

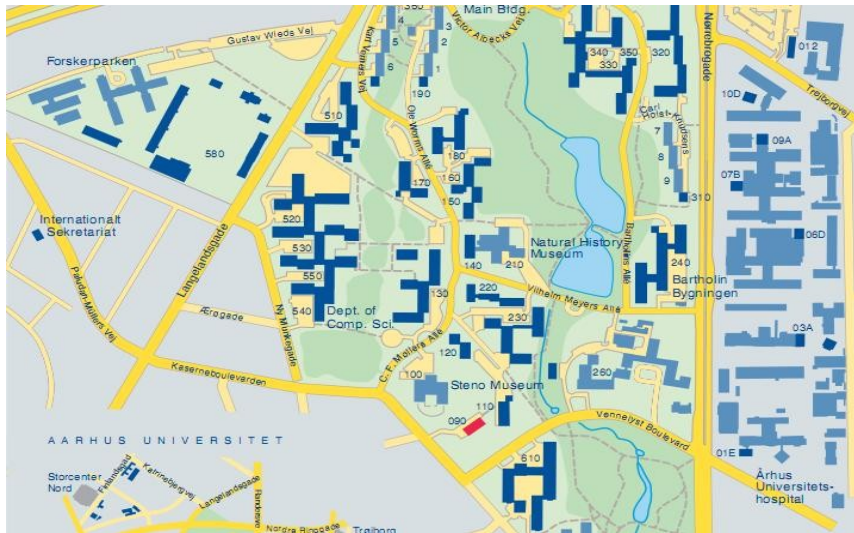
# Bioinformatics

Martin Simonsen

[zxr@cs.au.dk](mailto:zxr@cs.au.dk)

<http://www.birc.au.dk>

# Bioinformatics Research Center



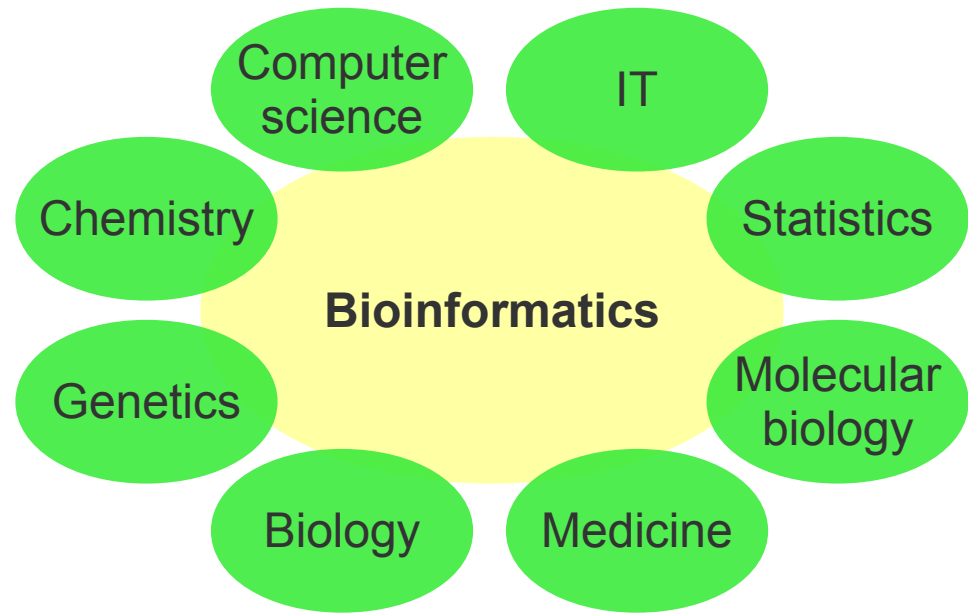
Established in 2001 by NAT and SUN. Currently 38 “BiRC-people”.

**VIP:** Christian Storm Pedersen (director of BiRC), Thomas Mailund.

**PhD-students:** Thomas Greve Kristensen, Jesper Nielsen, Peter Justesen, Christian Hachenberg, Martin Simonsen.

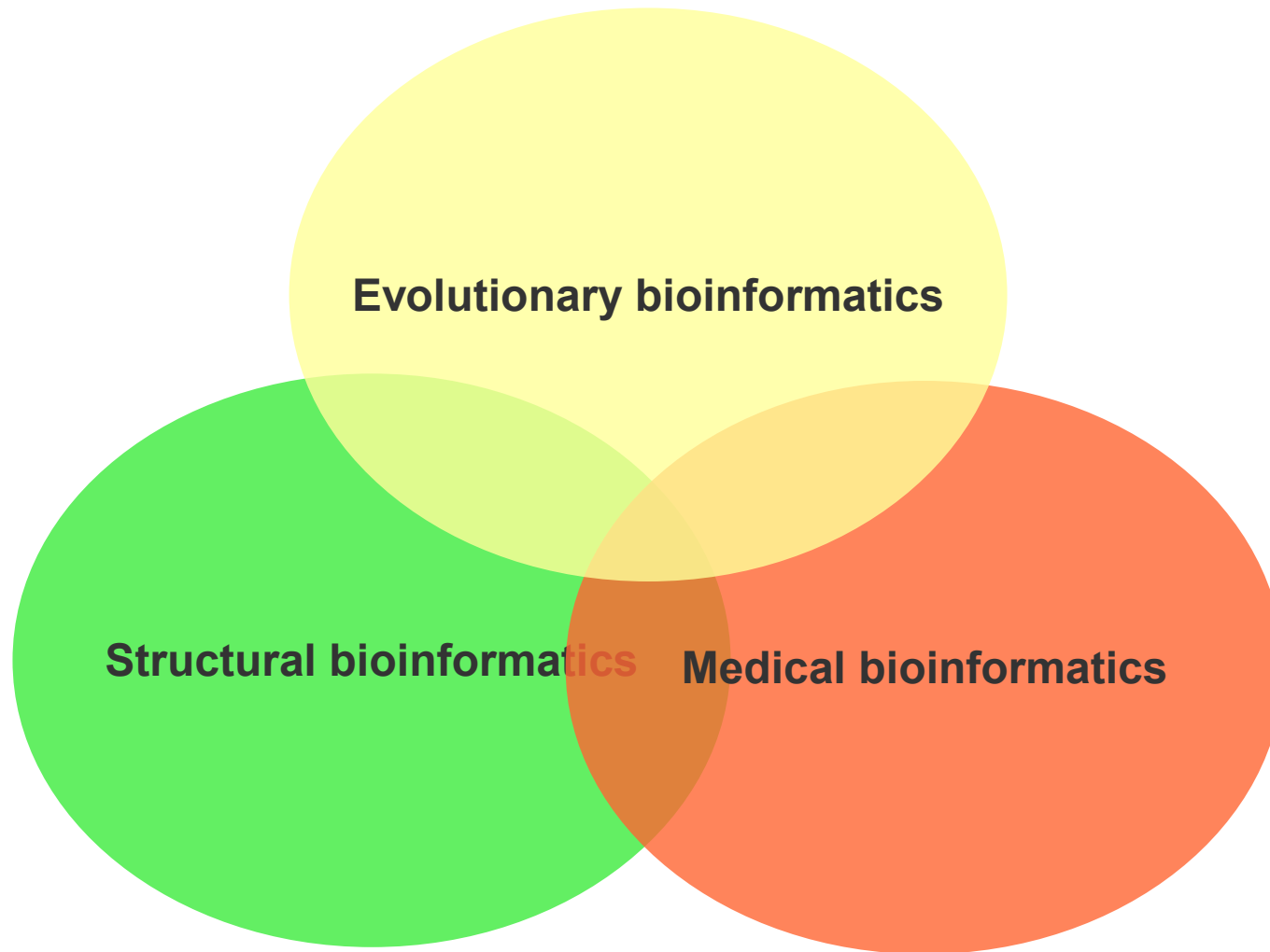
**Stud.progs:** Poul Liboriussen, Steffen Mikkelsen, Andreas Sand, Troels Toftebjerg, Anders Halager, Jørgen Fogh.

