生物信息学：导论与方法

Bioinformatics: Introduction and Methods

https://www.coursera.org/course/pkubioinfo

生物信息学：导论与方法

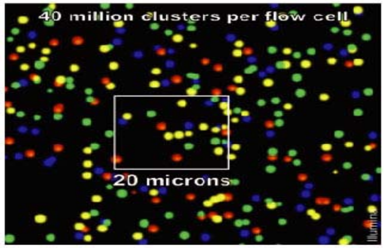
Bioinformatics: Introduction and Methods

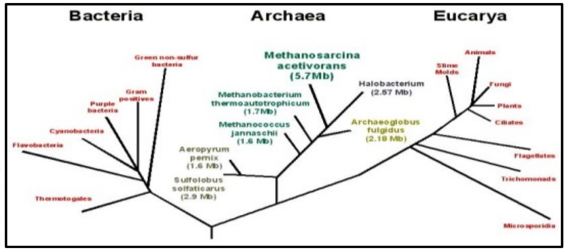
北京大学生物信息学中心 高歌、魏丽萍

Ge Gao & Liping Wei

Center for Bioinformatics, Peking University

```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



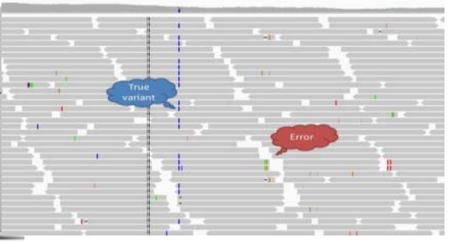

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Halobacterium (2.57 Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfaticarus (2.9 Mb)
Animals
Slime Molds
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

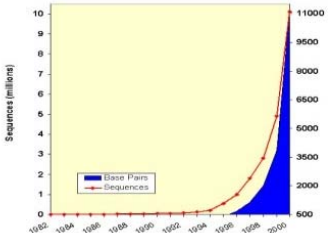
Sequence Alignment

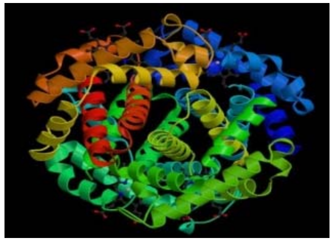
北京大学生物信息学中心 高歌

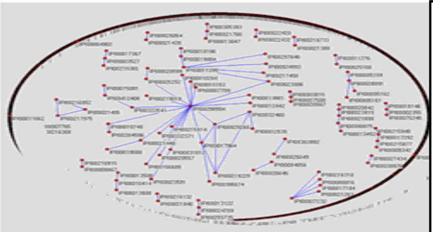
Ge Gao, Ph.D.

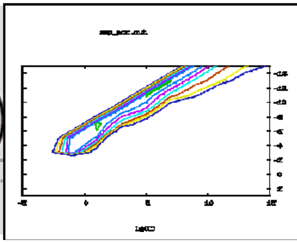
Center for Bioinformatics, Peking University



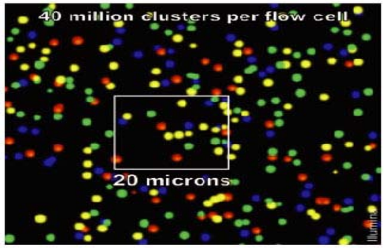


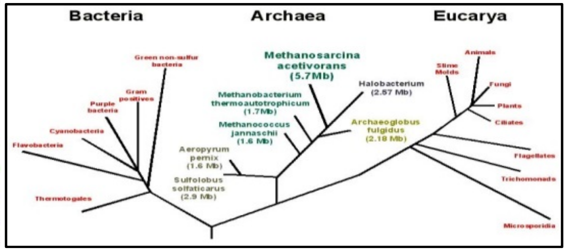






```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



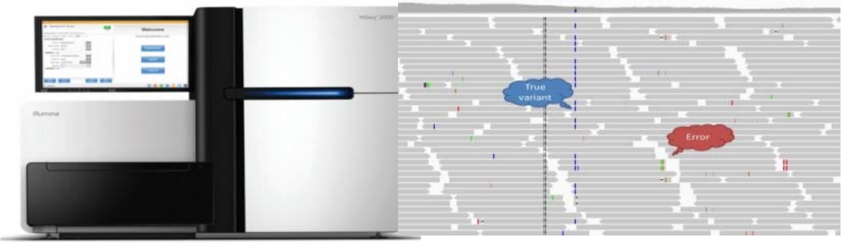

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfaticarus (2.9 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Slime Molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

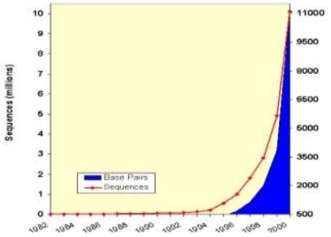
Unit 1: Essential Concepts

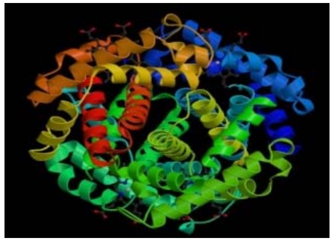
北京大学生物信息学中心 高歌

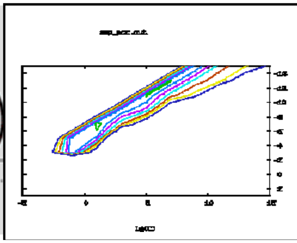
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



