RNA Matrices and RNA Secondary Structures

Institute for Mathematics and Its Applications: RNA in Biology, Bioengineering and Nanotechnology, University of Minnesota

October 29 – November 2, 2007

Asamoah Nkwanta, Morgan State University

Nkwanta@jewel.morgan.edu

RNA Secondary Structure Prediction

- **Given a primary sequence, we want to find the biological function of the related secondary structure. To achieve this goal we predict its' secondary structure using a lattice walk or path approach.**

- **This walk approach involves enumerative combinatorics and is connected to infinite lower triangular matrices called RNA matrices.**

RNA Secondary Structure

- **Primary Structure – The linear sequence of bases in an RNA molecule**
- **Secondary Structure – The folding or coiling of the sequence due to bonded nucleotide pairs: A-U, G-C**
- **Tertiary Structure – The three dimensional configuration of an RNA molecule. The three dimensional shape is important for biological function, and it is harder to predict.**

RNA Molecule

Ribonucleic acid (RNA) molecule: Three main categories

- **mRNA (messenger) – carries genetic information from genes to other cells**
- **tRNA (transfer) – carries amino acids to a ribosome (cells for making proteins)**
- **rRNA (ribosomal) – part of the structure of a ribosome**

RNA Molcule (cont.)

Other types (RNA) molecules:

- **snRNA (small nuclear RNA) – carries genetic information from genes to other cells**
- **miRNA (micro RNA) – carries amino acids to a ribosome (cells for making proteins)**
- **iRNA (immune RNA) – part of the structure of a ribosome (Important for HIV studies)**

Primary RNA Sequence

- **CAGCAUCACAUCCGCGGGGUAAACGCU**
- **Nucleotide Length, 27 bases**

Geometric Representation

- **Secondary structure is a graph defined on a set of n labeled points (M.S. Waterman, 1978)**

- **Biological**
- **Combinatorial/Graph Theoretic**
- **Random Walk**
- **Other Representations**

