Designing RNA Structures: Natural and Artificial Selection

Barry Cohen*
Dept. of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102
(212) 877-5654, fax (978) 359-5945
bcohen@cis.njit.edu

Steven Skiena†
Dept. of Computer Science
SUNY Stony Brook
Stony Brook, NY 11794-4400, USA
(631) 632 9026
skiena@cs.sunysb.edu

ABSTRACT

Messenger RNA (mRNA) sequences serve as templates for proteins according to the triplet code, in which each of the $4^3 = 64$ different codons (sequences of three consecutive nucleotide bases) in RNA either terminate transcription or map to one of the 20 different amino acids (or residues) which build up proteins. Because there are more codons than residues, there is inherent redundancy in the coding. Certain residues (e.g. tryptophan) have only a single corresponding codon, while other residues (e.g. arginine) have as many as 6 corresponding codons. This freedom implies that the number of possible RNA sequences coding for a given protein grows exponentially in the length of the protein.

Thus nature has wide latitude to select among mRNA sequences which are informationally equivalent, but structurally and energetically divergent. In this paper, we explore how nature takes advantage of this freedom, and how to algorithmically design structures more energetically favorable than have been built through natural selection. In particular:

- *Natural Selection* – We perform the first large-scale computational experiment comparing the stability of mRNA sequences from a variety of organisms to random synonymous sequences which respect the codon preferences of the organism. This experiment was conducted on over 27,000 sequences from 34 microbial species with 36 genomic structures. We provide evidence that in all genomic structures highly stable sequences are disproportionately abundant, and in 19 of 36 cases highly unstable sequences are disproportionately abundant. This suggests that the stability of mRNA sequences is subject to natural selection.
- *Artificial Selection* – Motivated by these biological results, we examine the algorithmic problem of designing the most stable and unstable mRNA sequences which code for a target protein. We give a polynomial-time dynamic programming solution to the *most stable sequence problem* (MSSP), which is asymptotically no more complex than secondary structure prediction. We show that the corresponding *least stable sequence problem* (LSSP) is NP-complete, but develop two heuristics for the construction of such sequences.

  We have implemented these algorithms, and present experimental results placing the high/low stability sequences in context with both wildtype and random encodings. Our implementation has already been applied to the design of RNA "code-words" creating little or no secondary structure in RNA computing [**?**, **?**], and we anticipate a variety of other applications of this work to sequence design problems [**?**].

1. NATURAL SELECTION ON MRNA STABILITY

We use newly-available whole genome data and computational methods to shed light on an interesting question. *Why, over the course of eons of evolution, has nature chosen one of all these possible RNA encodings?* Has it, indeed, "chosen," or is the encoding we see simply a snapshot of a particular moment in a random walk through this space, without functional significance?

We conjecture that the shape and stability of an mRNA molecule enhance or impede the chemical machinery which translates it into a protein, and so impact the chances of the organism for survival and reproduction. In particular, we believe that nature exerts varied selective pressures on mRNA sequences to form secondary structures, or to avoid them.

There is some biological evidence to support such a notion. Although the study of mRNA has focused largely on its protein coding function, as the putative aboriginal biotic material [5] RNA would have been subject to selection for structure well before its protein-coding role evolved [15]. The transcription and translation of mRNA in the course of protein production expose it to varied processes and environments. Its life cycle may require – depending on the organism – formation and breaking of secondary and tertiary structure, the excision of introns, passage through an organelle membrane and persistence in the cytoplasm. Each

*Corresponding author.

†This work is partially supported by NSF Grant CCR-9625669 and ONR Award N00149710589.

phase in its life cycle offers possibilities for structure-based selection. RNA structure is also known to play a regulatory role [13]. Flamm *et.al.* [3] have recently demonstrated the existence of bistable RNAs which are easily accessible in evolution, and which could serve as conformational switches.

1.1 Experimental Design

We compare the stability of actual ("wildtype") mRNA coding sequences with randomly-composed synthetic sequences in an *in silico* experiment. Skewed distribution of stability of wildtype sequences relative to random ones provides evidence of natural selection on this morphological characteristic of the molecular species.

Our randomized construction of synthetic sequences respects two constraints. First, the protein product of each coding sequence is fixed, so the codon at each position in the synthetic coding sequence is chosen from the set of synonymous codons for that amino acid. Second, in selecting among the synonymous codons, the probability of occurrence of each codon is determined by the codon preference displayed in the organism as a whole.

The former constraint reflects the fact that coding sequences change much more rapidly than proteins. The latter constraint follows from the fact that the pattern of codon use is fairly consistent within a given organism, but differs sharply between organisms. Furthermore, codon choice seems to reflect a set of constraints and biases which operate on codon choice at the organism level [4]. tRNA abundance is known to be a mechanism for such regulation of codon bias. There is a stochastic "hesitation" in the translation of a codon with a low tRNA level; the codon with the more abundant tRNA leads to a more effective expression of the gene. If no tRNA is present for a given codon, as is occasionally the case, the codon is effectively a stop codon.

Our experiment, therefore, reflects a model in which the coding targets are fixed, and evolutionary processes explore the coding space for these targets.

