Sequence Clustering Methods and Completeness of Biological Database Search

Qingyu Chen
The University of Melbourne
qingyuc1@student.unimelb.edu.au

Xiuzhen Zhang
RMIT University
xiuzhen.zhang@rmit.edu.au

Yu Wan
The University of Melbourne
wanyuac@gmail.com

Justin Zobel
The University of Melbourne
jzobel@unimelb.edu.au

Karin Verspoor
The University of Melbourne
karin.verspoor@unimelb.edu.au

Abstract

Sequence clustering methods have been widely used to facilitate sequence database search. These methods convert a sequence database into clusters of similar sequences. Users then search against the resulting non-redundant database, which is typically comprised of one representative sequence per cluster, and expand search results by exploring records from matching clusters. Compared to direct search of original databases, the search results are expected to be more diverse are also more complete. While several studies have assessed diversity, completeness has not gained the same attention. We analysed the BLAST results on non-redundant versions of the UniProtKB/Swiss-Prot database generated by clustering method CD-HIT. Our findings are that (1) a more rigorous assessment on completeness is necessary, as an expanded set can have so many answers that Recall is uninformative; and (2) the Precision of expanded sets on top-ranked representatives drops by 7%. We propose a simple solution that returns a user-specified proportion of top similar records, modelled by a ranking function that aggregates sequence and annotation similarities. It removes millions of returned sequences, increases Precision by 3%, and does not need additional processing time.

1 Introduction

Biological sequence databases accumulate a wide variety of observations of biological sequences and provide access to a massive number of sequence records submitted from individual labs [Baxevanis and Bateman, 2015]. Their primary application use is in sequence database search, in which: database users prepare query sequences such as uncharacterised proteins; perform sequence similarity search of a query sequence against deposited database records, often via BLAST [Altschul *et al.*, 1990]; and judge the output, that is, a ranked list of retrieved sequence records.

A key challenge for database search is redundancy, as database records contain very similar or even identical sequences [Bursteinas *et al.*, 2016]. Redundancy has two immediate impacts on database search: the top ranked retrieved sequences can be highly similar, and may not be independently informative (such as shown in Figure 1(a)); and it makes it difficult to find potentially interesting sequences that are distantly similar. A possible solution is to remove redundant records. However, the notion of redundancy is context-dependent; removed records may be redundant in some contexts but important in others [Chen *et al.*, 2017].

Machine learning techniques are often used to solve biological problems. In this case clustering methods have been widely applied [Fu *et al.*, 2012]. These cluster a sequence database at a user-defined sequence identity threshold, creating a *non-redundant database*. Users search against the non-redundant database and expand search results by exploring records from the same clusters. Thus it is expected that the search results will be more *diverse*, as retrieved representatives may be distantly similar. The results also will be more *complete*; the expanded search results should be similar enough to direct search of original databases that potentially interesting records will still be found. Existing studies measured search effectiveness primarily from the perspective of diversity [Fu *et al.*, 2012; Chen *et al.*, 2016a], but, largely, have not examined completeness. An exception is a study that measured completeness but did not address user behaviour or satisfaction [Suzek *et al.*, 2015].

We study search completeness in more depth by analysing BLAST results on non-redundant versions of the UniProtKB/Swiss-Prot. We find that a more rigorous assessment on completeness is necessary; for example, an expanded set brings 40 million more query-target pairs, making Recall uninformative. Moreover, Precision of expanded sets on top-ranked representatives drops by 7%. We propose a simple solution that returns a user-specified proportion of top similar records, modelled by a ranking function that aggregates sequence and annotation similarities. It removes millions of returned query-target pairs, increases Precision by 3%, and does not need additional processing time.

