HHS Public Access

Author manuscript

Res Comput Mol Biol. Author manuscript; available in PMC 2018 April 26.

Published in final edited form as:

Res Comput Mol Biol. 2017 May ; 10229: 389–390. doi:10.1007/978-3-319-56970-3.

Joker de Bruijn: Sequence Libraries to Cover All *k*-mers Using Joker Characters

Yaron Orenstein[1], **Ryan Kim**[2], **Polly Fordyce**[3], and **Bonnie Berger**[1]

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Research Science Institute, McLean, VA 22207, USA

[3]Stanford University, Stanford, CA 94305, USA

1 Introduction

Protein-DNA, -RNA and -peptide interactions drive nearly all cellular processes. Due to their high importance, high-throughput technologies using sequence libraries that cover all *k*-mers (i.e. words of length *k*) have been developed to measure them in a universal and unbiased manner [1]. These techniques all face a similar challenge: the space on the experimental device is limited, restricting the total sequence space that can be probed in a single experiment. While de Bruijn sequences cover all *k*-mers in the most compact manner, they remain $|\Sigma|^k$ characters long (where Σ is the alphabet, e.g. {A,C,G,T}). Here, we introduce a novel idea and algorithm for sequence design to cover all possible *k*-mers with a significantly smaller experimental sequence library by using *joker* characters, which represent all characters in the alphabet. Experimentally, such joker characters can be easily incorporated during oligonucleotide or peptide synthesis by using degenerate mixtures of nucleotides or amino acids, at no extra cost. However, joker characters introduce degeneracy which could potentially lower the statistical robustness of the measurements (as a measurement of a single oligonucleotide is now assigned to multiple sequences instead of just one). To address this challenge, we limit the use of joker characters to either one or two joker characters per *k*-mer, enabling the coverage of (*k*+2)-mers at the same cost and space of *k*-mers — a savings of a factor of $|\Sigma|^2$ in sequence length (16 and 400 for DNA and amino acid alphabets, respectively). We validate that the library remains capable of *de novo* identification of high-affinity *k*-mers by testing it on known DNA-protein binding data for hundreds of proteins. The implementation of our algorithm is freely available at `jokercake.csail.mit.edu`.

2 Methods

We propose a novel solution to the problem of generating a short sequence covering all *k*-mers using joker characters. The solution is based on two steps: (i) a greedy heuristic; (ii) and an ILP formulation. The greedy heuristic examines at each step an addition of $k - 1$ characters from Σ followed by a joker character. The addition that covers the most *k*-mers

Correspondence to: Bonnie Berger.

that are yet to be covered p times is chosen and added to the current sequence. The algorithm terminates when all k-mers have been covered at least p times. The ILP formulation minimizes the number of k-mers in the sequence under two sets of constraints. The first requires that all k-mers occur at least p times. The second guarantees that the k-mer occurrences can form a sequence. The ILP is solved using Gurobi ILP solver version 6.5.2 [2], where it is given the greedy solution as a starting solution.

3 Results

To test the performance of our algorithm, we ran it on different parameter combinations. We ran the greedy heuristic on $5 \leq k \leq 8$ for a DNA alphabet and $3 \leq k \leq 4$ for an amino acid alphabet, with $p = 1$. We then ran the ILP solver, starting from the greedy solution, with a time limit of 4 weeks. Results show that the greedy algorithm produces a sequence that is much smaller than the original de Bruijn sequence; i.e., less than 40% and 8% of the original for DNA and amino acid alphabets, respectively. Following the ILP solver, sequence length drops even further to less than 33% and 8% of the original, respectively, where the theoretical lower bounds are 25% and 5%, respectively. To test the performance of our algorithm in covering k-mers multiple times, we ran the greedy heuristic on $k = 6$, a DNA alphabet, and $1 \leq p \leq 16$. Here, we see that the greedy algorithm is producing a near-optimal sequence, less than 27% of the size of the original de Bruijn sequence for $p \geq 4$.

To demonstrate the utility of these libraries, we validated our performance when tested against a standard experimental 10-mer library of nearly 42,000 DNA sequences for which the binding affinities of hundreds of transcription factors is known [3]. Remarkably, our library correctly recovers the high-affinity target sites, despite a nearly 4-fold reduction in library size. We were able to handle 10-mer libraries due to a 100-fold speedup in implementation over a naive one for our joker library design.

4 Conclusion

We presented a new library design that covers all k-mers with a library of size that is almost $1/|\Sigma|$ (and possibly $1/|\Sigma|^2$) smaller than current libraries, making it possible to measure interactions of significantly longer k-mers while reducing both experimental footprint and cost. We have made the implementation and library designs freely available to others.

References

1. Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, Quake SR. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol. 2010; 28(9):970–975. [PubMed: 20802496]
2. Gurobi Optimization, I. Gurobi Optimizer Reference Manual. 2015. http://www.gurobi.com
3. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2014:gku1045.