1.2 Data Set

The data set consists of coding sequences taken from all 34 microbes whose complete genomes were available at the begining of our study. These include 36 genomic structures, since two organisms (*D. radiodurans* and *V. cholerae*) contain two structures each. All short coding sequences – those of less than 750 nucleotides – are used. The genomes were acquired from the NCBI database. Codon use tabulations are from `http://www.kazusa.or.jp/codon/`.

A total of 27,207 sequences, constituting approximately one-third of the coding sequences in these organisms, are included in the study. The organisms are all eubacteria or archae, and therefore enjoy similar transcription and translation mechanisms.

For each coding sequence, we construct five random synonymous sequences – a total of over 160,000 sequences.

Each sequence – both wildtype and randomly generated – is then computationally "folded." That is, a predicted minimum-free-energy secondary structure is calculated, according to the nearest neighbor dynamic programming algorithm. We used our implementation of the algorithm described in [19], which is a special case of the design programs in Section 2.

The restriction to short coding sequences and a relatively small number of random trials is due to the fact that the folding algorithm is cubic in the length of the sequence. The calculations, which were done on a network of Unix workstations, required about 300 cpu-days. Extending the study to the longer two-thirds of coding sequences would require approximately 100 times as much computation. The accuracy of the algorithm is also know to decline on longer inputs. For sequences of fewer that 700 nucleotides, recent studies [12] show that, on average, 73% of known base pairs are predicted correctly.

1.3 Experimental Results

First, we define the energy rank $R(S)$ of a wildtype sequence. For each wildtype sequence S_{wild}, we have constructed t random sequences, $S_{i.rnd}$, and computed the minimum free energy G of each. (In our experiment, $t = 5$.) The energy rank of a sequence $R(S_{wild})$ is the number of $S_{i.rnd}$ such that $G(S_{i.rnd}) < G(S_{wild})$. Hence, $R(S_{wild}) = 0$ if R_{wild} is more stable (less energetic) than all of the random sequences. $R(S_{wild}) = t$ if R_{wild} is less stable (more energetic) than all of the random sequences.

The energy ranks define the *stability signature* of the structure. Let $R_j(O)$ be the number of sequences of O of rank j divided by the expected number of sequences of rank j. $R_j(O)$ is a normalized function; its expected value is 1.0 for all $0 \leq j \leq t$. The stability signature $R(O)$ of structure O is the set of all $R_j(O)$, $0 \leq j \leq t$.

Several observations can be made about the observed stability signatures:

- A bias toward highly stable sequences is evident in all genomic structures. In each of the 17 genomic signatures which are unimodal, highly stable sequences are the mode. Among the 19 signatures which are bimodal, highly stable sequences are the primary mode in 10 and the secondary mode in 9. In 34 of 36 cases, the number of stable structures exceeds the expected value. The high value for R_0 – that is, the structure most strongly skewed toward stable sequences – belongs to *M. genitalium*, which also has the low value for R_5.

- Highly unstable structures are overrepresented in 19 of 36 cases. In 9 they are the primary mode; in 10 they are the secondary mode. The high value for R_5 – that is, the structures most strongly skewed toward unstable sequences – belongs to *X. fastidiosa*, which also has the low value for R_0.

- Nineteen stability signatures are bimodal, favoring both highly stable and highly unstable sequences. Among all structures there is a clear bimodal distribution; the most stable sequences (R_0) appear 1.544 times as frequently as in a neutral distribution; the least stable sequences (R_5) appear 1.092 times as frequently as expected.

- The skewing toward high and low energy structures is statistically significant. For each genomic structure, we calculate the probability that an equal or greater mode would occur by chance. For 31 of 36 structures, the probability is less than 0.001; for 33 it is less than 0.01. The highest probability is 0.2. In every case, the mode is either R_0, the most stable rank, or R_5, the least stable rank. It is unlikely (3^{-36}) that all modes would occupy one of these extremes by chance.