2 Sequence clustering methods

Clustering is an unsupervised machine learning technique that groups records based on a similarity function. It has wide applications in bioinformatics such as creation of non-redundant databases [Mirdita *et al.*, 2016] and classifying sequence records into Operational Taxonomic Units [Chen *et al.*, 2013]. Here we explain how CD-HIT, a widely-used clus-

Figure 1: Search of query sequences against original database vs. non-redundant database using search results of UniProtKB/Swiss-Prot record *A7FE15* on UniProtKB and UniRef50 (a clustered database) as an example. (a) The top retrieved results of original database may be highly similar or not independently informative; (b) The top retrieved results of the non-redundant version are more diverse; (c) The expanded set makes the search results more complete.

tering method, generates non-redundant databases. From an input sequence database and a user-defined sequence identity threshold, it constructs a non-redundant database in three steps [Fu *et al.*, 2012]: (1) Sequences are sorted by decreasing length. The longest sequence is by default the *representative* of the first cluster. (2) The remaining sequences are processed in order. Each is compared with the cluster representative. If the sequence identity for some cluster is no less than the user-defined threshold, it is assigned to that cluster; if there is no satisfactory representative, it becomes a new cluster representative. (3) Two outputs are generated, representatives and the complete clusters. These comprise the non-redundant database. As sequence databases are often large, greedy procedures and heuristics are used to speed up clustering. For example, a sequence will be assigned to a cluster immediately as long its sequence identity between the representative satisfies the threshold.

Sequence search on non-redundant databases consists of two steps. Users first search query sequences against the non-redundant database only, as shown in Figure 1(b). The retrieved records are effectively a ranked list of representatives in the non-redundant database. This step aims for diversity. Users then expand search results by looking at the complete clusters, that is, retrieved representatives and the associated member records, as shown in Figure 1(c). This step focuses on completeness.

3 Measurement of search effectiveness

To quantify whether clustering methods indeed achieve both diverse and complete search results, search effectiveness on the non-redundant databases has been measured. Many studies focus on diversity; for example, the remaining redundancy between representatives in CD-HIT has been considered [Fu *et al.*, 2012] and a recent study found that this remaining redundancy is higher as the identity threshold is reduced [Chen *et al.*, 2016a]. Completeness has been overlooked, despite its value to users as indicated by several studies:

- Suzek *et al.* constructed UniRef databases using CD-HIT at different thresholds [Suzek *et al.*, 2015]. They measured diversity of representatives in a case study of determining remote protein family relationship and measured the completeness of the expanded set in a case study of searching sequences against UniProtKB.
- Mirdita *et al.* constructed Uniclust databases using a similar clustering procedure to that of CD-HIT [Mirdita *et al.*, 2016]. They assessed cluster consistency by measuring Gene Ontology (GO) annotation similarity and protein-name similarity to ensure that users obtain consistent views when expanding search results.
- Cole *et al.* created a protein sequence structure prediction website that searches user submitted sequences against UniRef and selects the top retrieved representatives based on e-values [Cole *et al.*, 2008].
- Remita *et al.* searched against UniRef for miRNAs regulating glutathione S-transferases and expanded the results from the associated Uniref clusters to obtain alignment information, Gene Ontology (GO) annotations, and expression details to ensure they did not miss any other related data [Remita *et al.*, 2016].

The first two examples directly show that database staff care about diversity and completeness when creating non-redundant databases; the last two further illustrate that database users in practice may use only representatives for diversity or expand search results for completeness. There are many further instances [Capriotti *et al.*, 2012; Sato *et al.*, 2011; Liew *et al.*, 2016]. These examples demonstrate that both diversity and completeness are critical and the associated assessments are necessary. When UniRef staff measured search completeness, they used all-against-all BLAST search results on UniProtKB as a gold standard [Suzek *et al.*, 2015]. Then they evaluated the overall Precision and Recall of the expanded set (Formulas 1 and 5): Precision quantifies whether expanded records are identified as relevant in the gold standard and Recall quantifies whether the results in the gold standard can be found in the expanded set. UniRef is one of the best known clustered protein databases. The measurement shows that assessing search completeness is of value.

