

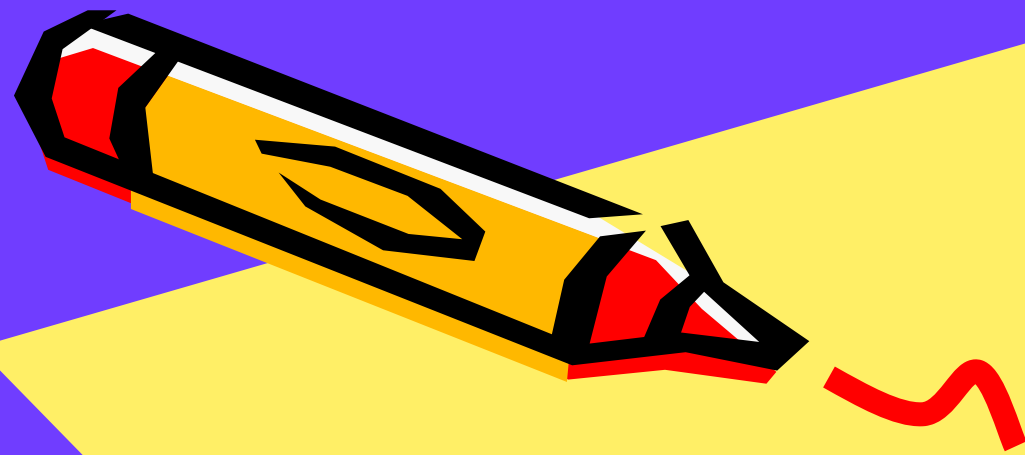
# ***Bioinformatics Tools for RNA Data Analysis***

Joseph Santos  
Bloomfield Tech High School  
Bloomfield, New Jersey

# Contents

- What is Bioinformatics?
- Vocabulary (with metaphors)
- 1D--Sequence (BLAST)
- 2D--Secondary Structure (RSmatch)





# *What is Bioinformatics?*

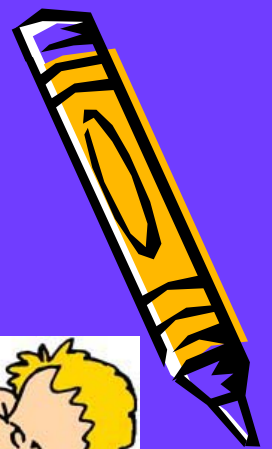
"Let's learn the Latin roots"



You failed your Latin exam!  
But Sweety, it's important to learn Latin:  
All your friends' names have Latin roots...



# Using Latin Roots to Define Bioinformatics



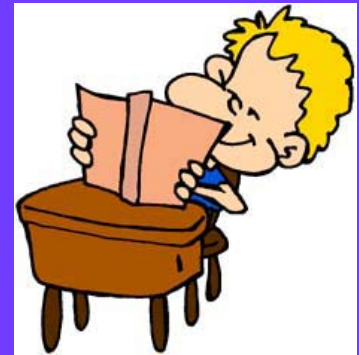
"Bio"--means "Life"

- Ex: Biology is the "study of life".

"Info"--explicitly detailed data

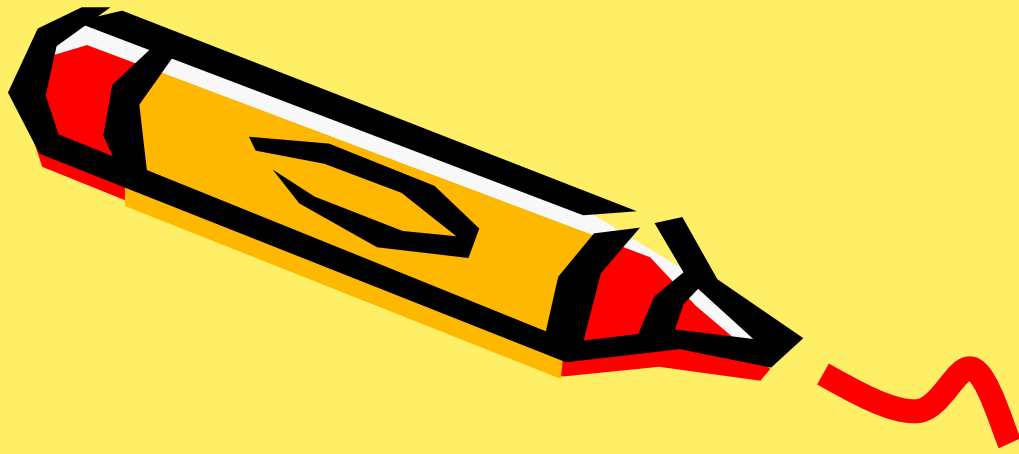
"-Matics"--refers to mechanical process or mechanism.

- Ex: Automatic--"mechanism of its own"
- Ex: Information--"data that has been mechanically processed" (in this case "mechanically" means it was worked on).



