E. Simoudis, J. Han, and U. Fayyad (eds.)
Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,
AAAI Press 1996, Menlo Park, CA, pp. 20-25.

Knowledge Discovery in RNA Sequence Families of HIV Using Scalable Computers

Ivo L. Hofacker
Beckman Institute
University of Illinois
ivo@ks.uiuc.edu

Martijn A. Huynen
Los Alamos National Lab
and Santa Fe Institute
mah@santafe.edu

Peter F. Stadler
University of Vienna
and Santa Fe Institute
stadler@santafe.edu

Paul E. Stolorz
Jet Propulsion Lab
California Inst. of Technology
pauls@aig.jpl.nasa.gov

Abstract

Secondary structure prediction is a standard tool in the analysis of RNA sequences. The prediction of RNA secondary structures is inherently non-local. This makes the analysis of long sequences (more than 4000 nucleotides) infeasible on present-day workstations. An implementation of the secondary structure prediction algorithm for hypercube-type parallel computers allows to compute efficiently the structure of complete RNA virus genomes such as HIV-1 and other lentiviruses.

Introduction

One of the major problems facing computational molecular biology is the fact that sequence information about important macromolecules such as proteins and RNA molecules exists in far greater quantities than information about the three-dimensional structure of these biopolymers. The development and implementation of computational methods capable of predicting structure reliably on the basis of sequence information will provide huge benefits in terms of our understanding of the relationship between sequence and structure. They will also help greatly in tasks such as drug discovery and verification, as well as in the study of molecular evolution. These methods can then be applied to the vast quantities of sequence information at our disposal to discover important motifs and trends within various macromolecules, without having to laboriously and expensively measure the 3D structure of each and every molecule by hand.

It turns out that the full-blown task of three-dimensional structure prediction is much too difficult to be solved with current knowledge and methods. A simpler problem, however, the prediction of secondary structure, is tractable. Functional secondary structures are conserved in evolutionary phylogeny, and they represent a qualitatively important description of the molecules, as documented by their extensive use for the interpretation of molecular evolution data.

The most popular computational approach to the prediction of RNA secondary structure from sequence information is based upon dynamic programming. The main difficulty with this algorithm is the fact that its computational complexity grows as the cubic power of the RNA chain length, and that its memory requirements grow quadratically with chain length. This drawback has limited its use in the past to RNA molecules containing up to a few thousand nucleotides. Unfortunately, many molecules of great biological interest, such as HIV molecules, contain 10,000 or more nucleotides. The genome of HIV is dense with information for the coding of proteins and biologically significant RNA secondary structures. The latter play a role in both the entire genomic HIV-1 sequence and in the separate HIV-1 messenger RNAs which are basically fragments of the entire genome. The total length of HIV-1 (about 9200 bases) makes biochemical analysis of secondary structure of the HIV-1 full genome infeasible. For RNAs of this size computer prediction of the folded structure is the only approach that is available at present.

The goal of this paper is to demonstrate the unique ability of concurrent computers to enable data-mining of families of RNA sequences of the size and scope of HIV, by allowing automatic identification of important motifs. Sequence data-mining problems of this magnitude, requiring secondary structure prediction for a number of long RNA sequences, have never before been tackled because of their severe computational demands. We report the fastest secondary structure predictions ever achieved, and for the largest sequences that have ever been analyzed ($\sim$ 10000 nucleotides). Our results show that concurrency can be applied in this problem domain to allow novel sequence analysis and knowledge discovery on a large scale. More importantly, massively parallel machines enable not just the prediction of secondary structure for long individual sequences, but also knowledge discovery in the form of comparisons between secondary structures for families of sequences. We have been able to exploit this power to allow the identification of prominent secondary structure motifs common to an entire set of HIV genomes. Our results point the way to a range of new sequence analysis possibilities in the future.

