about 13% less than that of the original simple query. This shows that it is more efficient to update this database by applying the methodology proposed than using the original simple query.

Complex Sequence Query. A total of 50 motifs are identified using MEME program. MFS according to different *minsup* are generated and submitted to BLAST for the retrieval of records from the protein database in NCBI and the retrieval results are listed in Table 1. The relationship between F-score and *minsup* for these retrievals is plotted in Figure 3b. The F-score for these retrieval are lower than that of the original query, with the highest being 39%. However, as the snake venom PLA2 was not initially retrieved using motif queries, comparison made with the original query may not be fair.

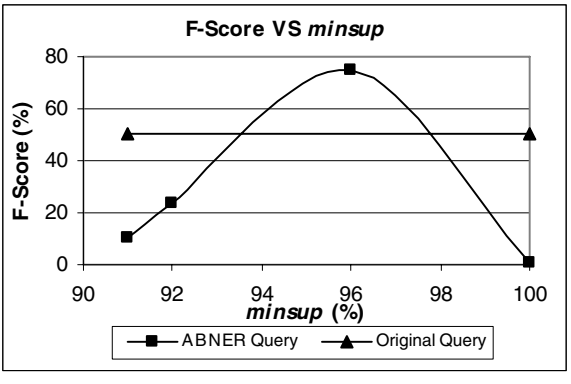
Also, the efficacy of the BLAST search is dependent on the choice other parameters, such as the substitution matrices and the gap cost. As the lengths of the motifs identified are mostly less than 35, we use PAM30 as the substitution matrix and select a gap cost of 10 so as discourage the introduction of gaps within the motifs. The selection of other matrices and gap cost combinations may result in further optimisation of the F-score. For this, further investigation may be required.

Combinatory Query. An investigation is made on the effect of combining both entities and motifs queries on the retrieval result. The queries that give the best F-score are chosen and are submitted to NCBI to retrieve the relevant records. The results are as listed in Table 1.

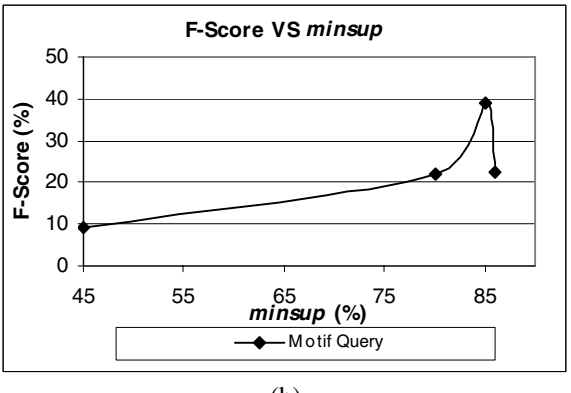
An F-score of 85% is achieved using a combination of entity and motif queries, giving a 35% improvement over the original query. This indicates that the method gives a much better result on retrieval using both textual or sequence queries, as opposed to using either only.

Table 1. Results of varying *minsup* using different queries for the svPLA$_2$ database. The simple query is the same query that is used during the creation of the database [4]. The bolded values denote the best F-Score achieved in the respective queries. These are then used to form the combinatory query.

	Complex textual query				Complex sequence query				Combinatory query	Simple query
minsup (%)	91	92	**96**	100	45	80	**85**	86	-	-
No. of records retrieved	49	53	**197**	26	53	135	**294**	139	**231**	226
Precision (%)	35	75	**92**	3.8	30	35	**38**	35	**95**	58
Recall (%)	5.9	14	**63**	0.35	5.5	16	**39**	17	**76**	45
F-score (%)	10	23	**74**	0.63	9.4	22	**39**	22	**85**	50



(a)



(b)

Fig. 3a and 3b. F-score of textual and sequence queries at varying *minsup* for svPLA$_2$

3.2 Food Allergen Database

Complex Textual Query. A similar experiment is carried out on the Food Allergen database. A total of 734 key terms are identified using ABNER program. The results

are listed in Table 2 and the corresponding Figure 4 shows the trend of F-score versus the varying *minsup*. From Table 2, we can see that in general, the recall of the retrieval is above 50%, with the exception when *minsup* is 12%. This shows that more that half of the positive records are retrieved using this textual-based queries. However, we are only able to achieve a maximum F-score of 4.5% when the *minsup* is 7%.