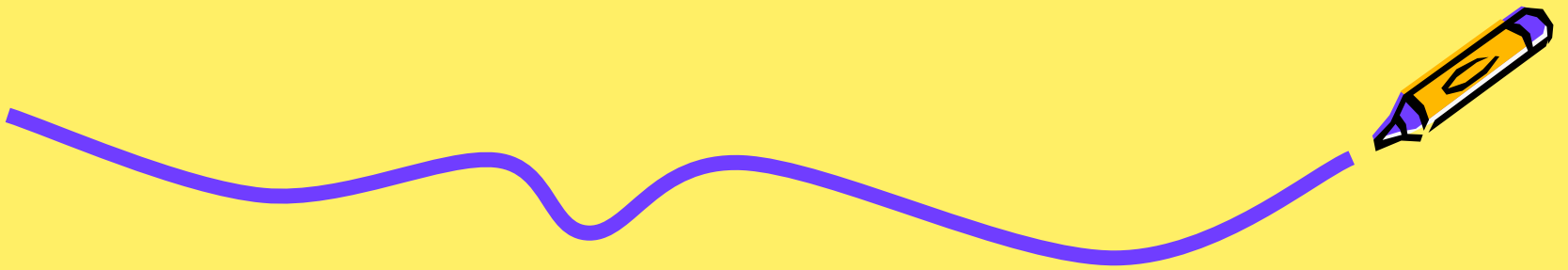
Hence the meaning of "Bioinformatics" is "computerization or mechanical processing of life data".



