PRACE Spring School 2014

Software Engineering for Supercomputers in Research & Industry

About the importance of HPC for Life Sciences

Oswaldo Trelles
University of Malaga, Spain
ortrelles@uma.es

Computer Architecture Dept.

Main Research Lines

- VLSI (ASIC integration)
- ‘Parallel’ Compilers
- Applications (sparse matrices → I-O bounded)

Bioinformatics: Computer sciences as applied to biological data

www.bitlab-es.com
Bioinformatics and Information Technologies Lab

BITLAB: Bioinformatics and Information Technologies Laboratory

Basic & applied Research

www.bitlab-es.com

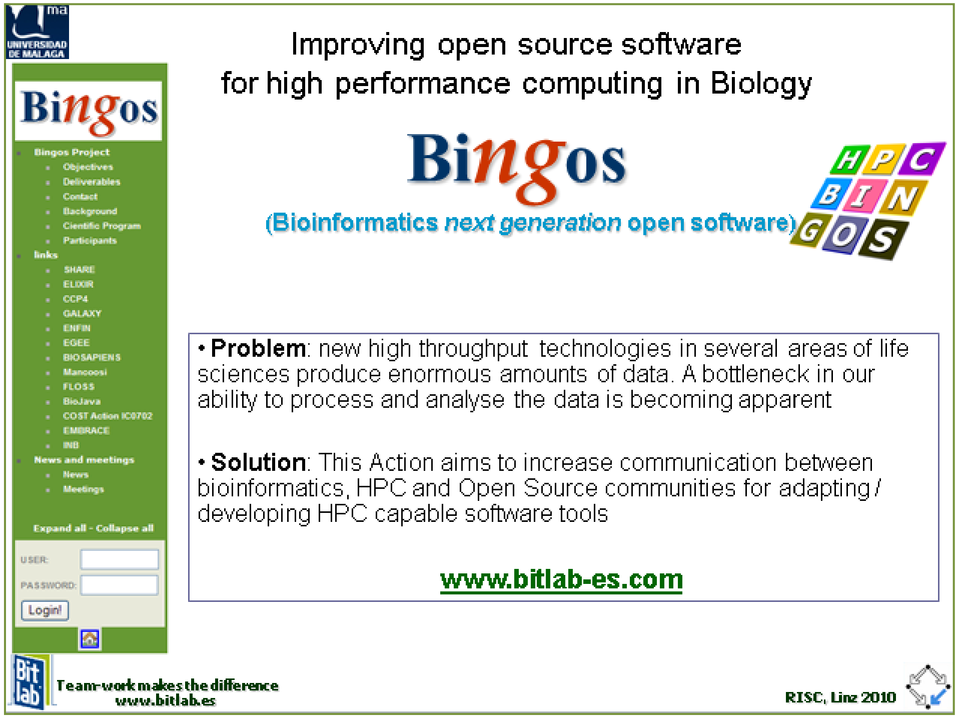
UNIVERSIDAD DE MÁLAGA

High Performance Computing applied to Life Sciences

RISC

JKU

UMA



Improving open source software
for high performance computing in Biology

Bingos

(Bioinformatics *next generation* open software)

- **Problem**: new high throughput technologies in several areas of life sciences produce enormous amounts of data. A bottleneck in our ability to process and analyse the data is becoming apparent

- **Solution**: This Action aims to increase communication between bioinformatics, HPC and Open Source communities for adapting / developing HPC capable software tools

www.bitlab-es.com

Bit lab Team-work makes the difference
www.bitlab.es

RISC, Linz 2010

PRACE

Targeting Big-Data problems in BI

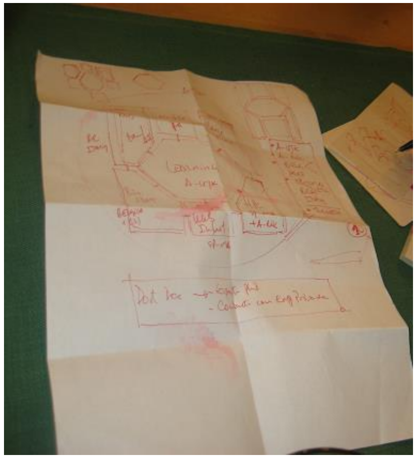
1995: 1 US$ per base (3.000M US$ the full human genome)
2000: 1 Mbp ≈ 10.000 US $ [1]
2008: Full human genome (3,2 Gbp) in 6 weeks, and ≈ $60,000 [2]
Predicted: US$1,000 genome in next 3 years.
2009: (October) nanopore DNA sequencing [3, 4]
2011: (Mar. 2011) 0,5 US$ per Mbp [5]
2012: (Feb. 2012) size: USB memory stick / $900 [6]

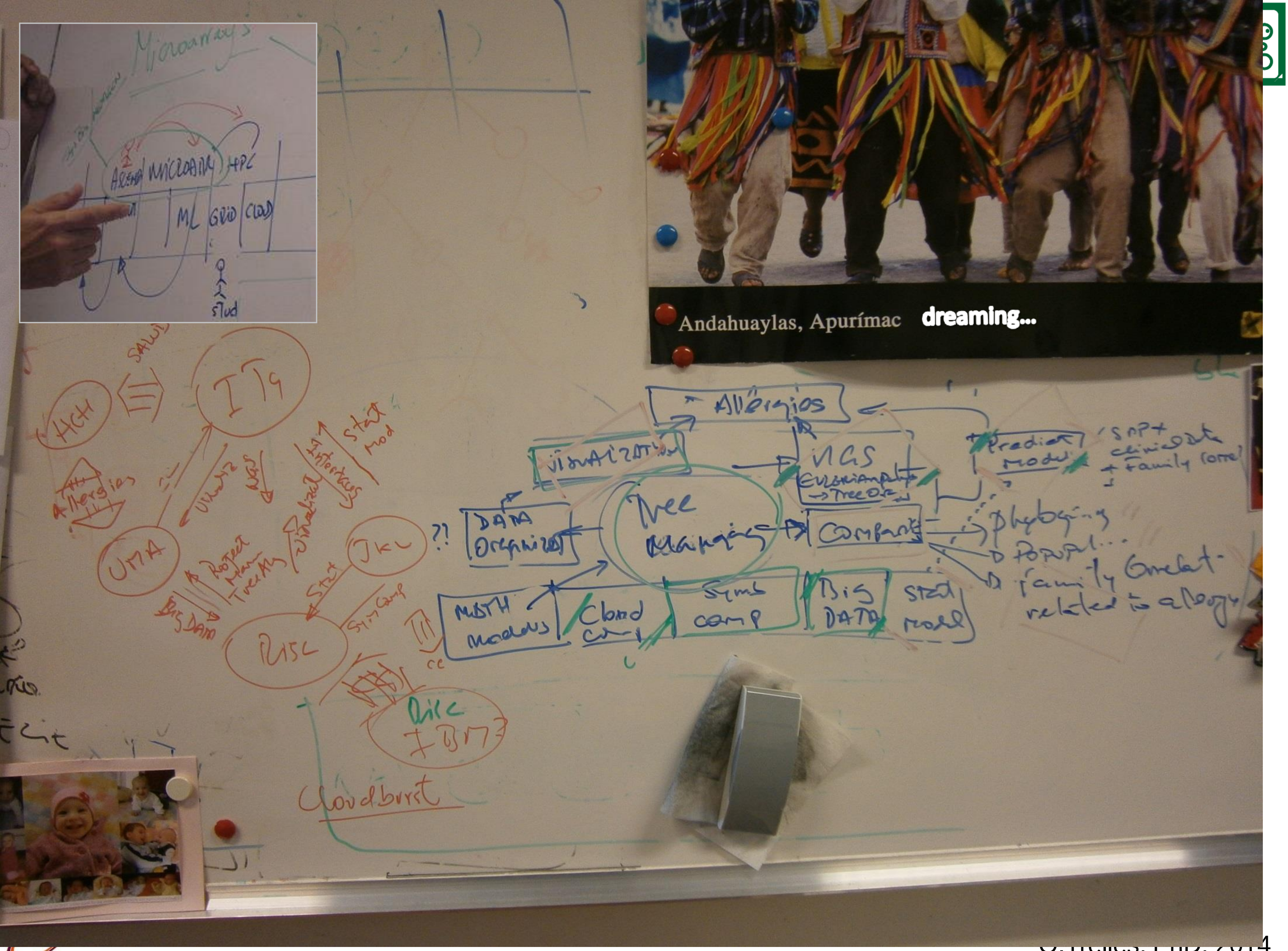
Allow individual to get his or her genome sequenced, thus truly ushering in the era of genetics based personalized medicine.

[1] http://www.genome.gov/11006943
[2] http://www.technologyreview.com/Biotech/20640/
[3] http://nextbigfuture.com/2009/10/ibm-targeting-100-dollar-genome.html
[4] http://www.nature.com/embor/journal/v8/n10/full/7401070.html
[5] http://singularityhub.com/2011/03/05/costs-of-dna-sequencing-falling-fast-look-at-these-graphs/
[6] http://www.nanoporetech.com/news/press-releases/view/39

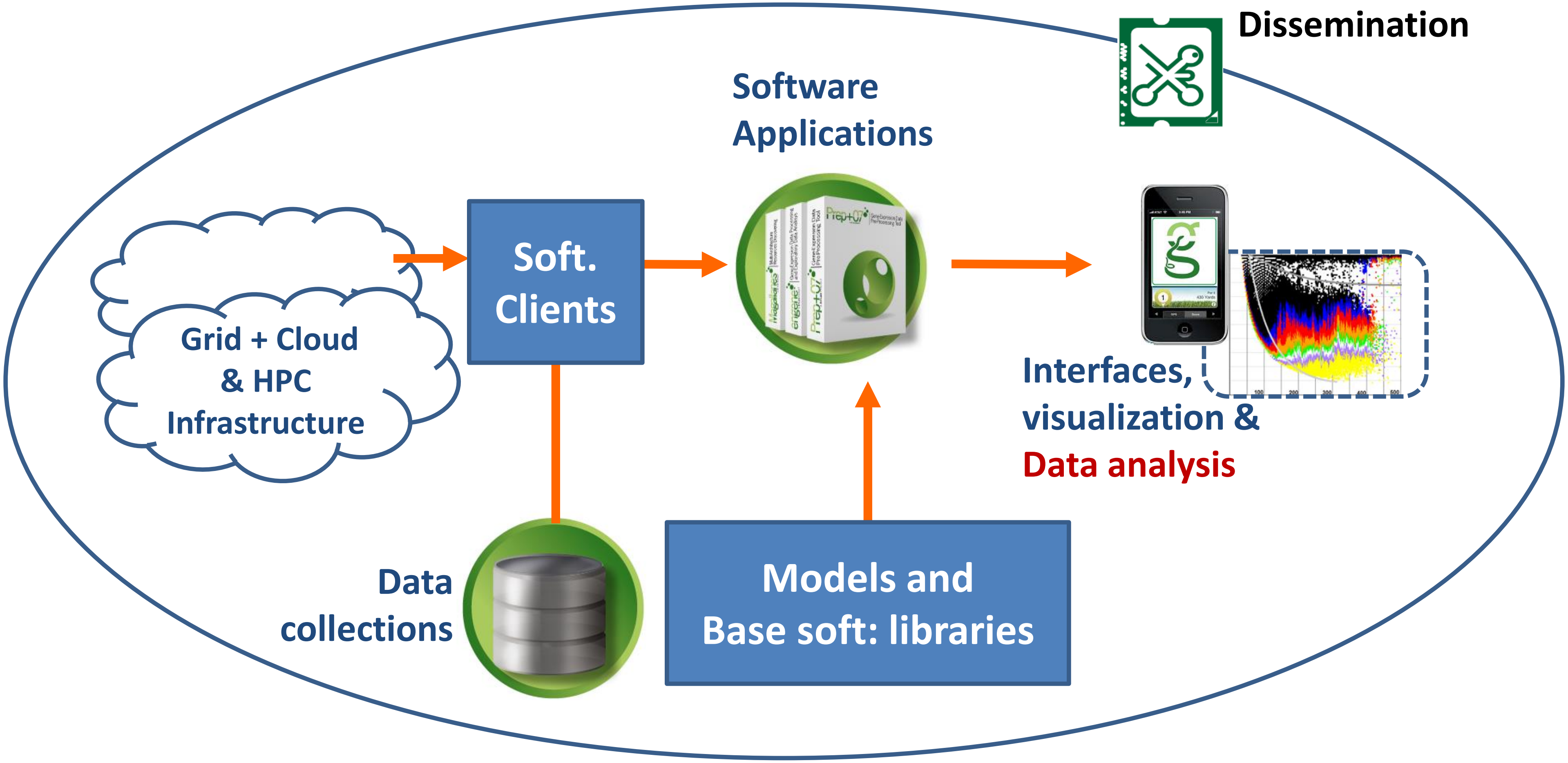
Trelles, O.; Prins P.; Snir M. and Jansen C.; (2011)
“Big data.. are we ready?”; Nature Reviews Genetics)

The first draft…



Microarrays
ARENA MICROARRAY HPC
ML
GRID
CLOUD
Stud
Andahuaylas, Apurímac
dreaming...
HCH
SAW
ITG
Allergias
UMA
Visualiz
stat mod
Interfaces
JKU
Project Man Tracking
Big DATA
RISC
Stat
Sym Comp
CC
QILC IBM
Cloudburst
Allergias
VISUALIZATION
DATA Organization
Tree Managing
NGS
Evaluation
→ Tree Dist
Predict model
SNP + clinical data + family correl.
Compare
Phylogeny
Popul...
family Correlat. related to allergy
MATH models
Cloud comp
Symb comp
Big DATA
Stat model

Mr.SymBioMath



Dissemination

Software Applications

Grid + Cloud & HPC Infrastructure

Soft. Clients

Interfaces, visualization & Data analysis

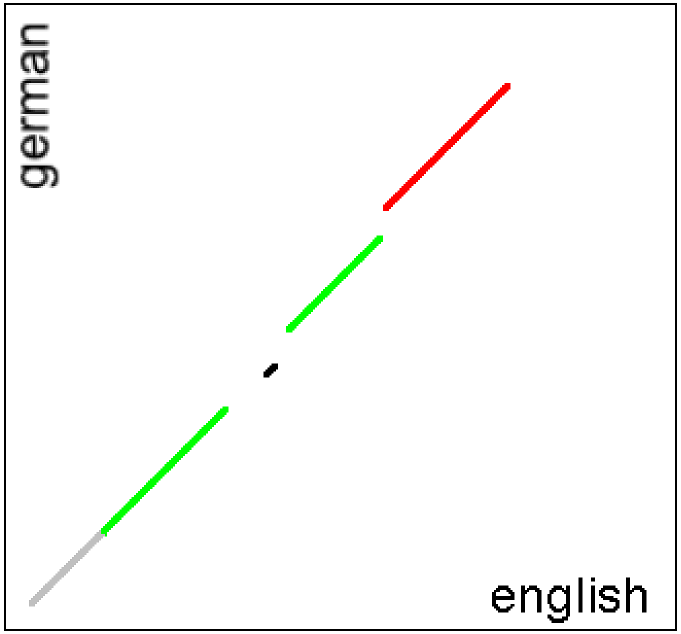
Data collections

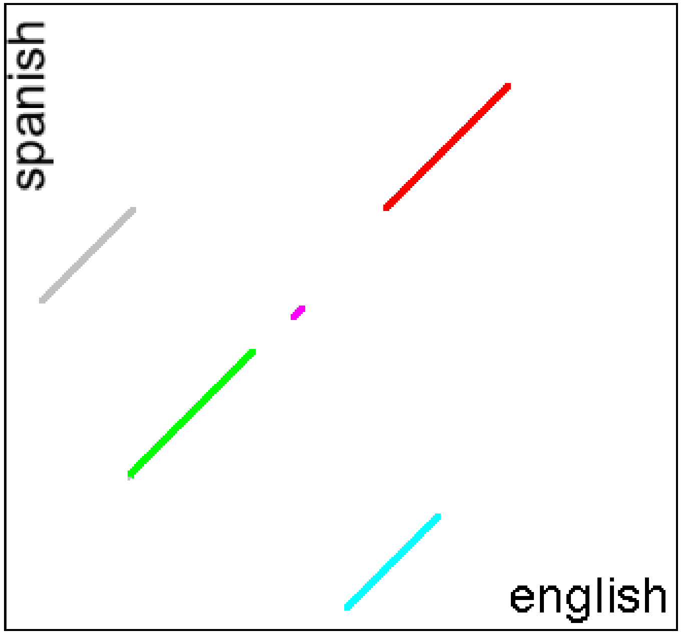
Models and Base soft: libraries

RISC (Cloud)	HCH + UMA	JKU + RISC+ LRZ	LRZ + ITG + ICG
UMA (clients)	LNCC	UMA+ ITG	HCH+ LNCC (users)
IBM (Big Data)	PUBLIC	IBM + ICG+ BAOBAD	

Mr.SymBioMath

```
     THECOMPUTERARCHITECTUREDEPARTMENTAT-THE--UNIVERSITY--OFMALAGA
       ||||||||||||||||.||||          |    |  |||||||||     ||||||
     DIECOMPUTERARCHITEKTUREABTEILUNG-AN--DER-UNIVERSITATVONMALAGA
                ||..|||.|||               ||  |||||||| |    ||||||
ELDEPARTAMENTODEARQUITECTURADECOMPUTADORESDELAUNIVERSIDAD-DEMALAGA
```

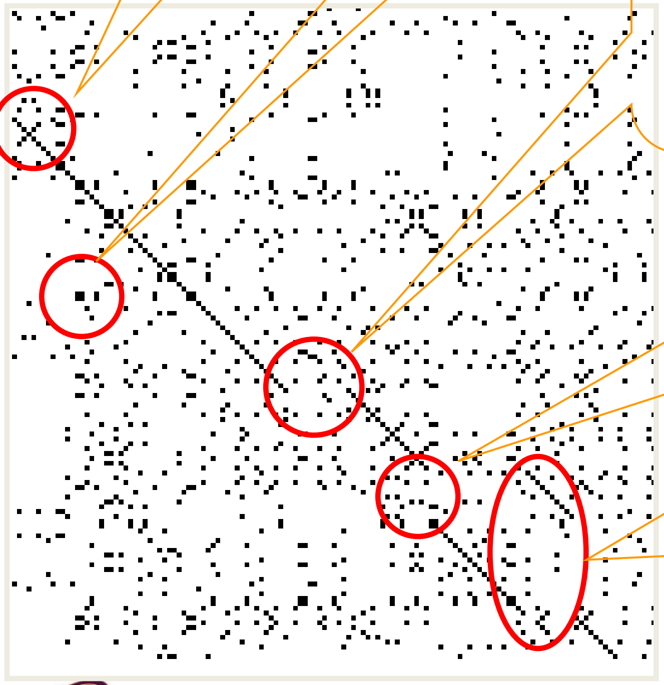




Visualization & Interpretation

Dortplots for DNA sequences can be noisy since there are only 4 symbols (each symbol in one sequence will match qith the 25% of the symbols in the other sequence

To avoid noise, instead of compare pair of symbols an sliding windows is used, and a minimal threshold or stringency level is used to assign a real match (e.g W=10, T=6).



Palindromes Inverted diagonals

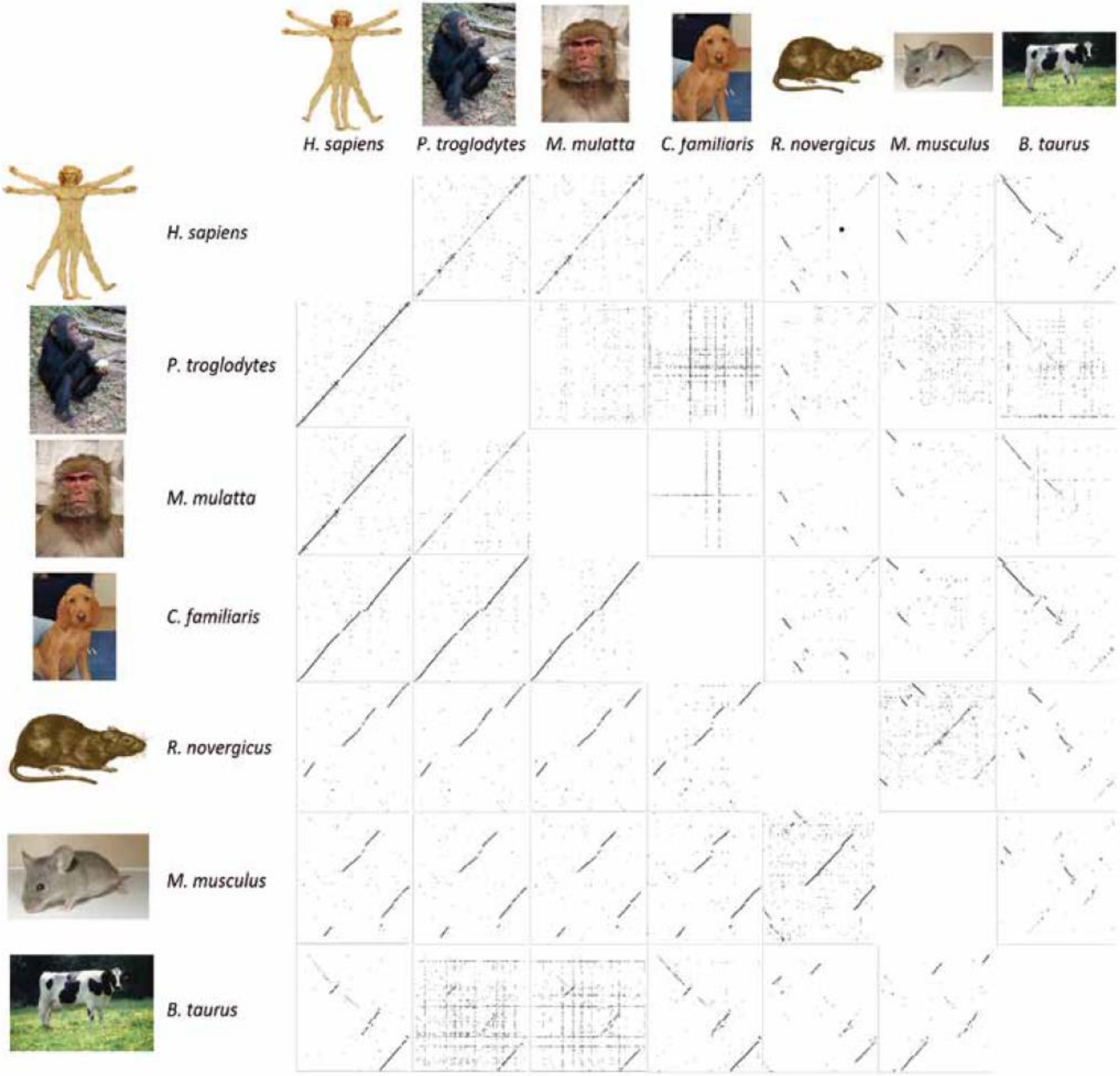
Low complexity zones (repetitions of the same symbol in both sequences

Residue deletion in the vertical sequence or insertion in the horizontal

Residue deletion in the horizontal or insertion in the vertical sequences

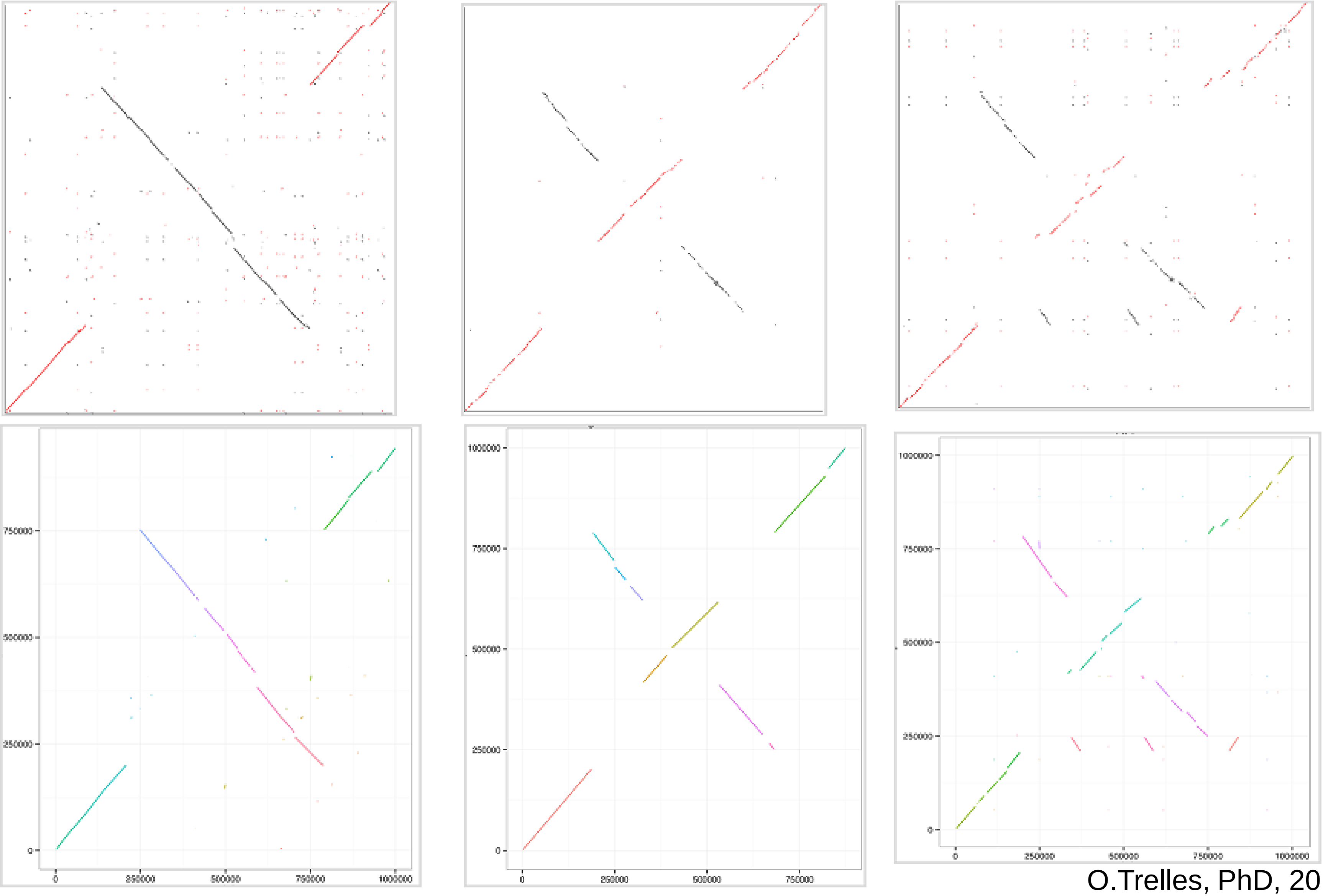
Repetitions: a zone in the horizontal sequence is similar to more than one in the vertical

