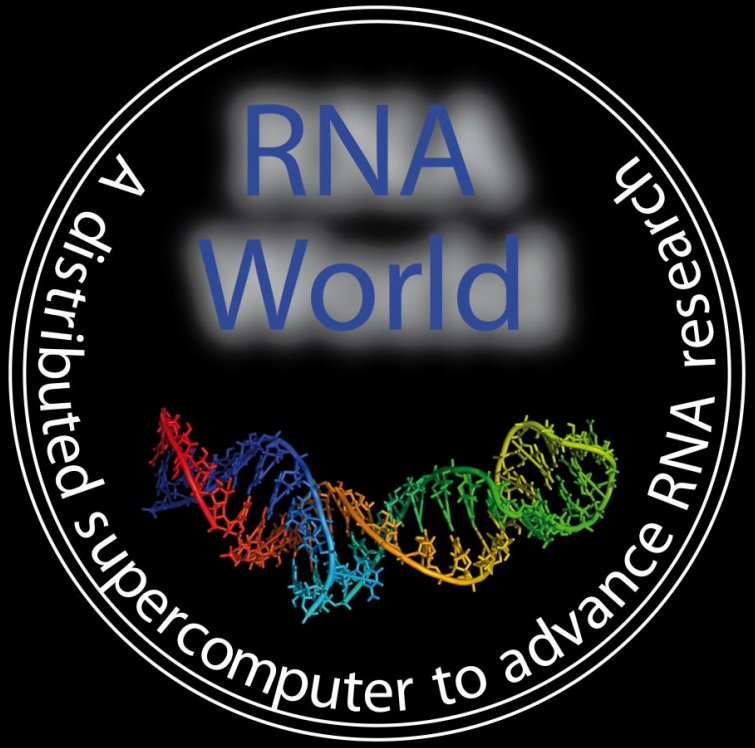
RNA World – A BOINC-based Distributed Supercomputer for High-Throughput Bioinformatic Studies to Advance RNA Research



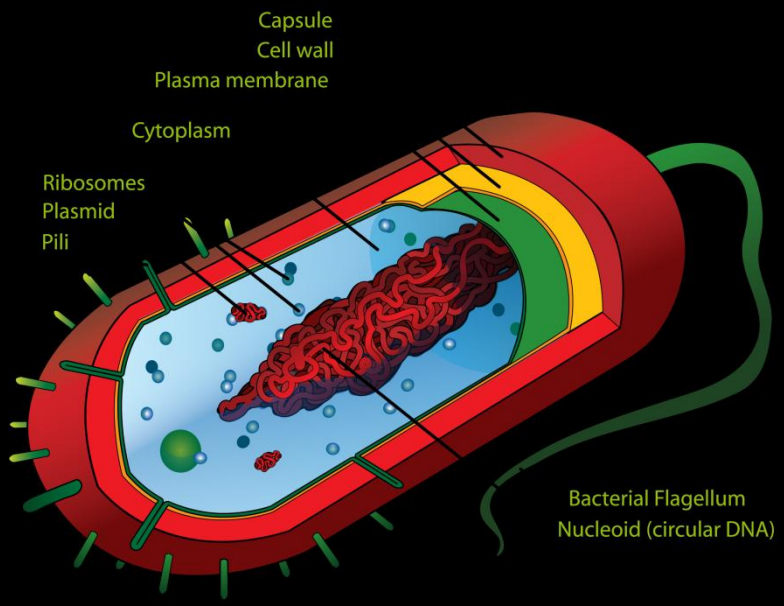
Michael H.W. Weber

5th Pan-Galactic BOINC Workshop
Barcelona 2009

Philipps Universität Marburg

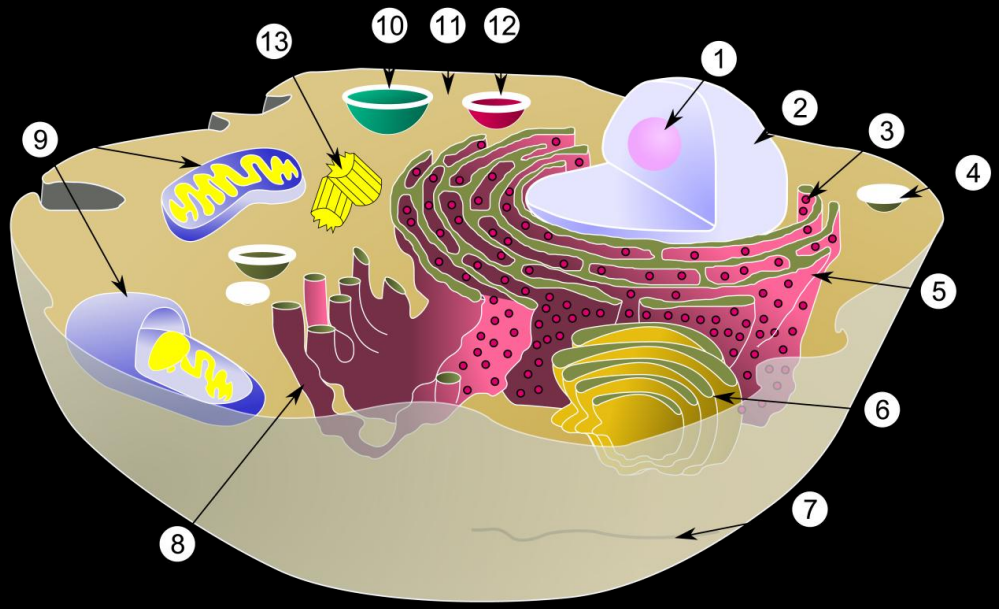
Rechenkraft.net e.V.
Verein zur Förderung von Bildung, Forschung und Wissenschaft durch Einsatz vernetzter Computer