# Bioinformatics?

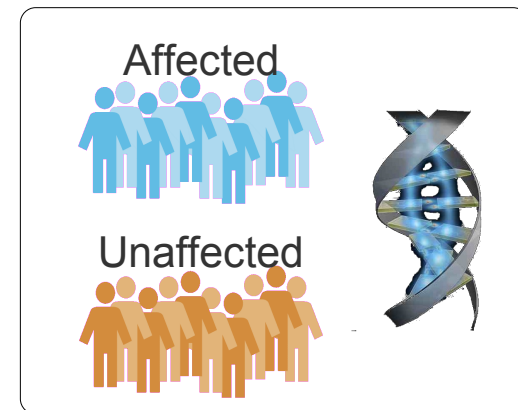
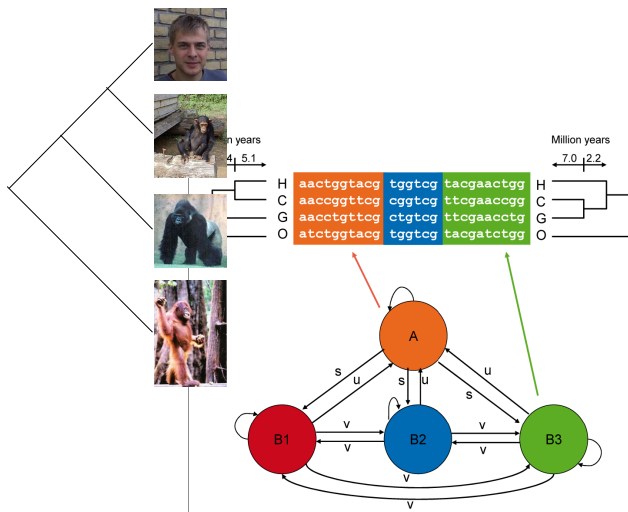


Developing and using algorithms and programs for collecting, handling, and analysis of biological and biomedical data

# Research at BiRC



# Research at BiRC



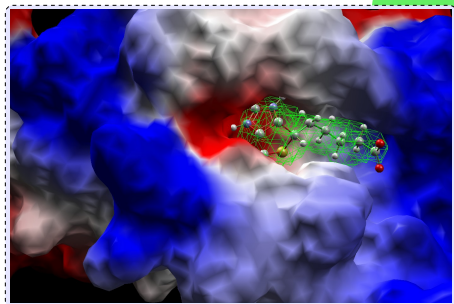
Evolutionary bioinformatics

Association mapping  
EU (POLYGENE), FTP and FNU

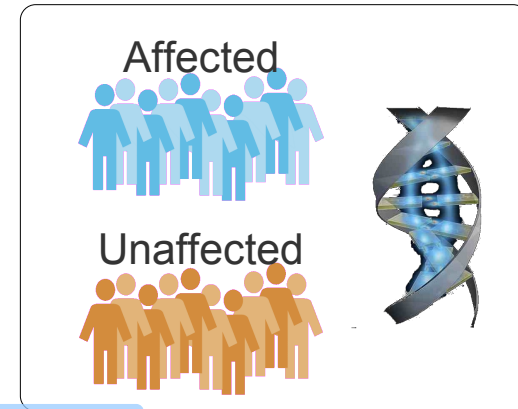
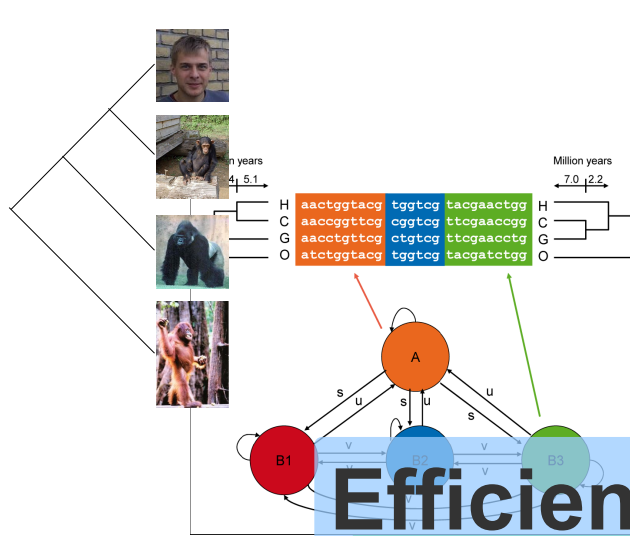
Genome analysis  
EU, HTF and FNU

