Thermodynamic modeling of donor splice site recognition in pre-mRNA

Jeffrey A. Garland and Daniel P. Aalberts*
Physics Department, Williams College, Williamstown, Massachusetts 01267, USA
(Received 6 October 2003; published 26 April 2004; publisher error corrected 30 April 2004)

When eukaryotic genes are edited by the spliceosome, the first step in intron recognition is the binding of a U1 small nuclear RNA with the donor (5′) splice site. We model this interaction thermodynamically to identify splice sites. Applied to a set of 65 annotated genes, our "finding with binding" method achieves a significant separation between real and false sites. Analyzing binding patterns allows us to discard a large number of decoy sites. Our results improve statistics-based methods for donor site recognition, demonstrating the promise of physical modeling to find functional elements in the genome.

DOI: 10.1103/PhysRevE.69.041903 PACS number(s): 87.15.−v, 87.10.+e, 34.10.+x, 82.60.−s

The vast majority of bioinformatics methods treat nucleic acids as simple strings of characters, abstracted of their complex physical properties. To find biologically relevant areas hidden in vast genomic sequences, such methods analyze patterns of base frequencies extracted from large databases of known signals. While such methods yield results for many important problems, there are areas in which they have so far proved insufficient.

One important example is RNA splicing. Before being translated into proteins, RNA is processed in the nucleus. The spliceosome directs precursor messenger RNA to remove intervening sequences (introns), and to splice the remaining expressed sequences (exons) back together to form mature mRNA [1]. The splicing is done with great specificity, even though the apparent splicing signals are rather weak: at either end of the intron, only two bases are conserved and only about 4 bits of additional information are contributed from neighboring positions. There are additional signals from features like the "branch point" and the composition and length of introns themselves [2], but this information is not enough for current statistics-based methods [3] to accurately detect the sites which cells find so routinely.

One of the first, and simplest, statistical approaches to be applied to splice site detection is the weight matrix method (WMM) [4]. Data from known splice sites are compiled to estimate the probability $p_i(N_i)$ of finding nucleotide $N_i \in \{A,C,G,U\}$ at position i. The net splice site probability is approximated as the product of the nucleotide probabilities,

$$p_{\mathrm{wmm}} = \prod_i p_i(N_i). \tag{1}$$

More probable splice sites typically have higher p_{wmm} values, but the WMM neglects correlations between positions.

Identifying all the genes and other functional elements hidden within a genome is the first step following its sequencing. The alternation of coding and noncoding regions makes eukaryotic genes difficult to predict from primary sequence alone, so the ability to correctly identify the intron-exon boundaries is crucial to gene finding. Accurately identifying splice sites and other such functional areas *in silico* would make this process more efficient and complete, and it is considered one of the grand challenges of computational biology [5]. It is a curious fact that cells know nothing of abstract statistics and yet are able to detect splice sites with terrific accuracy. How do cells do it?

Thermodynamics.

We approach splicing as cells must, from a physical perspective. Our method, which we call "finding with binding," models the binding of the spliceosome to the pre-mRNA. The spliceosome comprises five different small nuclear RNAs (snRNAs) and well over one hundred different proteins [1,6–10]. The primary step in this process is when the spliceosome component called U1 snRNA binds to the"donor" splice site at the start (5′end) of each intron, at and around the donor's conserved *GU* sequence. We hypothesize that proper binding of the U1 is a good predictor of donor splice sites and that the signal information at the splice sites arises from natural RNA-RNA binding rules [11].

To test our hypotheses, we first must find the optimal U1-donor bound conformation by minimizing the free energy. Fortunately, Turner and others have measured the interaction free energies for *AU*, *GC*, and wobble-*GU* base pairs; for bulges, mismatches, and interior loops; and for hairpin loops [11]. While Turner's standardized experimental conditions (1 *M* NaCl, 0 *M* Mg^{++}, and 37°C) differ somewhat from those at the spliceosome, the free energies of the Turner model provide a reasonable starting point for calculation. These physical properties "train" our method, unlike statistical methods which require databases of sequence data for training.