O.Trelles, PhD, 2014


H. sapiens
P. troglodytes
M. mulatta
C. familiaris
R. novergicus
M. musculus
B. taurus
H. sapiens
P. troglodytes
M. mulatta
C. familiaris
R. novergicus
M. musculus
B. taurus

Comparative Genomics

Detection (& sequence) of Evolution Events



O.Trelles, PhD, 2014

UNIVERSIDAD DE MÁLAGA

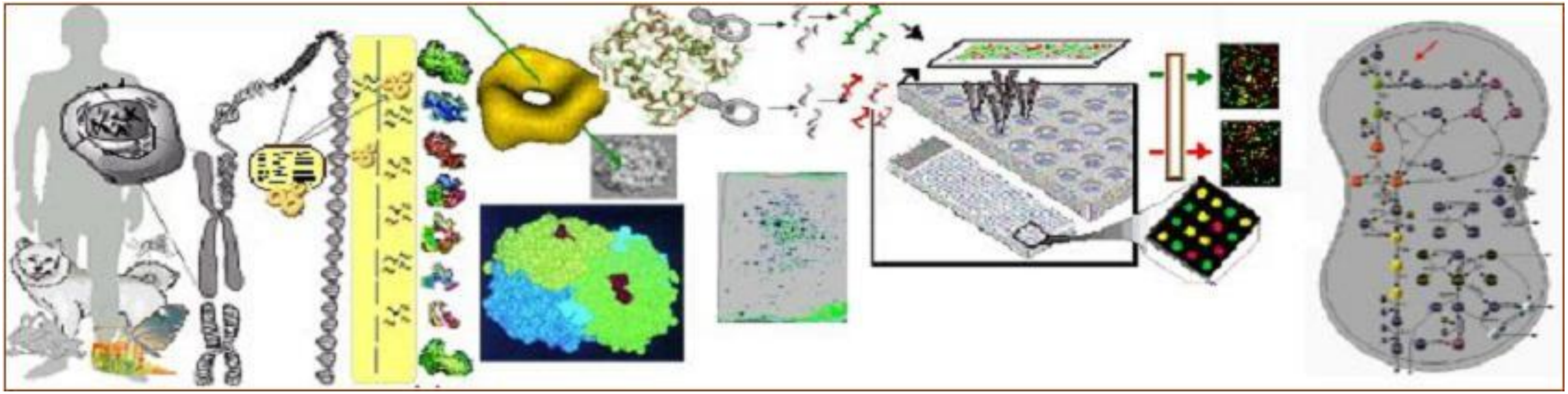
Introduction

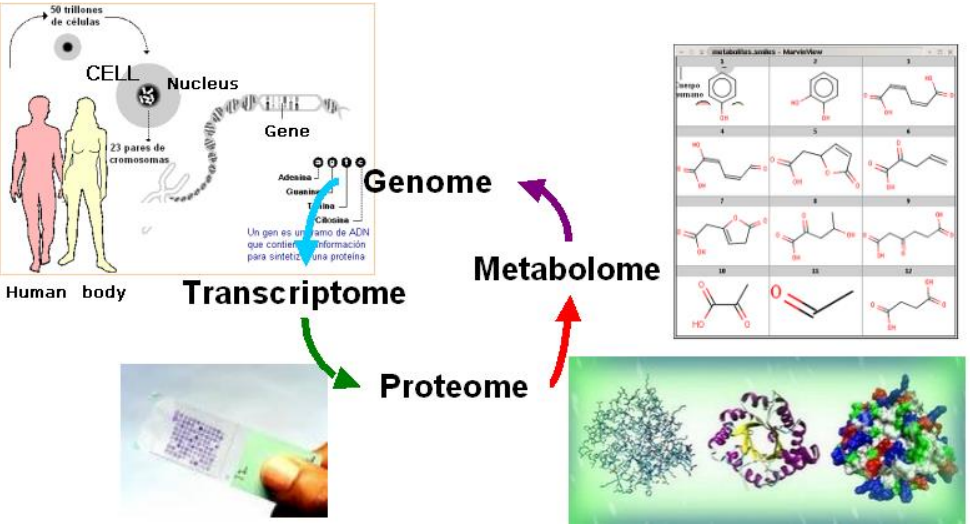
Survey on biology and bioinformatics

PRACE

O.Trelles, PhD, 2014

From genes to pathways



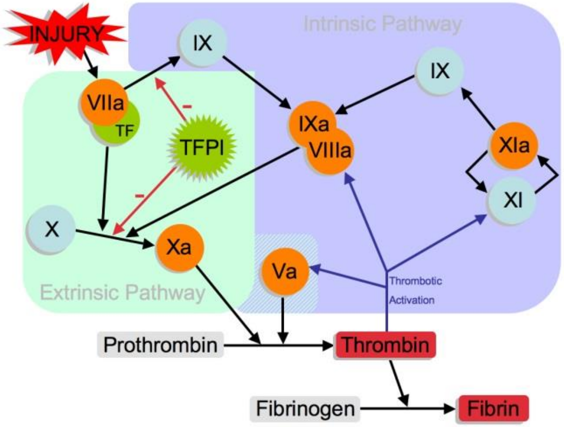

50 trillones de células
CELL
Nucleus
Gene
23 pares de cromosomas
Adenina
Guanina
Timina
Citosina
Un gen es un tramo de ADN que contiene información para sintetizar una proteína
Human body
metabolites.smiles - MarvinView

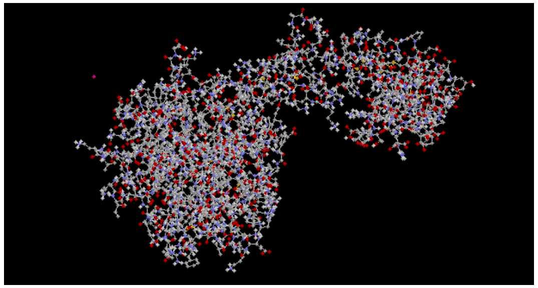
Genome

Transcriptome

Proteome

Metabolome

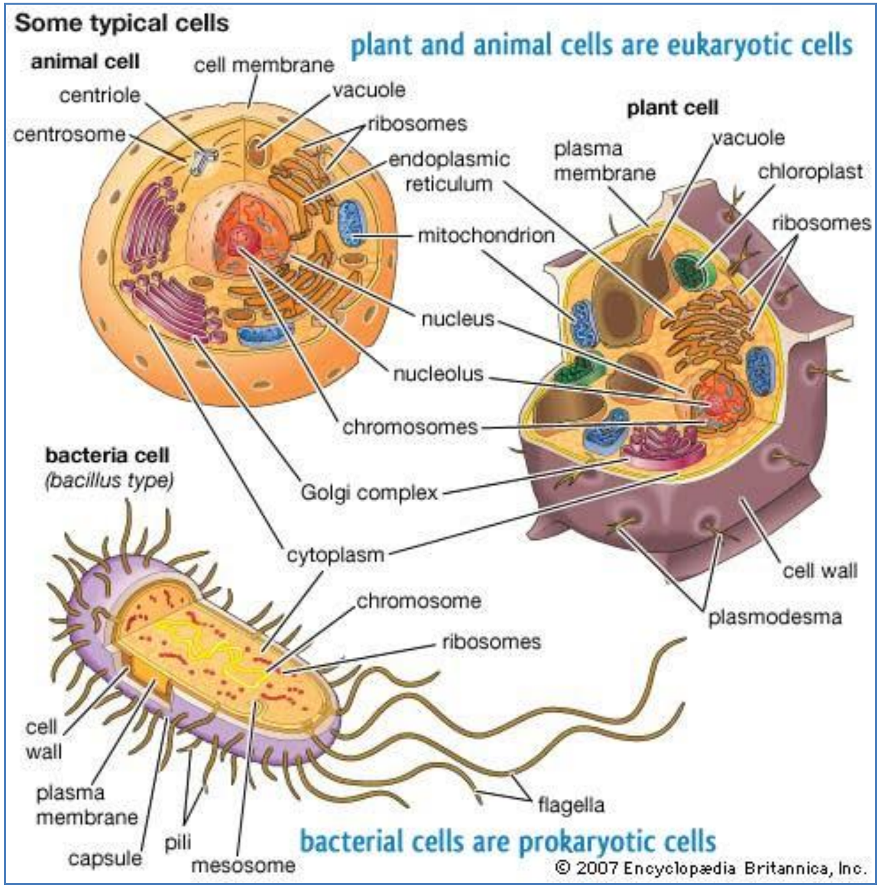




Cells and organisms

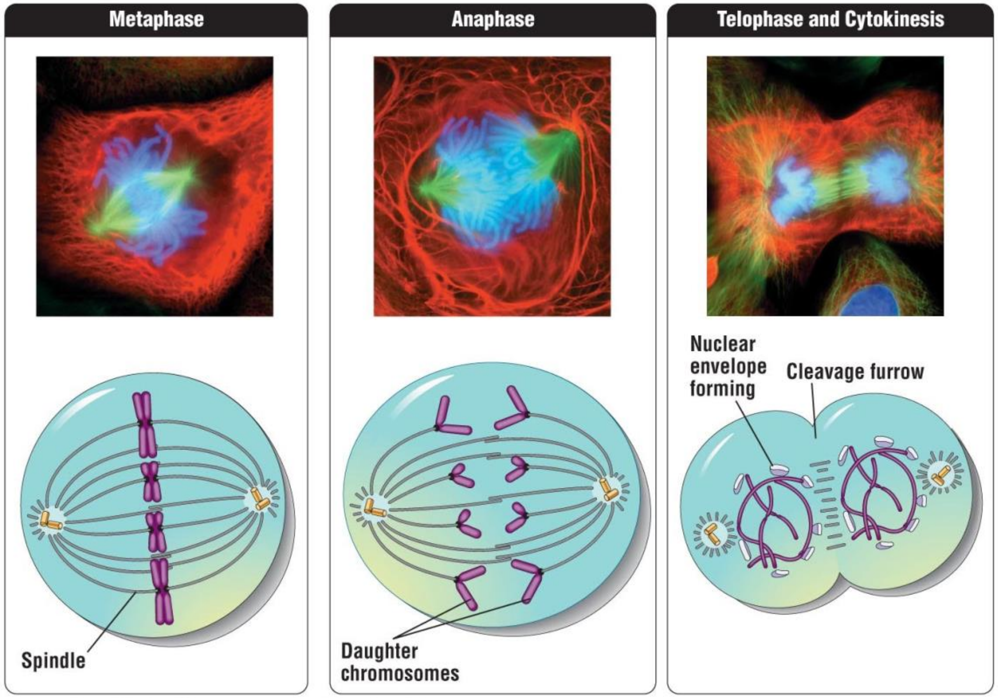
All living things are made of cells.

Prokaryotic & Eukaryotic



© 2007 Encyclopædia Britannica, Inc.

Cells reproduction



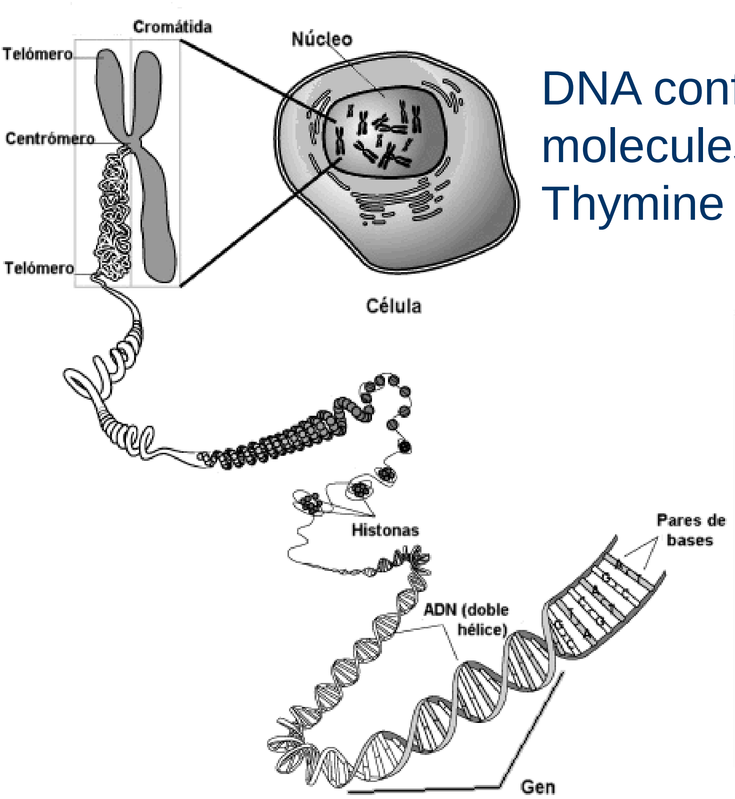
Cell reproduction is the process of a cell splitting and becoming two similar cells.

Prokaryotes by **binary fission**

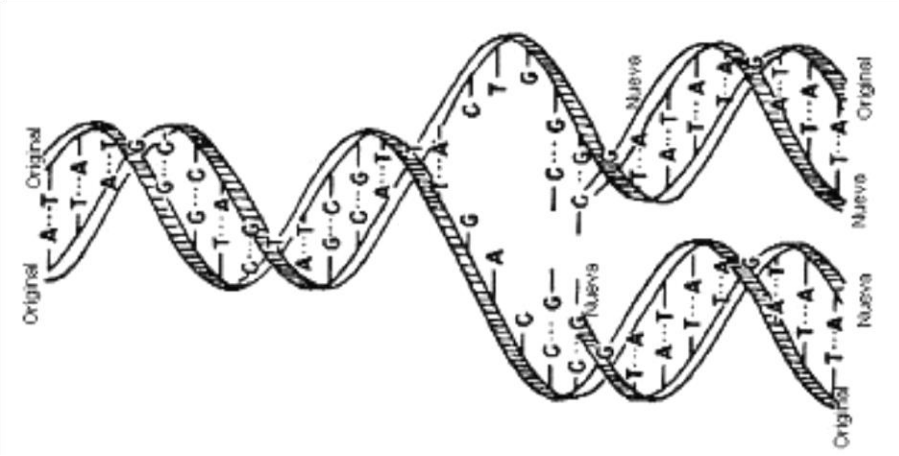
Eukaryotic cells reproduce using either **mitosis** (2) or **meiosis**. (4)

daughter cells have identical genetic composition, except for spontaneous **mutations**.

The DNA carries the hereditary information

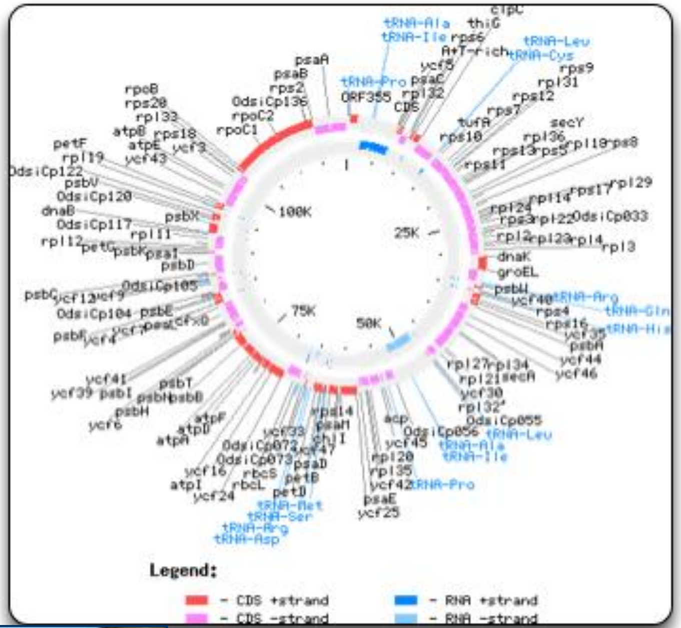


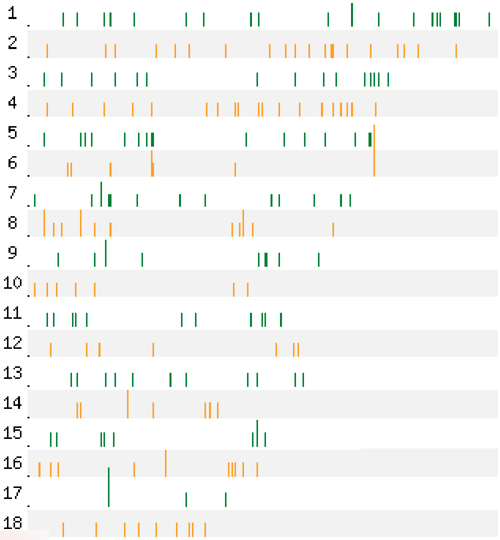
DNA conforms a long linear polymer using 4 different molecules or monomers: Adenine, Cytosine, Guanine y Thymine (A, C, G and T) also called nucleotides or bases.

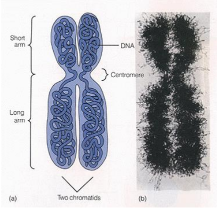


Chromosomes and Genes

DNA is organized in chromosomes
Genes carry out the instructions to synthetize proteins







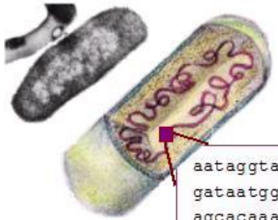
Mouse genome: 20 pairs of chromosomes

Circular chromosome of “*Odeontella*” with 119,704 base pairs / 174 genes

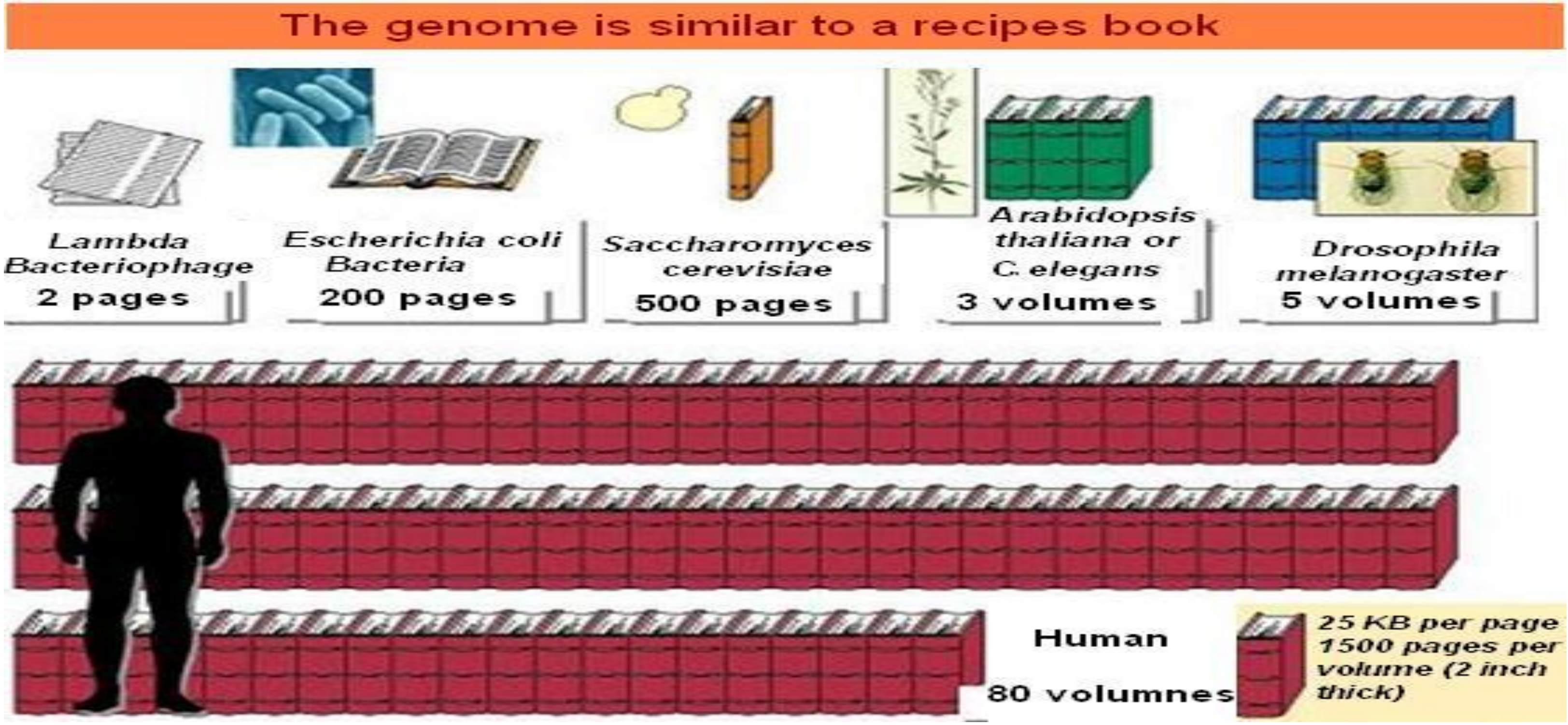
source: http://chloroplast.ocean.washington.edu/chloroplast_files/images/odontella_genome.png

Genome size

The genome is replicated in each cell
Size: from few thousands of bases in bacteria (viruses?)
To about some GBp (basepairs)

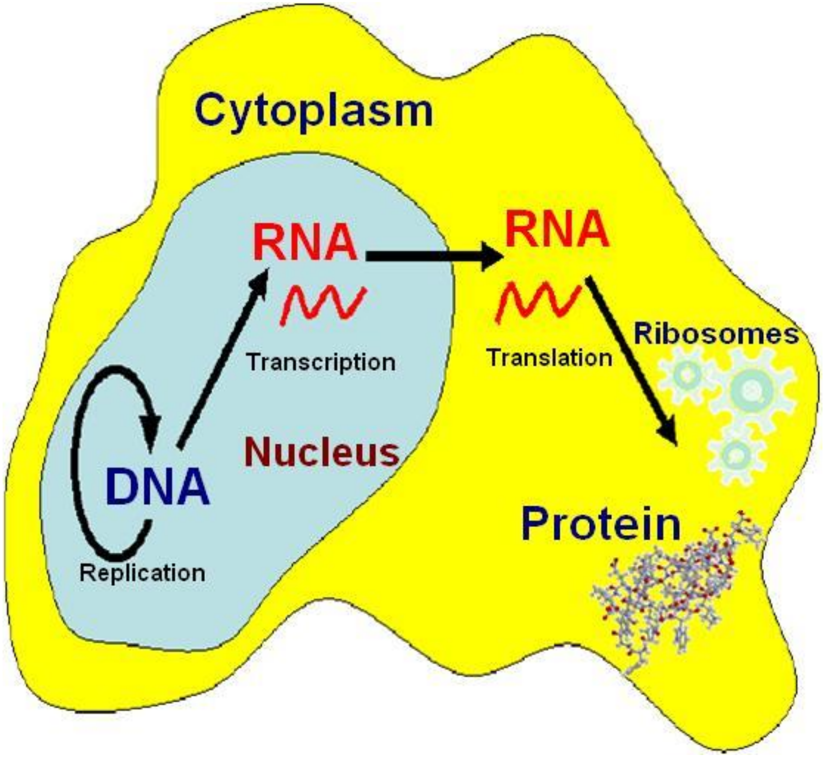


```
aataggtaaatctacaacaacacaaaacttaacatcagggcttgctacaatggacaagaa
gataatggtagtaggatgtgatcctaaggctgactcaacaaggttattactaggaggact
agcacaaaaaagtgttcttgatacattaagagaagaaggagatgacgtagatttagattc
aatcttaaagccaggatttagaggtataaaatgtgttgaatcaggcggtccagaaccagg
agttggatgtgcaggaagaggtataataacttcaatcaatatgctagagcaattaggtgc
ttacgaatcagatttagattatgttttctatgatgtattaggt
```



O.Trelles, PhD, 2014

Central Dogma of molecular biology

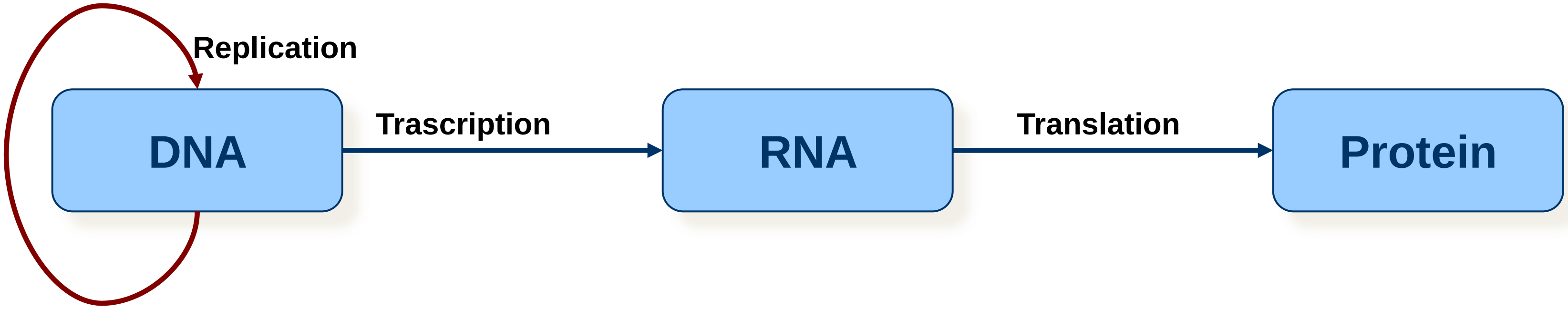


The **central dogma** of molecular biology states:

(1) DNA carries the genetic information of organisms and **replicates** during cell division to allow each daughter cell to contain a full complement of chromosomes.