Structural bioinformatics    Medical bioinformatics

Molecular docking and dynamics  
NABIIT, DCSC and DG (PUMPKIN)



# Research at BiRC



Evolutionary

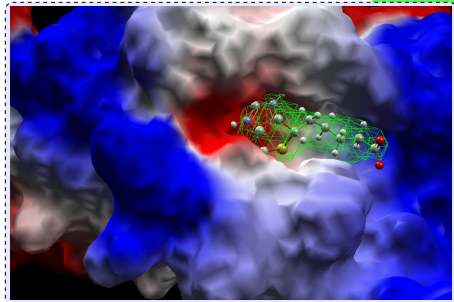
Association mapping  
EU (POLYGENE), FTD and FNU

Efficient algorithms in practice

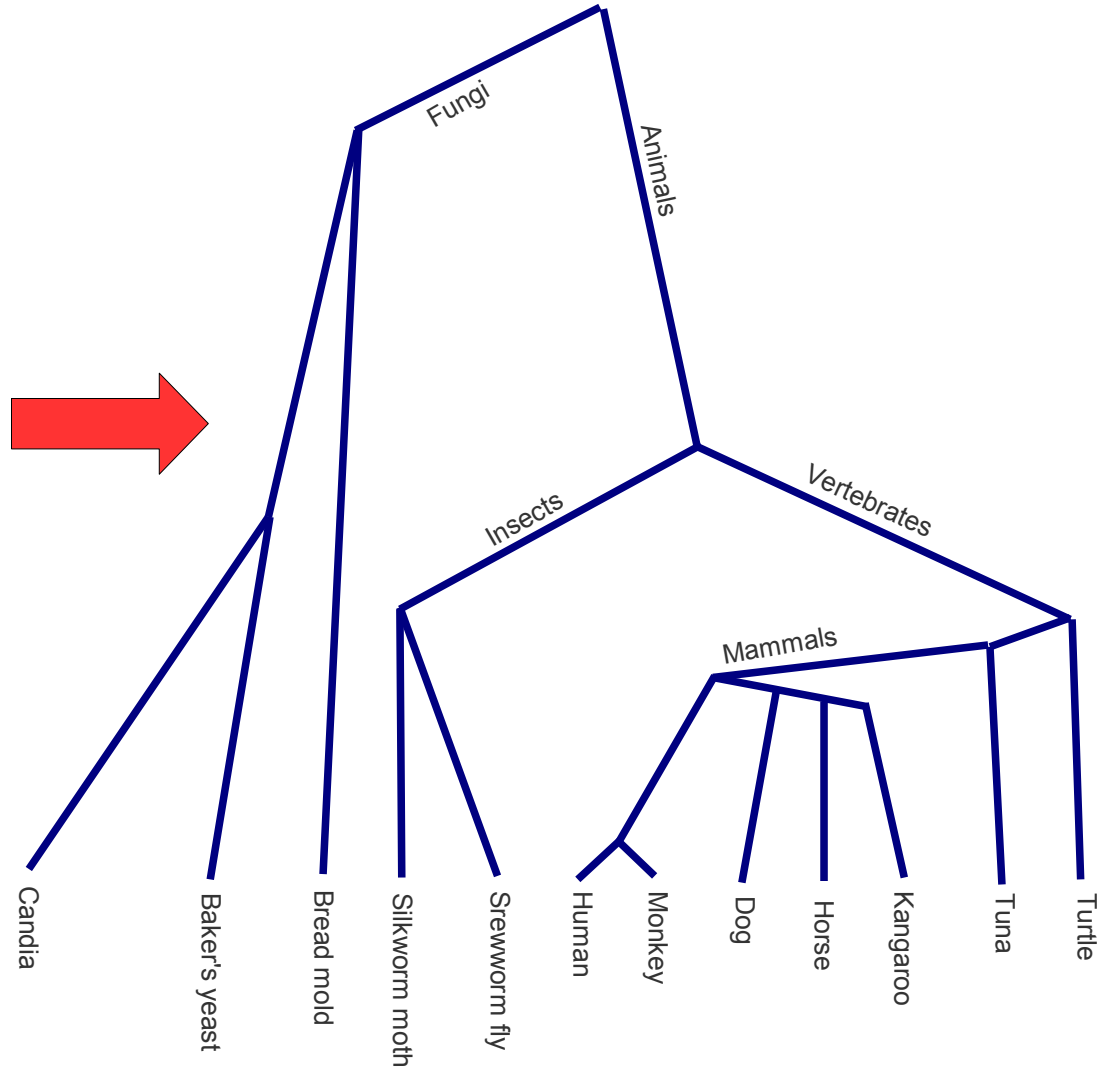
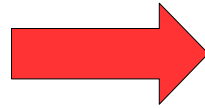
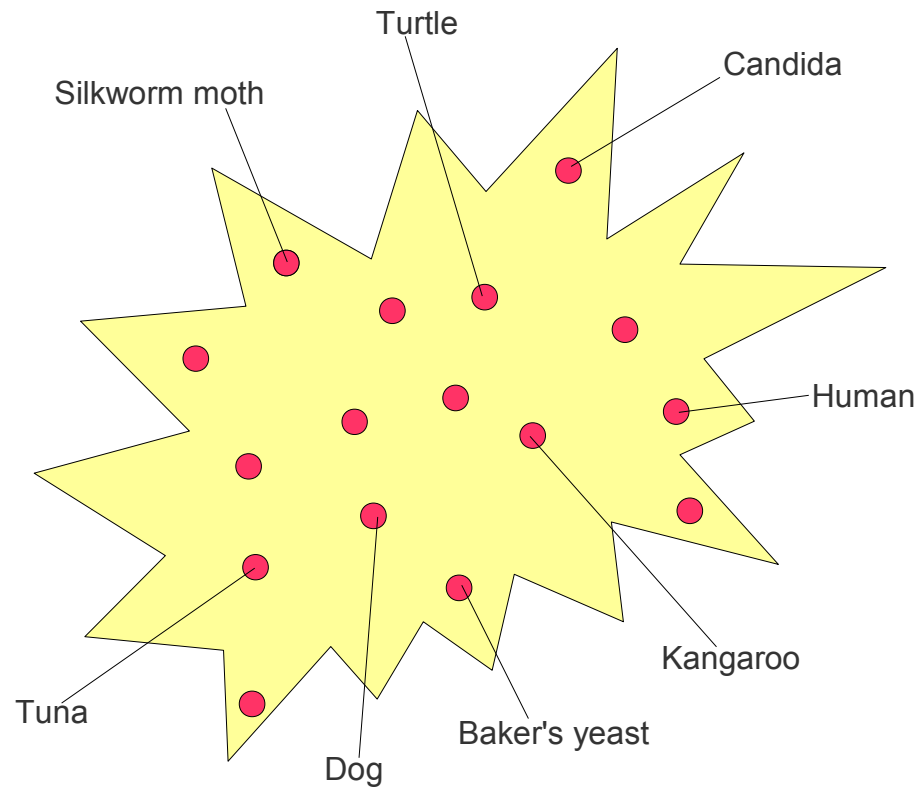
Genome analysis  
EU, HTF and FNU

