

Computational Bioinformatics: Software and Databases

Jason T. L. Wang, Professor

**Bioinformatics Program and Computer Science Department
New Jersey Institute of Technology**

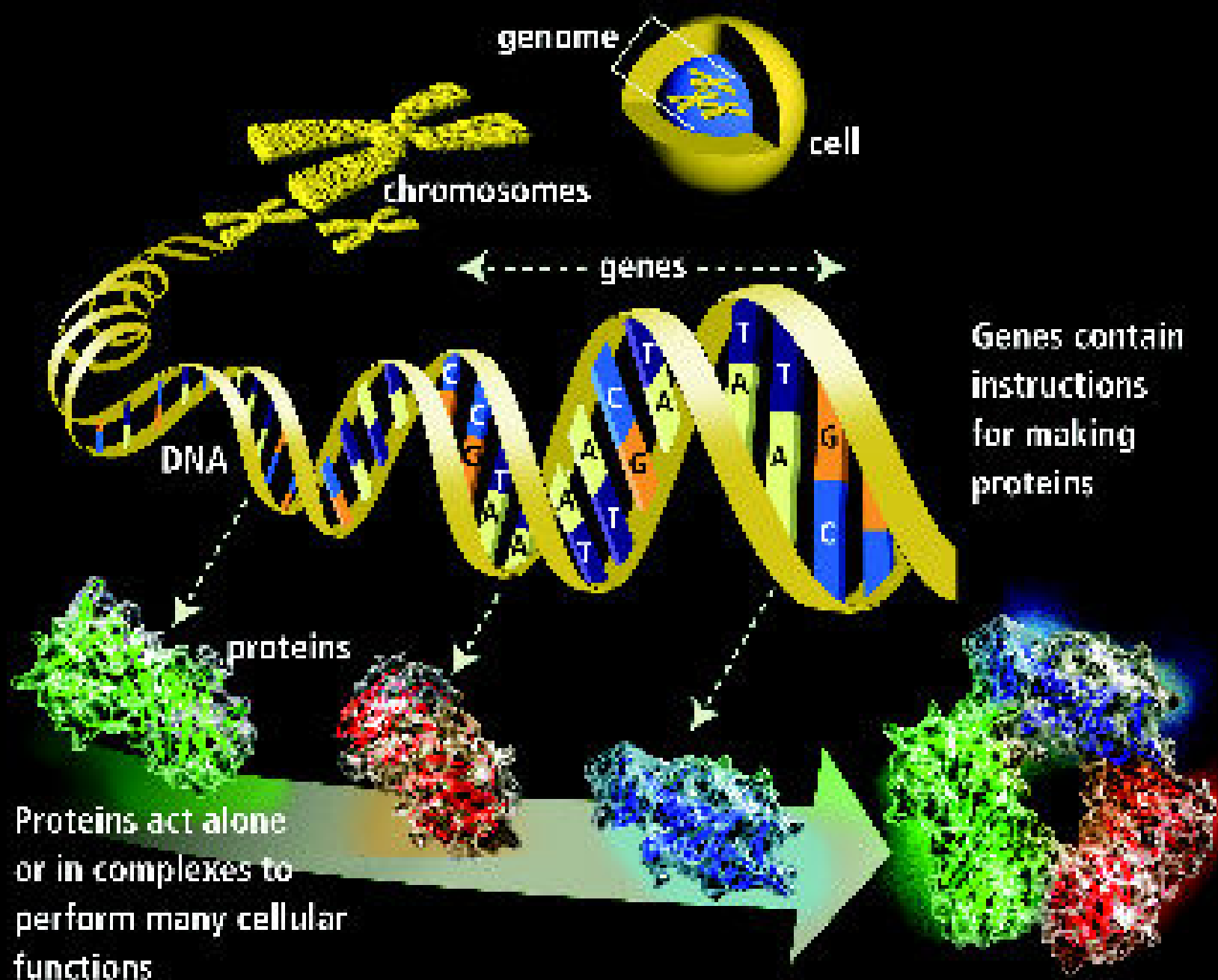
<http://web.njit.edu/~wangj>

Work supported by NSF grant IIS-0707571

Presentation for NSF-Sponsored C2PRISM Program

Outline

- Introduction to Bioinformatics
- Introduction to Computational RNA Genomics (Our Current Project)
- RNA Informatics Tools
- RNA Databases
- Bioinformatics Center
- Conclusion and Future Work