(2) The genetic information in the DNA is used in a process called **transcription** to produce a complementary one-strand messenger of mRNA

(3) mRNA is interpreted (**translation**) in the ribosomes using the genetic-code to produce a protein.



Replication

DNA

Trascription

RNA

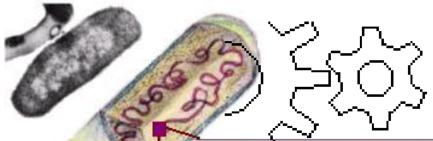
Translation

Protein

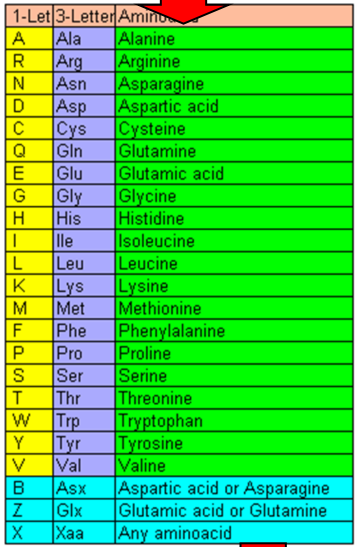
O.Trelles, PhD, 2014

from Genes to Proteins

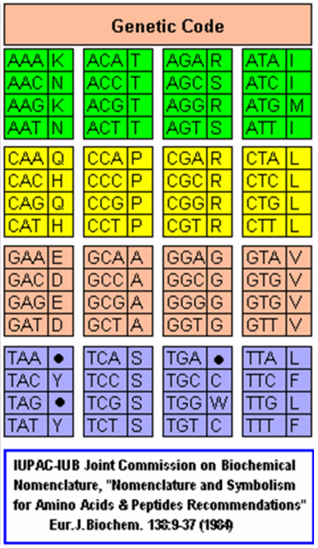
The Genetic Code



```
aataggtaaatctacaacaacacaaaacttaacatcagggcttgctacaatggacaagaa
gataatggtagtaggatgtgatcctaaggctgactcaacaaggttattactaggaggact
agcacaaaaaagtgttcttgatacattaagagaagaaggagatgacgtagatttagattc
aatcttaaagccaggatttagaggtataaaatgtgttgaatcaggcggtccagaaccagg
agttggatgtgcaggaagaggtataataacttcaatcaatatgctagagcaattaggtgc
ttacgaatcagatt[illegible]ttatgttttctatgatgtattaggt
```



1-Let	3-Letter	Amino[illegible]
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamic acid or Glutamine
X	Xaa	Any aminoacid



Genetic Code							
AAA	K	ACA	T	AGA	R	ATA	I
AAC	N	ACC	T	AGC	S	ATC	I
AAG	K	ACG	T	AGG	R	ATG	M
AAT	N	ACT	T	AGT	S	ATT	I
CAA	Q	CCA	P	CGA	R	CTA	L
CAC	H	CCC	P	CGC	R	CTC	L
CAG	Q	CCG	P	CGG	R	CTG	L
CAT	H	CCT	P	CGT	R	CTT	L
GAA	E	GCA	A	GGA	G	GTA	V
GAC	D	GCC	A	GGC	G	GTG	V
GAG	E	GCG	A	GGG	G	GTG	V
GAT	D	GCT	A	GGT	G	GTT	V
TAA	●	TCA	S	TGA	●	TTA	L
TAC	Y	TCC	S	TGC	C	TTC	F
TAG	●	TCG	S	TGG	W	TTG	L
TAT	Y	TCT	S	TGT	C	TTT	F

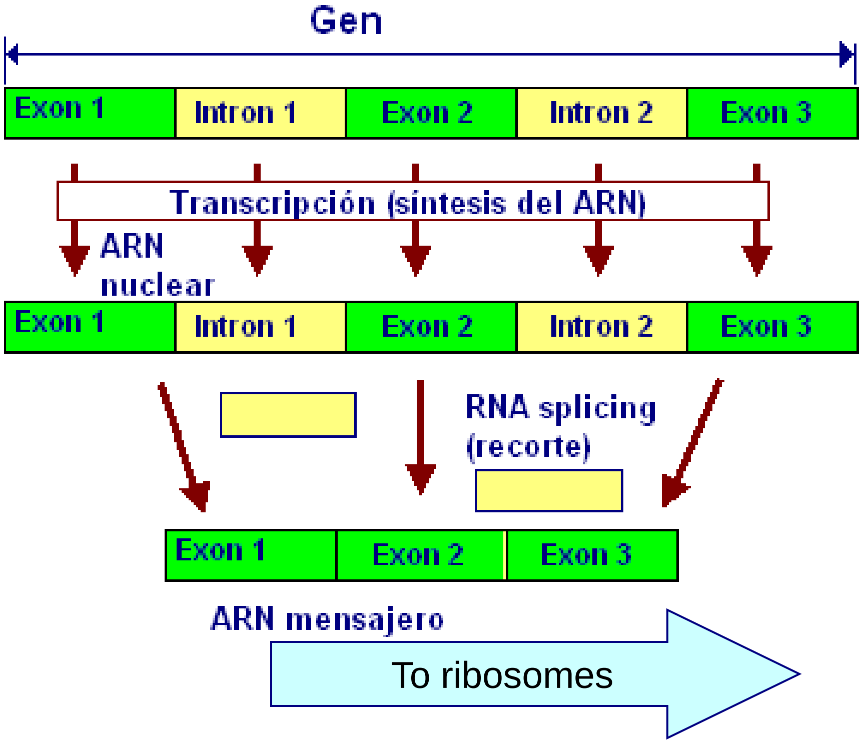
IUPAC-IUB Joint Commission on Biochemical Nomenclature, "Nomenclature and Symbolism for Amino Acids & Peptides Recommendations" Eur.J.Biochem. 138:9-37 (1984)

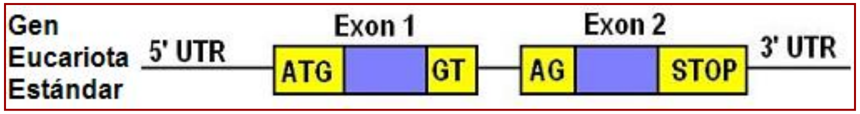
Genes contains the instructions for protein synthesis. That instructions are translated by the cellular machinery using the so called genetic code that translate each consecutive codon (DNA triple) into an specific amino acid

- **Codon**: 3 consecutive bases of DNA
- There are 6 (putative) different ways to read the DNA
- **ORF**: the frame o DNA with not stop codons

Some details….

(*i.e.* eukaryotic genes)





Protein synthesis start with a copy of one of the DNA strands into RNA inside the cellular nucleus. This RNA is spliced to remove the introns (mature mRNA).

Small signals for starting (donors) of introns and exons and ending points (acceptors) are used to identify the right cutting position, including the stop signals for ending the translation.

Differences between pro and eukaryotic cells

Proteins

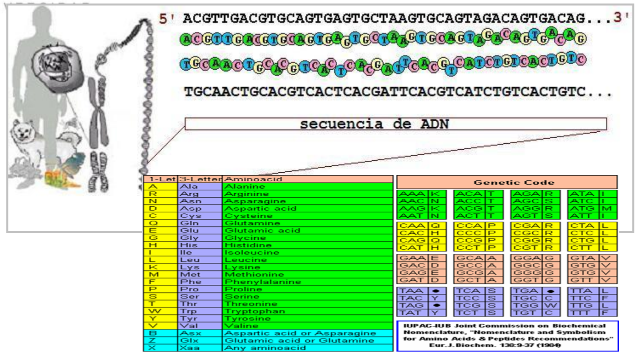
DNA: instructions to drive the synthesis of proteins

Proteins: organisms are made of proteins (bones, muscles, nervous…

Protein function is associated to its 3D spatial conformation

DNA is present in all cells

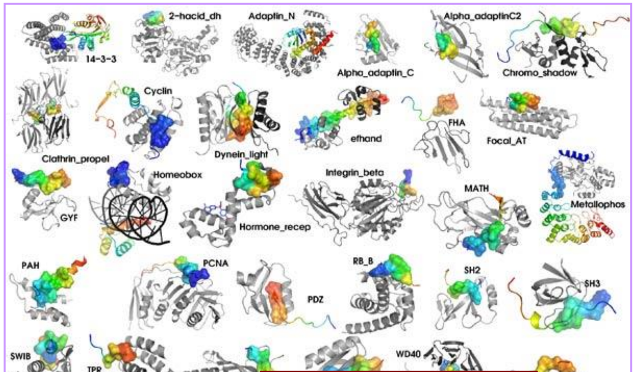
Proteins: each cell produces only those proteins the cell needs

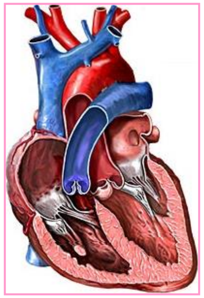

5' ACGTTGACGTGCAGTGAGTGCTAAGTGCAGTAGACAGTGACAG...3'
TGCAACTGCACGTCACTCACGATTCACGTCATCTGTCACTGTC...
secuencia de ADN

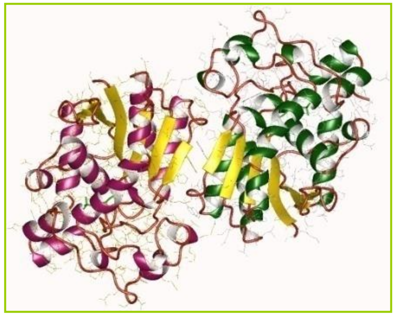
1-Let	3-Letter	Aminoacid
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamic acid or Glutamine
X	Xaa	Any aminoacid

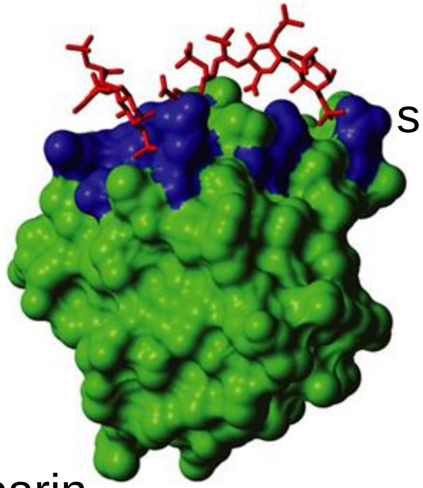
Genetic Code							
AAA	K	ACA	T	AGA	R	ATA	I
AAC	N	ACC	T	AGC	S	ATC	I
AAG	K	ACG	T	AGG	R	ATG	M
AAT	N	ACT	T	AGT	S	ATT	I
CAA	Q	CCA	P	CGA	R	CTA	L
CAC	H	CCC	P	CGC	R	CTC	L
CAG	Q	CCG	P	CGG	R	CTG	L
CAT	H	CCT	P	CGT	R	CTT	L
GAA	E	GCA	A	GGA	G	GTA	V
GAC	D	GCC	A	GGC	G	GTG	V
GAG	E	GCG	A	GGG	G	GTG	V
GAT	D	GCT	A	GGT	G	GTT	V
TAA	●	TCA	S	TGA	●	TTA	L
TAC	Y	TCC	S	TGC	C	TTC	F
TAG	●	TCG	S	TGG	W	TTG	L
TAT	Y	TCT	S	TGT	C	TTT	F

IUPAC-IUB Joint Commission on Biochemical Nomenclature, "Nomenclature and Symbolism for Amino Acids & Peptides Recommendations" Eur.J.Biochem. 138:9-37 (1984)





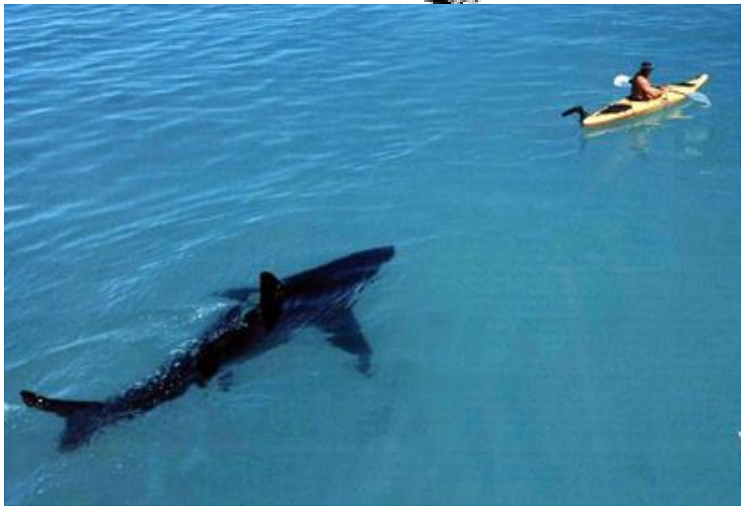




O.Trelles, PhD, 2014

Proteins levels

Throught Gene-Expression



Levels of proteins ⟷ Cellular state

Gene levels ≈ Protein levels

Env. stimuly ⟷ Change proteins levels

Change proteins levels ⟵ Change gene levels

⟶ Gene regulation mchanisms:

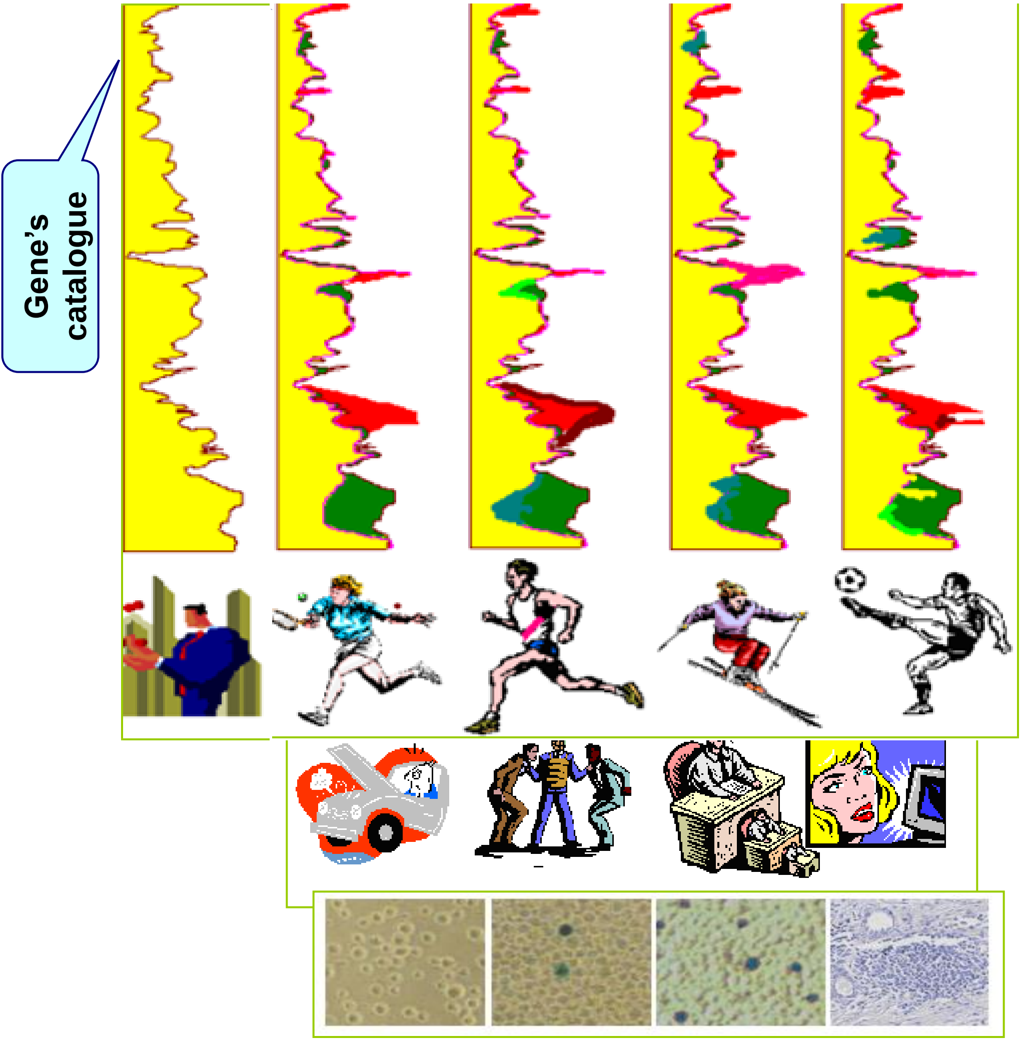
Changes in protein levels have profund effect in the biology of the organisms (even with phisiological and pathologic effects)

Gene-expression levels are used to determine the response of an organism to a particular event

Different developmental stage, tissues types, clinical conditions, organisms, etc

Gene Expression

Quantify the level at which a particular gene is expressed



GE levels are used to determine the response of an organism to a particular event

Simultaneosus anaysis of thousands of gene

Different developmental stage, tissues types, clinical conditions, organisms, etc

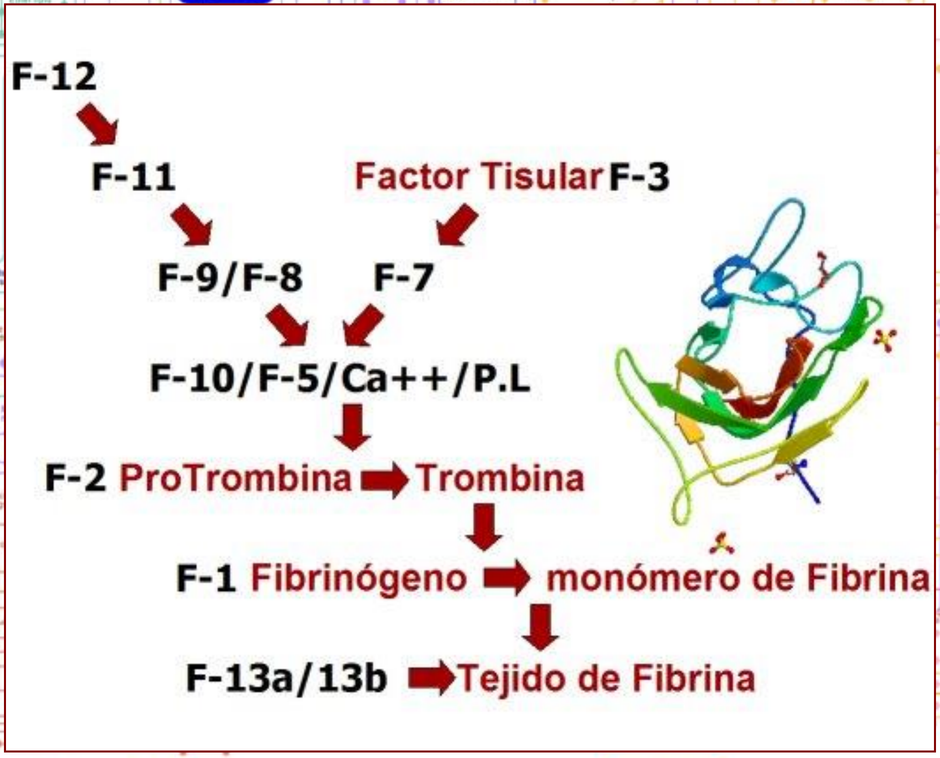
Metabolic Pathways

Series of chemical reactions ocurring within the cell

Enzymes: catalyze the reactions

Metabolite: intermediate and final product of metabolism

Metabolome: network of metabolic reactions



Bioinformatics

Source: ECCC'02 Web site

Featuring the application domain

O.Trelles, PhD, 2014

Bioinformatics (Computational biology)

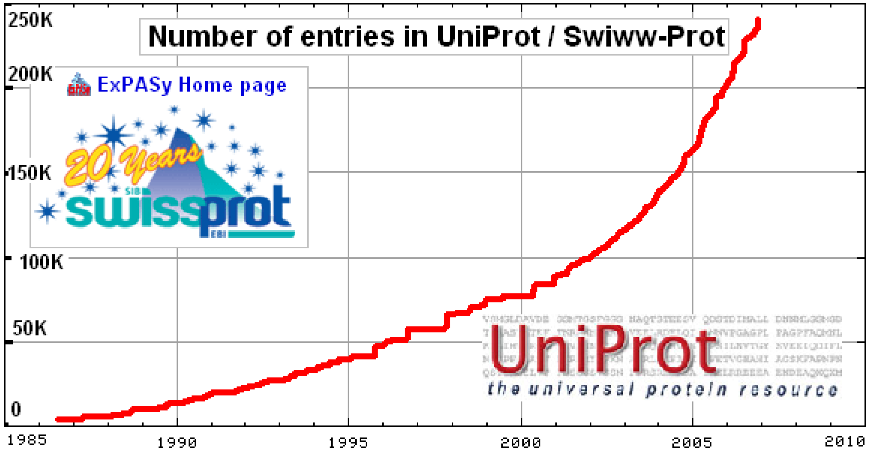
Computer sciences as applied to biological data

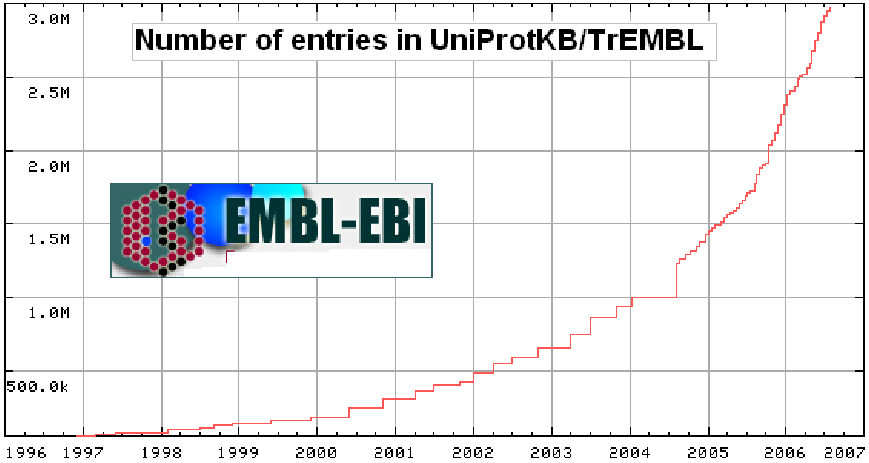
Computer sciences, statistics, physiscs, chemistry, IT, …

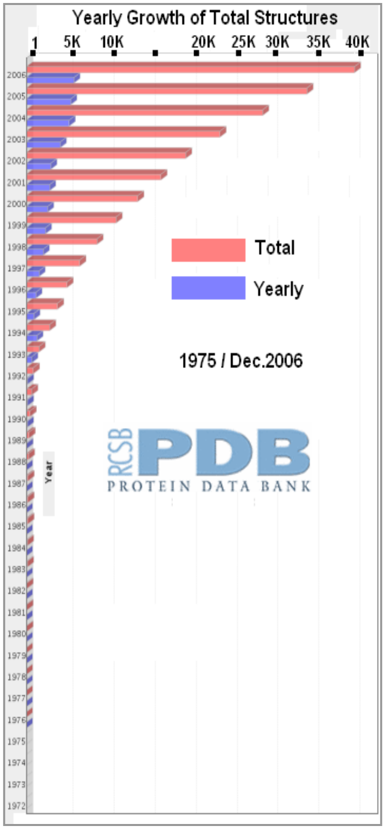
Molecular clínical, imaging, population, environmental, ...