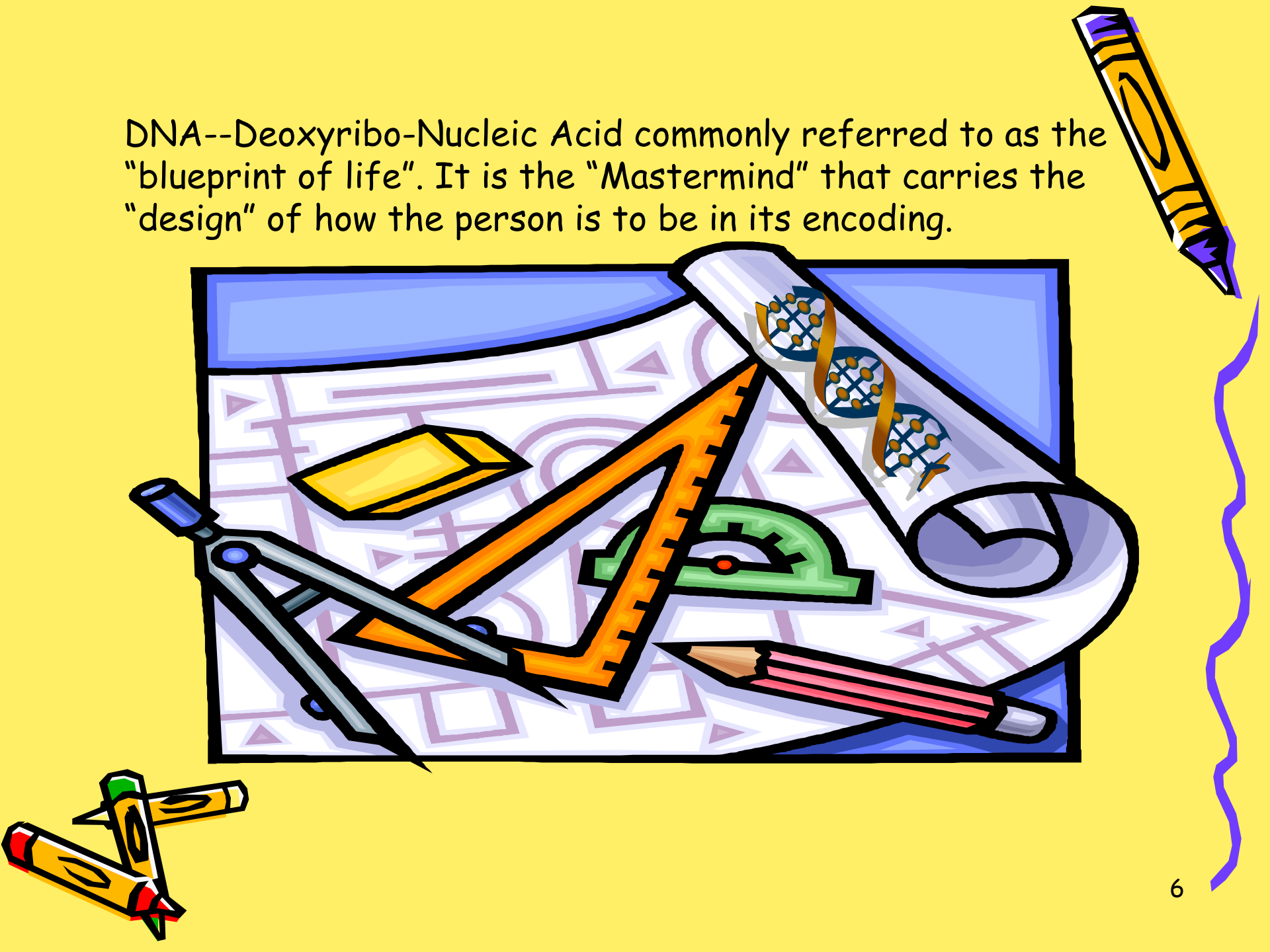


# *Vocabulary*

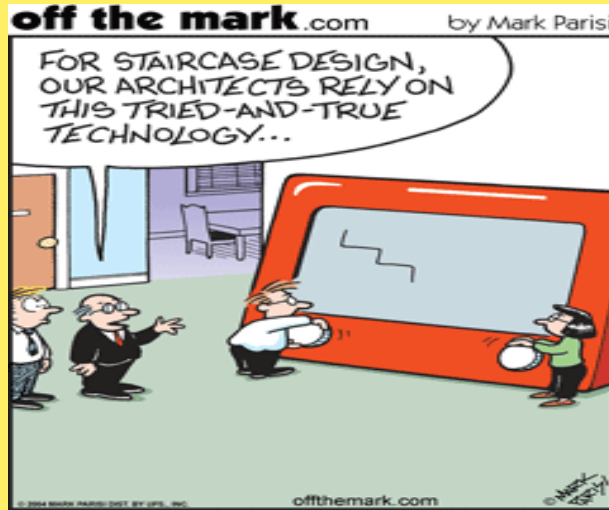
"Words to Know"



DNA--Deoxyribo-Nucleic Acid commonly referred to as the "blueprint of life". It is the "Mastermind" that carries the "design" of how the person is to be in its encoding.



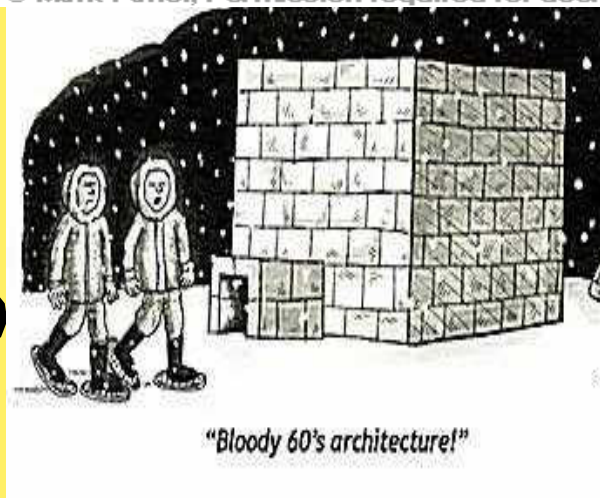
RNA--Ribo-Nucleic Acid is the "Architect" and the "messenger". It reads the "blueprint" and carries out the "written plan" and gets to work in the "construction" with the help of the ribosomes (AKA the "cement mixers").



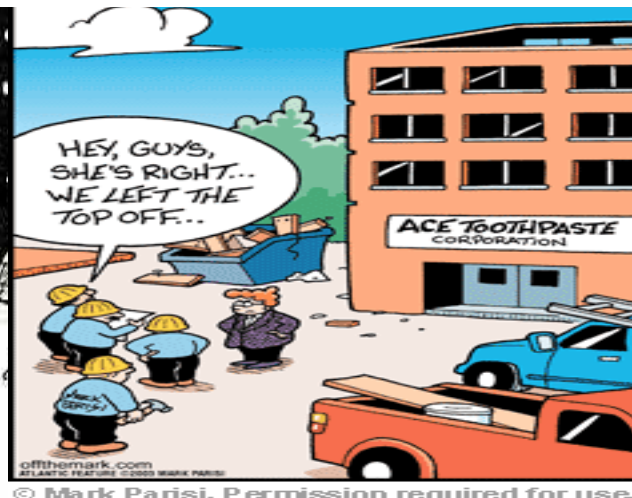
© Mark Parisi, Permission required for use.



The Leaning Tower of Pisa finally explained:  
The architect had astigmatism.



"Bloody 60's architecture!"



© Mark Parisi, Permission required for use.



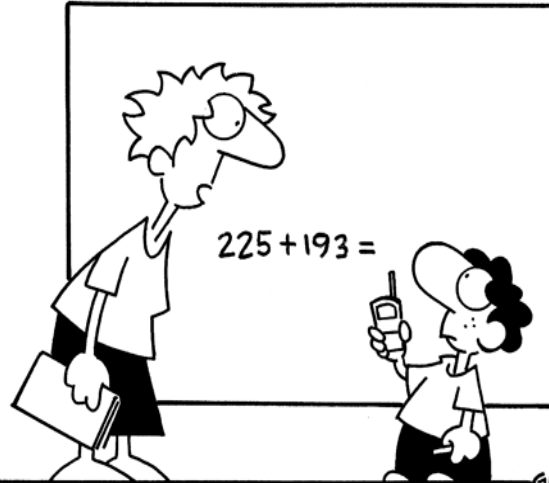
Nucleotides--the "numbers and variables" which the RNA has to "analyze" and use in order to make the "calculations and adjustments" which leads to making the "Mastermind's Design".