What then is the optimal U1-donor bound state? Finding this conformation is quite similar to the problem of calculating the optimal free energy of a single-strand RNA fold. The MFOLD program [11,12], among others, uses Turner's free energy parametrization to predict the optimum fold. We employ MFOLD to perform the computation of interest [13], with only one minor alteration.

The fact that MFOLD folds single RNA molecules, while the U1 and the donor site are two distinct molecules, can be taken into account quite simply by joining them. As

*Corresponding author. Email address: aalberts@williams.edu

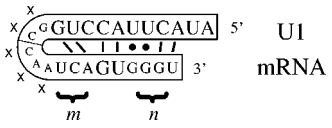

FIG. 1. A schematic for the "finding with binding" method. We model the association of a conserved 13 base region of the U1 snRNA (*AUACUUACCUGGC*) to possible precursor mRNA sites and compute the base-pairing conformation and free energy. Each possible pre-mRNA donor site includes the consensus *GU* plus the *n* bases after it, the *m* bases prior to it, and a three base contribution to the artificial linker region [shown is $(m,n)=(3,4)$]. The U1 and pre-mRNA sequences are concatenated into a single strand. Prohibiting base pairing (x) in the five-base artificial linker region, we find the optimal fold and its free energy. Note that bulges can shift the alignment of U1-mRNA pairing and that *GU* wobble base pairing (•) is included.

diagrammed in Fig. 1, the donor string contains (in 5′ to 3′ order) a three-base linker contribution which is prohibited from folding, *m* bases, the conserved *GU*, and *n* bases. Every pre-mRNA $(3+m+2+n)$-mer was concatenated to the relevant part of the conserved U1 snRNA sequence (*AUACUUACCUGGC*). Because of the high U1-donor complementarity and the unfavorability of folding each half independently (e.g., $\Delta G=+1$ kcal/mol for U1 folding alone), MFOLD folds the concatenated sequence into a hairpin structure. However, the free energy of this fold differs from the real U1-donor because of the hairpin loop formation penalty.

To eliminate the loop entropy contribution (which depends on the loop length), we modified the MFOLD input parameters, setting the loop entropy penalty to a constant value $\Delta G_{\mathrm{loop}}(N)=+5.4$ kcal/mol, for all loop lengths $N\geqslant 3$ (note that $N<3$ is too short to constitute a hairpin). Although we have taken the minimum ΔG_{loop} value [11] for allowed hairpins, conformations with multiple hairpins are strongly penalized and were not seen. We use the "prohibit folding" option of MFOLD to prevent the middle five connector nucleotides from pairing; however, the first and fifth do affect the free energy via dangling-end and terminal-mismatch bonuses.

To validate our method, we used the test set of Burge and Karlin [14], itself based on the Kulp/Reese set derived from GenBank Release 95 [15]. This set contains 65 genes, with 338 annotated real splice sites in coding (CDS) regions. In the same 330 kilobase CDS region, there are 16 961 appearances of the consensus *GU* sequence not annotated as splice signals, which are labeled "decoy" sites. The optimal folds and their free energies (with altered loop entropies, as described above) were then calculated using MFOLD. It is not known how much of the donor sequence is available to pair with the U1, so a range of different (m,n) values from $(0,2)$ to $(6,10)$ were analyzed.

