BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 16 2012, pages 2162–2168
doi:10.1093/bioinformatics/bts367

Data and text mining Advance Access publication July 2, 2012

AutoBind: automatic extraction of protein–ligand-binding affinity data from biological literature

Darby Tien-Hao Chang[1], Chao-Hsuan Ke[2], Jung-Hsin Lin[3,4] and Jung-Hsien Chiang[2,*]

[1]Department of Electrical Engineering, [2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, [3]School of Pharmacy, National Taiwan University, Taipei 10051 and [4]Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Determination of the binding affinity of a protein–ligand complex is important to quantitatively specify whether a particular small molecule will bind to the target protein. Besides, collection of comprehensive datasets for protein–ligand complexes and their corresponding binding affinities is crucial in developing accurate scoring functions for the prediction of the binding affinities of previously unknown protein–ligand complexes. In the past decades, several databases of protein–ligand-binding affinities have been created via visual extraction from literature. However, such approaches are time-consuming and most of these databases are updated only a few times per year. Hence, there is an immediate demand for an automatic extraction method with high precision for binding affinity collection.

Result: We have created a new database of protein–ligand-binding affinity data, AutoBind, based on automatic information retrieval. We first compiled a collection of 1586 articles where the binding affinities have been marked manually. Based on this annotated collection, we designed four sentence patterns that are used to scan full-text articles as well as a scoring function to rank the sentences that match our patterns. The proposed sentence patterns can effectively identify the binding affinities in full-text articles. Our assessment shows that AutoBind achieved 84.22% precision and 79.07% recall on the testing corpus. Currently, 13 616 protein–ligand complexes and the corresponding binding affinities have been deposited in AutoBind from 17 221 articles.

Availability: AutoBind is automatically updated on a monthly basis, and it is freely available at http://autobind.csie.ncku.edu.tw/ and http://autobind.mc.ntu.edu.tw/. All of the deposited binding affinities have been refined and approved manually before being released.

Contact: jchiang@mail.ncku.edu.tw

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on December 30, 2011; revised on June 8, 2012; accepted on June 25, 2012

1 INTRODUCTION

Determination of the binding affinity is essential for quantifying the protein interactions with other bio-molecules. For example, actually estimating the protein–DNA-binding affinity is an important step in understanding the regulation mechanisms of cells (Chen *et al.*, 2012; Chien *et al.*, 2012; Morozov and Siggia, 2007). In the field of drug design, the binding affinity of a ligand to a given protein target is used to judge whether this ligand can be considered as a potential drug candidate.

Several databases of protein–ligand-binding affinities have been proposed, including Ligand-Protein Database (LPDB) (Roche *et al.*, 2001), Protein Ligand Database (PLD) (Puvanendrampillai and Mitchell, 2003), PDBbind (Wang *et al.*, 2004), Binding MOAD (Hu *et al.*, 2005), AffinDB (Block *et al.*, 2006) and BindingDB (Chen *et al.*, 2002; Liu *et al.*, 2007). These databases depend on structure databases (Berman *et al.*, 2000; Chang *et al.*, 2012) and require much time and cost for the manual inspection of the literature. Until September 2011, more than 21 million articles indexed contain the descriptions of binding affinity in the MEDLINE database. The time-consuming manual extraction cannot timely accommodate the astonishingly increasing number of full-text articles. Therefore, there is an immediate demand for automatic extraction of protein–ligand-binding affinities.

Information Retrieval (IR) is a computational technique aiming at extracting information from the literature without manual intervention (Krallinger and Valencia, 2005). Summarization is a widely seen IR application in various areas, which can automatically compile extractive summaries that include the important aspects of full-text articles (Bhattacharya *et al.*, 2011). In the biomedical field, this technique is used to generate gene annotations (Chiang *et al.*, 2006) and to search bio-images (Agarwal and Yu, 2011). In recent years, IR has been used to extract biological entities of interest and their relationships (Rindflesch *et al.*, 2000). For example, 'activate' and 'interact' are two common relation terms that connect molecule entities. This technique has been applied to protein–protein interactions (Blaschke *et al.*, 1999; Koike and Takagi , 2005; Zhou and He, 2008), pathway networks (McDonald *et al.*, 2004) and drug interactions (Tari *et al.*, 2010).

However, so far there have been only a few studies that attempted to extract biological values or parameters. For example, Spasić *et al.* used terminology concept to index and select literature, then a score function was implemented to rank the retrieved articles by relevance of parameters for kinetic modeling of yeast metabolic pathways (Spasić *et al.*, 2009); Heinen *et al.* used a rule-based approach to collect kinetic information of enzymes (Heinen *et al.*, 2010). In general, the experimental procedures or results described in full-text articles are hard for researchers to re-use (Névéol *et al.*, 2011). The fact that comparisons with events, values and parameters are usually recorded in a more flexible way in the literature makes the

[*]To whom correspondence should be addressed.

development of value extraction approaches more difficult (Milo *et al.*, 2010).