RNA Secondary Structures

RNA structure can be broken down conceptually into a secondary structure and a tertiary structure. The secondary structure is a pattern of complementary base pairings, see Figure 1. The tertiary structure is the three-dimensional configuration of the molecule. As opposed to the protein case, the secondary structure of RNA sequences is well defined; it provides the major set of distance constraints that guide the formation of tertiary structure, and covers the dominant energy contribution to the 3D structure.

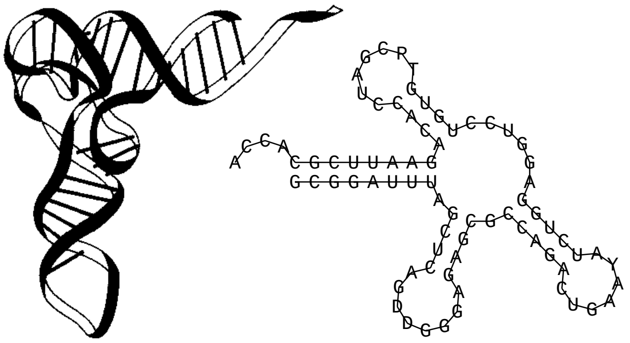

Figure 1: (l.h.s.) The spatial structure of the phenylalanine tRNA form yeast is one of the few known three dimensional RNA structures. (r.h.s.) The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.

A secondary structure of a sequence is a list of base pairs i, j with $i < j$ such that for any two base pairs i, j and k, l with $i \leq k$ holds (i) $i = k$ if and only if $j = l$ and (ii) $k < j$ implies $i < k < l < j$. The first condition says that each nucleotide can take part in not more that one base pair, the second condition forbids knots and pseudoknots[1]. Knots and pseudoknots are excluded by the great majority of folding algorithms which are based upon the dynamic programming concept.

A base pair k, l is *interior* to the base pair i, j, if $i < k < l < j$. It is *immediately interior* if there is no base pair p, q such that $i < p < k < l < q < j$. For each base pair i, j the corresponding *loop* is defined as consisting of i, j itself, the base pairs immediately interior to i, j and all unpaired regions connecting these base pairs. The energy of the secondary structure is assumed to be the sum of the energy contributions of all loops. (Note that two stacked base pairs constitute a loop of size 4; the smallest hairpin loop has three unpaired bases, i.e., size 5 including the base pair.)

Experimental energy parameters are available for the contribution of an individual loop as functions of its size and type (stacked pair, interior loop, bulge, multi-stem loop), of the type of its delimiting base pairs, and partly of the sequence of the unpaired strains (Turner, Sugimoto, & Freier 1988). Inaccuracies in the measured energy parameters, the uncertainties in parameter settings that have been inferred from the few known structures, and most importantly, effects that are not even part of the secondary structure model, limit the predictive power of the algorithms. Nevertheless, local structures can be computed in quite some detail, and a majority of the base pairs is predicted correctly.

A convenient way of displaying the size and distribution of secondary structure elements is the *mountain representation* introduced in (Hogeweg & Hesper 1984). In this representation a base paired to a base downstream is drawn as a step up, a base paired to a base upstream corresponds to step down, and an unpaired base is shown as horizontal line segment, see Figures 3 and 4. The resulting graph looks like a mountain-range where:

Peaks correspond to hairpins. The symmetric slopes represent the stack enclosing the unpaired bases in the hairpin loop, which appear as a plateau.

Plateaus represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively.

Valleys indicate the unpaired regions between the branches of a multi-stem loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The height of the mountain at sequence position k is simply the number of base pairs that enclose position k; i.e., the number of all base pairs (i, j) for which $i < k$ and $j > k$. The mountain representation allows for straightforward comparison of secondary structures and inspired a convenient algorithm for alignment of secondary structures (Konings & Hogeweg 1989).