The low F-score is mainly due to the diversity of textual information in the database. As there are no textual entities that occur in abundance (more than 20%) in the database, a relatively small *minsup* has to be used. This query using entities alone has retrieved a large number of records, thereby contributing to a higher number of chance retrieval and hence a low F-score.

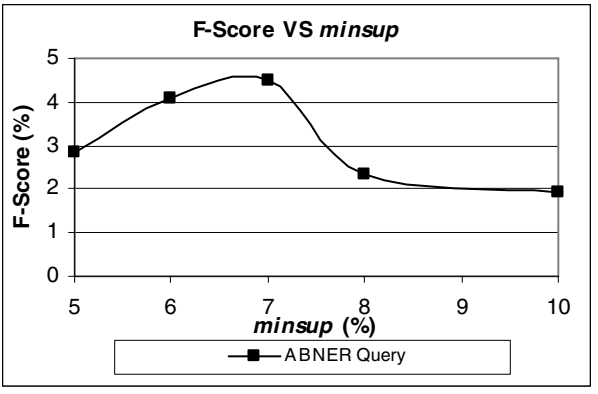
This experiment on the Food Allergen database exhibits the shortcoming of queries based on entities alone. Some specialist database, even though it is restricted to a certain domain, may contain very general textual information. If only textual information is used, one may be faced with a very large amount of chance matches, which amounts to 18K of records in the Food Allergen database that need to be reviewed manually. This may be improved with using queries based on motifs alone, as show in our investigation in the next section.

Complex Sequence Query. Although using query based on motifs alone we are able to obtain an F-score of 41%, the recall of the retrieval suffers. Less then half of the positive records are retrieved. However, using motifs alone, we are able to minimise the number of chance matches, i.e. decreasing the *FP* in our retrieval. Since there are shortcoming for both types of queries based on entities and motifs alone, we will further investigate if a combination of both would yield a better result in the next section.

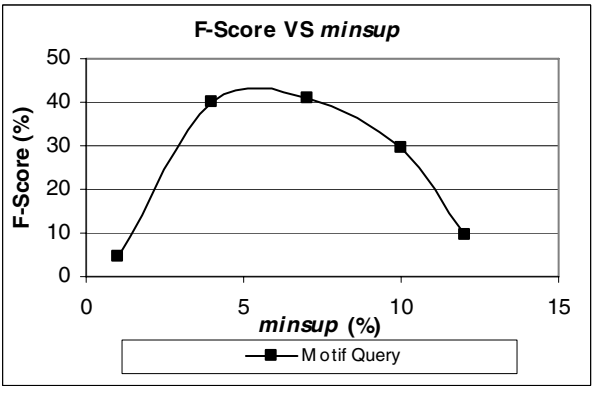
Combinatory Query. Textual query at *minsup* of 7% and sequence query at *minsup* of 7% are used together for combinatory retrieval. An F-score of 80% is achieved and both the precision and recall achieve a score of more than 50%. This demonstrates that the combination of both entity-based and motif-based queries gives a must better results than using either one alone.

Table 2. Results of varying ***minsup*** using different queries for the Food Allergen database. A comparison with simple query is not carried out as the simple query used for the creation of the original database is not availible. The bolded values denote the best F-Score achieved in the respective queries. These are then used to form the combinatory query.

	Complex textual query					**Complex sequence query**					**Combinatory query**
minsup (%)	5	6	**7**	8	10	1	4	**7**	10	12	-
No. of records retrieved	18K	10K	**9.6K**	18K	18K	13	245	**119**	187	41	**283**
Precision (%)	1.5	2.1	**2.3**	1.2	0.99	77	55	**93**	48	54	**100**
Recall (%)	63	52	**53**	52	43	2.4	32	**26**	21	5.2	**67**
F-score (%)	2.9	4.1	**4.5**	2.3	1.9	4.6	40	**41**	29	9.4	**80**



(a)



(b)

Fig. 4a and 4b. F-score of textual and sequence queries at varying *minsup* for the Food Allergen database

4 Conclusions

The task of database maintenance is a time-consuming process due to the ever-increase size of public data sources and the large number of chance matches using simply query method. In this paper we have proposed a methodology with can be used to formulate complex queries from a database based on the textual and sequence features that characterised the database. We have shown that these queries are able to reduce the number of *FP* records from the retrieval and hence reduce the amount of time and effort required to manually filter the retrieval results during database maintenance.

