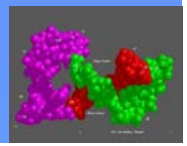
RNA Informatics: Discovery of RNA Structural Features by Computation

Shu-Yun Le
Jih-Hsiang Chen
Jacob V. Maizel, Jr



We know a number of important roles for RNA structure

Catalytic(e.g. ribosomal RNA is added to the list of ribozymes.)

Binding(RNA-protein, RNA-RNA, RNA-DNA and RNA-small molecule, other?) involved in regulation and localization.

Purely structural?

RNA Informatic issues

Some databases: Sequence database
Structural motif database
Functional structural database

Analysis program repositories and servers

Address http://www.imb-jena.de/RNA.html Go Links

Institut für Molekulare Biotechnologie

www.imb-jena.de/RNA.html

The RNA World Website

Databases, Web Tools	Software	Online Books and Tutorials	Meetings	Miscellaneous	Search

Welcome to ***The RNA World Website*** at **IMB Jena**. This web resource lists Internet links on RNA related topics.

- Have a look at a short article describing this site: J. Sühnel, *Trends in Genetics 1997, 13, 206-207*, Views of RNA on the World Wide Web (reprint version in PDF format, PubMed link).
- Read a WebWatch description of this website in *Nature Reviews: Molecular Cell Biology 2002, 3, 3-9*. [WebWatch is on p. 4; PDF].
- The RNA World Website has been included in the Web 's Best Sites collection of the *Encyclopedia Britannica*.

- Breakthrough of the Year: 2002 (20 December 2002 issue of Science)
  - D. Kennedy, Editorial, Science **298**, 2283 (2002)
  - J. Couzin, Breakthrough of the Year: Small RNAs Make Big Slash, Science **298**, 2296 (2002)

Science is making the full text of the online edition of the 20 December 2002 Science, including the Breakthrough of the Year section, available free of charge to all registered users of Science Online.

Databases, Web Tools

Three-dimensional structures (coordinates and images)

- *The Nucleic Acid Database (NDB)*
- *The Protein Data Bank (PDB)*
- *The RiboWeb Project (three-dimensional models of the E. coli 30 S ribosomal subunit and 16 S rRNA)*
- *RNase P 3D models*
- *IMB Jena Image Library of Biological Macromolecules*
  - (with a compilation of all RNA structures from the Protein Data Bank)
- *The RNA Structure Database*
- *SCOR: Structural Classification of RNA*
- *Visualization of Viruses (DNA and RNA)- University of Wisconsin, Madison*
- *Ribosome Images (Wadsworth Center Microscope 3D Database)*
- *Base pairs*
  - *Compilation by Tinoco*
  - *Compilation by Dirheimer et al.*
  - *Database of non-canonical base pairs found in known RNA structures (Fox Lab)*
  - *RNA base pair isostericity (Leontis, Westhof)*
  - *The Base Pair Directory of the IMB Jena Image Library of Biological Macromolecules*

Sequences, Secondary structures, Other

- *5S Ribosomal RNA Database*
- *Database of Ribosomal Crosslinks (DRC)*
- *Ribosomal Database Project II*
- *Ribosomal RNA Mutational Database*
- *European Large Subunit Ribosomal RNA Database*
- *European Small Subunit Ribosomal RNA Database*
- *Ribosomal Internal Spacer Sequence Collection (RISSC)*
- *Comparative RNA Web Site*
  - *Old RNA Secondary Structures Site*

- *tRNA and tRNA Gene Sequences*
- *GtRDB: The Genomic tRNA Database*
- *PLMItRNA: A Database for Plant Mitochondrial tRNA Genes and Molecules*
- *Aminoacyl-tRNA Synthetases Database (AARS)*

- *tmRNA Database*
- *tmRNA Website*

RNABase: The RNA Structure Database - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.rnabase.org/ **www.rnabase.org/** Go Links

RNABase.org

The RNA Structure Database

PDB or NDB ID Code Go

RNABase Home | Search RNABase | Help/About | Contact RNABase

Listing of RNABase Entries | Search RNABase | Analyze Your Structure | RNABase Meta-Analysis | Reference & Education | About RNABase

Listing of RNABase Entries

Complete Listing	A listing of all entries in RNABase with links to detailed records for each entry. All Entries
Technique Listing	A listing of all entries in RNABase by expermental technique. x-ray crystallography - NMR spectroscopy - all other methods
Category Listing	A listing of all entries in RNABase by structural or functional category. transfer RNAs - ribosomal RNAs - messenger RNAs - transcription-related RNAs - introns - splicing-related RNAs - signal recognition particle RNAs - ribozymes - RNase P - aptamers - pseudoknots - tetraloops - bulges - DNA-RNA hybrids - PNA-RNA hybrids - drug-RNA complexes - viral & phage RNAs
Outlier Rate Listing	A tabulation of error rate for each structure in RNABase organized by technique and category. All Entries

Search RNABase

Basic Search	Find the entry you are looking for by PDB or NDB code, author name, classification, experimental technique, resolution, or keywords.
Advanced Search	For those seeking more precise search capabilities.

Analyze Your Structure

Structure Analyzer *Coming Soon!*	Analyze your PDB format structure the same way as official RNABase entries have been analyzed.

RNABase Meta-Analysis

Done Internet

The European ribosomal RNA database - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://oberon.rug.ac.be:8080/rRNA/index.html **Oberon.rug.ac.be:8080/rRNA/indes.html** Go Links

The European Ribosomal RNA database

Department of Plant Systems Biology University of Gent, BELGIUM

- Home
- SSU rRNA
- LSU rRNA
- rRNA BLAST
- Quick phylogeny
- secondary structure
- variability maps
- primers
- software
- Help

Welcome to the European ribosomal RNA database. This database compiles all complete or nearly complete SSU (small subunit) and LSU (large subunit) ribosomal RNA sequences. Sequences are provided in aligned format. The alignment takes into account the secondary structure information derived by comparative sequence analysis of thousands of sequences. Additional information such as literature references, taxonomy, secondary structure modles and nucleotide variability maps, is also available.

The database was founded in the research group of Professor De Wachter at the university of Antwerp (BELGIUM) in 1984. Since 2002, the database is maintained at the university of Gent (BELGIUM) in the research group of Professor Van de Peer.

The database is maintained by Jan Wuyts and Yves Van de Peer.

Latest News:

Old server discontinued

Posted by Jan on Wed May 28 2003

The old web server at the University of Antwerpen has now been discontinued. All traffic to that computer will be redirected here.

Quick phylogeny and rRNA BLAST are back online

Posted by Jan on Tue Nov 19 2002

The Quick phylogeny and rRNA BLAST services were temporarily disabled yesterday because a bug in the scripts caused them to overload our server. This should be fixed now and the services are activated again. My apologies for any

SCOR - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://scor.lbl.gov/index.html **scor.lbl.gov/index.html** Go Links

SCOR

Structural Classification of RNA

History | Previous version SCOR 1.1 | Link to Programs | Help | PDB/NDB list

SCOR 1.2, 5 Jul 2002: 261 PDB entries (2 Oct 2000 and ribosomal RNA 1jj2 & 1j5e).
402 internal loops, 295 external loops.

- **RNA Three-Dimensional Motif Classification**
- **RNA Functional Classification**
- **RNA Tertiary Interactions**
- **Retrieve Structure Information**

PDB or NDB ID:

Sequence:

Key Word(s):

Retrieve Information

Reference: Klosterman PS, Tamura M, Holbrook SR, Brenner SE. 2002. SCOR: a structural classification of RNA database. *Nucleic Acids Res*. 30. 392-394. [pdf]

Synopsis

The SCOR database provides a survey of the three-dimensional motifs, currently internal and external loops, found in a comprehensive collection of NMR and X-ray RNA structures. The motifs are organized in a hierarchical classification; the RNA structures surveyed are also grouped by function. We also provide a classification of the well-characterized tertiary interactions found in the larger RNA structures, with examples. The functional classification and a list of motifs for structures whose motifs have been classified can also be retrieved directly, by entering the PDB identifier for the structure and clicking on Retrieve Information.

Gutell Lab Comparative RNA Web Site - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.rna.icmb.utexas.edu/ **www.rna.icmb.utexas.edu**

The Comparative RNA Web Site The Gutell Lab at The Institute for Cellular and Molecular Biology and The Section of Integrative	1. Comparative Structure Models (CSM)	2. Nucleotide Frequency and Conservation Information	3. Sequence and Structure Data	4. Data Access Systems

Comparative RNA Web Site - Table of Contents

Last updated on 04 April 2003.

Username:

Password:

Log In

Create a New Account
Account Management
(Lost Passwords)

Have rRNA, Will Travel

We are aware that some users have had trouble accessing parts of the CRW Site using the Microsoft Internet Explorer browser. We are exploring several possible solutions to these problems. We would recommend using alternate browsers in the meantime; we also appreciate bug reports to **web@www.rna.icmb.utexas.edu**.

It is appropriate and it is appreciated when you **cite** this work when information from this page benefits your research.

The link "The Comparative RNA Web Site" at the upper left of the screen will bring you back to this "Table of Contents" page at any time.

Number Codes for the Nucleotide Frequency Section: **1**, Single Position; **2**, Base Pairs (Model); **3**, Base Pairs (Tentative); **4**, Base Pairs (Lousy); **5**, Base Triples (Model); **6**, Base Triples (Tentative).

Information		Molecule					
		Ribosomal RNA (P)			Introns (P)		tRNA
Description	Highlights	5S	16S	23S	GrpI	GrpII	
Citation							
1) Comparative Structure Models (CSM)		5S	16S	23S	GrpI	GrpII	tRNA
A. Current Structure Models for Reference Organisms	(H)	X	X	X	X	X	X
B. Evolution of the							

http://www.rna.icmb.utexas.edu/FRAMES/csmframe.html Internet

WelcomeUTR - Microsoft Internet Explorer

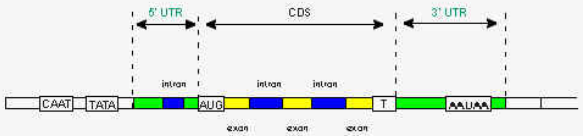
File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://bighost.area.ba.cnr.it/BIG/UTRHome/ **bighost.area.ba.cnr.it/BIG/UTRHome** Go Links

UTR Home Page

Internet Resources for sequence analysis of 5' and 3' Untranslated Regions of Eukaryotic mRNAs



Understanding the basic mechanisms of cell growth, differentiation and response to environmental stimuli, i.e. the program controlling the temporal and spatial order of molecular events, is becoming a real challenge in molecular biology. Indeed, although most of the regulatory elements are thought to be embedded in the non-coding part of the genomes, nucleotide databases are biased by the presence of expressed sequences mostly corresponding to the protein coding portion of the genes.

Among non-coding regions, the 5' and 3' untranslated regions (5'-UTR and 3'-UTR) of eukaryotic mRNAs have often been experimentally demonstrated to contain sequence elements crucial for many aspects of gene regulation and expression .

Dowonload
The collections UTRdb and UTRSite are available from : ftp://bighost.area.ba.cnr.it/pub/Embnet/Database/UTR/

Reference:
Pesole G., Liuni S., Grillo G., Licciulli F., Mignone F., Gissi C., and Saccone C. - "*UTRdb and UTRsite: specialized database of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs.Update 2002*". Nucleic Acids Res (2002), 30(1):335-340 (Full-text PDF)

Internet Resources for UTR analysis

SRS Retrieval of Specialized Databases

UTRdb	A specialized sequence collection, deprived from redundancy, of 5' and 3' UTR sequences from eukaryotic mRNAs
UTRSite	A collection of functional sequence patterns located in 5' or 3' UTR sequences

Sequences Analysis Tools

UTRScan	Looks for UTR functional elements by searching through user submitted query sequences for the patterns defined in the UTRsite collection
UTRBlast	Searches for similarity between a query sequence and 5' or 3' UTR sequences in UTRdb collections

Related Links:

Rfam : Home (Saint Louis) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://rfam.wustl.edu/ **Rfam.wustl.edu** Go Links

Washington University in St.Louis

Rfam :: Home

The Rfam database of RNA alignments and CMs

Rfam (St. Louis) | Rfam (Cambridge) | Infernal | WashU/Genetics
Home | Sequence search | Browse Rfam | Keyword search | Taxonomy search | Help

Rfam 4.0 (May 2003, 114 models)

Rfam is a large collection of multiple sequence alignments and covariance models covering many common non-coding RNA families.

Rfam makes use of a large amount of available data, especially published multiple sequence alignments, and repackages these data in a single searchable and sustainable resource. We have made every effort to credit individual sources on family pages, and have compiled a list of links to these sources here. If you find any of the data presented here useful, please also be sure to credit the primary source.

For more information on using this site, click here.

SEQUENCE SEARCH	Analyze a DNA query sequence to find Rfam family matches. (Links to Sanger Centre)
BROWSE RFAM	View Rfam annotation and alignments.
KEYWORD SEARCH	Query Rfam by keywords.
TAXONOMY SEARCH	Search for Rfam families by taxonomy.

More information and help on Rfam

FTP access to Rfam

Rfam alignments, CMs, and other files are also available by FTP.
Most large files are compressed with GNU gzip (suffix .gz).

- Browse the Rfam ftp site (St. Louis)
- README
- USERMAN - a description of the Rfam flatfile formats.
- Rfam.seed.gz - The seed alignments of the curated families.
- Rfam.full.gz - The full alignments of the curated families.
- Rfam.tar.gz - CMs for all families.
- Rfamseq.lst.gz - List of EMBL 74 sequence accessions covered by Rfam.
- Rfam.fasta.gz - FASTA format file of EMBL 74 sequence covered by Rfam, filtered to <90% identity

114 families

Washington University in St.Louis

rfam.wustl.edu

Rfam :: Browse Families

Browse available alignments and models

Rfam (St. Louis) | Rfam (Cambridge) | Infernal | WashU/Genetics
Home | Sequence search | Browse Rfam | Keyword search | Taxonomy search | Help

Browse Rfam

Name	Accession	No.seed	No.full	Av.len	Av.%id	Description
5S_rRNA	RF00001	606	4371	115	59%	5S ribosomal RNA
5_8S_rRNA	RF00002	64	40995	100	82%	5.8S ribosomal RNA
6S	RF00013	7	21	170	73%	6S / SsrS RNA
7SK	RF00100	5	229	326	70%	7SK RNA
CD_sno_Pyro	RF00095	38	57	56	58%	Pyrococcus C/D box small nucleolar RNA
CopA	RF00042	17	32	90	83%	CopA-like RNA
CRE	RF00048	56	105	60	81%	Enterovirus cis-acting replication element
CsrB	RF00018	6	15	351	76%	CsrB/RsmB RNA family
DicF	RF00039	6	19	52	82%	DicF RNA
DsrA	RF00014	3	12	86	94%	DsrA RNA
E2	RF00091	3	9	154	84%	E2 small nucleolar RNA
E3	RF00092	9	26	126	74%	E3 small nucleolar RNA
Entero_OriR	RF00041	35	84	71	77%	Enteroviral 3' UTR element
FinP	RF00107	6	12	77	92%	FinP
GcvB	RF00022	3	10	111	89%	GcvB RNA
HACA_sno_Snake	RF00098	22	48	140	91%	Snake H/ACA box small nucleolar RNA
Hammerhead	RF00008	68	313	35	68%	Hammerhead ribozyme
HBII-52	RF00105	23	122	81	93%	HBII-52 small nucleolar RNA
HBII-85	RF00108	7	72	93	86%	HBII-85 small nucleolar RNA
HCV_IRES	RF00061	18	815	299	91%	Hepatitis C virus internal ribosome entry site
HDV_ribozyme	RF00094	15	42	88	95%	Hepatitis delta virus ribozyme
HgcC	RF00062	22	39	169	57%	HgcC family RNA
HgcE	RF00060	4	5	111	90%	HgcE RNA
HgcF	RF00058	4	4	168	95%	HgcF RNA
HgcG	RF00064	3	4	225	90%	HgcG RNA
Histone3	RF00032	66	917	26	78%	Histone 3' UTR stem-loop
Intron_gpI	RF00028	30	2747	312	40%	Group I catalytic intron
Intron_gpII	RF00029	116	4875	75	56%	Group II catalytic intron
IRE	RF00037	41	175	29	68%	Iron response element
let-7	RF00027	12	28	79	69%	let-7 microRNA precursor

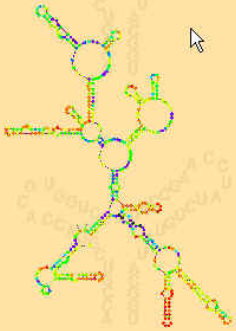
Algorithms, thermodynamics and databases for RNA secondary structure. - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

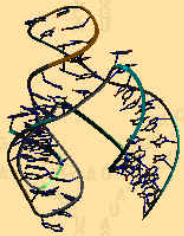
Address http://www.bioinfo.rpi.edu/~zukerm/rna/ **www.bioinfo.rpi.edu/~zukerm/rna/** Go Links

Welcome / ברוכים הבאים fcjake6.ncifcrf.gov | Dr. Zuker's Home Page | Leave Comments | View Guestbook | E-mail



The Zuker Group

Algorithms, Thermodynamics and databases for RNA Secondary Structure.



Left: A color annotated secondary structure prediction

Right: A 3D view of the *hammerhead rybozyme* by Pley *et al*. Click on the image for details.

This page was featured in NetWatch. (*Science* 287, #5450, pg. 7, 2000)

5′ 3′

Zuker Group: PI - Michael Zuker
Post-doctoral Research Associate - Roumen Dimitrov
Students - Nick Markham, Jalagani Srineel
Systems administrator - Alex Yu Web server development - Georgi Shablovsky
Factotum - Melissa Strope

WWW *mfold* servers: Fold your RNA (or DNA) sequences on our server.

- RNA folding server
- DNA folding server
- *Quikfold* server: Fold many short sequences at once.
- *efn* server: Enter a sequence and secondary structure and compute its free energy.
- View folding results for fcjake6.ncifcrf.gov: RNA & DNA, Quikfold

- RNA related publications
- Lectures on RNA Secondary Structure Prediction (incomplete)
- Thermodynamic considerations & rules for RNA folding
- RNA folding and search software
- Three dimensional RNA modeling
- RNA secondary structure display software
- RNA Databases
- The RNA World at IMB Jena
- Computational Structural Biology at the NIH
- Books, Journals, Societies & Meetings

Internet

RNA2D Web Interface - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://protein3d.ncifcrf.gov/shuyun/rna2d.html **protein3d.ncifcrf.gov/shuyun/rna2d.html** Go Links

RNA2D Web Interface

RNA2D: A suite of programs for discovering structural features in RNAs. The program EDscan and SigED are used to search for well-ordered folding segments in a genomic sequence. The program SigStb and SegFold are used to search for unusual folding regions in RNAs. Using the three programs we intend to discover local RNA elements with potentially structure-dependent functions, and to correlate them with known experimental properties and to suggest candidate sites for further experimental study. The program EFFold is used to predict a set of RNA secondary structures that are nearly as good by simulated energy rules and using dynamic programming algorithm. RNAGA employed a genetic algorithm to predict common secondary structures from a set of homologous RNA sequences without the need for their pre-aligned multiple sequences.

- Program SigED
- Program EDscan
- Program SigStb
- Program SegFold
- Program EFFold
- Program RNAGA
- References

- **RNA2D Server for Online Users**

Back to S.Y. Le's Home Page

Back to LECB

Please send questions or suggestions to shuyun@ncifcrf.gov
January 16, 2003

Our goal is to predict structure from sequence

Primary sequence

Secondary structure

Tertiary structure(e.g. pseudoknots, etc.)

Ultimately atomic scale models

Some Rules for RNA Folding (severely simplified)

Single stranded nucleic acids can fold back on themselves to form regions of typical duplex structure(called “stems”)

Watson-Crick rules: A:U, G:C, (G:U -”wobble”)are favorable

Thermodynamic refinements:

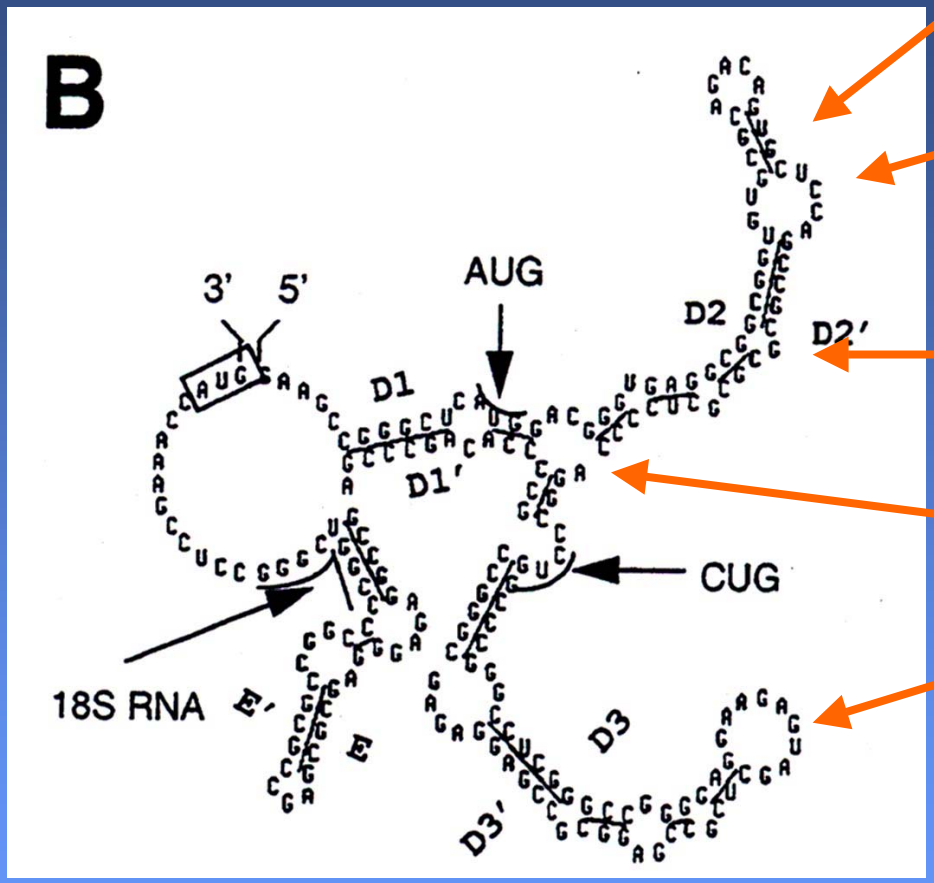
benefits for helical stacking

penalties for loops (hairpin, internal, and multi-branched).

Rules are undergoing long term, gradual refinement so that they can now correctly predict “most” of the base pairs observed in known structures.

BUT, they are not 100% correct.

An example of RNA secondary structure



stem

internal

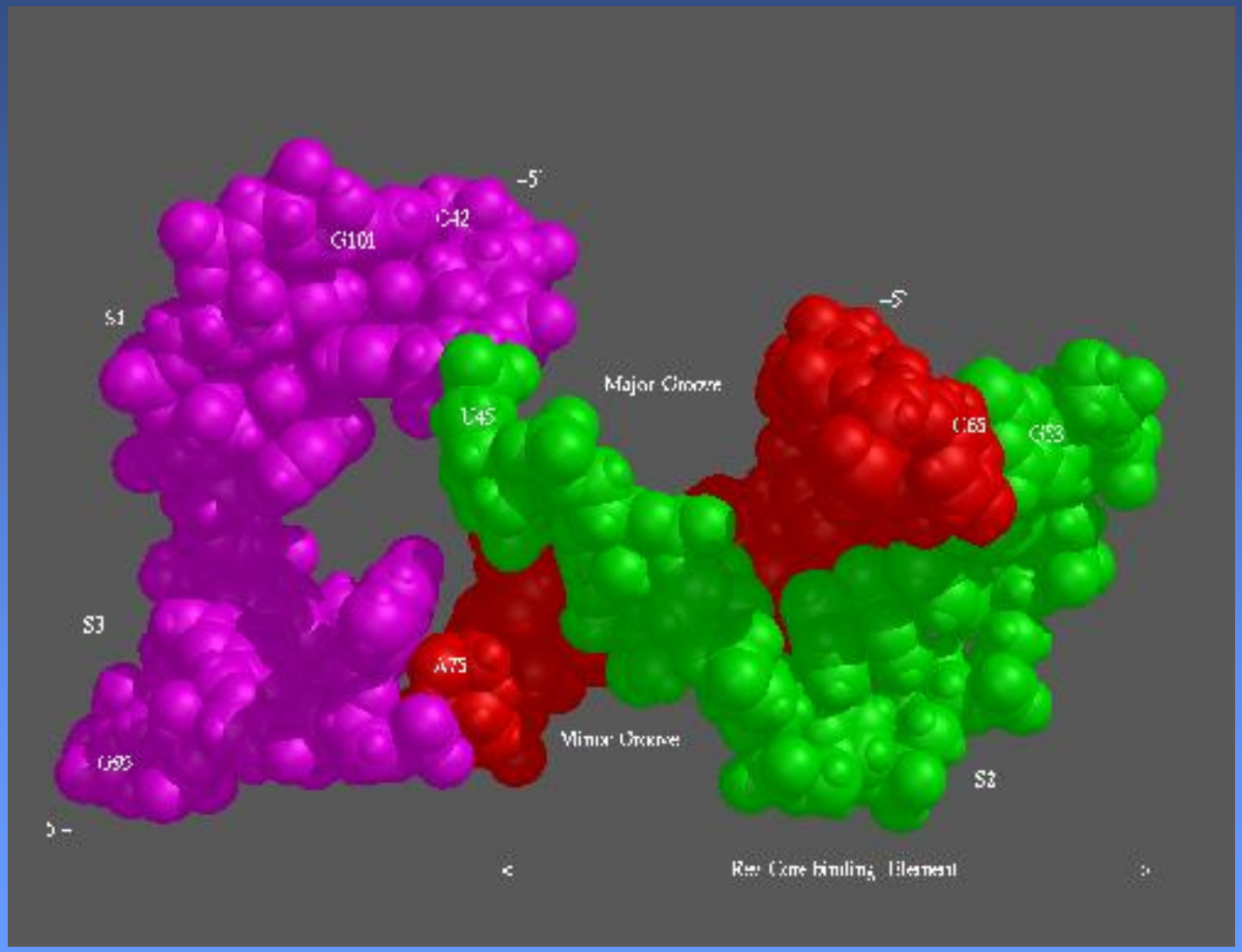
bulge

Mult-branch

hairpin

(Drawn with Shapiro et al STRUCTURELAB)

Atomic level structural model of HIV-1 RRE core



Maximally stable secondary structure

One of the most widely used approaches employs base pair and loop parameters, calibrated from thermodynamic experiments on small RNA molecules, to deduce the lowest free energy structure.

