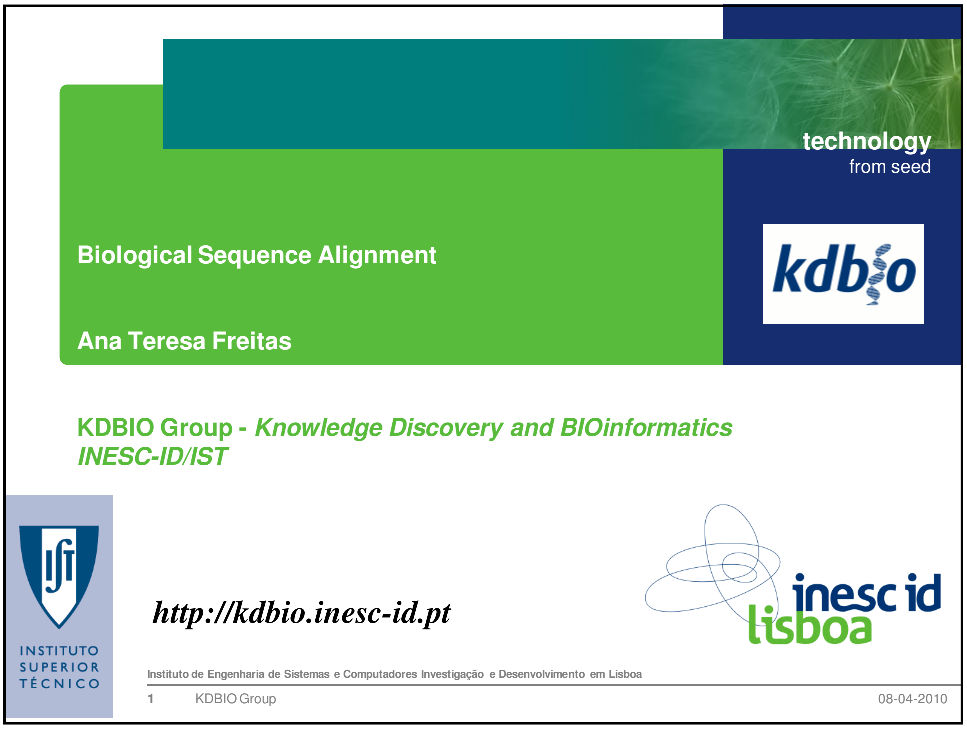


technology
from seed

Biological Sequence Alignment

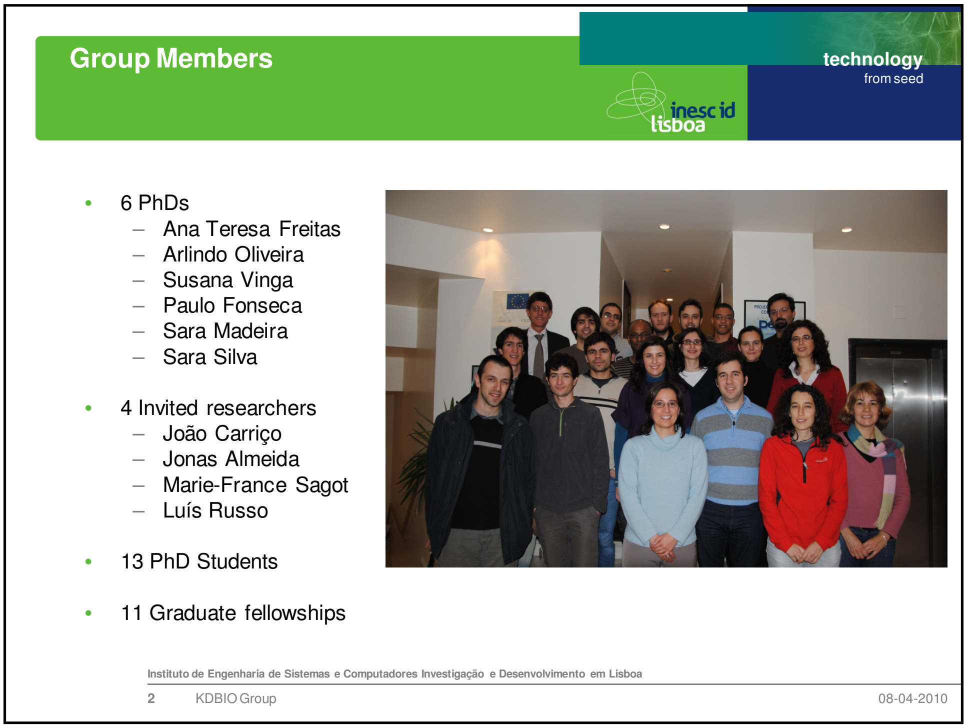
Ana Teresa Freitas

KDBIO Group - *Knowledge Discovery and BIOinformatics*
INESC-ID/IST

http://kdbio.inesc-id.pt

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

1 KDBIO Group 08-04-2010



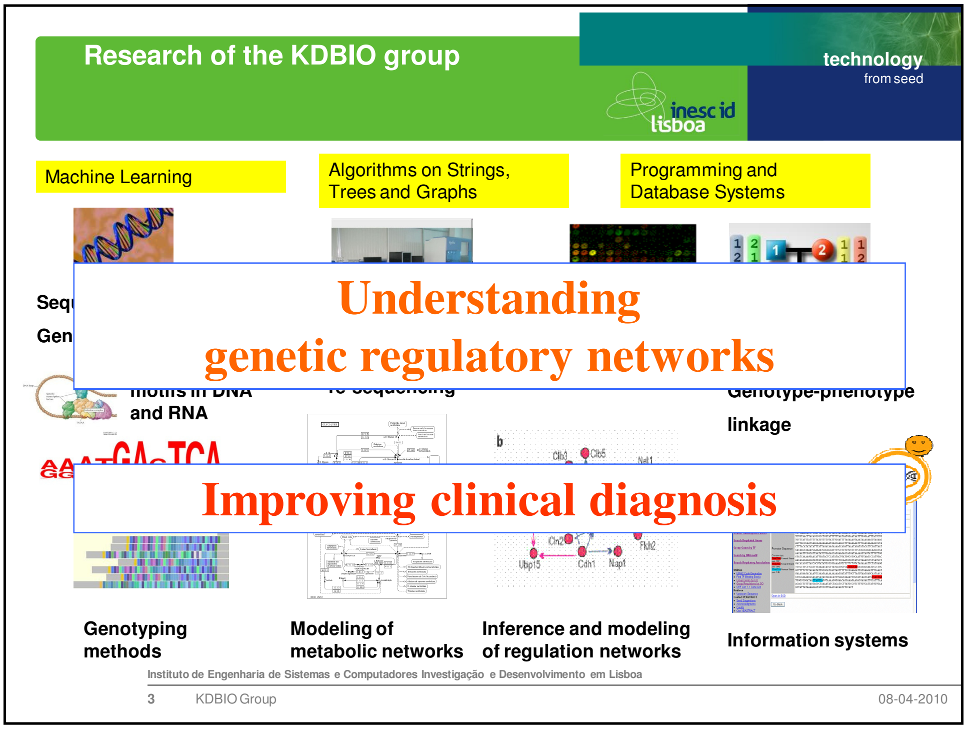
Group Members

technology
from seed

- 6 PhDs
  - Ana Teresa Freitas
  - Arlindo Oliveira
  - Susana Vinga
  - Paulo Fonseca
  - Sara Madeira
  - Sara Silva
- 4 Invited researchers
  - João Carriço
  - Jonas Almeida
  - Marie-France Sagot
  - Luís Russo
- 13 PhD Students
- 11 Graduate fellowships