Structural bioinformatics      Medical bioinformatics

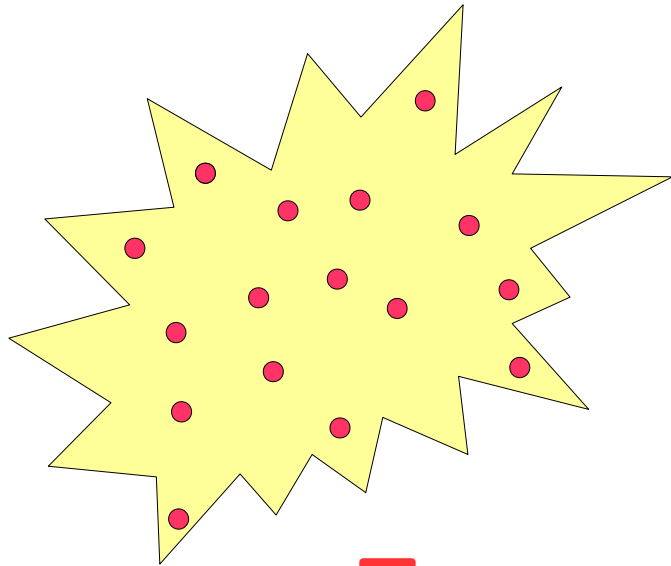
Molecular docking and dynamics  
NABIIT, DCSC and DG (PUMPKIN)



# Evolutionary Trees



# Distance Based Phylogenetic inference

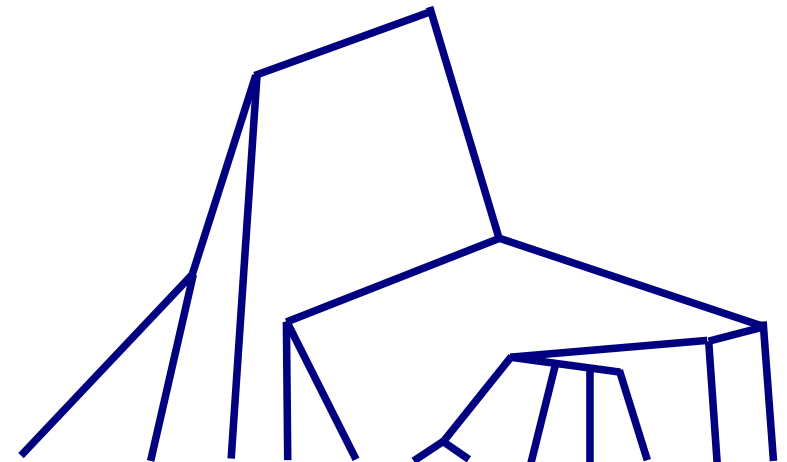


0.0									
0.96	0.0								
1.15	0.04	0.0							
0.87	1.14	0.87	0.0						
0.13	0.45	1.97	0.14	0.0					
0.69	0.47	0.36	0.44	0.66	0.0				
0.32	0.87	1.22	1.64	0.11	0.07	0.0			
1.77	0.38	0.84	0.72	1.73	0.47	0.13	0.0		
1.21	0.17	0.51	1.31	1.81	0.46	0.73	0.46	0.0	
0.91	1.54	1.15	1.47	1.46	1.48	1.48	1.67	0.23	0.0

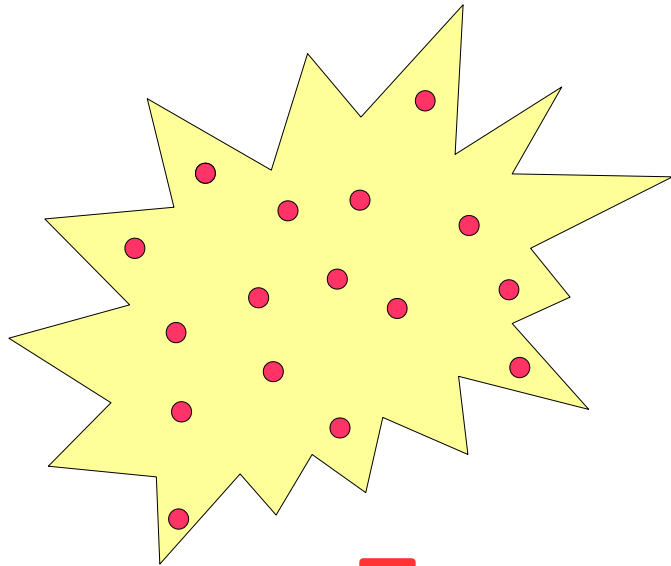


HUMAN  
CANDIDA  
TUNA  
KANGAROO  
HORSE  
:  
:  
:

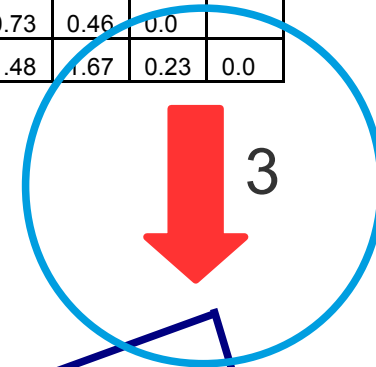
CT---AAACAACCTCTGTACAGTCTCCGAGCTAGCACGGC  
TAAGGTAACCGCGT---ATCTCT---GAGGTGCAG---A  
TCGCATACCACTCCATTGTT-----GAGAAGAGAGCTC  
AAAAC TAGGTGAACACTGAGCTACAGATGAAC---CTGA  
TT---GAGTACGCTACGA-----CGTACCGGTGT  
:  
:  
:  
:



# Distance Based Phylogenetic inference



0.0									
0.96	0.0								
1.15	0.04	0.0							
0.87	1.14	0.87	0.0						
0.13	0.45	1.97	0.14	0.0					
0.69	0.47	0.36	0.44	0.66	0.0				
0.32	0.87	1.22	1.64	0.11	0.07	0.0			
1.77	0.38	0.84	0.72	1.73	0.47	0.13	0.0		
1.21	0.17	0.51	1.31	1.81	0.46	0.73	0.46	0.0	
0.91	1.54	1.15	1.47	1.46	1.48	1.48	1.67	0.23	0.0



HUMAN  
CANDIDA  
TUNA  
KANGAROO  
HORSE  
:  
:  
:

CT---AAACAAC TCTGTACAGTCTCCGAGCTAGCACGGC  
TAAGGTAACCGCGT---ATCTCT---GAGGTGCAG---A  
TCGCATACCACTCCATTGTT-----GAGAAGAGAGCTC  
AAAAGTGGTGAACACTGAGCTACAGATGAAC---CTGA  
TT---GAGTACGCTACGA-----CGGTACCGGTGT  
:  
:  
:  
:

# The Neighbour-Joining Method



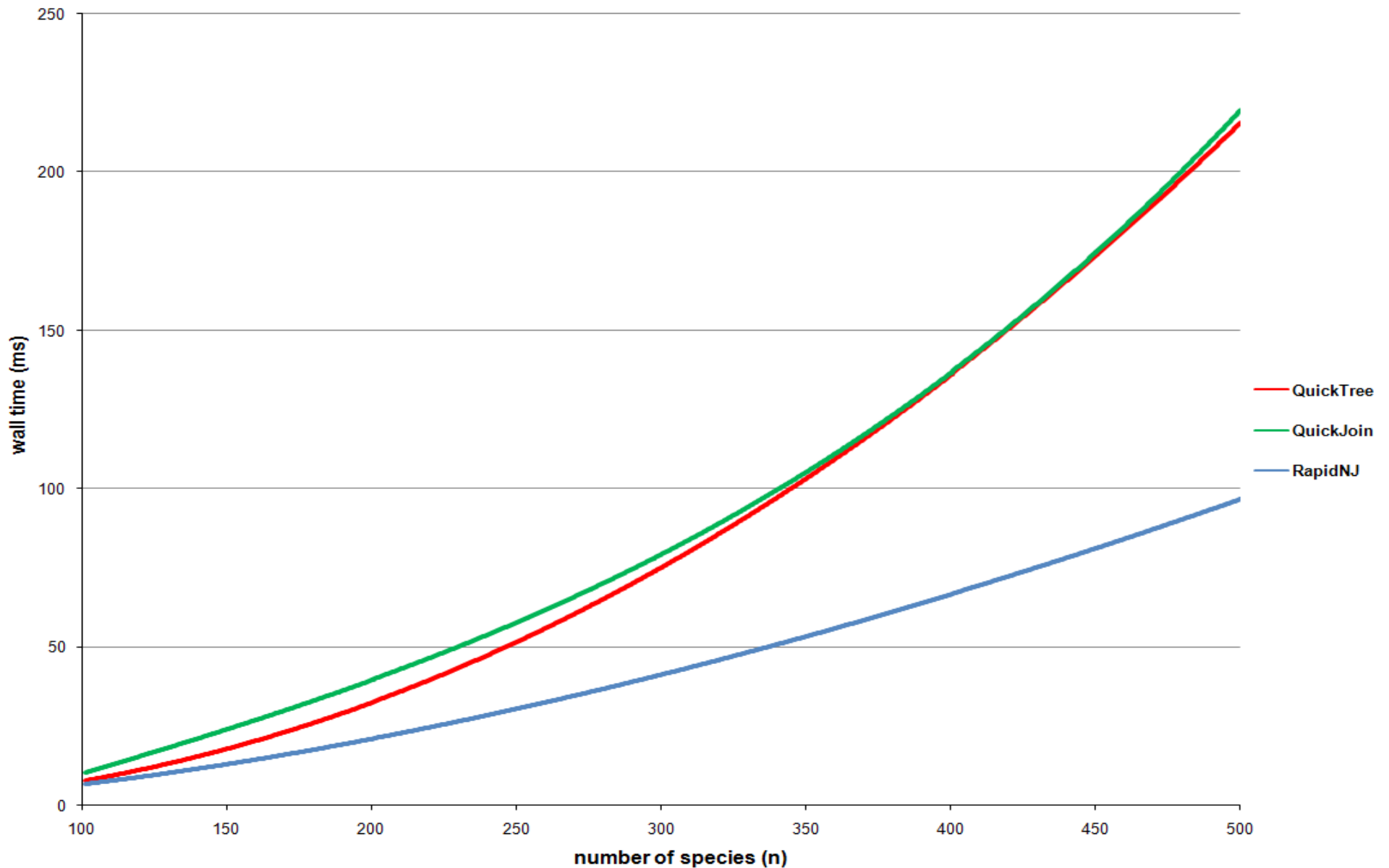
- Simple and fairly accurate method from 1987.
- Criticized by many but still widely used.
- Infers trees with  $n$  species in time  $O(n^3)$ .
- Infeasible to infer large trees.

