Finding synonymous coding DNA sequences with maximum base pairing

Claudio Arbib[1,2], Andrea D'Ascenzo[1] Andrea Manno[1,2], and Fabrizio Rossi[1]

[1] Dipartimento di Scienze/Ingegneria dell'Informazione e Matematica
Università degli Studi dell'Aquila, L'Aquila, Italia
claudio.arbib@univaq.it
andrea.dascenzo@graduate.univaq.it
andrea.manno@univaq.it
fabrizio.rossi@univaq.it,
[2] Center of Excellence DEWS
Università degli Studi dell'Aquila, L'Aquila, Italia

1 Motivation

A *Coding DNA Sequence* (CDS) is the portion of a gene's DNA or RNA that, organized into nucleotide triplets called codons, encodes for a protein. During RNA filament folding, the nucleotides aggregate into secondary structures whose chemical (in)stability and stereometry may influence the interaction between RNA and ribosomes. Stability is also relevant in RNA vaccination, where mRNA coding sequences are inoculated in the host organism to make it produce proteins that the immune system can recognize as potentially dangerous[3].

Of the many indicators associated with a CDS, the *Base Pair Number* (BPN) is often used in the literature to define properties relevant to global stability: this indicator is defined as the largest number of hydrogen bonds a CDS can form, according to Watson-Crick model or extensions, in a feasible folding [6, 9]. Incrementing the BPN can actually increase the chances nucleotide substrings have to mutually form hydrogen links and hence fold into more or less complex secondary structures. In fact, a correlation of BPN to other stability indicators can generally be observed.

But, as different codons can encode for the same amino acid, the same protein corresponds to a large number of synonymous CDSs. So, when cloning vectors are inserted within a non-native organism to produce a specific protein, a problem arises of choosing a "good" CDS out of many different ones. In our context, it is then not sufficient to maximize the BPN of a single CDS, but is necessary to solve the more complex problem of finding, among all the possible synonymous CDSs that encode for a target protein, one that maximizes the BPN. To this aim, we propose an exact algorithm that combines DP with implicit enumeration and that, in our tests, proved to definitely outperform ILP approaches.

[3] Such techniques have recently been proposed for vaccine in the current COVID-19 pandemic [2, 5]

2 The problem

A protein, that is a sequence $\pi = \pi_1 \cdots \pi_n$ of *amino acids*, is encoded by an RNA filament formed by a sequence $b_1 \cdots b_{3n}$ of *nucleotide bases*, with $b_i \in \{\text{A,C,G,U}\}$. The i-th base triplet in the filament encodes for the i-th amino acid of π, and is called a *codon.* Note that there exist $4^3 = 64$ codons but only 20 amino acids, thus the same amino acid can be encoded by a (known) set of different codons.

Once formed, the RNA filament folds onto itself due to hydrogen bonds involving canonical base pairs A-U or C-G: in a feasible folding those bonds must also observe the following structural constraints:

- Pairs are separated by at least 4 bases;
- Base pairs are no-crossing, that is, one never has $i < i' < j < j'$ for any two given pairs (b_i, b_j), $(b_{i'}, b_{j'})$.

A basic question to be answered is then:

Problem 1. Given an RNA filament $b_1 \cdots b_m$, $m = 3n$, what is the largest possible number of hydrogen bonds it can form in a feasible folding?

The BPN maximization problem we address in this paper is more complex as it involves a search among a very large number of synonymous RNA sequences:

Problem 2. Find an RNA filament that, among those encoding for a given protein π, can form the largest possible number of hydrogen bonds.

3 The algorithm

Problem 2, as well as Problem 1, can be formulated in terms of Integer Linear Programming (ILP). However, unlike other codon optimization problems [1, 7], ILP is in this case known to behave quite badly due to a poor linear relaxation and to the large number of inequalities needed to model no-cross constraints [4].

An $O(\frac{m^3}{\log m})$ time dynamic programming (DP) algorithm is on the other hand available for Problem 1 [3]. Extending it to Problem 2 requires to handle local information associated with protein encoding. In particular, for those amino acids whose codons do not share a prefix (collectively denoted as Class II amino acids), base pair information should be carried on through successive DP recursions.

To overcome this difficulty, our approach leverages the efficiency of the DP method for fixed sequences together with an implicit enumeration scheme. Namely, we devise a combinatorial branch-and-bound method where the decision tree has a level for each Class II amino acid of the input protein π. At each node of the tree we operate a binary choice on the prefix to be used in the encoding, and find specific nucleotide sequences by DP:

- To get an upper bound, we find by DP a max BPN sequence where all the Class II amino acids not yet fixed admit codons associated with different amino acids (potentially leading to an infeasible sequence);

– To get lower bounds, we run DP by fixing Class II codons in a feasible way, according to the branches followed by the enumeration scheme up to that node.

By the very structure of the problem, each leaf of the decision tree is associated with a feasible solution.

Experimentally, our method proved to be efficient for both synthetic and real-life RNA sequences of ~ 200 nt, always finding optimal solutions in less than 4 minutes on average, after the exploration of a limited amount of branch-and-bound nodes ($\leq 9\%$ the total search space). Moreover, the optimal sequences found sensibly reduce the *Minimum Free Energy* and *Average Unpair Probability* [8, 10] of the original RNA synonym, thus improving the global stability of the RNA filament as measured by those indicators.

References

1. Arbib C, Pınar MÇ, Rossi F, Tessitore A. Codon optimization by 0-1 linear programming. *Computers & Operations Research* **119** (2020)
doi: 10.1016/j.cor.2020.104932
2. Chauhan G, Madou MJ, Kalra S, Chopra V, Ghosh D, Martinez-Chapa SO. Nanotechnology for COVID-19: Therapeutics and Vaccine Research. *ACS Nano* **14**, 7 (2020) 7760-7782
3. Frid Y, Gusfield D. A simple, practical and complete $O(\frac{n^3}{\log n})$-time Algorithm for RNA folding using the Four-Russians Speedup. *Algorithms for Molecular Biology* **5**, 13 (2010)
4. Gusfield D. *Integer Linear Programming in Computational and System Biology.* Cambridge University Press 2019, 108-109
https://doi.org/10.1017/9781108377737
5. McKay PF, Hu K, Blakney AK, et al. Self-amplifying RNA SARS-CoV-2 lipid nanoparticle vaccine candidate induces high neutralizing antibody titers in mice. *Nature Communications* **11**, 1 (2020): e3523
doi: 10.1038/s41467-020-17409-9
6. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for loop matchings. *SIAM J. on Applied Mathematics*, **35**, 1 (1978) 68-82
7. Sen P, Kargar K, Akgün E, Pinar MÇ. Codon optimization: a mathematical programming approach. *Bioinformatics* **36** (2020)
doi: 10.1093/bioinformatics/btaa248
8. Torabi S-F, Chen Y-L, Zhang K, et al. Structural analyses of an RNA stability element interacting with poly(A). *PNAS* **118**, 14 (2021) 1-11
9. Waterman MS, Smith TF. RNA secondary structure: A complete mathematical analysis. *Math. Biosc.* **42** (1978) 257-266
10. Wayment-Steele HK, Kim DS, Choe CA, et al. Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acid Research* **49**, 18 (2021) 10604-10617