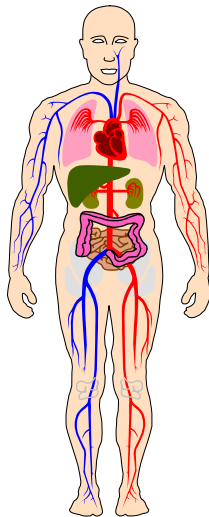


Genes contain instructions for making proteins

Proteins act alone or in complexes to perform many cellular functions

Gene:

- Genetic information-containing elements
- Distributed to each cell when cell divides
- Made of deoxyribonucleic acid --DNA



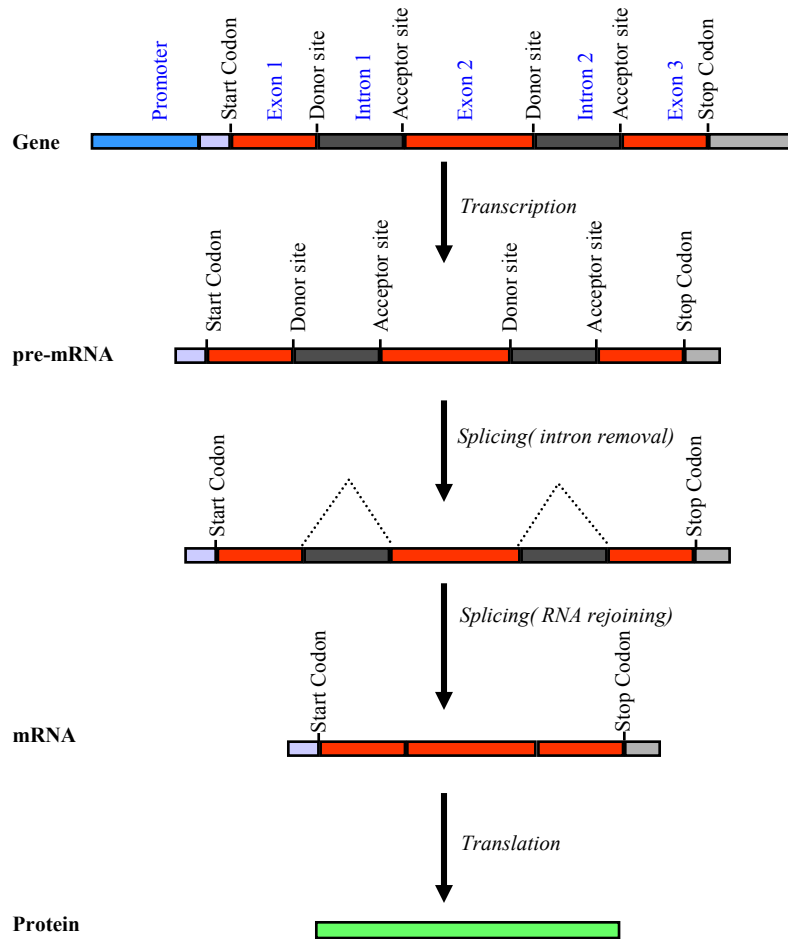
Gene Structure:

- Promoter
- Start codon
- Introns
- Exons
- Stop codon
- etc

Gene:

- Transcription : DNA to RNA
- RNA Splicing: Remove Intons--mRNA
- mRNA translation--Protein