General Cell Architectures



Prokaryote

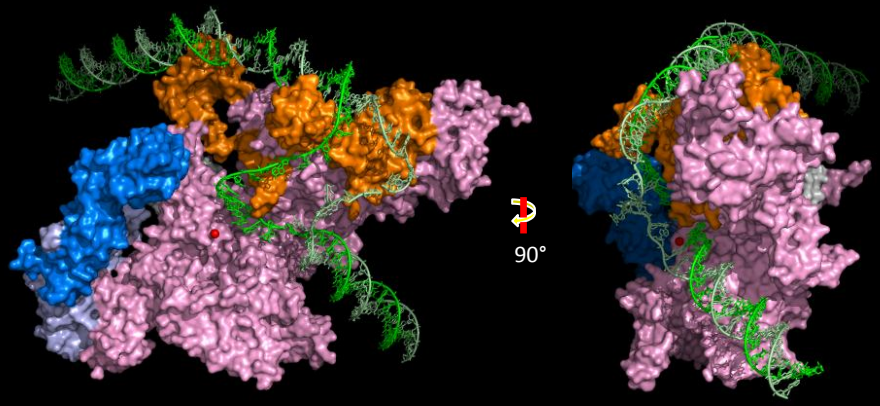
Eukaryote



(1) nucleolus, (2) nucleus, (3) ribosome, (4) vesicle, (5) rough endoplasmic reticulum (ER), (6) Golgi apparatus, (7) Cytoskeleton, (8) smooth endoplasmic reticulum, (9) mitochondria, (10) vacuole, (11) cytoplasm, (12) lysosome, (13) centrioles within centrosome

