



# Βιοπληροφορική

## Ενότητα 10:

Αναζήτηση Ομοιοτήτων σε  
ΒΔ Ακολουθιών - Blast,  
(2/2) 1ΔΩ

Τμήμα: **Βιοτεχνολογίας**

Όνομα καθηγητή: **Τ. Θηραίου**



Ευρωπαϊκή Ένωση  
Ευρωπαϊκό Κοινωνικό Ταμείο



ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ  
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης





# Μαθησιακοί Στόχοι

- Αναφορά στις παραλλαγές του BLAST.
- Εξοικείωση με τη διεπαφή του BLAST.



# Λέξεις Κλειδιά

- Λέξεις κλειδιά: Βάσεις δεδομένων BLAST, Παραλλαγές BLAST, Ανταποδοτικό BLAST.
- Key words: BLAST databases, BLAST family of programs, PSI-BLAST, PHI-BLAST, Reciprocal BLAST, BLAST interface.



# Παραλλαγές του BLAST 11/12

- **PHI-BLAST:**

- **μοτίβο (pattern).**

- χαρακτηρίζει μια οικογένεια πρωτεϊνών.

- π.χ. [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].



# Παραλλαγές του BLAST 12/12

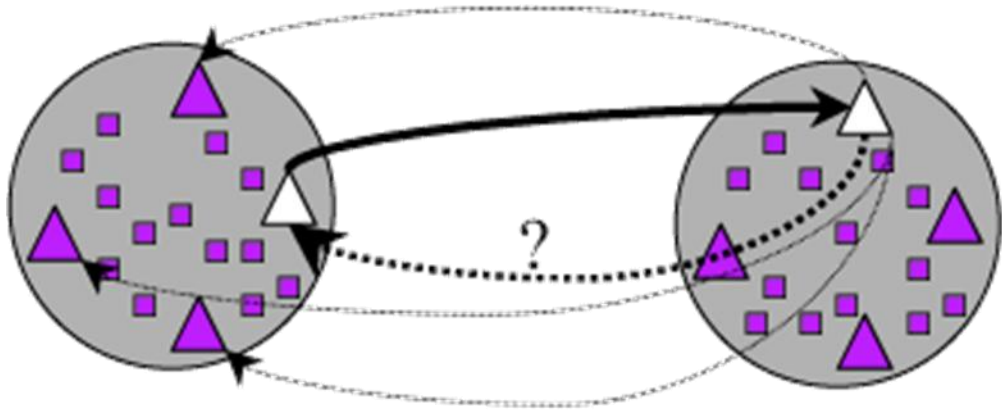
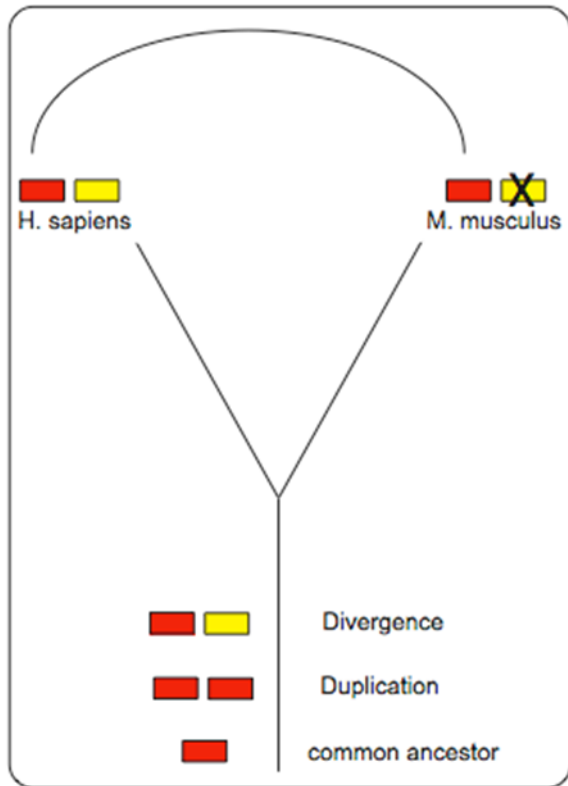
## ● PHI-BLAST:

- Δεδομένα εισόδου για την αναζήτηση:
  - ακολουθία επερώτησης.
  - μοτίβο (regular expression) που υπάρχει στην ακολουθία επερώτησης.
- Αναζήτηση ακολουθιών που **περιέχουν το μοτίβο** και έχουν **ομοιότητα** με την ακολουθία επερώτησης στη γειτονική περιοχή του μοτίβου.
- **Μείωση των hits** που **δεν έχουν πραγματική ομολογία** με την ακολουθία επερώτησης.