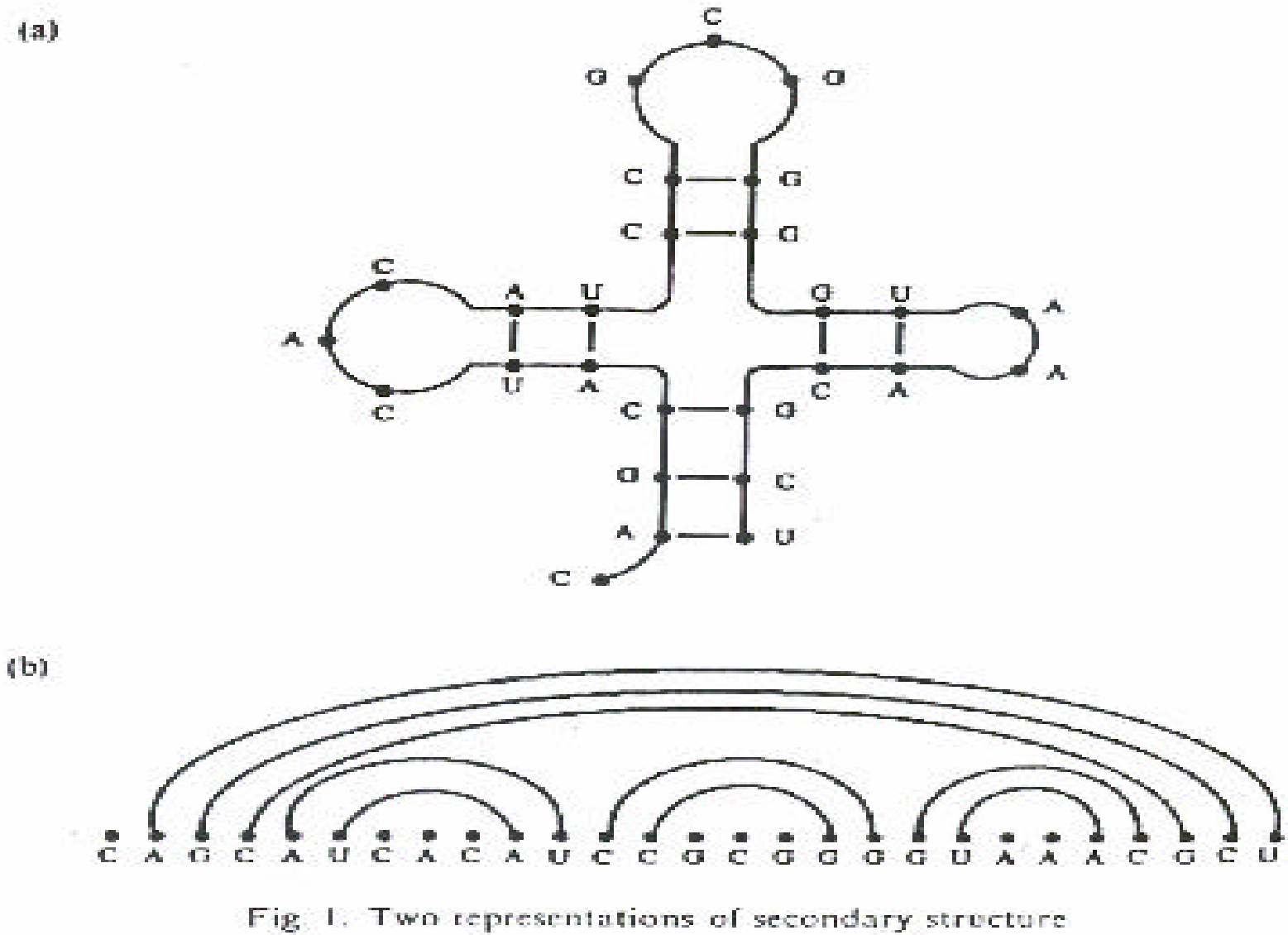
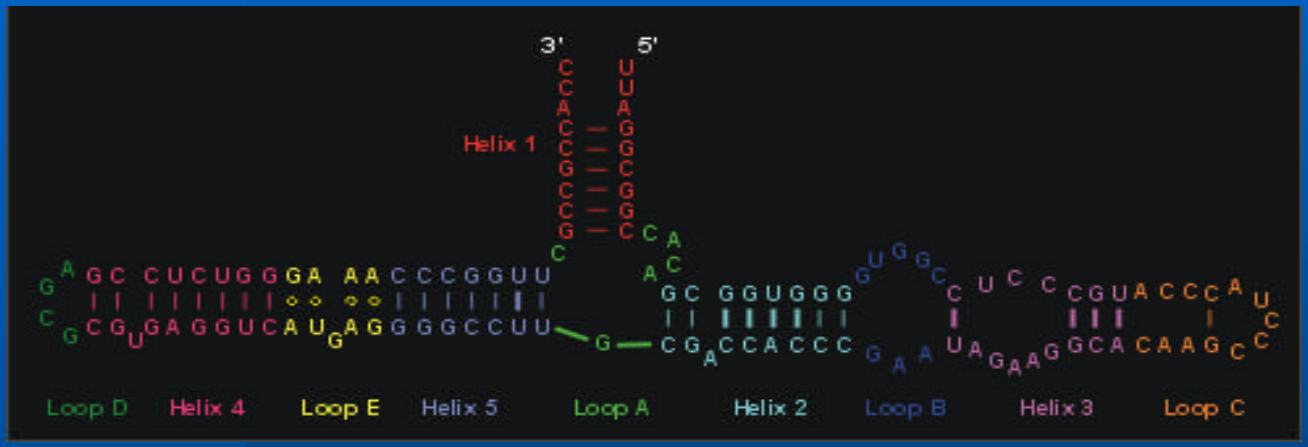