The Cellular Flow of Genetic Information

DNA

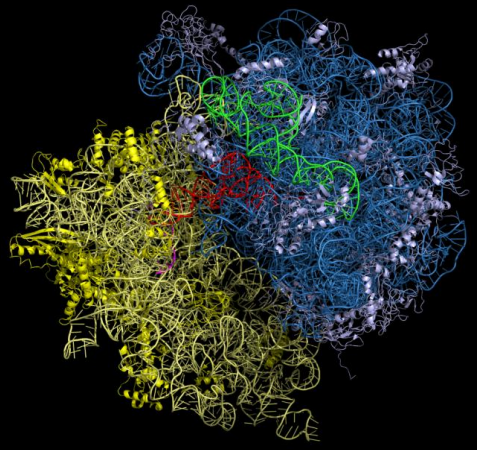


RNA polymerase

Transcription

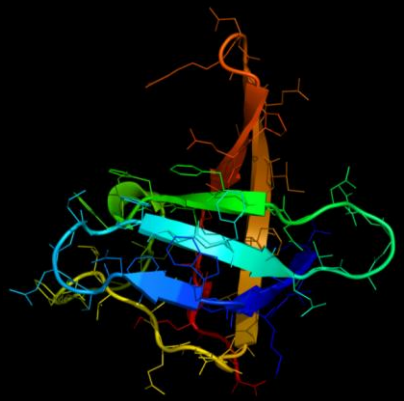
5´ A AGGAGG AUG UAA GGGAUACCCUU 3´

RNA



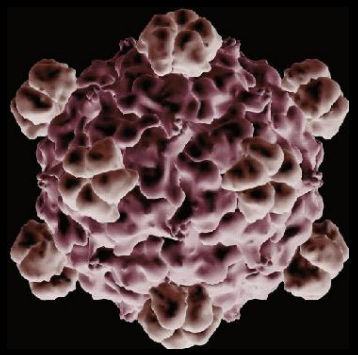
Ribosome

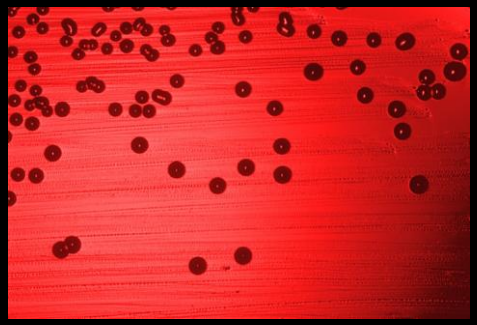
Translation

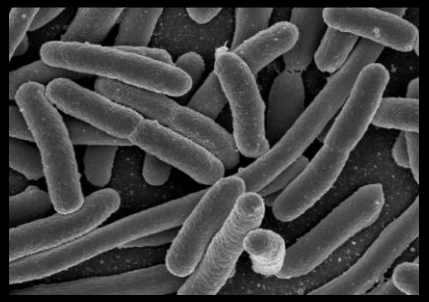


Protein

Genome Architectures: Information Content

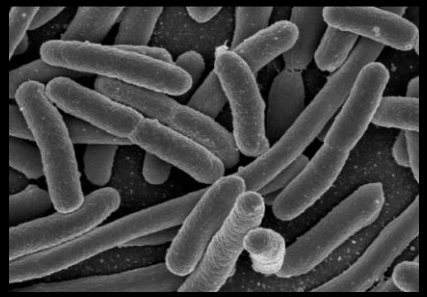


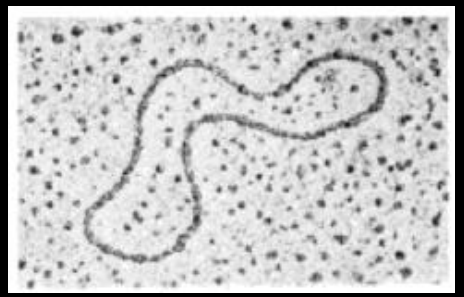


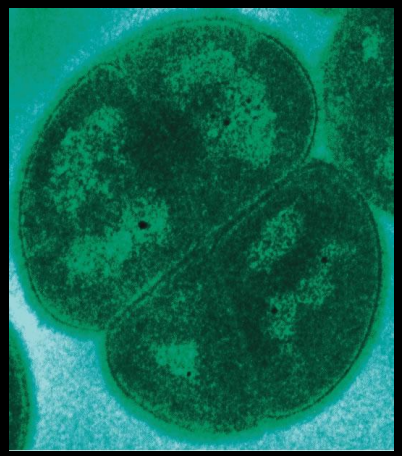


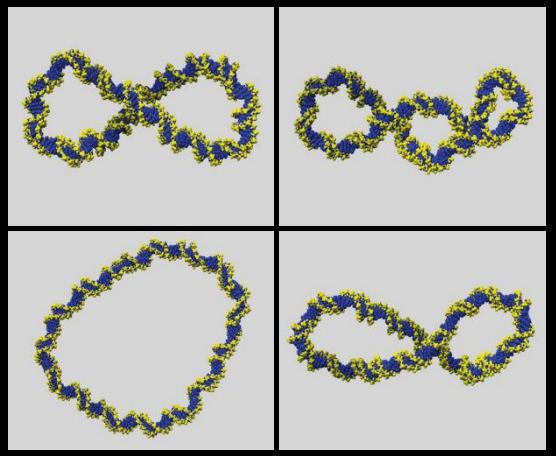
Organism	Genome size (bp)	Year	Remarks
Phage Φ-X174	5,386	1977	first DNA genome ever sequenced
Haemophilus influenzae	1,830,000	1995	first genome of living organism
Escherichia coli	4,600,000	1997	bacterial model organism #1
Caenorhabditis elegans	100,300,000	1998	first multicellular animal genome
Arabidopsis thaliana	157,000,000	2000	first plant genome sequenced
Homo sapiens	3,200,000,000	2001	first draft sequence
Polychaos dubium	670,000,000,000	2008	largest known genome

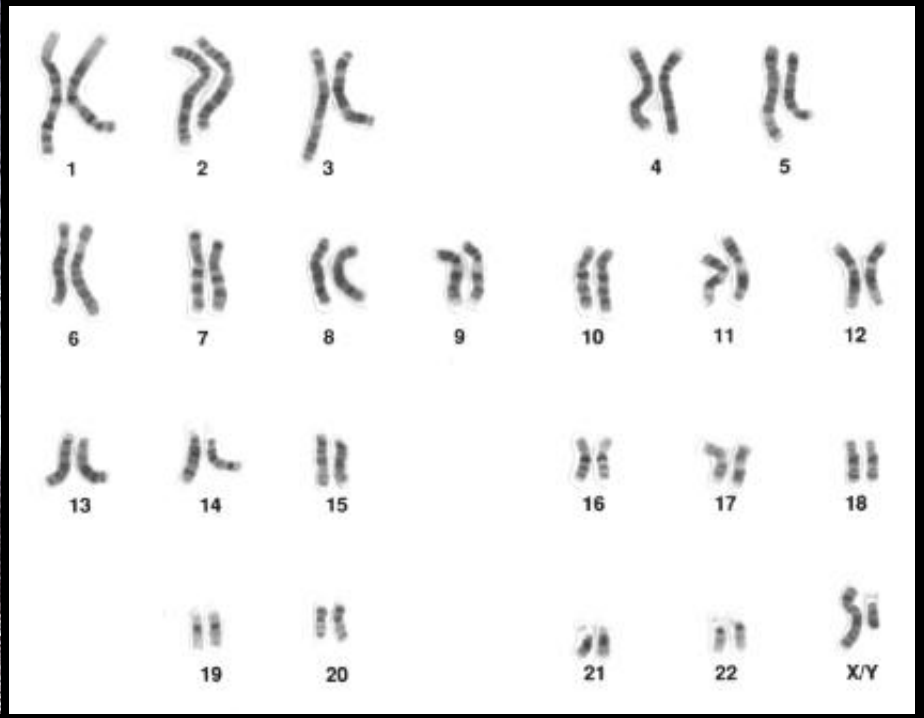
Genome Architectures: Information Distribution