Gene Structure and Gene Expression

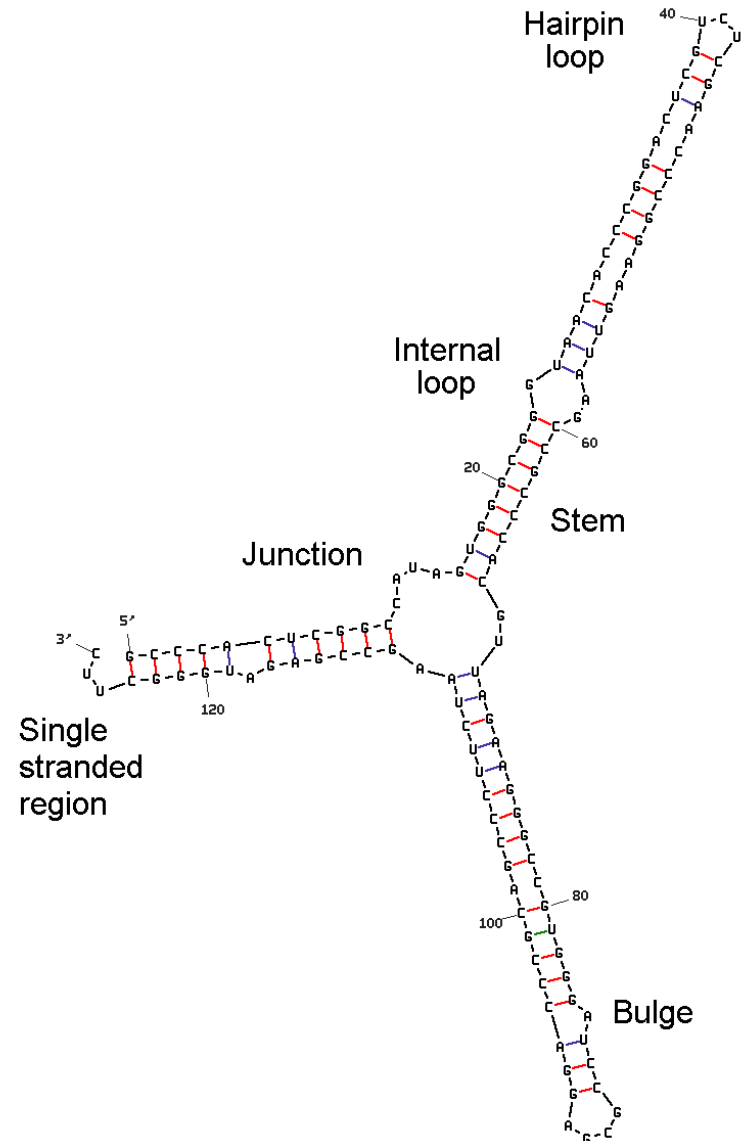


```

CCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCAGGAGCAGG
GAGGGCAGGAGCCAGGGCTGGGCATAAAAAGTCAGGGCAGAGCCATCTAT
Exon 1  TGCTTACATTTGCTTCTGCACACAACGTGTTCTACTAGCAACTCAAACAG
ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTACTGCCCT
GTGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGT
Intron 1 TGGTATCAAGGTTACAAGACAGGTTTAAAGGAGACCAATAGAAACTGGGC
ATGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCT
CTGCCATTGGTCTATTTTCCACCCTTAGGCTGCTGGTGGTCTACCT
Exon 2  TGGACCAGAGGTTCTTTGAGTCCCTTGGGGATCTGTCCACTCCTGATG
CTGTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTG
CCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAA
CCTTCAGGTTGAGTCTATGGGACCCTTGATGTTTTCTTTCCCTCTTTT
CTATGGTTAAGTTCATGTATAGGAAGGGGAGAAGTAACAGGGTACAGT
TtagAATGGGAAACAGACGAATGATGTCATCAGTGTGGAAGTCTCAGGA
TCGTTTTAGTTTTCTTTATGTCTGTTACATAACAAATGTTTTTTGTT
TAATCTTGCTTTCTTTTTTTTTTCTTCTCCGCAATTTTTACTATTATAC
TTAATGCCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATACA
TTAAGTAACTTAAAAAAACTTTACACAGCTGCTCCTAGTACATTACTA
TTTTGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTA
TTTTCTTTTTTTTTTAAATTGATACATAATCATTATACATATTTATGGGT
Intron 2 TAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAAATCAGGGTA
ATTTTGCATTTGTAAATTTAAAAAATGCTTTCTTTTAAATATACTTT
TTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTTCAGGG
CAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAA
CAGTGATAATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAAT
ATTTCTGCATATAAATTTGTAACCTGATGTAAGAGGTTTCAATTTGCTAAT
AGCAGCTACAATCCAGCTACCATTCTGCTTTTTATTTATGGTTGGGATA
AGGCTGGATTATCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTC
ATACCTCTATCTTCCCTCCACAGCTCCTGGGCAACGTGCTGGTCTGTGT
Exon 3  GCTGGCCCATCACTTTGGCAAGAATTCAACCACAGTGCAGGCTGCCT
ATCAGAAAGTGGTGGCTGGTGGCTAATGCCCTGGCCCAAGTATCA
CTAAGCTCGTTTCTTGCTGTCCAATTTCTATTAAAGTTCTTTGTTC
CCTAAGTCCAACTACTAAACTGGGGATATTATGAAGGGCTTGGCAT
CTGGATTCTGCCTAATAAAAAACATTTATTTTTCATTGCAATGATGTATT
TAAATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGGAGGTCAAGTG
CATTTAAAACATAAAGAAATGATGAGCTGTTCAAACCTTGGGAAAAATAC
ACTATATCTTAACTCCATGAAAGAAGGTGAGGCTGCAACCAGCTAATG
CACATTGGCAACAGCCCCGATGCCTAATGCACATTGGCAACAGCCCCCT
GATGCCATAGCCTTATTCATCCCTCAGAAAAGGATTCTTTGAGAGGCTT
GATTTGCAAGTTAAAGTTTTGCTATGCTGATTTTACATTACTTATTTG
TTTAGCTGTCTCATGAATGTCTTTTC
    
```

