Base Pairing Probabilities in a Complete HIV-1 RNA

Martijn A. Huynen[a,b,*], Alan S. Perelson[a],
Wayne A. Vieira[c], and Peter F. Stadler[d,e]

[a]Theoretical Division, Los Alamos Natl. Lab., Los Alamos, NM 87545
[b]Center for Non-Linear Studies, Los Alamos Natl. Lab., Los Alamos, NM 87545
[c]Parallel Applications Technology Program, Cray Research Inc., Los Alamos, NM
[d]Santa Fe Institute, Santa Fe, NM 87501
[e]Inst. f. Theoretische Chemie, Univ. Wien, Austria

[*]Mailing Address: Martijn A. Huynen
T-10, Mail Stop K-710, Los Alamos Natl. Lab., Los Alamos, NM 87545, U.S.A.
Phone:(505) 667 9455 Fax: (505) 665-3493
E-Mail: mah@t10.lanl.gov

Keywords

HIV — RNA Secondary Structure — Partition Function — RRE

Abstract

We have calculated the base pair probability distribution for the secondary structure of a full length HIV-1 genome using the partition function approach introduced by McCaskill (McCaskill, 1990). By analyzing the full distribution of base pair probabilities instead of a restricted number of secondary structures, we gain more complete and reliable information about the secondary structure of HIV-1. We introduce methods that condense the information in the probability distribution to one value per nucleotide in the sequence. Using these methods we represent the secondary structure as a weighted average of the base pair probabilities, and we can identify interesting secondary structures that have relatively well defined base pairing. The results show high probabilities for the known secondary structures at the 5' end of the molecule that have been predicted on the basis of biochemical data. The Rev response element (RRE) appears as a distinct element in the secondary structure. It has a meta-stable domain at the high affinity site for the binding of Rev. The overall structure decomposes into fairly small independent structures in the first 4000 bases of the molecule. The remaining 5000 bases (excluding the terminal repeat) form a single, large structure, on top of which the RRE is located.

1. Introduction

HIV-1 and HIV-2 are highly complex retroviruses. Their genomes are densely packed with information for the coding of proteins and biologically significant RNA higher order structures. Knowledge of all the functional secondary structures is a prerequisite for the understanding of the molecular biology of HIV. This article describes the probability distribution of secondary structures in a full length HIV-1 genome. Previously a number secondary structures have been determined that play a role during various stages of the virus life cycle:

In reverse transcription: Reverse transcription is primed by a tRNA that is bound to the primer binding site (PBS) at the 5' of the viral genome. The PBS is located in a conserved secondary structure (Berkhout & Schoneveld, 1993; Baudin *et al.*, 1993). RNA secondary structure might also be involved in regulating the number of mutations during reverse transcription. The introduction of a stem-loop structure in HIV leads to a relatively high local mutation rate in reverse transcription (Pathak & Temin, 1992).

In transcription and regulation of splicing: At the 5' end of the molecule resides the trans-activating responsive (TAR) element, which interacts with the regulatory Tat protein. The binding of the Tat protein to TAR is responsible for the activation of the pro-virus transcription (Feng & Holland, 1988; Klaver & Berkhout, 1994). The sequence in the main hair-pin loop of TAR is complementary to a strongly conserved sequence further downstream in the *gag* gene. These sequences might interact with each other, forming a so-called kissing motif (Chang & Tinoco Jr., 1994). The Rev response element (RRE), an RNA structure that is located within the *env* gene plays a regulatory role. The binding of the Rev protein to RRE regulates the transport of unspliced HIV sequences to the cytoplasm (Malim *et al.*, 1989; Malim & Cullen, 1989; Mann *et al.*, 1994; Kimura & Ohyama, 1994). Splicing itself appears dependent on the RNA secondary structure. The major-splice donor (SD) resides in a strongly conserved (Harrison & Lever, 1992) and biochemically stable (Baudin *et al.*, 1993) secondary structure.

In translation: RNA structure 3' from the *gag-pol* frameshift site contributes to ribosomal frameshift efficiency (Vickers & Ecker, 1992).

In packaging and dimerization: Late in infection full length HIV sequences need to be packed into virion particles. The packaging signal appears to form a stem-loop structure (Hayashi *et al.*, 1992). The sequence upstream of the packaging signal that is necessary for the dimerization of HIV-1 is proposed to form a separate hairpin (Laughrea & Jette, 1994). The sequence in the loop of this hairpin might interact with its counterpart in another HIV-1 sequence to start the dimerization.

The structures mentioned above were all determined using a combination of biochemical analysis, comparison of sequence data, and secondary structure prediction algorithms. A number of undiscovered functional secondary structures might still be present They could be involved in splicing, the regulation of translation of the various mRNAs, or regulation of the stability of the full length genome and its various splicing products. Prediction of RNA secondary structure using computer algorithms can be regarded as complementary to methods for the prediction of secondary structure that are based on sequence comparison or on biochemical analysis. It can point out regions of interest that can henceforth be analyzed in greater detail using biochemical and comparative analysis.

Almost all secondary structure predictions in the literature have only considered the minimum free energy structure and/or a fairly small sample of suboptimal structures, as provided, e.g., by Zuker's `mfold` package (Zuker & Sankoff, 1984; Zuker, 1989; Zuker, a). McCaskill's partition function approach (McCaskill, 1990), which allows for an exact computation of the complete matrix of all base pairing probabilities, provides more complete and reliable structural information. Biebricher has used McCaskill's algorithm to analyze the RNA secondary structure requirements necessary for replication by Qβ replicase (Biebricher, 1994), and implications for evolutionary dynamics of RNA secondary structure as a probability distribution have been studied (Bonhoeffer *et al.*, 1993). The large size of HIV genomes ($\approx$ 9200 nucleotides) implies that there is a huge number of low energy states. Hence the computation of the complete set of possible structures is too costly to be practical. For example, the frequency of the minimum energy structure in the ensemble of possible structures at thermodynamic equilibrium is in general smaller than 10^{-23} for RNAs of the size of a viral genome. Hence one needs more than 10^{23} of different structures to adequately describe the ensemble, and the direct generation and analysis of this amount of structural information is way beyond the capabilities of even the most modern computer systems. McCaskill's approach provides a computationally feasible alternative, which is in fact comparable to the requirements of the simple minimum free energy folding algorithm. The algorithm used in this paper is taken from the public domain `Vienna RNA Package` (Hofacker *et al.*, a). It has been used already successfully for computing the base pairing probability matrix of the complete genome of the coli-phage Qβ on a UNIX workstation (Hofacker *et al.*, 1994).

The length of the HIV-1 genome has heretofore been prohibitive for folding the entire genome. Secondary structure predictions were therefore based on folding subsequences in fairly small windows. This approach has two disadvantages: (i) by definition it cannot be used for long-range interactions that span more than the window size, and (ii) the results depend crucially on the exact location of the boundaries of the sequence window.

Subsequences can be folded independently of the rest of the structure only if they form a component by themselves, i.e., if there are no base pairs to the outside of the sequence window. The first 400 bases of the $\texttt{HIV-1}_{LAI}$ genome form two components which are well separated from the rest of the molecule. Consequently, the structure can be approximated quite well by independently folding this piece of sequence. In section 3 we give an example of the devastating effect of using sequence windows which do not coincide with components. Splitting the sequence in two halves gives a completely wrong structure. The only way of identifying the component boundaries is folding the sequence in its entirety.

We report here the computation of the base pairing probability matrix for the complete genome of an HIV-1 virus, Genbank entry $\texttt{HIV-1}_{LAI}$ (Wain-Hobson *et al.*, 1985). The sequence length is $n = 9229$. Both memory and CPU requirement for this task are formidable but not beyond the reach of a CRAY Y-MP supercomputer. The algorithm and its implementation are discussed in the appendix. In section 2 we describe the over-all structure of the viral genome along with useful condensed representations of the data. Section 3 compares our computational data to known secondary structure motifs of the HIV-1 virus. In particular, at the 5' end our method resolves the TAR region, the region containing the primer-binding site, the region containing the major splice donor and the packaging signal, and the hairpin 3' from the ribosomal frame-shift between *gag* and *pol*. The RRE is very well separated from other secondary structures. The RRE is particularly interesting since, whereas part of its structure is very well defined, the high affinity binding site for the binding of Rev within the RRE has a high structural versatility. We also show that the TAR secondary structure is repeated at the 3' end.

2. Representing the Structure

The complete base pairing probability matrix is a symmetric $n \times n$ matrix with the entry (i, j) being the probability that base i is paired with base j. For the HIV-1 genome analyzed here n=9229 and hence the matrix has approximately 8.5×10^7 entries. Below we discuss a number of ways of representing these data in a convenient way. For access to the full matrix see the appendix.

Dot Plots.

A dot plot is a two-dimensional graph in which the size of the dot at position i, j within the graph represents the probability of the i, j base pair. Thus, in principle, dot plots contain complete base pairing information. In practice, we suppress the dots corresponding to base pairs that occur with a probability of less than 10^{-5}. Figure 1 shows the 5'-end of $\texttt{HIV-1}_{LAI}$. The plot is divided into two triangles. The upper right triangle contains the base pairing probability matrix (p_{ij}); the size of the squares is proportional to the pairing probability. The lower-left triangle displays the minimum free energy structure (MFE) for comparison. Here only the base pairs that occur in the MFE are indicated. This way of presenting the data is similar to the dot-plot method used by Zuker (Jacobsen & Zuker, 1993). In Zuker's representation the dots in the upper right triangle correspond to base pairs that are part of secondary structures with free energies close to the free energy of the MFE. In our representation the sizes of the dots correspond to the actual probabilities of the base pairs.

The base pair probability matrix (p_{ij}) itself is not a very convenient data set for gaining an overview over the structure of a very large RNA molecule. We thus introduce some additional quantities that provide condensed information about the secondary structure. These measures depend only on the sequence position, instead of the base pair, and thus allow for a convenient graphical representation. They can also be used for comparing structures from different sequences by alignment-like procedure, see (Bonhoeffer *et al.*, 1993; Hofacker *et al.*, 1994) for the discussion of an alignment procedure based on a variant of the mountain representation.