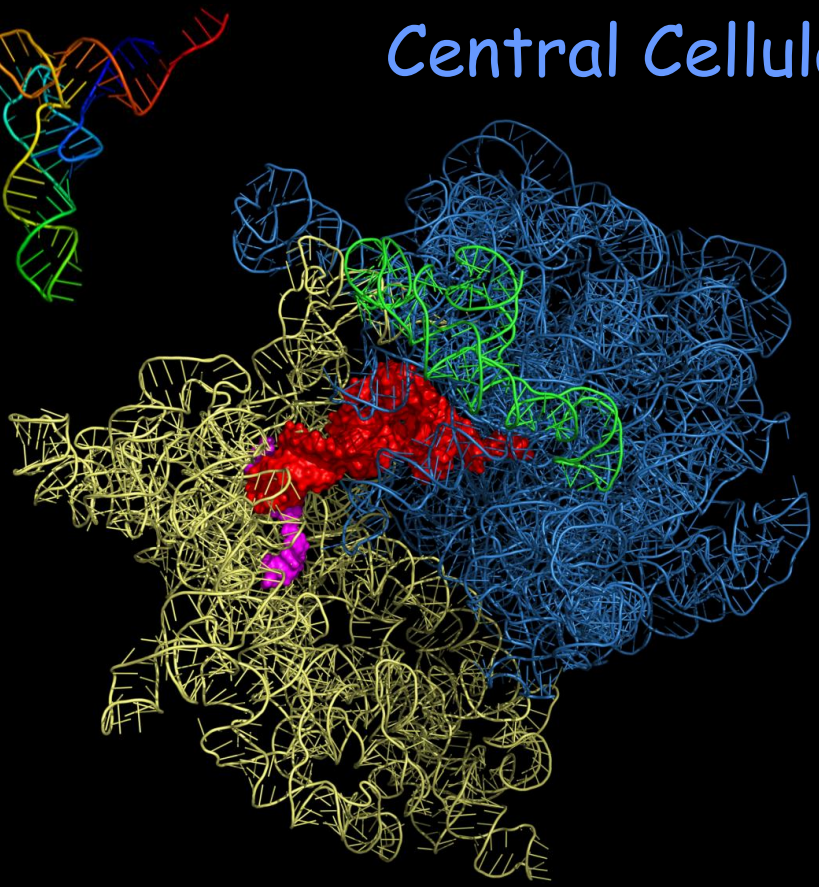




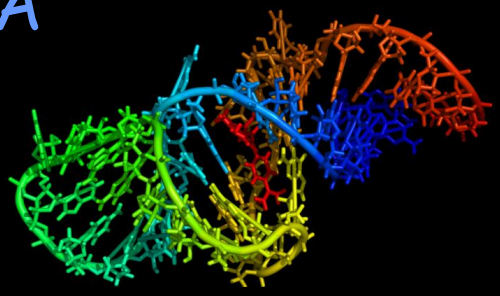




Central Cellular Roles of RNA

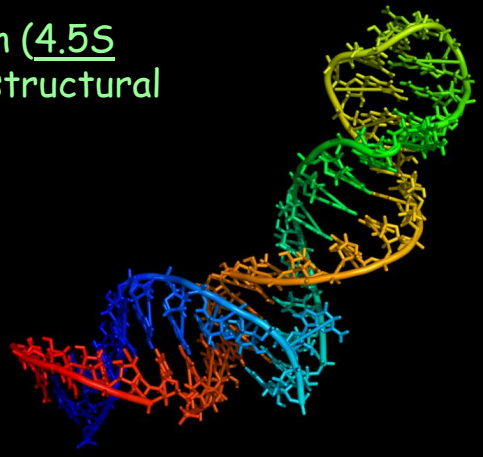


No protein coding without mRNAs, no eukaryotic mRNAs without the spliceosome

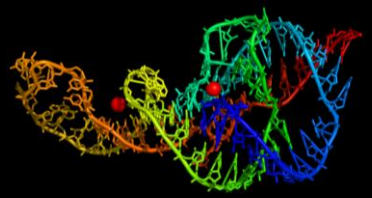


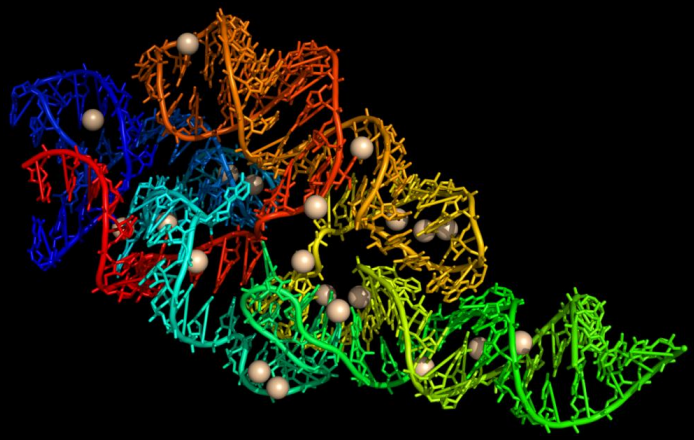
No metabolite detection without RNA aptamers

No protein secretion (4.5S RNA/SRP) without structural RNAs



No tRNA processing (RNase P) and protein synthesis (ribosome) without ribozymes





sRNA regulators: 6S RNA (binds RNA polymerase), miRNAs (regulate cell differentiation, cancer-involved)

Project Motivation: Making RNA Bioinformatic Tools Broadly Available to Non-IT-Specialized Scientists

1) Most RNA-related bioinformatic tools are available only for Linux but many scientists, especially in life-science research, are often not yet familiar with this smart OS

2) Many tools are computationally very expensive or difficult to handle in practice (command-line-based) and for many scientific aspects only few web servers are available

We like to not only follow up
our own scientific projects
but also allow others to use
our distributed system by
implementing appropriate
job submission forms ☺

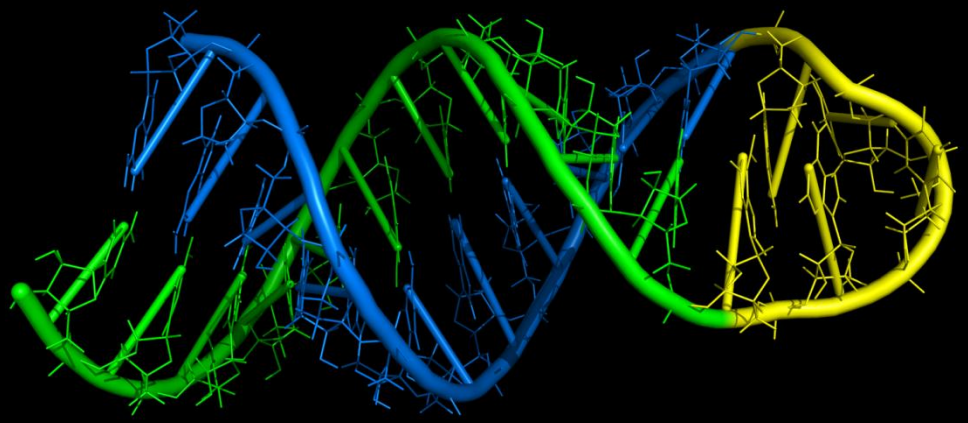
Our Initial Focus:
The Problem of Identifying RNA Homologs

Primary structure comparison: virtually no similarity

```
PDB 1YSV:    GGUAACAAUAU-GCUAA-AUGUUGUUACC