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

2 KDBIO Group 08-04-2010



Research of the KDBIO group

Machine Learning

Algorithms on Strings, Trees and Graphs

Programming and Database Systems

Seq

Gen

Understanding genetic regulatory networks

motifs in DNA and RNA

re-sequencing

Genotype-phenotype linkage

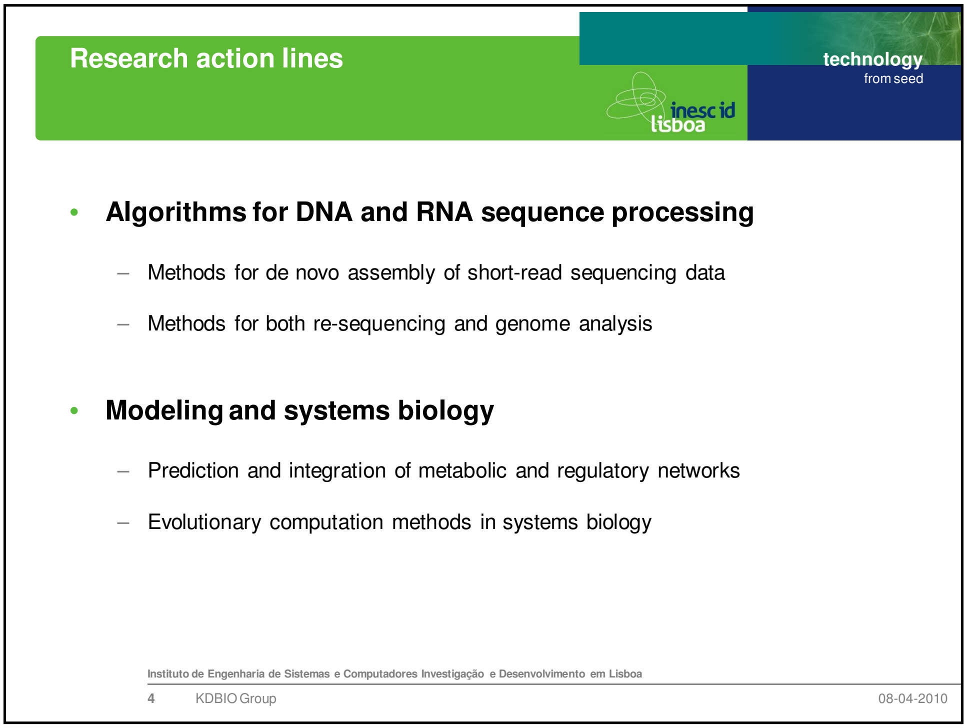
Improving clinical diagnosis

Genotyping methods

Modeling of metabolic networks

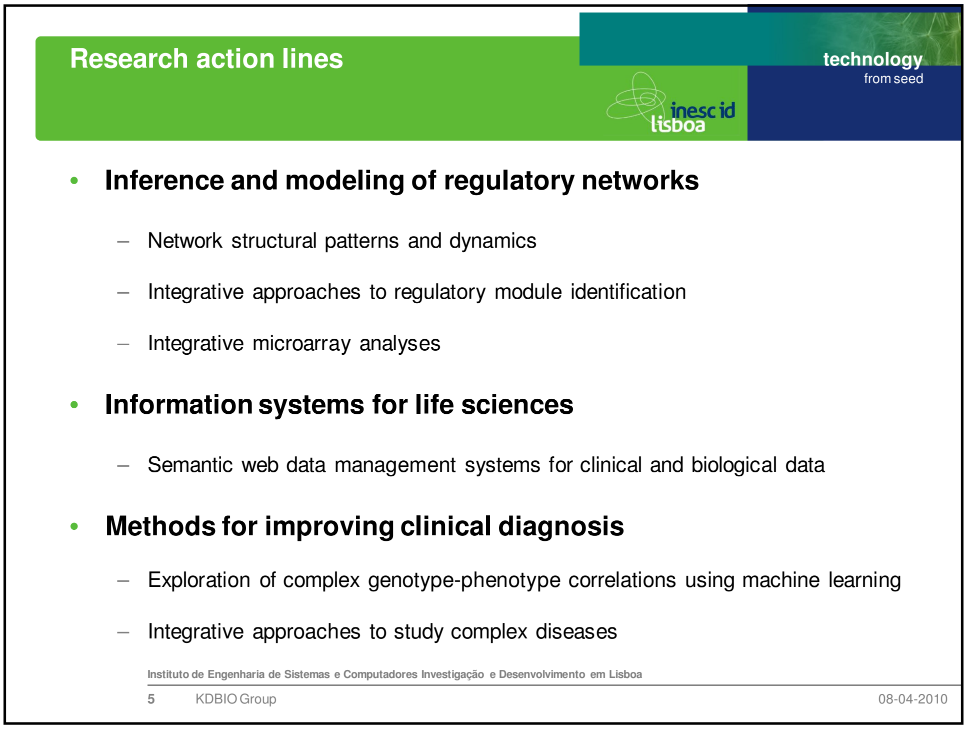
Inference and modeling of regulation networks

Information systems



Research action lines

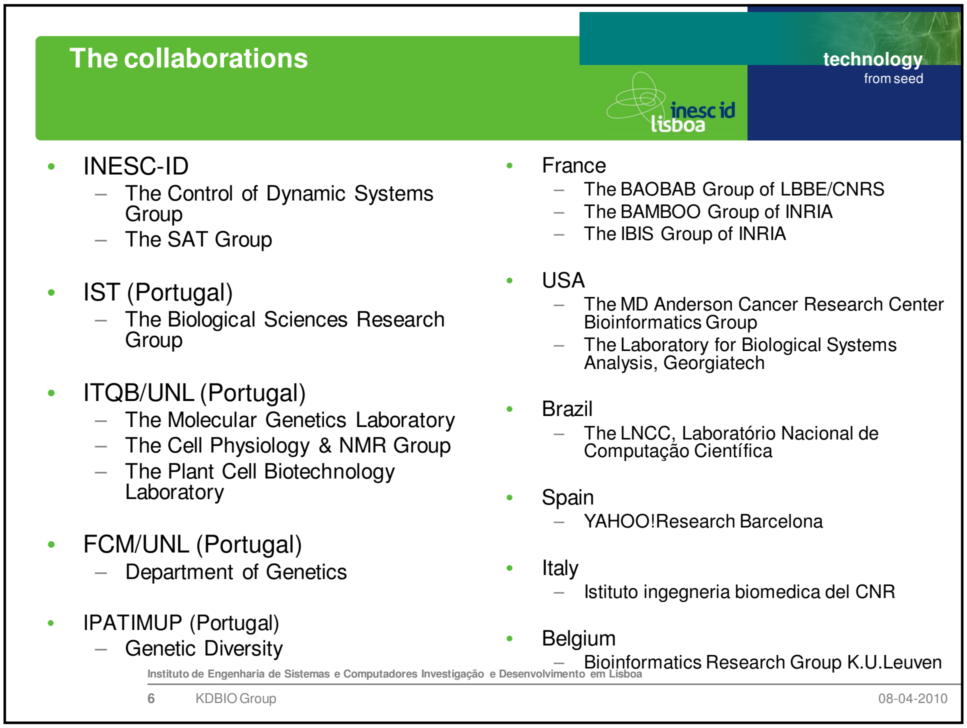
- **Algorithms for DNA and RNA sequence processing**
  - Methods for de novo assembly of short-read sequencing data
  - Methods for both re-sequencing and genome analysis
- **Modeling and systems biology**
  - Prediction and integration of metabolic and regulatory networks
  - Evolutionary computation methods in systems biology



Research action lines

technology
from seed

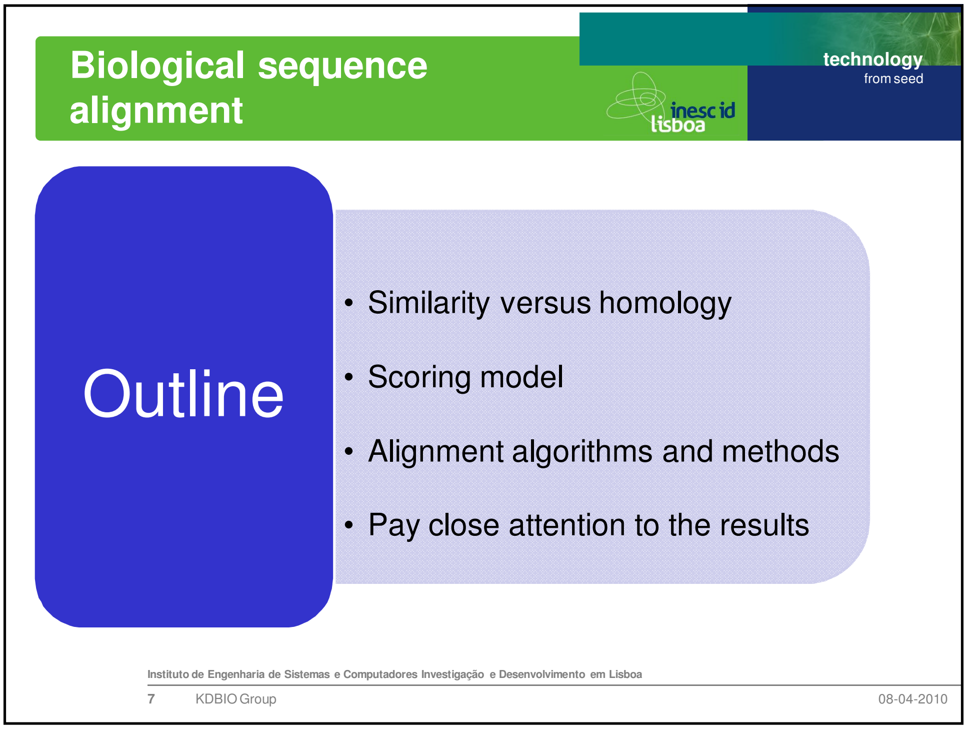
- **Inference and modeling of regulatory networks**
  - Network structural patterns and dynamics
  - Integrative approaches to regulatory module identification
  - Integrative microarray analyses
- **Information systems for life sciences**
  - Semantic web data management systems for clinical and biological data
- **Methods for improving clinical diagnosis**
  - Exploration of complex genotype-phenotype correlations using machine learning
  - Integrative approaches to study complex diseases



The collaborations

technology
from seed

- INESC-ID
  - The Control of Dynamic Systems Group
  - The SAT Group
- IST (Portugal)
  - The Biological Sciences Research Group
- ITQB/UNL (Portugal)
  - The Molecular Genetics Laboratory
  - The Cell Physiology & NMR Group
  - The Plant Cell Biotechnology Laboratory
- FCM/UNL (Portugal)
  - Department of Genetics
- IPATIMUP (Portugal)
  - Genetic Diversity
- France
  - The BAOBAB Group of LBBE/CNRS
  - The BAMBOO Group of INRIA
  - The IBIS Group of INRIA
- USA
  - The MD Anderson Cancer Research Center Bioinformatics Group
  - The Laboratory for Biological Systems Analysis, Georgiatech
- Brazil
  - The LNCC, Laboratório Nacional de Computação Científica
- Spain
  - YAHOO!Research Barcelona
- Italy
  - Istituto ingegneria biomedica del CNR
- Belgium
  - Bioinformatics Research Group K.U.Leuven



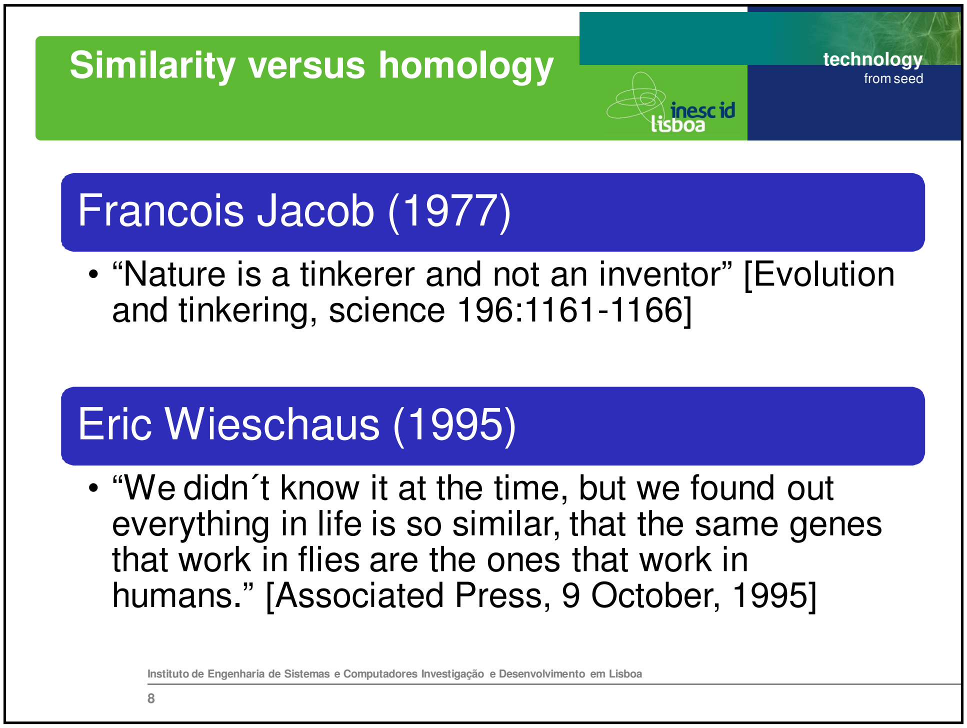
Biological sequence alignment

technology from seed

inesc id lisboa

Outline

- Similarity versus homology
- Scoring model
- Alignment algorithms and methods
- Pay close attention to the results