In this contribution we shall be interested in the secondary structure of the RNA genomes of a certain class of single-stranded RNA viruses. Lentiviruses such as HIV-1 and HIV-2 are highly complex retroviruses. Their genomes are dense with information for the coding of proteins and biologically significant RNA secondary structures. The latter play a role in both the entire genomic HIV-1 sequence and in the separate HIV-1 messenger RNAs which are basically fragments and combinations of fragments of the entire genome. By predicting the minimum free energy secondary structure of the full length HIV-1 and other known lentiviruses sequences (HIV-2, SIV, CAEV, visna, BIV and EIAV) and their various splicing products, and by comparison of the predicted structures, a first step can be made towards the unravelling of all important secondary structures in lentiviruses.

Elucidation of all the significant secondary structures is necessary for the understanding of the molec-

[1] A pseudoknot is a configuration in which a nucleotide that is inside a loop base pairs with a nucleotide outside that loop.

ular biology of the virus. So far a number of significant secondary structures have been determined that play a role during the various stages of the viral life cycle (see section 4). We expect a high number of undiscovered biologically functional secondary structures to be still present within the various transcripts. A systematic analysis of the 5' end of the HIV genome showed an abundance of functional secondary structures (Baudin *et al.* 1993). Secondary structures further downstream could well be involved in the splicing, regulation of translation of the various mRNAs, or regulation of the stability of the full length sequence and its various splicing products.

Parallel Decomposition Issues and Related Work

Dynamic programming, when applied to RNA folding, requires CPU time that scales roughly as the cubic power of the sequence length, and memory that scales quadratically with sequence length. Even so, sequences such as HIV that are approximately 10000 nucleotides in length still require only on the order of 35 minutes to fold on 256 nodes of the Intel `Delta` supercomputer. The same calculation would require on the order of 60 hours on a high-end workstation.

Memory requirements are a even more severe problem for RNA molecules the size of HIV. The simplest RNA folding calculation, which computes just the single minimum free energy structure, requires of the order of 1 Gigabyte of memory for a sequence of the length of HIV-1. More sophisticated algorithms that compute averages over a larger number of structures near the minimum free energy typically require upwards of 2 Gigabytes. Distributed massively parallel architectures can easily satisfy these memory requirements. These resources are the primary reason that scalable architectures are necessary for performing RNA folding computations on large macromolecules.

As a consequence of the additivity of the energy contributions, the minimum free energy of an RNA sequence can be calculated recursively by dynamic programming (Waterman 1978; Zuker & Sankoff 1984). This method is at the heart of our approach. The basic logic of the folding algorithm is derived from sequence alignment: In fact, folding of RNA can be regarded as a form of alignment of the sequence to itself. The implementation of sequence alignment algorithms on massively parallel architectures in discussed in detail in (Jones 1992).

The algorithm proceeds by calculating energies for every subsequence and can be parallelized very easily: all subsequences with a common length are independent of each other and can therefore be computed concurrently, as in the case of sequence alignment. The major computational difficulty in the case of folding, distinguishing it from standard sequence alignment, is the fact that each entry requires the *explicit* knowledge of a large number data belonging to smaller subsequences.

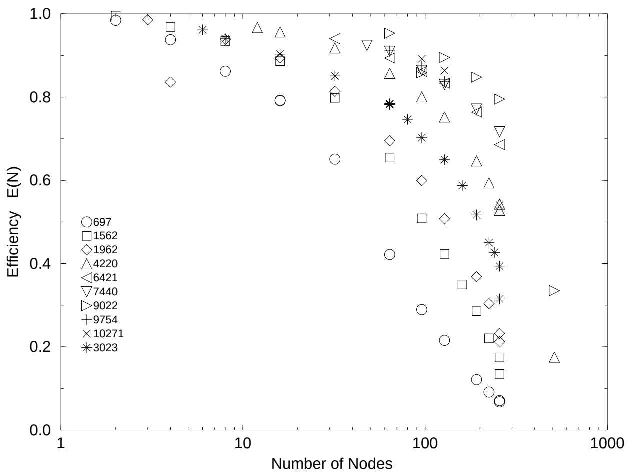

Figure 2: Efficiency of the parallelization versus the number N of nodes.