unknown:     GGGGCCCGGGG-AUACC-CCCCGGGCCCC
consensus:   GG---C-----  -----  -----G---CC
```

Secondary structure comparison: identical hairpin fold

```
              G-C
GGUAACAAUAU     \
||||||||||       U
CCAUUGUUGUA     /
              A-A


              A-U
GGGGCCCGGGG     \
||||||||||       A
CCCCGGGCCCC     /
              C-C
```



Tertiary structure: PDB 1YSV: similar

A Solution: INFERNAL 1.0*

1) INFERNAL supports searching genomes for non-coding RNAs using a combination of primary and secondary structure information (SCFG/HMM-based)

2) Due to its extreme compute requirements, for serious bioinformatic analyses, INFERNAL is currently executed on high-performance computing clusters, only (CMCALIBRATE run times on a 2.4 GHz Intel Centrino P8600 CPU vary between 14 min to 72 hrs with seed alignments taken from Rfam 9.1)

*Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**: 1335-7.

Achievements: Server Setup, Client Implementation, Alpha Testing, Screensaver Creation

RNA World (alpha) - Mozilla Firefox

http://www.rnaworld.de/rnaworld/

RNA World (alpha)

Please visit donation page to help the project cover running costs.

About RNA World (alpha)

RNA World (alpha) is a distributed supercomputer that uses Internet-connected computers to advance RNA-related research. You can participate by downloading and running a free program on your computer.

RNA World (alpha) is based at the Rechenkraft.net e.V. research facility located in Germany.

