EPiC Series in Computing

Volume 70, 2020, Pages 192–197

Proceedings of the 12th International Conference on Bioinformatics and Computational Biology

ASPECT, an LDA-Based Predictive Algorithm for In Vitro Selection

Puzhou Wang[1]

[1] Synthego Corporation, California, United States.

puzhou.wang@synthego.com

Abstract

In vitro selection enables the identification of functional DNA or RNA sequences (i.e., active sequences) out of entirely or partially random pools. Various computational tools have been developed for the analysis of sequencing data from selection experiments. However, most of these tools rely on structure-function relationship that is usually unknown for de novo selection experiments. This largely restricts the applications of these algorithms. In this paper, an active sequence predictor based on Latent Dirichlet allocation (LDA), ASPECT (Active Sequence PrEdiCTor), is proposed. ASPECT is independent of a priori knowledge on the structures of active sequences. Experimental results showed that ASPECT is effective.

1 Introduction

In vitro selection is the experiment process by which functional sequences of DNA or RNA are identified through parallelly examining random sequence pools. Those particular sequences that have a desired function are referred to as active sequences. This technology allowed the expansion of both catalytic and binding capabilities of DNA and RNA molecules. The scope of DNA and RNA enzymes include not only nucleotide acid cleavage [1,2], hydrolysis [3,4], and modifications [5,6], but also peptide modifications [7-9] and reactions between small molecules [10,11]. DNA and RNA aptamers are sequences that bind to specific target molecules. Aptamers have been found for many targets, ranging from small molecules such as kanamycin [12], to large protein complexes such as human thrombin [13], and even entire cells exemplified as human lung carcinoma cell A549 [14].

Many computational tools have also been developed and applied to in vitro selection experiments. Most of these tools are predictive algorithms of folding structures, such as mFold [15], DNA Software's OMP, ViennaRNA [16], and MC-Fold/MC-Sym [17]. They can be used to predict folding structures of DNA and RNA molecules and then to relate the predicted structures with potential functions. However, these tools rely highly on the structure-function relationship of desired sequences, which is usually not available for de novo selection experiments. Therefore, in most cases, the aforementioned tools cannot be used to predict active sequences in de novo selection experiments.

To overcome this limitation, a novel predictive algorithm based on Latent Dirichlet allocation (LDA), ASPECT (Active Sequence PrEdiCTor), was discussed in this work. ASPECT clusters the

Q. Ding, O. Eulenstein and H. Al-Mubaid (eds.), BICOB 2020 (EPiC Series in Computing, vol. 70), pp. 192–197

sequences observed in the selection pools from different rounds and highlights the clusters enriched through selection rounds. When tested with sequence data from Diel-Alderase (DAse) RNA enzyme selection [18], ASPECT showed robust prediction of active sequences.

2 The ASPECT Algorithm

In ASPECT, the selection pool from each round is considered as a corpus and each sequence in the pool is considered as a document (Figure 1). Every document is represented by a bag of words that are k-mer fragments of the DNA/RNA sequence. The vocabulary of a selection experiment is the collection of the k-mer words observed in the corpora of all selection rounds. The definition of a topic in the ASPECT algorithm is identical to its canonical definition in topic modeling, a probability distribution over the fixed vocabulary. With this representation, the enrichment of active sequences through selection experiments can be translated into the enrichment of "active" topics. Once active topics are predicted, active sequences can be easily identified from designated selection round pool by selecting the representative documents in each active topic. Therefore, the goal of the ASPECT algorithm is to (A) discover a set of active topics, T_1, …, T_l, from sequencing data of selection pools C_{r_1}, …, C_{r_n} from a series of selection rounds r_1, …, r_n; (B) select representative sequences, d_1, …, d_m, for discovered active topics from C_{r_s}, where $r_s \in \{r_1, \ldots, r_n\}$ is a selection round number designed by the user.

