Secondary Structure Prediction of Proposed RNAi Targets:

Can Current Energy Minimization Algorithms Be Used?

Laura Almstead
BIOC 218
June 6, 2002
almstead@stanford.edu

INTRODUCTION

Just as the three-dimensional shape of a protein can provide essential information in the determination of its function, the activity of an RNA molecule also depends strongly on its spatial organization. Though modeling of tertiary RNA structures such as pseudoknots has proved a challenge, relatively recent advances in modeling algorithms have allowed for the development of several web-based secondary structure modeling programs. Until now, most computationally-based RNA secondary structure prediction has relied on the use of comparative sequence analysis supported by additional information from RNase sensitivity and nucleotide chemical base reactivity assays. With the recent explosion of RNA interference (RNAi) as a technique for "non-genetic" gene knock-out and/or knock-down, determining the structure of relatively small segments of mRNA will become increasingly important. In RNAi, short segments of double stranded RNA called small interfering RNAs (siRNAs) are used to induce the degradation of a particular mRNA product. Specificity is achieved through the complementarity of the siRNA to the target mRNA. For example, in mammalian systems, siRNA sequences are chosen by scanning the mRNA of interest for AA's and then recording the next nineteen nucleotides to create a 21 nucleotide dsRNA molecule with a UU overhang. siRNAs have been found to have a wide range of efficiency depending on a multitude of parameters such as the degree of structural complexity found at that region. Arguments continue as to the importance of these parameters, but it is thought that highly organized secondary structure in the target could decrease siRNA effectiveness due to the presence double stranded regions and/or inhibited accessibility. Therefore, it is likely that scientists from diverse backgrounds will be looking for accurate methods of predicting the secondary structure of short segments of mRNA sequences without the benefit of multiple sequence alignments.

The aim of this exploration is to examine the effectiveness of two web-based RNA secondary structure prediction programs (***mfold*** and GeneBee) to meet this developing need. The internal ribosome entry sites (IRESs) of five ssRNA positive-strand viruses were chosen for the analysis. These IRESs are known to have high degrees of functionally significant secondary structure and, as they function to recruit translational machinery to the mRNA, are potential targets for RNAi. In addition,

published structures based on multiple methods of structure determination are available.

BACKGROUND

Historically, thermodynamic and topological rules have governed RNA secondary structure determination resulting in energetically favorable conformations. Early prediction methods were restricted to short sequences due to computational time limitations. In an attempt to improve the computational aspect, Nussinov et. al. designed a dynamic programming algorithm maximizing the base pairing in a structure. Though faster, thermodynamic data, stacking energies, and destabilizing energies could not be included. This was partially remedied by Waterman and Smith in a less efficient method that included the above parameters, but was limited to sequences shorter than two hundred nucleotides.

Diverging slightly from these methods, Hugo Martinez took a more biological perspective. To determine the best methods for predicting RNA secondary structure, it is important to consider ***in vivo*** RNA folding and stability. For mRNA molecules, sequential folding as the molecule is transcribed by the ribosome would seem logical. In addition, the function of RNA often depends strongly on its structure. Therefore, it is likely that evolutionary pressure has functioned to maintain this structure. Any drastic evolutionary changes, such as an increase in length would most likely lead to the addition of secondary structure based upon the original structure. RNA molecules must also be energetically stable to be able to maintain their structure and perform their functions. Consideration of these factors led to the proposal of a simple rule for RNA secondary structure prediction based on the sequential addition of stem-forming double stranded regions.[1] Specifically, the all of the potential stems are first determined and the structure is built by adding the next stem compatible with the existing structure that has the largest equilibrium constant.[1] Compatible stems are those whose bases are not already involved in stem formation and will not form a pseudoknot structure. Selection of the stem with the largest equilibrium constant was based on the fact that of the competing stems, this one would predominate.[1] A modification to improve this simple rule involved breaking each stem and allowing the structure to re-form until no

alterations occurred upon perturbation. This was expanded using a Monte Carlo method to select a range of possible stems at each step with equilibrium constants within a certain percentage of the maximum.[1] Further improvements introduced in a program called MONTECARLO included the ability to have stems with bulges and inner loops, limitation of the stem population to the best structures, and the option for reducing the destabilizing effect of hairpin loops.[2]

Without the benefit of sequence comparison in determining RNA secondary structure, energy minimization based algorithms remain the only option for those not intimately involved in the field. The two web-based RNA secondary structure algorithms selected for this analysis rely on different energy minimization strategies. The program maintained on the GeneBee server by the Belozersky Institute (hereafter referred to as GeneBee) can be found at http://www.genebee.msu.su/genebee.html. ***mfold***, designed by Michael Zucker, is located at http://bioinfo.math.rpi.edu/~***mfold***/rna/form1.cgi. These programs were selected over others as they could both accommodate sufficient sequence lengths such that all of the chosen sequences could be analyzed in their entirety. See Table 1 for a comparison of the two programs.

In the absence of a multiple alignment, the algorithm used by GeneBee produces structures through the minimization of the potential energy of the system.[3] The algorithm divides a sequence, S, into subsequences (S_{ij}) and calculates the minimal energy for two cases. W(i, j) defines the minimal free energy of all allowed structures formed from S_{ij}, and V(i, j) the minimum energy if S_i and S_j pair with each other.[4] The second term is set to infinity if S_i and S_j cannot base pair.[4] Starting with sequences of five nucleotides, W(i, j) and V(i, j) are calculated recursively for longer and longer sequences, selecting the optimal structure at each step.[4] The final computation considering the entire sequence is the result. W(i, j) and V(i, j) are calculated based upon the stability of base pairs formed and the resulting structures: hairpin loops, stacking regions, bulge loops, interior loops, and bifurcation loops as defined by Zuker and Steigler (1981).

Mfold uses an algorithm based upon nearest neighbor thermodynamic rules.[5] In this method, free energy is assigned to entire loops rather than single base pairs. To avoid the breakdown of energy rules, base pairs are limited to two nucleotides, hairpin

loops must have greater than three bases, and pseudoknots are forbidden. To meet these criteria, bases are considered accessible from a pair i•j if i<i'(,j)<j and no other pair k•l exists such that i<k<i'(<j')<l<j.[5] A loop closed by i•j is defined as L(i•j).[5] An exterior loop, L_e, is used to define inaccessible bases for a linear RNA sequence.[5] In building the secondary structure, S, of sequence R, the RNA is put into loops. Several classes of loops are defined. ***k***-loops (consisting of k base pairs plus the closing pair), interior loops, and multibranch loops.[5] Free energy is calculated based on the type of loop and the amount of single stranded region within the loop using adjusted sequence-sensitive thermodynamic parameters as described by Mathews et. al.[6] In addition, an asymmetry penalty is introduced for an odd number of single stranded bases. Hairpins are treated as special cases based upon their base pair composition. A minimum free energy, $\Delta G(i, j)$, is computed for every potential base pair, and a dot plot produced of all possible foldings within a defined distance of the minimum free energy. Users can introduce additional constraints by forcing and/or forbidding the pairing of certain bases and/or segments.

METHODS

For this analysis, five different internal ribosome entry sequences (IRESs) from various ssRNA positive-strand viruses were used to examine the accuracy of GeneBee and ***mfold*** secondary structure prediction programs. Viruses use IRES-mediated translation as an alternative to typical eukaryotic cap-mediated translation. IRESs are known to be highly structured and there is strong evidence that their activity is structure dependent.[7] Specifically, it appears that the translation initiation factor requirement for ribosome recruitment is linked to their secondary structure, rather than primary sequence.[7] The IRESs selected ranged from 184 to 820 base pairs beginning at either the extreme 5' end of the virus, at nucleotide 40, or at nucleotide 910. Table 2 contains a summary of viral taxonomy, size, and IRES characteristics for each of the selected viruses. Genomic sequences were obtained from NCBI (http://www. ncbi.nlm.nih.gov). Published structures were obtained from literature sources. For poliovirus, the literature structure was determined through dimethyl sulfate modification and nuclease treatments with consideration of both evolutionary and thermodynamic parameters.[13] Classic

swine fever virus and hepatitis C virus published structures were complied based upon comparative sequence analysis using multiple strains, specific mutations in predicted stem-loops, enzymatic probing, and RNAKNOT thermodynamic analysis to specifically focus on pseudoknot structures.[14, 15, 16] The only literature structure prediction partially dependent on mfold prediction was bovine enterovirus.[9] Phylogenetic sequence comparisons and structural predictions of related viruses were also utilized in the final determination.[17, 18] Comparative sequence analysis, dimethyl sulfate modification, and nuclease treatment data were used to determine the published structure for TMEV.[19]

For both GeneBee and ***mfold*** analysis, the nucleotides corresponding to the IRES sequences were cut and pasted from the genomic sequence. Initially, the default parameters were used for both programs (see Tables 3 and 4). With the ***mfold*** program, all jobs were submitted as batch jobs to allow for the use of the entire polio and BEV IRESs. As the purpose of this analysis is to test the ability of these algorithms to predict the secondary structure for sequences of RNA about which no structural information may exist, the default parameters were the most logical choice. In addition, most of the parameters that may be modified relate to predictions involving multiple alignments. After obtaining the initial data, additional analysis was performed on the BEV IRES using GeneBee and setting the maximum distance between paired bases to 300 bases. In each case, the lowest energy structure was selected, and the stem-loop structures characterized. A stem-loop was defined by the base pair enclosing ≥15 bases. Exceptions to this rule were made for the clover-leaf structure at the 5' end of the polio and BEV IRESs. For this analysis, the following terms were defined to describe the hierarchy of stem-loops: major stem-loop (MSL), internal stem-loop (ISL), secondary internal stem-loop (SISL), and tertiary internal stem-loop (TISL). Figure 1 presents a schematic of the various stem-loop types. To analyze the accuracy of the algorithms, the stem-loops from the published structures were similarly characterized.

RESULTS

Appendices A-E present the published secondary structures for polio, CSFV, BEV, HCV, and TMEV respectively. Graphical depictions of the minimum free energy structures produced by GeneBee and ***mfold*** are similarly depicted in Appendices F-J

and Appendices K-O respectively. Appendix P presents the secondary structure of the BEV IRES as predicted by ***mfold*** with the maximum distance between paired bases set to 300 bases. The stem-loops observed in polio secondary structure are summarized in Table 5. The closing base pair of each stem loop, L(i•J), as well as the size and class are listed. Tables 6-9 present similar data for CSFV, BEV, HCV, and TMEV. To take a broader perspective, the overall degree of secondary structure was analyzed by counting the number of each class of loops for the published, GeneBee, and ***mfold*** structures. This data is presented in Tables 10-14.