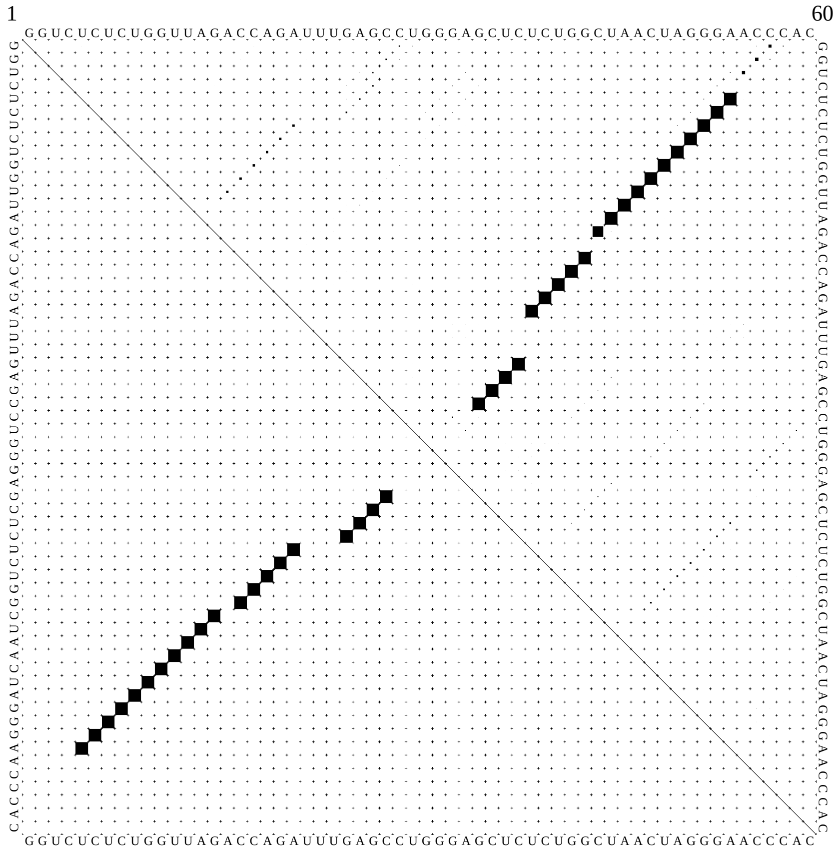

Figure 1: Dot plot of the TAR region at the 5' end of $\mathtt{HIV\text{-}1}_{LAI}$.
The upper right triangle contains the base pairing probability matrix (p_{ij}); the size of the squares is proportional to the pairing probability. The lower-left triangle displays the minimum free energy structure for comparison. Hairpin loops appear as diagonal patterns close to the separating line between the two triangles, with the distance from this line indicating the loop size. Internal loops and bulges appear as shifts and gaps in the diagonal patterns. Some biological information about the TAR region can be found in section 4.

Generalized Mountain Representation.

A convenient way of displaying the size and distribution of secondary structure elements is a generalized version of the *mountain representation* introduced by Hogeweg and Hesper (Hogeweg & Hesper, 1984). Let us define the number of base pairs that enclose position k as the number of base pairs (i, j) for which $i < k$ and $j > k$. In the original mountain representation a single secondary structure is represented in a two dimensional graph, with the x-coordinate being the position(k) in the sequence and the y-coordinate the number of base pairs that encloses k. The resulting graph looks like a mountain-range where:

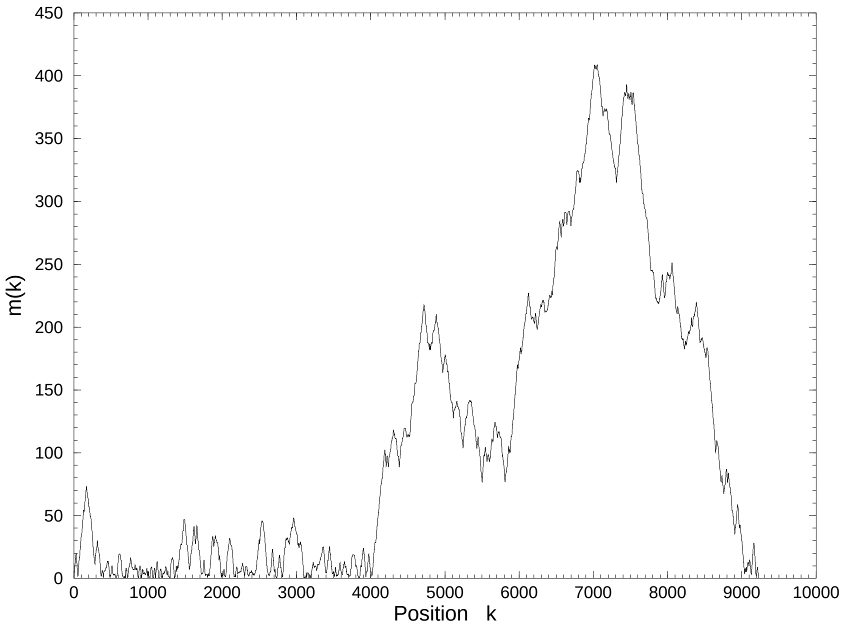

Figure 2: Generalized mountain representation m_k of $\texttt{HIV-1}_{LAI}$ obtained from the full base pair probability matrix. This representation gives a good impression on the average structure of the RNA molecule. The slope is proportional to the strength of the base pairing, the sign of the slope indicates whether the pairing is up-stream or down-stream. Unpaired regions appear as plateaus, peaks correspond to hairpin loops.

- *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-loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The mountain representation allows for straightforward comparison of secondary structures and inspired a convenient algorithm for alignment of secondary structures Hogeweg:94.

We construct a generalized version of the mountain representation as follows: Consider the numbers

$$m_k \stackrel{\text{def}}{=} \sum_{i<k} \sum_{j>k} p_{ij}$$

for all sequence positions k. By definition, m_k counts all base pairs which enclose k, weighted with their respective pairing probabilities.

In order to see that this measure is in fact a close relative of the mountain representation, assume for a moment that p_{ij} is the pairing matrix of a minimum free energy structure. Thus $p_{ij} = 0$ or 1. In this case m_k is the number of base pairs which contain k, i.e. it is constant for all positions in the same loop, increases by one at each paired position at the 5' side of a stack and decreases by one at each paired position at the 3' side of a stack. The generalized mountain representation gives a weighted average of the ensemble of secondary structures. The y-coordinate of base k corresponds to the number of base pairs that is expected to enclose k *on average*. In the original mountain representation the steepness of the slopes is either $+1$, -1 or 0 corresponding to whether a base is paired to a base downstream, to a base upstream, or whether it is not paired at all respectively. In the generalized mountain representation the slope can have any value between $+1$ and -1, where the steepness of the slope gives the probability of being paired to a base downstream minus the probability of being paired to a base upstream. Figure 2 is the generalized mountain representation of HIV-1_{LAI}. The first 4000 nucleotides of the sequence form a large number of independent secondary structures whereas the remainder of the sequence forms a single multi-stem structure.

Separation of Components.

Large RNA molecules in general decompose into components, that is, into contiguous sequence pieces that form base pairs only inside themselves and which are not interior to any other base pairs. Components play a special role in the calculation of the MFE structure: they can be folded independently from each other. The biophysical significance of components is that they form well separated substructures which are only loosely tied to each other. One therefore expects a fair amount of flexibility between components. Since m_k drops to zero at component boundaries this measure could be used to identify them. Unfortunately, when an ensemble of structures is considered, m_k does not have an immediate statistical interpretation in terms of the separation of components.

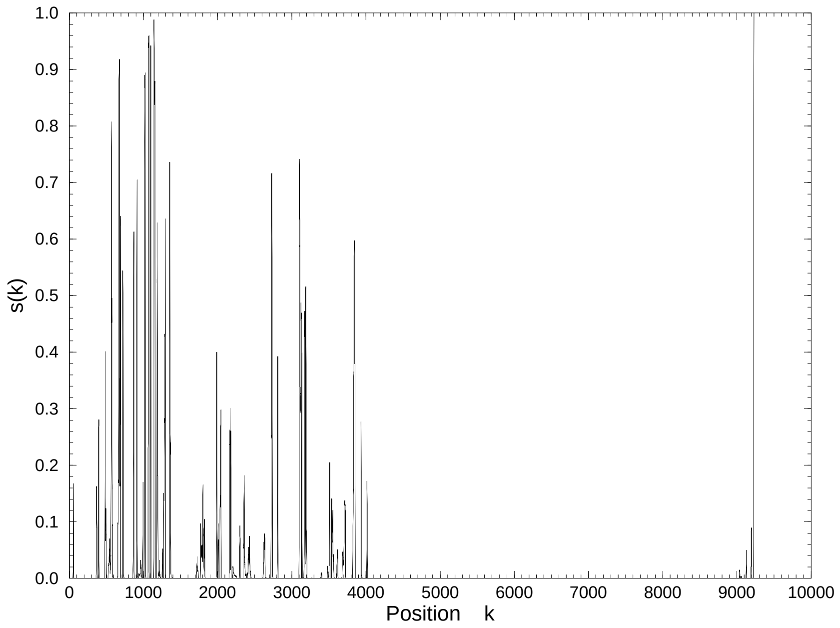

Figure 3: Components of $\texttt{HIV-1}_{LAI}$. The separation probabilities s_k show that the molecule consists of a large number of independent components in its 5' half. The portion beyond position 4000 seems to form a single huge component, see also Figure 2.

A more intuitive measure for identifying components is the probability

$$s_k \stackrel{\text{def}}{=} \prod_{i<k} \prod_{j>k} (1 - p_{ij})$$

that k is the boundary between components. Thus s_k is the probability that there is no base pair to which position k is interior. A plot of the *separation probabilities* s_k for $\texttt{HIV-1}_{LAI}$ is shown in Figure 3. The results are consistent with Figure 2: the first 4000 nucleotides form a large number of independent components, whereas the remaining 5000 form a single component. Component boundaries with $s_k > 0.7$ are the given in Table 1.