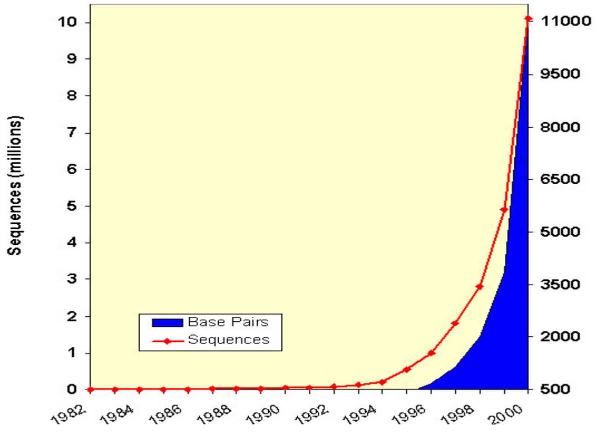




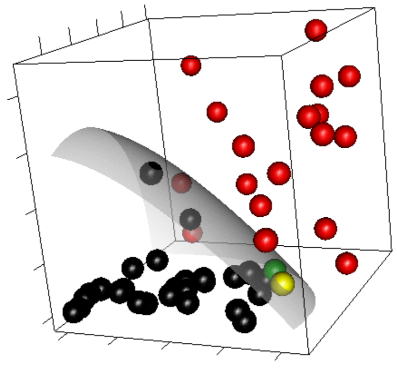


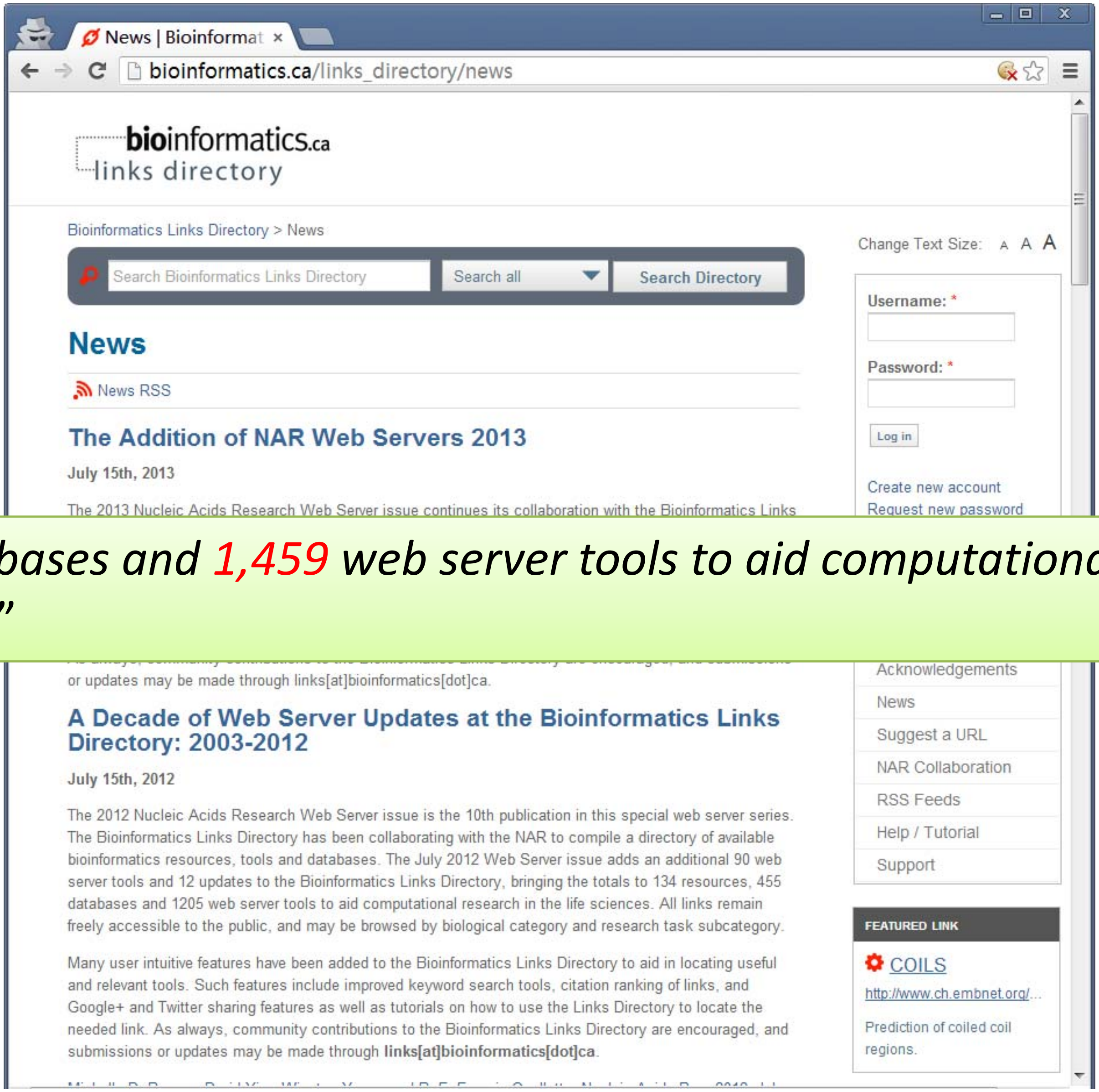
Opportunities and challenges hand-in-hand: the driving forces of bioinformatics

- High-throughput data
  - huge amount
  - explosive growth
  - noisy
  - multi-type
  - multi-scale
  - Heterogeneous



- Requirements for the methods
  - Data needs to be stored in efficient ontology-based database systems
  - The huge amount of data requires efficient methods
  - Exponential growth requires scalable methods
  - The low signal-to-noise ratio requires accurate methods
  - Multiple types of data requires data integrative methods





Prediction of coiled coil regions.

News

News RSS

The Addition of NAR Web Servers 2013

July 15th, 2013

The 2013 Nucleic Acids Research Web Server issue continues its collaboration with the Bioinformatics Links

As always, community contributions to the Bioinformatics Links Directory are encouraged, and submissions or updates may be made through links[at]bioinformatics[dot]ca.

A Decade of Web Server Updates at the Bioinformatics Links Directory: 2003-2012

July 15th, 2012

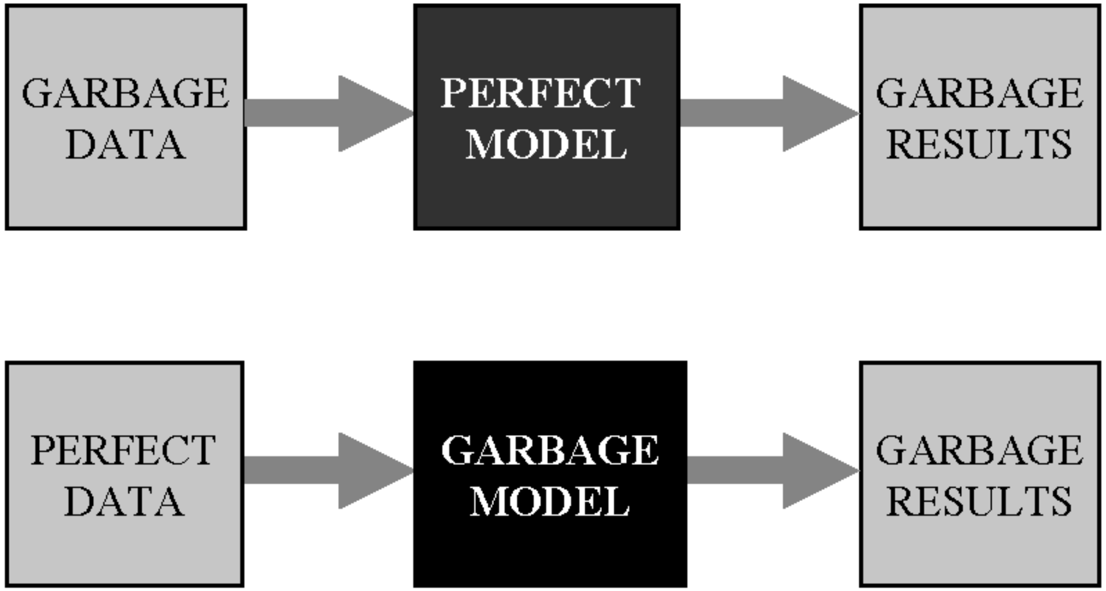
The 2012 Nucleic Acids Research Web Server issue is the 10th publication in this special web server series. The Bioinformatics Links Directory has been collaborating with the NAR to compile a directory of available bioinformatics resources, tools and databases. The July 2012 Web Server issue adds an additional 90 web server tools and 12 updates to the Bioinformatics Links Directory, bringing the totals to 134 resources, 455 databases and 1205 web server tools to aid computational research in the life sciences. All links remain freely accessible to the public, and may be browsed by biological category and research task subcategory.

Many user intuitive features have been added to the Bioinformatics Links Directory to aid in locating useful and relevant tools. Such features include improved keyword search tools, citation ranking of links, and Google+ and Twitter sharing features as well as tutorials on how to use the Links Directory to locate the needed link. As always, community contributions to the Bioinformatics Links Directory are encouraged, and submissions or updates may be made through **links[at]bioinformatics[dot]ca**.

