

Finding Recurrent Local and Composite Motifs in RNA 3D Structures

M. Sarver¹ C.L. Zirbel¹ J. Stombaugh² A. Mokdad²
N. Leontis³

¹Department of Mathematics and Statistics

²Department of Biology

³Department of Chemistry

Bowling Green State University

Michigan RNA Society, April 8, 2006

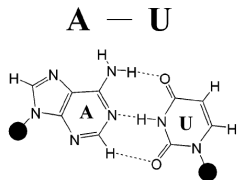
Outline

- 1 Basepairs and Recurrent Motifs
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Outline

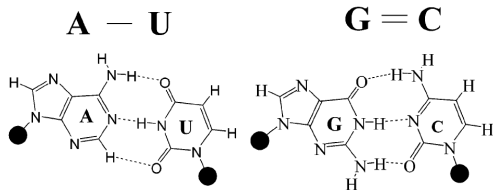
- 1 **Basepairs and Recurrent Motifs**
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Canonical basepairs



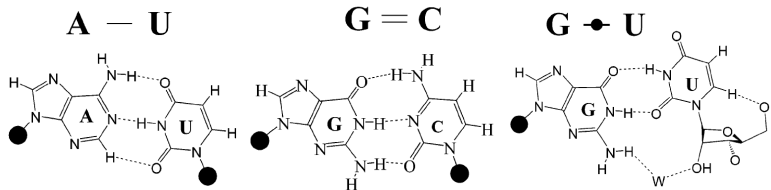
- These basepairs are very common in RNA.
- The interacting edge is the *Watson–Crick* edge.

Canonical basepairs



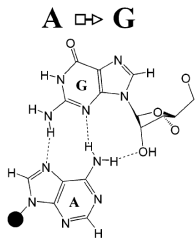
- These basepairs are very common in RNA.
- The interacting edge is the *Watson–Crick* edge.

Canonical basepairs



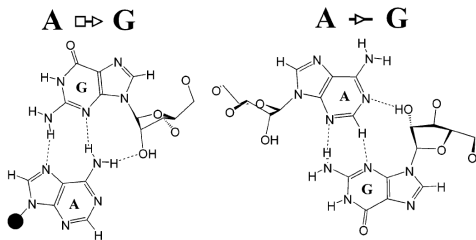
- These basepairs are very common in RNA.
- The interacting edge is the *Watson-Crick* edge.

Noncanonical basepairs



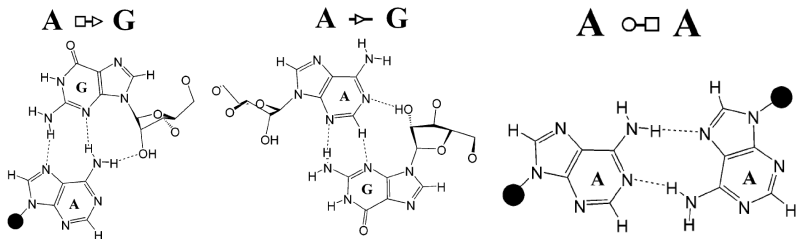
- These basepairs are less common, but still recur.
- They use different edges of the base.

Noncanonical basepairs



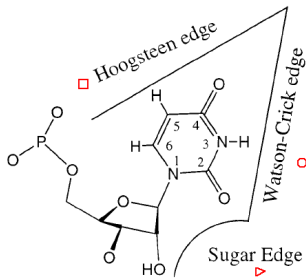
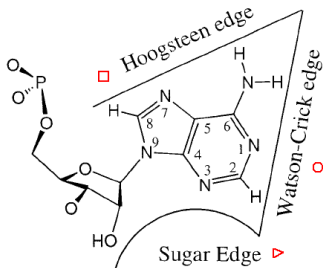
- These basepairs are less common, but still recur.
- They use different edges of the base.

Noncanonical basepairs



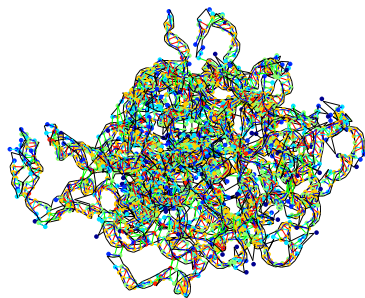
- These basepairs are less common, but still recur.
- They use different edges of the base.

