Math 8803/4803, Spring 2008: Discrete Mathematical Biology

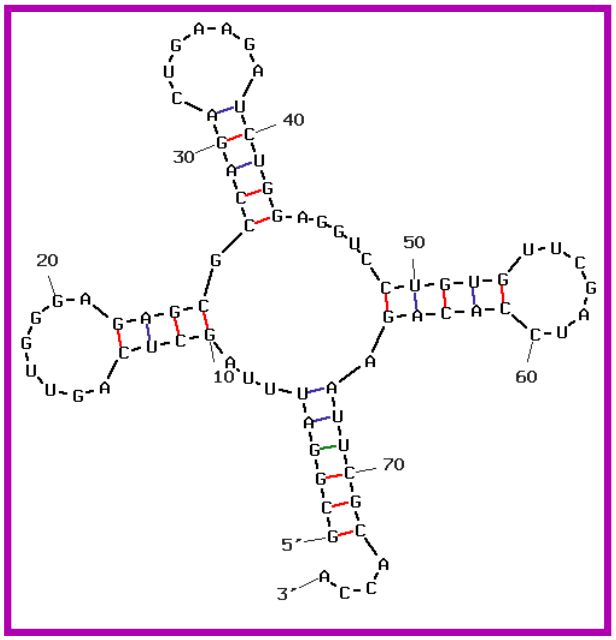
Prof. Christine Heitsch
School of Mathematics
Georgia Institute of Technology