In this article we describe the construction of AutoBind, an automatic information extraction system dedicated to protein–ligand-binding affinity. Our system includes four patterns to identify candidate sentences describing protein–ligand-binding affinity and a scoring function to rank the identified sentences for extracting binding affinities. We applied AutoBind on the primary citations of all entries in the Protein Data Bank (PDB). Many primary citations contain multiple affinities corresponding to different experimental conditions, such as wild-type and mutated proteins, but only one represents the actual binding affinity of the crystal structure. In this study, we focused on the affinity corresponding to the structure in each primary citation since 1996, whose full-text references can be easily retrieved. Finally, the extracted protein–ligand-binding affinities and the corresponding complex structures were organized as a public database named AutoBind. Note that AutoBind in this article refers both the extraction algorithm and the database. The extraction algorithm is completely automatic. However, the binding affinities deposited in the database have been approved manually, in order to refine them and eliminate possible errors and inconsistencies. On the other hand, all the reported performances (in Tables 3–6 and Sections 4.1 and 4.2) were evaluated on the results obtained before the manual approval. In short, this article presents an automatic extraction algorithm to expedite constructing and maintaining a database, but the database construction was not fully automatic.

The AutoBind database currently contains 13 616 protein–ligand complexes. To our knowledge, this is the largest collection of protein-binding affinities. The database will be updated regularly.

2 BINDING AFFINITY RECORD

In this section, we describe the definition of a binding affinity record in this study, which frames the data that AutoBind extracted. The binding affinity is a measure of the effectiveness of a compound in connecting other molecules. It is thermodynamically quantified as dissociation constant (K_d), inhibition constant (K_i) or half maximal (50%) inhibitory concentration (IC_{50}). The detailed definitions of these measures can be found in the Supplementary Data.

We carried out an analysis on a set of full-text articles collected from the PDBbind database (Wang *et al.*, 2004), denoted *PDBbind_corpus*, where the actual binding affinities as well as the associated protein–ligand complexes were known. In this corpus, we observed that a description of a binding affinity has four major objects:

- <PROTEIN>: functional polypeptides or macromolecules.
- <LIGAND>: the meta-ions and small organic/inorganic ions and organic solvent molecular.
- <EVENT>: the interaction/relation between the protein and ligand.
- <VALUE>: binding affinity.

Example 1 shows an example of describing protein-binding interaction and their binding affinity, where all the four objects (italicized) appeared in the same sentence.

Example 1 The lack of electron density for the adenosine moiety of the *TP4A*<LIGAND> molecule *bound to*<EVENT> *hTK1*<PROTEIN> is surprising, taking into account the very high affinity of the bisubstrate analog for *hTK1*, with a K_d *of 29 nM*<VALUE>. (PMID: 17407781)

However, authors frequently use variable writing styles to make the articles fluent. Example 2 shows a description of protein–ligand binding without precisely describing the <LIGAND>.

Example 2 For the wild-type *PARN-RRM*<PROTEIN>, a K_d *value of 6.94* ±*2.08* μM<VALUE> was determined. (PMID: 18694759)

Therefore in this study we have relaxed the requirement widely used in previous studies that an effective description of a relation between two entities must contain both entities (Bui *et al.*, 2011; Jang *et al.*, 2006). However, if only considering the sentence of Example 2, it is difficult to know the ligand. In AutoBind, the PDB TITLE, which was written by the author of the structure to describe the main molecules in the structure, is adopted to overcome this problem. The PDB TITLE corresponding to Example 2 (PDB ID: 3CTR) is ‘Crystal structure of the RRM-domain of the poly(A)-specific ribonuclease PARN bound to m7GTP’. The protein and ligand names are *PARN* and *m7GTP*, respectively. Thus, AutoBind can restore the missing ligand name in Example 2 to *m7GTP*.