Computational RNA Genomics

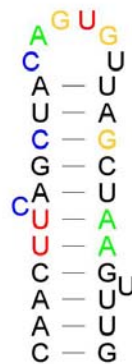
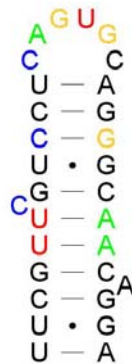
- Biochemical and genetic studies have demonstrated many functions associated with the UTRs in mRNAs.
- Unlike proteins, RNA sequence search is insufficient for detecting similarity.



Sequence Similarity vs. Structural Similarity

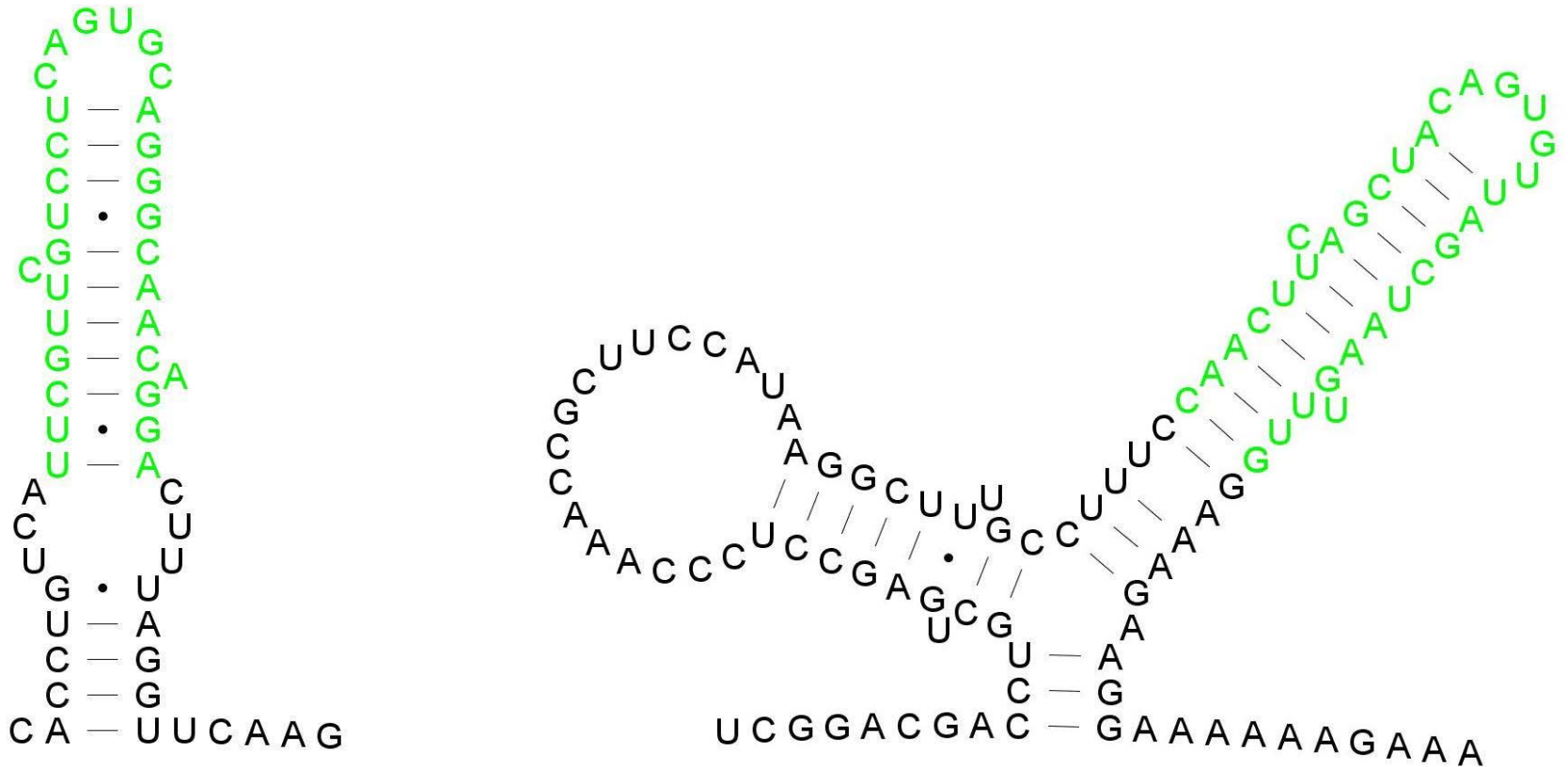
```

>NM_000032
UUCGUUCGUCCUCAGUGCAGGGCAACAGGA
(((((((.((((((. . . . .)))))))).)))
>NM_014585
CAACUUCAGCUACAGUGUUAGCUAAGUUUG
(((((((.((((((. . . . .)))))))).)))
    
```