"…620 databases and 1,459 web server tools to aid computational research in the life sciences"

MODEL CALCULATIONS

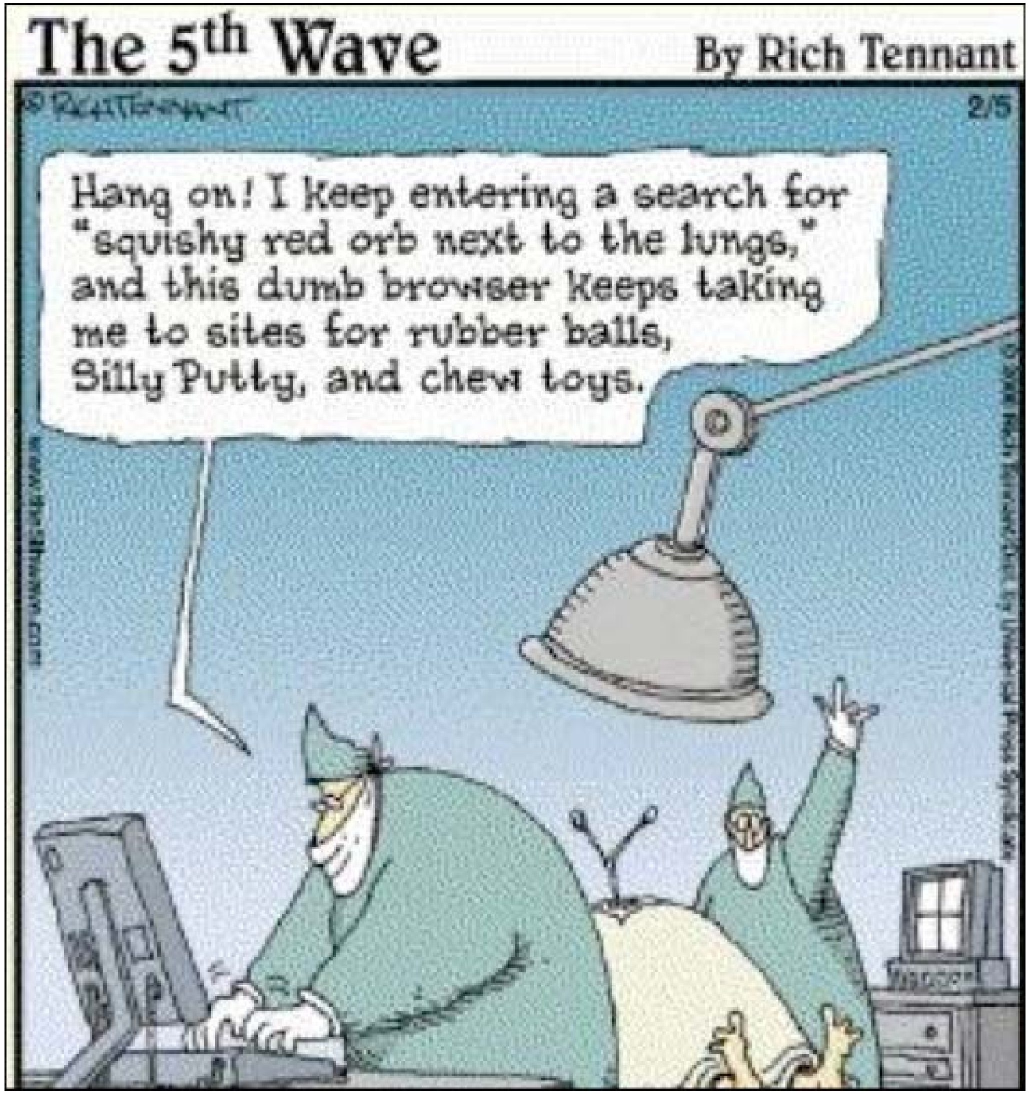
”Garbage In-garbage Out” Paradigm



GARBAGE DATA → PERFECT MODEL → GARBAGE RESULTS

PERFECT DATA → GARBAGE MODEL → GARBAGE RESULTS

(Source: http://blog.potterzot.com/wp-content/uploads/2007/09/garbage_paradigm.gif)


The 5th Wave
By Rich Tennant
2/5
Hang on! I keep entering a search for "squishy red orb next to the lungs," and this dumb browser keeps taking me to sites for rubber balls, Silly Putty, and chew toys.

(Source: http://performancemarketingassociation.com/new-working-group-data-feed-standard)

SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was o
a trajectory most young scientists only drean
about. In 1999, at the age of 28, the protei
crystallographer landed a faculty position a
the prestigious Scripps Research Institute i
San Diego, California. The next year, in a cer
emony at the White House, Chang received
Presidential Early Career Award
for Scientists and Engineers, the

Sour

Box 1

The good, the bad and the ugly

The good
In 1995, Fleischman *et al.* [34] were the first to succee
bacterium *Haemophilus influenzae* Rd. The group ident
represent genes. They translated the coding regions in
sequences in a protein database, identifying 1,007 clos
extensive annotation on the function of the entries, all
functions of most of the putative genes.

The bad
In 1997, the discovery of a new plant adenylyl cyclase
plants were not believed to have adenylyl cyclases. Th
for plants. The 'homology' (sequence similarity) they showed was not so weak: there was definitely some similarity, and the homology had a high 'score' (which by itself is not very meaningful) - but when their adenylyl cyclase was aligned to a profile for other known adenylyl cyclases, it was obvious to even first-year graduate students that the characteristics that are common to all other adenylyl cyclases were largely missing.

The ugly
The authors were later forced to retract their paper [36]. What might have saved them from public humiliation was a more careful analysis of their results.

Source: *Genome Biol 2*:reviews2002-review2002.10, 2001

(Source: http://cartoonmela.blogspot.com/2009_11_01_archive.html)

- **B**iology
  - What is the biological question or problem?
- **D**ata
  - What is the input data?
  - What other supportive data can be used?
- **M**odel
  - How is the problem formulated computationally?
  - Or, what’s the data model?
- **A**lgorithm
  - What is the computational algorithm?
  - How about its performance/limitation?

Sequence Alignment

Biological Question:

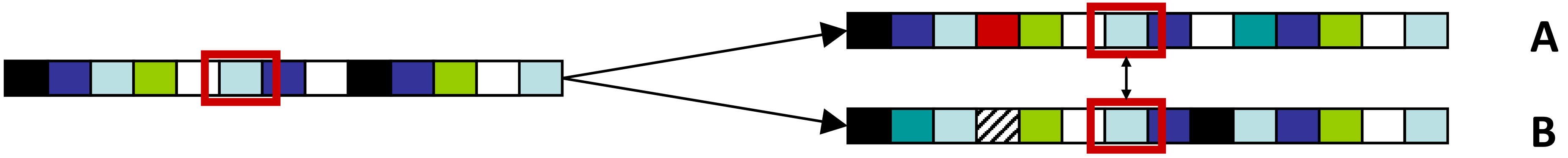
“How can we determine the similarity between two sequences?”