It is worthwhile to note that an article may contain multiple descriptions of binding affinities but some of them correspond to ligands not of interest. The biologically relevant ligand in this work is defined as (i) non-covalently bound to the protein and (ii) formed an unambiguously complex with the protein. However, in some cases, two ligands bind together inside the same binding site on the protein. In such cases, experimental measurement of the binding affinity of each individual ligand becomes much more complicated. In a PDB file, there could be some biologically irrelevant ligands [e.g. GOL (glycerol) in PDB entry 3GBA, where glycerol was added to the solution as a cryoprotectant] that were crystallized along with the main molecules. In the TITLE records of the 1486 PDB entries of the *PDBbind_corpus*, only 58 do not describe the biologically relevant ligands and none of them describes biologically irrelevant ligands. Thus in practice, we resorted to the TITLE record in a PDB file, which was written by the author of the structure to describe the main molecules in the structure, to identify the biologically relevant ligands. For example, Examples 3 and 4 are two descriptions of protein–ligand-binding affinity in the same primary citation of a complex structure (PDB ID: 2HAW; PMID: 17095506). Example 3 describe the affinities of *PPase* binding with Mg^{2+} and Mn^{2+} ions, both of which are not the ligand of interest for family II *PPase*.

Example 3 Interestingly, Mg^{2+} <LIGAND> confers lower activity but greater substrate-binding affinity *on*/PREP family II *PPases*<PROTEIN> than Mn^{2+} <LIGAND> (K_d *of* 60 μM versus 180 μM)<VALUE>, whereas, in family I enzymes, the reverse is true.

In contrast, Example 4 shows the K_d value (60 μM) of family II *PPase* binding with *PNP*, which is the ligand of interest in the

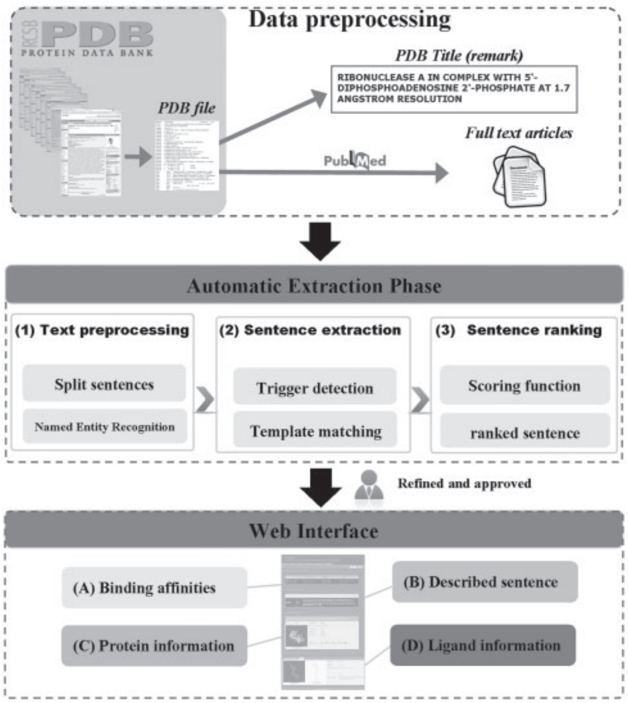

Fig. 1. System architecture of AutoBind. The extraction phase of AutoBind includes three stages: (1) text preprocessing, (2) sentence extraction and (3) sentence ranking. The result page of AutoBind includes four parts: (**A**) binding affinities extracted by AutoBind along with those deposited in other published databases, (**B**) full-sentenced descriptions where the binding affinity were extracted and the information of (**C**) the protein and (**D**) the ligand in the binding

article and conforms to the PDB TITLE 'Crystal structure of family II Inorganic pyrophosphatase in complex with PNP'.

Example 4 The dissociation constant for *PNP*<LIGAND> *binding*<EVENT> to family II *PPases*<PROTEIN> (12 *μM*)<VALUE> is reasonably close to that for *PPi*<PROTEIN> (*60 μM*)<VALUE>, in agreement with their very similar three-dimensional structures.

In this study, Example 4 was treated as an effective binding affinity and was collected in AutoBind; while Example 3 was not.

3 MATERIALS AND METHODS

The PDB release in September 2011 contained 76 114 entries from 32 986 primary citations, which are the original published articles of the affinity measurement for the corresponding protein–ligand complex. We collected 17 221 articles whose full text is publicly accessible. The binding affinities of 1586 articles of the 17 221 primary citations were annotated in the PDBbind database release of September 2011. We named the 1586 full-text articles as *PDBbind_corpus* and adopted the corresponding binding affinities as a gold standard to evaluate the performance of our automatic extraction method. We used 100 articles from *PDBbind_corpus* to construct the four patterns (see Section 3.2.2) and used the remaining 1486 articles as the testing corpus.

The workflow of constructing AutoBind is shown in Figure 1. In automatic extraction phase, it includes three major stages: (i) text preprocessing, (ii) sentence extraction and (iii) sentence ranking. AutoBind also provides a friendly web-based interface for users to query and view the identified binding affinity. The details of each stage in Figure 1 are described in the following subsections. The last subsection describes a naive extraction algorithm we implemented as a comparison baseline.