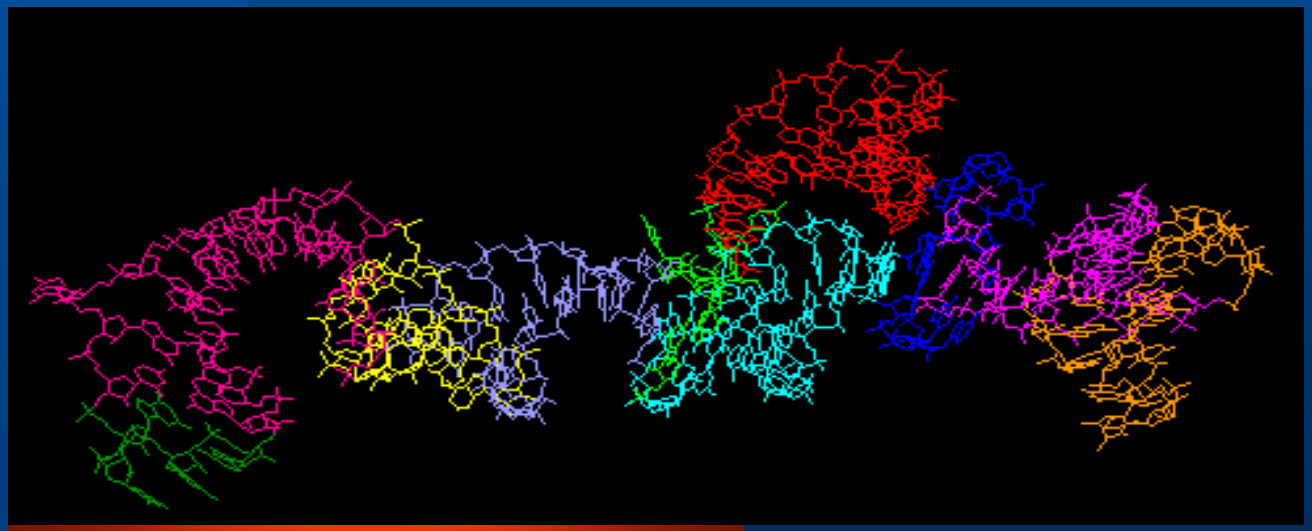


