Applying Sequence Alignment Algorithms to Data Compression

Ali Yahya

Scientists create a living organism

By **Clive Cookson**, FT
May 20, 2010 10:47 p.m. EDT

FINANCIAL TIMES

Genomics pioneer Craig Venter, pictured in 2008, said, "We have passed through a critical psychological barrier."

(FT) -- Scientists have turned inanimate chemicals into a living organism in an experiment that raises profound questions about the essence of life.

Craig Venter, the U.S. genomics pioneer, announced on Thursday that scientists at his laboratories in Maryland and California had succeeded in their 15-year project to make the world's first "synthetic cells" -- bacteria called *Mycoplasma mycoides*.

STORY HIGHLIGHTS

- Synthetic research behaves and divides in lab dishes like natural bacteria
- Independent scientists and philosophers hail research as a landmark

"We have passed through a critical psychological barrier," Dr. Venter told the FT. "It has changed my own thinking, both scientifically and philosophically, about life, and how it works."

The bacteria's genes were all constructed in the laboratory "from four bottles of chemicals on a chemical synthesizer, starting with information on a computer," he said.

Motivation

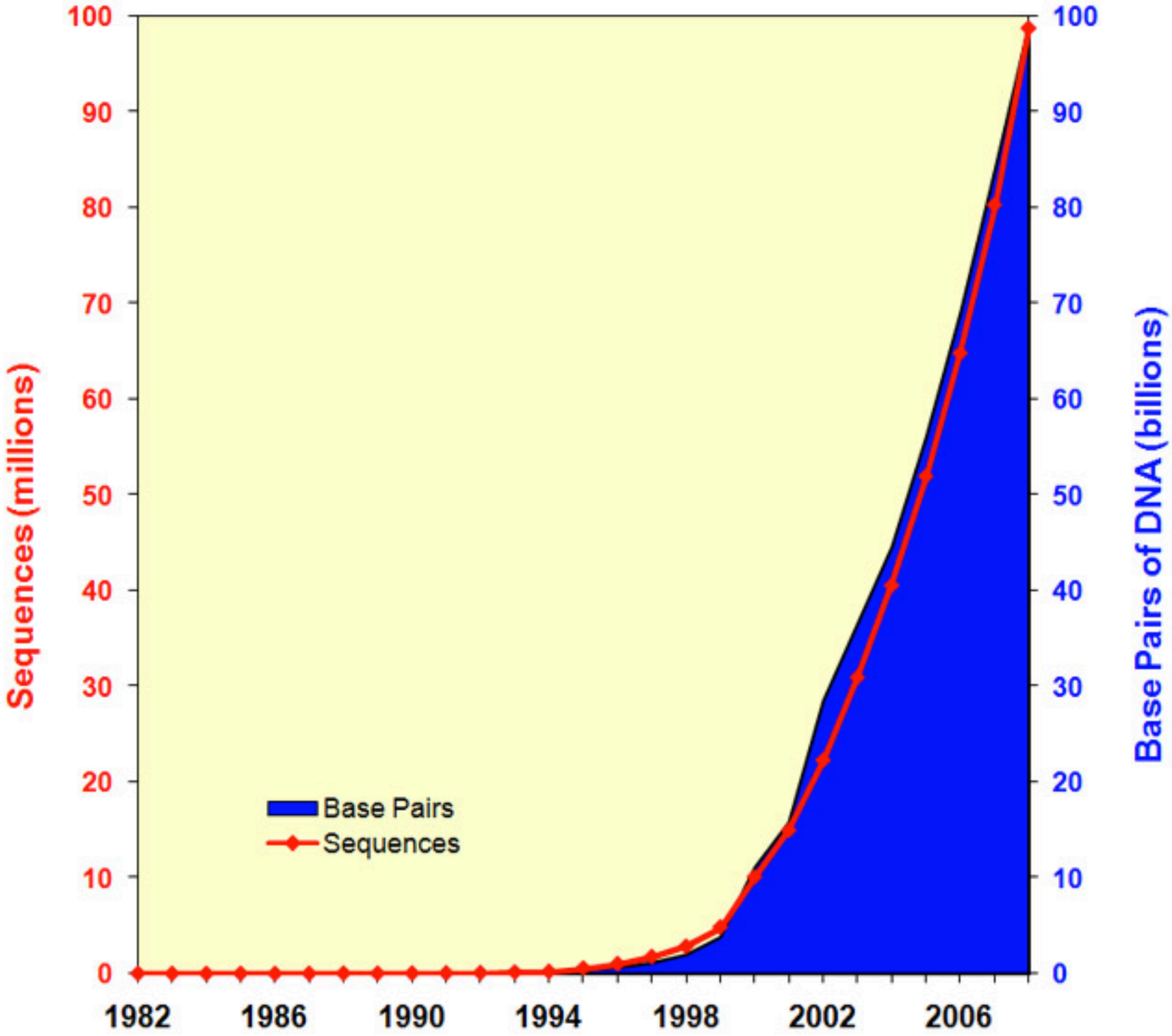
- Outlook on DNA Sequencing
  - From $500,000 to...
    - $50
  - From 3 months to...
    - < 1hr