3.1 Text preprocessing

AutoBind first splits full text into sentences and marks the protein and ligand names in sentences using named entity recognition (NER). We used a protein name recognizer, NLProt (Mika and Rost, 2004), and a lexicon, UniProt (Bairoch *et al.*, 2005), to recognize protein names. If a term is (i) recognized as protein name by NLProt or (ii) identical to a protein name/synonym in UniProt, it is marked as a protein name. Namely, AutoBind used the union of the results of NLProt and UniProt as the protein names. Similarly, we used a compound name recognizer, OSCAR4 (Jessop *et al.*, 2011), and a chemical molecules databases, PubChem (Wang *et al.*, 2009), to recognize ligand names. As protein recognition, the union of the results of OSCAR4 and PubChem was used as the ligand names. The detailed information about the used NER tools can be found in the supplementary data.

To solve the problem of name variants, including orthographic variation (e.g. 'EBNA3', 'EBNA-3' and 'EBNA 3'), Greek alphabet letter (e.g. 'p38 alpha' and 'p38 α') and extra words (e.g. 'TRAM' and 'TRAM protein'), AutoBind converted Greek alphabet letters to the corresponding English letters and removed three specific words 'protein', 'gene' and 'product'. Abbreviation discrimination is another critical problem in NER (Zhou *et al.*, 2006). In AutoBind, if a name is followed by a parentheses-surrounded term, this term is dynamically labeled as an abbreviation to that name. The protein and ligand abbreviations are considered identical to the protein and ligand names in the following stages.

Frequently, there are multiple binding affinities in a full-text article that, for example, correspond to the binding under different experimental conditions or the binding of the wild-type and mutated proteins. To identify the target protein and ligand of the extracted article, however, is difficult and requires much domain knowledge. Examples 3 and 4 in the previous section demonstrated that the title of the complex structure is a sentence well describing the crystal structure. In this study we focus on the affinity corresponding to the crystal structure and adopted the PDB TITLE to make the step of selecting protein and ligand in an article automated. Therefore, we also used NER to recognize protein and ligand names in the PDB TITLE. The recognized names were used to select binding affinity records later. The benefit of this strategy is that the extracted binding affinity is guaranteed to have a corresponding protein–ligand complex structure.

3.2 Sentence extraction

3.2.1 Trigger detection Trigger is usually a noun, noun phrase (NP), verb or verb phrase (VP), used to detect relation or event between entities in a sentence (Stapley and Benoit, 2000). Setting a collection of triggers is a useful practice to rapidly filter large sentences in a full-text article before deeper parsing. AutoBind defined two types of triggers to detect <VALUE>, T_{value}, and <EVENT>, T_{event}. Table 1 shows the 19 T_{value} and Table 2 shows the seven T_{event} identified by examining *PDBbind_corpus* manually. Each word in T_{event} stands for all its morphological variations.

In AutoBind, <VALUE> is more precisely defined by four components: trigger, sign, value and unit, such as '$K_d = 26\ \mu M$' ('K_d': trigger; '=': sign; '26': value and 'μM': unit). The four components are not required to be consecutive (e.g. 'shows a K_d value of 0.77 μM' in PMID: 18596699 is a valid <VALUE> in AutoBind). To simplify the following steps, AutoBind rearranges <VALUE> to its compact form as follows:

(1) Identify the position of T_{value} in the sentence.

(2) Extract the words from the position of T_{value} to the first encountered unit term (mM, μM, nM and pM).

(3) If all of trigger, value and unit are present in the extracted words, AutoBind rearranges them into the form of 'trigger = value unit'.

Here we use the sentence 'We first showed that Ap5G is an efficient inhibitor of EcGMPK, with an IC_{50} of about 0.5 μM' (PMID: 16690197) as

Table 1. A set of triggers (T_{value}) for recognition of protein-binding affinity data

K_d, K_i, IC_{50}, EC_{50}, K_{ii}, K_a, K_m, K_{diss}, pK_i, pK_d, pK_a, dissociation constant,
inhibition constant, association constant, 50% inhibitory concentrations,
half maximal effective concentration, binding constant(s), Michaelis constant(s),
Kinetic Parameter(s)

Table 2. Directions of semantic relations (T_{event})

Event word
Inhibit (ion, ed, ive) + with/at
Bind(s/ bound(s) + with/at/to
Interact (ion, ed, ive, ved) + with/at/to
Mutant + with/at/to
Complex + with/to
Activat (ion, e, ed)
React (ion, ed, ive) + with/at/on

an example. The trigger 'IC_{50}' is first identified. Then 'IC_{50} of about 0.5 μM' is extracted and parsed. The final <VALUE> in this example is rearranged as '$IC_{50} = 0.5\mu$M'.