Zuker's implementation of a dynamic programming algorithm first described by Ruth Nussinov and co-workers in the early 1970's is called mfold.

We use variations of this program extensively.

Other approaches exist. (genetic algorithms, phylogenetic, monte carlo, linguistic, etc.)

How we judge the significance of a structure

A random sequence of bases can fold into a structure that looks real.

We initially focus on the stability of a real sequence's structure compared to the stability of a large number of structures from randomly shuffled sequences of the same base composition.

Statistical significance is obtained, but since that may or may not reflect biological significance, we follow the suggestion of Karlin and refer to statistically significant folded structures as statistically "unusual" folding regions, or **UFRs.**

The main workhorses

SIGSTB scans a sequence for potential unusual folding regions(UFRs).

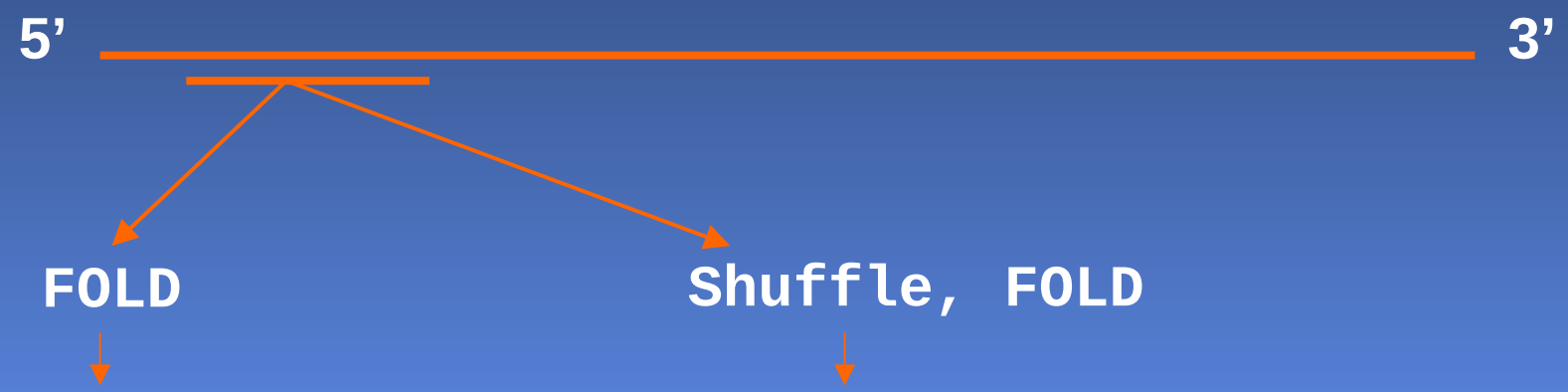
SEGFOLD employs a systematic search with automatically varying window size to define the limits of the UFRs.

EDscan looks for well-formed folding regions.

How SIGSTR and SEGFOLD Find Statistically Unusual Features

The Principle:

Choose successive, overlapping segments

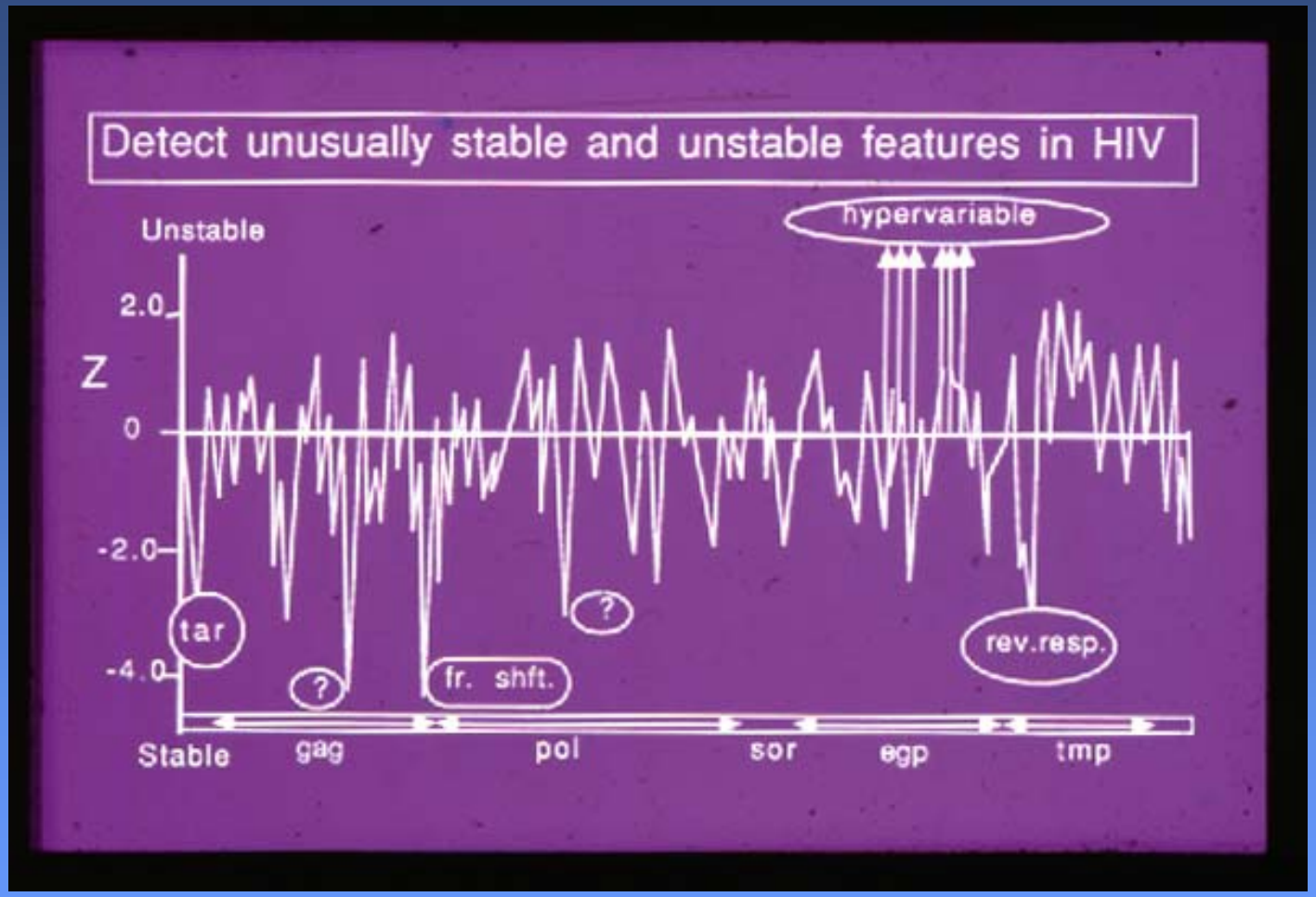


FOLD → Obtain stability of real segment

Shuffle, FOLD → Average stability of shuffled segments, and std. deviation

$$\text{Z-score} = \frac{(\text{stability of real}) - (\text{Av. Stability of shuffled})}{(\text{Std. Deviation of the shuffled})}$$

Example Z-score plot



How EDscan finds statistically well folded regions (WFR's)

The Principle:

Choose successive, overlapping segments

5' **3'**

FOLD

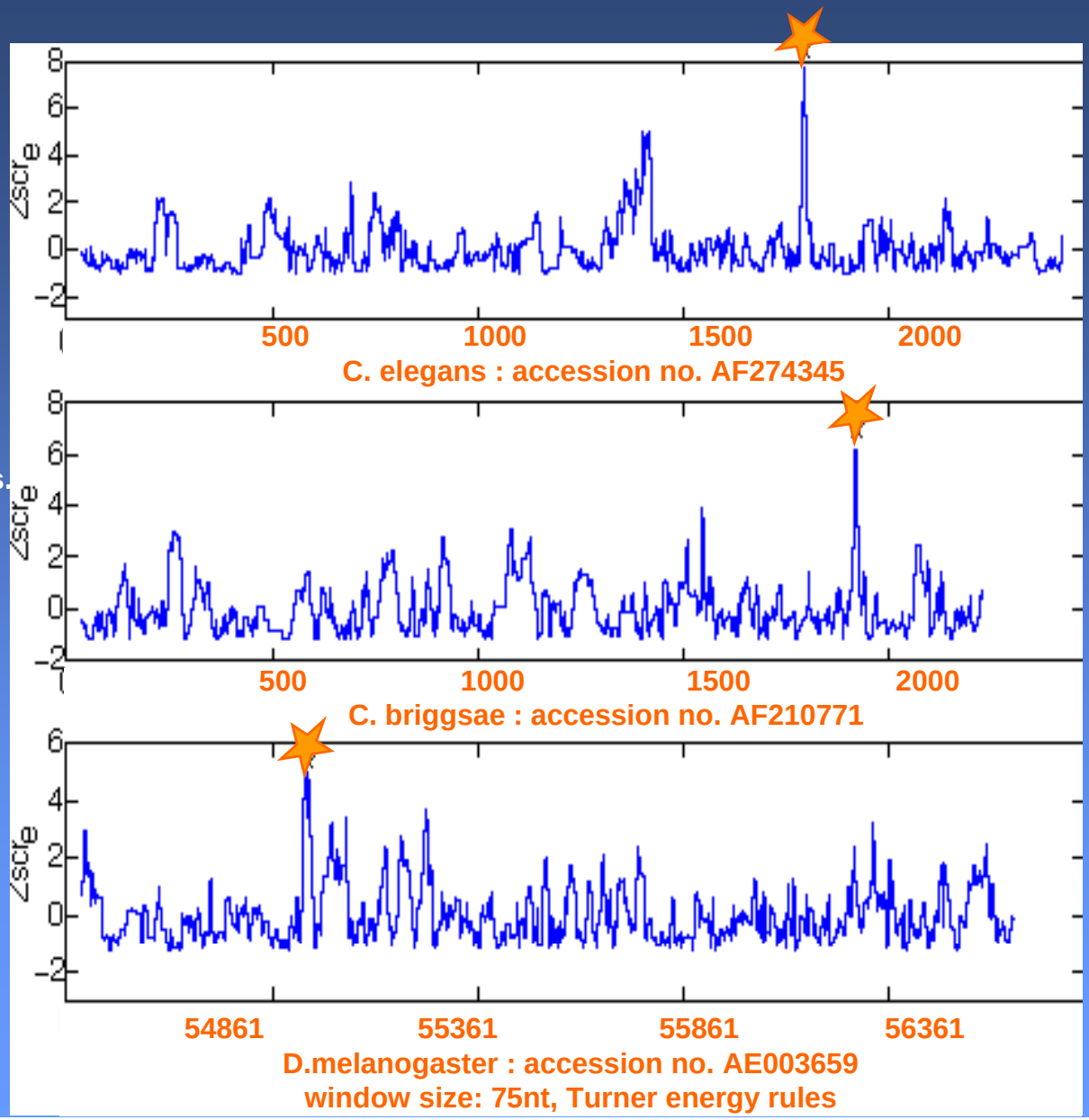
FOLD after excluding all pairs in the optimal structure

E_{diff} = difference between real fold and excluded fold

$E_{diff}(w)$ is mean of all segments in the sequence, and std_w is their standard deviation.

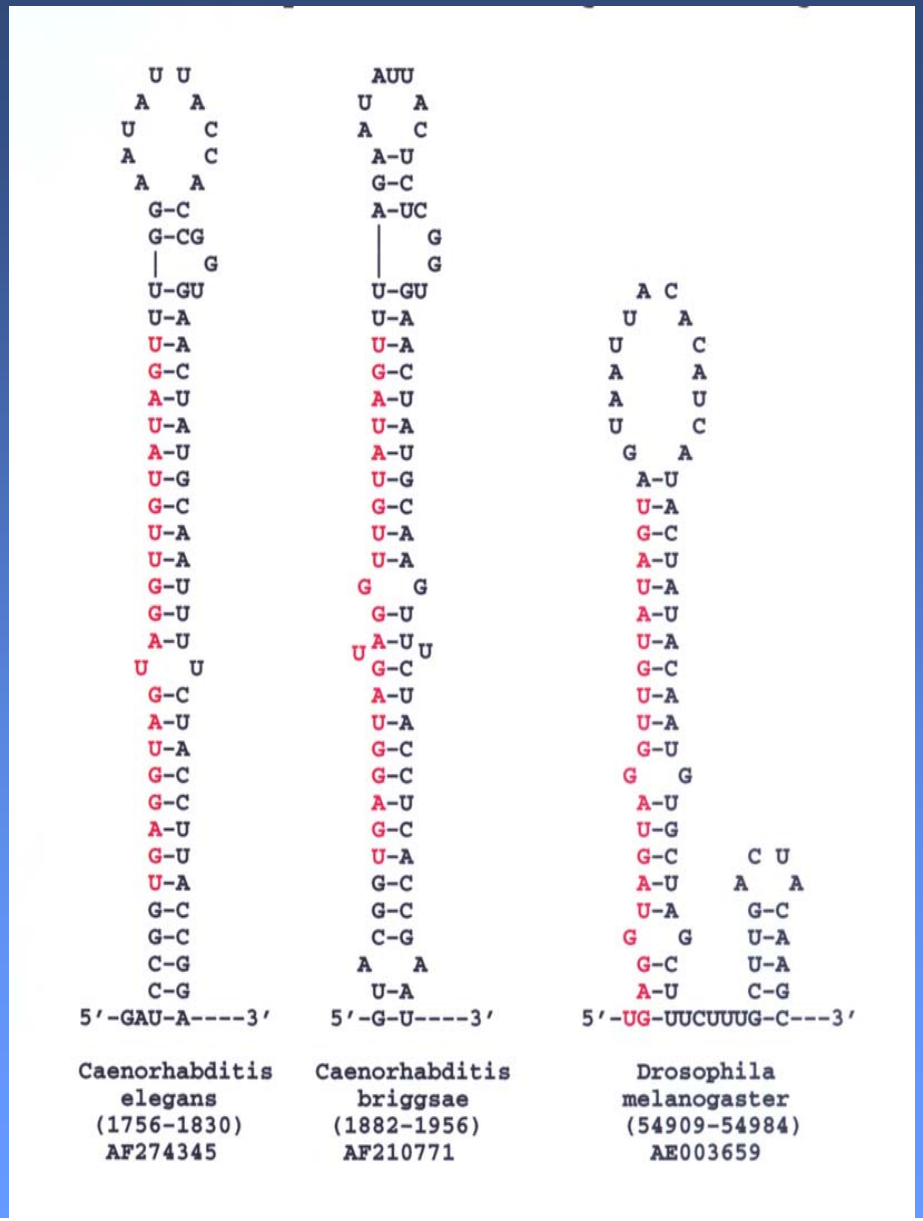
$$Zscr_e = (E_{diff} - E_{diff}(w)) / std_w$$

ED_SCAN finds let-7 gene features



Marks the location of known *let-7* features.

Well determined structures in let-7 genes



Early view of mRNA

But now it's more interesting

UTR = Untranslated Region

Discovering features in 5’ and 3’ UTRs

5’ UTR features

Internal ribosome entry sequence (IRES); may be developmentally controlled
Iron response element (IRE)
Neurovirulence mutations in poliovirus
(Others ???)

3’ UTR features

Poly A addition site
mRNA stability
Intracellular localization in development
e.g.anterior/posterior egg pole
Effects on translation; may be developmentally controlled (e.g. frog onconase-related RNAse)
(Others ???)

Utrdb15.0

Table 1. Number of entries (N) and nucleotide length (L) of UTRdb collections (release 15.0) after redundancy cleaning

Collection	N	L	Redundancy %N	%L
5′-UTR				
Fungi	2223	275 886	27.76	16.74
Human	30 922	4 515 966	40.06	22.33
Invertebrate	19 947	2 987 661	28.06	18.60
Other_mammal	5751	852 910	35.51	14.36
Other_vertebrate	7327	792 573	8.61	15.00
Plant	17 819	1 490 067	25.64	13.51
Rodent	19 759	2 518 594	36.76	20.22
Virus	14 663	3 402 809	81.71	73.82
Total	118 411	16 836 466	–	–
3′-UTR				
Fungi	2304	465 396	14.84	11.30
Human	36 015	18 906 357	41.72	29.83
Invertebrate	18 230	5 151 363	36.74	20.92
Other_mammal	6927	2 548 434	29.78	17.55
Other_vertebrate	8528	3 351 751	21.22	13.37
Plant	21 526	4 328 226	16.93	13.28
Rodent	21 489	9 113 464	37.32	23.30
Patent	14 118	3 359 534	77.40	69.42
Total	129 137	47 224 525	–	–

UTRdb 15.0 was generated from EMBL release 67. Relevant redundancy percentages calculated with respect to the number of entries (%N) and to the nucleotide length (%L) are also indicated.

5UTR db

Table 1. Distribution of the 5′UTR Length for Sequences Collected in the UTR-DB (Release 7.0, May 1998)

Files	Total No. of Sequence	No.of Seq (L >= 100) nt	No.of Seq (>= 200) nt	No.of Seq (>= 300) nt	No.of Seq (>= 600) nt	Mean of Length
Human (Hum_5utrnr.dat)	6669	3725	2055	1246	415	207-nt
Mammalian (Om_5utrnr.dat)	2106	833	348	202	56	138-nt
Rodent (Ro_5utrnr.dat)	7056	3844	2048	1165	345	192-nt
Vertebrate (Ov_5utrnr.dat)	2846	1265	610	311	81	152-nt
Invertebrate (In_5utrnr.dat)	4054	1975	1107	709	268	205-nt
Plant (Pl_5utrnr.dat)	5993	1826	592	298	76	106-nt
Fungal (Fun_5utrnr.dat)	931	406	215	153	63	174-nt

Drosophila upstream AUGs

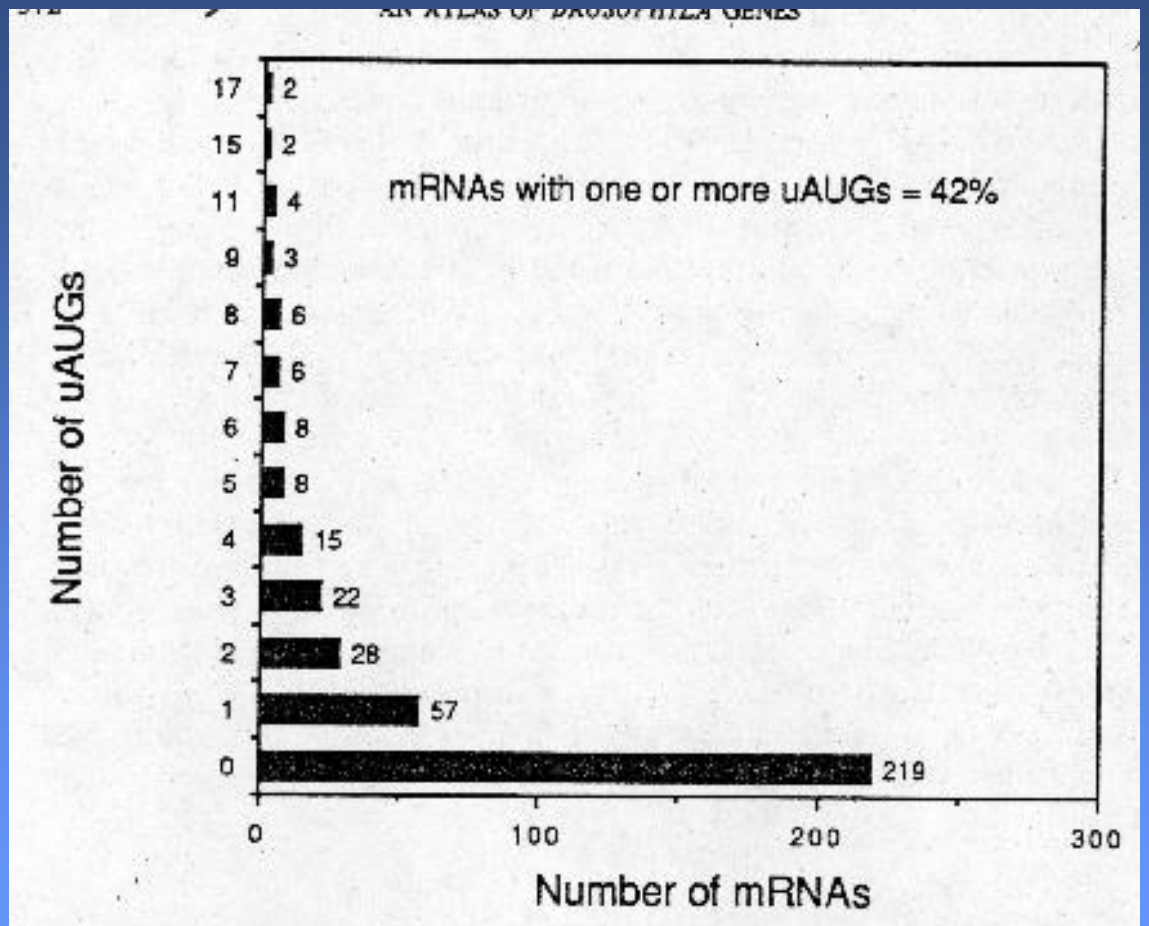


FIG. 36.2. Distribution of upstream AUGs in the 5′ untranslated leader sequences of *Drosophila* mRNAs.

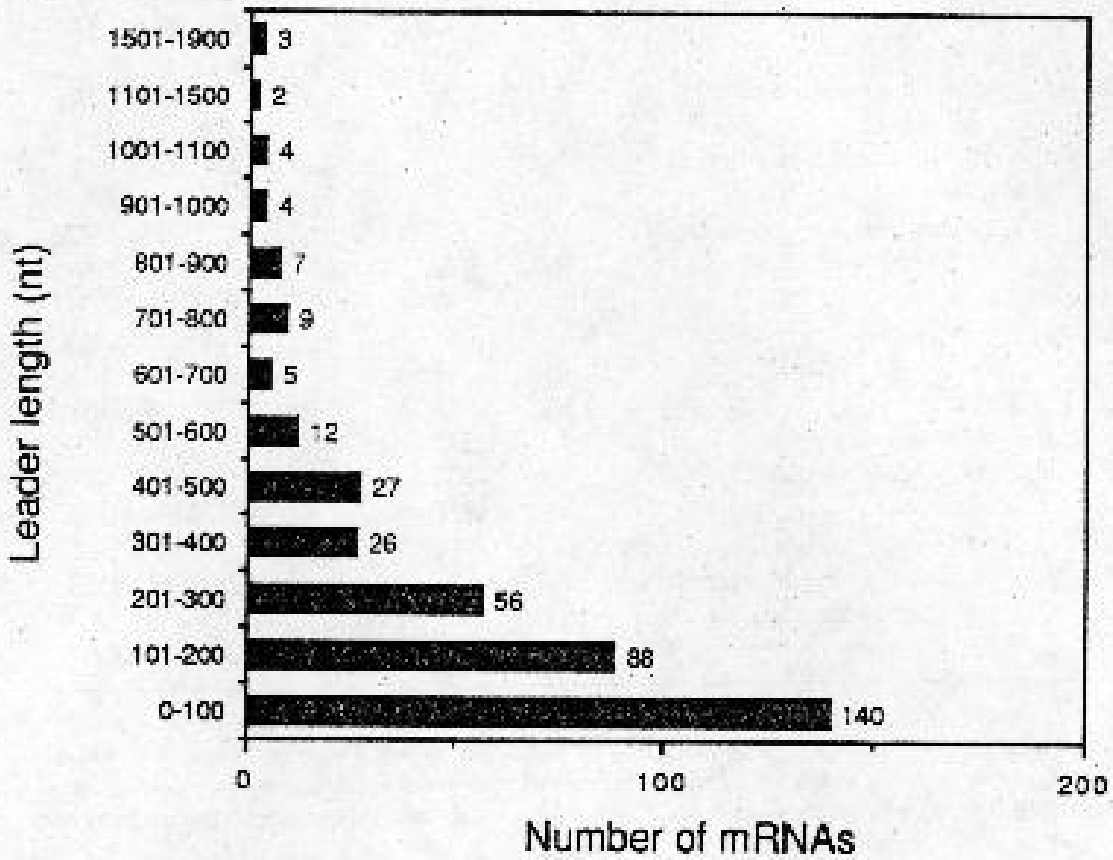


FIG. 36.1. Distribution of the number of nucleotides (nt) in the 5′ untranslated leader sequences of *Drosophila* mRNAs.

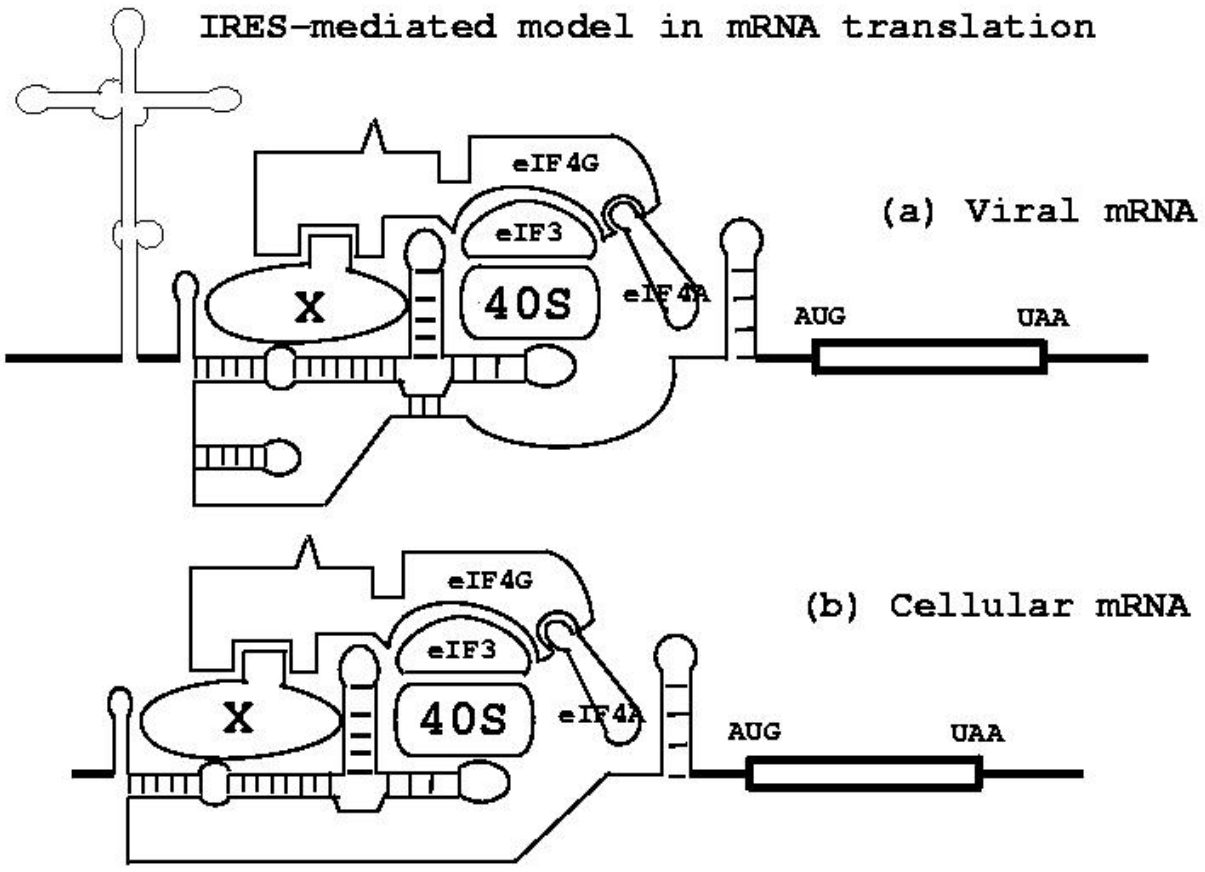
Today's talk(continued)

Chapter 1:
Computer prediction of RNA structure

Chapter 2:
Application of computer tools to 5'UTRs of virus and cellular mRNAs

Chapter 3:
A discovery of novel features in a long 3' UTR

IRES in translation



Internal Ribosome Entry Sequences(IRES)

The poliovirus genome, which is an mRNA, presented a paradox to Kozak's "scan to the first AUG" model. There are a number of upstream AUGs before the true initiator codon in the 743 nt, 5'UTR.