- Project description

Benutzer des Tages

domini1000

Ich engagiere m

Join RNA World (alph

- Unsere Regeln und Gru
- This project uses BOI
  Attach to Project. If
- When prompted, ente
  http://www.rnawor
- If you're running a co
  create an account firs
- If you have any probl

Returning participant

Fertig

Die besten Teams - Mozilla Firefox

http://www.rnaworld.de/rnaworld//top_teams.php

Diskussionsforen | Chat | Project Description | Statistics | Rechenkraft.net

RNA World (alpha)

Please visit donation page to help the project cover running costs.

Die besten Teams

Rang	Name	Mitglieder	Durchschnittliches Guthaben	Gesamtguthaben	Land	Type
1	Rechenkraft.net	27	426	701,632	Germany	None
2	Nordlichter	3	305	187,098	Germany	None
3	SETI.Germany	4	45	32,775	Germany	Local/regional
4	Erster Kontakt	1	20	18,309	Germany	None
5	BOINC Confederation	1	1	1,268	International	None

Server status page - Mozilla Firefox

http://www.rnaworld.de/rnaworld/server_status.php

Server status

Program	Host	Status
data-driven web pages	r094	Running
upload/download server	r094	Running
scheduler	r094	Running
feeder	r094	Running
transitioner	r094	Running
file_deleter	r094	Running
sample_work_generator	r094	Disabled
sample_bitwise_validator	r094	Running
sample_assimilator	r094	Running
cmsearch_validator	r094	Running
cmsearch_assimilator	r094	Running
cmcalibrate_validator	r094	Running
cmcalibrate_assimilator	r094	Running
db_purge	r094	Running

Running: Program is operating normally

Not Running: Program failed or ran out of work (or the project is down)

Disabled: Program has been disabled by staff (for debugging/maintenance)

Fertig

Workunits waiting for deletion	0
Results waiting for deletion	0
Transitioner backlog (hours)	0

Results by

Application	unsent	in progress	avg runt
cmsearch	0	0	0.06 (0.01-0.36
cmcalibrate	4	3	23.93 (0.23-13

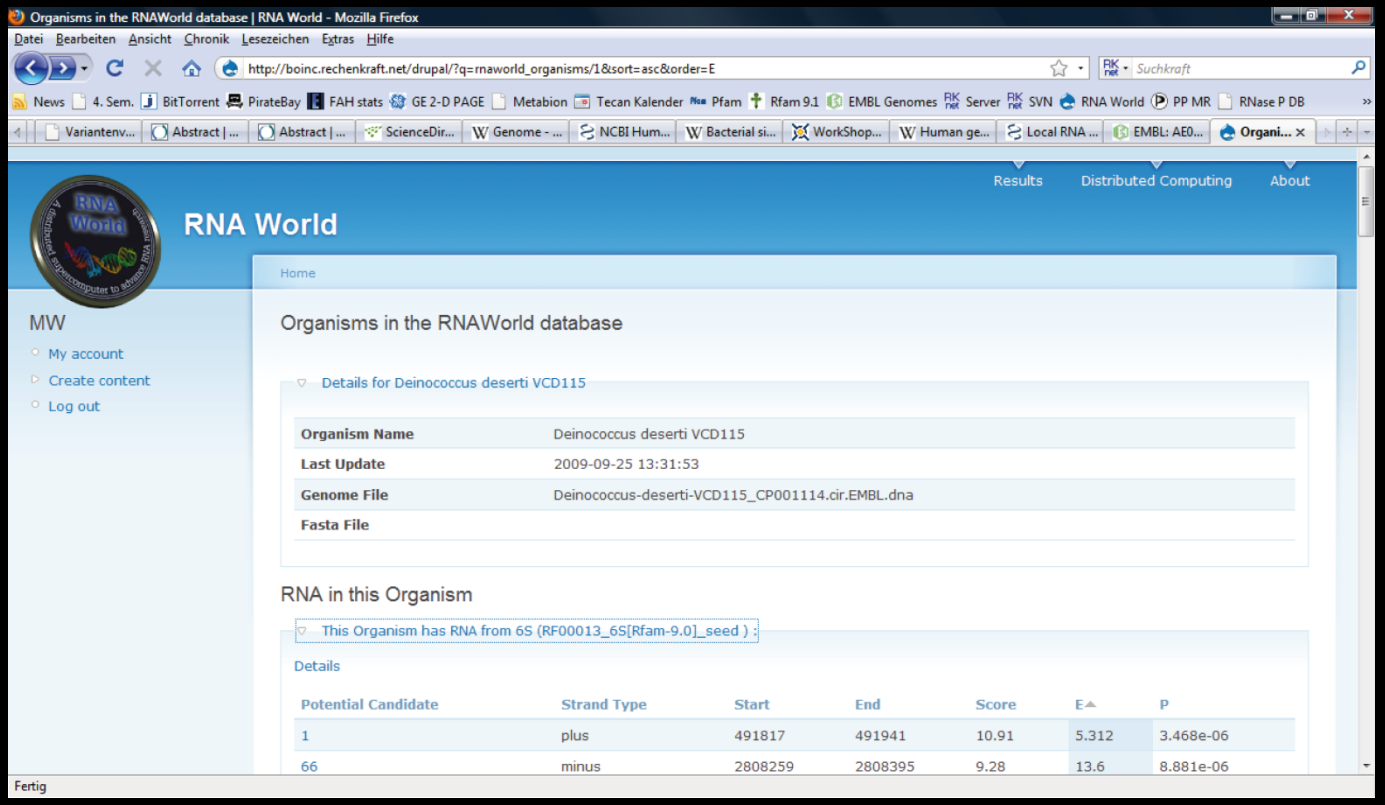
workunits todo	
cmc_RNaseP[Rfam-1.0-9.1]	1