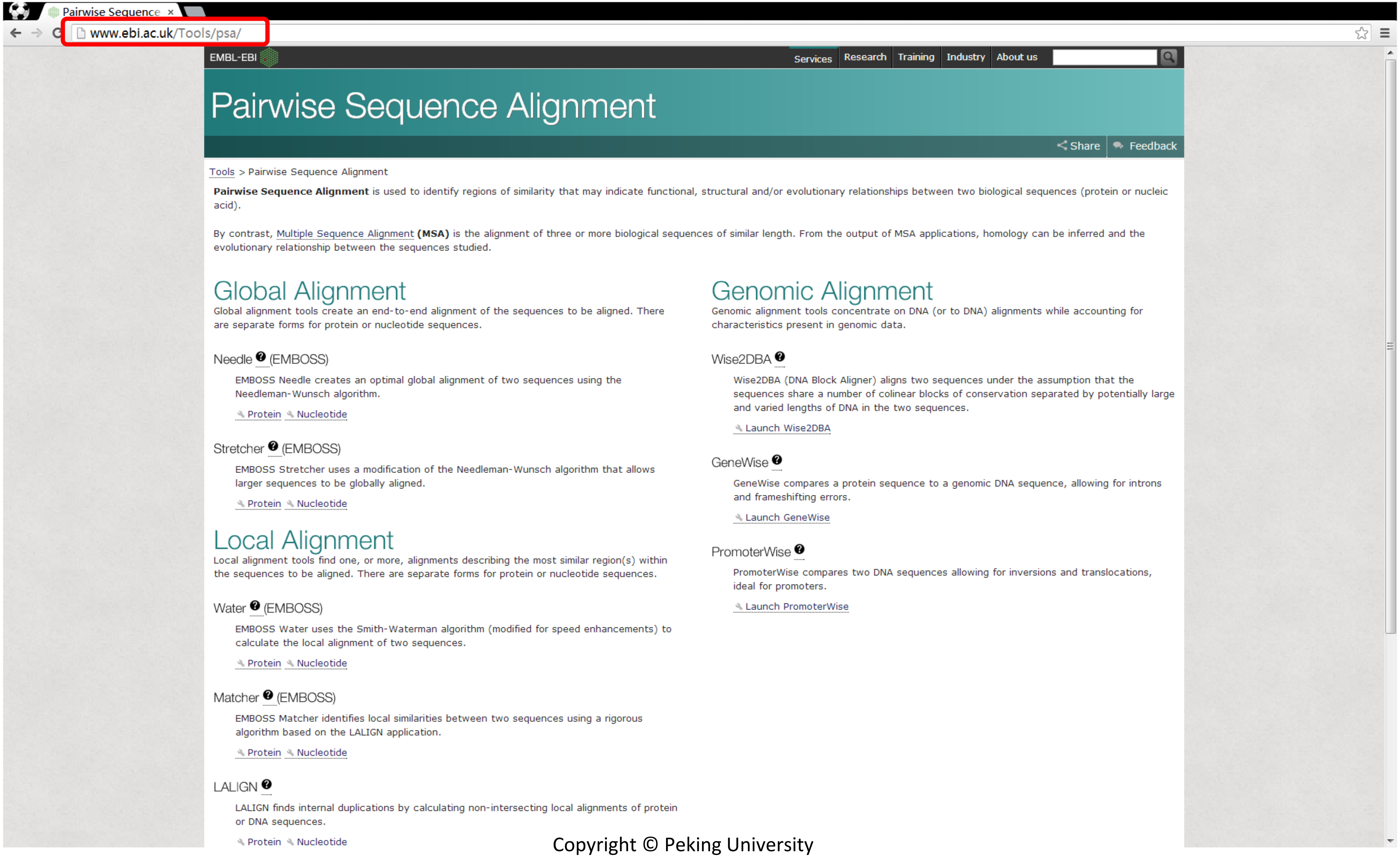
Why is it important?

- Similar sequence ➔ Similar structure ➔ Similar function (The “*Sequence-to-Structure-to-Function Paradigm*”)
- Similar sequence ➔ Common ancestor (“*Homology*”)

Sequence Alignment in Biology

The purpose of a sequence alignment is to line up all residues in the inputted sequence(s) for maximal level of similarity, in the sense of their functional or evolutionary relationship.





Pairwise Sequence Alignment

Tools > Pairwise Sequence Alignment

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, Multiple Sequence Alignment **(MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Needle (EMBOSS)

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

Protein Nucleotide

Stretcher (EMBOSS)

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

Protein Nucleotide

Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Water (EMBOSS)

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

Protein Nucleotide

Matcher (EMBOSS)

EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.

Protein Nucleotide

LALIGN

LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.

Protein Nucleotide

Genomic Alignment

Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.

Wise2DBA

Wise2DBA (DNA Block Aligner) aligns two sequences under the assumption that the sequences share a number of colinear blocks of conservation separated by potentially large and varied lengths of DNA in the two sequences.

Launch Wise2DBA

GeneWise

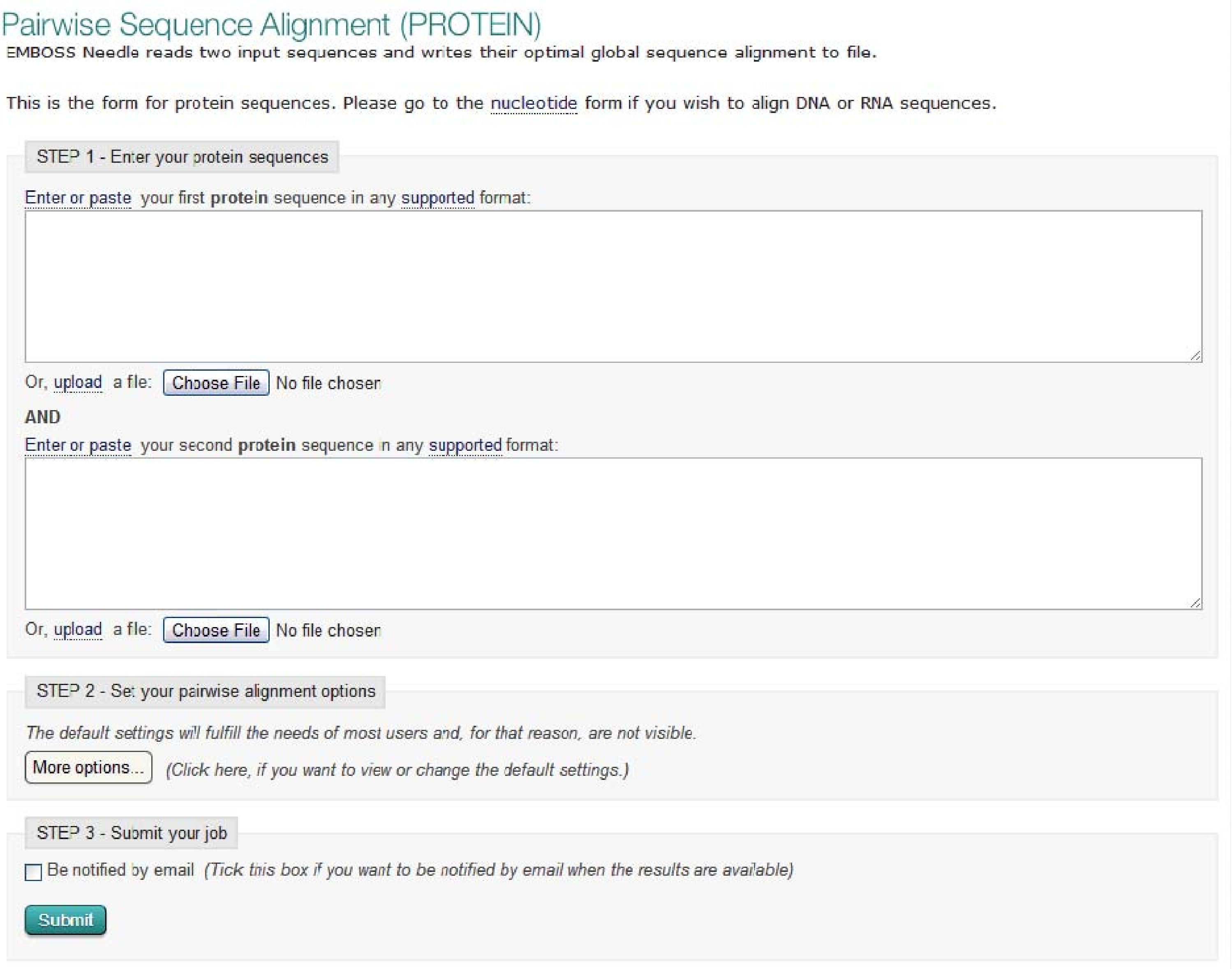
GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

Launch GeneWise

PromoterWise

PromoterWise compares two DNA sequences allowing for inversions and translocations, ideal for promoters.