The *efficiency* of the parallelization is measured by $\mathcal{E}(N) := T^*/(Nt)$, where T^* is the (hypothetical) single node execution time, N is as usual the number of nodes used for the calculation and t is real time used for the folding (including the backtracking step). The data in Figure 2 show that we achieve efficiencies of more than 90% when the smallest possible number of nodes is used for the computation.

Knowledge Discovery within the Secondary Structure of Lentiviruses

Retroviruses are viruses that in their life cycle alternate between a single stranded RNA stage and a double stranded DNA stage. The lentiviruses are a subclass of the retroviruses, characterized by long incubation times and a similar genomic organization. The genome of a lentivirus consists of a single RNA molecule with about 7000 to 10000 nucleotides. Almost all of this genome is used for coding for various viral proteins and RNA secondary structures. Below we highlight the role of some of the known functional secondary structures and describe the progress we have made towards the discovery of new scientific knowledge by mining information from a number of RNA sequences on massively parallel computers.

The minimum free energy structure was predicted for the 22 available full-length HIV-1 sequences: HIVANT70, HIVBCSG3C, HIVCAM1, HIVD31, HIVELI, HIVHAN, HIVHXB2R, HIVJRCSF, HIVLAI, HIVMAL, HIVMN, HIVMVP5180, HIVNDK, HIVNL43, HIVNY5CG, HIVOYI, HIVRF, HIVSF2, HIVU455, HIVYU10, HIVYU2, HIVZ2Z6) and for 9 sequences of related lentiviruses: EIAV (equine infectious anemia virus), CAEV (caprine arthritis encephalitis virus), BIV106 (bovine immunodeficiency virus), VLVCG (visna virus), three simian immunodeficiency viruses SIVMM239, SIVMM251 (from macaque) and SIVSYK (from Syke's monkey), and two HIV-2 sequences HIV2BEN and HIV2ST.

The majority of the secondary structures exhibit two distinct domains: whereas the 5' half consists of a large number of fairly small components, the 3' part is a single component (except for a region of about one hundred nucleotides). The boundary between the two structural domains coincides roughly with the end of the *pol* gene.

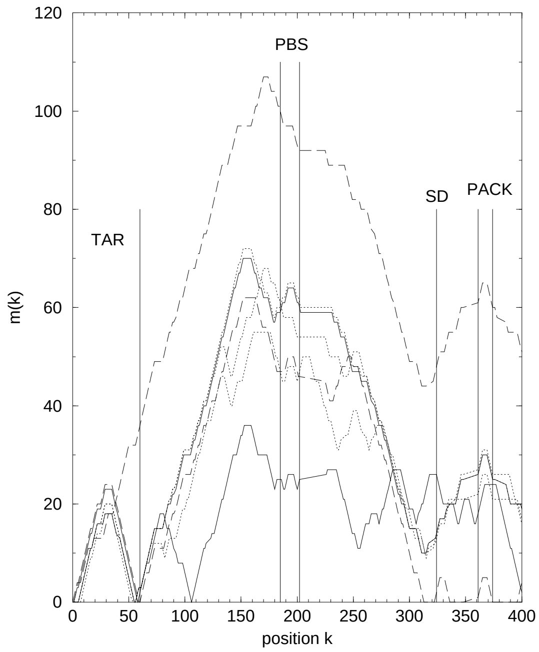

Figure 3: Mountain representation of the secondary structure of the 5' end of seven HIV-1 sequences (HIVLAI, HIVOYI, HIVBCSG3C: dotted line, HIVELI, HIVDNK: dashed line, HIVANT70: solid line, HIVMAL: long-dashed line) The secondary structures were aligned at the sequence level. Although the structures do show considerable variation, some features are conserved: (i) The TAR hairpin structure is present in six out of seven sequences. (ii) The center of the Primer Binding Site (PBS) is always single stranded (sometimes as a hairpin loop, sometimes as an internal loop), thus exposing this part of the sequence for base pairing with the tRNA primer. (iii) The center of the packaging signal (PACK) is always present as a hairpin.