By comparing the stem-loops predicted by each algorithm to the expected stem-loops with regards to the closing base pair, an approximate degree of similarity was determined. Due to the relatively inaccurate predictions, the criteria for correspondence were not particularly stringent. The class of stem-loop was not considered as the actual prediction of the formation of the stem-loop was the primary focus. Comparing the stem-loops predicted by each algorithm for polio, four out of the twelve (~33%) expected stem-loops were approximately predicted by both programs. These four loops reside at the extreme 5' end and form a classic clover-leaf structure of one MSL with three ISLs. With regards to CSFV, ***mfold*** predicted stem-loops corresponding to four of the seven expected (~57%), while GeneBee stem-loops did not show correlation with any of the expected. ***Mfold*** analysis resulted in three out of fourteen (~21%) expected stem-loops for BEV, while GeneBee correctly predicted only one (~7%). For HCV, ***mfold*** analysis resulted in one correlating stem-loop out of six (~17%) and GeneBee analysis produced none. All four of the stem-loops expected for TMEV were quite accurately predicted by both programs. As the published structure for TMEV was determined through comparative sequence analysis, base modification, and nuclease susceptibility, this is not due to a biased comparative structure.

DISCUSSION

Neither GeneBee nor ***mfold*** proved a very successful program for predicting the secondary structures of short segments of mRNA. Over prediction was observed using both algorithms. With the exceptions of the GeneBee prediction for BEV and CSFV, the

overall degree of secondary structure as measured by the total number of significant stem-loops was greater than or equal to the expected structure. SISLs and TISLs were found only in the structures predicted by GeneBee and ***mfold***. ***Mfold*** over predicted the degree of secondary structure of all of the IRESs except TMEV (Tables 10-14). GeneBee prediction resulted in over prediction only for TMEV, with a single additional stem-loop (Table 14). Overall, the TMEV IRES was most successfully predicted. ***Mfold*** produced an almost prefect structure, while GeneBee predicted the expected structure correctly but enclosed within an MSL (Table 9). ***Mfold*** also predicted the structure of the CSVF IRES fairly well, with good correspondence between the first two MSLs and two similar stem-loops farther downstream (Table 6). Though the majority of the sequence was incorrectly predicted as compared to the published structure, GeneBee and ***mfold*** predicted the clover-leaf structure of one MSL and three ISLs formed by the first 80 bases of the polio IRES with a high degree of precision (Table 5). For both programs, BEV and HCV yielded the worst predictions (Tables 7 and 8). The poor prediction of BEV by ***mfold*** confirms that the methods used to determine the "gold standard" structures did not significantly bias the comparison, as mfold data partially contributed to the development of the BEV literature structure.

Looking for connections between IRES size and/or position, and the accuracy of the prediction methods, a negative correlation can be drawn between size and prediction ability. TMEV (the shortest IRES) resulted in the best prediction by both programs, and BEV (the longest IRES) was among the worst predicted. HCV (the second shortest IRES) was also poorly predicted, though it is almost twice as long as the TMEV IRES. TMEV and HCV also represent the two interior IRESs. It does not appear, therefore, that interior sequences result in better predictions. This lack of correlation is supported by the almost perfect prediction of the first 80 base pairs of the polio IRES.

Overall, the published structures appeared to have more local base pairing as evidenced by smaller stem-loops and less internal branching. Based upon this observation, an attempt was made to improve the secondary structure prediction for BEV using ***mfold***. This program allows the user to set the maximum distance between base pairs. The stem-loop sizes in the published structure of BEV range from 11 to 205

bases. A maximum distance of 300 bases between pairings was selected as a constraint, and the secondary structure predicted. 300 bases was chosen to prevent complete biasing. The resulting structure contained three MSLs of approximately 260 bases and was actually a poorer prediction than the original structure (Appendix P as compared to Appendix M). Looking at the algorithms, there does not appear to be a strong a priori reason to expect larger stem-loops with more internal structures to be favored. One possibility is that the bigger, more structured stem-loops allow for more regions of base pairing, and thereby decrease the free energy. This, however, results in the formation of many very small hairpin loops within the larger stem-loop that may be unstable. Another factor to consider is the energy of single-stranded regions enclosed within loops versus external bases. Longer stretches of unpaired bases appear in the published structures as compared to those created by either of the two programs. The free energy of external bases is not specifically addressed by the algorithms.

Comparing the two programs, ***mfold*** structure prediction resulted in higher percentages of stem-loops corresponding to the expected structure for all of the IRESs (except TMEV where the GeneBee algorithm erred through the addition of a MSL). The relatively greater success of ***mfold*** is likely due to the more sequence-specific thermodynamic parameters, the asymmetry penalty, and the consideration of the amount of single stranded RNA contained within a loop.

CONCLUSIONS

As evidenced by this analysis, neither of the energy minimization algorithms used by GeneBee or ***mfold*** provides a viable method for predicting the secondary structure of short, highly organized sequences of mRNA. To best suit the needs of researchers desiring secondary structures for potential RNAi targeting regions, several issues must be addressed. Three main problems were brought to light by this analysis: (1) the introduction of large stem-loops due to base pairing between distant bases not seen in the published structures, (2) the inability of these methods to allow for regions of unstructured RNA, and (3) the tendency of these algorithms to over predict the degree of secondary structure.

The first issue was somewhat addressed in the analysis of the BEV IRES with ***mfold*** using 300 bases as the maximum distance allowed between pairs. This attempted fix was unsuccessful in that it merely created large stem-loops of sizes just under 300 bases. In addition, the aim of this exploration was based on the premise that researchers checking potential target sequences would not have the benefit of other structural knowledge, and so could not make an accurate estimate for this parameter. One potential solution would be to introduce penalties for stem-loops over a certain size, and/or that encompassed greater than a certain percentage of the overall sequence. Examination of currently available structures could provide information as to reasonable default values for this parameter.

Allowing for regions of unstructured RNA would be a particularly important feature in a successful algorithm for modeling RNAi targets. Though the sequences analyzed here are known to have high degrees of function-relevant secondary structure, this is not a universal characteristic. Here, it appears that the best solution would be to "break" the energy minimization rule. This could be accomplished by assigning free energy values to unpaired bases in such a way that they were not necessarily disfavored over small regions of distant interactions. A less direct solution would be to increase the penalty of unpaired bases within stem-loops.

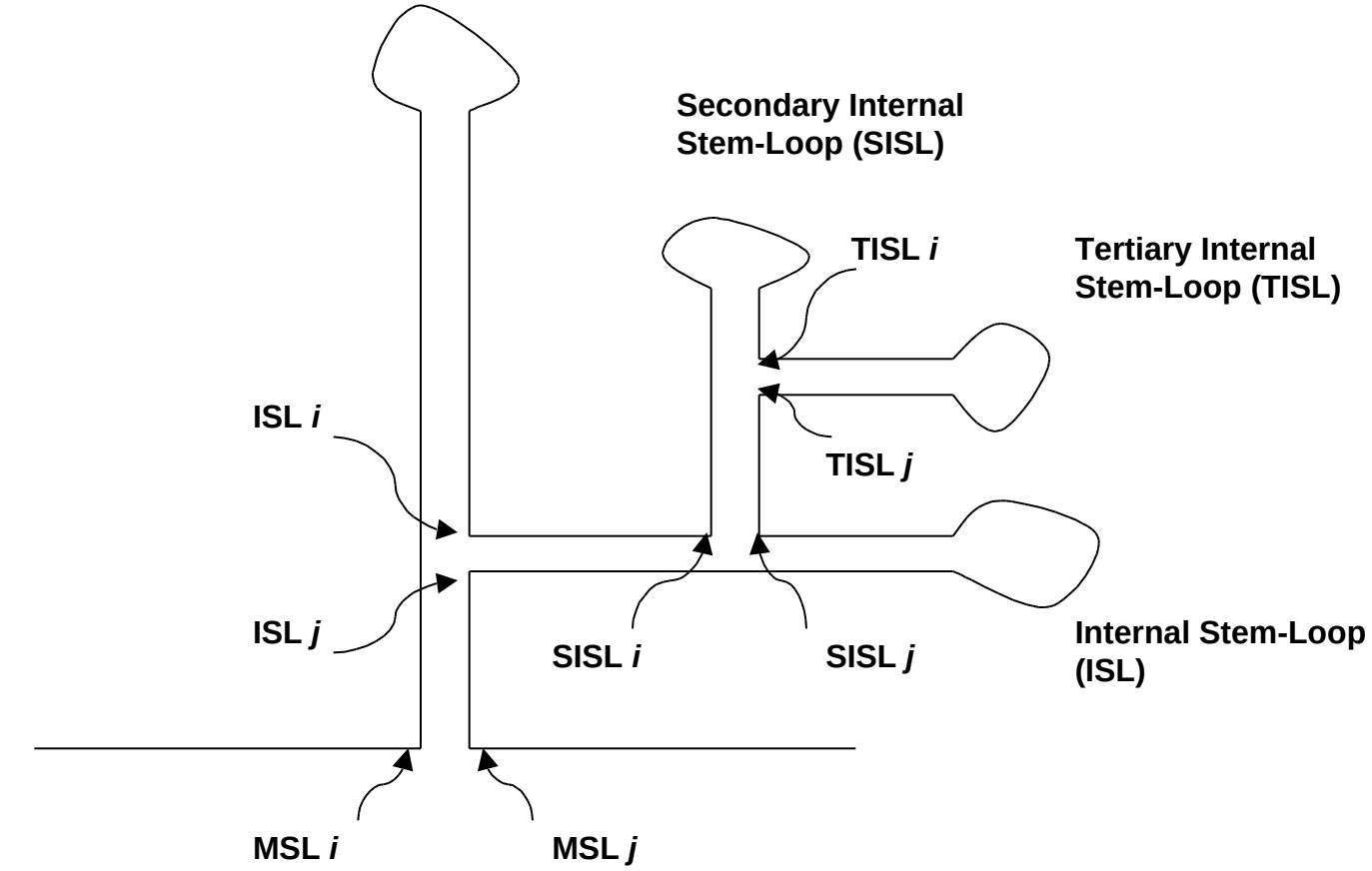
Biologically speaking, a major factor contributing to the poor performance of ***mfold*** and GeneBee is likely the fact that these structures are functionally important. Thermodynamic considerations are therefore not the only driving force in folding. The use of comparative sequence analysis can partially overcome this problem as the conserved regions are likely those that are functionally important. Correcting for the functional significance of the structures without the benefit of related sequences is a major obstacle. Some benefit might be derived from the above suggestions to reduce over prediction. The predicted structures would not be the lowest in energy, but might better reflect biological reality. In the case that related sequences were available, another possible method for overcoming this problem would be to look to see if the targeted region showed any sequence conservation. If none was found, the minimum energy prediction could be used, as functional significance would be unlikely. Other

cases would have to be analyzed more closely to determine the significance of the conserved region.

Considering that the five IRESs examined here are considered to be “highly structured,” the degree of over prediction by these algorithms is particularly concerning. It is possible that this could be remedied by addressing the other two issues. Prohibiting large stem-loops and allowing for unstructured regions of RNA would certainly be expected to reduce the overall degree of secondary structure. In summary, drastic modifications to the current energy minimization algorithms will be necessary before these web-based secondary structure prediction programs can be used to determine the secondary structure of RNAi target sequences.