Data production

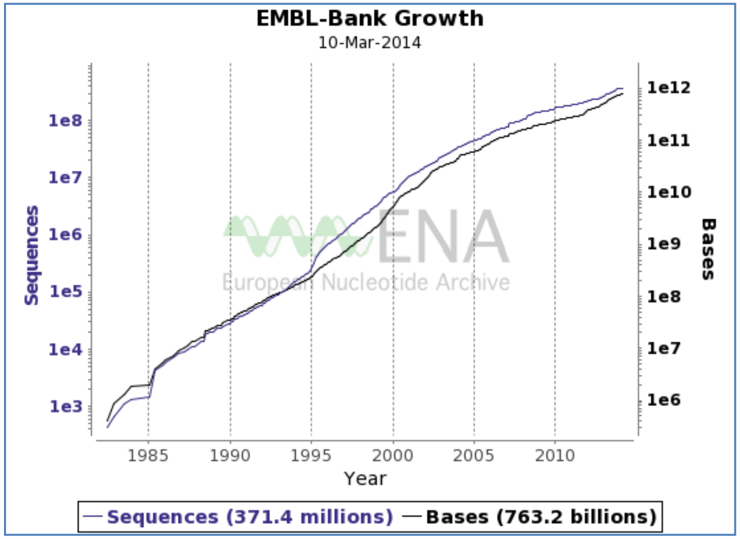
Huge data production at different levels

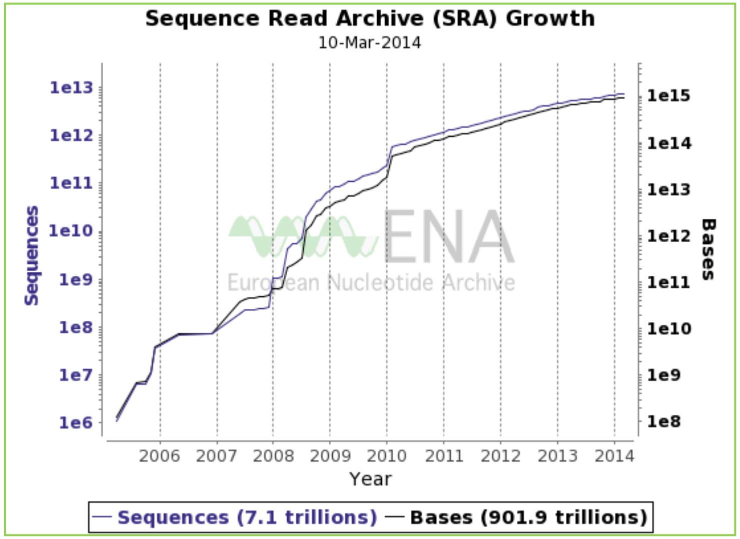






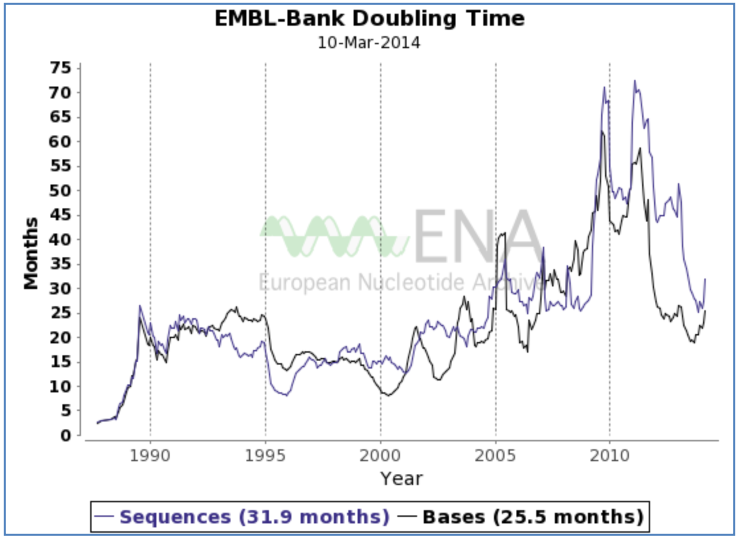
Assembled/annotated sequence growth

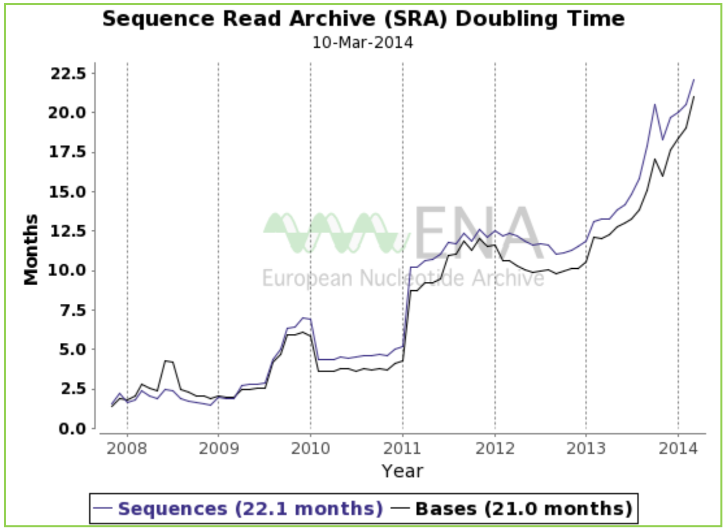




READs growth

Assembled/annotated sequence doubling time

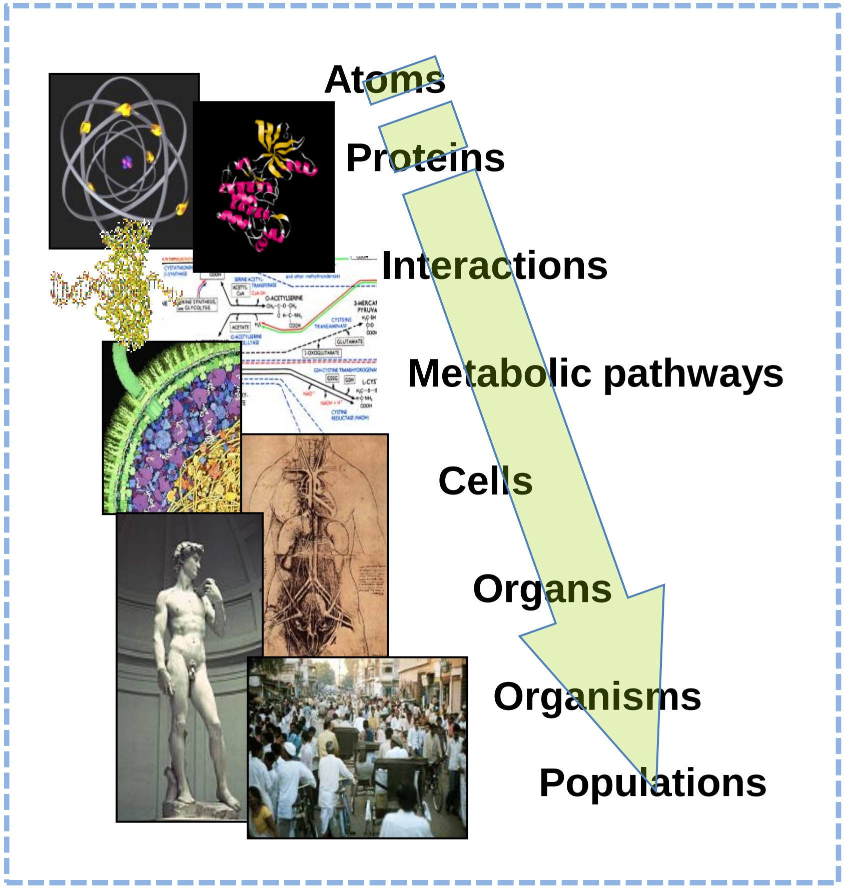




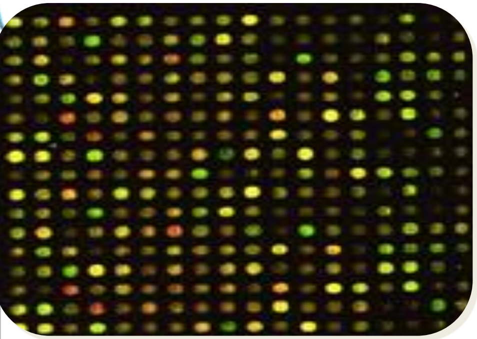
READs doubling time

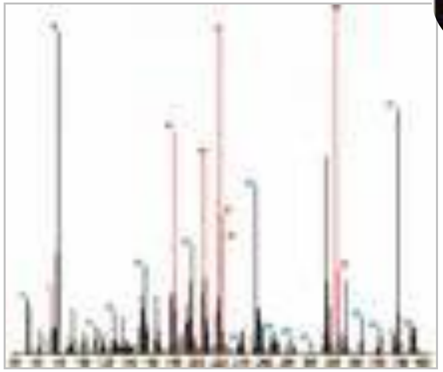
ENA statistics: https://www.ebi.ac.uk/ena/about/statistics

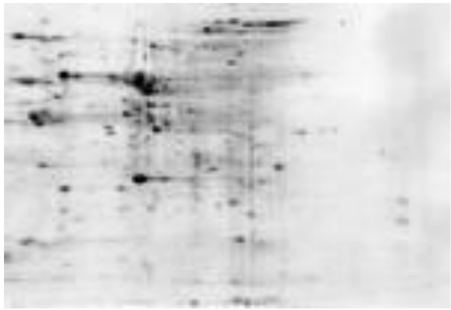
Diverse types of data



```
> E01306 229 bp  DNA linear
gaattctaac ggtcccgaaa ctctgtgcgg tgctgaactg gttgacgctc tgcagtttgt
ttgcggtgac cgtggttttt attttaacaa acccactggt tatggttctt cttctcgtcg
tgctccccag actcgta[illegible]tg ttgacgaatg ctgctttcgt tcttgcgacc tgcgtcgtct
ggaaatgta[illegible] tgcg[illegible]tc[illegible]cc [illegible]gaaacccgc taaatctgct tagaagctt
```







O.Trelles, PhD, 2014

Format heterogeneity

```
LOCUS       E01306      229 bp    DNA     linear   PAT 04-NOV-2005
DEFINITION  DNA encoding human insulin-like growth factor I(IGF-I).
ACCESSION   E01306
VERSION     E01306.1  GI:2169565
KEYWORDS    JP 1987190088-A/1.
SOURCE      synthetic construct
  ORGANISM  synthetic construct
            other sequences; artificial sequences.
REFERENCE   1  (bases 1 to 229)
  AUTHORS   Raasu,A., Toomasu,M., Berun,N. and Majiasu,U.
  TITLE     METHOD FOR TRANSPORTING GENE PRODUCT TO MEDIUM PROPAGATING GRAM
            NEGATIVE BACTERIA
  JOURNAL   Patent: JP 1987190088-A 1 20-AUG-1987;
            KABIGEN AB
COMMENT     OS   Artificial gene
            OC   Artificial sequence; Genes.
            OS   Homo sapiens
            PN   JP 1987190088-A/1
            PD   20-AUG-1987
            CC   strandedness: Single;
            CC   topology: Linear;
            CC   hypothetical: No;
            CC   anti-sense: No;
            FH   Key     Location/Qualifiers
            FT           /product='human insuline-Like growth factor I
            FT   CDS              >2..223
FEATURES             Location/Qualifiers
     source          1..229
                     /organism="synthetic construct"
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:32630"
ORIGIN
        1 gaattctaac ggtcccgaaa ctctgtgcgg tgctgaactg gttgacgctc tgcagtttgt
       61 ttgcggtgac cgtggttttt attttaacaa acccactggt tatggttctt cttctcgtcg
      121 tgctccccag actggtattg ttgacgaatg ctgctttcgt tcttgcgacc tgcgtcgtct
      181 ggaaatgtat tgcgctcccc tgaaacccgc taaatctgct tagaagctt
//
```

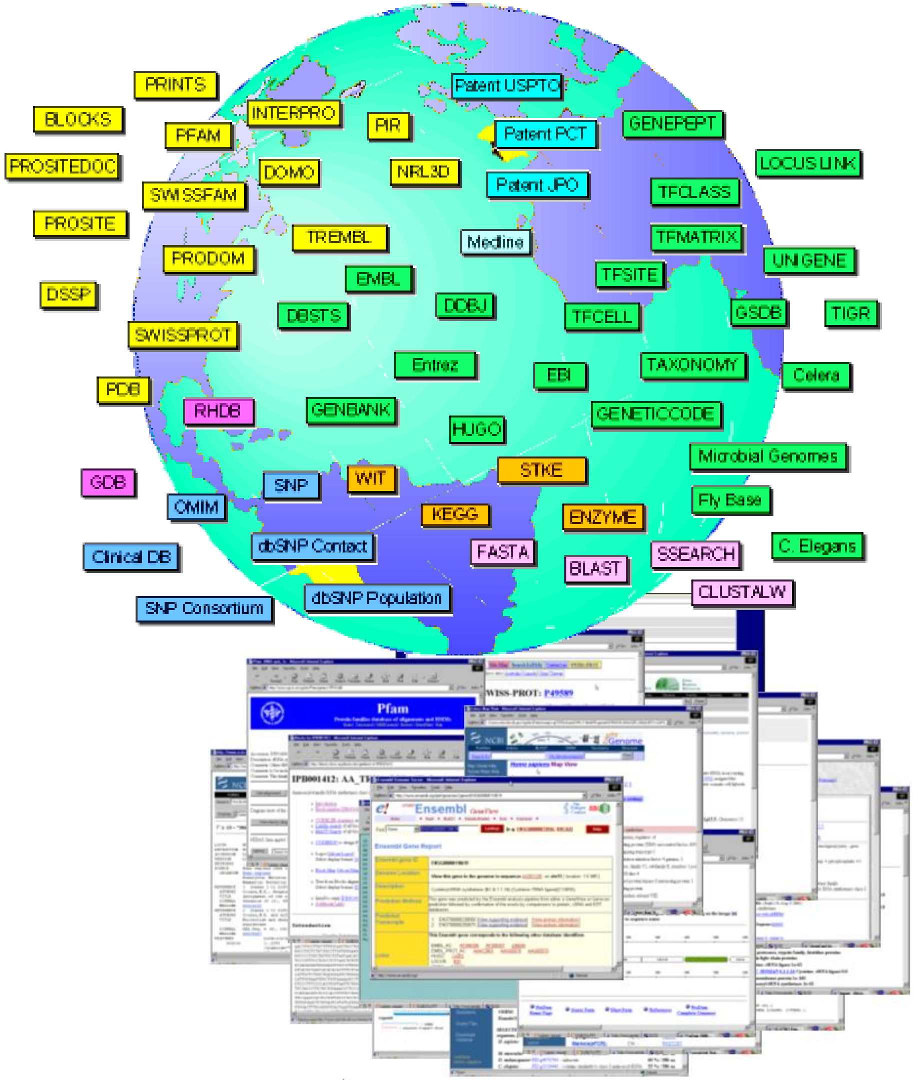
```
ID   E01306; SV 1; linear; unassigned DNA; PAT; SYN; 229 BP.
AC   E01306;
DT   07-OCT-1997 (Rel. 52, Created)
DT   09-NOV-2005 (Rel. 85, Last updated, Version 3)
DE   DNA encoding human insulin-like growth factor I(IGF-I).
KW   JP 1987190088-A/1.
OS   synthetic construct
OC   other sequences; artificial sequences.
RA   Raasu A., Toomasu M., Berun N., Majiasu U.;
RT   "METHOD FOR TRANSPORTING GENE PRODUCT TO MEDIUM PROPAGATING GRAM
RT   NEGATIVE BACTERIA";
RL   Patent number JP1987190088-A/1, 20-AUG-1987.
RL   KABIGEN AB.
CC   OS   Artificial gene
CC   OC   Artificial sequence; Genes.
CC   OS   Homo sapiens
CC   CC   strandedness: Single;
CC   CC   topology: Linear;
CC   CC   hypothetical: No;
CC   CC   anti-sense: No;
CC   FH   Key         Location/Qualifiers
CC   FT   mat_peptide      11..220
CC   FT   CDS         >2..223
CC   FT               /product="human insulin-like growth factor I"
FH   Key              Location/Qualifiers
FT   source           1..229
FT                    /organism="synthetic construct"
FT                    /mol_type="unassigned DNA"
FT                    /db_xref="taxon:32630"
SQ   Sequence 229 BP; 40 A; 57 C; 55 G; 77 T; 0 other;
     gaattctaac ggtcccgaaa ctctgtgcgg tgctgaactg gttgacgctc tgcagtttgt        60
     ttgcggtgac cgtggttttt attttaacaa acccactggt tatggttctt cttctcgtcg       120
     tgctccccag actggtattg ttgacgaatg ctgctttcgt tcttgcgacc tgcgtcgtct       180
     ggaaatgtat tgcgctcccc tgaaacccgc taaatctgct tagaagctt                   229
//
```

The DNA encoding human insulin-like growth factor I(IGF-I) available at GenBank: E01306.1 http://www.ncbi.nlm.nih.gov/

The same insulin (E01306) sequence at EBI www.ebi.ac.uk

(in both text-boxes some lines has been removed)

Dispersion of data sources

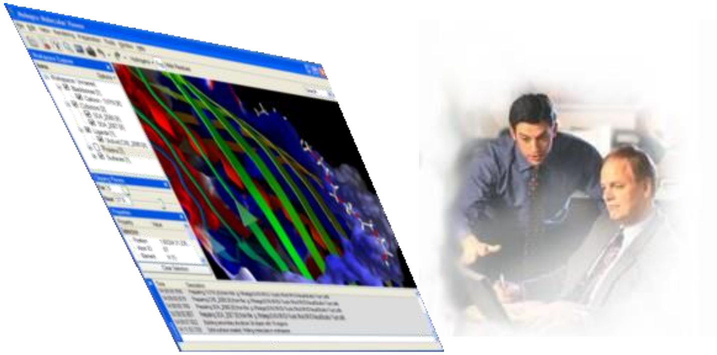


More than 1000 biological DB collections

Workflows: the usual way to work

See: [1] Infobiogen: Catalog of DBs:

http://www.infobiogen.fr/services/dbcat

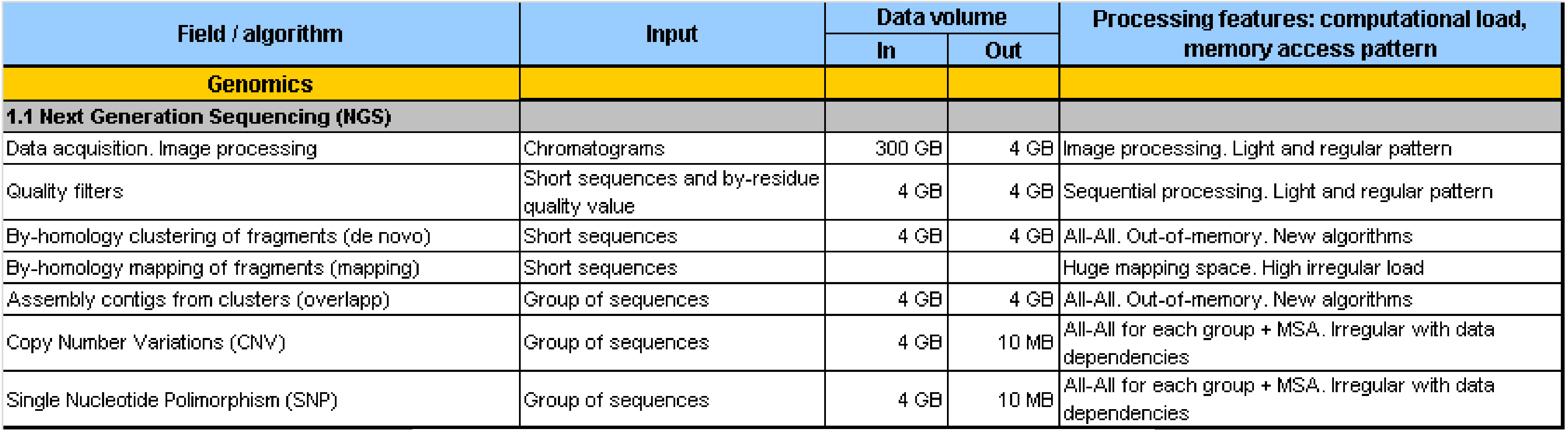


Bioinformatics: a web-based domain

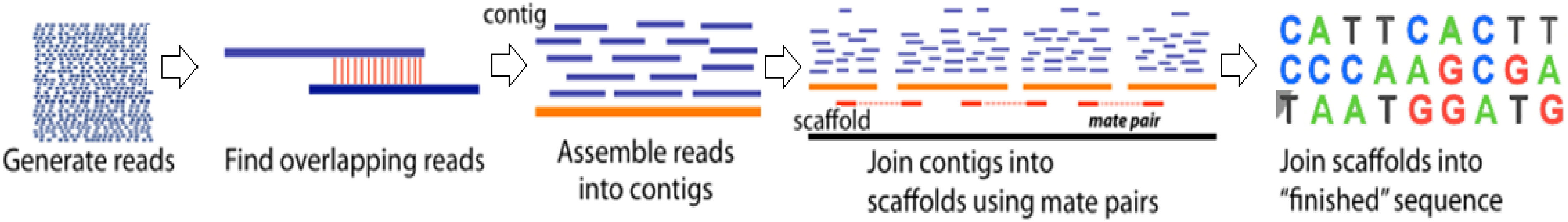
Applications

DNA Sequencing (*n*NGS) & *A*ssembly

(>> 10^9 sequence reads / 36bp to 1kb)



Field / algorithm	Input	Data volume		Processing features: computational load, memory access pattern
		In	Out	
Genomics				
1.1 Next Generation Sequencing (NGS)				
Data acquisition. Image processing	Chromatograms	300 GB	4 GB	Image processing. Light and regular pattern
Quality filters	Short sequences and by-residue quality value	4 GB	4 GB	Sequential processing. Light and regular pattern
By-homology clustering of fragments (de novo)	Short sequences	4 GB	4 GB	All-All. Out-of-memory. New algorithms
By-homology mapping of fragments (mapping)	Short sequences			Huge mapping space. High irregular load
Assembly contigs from clusters (overlapp)	Group of sequences	4 GB	4 GB	All-All. Out-of-memory. New algorithms
Copy Number Variations (CNV)	Group of sequences	4 GB	10 MB	All-All for each group + MSA. Irregular with data dependencies
Single Nucleotide Polimorphism (SNP)	Group of sequences	4 GB	10 MB	All-All for each group + MSA. Irregular with data dependencies



Generate reads

Find overlapping reads

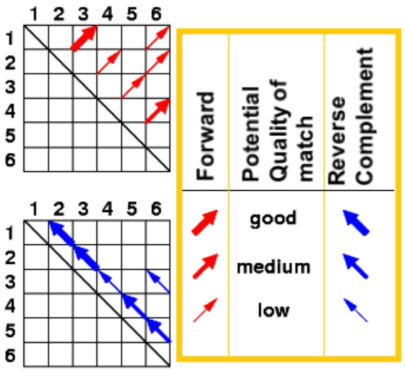
Assemble reads into contigs

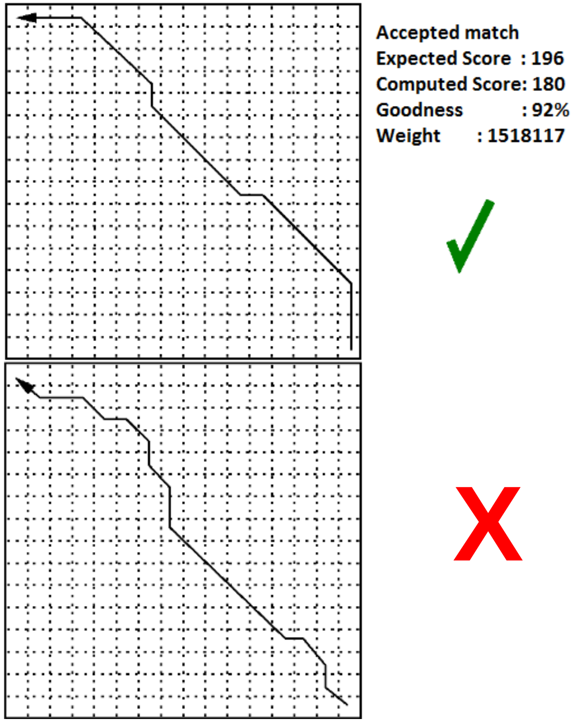
Join contigs into scaffolds using mate pairs

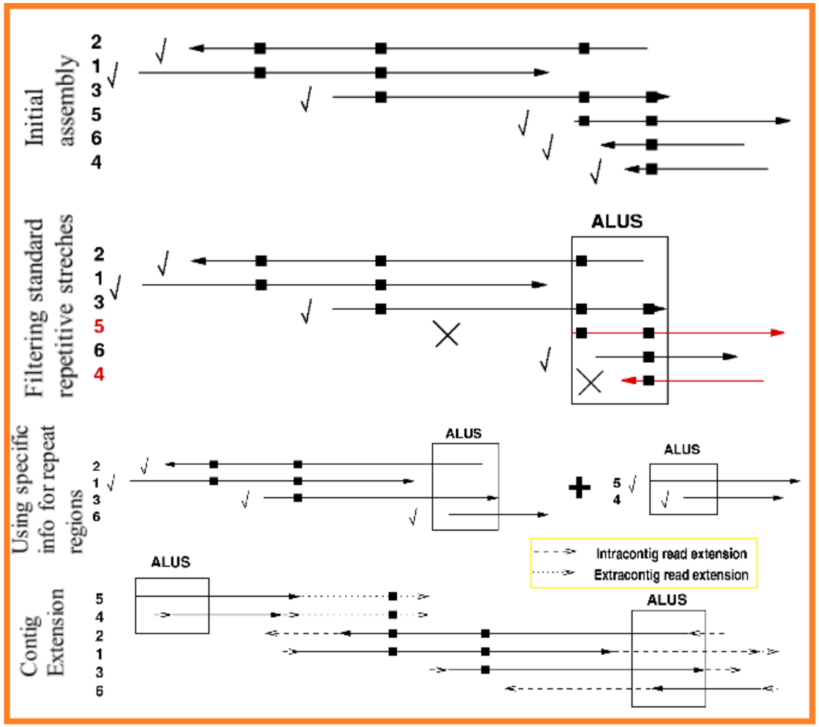
Join scaffolds into "finished" sequence

O.Trelles, PhD, 2014

General concepts for NGS assembly Algorithms (2)