Similarity versus homology

technology from seed

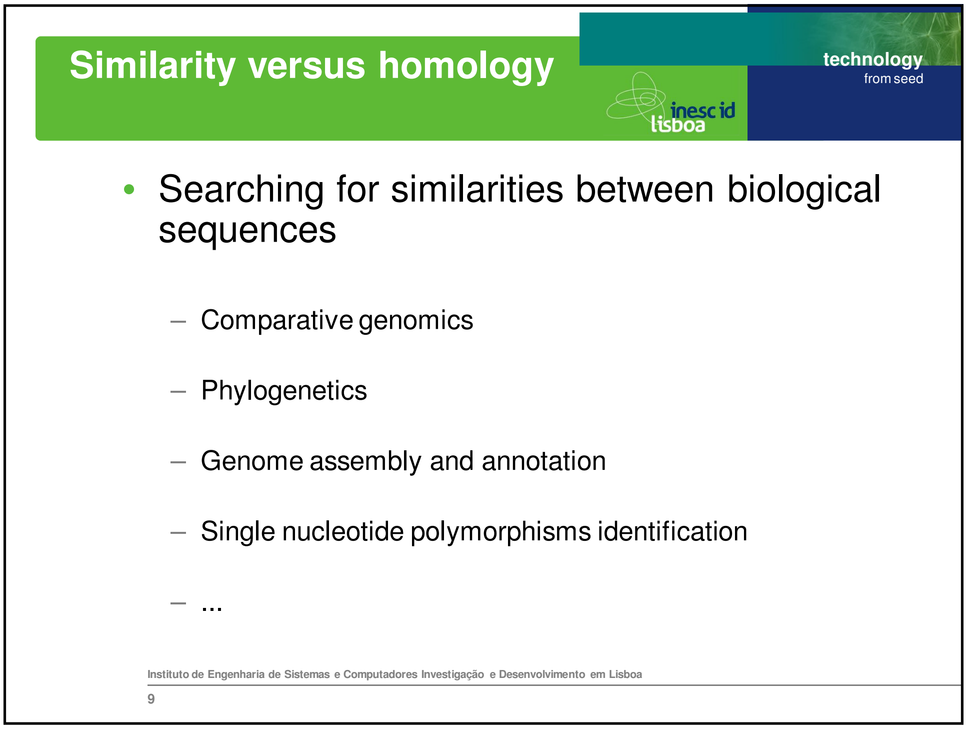
inesc id lisboa

Francois Jacob (1977)

- “Nature is a tinkerer and not an inventor” [Evolution and tinkering, science 196:1161-1166]

Eric Wieschaus (1995)

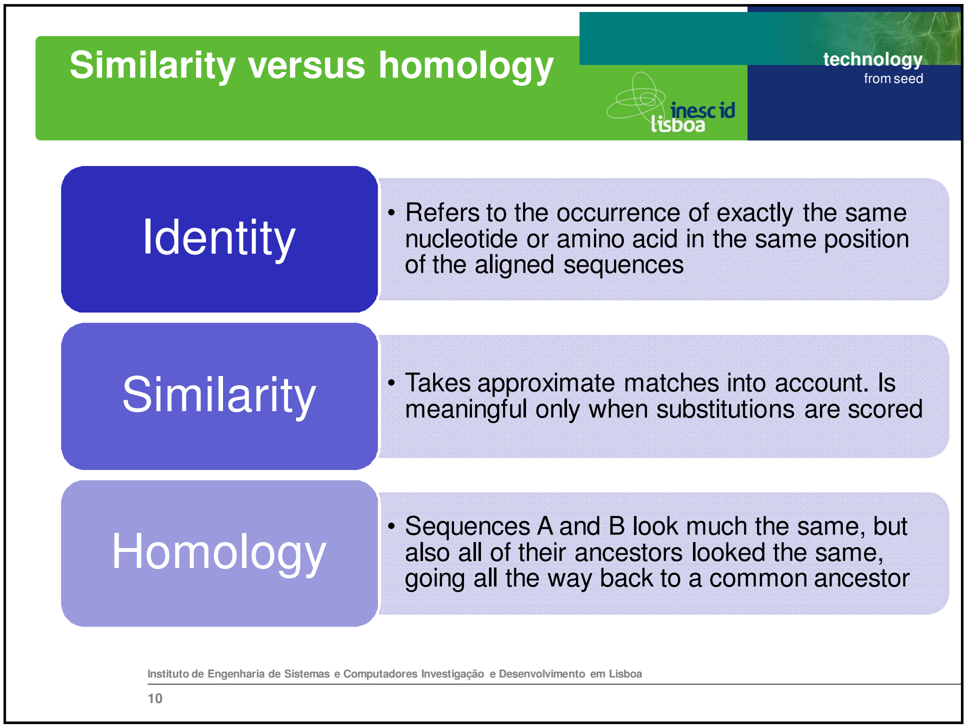
- “We didn´t know it at the time, but we found out everything in life is so similar, that the same genes that work in flies are the ones that work in humans.” [Associated Press, 9 October, 1995]



Similarity versus homology

technology
from seed
inesc id
lisboa

- Searching for similarities between biological sequences
  - Comparative genomics
  - Phylogenetics
  - Genome assembly and annotation
  - Single nucleotide polymorphisms identification
  - ...



Similarity versus homology

technology
from seed
inesc id
lisboa

Identity

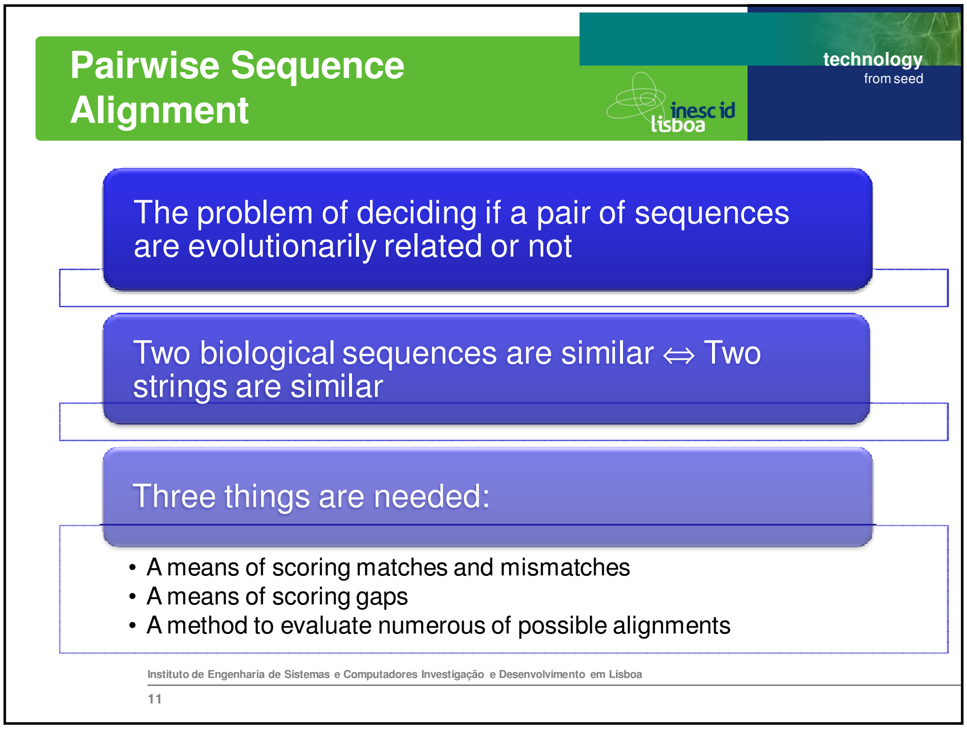
- Refers to the occurrence of exactly the same nucleotide or amino acid in the same position of the aligned sequences

Similarity

- Takes approximate matches into account. Is meaningful only when substitutions are scored

Homology

- Sequences A and B look much the same, but also all of their ancestors looked the same, going all the way back to a common ancestor



Pairwise Sequence Alignment

technology from seed

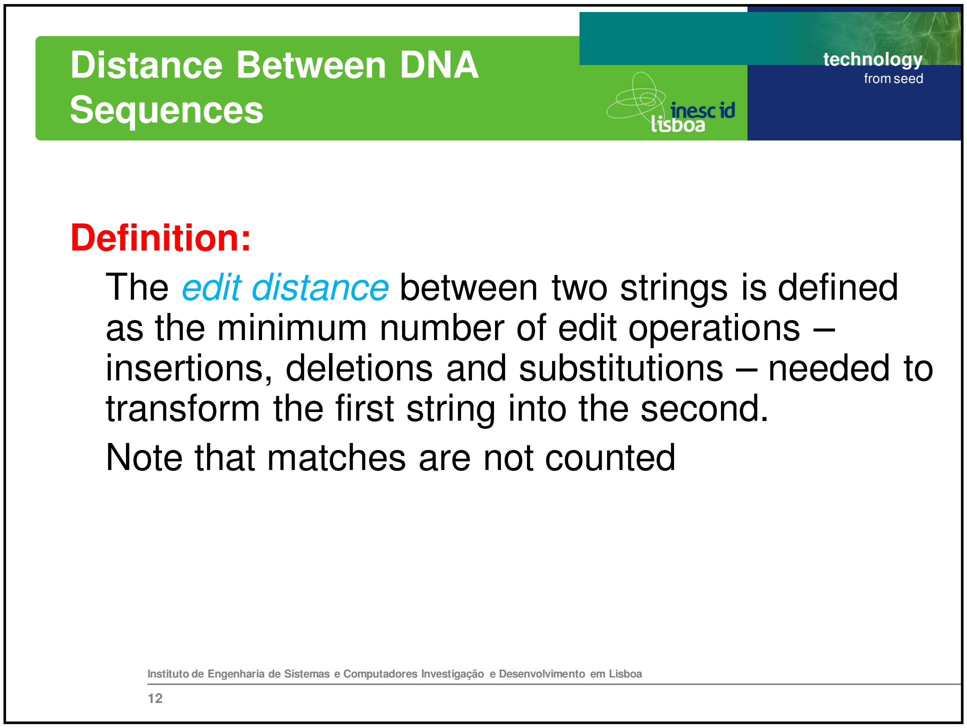
inesc id lisboa

The problem of deciding if a pair of sequences are evolutionarily related or not

Two biological sequences are similar ⇔ Two strings are similar

Three things are needed:

- A means of scoring matches and mismatches
- A means of scoring gaps
- A method to evaluate numerous of possible alignments



Distance Between DNA Sequences

technology from seed

inesc id lisboa

Definition:

The *edit distance* between two strings is defined as the minimum number of edit operations – insertions, deletions and substitutions – needed to transform the first string into the second.

Note that matches are not counted

Gaps

technology from seed

inesc id lisboa

- Gaps help create alignments that better conform to underlying biological models
- Mechanisms that make long insertions or deletions in DNA include: unequal crossing-over in meiosis; DNA slippage during replication; insertion of transposable elements into DNA string; insertions of DNA by retro-viruses; etc...