Interacting edges



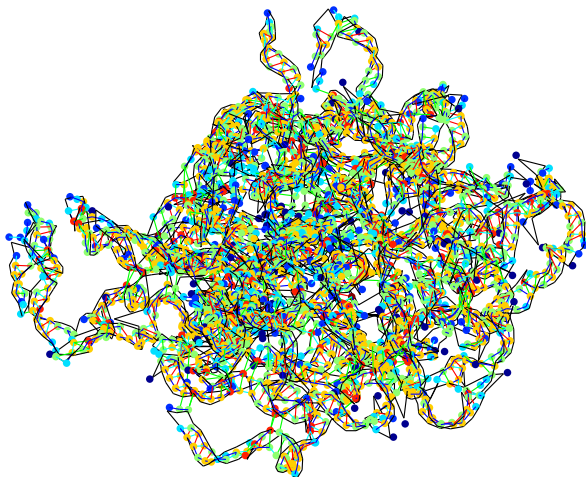
- 4 bases, each with 3 edges, and 2 possible orientations
- Some 180 basepair combinations have been identified and named.
- Classification of interactions by Leontis, Stombaugh, Westhoff (2002) NAR.

Interactions in the 23S RNA



- Bases represented by dots, interactions by edges.
- Red edges are Watson–Crick Watson–Crick interactions
- Roughly 70% of basepairs are canonical, 30% noncanonical.

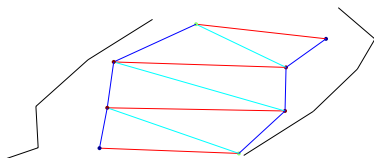
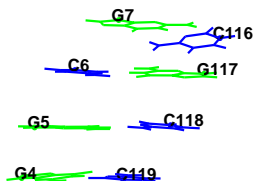
Interactions in the 23S RNA



Outline

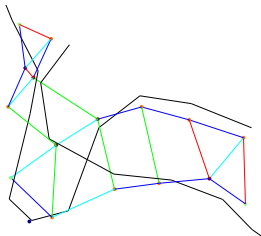
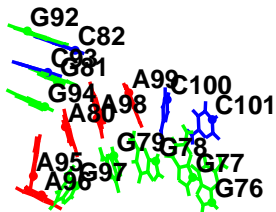
- 1 **Basepairs and Recurrent Motifs**
 - Canonical and noncanonical basepairs
 - **Recurrent motifs**
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Recurrent motifs



- Helices have cis WC–WC basepairs (red lines) and stacking (light and dark blue lines)

Recurrent motifs



- Kink–turns connect two helices at an angle
- Their internal structure requires non–canonical interactions, shown by green lines.
- Some bases make multiple basepairs.

Outline

- 1 Basepairs and Recurrent Motifs
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Symbolic versus geometric search

- You can search for a given pattern of basepairing and stacking interactions, which is very fast. We call these **symbolic constraints**. Such searches will miss certain candidates and may not do well with hairpins loops, for example.
- In a **Geometric search**, given a **query motif**, you find **candidate motifs** which are geometrically similar to the query motif, and you rank them according to degree of similarity.
- In a **combined** search, you screen the candidates found geometrically according to symbolic constraints.

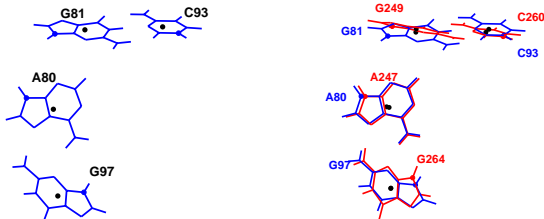
Symbolic versus geometric search

- You can search for a given pattern of basepairing and stacking interactions, which is very fast. We call these **symbolic constraints**. Such searches will miss certain candidates and may not do well with hairpins loops, for example.
- In a **Geometric search**, given a **query motif**, you find **candidate motifs** which are geometrically similar to the query motif, and you rank them according to degree of similarity.
- In a **combined** search, you screen the candidates found geometrically according to symbolic constraints.

Symbolic versus geometric search

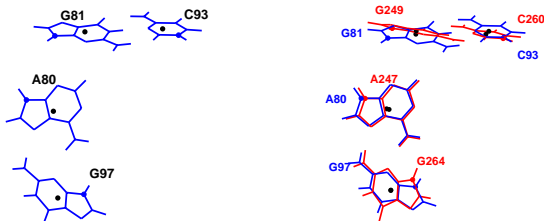
- You can search for a given pattern of basepairing and stacking interactions, which is very fast. We call these **symbolic constraints**. Such searches will miss certain candidates and may not do well with hairpins loops, for example.
- In a **Geometric search**, given a **query motif**, you find **candidate motifs** which are geometrically similar to the query motif, and you rank them according to degree of similarity.
- In a **combined** search, you screen the candidates found geometrically according to symbolic constraints.