Motivation

- Explosion of information
  - 750 Megabytes
    - $2_{\text{ bits}} * 3x10^{9}{}_{\text{ base pairs}} = 6x10^{9}{}_{\text{bits}}$
  - 100 million sequences and,

    99 billion base pairs in GenBank
  - Numbers will skyrocket as costs drop

Growth of GenBank
(1982 - 2008)



Opportunity

- Inevitable Trends
  - Genetic engineering will increasingly become a computing problem
  - Storage and analysis of data will occur in a fully distributed manner

Fundamental Problem

Limitations of today's representation/analysis techniques

- The inability to efficiently represent, manipulate, and analyze statistically significant samples of biological data within reasonable hardware limitations.
- Primary limiting factors:
  - memory
  - bandwidth

Steps to a Solution

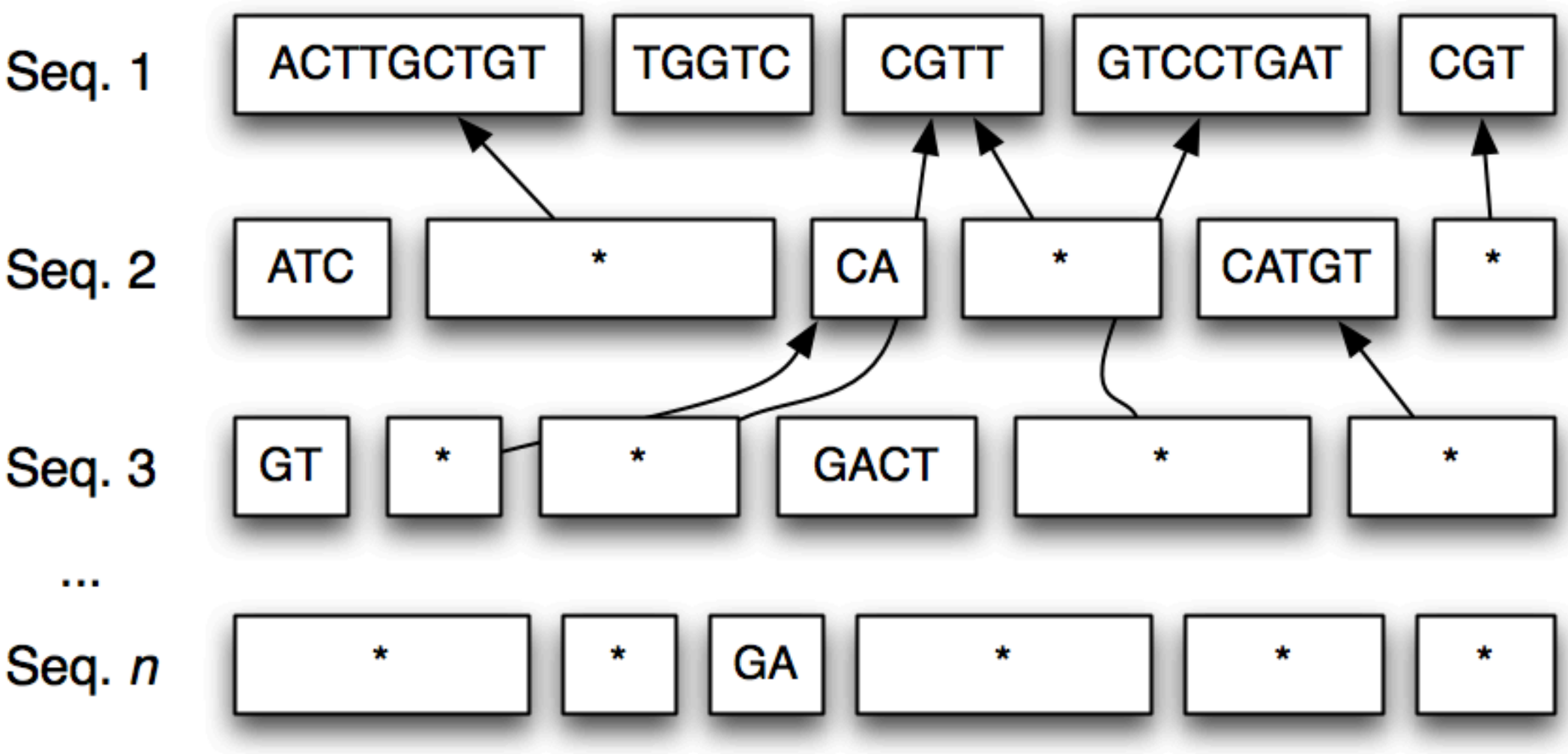
"Nature is a tinkerer, not an inventor" – François Jacob

- Evolution is inherently incremental
  - An organism's DNA is related to that of it's ancestors
  - Estimate: Only 0.05% of the genetic code of unrelated people differ

Approach


AAAGTTGCATTGGCATTG
ATACTGCACGTTGCGTTG
AAAGT--TGCA--TTG-GCATTG
A---TACTGCACGTTGCG--TTG
AAAG--TGCA--TTG-CA
A---TACTGCACGTTGCG--TTG