I'll never be able to understand math!



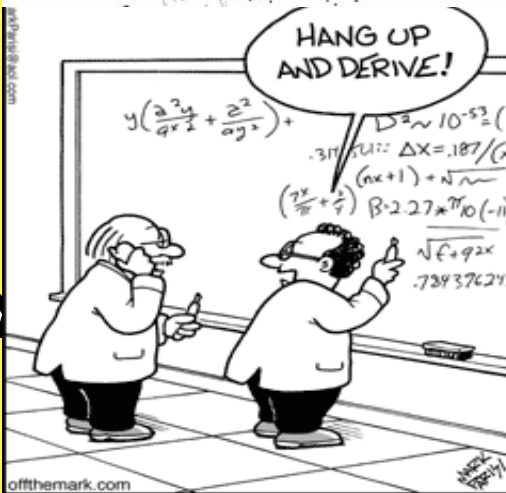
Copyright 2005 by Randy Glasbergen. www.glasbergen.com



GLASBERGEN

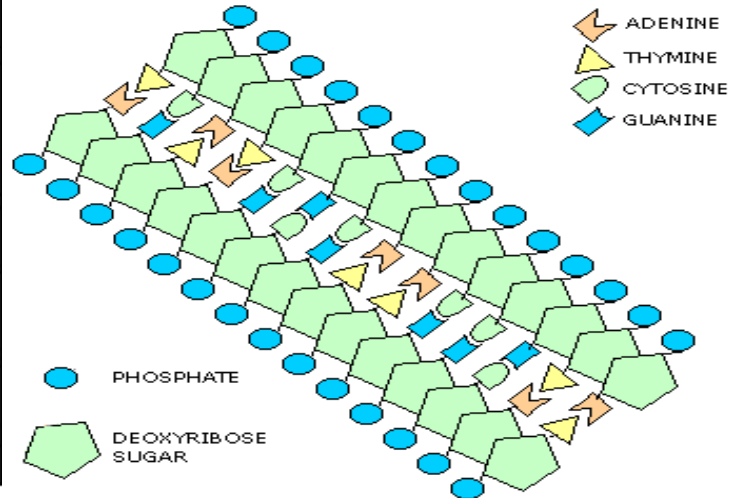
"You have to solve this problem by yourself. You can't call tech support."

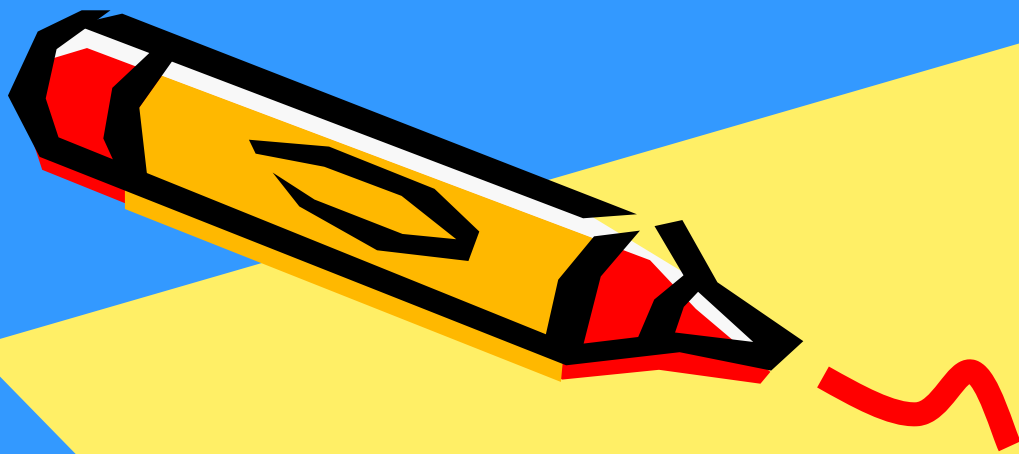
HANG UP AND DERIVE!



offthemark.com

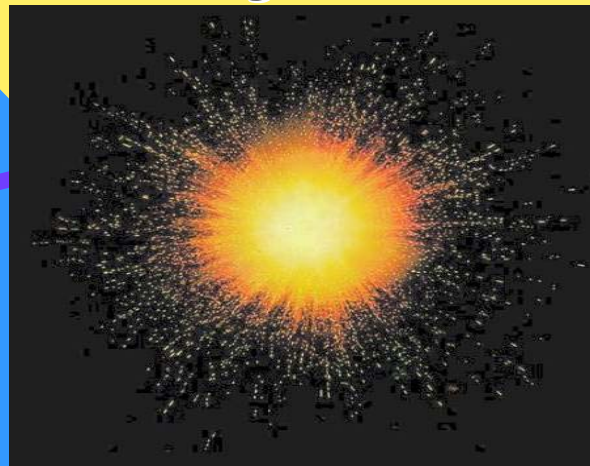
© Mark Parisi, Permission required for use.