Sonenberg and co-workers showed definitively that ribosomes initiated internally.

Structure predictions show similar features in a wide range of organisms from viruses to cellular proteins, including many oncogenes.

Now some examples.

Sequence Analysis Data of the 5′UTR and IRES of Viruses:

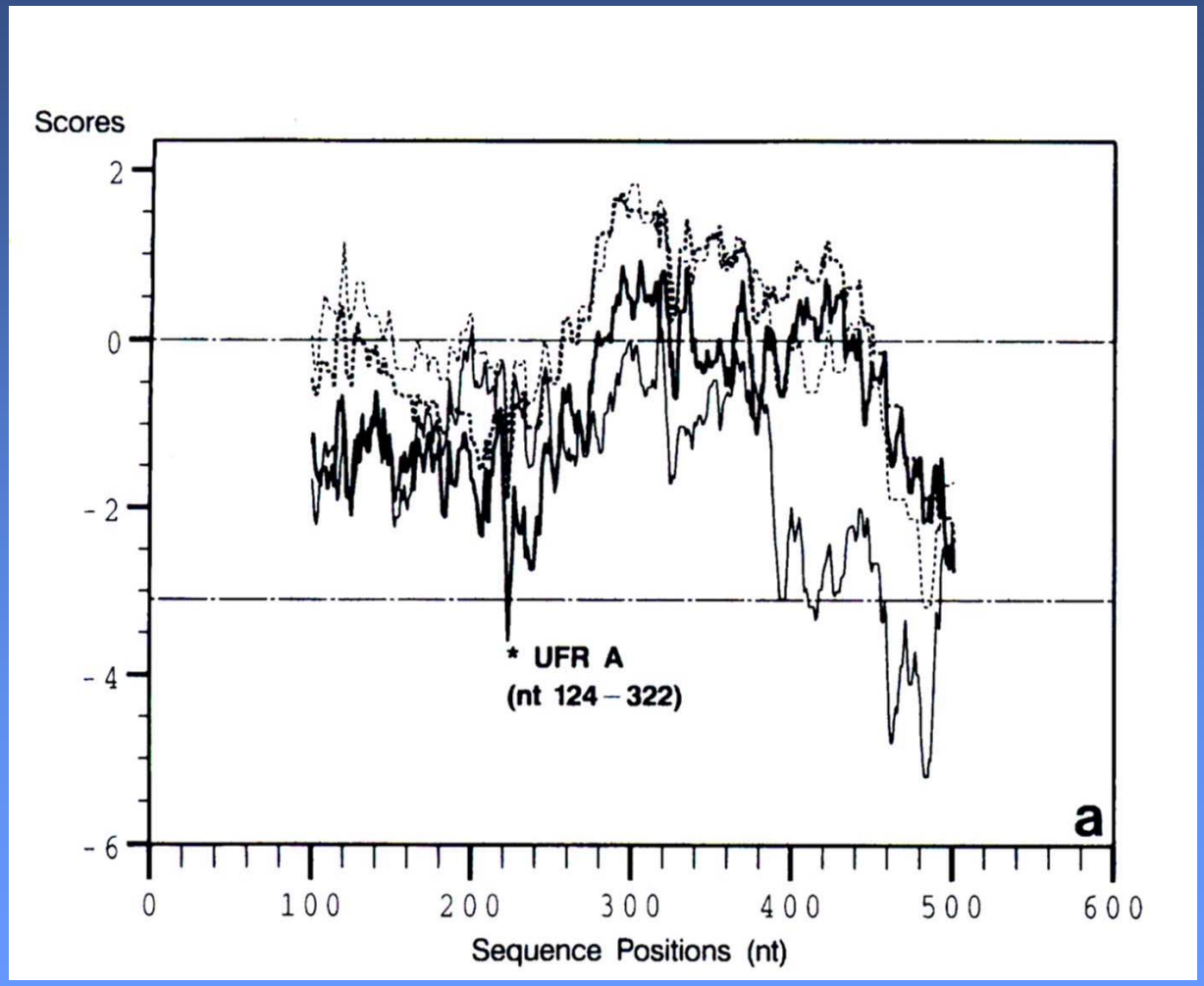
Picornavirus IRES sequences (137):

HETV:	human enterovirus (3)
Cox-A:	human coxsackievirus type A (18)
Cox-B:	human coxsackievirus type B (5)
Echo:	human echovirus (18)
SVDV:	swine vesicular disease virus (1)
PV1-3:	human poliovirus types 1-3 (31)
HRV:	human rhinovirus (6)
BEV:	bovine enterovirus (3)
EMCV:	encephalomyocarditis virus (6)
Mengo:	mengovirus (1)
TMEV:	Theiler's murine encephalomyelitis virus (11)
ERV:	equine rhinovirus (2)
FMDV:	foot-and-mouth disease virus (12)
HAV:	human hepatitis virus (20)

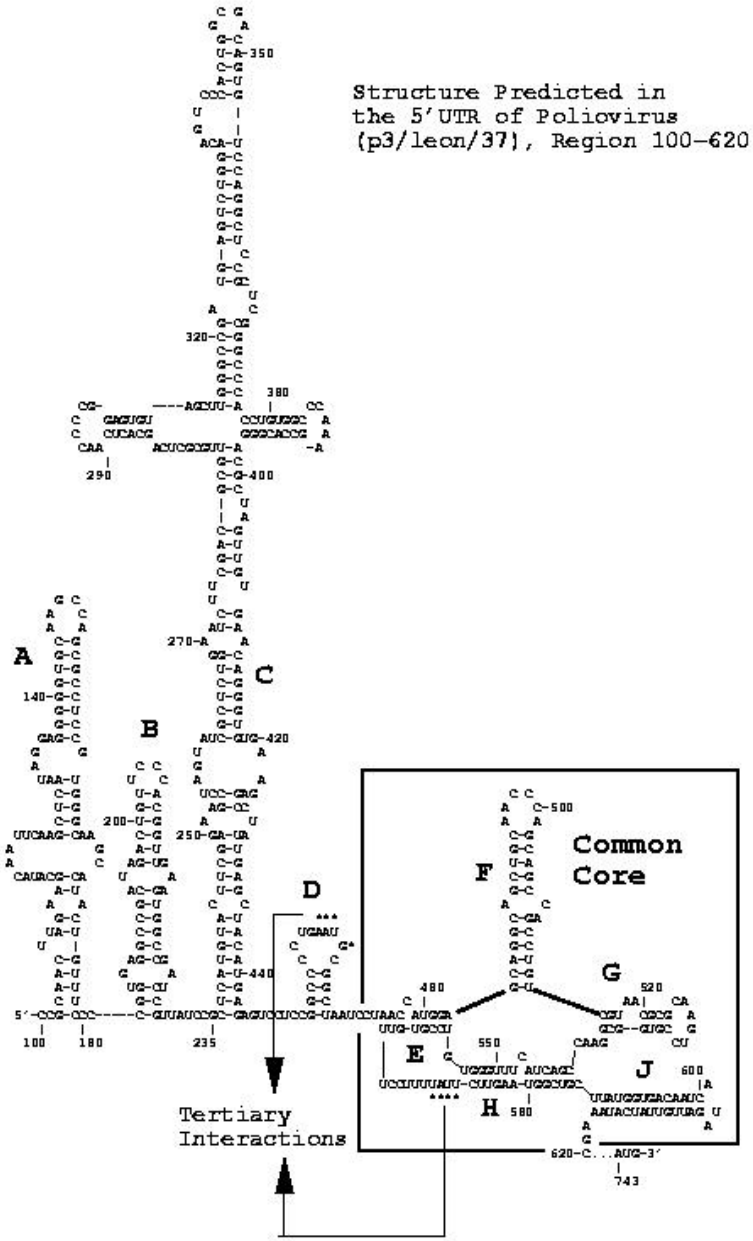
HCV type 1-11 (331)

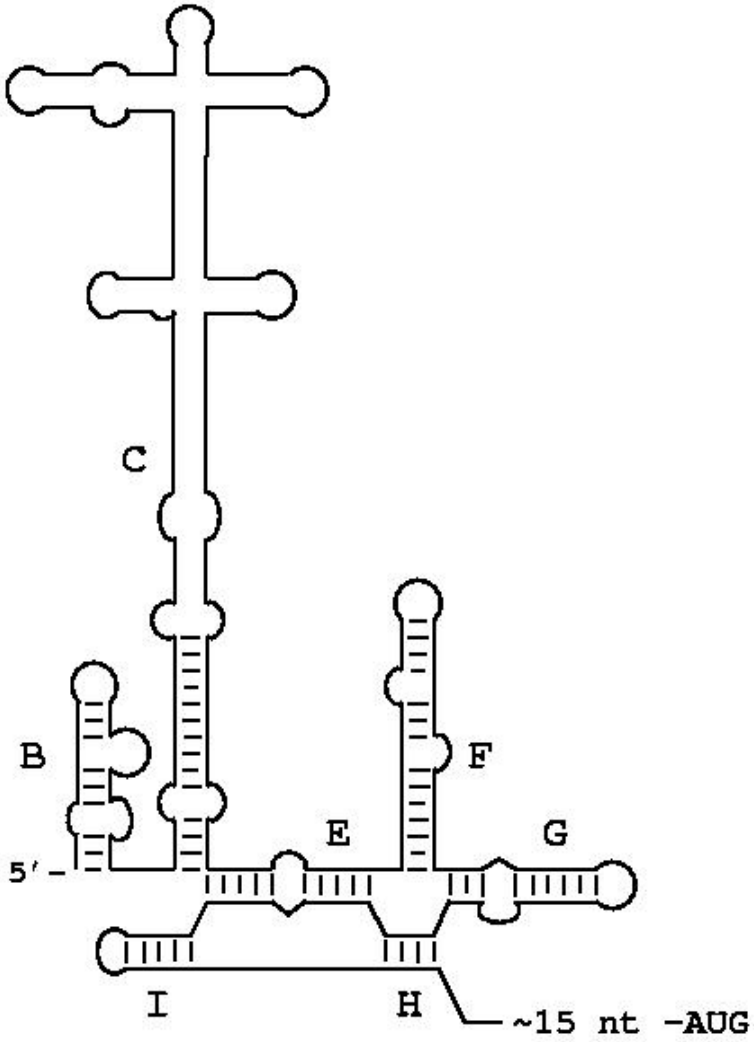
Pestivirus (BVDV, HoCV and others, 143)

SIGSTB plot for a picornavirus(HCV)



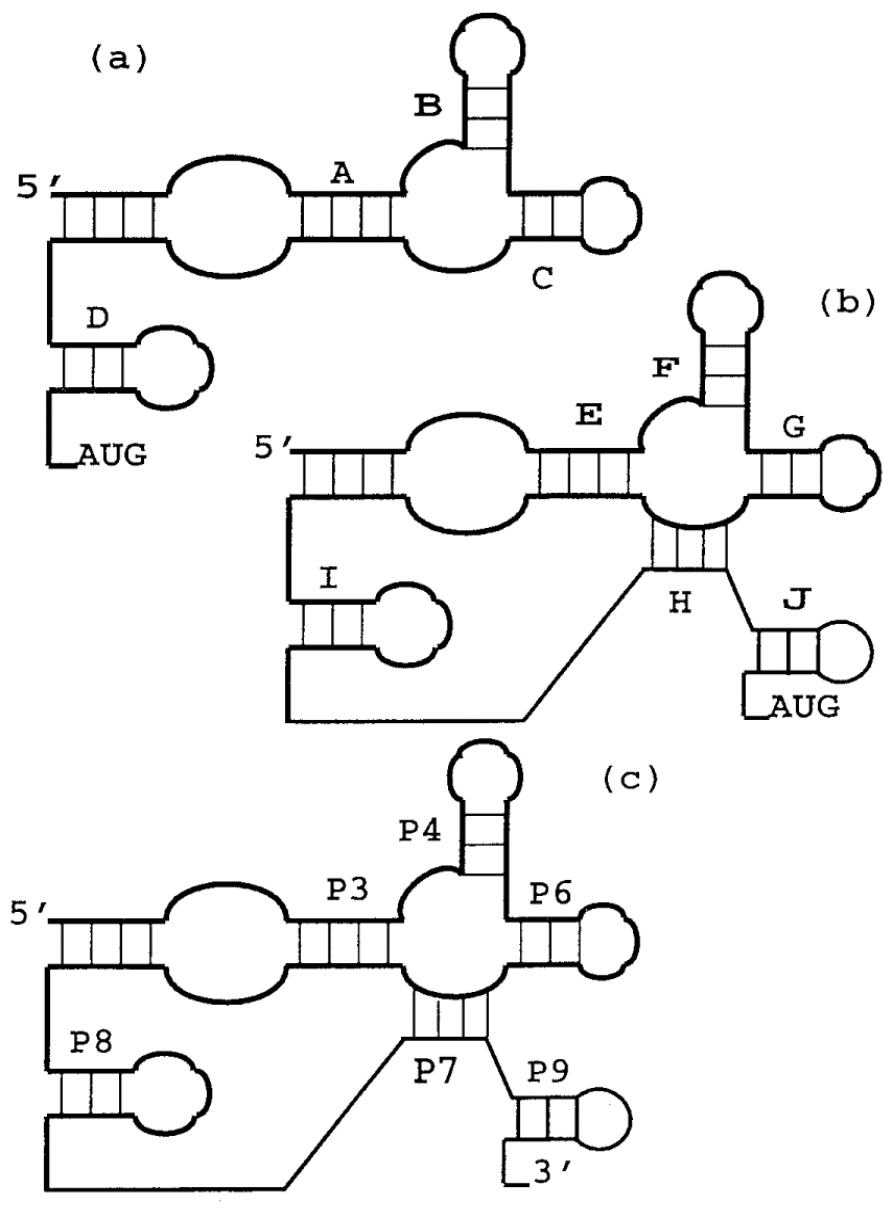
Structure Predicted in the 5′UTR of Poliovirus (p3/leon/37), Region 100–620




C
B
F
E
G
5'-
I
H
~15 nt -AUG

Cell , virus and intron IRES cores summary

Cellular IRES



Viral IRES

Group I intron

Some examples of cellular IRESs

Following the establishment of the phenomenon in viruses, a number of cellular mRNAs exhibiting long 5’UTRs with upstream AUGs, and IRESs were observed.

Reasonable structures could be predicted for those cases where experimental data existed.

This provides a rationale for extending the predictions to untested 5’UTRs.

5′UTR of Cellular mRNAs:

1580 nt (15 uAUG) AUG

human **AML-1** (Acute Myeloid Leukemias)

1022 nt AUG (3 uAUG)

human **PDGF2/c-sis** (platelet-derived growth factor B)

633 nt AUG 6 uAUG)

human **hcK** (hemopoietic cell kinase)

483 nt AUG (3 uAUG)

human **FGF-2** (fibroblast growth factor 2)

368 nt AUG (4 uAUG)

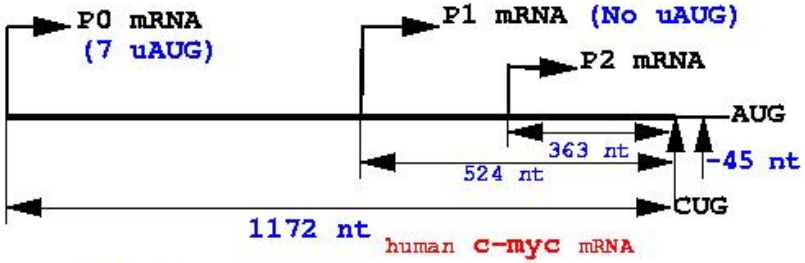
human **eIF4G** (eukaryotic initiation factor 4 gamma)

340 nt AUG (6 uAUG)

human **abl** (containing alternative first exon)

1038 nt AUG

human **VEGF** (Vascular Endothelial Growth Factor)



human **c-myc** mRNA

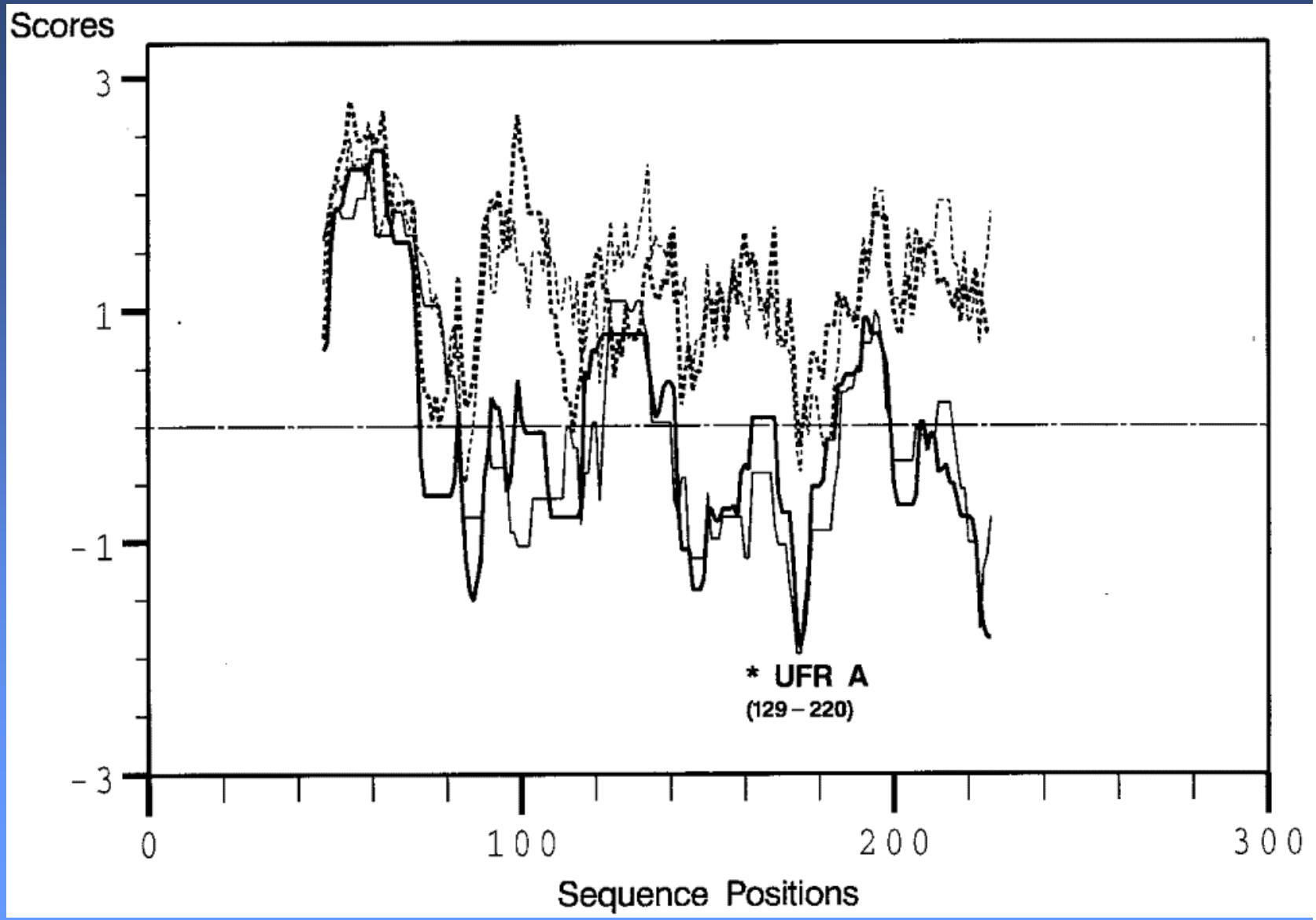
333 nt AUG (3 UAUG)

human **erb** (thyroid receptor hormore)

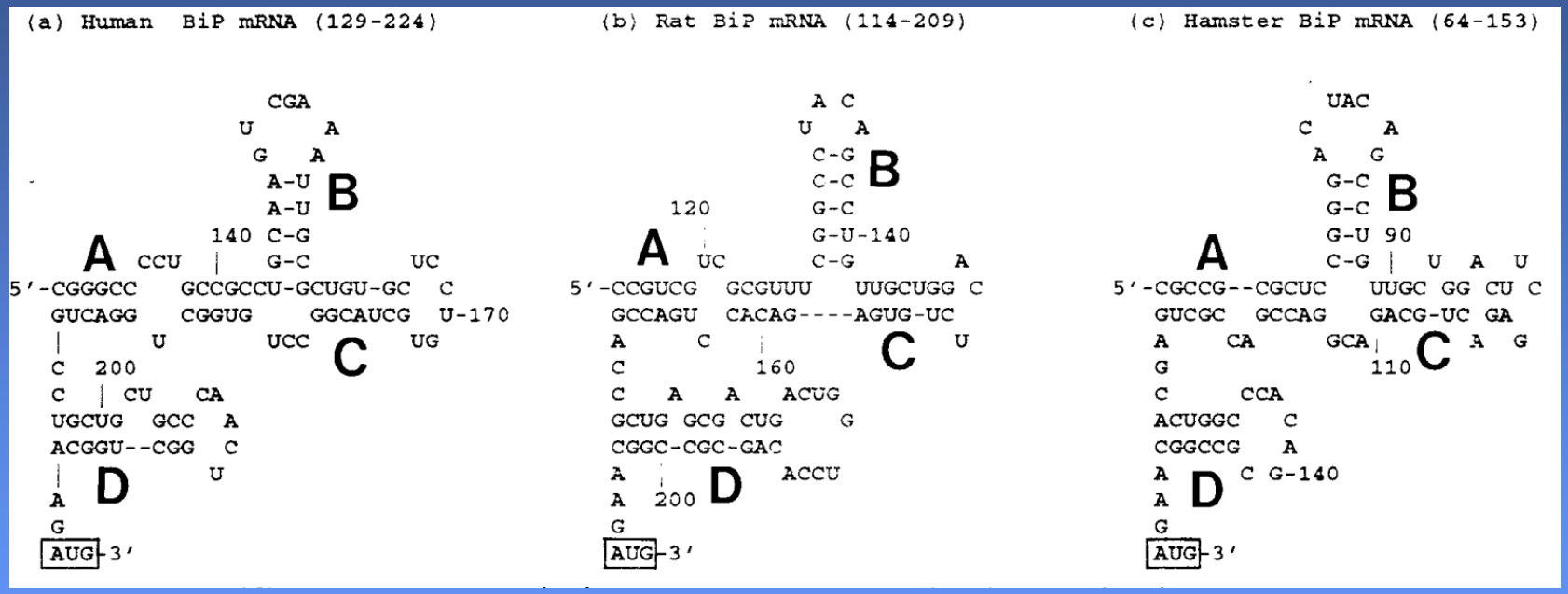
267 nt AUG (3 uAUG)

human **mas** (partial 5′UTR)

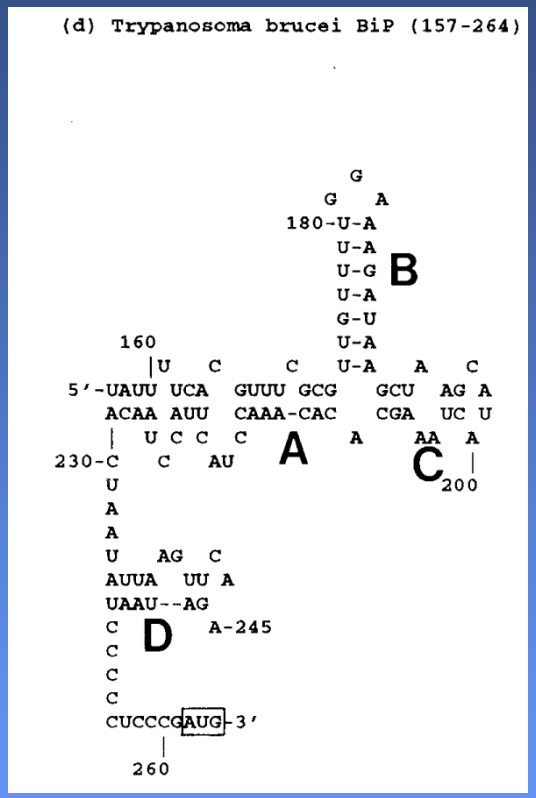
Distributions of significance scores (Sigscr, broken curve) and stability scores (Stbscr, continuous curve) in human BiP mRNA sequence.