Fig. 1. Two representations of secondary structure

RNA Structure





3-D structure of *Haloarcula marismortui* 5S ribosomal RNA in large ribosomal subunit

RNA NUMBERS

- 1,1,1,2,4,8,17,37,82,185,423,978,…
- These numbers count RNA secondary structures of length n.

RNA Combinatorics

- **Recurrence Relation:**

$$s(0) = s(1) = s(2) = 1$$

$$s(n+1) = s(n) + \sum_{j=1}^{n-1} s(j-1)s(n-j), (n \geq 2)$$

- **M. Waterman, Introduction to Computational Biology: Maps, sequences and genomes, 1995.**

- **M. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math. (suppl.) 1978.**

Counting Sequence Database

- **The On-line Encyclopedia of Integer Sequences: http:/www.research.att.com/njas/sequences/index.html**
- **N.J.A. Sloane & S. Plouffe, The Encyclopedia of Integer Sequences, Academic Press, 1995.**

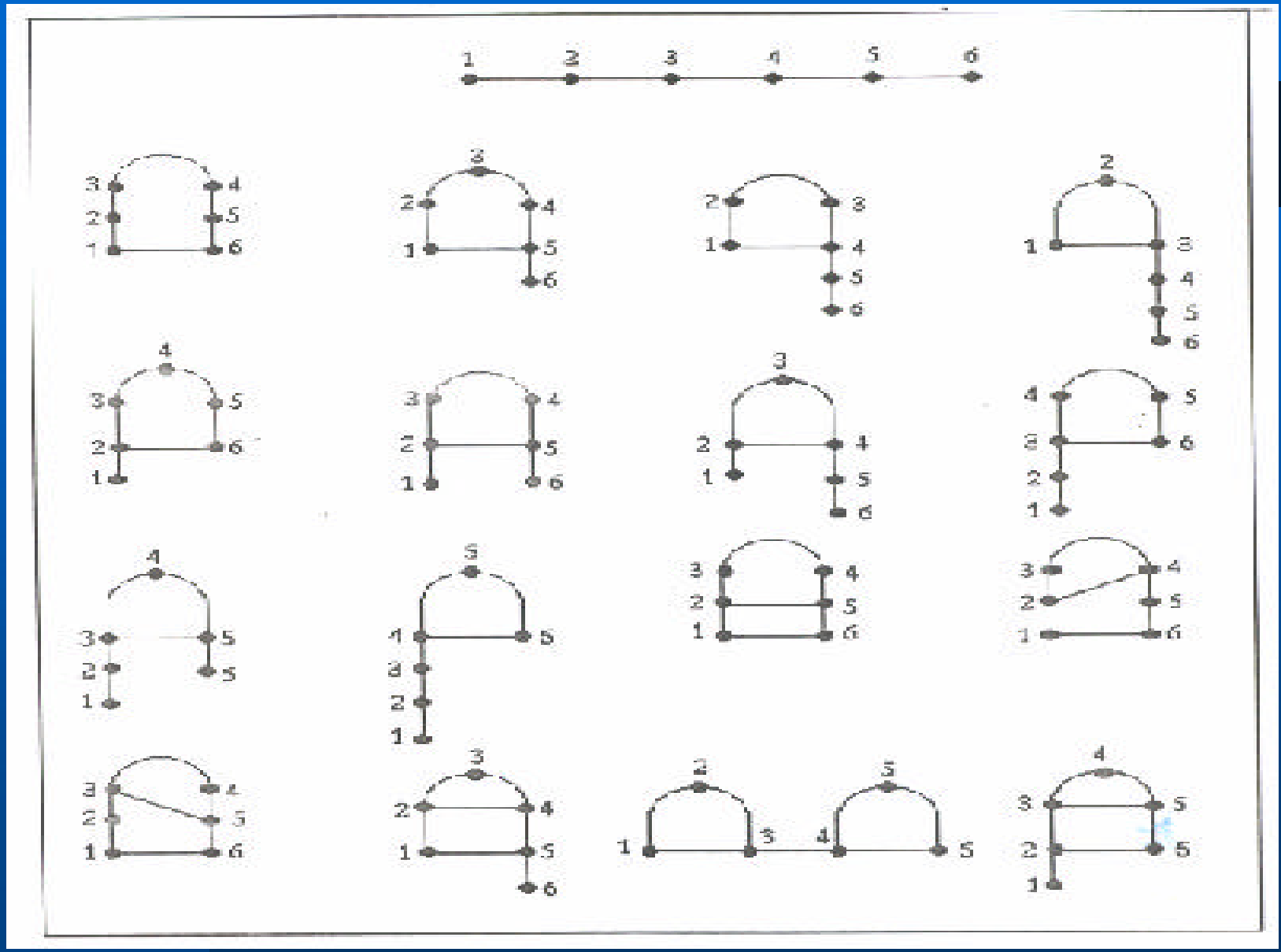
RNA Combinatorics (cont.)

- **The number of RNA secondary structures for the sequence [1,n] is counted by the coefficients of S(z):**

$$s(z) = 1 + z + z^2 + 2z^3 + 4z^4 + 8z^5 + 17z^6 + 37z^7 + \cdots$$

Coefficients of the power series:

- **(1,1,1,2,4,8,17,37,82,185,423,978,…)**



$$s(z) = 1 + z + z^2 + 2z^3 + 4z^4 + 8z^5 + 17z^6 + 37z^7 + \cdots$$

RNA Combinatorics (cont.)