Table 1. Most Prominent Component Boundaries.

k	s_k	k	s_k	k	s_k	k	s_k
567	0.808	677	0.919	1023	0.895	1074	0.960
1144	0.970	1356	0.737	2729	0.716	3102	0.742

The separation probabilities are of course closely linked to the generalized mountain representation. This can be seen from the following consideration:

$$\ln s_k = \ln \left(\prod_{i<k} \prod_{j>k} (1 - p_{ij}) \right) = \sum_{i<k} \sum_{j>k} \ln(1 - p_{ij}) \approx - \sum_{i<k} \sum_{j>k} p_{ij} = -m_k ,$$

provided all $p_{ij} \ll 1$ for all $i < k$ and $j > k$. In other words, we have $s_k \approx \exp(-m_k)$ for small m_k, and $s_k = 1$ for $m_k = 0$.

Well-Definedness.

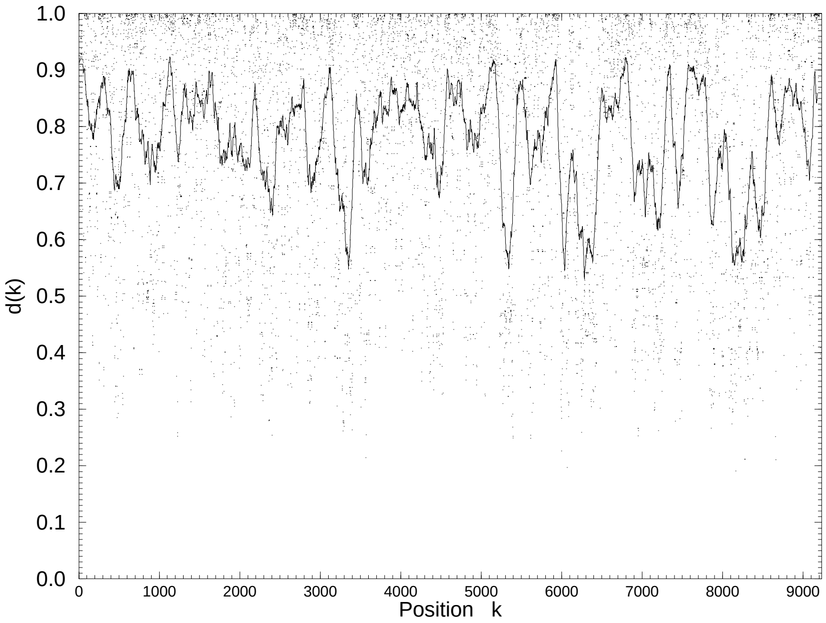

Figure 4: well-definedness of the `HIV-1`$_{LAI}$ structure. The solid line is a running average with a window size of 100.

Another interesting quantity is the well-definedness d_k of the structure in a certain region. Let

$$d_k \stackrel{\text{def}}{=\!=} \max \left\{ \max_i \{p_{ik}\},\, 1 - \sum_i p_{ik} \right\} ,$$

i.e., d_k is the probability of the most probable base pair involving k, or the probability that k is unpaired, whichever is larger. Thus d_k is high when a base either has a high probability of pairing with one specific other base *or* it has a high probability of not interacting at all.

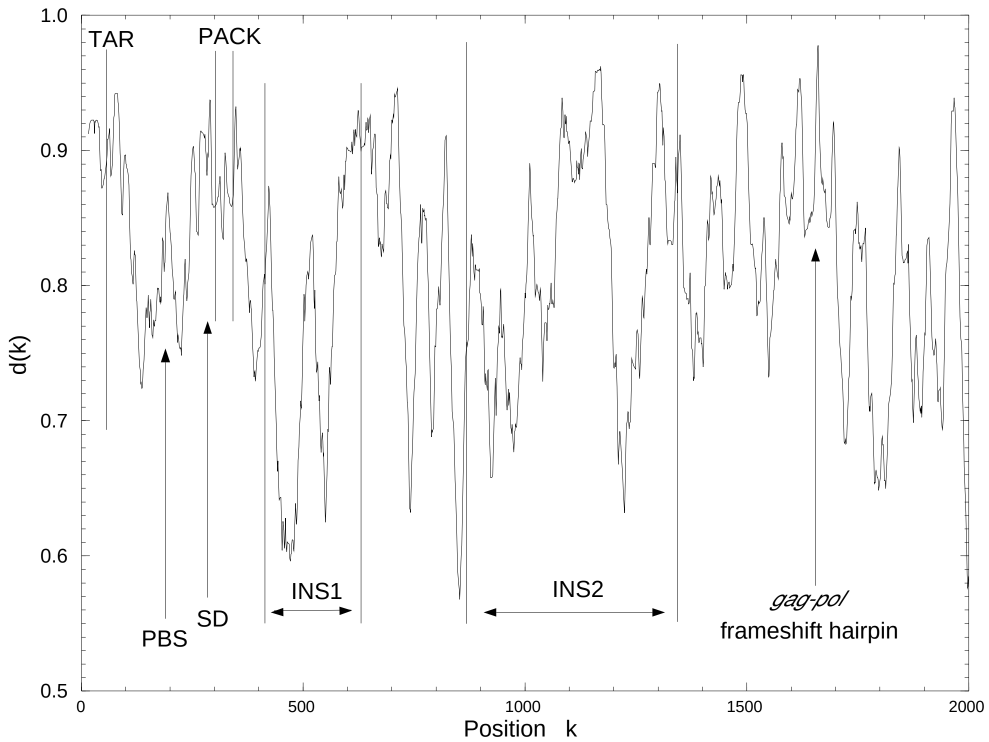

Figure 5: well-definedness of the first 2000 bases of the $\mathtt{HIV\text{-}1}_{LAI}$ structure. The solid line is a running average with a window size of 30. The positions of elements that are known to have a functional or at least a conserved RNA secondary structure are shown. The boundaries of TAR (only the 3' boundary, the 5' boundary is at the start of the sequence), the packaging signal (PACK, boundaries according to (Hayashi *et al.*, 1993))), and the instability sequences INS1 (Schwartz *et al.*, 1992) and INS2 (boundaries according to Ralf Schneider, pers. comm.) are indicated. The primer binding site (PBS), the splice donor (SD), and the *gag-pol* frameshift hairpin are indicated with arrows. The known secondary structure elements are further discussed in the next section.

A plot of d_k versus nucleotide position reveals information on the stability of small scale patterns. The idea behind measuring d_k is that the well-definedness of a region provides information on its functional significance. A secondary structure that is important for the function of a molecule should have a high probability of occuring in the thermodynamic ensemble of alternative secondary structures, and should not just be one of the many alternative structures that have a near equal probability of occuring. Note that the well-definedness measure does not depend on the well-definedness of *all* nucleotides within a secondary structure. As long as most of the nucleotides in a region of the sequence are well defined, the presence of a few nucleotides with variable base-paring behavior ("breathing") will not drastically decrease the well-definedness of the whole region.

3. Some Known Secondary Structure Patterns

The TAR element at the 5' End.

At the 5' end of HIV-1 resides the *trans*-activating responsive (TAR) element, which interacts with the regulatory Tat protein. The binding of the Tat protein to TAR increases transcription rates (Feng & Holland, 1988; Jeang *et al.*, 1991). The TAR hairpin presents itself in our folding data as a very well defined stem-loop structure (5-54), at about the 90% level in terms of our d-measure, see Figures 5 and 6. Comparing the two triangles in Figure 1 shows that the minimum free energy structure dominates the dot plot in this region.

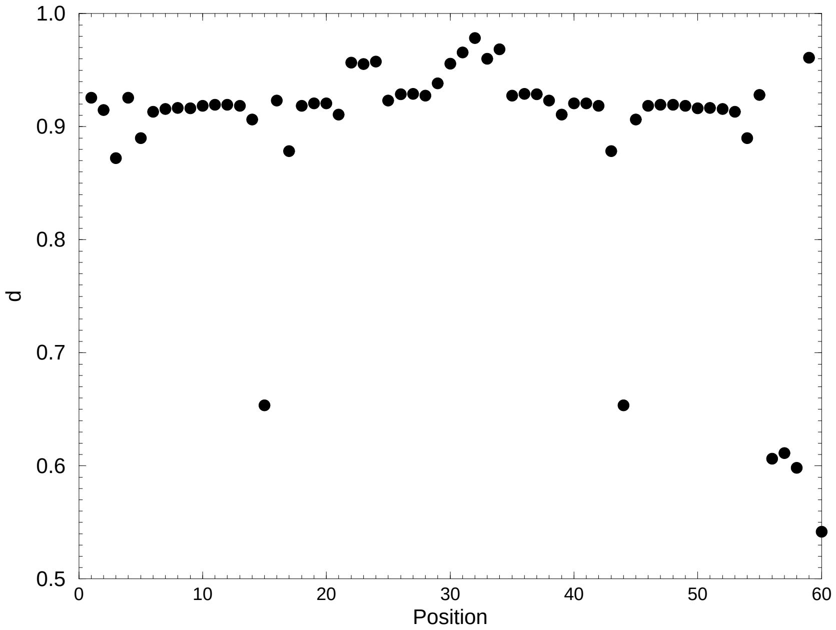

Figure 6: Well-definedness of the TAR hairpin motif.