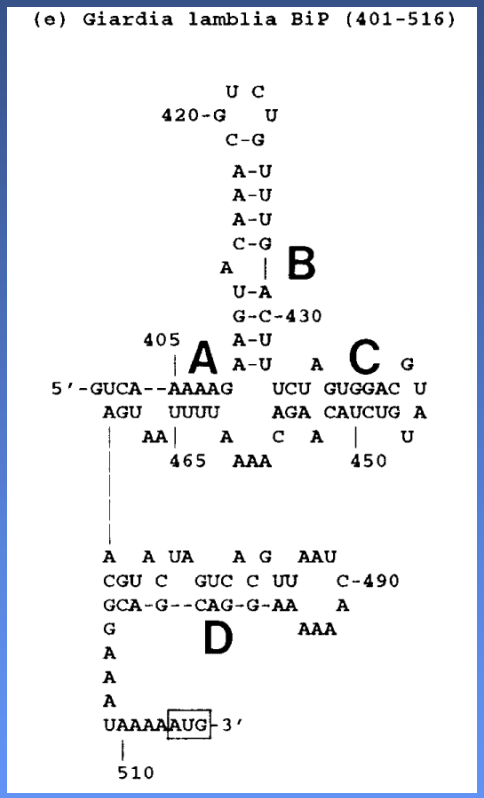


Human, rat and hamster BiP IRES

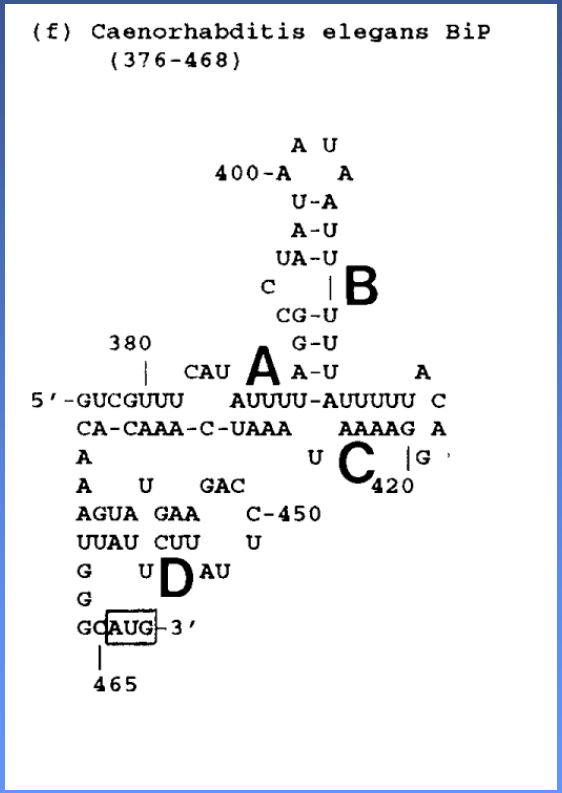


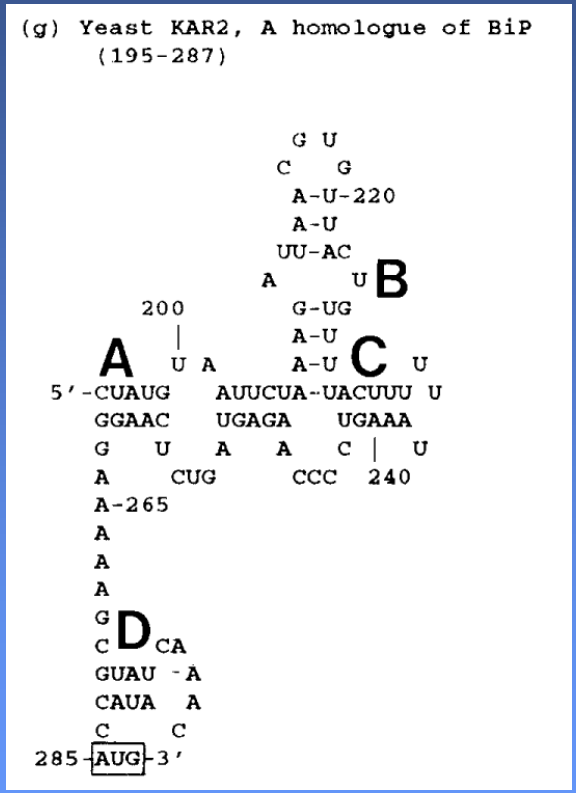
Trypanosome and giardia BiP IRES





C. elegans and yeast BiP IRES





Human and rat FGF-2 IRES

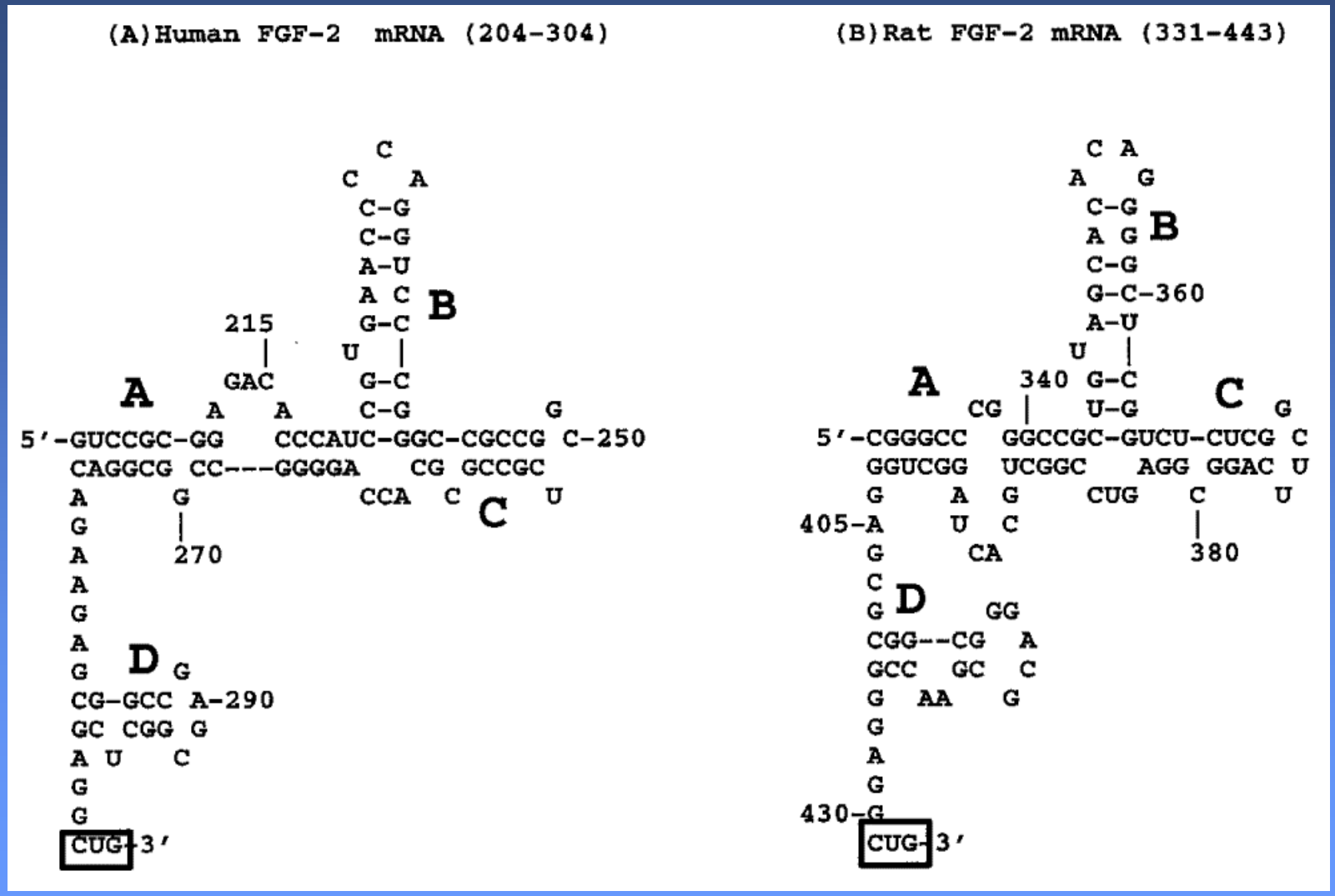
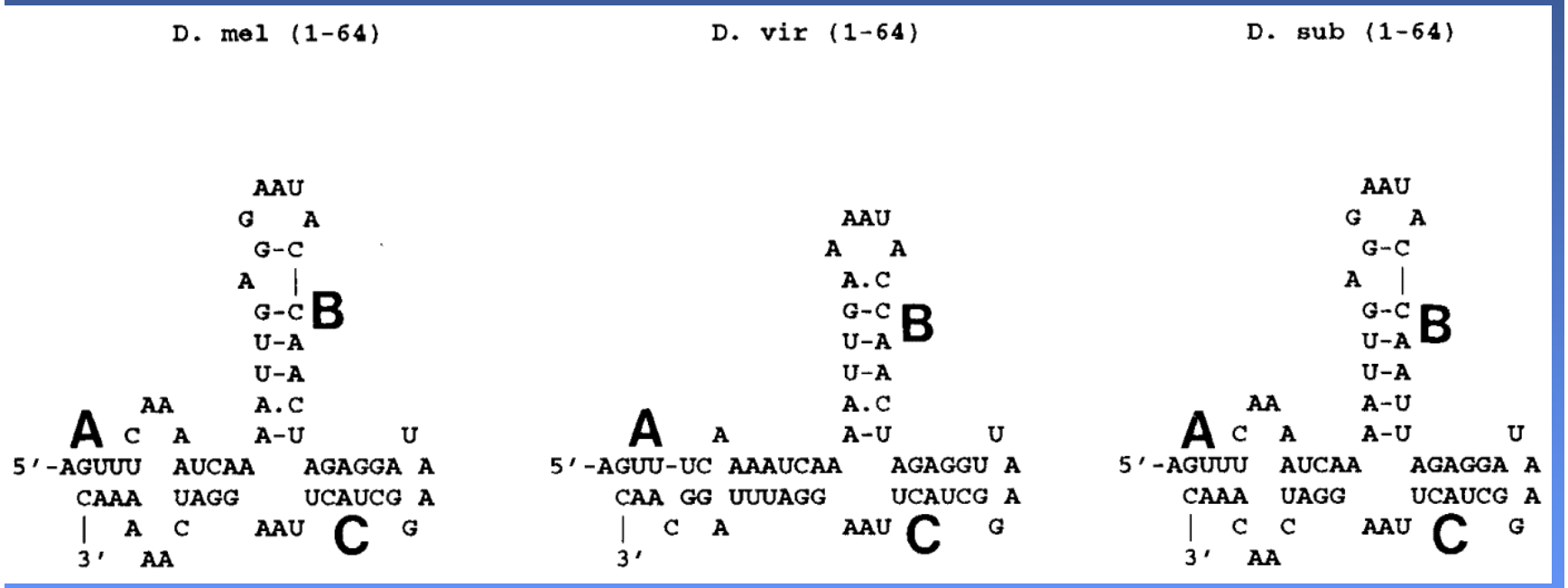
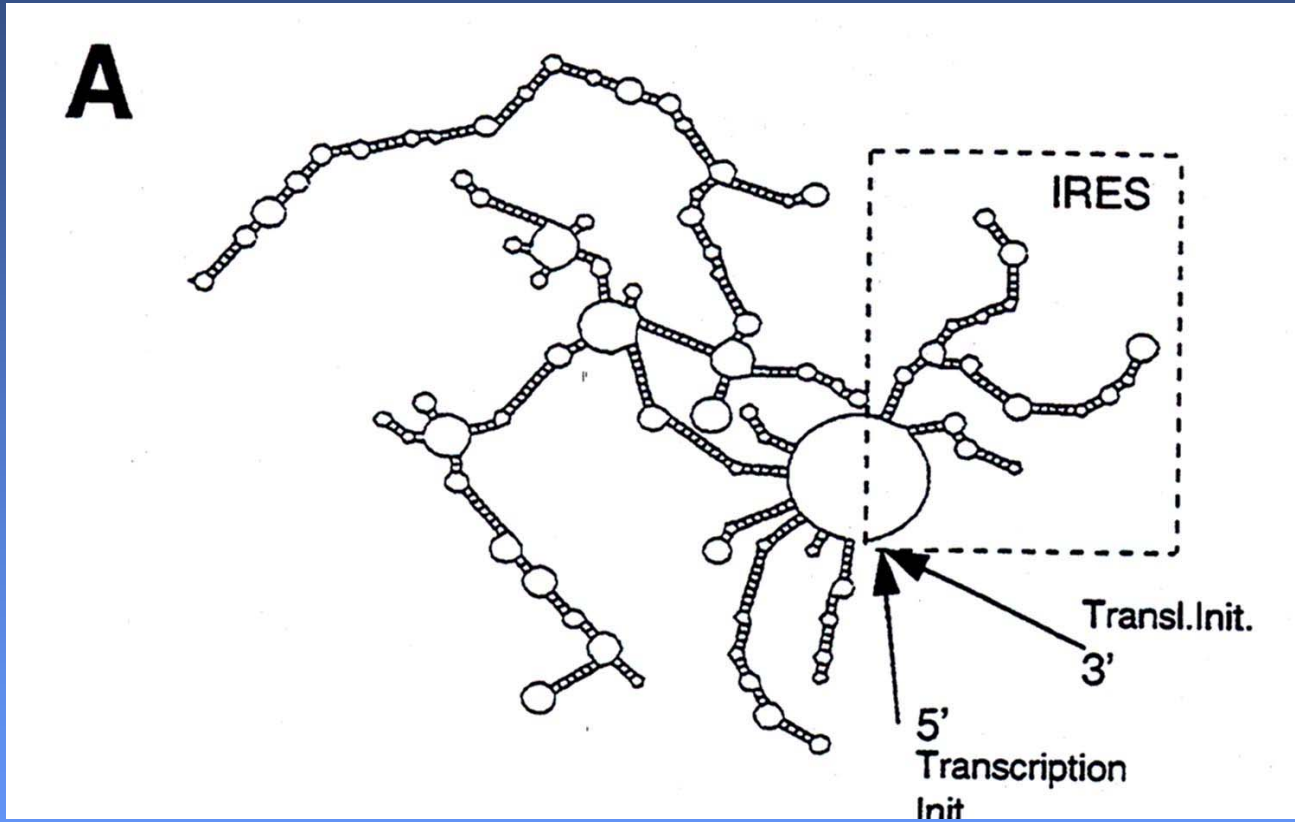


Fig 7

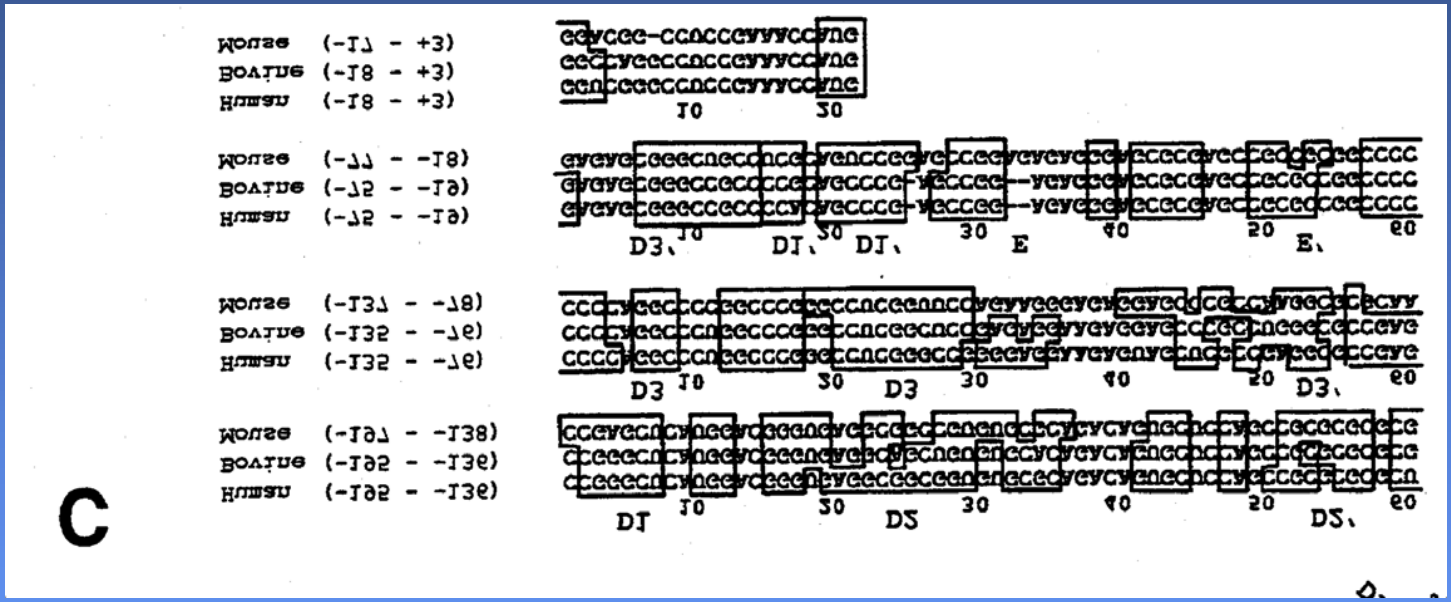

D. mel (1-64)
AAU
G A
G-C
A |
G-C B
U-A
U-A
AA A.C
A C A A-U U
5'-AGUUU AUCAA AGAGGA A
CAAA UAGG UCAUCG A
| A C AAU C G
3' AA
D. vir (1-64)
AAU
A A
A.C
G-C
U-A B
U-A
A.C
A A-U U
5'-AGUU-UC AAAUCAA AGAGGU A
CAA GG UUUAGG UCAUCG A
| C A AAU C G
3'
D. sub (1-64)
AAU
G A
G-C
A |
G-C
U-A B
U-A
AA A-U
A C A A-U U
5'-AGUUU AUCAA AGAGGA A
CAAA UAGG UCAUCG A
| C C AAU C G
3' AA

```
                                        10        20        30        40        50        60        70        80
#1  human Bip     (129)   ..CGGGCC..ccu...GCCGC.cUGCAA.....gucgaaa........UUGCGCUGU..GC..uccugu...GCu..ACGG
#2  rat Bip       (114)   ..CCGUCG..uc....GCGUU.uCGGGC.......uaca.........GCCUGUUGCu.GG...acu.....CU...GUGA
#3  hamster Bip   ( 64)   ...CGCCG........CGCUC..CGGG......acuacag.........CCUGUUGCu.GGaCUucg..AGaCU...GCAG
#9  yeast KAR2    (195)   ...CUAUG..ua....AUUCU..AAAGa.uUAA..cgug....UUAcugUUUUACUUU......uuu.........AAAGU
#10 CE Bip        (376)   .GUcGUUU..cau...AUUU...UAGGccuAAU..aaua....AUU...UUUAUUUUU......acag........GAAAA
#11 GL Bip        (401)   .....UCA........AAAA..gAAGUaCAAAC..gucu....GUUUG.ACUUUCUaGUGGAC.guau..GUCUAC.aAGA
#12 TB Bip        (157)   UAUUuUCAc.....GUUUcGCG.UUGUUUU......gga.......AAGAUAAGCUa..AG...caua....CU..aaAGC
#6  rat FGF-2 E1  (331)   ..CGGGCC....cg..GGCCG..CUGu.AGCAC...acag...GGGCU..CGGUCU.CUCG...gcuu....CAGG.cGGA
#4  H. FGF-2 E1   (204)   GUCCGCGG.agaca..CCCAU..CCGu.GAACC...cca....GGUCC..CGGGCCG.cCG...gcu.....CGc.CGcGC
#5  H. FGF-2 E2   (360)   ..GGGGCC........GUGC...CCCGGag...CG.ggu.CG.gagg.CCGGGGCCGggg.CCGggggaCGG.....CGGC
#8  B. FGF-2 E2   ( 16)   ..GGAG..........CGCG....UCGG........agg..........CCGGGGCCG......gggcg...cgg.CGGCU
#7  rat FGF-2 E2  (438)   ...CCCGC....g...GGCGA..GCCGCG..CU..gggg....GGc.CGAGGCCGGGG...UCGgggc.CGGggagCCCCG
#13 H. FGF-5      ( 26)   ..UAUGUCcacccgGUGcGGCGAGGCGGGCagaGCCagaGGCacGCagCCGCACAGGGgCU...acag....AG..CCCAG
#14 H. FGF-9      ( 81)   .GCAGUAA.gg.....GAGGG..GAGUUG......gaua........UACCUCGCCUA.....auaucucc.....UGGGU
#20 rat FGF-9     ( 82)   .GCAGCAA.gg.....GAGGG..GAGUUG..,...gaua........UACCUCGCCUAG.....ugucuc.....CUGGUU
#17 M. FGF-6      (  4)   ..UUUUAG........GGCCA..UUA......,..auuc...........UGACCACGUGCCU.gag...AGGCaAGGUGG
#18 M. FGF-8      (  2)   ....GCAC...c....UUCGGc.UUGucc..CC..ccgc....GGcc.ucCAGUGGG.aCG...gcg.....UGa.cCCCG
#15 H. KGF        (366)   GUcAAUGA........CCUAG..GAGU....aacaauca.........AC.UCAAGA.......uuca.........UUUU
#16 H. KSFGF      (221)   ...GGUCG.gcc.GUGCGCUCccGCG...ggacgccacagg.........CGCAGCU....cugcccccc.......AGCU
#19 M. KFGF       ( 33)   ....UUUG.cucuc..GCUACu.UAGGU..CUG..ugcg....CAGc.ACUCACCG......aacuca..........CGG
                          ____________________  __________  ____________/\___________  ____________
                                   A                 B            B'          C              C'


                                        10        20        30        40        50        60        70        80
#1  human Bip     (180)   ccu..GUGGC.u......GGACUG..cc.....UGCUGcuGCC.....caacu......GGCUGGCA..........agAUG
#2  rat Bip       (158)   .....GACAC.c......UGACCG.acc......GCUGa.GCGaCUGacugguccaCAGCGCCGGCa..........agAUG
#3  hamster Bip   (111)   acg..GACCGac......CG.CUG.agc......ACUG..GC......ccacagc.....GCCGGCa..........agAUG
#9  yeast KAR2    (243)   ccccaAGAGUagucu...CAAGG..gaaaaagc.GUAU...........caaac........AUAC............cAUG
#10 CE Bip        (425)   u.....AAAU.c......AAACAC.aa......AGUAu..GAA.....gaccuua....UUCuUAUU........gggcAUG
#11 GL Bip        (457)   caaaa.UUUUaa......UGA....a.......CGUaCuaGUCaCgUUaaucaaaaAAGGACG.ACG...gaaauaaaaAUG
#12 TB Bip        (208)   a...CAC.AAACcuac..UUAccuAACAcuaau.AUUAagUU.........caa......GAUAAU...cccccucccgAUG
#6  rat FGF-2 E1  (384)   guc..CGGCUgcacua..GGCUGGgagcg.....CGG....CG.....ggacg......CGaaCCG........ggaggCUG
#4  H. FGF-2 E1   (260)   acc..AGGGG........CCgGCGGACagaagag.CG...GCC.....gagc.......GGCuCG...........aggCUG
#5  H. FGF-2 E2   (415)   ucccc.GCGC........GGCUCCagc..GGCUCGGGGauCCC.....ggcc.......GGGCCCCGcAGggaCC....AUG
#8  B. FGF-2 E2   ( 53)   ccccg.CGCGg.........CUCCagg..GGCUCGGGGacCCC.....gcca.......GGGCCUUGguggGGCC....AUG
#7  rat FGF-2 E2  (494)   agag.CUGCCgca.....GCGGG..........GUCCC..........ggggccgcgga...GGGGC...........cAUG
#13 H. FGF-5      ( 96)   aaUCaGCCcUACaa....GAUGCA..c.......UUAGgaCC......cccgc.......GGCUGGa.........agaAUG
#14 H. FGF-9      (129)   ugacaCCAUCa.......UUAUUGUuuauu....CUUGu.GCU.....ccaaa......AGCCGAG......uccucugAUG
#20 rat FGF-9     (130)   .gauaCCAUCa.......UUAUUGUuuauu....CUUGu.GAU.....uaaa.......AGCCGAG......uccucugAUG
#17 M. FGF-6      ( 49)   a....UGGCC........CUGGGA.....ca...GAG...........gcug...........UUCa......ucacu.AUG
#18 M. FGF-8      ( 52)   c....UCGGGcucuca..GUGC....ucccgg..GGC...........cgcgc..........GCC.............AUG
#15 H. KGF        (407)   ca...UUAUG........UUAUUcAUgaacacccGGAG..........cacua.........CACUa..........uaAUG
#16 H. KSFGF      (274)   uccc.GGGCGCAC.....UGACC...g.......CCUGa.CCGAcG..cacgc.CcUCGGgcCGGG.............AUG
#19 M. KFGF       ( 81)   ccc..GCAGC.u......CGAA............CUCA..CG......cacggcc.....CGCGGG.......cc.gggAUG
                          ____________________  ____________________  ____________________  ____
                                   A'                   D                    D'            Initiator
```

Vegf 5utr



Vegf aln

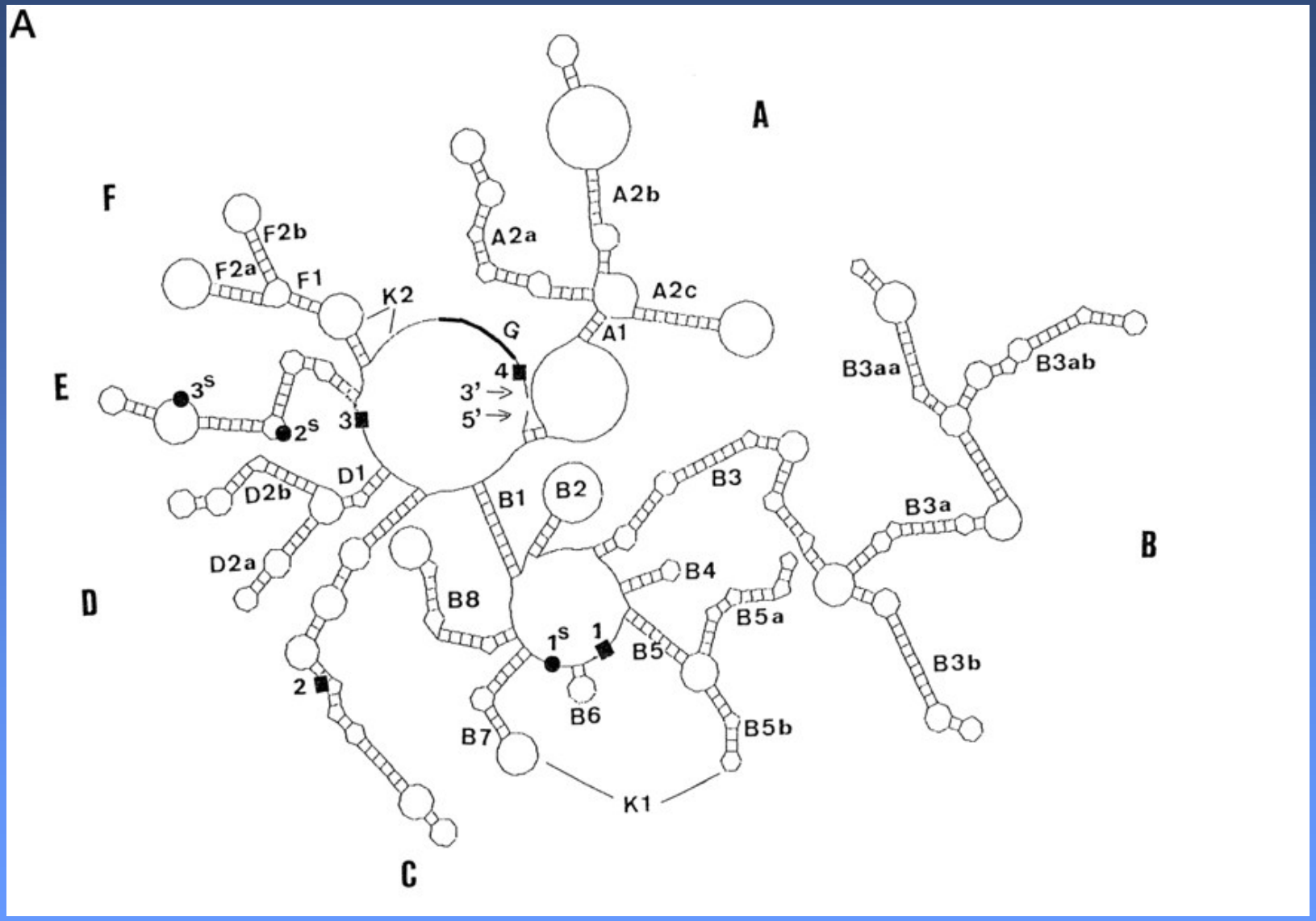


Platelet derived growth factor(PDGF) provides the interesting first example of a differentiation specific IRES