FIGURES

Figure 1. Stem-Loop Classification



TABLES

Table 1. GeneBee and ***Mfold*** Comparison

	GeneBee	***Mfold***
User Input	Single letter nucleotide sequence	Single letter nucleotide sequence
Algorithm	Potential energy minimization	Nearest neighbor thermodynamic rules
User Defined Parameters (relevant for single sequences)	– Method – Energy threshold for helices – Greedy parameter – Subsection of sequence	– Constraints on base pairing – Linear or Circular – Percent suboptimaility – Upper bound for number of foldings – Maximum distance between paired bases – Window
Sequence Size Limitations	None indicated	- 500 bases for an immediate job - 3000 bases for a batch job

Table 2. Summary of Virus Characteristics

Virus	Abbreviation	Taxonomic Classification	NCBI AC#	Genome Size (bp)	IRES Location
Poliovirus	polio	Picornavirus, enterovirus	V01150.1	7441	1-743[8]
Bovine Enterovirus	BEV	Picornavirus, enterovirus	NC_001859	7414	1-820[9]
Theiler's Murine Encephalomyelitis Virus	TMEV	Picornavirus, cardiovirus	NC_001366	8101	910-1094[10]
Classical Swine Fever Virus	CSFV	Flaviviridae, pestivirus	AF407339	12297	1-374[11]
Hepatitis C Virus	HCV	Flaviviridae, hepacivirus	AF139594	9616	40-372[11, 12]

Table 3. GeneBee Default Parameters

Parameter	Description	Value
Method	greedy or dynamic programming	Greedy
Energy threshold	energy threshold for helices	-4.0 kcal/mol
Conserved factor	“premium” coefficient of increasing free energy for conservativeness of a pair of complementary positions	2 kcal/mol
Compensated factor	same as above, but including compensatory changes in the alignment	4 kcal/mol
Cluster factor	“demanded” coefficient of increasing free energy when grouping a cluster of two local secondary structures	2 kcal/mol
Conservativity	“demanded” percent of correct pairs in the alignment in order to consider two positions “conservative” or “compensative” complementary	0.8
Greedy parameter	number of variants tried for inclusion at each stem	2 kcal/mol
Part of sequence	nucleotides to be structured	1-1000
Treated sequence	number of the sequence to be structured in multiple alignments	1

Table 4. ***Mfold*** Default Parameters

Parameter	Description	Value
Constraint information	forced and/or forbidden base and/or segment pairing	none
Linear/Circular	type of RNA sequence	linear
Percent suboptimality number	consider only foldings this percent from the minimum free energy	5
Upper bound	maximum number of foldings (noted not necessary)	50
Window	number of foldings automatically computed (default based upon sequence length)	default
Max. distance between paired bases	maximum number of bases allowed between pairs	no limit

Table 5. Poliovirus Structure Stem-Loop Comparison*

EXPECTED				GENEBEE				MFOLD			
Closing Base Pair		Size	Class	Closing Base Pair		Size	Class	Closing Base Pair		Size	Class
i	*j*			*i*	*j*			*i*	*j*		
1	88	87	MSL	5	91	86	MSL	2	88	86	MSL
10	30	20	ISL	13	37	24	ISL	10	34	24	ISL
35	45	10	ISL	38	48	10	ISL	35	45	10	ISL
50	80	30	ISL	50	83	33	ISL	53	78	25	ISL
124	162	38	MSL	93	309	216	MSL	101	705	604	MSL
180	220	40	MSL	98	138	40	ISL	120	580	460	ISL
234	440	206	MSL	157	212	55	ISL	148	244	96	SISL
285	311	26	ISL	216	239	23	ISL	210	239	29	TISL
311	375	64	ISL	323	744	421	MSL	248	554	306	SISL
375	394	19	ISL	350	452	102	ISL	305	437	132	TISL
448	556	108	MSL	507	635	128	ISL	445	491	46	TISL
561	625	64	MSL	664	718	54	ISL	580	617	37	ISL
								622	681	59	ISL
								712	737	25	MSL

* Red closing base pairs denote correlation with a predicted stem-loop for both programs. Green closing base pairs denote correlation with a predicted stem-loop for GeneBee. Blue closing base pairs denote correlation with a predicted stem-loop for ***mfold***.

Table 6. Classic Swine Fever Virus Structure Stem-Loop Comparison

EXPECTED				GENEBEE				MFOLD			
Closing Base Pair		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**
i	***j***			***i***	***j***			***i***	***j***		
3	24	21	MSL	6	30	24	MSL	1	29	28	MSL
29	51	22	MSL	69	374	305	MSL	32	57	25	MSL
53	128	75	MSL	72	133	61	ISL	64	333	269	MSL
129	334	205	MSL	134	349	215	ISL	70	129	59	ISL
143	307	164	ISL					134	320	186	ISL
162	181	19	ISL					259	281	22	ISL
255	280	25	ISL					342	371	29	MSL

* Red closing base pairs denote correlation with a predicted stem-loop for both programs. Green closing base pairs denote correlation with a predicted stem-loop for GeneBee. Blue closing base pairs denote correlation with a predicted stem-loop for ***mfold***.

Table 7. Bovine Enterovirus Structure Stem-Loop Comparison

EXPECTED				GENEBEE				MFOLD			
Closing Base Pair		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**
i	***j***			***i***	***j***			***i***	***j***		
1	90	89	MSL	49	233	184	MSL	5	21	16	MSL
10	34	24	ISL	60	73	13	ISL	30	260	230	MSL
34	45	11	ISL	84	103	19	ISL	56	71	15	ISL
45	80	35	ISL	123	211	88	ISL	76	214	138	ISL
110	200	90	MSL	248	372	124	MSL	81	100	19	SISL
120	143	23	ISL	390	774	384	MSL	107	203	96	SISL
156	189	33	ISL	462	500	38	ISL	262	800	538	MSL
215	285	70	MSL	509	580	71	ISL	285	533	248	ISL
291	306	15	MSL	581	612	31	ISL	365	389	24	SISL
315	520	205	MSL	705	753	48	MSL	390	455	65	SISL
362	389	27	ISL					457	472	15	SISL
456	473	17	ISL					536	778	242	ISL
528	639	111	MSL					550	617	67	ISL
663	694	31	MSL					627	676	49	ISL
								691	761	70	ISL

* Red closing base pairs denote correlation with a predicted stem-loop for both programs. Green closing base pairs denote correlation with a predicted stem-loop for GeneBee. Blue closing base pairs denote correlation with a predicted stem-loop for ***mfold***.

Table 8. Hepatitis C Virus Structure Stem-Loop Comparison

EXPECTED				GENEBEE				MFOLD			
Closing Base Pair		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**
i	***j***			***i***	***j***			***i***	***j***		
44	117	73	MSL	52	368	316	MSL	46	368	322	MSL
49	64	15	ISL	68	83	15	ISL	59	329	270	ISL
123	331	208	MSL	120	360	240	ISL	61	121	60	SISL
155	171	16	ISL	139	217	78	SISL	126	324	198	SISL
251	282	31	ISL	240	305	65	SISL	132	300	168	TISL
				338	354	16	SISL	302	317	15	TISL
								334	351	17	ISL

* Red closing base pairs denote correlation with a predicted stem-loop for both programs. Green closing base pairs denote correlation with a predicted stem-loop for GeneBee. Blue closing base pairs denote correlation with a predicted stem-loop for ***mfold***.

Table 9. Theiler's Murine Encephalomyelitis Structure Stem-Loop Comparison

EXPECTED				GENEBEE				MFOLD			
Closing Base Pair		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**	**Closing Base Pair**		**Size**	**Class**
i	***j***			***i***	***j***			***i***	***j***		
910	1024	114	MSL	916	1077	161	MSL	910	1022	112	MSL
920	965	45	ISL	922	1025	103	ISL	925	966	41	ISL
966	1007	41	ISL	928	969	41	SISL	967	1002	35	ISL
1025	1044	19	MSL	971	1004	33	SISL	1025	1045	20	MSL
				1028	1047	19	MSL				

* Red closing base pairs denote correlation with a predicted stem-loop for both programs. Green closing base pairs denote correlation with a predicted stem-loop for GeneBee. Blue closing base pairs denote correlation with a predicted stem-loop for ***mfold***.

Table 10. Degree Secondary Structure Summary: Poliovirus

	Expected	GeneBee	Mfold
# MSL	6	3	3
# ISL	6	9	6
# SISL	0	0	2
# TISL	0	0	3
Total	***12***	***12***	***14***

Table 11. Degree Secondary Structure Summary: Classic Swine Fever Virus

	Expected	GeneBee	Mfold
# MSL	4	2	4
# ISL	3	2	3
# SISL	0	0	0
# TISL	0	0	0
Total	***7***	***4***	***7***

Table 12. Degree Secondary Structure Summary: Bovine Enterovirus

	Expected	GeneBee	Mfold
# MSL	7	4	3
# ISL	7	6	7
# SISL	0	0	5
# TISL	0	0	0
Total	***14***	***10***	***15***

Table 13. Degree Secondary Structure Summary: Hepatitis C Virus

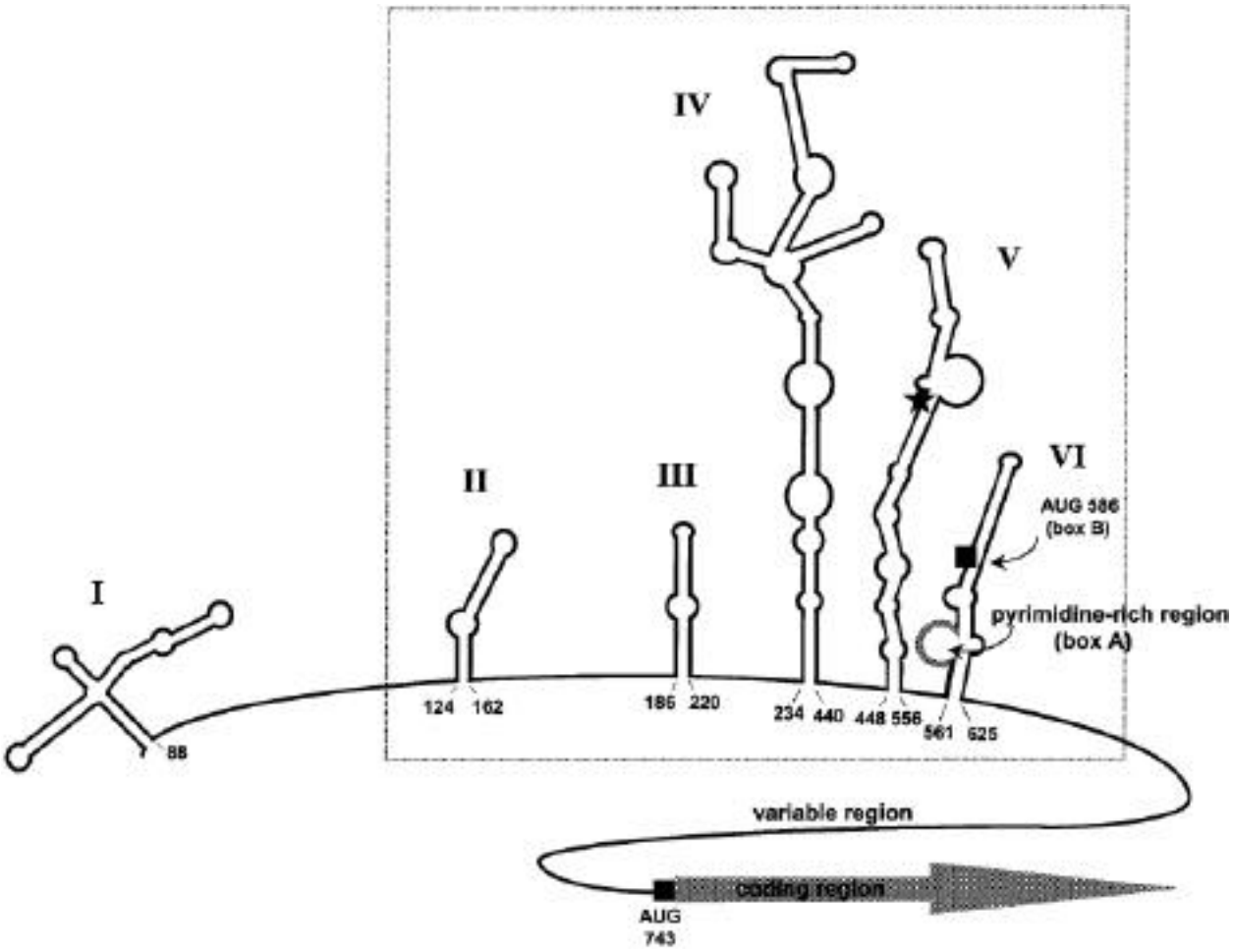
	Expected	GeneBee	Mfold
# MSL	2	1	1
# ISL	3	2	2
# SISL	0	3	2
# TISL	0	0	2
Total	***5***	***6***	***7***