The entire TAR motif appears to be well separated from the rest of the structure (Figure 7). The numerical values of the separation probabilities at the end of the TAR site are:

$$s_{54} = 0.1492 \qquad s_{55} = 0.1572 \qquad s_{56} = 0.1683 \qquad s_{57} = 0.0645\,.$$

These separation probabilities seem to be quite low at first glance. Note that for small m_k, $s_k = e^{-m_k}$. Thus $s_k = 0.14$ if $m_k = 2$. Furthermore, one has to take into account that there is a base pair $(1-57)$ with $p_{1,57} = 0.0529$, which decreases s_{56}, but in fact separates the TAR region from the rest of the molecule. This indicates that s_k should be regarded as a conservative measure of separation. We can say that the TAR motif is completely isolated in at least 16.8% of the thermodynamic ensemble. Nevertheless, there are many alternative structures that have pairs from the TAR region further upstream into the 3' region, mostly these pairs link various portions of the TAR region with a region in the beginning of the *gag* coding region, positions 360-400. Most of the pairs have very low probabilities. A few short stacks occur with probabilities of about one percent; a typical example is

$$p_{33,378} = 0.0144 \qquad \text{and} \qquad p_{34,377} = 0.0142.$$

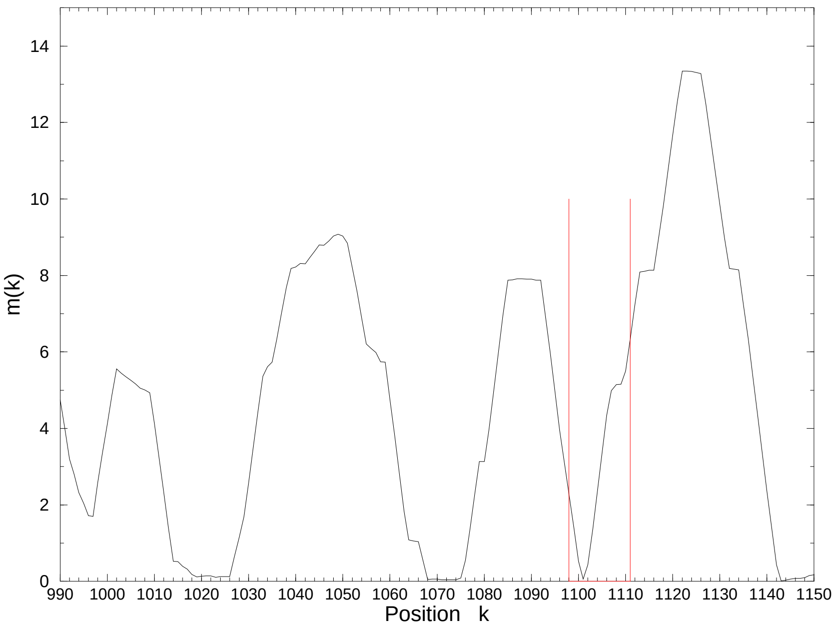

Figure 8: Mountain representation of the region hypothetically interacting with TAR (Chang & Tinoco Jr., 1994). The viral RNA decomposes into a number of small well separated components in this region. The proposed hairpin (bases 1098-1111, indicated with the dashed lines) does not form due to competition with the much more stable hairpin-loop elements on both sides.

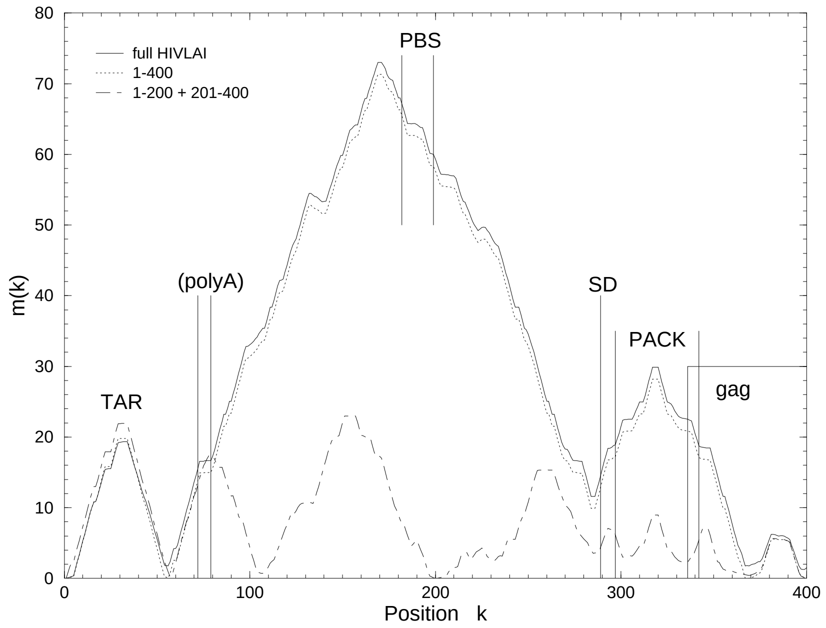

Figure 7: Generalized mountain representation of the 5' end of $\mathtt{HIV\text{-}1}_{\mathrm{LAI}}$ RNA.
The solid line is obtained from the base pairing probability matrix of the complete $\mathtt{HIV\text{-}1}_{\mathrm{LAI}}$ sequence. The dotted line was computed by considering the sequence fragment 1-400 only. The structure predictions are very close to each other because there is a component boundary at positions 396-401. The dot-dashed line was obtained from splitting the region into two windows of length 200 and folding them separately. It clearly shows the dramatic effect of using sequence windows on the predicted secondary structure. Except for the TAR region and the isolated hairpin between positions 375-400 which form components on their own the prediction has almost nothing in common with the structure obtained from folding the full genome. The begin of the *gag* coding region and a number of biologically significant sites are indicated in the figure.

The sequence in the hairpin loop of TAR (CUGGGA) is complementary to a strongly conserved sequence further downstream (UCCCAG) that is located in the loop of a postulated hairpin in the *gag* gene, positions 1098 to 1111. Biochemical analysis shows that these sequences might indeed interact with each other, forming a so-called "kissing hairpin" motif (Chang & Tinoco Jr., 1994). The kissing hairpin motif is a pseudoknot; i.e., it requires the interaction between bases that are internal to different secondary structure elements. Our analysis does not include the search for potential pseudoknots. What we can look for, however, is the presence of the hairpins in the loop of which the complementary sequences are located. In other words, are the nucleotides accessible for base pairing ? The TAR hairpin is present with a high probability as discussed above. The possible hairpin in *gag* with the UCCCAG motif however, occurs only with very low probability in the thermodynamic ensemble. The individual base pairing probabilities are

$$p_{1098-1111} \approx p_{1099-1110} \approx p_{1100-1109} \approx p_{1101-1108} \approx 5 \times 10^{-5}.$$

The reason why this hairpin does not form more frequently is competition with more stable hairpin structures which contain base pairs both upstream and downstream from the region between positions 1098 and 1111, see Figure 8. For instance, the stack on the 5' side contains the base pair (1077-1099) with probability $p_{1077,1099} = 0.856$.

The primer binding site region and the packaging signal region.

Reverse transcription of HIV-1 RNA into DNA is primed by a lysine tRNA that is bound to a region of 18 nucleotides in the 5' LTR, positions 182-199. The nucleotide sequence in this region, the so called the primer binding site (PBS), is complementary to the nucleotides at the 3' end of the tRNA. Figure 7 shows that the PBS is located in a partly unpaired region belonging to an interior loop (the mountain representation is partly horizontal in the PBS region). The results for the overall structure of the PBS region are not in strict agreement with those of (Baudin *et al.*, 1993) for HIV-1_{MAL}. HIV-1_{MAL} has an insertion relative to HIV-1_{LAI} of 23 nucleotides after position 221, which can explain the observed difference in predicted structure. Note however that the PBS lies in a not well defined region, it is located in the "dip" in Figure 5 that lies between the first and second peak which represent TAR and the packaging signal respectively. The specific secondary structure conformation of the PBS region therefore does not seem to be very significant for its function.

The packaging signal region (positions 285-367) forms a structure consisting of three or four consecutive stems separated by interior loops. Figures 5, 7 and 9 show this structural motif very clearly. A well defined component boundary with $s_{399} = 0.2812$ separates this region from the rest of the genome. Both the "bottom" stem in which this closing base pair is located and the "top" stem that contains the hairpin-loop GGAG are very well defined, whereas the two stems that lie in between are ill-defined (one of them is not present in the minimum free energy structure). The major-splice donor is located at the 5' end of the region,

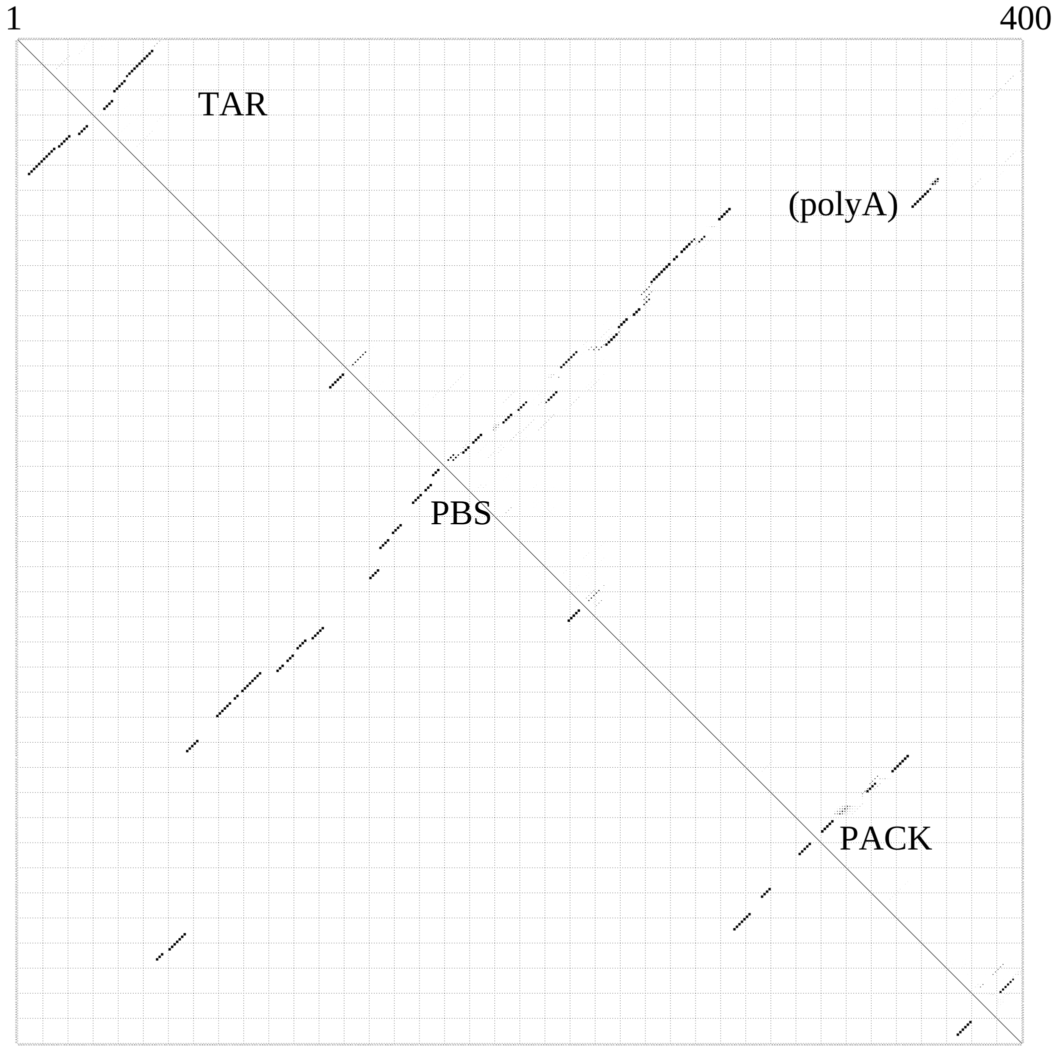

Figure 9: Dot plot of the 5' end of $\texttt{HIV-1}_{\mathrm{LAI}}$. The generalized mountain representation corresponding to this dot-plot is shown in Figure 7. The TAR region with its well defined stacks is clearly visible in the upper left corner. The potential base pairing between the TAR region and the area at the beginning of *gag*, between positions 360 and 400, appears as shaded rectangle in the upper right corner. The primer binding site (PBS) is located some 10 bases upstream from the hairpin in the middle of the plot in a quite flexible region. This can be seen more impressively in the well-definedness plot and the generalized mountain representation, see Figures 5 and 7. The non-functional repeat of the polyadenylation signal (polyA) is located in an unpaired portion of the multi-loop which carries both the PBS region and the linear stem-loop structure of packaging signal. This multi-loop is also present in the minimum free energy structure. The packaging signal (PACK) forms a linear stem-loop structure consisting of three or four stems. One of them appears to be meta-stable.