Figure 2 shows the results of folding real and decoy sites, arranged by free energy. Clearly, real sites bind to the U1 at a lower average free energy than decoy sites. Real and decoy distributions do still overlap. Next, we examine the folding of the bound pairs and reject all sites in which the *GU* in the potential donor site does not pair with the corresponding *AC*

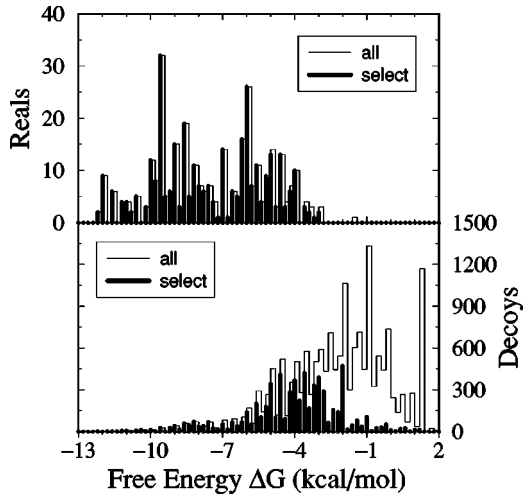

FIG. 2. Histograms of the free energy of U1-mRNA binding. Binding energies for all 338 real and 16 961 decoy sites are given, as well as the select subset whose *GU* consensus pairs with the corresponding *AC* in the U1 snRNA. Notice that the selection process rejects 64% of the decoy sequences, while rejecting only 2% of real sequences. The data shown are for $(m,n)=(2,6)$.

in the U1 sequence. This simple check eliminates roughly two-thirds of the decoy sites at the cost of as little as 2% of the reals, as shown in Fig. 2. For the selection step, we found that using $(m,n)=(2,6)$ gave the best results.

Requiring the U1 to bind optimally to a donor subsequence forces the correct alignment to compete against alternate alignments. Only good candidates advance for further screening, creating a more favorable data set with vastly fewer decoys. Any number of techniques, statistical or physical, can then be employed to score the remaining candidates.

We score the sequences both by their binding free energies, and by the WMM of Eq. (1) which is trained statistically [16]. By cross-correlating these methods, we are also able to investigate the claim that nucleotide biases in sequences might be interpreted as an effective free energy [17]. In Fig. 3, we compare $-RT\ln(p_{\mathrm{wmm}})$ with ΔG. It is interesting to see how poorly correlated the log of the WMM's "Boltzmann weight" is to the binding free energy. The scatter of the points is significant, with correlation coefficient *r*

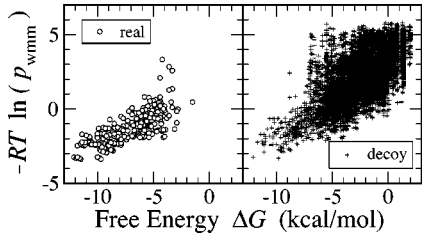

FIG. 3. Comparing $-RT\ln(p_{\mathrm{wmm}})$ and ΔG at $T=37$ °C allows us to evaluate the claim that biases in sequences represent a free energy. The log of the WMM's "Boltzmann weight" shows significant scatter when plotted against the binding free energy. The best-fit slopes are inconsistent with the hypothesis that p_{wmm} can effectively estimate binding free energies. The data shown are for all 338 real and 16 961 decoy sequences.

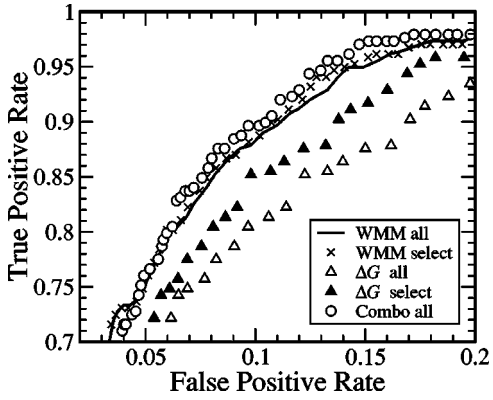

FIG. 4. The receiver operating curves for the weight matrix method (WMM), the finding with binding free energy method (ΔG), and the combined method [Eq. (2)] are shown for all data, and after selecting for pairing at the *GU*. Interestingly, the accuracy of the combo method did not improve with selection so we do not present values for the combo-select method.