RSmatch and RADAR

(BMC Bioinformatics 2005)
(Nucleic Acids Research 2007)



Alignment of two RNA secondary structures where the local matches found by RSmatch are in green.

RADAR - Windows Internet Explorer

http://datalab.njit.edu/biodata/rna/RSmatch/server.htm

Links Customize Links Free AOL & Unlimited Internet Free Hotmail Windows Windows Marketplace Windows Media

Search web... Favorites Maps Spaces

Google Go Bookmarks PageRank Popups okay Check AutoLink AutoFill Send to Settings

☆ + RADAR

RADAR

RNA Data Analysis and Research

[\[Home\]](#) [\[Web Server\]](#) [\[Help\]](#) [\[Download Software\]](#)

RADAR is a web-based server for RNA data analysis and research. RADAR can align structure-annotated RNA sequences so that both sequence and structure information are taken into consideration. This server is capable of performing database search, multiple structure alignment, and pair-wise structure comparison. In addition, RADAR provides two salient features: (1) constrained alignment of RNA secondary structures, and (2) prediction of the consensus structure for a set of RNA sequences. RADAR assists scientists in performing many important RNA mining operations, including understanding the functionality of RNA sequences, the detection of structural RNA motifs and the clustering of RNA molecules, among others. RADAR uses the RSmatch algorithm for alignment. Standalone [RSmatch v2.0](#), which is RADAR, is freely available for interested researchers.

References:

1. J. Liu, J.T.L. Wang, J. Hu and B. Tian. [A method for aligning RNA secondary structures and its application to RNA motif detection](#). *BMC Bioinformatics*, 6:89, 2005.
2. M. Khaladkar, V. Bellofatto, J.T.L. Wang, B. Tian and B.A. Shapiro. [RADAR: a web server for RNA data analysis and research](#). *Nucleic Acids Research*, 35:W300-W304, 2007.

[A map of 20 most recent site visitors](#)

For any suggestions, comments or queries about this website, please contact jason.t.wang@njit.edu.

Done Internet 100%

RADAR

RNA Data Analysis and Research

[Home] [Consensus Structure Prediction] [Clustering] [Sequence Folding]

This method compares the given RNA structures against one another and outputs the similarity matrix which consists of the pair-wise alignment scores. This matrix can be used for clustering the RNA structures.

Subject:

Sample RNA structure dataset

Paste input below OR Upload from file

```
>NM_000032:1-52 Homo sapiens aminolevulinate, delta-, synthase 2 (sideroblastic/hypochromic anemia) (ALAS)
CACCCUGUCAUUCGUUCGUCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG
..(((((((.....))))))..)).....
>NM_014585:1-100 Homo sapiens solute carrier family 40 (iron-regulated transporter), member 1 (SLC40A1),
AGCGGGACGCCCCGGGCGGCCUGAAGGGGACGGGGCGGCCAGUCGGAGGUCGCAGGGAGCUCGCCCGCCCGACUCGGUAUAAGAGCUGGGCCCCGCCCCA
..(((((((.....))))))..))..(((((((.....))))))..))..(((((((.....))))))..))..(((((((.....))))))..))..
>NM_002081:51-150 Homo sapiens glypican 1 (GPC1), mRNA
CGCCCGCCCGCCGCUUUGUUGUCUCCGCCUCCUCGGCCGCGCCCGCCUUGGACCGCGAGCCGCGCGCGCCGGGACCUUGGCUCUGCCCUUCGCGGGCGG
..((((.....)))).....(((((((.....))))))..))..(((((((.....))))))..))..(((((((.....))))))..))..
>NM_003449:1001-1100 Homo sapiens tripartite motif-containing 26 (TRIM26), mRNA
GGCAGUACAUUGUGGCUGAGUUUGAGCAGGGUCAUCAGUUCUGAGGGAGCGGGAGGAACACCUCUGGGAACAGCUGGGCGAAGCUGGAGCAGGAGCUCAC
..((((.....))))..(((((((.....))))))..))..(((((((.....))))))..))..(((((((.....))))))..))..(((((((.....))))))..
>NM_018992:451-550 Homo sapiens potassium channel tetramerisation domain containing 5 (KCTD5), mRNA
```

Gap penalty (default -2)
Alignment type Global Local

Score matrix

```
>single-base scoring matrix:
      A  C  G  U
A     1 -1 -1 -1
C    -1  1 -1 -1
G    -1 -1  1 -1
U    -1 -1 -1  1
```

Multiple Structural Alignment

The screenshot shows a web browser window with the address bar displaying `http://aria.njit.edu/biodata/cgi-bin/R5match/entry.cgi`. The main content area displays a multiple structural alignment of five mRNA sequences. The sequences are:

- >NM_014585:1-100 Homo sapiens solute carrier family 40 (iron-regulated transporter), member 1 (SLC40A1), mRNA
- >NM_002081:51-150 Homo sapiens glypican 1 (GPC1), mRNA
- >NM_003449:1001-1100 Homo sapiens tripartite motif-containing 26 (TRIM26), mRNA
- >NM_018992:451-550 Homo sapiens potassium channel tetramerisation domain containing 5 (KCTD5), mRNA
- >NM_000146:1-100 Homo sapiens ferritin, light polypeptide (FTL), mRNA

The alignment is shown as a block of text with dots representing gaps. Below the alignment, a section titled "The structural alignment result" provides summary statistics:

```
=== Result of Multiple Alignment ===
Min: 10.75      Max: 13.78      Avg: 12.77
# STOCKHOLM 1.0

NM_000032:1-52      10      UU-CG-UUCGUCCUCA-GUGCAGGGCAA-CAGGA      39
NM_014585:151-250  52      CA-AC-UUCAGCUACA-GUGUUAGCUAA-GUUUG      81
NM_000146:1-100    14      GUCUCUUGCUUCAACA-GUGUUUGGACG-G-AAC      44
NM_014585:1-100    13      GG-GCG-G-CCUGAAGGGGACGGGGC-GG-CCC      41
#=GC SS_cons      ((.((.(.(((.....)))))).))
//
```

The browser window also shows standard navigation buttons and a status bar at the bottom indicating "Done".

GLEAN-UTR Database

(BMC Genomics 2008)

- Use RADAR, hierarchical clustering and Gene Ontology to mine RNA motifs in the UnTranslated Regions (UTRs) conserved between human and mouse orthologs in multiple genes sharing common biological pathways.
- GLEAN-UTR DB contains 90 RNA motifs (structure groups) from 698 genes. Top two motifs are Iron response element (IRE) and histone 3'-UTR stem-loop structure.

<http://datalab.njit.edu/biodata/GLEAN-UTR-DB/>

GLEAN-UTR-DB - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://datalab.njit.edu/biodata/GLEAN-UTR-DB/

Getting Started Latest Headlines

GLEAN UTR Database

Grouping of Structurally Related RNAs with Gene Ontology

[\[Help\]](#)

Search (by entering a query from the [gene list](#))

RefSeq ID:

OR

Gene ID:

Search (by entering a GO ID)

GO:

Search (by entering a Group ID)

GID:

Done

GLEAN-UTR-DB - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://aria.njit.edu/biodata/cgi-bin/GLEAN-UTR-DB/test.pl?searchname=154&species=human&cv=15.4

Getting Started Latest Headlines

Group from human homolog sequences

Note: The sequences which are shown in brown color are from the 3' UTR and those in green are from the 5' UTR.