Geometric discrepancy



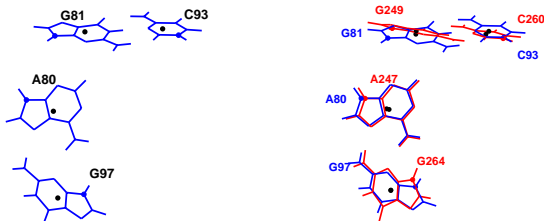
- Consider a query motif (blue) and a candidate motif (red).
- Rigidly move candidate to align base centers (black dots).
- The **fitting error** L is the RMS sum of distances between corresponding base centers.
- The **orientation error** A is the RMS sum of angles required to rotate candidate bases onto query bases.
- **Geometric discrepancy** $D = \frac{1}{m} \sqrt{L^2 + A^2}$.

Geometric discrepancy



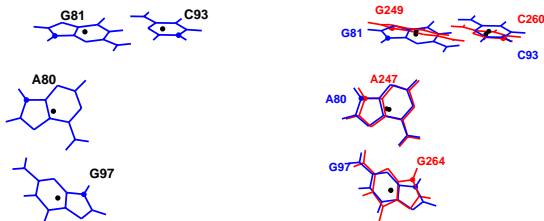
- Consider a query motif (blue) and a candidate motif (red).
- Rigidly move candidate to align base centers (black dots).
- The **fitting error** L is the RMS sum of distances between corresponding base centers.
- The **orientation error** A is the RMS sum of angles required to rotate candidate bases onto query bases.
- **Geometric discrepancy** $D = \frac{1}{m} \sqrt{L^2 + A^2}$.

Geometric discrepancy



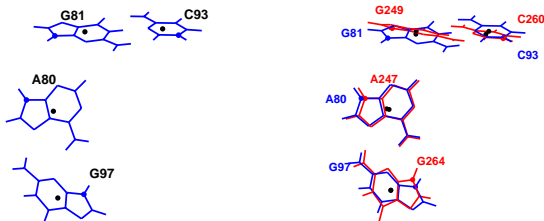
- Consider a query motif (blue) and a candidate motif (red).
- Rigidly move candidate to align base centers (black dots).
- The **fitting error** L is the RMS sum of distances between corresponding base centers.
- The **orientation error** A is the RMS sum of angles required to rotate candidate bases onto query bases.
- **Geometric discrepancy** $D = \frac{1}{m} \sqrt{L^2 + A^2}$.

Geometric discrepancy



- Consider a query motif (blue) and a candidate motif (red).
- Rigidly move candidate to align base centers (black dots).
- The **fitting error** L is the RMS sum of distances between corresponding base centers.
- The **orientation error** A is the RMS sum of angles required to rotate candidate bases onto query bases.
- **Geometric discrepancy** $D = \frac{1}{m} \sqrt{L^2 + A^2}$.

Geometric discrepancy



- Consider a query motif (blue) and a candidate motif (red).
- Rigidly move candidate to align base centers (black dots).
- The **fitting error** L is the RMS sum of distances between corresponding base centers.
- The **orientation error** A is the RMS sum of angles required to rotate candidate bases onto query bases.
- **Geometric discrepancy** $D = \frac{1}{m} \sqrt{L^2 + A^2}$.

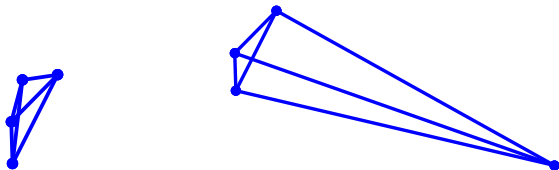
Too many candidates!

- In the 23S, there are 2754 bases. For a 4–nucleotide query motif, that makes for $2754 \cdot 2753 \cdot 2752 \cdot 2751 = 57,399,639,825,024$ possible candidate motifs.
- You cannot calculate the discrepancy for every conceivable candidate motif.
- Instead, set a **cutoff discrepancy** D_0 and find all candidates whose discrepancy with the query motif is smaller than D_0 .