# *First Dimension: Sequence (BLAST!)*

(Basic Local Alignment Search Tool)



# What is BLAST?

Here is a hint: BLAST is not a huge explosion. BLAST is a program used to analyze DNA, RNA, and proteins and compare similarities in nucleotides' patterns by pairing them up side by side.

It will notify you of the alignment that has been isolated for analysis and to what degree it matches by percentages and by matrices.

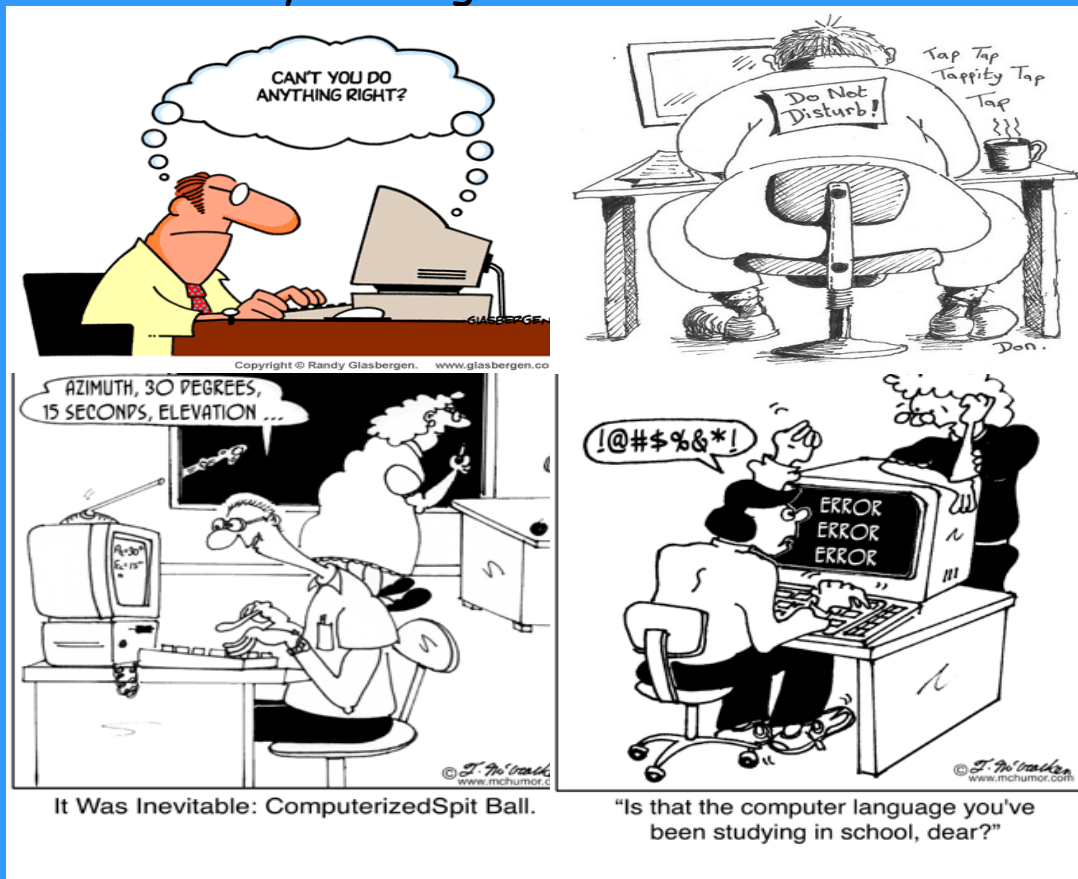
*How does it work?*





# "The Answer to all your Troubles"

Like all machines it only serves by reaction. We give it an input and it gives us an output. We just feed it data and it gives us detailed information. We give it the broken down pieces and the computer glues it together in the right spots. Still we have to keep in mind that "the creation could only be as good as its creator allows it to be".



# Examples of Input and Output



## Input:

BLAST Human Sequences.