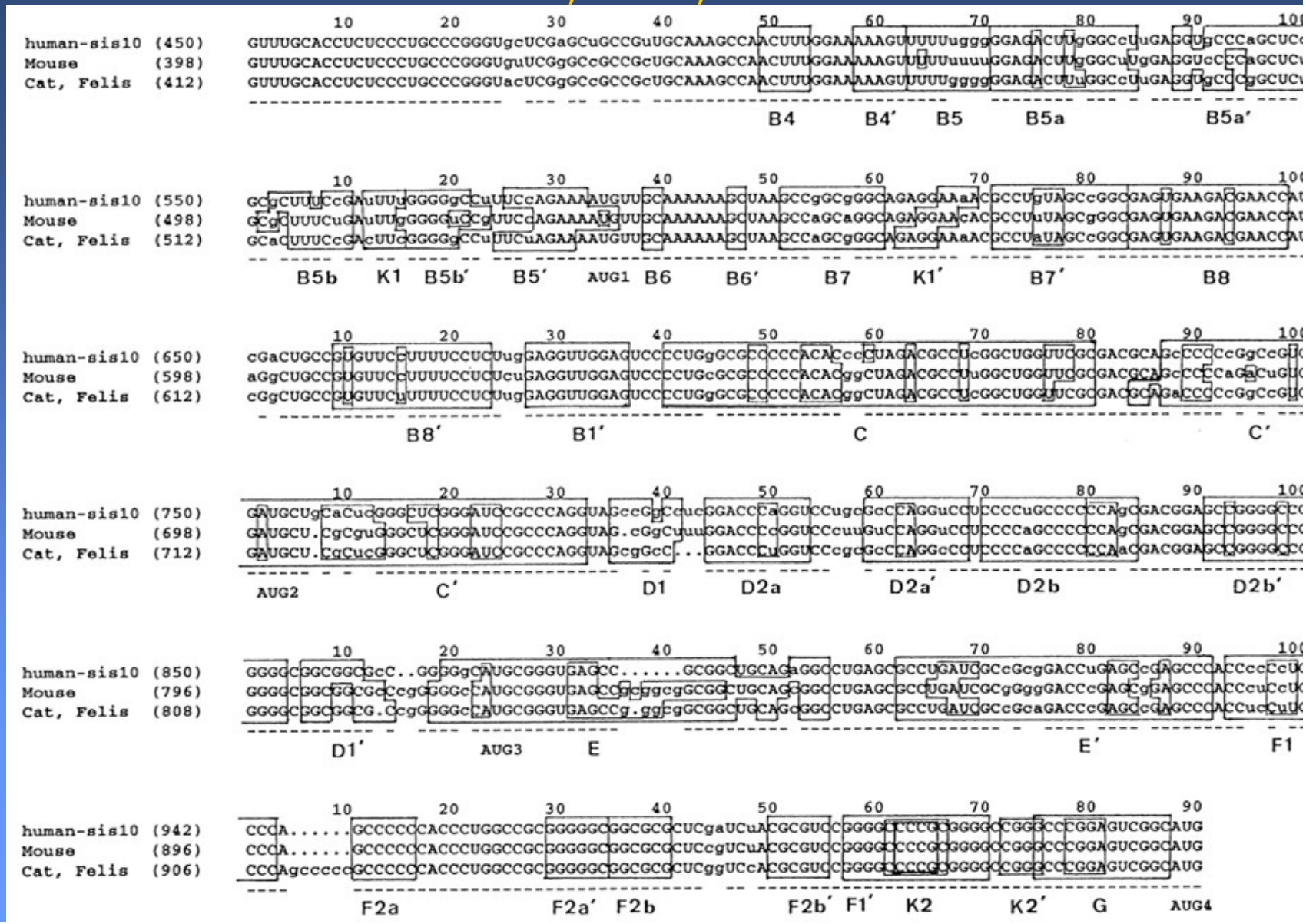
This is a collaboration between the laboratory of Orna Elroy-Stein at Tel Aviv University and Le of the LECB.

Normally the 5’UTR inhibits translation, but during differentiation of megakaryocytes in bone marrow, or in phorbol ester stimulation, translation is elevated.

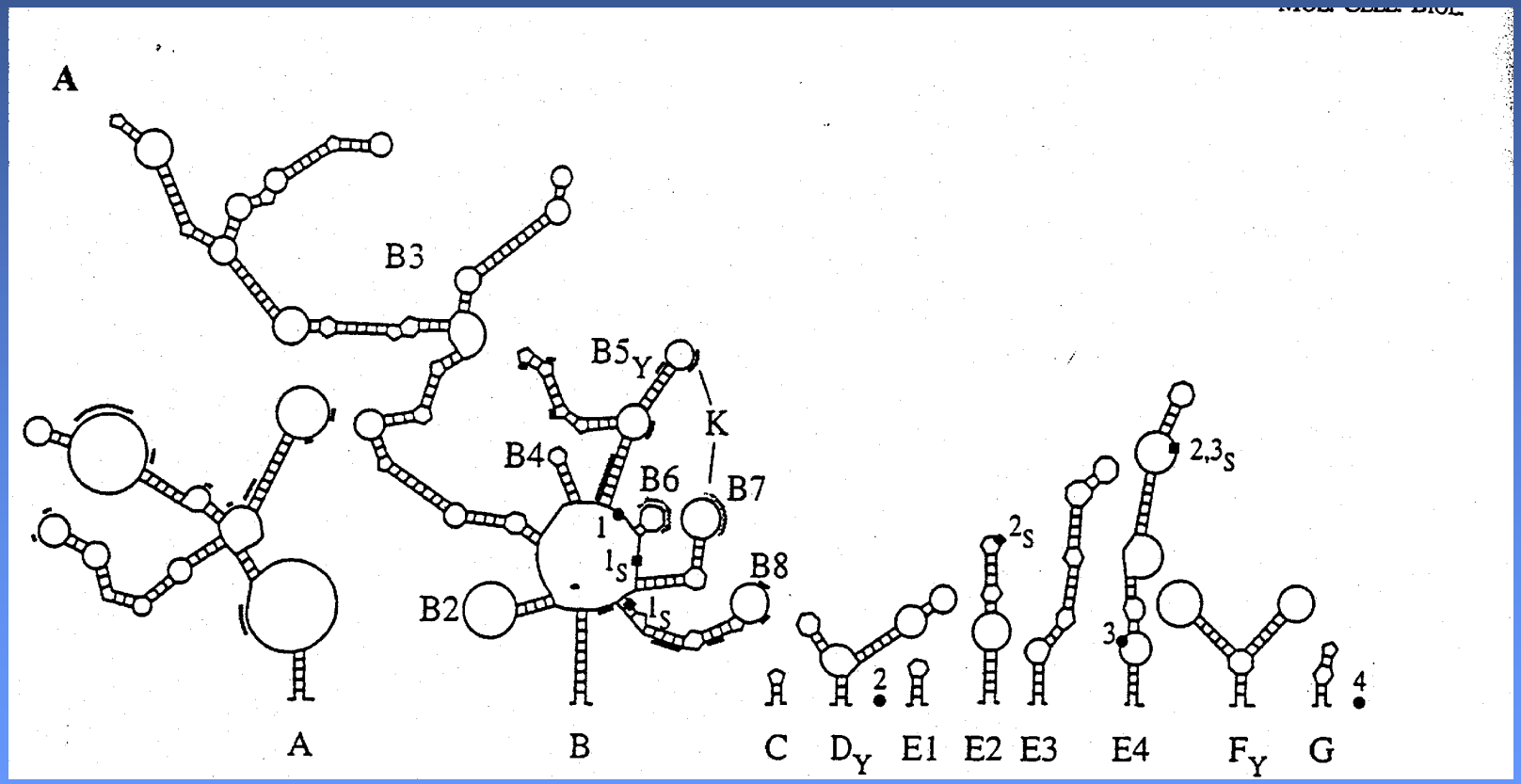
Secondary structure model of the 5-UTR of human PDGF



Sequence alignment of the 3′ portion of PDGF2/c-*sis* 5-UTR. Sequences of human, mouse, and cat



PDGF stem-loops tested for IRES function by deletion mutations



C-sis mutants

	nt. no.	
1	(1-1022)	5' A B2 B3 B4 Y_{B5} K B6 B7 B8 C Y_D E1 E2 E3 E4 Y_F G 3'
2	(1-227)	A B2
3	(215-497)	B3
4	(601-1022)	B7 *B8* C *Y_D E1 E2* E3 *E4 Y_F* G
5	(475-649)	B4 Y_{B5} K* *B6* B7
6	(475-1022)	B4 Y_{B5} K* *B6* B7 *B8* C *Y_D E1 E2 E3* E4 *Y_F* G
7	(475-685)	B4 Y_{B5} K *B6* B7 *B8*
8	(1-685)	A B2 B3 B4 Y_{B5} K B6 B7 B8
9	(601-846)	B7 B8 *C* Y_D E1 E2 E3
10	(475-797)	B4 *Y_{B5} K* B6 B7 B8 *C* Y_D *E1 E2*
11	(215-846)	B3 B4 *Y_{B5} K* B6 B7 B8 *C* Y_D E1 E2 E3
12	(1-846)	A B2 B3 B4 *Y_{B5} K* B6 B7 B8 *C* Y_D *E1 E2 E3*
13	(215-1022)	B3 B4 *Y_{B5} K* B6 B7 B8 *C* Y_D *E1 E2 E3* E4 *Y_F* G
14	(862-1022)	E4 Y_F *G*

Table 2 Effect of Differentiation on IRES Activity Conferred by Truncated PDGF2/c-sis 5'UTR and their RNA Structural Motif

Fragment	D-IRES (%)	RNA Structural motif
S1(1-1020)	100	B5(512—581), D(700—750) and F(939—993) B5, D and F are Y-shape structure followed by AUG and a pseudoknot K (UUUU//AGAG).
S2 (1-848)	53	D (Y-shape structure followed by AUG) and B5 Y-shape structure was destroyed.
S3 (1-685)	77	B5 (Y-shape structure followed by AUG) and pseudoknot K.
S4 (1-215)	0	No Y-shape structure motifs.
S5(215-1020)	46	B5 (Y-shape structure followed by AUG) and pseudoknot K. No Y-shape structures of D and F.
S6(475-1020)	0	B5 (Y-shape structure followed by AUG), and a changed pseudoknot K.
S7(601—1020)	0	No Y-shape structure motifs of D and F.
S8(861-1020)	37	F (Y-shape structure followed by AUG) and base complementarity between the upstream of AUG and 18S rRNA.
S9 (475-685)	58	B5 (Y-shape structure followed by AUG) and a pseudoknot K.
S10(601-846)	71	D (Y-shape structure followed by AUG)
S11(215-497)	0	No Y-shape structure motifs
S12(475-796)	72	D (Y-shape structure followed by AUG)
S13(215-846)	96	B5 and D (Y-shape structure followed by AUG) and a pseudoknot K.
S14(475-649)	0	B5 (Y-shape structure followed by AUG), and a changed pseudoknot K.

Today's talk(continued)

Chapter 1:
Computer prediction of RNA structure

Chapter 2:
Application of computer tools to 5'UTRs of virus and cellular mRNAs

Chapter 3:
Discovery of novel features in a long 3' UTR

Known patterns in UTRs

Table 2. Functional patterns included so far in UTRsite (v3.0)

Functional patterns	Reference	Hits found in UTRdb 12.0
Iron-responsive element (IRE)	23	65
Histone 3′UTR stem–loop structure	24	27
AU-rich class II destabilising element	25	175
TGE translational regulation element	26	45
Selenocysteine insertion sequence (SECIS)	27,28	189
APP 3′-UTR stability control element	29	7
Cytoplasmatic polyadenylation element (CPE)	30	4614
Nanos	31	397
ribosomal protein mRNA 5′ TOP	32–34	298
TNF mRNA translation repression element	35	14
Vimentin 3′UTR mRNA element	36	12
GLUT1 mRNA stabilising element	37	48
15-LOX-DICE	38	83
Repetitive elements		44 806

For each pattern the number of hits with UTRdb entries is also reported.

Function and structure in a 3’UTR from *Rana pipiens*

This is from collaboration with the Rybak lab where cytotoxic RNAse-type proteins are being investigated as anti-tumor immuno-conjugates.

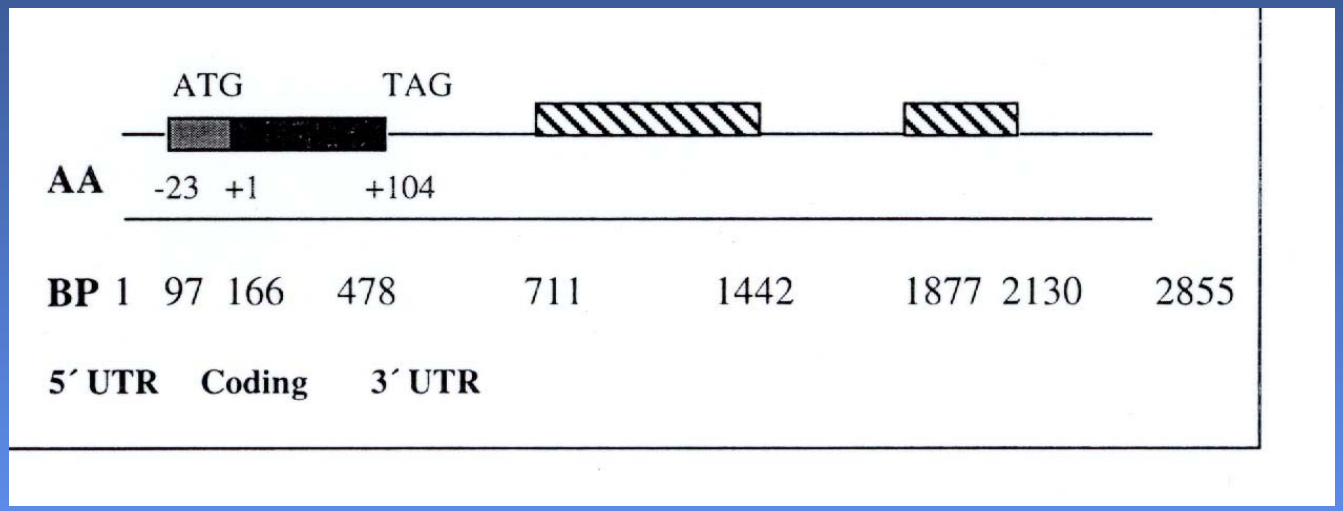
While cloning the cytotoxic RNAse an mRNA was found that had a very long 3’UTR of 2377nt.

Although present as an mRNA in the liver in high levels, it was not translated.

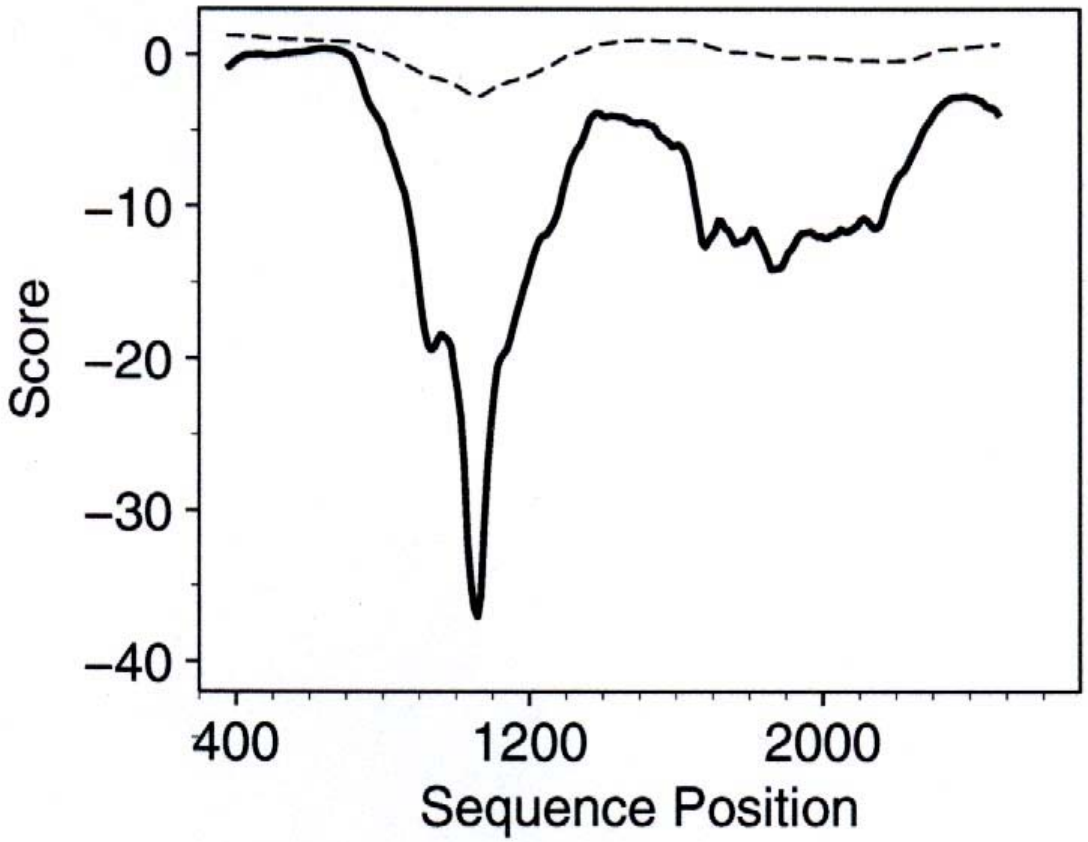
When the 3’UTR was truncated it was efficiently translated *in vitro*.

This led to examining the sequence for potential structures.

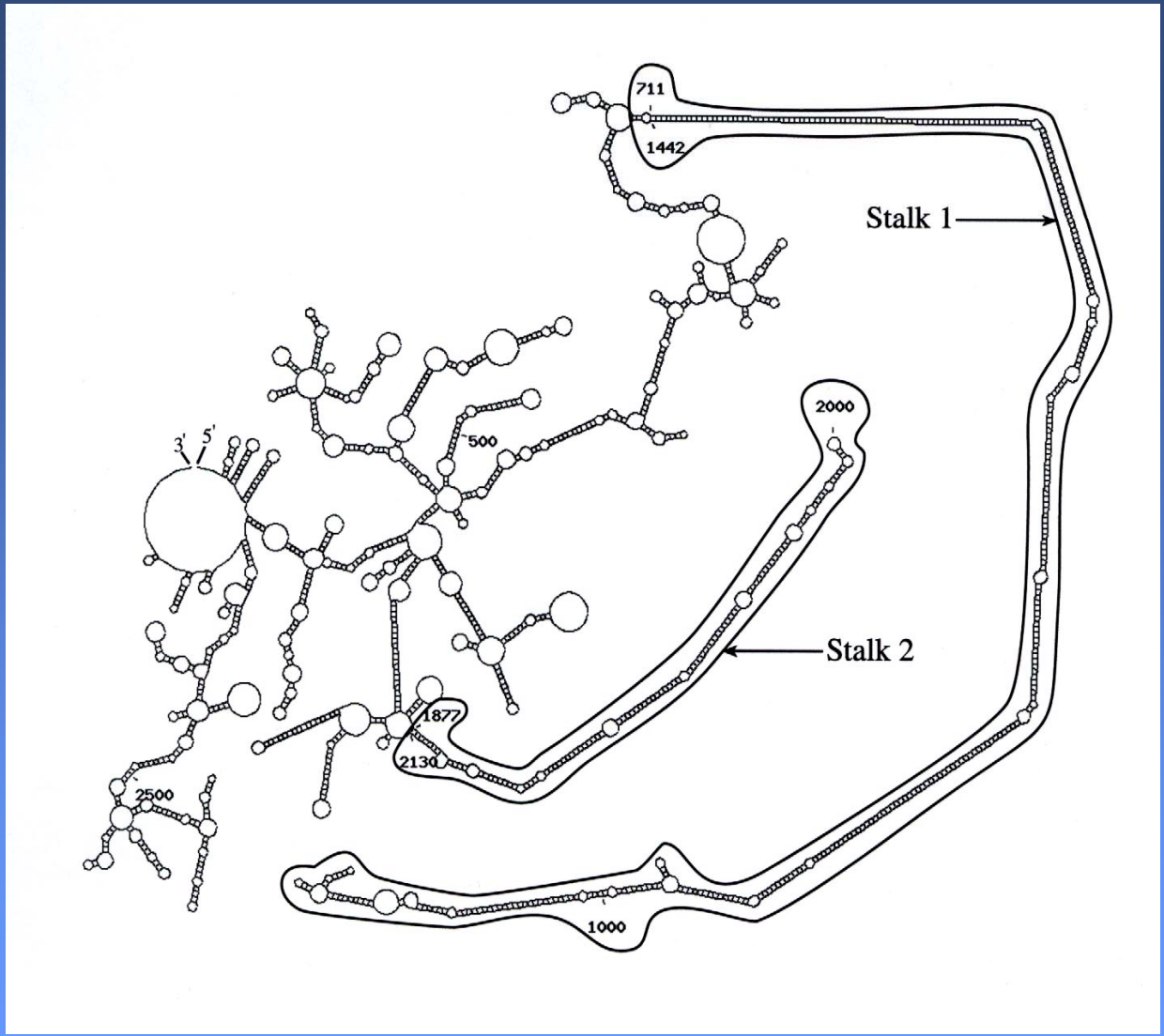
*rap*LR1 map


ATG
TAG
AA
-23 +1
+104
BP 1 97 166 478 711 1442 1877 2130 2855
5´ UTR
Coding
3´ UTR

*rap*LR1 z-score



*rap*LR1 stalk-like UFRs



Extension to other 3'UTRs

Work will continue on *rap*LR1 RNA.

Meanwhile we used the structure of the stalk to create a template for the RNAMOT motif searching program.

There are more UTRs to be studied

There a number of known mRNAs with long UTRs.

The Pesole *et al* database (UTRdb12.0) reports 38498 5'- and 46881 3'-UTRs, and counting.

http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/

The observation and database of *Drosophila* by Cavener and Cavener suggests that more than 40% of mRNAs have long 5'UTRs. This pre-dates the availability of the complete genome.