Outline

- 1 Basepairs and Recurrent Motifs
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - **Screening algorithm**
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Rejecting candidates quickly



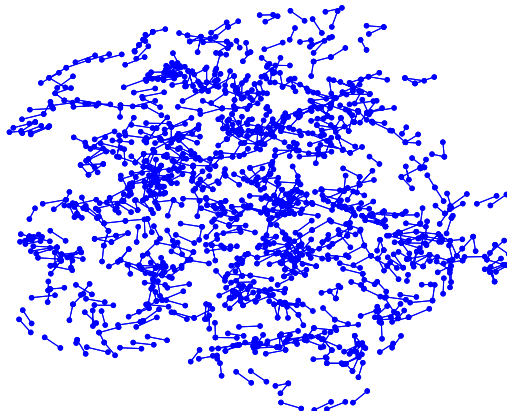
- Many candidates are nowhere close to the query motif.
- We derive the inequality:

$$D \geq \frac{1}{m} \sqrt{\frac{1}{\sum_{i \in I} w_i} \left(\sum_{\substack{i, j \in I \\ i < j}} w_i w_j (Q_{ij} - C_{ij})^2 \right)}$$

where Q_{ij} is the distance between centers of bases i and j in the query motif, and C_{ij} for the candidate.

Screening algorithm

- Focus on bases 1 and 2 in the query motif.
- Find all pairs in the structure whose distances are similar.



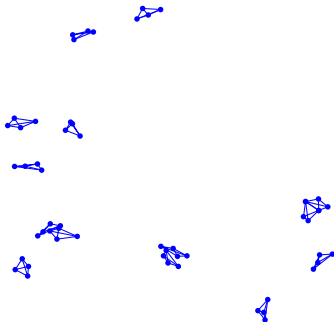
Screening algorithm

- Focus on bases 1, 2, and 3 in the query motif.
- Find all triples in the structure whose distances are similar.



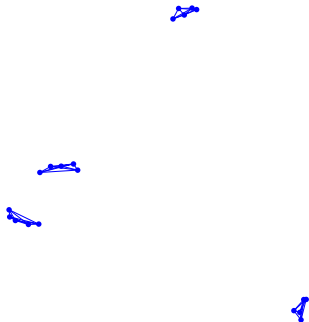
Screening algorithm

- Focus on bases 1, 2, 3, and 4 in the query motif.
- Find all quadruples in the structure whose distances are similar.



Screening algorithm

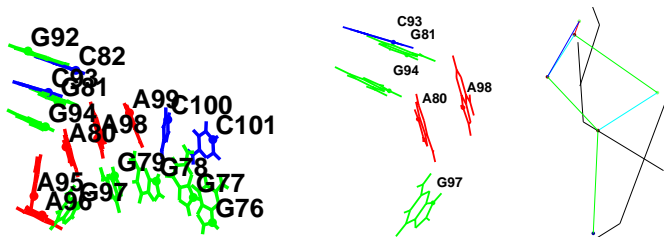
- Look at all bases in the query motif.
- Find all quintuples in the structure whose distances are similar.



Outline

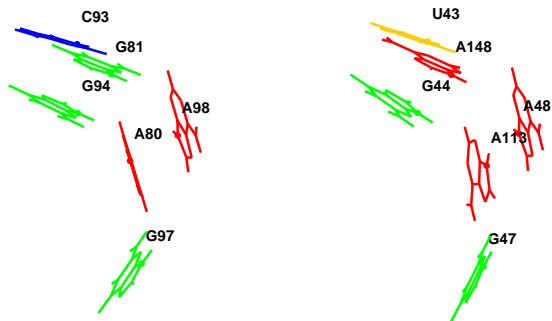
- 1 Basepairs and Recurrent Motifs
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - **Kink–turn central basepair search**
 - Kink–turn closing basepair search

Kink–turn central basepair combined search



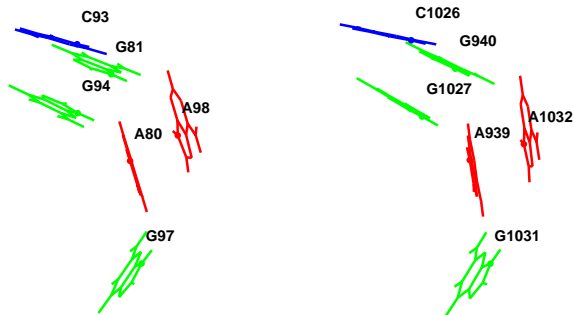
- Query motif (center) focuses on the **central bases** in the turn
- Require only the interaction A80-G97 trans Hoogsteen–Sugar (symbolic constraint)