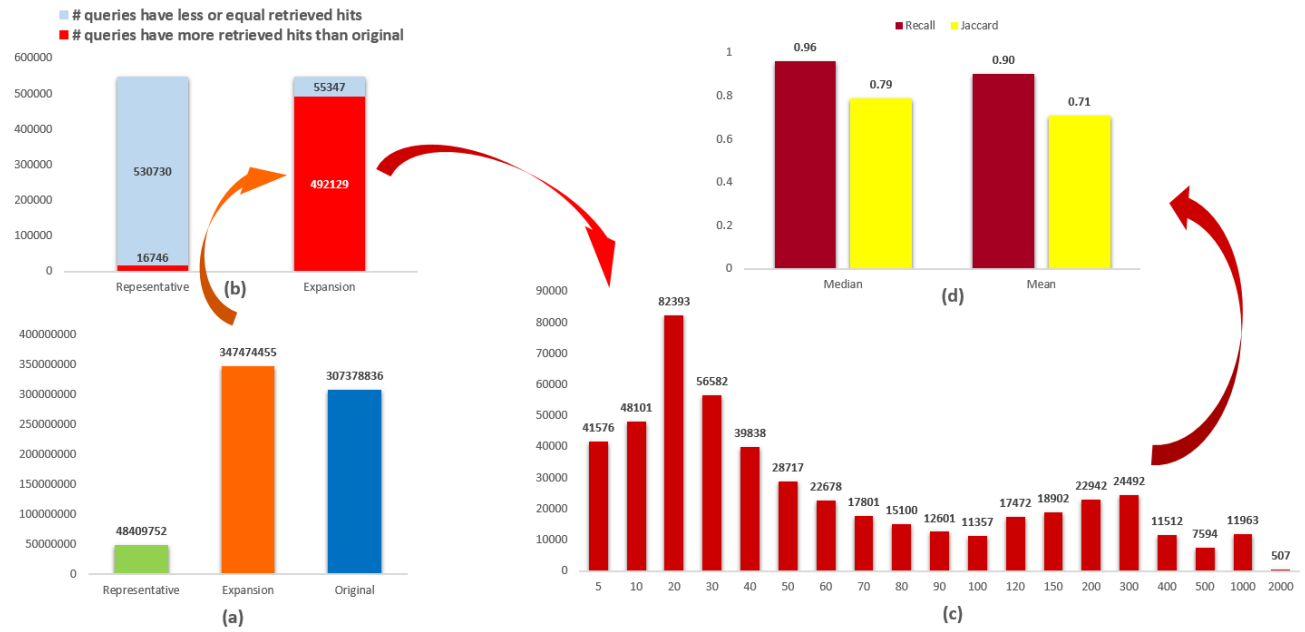

Figure 2: (a) Expansion brings more hits than original search. (b) After expansion, ≈90% of queries have more hits than search on the original database. (c) Those ≈90% of queries have a median of 34 more hits than original search. (d) Recall is high but at the cost of returning more hits than original search. Jaccard similarity is lower than Recall, showing the results of the expanded set are not similar to those of the original database.

However, its measurement on completeness does have limitations. A major limitation is that database user behaviour or user satisfaction are not examined. Given a query, the adopted overall Precision measures all the records in the expanded set. However, users may only examine retrieved representatives without expanding the search results [Sato *et al.*, 2011]. Also, they may only examine the top-ranked representatives and expand the associated search results [Remita *et al.*, 2016]. Measuring only overall Precision on an expanded set fails to reflect this behaviour. The proposed metrics should reflect user satisfaction [Moffat *et al.*, 2013].

The adopted measure of Recall also has failings. It has been a long-term concern that Recall may not be effective for information retrieval measurement [Zobel, 1998; Webber, 2010; Walters, 2016]. In this case the Recall might be higher if the expanded set has more records than the gold standard. But this means users will have to browse more results. Also users may only examine and expand the top retrieved representatives so the associated expanded set will be always a small subset of the complete search results. Recall is not applicable in those cases. We proposed a more comprehensive approach below.