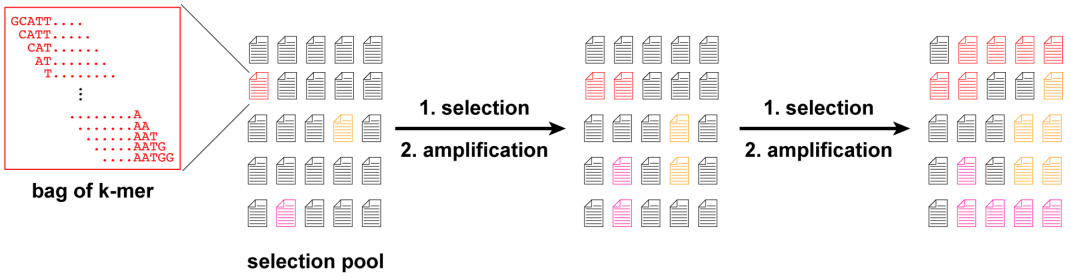

Figure 1: In vitro selection from the aspect of ASPECT. Selection pool from each round is considered as a corpus. DNA/RNA sequences are documents composed by bags of k-mer fragments.

Input: sequencing data of selection pools $C = \{C_{r_1}, \ldots, C_{r_n}\}$ from a series of selection rounds r_1, …, r_n, LDA-based topic modeling method M, four parameters k (fragment size), tn (how many topics to model), μ (active topic threshold), and r_s (designed selection round for active sequence identification)
Output: predicted active sequences, d_1, …, d_m, from C_{r_s}
Workflow:

1. Apply M to C to generate tn topics T_1, …, T_{tn}
2. Calculate the correlations between topics and round numbers, and select active topics T_1, …, T_l, which have correlation coefficients above the threshold μ (e.g., 0.9)
3. For each active topic, select representative sequences from selection pool C_{r_s} of round r_s

3 Evaluation

ASPECT was evaluated with the Next-Generation Sequencing (NGS) data from the DAse selection experiment [10]. An RNABOB-based secondary structure prediction reported for DAse was used as a surrogate for true positive of active sequence prediction, as real positive controls based on activity assays are not accessible [18]. Both precision and recall of the ASPECT algorithm were reported, compared to the RNABOB-based prediction, as well as the correlation coefficients for mined topics.

3.1 DAse Data Set

The DAse RNA enzymes were identified from N120 pool to catalyze carbon-carbon bond formation by Diels-Alder reaction of maleimide and anthracene [10]. The structure-function relationship for DAse was thoroughly studied [19-22]. The DNA pools from the individual rounds of original selection experiment were also sequenced with NGS to further understand the evolution pathway [18]. In this work, the top 2000 sequences from each round pool were used to validate ASPECT.

3.2 Experiment Results

In the initial experiment, sequencing data from round 4 – 7 were used as the set of corpora in ASPECT, since the pool activity reached maximum in round 7 during the original selection experiment. After the correlation coefficients of generated topics were examined (k = 16, tn = 10, Figure 2a), topics 3, 8, and 9 were selected as the "active" topics (μ = 0.9). From round 7 pool, 1294 sequences were predicted as active by ASPECT. Compared with RNABOB-based prediction, the precision and recall of ASPECT were 0.63 and 0.65, respectively (Figure 2b).

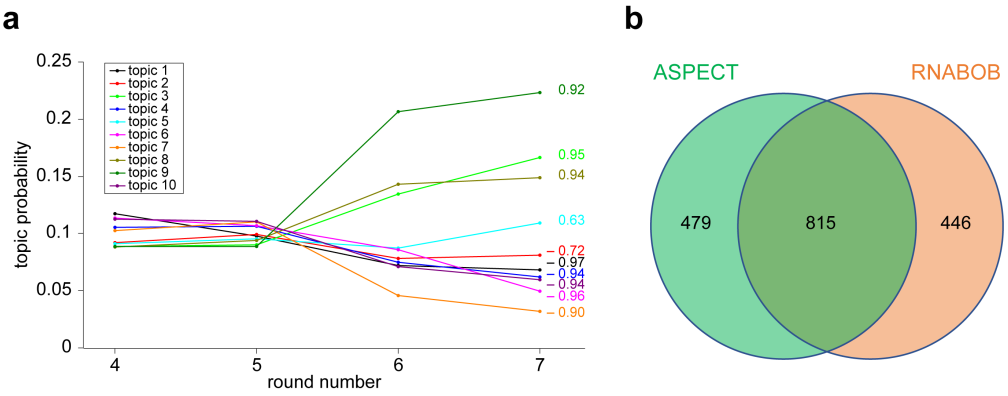

Figure 2: ASPECT with DAse dataset. (a) Topic possibility against selection rounds in DAse dataset. The correlation coefficients are labeled besides the data points of round 7. (b) Venn diagram of active sequences predicted by ASPECT and RNABOB.