Definition: A *gap* is any maximal, consecutive run of spaces in a single string of a given alignment

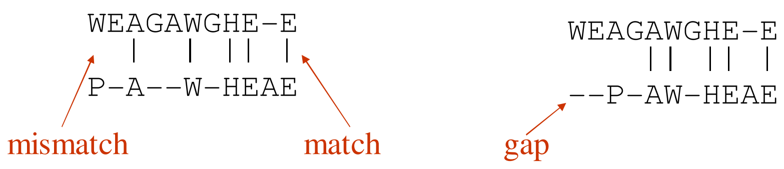
Pairwise Sequence Alignment

technology from seed

inesc id lisboa

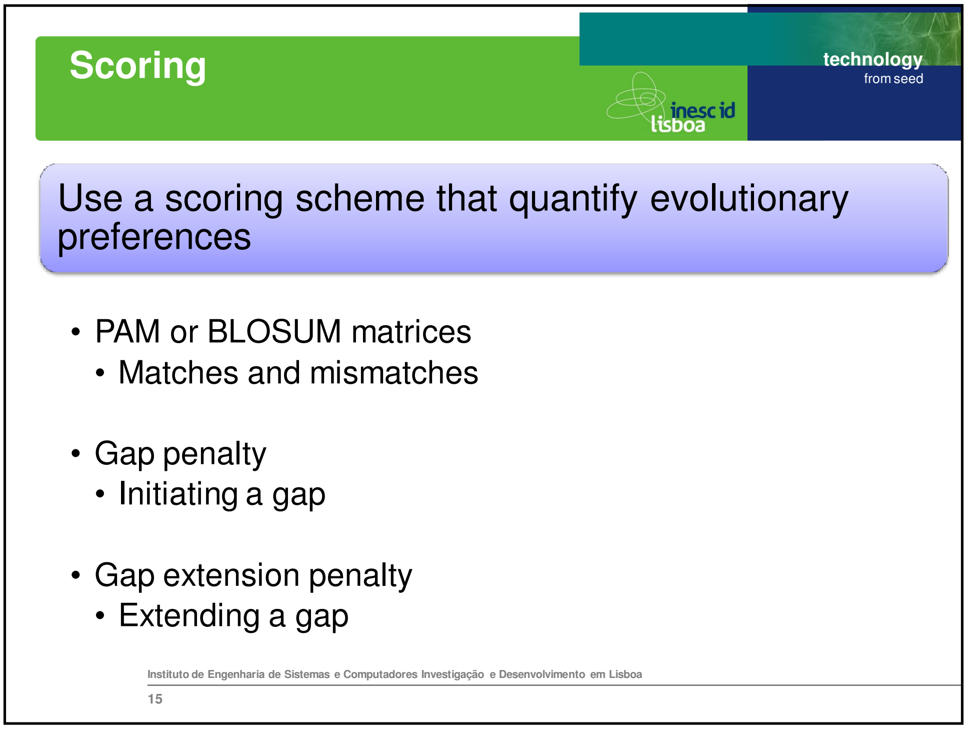
- Example:

```
WEAGAWGHEE
PAWHEAE
```



```
WEAGAWGHE-E              WEAGAWGHE-E
  |  | || |                  || || |
P-A--W-HEAE              --P-AW-HEAE
```

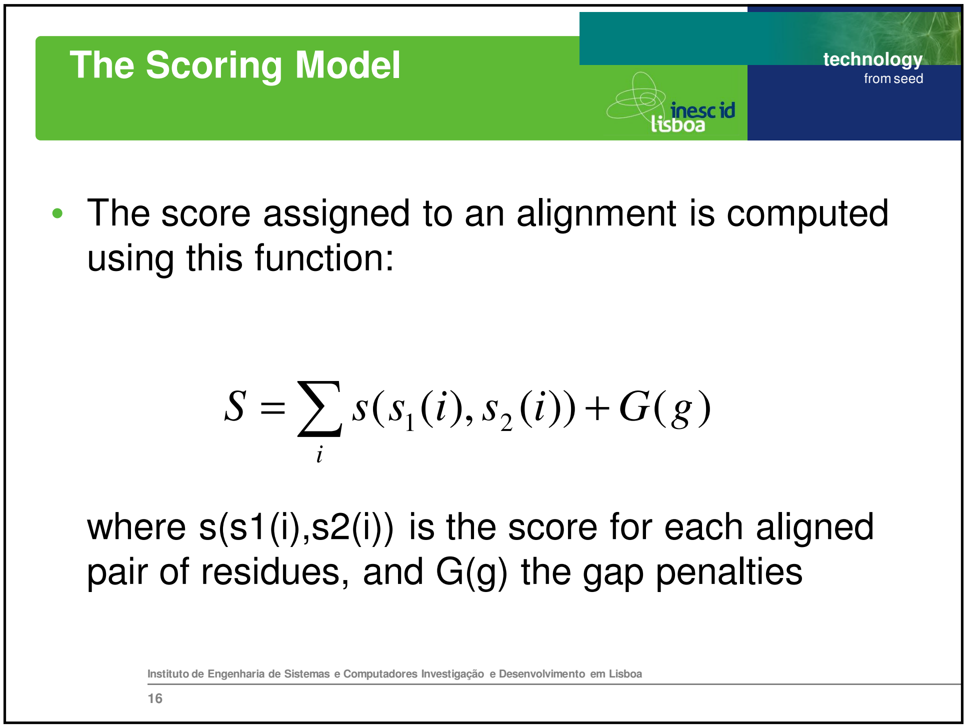
  - Which one is better?
  - Is it a true or a spurious alignment?



Scoring

Use a scoring scheme that quantify evolutionary preferences

- PAM or BLOSUM matrices
  - Matches and mismatches
- Gap penalty
  - Initiating a gap
- Gap extension penalty
  - Extending a gap

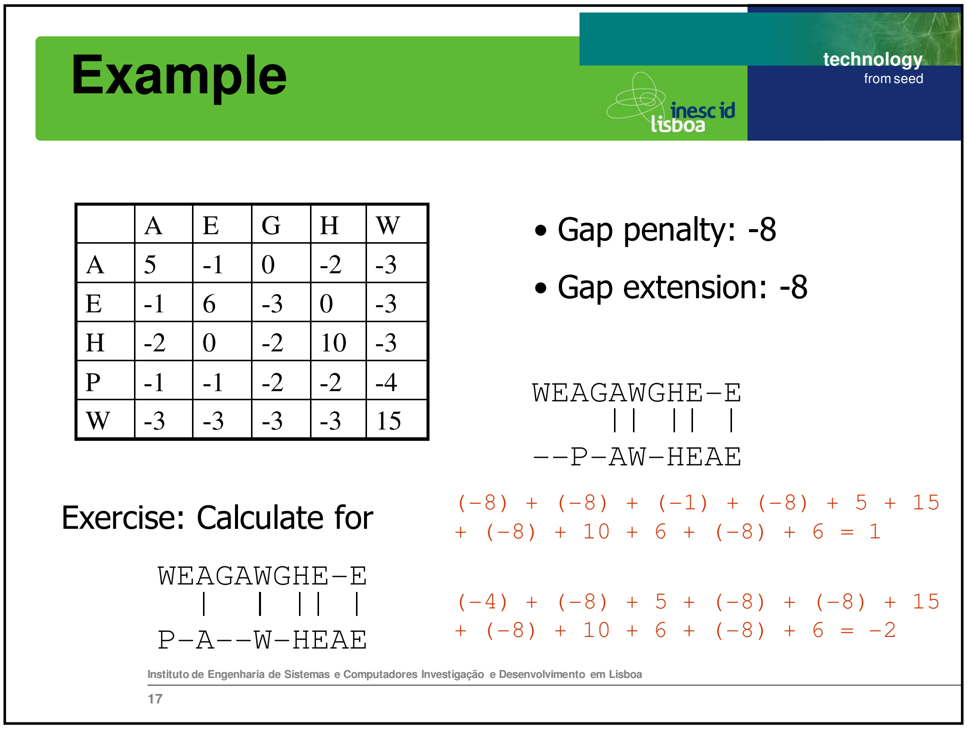


The Scoring Model

- The score assigned to an alignment is computed using this function:

$$S = \sum_i s(s_1(i), s_2(i)) + G(g)$$

where s(s1(i),s2(i)) is the score for each aligned pair of residues, and G(g) the gap penalties



Example

technology from seed

inesc id lisboa

	A	E	G	H	W
A	5	-1	0	-2	-3
E	-1	6	-3	0	-3
H	-2	0	-2	10	-3
P	-1	-1	-2	-2	-4
W	-3	-3	-3	-3	15

- Gap penalty: -8
- Gap extension: -8

```
WEAGAWGHE-E
    ||  ||  |
--P-AW-HEAE
```

(-8) + (-8) + (-1) + (-8) + 5 + 15 + (-8) + 10 + 6 + (-8) + 6 = 1

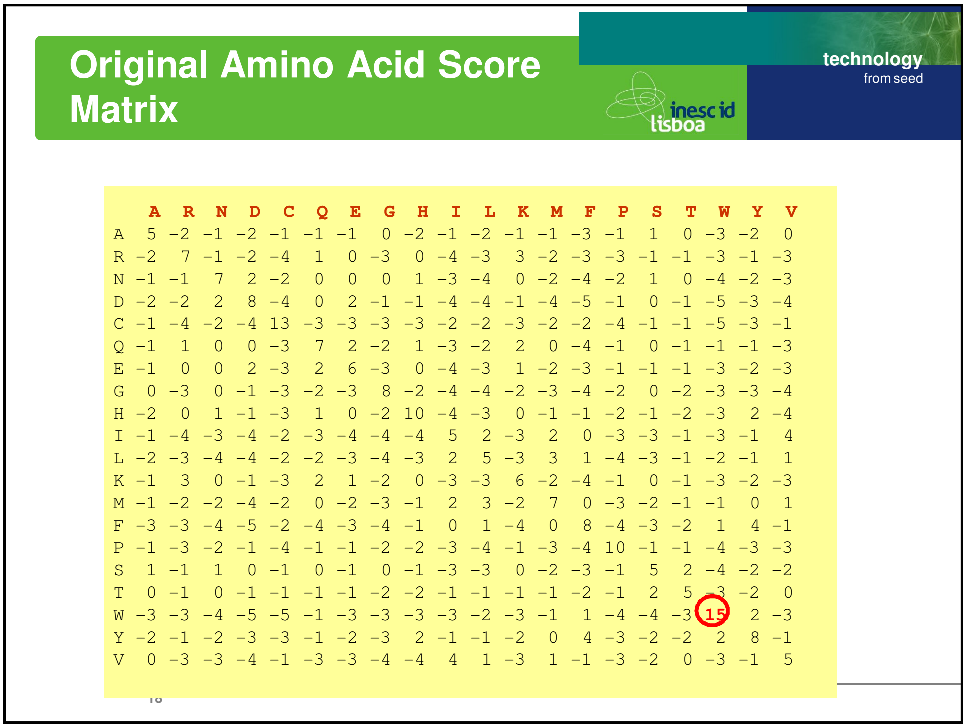
Exercise: Calculate for

```
WEAGAWGHE-E
  |   |  ||  |
P-A--W-HEAE
```