Distribution of 3'UTR Lengths for Sequences in the UTR-DB (Release 7.0, May 1998)

Table 1

Files	Total No. of Sequence	No. of Seq (Len > 100-nt)	No. of Seq (> 200-nt)	No. of Seq (> 300-nt)	No. of Seq (> 600-nt)	Mean of Length
Hum_3utrnr.dat (Human)	7503	6767	5726	4813	3062	738-nt
Om_3utrnr.dat (Mammalian)	2457	2021	1513	1183	636	477-nt
Ro_3utrnr.dat (Rodent)	7633	6783	5670	4614	2687	620-nt
Ov_3utrnr.dat (Vertebrate)	3499	3008	2306	1817	959	485-nt
In_3utrnr.dat (Invertebrate)	5067	3841	2551	1917	989	396-nt
Pl_3utrnr.dat (Plant)	8116	7574	4617	1585	192	242-nt
Fun_3utrnr.dat (Fungal)	1154	904	492	278	74	248-nt

Table 2a. Statistical significance of the unusual folding region (UFR) enclosed the stalk-like stem-loop structure in the 3'UTR of 7503 human mRNAs

mRNA Sequences	UFR location in mRNA and (in 3'UTR)	SIGSCR (Mean and STD) (in corresponding mRNA)	STBSCR	Base composition in UFR A	G	C	U(%)
pH-sensing regulator	1097-1703 (404-1010)	-10.80 (-1.98, 1.852)	-4.63	24.1	24.5	23.9	27.5
KIAA0432 gene	4910-5702 (2658-3450)	-32.76 (-2.40, 5.903)	-5.06	27.5	22.1	19.3	31.1
VHL gene	3224-4021 (2510-3307)	-25.56 (-6.88, 6.171)	-2.69	23.4	26.3	25.2	25.1
AIK kinase	1477-1804 (99-426)	-39.84 (-4.03, 9.687)	-3.81	29.6	20.7	20.4	29.3
alpha (1, 2) fucosyltransferase	1532-2312 (389-1169)	-7.35 (-1.96, 1.355)	-0.47	28.0	22.5	23.7	25.7
Adipogenesis inhibitory factor	1473-2016 (810-1353)	-8.07 (-2.73, 1.751)	-1.40	27.6	26.3	22.6	23.5
Ring zinc-finger protein	2368-3147 (568-1347)	-12.25 (-2.24, 2.200)	-1.52	24.1	25.6	19.5	30.8
Obese protein	1061-1655 (501-1095)	-16.02 (-2.77, 2.182)	-4.06	22.2	28.2	19.2	30.4
KIAA0349 gene	4557-4847 (729-1019)	-7.0 (-0.97, 1.402)	-2.04	28.5	17.2	13.7	40.5
GM2 activator protein	1329-1928 (689-1288)	-6.59 (-2.0, 1.709)	-2.08	25.0	22.3	24.0	28.7
KIAA0186 gene	1893-2679 (1209-1995)	-12.40 (-3.73, 4.313)	-2.36	22.5	22.5	19.7	35.3
MACH-alpha-1 protein	1994-2449 (263-718)	-15.29 (-2.05, 2.562)	-3.97	29.8	17.5	19.7	32.9
NAD+ dependent 15-hydroxyprostaglandin dehydrogenase	2078-2412 (1260-1594)	-31.08 (-1.75, 3.031)	-6.34	34.9	15.8	14.3	34.9

Table 2b. Statistical significance of the unusual folding region (UFR) enclosed the stalk-like stem-loop structure in the 3'UTR of mammalian (2457 mRNAs) and rodent (7633 mRNAs)

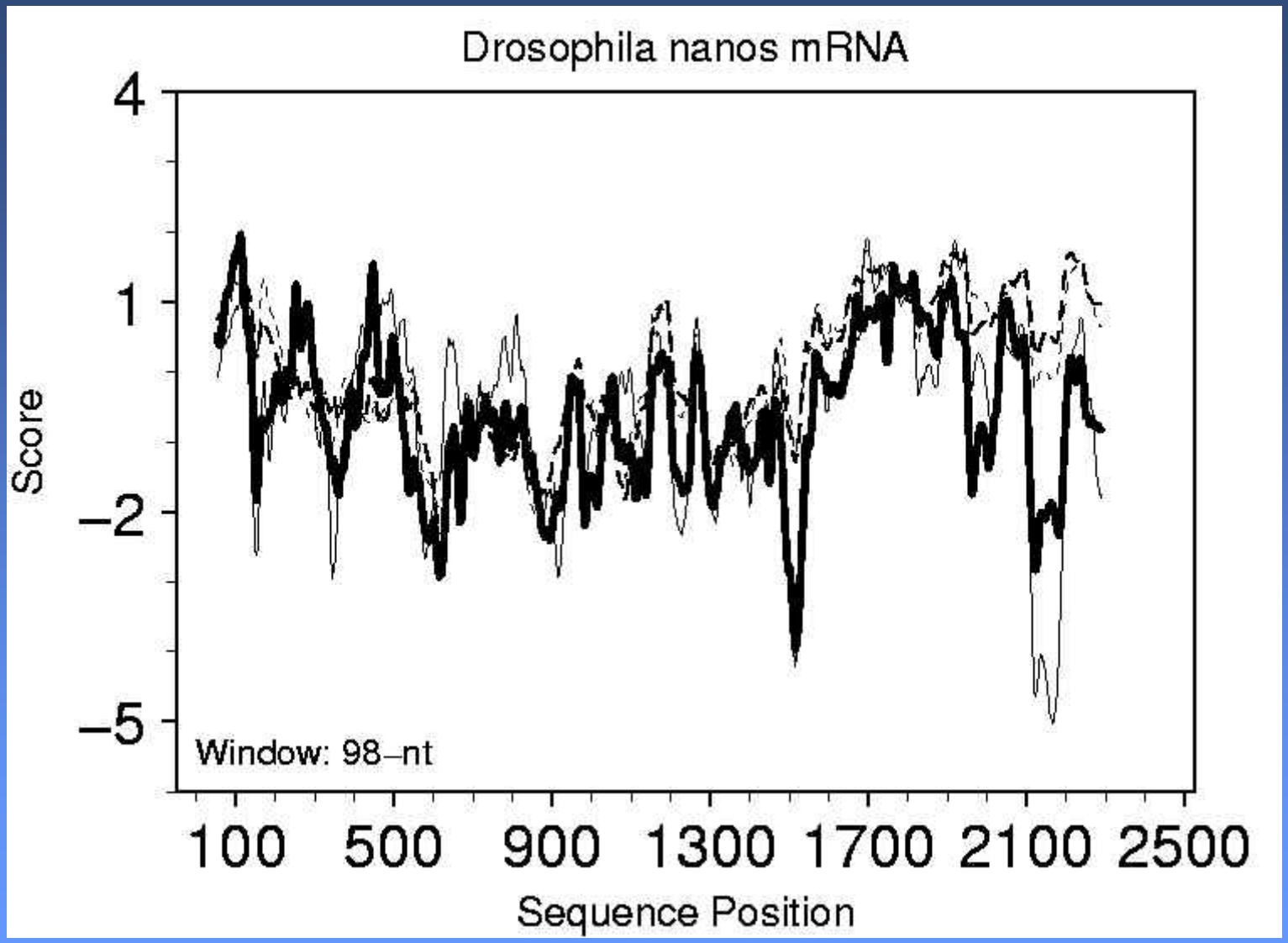
mRNA Sequences	UFR location in mRNA and (in 3'UTR)	SIGSCR (Mean and STD) (in corresponding mRNA)	STBSCR	Base composition in UFR A	G	C	U(%)
Bovine ACTH receptor	1314-1998 (295-979)	-11.58 (-1.91, 1.958)	-4.47	25.5	23.8	23.9	26.7
Bovine cyclase OS-GC2	4476-5101 (837-1462)	-17.66 (-1.99, 2.790)	-4.48	23.6	25.9	24.9	25.6
Dog desmosome	3151-3442 (775-1066)	-22.68 (-0.46, 2.871)	-6.08	36.0	16.1	17.8	30.1
Rodent							
Mouse NDPK-A	702-886 (203-387)	-22.28 (-1.38, 3.653)	-5.83	30.3	18.4	20.0	31.4
Mouse caspase-3	1073-1297 (172-396)	-32.83 (-1.76, 6.036)	-5.36	34.7	15.6	14.2	35.6
Mouse rds protein	2012-2345 (759-1092)	-13.32 (-1.18, 1.916)	-5.20	21.9	29.3	25.4	23.4
Rat collagenase	2878-3228 (416-766)	-22.80 (-0.74, 2.689)	-6.19	23.1	23.9	27.9	25.1

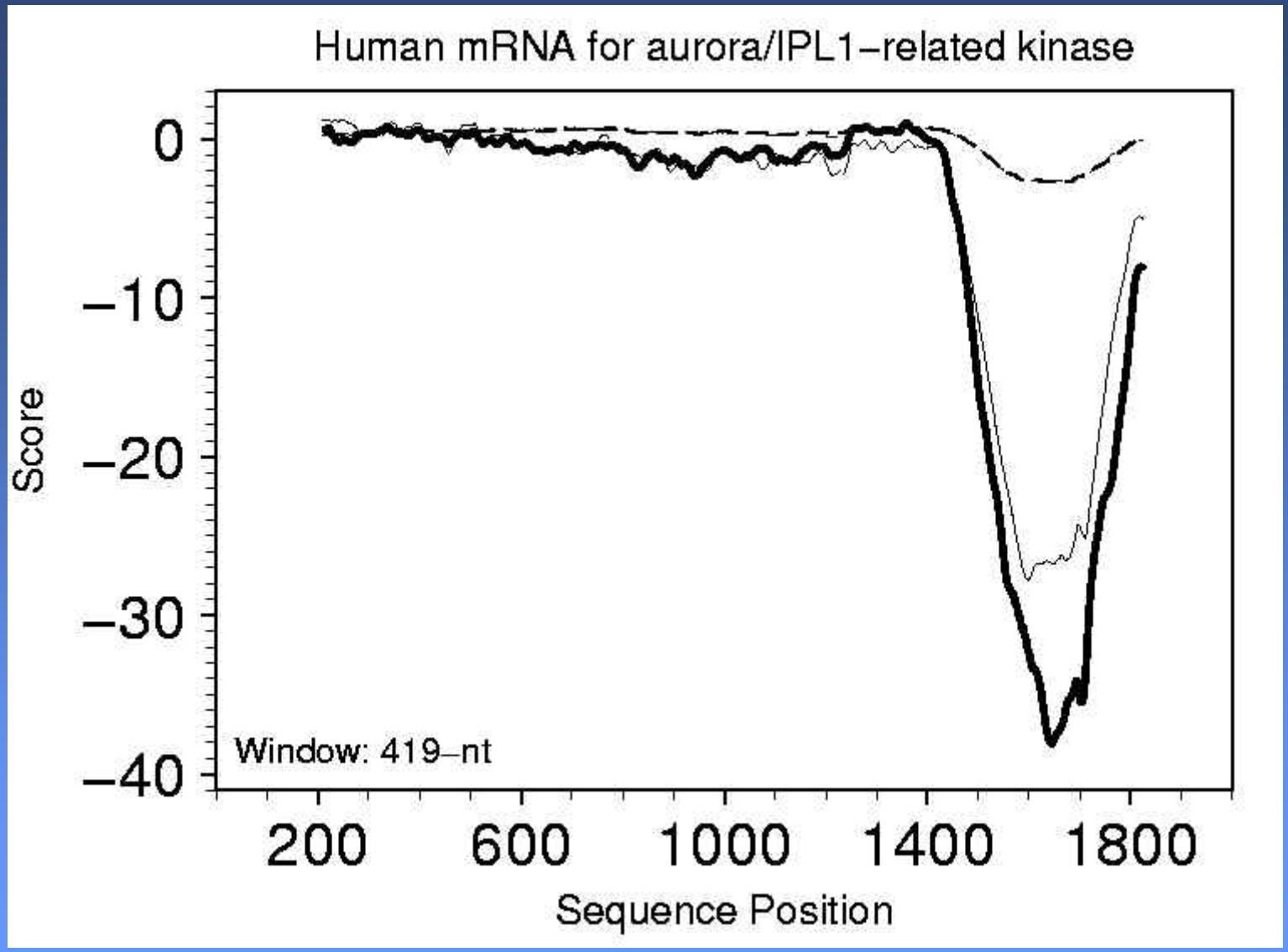
Table 2c. Statistical significance of the unusual folding region (UFR) enclosed the stalk-like stem-loop structure in the 3'UTR of 3499 Vertebrate mRNAs

mRNA Sequences	UFR location in mRNA and (in 3'UTR)	SIGSCR (Mean and STD) (in corresponding mRNA)	STBSCR	Base composition in UFR A	G	C	U(%)
Goldfish aromatase	1781-2290 (162-671)	-13.42 (-2.09, 2.709)	-1.89	30.2	21.2	16.7	32.0
Zebrafish PAX7A	1679-1933 (499-753)	-18.35 (-1.25, 3.315)	-4.61	30.6	17.6	18.0	33.7
Goldfish CYPXIX aromatase	1836-2286 (184-634)	-21.13 (-2.04, 3.514)	-3.13	27.3	23.1	16.2	33.5
Rainbow trout p53	1443-1918 (45-520)	-14.47 (-1.75, 3.086)	-2.33	30.0	18.3	17.6	34.0
Torpedo marmorate vesamicol binding protein	2063-2358 (287-582)	-16.12 (-1.17, 3.995)	-4.41	19.9	32.8	21.3	25.7
Xenopus transcription factor A	1218-1581 (145-508)	-29.37 (-3.66, 6.942)	-3.46	17.9	33.0	18.4	30.8
Xenopus thymocyte marker CTX	1328-1936 (295-903)	-30.98 (-3.92, 6.353)	-3.36	28.4	22.2	16.4	33.0
Xenopus DNA polymerase gamma	3904-4348 (89-533)	-20.41 (-1.64, 3.567)	-5.51	22.2	27.2	18.9	31.5
Xenopus myosin light chain	686-1199 (140-653)	-28.14 (-6.17, 8.845)	-2.44	24.1	25.1	18.9	31.9
Xenopus laevis tanabin	6266-6778 (954-1466)	-12.38 (-1.37, 1.846)	-4.14	32.6	18.5	23.4	25.5
Xenopus Xtl1	2299-2552 (676-929)	-26.52 (-1.96, 4.011)	-5.63	19.7	28.7	22.0	29.5

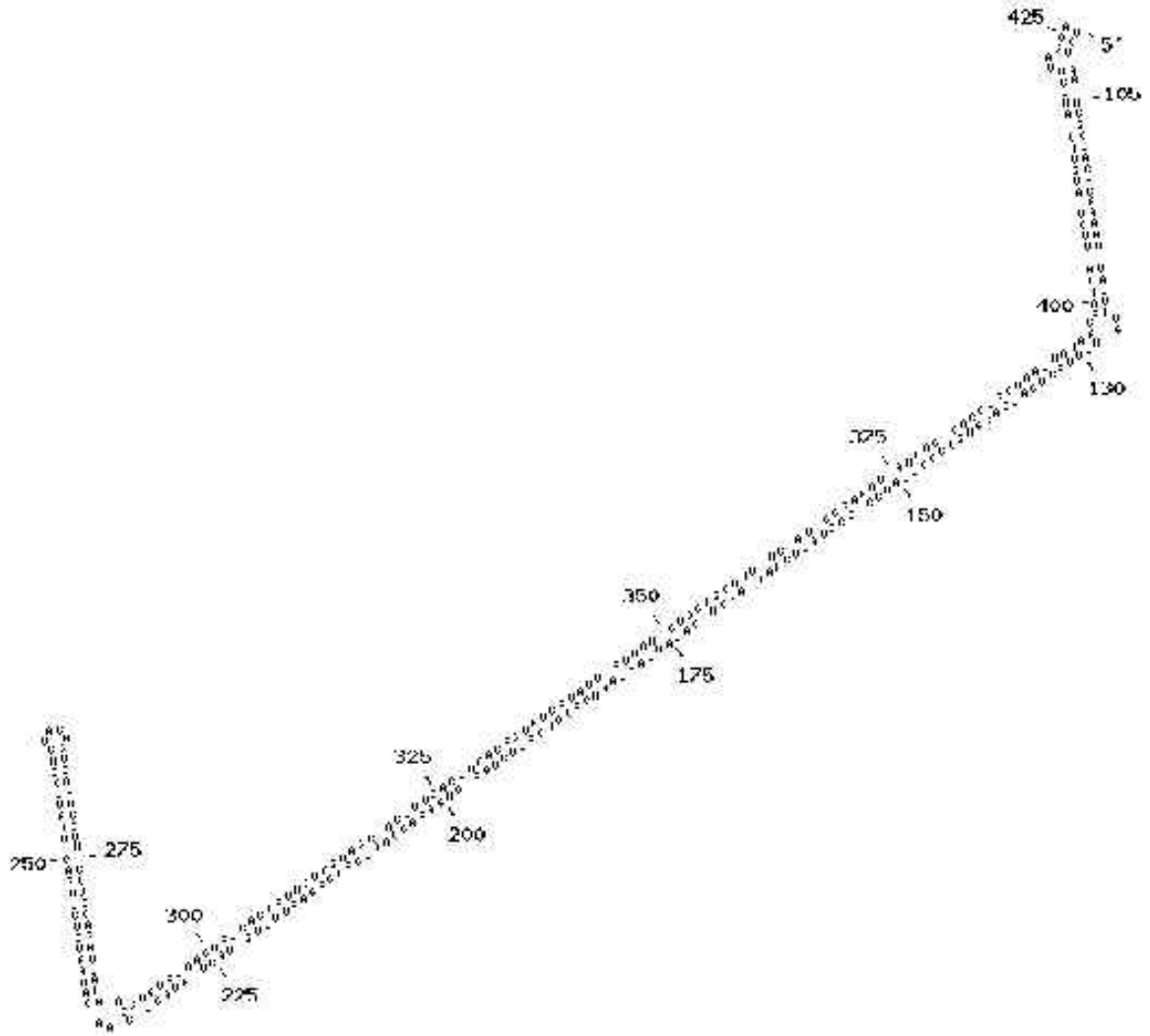
Table 2d. Statistical significance of the unusual folding region (UFR) enclosed the stalk-like stem-loop structures in the 3'UTR of 5067 invertebrate, 8116 plant and 1154 fungi mRNAs

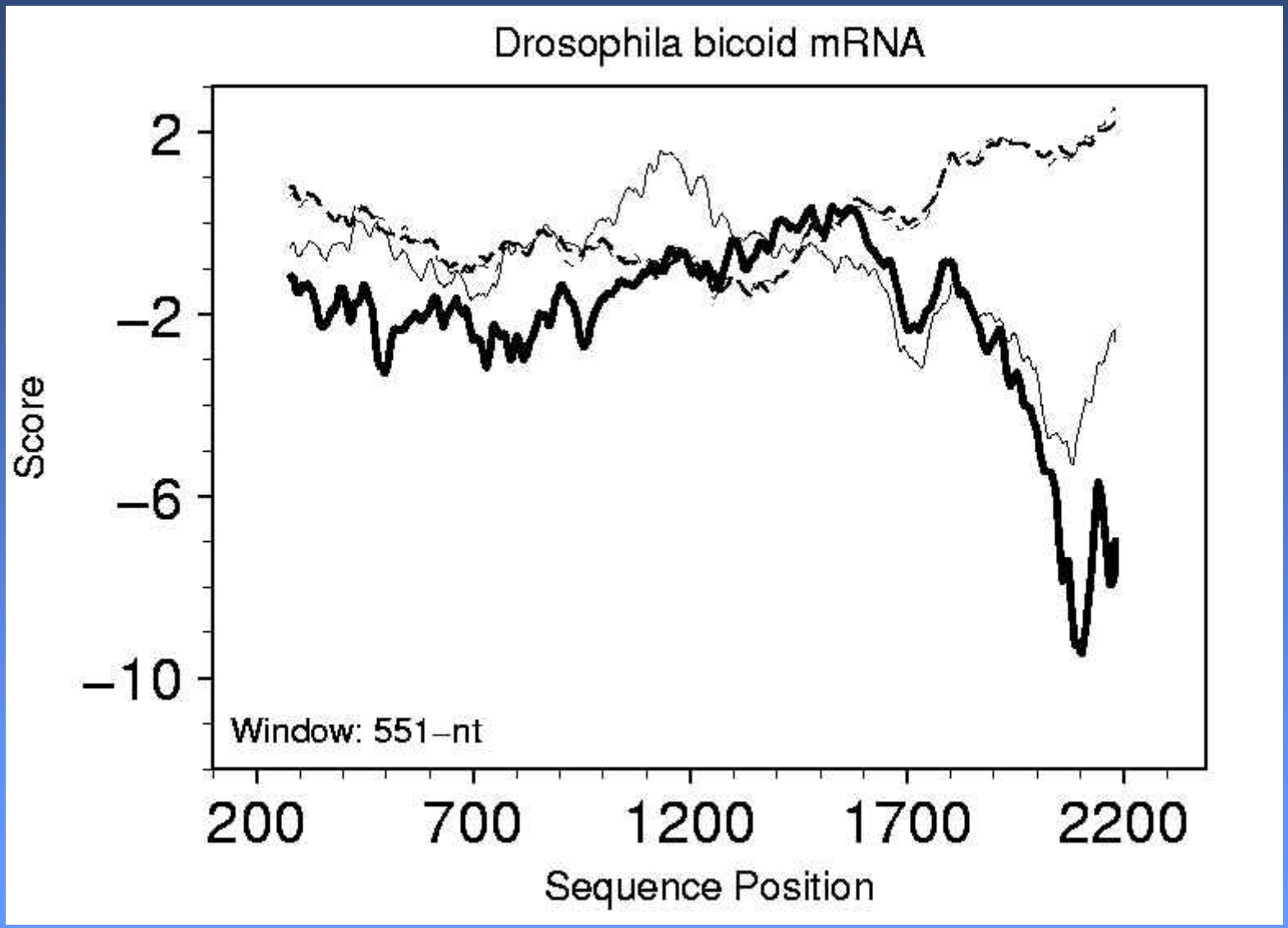
mRNA Sequences	UFR location in mRNA and (in 3'UTR)	SIGSCR (Mean and STD) (in corresponding mRNA)	STBSCR	Base composition in UFR A	G	C	U(%)
Invertebrate							
Lancelet Phe hydroxylase	1657-2411 (339-1093)	-75.05 (-11.66, 19.704)	-3.34	33.0	17.2	15.8	34.0
Bombyx mori invected homeodomain	1763-2499 (82-818)	-31.27 (-4.91, 9.917)	-3.70	25.5	22.0	22.0	30.5
Bombyx mori lysozyme	506-680 (54-228)	-15.98 (-1.09, 2.696)	-3.97	32.6	19.4	17.7	30.3
C. elegans pha-1	1672-1999 (165-492)	-14.15 (-2.03, 2.726)	-4.99	36.0	17.1	16.5	30.5
C. elegans tra-1	3643-3951 (151-459)	-10.33 (-0.73, 1.829)	-4.13	24.6	26.5	25.6	23.3
S. franciscanus bindin	1739-2296 (88-645)	-35.11 (-1.89, 5.881)	-4.57	27.1	24.0	16.5	32.4
Plant							
Fava bean pyrophosphorylase	1617-1944 (1-328)	-34.25 (-2.19, 6.648)	-4.88	32.6	18.9	14.6	33.8
Fungi							
Yeast virus 1 internal repeat	75-594 (1-520)	-20.07 (-7.71, 6.233)	-2.06	25.6	22.9	25.6	26.0

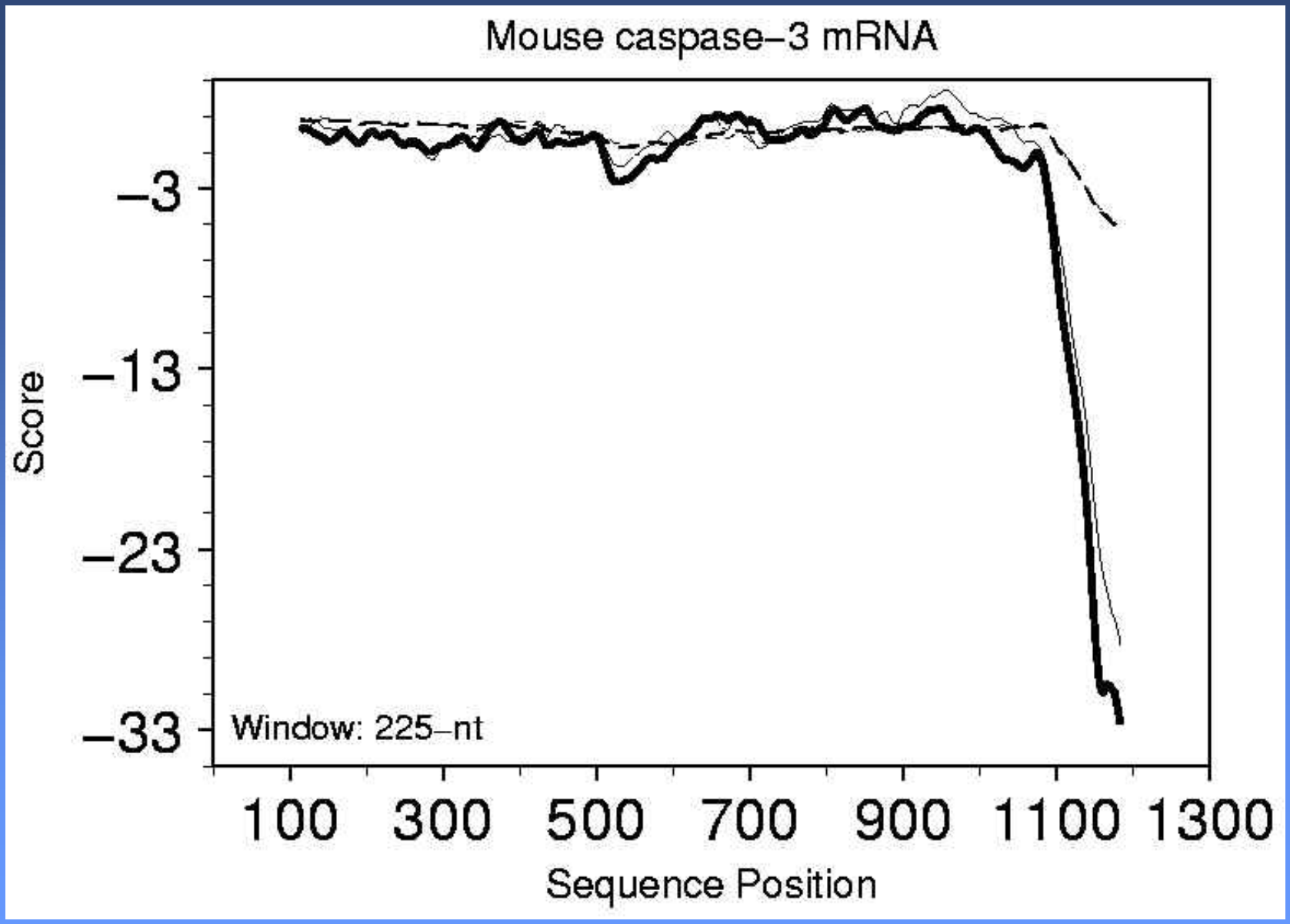

Drosophila nanos mRNA
4
1
−2
−5
Score
Window: 98−nt
100
500
900
1300
1700
2100
2500
Sequence Position

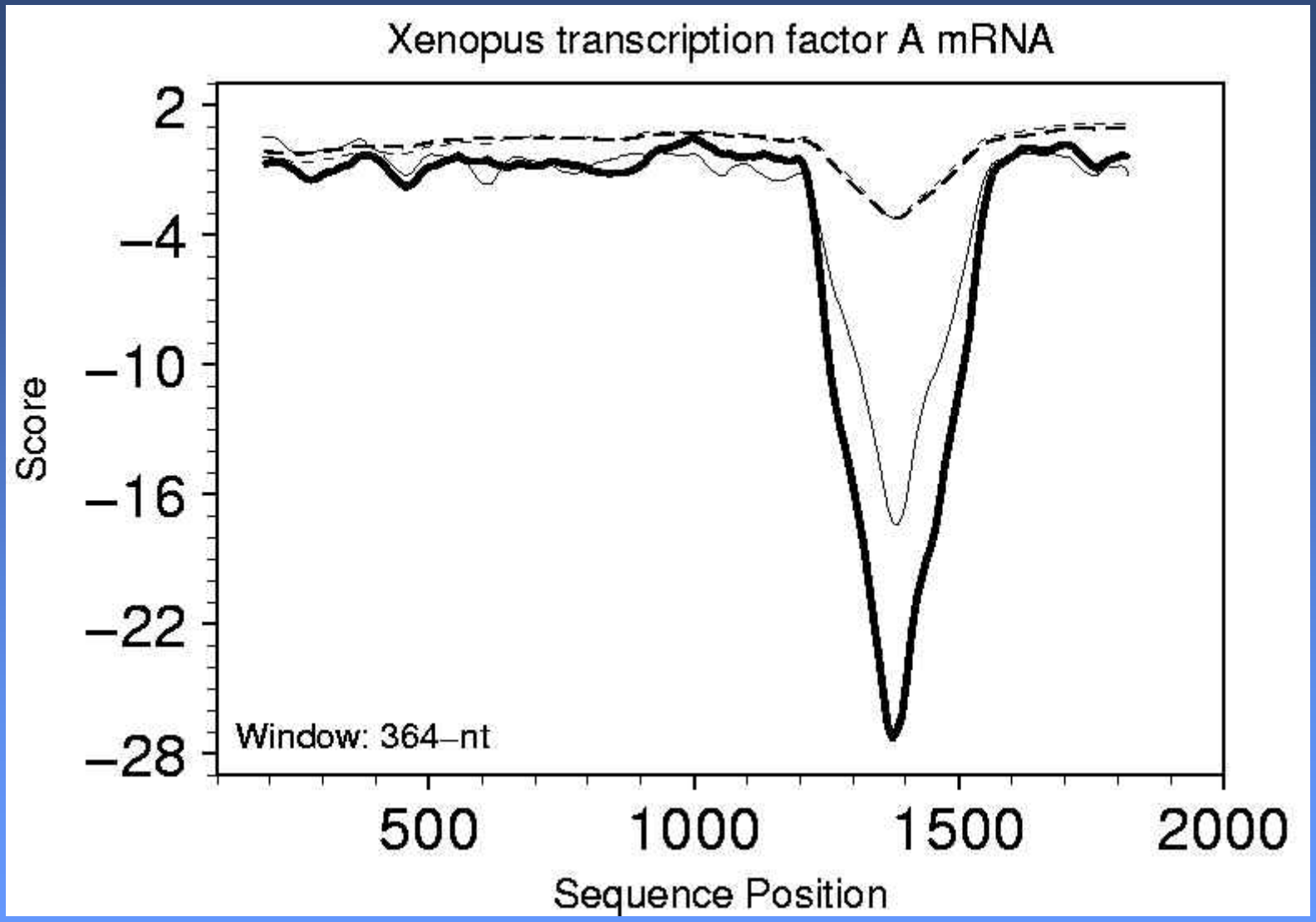

Human mRNA for aurora/IPL1−related kinase
Score
0
−10
−20
−30
−40
Window: 419−nt
200
600
1000
1400
1800
Sequence Position