Accepted match
Expected Score : 196
Computed Score: 180
Goodness : 92%
Weight : 1518117


Initial assembly
2
1
3
5
6
4
Filtering standard repetitive streches
ALUS
Using specific info for repeat regions
+
Intracontig read extension
Extracontig read extension
Contig Extension

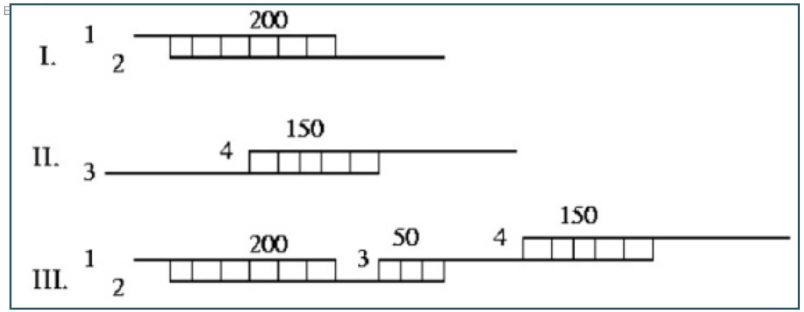
All-vs-All + reversed complement

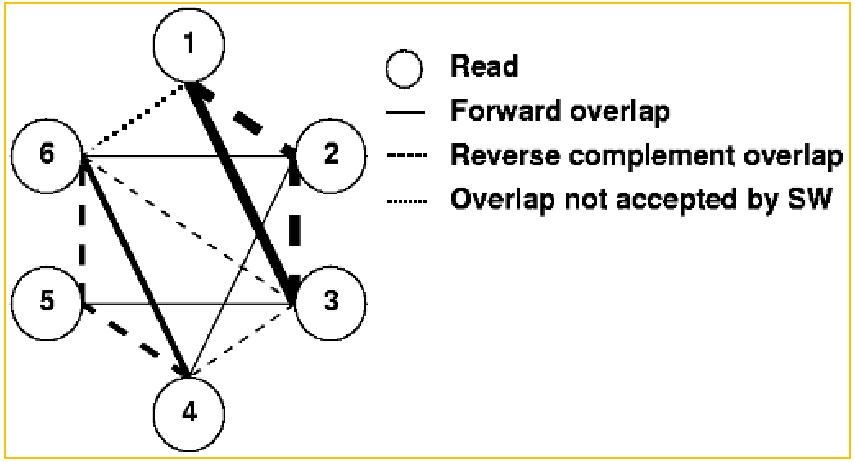
Dynamic Programming bounded Gaps

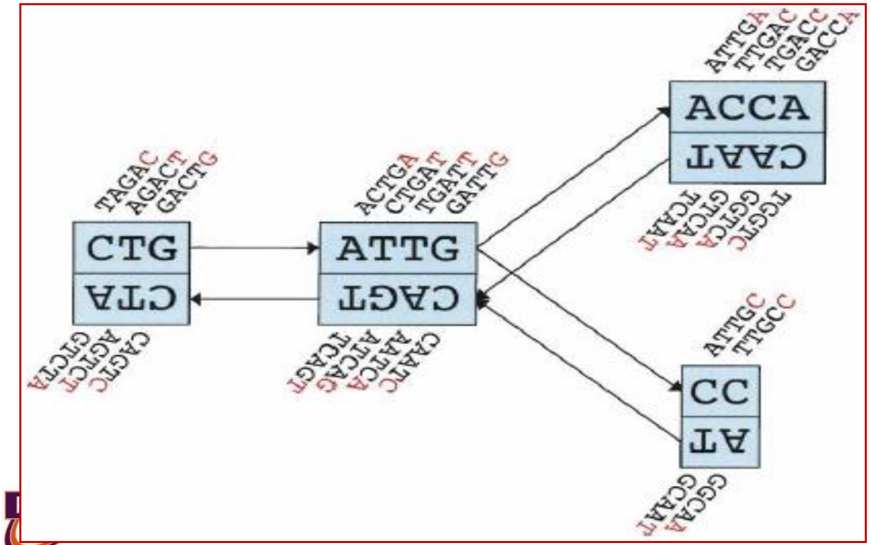
Built-up Contigs and extensions

O.Trelles, PhD, 2014

Assembly Algorithms: Data Management







Greedy assembly: progressive joint of overlapping fragments.

Overlap Layout consensus:
reads are nodes and overlaps are edges. Identify a Hamiltonian path through the graph that contains all the nodes

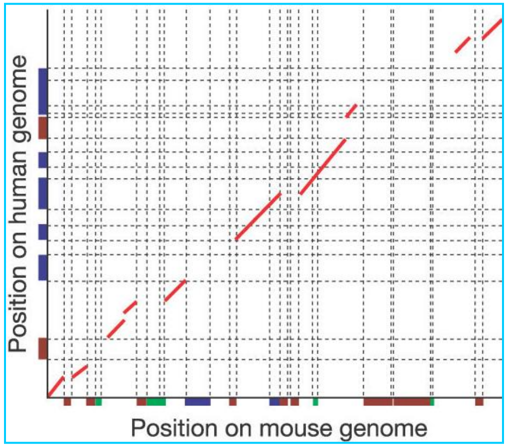
Eulerian path approaches breaks up each read into their overlapping k-mers. Each k-mer is and edge connecting two nodes of its k-1 prefix and suffix respectively. The assembly solution is a path in the graph that uses all the edges - an Eulerian path.

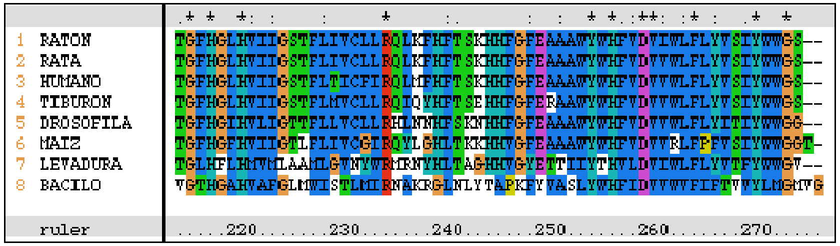
(see also: Bruijin graphs (Velvet)
http://en.wikipedia.org/wiki/De_Bruijn_graph

Align-layout-consensus - Mapping of reads over a related genome (or reference)

Sequence Analysis & Phylogeny

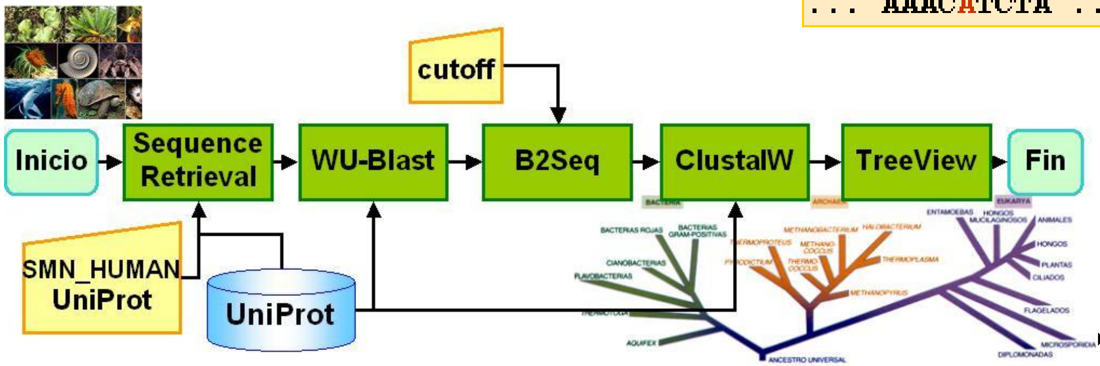
Field / algorithm	Input	Data volume		Processing features: computational load, memory access pattern
		In	Out	
Genomics				
1.2 Sequence analisis and large scale phylogeny				
Gene identification	Large sequences, full genomes	1 GB	10 MB	E/S (local) busquedas intensivas por semejanza, tareas independientes, livianas
Searching by homology (Blast, Fasta, Dyn.Prog.)	Query and Sequences DB	4 GB	1 MB	E/S (local) intensiva, tareas independientes, livianas
Genome scale comparissons (dotplots)	2 Genomes	6 GB	200 MB	Gran demanda de memoria (alg. fuera de memoria)
Comparative genomics	Tens of genomes	30 GB	4 GB	Gran demanda de memoria y de E/S, nuevos algoritmos
Multiple Sequence Alignments (MSA)	Groups of sequences	10 MB	1 MB	Todos vs. Todos + resolución de arbol de alineamiento (irregular, dependencias) y diferentes tipos de tareas
Phylogeny (by parsimony)	Groups of sequences	10 MB	1 MB	Todos vs. Todos + resolución de arbol de alineamiento (irregular, dependencias) y diferentes tipos de tareas
Phylogeny (maximum likelihood)	Groups of sequences	10 MB	1 MB	Patrón irregular y dependencias de datos. Tareas pesadas





```
            .* *  *: :    :     *     :.        :   :   * *.:**: :* :  .*  *
1 RATON     TGFHGLHVIIGSTFLIVCLLRQLKFHFTSKHHFGFEAAAWYWHFVDVIWLFLYVSIYWWGS--
2 RATA      TGFHGLHVIIGSTFLIVCLLRQLKFHFTSKHHFGFEAAAWYWHFVDVVWLFLYVSIYWWGS--
3 HUMANO    TGFHGLHVIIGSTFLTICFIRQLMFHFTSKHHFGFEAAAWYWHFVDVVWLFLYVSIYWWGS--
4 TIBURON   TGFHGLHVIIGSTFLMVCLLRQIQYHFTSEHHFGFERAAWYWHFVDVVWLFLYVSIYWWGS--
5 DROSOFILA TGFHGIHVLIGTTFLLVCLLRHLNNHFSKNHHFGFEAAAWYWHFVDVVWLFLYITIYWWGG--
6 MAIZ      TGFHGFHVIIGTLFLIVCGIRQYLGHLTKKHHVGFEAAAWYWHFVDVVRLFPFVSIYWWGGT-
7 LEVADURA  TGLHFLHMVMLAAMLGVNYWRMRNYHLTAGHHVGYETTIIYTHVLDVIWLFLYVTFYWWGV--
8 BACILO    VGTHGAHVAFGLMWISTLMIRNAKRGLNLYTAPKFYVASLYWHFIDVVWVFIFTVVYLMGMVG

ruler     .....220.......230.......240.......250.......260.......270.....
```

... AAACGTCTA ...
... AAAC-TCTA ...
... AAACGTCTA ...
... AAAGCTCTA ...
... AAACGTCTA ...
... AAACATCTA ...



hD, 2014

K-mers numbers

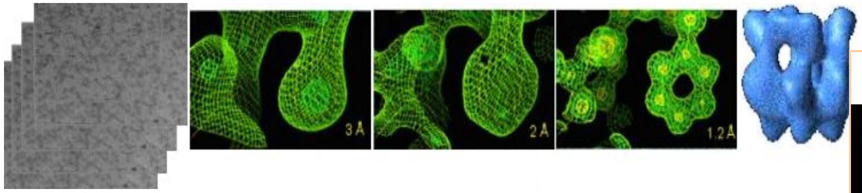
K	number of combinations in DNA			Number of combinations in Proteins	Aprox.
1	4			20	
2	16			400	
3	64			8.000	8 KB
4	256			160.000	
5	1.024	1 KB		3.200.000	3 MB
6	4.096			64.000.000	
7	16.384			1.280.000.000	1,2 GB
8	65.536			25.600.000.000	
9	262.144			512.000.000.000	
10	1.048.576	1 MB		10.240.000.000.000	10 TB
11	4.194.304			204.800.000.000.000	
12	16.777.216			4.096.000.000.000.000	4 PB
13	67.108.864			81.920.000.000.000.000	
14	268.435.456			1.638.400.000.000.000.000	1,6 EXA
15	1.073.741.824	1 GB		32.768.000.000.000.000.000	
16	4.294.967.296			655.360.000.000.000.000.000	
17	17.179.869.184			13.107.200.000.000.000.000.000	13 Zetta
18	68.719.476.736			262.144.000.000.000.000.000.000	
19	274.877.906.944			5.242.880.000.000.000.000.000.000	
20	1.099.511.627.776	1 TB		104.857.600.000.000.000.000.000.000	100 YottaB
25	1.125.899.906.842.620	1 PETA			
30	1.152.921.504.606.850.000	1 EXA			
32	18.446.744.073.709.600.000				

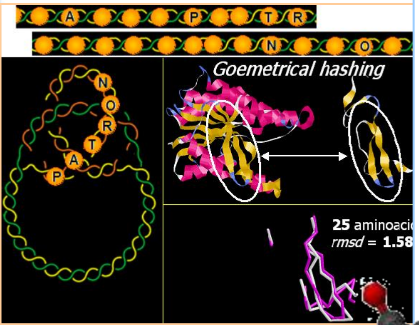
Computational space reduction

O.Trelles, PhD, 2014

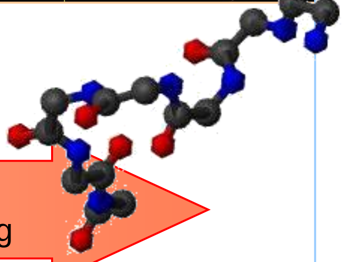
Structural Analysis: Proteins

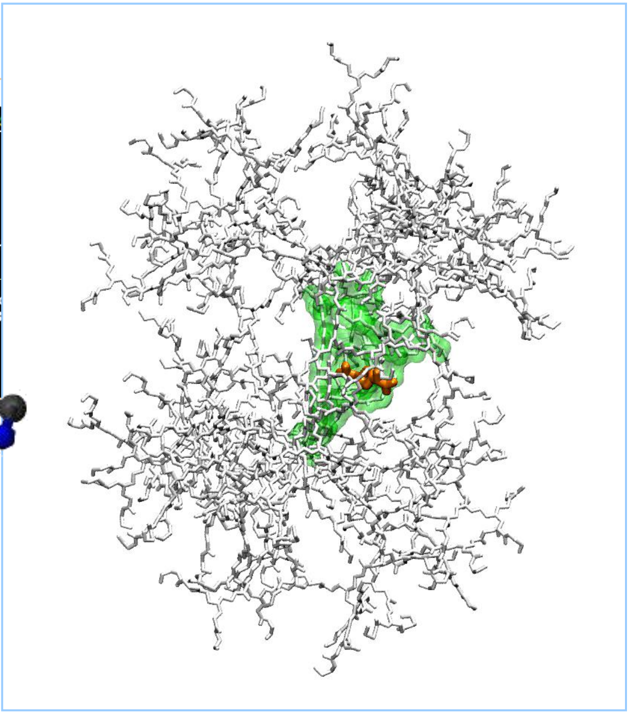
Field / algorithm	Input	Data volume		Processing features: computational load, memory access pattern
		In	Out	
Proteomics				
1.2 Sequence analisis and large scale phylogeny				
Structure prediction	3D protein BD and query seq.	100 MB	1 MB	Diversos tipos de tareas (búsquedas BD, comparación de estructuras, refinamiento de topologías), diferente carga
By-structural-homology database searching	3D protein BD and query seq.	100 MB	4 MB	E/S, tareas homogéneas, pesadas
Structural pattern matching	3D protein BD and query seq.	100 MB	4 MB	E/S, tareas heterogeneas, livianas (conteo)
Dynamic protein folding	Query sequence	4 MB	4 MB	Heavy tasks with data dependencies
Molecular inteactions and docking	Query sequence	4 MB	4 MB	Heavy tasks with data dependencies





```
HEADER    HORMONE                                08-OCT-96   2HIU
TITLE     NMR STRUCTURE OF HUMAN INSULIN IN 20% ACETIC ACID,
COMPND    MOLECULE: INSULIN;
COMPND    3 BIOLOGICAL_UNIT: HETERODIMER
SOURCE    ORGANISM_SCIENTIFIC: HOMO SAPIENS;
KEYWDS    INSULIN, HORMONE, GLUCOSE METABOLISM
AUTHOR    Q.X.HUA,S.N.GOZANI,R.E.CHANCE,J.A.HOFFMANN,B.
MODEL        1
ATOM      1  N   GLY A   1      -6.132   6.735   1.016  1.00  0.00
ATOM      2  CA  GLY A   1      -4.686   6.753   1.376  1.00  0.00
ATOM      3  C   GLY A   1      -3.864   6.149   0.235  1.00  0.00
```





Sequence homology → Structural similarity → Physico-chemical properties → Building-up the model → Fine-tuning

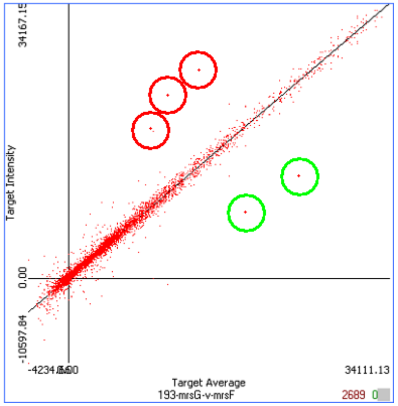
Medium/large protein: 3-4 weeks of CPU per 1 nanosecond of simulation
[128 or 256 cores] : 2-3 ns per day.
Biological processes : rank [micro to millisecond time scale]

O.Trelles, PhD, 2014

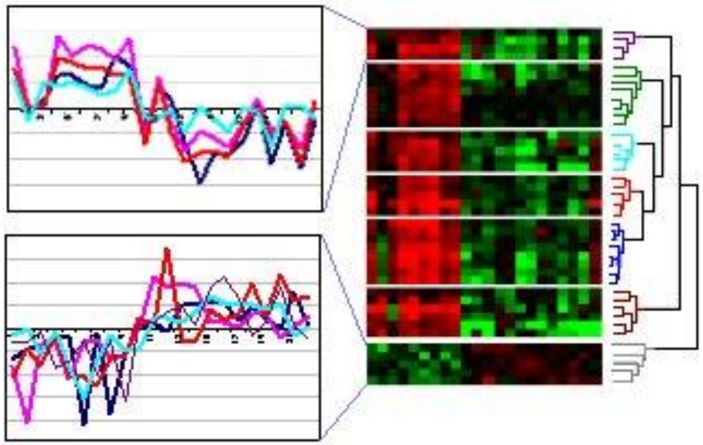
Gene-expression

Field / algorithm	Input	Data volume		Processing features: computational load, memory access pattern
		In	Out	
Transcriptomics				
1.4 Gene-expression analysis				
Data acquisition. Image processing	100 Exp. & 6 M samples	10 GB	10 GB	Image processing. Light and regular pattern
Data Quality and normalization	100 Exp. & 6 M samples	10 GB	10 GB	Image processing. Heavy, regular pattern
Clustering of gene-expression profiles	100 Exp. & 6 M samples	10 GB	10 GB	Out-of-memory, data dependencies, lighted tasks
Marker genes identification	100 Exp. & 6 M samples	10 GB	10 GB	Heavy I/O, out-of-memory, light tasks

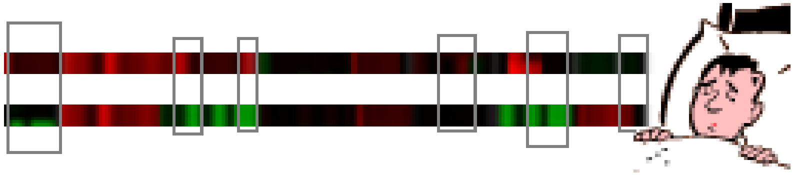
Diferential Expression

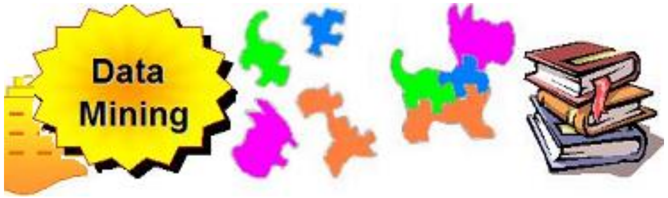


Clustering



Clasification





KDD: Association studies

Illustrative use cases

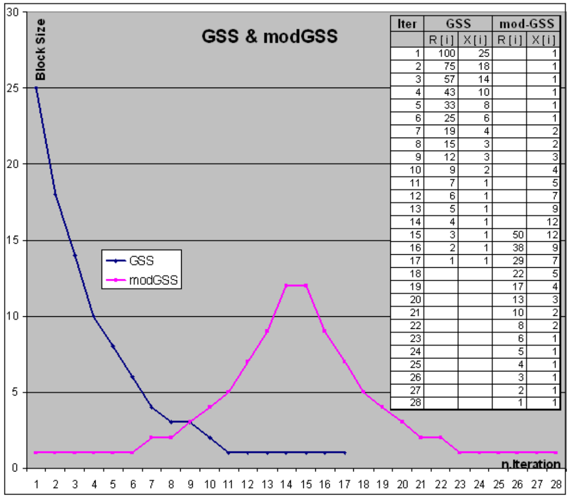
HPC: The basic model

DB-searching Applications

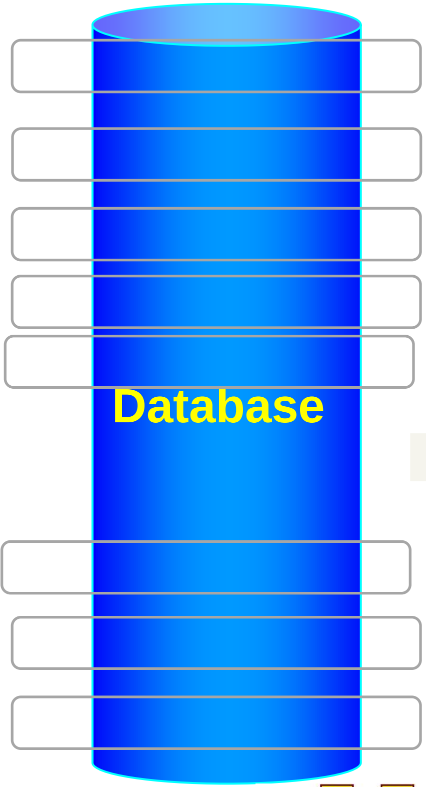
- High number of tasks
- Heterogeneity (tasks & CPU-power)
- Network overload
- Scheduling / distribution overload
- Task synchronization
- Fault tolerance
- Portability

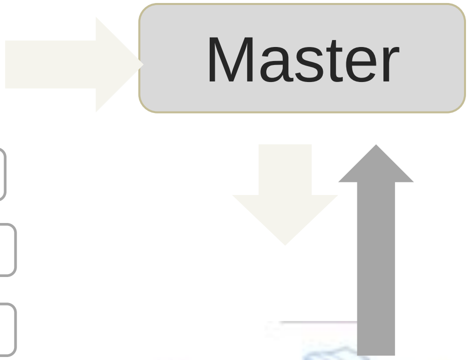
The model

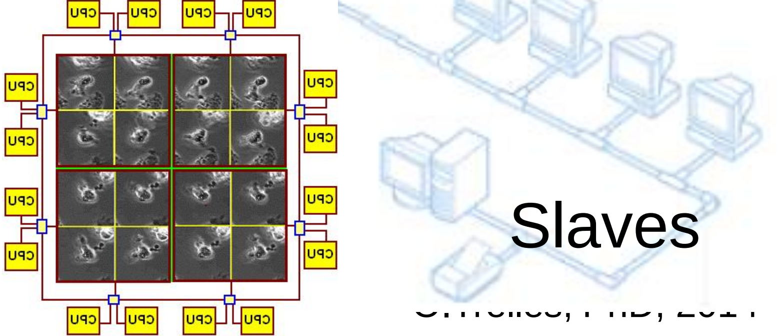
- Task parallel (coarse grained)
- Dynamic load balancing
- Network optimization (message size)
- Minimize number of messages
- Buffering (speculative scheduling)
- Check-points
- SM, DM, D&SM architectures



Iter	GSS		mod-GSS	
	R [i]	X [i]	R [i]	X [i]
1	100	25		1
2	75	18		1
3	57	14		1
4	43	10		1
5	33	8		1
6	25	6		1
7	19	4		2
8	15	3		2
9	12	3		3
10	9	2		4
11	7	1		5
12	6	1		7
13	5	1		9
14	4	1		12
15	3	1	50	12
16	2	1	38	9
17	1	1	29	7
18			22	5
19			17	4
20			13	3
21			10	2
22			8	2
23			6	1
24			5	1
25			4	1
26			3	1
27			2	1
28			1	1