Approach


Seq. 1
ACTTGCTGT
TGGTC
CGTT
GTCCTGAT
CGT
Seq. 2
ATC
*
CA
*
CATGT
*
Seq. 3
GT
*
*
GACT
*
*
...
Seq. n
*
*
GA
*
*
*

Straw Man Implementation

Dynamic Programming

A T C C G A C G T C G G and A T C C G T C G G

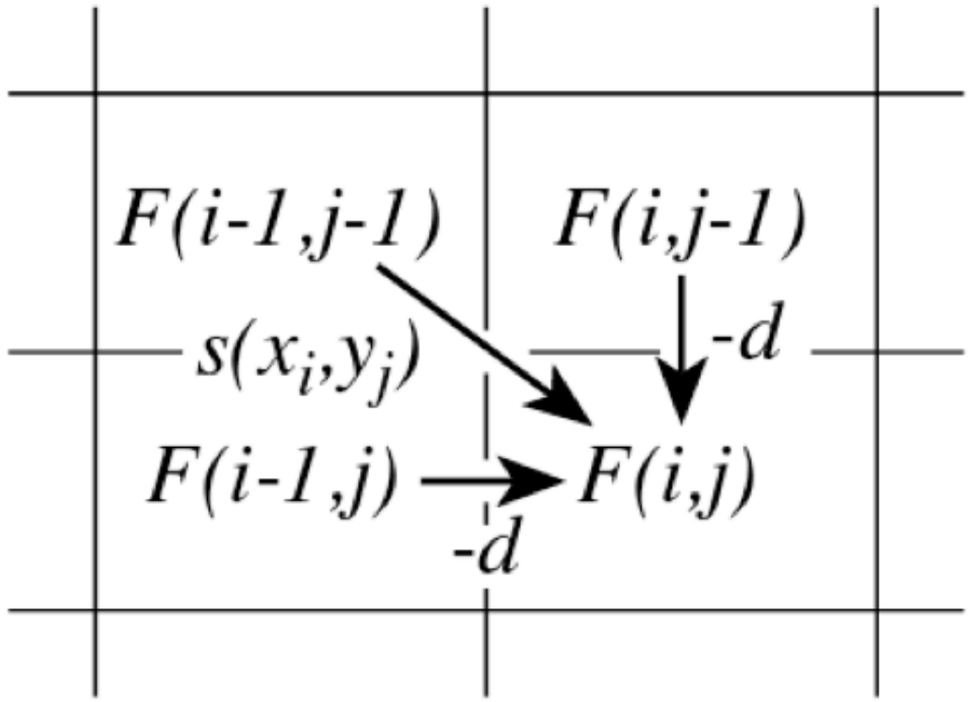
		A	T	C	C	G	T	C	G	G
	0	0	0	0	0	0	0	0	0	0
A	0	**1**	1	1	1	1	1	1	1	1
T	0	1	**2**	2	2	2	2	2	2	2
C	0	1	2	**3**	3	3	3	3	3	3
C	0	1	2	3	**4**	4	4	4	4	4
G	0	1	2	3	4	**5**	5	5	5	5
A	0	1	2	3	4	**5**	5	5	5	5
C	0	1	2	3	4	**5**	5	6	6	6
G	0	1	2	3	4	**5**	5	6	7	7
T	0	1	2	3	4	**5**	6	6	7	7
C	0	1	2	3	4	**5**	6	7	7	7
G	0	1	2	3	4	**5**	6	6	8	8
G	0	1	2	3	4	**5**	6	6	8	9

Alignment

```
A T C C G A C G T C G G
A T C C G - - - T C G G
```

[1] Dynamic Programming, Needleman-Wunsch

Implementation



$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

[2] Durbin et al.

Implementation

A T C C G A C G T C G G and A T C C G T C G G

		A	T	C	C	G	T	C	G	G
	0	0	0	0	0	0	0	0	0	0
A	0	**1**	1	1	1	1	1	1	1	1
T	0	1	**2**	2	2	2	2	2	2	2
C	0	1	2	**3**	3	3	3	3	3	3
C	0	1	2	3	**4**	4	4	4	4	4
G	0	1	2	3	4	**5**	5	5	5	5
A	0	1	2	3	4	**5**	5	5	5	5
C	0	1	2	3	4	**5**	5	6	6	6
G	0	1	2	3	4	**5**	5	6	7	7
T	0	1	2	3	4	**5**	6	6	7	7
C	0	1	2	3	4	**5**	6	7	7	7
G	0	1	2	3	4	**5**	6	6	8	8
G	0	1	2	3	4	**5**	6	6	8	9

Valid Alignments

```
A T C C G A C G T C G G
A T _ C _ _ C G T C G G

A T C C G A C G T C G G
A T C C _ _ _ G T C G G

A T C C G A C G T C G G
A T C _ _ _ C G T C G G
```