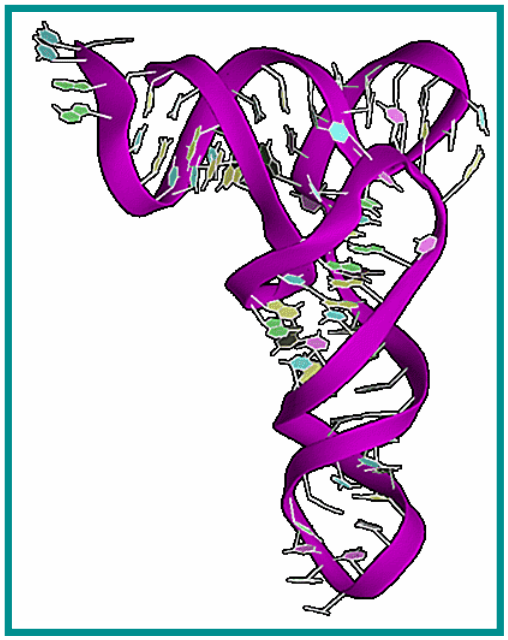
Lecture 12 – February 4, 2008

Levels of RNA structure

Selective base pair hybridization $\Longleftrightarrow$ structure and function

GCGGAUUUAG
CUCAGUUGG
GAGAGCGCCA
GCCUGAAGA
UCUGGAGGUC
CUGGUUCGA
UCCACAGAAU
UCGCACCA



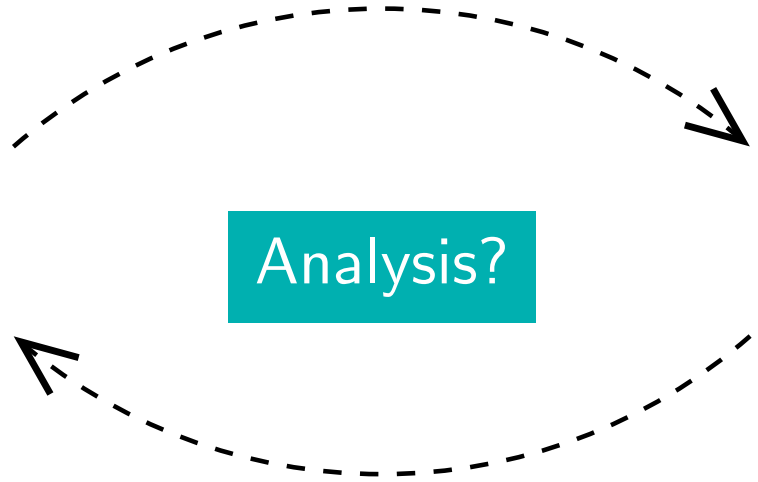


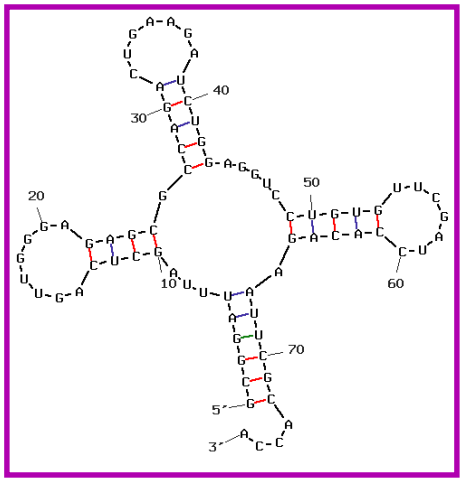
Primary sequence $\longrightarrow$ secondary structure $\longrightarrow$ 3D molecule

Important biomathematical questions

Prediction?

GCGGAUUUAG
CUCAGUUGG
GAGAGCGCCA
GCCUGAAGA
UCUGGAGGUC
CUGGUUCGA
UCCACAGAAU
UCGCACCA





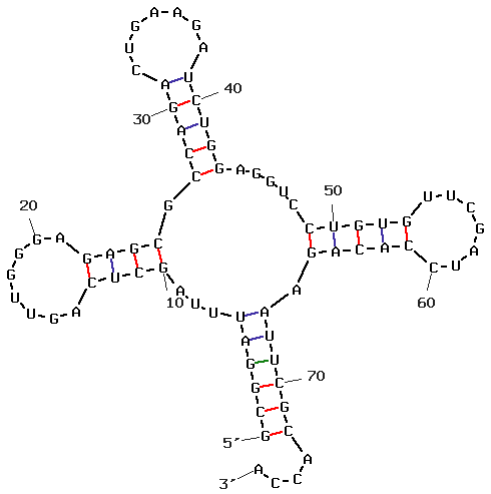
Design?

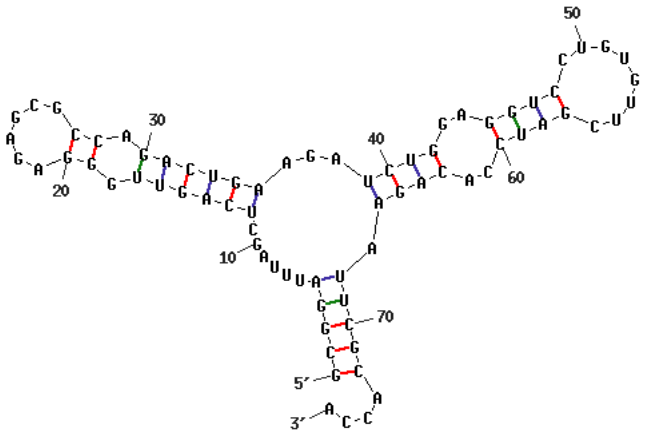
How do RNA sequences encode secondary structures?

Sequence to structure: a one-to-many mapping

$R =$ GCGGA UUUAGCUC AGUUGGGA GAGC G CCAGA CUGAA

GAUCUGG AGGUC CUGUG UUCGAUC CACAG A AUUCGC ACCA

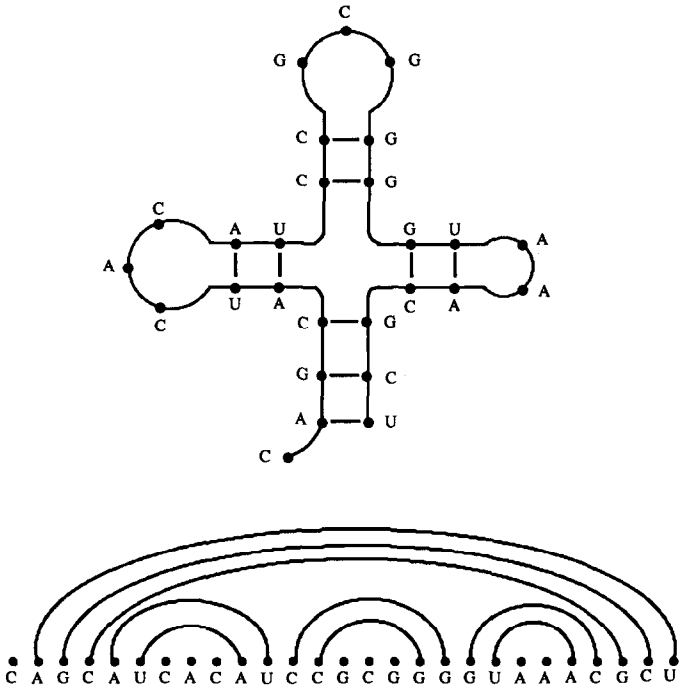
$S_1(R) =$ 

$S_2(R) =$ 

Hypothesis: RNA sequences fold with minimal free energy.