4 Data and Methods

Dataset, tools, and experiments

We used full-size UniProtKB/Swiss-Prot Release 2016-15 as our experimental dataset. It consists of 551,193 protein sequence records. CD-HIT (4.6.5) was used to construct the associated non-redundant UniProtKB/Swiss-Prot; NCBI BLAST (2.3.0+) was used to perform all-against-all searches.

CD-HIT by default removes sequences of length no greater than 10 since such short sequences are generally not informative. We removed those records correspondingly in full-size UniProtKB/Swiss-Prot. The updated dataset has 550,047 sequences. We used them as queries and performed BLAST searches on the updated UniProtKB/Swiss-Prot and its non-redundant version at 50% threshold generated by CD-HIT. The non-redundant database at 50% consists of 120,043 sequences. 547,476 out of 550,047 query sequences have at least one retrieved sequence in both databases. The BLAST results are commonly called *query-target pairs* or *hits*. We removed two types of query-target pairs: where the target is the query itself; and the same sequence retrieved more than once for a query. BLAST performs local alignment; it is reasonable that multiple regions of a sequence are similar as the query sequence. However repeated query-target pairs in this case bias statistical analysis.

The commands for running CD-HIT[1] and BLAST[2] strictly follow user guidance. NCBI BLAST staff (personal communication via email) advised on the maximum number of output sequences, to ensure sensible results. Note also that this study focuses on general uses of the tools, while, for instance, UniRef and Uniclust may use different parameters to construct non-redundant databases for specific purposes.

[1] *./cd-hit -i input_path -o output_path -c 0.5 -n 2*, where -i and -o stand for input and output path. -c stands for identity threshold, -n specifies word size recommended in the user guide.

[2] *./blastp -task blastp -query query_path -db database_path -max_target_seqs 100000*, where blastp specifies protein sequence, -query and -db specifies query and database path. -max_target_seqs is the maximum number of returned sequences for a query.

Figure 3: Proportion of queries having higher Precision in representatives than in the expanded set. We removed queries that have same number of hits in both (it means retrieved representatives do not have any member records). The first row compares unranked expanded set (a) with our proposed ranked model (b) using the metric $P@K_{equal}$; the second row compares unranked expanded set (c) with our proposed ranked model (d) using $P@K_{weight}$.

Assessing search effectiveness

We measured the search effectiveness on the non-redundant data set as follows. Given a query Q, let F be the list of fetched (retrieved) representatives from the non-redundant database, E its expanded set, and R the set of relevant sequences. Here, F is a ranked list, consisting of representatives ordered by BLAST scores, whereas E contains representatives and the associated cluster members, which may not have a particular order. R in this case stands for all the fetched sequences for Q from the original UniProtKB/Swiss-Prot as the gold standard. Each sequence, either in F or E, is scored by a function S: 0 if it is not in R, 1 otherwise. We compared the number of query-target pairs in F, E and R respectively. This examines how many retrieved results users need to browse in the non-redundant version compared with original database. We also employed standard evaluation metrics from information retrieval, adapted specifically for our study, as below.

Since users may or may not expand the search results, we measured Precision of both representatives and expanded set:

$$Precison(F) = \frac{|F \cap R|}{|F|} \qquad Precision(E) = \frac{|E \cap R|}{|E|} \quad (1)$$

Users may focus on top-ranked retrieved representatives and expand only those. Overall Precision cannot capture such cases. We therefore measured $P@K$, Precision at top K retrieved sequences. $P@K$ for R measures the Precision at K representatives, which is a standard metric used in Information Retrieval evaluation [Webber, 2010]:

$$P@K(F) = \frac{1}{K}\sum_{i=1}^{K} S(F_i) \quad (2)$$

$P@K$ for E, however, is not straightforward. K in this context refers to K clusters, which contain many more than K records; thus is not directly comparable. We propose two $P@K$ metrics for E, summarised in Formula 3 and 4. In this formula, C_i, $|C_i|$, $C_{i,j}$ are an expanded cluster, the expanded cluster size, and a sequence in the expanded cluster, respectively. The idea is to transform the score of a sequence relative to the cluster size; for example, the score of a sequence in a cluster of 10 records will be $\frac{1}{10}$. The former formula treats every cluster equally, that is, ($\frac{1}{K}$). The latter weights clusters such that larger clusters have higher weights.