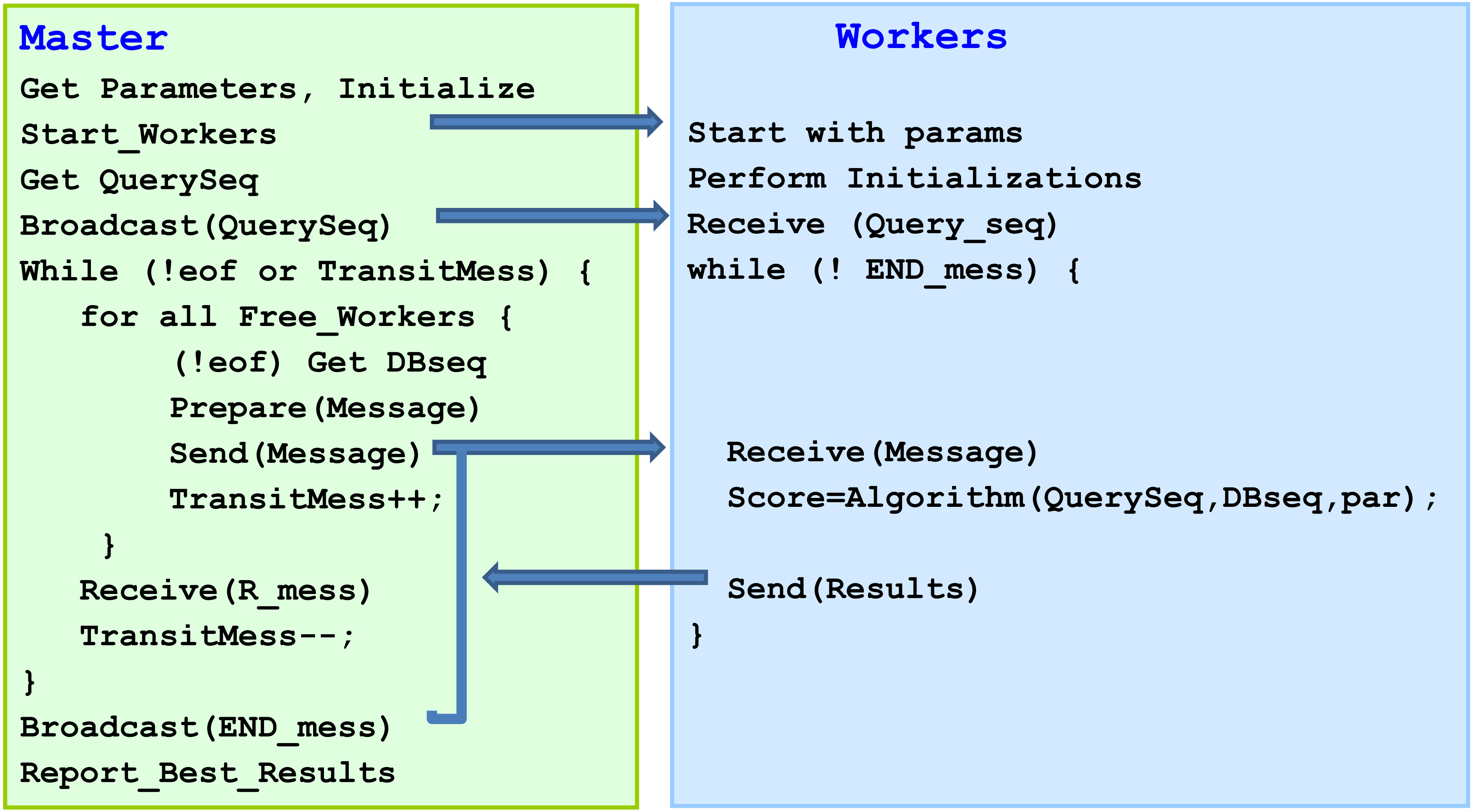






Sequence DBsrch with Dynamic programming

Coarse grained



Master

```
Get Parameters, Initialize
Start_Workers
Get QuerySeq
Broadcast(QuerySeq)
While (!eof or TransitMess) {
   for all Free_Workers {
        (!eof) Get DBseq
        Prepare(Message)
        Send(Message)
        TransitMess++;
    }
   Receive(R_mess)
   TransitMess--;
}
Broadcast(END_mess)
Report_Best_Results
```

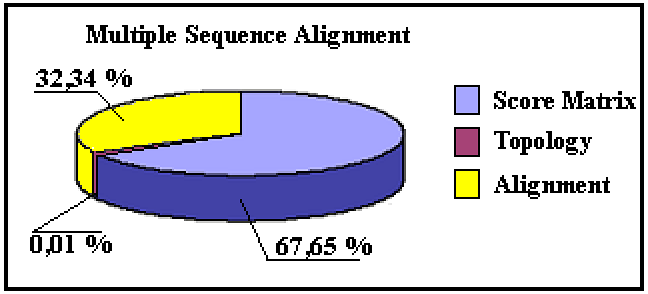
Workers

```
Start with params
Perform Initializations
Receive (Query_seq)
while (! END_mess) {
  Receive(Message)
  Score=Algorithm(QuerySeq,DBseq,par);

  Send(Results)
}
```

ClustalW overview: (Thompson J. *et al*, NAR, 1994, 2003, 2007)

- **Pairwise (PW) alignment matrix**

  average alignment calculation spends most of its time here easy to parallelize as all *N*(N-1)/2* elements are independent

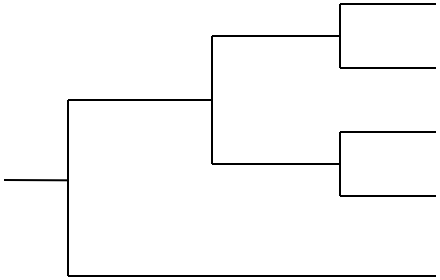


Cross Similarity Matrix

	[0]	[1]	[2]	[3]	[4]	[5]	[6]
[0]	-	-	-	-	-	-	-
[1]	82	-	-	-	-	-	-
[2]	52	54	-	-	-	-	-
[3]	60	62	86	-	-	-	-
[4]	22	24	18	24	-	-	-
[5]	26	20	12	16	78	–	-
[6]	22	14	10	8	46	48	-

- **Guide tree calculation**

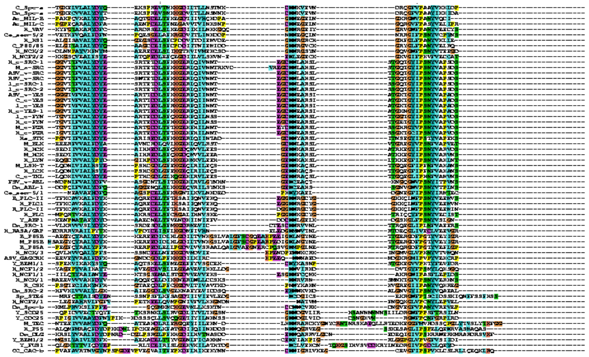
  Calculation of closest sequences (branch) is a relatively light task, that can be solved sequentially.



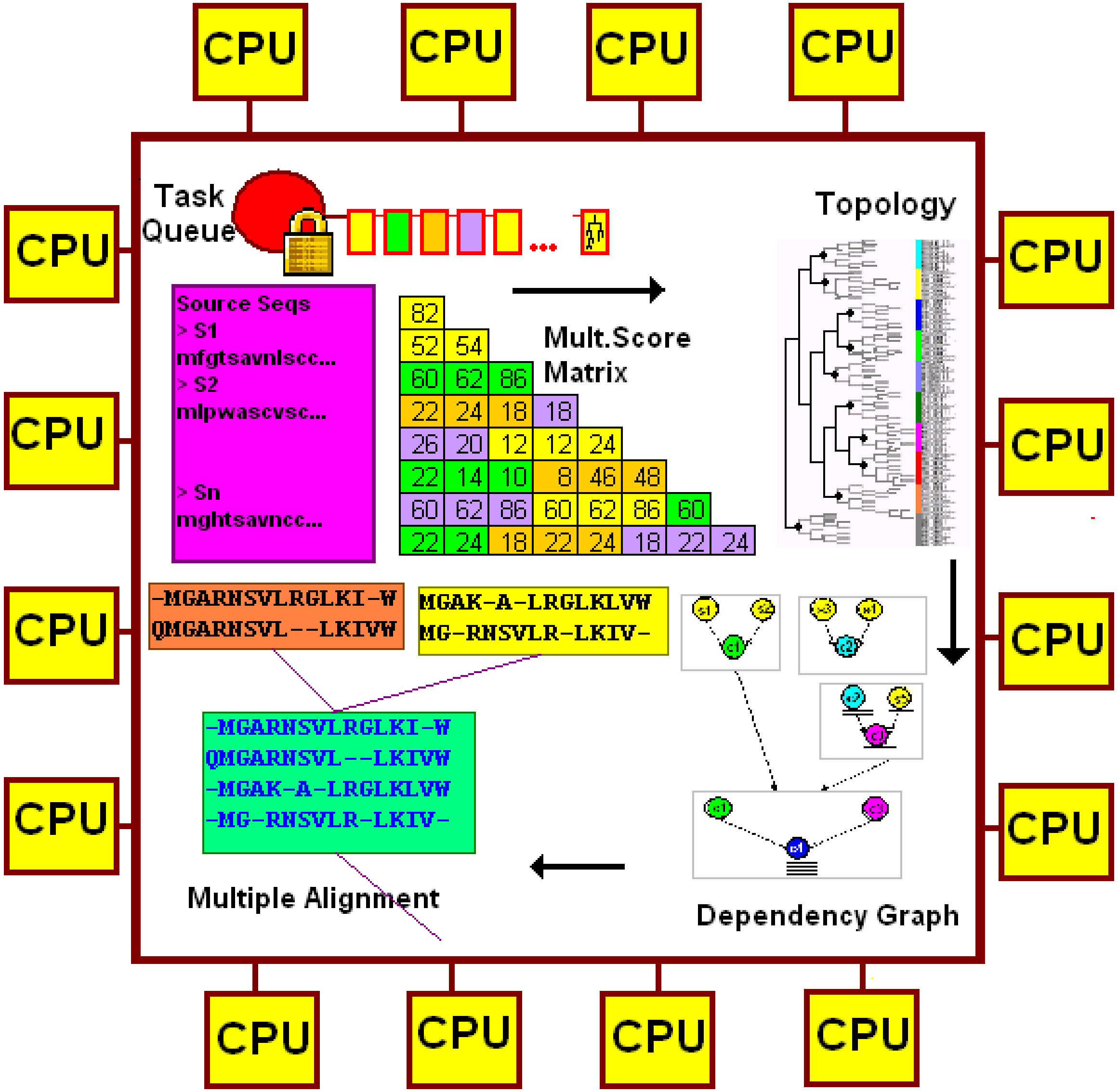
- **Progressive alignment**

  Remaining ~30% of the code can be parallelized at this stage by calculating profile scores in parallel, and by solving data dependencies. *(N-1) cluster vs cluster* alignments must be solved.

  As a result the whole application is ~90% parallel depending on a size of a problem



Shared Memory Parallel Model



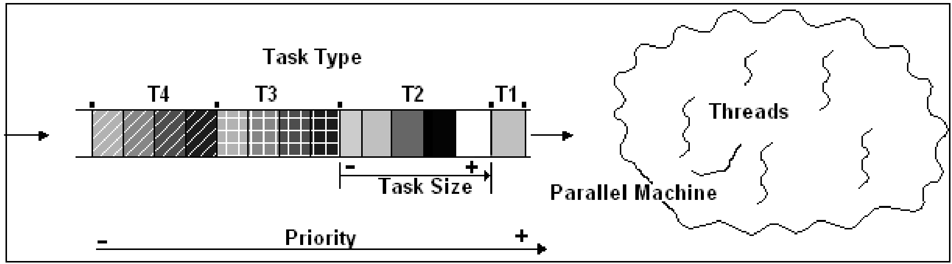
Source Seqs
> S1
mfgtsavnlscc...
> S2
mlpwascvsc...
> Sn
mghtsavncc...

82							
52	54						
60	62	86					
22	24	18	18				
26	20	12	12	24			
22	14	10	8	46	48		
60	62	86	60	62	86	60	
22	24	18	22	24	18	22	24

```
-MGARNSVLRGLKI-W
QMGARNSVL--LKIVW
```

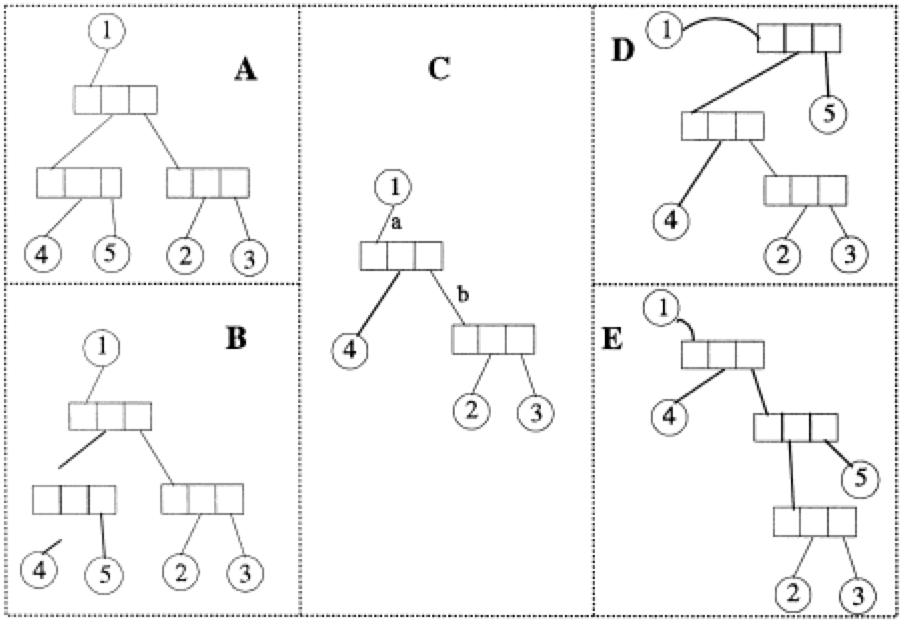
```
MGAK-A-LRGLKLVW
MG-RNSVLR-LKIV-
```

```
-MGARNSVLRGLKI-W
QMGARNSVL--LKIVW
-MGAK-A-LRGLKLVW
-MG-RNSVLR-LKIV-
```



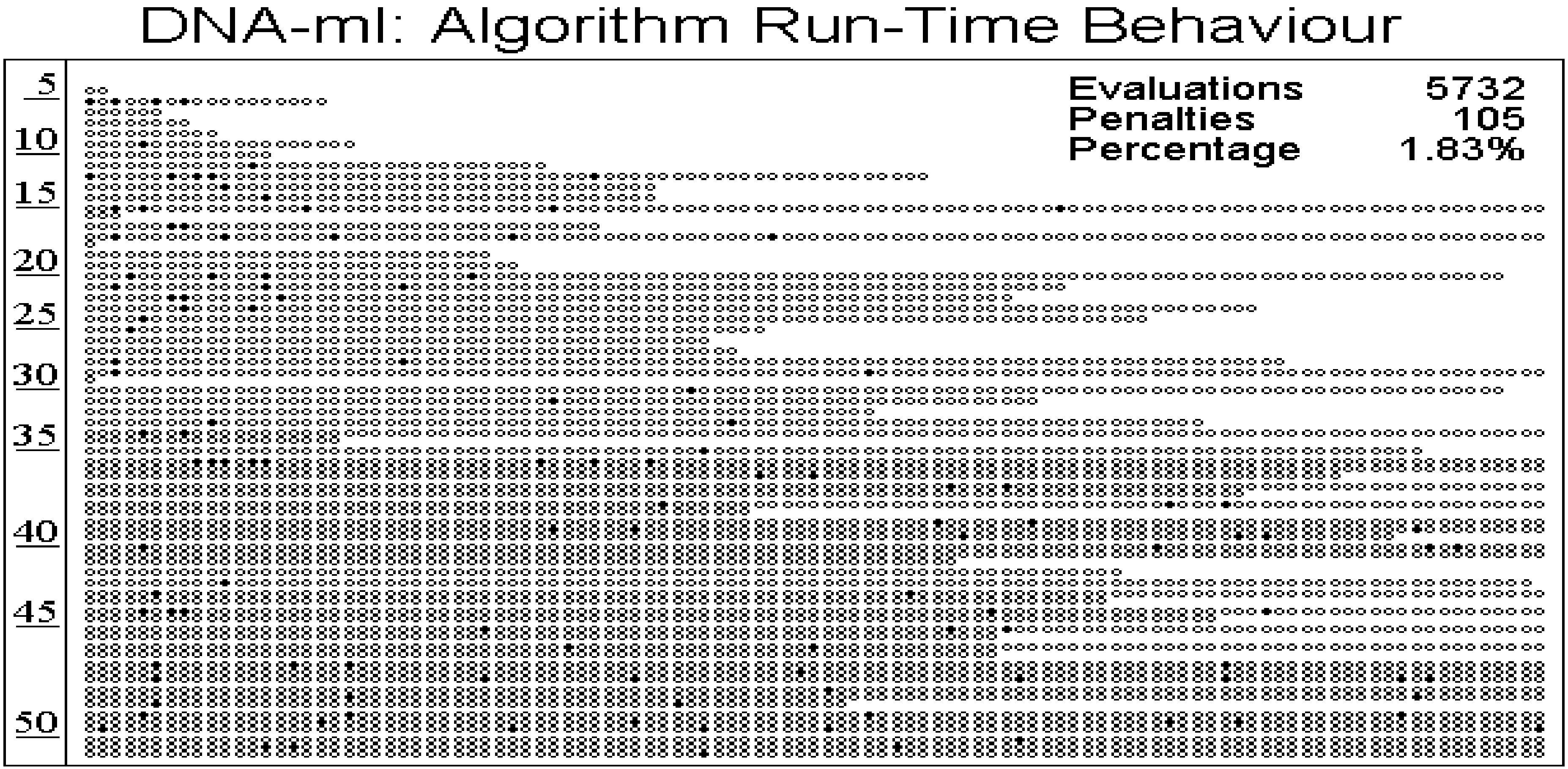
Irregular algorithms: DNAml

```
Current-best-tree $T_k$ ($L_k$) [from insertion step]
for i = 1 to n-tasks
    Remove sub-tree i from $T_k$ and produce $T_{k1}$ and $T_{k2}$
    Likelihood evaluation for $T_{k1}$ and $T_{k2}$ ($L_{k1}$ and $L_{k2}$)
    Current-best-tree $T_k$ = tree with greater likelihood ($T_k$, $T_{k1}$, $T_{k2}$)
end for
```



Irregular algorithms

Speculative computing



O.Trelles, PhD, 2014

Open & Provoking questions

Scenarios: the real work

	Comparative genomics scenarios (CG)
CG1	multi-genome comparison on higher mammalians
CG2	Multi-genome comparison and phylogenomics.
CG3	Symbionts study case
CG4	Metagenome analysis
	Biomedical scenarios (BM)
BM1	access to summarized information of the clinical DB through mobiles
BM2	Data analysis: Combining protein interaction and pathway data
BM3	Discovering correlations between clinical and molecular patient data

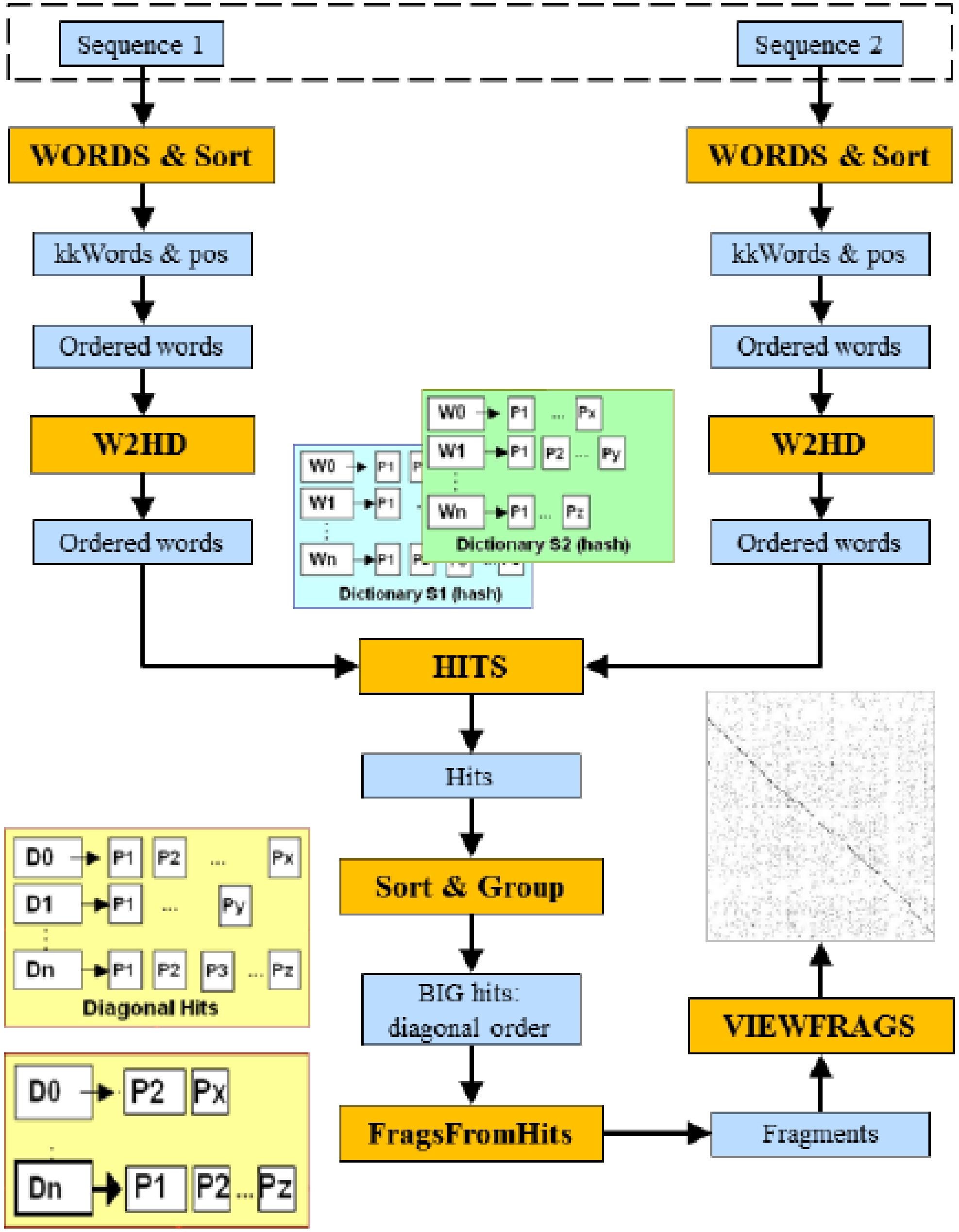
Aims

(1) big-data, HPC, Grid & Cloud, visualization

(2) Security, data sensitivity, data analysis

(3) New statatistic, math & biological models

The global idea: HSPs out-of-core



Sequence 1

Sequence 2

WORDS & Sort

kkWords & pos

Ordered words

W2HD

Ordered words

WORDS & Sort

kkWords & pos

Ordered words

W2HD

Ordered words

HITS

Hits

Sort & Group

BIG hits:
diagonal order

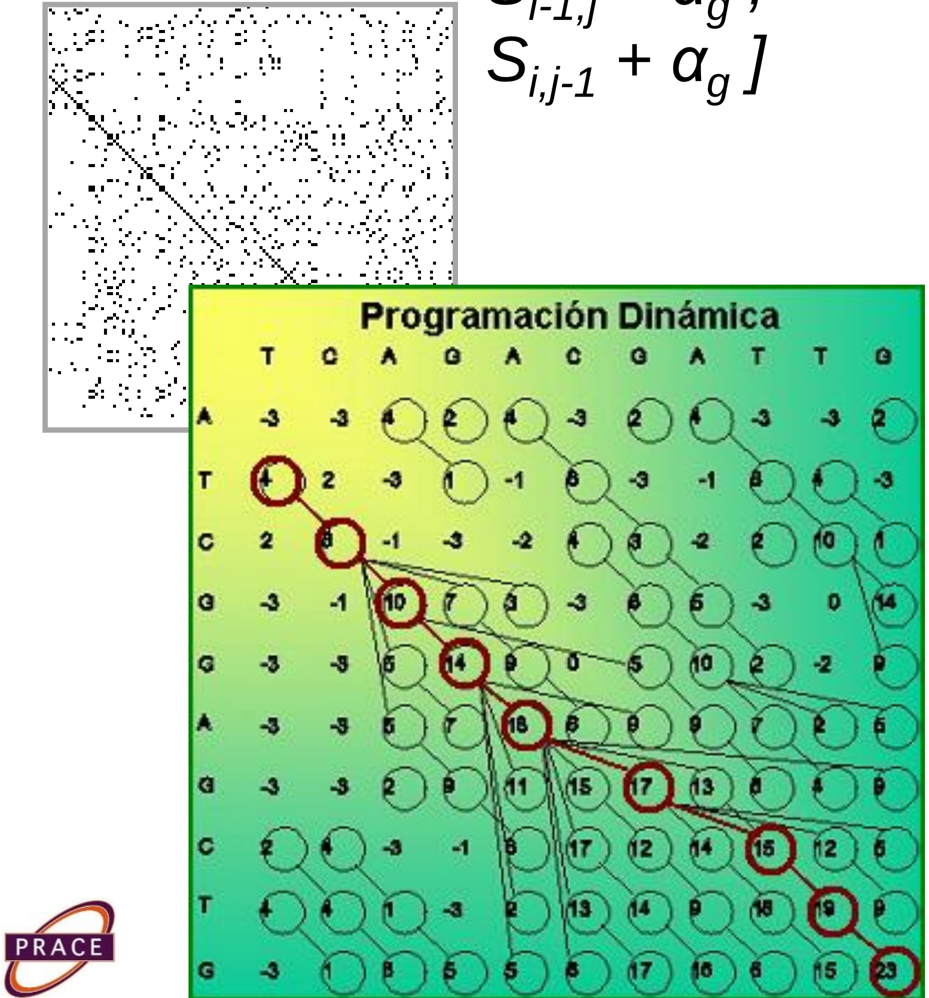
FragsFromHits

Fragments

VIEWFRAGS

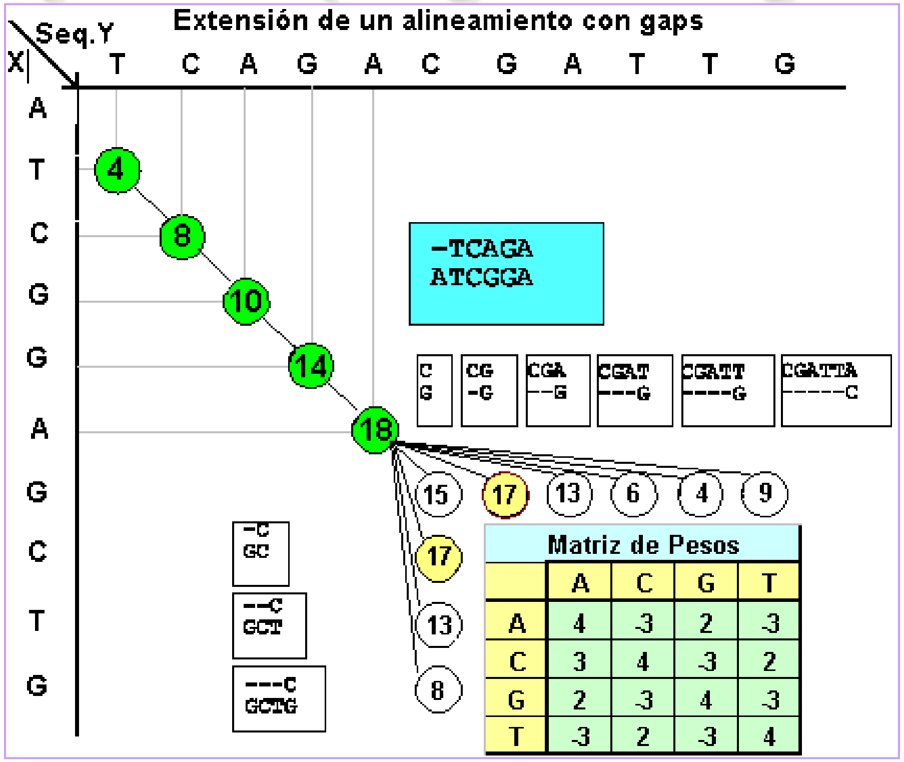
Pairwise sequence/genome comparison
Sequence DBsrch with Dynamic programming

$$S_{i,j} = \max [\; S_{i-1,j-1} + w(x_i,y_j),\; S_{i-1,j} + \alpha_g ,\; S_{i,j-1} + \alpha_g \;]$$