Stalk-like stem-loop in the 3'UTR of human AIK mRNA (aurora/IPL1-related kinase), D84212

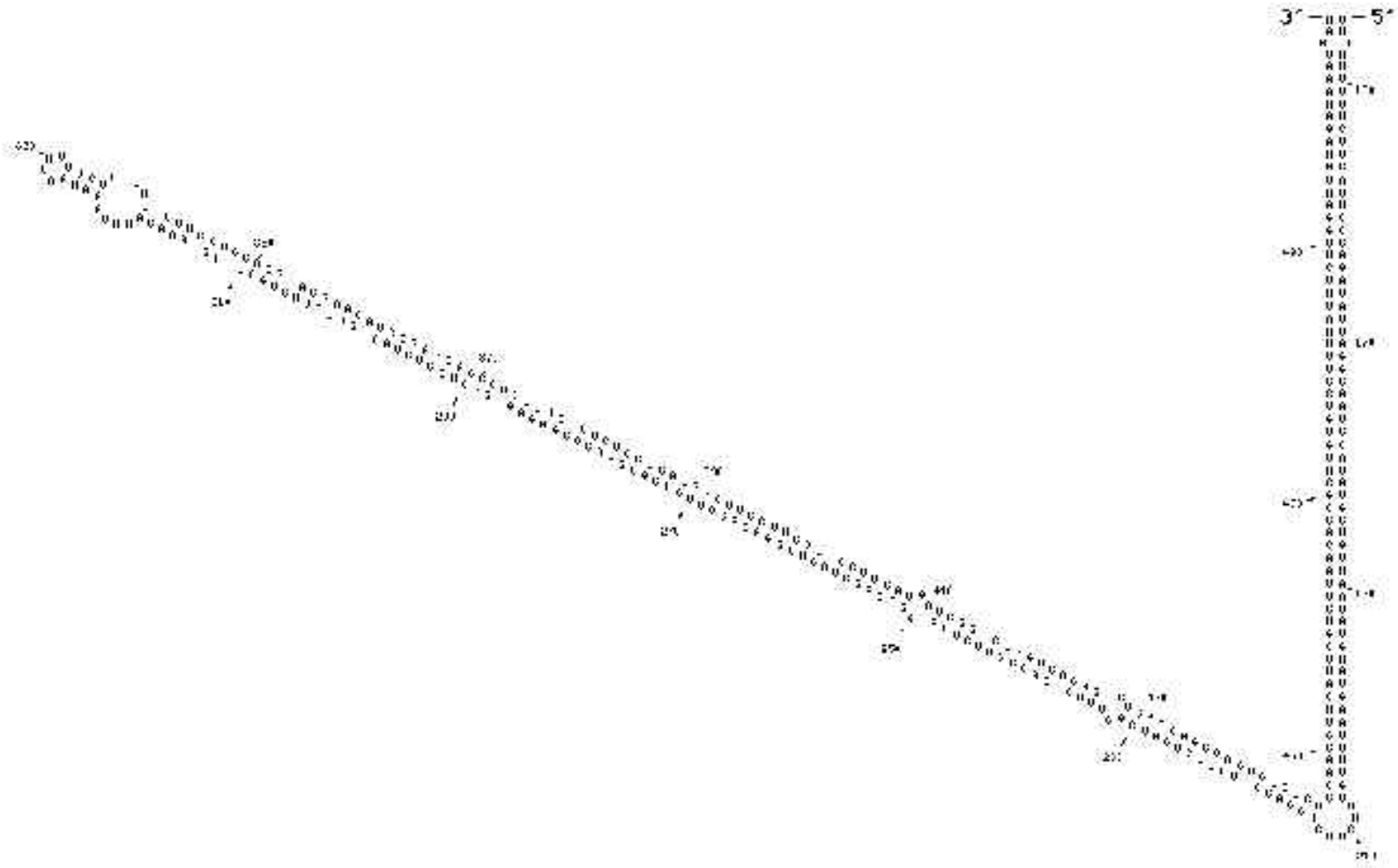


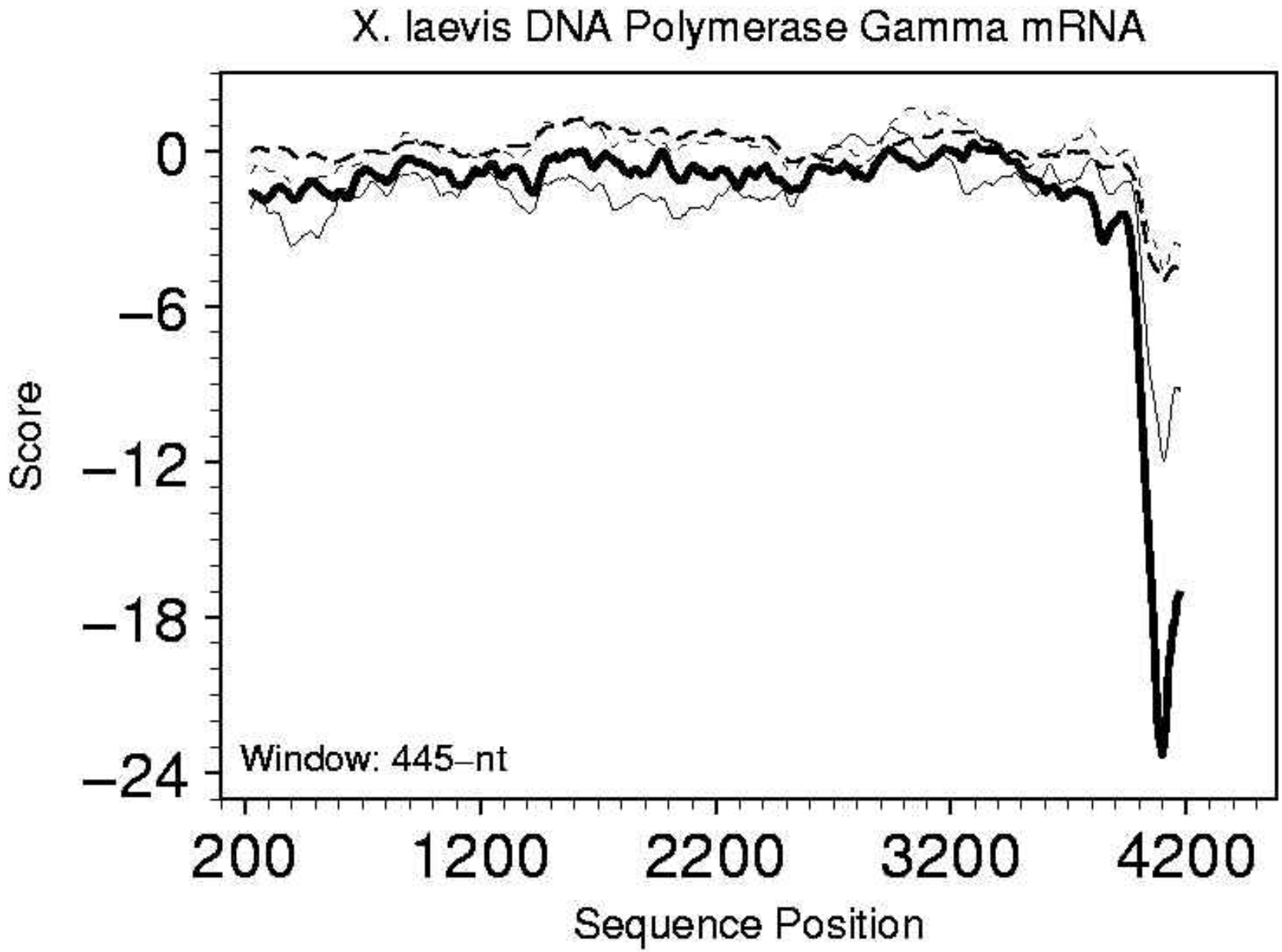

Drosophila bicoid mRNA
Score
2
−2
−6
−10
Window: 551−nt
200
700
1200
1700
2200
Sequence Position

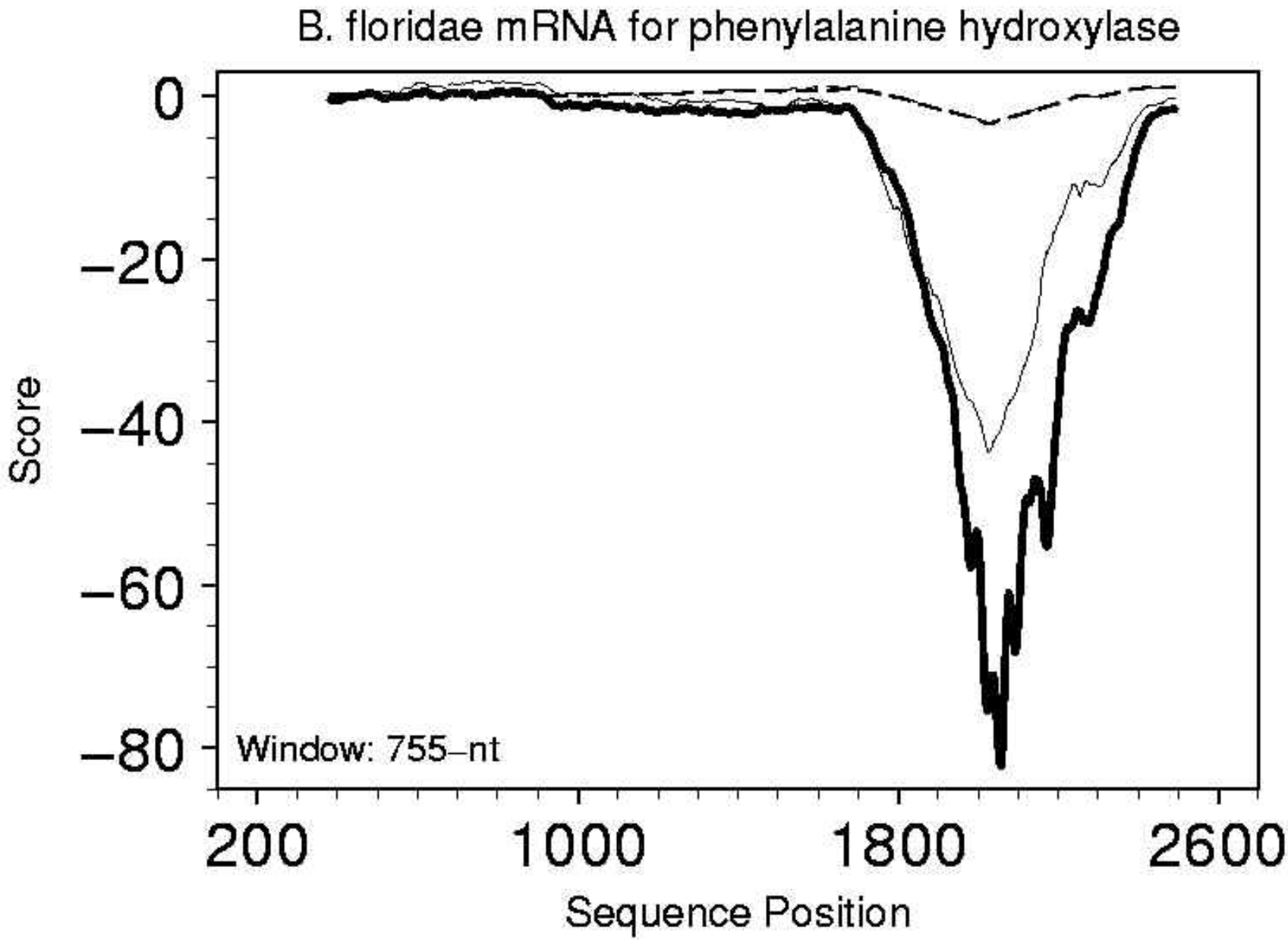

Mouse caspase−3 mRNA
Score
−3
−13
−23
−33
Window: 225−nt
100
300
500
700
900
1100
1300
Sequence Position


Xenopus transcription factor A mRNA
Score
2
−4
−10
−16
−22
−28
Window: 364−nt
500
1000
1500
2000
Sequence Position

Stalk-like stem-loop in the 3'UTR of X. laevis transcription factor A mRNA, U35728

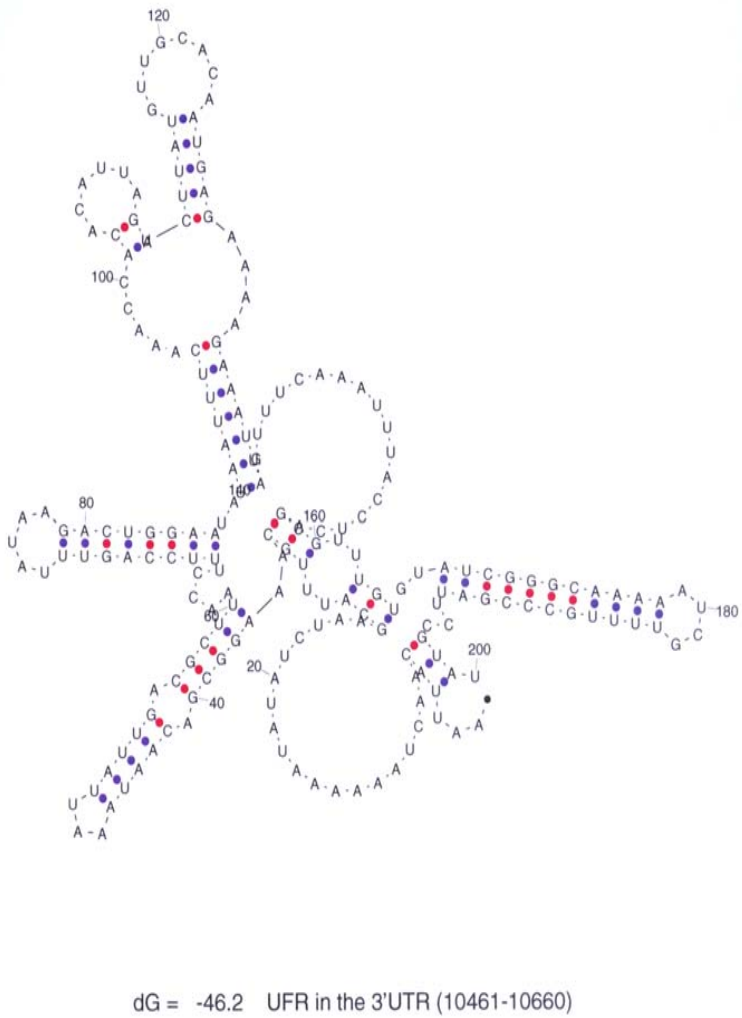



X. laevis DNA Polymerase Gamma mRNA
0
−6
−12
−18
−24
Score
Window: 445−nt
200
1200
2200
3200
4200
Sequence Position


B. floridae mRNA for phenylalanine hydroxylase
0
−20
−40
−60
−80
Score
Window: 755–nt
200
1000
1800
2600
Sequence Position

Minimum energy folding of the highly significant UFR in the 3'-untranslated region of BRCA2 mRNA

This structure has a highly significant score of 2.65 when tested for non-random morphology by ST_COMP.



dG = -46.2 UFR in the 3'UTR (10461-10660)

Combinined phylogenetic alignment and energy minimization in RNAGA

Obtaining a refined higher order structure

EFFOLD is used to explore alternative structures by producing maximally stable structures when the energy rules are perturbed slightly. Zuker's mfold can also produce sub-optimal structures.

COMFOLD examines the covariation in paired bases between foldings of phylogenetically related molecules or alternative energy foldings to produce a "consensus" folded structure.

RNAKNOT looks for favorable potential pairings between the unpaired bases between 2 loops or between a loop and an open strand. These are often referred to as pseudoknots.

RNAMOT(Gautheret and Laferriere) can search for related structural motifs in data bases using patterns derived from the predicted structures.

Prediction of common secondary structures of RNAs: a genetic algorithm approach
Jih-H. Chen, Shu-Yun Le, and Jacob V. Maizel
Nucleic Acids Research 2000, Vol 28, No.4, 991-999.

Predicts a common RNA secondary structure without knowing or finding the alignment of sequences using a genetic algorithm(GA) approach.

GAs operate on a population of tentative solutions in an encoded representation from a list of potential stems in each sequence.

Randomly change some solutions(A mutation) and recombine parts of solutions(GA crossover), then select survivors by scoring function(energy and structural conservation).

Rank the solutions by a score closely related to the selection function.

In four test cases(20 tRNAs, 25 5S rRNAs, seven *rev* response elements of HIV-1, and 10 *rev* response elements of HIV-2) fairly convincing common structures were obtained by RNAGA in the top 10 ranked solutions.

It is compute intensive. 2, 9.5, 30, and 40 hrs on a Dec alpha processor were needed, respectively, for the test cases.

Memory requirements are $O(n^2m^2N^2)$.n = number of all stems, m = max number of structures among N sequences.

Available by anonymous ftp at ftp.ncifcrf.gov/pub/users/chen/rnaga.tar.Z

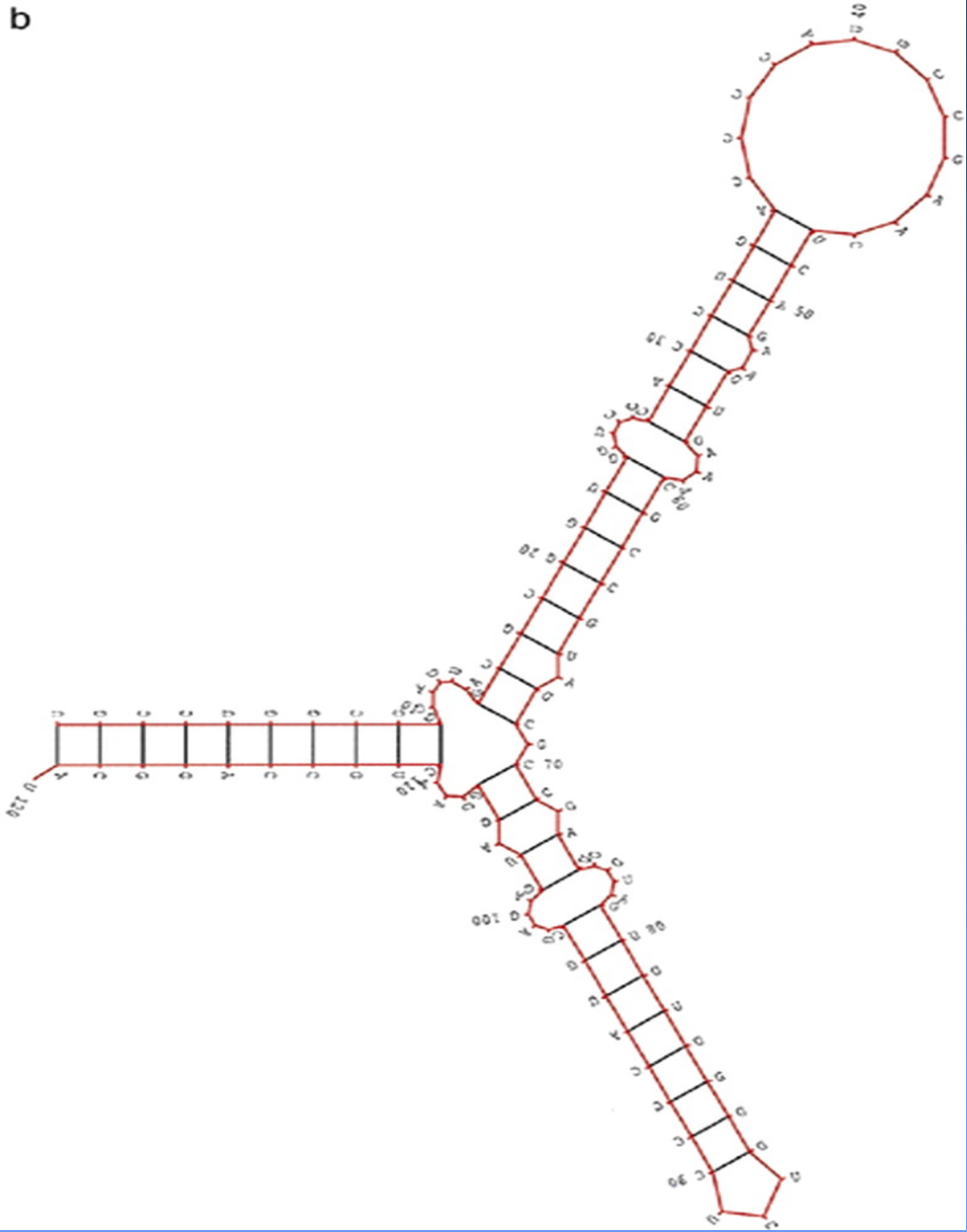
Accuracy of RNAGA predictions

Table 1. Accuracy of a genetic algorithm for RNA common secondary structure prediction

RNA	Nucleotides	Base pair	Correctly predicted base pair (%)			
			Rank 1	Rank 10	Best structure	Any structure
tRNA	1556	432	87.7 ± 12.4	81.2 ± 12.5	98.8 ± 2.7	99.8
5S rRNA	3004	910	95.3 ± 7.0	87.9 ± 7.3	98.6 ± 4.3	98.7

Only the first 10 ranked ordered structures were considered in assessing the accuracy. The accuracy was determined for: the structure ranked first (i.e. with highest adjusted conservation score); the structure ranked tenth; the single best structure of the first 10 ranked ordered structures (column 6); the base pairs correctly predicted in at least one structure (column 7). The accuracy was determined by counting correctly predicted base pairs. Standard deviations are given with the percentages to demonstrate the range of accuracy. Only tRNA and 5S rRNA are listed since there are no known standard structures in the RREs of HIV-1 and HIV-2.