Table 14. Degree Secondary Structure Summary: Theiler's Murine Encephalomyelitis Virus

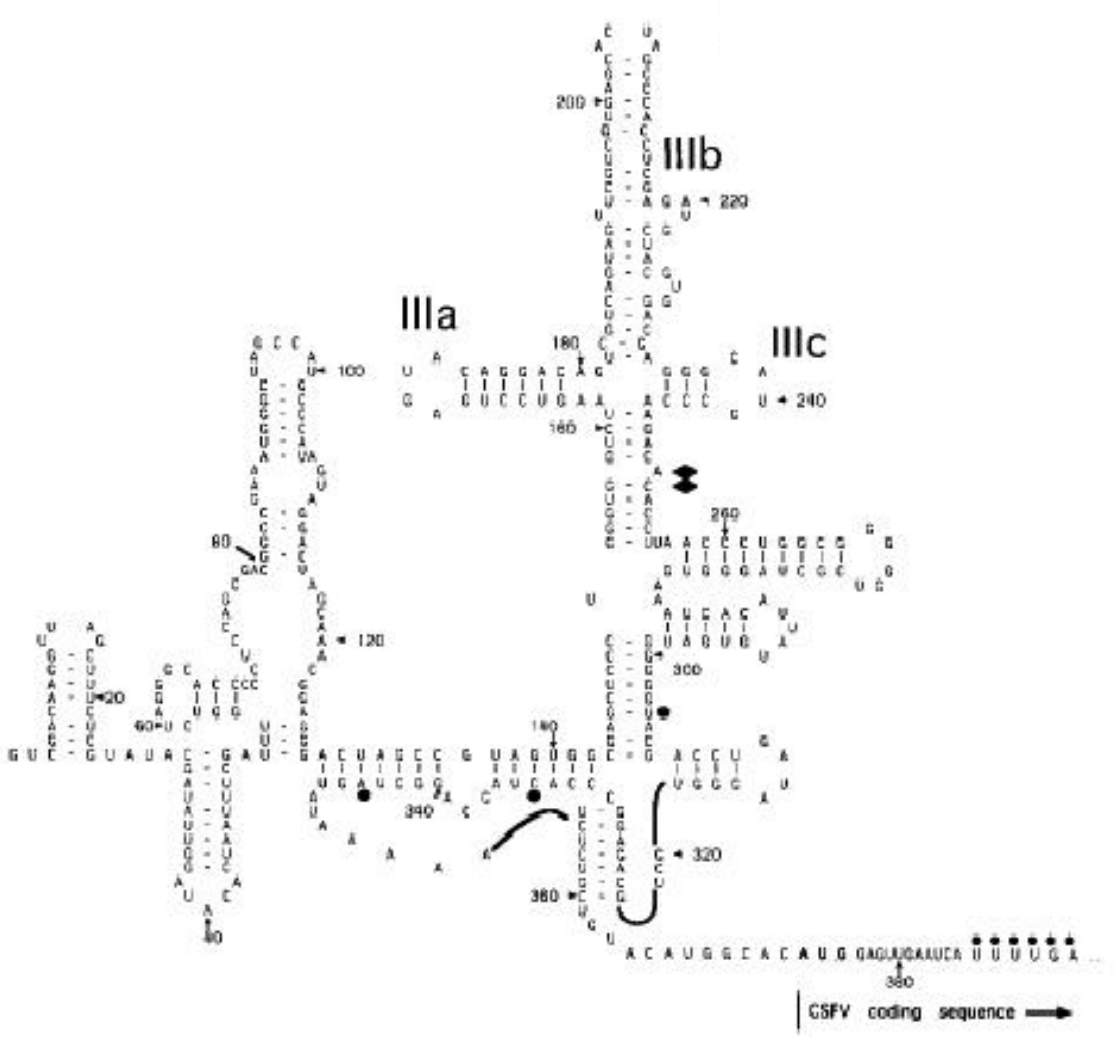
	Expected	GeneBee	Mfold
# MSL	2	2	2
# ISL	2	1	2
# SISL	0	2	0
# TISL	0	0	0
Total	***4***	***5***	***4***

APPENDICES

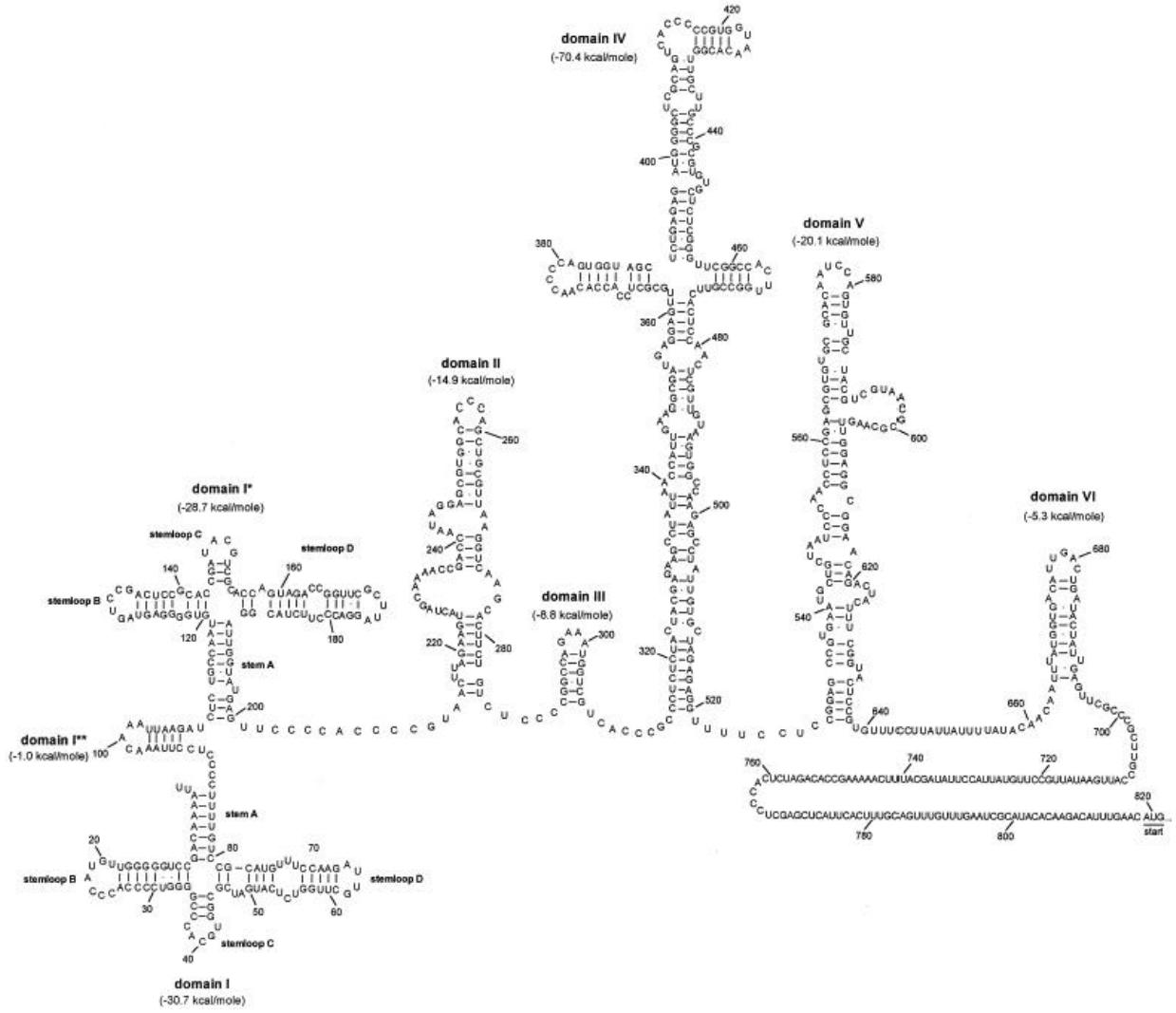
Appendix A. Poliovirus Published Strucutre[8 *]
*base pair numbering obtained from Agol, 1991



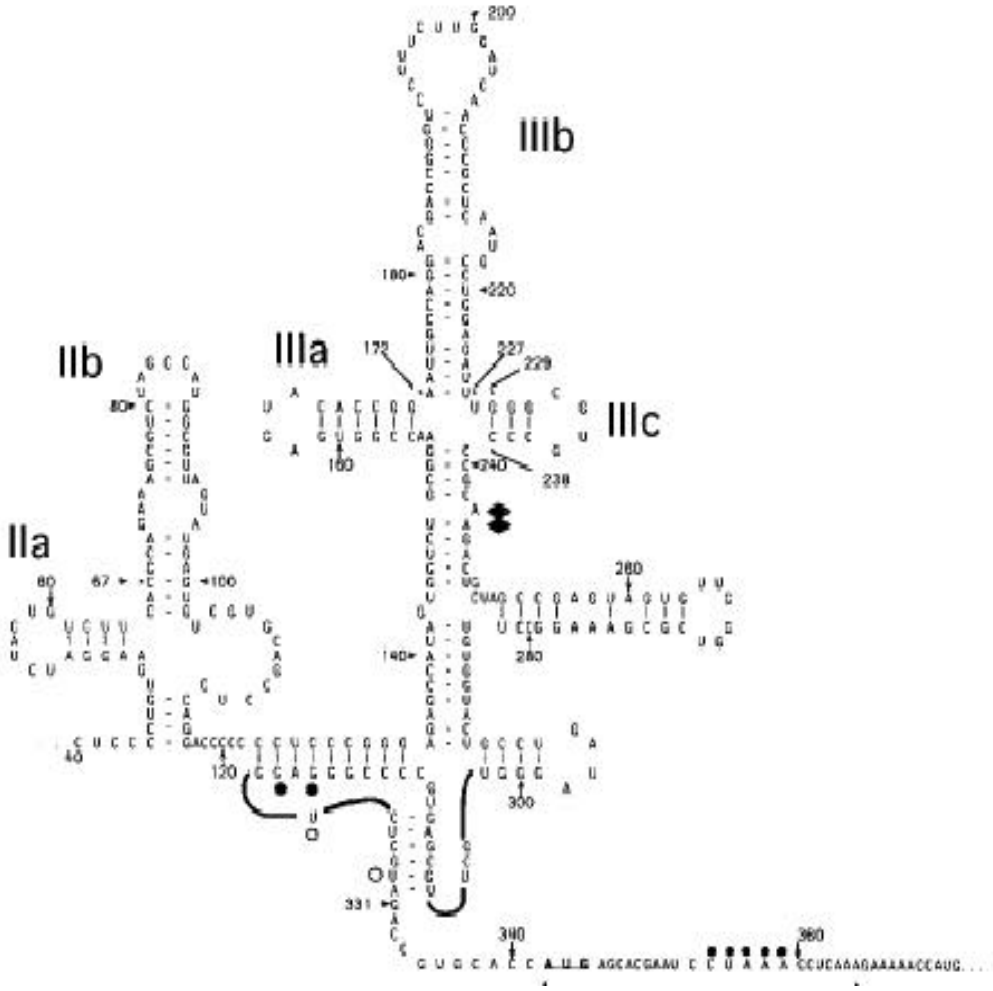
Appendix B. Classic Swine Fever Virus Published Strucutre[11]