Programación Dinámica

	T	C	A	G	A	C	G	A	T	T	G
A	-3	-3	4	2	4	-3	2	4	-3	-3	2
T	4	2	-3	1	-1	6	-3	-1	8	4	-3
C	2	8	-1	-3	-2	4	3	-2	2	10	1
G	-3	-1	10	7	3	-3	6	5	-3	0	14
G	-3	-3	5	14	9	0	5	10	2	-2	9
A	-3	-3	5	7	18	6	9	9	7	2	6
G	-3	-3	2	9	11	15	17	13	6	4	9
C	2	4	-3	-1	6	17	12	14	15	12	6
T	4	4	1	-3	2	13	14	9	18	19	9
G	-3	1	8	5	5	6	17	16	6	15	23

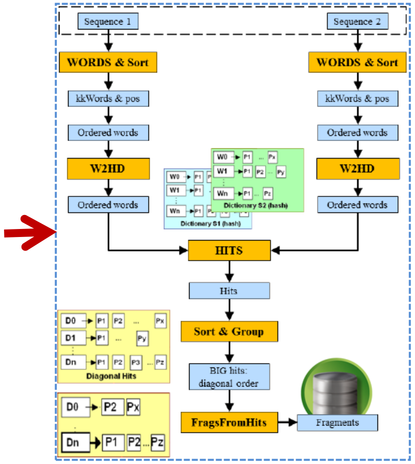


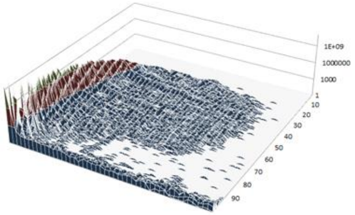
Matriz de Pesos

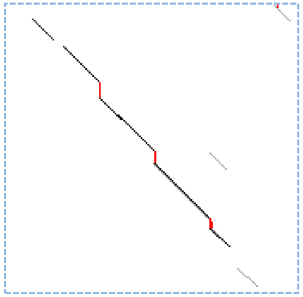
	A	C	G	T
A	4	-3	2	-3
C	3	4	-3	2
G	2	-3	4	-3
T	-3	2	-3	4

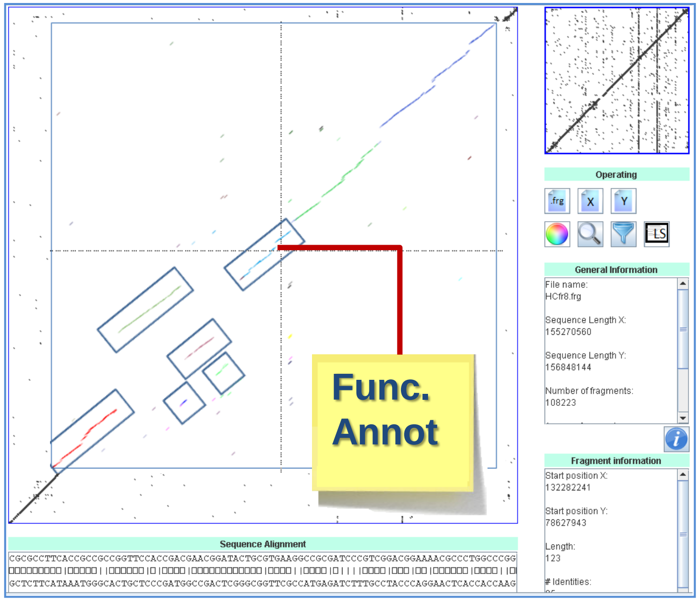
```
qNEW-SEQ   --SARGDFLNAA YALFFMRSHN FGHSDVLPVL
             ||||||||   |||  |||||    |||||||
KNOWN-SEQ  MMSARGDFLN-- YALSLMRSHN DEHSDVLPVL

qNEW-SEQ  --CSLKHVAY WDAYQALIYW IKAMNQQTDTSI
            ||||||||     |||||| | |||||||||
KNOWN-SEQ  DVCSLKHVAY -VFQALIYW IKAMNQQTTLDT

qNEW-SEQ  --RPPDDQAF GHHHLPQAMH --SRLYVPS-SK
            |||      |   || |     ||||||| ||
KNOWN-SEQ TIRPPA---- GAFGLPTANT CISRLYVPSMSK
```

Multi-genome comparison (CG1)









```
CGCGCCTTCACCGCCGCCGGTTCCACCGACGAACGGATACTGCGTGAAGGCCGCGATCCCGTCGGACGGAAAACGCCCTGGCCCGG
□□□□□□□□□ |□□□□□ | |□□□□□□ |□ |□□□□ |□□□□□□□□□□□□ |□□□□ | |□□□□ |□ | | | |□□□□ |□□□ |□□ |□□□□□□ |□□□□ | |□
GCTCTTCATAAATGGGCACTGCTCCCGATGGCCGACTCGGGCGGTTCGCCATGAGATCTTTGCCTACCCAGGAACTCACCACCAAG
```

File name:
HCfr8.frg

Sequence Length X:
155270560

Sequence Length Y:
156848144

Number of fragments:
108223

Start position X:
132282241

Start position Y:
78627943

Length:
123

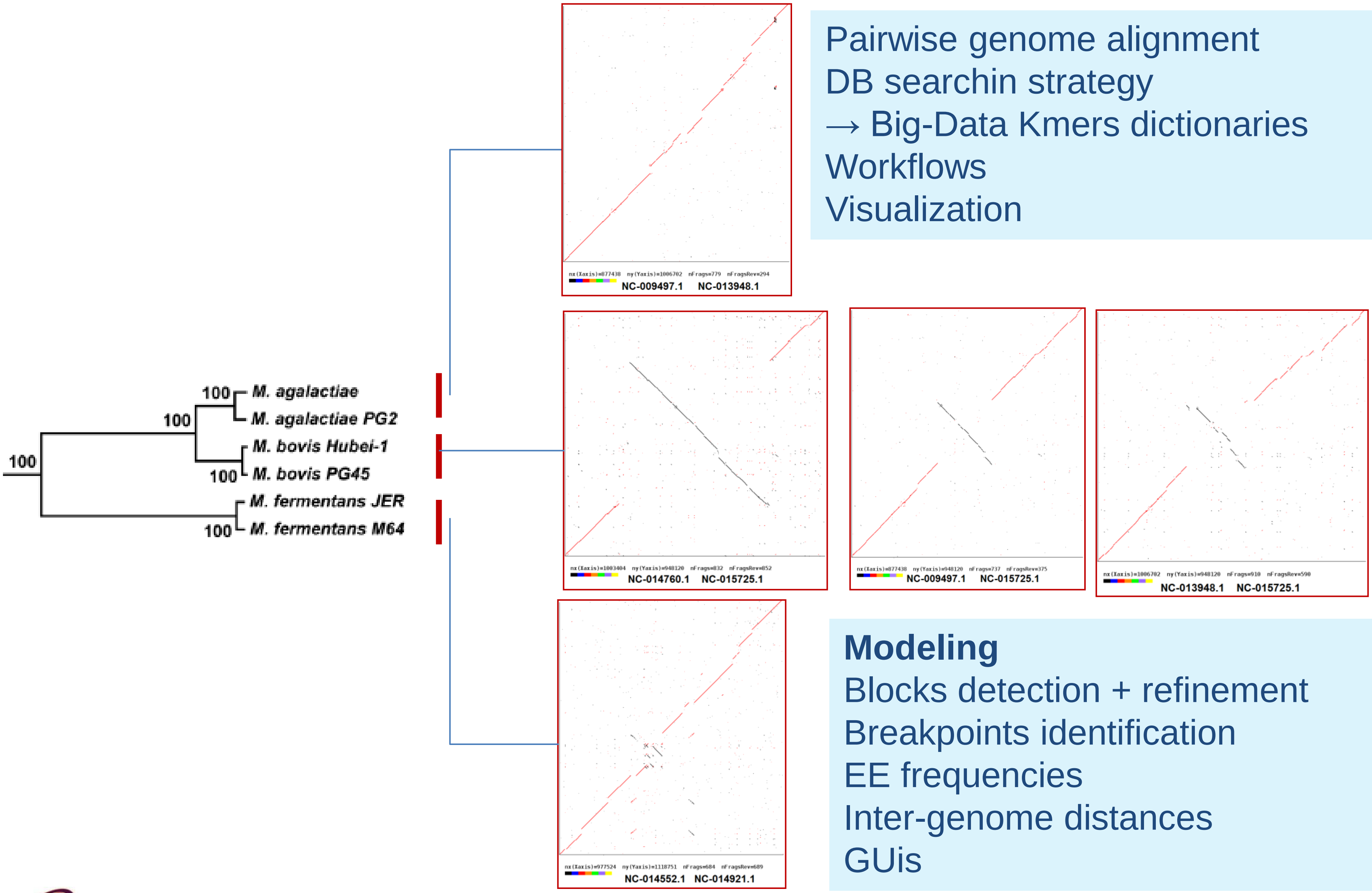
Identities:

Big-data
HPC
Modeling
Visualization Data analysis
GUIs

O.Trelles, PhD, 2014

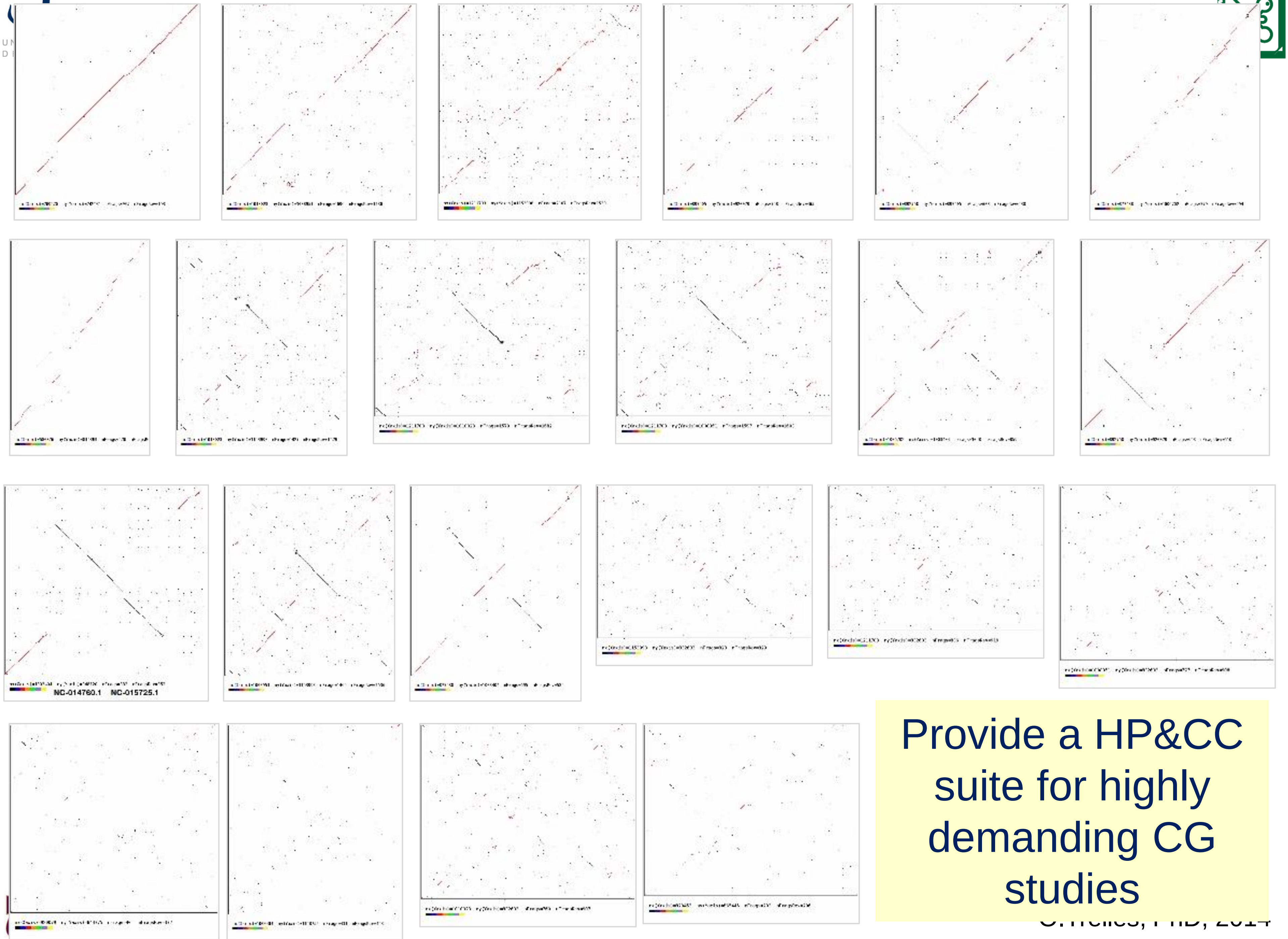
MG comparison & phylogeny (CG2)

Pairwise genome alignment
DB searchin strategy
→ Big-Data Kmers dictionaries
Workflows
Visualization


100
M. agalactiae
100
M. agalactiae PG2
100
M. bovis Hubei-1
100
M. bovis PG45
M. fermentans JER
100
M. fermentans M64
nx(Xaxis)=877438 ny(Yaxis)=1006702 nFrags=779 nFragsRev=294
NC-009497.1 NC-013948.1
nx(Xaxis)=1003404 ny(Yaxis)=948120 nFrags=832 nFragsRev=852
NC-014760.1 NC-015725.1
nx(Xaxis)=877438 ny(Yaxis)=948120 nFrags=737 nFragsRev=375
NC-009497.1 NC-015725.1
nx(Xaxis)=1006702 ny(Yaxis)=948120 nFrags=910 nFragsRev=590
NC-013948.1 NC-015725.1
nx(Xaxis)=977524 ny(Yaxis)=1118751 nFrags=684 nFragsRev=689
NC-014552.1 NC-014921.1

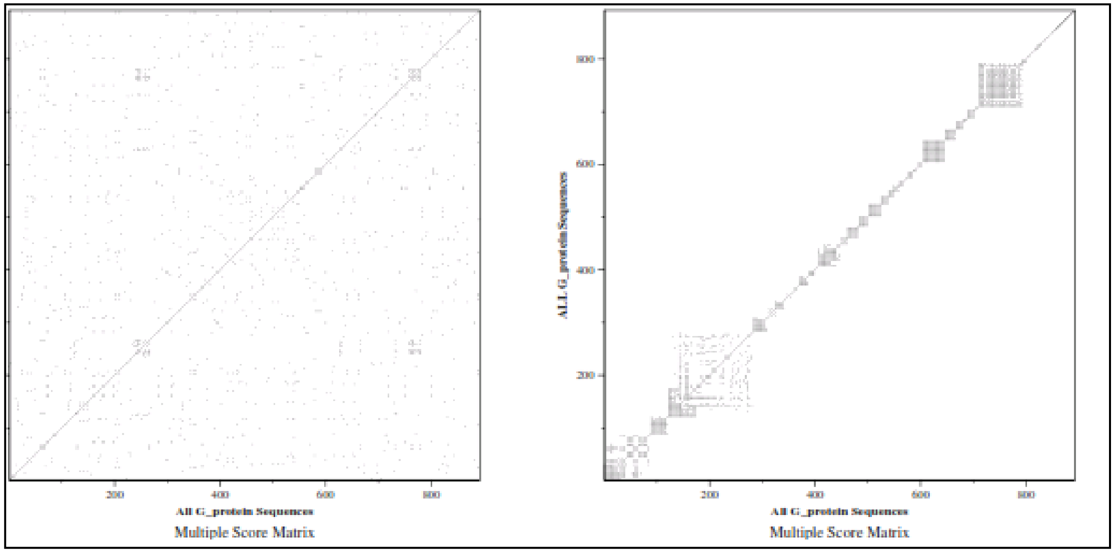
Modeling
Blocks detection + refinement
Breakpoints identification
EE frequencies
Inter-genome distances
GUIs

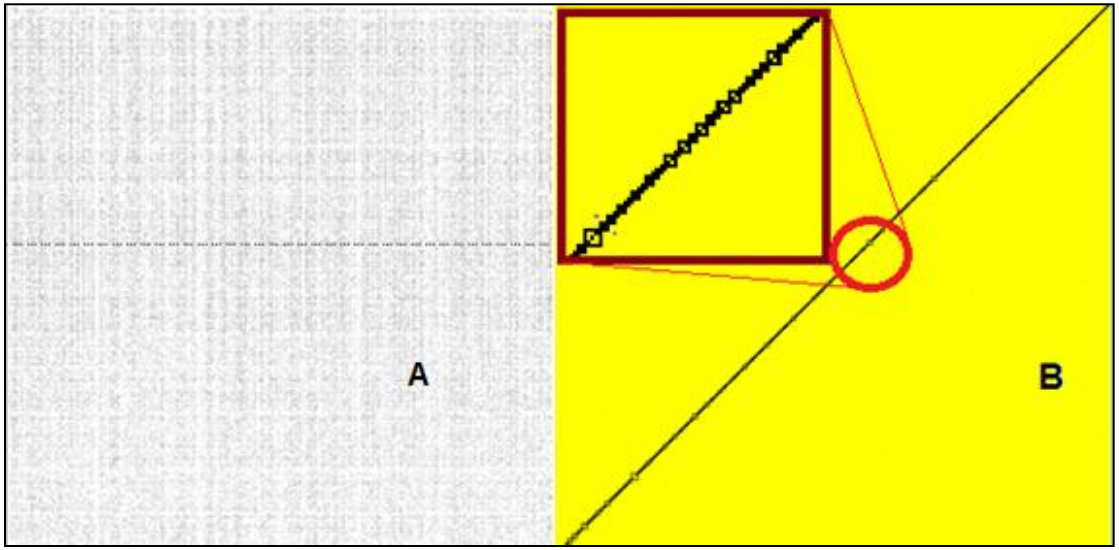
O.Trelles, PhD, 2014



Provide a HP&CC suite for highly demanding CG studies

Meta-genomes comparison



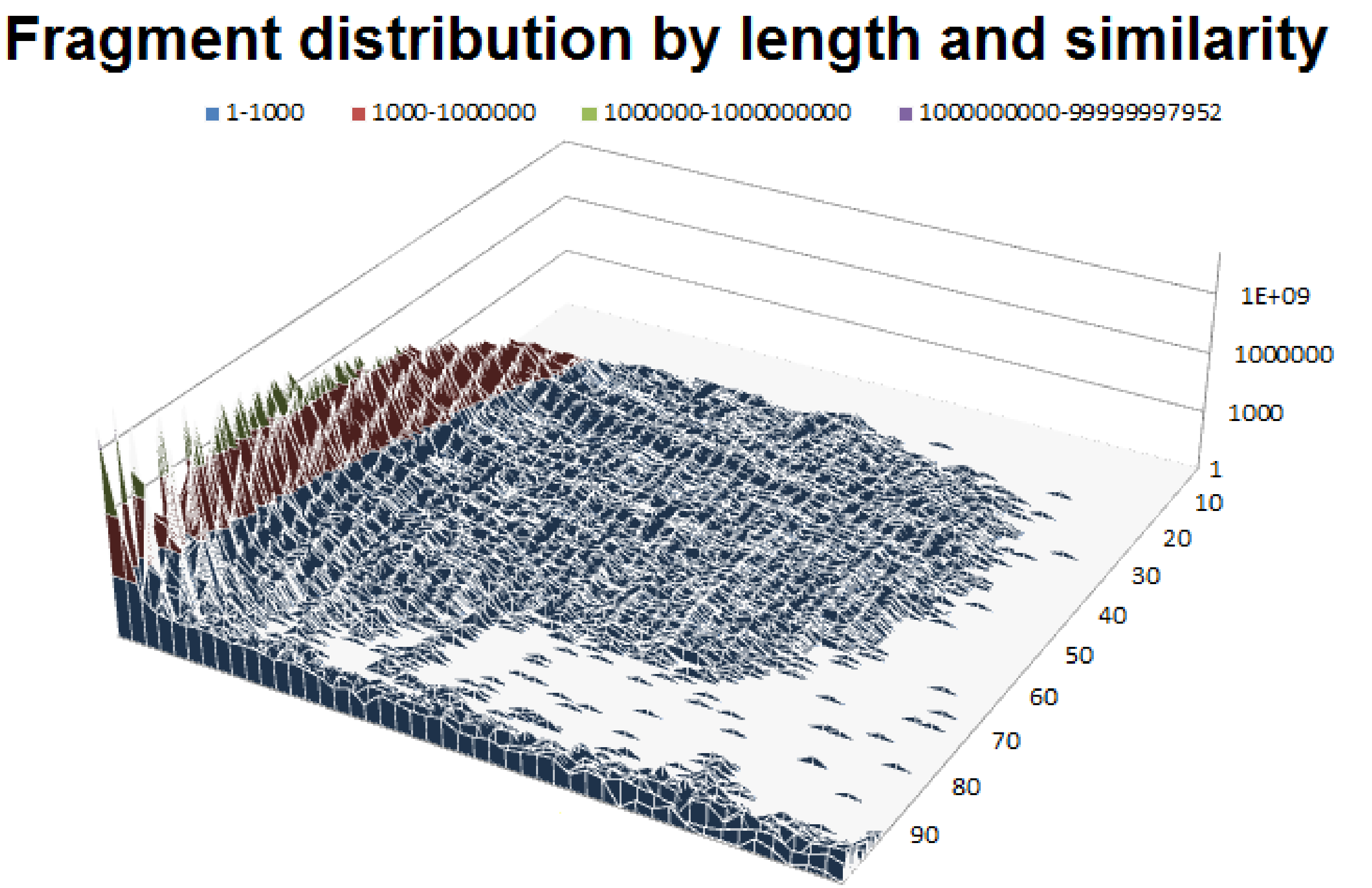


Computational space reduction based on the fast identification of matching reads

Trelles O. et al. Computational Space Reduction and Parallelization of a new Clustering Approach for Large Groups of Sequences"; Bioinformatics vol.14 no.5 1998 (pp.439-451)

Human vs Chimpanzee

(close related organisms)



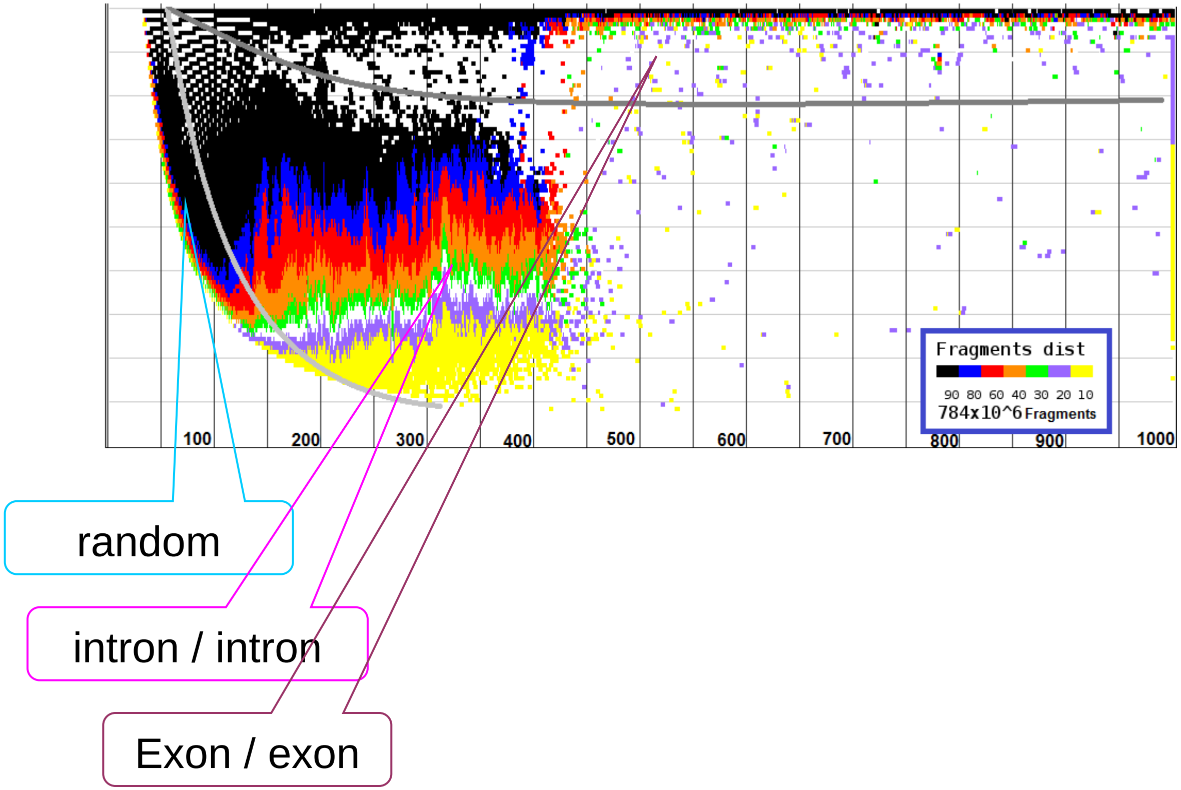
33	41	51	61	71	81	91	101	111	121	131	141	151	161	171	181	191	201	211	221	231	241						
40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250						
			251	261	271	281	291	301	311	321	331	341	351	361	371	381	391	401	411	421	431	441	451	461	471	481	491
			260	270	280	290	300	310	320	330	340	350	360	370	380	390	400	410	420	430	440	450	460	470	480	490	500

Different distributions?