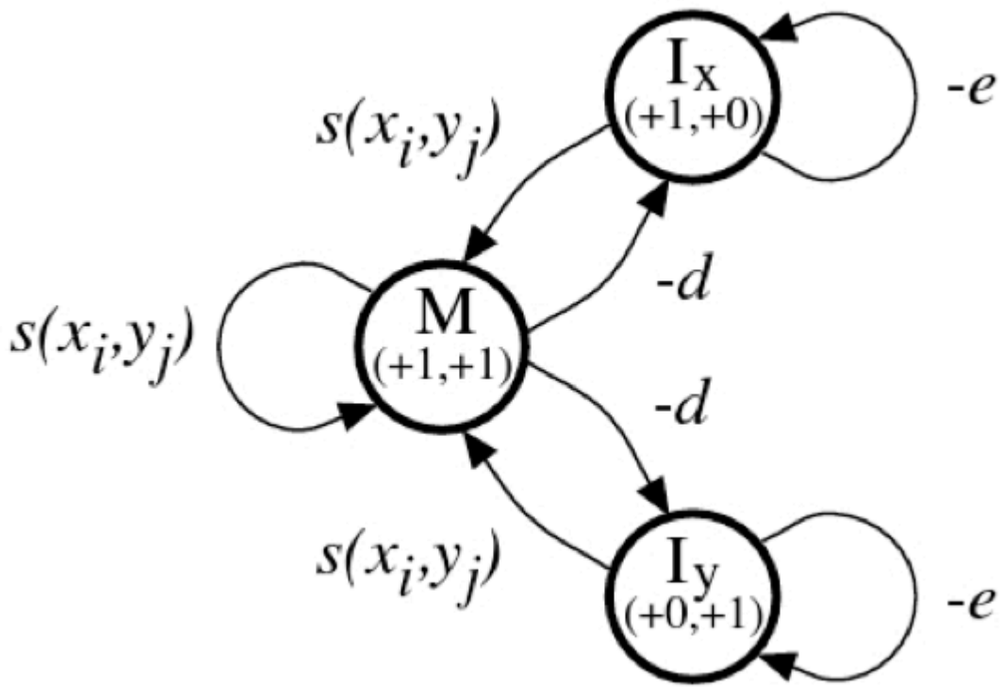
[1] Dynamic Programming, Needleman-Wunsch

Implementation 2.0

- Objective: Cluster the occurrence of gaps
- Approach: Penalize more for opening gap, and less for subsequent gaps
- Solution:
  - Use three matrices instead of one
  - Maintain state across gaps to coalesce them into “gap segments”

Implementation 2.0

State Machine



Equations

$$M(i,j) = \max \begin{cases} M(i-1,j-1)+s(x_i,y_j), \\ I_x(i-1,j-1)+s(x_i,y_j), \\ I_y(i-1,j-1)+s(x_i,y_j); \end{cases}$$

$$I_x(i,j) = \max \begin{cases} M(i-1,j)-d, \\ I_x(i-1,j)-e; \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i,j-1)-d, \\ I_y(i,j-1)-e. \end{cases}$$

[2] Durbin et al.

Results

Sequence Test Run 3.4		*Organism*	*Region*	*Span*
Sequence A	Accession FJ919248 http://bit.ly/7GDXy	Human rotavirus A	outer capsid protein (VP4)	Bases 0 - 100
	ATTTATAGACAACTTCTCACTAATTACTATTCGGTAGACTTGCATGACGAAATAGAACAGATTGGATCGGAGAAAACTCAAAATGTGACGGTAAATCCAG			
Sequence B	GGCTATAAAATGGCTTCGCTCATTTATAGACAGCTTCTCACTAATTCATATTCAGTAGATTTATATGATGAAATAGAGCAAATTGGATCAGAAAAAACTC			
	Accession EU839962 http://bit.ly/M5ibV	Human rotavirus G9P[8]	outer capsid protein (VP4)	Bases 0 - 100

Percent compression on sequence B given sequence A	**41.00%**	39.00%	**5.13%**

Alignment

```
A: --------------------ATTTATAGACAA-CTTCTCACTAATTAC-TATTC-GGTAGACTT-- ...
B: GGCTATAAAATGGCTTCGCTCATTTATAGAC-AGCTTCTCACTAATT-CATATTCAG-TAGA-TTTA ...

... ---GCATGACGAAATAGAACAG---ATTGGATCGGAGA--AAACTCAAAATGTGACGGTAAATCCAG
... TATG-ATGA---AATAGA---GCAAATTGGATC--AGAAAAAACTC--------------------
```