>NM\_003234:3394-3493 Homo sapiens transferrin receptor

AGCTTTCTGTCCTATTGGCACTGAGATATTTATTGTTTATTTATCAGTGACAGAGTTCACTATAAAATGGTGTGTTATTTAATA  
GAATATAATTATCGGAAG

## Output:

### Descriptions

Score E Sequences producing significant alignments: (Bits) Value

[ref|NT\\_029928.13|](#) Homo sapiens chromosome 3 genomic contig, G... [174](#) 2e-41

[ref|NW\\_001838889.1|](#) Homo sapiens chromosome 3 genomic contig,... [174](#) 2e-41

[ref|NW\\_921873.1|](#) Homo sapiens chromosome 3 genomic contig, al... [174](#) 2e-41

Score = 174 bits (94),

Expect = 2e-41

Identities = 98/100 (98%),

Gaps = 0/100 (0%)

Strand=Plus/Minus

### Query 1

AGCTTTCTGTCCTATTGGCACTGAGATATTTATTGTTTATTTATCAGTGACAGAGTTCAC 60

|||||  
|||||-----|||||  
AGCTTTCTGTCCTTTTGGCACTGAGATATTTATTGTTTATTTATCAGTGACAGAGTTCAC 1730656 Sbjct 1730715

### Query 61

TATAAATGGTGTGTTTATTTAATAGAATATAATTATCGGAAG 100

|||||  
|||||-----||||| Sbjct 1730655  
TATAAATGGTGTGTTTATTTAATAGAATATAATTATCGGAAG 1730616





**Second Dimension:  
Secondary Structure  
(RSmatch)**

(RNA Secondary Structure Matching)



# What is RSmatch?

Simple answer is that it's a program that helps juxtapose two secondary structures of RNA. It identifies similarities as well as the differences amongst them.

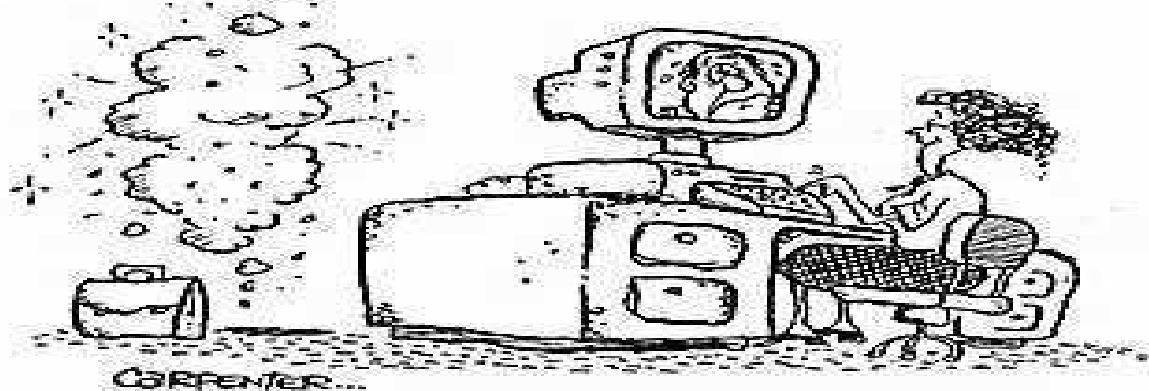


*How does it work?*

# Once Again...

It works the same way as BLAST: by input.

© Original Artist  
Reproduction rights obtainable from  
[www.CartoonStock.com](http://www.CartoonStock.com)

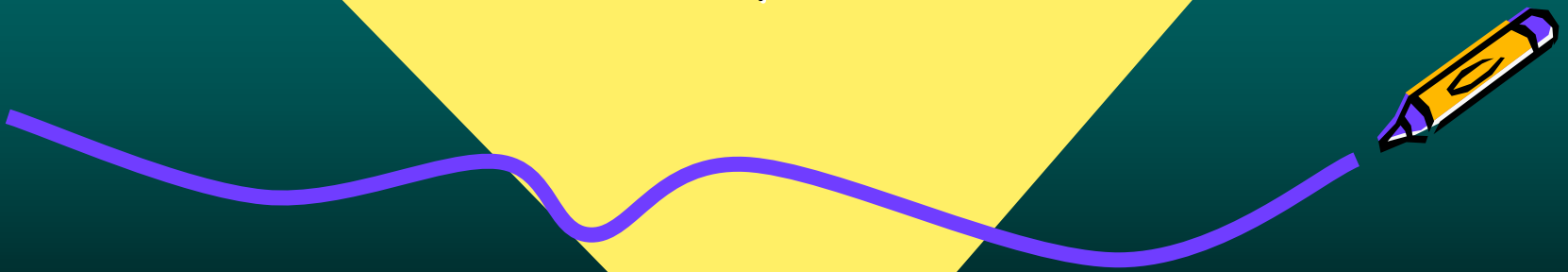


"THERE, MR. JASON. YOU ARE NOW ENTERED INTO OUR COMPUTER."