3.2.2 Pattern matching After trigger detection, the remaining sentences that contain triggers are matched to the following patterns. The four patterns were based on manually reviewing 100 articles randomly selected from the *PDBbind_corpus*. In each of the pattern-based sentence patterns listed below, a symbol '-' indicates a word gap that can be filled with any number of words within the sentence. A token followed by a symbol '?' indicates that the token is optional.

Pattern 1	<P/L> – <EVENT> – <P/L> - BEV/PREP - <VALUE>
Example 5	The enzyme from *S. pombe* was the first described *lumazine synthase*<PROTEIN> that was found to *bind*<EVENT> *riboflavin* <LIGAND> *with*/PREP relatively high affinity (K_D of 1.2 μM) <VALUE>. (PMID: 12083520)
Example 6	Kinetic assays showed that *pteroic acid*<LIGAND> could *inhibit*<EVENT> *RTA*<PROTEIN> activity *with*/PREP an apparent K_i *of 0.6 mM* <VALUE>. (PMID: 9086280)

Pattern 1 accommodates for a typical description of a protein combined with a small molecular. The <P/L> (short of <PROTEIN>/<LIGAND>) indicates the protein/ligand names recognized by NER, <EVENT> indicates triggers recognized by T_{event} and <VALUE> indicates triggers recognized by T_{value}. A specific rule for the patterns containing two <P/L>s is that one must be <PROTEIN> and the other must be <LIGAND> with no order restrictions. BEV (to be verb) or PREP (preposition) connects <PROTEIN> and <LIGAND> to <VALUE>. Example 5 describes that *lumazine synthase* binds with *riboflavin*, where the binding affinity evaluated by the dissociation constant (K_d) is 1.2 μM. Example 6 shows an example where <LIGAND> appeared before <PROTEIN>.

Pattern 2 is designed to identify descriptions of passive form. In this pattern, <EVENT> is in between <PROTEIN> and <LIGAND>, and <VALUE> is at end of the sentence.

Pattern 2	<P/L> - BEV - <EVENT> - PREP - <P/L> – <VALUE>
Example 7	*OXA-13*<PROTEIN> was found to *be*/BEV *inhibited*<EVENT> efficiently *by*<PREP> *imipenem* and *meropenem*<LIGAND> (*IC_{50} = 0.02 μM*)<VALUE>. (PMID: 11453693)

Pattern 3 accommodates sentences where <PROTEIN> and <LIGAND> are connected with a BEV or PREP rather than <EVENT>. The design of Pattern 3 is distinct to previous IR studies requiring a verb to link two entities (Bui *et al.*, 2011; Fundel *et al.*, 2007). However, we observed that some sentences contain binding affinities even without any T_{event} (Example 8).

Pattern 3	<P/L> - BEV/PREP - <P/L> - (BEV)? - <VALUE>
Example 8	The affinity of I219A *hGST A1-1*<PROTEIN> *for*/PREP *GTX*<LIGAND> (K_d = 1.3 μM, 25 °C)<VALUE> is an order of magnitude weaker than that of the wild-type protein (K_d = 0.07 μM, 25 °C)<VALUE>. (PMID: 15893769)

Pattern 4 is a commonly observed writing style which describes the result (<VALUE>) first and then the details (<P/L>s and <EVENT>). Example 9 is an example of this pattern. <EVENT> in sentences of this pattern is not used to connect entities and could be omitted.

Pattern 4	<VALUE> - PREP - <P/L> – <EVENT/BEV>?
Example 9	The K_i *values of 44 and 220 μM*<VALUE> *for*/PREP *(S)-and (R)-propranolol*<LIGAND>, respectively, show that the (S)-form *binds*<EVENT> almost as strong as the natural product cellobiose [$K_i 20\,\mu M$], whereas the binding of the (R)-form is about 5 times weaker. (PMID: 11114249)

3.3 Sentence ranking

The sentences extracted from a full-text article in the previous stages were denoted $S = C_{\{S_i\}}$, where s_i is the ith extracted sentence matching at least one of the defined patterns. In this stage, AutoBind ranks S with a scoring function shown below. In evaluation, a successful extraction is counted if it is retrieved in the three highest ranked sentences. The experimental results shown in the next section reveal that the correct binding affinity appeared in the top three sentences in 79.07% articles.

$$f_i = \frac{\omega - |\{e_i | e_i \notin \gamma\}|}{\omega} \quad (1)$$

$$g_i = \frac{e_i \cap \gamma}{e_i} \quad (2)$$

$$F_i = f_i + g_i, \quad (3)$$

where ω is the number of distinct protein and ligand names in all extracted sentences in the article, γ is the set of protein and ligand names in the corresponding PDB TITLE, f_i measures the overlap of protein and ligand names of S to the corresponding PDB TITLE, e_i is the number of protein and ligand names ins_i and g_i is the ratio of entities in s_i that are also in the corresponding PDB TITLE. In other words, f_i measures the correlation of the article to the PDB TITLE, g_i measures the correlation of s_i to the PDB TITLE, while F_i is the sum of the two measures. The final score, F_i, is the sum of f_i and g_i.