This is the first time ARM is used in formulating complex queries for the purpose of incremental maintenance of specialised biological databases. Tested on two real-world databases, our methodology shows that an F-score of up 80% is achieved.

At this current stage, many of the parameters such as *minsup* and the minimum and maximum number of items per MFS are determined empirically. Further work can be

done on finding these values by machine learning techniques. For example, the optimal *minsup* can be found automatically by using internal cross validation. This would ease the burden from the user to determine a few arbitrary *minsup* and observe from the result which one is closer to the optimal value.

This methodology can be integrated in a biological data warehousing system such as BioWare for the incremental maintenance of the existing databases. Furthermore, information retrieval of PubMed records based on a sample of PubMed articles can be carried out more efficiently using the entity query formulation portion of our methodology. This is useful for initial research work where information retrieval from the public domain is essential.

References

1. Siew, J.P., Khan, A.M., Tan, P.T., Koh, J.L. , Seah, S.H., Koo, C.Y., Chai, S.C., Armugam, A., Brusic, V., Jeyaseelan, K., "Systematic analysis of snake neurotoxins functional classification using a data warehousing approach", *Bioinformatics*, 20(18), 2004, pp. 3466-3480.
2. Wang, Z. and Wang, G., "APD: the Antimicrobial Peptide Database", *Nucleic Acids. Res.* 32, 2004, pp. 590-592.
3. Szymanski, M. and Barciszewski, J., "Aminoacyl-tRNA synthetases database Y2K", *Nucleic Acids Res.* 28, 2000, pp. 326–328.
4. Tan, P.T.J., Khan, A.M. and Brusic, V., "Bioinformatics for venom and toxin sciences", *Brief Bioinform.* 1, 2003, pp. 53-62.
5. Gendel, S.M., "Sequence Databases for Assessing the Potential Allergenicity of Proteins Used in Transgenic Foods", *Advances in Food and Nutrition Research*, v42 1998, pp. 63-92.
6. Koh, J.L.Y., Krishnan, S.P.T, Seah, S.H., Tan, P.T.J., Khan, A.M., Lee, M.L., Brusic, V., "BioWare: A framework for bioinformatics data retrieval, annotation and publishing", *SIGIR'04 workshop on Search and Discovery in Bioinformatics*, July 29, 2004, Sheffield, UK
7. Agrawal, R., Imielinski, T., Swami, A., "Mining association rules between sets of items in large databases", *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Washington, D.C., United States, 1993, pp. 207-216.
8. Creighton, C. and Hanash, S., "Mining gene expression databases for association rules", *Bioinformatics*, 19(1), 2003, pp. 79-86.
9. Agrawal, R., Srikant, R., "Fast algorithms for mining association rules", *The International Conference on Very Large Databases*, 1994, pp. 487-499.
10. Borgelt, C., Kruse, R., "Induction of Association Rules: Apriori Implementation", *15th Conference on Computational Statistics*, Physica Verlag, Heidelberg, Germany, 2002.
11. Ananiadou, S., Friedman, C., Tsujii, J., "Introduction: named entity recognition in biomedicine", *Journal of Biomedical Informatics*, 37(2004), pp. 393-395.
12. Zhou, G.D., Zhang, J., Su, J., Shen, D., Tan, C.L., "Recognizing Names in Biomedical Texts: a Machine Learning Approach", *Bioinformatics*, v20(7) 2004, pp. 1178-1190.
13. Settles, B., "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text", *Bioinformatics*, v21(14) 2005, pp. 3191-3192.
14. Ohta, T., Tateisi, Y., Kim, J., Mima, H., Tsujii, J., "The GENIA corpus: an annotated research abstract corpus in molecular biology domain", *Proceedings of Human Language Technology (HLT' 2002)*, San Diego, pp. 489-493.

15. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N., "Introduction to the bio-entity recognition task at JNLPBA", *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA),* Geneva, Switzerland, 2004, pp. 70-75.
16. Yeh, A., Hirschman, L., Morgan, A., Colosimo, M., "BioCreAtIve Task 1A: gene mention finding evaluation", *BMC Bioinformatics*, 6(Suppl 1):S2, 2005.
17. Bailey, T.L., Elkan, C., "The Value of Prior Knowledge in Discovering Motifs with MEME", *ISMB*, v3 1995, pp. 21-29.