RNA secondary structures as nested arcs

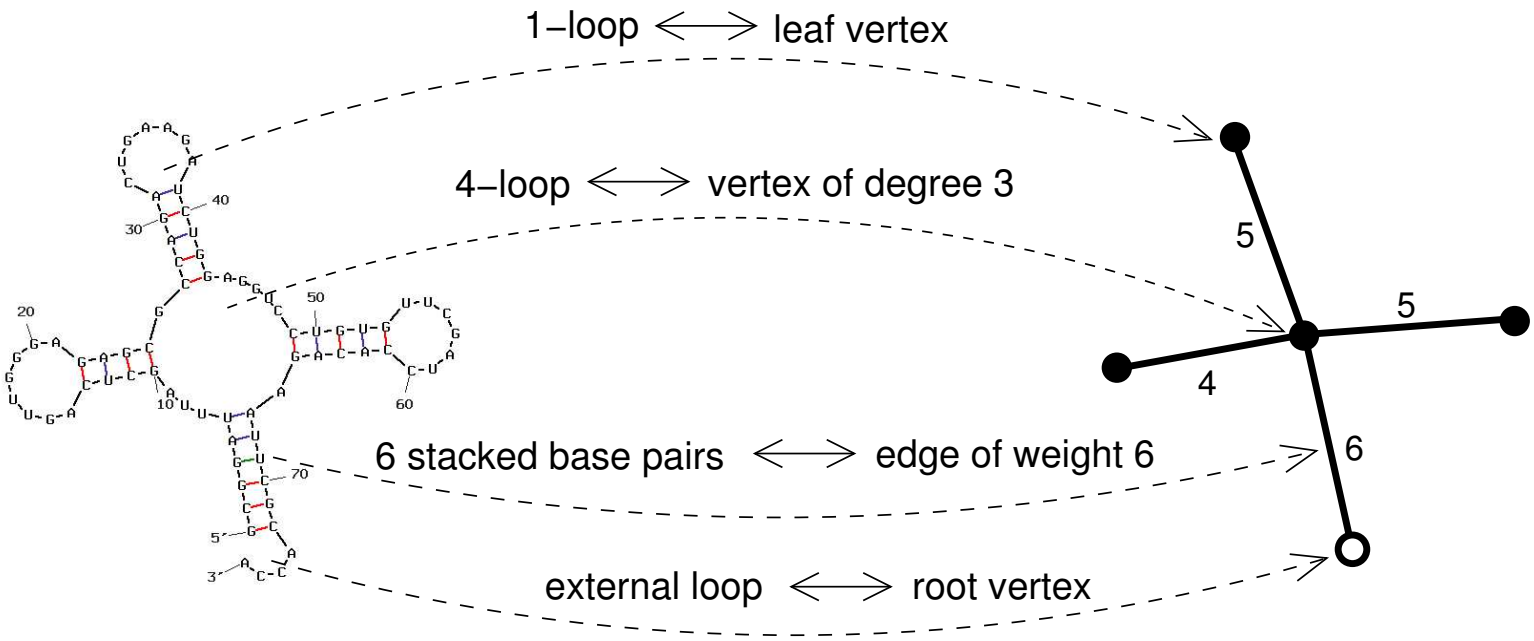
$R = \text{CAGCAUCACAUCCGCGGGGUAAACGCUAAACGCU}$



How many possible $S(R)$ for a sequence R of length n?