Launch PromoterWise



Pairwise Sequence Alignment (PROTEIN)

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

Or, upload a file:

AND

Enter or paste your second **protein** sequence in any supported format:

Or, upload a file:

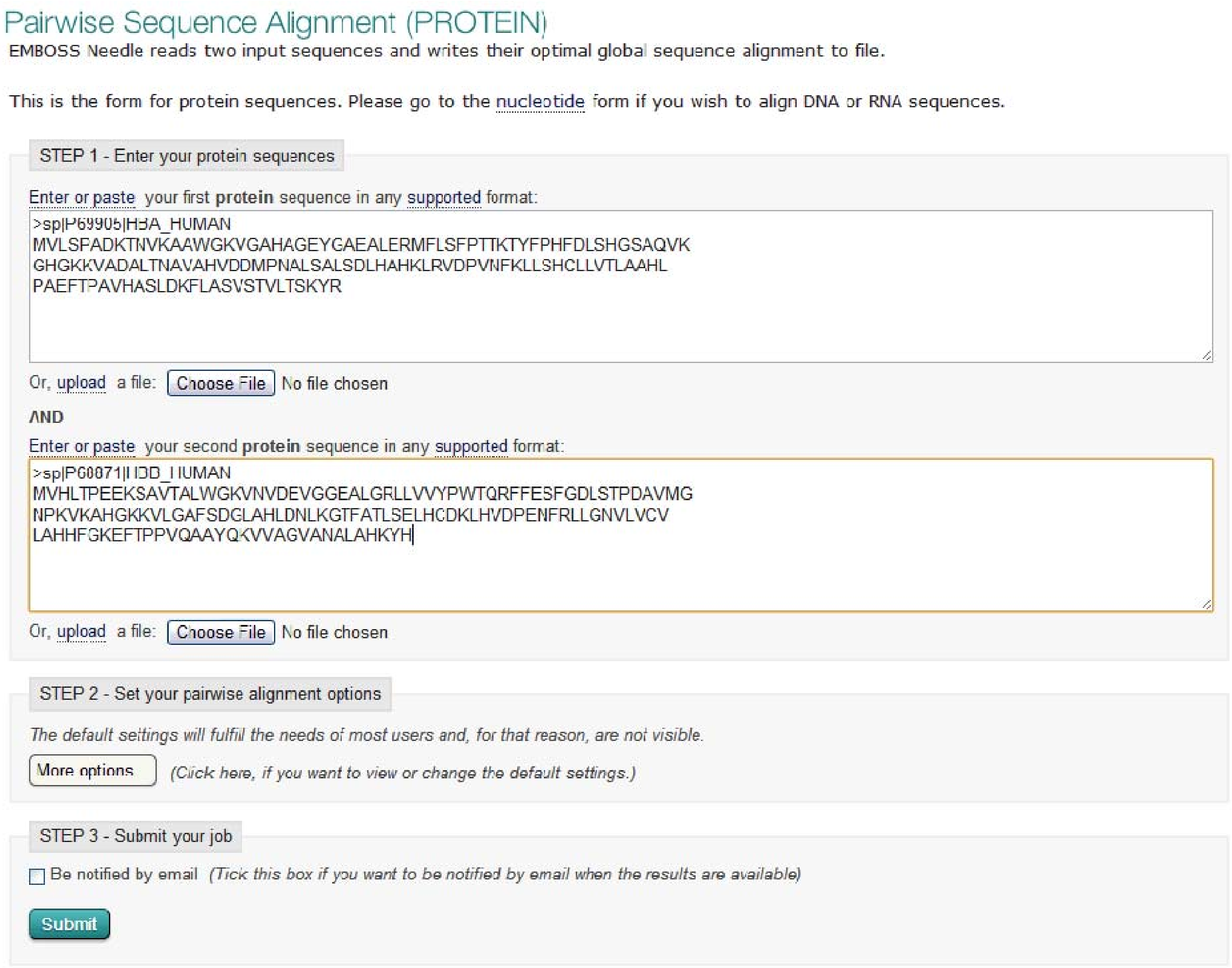
STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*



Pairwise Sequence Alignment (PROTEIN)

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

```
>sp|P69905|HBA_HUMAN
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVK
GHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
PAEFTPAVHASLDKFLASVSTVLTSKYR
```