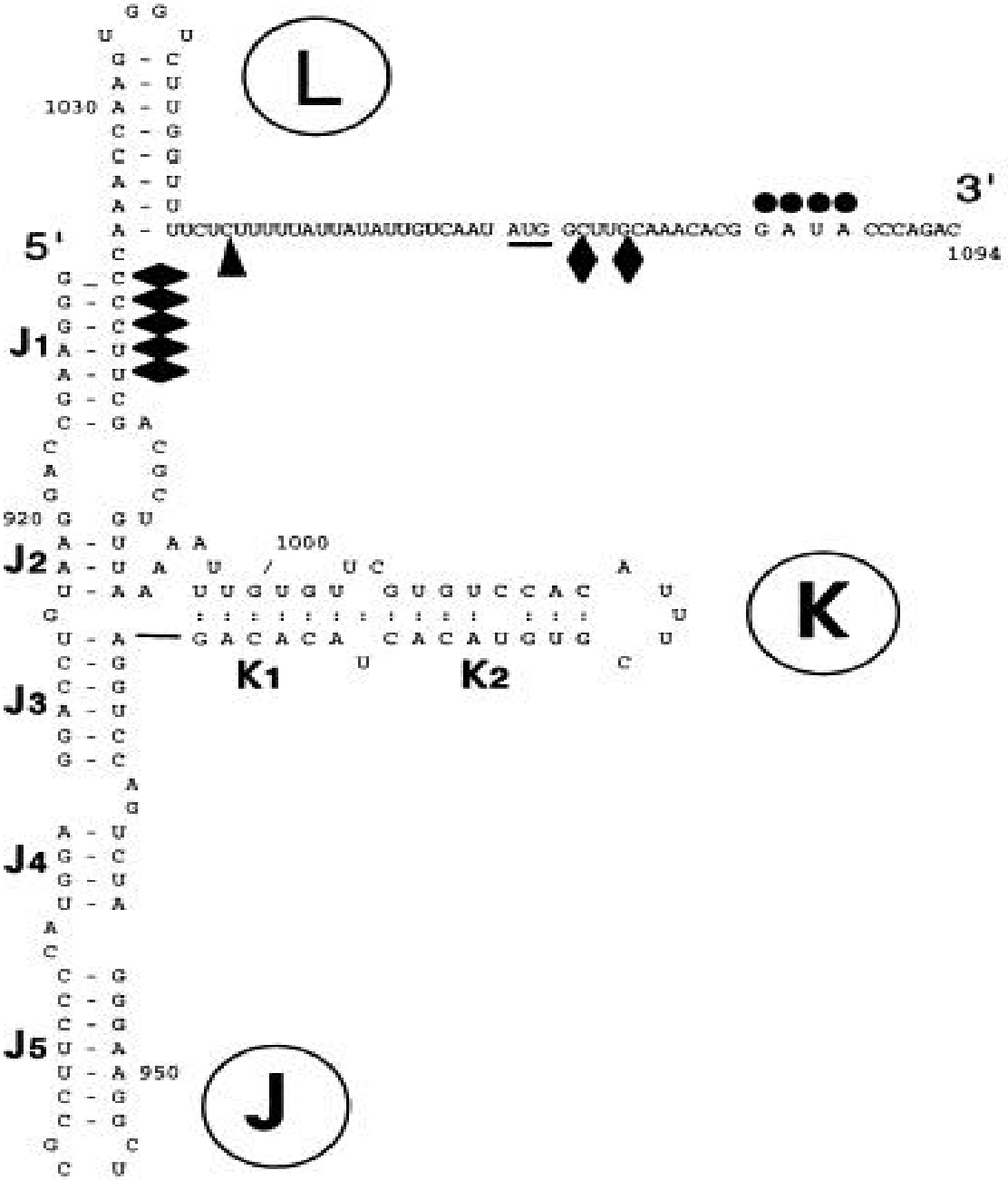
Appendix C. Bovine Enterovirus Published Structure[9]



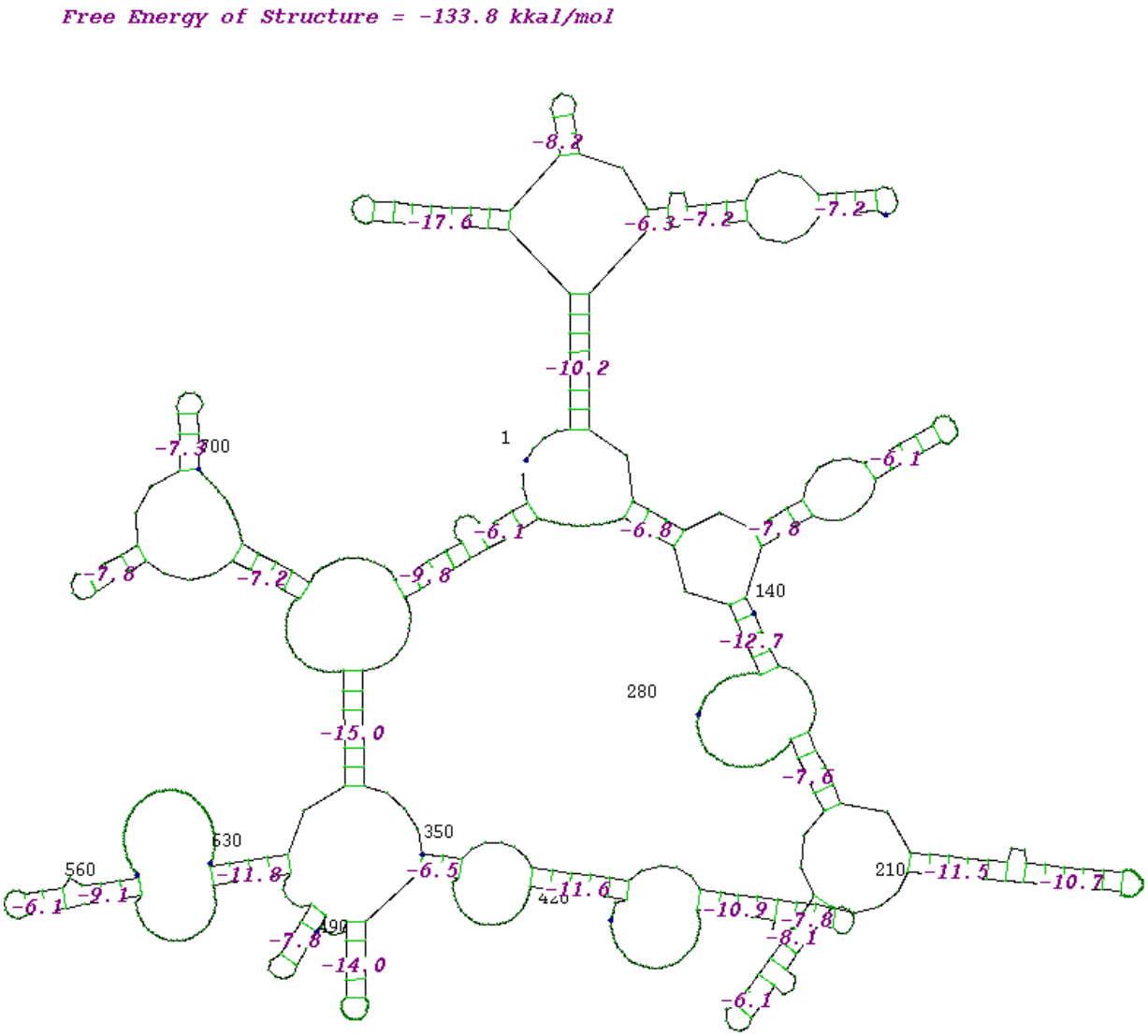
Appendix D. Hepatitis C Virus Published Structure[11]