position 289. It resides in a strongly conserved (Harrison & Lever, 1992) and stable (Baudin *et al.*, 1993) secondary structure a few positions up-stream from the packaging signal. Figure 7 shows that it is located close to a minimum in the generalized mountain representation, enclosed by on average 11 base pairs, which form a stem that carries both the packaging signal region and the PBS region.

The hairpin loop 3' of the *gag-pol* frameshift site and the instability sequences.

The hairpin loop that is associated with the ribosomal frameshift between *gag* and *pol* is located between positions 1645 and 1672 in HIV-1$_{LAI}$. As is show in Figure 5 this hairpin is very well defined. This corresponds to the results of Le and co-workers (Le *et al.*, 1989), who showed for a large set of retroviruses that this hairpin has relatively low free energy and hence high thermodynamic stability.
A number of so-called instability elements have been identified in the *gag* and *pol* gene, which reduce HIV-1 protein expression in the absence of Rev (Schwartz *et al.*, 1992). Although the instability elements INS1 and INS2 do contain very well defined secondary structures (see Figure 5), the significance of this is not clear, given the regions over which the elements are defined are hundreds of nucleotides. Also in random sequences of that length one expects to find well defined structures.

The RRE Region.

The Rev response element (RRE, also called CAR), is an RNA structure that is located within the *env* gene (positions 7342-7695 in the definition in (Mann *et al.*, 1994)). The binding of the Rev protein to RRE promotes the transport of unspliced HIV transcripts to the cytoplasm (Malim *et al.*, 1989; Malim & Cullen, 1989; Mann *et al.*, 1994; Kimura & Ohyama, 1994).

The RRE region forms a well defined structure on the outside of a large bulk of secondary structure, enclosed by more than 350 base pairs. The stem-root structure (I) contains a total of 32 base pairs that do not show any significant alternative structures. It well separates the hairpins of the RRE from the rest of the RNA. The long stem-loop structure furthermore indicates that the RRE is easily accessible. There is very little interaction from the outside into the RRE region. This can be seen, for example, from the fact that $p_{7379,7576} \approx 0.943$, i.e., this base pair almost always encloses the RRE region.

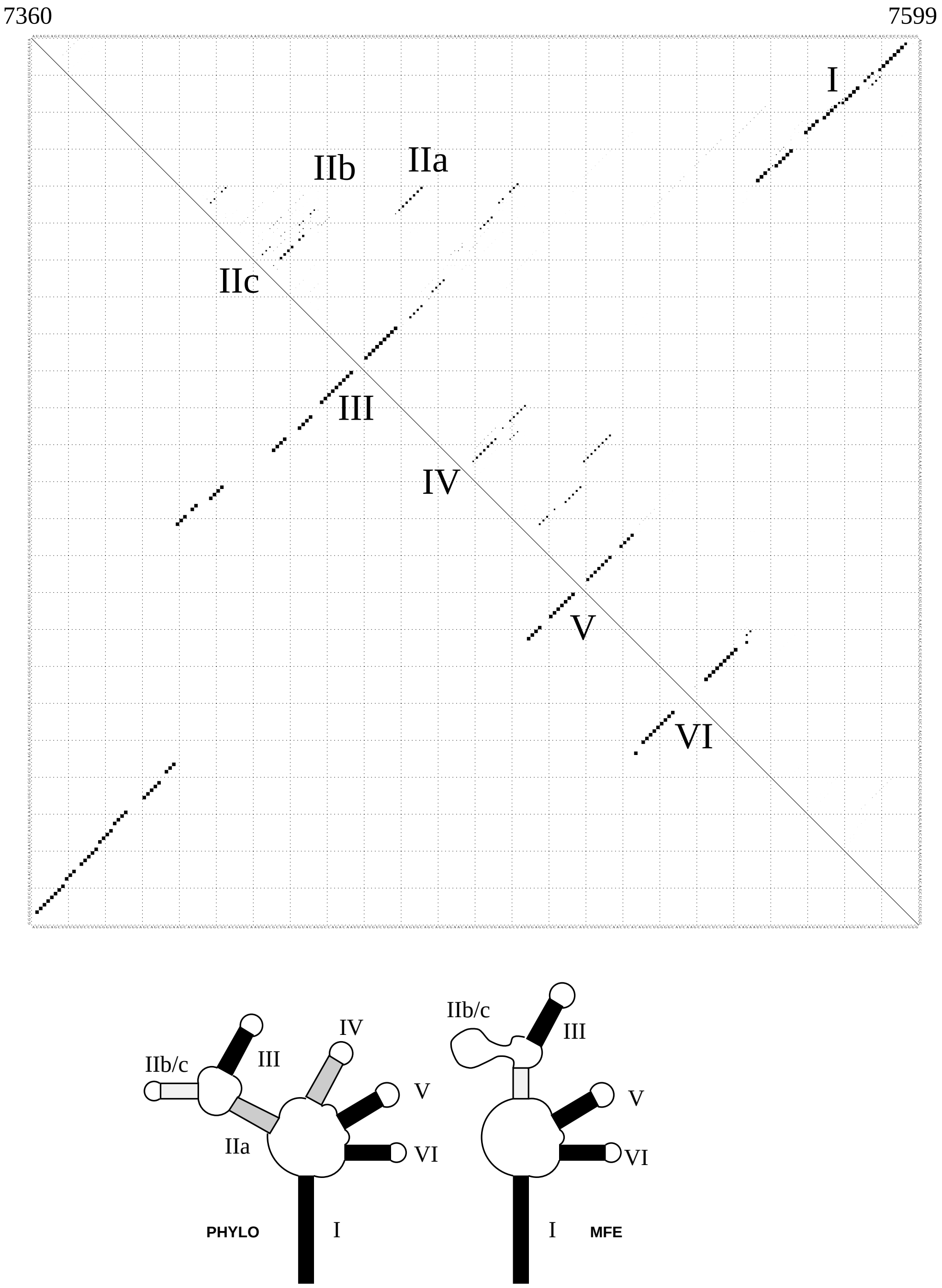

Figure 10: a) The RRE locus of $\texttt{HIV-1}_{LAI}$. **b)** Sketch of the two most important alternative structures. The structure on the left has been inferred from comparisons between of a number of HIV-1 RRE's (Konings, 1992). The right structure occurs as minimum free energy structure in our computation. Stacks occurring in both structures are shown in black.

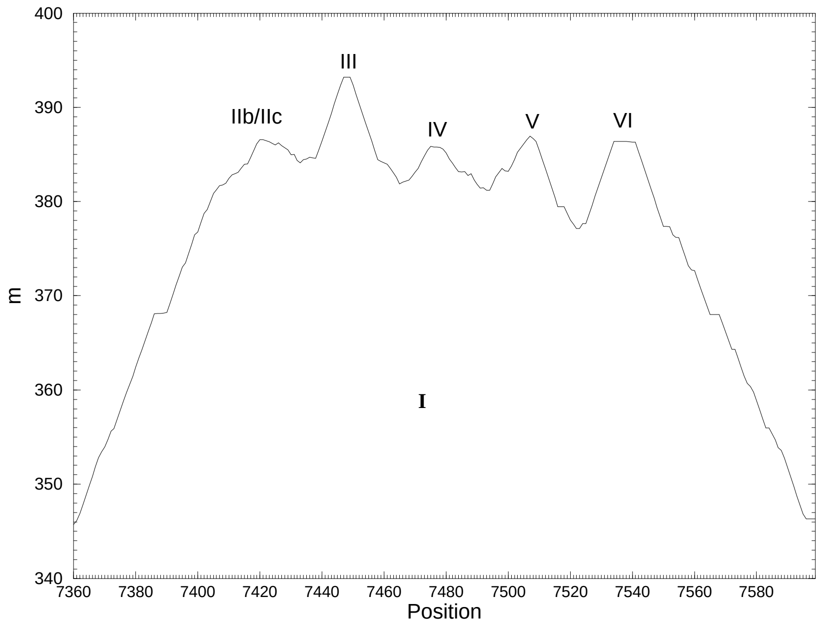

Figure 11: Generalized mountain representation of the RRE locus of $\mathtt{HIV\text{-}1}_{LAI}$. The five-fingered structural motif is clearly visible.

The consensus secondary structure for the RRE in HIV-1 consists of 5 hairpins in a multiple branched conformation closed by a single stem structure (Konings, 1992). An alternative structure of only 4 hairpins, in which the hairpins IV and V of the consensus model (nomenclature of hairpins according to (Dayton *et al.*, 1989)) merge to form one hairpin, has also been proposed (Mann *et al.*, 1994).

Extensive computer analysis has shown that the alignment of the RRE at the level of the sequence does not coincide with the alignment at the level of the secondary structure(Konings, 1992). This has two important implications: 1) methods that predict secondary structure of RNA on the basis of co-variation of positions within the sequence (Gutell *et al.*, 1993b) can not be used here, and 2) the RRE has structural versatility. The structural versatility could also play a role in a single HIV clone; i.e. as long as the structural conformation is close to some consensus conformation the RRE is functional. Such structural versatility is exactly the motivation for analyzing and presenting the secondary structure as an ensemble of base pair probabilities instead of a single or a few (alternative) structures.