$$P@K_{equal}(E) = \sum_{i=1}^{K} \frac{1}{K|C_i|} \sum_{j=1}^{|C_i|} S(C_{i,j}) \quad (3)$$

$$P@K_{weight}(E) = \sum_{i=1}^{K} \frac{|C_i|}{\sum_{i=1}^{K}|C_i|} \sum_{j=1}^{|C_i|} S(C_{i,j}) \quad (4)$$

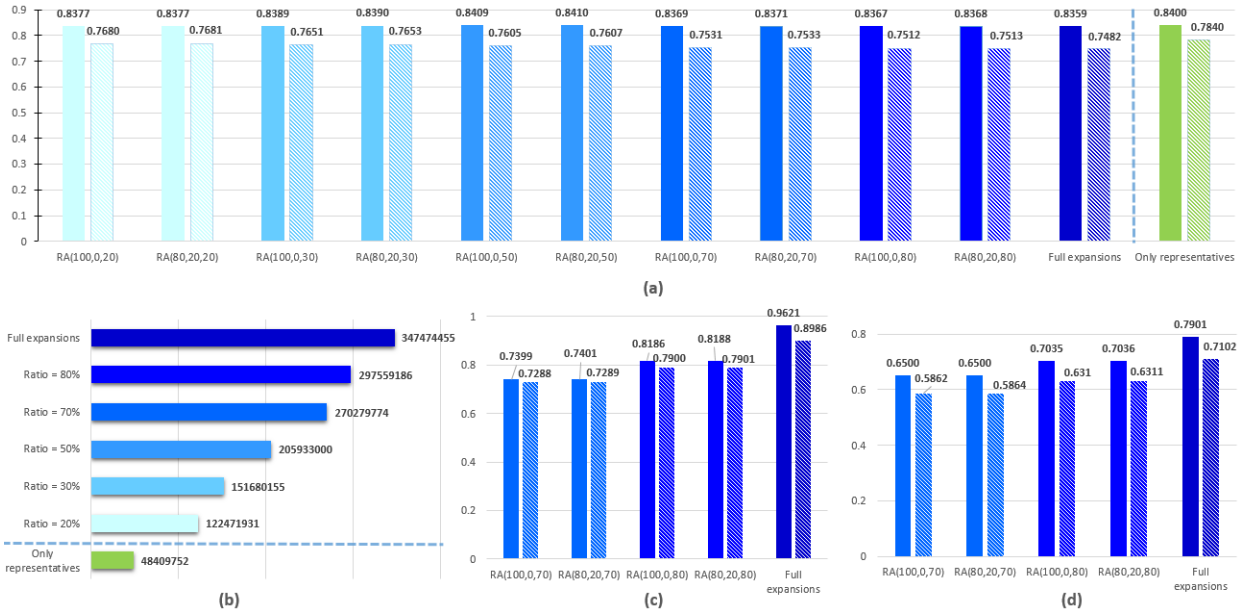

Figure 4: Comparative results for original (unranked) expanded set and our proposed ranked model. Sub-graphs (a): $P@K$ measures; (c): Recall results; and (d): Jaccard results. Each of them shows the mean and median result of the metrics, where median is represented in dash. (b) presents Number of retrieved hits. $RA(seq, annotation, proportion)$ refers the ranked model summarised in Section 5, where seq and $annotation$ refer to the weight of sequence identity and annotation similarity, effectively α and β in Formula 6 and $proportion$ refers to the proportion specified by users to expand search results.