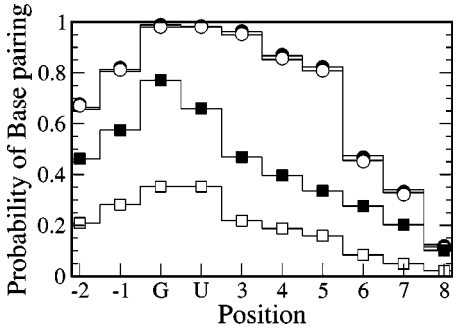

FIG. 5. The probability of base pairing is position dependent for real (circles) and decoy (box) sequences (filled, all; open, after selecting). Selecting for proper base pairing at the *GU* decreases the number of decoy sequences at a cost of relatively few real sites. The physical selection procedure improves the results of statistics-based approaches (see Fig. 4). Data are given for $(m,n)=(2,6)$.

$=0.79$ for the reals and $r=0.69$ for the decoys. Furthermore, instead of being related with a slope $m=1$, one finds $m=0.388\pm0.016$ for the reals and $m=0.498\pm0.004$ for the decoys.

To make predictions about the reality of a sequence, a cutoff is chosen. Reducing the cutoff probability score, or increasing the free energy cutoff, increases the number of both real and decoy sites identified. The WMM method corresponds to choosing horizontal lines in Fig. 3; the free energy method corresponding to choosing vertical lines. We also combined these methods, employing a quadratic discriminant analysis cutoff rule which can be visualized as a circle in Fig. 3, in which the algorithm marks as real all candidate sequences satisfying

$$[\Delta G-\Delta G_{\min}]^2+\left[-RT\ln\left(\frac{p_{\mathrm{wmm}}}{p_{\mathrm{wmm,min}}}\right)\right]^2<C^2, \qquad (2)$$

for different cutoff values C.

In finding with binding, the cutoff energy can be understood in physical terms. Each splice site can be occupied by either zero or one U1 molecules. This condition is reminiscent of Pauli exclusion. The probability of occupying a particular binding site can be estimated with Fermi-Dirac statistics as

$$p_{\mathrm{occ}}=\frac{1}{\exp\{(\Delta G-\mu)/RT\}+1}, \qquad (3)$$

where μ is the chemical potential of U1 factors with a logarithmic dependence on the U1 concentration. Equation (3) is often approximated as a step function.

To assess the accuracy of our methods, we measure the (true positive rate) = (true positives)/(all reals) and (false positive rate) = (false positives)/(all decoys). In Fig. 4, results are shown for the WMM, the finding with binding free energy method, and the Eq. (2) combination method. Scoring with free energies alone does not perform as well as WMM, though it is possible, by improving thermodynamic parameters, reshaping Eq. (2) as an ellipse, or employing other selection criteria, that this could improve. For example, the lowest ΔG may not be optimal for splicing. Any $\Delta G<\mu$ will have a significant p_{occ}, but tighter binding may slow later reactions. Indeed, in this data set, the consensus sequence is no more likely than a number of other sequences with one to three mutations. The fact that a number of identical sequences appear in both the real set and decoy set indicates a greater role for the bases farther away from the splice site and for the secondary structure [18].

Because its engine is an RNA folding algorithm, the finding with binding method naturally accommodates the effects of secondary structure. It is even possible to calculate the free energy of refolding the pre-mRNA to expose the binding sites. We hypothesize that differences in preexisting secondary structure may separate sites with identical primary sequence. The interactions between donor and branch sites can also be included via polymer physics modeling.

A physical modeling approach also provides detailed predictions about base pairing (see, for example, Fig. 5). It is not a surprise to see strong evidence that there must be base-pairing at the consensus sequence in order for the spliceosome to function. Bulges and mismatches are costly, making it difficult to resume base pairing after a duplex is disrupted. Furthermore, since our method predicts exactly how the U1 and mRNA base pair, a more thorough analysis of these binding patterns could suggest exactly which contacts help the spliceosome recognize real splice sites.