structure	abbr	taxonomy	# cds	% GC
Aeropyrum pernix K1	aero	archae	1666	0.578
Aquifex aeolicus DSM4304	aquae	bacteria	594	0.431
Archaeoglobus fulgidus VF5	aful	archae	1181	0.482
Bacillus halodurans C-125	bhal	bacteria	1676	0.430
Bacillus subtilis 168	bsub	bacteria	447	0.416
Borrelia burgdorferi B31	bbur	bacteria	320	0.278
Campylobacter jejuni NCTC 11168	cjej	bacteria	256	0.298
Chlamydia pneumoniae AR39	cpneuA	bacteria	318	0.399
Chlamydia pneumoniae CWL029	cpneu	bacteria	401	0.402
Chlamydia trachomatis MoPn	ctraM	bacteria	329	0.406
Chlamydia trachomatis serovar D	ctra	bacteria	346	0.413
Deinococcus radiodurans R1 chr 1	dra1	bacteria	1075	0.668
Deinococcus radiodurans R1 chr 2	dra2	bacteria	113	0.670
Escherichia coli K12-MG1655	ecoli	bacteria	1799	0.501
Haemophilus influenzae KW20	hinf	bacteria	746	0.378
Halobacterium sp. NRC-1	hbsp	archae	558	0.671
Helicobacter pylori 26695	hpyl	bacteria	673	0.387
Helicobacter pylori J99	hpyl99	bacteria	596	0.392
M. jannaschii DSM 2661	mjan	archae	744	0.312
M. thermoautotrophicum delta H	mthe	archae	900	0.494
Mycobacterium tuberculosis H37Rv	mtub	bacteria	1345	0.650
Mycoplasma genitalium G37	mgen	bacteria	177	0.315
Mycoplasma pneumoniae M129	mpneu	bacteria	187	0.398
Neisseria meningitidis NC58	nmen	bacteria	1001	0.500
Neisseria meningitidis Z2491	nmenA	bacteria	959	0.504
Pseudomonas aeruginosa PAO1	paer	bacteria	1896	0.661
Pyrococcus abyssi GE5	pabyssi	archae	654	0.446
Pyrococcus horikoshii (shinkaj) OT3	pyro	archae	1023	0.420
Rickettsia prowazekii Madrid E	rpxx	bacteria	336	0.295
Thermoplasma acidophilum DSM 1728	tacid	archae	314	0.464
Thermotoga maritima MSB8	tmar	bacteria	600	0.455
Treponema pallidum Nichols	tpal	bacteria	292	0.534
Ureaplasma urealyticum serovar 3	uure	bacteria	253	0.257
Vibrio cholerae El Tor N16961 chr 1	vcho1	bacteria	1149	0.469
Vibrio cholerae El Tor N16961 chr 2	vcho2	bacteria	573	0.451
Xylella fastidiosa 9a5c	xfas	bacteria	1537	0.517

Table 1: Organisms used in the study

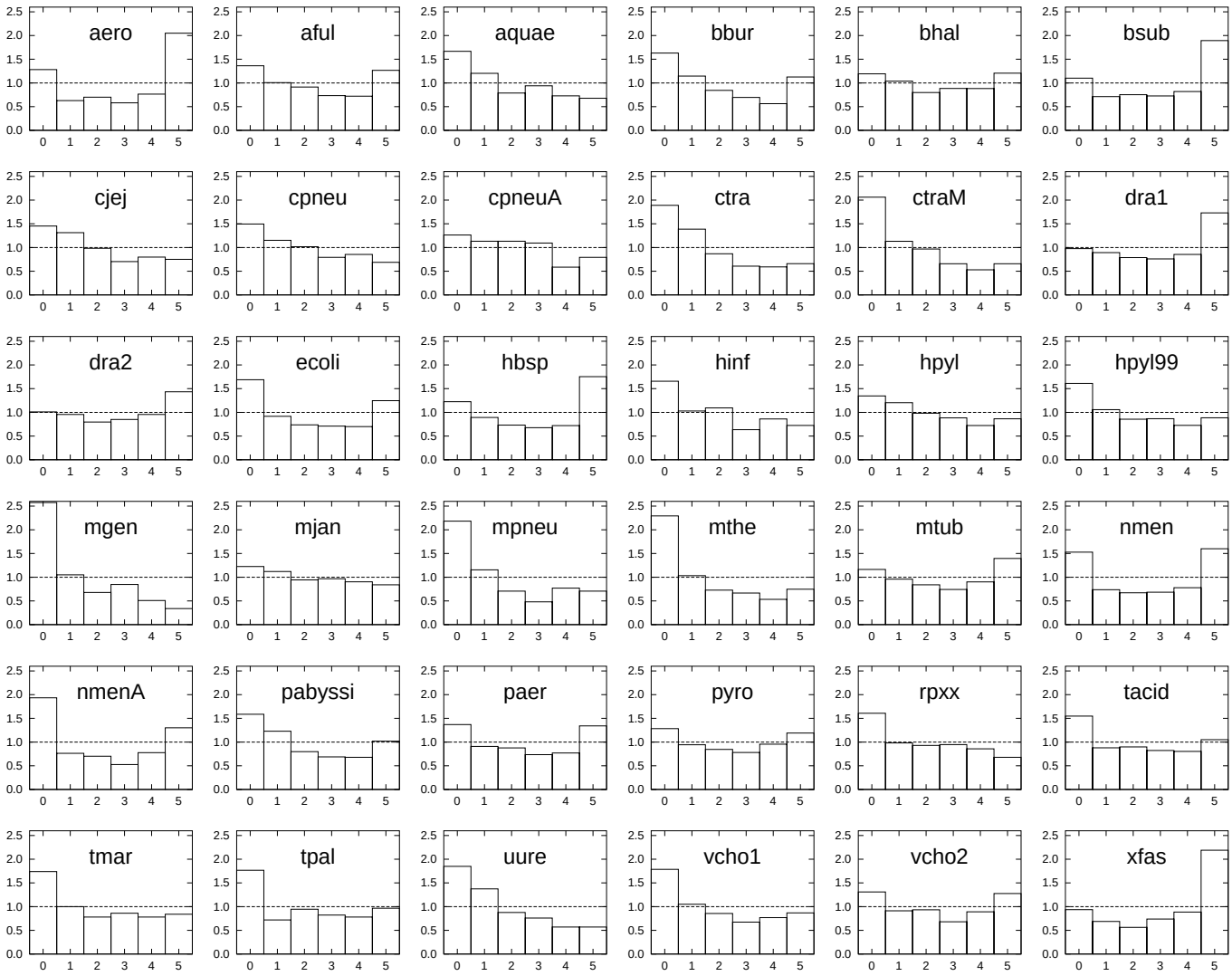

Table 2: Stability signature of genomic structures. Each wildtype sequence occupies one of six possible "energy ranks" – the number of synthetic sequences which are more stable than it. The histogram of the count of sequences in these ranks provides a stability signature for the structure. Horizontal dotted lines show the signature expected in the absence of selective pressures on stability. The leftmost (0) bar shows the frequency of highly stable sequences; the rightmost (5) bar shows the frequency of unstable sequences.