We also measured Recall and Jaccard similarity to assess whether E is (near) identical to R. Recall is used in the previous study. However, it may be biased if an expanded set has more hits than original search. Jaccard similarity is thus used as a complementary metrics because it can better illustrate the differences between two sets of results. Note that those two metrics are not applicable for F, since F are intended to only retrieve a subset of the complete results.

$$Recall(E) = \frac{|E \cap R|}{|R|} \qquad Jaccard(E) = \frac{|E \cap R|}{|E \cup R|} \tag{5}$$

5 Results and Discussion

Our experiments on the number of query-target pairs in the clustered non-redundant data as compared with original database demonstrate that Recall is over-estimated and in turn is not informative, due to the expanded set having even more query-target pairs than the original dataset. Figure 2(a) compares the number of query-target pairs. The retrieved pairs among representatives include only about 15% of the pairs from the original dataset. On the one hand this indicates that users can browse the search results more efficiently. On the other hand it shows that expansion of results is valuable since potential interesting records may be in the other 85%. However, the expanded set produces 40,095,619 more pairs than the original. Figure 2(b) further shows that the expanded set produces more pairs on over 89% of queries (492,129 out of 547,476), and on average produces about 10 pairs per query (Figure 2(c)). Having more pairs results in high Recall. Both median and mean Recall (Figure 2(d)) are above 90%, but this comes with the cost of producing more 40 million pairs. Jaccard similarity by comparison is almost 20% lower than Recall, which clearly shows the results of the expanded set are not similar to those of the original database.

In addition, the Precision of the expanded set distinctly degrades at top-ranked hits. Table 1 shows different levels of Precision on representatives and the expanded sets. We assessed both measures at depth 10, 20, 50, 100, and 200 respectively to quantify the Precision of the top-ranked hits that are more likely examined by users. In general, top-ranked hits from representatives are valuable: Precision is over 96% across different K. The Precision of the expanded set, either $P@K_{equal}$ or $P@K_{weight}$, is always lower than that of representatives, with degradation of up to 7% at $K = 200$. It may be argued that, for a representative, if its relevance is 1, the relevance of the associated expanded set will almost be lower, since each record in the expanded set would also have to be relevant. Conversely, the relevance of the expanded set is likely to be higher if the relevance of the representative is 0, since a single relevant record will improve on this.

We further compared Precision in detail on an individual query level, as summarised in Figure 3. The Precision of representatives at the top K positions is higher than that of the expanded sets for at least 80% of the queries; the proportion increases as K grows.

Driven by these observations, we propose a simple solution that ranks records in terms of their similarity with cluster representatives and only returns the top X%, a user-defined proportion, when they expand search results. To our knowledge, existing databases such as UniRef select representatives based on whether a record is reviewed by biocurators,

	$P@K$				
	K=10	20	50	100	200
Representatives	0.968	0.977	0.983	0.985	0.983
$P@K_{equal}$ original	0.938	0.951	0.958	0.980	0.952
Ranked sequence	0.938, 0.946	0.952, 0.960	0.958, 0.966	0.959, 0.967	0.952, 0.963
Ranked seq & annotation	0.938, 0.947	0.952, 0.960	0.959, 0.967	0.959, 0.968	0.953, 0.953
$P@K_{weight}$ original	0.924	0.935	0.938	0.929	0.917
Ranked sequence	0.926, 0.940	0.937, 0.952	0.940, 0.957	0.933, 0.953	0.922, 0.947
Ranked seq & annotation	0.926, 0.940	0.938, 0.952	0.941, 0.957	0.933, 0.954	0.923, 0.947

Table 1: $P@K$ measure results. Representatives: $P@K$ for representatives (Formula 2); $P@K_{equal}$ and $P@K_{weight}$ are $P@K$ for expanded sets (Formulas 3 and 4 respectively); Original refers to expanded whole records and Ranked refers to our ranked model (Formula 6). *Ranked sequence* takes sequence identity only; *Ranked seq & annotation* takes sequence identity weighted 80% and annotation similarity weighted 20%. The results of the ranked model were measured at 20%, 30%, 50%, 70% and 80%, the user-specified proportion to expand search results, summarised in the form of *min*,*max*.

is from a model organism and other such record-external factors. They do not compare and rank the similarity between records. Also they expand all the records in a cluster rather than choosing only a subset.