AND

Enter or paste your second **protein** sequence in any supported format:

```
>sp|P68871|HBB_HUMAN
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMG
NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCV
LAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

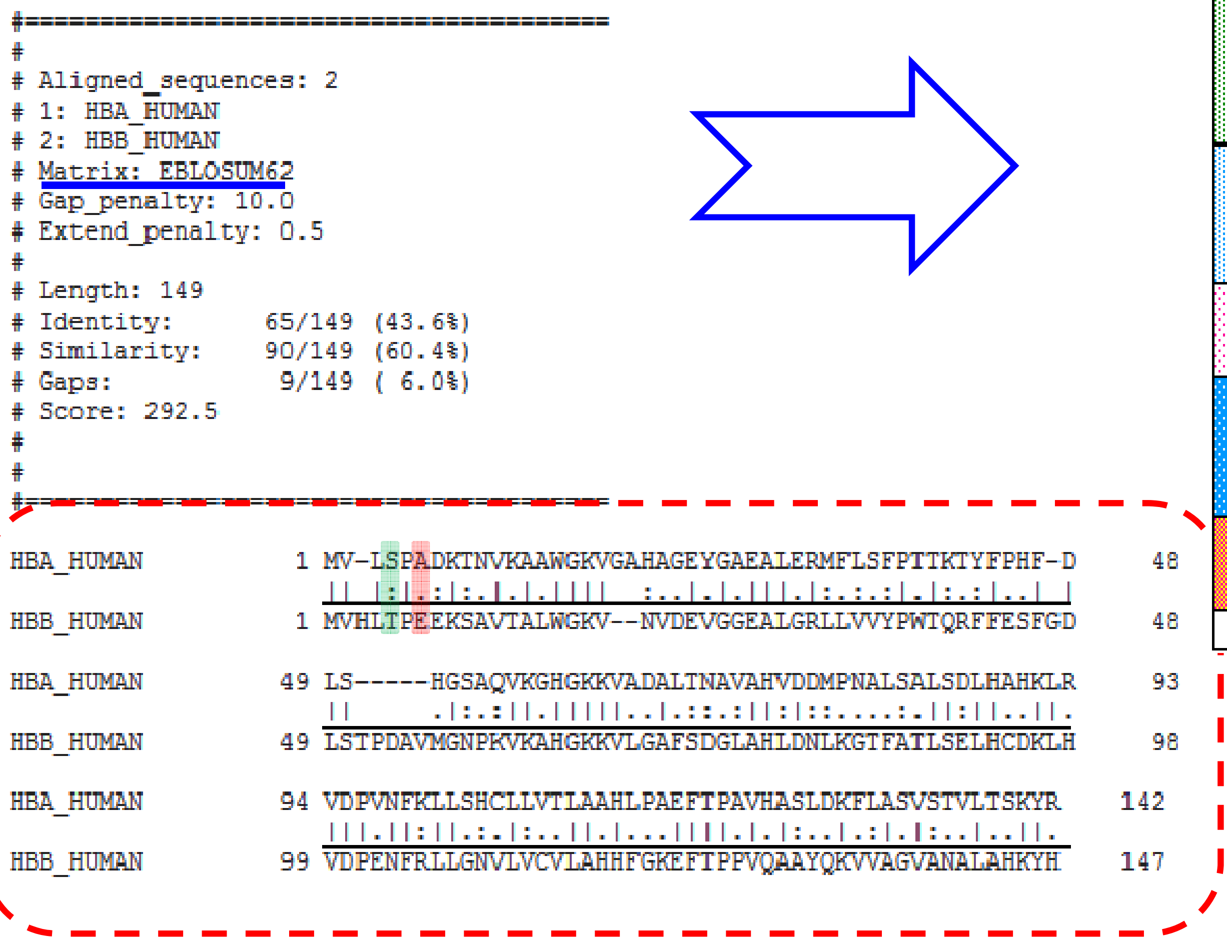
STEP 2 - Set your pairwise alignment options

The default settings will fulfil the needs of most users and, for that reason, are not visible.

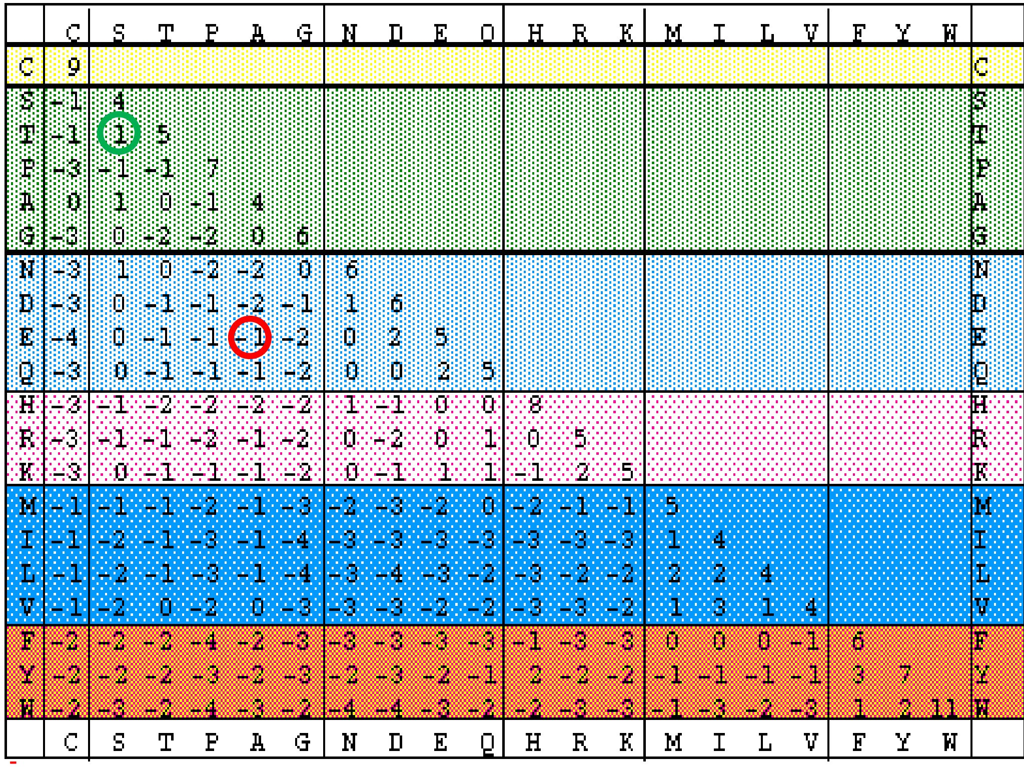
(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

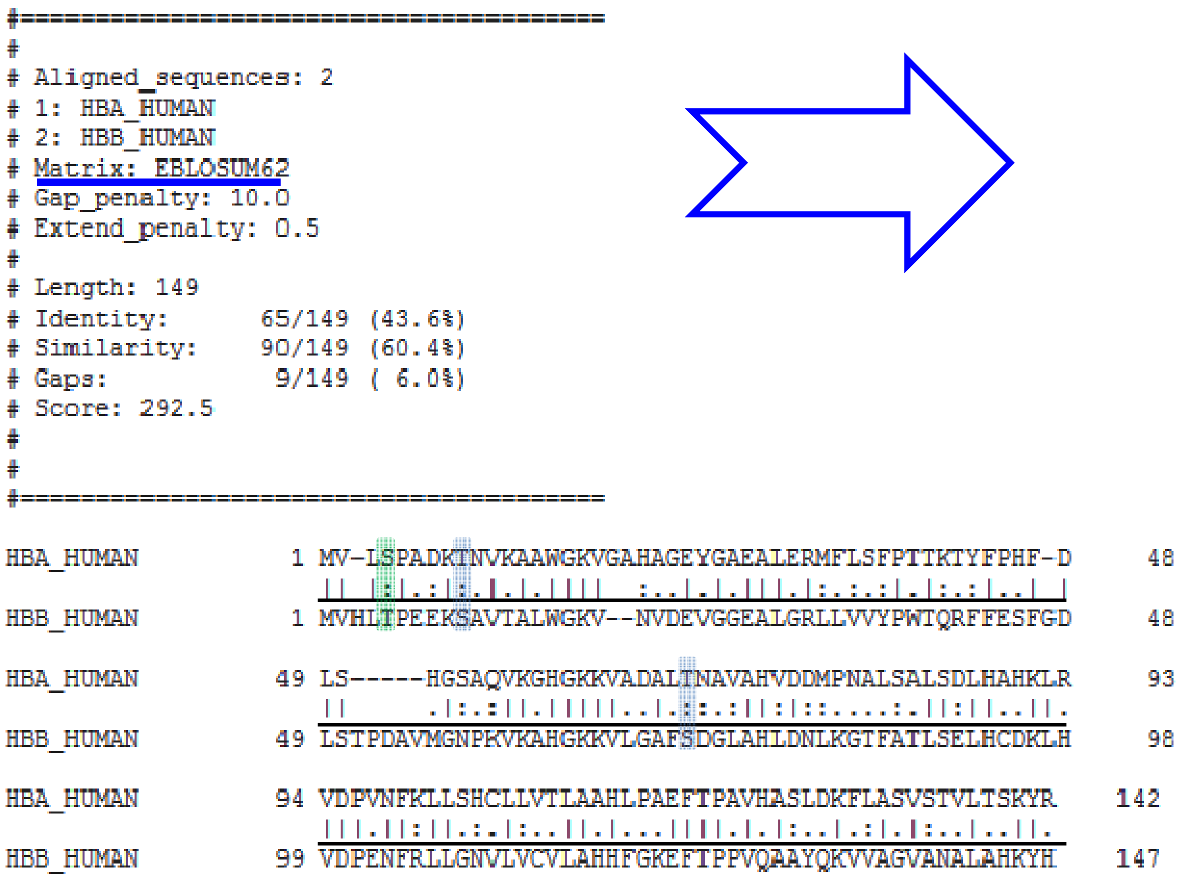
☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*



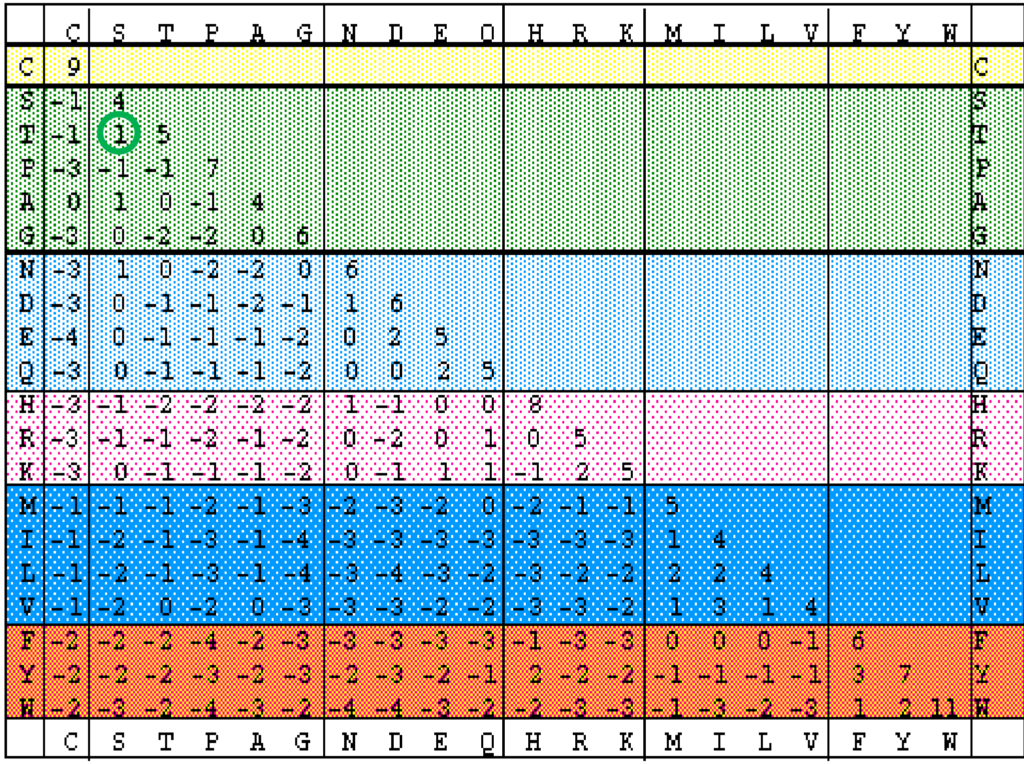
```
#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D      48
                       ||  |:||.:|:.|.|.||||   :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR      93
                       ||      .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH      98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR     142
                       |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH     147
```



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	



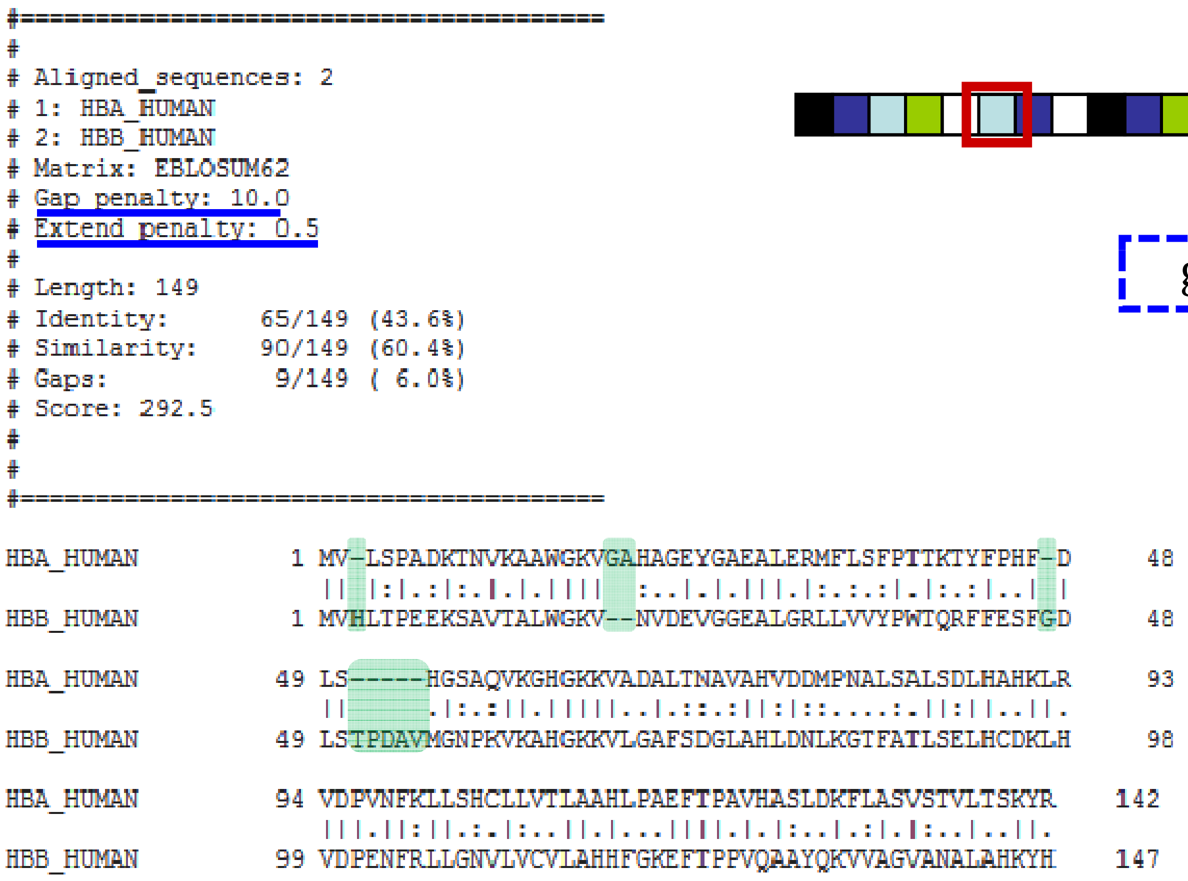
```
#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||   :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.:.:.:||:|::....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```