Results of Kink-turn central basepair search



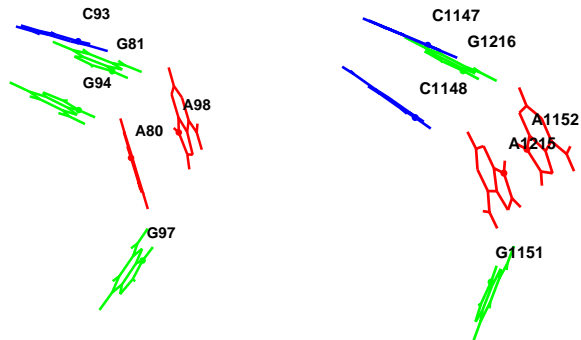
- Candidate 1 has discrepancy 0.1720. Composite.

Results of Kink–turn central basepair search



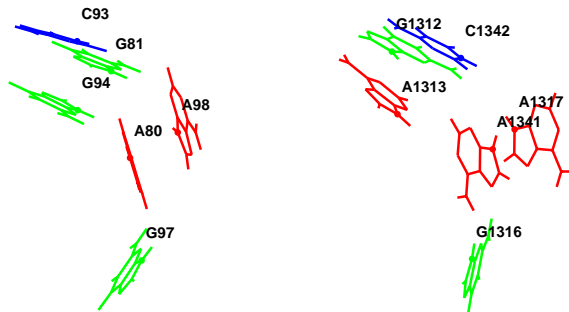
- Candidate 2 has discrepancy 0.2196. Local.

Results of Kink–turn central basepair search



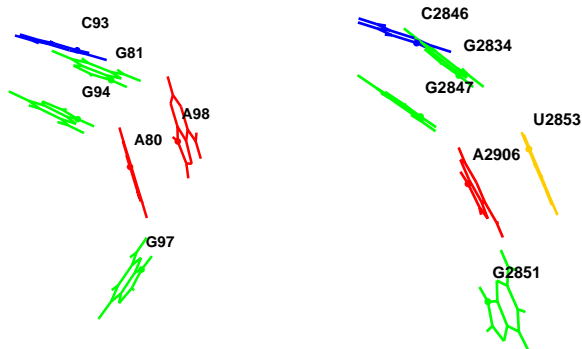
- Candidate 3 has discrepancy 0.2402. Local.

Results of Kink-turn central basepair search



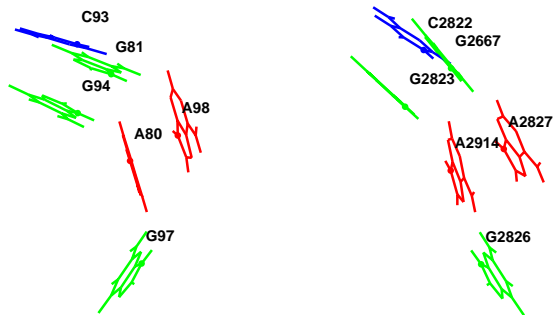
- Candidate 4 has discrepancy 0.3874. Local.

Results of Kink-turn central basepair search



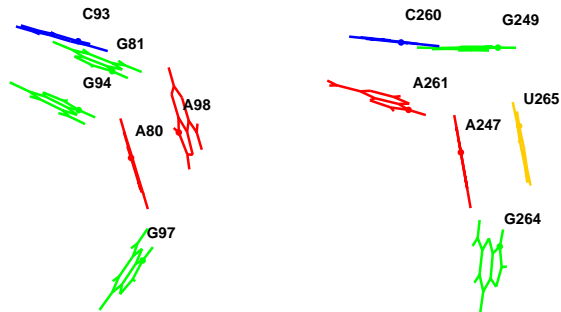
- Candidate 5 has discrepancy 0.5186. Composite and new.

Results of Kink-turn central basepair search



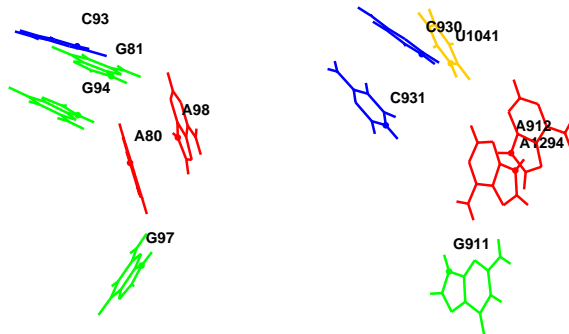
- Candidate 6 has discrepancy 0.5259. Composite.