In our proposal, the notion of similarity between a record and its cluster representative is modelled based on sequence identity and annotation similarity. This similarity function is shown in Formula 6, where R and M refer to a representative and an associated cluster member record. Sim_{seq} and $Sim_{annotation}$ stand for their sequence identity and annotation similarity respectively. Annotations are based on record metadata, such as GO terms, literature references and descriptions. Sequence identity is arguably the dominant feature, but existing studies for other tasks demonstrate that combining sequence identity and metadata similarity is valuable [Chen *et al.*, 2016b]. α and β refer to their corresponding weights; for example, sequence identity accounts for 80% of the aggregated similarity and annotation similarity accounts for another 20% when α is 0.8 and β is 0.2.

$$Sim(R, M) = \alpha Sim_{seq}(R, M) + \beta Sim_{annotation}(R, M) \quad (6)$$

The records in each cluster are thus ranked by this similarity function in descending order. The top-ranked X% records, with X specified by a user, will be presented when the user expands search results. The ranked model can be adjusted by both database staff and database users. On the one hand, database staff can customise the ranking function, such as adjusting weights and selecting different types of annotations, when creating non-redundant databases. On the other hand, database users can select how many records to browse rather than seeing all records when expanding search results.

In this study, we used sequence identity reported by CD-HIT and Molecular Function (MF) GO term similarities as annotation similarity. MF GO terms are extracted from UniProt-GOA dataset [Courtot *et al.*, 2015] and the similarity is calculated using the well-known *LinAVG* metric [Lin, 1998]. We applied the ranking function with two sets of weights: the first is when α = 100% and β = 0%, i.e., only rank based on sequence identity, whereas the second is α = 80% and β = 20%. We then measured in different proportions 20%, 30%, 50%, 70%, and 80% to reflect how much proportion users want to expand. $RA(seq, annotation, proportion)$ used in Figure 4 shows the values of α, β and the returned proportion, respectively.

Table 1 compares detailed $P@K$ measures for the ranked model with the original unranked expanded set. The ranked model always has higher Precision across different ratios and values of K. Figure 3 shows that over 85% queries have higher Precision in representatives than the expanded set. The ranked model decreases this dramatically, to about 35%, showing that the ranked model has the potential to maintain Precision over expanded search results. Results in Figure 4 further confirmed the findings. Figure 4(b) illustrates that user-defined proportions can significantly reduce the number of expanded query-target pairs: even the highest proportion 80% has about 50 million fewer query-target pairs than the full expanded set, and its median and mean Precision are higher than that of the full expanded set (shown in Figure 4(a)). This shows that in practice users can browse many fewer results. This shows the plausibility of our solution and also demonstrates that metadata is effective in the context of sequence search. Another advantage of our solution is that it does not require additional time in sequence searching: CD-HIT by default reports the identities between representatives and members; MF GO terms similarities can also be pre-computed.

A limitation of the approach is that it has lower Recall and Jaccard similarity than the full expanded set (shown in Figure 4(c,d)). However, it is our view that the number of expanded query-target pairs and Precision measures are more critical to user satisfaction. For instance, proportion at 20% produces around 200 million fewer query-target pairs and has 2% higher $P@K$ and mean Precision. Users may already find enough interesting results from the expanded 20% results.

6 Conclusion

We have analysed the search effectiveness of sequence clustering from the perspective of completeness. The detailed assessment results illustrate that the Precision of representatives is high, but that expansion of search results can degrade Precision and reduce user satisfaction by producing large numbers

of additional hits. We proposed a simple solution that ranks records in terms of sequence identity and annotation similarity. The comparative results show that it has the potential to bring more precise results while still providing users with expanded results.

Acknowledgments

We appreciate the advice of the NCBI BLAST team on BLAST related commands and parameters. Qingyu Chen's work is supported by Melbourne International Research Scholarship from the University of Melbourne. The project receives funding from the Australian Research Council through a Discovery Project grant, DP150101550.