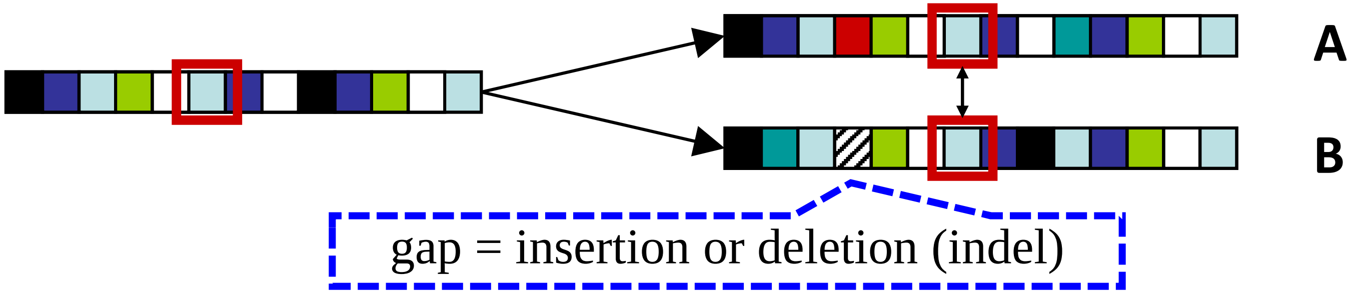
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

