

Lecture 15

Start Codon Prediction

February 24, 2000
Notes: Mingzhou Song

15.1. Experimental Results of Glimmer

The experimental results of Glimmer were presented by Delcher *et al.* [1]. They used the method described in Section 14.2 to predict genes in ten sequenced microbial genomes. The procedure was automated so as not to require human intervention. For each of the ten microbial genomes, the procedure was as follows.

In the training phase, they constructed a training set consisting of all ORFs longer than 500 bp with no overlap. The authors state that this set has more than enough data to train the interpolated context model accurately. They then trained the interpolated context model on the training set, as described in Section 14.2.1.

The identification phase and overlap resolution were then carried out as described in Sections 14.2.2 and 14.2.3. In each of the ten genomes, 99% of the annotated genes were correctly identified. The authors did not mention whether the start codons had been correctly identified in all cases.

Glimmer thus achieved a false negative rate of 1%, but also a false positive rate of 7–25% on each genome. The false negative rate is the percentage of annotated genes that were not identified by Glimmer. The false positive rate is the percentage of ORFs identified by Glimmer as genes, but not so annotated in the database. Of course, some of the annotations could be incorrect.

15.2. Start Codon Prediction

The accurate prediction of the translation start site, that is, the correct start codon, is important in order to analyze the putative protein product of a gene. Given the quality of the rest of the process, accurate start codon prediction is the most difficult remaining part of prokaryotic gene prediction. The gene-finding techniques discussed so far do little to predict the correct start codon among all the candidates.

Among the possible start codon candidates, what extra evidence can be used to identify the true translation start site? Recall from Section 2.2 that the ribosome is the structure that translates mRNA into protein and, at the initiation of that translation, is responsible for identifying the true translation start site. How does the ribosome perform this identification? Can we improve start codon prediction by mimicking the ribosome's process?

At the initiation of protein synthesis, the ribosome binds to the mRNA at a region near the 5' end of the mRNA called the *ribosome binding site*. This is a region of approximately 30 nucleotides of the mRNA that is protected by the ribosome during initiation. The ribosome binding site is approximately centered on the

<i>Bacillus subtilis</i>	5' ... CUGGAUCACCUCCUUUCUA	3'
<i>Lactobacillus delbrueckii</i>	5' ... CUGGAUCACCUCCUUUCUA	3'
<i>Mycoplasma pneumoniae</i>	5' ... GUGGAUCACCUCCUUUCUA	3'
<i>Mycobacterium bovis</i>	5' ... CUGGAUCACCUCCUUUCU	3'
<i>Aquifex aeolicus</i>	5' ... CUGGAUCACCUCCUUUA	3'
<i>Synechocystis sp.</i>	5' ... CUGGAUCACCUCCUUU	3'
<i>Escherichia coli</i>	5' ... UUGGAUCACCUCCUUA	3'
<i>Haemophilus influenzae</i>	5' ... UUGGAUCACCUCCUUA	3'
<i>Helicobacter pylori</i>	5' ... UUGGAUCACCUCCU	3'
<i>Archaeoglobus fulgidus</i>	5' ... CUGGAUCACCUCCU	3'
<i>Methanobacterium thermoautotrophicum</i>	5' ... CUGGAUCACCUCCU	3'
<i>Pyrococcus horikoshii</i>	5' ... CUCGAUCACCUCCU	3'
<i>Methanococcus jannaschii</i>	5' ... CUGGAUCACCUCC	3'
<i>Mycoplasma genitalium</i>	5' ... GUGGAUCACCUCC	3'

Table 15.1: 3' end of 16S rRNA for various prokaryotes

start codon (usually AUG). That is, the ribosome binding site contains not only the first few codons to be translated, but also part of the 5' untranslated region of the mRNA (see Section 2.3).

The ribosome identifies where to bind to the mRNA at initiation not only by recognizing the start codon, but also by recognizing a short sequence in the 5' untranslated region within the ribosome binding site. This short mRNA sequence will be called the *SD site*, for reasons that will become clear below. The mechanism by which the ribosome recognizes the SD site is relatively simple base-pairing: the SD site is complementary to a short sequence near the 3' end of the ribosome's 16S rRNA, one of its ribosomal RNAs.

The SD site was first postulated by Shine and Dalgarno [7] for *E. coli*. Subsequent experiments demonstrated that the SD site in *E. coli* mRNA usually matches at least 4 or 5 consecutive bases in the sequence AAGGAGG, and is usually separated from the translation start site by approximately 7 nucleotides, although this distance is variable. Numerous other researchers such as Vellanoweth and Rabinowitz [9] and Mikkonen *et al.* [6] describe very similar SD sites in the mRNA of other prokaryotes. It is not too surprising that SD sites should be so similar in various prokaryotes, since the 3' end of the 16S rRNA of all these prokaryotes is well conserved (Mikkonen *et al.* [6]). Table 15.1 shows a number of these rRNA sequences. Note their similarity, and in particular the omnipresence of the sequence CCUCCU, complementary to the Shine-Dalgarno sequence AGGAGG.

This SD site can be used to improve start codon prediction. The simplest way to identify whether a candidate start codon is likely to be correct is by checking for approximate base pair complementarity between the 3' end of the 16S rRNA sequence and the DNA sequence just upstream of the candidate codon. We say "approximate" complementarity because the ribosome just needs sufficient binding energy between the 16S rRNA and the mRNA, not necessarily perfect complementarity.