- **Based on the coefficients of the generating function there are approximately 1.3 billion possible RNA structures of length n = 27.**

$$s(z) = 1 + z + z^2 + 2z^3 + \cdots + 1{,}392{,}251{,}012 z^{27} + \cdots$$

RNA Combinatorics (cont.)

- **Using the recurrence relation we can find the closed form generating function associated with the RNA numbers.**

$$s(z) = \frac{\left(1 - z + z^2\right) - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

$$s(z) = 1 + z + z^2 + 2z^3 + 4z^4 + 8z^5 + 17z^6 + \cdots + 1{,}392{,}251{,}012z^{27} + \cdots$$

RNA Combinatorics (cont.)

- **Exact Formula, and Asymptotic Estimate (as n grows without bound):**

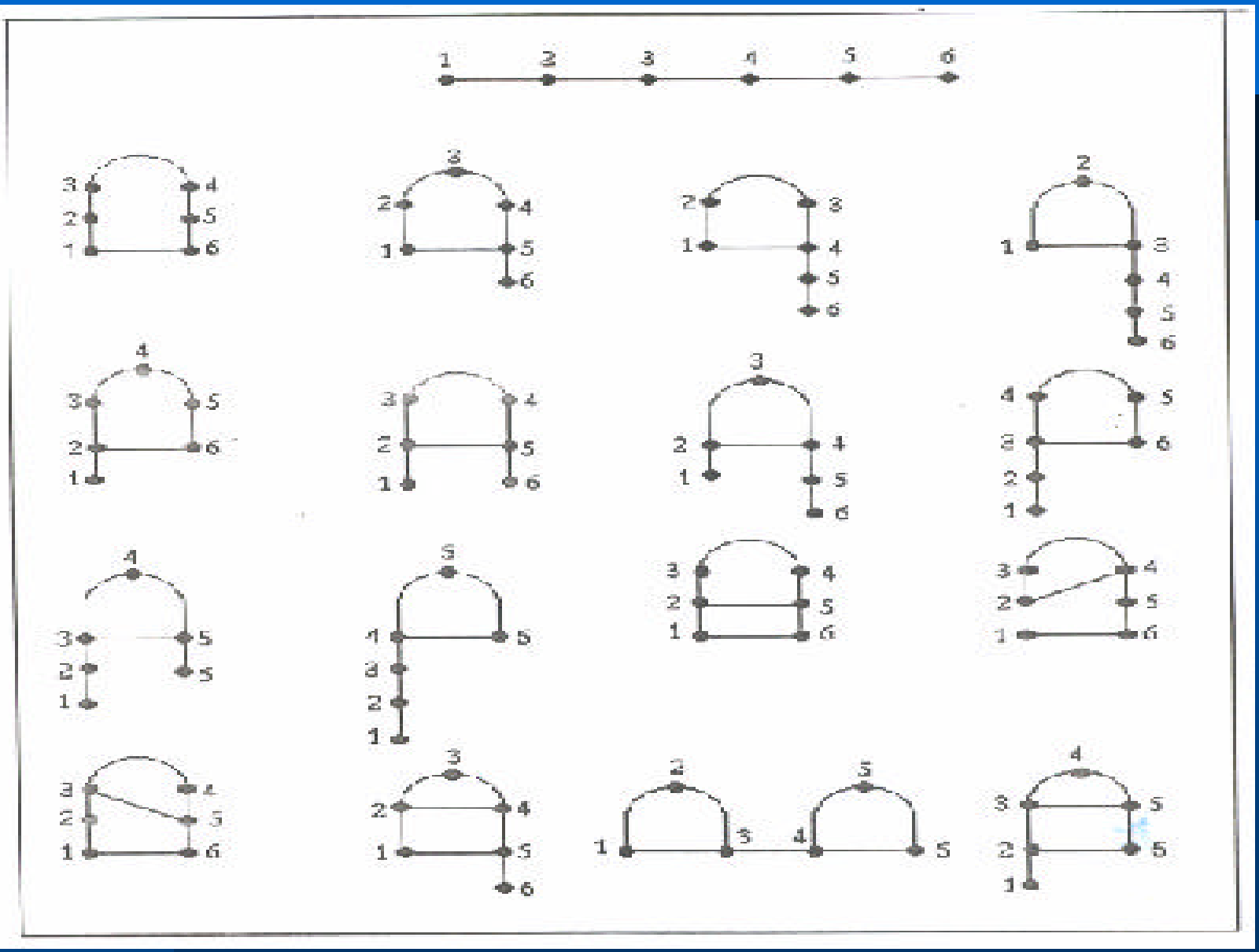
$$s(n) = \sum_{k \geq 1} \frac{1}{(n-k)} \binom{n-k}{k-1} \binom{n-k}{k}$$