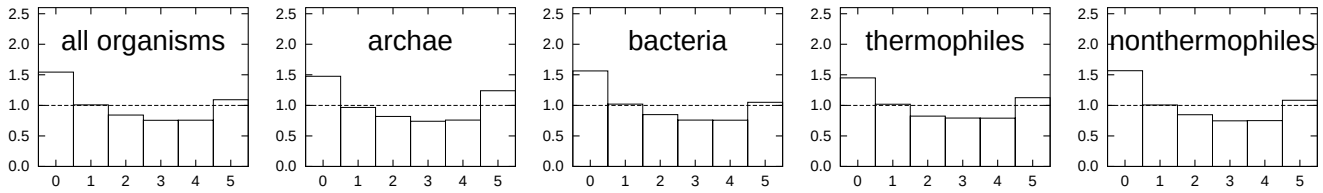

Table 3: Stability signature of groups of organisms.

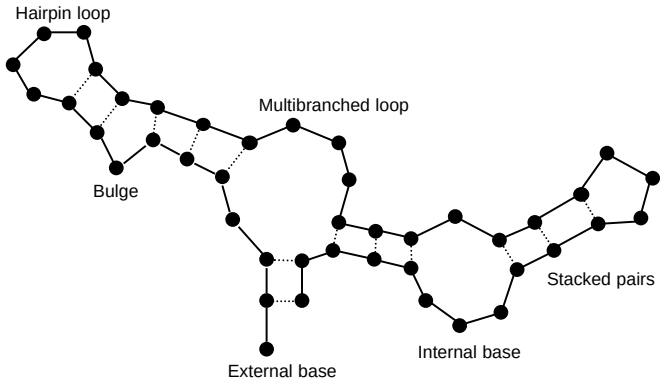

Figure 1: Illustration of substructures in an example RNA structure. Bases are depicted by dots, the RNA backbone by solid lines, and base pairings by dotted lines.

- We initially conjectured that selection for stable structure might occur more intensely under extreme temperatures and pressures. The data does not support this. There are seven thermophiles in our data set (*A. pernix*, *A. fulgidus*, *A. aeolicus*, *M. jannaschii*, *P. abyssi*, *P. horikoshii*, *T. maritima*). They do not display any markedly different pattern of stability the entire data set. No discernable bias is reflected by the Archae or any other subclass we identified.

2. RNA OPTIMIZATION

In view of this evidence, we seek to design the most stable and unstable RNA sequences which codes for a specifies protein. Our work is related to the *inverse folding problem*, one is given a structure and asked to design a sequence which produces it [1].

In this section, we describe an efficient algorithm for finding the RNA sequence coding for a given protein P which has the minimum energy (most stable structure) over all encodings of P, and describe an implementation of our algorithm. Our algorithm has the same asymptotic complexity as standard algorithms which simply fold a given RNA encoding.

We also demonstrate that the analogous problem of finding the RNA sequence coding for a given protein P which has *maximum* energy (minimum structure) over all encodings of P is NP-complete. This implies that we are unlikely to find efficient algorithms to assure finding either an RNA sequence with minimum structure or one realizing an arbitrary target shape. Still, we develop two efficient heuristics for constructing high energy (unstable) coding sequences for a target protein, and present extensive experimental results from implementations of these algorithms.

2.1 Secondary Structure Prediction Models

Ribonucleotides tend to pair with each other according to the same pattern as the helices formed by antiparallel strands of DNA – U with A, and C with G. Wobble pairs – GU and UG – are also included in the canonical set of pair bonds. Nucleotides which are not paired are single stranded. These pair bonds define the secondary structure of an RNA strand. Its substructures are defined in Figure 1.

In the most widely used folding prediction model [8, 14], such pairs demarcate the RNA strand into nested loops – hairpin turns, helices, bulges, two-sided internal loops and multibranched internal loops. These loop structures are illustrated in Figure 1.

Zuker, et.al [10, 19] has developed algorithms which achieve substantial success in RNA secondary structure prediction. Recent studies [12] show, on average, that 73% of known base pairs on domains of fewer than 700 nucleotides are correctly predicted by these methods. Measurements in Turner's laboratory [9] of energy involved in various nucleotide interactions form the basis of the energy function employed. Dynamic programming is then used to optimize the secondary structure under the observed energy functions. The Zuker-Turner model forms the framework for the problems we are working on.

In the model employed here, all loops are either nested or disjoint; overlapping is not permitted. Such overlapping loops, called pseudoknots [11], occur in nature but considerably complicate the task of structure prediction.

Under this model [8, 17], the energy of an RNA strand is the sum of the energies of the loops of its secondary structure. This provides a sufficiently close approximation to permit significant predictive success, and it allows for a dynamic programming computation.