The present results demonstrate that physical modeling enhances splice site detection, complementing mature statistics methods [3] and providing mechanistic insights. Association of nucleic acids is central to many other important biological processes, including gene expression. We believe it will be advantageous to include the physical interactions between the biopolymers and to approach many genomic problems from a physical perspective. While this is more difficult when proteins are involved, predicting the affinities of nucleic acids can and should be done.

What is most promising is that physical methodologies

can be applied to a wide variety of related problems, such as the recognition events of the following: the U2-branch splice site, alternative splicing [19], retrotransposons [19,20], short interfering RNAs [19,21,22], Shine-Dalgarno sequences [1,19], and the snoRNA-rRNA associations which guide methylation and pseudouridylation [1,19,23].

The authors thank Michael Zuker and Washington University for making available MFOLD 3.0 to run on our workstations, and thank Eric Daub, Nathan Hodas, and Lois Banta for assistance. This research was supported by Research Corporation and a grant by the National Institutes of Health Grant No. (GM068485).

[1] H. Lodish , *Molecular Cell Biology*, 4th ed. (W. H. Freeman, New York, 2000).
[2] L. P. Lim and C. B. Burge, Proc. Natl. Acad. Sci. U.S.A. **98**, 11 193 (2001).
[3] *Computational Methods in Molecular Biology*, edited by S. L. Salzberg, D. B. Searls, and S. Kasif (Elsevier, Amsterdam, 1998).
[4] R. Staden, Nucleic Acids Res. **12**, 505 (1984).
[5] F. S. Collins, , Nature (London) **422**, 835 (2003).
[6] Y-T. Yu , in *The RNA World, Second Edition*, edited by R. F. Gesteland, T. R. Cech, and J. F. Atkins (Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY, 1999), pp. 487–524.
[7] C. B. Burge, T. Tuschl, and P. A. Sharp, in *The RNA World, Second Edition*, edited by R. F. Gesteland, T. R. Cech, and J. F. Atkins (Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY, 1999), pp. 525–560.
[8] J. A. Wise, Science **262**, 1978 (1993).
[9] J. P. Staley and C. Guthrie, Cell **92**, 315 (1998).
[10] Z. Zhou , Nature (London) **419**, 182 (2002).
[11] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).
[12] M. Zuker, D. H. Mathews, and D. H. Turner, in *RNA Biochemistry Biotechnology*, edited by J. Barciszewski and B. F. C. Clark (Kluwer, Dordrecht, 1999), pp. 11–43.
[13] Other algorithms which calculate free energies for oligomeric associations include D. H. Mathews , RNA **5**, 1458 (1999); M. Zuker, Nucleic Acids Res. **31**, 3406 (2003); M. Andronescu , *ibid.* **31**, 3416 (2003); N. O. Hodas and D. P. Aalberts (unpublished).
[14] C. Burge and S. Karlin, J. Mol. Biol. **268**, 78 (1997).
[15] http://www.fruitfly.org/seq_tools/datasets/Human/GENIE_95/
[16] The WMM is trained on the donor sites of the 380-gene Burge/Karlin learning set [14], calculating base-frequencies at positions starting three bases before and ending four bases after the consensus *GU* splice site.
[17] M. Q. Zhang and T. G. Marr, CABIOS, Comput. Appl. Biosci. **9**, 499 (1993).
[18] Although many believe that the secondary structure of the pre-mRNA affects splicing, methods to predict splicing using this information have thus far had only limited success: D. J. Patterson, K. Yasuhara, and W. L. Ruzzo, in *Pacific Symposium on Biocomputing 2002*, eds. R. B. Altman, K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein (World Scientific, Singapore, 2002), pp. 223–234.
[19] T. A. Brown, *Genomes*, 2nd ed. (Wiley, New York, 2002).
[20] K. Ichiyanagi , Mol. Microbiol. **46**, 1259 (2002).
[21] J. Couzin, Science **298**, 2296 (2002).
[22] R. F. Ketting , Genes Dev. **15**, 2654 (2001).
[23] T. M. Lowe and S. R. Eddy, Science **283**, 1168 (1999).