$$s(n) \sim \sqrt{\frac{15 + 7\sqrt{5}}{8p}} \left(n^{-\frac{3}{2}} \right) \left(\frac{3 + \sqrt{5}}{2} \right)^{n}$$

RNA Combinatorics (cont.)

- **S(n,k) is the number of structures of length n with exactly k base pairs: For n,k > 0,**

$$s(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1}$$



$$s(6,1) = 10 \text{ and } s(6,2) = 6$$

RNA Combinatorics (cont.)

- **RNA hairpin combinatorics.**

$h(n) = 2^{n-2} - 1$, number of hairpins of length n

$H(n) = 2^{n-m-1} - 1$, number of hairpins with m or more bases in the loop, $n \geq m+1$

$k(a,b) = \sum_{k \geq 0} \binom{a+b-2}{a-1}$, number of stems with topmost bases bound

$K(a,b) = \sum_{k \geq 0} \binom{a+b-4}{a-2}$, number of stems with first and last bases bound

Random Walk

- A random walk is a lattice path from one point to another such that steps are allowed in a discrete number of directions and are of a certain length

RNA Walk – Type I

- **NSE* Walks – Unit step walks starting at the origin (0,0) with steps up, down, and right**

- **No walks pass below the x-axis and there are no consecutive NS steps**

Type I, RNA Array (n x k)

1	0	0	0	0	0	0	.
1	1	0	0	0	0	0	.
1	2	1	0	0	0	0	.
2	3	3	1	0	0	0	.
4	6	6	4	1	0	0	.
8	13	13	10	5	1	0	.
17	28	30	24	15	6	1	.
.	.	.	.	.	.	.	.

Type I, RNA Array (n x k)

1	0	0	0	0	0	0
1	1	0	0	0	0	0
1	2	1	0	0	0	0
2	3	3	1	0	0	0
4	6	6	4	1	0	0
8	13	13	10	5	1	0
17	28	30	24	15	6	1
.	.	.	.	.	.	.

Type I, Formation Rule (Recurrence)

$$m(0,0)=1$$

$$m(n+1,0)=\sum_{j\geq 0} m(n-j,\, j)$$

$$m(n+1,k)=m(n,k-1)+\sum_{j\geq 0} m(n-j,k+j)$$

$$m(n+1,k)=0, k>n+1$$

Note. S(z) can be derived using this recurrence.

First Moments/Weighted Row Sums

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdot \\ 1 & 1 & 0 & 0 & 0 & 0 & \cdot \\ 1 & 2 & 1 & 0 & 0 & 0 & \cdot \\ 2 & 3 & 3 & 1 & 0 & 0 & \cdot \\ 4 & 6 & 6 & 4 & 1 & 0 & \cdot \\ 8 & 13 & 13 & 10 & 5 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \cdot \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 8 \\ 21 \\ 55 \\ 144 \\ \cdot \end{pmatrix}$$

Computing the average height of the walks above the x-axis is given by the alternate Fibonacci numbers

RNA Walk – Type II

- **NSE** Walks – Unit-step walks starting at the origin (0,0) with steps up, down, and right such that no walks pass below the x-axis and there are no consecutive SN steps**

Type II , RNA Array (n x k)

1	0	0	0	0	0	0 .
1	1	0	0	0	0	0 .
2	2	1	0	0	0	0 .
4	4	3	1	0	0	0 .
8	9	7	4	1	0	0 .
17	20	17	11	5	1	0 .
37	45	41	29	16	6	1 .
.	.	.	.	.	.	. .