Results of Kink–turn central basepair search



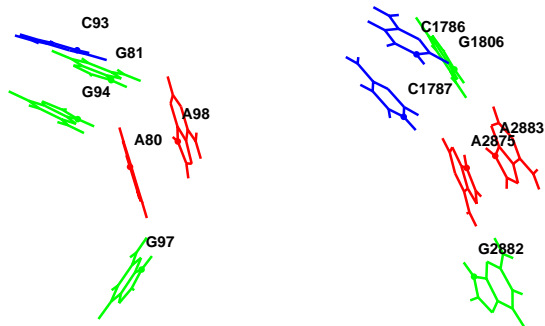
- Candidate 7 has discrepancy 0.5335. Local.

Results of Kink-turn central basepair search



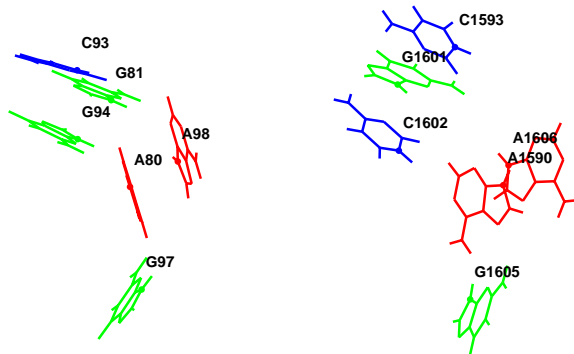
- Candidate 8 has discrepancy 0.5373. Tertiary.

Results of Kink-turn central basepair search



- Candidate 9 has discrepancy 0.5617. Tertiary.

Results of Kink–turn central basepair search

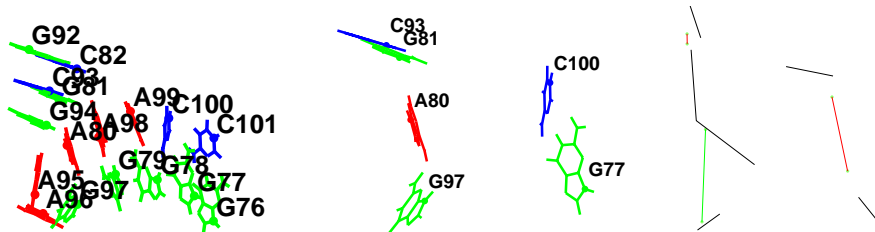


- Candidate 10 has discrepancy 0.6406. Local.

Outline

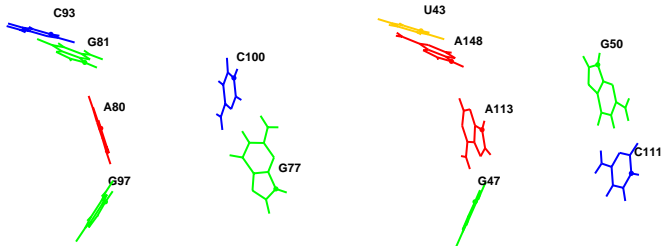
- 1 Basepairs and Recurrent Motifs
 - Canonical and noncanonical basepairs
 - Recurrent motifs
- 2 Geometric search for motifs
 - Geometric search and geometric discrepancy
 - Screening algorithm
- 3 Examples of searches
 - Kink–turn central basepair search
 - Kink–turn closing basepair search

Kink–turn closing basepair combined search



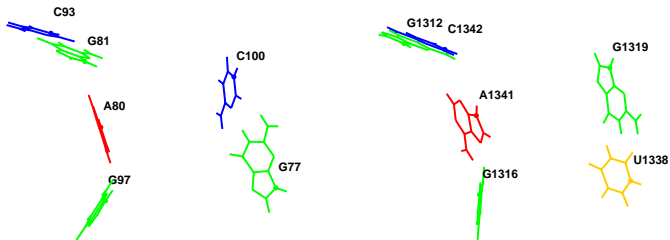
- Query motif (center) focuses on the **closing basepairs**
- Require only the interaction A80-G97 trans Hoogsteen–Sugar (symbolic constraint)

Results of Kink–turn central basepair search



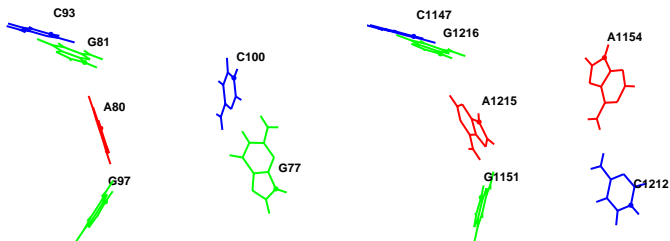
- Candidate 1 has discrepancy 0.2166. Composite.