At the 5' end of the viral HIV-1 RNA molecule resides the *trans*-Activating Responsive (TAR) element (Berkhout 1992); binding of the TAT protein to TAR is necessary for high levels of transcription. On the basis of biochemical analysis (Baudin *et al.* 1993) and computer prediction of the 5' end of the genome it is known that the TAR region in HIV-1 forms a single, isolated stem loop structure of about 60 nucleotides with about 20 base pairs interrupted by two bulges. This structure is indeed predicted in the minimum free energy structures of six of the seven sequences in Figure 3. Besides in HIV-1, a functional TAR structure has also been observed in HIV-2 and various SIV types while all other known lentiviruses have a *tat* gene. Although the secondary structure of TAR is strongly conserved within HIV-1, it varies considerably between the various human (HIV-1 and HIV-2) and simian (SIV) lentiviruses, as is also reflected in the minimum free energy foldings. Our analysis shows that CAEV, visna, EIAV and BIV all have a short hairpin structure at their 5' end.

The packaging signal is essential for the packaging of full length genomes into new virion particles. All analyses of its secondary structures are consistent with a short (5 base pairs) hairpin structure that carries a **GGAG** loop (Harrison & Lever 1992). Indeed, this feature is shared by all the sequences in Figure 3. However, the predictions in the literature for the more global secondary structure of this region of the RNA (beyond the 6 base pair hairpin) vary considerably. A large variation in the predicted secondary structures is also present in the minimum free energy structures of the various HIV-1 sequences.

The Primer Binding Site (PBS) at the 5' of the viral genome (Baudin *et al.* 1993) is necessary for the initiation of reverse transcription of the HIV genomic RNA into DNA. It is a sequence of 18 nucleotides that is complementary to the nucleotides at the 3' end of the tRNA with which it base pairs. The tRNA serves as a primer to initiate the reverse transcription of the viral RNA. In absence of the primer, part of the Primer Binding Site is paired to bases outside the PBS. The binding of the primer could therefore lead to a rearrangement of the secondary structure of the 5' end of the molecule. Indeed, such rearrangements were observed up to 69 nucleotides upstream and 72 nucleotides downstream of the PBS after the binding of the primer (Isel *et al.* 1995). Computer prediction of the secondary structure of RNA can play a role in guiding these types of experiments and explaining their results.

Within the *env* gene of lentiviruses resides the Rev response element (RRE). The consensus secondary structure of the RRE in HIV-1 is a multi-stem loop structure consisting of five hairpins supported by a large stem structure (Konings 1992). The interaction of RRE with the Rev protein reduces splicing and increases the transport of unspliced and single-spliced transcripts to the cytoplasm, which is necessary for the formation of new virion particles (Malim *et al.* 1989). Figure 4 shows an alignment of the RRE region of 17 out of the 22 HIV-1 sequences based entirely on the predicted secondary structures and without gaps. Most of the secondary structures show the five-fingered hairpin motif. An alternative structure is present in which hairpin III is relatively large and a few of the other hairpins have disappeared from the