Examples

- **Type I: ENNESNESSE**
- **Type II: NEEENSEEES**

Note. Some Type II walks are not associated with RNA. Thus we have two class of walks to work with for RNA prediction.

RNA Walk Bijection

- **Theorem: There is a bijection between the set of NSE* walks of length n+1 ending at height k = 0 and the set of NSE** walks of length n ending at height k = 0.**

- **Source: Lattice paths, generating functions, and the Riordan group, Ph.D. Thesis, Howard University, Washington, DC, 1997**

Main Theorem

- **Theorem: There is a bijection between the set of RNA secondary structures of length n and the set of NSE* walks ending at height k = 0.**

- **Source: Lattice paths and RNA secondary structures, DIMAC Series in Discrete Math. & Theoretical Computer Science 34 (1997) 137-147. (CAARMS2 Proceedings)**

Main Theorem (cont.)

- **Proof (sketch): Consider an RNA sequence of length n and convert it to the non-intersecting chord form. Consider the following rules:**

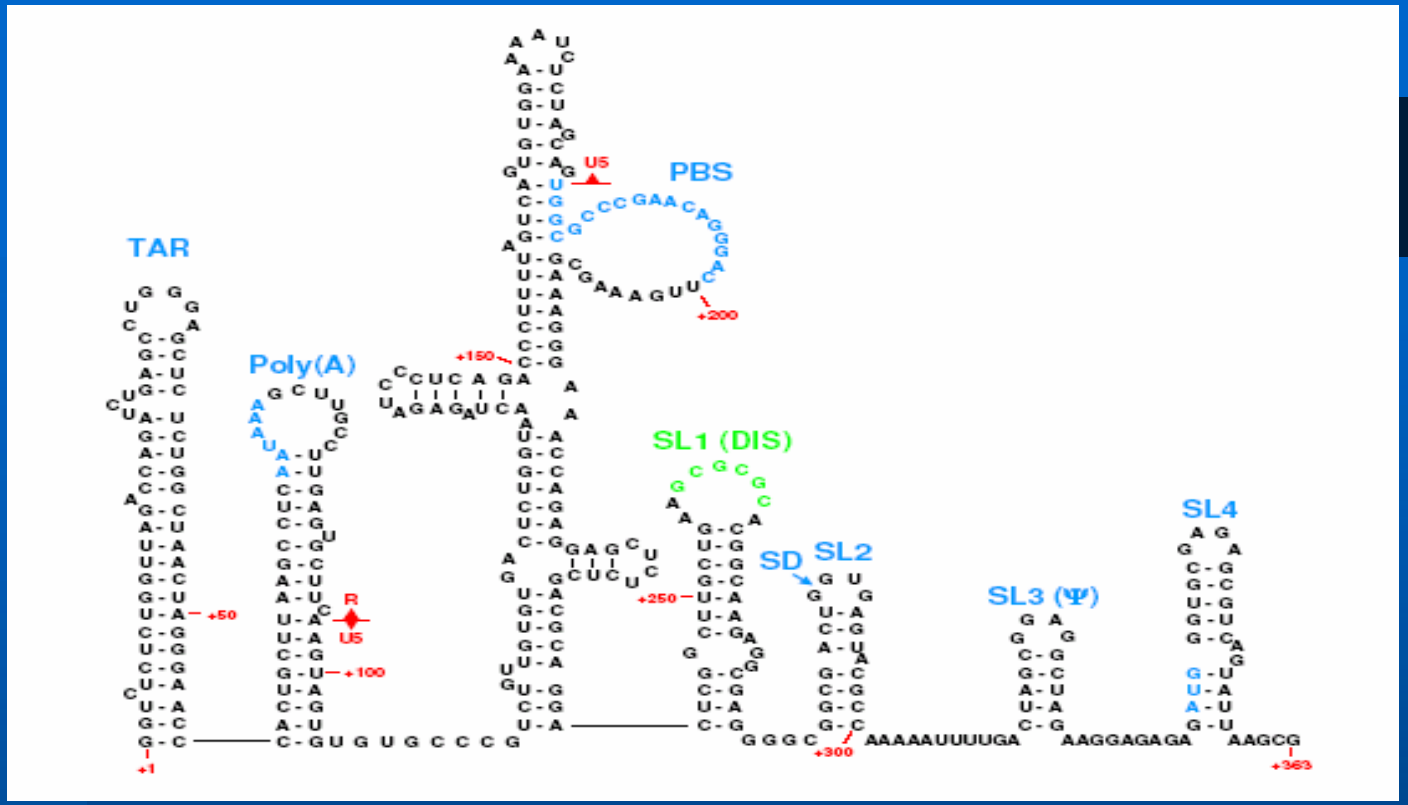
$$k \leftrightarrow E$$

$$(i, j) \leftrightarrow (N, S)$$

Application: HIV-1 Prediction

- **Given primary RNA sequences and using RNA combinatorics, the goal of this project is to model components of an HIV-1 RNA secondary structure (namely SL2 and SL3 domains). The major concentration of this project is on reducing the minimum free energy to form an optimum HIV-1 RNA secondary structure.**

- **Source: HIV-1 sequence prediction, 2007, in progress**