Results of Kink–turn central basepair search



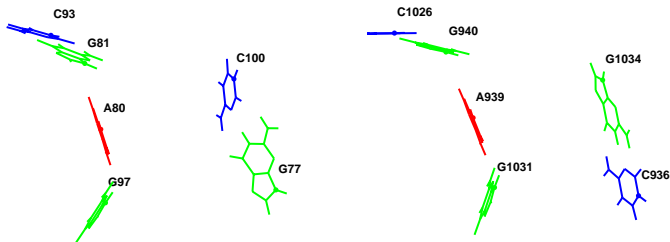
- Candidate 2 has discrepancy 0.5406. Local.

Results of Kink–turn central basepair search



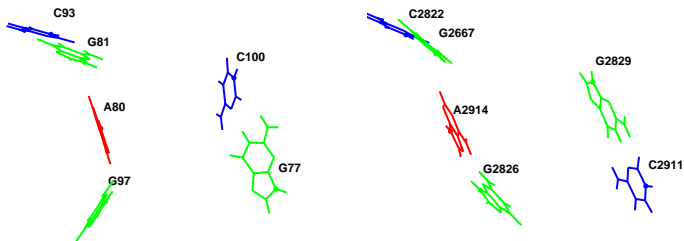
- Candidate 3 has discrepancy 0.5887. Local.

Results of Kink–turn central basepair search



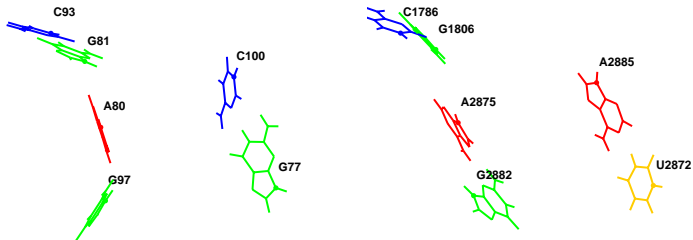
- Candidate 4 has discrepancy 0.6137. Local.

Results of Kink–turn central basepair search



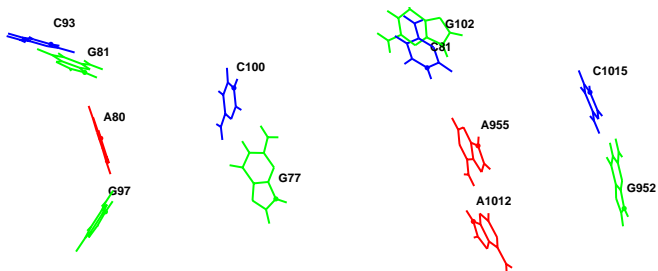
- Candidate 5 has discrepancy 0.6735. Composite.

Results of Kink–turn central basepair search



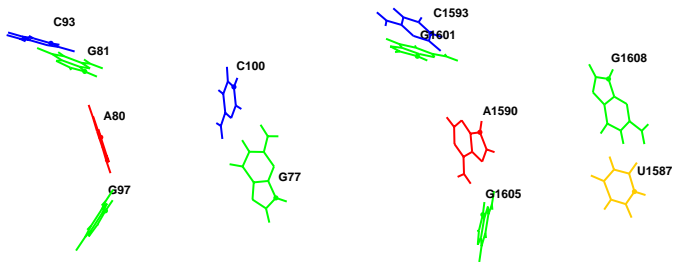
- Candidate 6 has discrepancy 0.6814. Tertiary.

Results of Kink–turn central basepair search



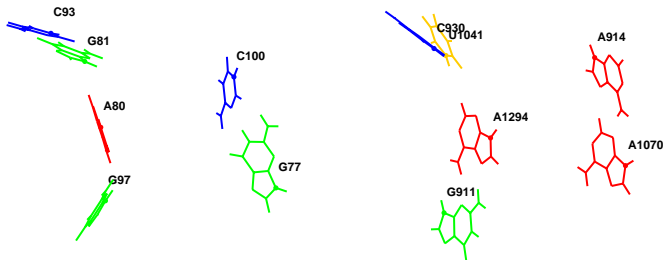
- Candidate 7 has discrepancy 0.6912. Tertiary.