1. Symmetry

2. Context-insensitive



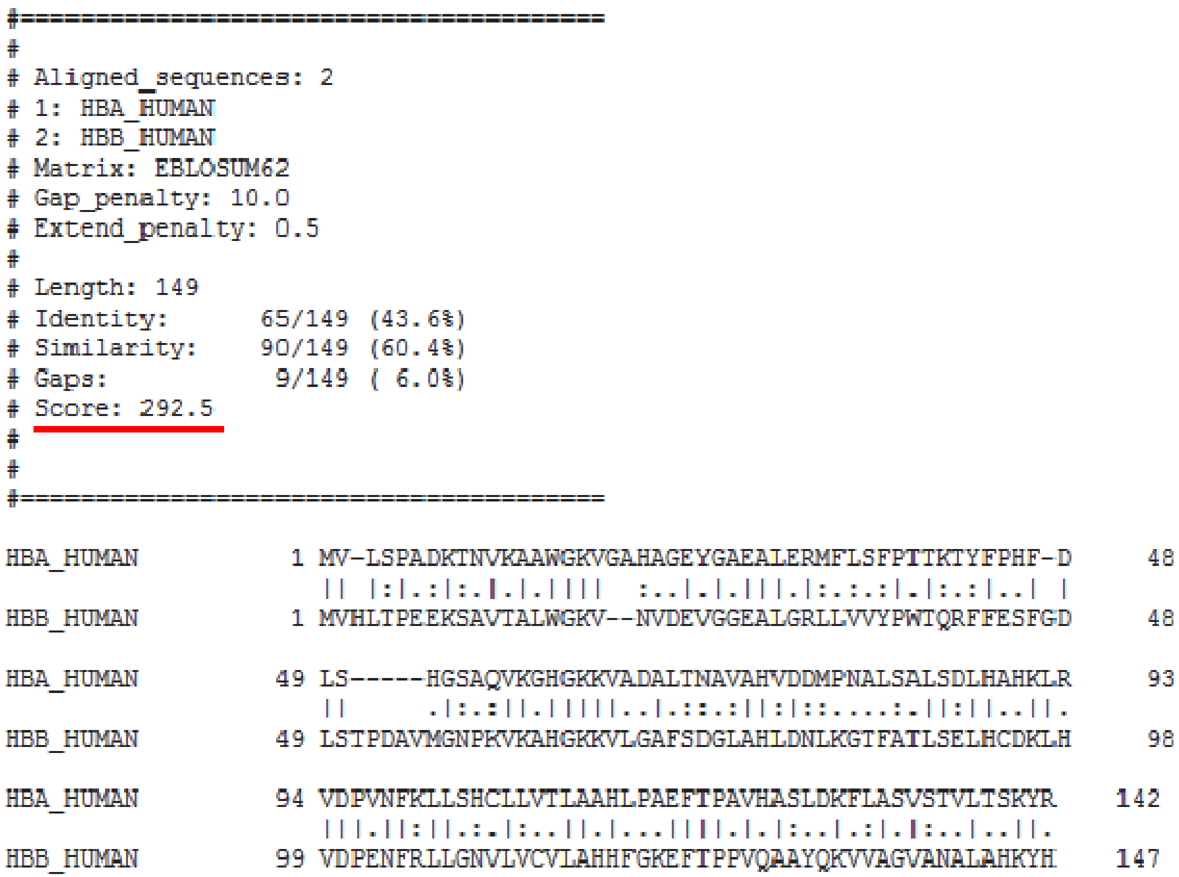
```
#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```



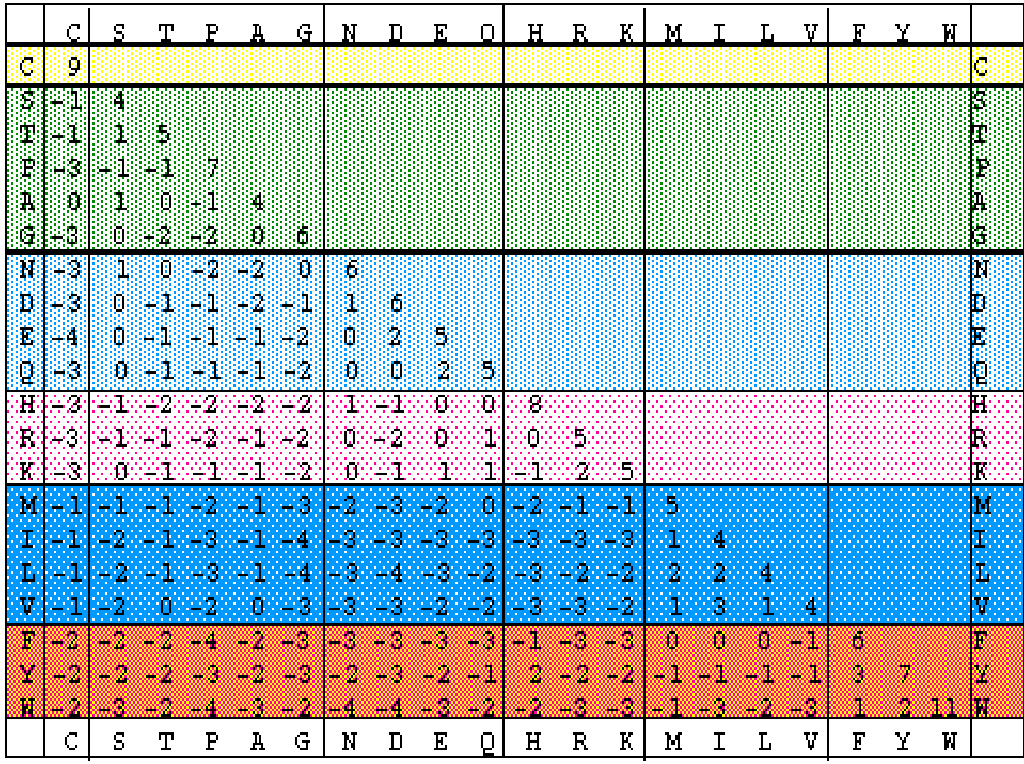
gap = insertion or deletion (indel)

Affine gap penalty: opening a gap receives a penalty of d; extending a gap receives a penalty of e. So the total Penalty for a gap with length n would be:

Penalty = d + (n-1)* e



```
#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Affine gap penalty: opening a gap receives a penalty of d; extending a gap receives a penalty of e. So the total Penalty for a gap with length n would be:

Penalty = d + (n-1)* e

Final Score = (sum of substitution scores) + (-1) * (sum of Gap Penalty)

Summary Questions

- Why do we do sequence alignment?

- How can we score a (pairwise) alignment?
  - (Why can we do so?)

生物信息学：导论与方法

Bioinformatics: Introduction and Methods

https://www.coursera.org/course/pkubioinfo