# Ανταποδοτικό BLAST 1/3

## ● Best Reciprocal (BLAST) Hit:





# Ανταποδοτικό BLAST 2/3

- Εύρεση **ορθόλογων** γονιδίων / πρωτεϊνών.
- Αναζήτηση BLAST με την ακολουθία α του οργανισμού A στις ακολουθίες του οργανισμού B.
  - καλύτερο hit η ακολουθία β.
- Αναζήτηση BLAST με την ακολουθία β του οργανισμού B στις ακολουθίες του οργανισμού A.
  - καλύτερο hit η ακολουθία α.
- Οι ακολουθίες α και β είναι ορθόλογες.



# Ανταποδοτικό BLAST 3/3

- Κρίσιμες παράμετροι:
  - είδος φίλτρου: soft filtering vs hard filtering.
  - αλγόριθμος τελικής στοίχισης: BLAST vs Smith-Waterman.
  - τιμές κατωφλίου: E-value ή bit-score, μήκος στοίχισης.
- Σφάλματα:
  - πρόσφατος εκτεταμένος γονιδιακός διπλασιασμός.
  - γονιδιακή σύντηξη.
  - domain rearrangements.





# Διεπαφή BLAST 1/6

## Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

## Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search using [SNP flanks](#)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)
- Search [RefSeqGene](#)



# Διεπαφή BLAST 2/6

[blastn](#) **[blastp](#)** [blastx](#) [tblastn](#) [tblastx](#)

BLASTP programs search protein databases using a protein query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

[Clear](#) **Query subrange** [?](#)

From

To

Or, upload file  [Browse...](#) [?](#)

**Job Title**

Enter a descriptive title for your BLAST search [?](#)

**Align two or more sequences** [?](#)

**Choose Search Set**

**Database**  [?](#)

**Organism** [Optional](#)

**Exclude** [+](#)

[?](#)

**Exclude** [Optional](#)

**Entrez Query** [Optional](#)

Enter an Entrez query to limit search [?](#)

**Program Selection**

**Algorithm**

**blastp** (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)



# Διεπαφή BLAST 3/6

**Algorithm parameters**

**General Parameters**

**Max target sequences**    
Select the maximum number of aligned sequences to display

**Short queries**  Automatically adjust parameters for short input sequences

**Expect threshold**

**Word size**

**Max matches in a query range**

**Scoring Parameters**

**Matrix**

**Gap Costs**

**Compositional adjustments**

**Filters and Masking**

**Filter**  Low complexity regions

**Mask**  Mask for lookup table only   
 Mask lower case letters

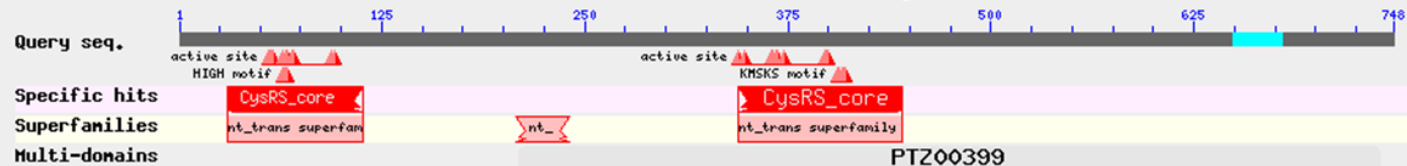
**BLAST** Search **database Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**   
 Show results in a new window



# Διεπαφή BLAST 4/6

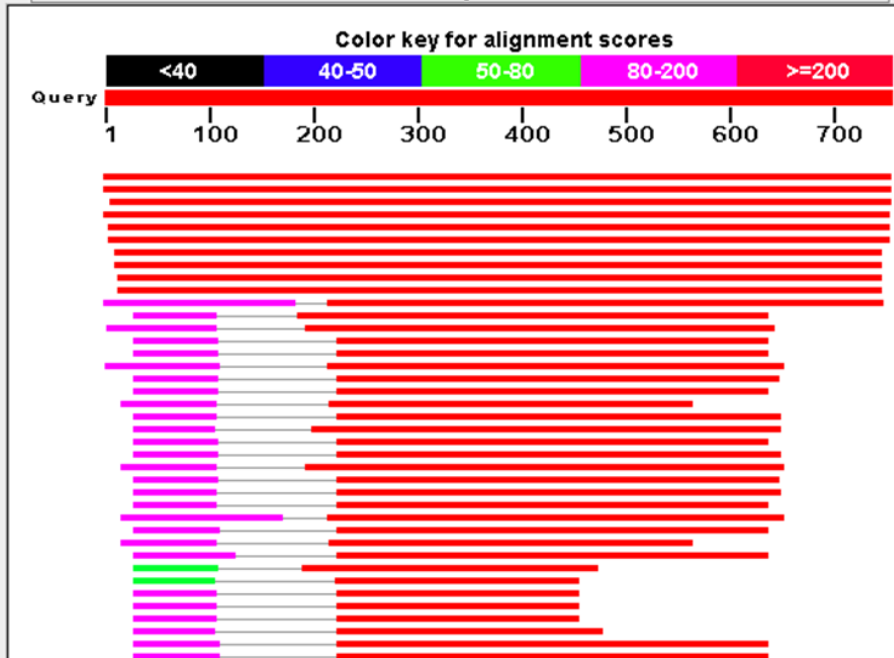
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 189 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments





