UNIVERSIDADE NOVA DE LISBOA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

NOVA School of Science and Technology
NOVA University Lisbon

Visual representations of consensus sequences

Gil Oliveira
MSc in Biotechnology student

Submitted in partial requirements of the course of
Applied Bioinformatics

April 2020
Caparica, Portutugal

1 Introduction

Consensus sequences are sequences of nucleotides (on DNA and RNA) or amino acids (in protein), which are comprised of the most commonly encountered letters at that postion [**pierce2012**]. These sequences are generally associated with inter- or intramolecular interactions [**liljas2001**]. A prime example is the Shine-Dalgarno (SD) sequence in prokaryotes, which is involved in the binding of the ribossome to the mRNA.

Even though these sequences are highly conserved, they do present some variations in between organisms and thus, mearly writing out the consensus sequence leaves out information about the frequence and variation of each of the nucleotides or aminoacids in the sequence, which can be very important in Molecular Biology or Bioinformatics analysis.

Let's take the example of the aforementioned consensus sequence. In *E. coli* the SD sequence is **5'-AGGAGG-3'**. This sequence, however, has been shown to have slight variations in different bacteria [**ma2002**] and thus a bioinformatician, when programming a gene prediction tool, may be misled to believe that that's the full extent of the SD sequence, when it's not. Moreover, one may argue that studying the variety of consensus sequences is important in undertanding intermollecular interactions.

We then arrive at the logical conclusion that the visualisation of consensus sequences plays a very important role in various Molecular Biology and Bioinformatics studies and is therefore relevant to study better ways to display the available data in a way that's informative, visually appealing and easy to understand.

2 Sequence logos

Faced with the challanges of creating a visual representation of consensus sequences that provides more information than just the sequence of letters, Schnider and Stephens came forward with a proposal, for which they called Sequence Logos (Fig. 1) [**schneider1990**]. According to the authors, the sequence logos are able to provide the following information:

- The general consensus of the sequences;
- The order of predominance of the residues at every position;
- The relative frequency of the residues at every position;
- The amount of information present at every postiion in the sequence, measured in bits;
- The initiation point, cut point, ot other significate location (if appropriate).

In order to assemble these logos, one has to first align the sequences, so that a table of frequences can than be constructed [**schneider1990**]. That table

As we can see, this graphical method of representing data is a major step forward in relation with what existed before, as this logos allowed us to get a better feel for the relative frequences of the residues at every position, which is crucial for a more complete understanding of such sequences.

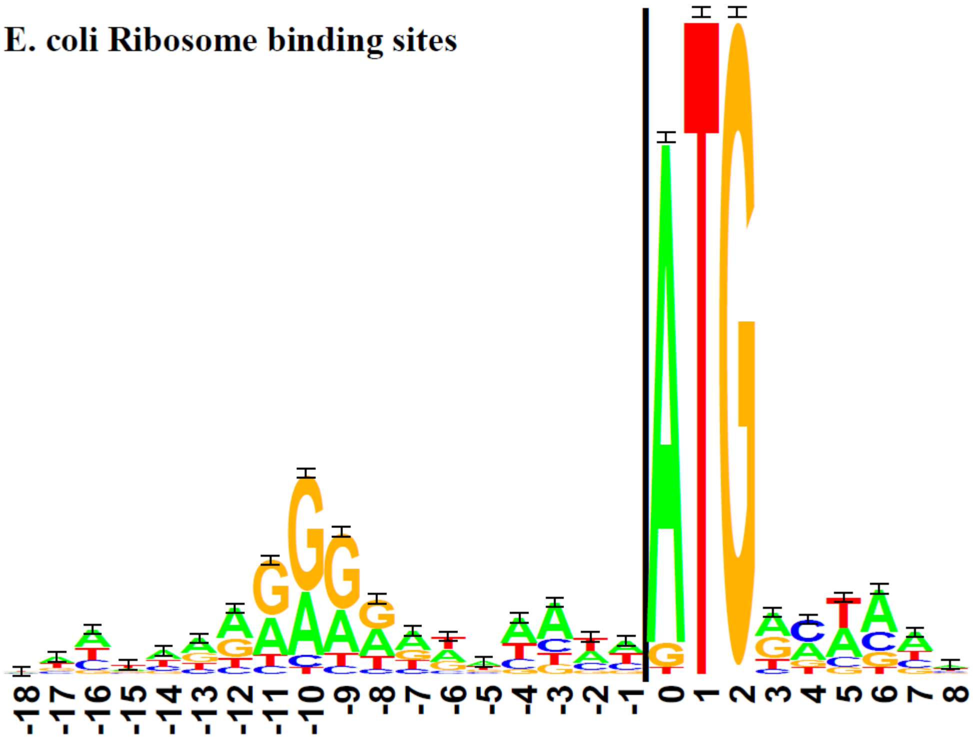

Figure 1: The Sequence Logo representation of the *E. coli* binding site [**schneider1990**].