Note: Cohesive Value is the average of all the pairwise similarity scores among all the structures in the same group.

Cohesive Value = 15.4

| Gene ID | Gene Symbol | Structure ID | Alignment | mfe (Kcal/Mol) | Gene Name |
|------------------------|-------------|-------------------------------------|---------------------------|----------------|---|
| 4254 | KITLG | NM 000899:1060-1087 | TTTGCTTCATAAAATGAAGCAGC | -8.5 | KIT ligand |
| 23512 | SUZ12 | NM 015355:3606-3643 | TATTCCTTTATTTATAAAGGATC | -3.6 | suppressor of zeste 12 homolog (Drosophila) |
| 6616 | SNAP25 | NM 003081:1331-1430 | T-TTATGCATTATGCATGA-- | -5.4 | synaptosomal-associated protein, 25kDa |
| 204851 | HIPK1 | NM 181358:2687-2710 | T-GACTGCATTGTTGTAGTC-- | -7.8 | homeodomain interacting protein kinase 1 |
| 5931 | RBBP7 | NM 002893:1613-1645 | --G-CTTGATTTATCAAG-C-- | -5.6 | retinoblastoma binding protein 7 |
| | | | .((((((((((.....)))))))). | | |

Related Gene Ontology entries

| GO ID | p-value | Type | Description |
|----------------------------|----------|--------------------|---|
| go:0006350 | 0.018499 | biological process | transcription |
| go:0016568 | 0.0013 | biological process | chromatin modification |
| go:0006355 | 0.033775 | biological process | regulation of transcription\, DNA-dependent |
| go:0008283 | 0.004459 | biological process | cell proliferation |

Done

RSpredict - RNA Secondary Structure Prediction of Multiple Sequence Alignments - Windows Internet Explorer

http://datalab.njit.edu/biology/RSpredict/

Links Customize Links Free AOL & Unlimited Internet Free Hotmail Windows Windows Marketplace Windows Media

Search web... Favorites Maps Spaces

Google Go Bookmarks PageRank Popups okay Check AutoLink AutoFill Send to Settings

RSpredict - RNA Secondary Structure Prediction of M...

RSpredict

[Home](#) [Download](#) [Documentation](#) [Help](#)

RSpredict is an RNA secondary structure prediction tool that works on multiple sequence alignments. It takes into account sequence covariation and employs effective heuristics for accuracy improvement. RSpredict accepts, as input data, a multiple sequence alignment in the Fasta format and outputs the consensus secondary structure of the input sequences in both the Vienna style Dot Bracket format and the Connectivity Table (CT) format.

Please enter the multiple sequence alignment in FASTA format below ([example](#)):

```
>M10740/1-73 Yeast-PHE
GCGGAUUUAGCUCAGUU-GGGAGAGCGCCAGACUGAAGAUUUGGAGGUCCUG-UGUUCGA
UCCACAGAAUUCGCA
>K00349/1-73 Drosophila-PHE
GCCGAAAUAGCUCAGUU-GGGAGAGCGUAGACUGAAGAUCAAAGGUCCCC-GGUCAA
UCCCGGUUUCGGCA
>K00283/1-74 Halobacterium volcanii Lys-tRNA-1
GGGCCGGUAGCUCAUUUAGGCAGAGCGUCUGACUCUUAUCAGACGGUCCG-UGUUCGA
AUCGCGUCCGCCCA
>K00354/1-74 Bacteriophage T4 Pro-tRNA
CUCCGUGUAGCUCAGUUUGGUAGAGCGCCUGAUUUUGGAUCAGGAGGUCCAA-GGUCAA
AUCCUUGUAGGAGA
>X02682/1-73 E. coli transfer RNA-Val-1(CAC)
GGGUGAUUAGCUCAGUU-GGGAGAGCACCUCCUUACAAGGAGGGGUCCGC-GGUUCGA
```

Done Internet 100%

RSpredict - Results - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://web.njit.edu/~js9/RSpredict/RSpredictServer/RSpredict.php

Most Visited Getting Started Latest Headlines

RSpredict

The input alignment has 11 sequences.