The 3' End.

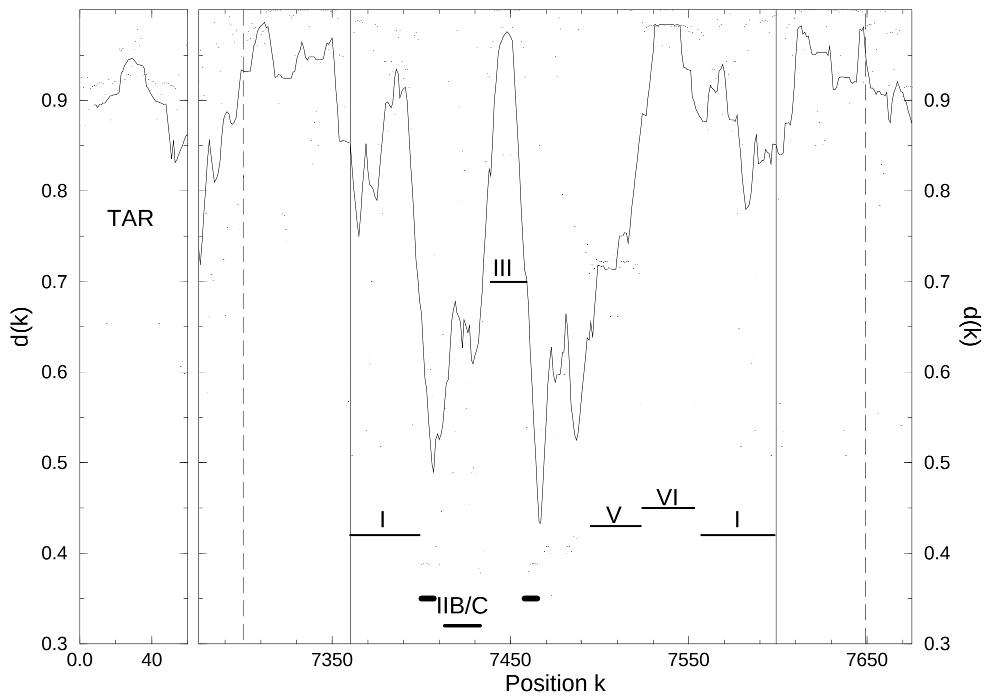

Figure 12: Well-definedness of the RRE locus (on the right) and of TAR (on the left) of HIV-1_{LAI}. The solid line shows a running average with window size 15. The stability of the stack (I) in the RRE is clearly visible. The binding site for the Rev protein (IIB/C) corresponds to a relatively ill-defined region of the secondary structure, especially in comparison to an other protein binding site, the TAR region. The minima in the Figure (located above the two short line-segments)) correspond to two regions that in the consensus model of RRE form the stem IIA. The two vertical lines indicate the "minimum" definition of the RRE of about 230 nucleotides (Malim *et al.*, 1989). The new definition as given by (Mann *et al.*, 1994) extends 60 nucleotides to the right and to the left and is marked by the dashed lines. The extension of stem structure I that corresponds to this definition has a high probability.

The 3' end of the HIV-1 viral genome consists of a repeat of about the first 100 nucleotides from the 5' end of the sequence. The mountain plot in Figure 13 clearly shows the repeat of the TAR region, which has no known function in this position. Its structure is very similar to the functional copy at the 5' end of the genome, this is in agreement with (Wang *et al.*, 1993). The polyadenylation signal (AAUAAA) lies in a small, weakly determined hairpin. The polyadenylation signal interacts with the cleavage and polyadenylation specificity factor (reviewed in (Wahle & Keller, 1992)). This interaction does not depend on a specific secondary structure formation of the polyadenylation signal. It is not surprising therefore that the secondary structure containing the sequence AAUAAA is not well defined.

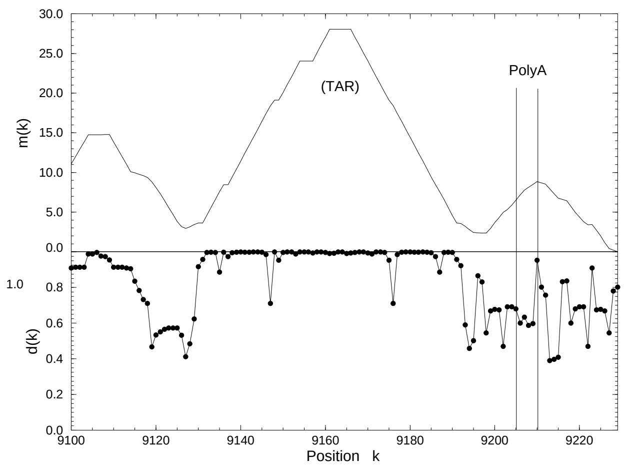

Figure 13: The 3' end of $\texttt{HIV-1}_{LAI}$. The upper part of the figure shows the generalized mountain representation, the lower part of the figure shows the well-definedness of the base pairing. (TAR) indicates the repeat of the TAR signal, the vertical lines indicate the AAUAAA sequence. The 3' repeat of TAR is very well defined, the polyadenylation signal is located in a not well defined hairpin.

4. Conclusions and Discussion

Global secondary structure of HIV-1.

We have computed and analyzed the base pair probability distribution in the full length genome of one HIV-1 sequence, HIV-1_{LAI}. The secondary structure of the HIV-1 genome shows a striking difference between the 5' half and the 3' half of the molecule. Whereas the first 4000 nucleotides form a large number of well separated elements, the second 5000 nucleotides (excluding the terminal repeat) form a single multi-stem secondary structure, with pairing between bases that are thousands of nucleotides apart. The first 4000 nucleotides of HIV-1 are rich in instability sequences (INS, CRS), and contain few splice-sites, whereas the 3' region of the molecule shows an abundance of splice sites. To the extent that RNA secondary structure is important for splicing or the functioning of the instability sequences the difference between the 5' and the 3' region of the molecule could be significant. Pairing between bases that are far apart in the sequence has been observed in Hepatitis Delta Viruses (maximum distance between paired bases about 1600, Andrew Ellington, pers. communication; unpublished data) and in 23s ribosomal RNA (maximum distance about 2900) (Gutell *et al.*, 1993a).

Representing the information in the base pair probability distribution.

Although the base pair probability distribution contains the most complete information about the secondary structure of any RNA molecule, the interpretation of the results for long sequences is hard, given the enormous amount of data. We have introduced methods that condense the information from the probability distribution to a single value per nucleotide. Our modification of Hogeweg and Hesper's mountain representation (Hogeweg & Hesper, 1984) gives a representation of the average expected structure of the whole molecule. Our method for calculating the separation probabilities indicates how the secondary structure separates into independent elements. Especially when the number of alternative secondary structures becomes large, as is the case for long sequences, these methods give a better impression of the predicted secondary structure than the full base pair probability matrix. When our method detects interesting regions, it is then possible to "zoom in" at a higher level of resolution for small regions of the sequence, e.g. by looking at the dot-plot representation or the minimum free energy secondary structure. However, if the number of alternative secondary structures remains large for small regions of the sequence, condensing the information in the base pair probability distribution becomes a much more meaningful way of describing secondary structure than just enumerating all the different secondary structures [see also the discussion in (Jacobsen & Zuker, 1993)].

Well-definedness of secondary structure.

Almost all RNA molecules — and consequently also almost all subsequences of a large RNA molecule — form secondary structures. The presence of a secondary structure in itself therefore does not imply any functional significance. A method has been developed by Le et al. (Le *et al.*, 1988) that identifies unusual secondary structures in a single sequence on the basis of their (predicted) free energy.

Although the method has been successful in finding significant structures, like the RRE in CAEV (Saltarelli *et al.*, 1988), it has definitely disadvantages. It is a method that works with a sliding window, and hence can give rise to artificial boundary effects. Furthermore, functional secondary structures do not necessarily have relative low free energies. A better clue for the presence of a functional secondary structure is its "well-definedness", that is, does the minimum free energy structure (MFE) have a relatively high probability of occurring (Huynen, 1993). A clue that such a method might work is the finding that tRNA secondary structures have, within a small percentage of free energy, fewer alternative secondary structures than do the secondary structures of random sequences (Marliere, 1983; Higgs, 1993). The most straightforward parameter for the well-definedness of a secondary structure is the frequency of the MFE in the partition function. This measure does however have the disadvantage that it only reflects the probability of a single secondary structure, and thus becomes very small if there are a few nucleotides that are ill defined within an otherwise well defined structure. Furthermore this measure would have to be calculated independently for different window sizes.

By calculating the probability distribution of all base pair interactions, we have the access to a better tool for the prediction of structures that occur with a relatively high probability. In this paper we have introduced a measure for well-definedness of secondary structure that is based on the probabilities of base pair interaction of single nucleotides. In this method every nucleotide has a value that is based on either the maximum probability it has of interacting with a specific other nucleotide, or the probability of not interacting at all, whichever is bigger. By taking running averages of the well-definedness we can identify regions in the sequence that give rise to well defined secondary structures.

Well-definedness of some known secondary structures in HIV-1.

We have shown that the TAR element is well defined, although there are a few low probability interactions between nucleotides in TAR and the beginning of the *gag* coding region. These long range interactions might not play a role in the biotic system. If the Tat-TAR interaction is an elongation signal (Kao *et al.*, 1987) it could stabilize the TAR secondary structure before the nucleotides that potentially interact with TAR are transcribed.

The splice donor and the hairpin 3' of the *gag-pol* frameshift have relatively well defined base pairing. The nucleotides in the primer binding site do have well defined base pairing, but lie within an otherwise less well defined region.