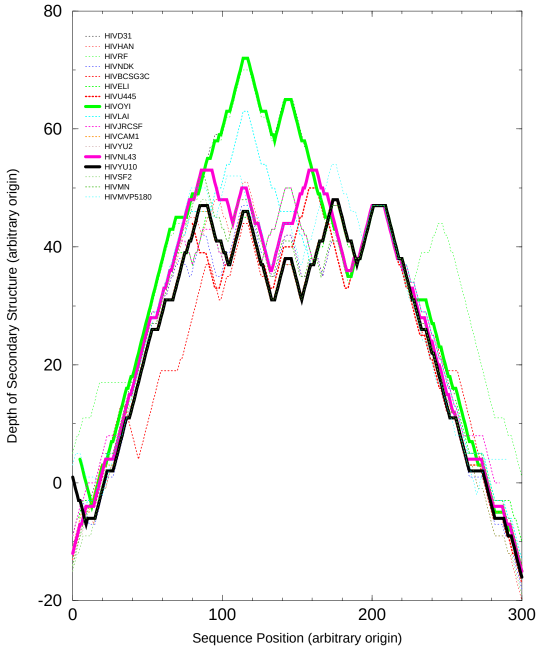

Figure 4: Alignment of the RRE regions of 17 sequences based solely on the minimum free energy secondary structure. The mountain representation reveals the five-fingered motif, the Roman numerals correspond to the numbering of the hairpins in (Dayton, Powell, & Dayton 1989). 5 out of 22 sequences showed a different pattern here. We find three different folding patterns each highlighted by one example. The first one (thick black line) corresponds to the consensus five-fingered motif that is presented in (Konings 1992). The second one (light gray) is present among other in HIVLAI. The third (dark gray) corresponds to the structure proposed (Mann *et al.* 1994).

minimum free energy structure. A comprehensive analysis of the base pairing probabilities in the RRE shows that the hairpins II, IV, and V, as well as the basis of hairpin III are meta-stable in the sense that they allow for different structures with nearly equal probabilities (Huynen *et al.* 1996). This structural versatility within a single sequence is here reflected in the variation in the minimum free energy secondary structure of closely related sequences. Although there is structural versatility in the hairpins, the stem structure on top of which the hairpins are located is generally present in the minimum free energy folding. The comparisons of the prediction obtained for different, evolutionarily related RNAs can be used to identify local misfoldings in the same way as a comparative analysis can be used to infer the structure from the phylogeny.

Discussion

Our implementation of dynamic programming RNA folding algorithms on up to 512 nodes of the `Delta` supercomputer demonstrates that massively parallel distributed memory computer architectures are well-suited to the problem of folding the largest RNA sequences available. With sequences comprising several thousand nucleotides, efficiencies above 80% are obtained on partitions of the machine containing about 100 nodes. As the partition size increases beyond 100 nodes the efficiency deteriorates to 20-40%, even for the larger sequences studied. Not surprisingly, the optimal partition sizes are those for which the total available memory on each node is utilized. These results are extremely encouraging. Apart from the insight that can be shed on the HIV virus itself, they indicate that even longer virus genomes containing up to 30000 nucleotides can be folded on the existing `Delta` architecture, with future scalable machines promising to extend this range even further. One long molecule of special interest is the Ebola virus, which contains roughly 20000 nucleotides.

The minimum free energy structure of a set of HIV-1, HIV-2, and related lentiviruses has been determined. The results show the presence of known secondary structures such as TAR, RRE, and the packaging signal that have been predicted on the basis of biochemical analysis, phylogenetic analysis, and the folding of small fragments of the sequence. In HIV-1 we observe a striking difference between the secondary structures of the first half and the second half of the molecule. Whereas the first 4000 nucleotides form a large number of independent components, the second 5000 nucleotides form a single huge component, on top of which the RRE is located. In general, although some relatively local patterns and the overall pattern with short range interactions in the 5' end and long range interactions at the 3' end appear conserved, there is extensive variation in the secondary structure between the various HIV-1 sequences.

The folding algorithm discussed in this paper predicts only the thermodynamically most stable secondary structure. Under physiological conditions, i.e., at or above room temperature, however, RNA molecules do not take on only the most stable structure, they seem to rapidly change their conformation between structures with similar free energies. A realistic investigation of RNA structures has to account for this fact which is of utmost biological importance. The simplest way to do this is to compute not only the optimal structure but all structures within a certain range of free energies (Waterman & Byers 1985). A more recent algorithm (McCaskill 1990) is capable of computing physically-relevant averages over all possible structures, by calculating an object known as the partition function. From it, the full matrix $P = \{p_{ij}\}$ of base pairing probabilities, which carries the biologically most relevant information about the RNA structure, can be obtained. In fact, a sequential implementation (Hofacker *et al.* 1993) has been ported recently to a `CRAY-Y-MP` and has been successfully applied to analyzing the base pair probabilities of a complete HIV-1 genome (Huynen *et al.* 1996). A comparative analysis of base-pair probabilities of RNA viruses requires an implementation of the partition function algorithm on massively parallel computers. Work in this direction is in progress.