Parameter optimization in ASPECT was also performed with the same set of corpora. For the representation of DNA sequences, different k-mer lengths were tested (k = 2, 4, 8, 16, and 32, Figure 3a). In general, the parameter k controls the vocabulary size (4^k) of the topic model used in ASPECT. On one side, when k is too small, the vocabulary won't be complicated enough to represent any functional motif in active sequences. On the other side, when k is too large, documents won't share enough words from the vast vocabulary, which makes topics hard to form. Our observation with the DAse data was consistent with the theoretical analysis. When k was 2, documents in the corpora were

over-simplified as bag of 2-mers, and ASPECT failed to predict any active sequence. For k values larger than 2, precision of ASPECT prediction was not sensitive to the change. Recall of ASPECT prediction showed an optimum for k = 16. Performance of ASPECT was then evaluated with different topic numbers, tn, while k was fixed as 16 (tn = 5, 10, 20, 30, 40, and 50, Figure 3b). Topic models with large topic numbers usually result with more specific and more coherent topics. However, too specific topics also lead to data sparseness that reduces the power of statistics. On the contrary, small topic numbers give topics with higher correlation but lower coherence. In the test case with DAse, it was observed that recall of ASPECT algorithm generally decreased with increasing tn, with the only exception of tn = 5. When tn = 5, the coherence of topics identified was too low that none of the topics passed the correlation threshold (μ = 0.9), leading to the failure of identifying any active sequences. In practice, it is recommended for the users to try a range of values for both k and tn to obtain the best prediction from ASPECT.

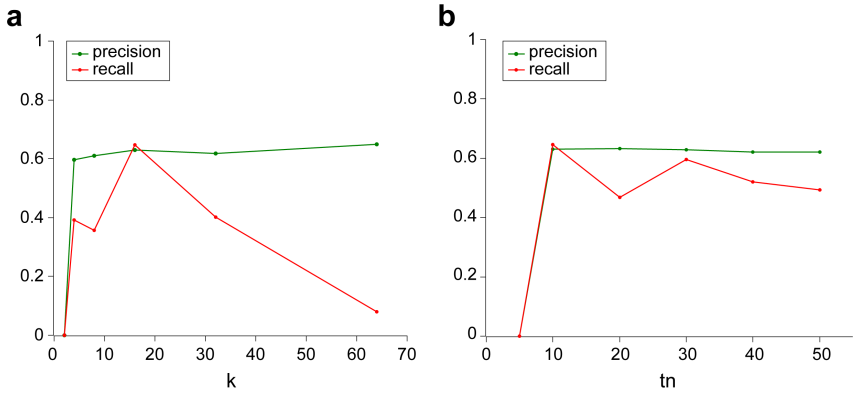

Figure 3: ASPECT parameter optimization with DAse dataset. (a) Evaluation of ASPECT with different k values. (b) Evaluation of ASPECT with different tn values.

4 Conclusion

An LDA-based active sequence predictive algorithm, ASPECT, was proposed in this paper. ASPECT regards the enrichment of active sequences in selection experiments as the enrichment of "active" topics in a series of corpora. The initial experiment of the proposed algorithm on DAse sequence data set showed effective precision and recall. Independent to structure-function relationship, ASPECT can facilitate the identification of DNA/RNA enzymes and aptamers from de novo selection experiments, by serving as a guidance for choosing candidates from sequencing data for downstream activity assay. Moreover, ASPECT can also be used to suggest active sequences from selection pool of early rounds, which might recover some active sequences lost during selection rounds due to PCR bias and other physical reasons. There is definitely still a lot of room for the ASPECT algorithm to upgrade. For example, the information of read numbers from sequencing data can be used as a priori knowledge for "active" topics in ASPECT. It is also possible to perform iterative refine of identified topics to increase the precision of ASPECT, as proposed in some casual topic mining works [23].

Availability

https://github.com/puzhou-wang/ASPECT

Reference:

[1] Pan, T.; Uhlenbeck, O. C. A small metalloribozyme with a two-step mechanism. *Nature* **1992**, *358*, 560-563.

[2] Lee, Y.; Klauser, P. C.; Brandsen, B. M.; Zhou, C.; Li, X.; Silverman, S. K. DNA-catalyzed DNA cleavage by a radical pathway with well-defined products. *J. Am. Chem. Soc.* **2017**, *139*, 255-261.

[3] Chandra, M.; Sachdeva, A.; Silverman, S. K. DNA-catalyzed sequence-specific hydrolysis of DNA. *Nat. Chem. Biol.* **2009**, *5*, 718-720.