Several papers do use this SD site information to improve translation start site prediction. These papers are described briefly below.

Hayes and Borodovsky [4] found candidate SD sites by running a Gibbs sampler (Section 10.2) on the DNA sequences just upstream of a given genome's purported start codons. They then used the 3' end of the genome's annotated 16S rRNA sequence to validate the SD site so found.

Frishman *et al.* [2] used a greedy version of the Gibbs sampler to find likely SD sites. In addition, they

took into account the distance from the SD site to the start codon, which should be about 7 bp.

Hannenhalli *et al.* [3] used multiple features to score potential start codons. The features used were the following:

1. the binding energy between the SD site and the 3' end of the 16S rRNA, allowing “bulges” (that is, insertions and deletions) in the binding,
2. the identity of the start codon (AUG, UUG, or GUG),
3. coding potential downstream from the start codon and noncoding potential upstream, using GeneMark’s scoring function (Section 13.3),
4. the distance from SD site to start codon, and
5. the distance from the start codon to the maximal start codon, which is as far upstream in this ORF as possible.

They took the score of any start codon to be a weighted linear combination of the scores on these five features. The coefficients of the linear combination were obtained using mixed integer programming.

15.3. Finding SD Sites

Do all prokaryotes have SD sites very similar to the Shine-Dalgarno sequences of *E. coli*? Given the collection of DNA sequences upstream from its putative genes, how can we identify a prokaryote’s SD site, without reliance on the annotation of its 16S rRNA?

Tompa [8] proposed a method to discover SD sites by looking for statistically significant patterns (or *motifs*) in the sequences upstream from the putative genes. The method is reminiscent of the relative entropy site selection problem of Lecture 10 but, unlike the algorithms discussed there, this one is exhaustive, and guaranteed to find the most statistically significant motif. The statistical significance is measured by the “*z*-score”, defined below. The sites with the highest *z*-scores are very unlikely to be from the background and very likely to be potential SD sites.

For each possible *k*-mer *s*, this approach takes into account both the absolute number N_s of upstream sequences containing (an approximation of) *s*, and the background distribution. It then calculates the unlikelihood of seeing N_s such occurrences, if the sequences had been drawn at random from the background distribution. The random process used in this calculation is a 1st order Markov chain based on the sequences’ dinucleotide frequencies. (See Section 13.2.)

The measure of unlikelihood used is based on the *z*-score, defined as follows. Let N be the number of upstream sequences that are input, and p_s the probability that a single random upstream sequence contains at least one occurrence of (an approximation of) *s*. (See Tompa [8] for a description of how to compute p_s .) Then Np_s is the expected number of input sequences containing *s*, and $\sqrt{Np_s(1-p_s)}$ is its standard deviation. The *z*-score is defined as

$$z_s = \frac{N_s - Np_s}{\sqrt{Np_s(1-p_s)}}.$$

The measure z_s is the number of standard deviations by which the observed value N_s exceeds its expectation, and is sometimes called the “normal deviate” or “deviation in standard units”. See Leung *et al.* [5] for a

detailed discussion of this statistic. The measure z_s is normalized to have mean 0 and standard deviation 1, making it suitable for comparing different motifs s .

The algorithm was run on fourteen prokaryotic genomes. Those motifs with highest z -score showed a strong predominance of motifs complementary to the 3' end of their genome's 16S rRNA. For the bacteria, these were usually a standard Shine-Dalarno sequence consisting of 4–5 consecutive bases from AAG-GAGG. For the thermophilic archaea *A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii*, however, the significant SD sites uncovered were somewhat different. What is interesting about these is that their highest scoring sequences display a predominance of the pattern GGTGA or GGTG, which satisfies the requirement of complementarity to a substring near the 3' end of the 16S rRNA (see Table 15.1). However, that 16S substring is shifted a few nucleotides upstream compared to the bacterial sites discussed above.

References

- [1] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999.
- [2] D. Frishman, A. Mironov, H.-W. Mewes, and M. Gelfand. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Research*, 26(12):2941–2947, 1998.
- [3] S. S. Hannenhalli, W. S. Hayes, A. G. Hatzigeorgiou, and J. W. Fickett. Bacterial start site prediction. 1999.
- [4] W. S. Hayes and M. Borodovsky. Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction. In *Pacific Symposium on Biocomputing*, pages 279–290, 1998.
- [5] M.-Y. Leung, G. M. Marsh, and T. P. Speed. Over- and underrepresentation of short DNA words in herpesvirus genomes. *Journal of Computational Biology*, 3(3):345–360, 1996.
- [6] M. Mikkonen, J. Vuoristo, and T. Alatosava. Ribosome binding site consensus sequence of *Lactobacillus delbrueckii* subsp. *lactis*. *FEMS Microbiology Letters*, 116:315–320, 1994.
- [7] J. Shine and L. Dalgarno. The 3'-terminal sequence of *E. coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Science USA*, 71:1342–1346, 1974.
- [8] M. Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 262–271, Heidelberg, Germany, Aug. 1999. AAAI Press.
- [9] R. L. Vellanoweth and J. C. Rabinowitz. The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli* *in vivo*. *Molecular Microbiology*, 6(9):1105–1114, 1992.