Table 3. Performance of binding affinity extraction

	TP + FN	TP	TP + FP	Precision	Recall	F-score
AutoBind	1486	1175	1395	84.22	79.07	81.56
Co-occurrence	1486	757	1081	70.03	50.94	58.98

The P-value between AutoBind and co-occurrence is $< 1E-10$.

Table 4. Recognition performance of protein names

Method	Precision	Recall	F-score
UniProt	77.66	61.93	68.91
NLProt	62.82	70.59	66.47
Hybrid	72.87	74.61	73.73

The P-values of NLProt to the other two methods are $< 1E-10$; the P-value between UniProt and Hybrid is 0.0025.

Table 5. Recognition performance of ligand names

Method	Precision	Recall	F-score
PubChem	85.37	51.79	64.47
OSCAR4	63.43	64.08	63.75
Hybrid	68.29	68.89	68.59

The P-values of any two methods are $< 1E-10$.

3.4 Co-occurrence method

As AutoBind is the first automatic extraction system dedicated for protein–ligand binding affinity, we implemented an algorithm that simply identifies sentences containing both the target protein and ligand. This algorithm, named sentence-level co-occurrence method in other studies (He *et al.*, 2009), was adopted as a comparison baseline.

4 BENCHMARKING

4.1 Evaluation of binding affinity extraction

Table 3 shows that AutoBind could extract sentences containing the correct binding affinity with 84.22% precision and 79.07% recall, where the P-value of the Wilcoxon signed ranked sum test (Wilcoxon, 1947) to the baseline is $< 1E-10$ (the evaluation metrics are explained in the Supplementary Data). Comparing with AutoBind, the low recall of the co-occurrence method is mainly due to requiring both protein and ligand to appear in the sentence. The results indicate that more than half of the binding affinity articles describe the correct binding affinities in sentences not containing either the target protein or ligand. This reveals the importance of relaxing the requirement of the presence of entities in the proposed patterns. An analysis of the errors of AutoBind and a discussion of the limitation of AutoBind can be found in the Supplementary Data.

4.2 Evaluation of NER

The NER module of AutoBind was actually a union of a dictionary- and a learning-based method (see Section 3.1 for details). In this experiment, we compared the performance of the dictionary-based method, the learning-based method and their union (denoted Hybrid in the context). The results of NER performance for protein and ligand names are shown in Tables 4 and 5, respectively. Table 4 shows that using both techniques outperformed the protein recognition performance over those using individual techniques. The P-values show that the differences of NLProt to UniProt and Hybrid were relatively significant than the difference between UniProt and Hybrid. Namely, the good performance of Hybrid came mainly from UniProt with a little Hybrid came mainly from UniProt with a little help from NLProt in protein recognition. This enhancement was also observed in ligand recognition (Table 5). The P-values show that the differences of Hybrid to the two individual methods were significant. Namely, combining PubChem and OSCAR4 significantly helps ligand recognition. In both protein and ligand name recognition, the dictionary-based methods achieved better precision than that using learning-based methods. On the other hand, dictionary-based methods may have many false negatives because some new compound names have not been collected in the adopted dictionary. For example, ligand recognition using PubChem had lower recall than that using OSCAR4. This problem was relatively moderate in protein recognition (comparing NLProt and UniProt in Table 4) since the catalogue of proteins is more mature than that of ligands. Our results concur with previous studies that hybridizing both techniques enhances recognition performance (Wermter *et al.*, 2009).

The next question is whether the improvement of hybridizing dictionary- and learning-based techniques in entity recognition benefits the performance of affinity extraction. Many studies that used IR technologies to extract biomedical information, e.g. protein–protein interaction (Zhou and He, 2008) and protein phosphorylation (Hu *et al.*, 2005), observed that a large amount of incorrect extractions were owing to the poor recognition of protein entities. Many incorrect words or phrases tagged as entities will largely increase the false-positive extractions and reduce the entire extraction performance. We carried out an evaluation of the binding affinity extraction using the dictionary-based, the learning-based and the hybrid method. The experiment shows the hybrid recognition scheme used in this study can effectively improve the performance of binding affinity extraction. The results are shown in Table 6. Note that the precisions, recalls and F-scores in Table 6 are the performance of affinity extraction, namely comparable with those in Table 3, and should not be compared with those in Tables 4 and 5. The results reveal that accurate NER is critical to affinity extraction. This conclusion is similar to the study of Kim *et al.* (2009). Furthermore, the P-values show that the differences of Learning to Dictionary and Hybrid were relatively significant than the difference between Dictionary and Hybrid. The results in Tables 4 and 5 suggest that the relatively small difference between Dictionary and Hybrid was due to the small difference in protein recognition.