(-4) + (-8) + 5 + (-8) + (-8) + 15 + (-8) + 10 + 6 + (-8) + 6 = -2

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

17



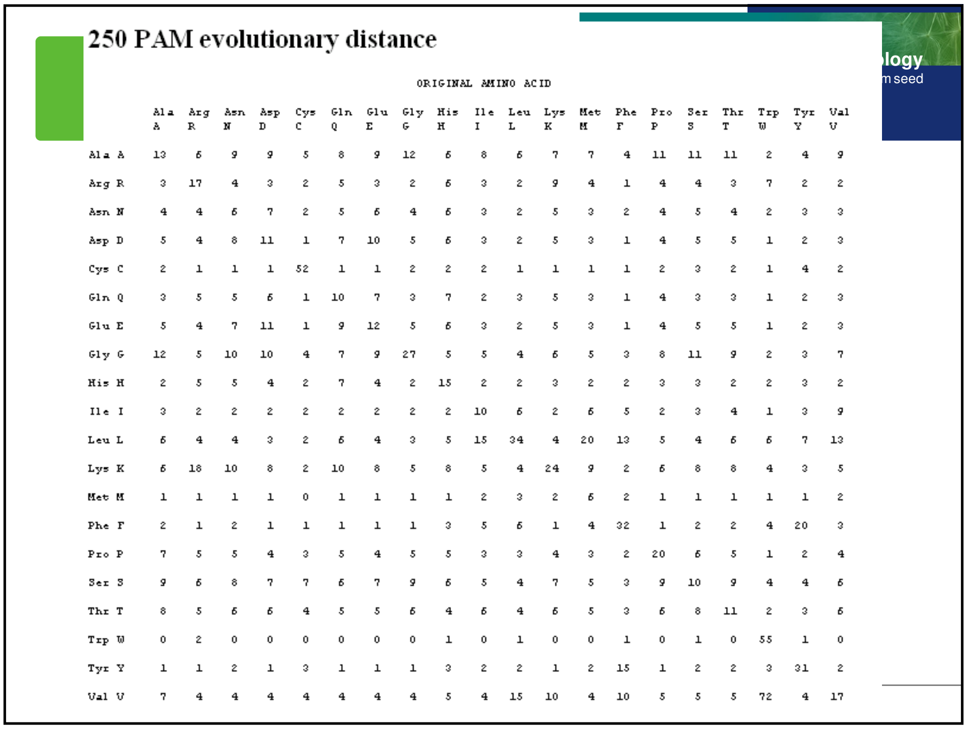
Original Amino Acid Score Matrix

technology from seed

inesc id lisboa

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

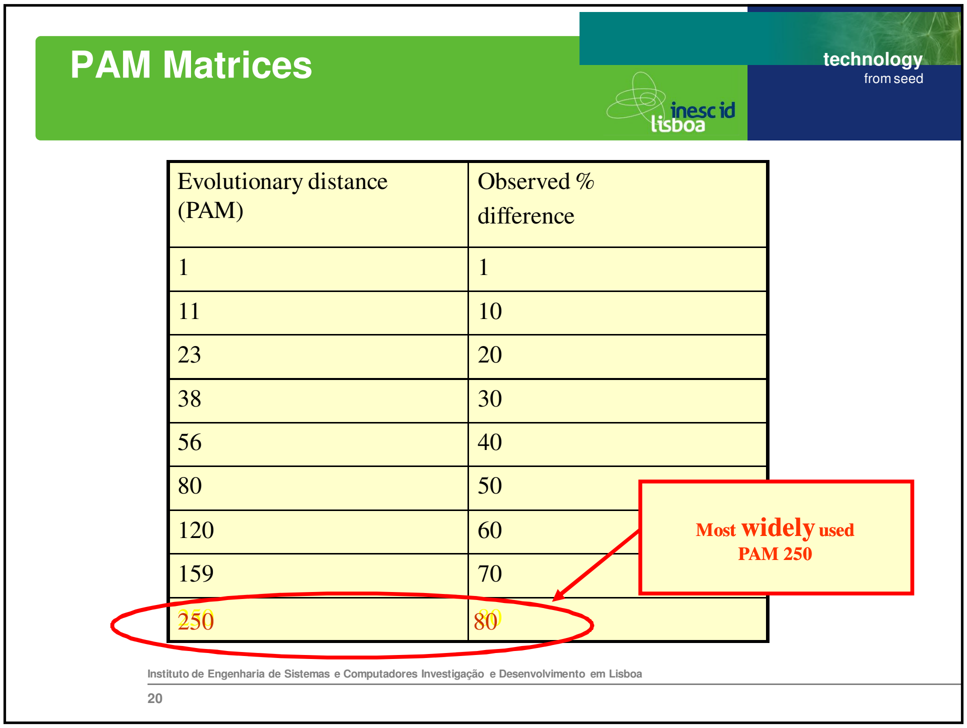
18



250 PAM evolutionary distance

ORIGINAL AMINO ACID

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17



PAM Matrices

Evolutionary distance (PAM)	Observed % difference
1	1
11	10
23	20
38	30
56	40
80	50
120	60
159	70
250	80

Most widely used PAM 250

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

20

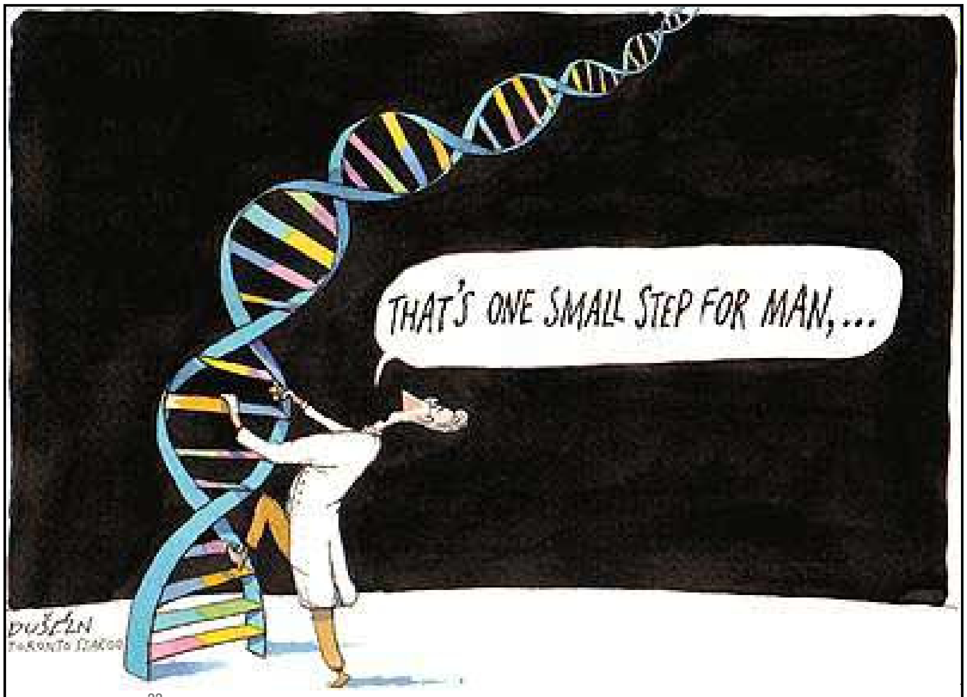
PAM vs. BLOSUM

technology from seed

- PAM model is designed to track evolutionary origin of proteins
- Blosum model is designed to find conserved domains of proteins

Thumb rules

- Lower PAMs and higher Blosums find short local alignment of highly similar sequences
- Higher PAMs and lower Blosums find longer weaker local alignment



Alignment Algorithms
How difficult is this?

technology from seed

inesc id lisboa

- Consider two sequences of length *n*
- There are

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

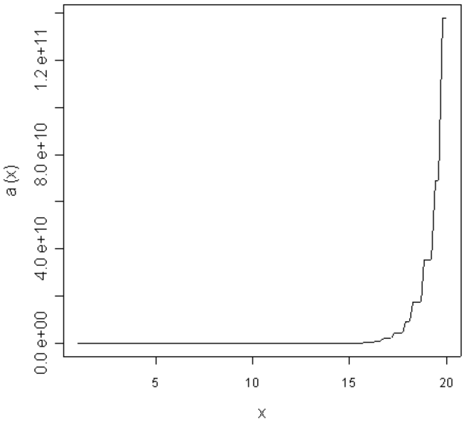
possible global alignments, and we need to find an optimal one from amongst those

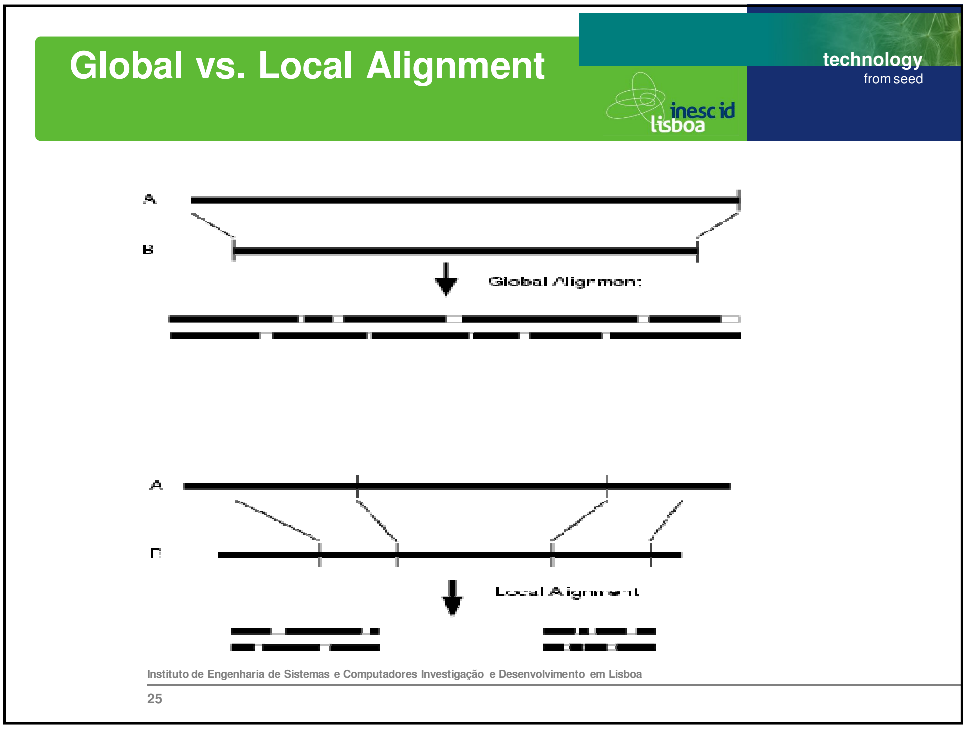
So what?

technology from seed

inesc id lisboa

- So at *n* = 20, we have over 120 billion possible alignments
- We want to be able to align much, much longer sequences
  - some proteins have 1000 amino acids
  - genes can have several thousand base pairs