Energy is assigned to pairs within loops according to nearest neighbor rules. That is, an ordered quartet of the two nucleotides in a pair and the adjacent bases is assigned a value, determined by laboratory measurements. More elaborate energy rules involving nearby but not necessarily adjacent bases are being formulated.

Hairpin loops, bulges and internal loops are charged energy penalties based on their size (number of unpaired bases). Each multiloops is penalized a fixed amount a, plus penalty b for each internal pair, plus penalty c for each unpaired base. These and a variety of other special rules together comprise the energy function over which we will seek to optimize.

2.2 Minimum Energy (Most Stable) Sequence Design

We have developed an efficient dynamic programming algorithm for the *minimum energy RNA sequence problem*, in which we are given a target protein P, and seek to find the most stable (lowest energy) RNA sequence which codes for P. The algorithm also permits one to constrain the codon set which may be used at each site in the sequence. Hence, as a special case, it predicts the folding of a specified RNA sequence.

Asymptotically, our algorithm runs in $O(n^3)$ time, the same time as state-of-the-art programs [10] which only fold a given RNA sequence. It exploits the fact that each residue can be coded in only a constant number of ways. The solution to the minimum energy problem is not in general unique, and our algorithm identifies which codons/positions in an optimal RNA sequence are undetermined or partially determined.

To compute the minimum energy RNA sequence for a given protein, we extend the recurrences which form the basis of the Zuker folding algorithms for minimum energy secondary structure in two different ways:

1. For each position (index) in the RNA coding sequence, we consider all possible codons for the respective residue in the target protein, and

2. A set of companion recurrences specify the set $req(S(i \cdot j))$ of required nucleotides at specified positions in the RNA sequence for structure S closed by $i \cdot j$.

N_j denotes the jth nucleotide of sequence S and C_i denotes a codon which spans N_i.

The energy of any structure $S(i \cdot j)$ is a function of a set of associated nucleotides P_{ij}, plus the energy of the enclosed structures S'. For example, if $i \cdot j$ closes a hairpin loop, then $P = \{N_i, N_j, N_{i+1}, N_{j-1}\}$, and $S' = S(i+1, j-1)$. In general, the form of the recurrence for the required nucleotides $req(S(i \cdot j))$ is:

$$req(S(i \cdot j)) = P_{ij} \cup req(S') \quad (1)$$

2.2.1 The Optimal Structure Recurrence

We employ two main arrays in our dynamic programming algorithm:

- $V[i, j, C_i, C_j]$ contains the minimum energy of the structure closed by $i \cdot j$, given corresponding codons C_i and C_j.
- $WM[i, j, C_i, C_j]$ contains the contribution of subsequence i to j, given codons C_i and C_j, to a multiloop which contains i to j. Since the sizes of the third and fourth dimensions of these arrays (C_i and C_j) are constants, the space utilized is $O(n^2)$.

If N_i, N_j, N_{i+1}, N_{j-1} denote nucleotides in positions i, j, $i+1$ and $j-1$ respectively, $stack(N_i, N_j, N_{i+1}, N_{j-1})$ denotes the energy from stacking $N_i \cdot N_j$ directly on the adjacent pair $N_{i+1} \cdot N_{j-1}$ These stacking energies have been measured in the laboratory, and represented by a table[18]. We are also given various bonuses and penalties based on the size of the loop and other specially defined features. $hPenalty(n)$ is the penalty assigned to a hairpin loop of size n.

The energy of a hairpin loop closed by $i \cdot j$, for codons C_i and C_j, is given by Equation 2.

The energy of the structure closed by pair $i \cdot j$ in a helix, given codons C_i and C_j, is given by Equation 3.

The energy of an internal loop with external closing pair $i \cdot j$ and internal closing pair $k \cdot l$ is given by Equation 4.

2.2.2 The WM Recurrence

Let $WM(i, j, C_i, C_j)$ denote the energy contribution of sequence from i to j to a multiloop, for codons C_i and C_j. The recurrence is given in Equation 5.

Note that in all cases the choice of codons must respect the coding logic of RNA. That is, if x and y are two nucleotide indices belonging to the same triplet, then C_x and C_y must be the same as well.

2.2.3 VM recurrence

Let $VM(i, j, C_i, C_j)$ denote the energy of a multiloop closed by $i \cdot j$, for codons C_i and C_j. The recurrence for it is Equation 6.

That is, VM is the minimum over all possible ways of composing the multiloop contained within $i \cdot j$, exclusive of the closing pair, from the prefix $(i+1) \cdot (k-1)$ and the suffix $k \cdot (j-1)$. The constant a is the fixed cost assigned to each multiloop.

The complete recurrence also has terms for possible terminal mismatches (dangles) for i, for the first structure in $i \cdot (k-1)$, for the last structure in $k \cdot (j-1)$ and for j. These expressions are omitted for brevity.

2.2.4 Composite recurrence

Combining these recurrences, we arrive at the composite recurrence for the minimum energy, Equation 7.

2.2.5 Complexity analysis

Since the number of codons representing any amino acid is bounded by a constant (specifically six), the minimization over all codons in these recurrences does not affect the asymptotic time or space complexity of the resulting algorithms. Thus all of these recurrences can be computed in $O(n^3)$, except for internal loops.