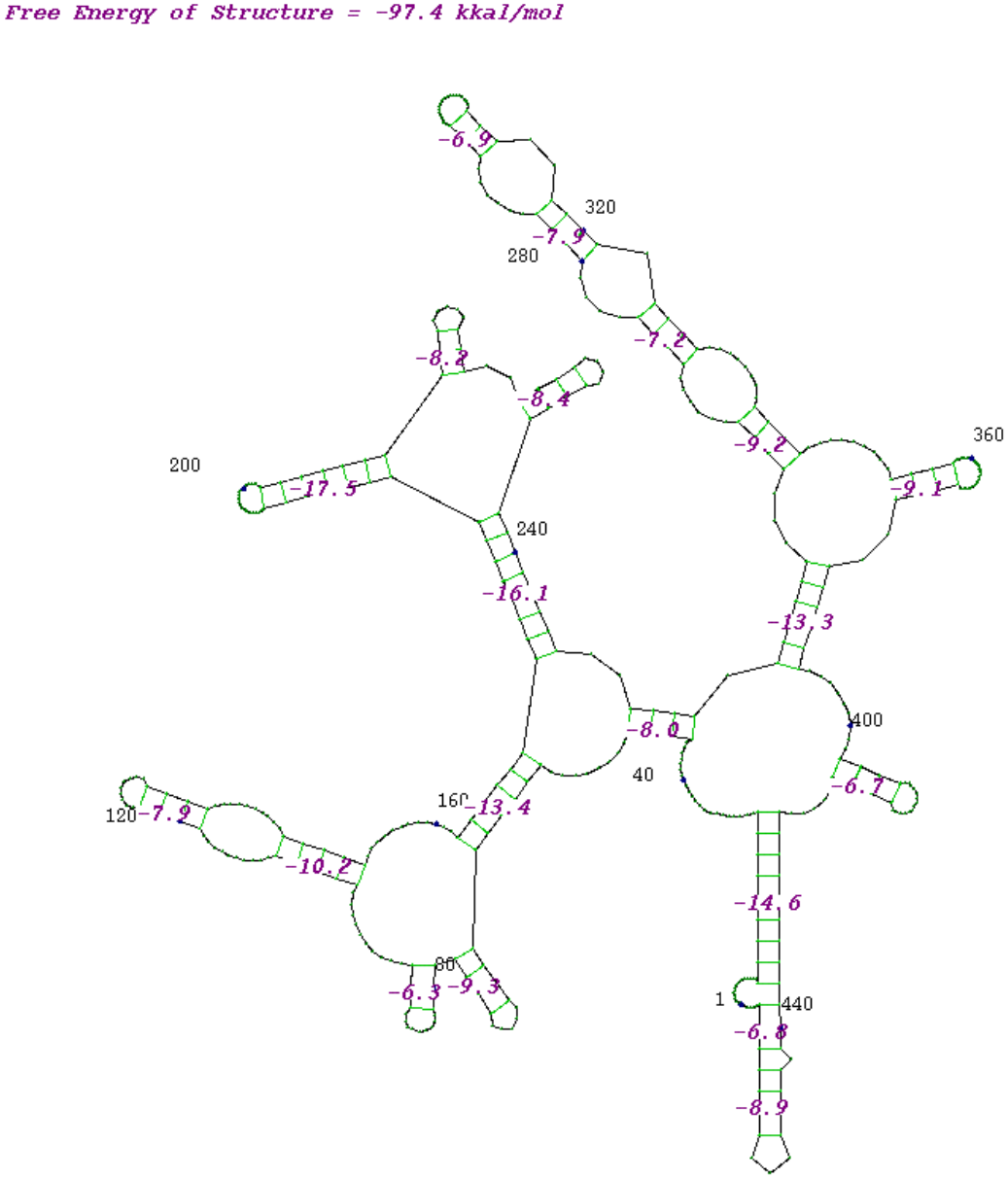
Appendix E. Thelier's Murine Encephalomyelitis Published Structure[10]



Appendix F. Poliovirus GeneBee Predicted Structure

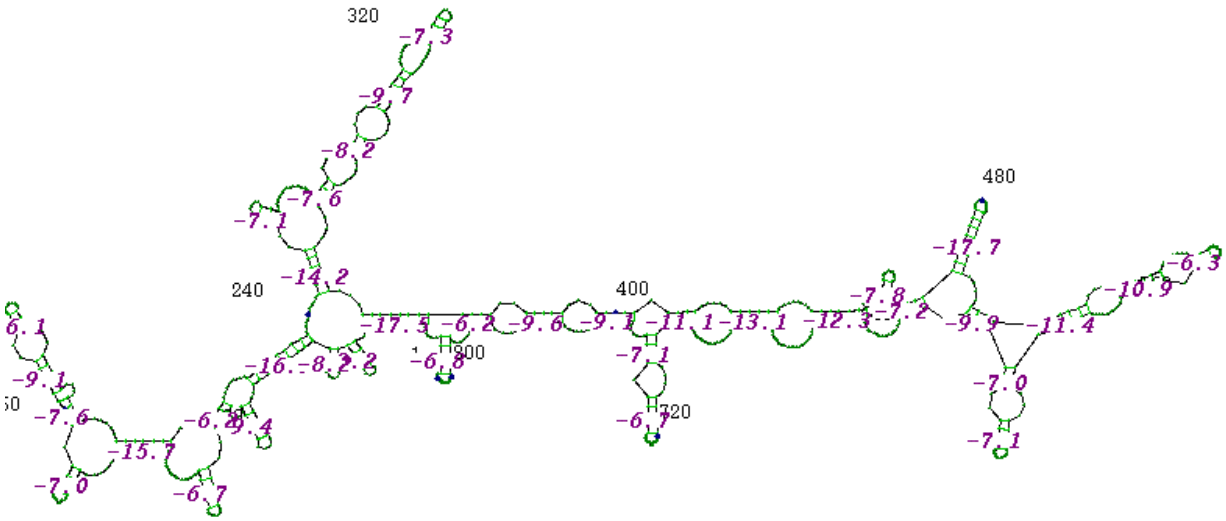


Appendix G. Classic Swine Fever Virus GeneBee Predicted Structure

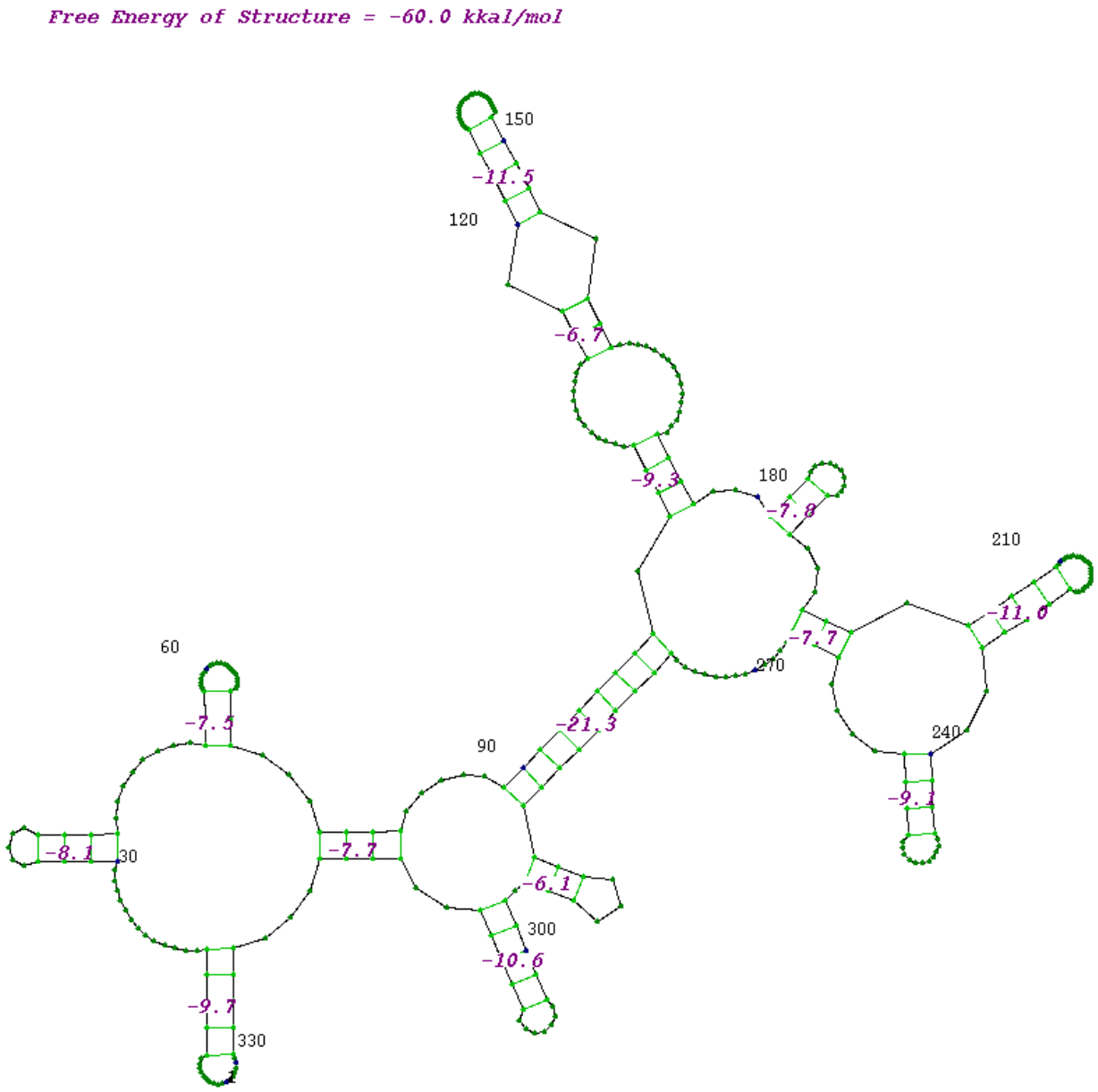


Appendix H. Bovine Enterovirus GeneBee Predicted Structure

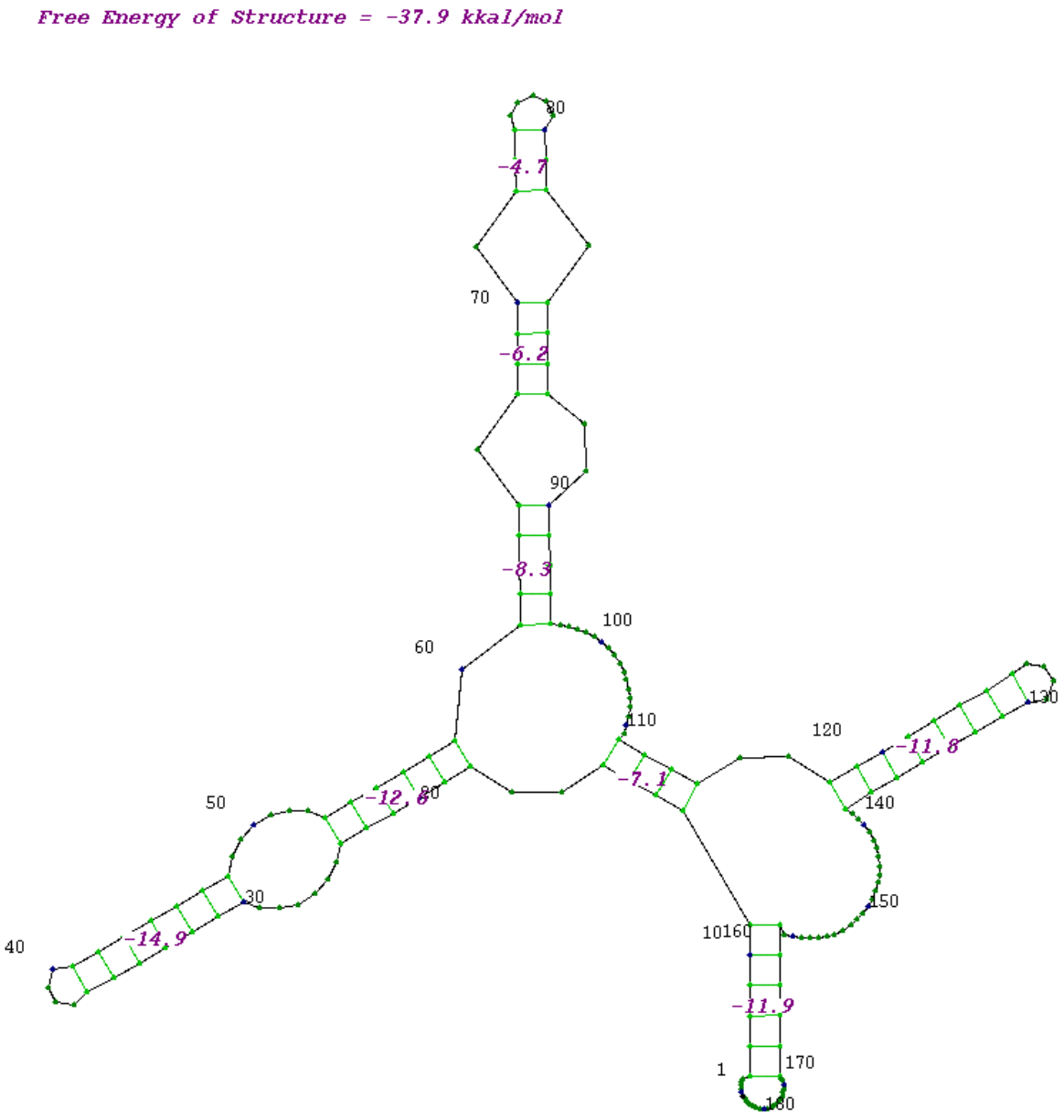
Free Energy of Structure = -183.7 kkal/mol