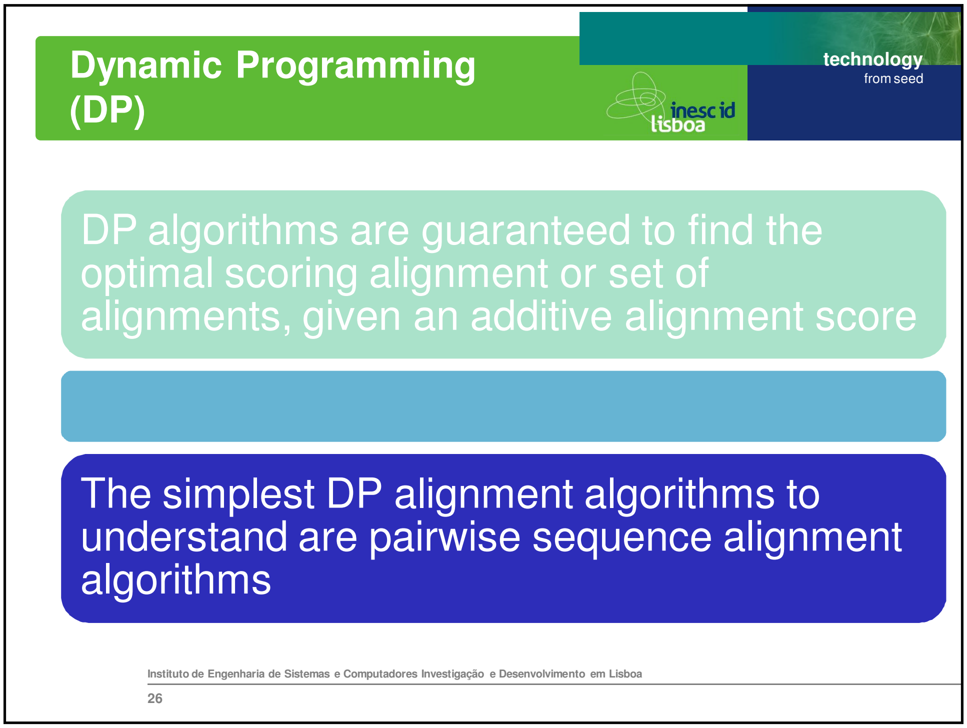
Global vs. Local Alignment

technology from seed

inesc id lisboa

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

25



Dynamic Programming (DP)

technology from seed

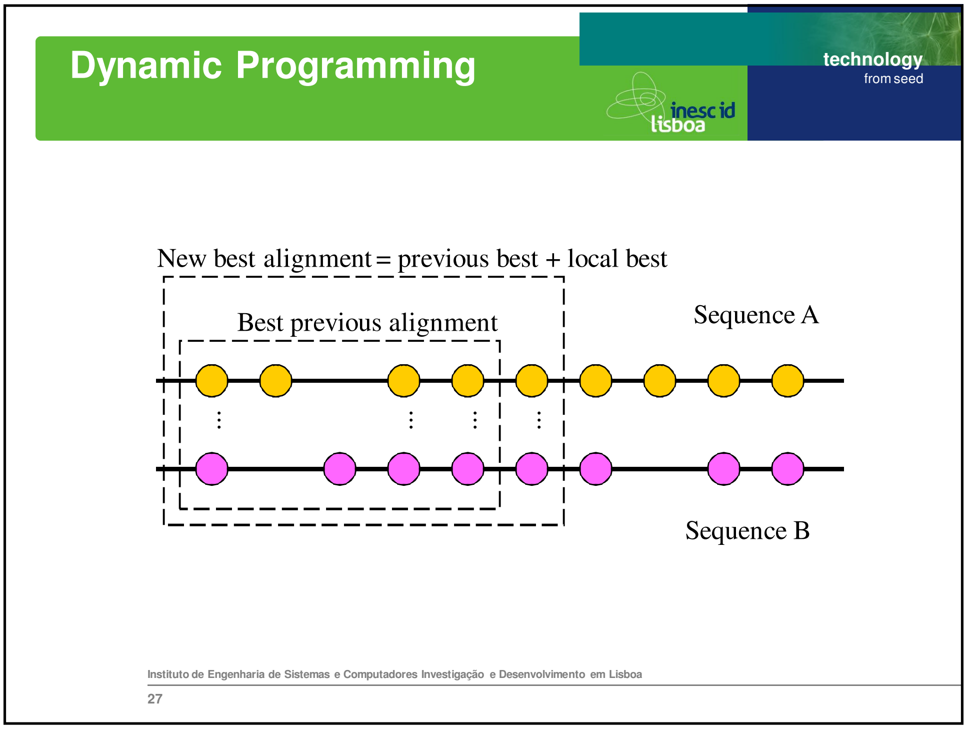
inesc id lisboa

DP algorithms are guaranteed to find the optimal scoring alignment or set of alignments, given an additive alignment score

The simplest DP alignment algorithms to understand are pairwise sequence alignment algorithms

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

26



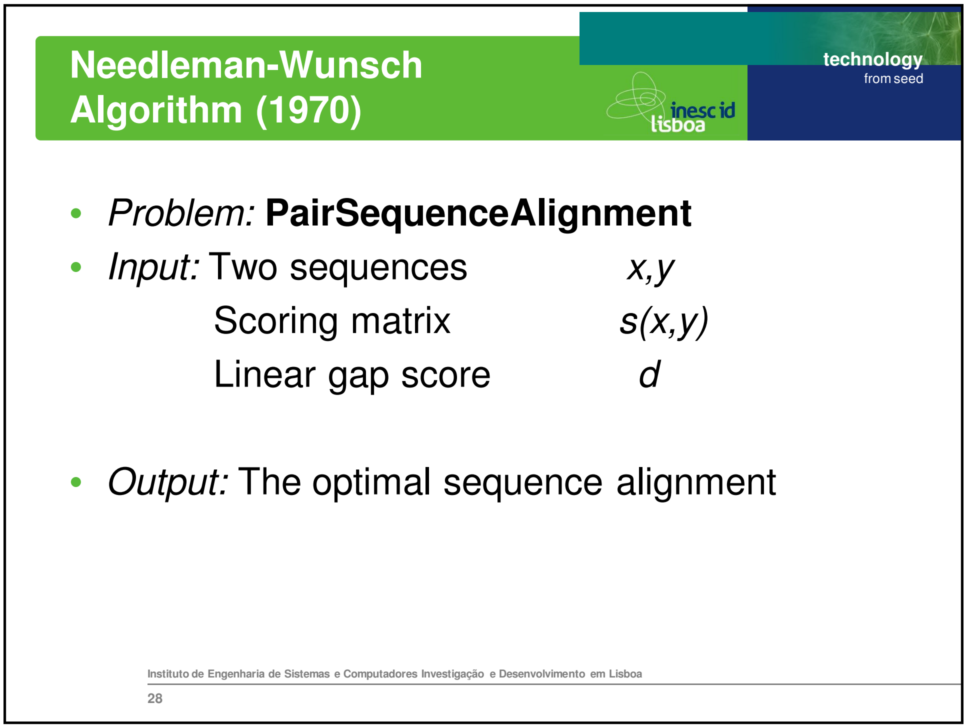
Dynamic Programming

New best alignment = previous best + local best

Best previous alignment

Sequence A

Sequence B

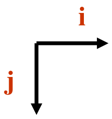


Needleman-Wunsch Algorithm (1970)

technology
from seed
inesc id lisboa

- *Problem:* **PairSequenceAlignment**
- *Input:* Two sequences x,y
  Scoring matrix $s(x,y)$
  Linear gap score d
- *Output:* The optimal sequence alignment

F matrix



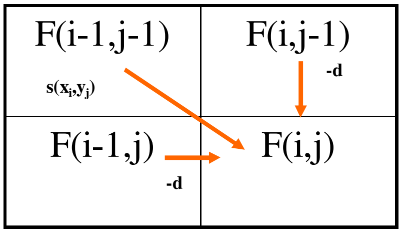
		H	E	A	G	A	W	G	H	E	E
P											
A											
W					X						
H											
E											
A											
E											

Three ways to obtain the best score F(i,j)

- x_i is aligned to y_j
- x_i is aligned to a gap
- y_j is aligned to a gap

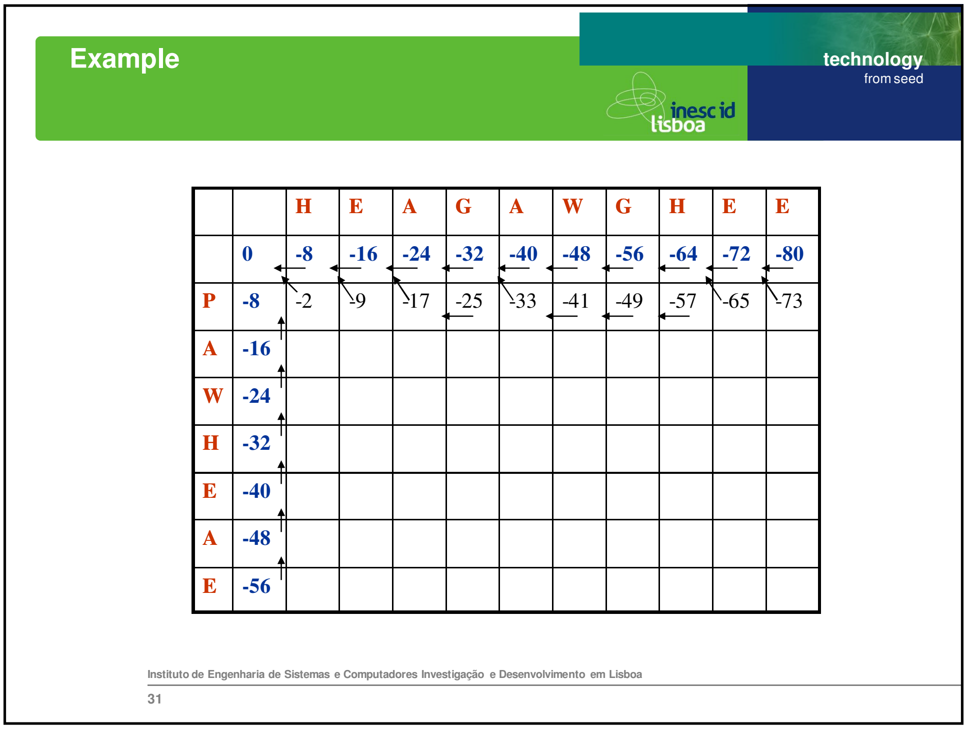
HEA x_i
-PA y_i

$$F(i,j)=F(i-1,j-1)+s(x_i,y_i)$$



- While building the table, keep track of where optimal score came from
- Initialize: F(0,0) = 0, F(i,0) = -id, F(0,j)=-jd
- Fill from top left to bottom right using the recursive relation