If loops of unlimited size are allowed, the computation of internal loops requires evaluation of all values of i, k, l, and j such that $i < k < l < j$, leading to $O(n^4)$ time. However, in nature, loops of size greater than 30 are not encountered. Consequently, following [7], we restrict the size of loops to 30. This leads to an overall time of $O(n^3)$.

Our algorithm is implemented in about 12,000 lines of C++ code. Running on a 300 Mz Pentium II, it computes the minimum energy sequence of a 150 residue protein in about three minutes. Achieving this performance requires three types of algorithmic speedups (to be discussed in the full paper), which together improve the runtime by about two orders of magnitude.

2.3 Maximum Energy (Least Stable) Sequence Design

The problem of finding a maximum energy coding for a given protein p, i.e. the RNA coding for p which has the smallest amount of secondary structure, is also of interest. The results of Section 1 demonstrate that there is a statistical bias toward low-energy sequences, which is evidence of a process of natural selection. One can theorize that nature prefers to minimize RNA secondary structure since energy is required to break these bonds during transcription.

However, the algorithmics of finding the maximum energy coding are considerably different than those of finding a minimum energy coding. While dynamic programming gave an efficient way to find the minimum energy coding, the maximum energy coding problem is hard.

To minimize the complications of specific energy functions, we consider the *string with no k-repeat problem*, which captures the essence of maximum energy coding, specifically the case were we seek to design a coding which has no secondary structure at all:

Input: An alphabet Σ, an integer k, and an ordered collection of subsets $S = \{s_1, \ldots, s_n\}$, where $s_i \subset \Sigma$, for all $1 \leq i \leq n$.

Question: Does there exist a string T such that $|T| = n$, $T[i] \in s_i$ for all $1 \leq i \leq n$, and there does not exist a $1 \leq u < v \leq n$ such that $T[u+j] = T[v+j]$ for all $0 \leq j < k$?

For the special case when $s_i = \Sigma$ for all $1 \leq i \leq n$, and $n = |\Sigma|^k + k - 1$, the *string with no k-repeat problem* reduces to the problem of constructing an order k-de Bruijn sequence on Σ [2]. Such sequences can be efficiently constructed by finding an Eulerian cycle of an appropriate shift-register or de Bruijn graph [6].

We claim that the *string with no k-repeat problem* provides

$$eH(i,j,C_i,C_j) = \min_{\substack{C_{i+1} \\ C_{j-1}}} \Big\{ stack(N_i,N_j,N_{i+1},N_{j-1}) + hPenalty(j-i-1) \Big\} \tag{2}$$

$$eS(i,j,C_i,C_j) = \min_{\substack{C_{i+1} \\ C_{j-1}}} \Big\{ stack(N_i,N_j,N_{i+1},N_{j-1}) + V(i,j,C_{i+1},C_{j-1}) \Big\} \tag{3}$$

$$eI(i,j,C_i,C_j) = \min_{\substack{i<k<l<j \\ C_{i+1},C_{j-1} \\ C_k,C_{k-1} \\ C_l,C_{l+1}}} \Big\{ stack(N_i,N_j,N_{i+1},N_{j-1}) + stack(N_l,N_{l+1},N_{k-1},N_k) + V(k,l,C_k,C_l) \Big\} \tag{4}$$

$$WM(i,j,C_i,C_j) = \min \Bigg(V(i,j,C_i,C_j) + b, \min_{C_{j-1}} \{WM(i,j-1,C_i,C_{j-1})\} + c,$$

$$\min_{C_{i+1}} \{WM(i+1,j,C_{i+1},C_j)\} + c, \min_{i<k\le j} \Big\{ \min_{C_{k-1}} WM\{(i,k-1,C_i,C_{k-1})\} + \min_{C_k} \{WM(k,j,C_k,C_j)\} \Big\} \Bigg) \tag{5}$$

$$VM(i,j,C_i,C_j) = \min_{\substack{i+1<k\le j-1 \\ C_k,C_{k-1}}} \Big\{ WM(i+1,k-1,C_{i+1},C_{k-1}) + WM(k,j-1,C_k,C_{j-1}) + a \Big\} \tag{6}$$

$$V(i,j,C_i,C_j) = \min \Big\{ eH(i,j,C_i,C_j), eS(i,j,C_i,C_j), eI(i,j,C_i,C_j), VM(i,j,C_i,C_j) \Big\} \tag{7}$$

a reasonable model for maximum energy coding because (1) both problems seek to build a string where each character position is restricted to certain subset of possible characters, and (2) both problems seek a string which avoids a pair of complementary patterns of sufficient length to bind. Although the binding patterns to avoid in minimizing RNA secondary structures are in fact reversed and complemented instead of repeated, this is not an algorithmically significant complication. Indeed, simply reversing and complementing the header in the reduction (given in the full paper) suffices to prove the hardness of this alternate formulation:

THEOREM 1. *The string with no k-repeat problem is NP-complete, even if* $k = 3$.

2.3.1 A-Substitution (ASUB) method

Because CG bonds are stronger than AU or GU bonds, it is reasonable to anticipate that, within a given coding space, the sequences with the lowest CG content will tend to congregate at the highest energy (least stable) region. Consequently, a simple substitution of As in a sequence whereever possible has been suggested by Zuker as a means of constructing high energy sequences (personal communication).