The alignment length is: 74

The average pairwise sequence identity is: 60%

The pairing threshold is: -0.34

The energy density of the predicted structure is: -9.87

The predicted structure is:

```
GCGGAAUUAGCUCAGUUUGGUAGAGCGCCAGACUGAAAAUCUGGAGGUCCCCGGUUCGAAUCCCGAAUCCGCA
(((((((...(((.....))))).((((.....))))). ....((((.....)))))))).
```

The input alignment is:

```
>M10740/1-73 Yeast-PHE
GCGGAUUUAGCUCAGUU-GGGAGAGCGCCAGACUGAAGAUUUGGAGGUCCUG-UGUUCGA
UCCACAGAAUUCGCA
>K00349/1-73 Drosophila-PHE
GCCGAAUAGCUCAGUU-GGGAGAGCGUUAGACUGAAGAUCAAAGGUCCCC-GGUUCAA
UCCCGGGUUUCGGCA
>K00283/1-74 Halobacterium volcanii Lys-tRNA-1
GGGCCGGUAGCUCAUUUAGGCAGAGCGUCUGACUCUUAUACAGACGGUCGCG-UGUUCGA
AUCGCGUCCGGCCCA
>K00354/1-74 Bacteriophage T4 Pro-tRNA
CUCCGUGUAGCUCAGUUUGGUAGAGCGCCUGAUUUGGGAUCAGGAGGUCCAA-GGUUCAA
AUCCUUGUAUGGAGA
>X02682/1-73 E. coli transfer RNA-Val-1(CAC)
GGGUGAUUAGCUCAGCU-GGGAGAGCACCUCCUUACAAGGAGGGGGUCGGC-GGUUCGA
UCCCGUCAUACCCCA
>J01624/1-73 E.coli glyW gene, a duplicate gene for gly-tRNA-3
GCGGGAAUAGCUCAGUU-GGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCG-AGUUCGA
GUCUCGUUUCCCGCU
```

Done

Bioinformatics Center - Windows Internet Explorer

http://bioinformatics.njit.edu/

Links: Customize Links, Free AOL & Unlimited Internet, Free Hotmail, Windows, Windows Marketplace, Windows Media

Search web... Favorites, Maps, Spaces, Settings

Bioinformatics Center

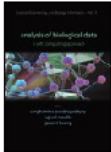
Bioinformatics Center

- Home
- Seminars
- Meetings
- Publications
- Resource
- Software
- Links

Welcome to the Bioinformatics Center, devoted to the areas of bioinformatics and life science informatics. We are proud of our strong research and educational connections to departments and schools at University Heights, Newark, including the Bioinformatics Program, the Information Technology Program, the Division of Biological Sciences, and the College of Computing Sciences at NJIT; the Public Health Research Institute; the University of Medicine and Dentistry of New Jersey; and Rutgers University.


If you cannot find the information you are looking for on our Web site, please [send us a message](#) and we will do our best to provide it.

Highlights



Bioinformatics, a field devoted to the interpretation and analysis of biological data using computational techniques, has evolved tremendously in recent years due to the explosive growth of biological information generated by the scientific community. Soft computing is a consortium of methodologies that work synergistically and provides, in one form or another, flexible information processing capabilities for handling real-life ambiguous situations. Several research articles dealing with the application of soft computing tools to bioinformatics have been published in the recent past; however, they are scattered in different journals, conference proceedings and technical reports, thus causing inconvenience to readers, students and researchers. This book, unique in its nature, is aimed at providing a treatise in a unified framework, with both theoretical and experimental results, describing the basic principles of soft computing and demonstrating the various ways in which they can be used for analyzing biological data in an efficient manner. Interesting research articles from eminent scientists around the world are brought together in a systematic way such that the reader will be able to understand the issues and challenges in this domain, the existing ways of tackling them, recent trends, and future directions. This book is the first of its kind to bring together two important research areas, soft computing and bioinformatics, in order to demonstrate how the tools and techniques in the former can be used for efficiently solving several problems in the latter.

Links: [SOFTBIO Book](#)



Bioinformatics is the science of managing, mining, integrating, and interpreting information from biological data at the genomic, metabolomic, proteomic, phylogenetic, cellular, or whole organism levels. The need for bioinformatics tools and expertise has increased as genome sequencing projects have

Events

October 6, 2008
c-Myc Target Prediction
Dr. Yili Chen
2:30 PM, GITC 4415

Internet 100%

Conclusion

- We have developed a warehouse of informatics tools and databases for RNA genomics.
- We want to invite high school students to our research team to conduct interesting research (Liberty Science Center Model)
- Contact Dr. Jason Wang (wangj@njit.edu)
- <http://web.njit.edu/~wangj>