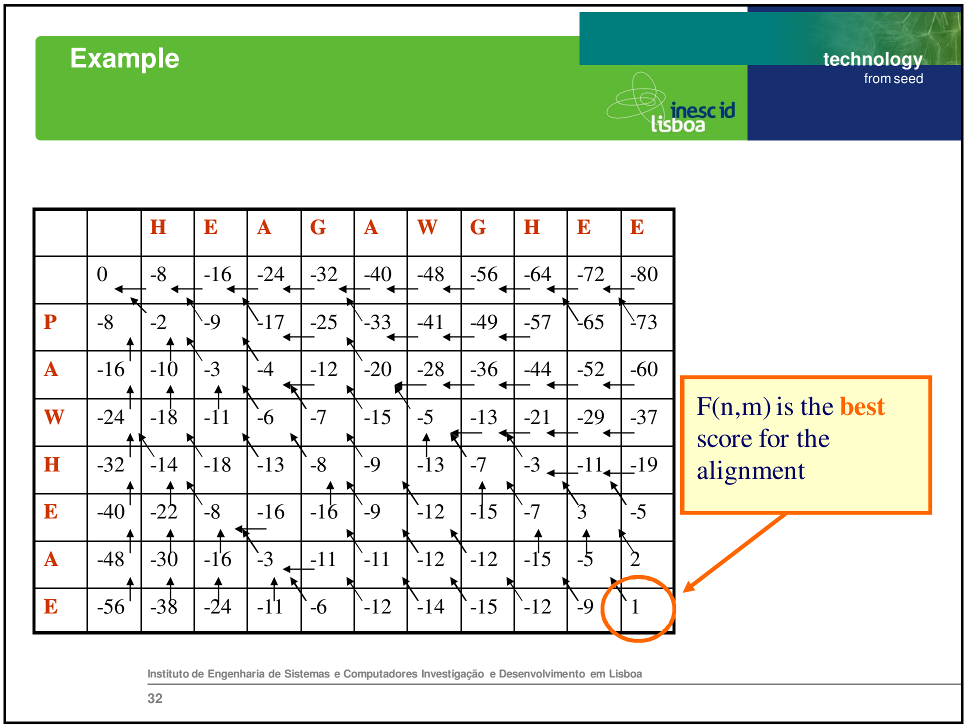
$$F(i,j)=\max\begin{cases}F(i-1,j-1)+s(x_i,y_j)\\F(i-1,j)-d\\F(i,j-1)-d\end{cases}$$



Example

technology
from seed
inesc id
lisboa

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-41	-49	-57	-65	-73
A	-16										
W	-24										
H	-32										
E	-40										
A	-48										
E	-56										

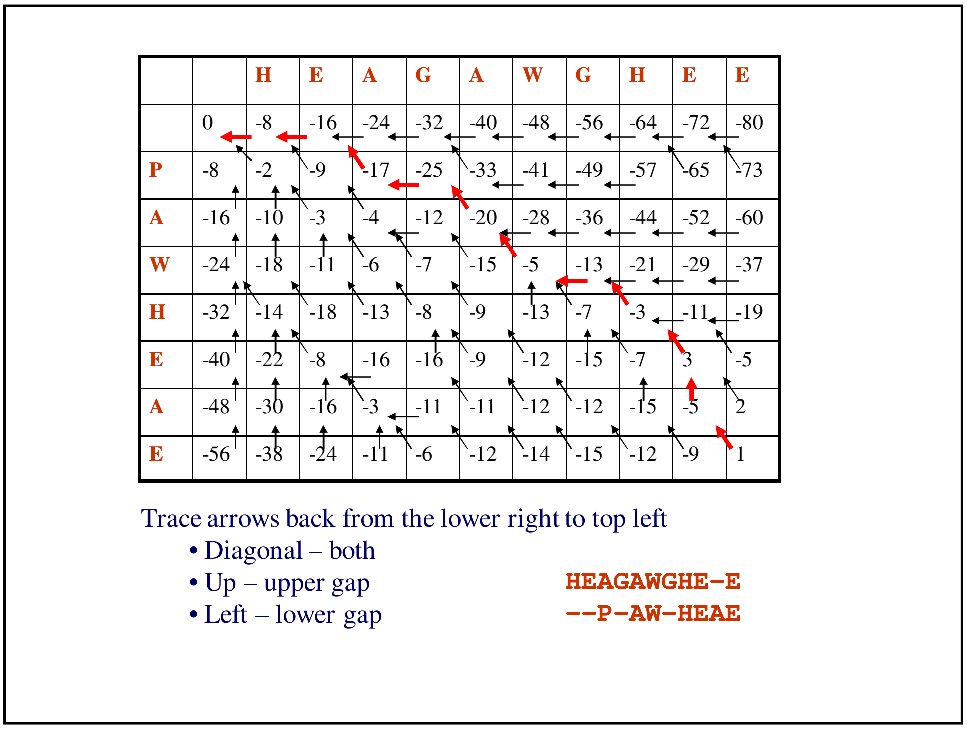


Example

technology
from seed
inesc id
lisboa

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-41	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

F(n,m) is the **best** score for the alignment

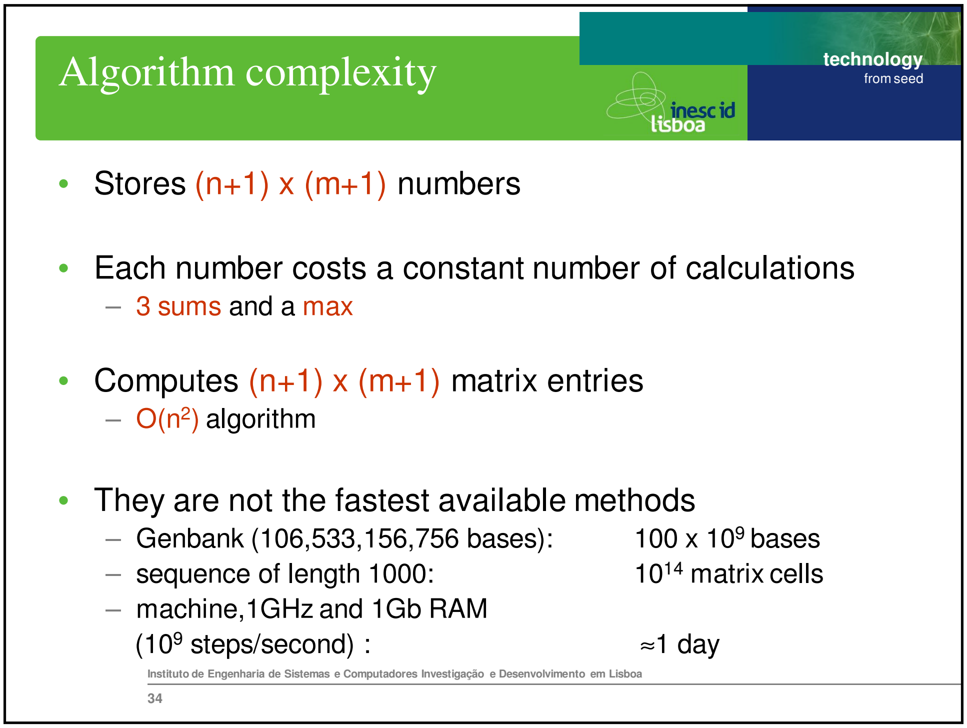


		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-41	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Trace arrows back from the lower right to top left

- Diagonal – both
- Up – upper gap
- Left – lower gap

```
HEAGAWGHE-E
--P-AW-HEAE
```



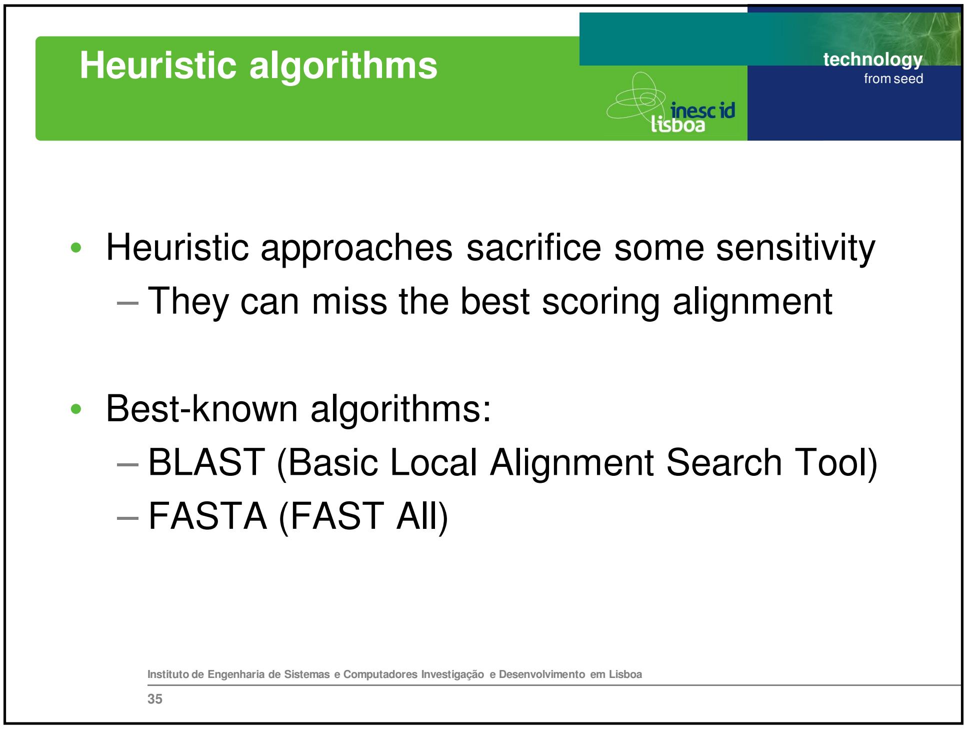
Algorithm complexity

technology
from seed
inesc id
lisboa

- Stores (n+1) x (m+1) numbers
- Each number costs a constant number of calculations
  - 3 sums and a max
- Computes (n+1) x (m+1) matrix entries
  - O(n^2) algorithm
- They are not the fastest available methods
  - Genbank (106,533,156,756 bases): 100 x 10^9 bases
  - sequence of length 1000: 10^{14} matrix cells
  - machine,1GHz and 1Gb RAM
    (10^9 steps/second) : ≈1 day