4.3 Comparison with other databases

During the past decade, several protein–ligand binding affinity databases were published, in which all of the collected data were based on manual reading (Table 7). LPDB is the earliest database, containing 195 complexes with binding affinities. A main feature of LPDB is providing computer-generated ‘docking decoys’ for developing and testing docking methods. BindingDB, first published in 2001, collected experimentally determined binding affinities of small molecules used for drug discovery. It provides various search

Table 6. Performance of binding affinity extraction using different NER schemes

NER schemes	Precision	Recall	*F*-score
Dictionary[a]	79.93	65.21	71.87
Learning[b]	66.52	72.69	69.47
Hybrid[c]	84.22	79.07	81.56

The *P*-value between dictionary and learning is $3E-5$; the *P*-value between dictionary and hybrid is 0.0496; the *P*-value between learning and hybrid is $1E-10$.

[a] The dictionary-based NER tools (UniProt for protein recognition and PubChem for ligand recognition) were used.

[b] The learning-based NER tools (NLProt for protein recognition and OSCAR4 for ligand recognition) were used.

[c] Both dictionary- and learning-based NER tools were used.

Table 7. Comparison of AutoBind to other databases

	Protein complex	Number of entries (affinity data)	Last updated date	First published year
AutoBind	37 929	13 616	December 1, 2011	This work
PDBBind	7986	7986	September 22, 2011	2004
Binding MOAD	16 955	5630	2010	2005
BindingDB	3056	1817	March, 2010	2001
AffinDB	474	748	2006	2006
PLD	149	149	2003	2003
LPDB	195	195	2001	2001

facilities of, for example, compound names, chemical structures and binding affinities, for accessing the collected data. AffinDB collected binding affinities from two kinds of scientific literatures: (i) primary sources describing the original experiments of affinity determination and (ii) secondary references collecting many binding affinities from primary references. It also provided experiment details such as pH values and temperature. Before 2010, Binding MOAD was the largest database that screened the entire PDB for primary citations containing binding affinities, but it has not been updated since 2010. It did not include small organic molecules or cofactors. PDBbind was another database that performed PDB-wide extraction. The most recent update of PDBbind is on September 22, 2011 (version 2011). Except PDBbind, other databases are rarely updated. AutoBind contains 13 616 protein–ligand-binding affinities of which 6217 were automatically extracted (45.66%) and 7399 were curated by experts. It contains almost twice binding affinities compared to other databases of similar data. The guideline of AutoBind to the experts can be found in the Supplementary Data.

Furthermore, AutoBind is updated every month. The higher update frequency is achieved by the automatic extraction system that alleviates manual loading. For each update, the newly released structures and the primary citations will be automatically downloaded from PDB and PubMed. The processing binding affinity extraction of each article takes 1 min in average using an Intel Xeon 2.5 GHz processor. Additionally, we adopted AutoBind early release version (updated on September 2011, 11 897 affinities were collected) to carry out a simple test to estimate the time needed for a biologist to review and record binding affinity data from protein structure-related articles. Our curators took ∼20 days to confirm the 5160 automatically extracted binding affinities while took about 3 months without the assistance of the automatic system to extract the other 6737 binding affinities. This reveals the importance of automatic extraction, which facilitated manual curation, hence drastically reducing the required time.

5 CONCLUSION

This work presents a new automatic system, AutoBind, for extracting protein-binding affinity from literature. Overall, our automatic extraction method collected almost twice binding affinities compared to other databases of similar data. The automatic extraction process contains four NER methods for protein and ligand names, four sentences patterns for sentence extraction and a ranking method to pick sentence that correctly describes binding affinities. The experimental results show that hybridizing multiple NER methods improved the performance of binding affinity extraction. These sentence patterns can quickly identify candidate sentences that may contain a protein-binding affinity. These patterns were designed to ensure minimal false negatives for, as shown in the experimental results, high extraction recall. Finally, the proposed ranking method ensures minimal false positives for high precision. The rank-based design facilities the following manual checking step. Biologists can start from the highest ranked sentences and stop immediately after a correct sentence is found. The entire system significantly reduces the extraction time while preserving the quality of the extracted binding affinities.