**QuickTree (2003)** by Howe et al. is an efficient implementation of the naive algorithm.

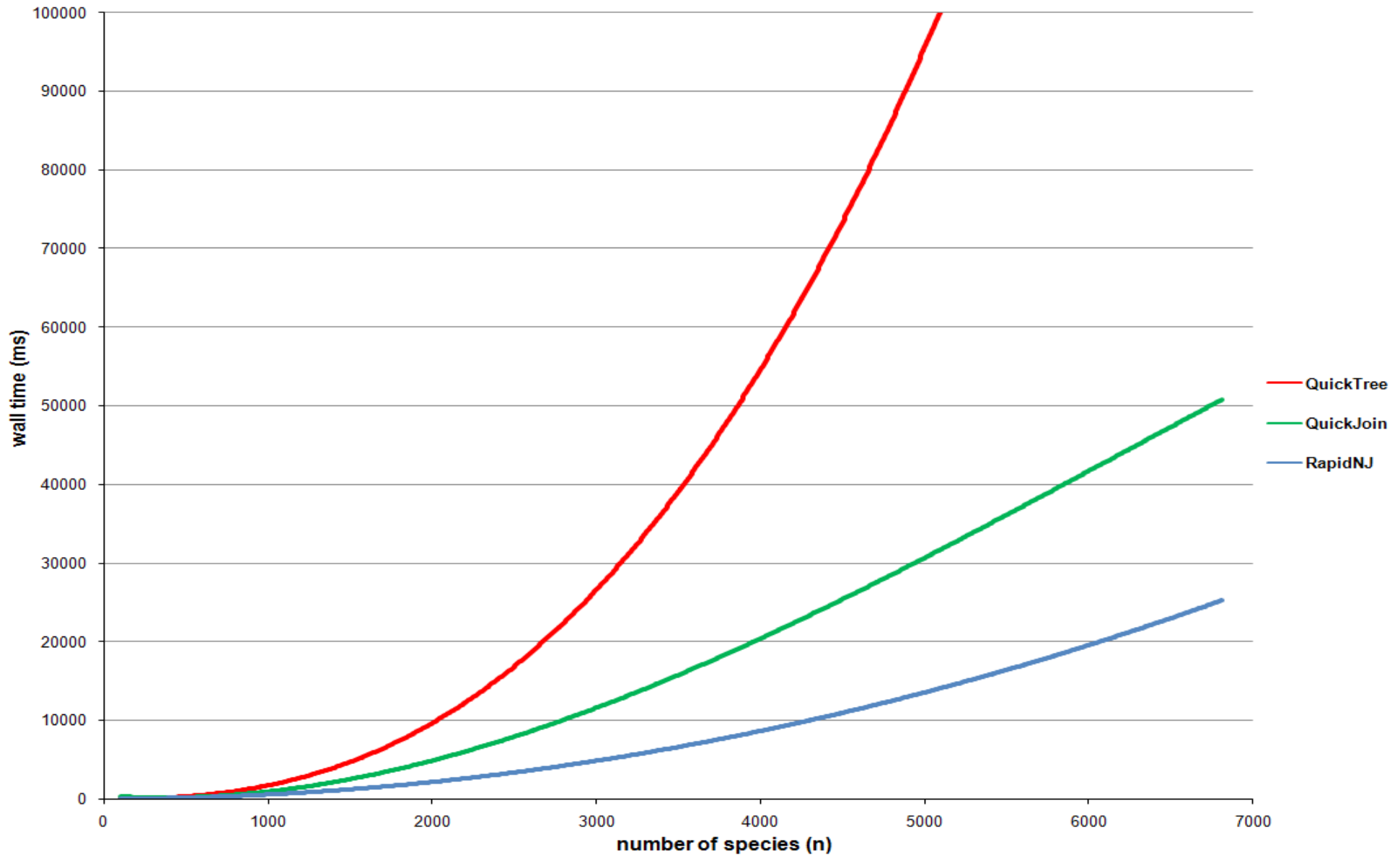
**QuickJoin (2006)** by Mailund et al. uses heavy algorithmic machinery to reduce the running time in practice for large inputs.

**RapidNJ (2008)** by Simonsen et al. uses a simple heuristic to reduce the running time on all input sizes. Published at WABI 2008.