Introns, exons, intergenic… produce the same type of fragments?

Statistical Significance of HSPs

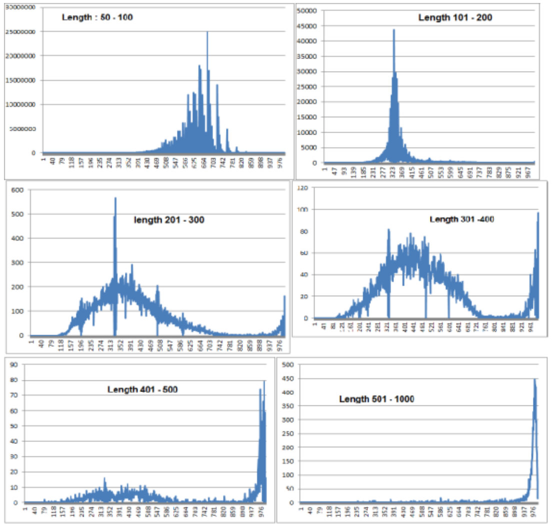
New models are needed



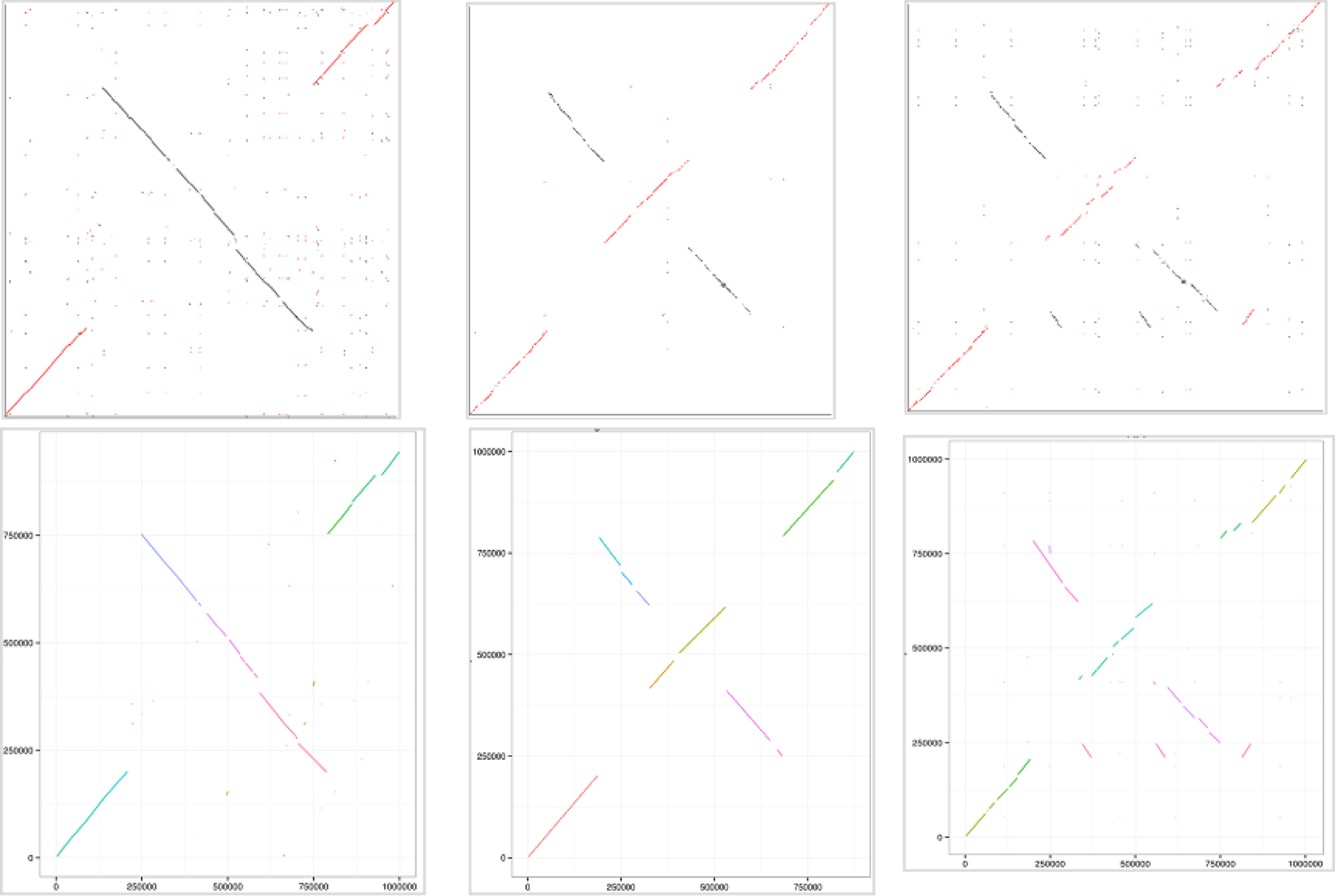
random

intron / intron

Exon / exon



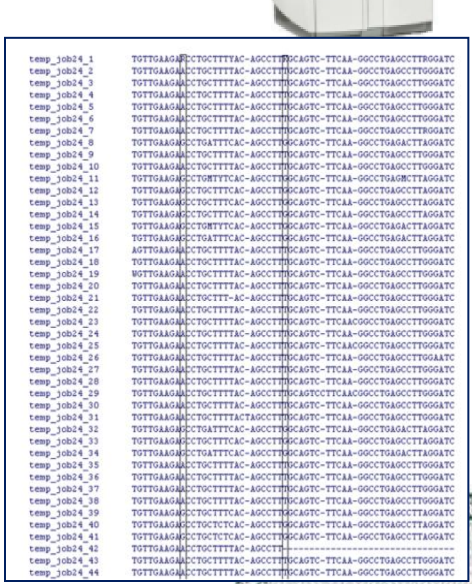
Evolution events in full genomes



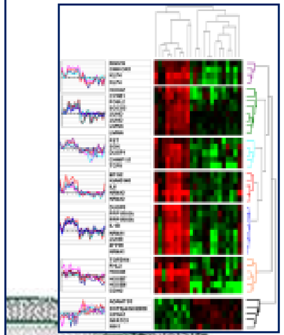
O.Trelles, PhD, 2014

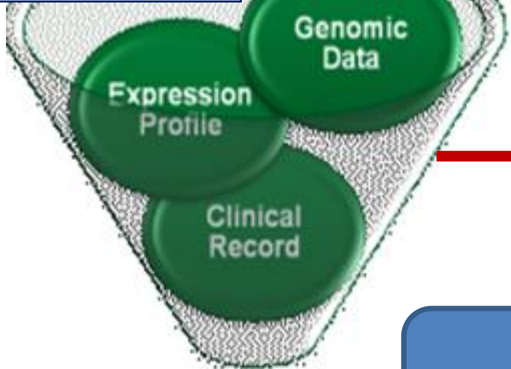
Bio-Medical scenarios

(Allergies)

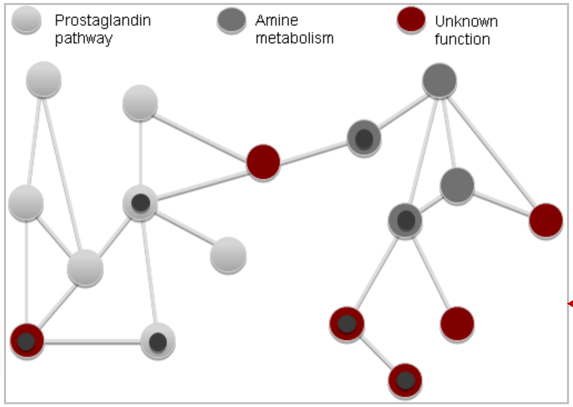


```
temp_job24_1      TGTTGAAGARCCTGCTTTYAC-AGCCTTRGCAGTC-TTCAA-GGCCTGAGCCTTRGGATC
temp_job24_2      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_3      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_4      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_5      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_6      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_7      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTRGGATC
temp_job24_8      TGTTGAAGAGCCTGATTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGACTTAGGATC
temp_job24_9      TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_10     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_11     TGTTGAAGAGCCTGMTYTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGMCTTAGGATC
temp_job24_12     TGTTGAAGAGCCTGCTTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_13     TGTTGAAGAGCCTGCTTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_14     TGTTGAAGAGCCTGCTTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_15     TGTTGAAGAGCCTGMTYTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGACTTAGGATC
temp_job24_16     TGTTGAAGAGCCTGATTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGACTTAGGATC
temp_job24_17     AGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_18     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_19     WGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_20     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_21     TGTTGAAGAACCTGCTTT-AC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_22     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_23     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAACGGCCTGAGCCTTGGGATC
temp_job24_24     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_25     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAACGGCCTGAGCCTTGGGATC
temp_job24_26     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGAATC
temp_job24_27     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_28     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_29     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTCCTTCAACGGCCTGAGCCTTGGGATC
temp_job24_30     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_31     TGTTGAAGAACCTGCTTTTACTAGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_32     TGTTGAAGAGCCTGATTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGACTTAGGATC
temp_job24_33     TGTTGAAGAGCCTGCTTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_34     TGTTGAAGAGCCTGATTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGACTTAGGATC
temp_job24_35     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_36     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_37     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_38     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_39     TGTTGAAGAGCCTGCTTTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_40     TGTTGAAGAGCCTGCTCTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_41     TGTTGAAGAGCCTGCTCTCAC-AGCCTTGGCAGTC-TTCAA-GGCCTGAGCCTTAGGATC
temp_job24_42     TGTTGAAGAACCTGCTTTTAC-AGCCTT--------------------------------
temp_job24_43     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
temp_job24_44     TGTTGAAGAACCTGCTTTTAC-AGCCTTTGCAGTC-TTCAA-GGCCTGAGCCTTGGGATC
```



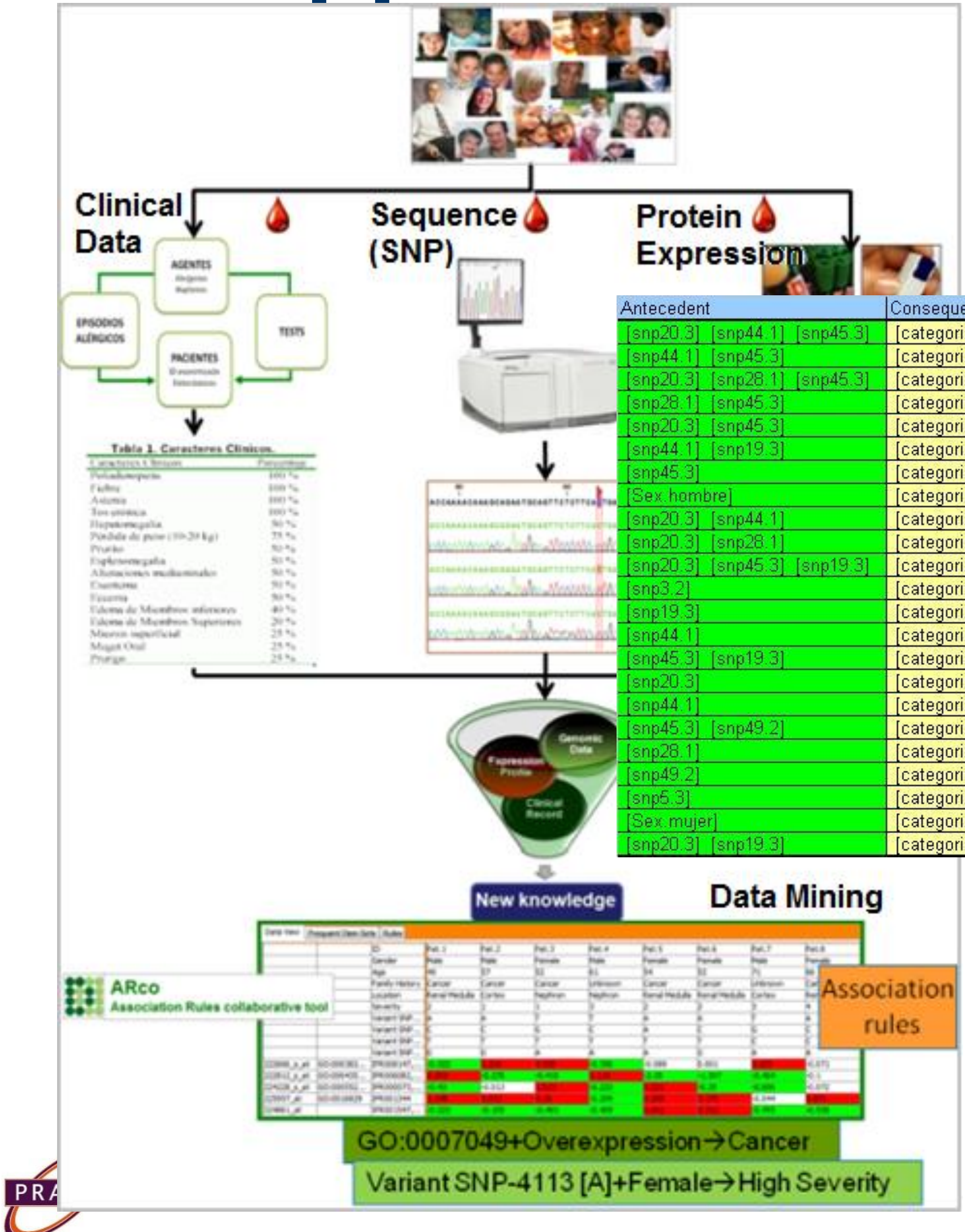


Pheno-geno Correlation Models



hD, 2014

Apps: Patient + Genomic data



Antecedent	Consequent	Confidence	Support	SupportAb	Coverage	Improveme	Leverage	Conviction
[snp20.3] [snp44.1] [snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	76,64	40,45	45	33,18	1,46	6,43	150,5
[snp44.1] [snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	73,24	44,09	53	41,36	1,38	6,61	138,24
[snp20.3] [snp28.1] [snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	71,79	40,91	46	36,82	1,34	5,35	133,6
[snp28.1] [snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	69,84	43,18	51	42,27	1,3	5,31	127,82
[snp20.3] [snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	69,62	52,27	71	59,09	1,29	7,29	127,2
[snp44.1] [snp19.3]	[categori.CONTROL]	69,26	43,18	51	42,73	1,31	5,51	128,18
[snp45.3]	[categori.URTICARIA+ANGIOEDEMA]	66,28	56,36	80	70,91	1,21	6,39	118,49
[Sex.hombre]	[categori.URTICARIA+ANGIOEDEMA]	65,55	40,91	46	41,36	1,2	3,42	116,74
[snp20.3] [snp44.1]	[categori.URTICARIA+ANGIOEDEMA]	64,5	42,73	50	45,91	1,17	3,32	114,32
[snp20.3] [snp28.1]	[categori.URTICARIA+ANGIOEDEMA]	64,49	42,27	49	45	1,17	3,25	114,3
[snp20.3] [snp45.3] [snp19.3]	[categori.URTICARIA+ANGIOEDEMA]	64,44	40	44	40,45	1,17	2,9	114,17
[snp3.2]	[categori.CONTROL]	63,39	40,45	45	42,27	1,17	2,97	113,61
[snp19.3]	[categori.CONTROL]	62,52	50,45	67	64,09	1,15	3,94	111,73
[snp44.1]	[categori.CONTROL]	61,76	49,55	65	63,18	1,13	3,41	110,14
[snp45.3] [snp19.3]	[categori.URTICARIA+ANGIOEDEMA]	61,67	42,27	49	47,73	1,1	2,1	108,24
[snp20.3]	[categori.URTICARIA+ANGIOEDEMA]	61,39	55	77	75,45	1,1	3,1	107,67
[snp44.1]	[categori.URTICARIA+ANGIOEDEMA]	61,04	49,09	64	63,18	1,09	2,38	106,99
[snp45.3] [snp49.2]	[categori.URTICARIA+ANGIOEDEMA]	59,9	40	44	44,55	1,06	1,17	104,76
[snp28.1]	[categori.URTICARIA+ANGIOEDEMA]	59,62	46,36	58	59,09	1,06	1,38	104,23
[snp49.2]	[categori.CONTROL]	59,44	47,27	60	61,36	1,07	1,89	105,55
[snp5.3]	[categori.CONTROL]	58,81	40,91	46	47,73	1,06	1,17	104,35
[Sex.mujer]	[categori.CONTROL]	58,41	45,45	56	58,64	1,05	1,2	103,62
[snp20.3] [snp19.3]	[categori.CONTROL]	57,86	41,82	48	50,91	1,04	0,76	102,61

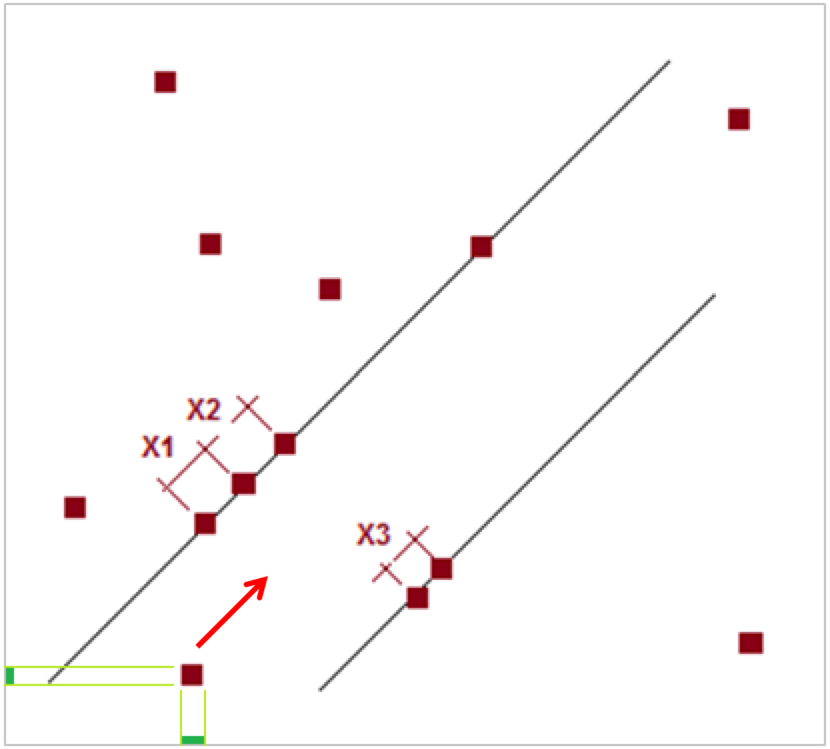
GWAS Analysis

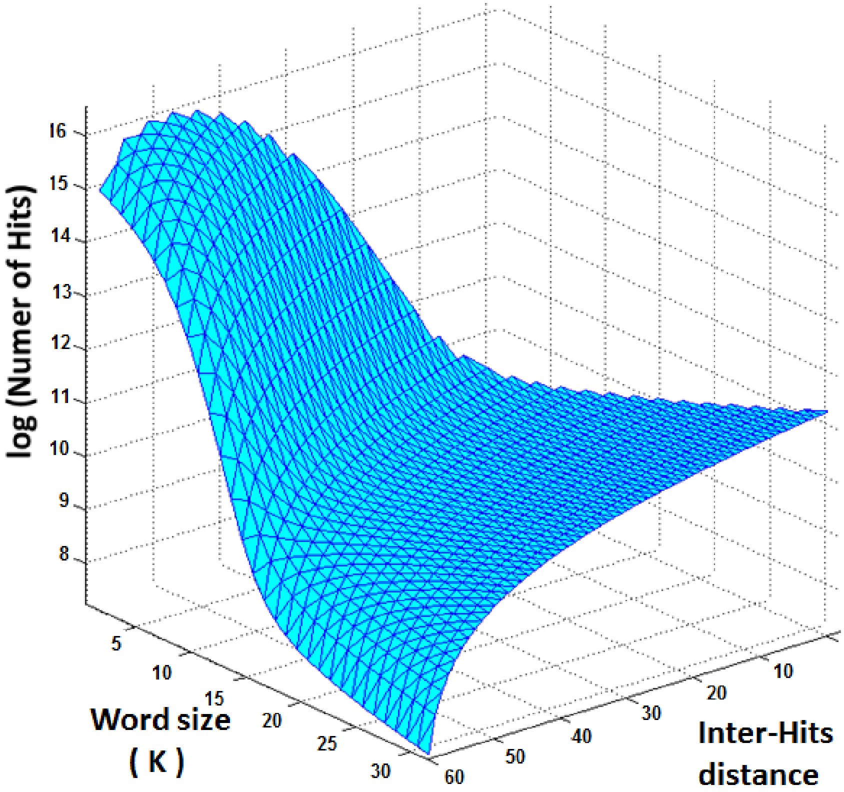
O.Trelles, PhD, 2014

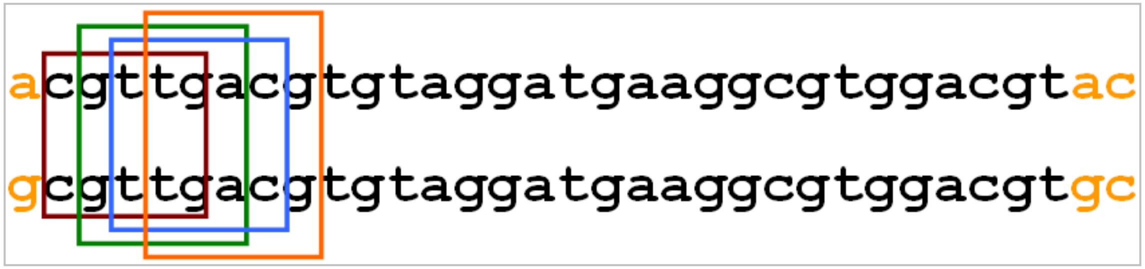
Computational space reduction

Retrieve compounds in a database that are similar to a query compound

It’s not only a problem of siize but complexity







acgttgacgtgtaggatgaaggcgtggacgtac

gcgttgacgtgtaggatgaaggcgtggacgtgc

O.Trelles, PhD, 2014

Thank you

O.Trelles, PhD, 2014