Results of Kink–turn central basepair search



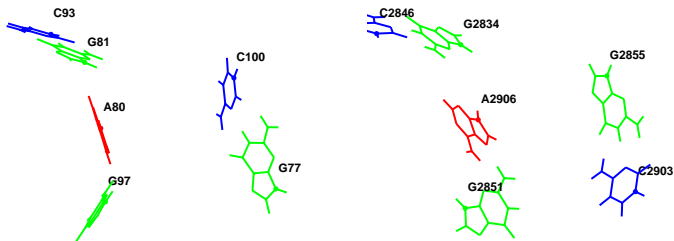
- Candidate 8 has discrepancy 0.7125. Local.

Results of Kink–turn central basepair search



- Candidate 9 has discrepancy 0.7337. Tertiary.

Results of Kink–turn central basepair search



- Candidate 10 has discrepancy 0.7340. Local.

Summary

- Noncanonical basepairs and recurrent motifs hold RNA together
- We can find and rank motifs similar to a given query motif
- We can apply symbolic constraints to narrow the search and reduce search time
- The program FR3D is available at <http://rna.bgsu.edu/FR3D>
Soon there will be a Graphical User Interface (GUI)

Search tables I

		Query Motif 1 Search							Type
		Discrepancy	Motifs found by FR3D						
Query Motif (Kink-turn):		0.0000	A 80	G 97	G 81	C 93	G 94	A 98	L
PDB File:	1S72	0.1720	A 113	G 47	A 148	U 43	G 44	A 48	C
Number of Search Nucleotides:	6	0.2196	A 939	G1031	G 940	C1026	G1027	A1032	L
Motifs Identified:	8 (All)	0.2402	A1215	G1151	G1216	C1147	C1148	A1152	L
Basepairs Constrained:	1	0.3874	A1341	G1316	C1342	G1312	A1313	A1317	L
Guaranteed Cutoff D_0 :	0.7	0.5186	A2906	G2851	G2834	C2846	G2847	U2853	C
Relaxed Cutoff D_1 :	0.7	0.5259	A2914	G2826	G2667	C2822	G2823	A2827	C
Exclude Redundant Candidates:	Yes	0.5335	A 247	G 264	G 249	C 260	A 261	U 265	L
		0.5373	A1294	G 911	U1041	C 930	C 931	A 912	T
		0.5617	A2875	G2882	G1806	C1786	C1787	A2883	T
		0.6406	A1590	G1605	C1593	G1601	C1602	A1606	L

Search tables I

		Query Motif 2 Search							Type
		Discrepancy	Motifs found by FR3D						
Query Motif (Kink-turn):		0.0000	A 80	G 97	G 81	C 93	C 100	G 77	L
PDB File:	1S72	0.2166	A 113	G 47	A 148	U 43	G 50	C 111	C
Number of Search Nucleotides:	6	0.5406	A1341	G1316	C1342	G1312	G1319	U1338	L
Motifs Identified:	8 (All)	0.5887	A1215	G1151	G1216	C1147	A1154	C1212	L
Basepairs Constrained:	1	0.6137	A 939	G1031	G 940	C1026	G1034	C 936	L
Guaranteed Cutoff D_0:	0.9	0.6735	A2914	G2826	G2667	C2822	G2829	C2911	C
Relaxed Cutoff D_1:	0.9	0.6814	A2875	G2882	G1806	C1786	A2885	U2872	T
Exclude Redundant Candidates:	Yes	0.6912	A 955	A1012	C 81 (5S)	G 102 (5S)	C1015	G 952	T
		0.7125	A1590	G1605	C1593	G1601	G1608	U1587	L
		0.7337	A1294	G 911	U1041	C 930	A 914	A1070	T
		0.7340	A2906	G2851	G2834	C2846	G2855	C2903	C
		0.7631	A 520	G 23	A 639	G1363	U 26	U 517	T
		0.7820	A 80 (5S)	G 102 (5S)	G 956	A1012	C 106 (5S)	G 75 (5S)	T
		0.7859	A1318	G1339	U 27	A 516	C1342	A1313	T
		0.8003	A1459	G1484	A 784	U 862	G1489	C1456	T
		0.8126	A1294	G 911	A1040	C 931	A 913	A1070	C
		0.8274	A 666	G 680	G 209	C 230	G 684	C 663	T
		0.8298	A 247	G 264	G 249	C 250	G 267	C 244	L
		0.8582	A 242	G 269	C 377	G 273	G 379	G 431	T
		0.8731	A1626	G1571	G1510	G1496	C1574	G1621	T