The packaging signal is in our analysis located in 3 to 4 consecutive stems that are separated by internal loops. Of these only the "top" stem, which contains the GGAG hairpin-loop, and the "bottom" stem are well defined. The GGAG hairpin is included in all previously proposed secondary structures of the packaging signal (Harrison & Lever, 1992; Sakaguchi *et al.*, 1993; Hayashi *et al.*, 1993; Baudin *et al.*, 1993; Clever *et al.*, 1995). Our overall model of the secondary structure (a single hairpin-loop located on top of three to four consecutive stems) is equal to the model proposed in (Hayashi *et al.*, 1993) and to one of the alternative structures proposed in (Baudin *et al.*, 1993) but is significantly different from the other models which include two or more hairpin-loops (Harrison & Lever, 1992; Sakaguchi *et al.*, 1993; Clever *et al.*, 1995). These multi hairpin-loop models are all based on an analysis of the secondary structure of only the region of the sequence that contains the packaging signal, and do exclude potential pairing between bases in this region and bases outside this region. As we have shown, analyzing the secondary structure between position 200 and 400 independently from the rest of the genome dramatically alters the secondary structure of the packaging region. It gives rise to a secondary structure with four hairpin loops that is equal to the structure proposed in (Clever *et al.*, 1995). This restates the importance of folding entire sequences, or at least isolating separate regions that only have local base pairing before analyzing them in detail (Konings *et al.*, 1992). The overall structure of the 5' region (including the beginning of the *gag* coding region) agrees fairly well with the structure in (Baudin *et al.*, 1993), the major difference being that in our analysis the structure containing the PBS and the structure containing the packaging signal are not completely independent, but are connected by a on average an 11 base pair stem structure.

The instability sequences INS1 and INS2 do contain well defined secondary structures, but are defined over too long ranges of sequence for this observation to be meaningful. The 3' copy of the direct repeat does show a well defined TAR element. The polyA sequence lies in a separate, not well defined structure.

For the RRE we find that the main stem structure (stem I) is very well defined and so is the the hairpin loop at its 3' end (stem VI). The other hairpin loops (IIb/c, IIa, IV, V and part of III) are much less well defined and show structural versatility. A nice example is the IIb/c stem which contains the high affinity site for the binding of Rev. The IIb/c stem is not part of the minimum free energy folding, but is clearly visible in the dot-plot and in our generalized mountain representation. It is puzzling that the high affinity binding site for the Rev protein lies in a relatively ill defined secondary structure, especially in comparison with TAR, the other site known to interact with a protein. Apparently the local secondary structure of the sequence is not that relevant for the binding of Rev. An alternative hypothesis is that the structural versatility of this region actually has a function. The structural versatility of RNA in general can lead to a high specificity of protein-RNA interactions (Ellington, 1993), specifically if the RNA has to sacrifice binding energy in the form of a conformational reorganization upon the binding of a protein. The binding of Rev proteins is cooperative process; the binding of the first Rev at the high affinity site facilitates the binding of other Rev proteins along stem IIa and stem I (Mann *et al.*, 1994). This process has been attributed to protein-protein interactions. An alternative possibility is that the binding of Rev to the high affinity site stabilizes a specific conformation of the secondary structure, therewith giving the other binding sites, in particular the one on stem IIa, the right secondary structure for the binding of other Rev proteins. This idea is supported by the findings of Tan and Frankel (Tan & Frankel, 1994) who show that the RNA in the RRE undergoes a conformational change upon the binding of Rev, and who hypothesize that the binding of Rev is needed to stabilize the RRE-IIb RNA structure.

A multitude of other well defined secondary structures of unknown function are present in the HIV sequence. However not every high score necessarily corresponds to a functionally significant secondary structure, since randomly generated sequences can also give rise to well defined secondary structures. To make reliable predictions for the significance of new, as yet undiscovered secondary structure elements we need to analyze the base pair probability distribution for a number of HIV-1 sequences. Our method can easily be extended to the analysis of a set of related sequences by taking the (geometric) mean of the well-definedness score for nucleotides in the same positions in the aligned sequences. A related question is the stability of structural motifs against mutations, both in the active secondary structure elements and in its flanking regions. One would speculate that biologically important motifs have evolved towards a higher than average tolerance against mutations in their neighborhood (Huynen & Hogeweg, 1994). The computational requirements for investigating these questions are tremendous and will require the use of massively parallel computers. A first step in this direction is a comparison of complete lentivirus minimum free energy structures (Hofacker *et al.*, 1995) which has been performed on a hypercube-type parallel computer. A parallel computer implementation of the partition function algorithm is currently under construction.

Acknowledgments

We thank Geert Wenes, Cray Research Inc., for arranging computing time on the `CRAY Y-MP`. We thank Andrew Ellington for comments on the original manuscript. This 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.

Appendix: Folding RNA on the CRAY Y-MP

RNA Secondary Structure Prediction.

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 $i \leq k$, the following holds (i) $i = k \iff j = l$, and (ii) $k < j \implies i < k < l < j$. The first condition implies that each nucleotide can take part in not more that one base pair, the second condition forbids knots and pseudoknots. 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[1].

As a consequence of the additivity of the energy contributions, the minimum free energy can be calculated recursively by dynamic programming (Waterman, 1978; Waterman & Smith, 1978; Zuker & Stiegler, 1981; Zuker & Sankoff, 1984). Experimental energy parameters are available for the contribution of an individual loop as functions of its size, of the type of its delimiting base pairs, and partly also as a function of the sequence of the unpaired strains. The energy parameters are extracted from measurements at (usually) $T = 37°C$ in $1M$ NaCl solution (Freier *et al.*, 1986; Jaeger *et al.*, 1989). For the base pair stacking both the enthalpic and the entropic contributions are known separately. Contributions from all other loop types are assumed to be purely entropic. Therefore one can estimate the actual free energy contributions for a temperature interval around $T = 37°C$:

$$\Delta G_{\text{stack}} = \Delta H_{\text{stack}} - T\Delta S_{\text{stack}} \qquad \Delta G_{\text{loop}} = -T\Delta S_{\text{loop}}$$

The free energy of a secondary structure is then

$$\Delta G = \sum_{\text{all stacks}} \Delta G_{\text{stack}} + \sum_{\text{all loops}} \Delta G_{\text{loop}}.$$

[1]Note that a stack constitutes an interior loop without unpaired bases.

We use a recent version of the parameter set published by Freier *et al.* (Freier *et al.*, 1986), which was supplied in an updated version (Jaeger *et al.*, 1989) by Danielle Konings. The current implementation does not consider dangling ends.

McCaskill's Algorithm.

The partition function for the ensemble of all possible secondary structures,

$$Q = \sum_{\text{all structures } S} e^{-\frac{\Delta G(S)}{RT}} ,$$

can be calculated analogously to the minimum free energy structure (McCaskill, 1990). Let Q_{ij} denote the partition function for the sub-sequence from i through j; thus $Q \equiv Q_{1n}$. Auxiliary arrays, such as the partition function Q^b_{ij} of subsequence i through j subject to the constraint that (i, j) is a base pair, are defined in the caption of Figure 14 which gives a pseudocode for the computation of Q.

```
for(d=1...n)
   for(i=1...d)
      j=i+d
      QB[i,j] = EHairpin(i,j) +
         SUM( i<p<q<j : EInterior(i,j;p,q)*QB[p,q] ) +
         SUM( i<k<j : QM[i+1,k-1]*QM1[k,j-1]*Ecc )
      QM[i,j] =
         SUM( i<k<j : (Ecu^(k-i)+QM[i,k-1])*QM1[k,j] )
      QM1[i,j]= SUM( i<k<=j : QB[i,k]*Ecu^(j-k)*Eci )
      Q[i,j]  = 1 + QB[i,j] +
         SUM( i<p<q<j : Q[i,p-1]*QB[p,q] )

partition_function = Q[1,n]
```

Figure 14: Pseudocode of McCaskill's partition function algorithms.
Ex:=exp($-x/RT$) denotes the Boltzmann weights corresponding to the energy contribution x. In particular, `cu`, `ci`, and `cc` are parameters referring to multi-loops, `Hairpin` is the hairpin contribution and `Interior` is the energy contribution of interior loops and stacks.
`Q[i,j]` denotes the partition function Q_{ij} of the subsequence i through j. The array `QM` contains the partition function Q^b_{ij} of the subsequence subject to the fact that i and j form a base pair. `QM` and `QM1` are used for handling the multiloop contributions. `x^y` means x^y. For details see (McCaskill, 1990).

Clearly the algorithm of Figure 14 does not predict a unique secondary structure. One can, however, extract the probability p_{kl} for the formation of a base pair (k,l) from the partition function via the following backtracking recursion:

$$p_{kl} = \frac{Q_{1,k-1}Q^b_{kl}Q_{l+1,n}}{Q} + \sum_{\substack{i,j \\ i<k<l<j}} \frac{p_{ij}Q^b_{kl}\texttt{EInterior}(i,j;k,l)}{Q^b_{ij}} +$$

$$\texttt{Ecc}\cdot\texttt{Eci} \sum_{\substack{i,j \\ i<k<l<j}} \frac{p_{ij}}{Q^b_{kl}} \left[\texttt{Ecu}^{k-i-1}Q^m_{l+1,j-1} + Q^m_{i+1,k-1}\texttt{Ecu}^{j-l-1} + Q^m_{i+1,k-1}Q^m_{l+1,j-1}\right]$$

Being products of a large number of terms the partition functions Q_{ij} become very large. In order to reduce the numerical problems arising for long sequences we rescale the partition function of a subsequence of length ℓ by a factor $\tilde{Q}^{\ell/n}$, where $\tilde{Q}$ is a rough estimate of the partition function Q_{1n}. This procedure works well as long as the composition of the RNA sequence is reasonably close to a random sequence. In this case one can use the expression

$$\tilde{Q}^{1/n} = \exp\left(\frac{184.3 - 7.27(T - 37.)}{1.98T}\right)$$

as a good estimate, see (Hofacker *et al.*, 1994).

Implementation and Performance.

Both a minimum free energy folding algorithm and McCaskill's partition function algorithm have been integrated into a single interactive program including post-script output of the minimum energy structure and the base pairing probability matrix. This program is part of the `Vienna RNA Package` (Hofacker *et al.*, a).