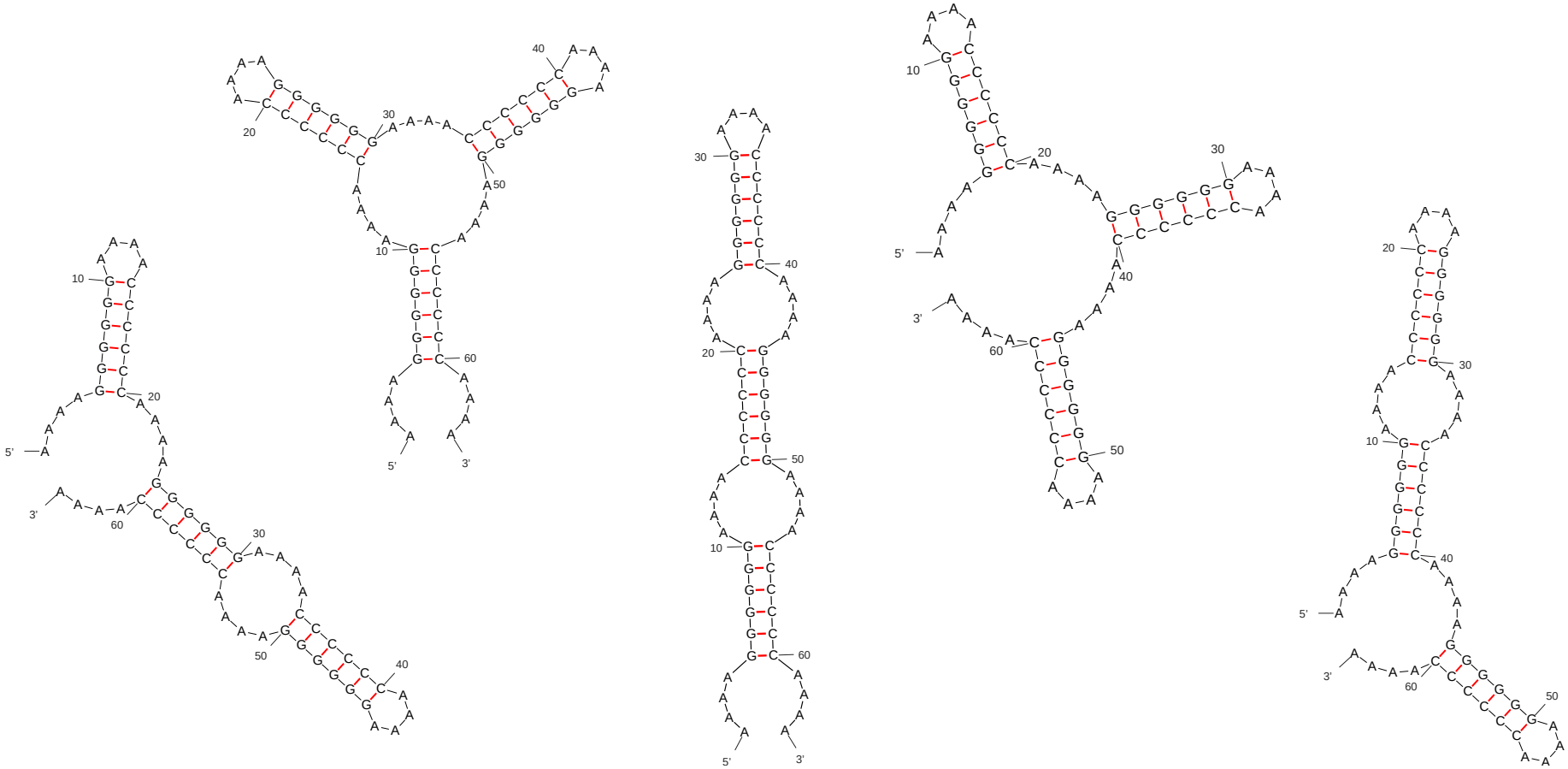
RNA foldings as plane trees

Abstract folded sequence to its "skeleton":
stacked base pairs $\longrightarrow$ edges, single-stranded regions $\longrightarrow$ vertices.