References

[Altschul *et al.*, 1990] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[Baxevanis and Bateman, 2015] Andreas D Baxevanis and Alex Bateman. The importance of biological databases in biological discovery. *Current protocols in bioinformatics*, pages 1–1, 2015.

[Bursteinas *et al.*, 2016] Borisas Bursteinas, Ramona Britto, Benoit Bely, Andrea Auchincloss, Catherine Rivoire, Nicole Redaschi, Claire O'Donovan, and Maria Jesus Martin. Minimizing proteome redundancy in the uniprot knowledgebase. *Database: The Journal of Biological Databases and Curation*, 2016.

[Capriotti *et al.*, 2012] Emidio Capriotti, Nathan L Nehrt, Maricel G Kann, and Yana Bromberg. Bioinformatics for personal genome interpretation. *Briefings in bioinformatics*, 13(4):495–512, 2012.

[Chen *et al.*, 2013] Wei Chen, Clarence K Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. A comparison of methods for clustering 16s rrna sequences into otus. *PloS one*, 8(8):e70837, 2013.

[Chen *et al.*, 2016a] Qingyu Chen, Yu Wan, Yang Lei, Justin Zobel, and Karin Verspoor. Evaluation of cd-hit for constructing non-redundant databases. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 703–706. IEEE, 2016.

[Chen *et al.*, 2016b] Qingyu Chen, Justin Zobel, Xiuzhen Zhang, and Karin Verspoor. Supervised learning for detection of duplicates in genomic sequence databases. *PloS one*, 11(8):e0159644, 2016.

[Chen *et al.*, 2017] Qingyu Chen, Justin Zobel, and Karin Verspoor. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database: The Journal of Biological Databases and Curation*, 2017(1), 2017.

[Cole *et al.*, 2008] Christian Cole, Jonathan D Barber, and Geoffrey J Barton. The jpred 3 secondary structure prediction server. *Nucleic acids research*, 36(suppl 2):W197–W201, 2008.

[Courtot *et al.*, 2015] Mélanie Courtot, Aleksandra Shypitsyna, Elena Speretta, Alexander Holmes, Tony Sawford, Tony Wardell, Maria Jesus Martin, and Claire O'Donovan. Uniprot-goa: A central resource for data integration and go annotation. In *SWAT4LS*, pages 227–228, 2015.

[Fu *et al.*, 2012] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[Liew *et al.*, 2016] Yi Jin Liew, Taewoo Ryu, Manuel Aranda, and Timothy Ravasi. mirna repertoires of demosponges stylissa carteri and xestospongia testudinaria. *PloS one*, 11(2):e0149080, 2016.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[Mirdita *et al.*, 2016] Milot Mirdita, Lars von den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):170–176, 2016.

[Moffat *et al.*, 2013] Alistair Moffat, Paul Thomas, and Falk Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 659–668. ACM, 2013.

[Remita *et al.*, 2016] Mohamed Amine Remita, Etienne Lord, Zahra Agharbaoui, Mickael Leclercq, Mohamed A Badawi, Fathey Sarhan, and Abdoulaye Baniré Diallo. A novel comprehensive wheat mirna database, including related bioinformatics software. *Current Plant Biology*, 7:31–33, 2016.

[Sato *et al.*, 2011] Shusei Sato, Hideki Hirakawa, Sachiko Isobe, Eigo Fukai, Akiko Watanabe, Midori Kato, Kumiko Kawashima, Chiharu Minami, Akiko Muraki, Naomi Nakazaki, et al. Sequence analysis of the genome of an oil-bearing tree, jatropha curcas l. *DNA research*, 18(1):65–76, 2011.

[Suzek *et al.*, 2015] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, and Cathy H Wu. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[Walters, 2016] William H Walters. Beyond use statistics: Recall, precision, and relevance in the assessment and management of academic libraries. *Journal of Librarianship and Information Science*, 48(4):340–352, 2016.

[Webber, 2010] William Edward Webber. *Measurement in information retrieval evaluation*. PhD thesis, 2010.

[Zobel, 1998] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314. ACM, 1998.