INFERNAL Output Post-Processing: InReAlyzer*

```
CM: 6S RNA
>gi|50812173|ref|NC_000964.2| B. subtilis

  Plus strand results:

 Query = 62 - 130, Target = 835746 - 835799
 Score = 16.93, E = 0.1324, P = 5.802e-08, GC =  56

          <-<<<<<----<<<<<<<------<<---<<<<<_____>>>>>>-->>----->>>>>>>
       62 GagcccucucUuuucagcgGuGuGcAuGCCcgcCUuGuAgcgGGAAgCcuaAAgcugaaa 121
          GAG CC  UCU ::         GC +GCC:G:CUUG  :C:GGAAGC U+A     ::
   835746 GAGUCCAUUCUAAA---------GCUGGCCGGUCUUGA-ACCGGAAGCGUUA-----UUG 835790

          -->>>>>>->
      122 auagggcaC 130
          A+ GG CAC
   835791 ACCGGGCAC 835799


  Minus strand results:

 Query = 1 - 188, Target = 2813908 - 2813716
 Score = 107.57, E = 1.339e-25, P = 5.869e-32, GC =  42

          :<<<<<<<<<<<<<<-<<<-------------<<<<-<<<<<<----------------.
        1 aaagccCUgcggUGUUCGucAguugcuuauaaguccCuGAgCCgAuaauuUuuauaaau. 59
          AAAG:CCU:::GUGUU GU      C+UA   GU:: UGA CCGA+ AUUUUU+U A+U
  2813908 AAAGUCCUGAUGUGUUAGUUGUACACCUA---GUUU-UGA-CCGAACAUUUUUUUGAUUu 2813854

          <<<-<<<<<----<<<<<<<------<<---<<<<<.....______.._>>>>>-->>---
       60 GGGagcccucucUuuucagcgGuGuGcAuGCCcgc.....CUuGu..AgcgGGAAgCcua 112
          GGGAGCCC:C +UUUU:A::GG+GU: AUGCC:::      U+G   A:::GGA  :  A
  2813853 GGGAGCCCGCAUUUUUAAAUGGCGUACAUGCCUCUuuucaUUCGGuaAAGAGGACUUACA 2813794

          -->>>>>>>>-->>>>>->>>------.------->>>>>>->>>>>-..------------
      113 AAgcugaaaauagggcaCCCACCUgg.aAcagcaGGuUCaAggacu..uaaugacgucaA 169
          A ::U:AAAA :GGGCACCCACCUG+  A AGC+GGUUCA ::AC     A++ C  CA
  2813793 AGAUUUAAAAGAGGGCACCCACCUGCuGAGAGCGGGUUCA-AAACAaaGGAAAGCUGCA- 2813736

          >>>>>>>.>>>>>>>>>>::
      170 aCGGCAc.ugcGGggcuuuu 188
          AC GCAC :::GGG:CUUU+
  2813735 ACGGCACuAUUGGGACUUUA 2813716
```

*Hatzenberger V, Hartmann RK, Weber MHW (2009) InReAlyzer: A fully automated graphical visualization pipeline for the convenient output file interpretation of INFERNAL-based RNA covariance analyses. *In preparation.*

Automated Results Archiving in a Publically Accessible Drupal/MySQL-based Web Database, OpenMPI Implementation, Construction of User Job Submission Forms



OpenMPI: searching DsrA in *M. tuberculosis* on a Quad-Opteron/2.6 GHz/Linux-32:

```
# of cores: 1, total actual time for CMCALIBRATE: 02:18:27, CMSEARCH: 00:28:08
# of cores: 2, total actual time for CMCALIBRATE: 01:33:18, CMSEARCH: 00:28:08
# of cores: 3, total actual time for CMCALIBRATE: 00:39:50, CMSEARCH: 00:14:05
# of cores: 4, total actual time for CMCALIBRATE: 00:26:45, CMSEARCH: 00:09:41
```