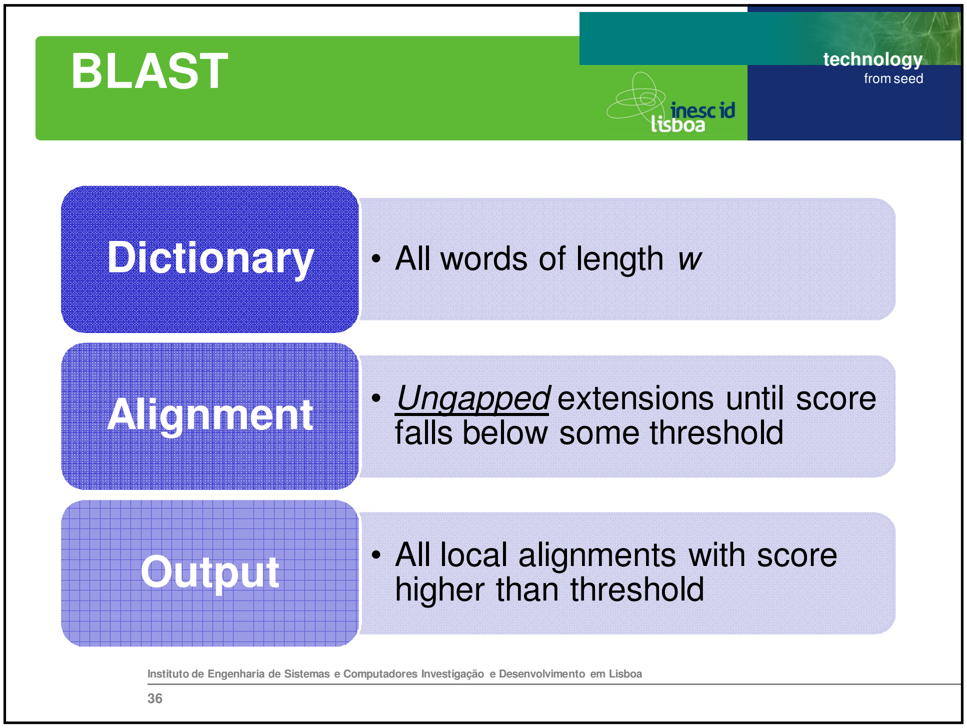
Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

34



Heuristic algorithms

- Heuristic approaches sacrifice some sensitivity
  - They can miss the best scoring alignment
- Best-known algorithms:
  - BLAST (Basic Local Alignment Search Tool)
  - FASTA (FAST All)



BLAST

Dictionary
- All words of length *w*

Alignment
- *Ungapped* extensions until score falls below some threshold

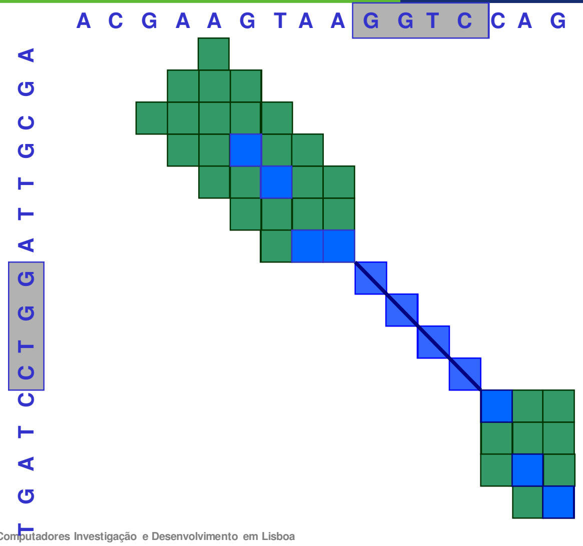
Output
- All local alignments with score higher than threshold

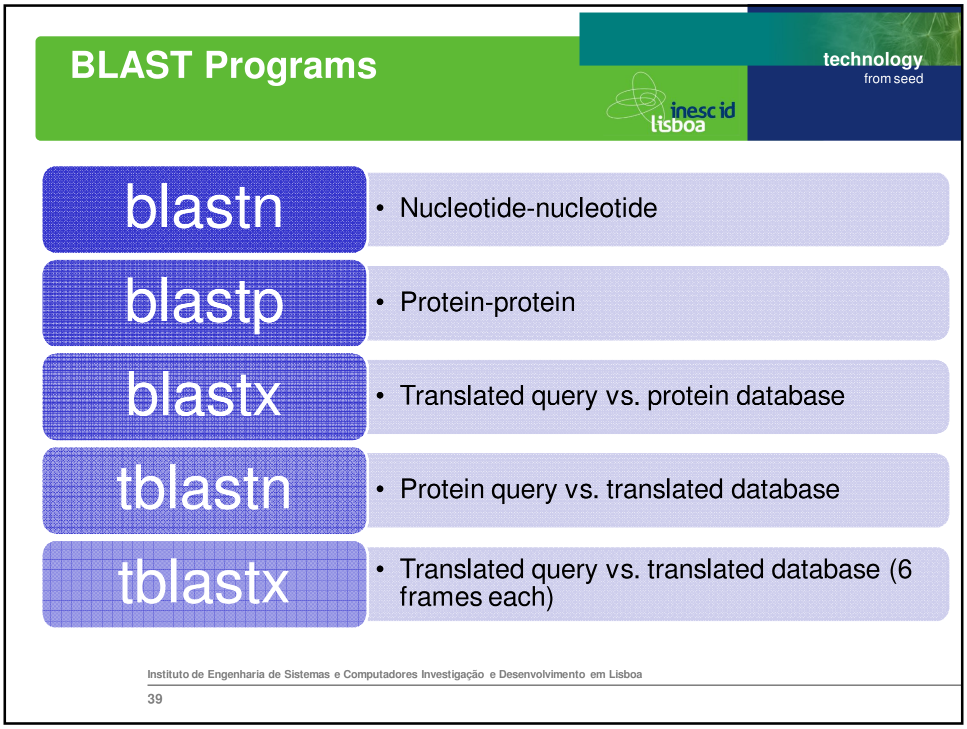
BLAST

- Make a list of *neighborhood words*
  - length 3 for proteins, 11 for nucleic acids
- Match query with score higher than some threshold
  - usually 2 bits per residue
- Scans database for words
- When a hit is obtained, extends the match in both direction as ungapped alignment

BLAST

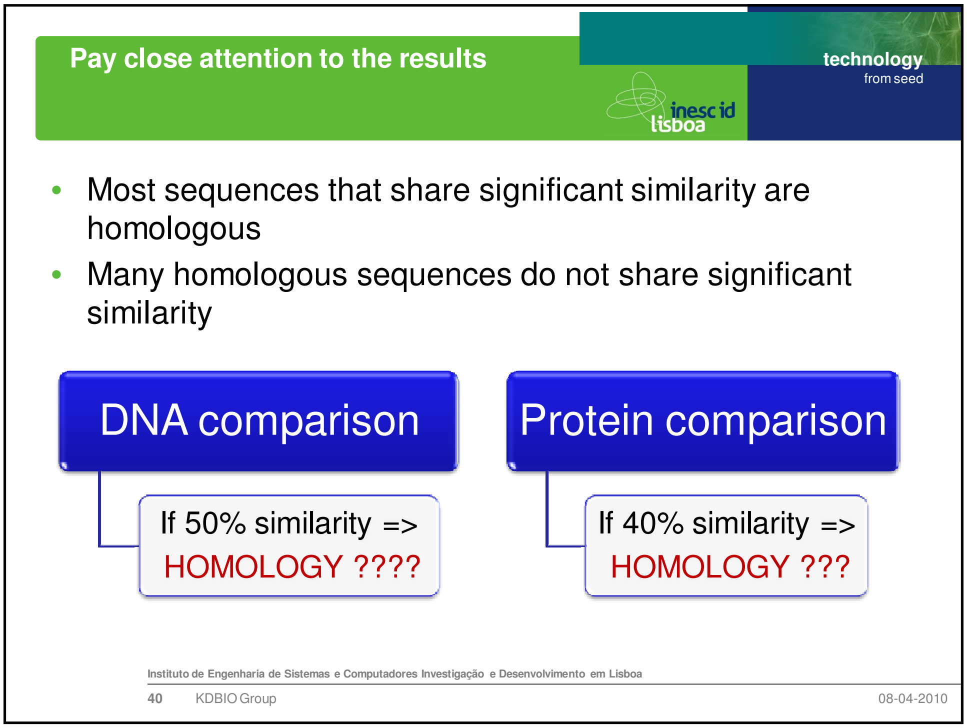
- Original BLAST exact keyword search
- Extend with gaps in a zone around ends of exact match





BLAST Programs

blastn	Nucleotide-nucleotide
blastp	Protein-protein
blastx	Translated query vs. protein database
tblastn	Protein query vs. translated database
tblastx	Translated query vs. translated database (6 frames each)



Pay close attention to the results

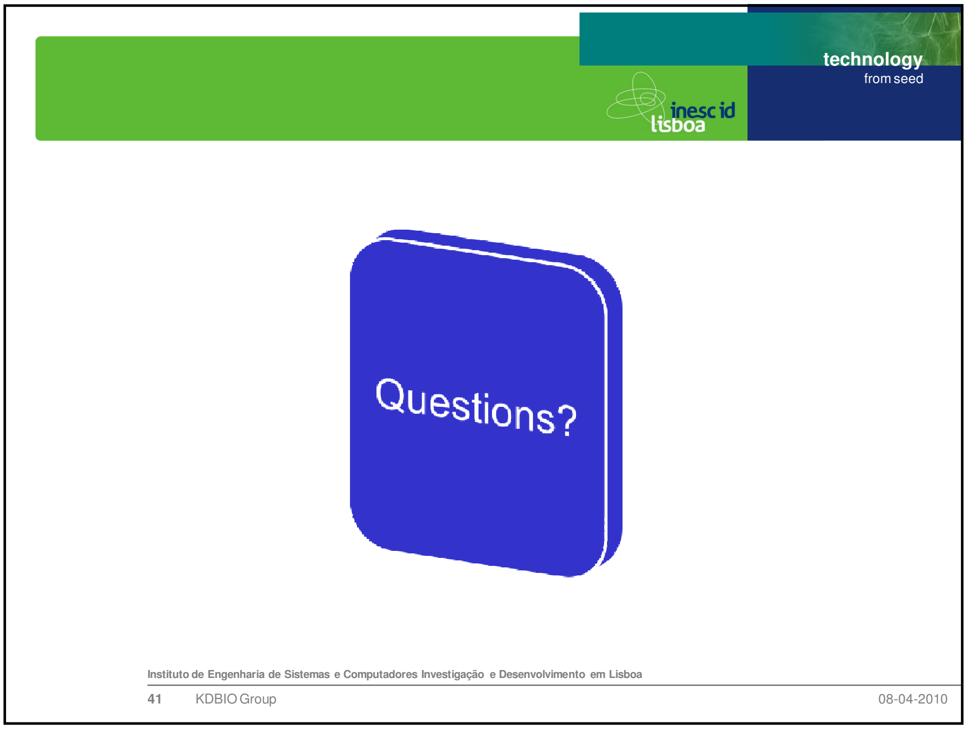
- Most sequences that share significant similarity are homologous
- Many homologous sequences do not share significant similarity

DNA comparison

If 50% similarity => HOMOLOGY ????

Protein comparison

If 40% similarity => HOMOLOGY ???



technology
from seed

inesc id lisboa

Questions?

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

41 KDBIO Group 08-04-2010