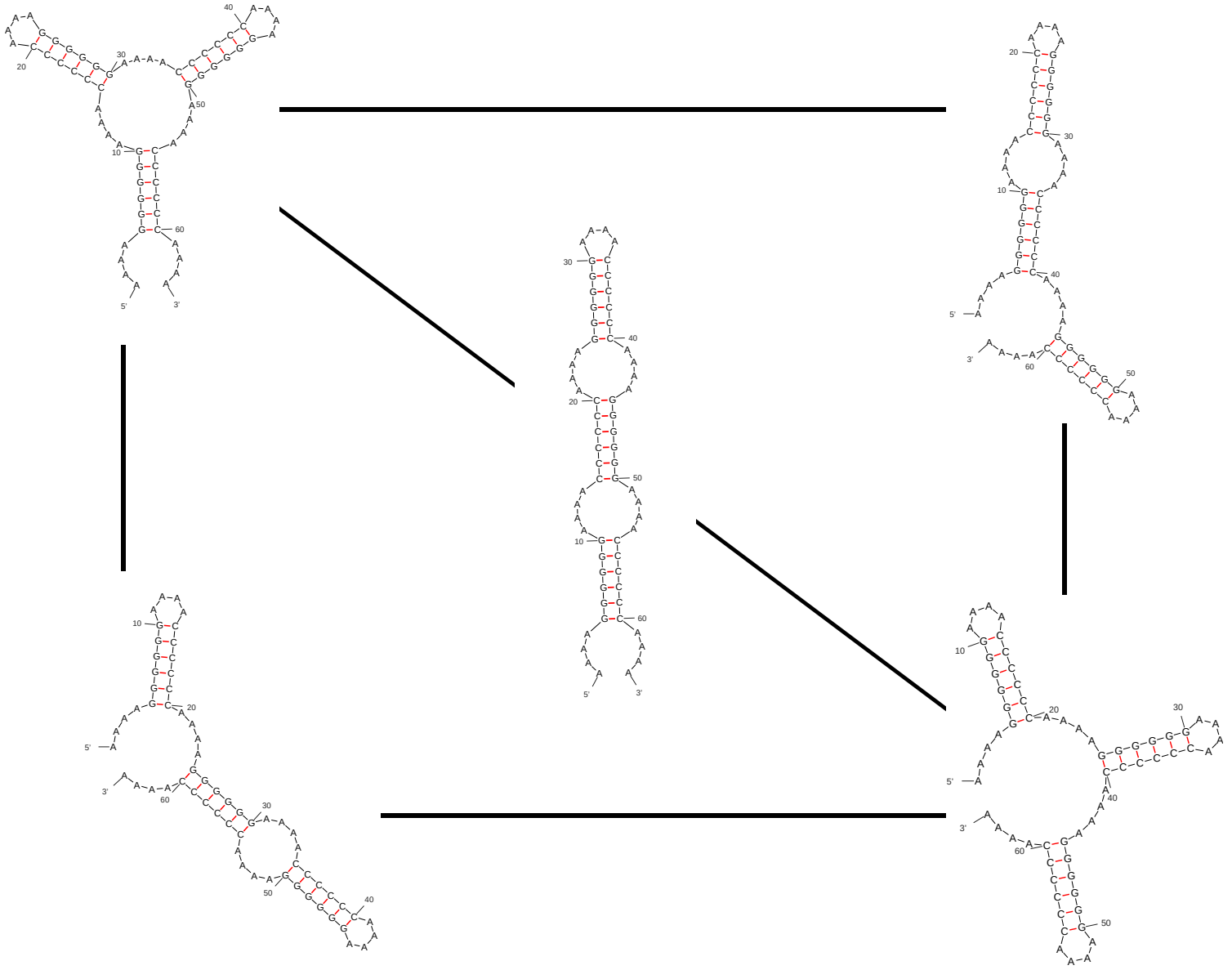
How many possible plane trees T with n edges?

How to compare RNA secondary structures?



Secondary structures for the combinatorial RNA sequence $R =$ AAAA GGGGGG AAAA CCCCCC AAAA GGGGGG AAAA CCCCCC AAAA GGGGGG AAAA CCCCCC AAAA.

What is the space of possible RNA configurations?



Metric: generalized distance

Definition. *Let X be a set. A metric d on X is a function $d : X \times X \to \mathbb{R}$ where for all $x, y, z \in X$*

1. $d(x, y) \geq 0$, *(nonnegativity)*

2. $d(x, y) = 0$ *if and only if* $x = y$, *(with nonnegativity gives positive definiteness)*

3. $d(x, y) = d(y, x)$, *and* *(symmetry)*

4. $d(x, z) \leq d(x, y) + d(y, z)$. *(triangle inequality)*

Two metrics on strings

Let A be a finite set of symbols, and $x, y \in A^+$.