This idea may be generalized by introducing a ranking among nucleotides which leads to a ranking among synonymous codons. First, the nucleotides are ranked in order of declining preference. A codon with the highest number of the highest ranked nucleotide is preferred over all its synonyms. If two synonymous codons have equal numbers of the preferred nucleotide, the tie is broken by the number of next-most-preferred nucleotide, and so on. We seek the nucleotide ranking which yields the preferred codon which, on average, leads to formation of the weakest bonds.

It is a curious consequence of the structure of the genetic code that prioritizing the nucleotides is sufficient to unambiguously order each set of synonymous codons. In other words, no codon for any of the 20 amino acids is a permutation of any of its synonyms. We don't have a basis for suggesting whether these are simply accidental byproducts of the factors which have been suggested as featuring in the structuring of the genetic code [16].

Figure 2 (left) shows the energy density of six sequences resulting from all possible orderings of nucleotide preference. No one ordering gives the best result for all sequences, but we employ the $AUCG$ ordering in experiments below.

2.3.2 Local structure avoidance (LSA) method

We have also developed and implemented a second approach, the *local structure avoidance* method. It seeks to construct a sequence S_n with a high free energy by starting with a short sequence which has minimum structure, identified through an exhaustive search. It then incrementally extends the sequence by a procedure that minimizes local secondary structure within a fixed-length suffix of S.

In other words, we construct S by sliding a window W along its length, within which we seek to minimize secondary structure. W consists of a fixed prefix F and a variable suffix V. Using a backtracking search, we examine all possible variable suffixes to find the optimal V which minimizes secondary structure within W. The first nucleotide of the optimal sequence of V is then appended to F, sliding W forward one position.

In this procedure, computations of secondary structure are performed only on a fixed-length window W, and there are only $O(n)$ possible positions for W. Hence, for constant $|F|$ and $|V|$, the time complexity of LSA is linear in n.

Due to the exponential dependence of time on $|V|$, variable window sizes of more than about 20 are impractical. Figure 2 shows the time/energy tradeoffs on $|V|$ for $|F| = 12$.

2.4 Experimental Results

2.4.1 Minimum energy experimental results

We applied our minimum energy algorithm to 200 short microbial RNA sequences, and compared the energy of the

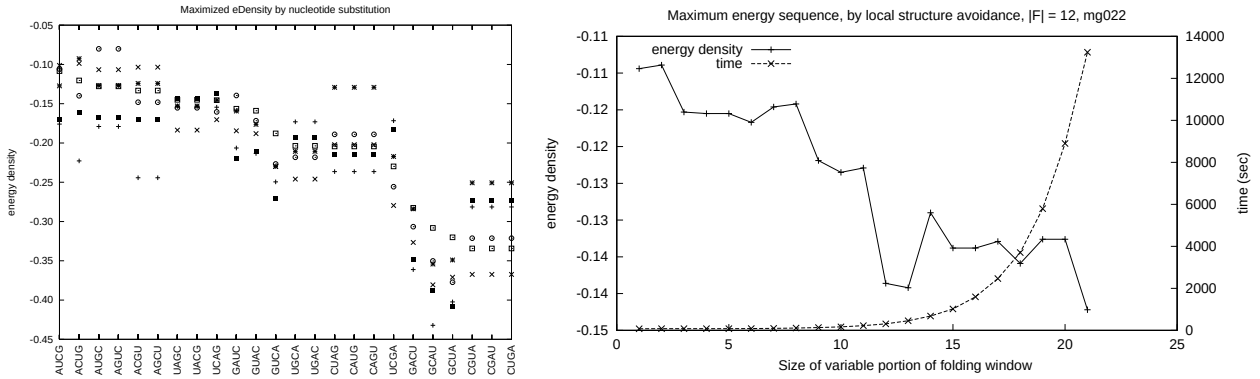

Figure 2: Energy density of six sequences for all possible orderings of nucleotide preference (left), with time/energy tradeoffs for the LSA algorithm. (right)

minimized sequences to the energy of the corresponding wild-type sequences.

The average energy density (energy/nucleotide) of wild-type sequences is -0.241/kcal-mol; for the minimum energy sequences it is -0.626. A minimized sequence has, on average, a minimum free energy 2.657 times as great as the naturally-occurring sequence.

It it noticeable that the greater stability of these sequences is achieved only in part by a increase in CG content. Even sequences with low CG content have low energies, due to the optimal arrangements of nucleotides, involving not only pairs but nearest neighbor quartets.

2.4.2 *Maximum energy experimental results*

Experiments indicate that the best results are usually achieved with a variable suffix size $V = 1$. The optimal fixed prefix size F is different for different sequences; below we have set $F = 50$. We randomly selected a set of 200 sequences from among the 27,000 examined in Section 1. High energy representatives of each of the coding spaces were then constructed by the ASUB and LSA methods.

Both methods prove effective in producing low-structure sequences. LSA produces better results in 73% of the sequences. The average energy density of ASUB sequences relative to native sequences is 0.561; the average energy density of LSA sequences relative to native sequences is 0.519.