# ***RADAR***

Examples



# Examples of Input: RADAR

## Your input

Subject:

```
>NM_000032:1-52 Homo sapiens aminolevulinate, nuclear gene encoding mitochondrial protein, mRNA
CACCCUGUCAUUCGUUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG
.((((...(((((((.....))))))))))....
>NM_014585:0-100 Homo sapiens solute carrier family 40, member 1 (SLC40A1), mRNA
AGCGGGACGCCGGCGGCCUGAAGGGACGGGGCGCCCGCAGUCGGAGGUCGACGGGAGCUCGCCCGACUCGGUAUAAGAGCUGGGCCCGGCCCA
...(((.(.....))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).
>NM_002081:50-150 Homo sapiens glypican 1 (GPC1), mRNA
CGCCCGCCCGCCUUUUGUUCUCCGCCUCCUGGCCCGCCCGCCCGCUCUGGACCGCGAGCCGCGCGCCGGGACCUUGGCUCUGCCCUUCGCGGGCGG
.((((.....)))).....(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).
>NM_003449:1000-1100 Homo sapiens tripartite motif-containing 26 (TRIM26), mRNA
GGCAGUACAUUGUGGCUGAGUUUGAGCAGGGUACUAGUUCUGAGGGAGCGGGAGAACACCCUGCUGGAACAGCUGGCAGAGCAGGAGCUCAC
..((((.....))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).
>NM_018992:450-550 Homo sapiens potassium channel tetramerisation domain containing 5 (KCTD5), mRNA
CAUUAUAAAAACUUGUAAAGGACAAAUAAGAGAACGAGACAGCAAAAACUCCAGGUGCCUGUGAAGCAUGUGUACCGUGUGCAGUGCCAGGAGGA
((((.....))))).((((.....))))).((((.....))))).((((.....))))).((((.....))))).((((.....))))).
>NM_014585:151-250 Homo sapiens solute carrier family 40, member 1 (SLC40A1), mRNA
UCGGACGACCUGCUGAGCCUCCAAACCGCUUCCAUUAGGCUUUGCCUUUCCAAUUCAGCUACAGUGUAGCUAAGUUUGGAAAGGAAAAAGAAA
.....(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).
>NM_000146:1-100 Homo sapiens ferritin, light polypeptide (FTL), mRNA
GUCCCGCGGGUCUGUCUUCUUCUUAACAGUGUUUGGACGGAACAGAUCCGGGGACUCUCUCCAGCCUCCGACCGCCUCCGAAUUUCCUCUCCGCUUGC
(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).(((((((.....)))))).
```

Query:

```
>NM_000032:1-52 Homo sapiens aminolevulinate, delta-, synthase 2 (sideroblastic/hypochromic anemia) (ALAS2), nuclear gene encoding mitochondrial protein, mRNA
CACCCUGUCAUUCGUUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG
.((((...(((((((.....))))))))))....
```

# Examples of Output: RADAR

#=== Hits ===#

Rank	Score	Query-offset	DB Str/seq	Offset	Annotation
1	68	1-52	NM_000032:1-52	1-52	Homo sapiens aminolevulinate, nuclear gene encodin
2	21	10-39	NM_014585:151-250	52-81	Homo sapiens solute carrier family 40, member 1 (S
3	16	14-34	NM_000146:1-100	20-40	Homo sapiens ferritin, light polypeptide (FTL), mR
4	8	15-33	NM_018992:450-550	51-66	Homo sapiens potassium channel tetramerisation dom
5	7	13-35	NM_014585:0-100	14-40	Homo sapiens solute carrier family 40, member 1 (S
5	7	17-32	NM_003449:1000-1100	38-51	Homo sapiens tripartite motif-containing 26 (TRIM2
7	6	17-32	NM_002081:50-150	85-99	Homo sapiens glypican 1 (GPC1), mRNA

```

=====
Rank: 1  Score: 68  p-value: 2.8E-02  Query: 52 (ss:20, ds:32)
Identity: str: 100%; seq:100% (ss:100%, ds:100%)
Gap: 0 (ss:0, ds:0)  Mismatch: 0 (ss:0, ds:0)
      .((((((...(((((((((.....)))))))))).)))...))))).
      .((((((...(((((((((.....)))))))))).)))...))))).
NM_000032:1-52:  1 CACCUGUCAUUCGUUCGUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG 52
      |||
NM_000032:1-52:  1 CACCUGUCAUUCGUUCGUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCAAG 52
=====

```

```

=====
Rank: 2  Score: 21  p-value: 21E-02  Query: 30 (ss:8, ds:22)
Identity: str: 100%; seq:40% (ss:75%, ds:27%)
Gap: 0 (ss:0, ds:0)  Mismatch: 18 (ss:2, ds:16)
      ((((((...(((((((((.....)))))))))).)))
      ((((((...(((((((((.....)))))))))).)))
NM_000032:1-52:  10 UUCGUUCGUCCUCAGUGCAGGGCAACAGGA 39
      ::::|||::|::| |||| ::|::||: ::
NM_014585:151-250: 52 CAACUUCAGCUACAGUGUUAGCUAAGUUUG 81
=====

```

The score matrices used are :