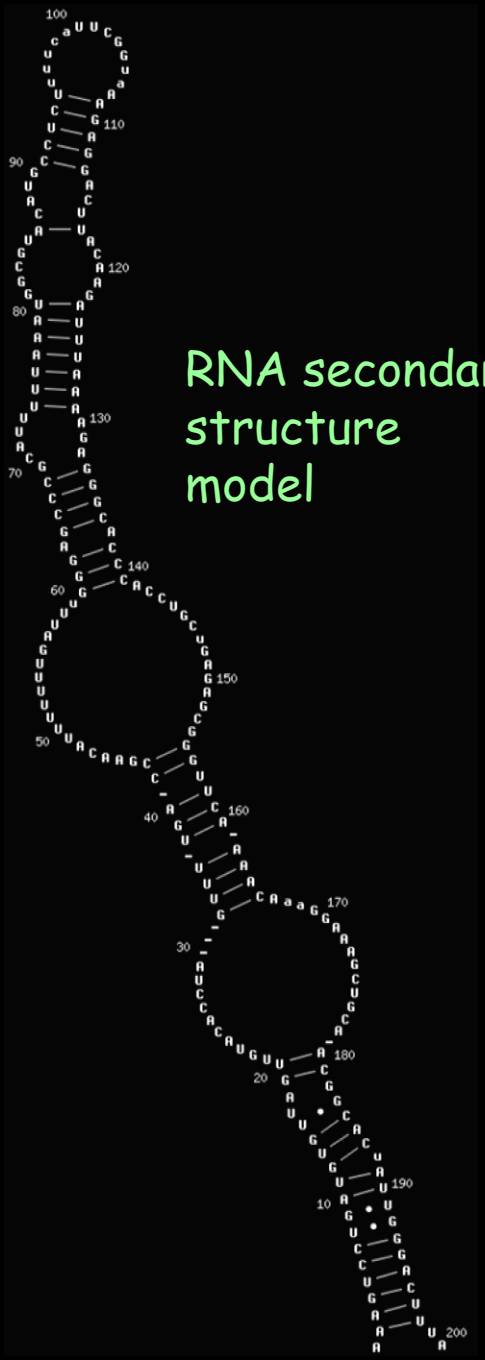
Problems & Useful Improvements

1) Initial (funny) validation issues: rounding is different in Linux & Windows: ASCII files containing floating point numbers cannot be validated when the WU is computed once under Linux and the other time under Windows

2) RNA World checkpointing currently works exclusively for Linux-32 machines and requires manual adjustments from a superuser: if BOINC could in the future run as a virtual machine, universal checkpointing would be possible where the science application has to take no measures to achieve this (most existing science applications cannot support checkpointing without entire re-coding, including INFERNAL) ☺

3) RNA World screensaver is currently implemented as a series of randomly selected flash movies: a universal (cross-OS) movie template/player would be very helpful to avoid diving deeper into graphics programming ☺

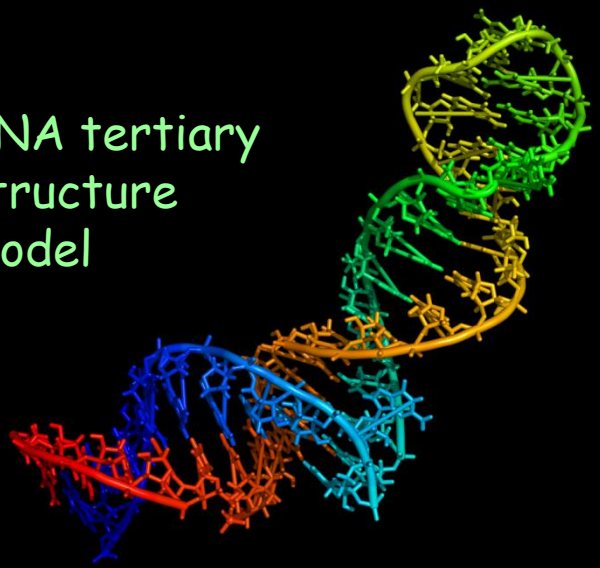
Future Perspectives



RNA secondary structure model

fully automated

RNA tertiary structure model



Project Team & Acknowledgements

RNA World project personnel

Server administrator: Uwe Beckert
Software development: Martin Bertheau
Volker Hatzenberger
Nico Mittenzwey
Graphics & design: Lasse J. Kolb
Rebirther
Michael H.W. Weber

Project leader & contact: Michael H.W. Weber
mw@rnaworld.de

Rechenkraft.net e.V.
Verein zur Förderung von Bildung, Forschung und Wissenschaft durch Einsatz vernetzter Computer

Philipps Universität Marburg

RNA World project cooperation partner laboratories

Germany: Roland K. Hartmann (Philipps-Universität Marburg)
India: Srinath Thiruneelakantan (Indian Institute of Science, Bangalore)

National Aeronautics and Space Administration

RCSB PDB PROTEIN DATA BANK

EOL Encyclopedia of Life

WIKIPEDIA The Free Encyclopedia