2.4.3 *Energy profile of a coding space*

An energy profile of a synonymous coding space can be constructed by finding its bounds – the least and most energetic sequences in it – and the frequency distribution of randomly selected members. Such profiles for two sequences are shown in Figure 3.

In both these sequences, the randomly sampled instances are tightly clustered, roughly symmetrically, around a central value, and fall off sharply. All 1,000 of the sample sequences fall within a range of 0.1 kCal/mole-nucleotide. Both the high and low energy sequences fall far outside this range.

3. REFERENCES

[1] B.I. Dahiyat and S.L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278:82–87, 1997.

[2] N. G. de Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946.

[3] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, and M. Zehl. Design of multistable rna molecules. *RNA*, 7:254–265, 2000.

[4] D.R. Forsdyke and J.R. Mortimer. Chargaff's legacy. *Gene*, 261:127–137, 2000.

[5] R.F. Gesteland, T.R. Cech, and J.F.Atkins, editors. *The RNA World*. Cold Spring Harbor Laboratory Press, 1999.

[6] I.J. Good. Normal recurring decimals. *J. London Math. Soc.*, 21:167–172, 1946.

[7] I.L. Hofacker, W. Fontana, P.F. Stadler, and L.S. Bonhoeffer. Fast folding and comparison of rna secondary structures. *Monatshefte fur Chemie*, 125:167–188, 1994.

[8] Jr. Tinoco I., P.N. Borer, B. Dnegler, and M.D. Levine. Improved estimation of secondary structure in ribonucleic acids. *Nat. New Biol.*, 246:40–41, 1973.

[9] J. Jaeger, D.H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[10] R. Lyngso, M. Zucker, and C. Pedersen. Internal loops in rna secondary structure prediction. In *Proc. Third Int. Conf. Computational Molecular Biology (RECOMB '99)*, pages 260–267, 1999.

[11] R. B. Lyngso and C. N. S. Pederson. Pseudoknots in rna secondary structures. In *Proc. Fourth Int. Conf. Computational Molecular Biology (RECOMB '00)*, 2000.

[12] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of rna secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[13] D.L. Oxender, G. Zurawski, and C.Yanofsky. Attenuation in the *escherichia coli* tryptophan operon: Role of rna secondary structure involving the tryptophan codon region. *Proc. Natl. Acad. Sci. USA*, 76:5524–5528, 1979.

[14] C. Papanicolaou, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. *Nucleic Acids Research*, 12, 1984.

[15] J.F. Smith and E. Szathmary. *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford University Press, 1999.

[16] E.N. Trifonov. Consensus temporal order of amino acids and evolution of the triplet code. *Gene*, 261:139–151, 2000.

[17] D. Turner, N. Sugimoto, and S.M. Freier. RNA structure prediction. *Annual Rev. Biopys. Biophys. Chem.*, 17:167–192, 1988.

[18] M. Zucker. RNA folding server. http://www.ibc.wustl.edu/~zuker/rna/, 2000.

[19] M. Zuker, D.H. Mathews, and D.H. Turner. *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide*. 1999.

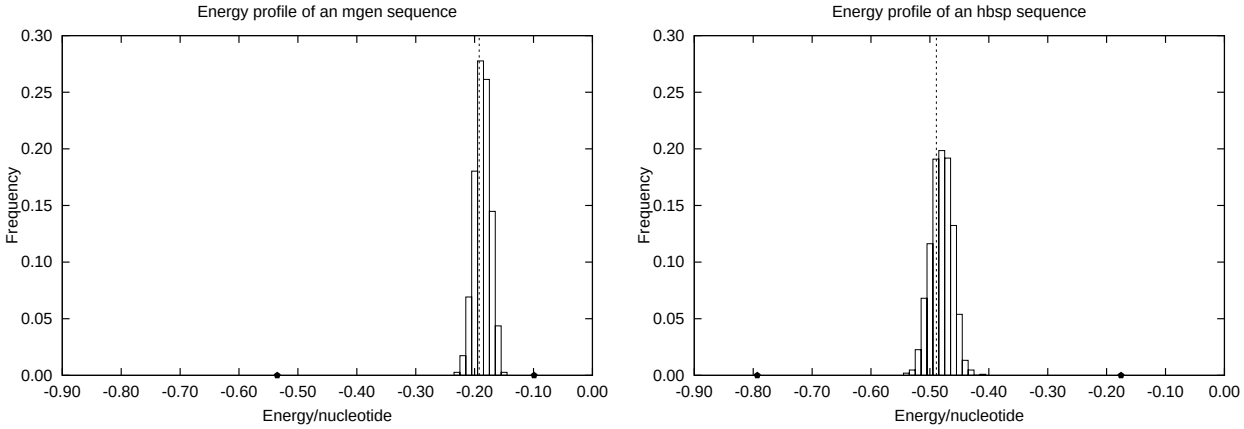

Figure 3: Energy profiles of the synonymous coding spaces of a 438-base sequence from *M. genitalium* (l) and a 435-base sequence from *Halobacterium sp. NRC-1* (r). Energy distribution is shown for 1,000 randomly composed synonyms for each sequence. Vertical dashed line is the energy of the wildtype sequence. Solid circles are energetically minimum and maximum sequences.