ACKNOWLEDGEMENTS

We would like to thank all the participants for their contribution to the manual evaluation. We also would like to thank Jing-Yeh Chen for his suggestions in biology and Patrick Chiang for professional proofreading.

Funding: This work was supported by National Science Council, Taiwan [NSC99-2221-E-006-127-MY3 and NSC100-2627-B-006-011].

Conflict of Interest: none declared.

REFERENCES

Agarwal,S. and Yu,H. (2011) Figure summarizer browser extensions for PubMed Central. *Bioinformatics*, **27**, 1723–1724.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhattacharya,S. *et al.* (2011) MeSH: a window into full text for document summarization, *Bioinformatics*, **27**, i120–i128.

Blaschke,C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 60–67.

Block,P. *et al.* (2006) AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.*, **34**, D522–D526.

Bui,Q.-C. *et al.* (2011) A hybrid approach to extract protein–protein interactions. *Bioinformatics*, **27**, 259–265.

Chang,D.T.-H. *et al.* (2012) AH-DB: collecting protein structure pairs before and after binding. *Nucleic Acids Res.*, **40**, D472–D478.

Chen, C.-Y. *et al.* (2012) Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PLoS One*, **7**, e30446.

Chen,X. *et al.* (2002) The Binding Database: data management and interface design. *Bioinformatics*, **18**, 130–139.

Chiang,J.-H. *et al.* (2006) GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformatics*, **7**, 392.

Chien,T.-Y. *et al.* (2012) DBD2BS: connecting a DNA-binding protein with its binding sites. *Nucleic Acids Research*, **40**, W173–W179.

Fundel,K. *et al.* (2007) RelEx—relation extraction using dependency parse trees. *Bioinformatics*, **23**, 365–371.

He,M. *et al.* (2009) PPI finder: a mining tool for human protein-protein interactions. *PLoS One*, **4**, e4554.

Heinen,S. *et al.* (2010) KID—an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes. *BMC Bioinformatics*, **11**, 375.

Hu,L. *et al.* (2005) Binding MOAD (Mother Of All Databases), *Prot. Struct. Funct. Bioinformatics*, **60**, 333–340.

Hu,Z.Z. *et al.* (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.

Jang,H. *et al.* (2006) Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics*, **22**, e220–e226.

Jessop,D. *et al.* (2011) OSCAR4: a flexible architecture for chemical text-mining, *J. Cheminform.*, **3**, 41.

Kim,J.-D. *et al.* (2009) Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Association for Computational Linguistics, Boulder, Colorado, pp. 1–9.

Koike,A. and Takagi,T. (2005) PRIME: automatically extracted protein interactions and molecular information database. *In Silico Biol.*, **5**, 9–20.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, **6**, 224.

Liu,T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

McDonald,D.M. *et al.* (2004) Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, **20**, 3370–3378.

Mika,S. and Rost,B. (2004) NLProt: extracting protein names and sequences from papers. *Nucleic Acids Res.*, **32**, W634–W637.

Milo, R. *et al.* (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.*, **38**, D750–D753.

Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl Acad. Sci. USA*, **104**, 7068–7073.

Névéol,A. *et al.* (2011) Extraction of data deposition statements from the literature: a method for automatically tracking research results. *Bioinformatics*, **27**, 3306–3312.

Puvanendrampillai,D. and Mitchell,J.B.O. (2003) Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. *Bioinformatics*, **19**, 1856–1857.

Rindflesch,T.C. *et al.* (2000) Extracting molecular binding relationships from biomedical text. In *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, Seattle, WA, pp. 188–195.

Roche,O. *et al.* (2001) Ligand-Protein DataBase: linking protein-ligand complex structures to binding data. *J. Med. Chem.*, **44**, 3592–3598.

Spasić,I. *et al.* (2009) KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics*, **25**, 1404–1411.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Proceedings of the fifth Pacific Symposium on Biocomputing*. Hawaii, pp. 529–540.

Tari,L. *et al.* (2010) Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, **26**, i547–i553.

Wang,R. *et al.* (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.

Wang,Y. *et al.* (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.

Wermter,J. *et al.* (2009) High-performance gene name normalization with GeNo. *Bioinformatics*, **25**, 815–821.

Wilcoxon,F. (1947) Probability tables for individual comparisons by ranking methods. *Biometrics*, **3**, 119–122.

Zhou,D. and He,Y. (2008) Extracting interactions between proteins from the literature. *J. Biomed. Informatics*, **41**, 393–407.

Zhou,W. *et al.* (2006) ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, **22**, 2813–2818.