For long sequences both folding algorithms have complexity $\mathcal{O}(n^3)$. For instance, folding the complete genome of the bacteriophage Qβ took about 10 hours on an `IBM Risc6000/550` workstation (Hofacker *et al.*, 1994). The limiting factor is not CPU time, however, but memory. The folding algorithm as implemented in the `Vienna RNA Package` requires $6\,n^2$ bytes of memory for the minimum energy folding and $10\,n^2$ bytes for the calculation of the partition function on machines with 32 bit integers and single precision floating point numbers. The large numbers occurring in the computation of the partition function forced us to use double precision floating point numbers, thereby doubling the memory requirements. The

```
Operating System           : sn2402 rain 7.C.3 prd.3 CRAY Y-MP
User CPU Time              :       227502.3132 Seconds
System CPU Time            :         6921.4689 Seconds
CPU Time Memory Integral   :     37327121.1483 Mword-seconds
Maximum memory used        :          203.3047 MWords
```

Figure 15: Performance Log for the folding on the CRAY Y-MP. A MegaWord (Mword) on a CRAY Y-MP is eight MegaBytes. The computation required 63 hours CPU time and 1.6 GigaByte of memory.

`C` source code of the `Vienna RNA package` could be compiled on the `CRAY-M90` without further alterations. Subsequently, a number of modifications have been introduced in order to optimize the code. Two approaches were successful:

(i) Most of the code is by itself highly parallel, since all entries Q_{ij} with a common value of $i-j$ can be computed in parallel because they are independent from each other. Adding various tasking directives resulted in a 6 to 7 times speedup in certain sections, and in an approximately 2.5 times speedup overall. On a CRAY system, the sections that remained serial did not hold up the 8 processor `M90` which was used for the computation. The remaining 7 processors were still available for other users. The routines that were not so easily parallelized probably still could be parallelized with some effort.

(ii) CRAY compilers have an efficient facility for automatic inlining. Depending on the size of a function, and its frequency of use, the function's code can be inserted into the place where it was called. As a result, the overhead of a given calling sequence (allocating memory, placing parameters on a stack, etc.) can be eliminated. In addition, this makes it easier for the compiler to vectorize and parallelize loops surrounding that function. By encouraging this process (making *all* source files visible to the compiler at the same time) a 20-25% performance improvement was obtained.

Access to the Data.

The complete base pairing probability matrix contains too much information to be conveniently displayed as a single figure. In fact, a printout on letter-size paper does not reveal much more than the concentration of interactions along the diagonal, indicating the well known fact that most interactions in RNA molecules are quite local, i.e., span no more than a few hundred bases. Enlarged details of the base pairing probability matrix are show in the second and third section, see Figure 1, 9 and 10. The complete data set is available electronically. A `PostScript` plot of the base pair probability matrix and an `ASCII` file containing the probabilities of all base pairs down to a pairing probability of 10^{-5} can obtained by anonymous `ftp` from the following locations:

- `ftp.itc.univie.ac.at` directory: `pub/HIV`
- `ftp.santafe.edu` directory: `ftp/mah/HIV`

References

Baudin, F., Marquet, R., Isel, C., Darlix, J., Ehresmann, B., & 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. & Schoneveld, I. (1993). Secondary structure of the HIV-2 leader RNA comprising the tRNA-primer binding site. *Nucl. Acids Res.* **21**, 1171–1178.

Biebricher, C. (1994). The role of RNA structure in RNA replication. *Ber. Bunsenges. Phys. Chem.* **98**, 1122–1126.

Bonhoeffer, S., McCaskill, J., Stadler, P. F., & Schuster, P. (1993). RNA multi-structure landscapes, a study based on temperature dependent partition functions. *Eur. Biophys. J.* **22**, 13.

Chang, K. & Tinoco Jr., I. (1994). Characterization of a "kissing" hairpin complex derived from the HIV genome. *Proc. Natl. Acad. Sci., USA,* **91**, 8705–8709.

Clever, J., Sassetti, C., & Parslow, T. (1995). RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1. *J. Virol.* **69**, 2101–2109.

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

Ellington, A. E. (1993). Out of shape but fit for recognition. *Current Biology,* **3**, 375–377.

Feng, S. & Holland, E. (1988). HIV-1 Tat trans-activation requires the loop sequence within TAR. *Nature,* **334**, 165–167.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA,* **83**, 9373–9377.

Gutell, R. R., Gray, M. W., & Schnare, M. N. (1993a). A compilation of large subunit (23s and 23s-like) ribosomal rna structures. *Nucl. Acids Res.* **21**, 3055–3074.

Gutell, R. R., Power, A., Hertz, G., Putz, E., & Stormo, G. (1993b). Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Research,* **20**, 5785–5795.

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

Hayashi, T., Shioda, T., Iwakura, Y., & Shibuta, H. (1992). RNA packaging signal of human immunodeficiency virus type 1. *Virology,* **188**, 590–599.

Hayashi, T., Ueno, Y., & Okamoto, T. (1993). Elucidation of a conserved RNA stem loop structure in the packaging signal of human immunodeficiency virus type 1. *FEBS,* **327**, 213–218.

Higgs, P. (1993). RNA secondary structure: a comparison of real and random sequences. *Journal de Physique,* **3**, 43–59.

Hofacker, I., Huynen, M., Stadler, P., & Stolorz, P. (1995). The minimum free energy structures of complete HIV genomes. In preparation.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (a). `Vienna RNA Package`. `pub/RNA/ViennaRNA-1.03 @ ftp.itc.univie.ac.at`. (Public Domain Software).

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie,* **125** (2), 167–188.

Hogeweg, P. & Hesper, B. (1984). Energy directed folding of RNA sequences. *Nucleic Acids Research,* **12**, 67–74.

Huynen, M. A. (1993). *Evolutionary Dynamics and Pattern Generation in the Sequence and Secondary Structure of RNA*. PhD thesis Utrecht University.

Huynen, M. A. & Hogeweg, P. (1994). Pattern generation in molecular evolution: exploitation of the variation in RNA landscapes. *Journal of Molecular Evolution,* **39**, 71–79.

Jacobsen, A. B. & Zuker, M. (1993). Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.* **46**, 261–269.

Jaeger, J. A., Turner, D. H., & Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci., USA,* **86**, 7706–7710.

Jeang, K.-T., Chang, Y., Berkhout, B., Hammarskjöld, M.-L., & Rekosh, D. (1991). Regulation of HIV expression: mechanisms of action of Tat and Rev. *AIDS,* **5 (suppl 2)**, 3–14.

Kao, S., Calman, A., Luciw, P., & Peterlin, B. (1987). Anti-termination of transcription within the long terminal repeat of HIV-1 by *tat* gene product. *Nature,* **330**, 489–493.

Kimura, T. & Ohyama, A. (1994). Interaction with the Rev response element along an extended stem I duplex structure is required to complete human immunodeficiency virus type 1 Rev-mediated trans-activation in vivo. *J. Biochemistry,* **115**, 945–952.

Klaver, B. & Berkhout, B. (1994). Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. *Embo J.* **13**, 2650–2659.

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

Konings, D., Nash, M., Maizel, J., & Arlinghaus, R. (1992). Novel GACG-hairpin pair motif in the 5' untranslated region of type C retroviruses related to murine leukemia virus. *J. Virol.* **66**, 632–640.

Laughrea, M. & Jette, L. (1994). A 19-nucleotide sequence upstream of the 5' major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. *Biochemistry,* **33**, 13464–13474.

Le, S.-Y., Chen, J.-H., Currey, K., & Maizel, J. (1988). A program for predicting significant RNA secondary structures. *CABIOS,* **4**, 153–159.

Le, S.-Y., Chen, J.-H., & Maizel, J. (1989). Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucl. Acids Res.* **17**, 6143–6152.

Malim, M. & Cullen, B. (1989). HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: implications for HIV-1 latency. *Cell,* **65**, 241–248.

Malim, M., Hauber, J., Le, S., Maizel, J., & Cullen, B. (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., Mikaelian, I., Zemmel, R., Green, S., Lowe, A., Kimura, T., Singh, M., Butler, P., Gait, M., & 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.

Marliere, P. (1983). Computer building and folding of fictitious transfer-RNA sequences. *Biochimie,* **65**, 267–273.

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

Pathak, V. & Temin, H. (1992). 5-azacytidine and RNA secondary structure increase the retrovirus mutation rate. *J. Virol.* **66**, 3093–3100.

Sakaguchi, K., Zambrano, N., Baldwin, E., Shapiro, B., Erickson, J., Omichinski, J., Clore, G., Gronenborn, A., & Appella, E. (1993). Identification of a binding site for the human immunodeficiency virus type 1 nucleocapsid protein. *Proc. Natl. Acad. Sci., USA,* **90**, 5219–5223.

Saltarelli, M., Querat, G., Konings, D., Vigne, R., & Clements, J. (1988). Nucleotide sequence and transcriptional analysis of molecular clones of HIV which generate infectious virus. *Virology,* **179** (3), 387–393.

Schwartz, S., Campbell, M., Nasioulas, G., Harrison, J., Felber, B., & Pavlakis, G. (1992). Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent *gag* expression. *J. Virol.* **66** (12), 7176–7182.

Tan, R. & Frankel, A. (1994). Costabilization of peptide and RNA structure in an HIV rev peptide-RRE complex. *Biochemistry,* **33**, 14579–14585.

Vickers, T. & Ecker, D. (1992). Enhancement of ribosomal frameshifting by oligonucleotides targeted to the HIV *gag-pol* region. *Nucl. Acids Res.* **20**, 3945–3953.

Wahle, E. & Keller, W. (1992). The biochemistry of 3' cleavage and polyadenylation of messenger RNA precursors. *Ann. Rev. Biochem.* **61**, 419–440.

Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S., & Alizon, M. (1985). Nucleotide sequence of the AIDS virus, LAV. *Cell,* **40**, 9–17.

Wang, P., Rouyey, M.-C., Ducamp, S., Saragosti, S., & Ventura, M. (1993). Similarity of the 5' and 3'-tar secondary structures in HIV-1. *Biochem. Biophys. Res. Commun.* **195**, 565–573.

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.

Waterman, M. S. & Smith, T. F. (1978). RNA secondary structure: A complete mathematical analysis. *Math. Biosc.* **42**, 257–266.

Zuker, M. (1989). The use of dynamic programming algorithms in RNA secondary structure prediction. In: *Mathematical Methods for DNA Sequences*, (Waterman, M. S., ed) pp. 159–184. CRC Press.

Zuker, M. (a). `mfold-2.0. pub/mfold.tar.Z @ nrcbsa.bio.nrc.ca`. (Public Domain Software).

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

Zuker, M. & Stiegler, P. (1981). Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids. Res.* **9**, 133–148.