[4] Parker, D. J.; Xiao, Y.; Aguilar, J. M.; Silverman, S. K. DNA catalysis of a normally disfavored RNA hydrolysis reaction. *J. Am. Chem. Soc.* **2013**, *135*, 8472-8475.

[5] Hager, A. J.; Szostak, J. W. Isolation of novel ribozymes that ligate AMP-activated RNA substrates. *Chem. Biol.* **1997**, *4*, 607-617.

[6] Camden, A. J.; Walsh, S. M.; Suk, S. H.; Silverman, S. K. DNA oligonucleotide 3'-phosphorylation by a DNA enzyme. *Biochemistry* **2016**, *55*, 2671-2676.

[7] Walsh, S. M.; Sachdeva, A.; Silverman, S. K. DNA catalysts with tyrosine kinase activity. *J. Am. Chem. Soc.* **2013**, *135*, 14928-14931.

[8] Chandrasekar, J.; Silverman, S. K. Catalytic DNA with phosphatase activity. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 5315-5320.

[9] Wang, P.; Silverman, S. K. DNA-catalyzed introduction of azide at tyrosine for peptide modification. *Angew. Chem. Int. Ed.* **2016**, *55*, 10052-10056.

[10] Seelig, B.; Jäschke, A. A small catalytic RNA motif with Diels-Alderase activity. *Chem. Biol.* **1999**, *6*, 167-176.

[11] Fusz, S.; Eisenführ, A.; Srivatsan, S. G.; Heckel, A.; Famulok, M. A ribozyme for the aldol reaction. *Chem. Biol.* **2005**, *12*, 941-950.

[12] Lato, S. M.; Boles, A. R.; Ellington, A. D. In vitro selection of RNA lectins: using combinatorial chemistry to interpret ribozyme evolution. *Chem. Biol.* **1995**, *2*, 291-303.

[13] Wang, J.; Gong, Q.; Maheshwari, N.; Eisenstein, M.; Arcila, M. L.; Kosik, K. S. et al. Particle display: a quantitative screening method for generating high-affinity aptamers. *Angew. Chem. Int. Ed.* **2014**, *53*, 4796-4801.

[14] Zhang, R.; Gu, Y.; Wang, Z.; Li, Y.; Fan, Q.; Jia, Y. Aptamer cell sensor based on porous graphene oxide decorated ion-selective-electrode: Double sensing platform for cell and ion. *Biosens. Bioelectron.* **2018**, *117*, 303-311.

[15] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic. Acids. Res.* **2003**, *31*, 3406-3415.

[16] Lorenz, R.; Bernhart, S. H.; Honer Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26.

[17] Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51-55.

[18] Ameta, S.; Winz, M. L.; Previti, C.; Jaschke, A. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic. Acids. Res.* **2014**, *42*, 1303-1310.

[19] Keiper, S.; Bebenroth, D.; Seelig, B.; Westhof, E.; Jaschke, A. Architecture of a Diels-Alderase ribozyme with a preformed catalytic pocket. *Chem. Biol.* **2004**, *11*, 1217-1227.

[20] Kraut, S.; Bebenroth, D.; Nierth, A.; Kobitski, A. Y.; Nienhaus, G. U.; Jaschke, A. Three critical hydrogen bonds determine the catalytic activity of the Diels-Alderase ribozyme. *Nucleic. Acids. Res.* **2012**, *40*, 1318-1330.

[21] Serganov, A.; Keiper, S.; Malinina, L.; Tereshko, V.; Skripkin, E.; Hobartner, C. et al. Structural basis for Diels-Alder ribozyme-catalyzed carbon-carbon bond formation. *Nat. Struct. Mol. Biol.* **2005**, *12*, 218-224.
[22] Wombacher, R.; Keiper, S.; Suhm, S.; Serganov, A.; Patel, D. J.; Jaschke, A. Control of stereoselectivity in an enzymatic reaction by backdoor access. *Angew. Chem. Int. Ed.* **2006**, *45*, 2469-2472.
[23] Kim, H. D.; Castellanos, M.; Hsu, M.; Zhai, C.; Rietz, T.; Diermeier, D. Mining causal topics in text data: iterative topic modeling with time series feedback; Kim, H. D.; Castellanos, M.; Hsu, M.; Zhai, C.; Rietz, T.; Diermeier, D., Ed.; ACM, 2013, pp 885-890.