# Διεπαφή BLAST 5/6

## Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

**NEW** - alignment score below the threshold on the previous iteration

- alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max

## Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">P49589.3</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1489</a>	1489	100%	0.0	100%	<a href="#">G</a> <a href="#">M</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q4R550.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1447</a>	1447	100%	0.0	97%	
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q9ER72.2</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1360</a>	1360	99%	0.0	90%	<a href="#">G</a> <a href="#">M</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q5F408.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1246</a>	1246	99%	0.0	82%	<a href="#">G</a> <a href="#">M</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q5M7N8.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1158</a>	1158	99%	0.0	77%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q7ZWR2.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">1153</a>	1153	99%	0.0	77%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q7KN90.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">821</a>	821	97%	0.0	59%	<a href="#">G</a> <a href="#">M</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q291L4.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">821</a>	821	97%	0.0	59%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">P53852.1</a>	RecName: Full=Cysteine--tRNA ligase; AltName: Full=CysteinyI-tRNA	<a href="#">589</a>	589	97%	0.0	45%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q09860.1</a>	RecName: Full=Probable cysteine--tRNA ligase; AltName: Full=Cysteir	<a href="#">561</a>	561	97%	0.0	44%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q54KR1.1</a>	RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=Cy	<a href="#">473</a>	580	94%	5e-155	47%	



# Διεπαφή BLAST 6/6

> [sp|Q54KR1.1|SYCC\\_DICDI](#) RecName: Full=Cysteine--tRNA ligase, cytoplasmic; AltName: Full=CysteinyI-tRNA synthetase; Short=CysRS  
Length=660

Sort alignments for this subject sequence by:

E value Score Percent identity

Query start position Subject start position

Score = 473 bits (1216), Expect = 5e-155, Method: Compositional matrix adjust.

Identities = 250/532 (47%), Positives = 350/532 (66%), Gaps = 17/532 (3%)

Query	213	SKLPKFWEGDFHRDMEALNVLPPDVLTRVSEYVPEIVNFVQKIVDNGYGYVSNNGSVYFDT	272
		S L K WE F DM+ LNVLPDP LTRV+EYVP+IV +V+KI+ NG+ Y SNGSVYFDT	
Sbjct	124	SDLKSKWETAFFEDMKLLNVLPPDALTRVTEYVVPQIVEYVEKIIISNGFAYESNGSVYFDT	183
Query	273	AKFASSEKHSYGKLVPEAVGDQKALQEGERDLSISADRLSEKRSFNDLWASKPGEPS	332
		F S+ H YGKL P +VG++K EGEK L+ ++ +SEKRS DFALWK SKPGEP	
Sbjct	184	VAF--SKAHDYGKLEPNSVGNELAAEGERGLTATS-AVSEKRSQDFALWKKSKPGEPV	240
Query	333	WPCPWGKGRPGWHIECSAMAGTLLGASMDIHGGGFDLRFPHHDNELAQSEAYFENDCNVR	392
		W PWG+GRPGWHIECSAMA LLG ++DIH GG DL+FPHHDNELAQSEA++ N W+	
Sbjct	241	WNSPWGEGRPGWHIECSAMASDLLGGNIDIHSGGSDLKFPHHDNELAQSEAFYGNRQWIN	300
Query	393	YFLHTGHLTIAGCKMSKSLKNFITIKDALKKHSARQLRLAFLMHSWKDTLDYSSNTMESA	452
		YF+H+GHL I G KMSKSLKNFITIK AL+K+++RQ+R+ F++H + ++YS +M A	
Sbjct	301	YFVHSGHLLIDGLKMSKSLKNFITIKQALEKYTSRQMRMFFILHKYDKAMNYSPEMGYA	360
Query	453	LQYEKFLNEFFLNVDILR-APVDITGQFEKUGEEAEELNKNFYDKKTAIHKALCDNVDT	511
		++ EK EFF K ILR +P+ + QF W + E +LNK+ + +H+ + DN +T	
Sbjct	361	IEMKTFVEFFHTAKQILRDSPLSLP-QF--WTQAEKDLNKHQLQANDQVHQFILDNFMT	417
Query	512	RTVMEEMRALVSQLNLYMAARKAVRKRPNQALLENIALYLTHMLKIFGAVEEDSS---LG	568
		++ + LV++ N+Y+ + + P L+ IA Y+T++ +FG E ++ +G	
Sbjct	418	SDALKTLSDLVNKTNVYIRSCAEQKTNPRNLISATAEYIYIFSVFGLTESSTASSMIG	477
Query	569	FPVGGPGTSLSEATVMPYLQVLSEFREGVRKIAREQKVPEILQLSDALRDNILPELGVR	628
		F G G ++E + P L L++FR VR A + IL+ D LRD +LP LGV+	
Sbjct	478	FGSAGKG---NIEEEMTPIALNALTQFRSEVRASAIKDDTTSILKTCMDLRDEVLPLLGVK	534
Query	629	FEDHEGLPTVVKLVDRNTLLkerekrrveeekrkkkeeaarrkqeqeaaakLAKMKIPPS	688
		+D + K D+ TL +++ ++E KKK+ K+++ K K KIPP	
Sbjct	535	IDDKSATTAMWKFEDKETL----KKEIEQKKEIEKKKQADKEEKEKLLKEKFEKSKIIPPQ	590
Query	689	EMFLSETDKYSKFDENGLPTHDMEGKELSKGQAKKLLKLFQAQEKLYKEYLQ	740
		++F++ETDKYSKF+E G+PTHG EG E++K Q KKL+K ++ Q K + YL+	
Sbjct	591	QLFINETDKYSKFNELGMPTHDKEGVEITKSQLKLLQKEYDNTKEHMNYLK	642