E. coli 5S rRNA

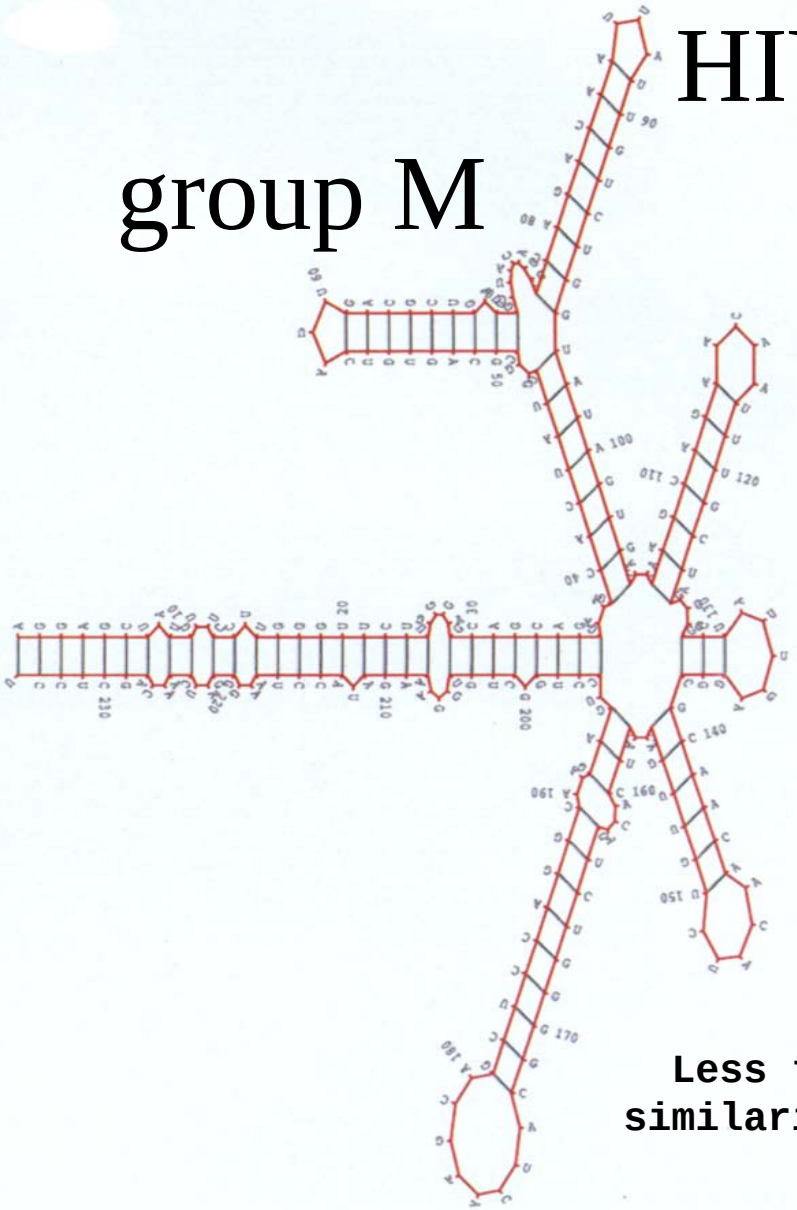


Alignment of 25 5S rRNAs

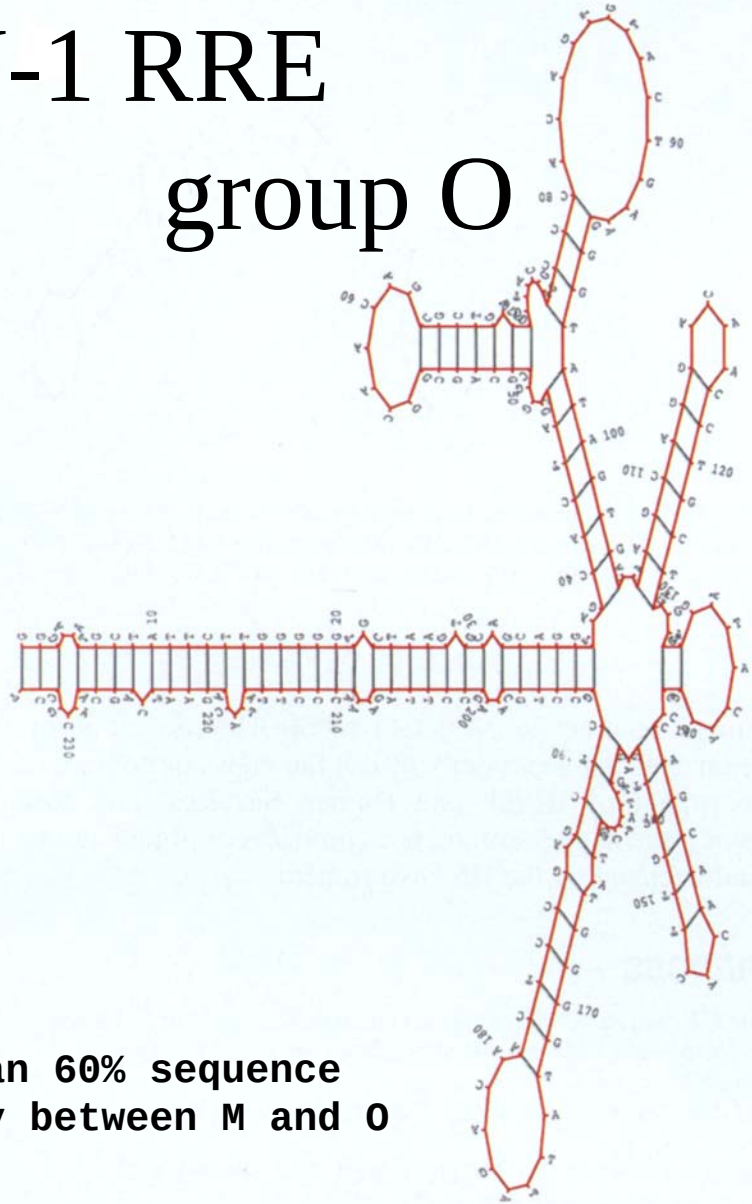
```
                A                   B                                   B'
E.coli      UGCC.UGGCGGcaGUAGCGCGGUGgucccCACCUGA.cccCAUGCcgaacUCAGaaGUGaaaCGC
Photobac    UGCU.UGGCGAccAUAGCGUUAUGgacccCACCUGA.UCcCUUGCcGAacUCAGuaGUGaaaCGU
Beneckia    uGCU.UGGCGAccAUAGCGAUUUGgacccCACCUGAcUUcCAUUCcGAacUCAGaaGUGaaaCGA
P. fluor    uUCU.UGACGAccAUAGAGCAUUGgaacCACCUGA.UCcCAUCCcGAacUCAGuaGUGaaaCGA
Azo.vine    UGCU.UGACGAUcAUAGAGCGUUGgaacCACCUGA.UCcCAUCCcGAacUCAGcaGUGaaaCGA
P.aerugi    uGCU.UGACGAUcAUAGAGCGUUGgaacCACCUGA.UCcCUUCCcGAacUCAGaaGUGaaaCGA
R.rubrum   UGGCC.UGGUGGUcAUUGCGGGCUCgaaaCACCCGA.UCcCAUCCcGAacUCGGccGUGaaaGAG
Wicrococ   G.UUA.CGGCGGCuAUAGCGUGGGGgaaacGCCCGG.cCGUAUAUCGaacCCGGaaGCuaagCCC
Streptom   G.uUU.CGGUGGUcAUAGCGUGAGGgaaacGCCCGG.UUaCAUUCcGAacCCGGaaGCuaagCCU
Th.aquat   AAUCCCCCGUGCCcUUAGCGGCGUGgaacCAC.CCGUUCcCAUUCcGAacACGGaaGUGaaaCGC
Th.therm   AAUCCCCCGUGCCcAUAGCGGCGUGgaacCAC.CCGUUCcCAUUCcGAacACGGaaGUGaaaCGC
Prochlor    uUCC.UGGUGUCuCUAGCGCuuuggaacCACuUCGAUUcCAUCCcGAacUCGAuuGUGaaacuu
Hb.salin    uu..aAGGCGGCcAUAGCGGUGGGguuacUC.CCGUaccCAUCCcgaacACGGaaGAuaagCCC
Hc.morrh    uu..aAGGCGGCcACAGCGGCGGGgcgacUC.CCGUaccCAUCCcgaacACGGcaGAuaagCCC
T.acidop    ...GGcAACGGUcAUAGCAGCAGGgaaacAC.CaGAUCcCAUUCcGAacUCgacGGUuaagCCU
S.acidoc    GCCCA.CCCGGUcACAGUGAGCGGgcaacAC.CCGGacuCAUUUcgaacCCGGaaGUuaagCCG
Wsp.hung    .uCAAUAGCGGCcACAgcAGGUGUgucacAC.CCGUUCcCAUUCcGAAcaCGGaaGUuaagACA
Paraco      .GUC.UGGUGGccAAAGCACGaGCAaaacAC.CCGAUCcCAUCCcGAacUCGGccGUuaagUGC
B.stearo    ..CC.UAGUGACaAUAGCGGAGAGgaaacAC.CCGUUCcCAUCCcGAacACGGaaGUuaagCUC
B.acidoc    ..UC.UGGUGACuAUAGCGGAGGggcaaCAC.CCGUaccCAUCCcgaacACGGacGUGaagaCC
Strept.     ..UG.UGGUGGCGAUAGCGAGAAGgauacAC.CUGUaacCAUGCcgaacACAGaaGUuaagCUU
B.brevis    ..UC.UGGUGAUGAUGGCGGAGGGgacacAC.CCGUUCcCAUACcGAacACGGccGUuaagCCC
C.pasteu    ..UC.CAGUGUCuAUGACUUAGAGguaacAC.UCCUUCcCAUUCcGAacAGGcaGGUuaagCUC
B.sub       ..UU.UGGUGGCGAUAGCGAAGAGgucacAC.CCGUUCcCAUACcGAacACGGaaGUuaagCUC
B.lichen    ..UU.UGGUGGCGAUAGCGAAGAGgucacAC.CCGUUCcCAUGCcGAacACGGaaGUuaagCUC

                        C                   C'                A'
E.coli      CGUaGC..gCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGCgaga.GUaGGg.aaCUGCCA.GGCAU
Photobac    AAUaGC..gCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGUgaga.GUaGGa.caUCGCCA.GGCAU
Beneckia    AUUaGC..gCCgAUgg.uaGUGUGG.GGCUU.CCCCAUGUgaga.GUaGGa.caUCGCCA.GGCuU
P. fluor    UGCaUC..GCCgAUgg.uaGUGUGG.GGUUU.CCCCAUGUgaga.GUaGGU.caUCGUCA.AGAuU
Azo.vine    CGCaUC..GCCgAUgg.uaGUGUGG.GGUUU.CCCCAUGUaaga.GUaGGU.cAUCGUCA.GGCGC
P.aerugi    CGCaUC..GCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGUgaga.GUaGGU.cAUCGUCA.AGCuC
R.rubrum    CCCuGC..GCCaAUgg.uaCUGCGU..CUUAAGGCG.UGGgaga.GUaGGU.cGCCGCCA.GGCCU
Wicrococ    CAUaGC..GCCgAUgguuaCUGUAA.CCGGGAGGUUGUGGgaga.GUaGGU.cGCCGCCG.UGA..
Streptom    UACaGC..gCCgAUgg.uaCUGCAG.GGGGGACCCUGUGGgaga.GUaGGa.cGCCGCCG.AAc.U
Th.aquat    GCCaGC..GCCgAUgg.uaCUGGGA.CCGCAGGGUCCUGGgaga.GUaGGU.cGGUGCGGGGGAU
Th.therm    GCCaGC..GCCgAUgg.uaCUGGGC.GGGCGACCGCCUGGgaga.GUaGGU.cGGUGCGGGGGAU
Prochlor    uGCuGC..GGCUA.agauaCUuGCU.GGGUUGCUGGCuGGgaaa.aUAGCU.cGAUGCCA.GGAuU
Hb.salin    GCCuGCGUuCCGGucaguaCUGGAGuGCGCGAGCCUCUGGgaaa.uCCGGuUcGCCGCCU...acU
Hc.morrh    GCCaGCGUuCCAGcgaguaCUGGAGuGUGCGAACCUCUGGgaaa.aCUGGuUcGCCGCCU...ccC
T.acidop    GCU.GCGUAUUGCGUUguaCUGUAUgCCGCGAGGGUACGGgaAGC.GCAAUAuGCUGUUaCC...ACU
S.acidoc    CUC.ACG.UUAGUGGGgccGUGGAUaCCGUGAGGAUCCGCagCCCcACUAA..GCUGGGaUGGGUUUU
Wsp.hung    CCUcacGUGGAUGAcgguaCUGAGGuACGCGAGUCCUCGGgaaa.UCAUCCUcGCUGCUAUUGu.U
Paraco      CGUaGC..GCCaAUgg.uaCU.GCG.UCAAAAGACGU.GGgaga.GUaGGU.caCCGCCA.GAC.C
B.stearo    UCCaGC..GCCgAUgg.uaGUuGGG.GCCAGCGCCCCuGCaaga.GUaGGU.cGUUGCUA.GG..C
B.acidoc    UCCaGC..GCCgagaa.uaCUGGGA.GGGCAGCCUCCUGGgaaa.guaGGU.cGUUGCCA.GG..C
Strept.     CUUaGC..gCCgAUUg.uaGUGAGG.GGGUUGCCCCUUGUgagA.GUaGGa.CGUCGCCA.CG..C
B.brevis    UCCaGC..gCCaAUgg.uaCUuGCU.CCGCAGGGAGCcGGgaga.GUaGGa.CGUCGCCA.GG..C
C.pasteu    UAAuGU..GCUgAUgg.uaCUGCAG.GGGAAGCCCUGUGGaaga.GUaGGU.cGACGCUG.GG..U
B.sub       UUCaGC..gCCgAUgg.uaGUcGGG.GG.UUUCCCCCuGUgaga.GUaGGa.CGCCGCCA.AG..C
B.lichen    UUCaGC..gCCgAUgg.uaGUuGGG.GG.CUUCCCCCuGUgaga.GUaGGa.CGCCGCCA.AG..C
```

HIV-1 RRE

group M



group O



Less than 60% sequence similarity between M and O

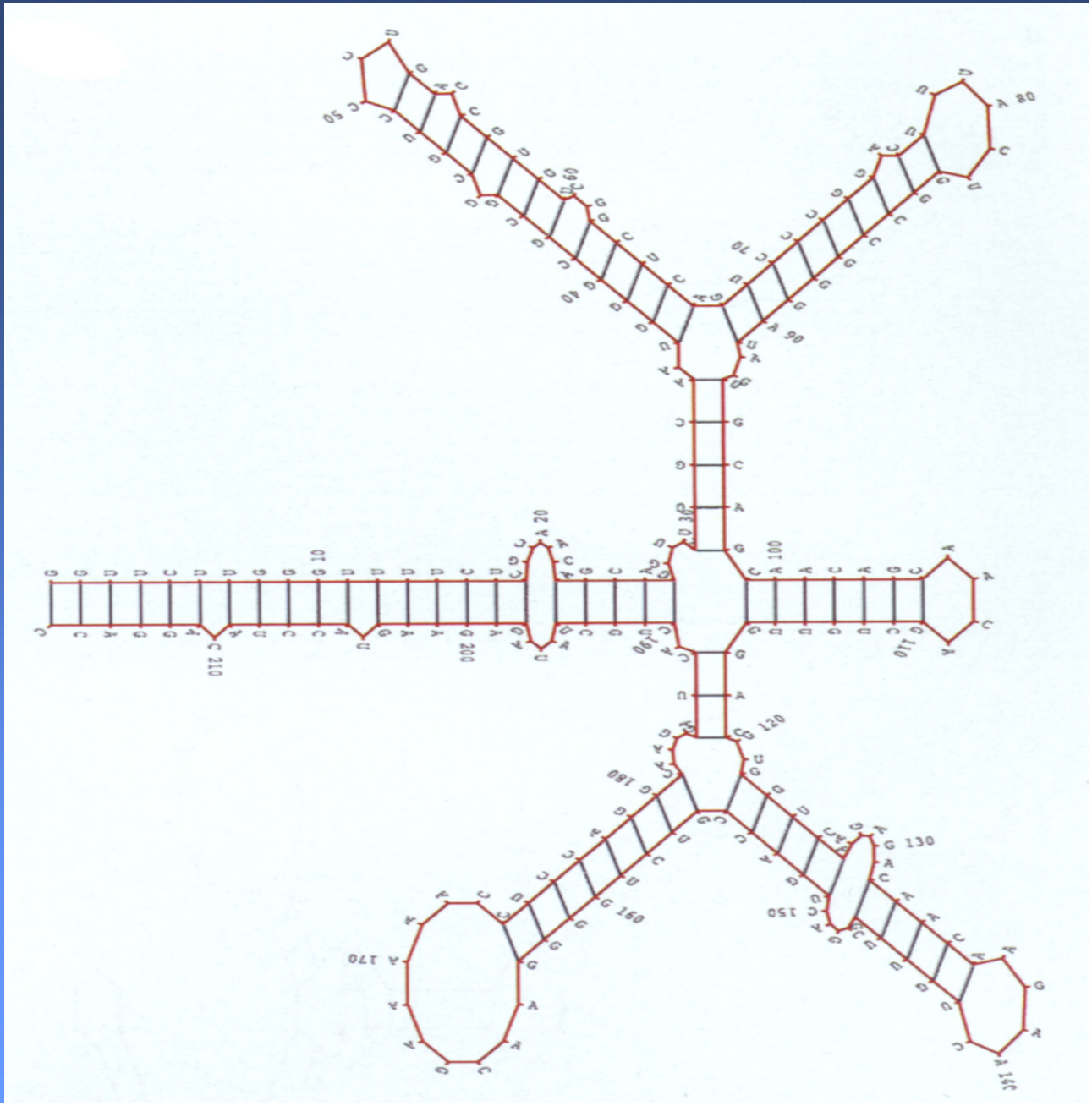
Structural alignment of seven HIV-1 RRE sequences

```
                                A                                   B1        B2
SF2        AGGA..GCUaUGuuCCuUG.GG.UUCUUgggaGCAGCAGGAAGCACUAUGGGCGCAGUG.UCAUU
HXB2       AGGA..GCUUUGuUCCUUG.GG.UUCUUgggaGCAGCAGGAAGCACUAUGGGCGCAGCc.UCAAU
MAL        AGGA..GCCaUGuUCCUUG.GG.UUCUUgggaGCAGCAGGAAGCACGAUGGGCGCAGCG.UcACU
ELI        AGGA..GCUauGUUCcuUG.GG.UUCUUgggaGCAGCAGGAAGCACgAUGGGCGCA.CGgUCAGU
HIVU455    AGGA..GCUaucUUCcuUG.GG.UUCUUaggaGCAGCuGGAAGCACaAUGGGCGCGGCG.UcAAU
HIVANT70   GGGAaUGCUA...UUCUUGGGGgUUCUAAGu.GCAGCAGGUAGCACUAUGGGCGCAGCG.gcAAC
MVP5180    GGGaaUGCUA...UUCUUGGGGGUgCUAAGu.GCaGCAGGUAGCACUAUGGGCGCAGCG.gcAAC

                B2'                B3          B3'       B1'         C

SF2        GACGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUAGUGCAACAGCAGAACAAUUU
HXB2       GAcGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUAGUGCAGCAGCAGAACAAUUU
MAL        aACGCUGaCGGUACAGGCCAGACAGUUACUGUCUGG UAUaGUGCAACAGCAAAACAAUUU
ELI        GACGcUGaCGGUACAGGCCAGACAaUUAaUGUCUGG UAUaGUGCAACAGCAAAACAAUUU
HIVU455    aACGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUaGUGCAACAGCAGAGCAAUCU
HIVANT70   aaCGCUGgCGGUACAGACCCaCAcuUUGcUGaaGGG UAUAGUGCAAcAGCAGGACAACCU
MVP5180    agCGCUGaCGGUACGGACCCaCAGuGUACUGaaGGG UAUAGUGCAACAGCAGGACAACCU

           C'      D        D'    E         E'            F

SF2        GCUGAGGGCUAUUGAGGC GCAACAaCAUcUGUUGCAA..CUCacaGUCUGGGGCAUCAAGCA
HXB2       GCUGAGGGCUAUUGAGGC GCAACAGCAUCUGUUGCAA..cUCacaGUcUGGGGCAUCAAGCA
MAL        GCUGAGGGCUAUAGAGGC GCAACAGCAUCUGUUGCAA..CUCacgGUCUGGGGCAUUAAACA
ELI        GCUGAGGGCUAUAGAGGC GCAACAGCAUCUGUUGCAA..cUCacgGUCUGGGGCAUUAAACA
HIVU455    GCUGAGGGCUAUAGAGGC UCAACAGCAUCUGUUGAAA..CUCacUGUCUGGGGCAUUAAACA
HIVANT70   GCUaAGAGCaAUACAgGC cCAGCAGCAAUUGCUGa..GGCUaucuxuaUGGGGUAUCAGACA
MVP5180    GCUGAGAGCgAUACAgGC cCAGCAACACUUGCUGa..GGuuaucuGUaUGGGGUAUUAGACA

                F'                          A'

SF2        GCUCCAGGCaaGAG UCCUgGCU.GUggaAAGAuA.CC.UAagGGauCAacAGC...UCCU
HXB2       GCUCCAaGCaaGAa UCCUaGCU.GUggaAAGAuA.CC.UAaAGGAuCAAcAGC...UCCU
MAL        GCUCCAGGCaaGAG UCCUgGCU.GUggaAAGAuA.CC.UAaAGGAuCAacGGC...UCCU
ELI        GCUCCAGGCaaGAa UCCUgGCU.GUggaAAGAuA.CC.UAaaGGAUcaacAGC...UCCU
HIVU455    GCUCCAGGCAaGAG UCCugGCU.GUggaAAGAuA.CC.UAcaGGAucaacAGC...UCCU
HIVANT70   ACUCCGAGCUc... GCCU.GCUaGC.CUUAGA.AaCCUUAcuacAGAAUcAGCAacUCCU
MVP5180    ACUCCGaGCUC... GCCU.GCaaGC.CUUAGaaACCCUUA.uacAGAAUcAGCAacgCCU
```

Figure 2

Figure 2. (Opposite and above) (**a**) A structural alignment for seven HIV-1 RRE sequences based on the structural information obtained from the method. Each structure in the alignment was directly obtained from the method. The structure of the first sequence in the alignment is the most favorable with respect to the adjusted conservation score. The structure for other sequences was selected from the best 10 structures of the sequence and most closely resembles the structure of the first sequence in the alignment. Deletions and insertions are needed to align structures between major group M and outlier group O. (**b** and **c**) A predicted common secondary structure of the HIV-1 RRE region for (b) SF2, an isolate from major group M, and (c) MVP5180, an isolate from outlier group O. The two structures are similar to each other even though the two sequences share less than 60% of nucleotides in the RRE region. The seven HIV-1 RRE sequences (accession nos are given in parentheses) used in this study were: HIVSF2 (K02007), HIVHXB2 (K03455), HIVMAL (K03456), HIVELI (K03454), HIVU455 (M62320), HIVMVP5180 (L20571) and HIVANT70 (L20587).

Consensus RRE for HIV-2/SIV



Structural alignment of 10 RRE sequences of HIV-2/SIV

```
                    A                    B1        B2              B2'
HIV2ROD   GGUUCUUGGGUUUUCUCgcaacAGCAGGUUCUGCAAUGGGCGCGgCGUCCCUGAcCGUGU
SIVMM251  GGUUCUUGGGUUUUCUCGcaacGGCAGGUUCUGCAAUGGGCGCGGCCUcGUUcAGGCUGa
SIVMM142  GGUUCUUGGGUUUUCUCGcaacGGCAGGUUCUGCAAUGGGCGCGgCGUCGUUGAcCGUGa
HIV2BEN   GGUuCUUGGGUUUUCUCGcgacAGCAGGUUCUGCAAUGGGCGCGCGgUCCCUGACGCuGU
HIV2CAM2  GGUUCUUGGGUUUUCUCacaacAGCAGGAGUUGCAAUGGGCACGGCGUCCUUGACGCUGU
HIV2D194  GGUUCUUGGGUUUUCUCGcgacAGCAGGUUCUGCAAUGGGCGGCGCGUCCUUGACGCuGU
HIV2ISY   GGUUcCUAGGUUUUCUCAcgacAGCAGGUGCUGCAAUGGGGGCGGCGUCUCUGACGCUGU
HIV2NIHZ  GGUUCcuAGGUUUUCUCgcaacAGCAGGUUCUGCcAUGGGCGCGGCGUCCUUGACGCUGU
HIV2UC1   GGUUCUUGGGacUUCUUGcaaUGGCAGGUUCUGCAAUGGGCGCAaCGUCCUUGACGcUGU
HIV2ST    GGUUCUUAGGUUUUCUCacgacAGCAGGAGCUGCAAUGGGCGCGGCGUCCUUGACGCUGU

                B3              B3'        B1'     C          C'   D1
HIV2ROD   cgGCUCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
SIVMM251  cCGCUCA GUCCCGGaCUUUAUuGGCUGGGAUAGUGCAG CAACAGCAACAGCUGUUG GGCG
SIVMM142  cCGCUCA GUCCCGGaCUUUAUuGGCUGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2BEN   caGCCCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2CAM2  caGCCCA GUCUCGGaCUUUAUuGGCCGGGAUAGUGCAG cAACAGCAACAGCUGUUa GACG
HIV2D194  CgGCUCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2ISY   CggCUCA GUCUCGGAcuUUAUUCCGuGGgAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2NIHZ  caGCUCA GUCUCGGaCUUUAUuGGCCGGGAUAGuGCAG CAACAGCAACAGCUGUUG GAUG
HIV2UC1   caGCUCA GUCCCGGaCUUUACuGGCUGGGAUAGUGCAG CAGCAGCAACAGCUGCUG GACG
HIV2ST    cgGCUCA GUCUCGGaCUUUAUuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG

                D2                    D2'          D3            D3'
HIV2ROD   UGGUCAagagaCAACAaGAAcUGUUGcgacUGACC GUCUGGGGaaCGAAAAacCUCCAGG
SIVMM251  UGGUCAAgagaCAACAAGAAUUGUUGcgaUUGACC GUCUGGGGaaCAAAGAacCUCCAGA
SIVMM142  UGGUCAagagaCAACAAGAAUUGUUGcgacUGACC GUCUGGGGaaCAAAGAacCUCCAGA
HIV2BEN   UaGUCAagagaCAACAaGAAaUGUUGcgacUGACc GUCUGGGGaaCGAAAAacCUCCAGG
HIV2CAM2  UGGUCAagagaCAACAAGAAUUGUUGcgacUGACC GUCUGGGGAACAAAAAUUCUCCAGG
HIV2D194  UGGUCAAgagaCAACAaGAAaUGUUGcgaUUGACC GUCUGGGGAaCGAAAAaUCUCCAGG
HIV2ISY   UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGaaCUAAAAacCUCCAGG
HIV2NIHZ  UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGAaCAAAAAaUCUCCAGG
HIV2UC1   UGGUCAaaagaCAACAGGAACUGUUGcggcUGACC GUCUGGGGaaCGAAAAacCUCCAGA
HIV2ST    UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGAaCAAAAAaUCUCCAGG

                D1'          A'
HIV2ROD   CAAGAGUCACUGCUauaGAGAAGuACCUAcAGGACC
SIVMM251  CUAGGGUCACUGCCauCGAGAAGuACUUAgAGGACC
SIVMM142  CUAGGGUCUCUGCCauCGAGAAGuACUUAaAGGACC
HIV2BEN   CAAGAGUCACUGCUauCGAGAAGuACCUAaAGcAUC
HIV2CAM2  CAAGAGUCACUGCUauaGAGAAGuACCUAaAGGAUC
HIV2D194  CAAGAGUCACUGCUauCGAGAAAuACUUAAaGGACC
HIV2ISY   CAAGAGUCACUGCUauUGAGAAGuACCUAGcAGaCC
HIV2NIHZ  CAAGAGUCACUGCUauaGAGAAGuCACUaaaGGACC
HIV2UC1   CAAGAGUCACUGCCAuCGAGAAauaCCUAAaGGACC
HIV2ST    CAAGAGUCACUGCUaucGAGAAAuACUUaAAGGACC
```

Atomic-level structural models

Atomic-level models can be derived by manual model building, or by programs.

We use Hugo Martinez's program RNA2D3D which literally folds a planar secondary structure model into 3D.

Refinements are done with molecular mechanical/dynamical programsl, mainly using the Kollman lab's AMBER.

Looking forward -- Implications

UTRs will continue to be a gold mine of interesting structure/function phenomena for computational and experimental approaches.

There are already findings that RNA and protein expression are often not directly proportional.
For example, Rudi Aebersold and co-workers find that yeast mRNA levels, as measured by SAGE analysis do not successfully predict protein levels and vary as much as 30-fold. This has serious implications for interpreting nucleic acid micro-array results.

Forward (continued)

Secondary structure parameters slowly improve, but there is still a lot that can be done.

Higher order structure prediction is beginning to be explored.

Excitement is high as more experimental structures appear, but size is limited. NMR can do 50-60nt, X-ray can do more, computer prediction of secondary structure can do 7000, but not at atomic resolution.

Computer predictions are receiving some kind words. See a paper by Tinoco and Bustamante(JMB 293, 271, 1999) discussing agreement between predictions from the Zuker mfold program and some experimental results.

Computers are cost effective and rapidly increasing in power.

Acknowledgments

Shu-yun Le (LECB) and Jih-Hsiang Chen(ABCC)

LECB
John Owens
Bruce Shapiro
Ruth Nussinov
Hugo Martinez
Kathy Currey

Sonnenberg Lab
Nahum Sonenberg

Siddiqui Lab
A. Siddiqui

ABCC
N. Pattabiraman

Rybak Lab
Susanna Rybak
Dianne Newton
Shin-Lin Chen

Elroy-Stein Lab
Orna Elroy-Stein
and co-workers