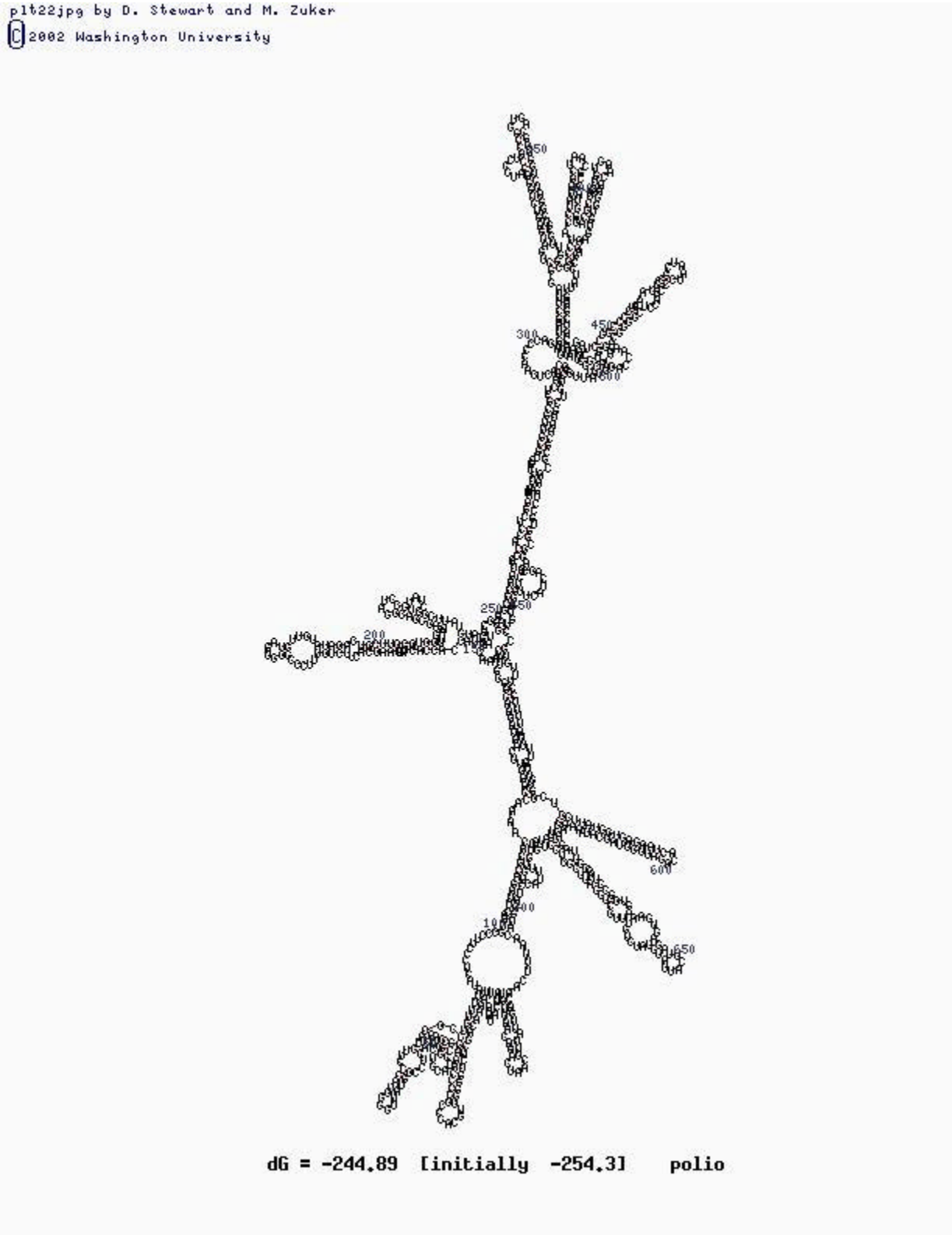
Appendix I. Hepatitis C Virus GeneBee Predicted Structure



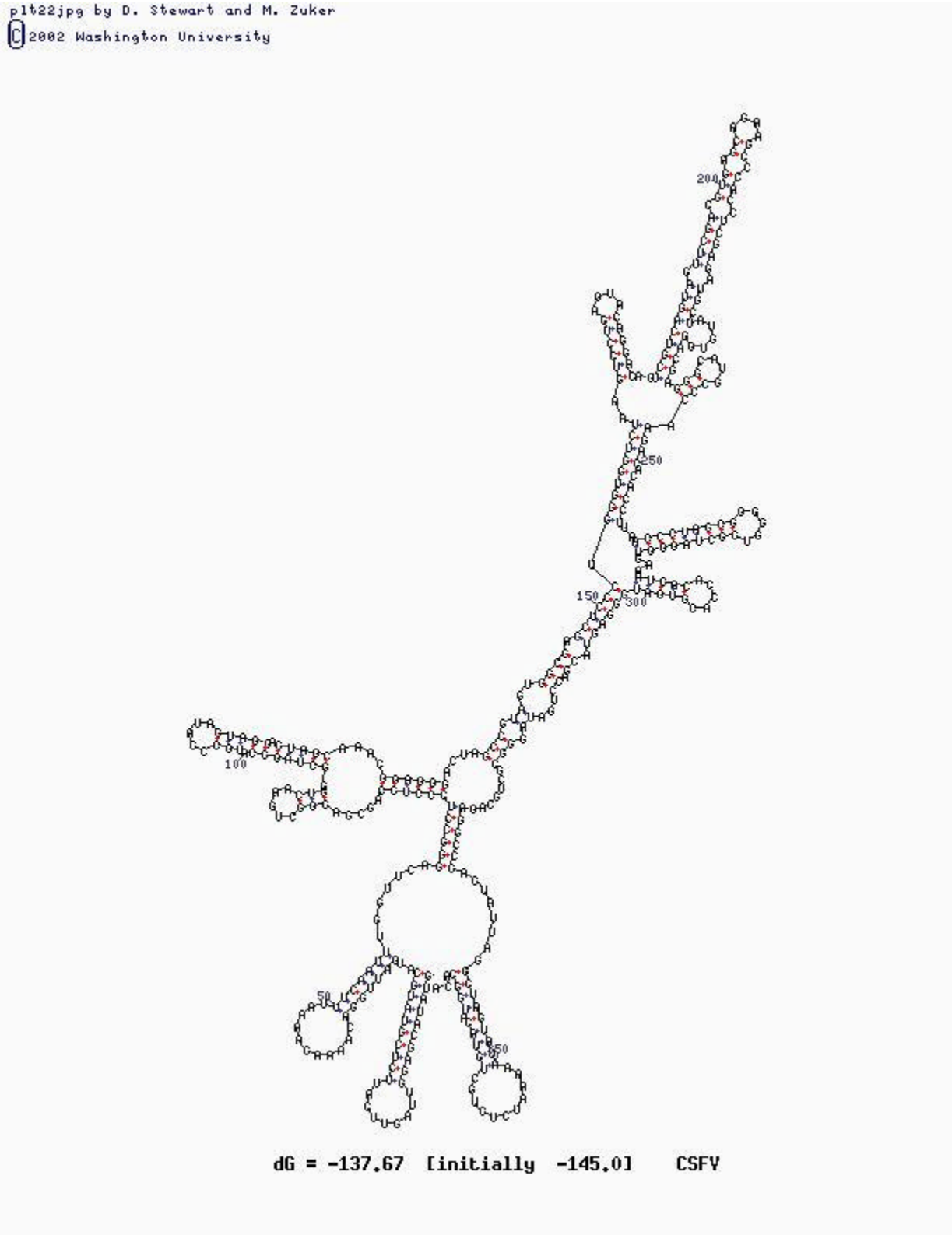
Appendix J. Theiler's Murine Encephalomyelitis Virus GeneBee Predicted Structure



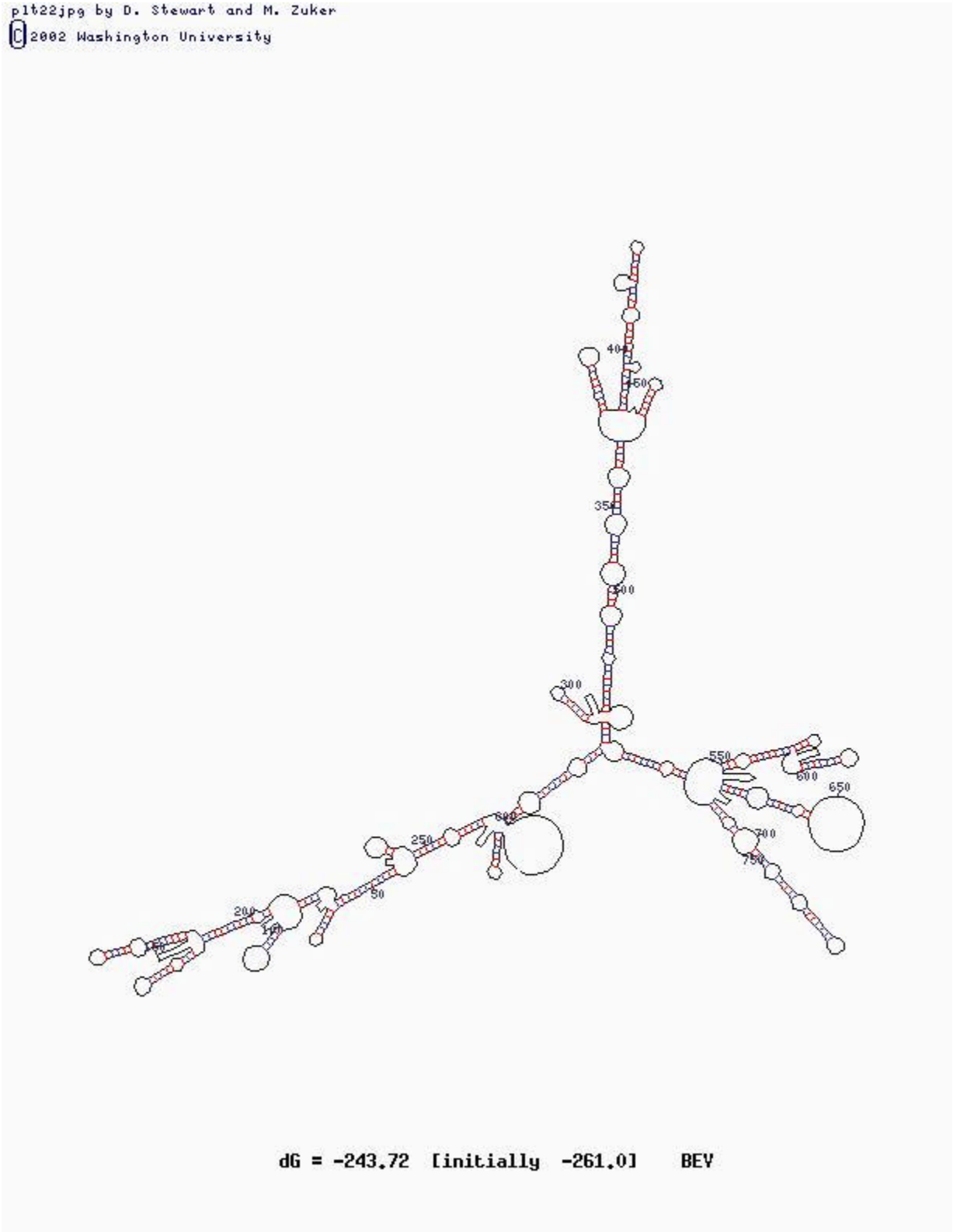
Appendix K. Poliovirus Mfold Predicted Structure