# Έλεγχος Αποτελεσμάτων

- Ομοιότητα σε επαρκές **μήκος** των ακολουθιών.
- Υψηλό ποσοστό **ταυτόσημων** καταλοίπων.
- Εμφάνιση χαρακτηριστικών **δομικών / λειτουργικών μοτίβων**.
- Ποιότητα δεδομένων στις βάσεις.
- Υποθετικά γονίδια.



# Βιβλιογραφία

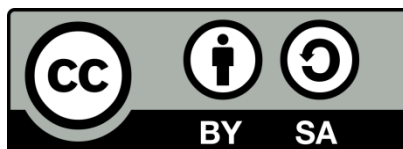
- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2<sup>nd</sup> edition (March 12, 2013).
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2<sup>nd</sup> edition (May 4, 2009).
- Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", Wiley-Interscience; 3<sup>rd</sup> edition (October 29, 2004).





# Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδεια χρήσης, η άδεια χρήσης αναφέρεται ρητώς.





# Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα Γεωπονικού Πανεπιστημίου Αθηνών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.



Ευρωπαϊκή Ένωση  
Ευρωπαϊκό Κοινωνικό Ταμείο



ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ  
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ  
*επένδυση στην κοινωνία της γνώσης*

ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ  
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΕΣΠΑ  
2007-2013  
πρόγραμμα για την ανάπτυξη  
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ



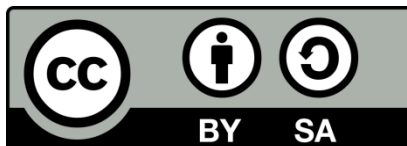
# Σημείωμα Αναφοράς

Copyright Γεωπονικό Πανεπιστήμιο Αθηνών 2015. Τμήμα Βιοτεχνολογίας, Θηραίου Τριάς. «Βιοπληροφορική». Έκδοση: 1.0. Αθήνα 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:  
<https://mediasrv.aua.gr/eclass/courses/OCDB100/>



# Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Παρόμοια Διανομή 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων, π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».



Η άδεια αυτή ανήκει στις άδειες που ακολουθούν τις προδιαγραφές του Ορισμού Ανοικτής Γνώσης [2], είναι ανοικτό πολιτιστικό έργο [3] και για το λόγο αυτό αποτελεί ανοικτό περιεχόμενο [4].

[1] <http://creativecommons.org/licenses/by-sa/4.0/>

[2] <http://opendefinition.org/okd/ellinika/>

[3] <http://freedomdefined.org/Definition/EI>

[4] <http://opendefinition.org/buttons/>



# Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
  - το Σημείωμα Αδειοδότησης
  - τη δήλωση Διατήρησης Σημειωμάτων
  - το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)
- μαζί με τους συνοδευόμενους υπερσυνδέσμους.