# Running times on Small Data Sets



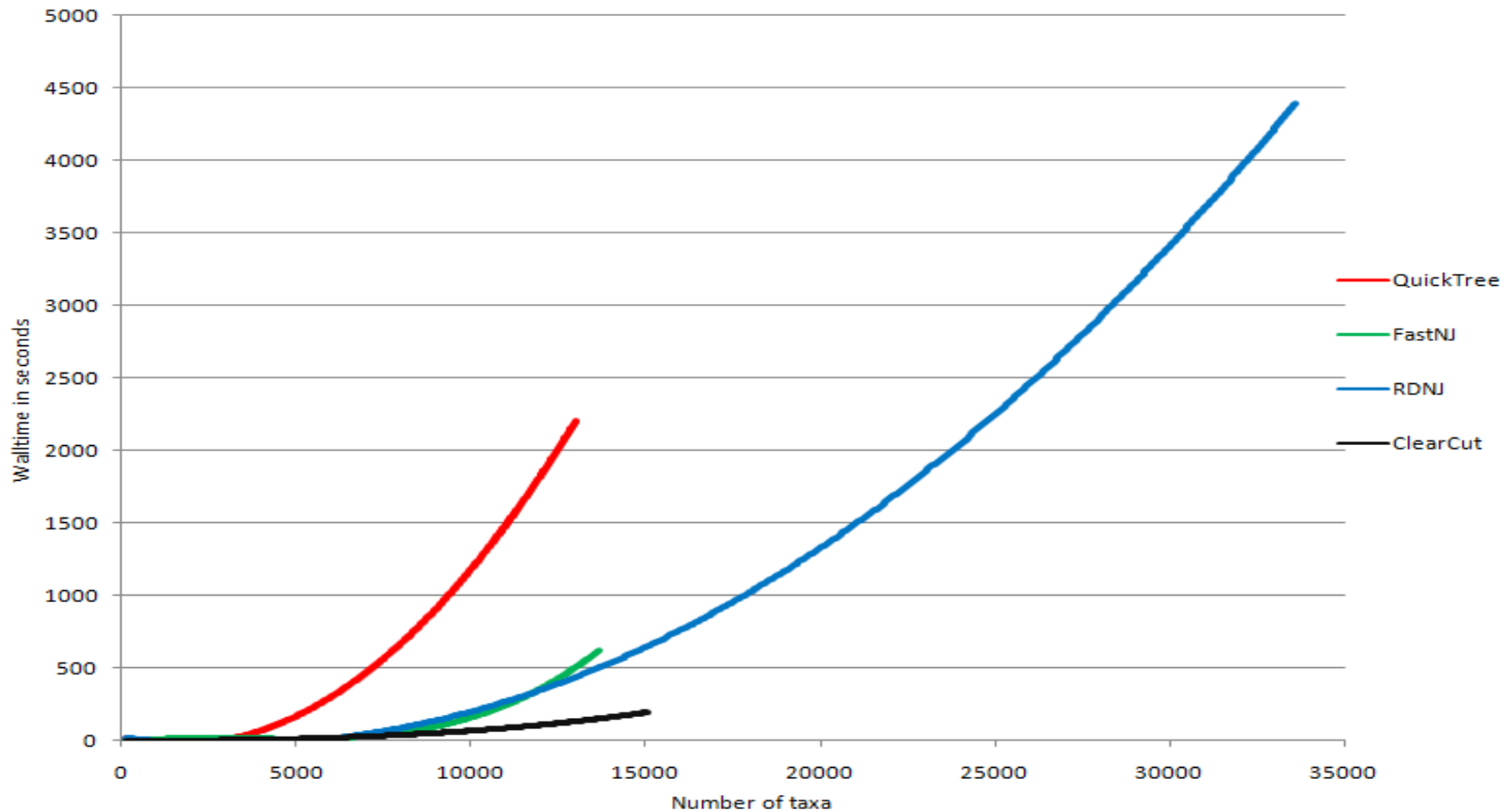
# Running times on Large Data Sets



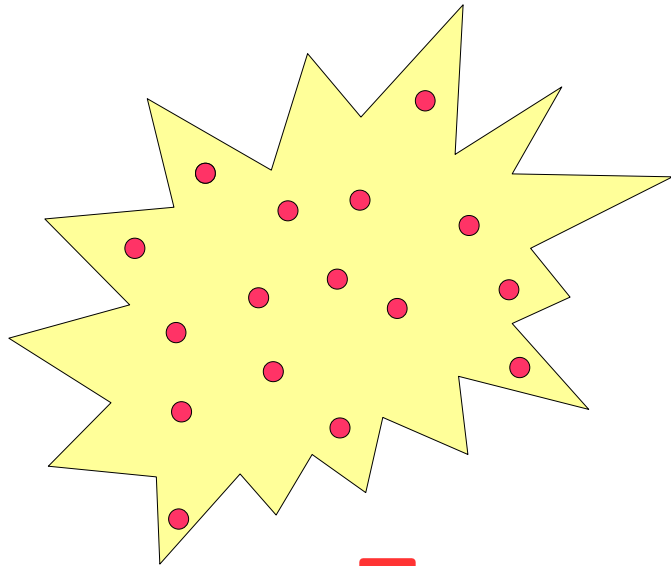
# Next Step – I/O Efficiency

The  $O(n^2)$  space consumption of the Neighbour-Joining method is a problem for data sets with 10000+ species.

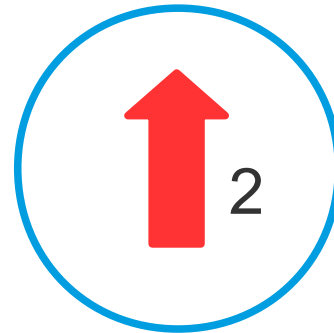
RapidDiskNJ is an I/O efficient version of RapidNJ.



# Distance Based Phylogenetic inference

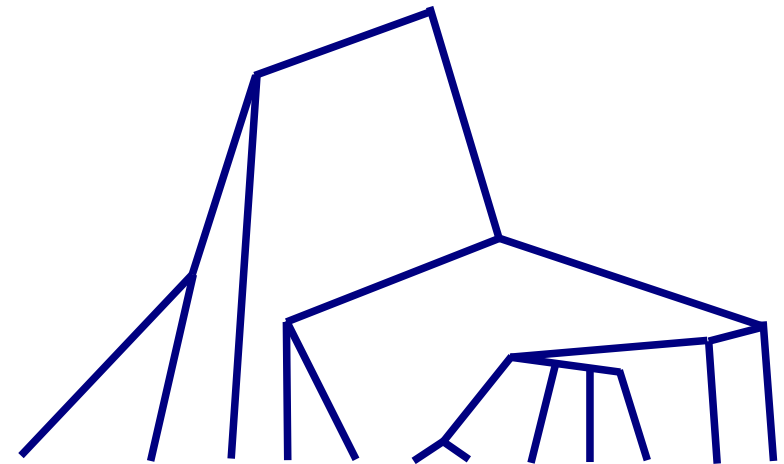


0.0									
0.96	0.0								
1.15	0.04	0.0							
0.87	1.14	0.87	0.0						
0.13	0.45	1.97	0.14	0.0					
0.69	0.47	0.36	0.44	0.66	0.0				
0.32	0.87	1.22	1.64	0.11	0.07	0.0			
1.77	0.38	0.84	0.72	1.73	0.47	0.13	0.0		
1.21	0.17	0.51	1.31	1.81	0.46	0.73	0.46	0.0	
0.91	1.54	1.15	1.47	1.46	1.48	1.48	1.67	0.23	0.0



HUMAN  
CANDIDA  
TUNA  
KANGAROO  
HORSE  
:  
:  
:

CT---AAACA  
ACTCTGTACAGTCTCCGAGCTAGCACGGC  
TAAGGTAACCGCGT---ATCTCT---GAGGTGCAG---A  
TCGCATACCACTCCATTGTT-----GAGAAGAGAGCTC  
AAAAGTGGTGAACACTGAGCTACAGATGAAC---CTGA  
TT---GAGTACGCTACGA-----CGGTACCGGTGT  
:  
:  
:  
:



# Counting mutational events in DNA sequences

```

HUMAN      CT---AACAACCTCTGTACAGTCTCCGAGCTAGCACGGC
CANDIDA    TAAGGTAAACCGCT---ATCTCT---GAGGTGCAG---A
TUNA       TCGCATACCACTCCATTGTT-----GAGAAGAGAGCTC
KANGAROO   AAAACTAGGTGAACACTGAGCTACAGATGAAC---CTGA
HORSE      TT---GAGTACGCTACGA-----CGGTACCGGTGT
:          :          :          :          :
:          :          :          :          :
:          :          :          :          :

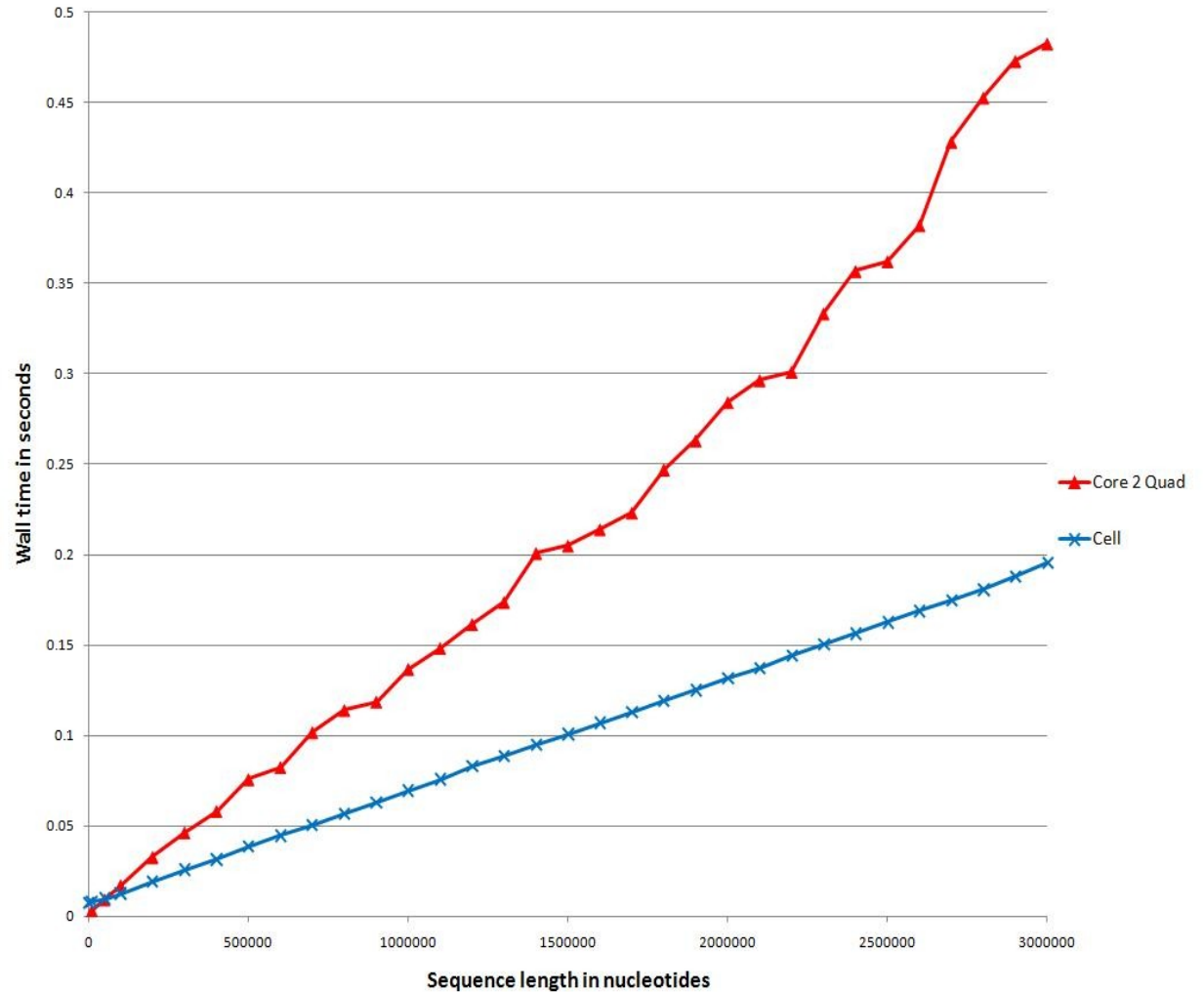
```



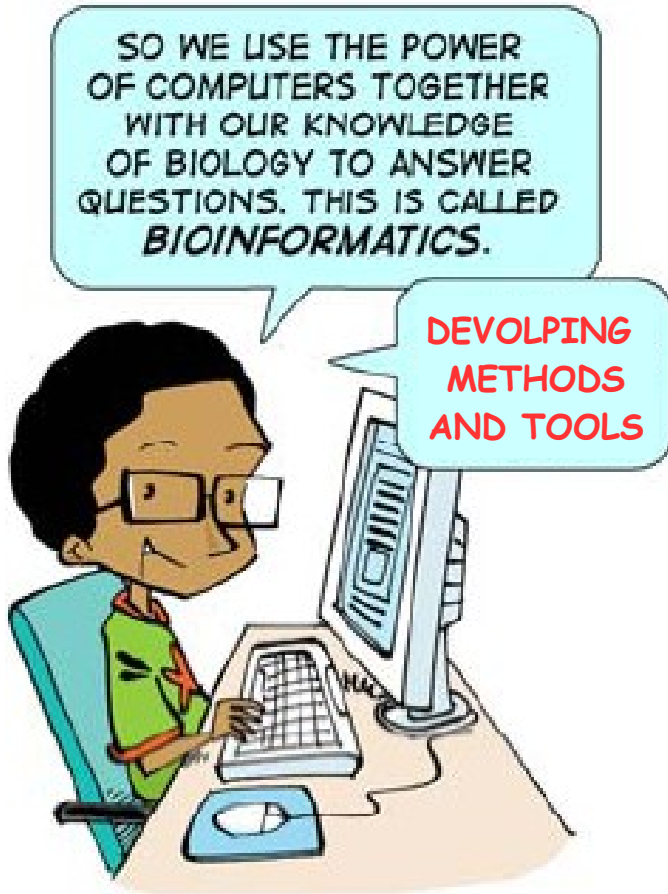
- Time  $O(n^2 l)$  to compare all  $n^2$  sequence pairs of length  $l$ .
- Too slow in practice for large  $l$ .
- Using a binary encoding of the DNA sequences and **the SIMD capabilities of modern CPU's** we can speed up this step.

0.0									
0.96	0.0								
1.15	0.04	0.0							
0.87	1.14	0.87	0.0						
0.13	0.45	1.97	0.14	0.0					
0.69	0.47	0.36	0.44	0.66	0.0				
0.32	0.87	1.22	1.64	0.11	0.07	0.0			
1.77	0.38	0.84	0.72	1.73	0.47	0.13	0.0		
1.21	0.17	0.51	1.31	1.81	0.46	0.73	0.46	0.0	
0.91	1.54	1.15	1.47	1.46	1.48	1.48	1.67	0.23	0.0

# Using Cell processors for counting mutational events



# More information



<http://www.birc.au.dk>