HIV-I 5' RNA Structural Elements. Illustration of a working model of the HIV-I 5' UTR showing the various stem-loop structures important for virus replication. These are the TAR element, the poly(A) hairpin, the U5-PBS complex, the stem-loops 1-4 containing the DIS, the major splice donor, the major packaging signal, and the gag start codon, respectively. Nucleotides and numbering correspond to the HIV-I HXB2 sequence. (Adapted from Clever et al. (1995) and Berkhout and van Wamel (2000))

Application: HIV-1 Prediction (cont.)

- **The following sequence was obtained from the NCBI website. The first 363 nucleotides were extracted from the entire HIV-1 RNA genomic sequence:**

- **GGUCUCUCUGGUUAGACCAGAUCUGAGCCUGGGAGCUCUCU GGCUAACUAGGGAACCCACUGCUUAAGCCUCAAUAAAGCUU GCCUUGAGUGCUUCAAGUAGUGUGUGCCCGUCUGUUGUGU GACUCUGGUAACUAGAGAUCCCUCAGACCCUUUUAGUCAGU GUGGAAAAUCUCUAGCAGUGGCGCCCGAACAGGGACCUGA AAGCGAAAGGGAAACCAGAGGAGCUCUCUCGACGCAGGAC UCGGCUUGCUGAAGCGCGCACGGCAAGAGGCGAGGGGCGG CGACUGGUGAGUACGCCAAAAAUUUUGACUAGCGGAGGCUA GAAGGAGAGAGAUGGGUGCGAGAGCGUCAGUAUUAAGCG**

- **Color key:**
  SL2 – yellow
  SL3 - red

Future Research: Centers For

- **Biological and Chemical Sensors Research**
- **Environmental Toxicology and Biosensors Research**
- **The mission is to advance the fundamental scientific and technological knowledge needed to enable the development of new biological and chemical sensors.**

Math-Bio Collaborators

- **Dwayne Hill, Biology Dept., MSU**
- **Alvin Kennedy and Richard Williams, Chemistry Dept., MSU**
- **Wilfred Ndifon, Ecology and Evolutionary Biology Dept., Princeton U.**
- **Boniface Eke, Mathematics Dept., MSU**