Definition. *If $|x| = |y|$, then the Hamming distance of x and y is*

$$d_H(x, y) = |\{i : 1 \leq i \leq n,\ x_i \neq y_i\}|$$

Definition. *The Levenshtein distance d_L of x and y is the minimum number of insertions, deletions, or substitutions to transform x into y.*

Example. Let $A = \{0, 1\}$, $x = 101010$, and $y = 010101$. Then $d_H(x, y) = 6$ and $d_L(x, y) = 2$.

Norm: generalized size

Definition. *Let V be an n dimensional vector space over a field F. A norm $||\cdot||$ on V is a function $||\cdot|| : V \to \mathbb{R}$ where for all $v, u \in V$ and $a \in F$*

1. $||v|| > 0$ *when* $v \neq 0$, *(nonnegativity)*

2. $||v|| = 0$ *if and only if* $v = 0$,
(with nonnegativity gives positive definiteness)

3. $||av|| = |a|||v||$, *and* *(positive homogeneity)*

4. $||v + u|| \leq ||v|| + ||u||$. *(triangle inequality)*

Metrics and norms

Let V be an n dimensional vector space over a field F with $v, u, w \in V$ and $a \in F$.

If V has a norm $||\cdot||$, define a metric d on V by

$$d(v, u) = ||v - u|| \text{ for all } v, u \in V.$$

Conversely, if V has a metric d which satisfies

1. $d(v, u) = d(v + w, u + w)$, and (translation invariance)
2. $d(av, au) = |a|d(v, u)$, (homogeneity)

then define a norm on V by $||v|| = d(v, 0)$.

Norms on spaces

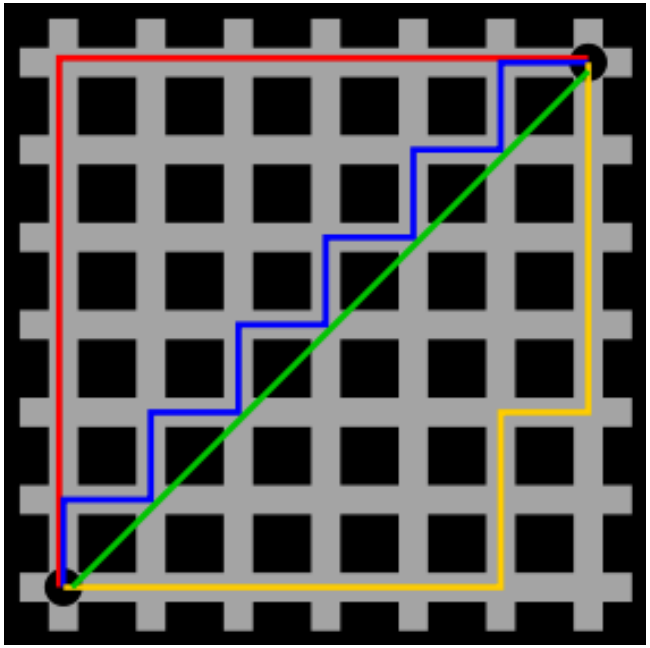
Let V be an n dimensional vector space over a field F with $v = (v_1, v_2, \ldots, v_n) \in V$ and $v_i \in F$.

For $p = 1, 2, \ldots$, the L^p-norm is defined as

$$||v||_p = (\sum_{i=1}^{n} |v_i|^p)^{1/p} \text{ with } ||v||_\infty = \max i|v_i|$$

Example. Let $F = \mathbb{R}$, $n = 3$, and $v = (1, 2, 3)$. Then $||v||_1 = 6$, $||v||_2 = \sqrt{14}$, $||v||_3 = 6^{2/3}$, $||v||_4 = 2^{1/4}\sqrt{7}$, $\ldots$, $||v||_\infty = 3$.

Metrics on spaces



L_1-norm $\longrightarrow$ Minkowski / rectilinear / Manhattan distance

L_2-norm $\longrightarrow$ Euclidean distance

$\vdots$

L_∞-norm $\longrightarrow$ Chebyshev distance

Acknowledgments

- Predicted RNA foldings courtesy of Michael Zuker's `mfold` algorithm, available online through `bioinfo.math.rpi.edu/~zukerm/`.
- Taxicab geometry `http://en.wikipedia.org/wiki/Taxicab_geometry`.