Acknowledgements

This research was performed in part using the facilities of the Center for Advanced Computing Research at Caltech. Access to this facility was provided by the California Institute of Technology. Partial financial support by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. P 9942 PHY, is gratefully acknowledged. PES acknowledges the Aspen Center for Physics at which part of this research was performed. PES and PFS also thank P. Messina and the hospitality of CACR, where much of the work was conducted. The work was supported by the Los Alamos LDRD program and by the Santa Fe Institute Theoretical Immunology Program through a grant from the Joseph P. and Jeanne M. Sullivan Foundation.

References

Baudin, F.; Marquet, R.; Isel, C.; Darlix, J. L.; Ehresmann, B.; and Ehresmann, C. 1993. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.* 229:382–397.

Berkhout, B. 1992. Structural features in TAR RNA of human and simian immunodeficiency viruses: a phylogenetic analysis. *Nucl. Acids Res.* 20:27–31.

Dayton, E.; Powell, D. M.; and Dayton, A. I. 1989. Functional analysis of CAR, the target sequence for the Rev protein of HIV-1. *Science* 246:1625–1629.

Harrison, G. P., and Lever, A. M. 1992. The human immunodeficiency virus type 1 packaging signal and major splice donor region have a conserved stable secondary structure. *J. Virology* 66:4144–4153.

Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; and Schuster, P. 1993. `Vienna RNA Package`(public domain software). available from
`ftp://ftp.itc.univie.ac.at/pub/RNA/`.

Hogeweg, P., and Hesper, B. 1984. Energy directed folding of RNA sequences. *Nucl. acids res.* 12:67–74.

Huynen, M. A.; Perelson, A. S.; Viera, W. A.; and Stadler, P. F. 1996. Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.* 3(2):253–274.

Isel, C.; Ehresmann, C.; Keith, G.; Ehresmann, B.; and Marquet, R. 1995. Initiation of reverse transcription of HIV-1: secondary structure of the HIV-1 RNA/tRNA(3Lys) (template/primer). *J. Mol. Biol.* 247:236–250.

Jones, R. 1992. Protein sequence and structure alignments on massively parallel computers. *Int. J. Supercomp. Appl.* 6:138–146.

Konings, D. A. M., and Hogeweg, P. 1989. Pattern analysis of RNA secondary structure similarity and consensus of minimal-energy folding. *J. Mol. Biol.* 207:597–614.

Konings, D. A. M. 1992. Coexistence of multiple codes in messenger RNA molecules. *Comp. & Chem.* 16:153–163.

Malim, M. H.; Hauber, J.; Le, S. Y.; Maizel, J. V.; and Cullen, B. R. 1989. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* 338:254–257.

Mann, D. A.; Mikaelian, I.; Zemmel, R. W.; Green, S. M.; Lowe, A. D.; Kimura, T.; Singh, M.; Butler, P. J.; Gait, M. J.; and Karn, J. 1994. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J.Mol. Biol.* 241:193–207.

McCaskill, J. S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–1119.

Turner, D. H.; Sugimoto, N.; and Freier, S. 1988. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry* 17:167–192.

Waterman, M. S., and Byers, T. H. 1985. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.* 77:179–188.

Waterman, M. S. 1978. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.* 1:167 – 212.

Zuker, M., and Sankoff, D. 1984. RNA secondary structures and their prediction. *Bull. Math. Biol.* 46(4):591–621.