Results

Sequence Test Run 3.1		*Organism*	*Region*	*Span*
Sequence A	Accession AF304073 http://bit.ly/10deJe	Physeter catodon (Sperm Whale)	cytochrome b, mitochondrial protein	Bases 0 - 100
	ATGACCAACATCCGAAAATCACACCCATTAATAAAAATCATTAACAATGCATTCATCGACCTCCCTACCCCATCAAACATTTCCTCATGATGAAACTTCG			
	ATGACCAACATCCGAAAAACACACCCATTGATAAAAATCGTCAACAACGCATTCATCGACCTCCCTACTCCATCAAACATCTCCTCATGATGAAATTTCG			
Sequence B	Accession U72040 http://bit.ly/iube4	Kogia breviceps (Pygmy Sperm Whale)	cytochrome b, mitochondrial protein	Bases 0 - 100

Percent compression on sequence B given sequence A	**69.00%**	69.00%	**0.00%**

Alignment

```
A: ATGACCAACATCCGAAA-ATCACACCCATT-AATAAAAAT------CATTAACAAT ...
B: ATGACCAACATCCGAAAAA-CACACCCATTGA-TAAAAATCGTCAAC---AAC--- ...

... GCATTCATCGACCTCCCTAC-CCCATCAAACATT-TCCTCATGATGAAACTT-CG
... GCATTCATCGACCTCCCTACTC-CATCAAACA-TCTCCTCATGATGAAA-TTTCG
```

Results

Sequence Test Run 3.2		*Organism*	*Region*	*Span*
Sequence A	Accession EF093041 http://bit.ly/NmVDT	Lagenorhynchus obliquidens (Pacific white-sided dolphin)	cytochrome b, mitochondrial protein	Bases 0 - 100
	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCTCAATGACGCATTCATCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG			
Sequence B	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATTATCAATAACGCATTCATTGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG			
	Accession AF084057 http://bit.ly/bmrCE	Pseudorca Crassidens (False Killer Whale)	cytochrome b, mitochondrial protein	Bases 0 - 100

Percent compression on sequence B given sequence A	**87.00%**	87.00%	**0.00%**

Alignment

```
A: ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCT--CAATGA-CG ...
B: ATGACCAACATCCGAAAAACACACCCACTAATAAAAAT--TATCAAT-AACG ...

... CATTCA-TCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG
... CATTCATT-GACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG
```

Results

Sequence Test Run 3.3		*Organism*	*Region*	*Span*
Sequence A	Accession EF093025 http://bit.ly/16E6VM	Lissodelphis Borealis (Northern Right Whale Dolphin)	cytochrome b, mitochondrial	Bases 0 - 100
	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCTCAATGACGCATTCATCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG			
Sequence B	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCTCAATAACACATTCATCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG			
	Accession AY257155 http://bit.ly/fjF7g	Lagenorhynchus Obscurus (Dusky Dolphin)	cytochrome b, mitochondrial	Bases 0 - 100
Percent compression on sequence B given sequence A		**92.00%**	92.00%	**0.00%**

Alignment

```
A: ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCTCA-----ATGAC ...
B: ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCCTCAATAACA---- ...

... GCATTCATCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG
... -CATTCATCGACCTACCCACTCCATCTAACATCTCCTCATGATGAAACTTTG
```

Current Ambitions

- Considering applicability to distributed filesystems
  - LBFS, Mazieres. Summer project.
  - Implement support of compression across *n* sequences rather than just two.
  - Test on larger scale sequence sets and measure performance

References

- [1] Needleman, S., Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. 1970.

- [2] Durbin, R., Eddy, S., Krogh A., Mitchison G. *Biological Sequence Analysis*. Cambridge University Press, 1998.