Appendix L. Classic Swine Fever Virus Mfold Predicted Structure

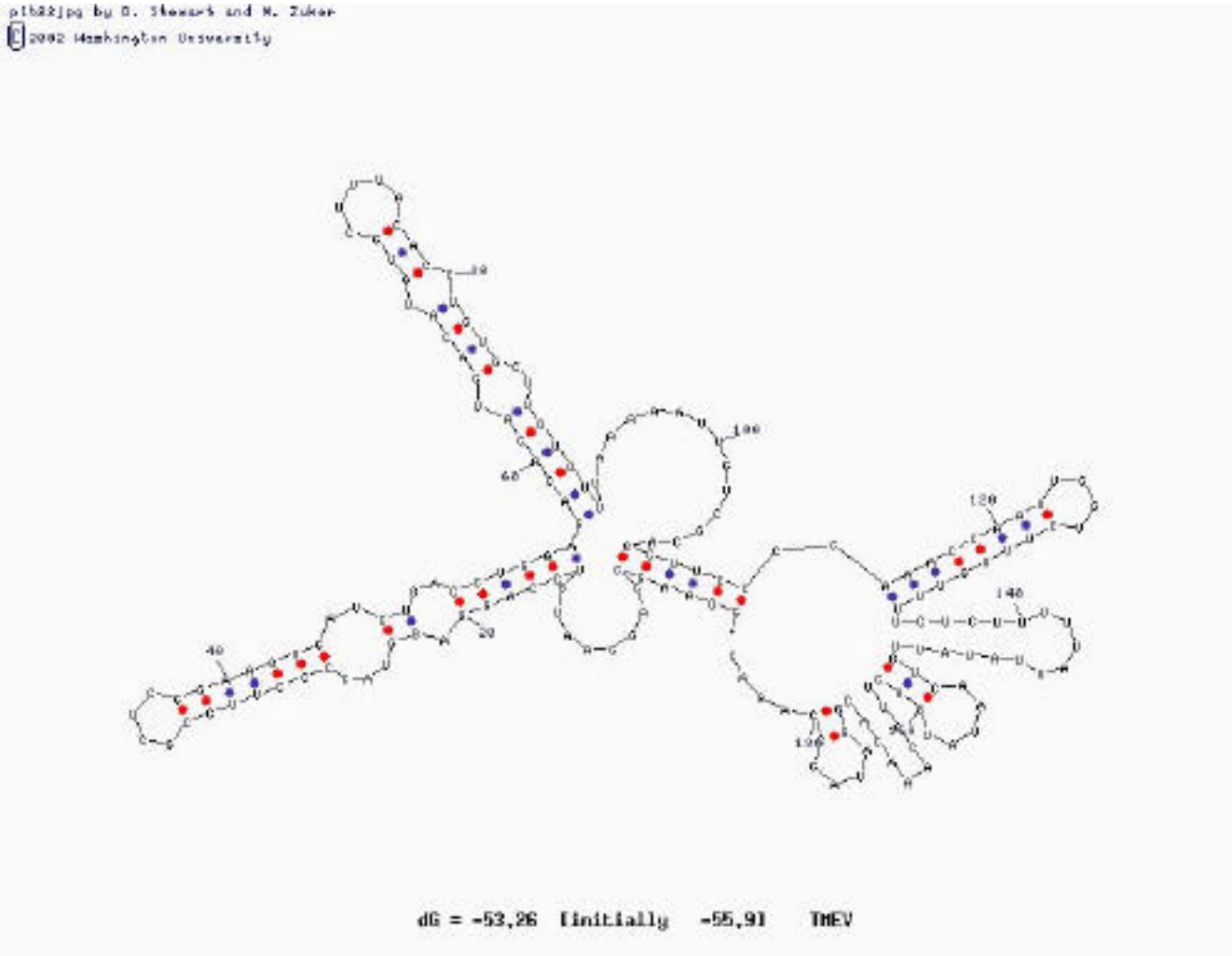


Appendix M. Bovine Enterovirus Mfold Predicted Structure

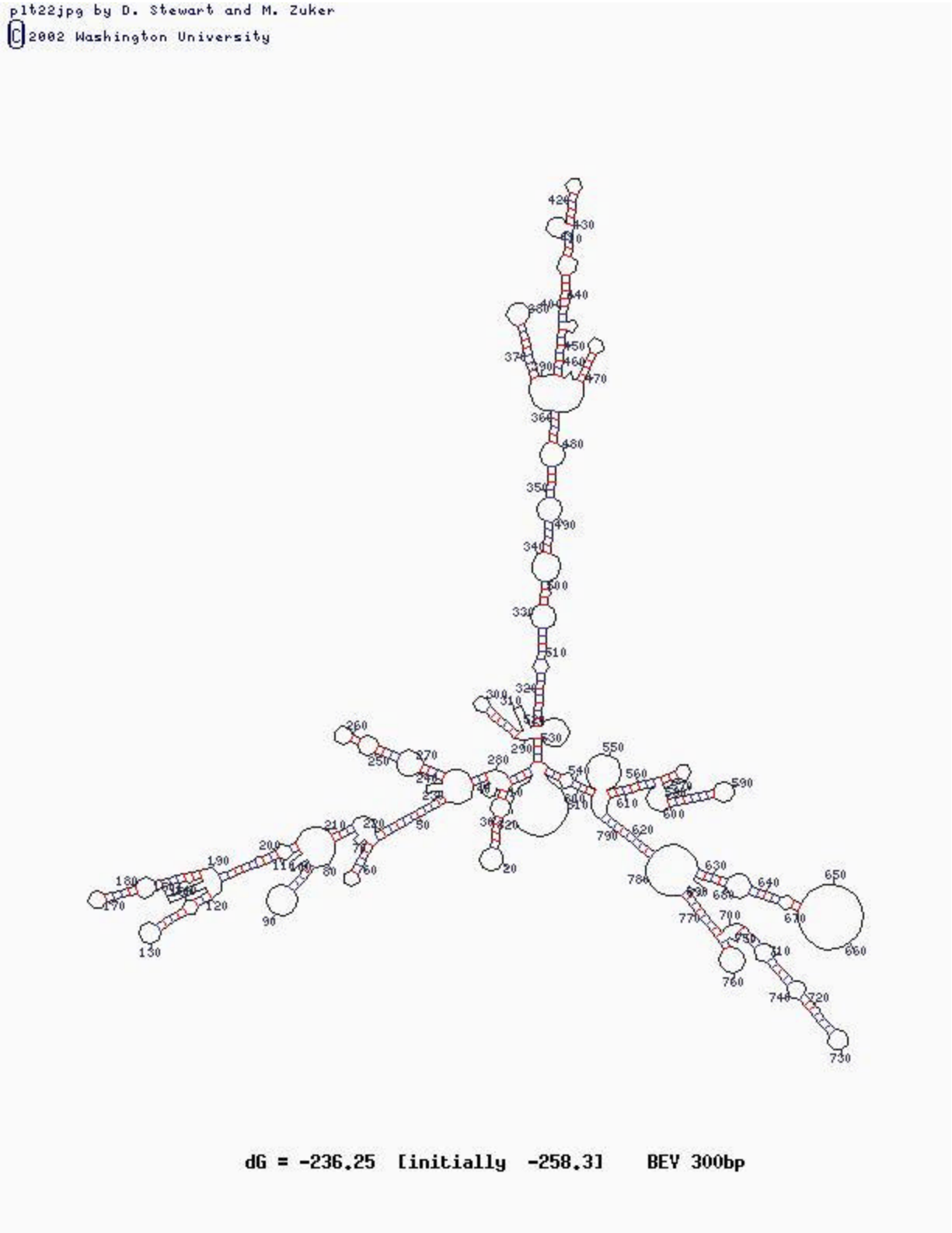


Appendix N. Hepatitis C Virus Mfold Predicted Structure

Appendix O. Theiler's Murine Encephalomyelitis Virus Mfold Predicted Structure



Appendix P. Bovine Enterovirus Mfold Predicted Structure with 300 Base Constraint



REFERENCES

1. Martinez, H.M. (1984) ***Nucleic Acids Research***. **12**: 323-34.
2. Martinez, H.M. (1988) ***Nucleic Acids Research***. **16**: 1789-98.
3. Brodsky, I.L. et. al. (1995) Biochemistry. 8: 923-28.
4. Zuker, M. and Stiegler, P. (1981) ***Nucleic Acids Research***. **9**: 133-48.
5. Zuker, M., Mathews, D.H., and Turner, D.H. In ***RNA Biochemistry and Biotechnology***. J. Barciszewski and B.F.C. Clark eds. Nato ASI Series. Kluwer Academic Publishing, 1999.
6. Mathews, D.H. et. al. (1999) ***Journal of Molecular Biology***. **288**: 911-40.
7. Martinez-Salas, E. et. al. (2001) ***Nucleic Acids Research***. **12**: 323-334.
8. Agol, V.I. (1991) ***Advances in Virus Research***. **40**: 103-80.
9. Zell, R. et. al. (1999) ***Journal of General Virology***. **80**: 2299-2309.
10. Pilipenko, E.V. et. al. (2000) ***Genes and Development***. **14**: 2028-45.
11. Pestova, T.V. et. al. (1998) ***Genes and Development***. **12**: 67-83.
12. Feltcher, S.P. and Jackson, R.J. (2002) ***Journal of Virology***. **76**: 5024-33.
13. Pilipenko, E.V. et al. (1989) ***Virology***. **168**: 201.
14. Brown, et. al. (1992) ***Nucleic Acids Research***. **20**: 5041.
15. Honda et. al. (1996) ***RNA***. **2**: 955.
16. Wang, et. al. (1995) ***RNA***. **1**: 526.
17. Skinner et. al. (1989) ***Journal of Molecular Biology***. **207**: 379-392.
18. Zell, R. and Stelzner. (1997) ***Virus Research***. **51**: 213-229.
19. Pilipenko, E.V. et. al. (1989) ***Nucleic Acids Research***. **17**: 5701-11.