	A	C	G	U
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
U	-1	-1	-1	1

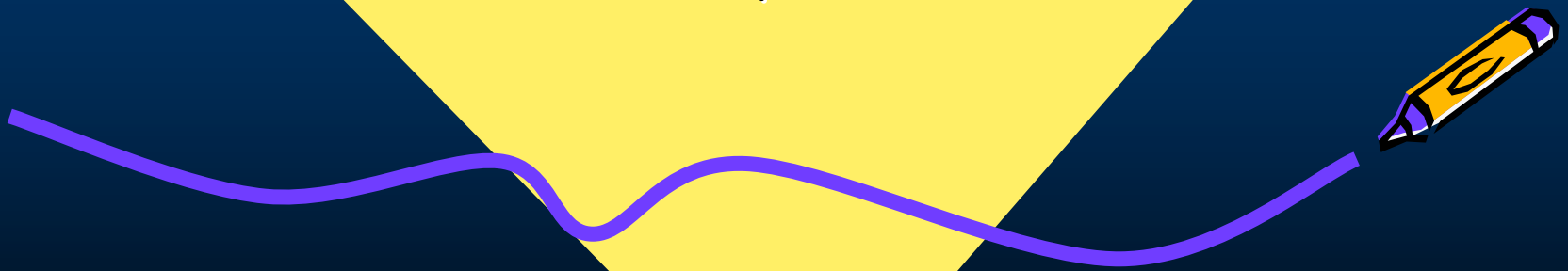
	AU	CG	GC	GU	UA	UG
AU	3	1	1	1	1	1
CG	1	3	1	1	1	1
GC	1	1	3	1	1	1
GU	1	1	1	3	1	1
UA	1	1	1	1	3	1
UG	1	1	1	1	1	3

Gap penalty is set as: -2.0



# *RmotifDB*

Examples



# Examples of Input: RmotifDB

## Input:

```
#=== Query ===#  
  
>NM_000032:1-52 Homo sapiens aminolevulinate, delta-, synthase 2 (sideroblastic/hypochromic anemia) (ALAS2), nuclear gene encoding mitochondrial protein, mRNA  
  
..((((((...(((((((((.....)))))))))).)))...)))))...  
1 CACCUGUCAUUTGUUCGUCCUCAGUGCAGGGCAACAGGACUUUAGGUUCA 50  
  
..  
51 AG 52
```

Compared with:  
18,233 RNA secondary structures  
taken from the 603 [Rfam](#) seed alignments (version 9.0)



# Examples of Output: RmotifDB

## Output:

```
#=== Hits ===#  
Rank Score Query-offset DB Str/seq Offset Annotation  
-----  
1 32 13-36 AF086786.1/2-28 3-26 %RF00037.0%  
1 32 13-36 AF171078.1/1416-1442 3-26 %RF00037.31%  
3 22 16-32 S57280.1/391-417 6-22 %RF00037.36%  
4 21 10-39 A&GW01600212.1/195-326 83-110 %RF00265.7%  
5 20 2-47 AACZ02117923.1/3779-39 5-55 %RF00413.17%  
6 19 11-38 X01060.1/3482-3508 1-27 %RF00037.6%  
6 19 11-38 AJ426432.1/1658-1684 1-27 %RF00037.11%  
6 19 11-38 X13753.1/1434-1460 1-27 %RF00037.15%  
9 18 2-47 Z12832.1/542-622 10-53 %RF00170.1%  
9 18 9-39 AF083002.1/1-548 382-411 %RF00177.217%  
9 18 10-42 BA000012.4/1306844-130 341-369 %RF00010.256%  
9 18 10-39 AF285177.1/3-32 1-30 %RF00037.1%  
9 18 13-36 AY112742.1/12-41 5-28 %RF00037.2%  
14 17 13-35 X01060.1/3950-3976 3-25 %RF00037.16%
```

```
=====  
Rank: 1 Score: 32 p-value: 2.1E-02 Query: 24 (ss:8, ds:16)  
Identity: str: 100%: seq:100% (ss:100%, ds:100%)  
Gap: 0 (ss:0, ds:0) Mismatch: 0 (ss:0, ds:0)  
(((.((((.(.....)))))))).  
(((.((((.(.....)))))))).  
NM_000032:1-52: 13 GUUCGUCCUCAGUGCAGGGCAACA 36  
|||||  
AF086786.1/2-28: 3 GUUCGUCCUCAGUGCAGGGCAACA 26
```

# References

- [Dongrong Wen and Jason T. L. Wang](#), "Design of an RNA Structural Motif Database," *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 1:32-41, 2009.
- [Mugdha Khaladkar, Vivian Bellofatto, Jason T. L. Wang, Bin Tian and Bruce A. Shapiro](#), "RADAR: A Web Server for RNA Data Analysis and Research," *Nucleic Acids Research*, 35:W300-W304, 2007.
- [Jianghui Liu, Jason T. L. Wang, Jun Hu and Bin Tian](#), "A Method for Aligning RNA Secondary Structures and Its Application to RNA Motif Detection," *BMC Bioinformatics*, 6:89, 2005.



***The End***

---